DEPARTMENT OF MATHEMATICS

MASTER THESIS

STATISTICAL SCIENCE FOR THE LIFE AND BEHAVIOURAL SCIENCES

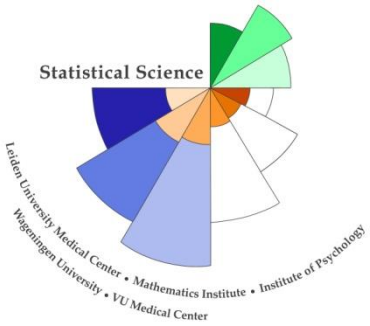
Multiple imputation with chained equations and survival outcomes

A simulation study

Monique van der Kruijk

July 2015

Universiteit Leiden



Thesis advisors:
Prof. Dr. H. Putter (LUMC)
Prof. Dr. S. van Buuren (TNO, UU)

Supervisor:
Dr. M. Fiocco (MI, LUMC)

Abstract

Multiple imputation is a commonly used technique to impute missing covariate values for incomplete data. Imputation of the incomplete covariates may be performed by using a regression model on both the other covariates and the outcome. With a survival outcome, several compositions of the status indicator δ and the observed event or censoring time T included in the imputation model can be found in the literature. In this study we evaluate different manners of inclusion of the survival outcome in the imputation model, using simulation studies. The different procedures are compared in case of an ordinary single survival outcome and in the context of competing risks. Furthermore, the different procedures are applied to a dataset provided by the European Society for Blood and Marrow Transplantation, which we further analyze after imputing the missing covariate values with the best found procedure.

Symbol description

X	survival time
C	censoring time
T	observed event or censoring time, $\min(X, C)$
δ	status indicator, $1\{X \leq C\}$
Y	n x p matrix of partially observed sample data
R	n x p matrix, binary response indicator of Y (0=unobserved, 1=observed)
Y_{obs}	observed sample data, (R = 1)
Y_{mis}	unobserved sample data, (R = 0)
Z	n x q matrix of observed covariates
Z_{inc}	incomplete covariate
Q	k x 1 vector with k parameters
U	k x k matrix, within-imputation variance
B	k x k matrix, between-imputation variance
V	k x k matrix, total variance obtained by combining U and B

Contents

1. Introduction

Researchers in different fields often face the problem of incompletely collected data. In prognostic research, completely collected data without missing covariate values is rare. Whether missing information is due to refusal of respondents in social sciences, attrition of patients in medical studies or other unfortunate reasons, gathering incomplete data seems to be inevitable [1]. Although the causes of missing values may differ from case to case, the difficulties that come as a result of them are similar. Naturally, collecting data with missing values is not intended and analyzing them is far from optimal. Given the expense of collecting data, starting over again is practically not feasible. Therefore, handling missing data in a proper manner is crucial to the analysis.

A common approach to deal with incomplete data is based on multiple imputation of the missing covariate values. Multiple Imputation has the ability to incorporate all sources of variability and uncertainty. By combining complete data inferences, Multiple Imputation is able to make valid inferences for the incomplete data [2]. In prognostic research, Multivariate Imputation with Chained Equations (MICE) is currently the golden standard. MICE is a special Multiple Imputation technique that handles multivariate missing data in a clever and flexible manner. In this iterative approach, each of the incomplete covariates is imputed based on a unique regression model specified by the user and using the other covariates as predictors [1]. In case the incomplete data are covariates in the analysis model, it is essential to include the analysis model outcome as predictor in the regression model as well [3].

As in many fields of research, missing covariate values may also occur in survival data. Data from the European Society for Blood and Marrow Transplantation (EBMT) database serves as an example of survival data, where in about half of the cases information on one or more covariates is missing. Multiple Imputation can be used for making valid inferences on this incomplete survival data. Specifying an imputation model for the survival data, implies including both the covariates that have any association with the incomplete data assessed in the analysis model [4] and the survival outcome [3]. The survival outcome must be correctly included in the imputation model, as the association between the incomplete covariate and survival may become weakened otherwise [5]. The question then becomes: how should the survival outcome be used in the regression model for imputation of missing covariate values? In literature, the survival outcome is included in several manners, varying from including the status indicator δ, observed event or censoring time T and the log of T as predictors in the imputation model [6], including only δ and the log of T [7], just δ and T [8] or the most recent suggestion of including δ and the cumulative baseline hazard $H_0(T)$ [5]. The aim of this study is to evaluate different manners of inclusion of the survival outcome in the imputation model by means of a simulation study. Afterwards an application to the motivating data from EBMT is conducted, to see how the different methods compare in a given data set.

In Chapter 2, the motivating dataset from EBMT is presented and analyzed for better insight in the process behind the missing covariate values. The basic concepts of missing values and the underlying theory are described in Chapter 3. In Chapter 4 a brief introduction of Multiple Imputation and Multivariate Imputation with Chained Equations is given. Different manners in which the survival outcome can be used for imputation of missing covariate values are described in Chapter 5. In Chapter 6, the design of the simulation study is described along with the results of the simulation. An extension of the study in case of competing risks is described in Chapter 7, where another simulation

is conducted and the results are presented. An application to the EBMT dataset is conducted and the results are reported in Chapter 8. Final results and conclusions on the different methods are discussed in Chapter 9. All R-code written for this thesis is provided separately.

2. EBMT data

Recently a retrospective multicenter study was performed on the cytogenetic classification for patients with myelodysplastic syndromes (MDS) or secondary acute myeloid leukemia (AML) evolving MDS (sMDS). Data was collected for patients who were reported to the European Society for Blood and Marrow Transplantation (EBMT) database. This recent study evaluated the impact of the revised International Prognostic Scoring System (IPSS-R) 5-group cytogenetic classification at time of transplantation for the outcome of the patients after allogeneic stem cell transplantation. The study consists of patients for which sufficient cytogenetic information was available and who received transplantation as treatment between 1982 and 2010, resulting in a total of 903 patients. After analysis, the study has shown that the recent 5-group cytogenetic IPSS-R classification at time of transplantation has added value in the prediction of patient outcome after transplantation compared to prediction models with only traditional risk factors or the older 3-group IPSS cytogenetic classification. It was also shown in the study that a model with a simplified 5-group IPSS-R cytogenetic classification (where the groups of very good-, good- and intermediate risk patients were merged) performed best for predicting outcomes after transplantation [9].

Using the EBMT data, a prognostic model will be built, where the simplified 5-group IPSS-R cytogenetic classification (*cytogenetics*) is included as a covariate. Several other prognostic variables were available, from which we selected the stage of the disease (*staging*), the calendar year of transplantation (*year*) and the patients age at transplantation (*age*), as these covariates were found to be good predictors for relapse-free survival [9]. For about half of the patients in the EBMT data information is incomplete; in total, information is missing on either *staging*, *cytogenetics* or both in 694 (51%) patients and all information is available in 660 (49%) patients. Question of interest is how to correctly handle the missing data, in order to make valid inference. Before we start reasoning on this question, we must first understand the process behind the missing data and explore whether the data is missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR) (see Chapter 3 for details on this subject).

Descriptives

Information is available for 1354 patients who all received transplantation for the treatment between 1981 and 2010. The age of the patients ranges from 18 – 74 years at time of transplantation, with a median of 50 years. At time of transplantation, 339 (25%) patients had untreated RA/RARS/del5q/RCDM-RS, 219 (16%) patients had RAEB(t)/tAL/sMDS/CMML in CR, 74 (5%) patients had untreated RAEB(t)/tAL/sMDS/CMML, 303 (22%) patients had RAEB(t)/tAL/sMDS/CMML not in CR and for 419 (31%) patients information concerning the stage of the disease was not available. Within 1 year before transplantation 662 (49%) of the patients were classified as very good-, good- or intermediate risk patients, 178 (13%) patients were classified as poor risk patients, 64 (5%) patients were classified as very poor risk patients and for 450 (33%) this information was not available. An overview of these disease and patients' characteristics from the EBMT database are given in table 2.1.

Table 2.1 Disease and patients' characteristics from the EBMT database

	No. (%)	Median *(range)*
Cytogenetics		
Very good/ normal/ intermediate	662 *(48.9)*	
Poor	178 *(13.1)*	
Very poor	64 *(4.7)*	
Missing	450 *(33.3)*	
Staging		
RA/RARS/del5q/RC DM-RS untreated	339 *(25.0)*	
RAEB(t)/tAL/sMDS/CM ML in CR	219 *(16.2)*	
RAEB(t)/tAL/sMDS/CM ML untreated	74 *(5.5)*	
RAEB(t)/tAL/sMDS/CM ML not in CR	303 *(22.4)*	
Missing	419 *(30.9)*	
Age	1354 *(100)*	50 *(18 - 74)*
Year	1354 *(100)*	2002 *(1981 – 2010)*

Exploring the process behind the missing information

For exploring the process behind the missing information, we start by comparing patients with all information available (complete cases) with patients from whom information is missing on either *staging*, *cytogenetics* or both (incomplete cases). We mainly focus on relapse-free-survival after transplantation, which is defined as the time from transplantation to death or relapse (whichever occurred first), censoring patients that have been reported alive and disease-free at the last time-point [9]. In figure 2.1 a Kaplan-Meier plot is shown that presents the relapse-free survival curves of both the complete case patients and the incomplete case patients. From the plot it is shown that the two groups have similar survival curves. A log-rank test indicates the complete case patients and the incomplete case patients do not significantly differ in their probability of relapse-free survival (p-value = 0.269).

Several tests are performed to see whether a relation between the missingness of *cytogenetics* or *staging* with (one of) the covariates exists. Figure 2.2 displays histograms presenting the distribution of *age* for patients for which the stage of the disease is non-missing and missing (top left and right, respectively) and also for patients for which the cytogenetic classification is non-missing and missing (bottom left and right, respectively). The age of the patients seems approximately normal distributed for all four groups. In order to test if the missingness across *staging* or *cytogenetics* is related to the age of the patients, t-tests were performed. The tests indicated that both missingness across *staging* or *cytogenetics* is unrelated to the age of the patients, as the groups with missing information do not significantly differ from the groups without missing information with respect to the age of the patients (p-value=0.189 for *staging* and p-value=0.396 for *cytogenetics*).

The covariate *year* is continuous as well, but unlike *age*, the year of transplantation is quite skewed to the right and therefore not normally distributed. To test if missingness across *staging* or *cytogenetics* is related to the year of transplantation, the covariate is split into three approximately equally sized groups and an overview is presented in tables 2.2 and 2.3 respectively. According to the performed χ^2-tests, the year of transplantation significantly relates to the missingness in *staging* (p-value < 0.001) and it also significantly relates to the missingness in *cytogenetics* (p-value < 0.001). These significant relations might indicate missing information in both *staging* and *cytogenetics* is not MCAR, as it is significantly related to the year of transplantation.

Furthermore, table 2.2 presents the relations between the cytogenetic classification and patients for which the stage of the disease is known (non-missing) or unknown (missing) and the relations between the stage of the disease and patients for which the cytogenetic classification is known (non-missing) or unknown (missing). We see from the table that the cytogenetic classification is unknown relatively more often for patients for which the stage of the disease is unknown (41.8%), than for patients for which the stage of the disease is known (29.4%). The other way around shows a similar result; the stage of the disease is unknown relatively more often for patients for which the cytogenetic classification is unknown (38.9%), than for patients for which the cytogenetic classification is known (27.0%). The χ^2-tests indicate that cytogenetic classification significantly relates to the missingness in *staging* (p-value $<$ 0.001) and that the stage of the disease is significantly related to the missingness in *cytogenetics* (p-value $<$ 0.001). From these results we see missing information in both *staging* and *cytogenetics* is not only significantly related to the year of transplantation, but also to each other.

To see how the missing information relates to the survival, two Kaplan-Meier plots are generated. In figure 2.3 the relapse-free survival curves are presented in a Kaplan-Meier plot for cytogenetic classification. The survival curves are presented for patients with very good/normal/intermediate cytogenetic classification (0; red line), patients with poor cytogenetic classification (1; yellow line), patients with very poor cytogenetic classification (2; blue line) and patients for which the cytogenetic classification is unknown (M; purple line). From the plot we see very good/normal/intermediate risk patients have a better chance of surviving than poor risk or very poor risk patients. Also, poor risk patients have better chance of surviving than very poor risk patients. The relapse-free survival curve of patients for which the cytogenetic classification is unknown, seems to be right in between the very good/normal/intermediate risk patients and the poor risk patients and, therefore, does not stand out with respect to the other survival curves.

Another Kaplan-Meier plot is generated to explore the relapse-free survival curves for the different types of staging and is shown below in figure 2.4. The curves are presented for patients with RA/RARS/del5q/RCDM-RS untreated (0; red line), patients with RAEB(t)/tAL/sMDS/CMML in CR (1; yellow line), patients with RAEB(t)/tAL/sMDS/CMML untreated (2; green line), patients with RAEB(t)/tAL/sMDS/CMML not in CR (3; blue line) and patients for which the stage of the disease is unknown (M; purple line). From the plot we see the probability of relapse-free survival decreases in this order as well, except for the patients for which the stage of the disease is unknown. The survival curve of the group with an unknown stage of the disease seems to be rather close to the survival curve of the patients with RAEB(t)/tAL/sMDS/CM ML untreated and does not stand out with respect to the other survival curves.

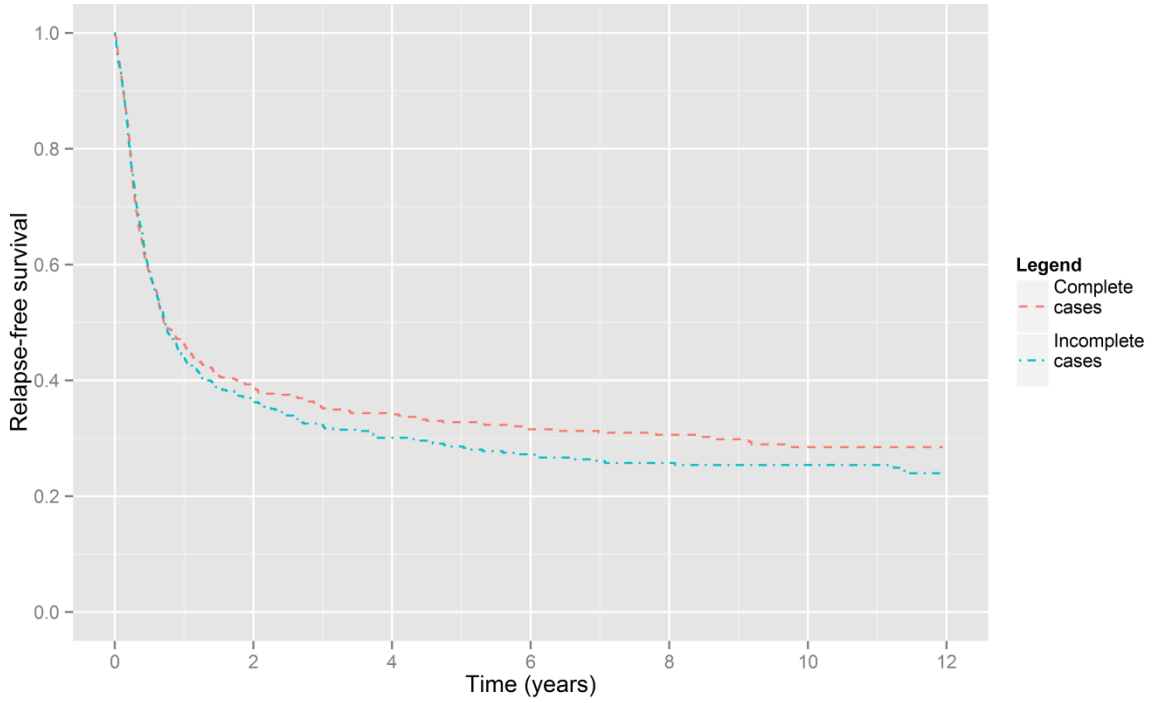


Figure 2.1 Kaplan-Meier curve of the survival times of the complete- and incomplete case individuals

Table 2.2 Relation between missingness in *staging* and *cytogenetics* and *year*

	Staging		
	Non-missing *(%)*	Missing *(%)*	P-value
Cytogenetics			< 0.001
Very good/ normal/ intermediate	479 *(51.2)*	183 *(43.7)*	
Poor	130 *(13.9)*	48 *(11.4)*	
Very poor	51 *(5.5)*	13 *(3.1)*	
Missing	275 *(29.4)*	175 *(41.8)*	
Year			< 0.001
1981-1999	352 *(37.7)*	125 *(29.8)*	
2000-2006	361 (38.6)	90 (21.5)	
2007-2010	222 (23.7)	204 (48.7)	

Table 2.3 Relation between missingness in *cytogenetics* and *staging* and *year*

	Cytogenetics		
	Non-missing *(%)*	Missing *(%)*	P-value
Staging			< 0.001
RA/RARS/del5q/RC DM-RS untreated	235 *(26.0)*	104 *(32.1)*	
RAEB(t)/tAL/sMDS/CM ML in CR	181 *(20.0)*	38 *(8.5)*	
RAEB(t)/tAL/sMDS/CM ML untreated	54 *(6.0)*	20 *(4.4)*	
RAEB(t)/tAL/sMDS/CM ML not in CR	190 *(21.0)*	113 *(25.1)*	
Missing	244 *(27.0)*	175 *(38.9)*	
Year			< 0.001
1981-1999	272 *(30.1)*	205 *(45.6)*	
2000-2006	324 *(35.8)*	127 *(28.2)*	
2007-2010	308 *(34.1)*	118 *(26.2)*	

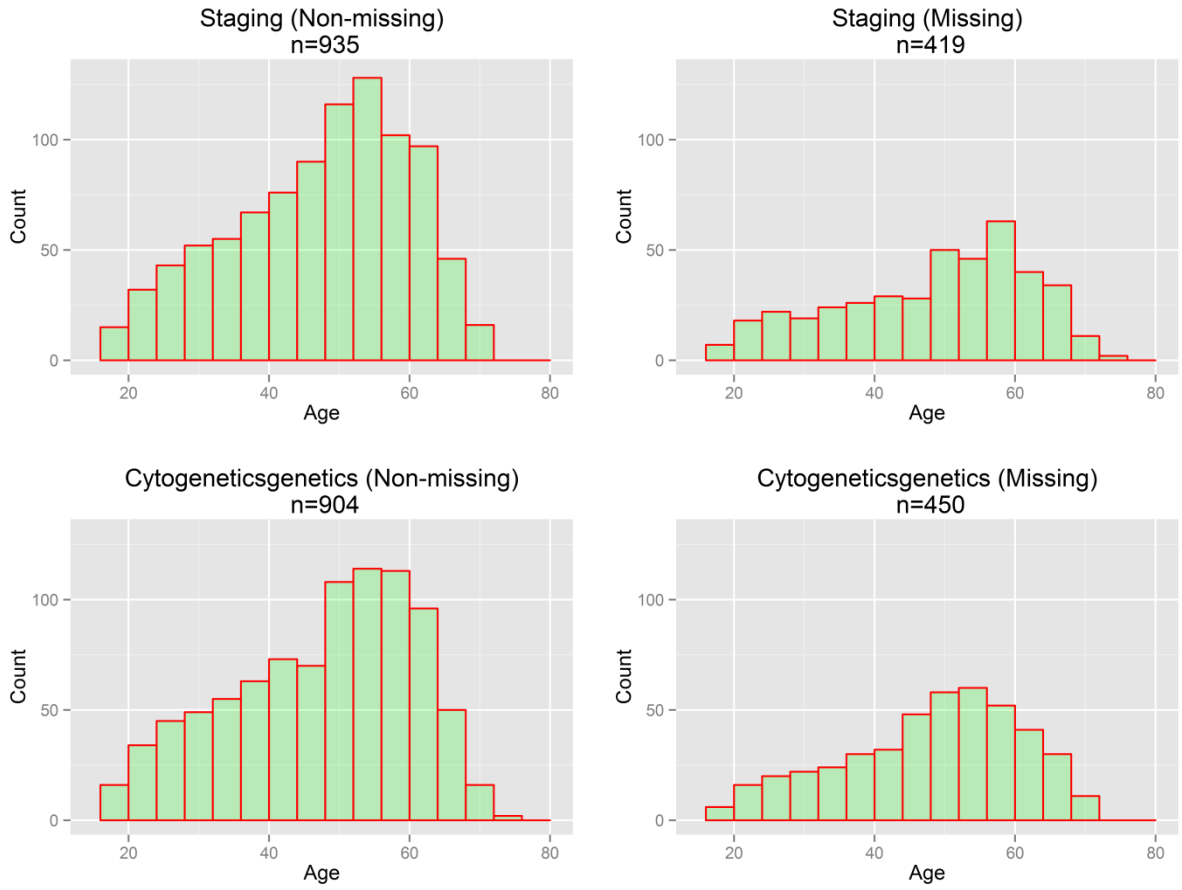


Figure 2.2 Histograms indicating the distribution of *age* for *staging* and *cytogenetics*

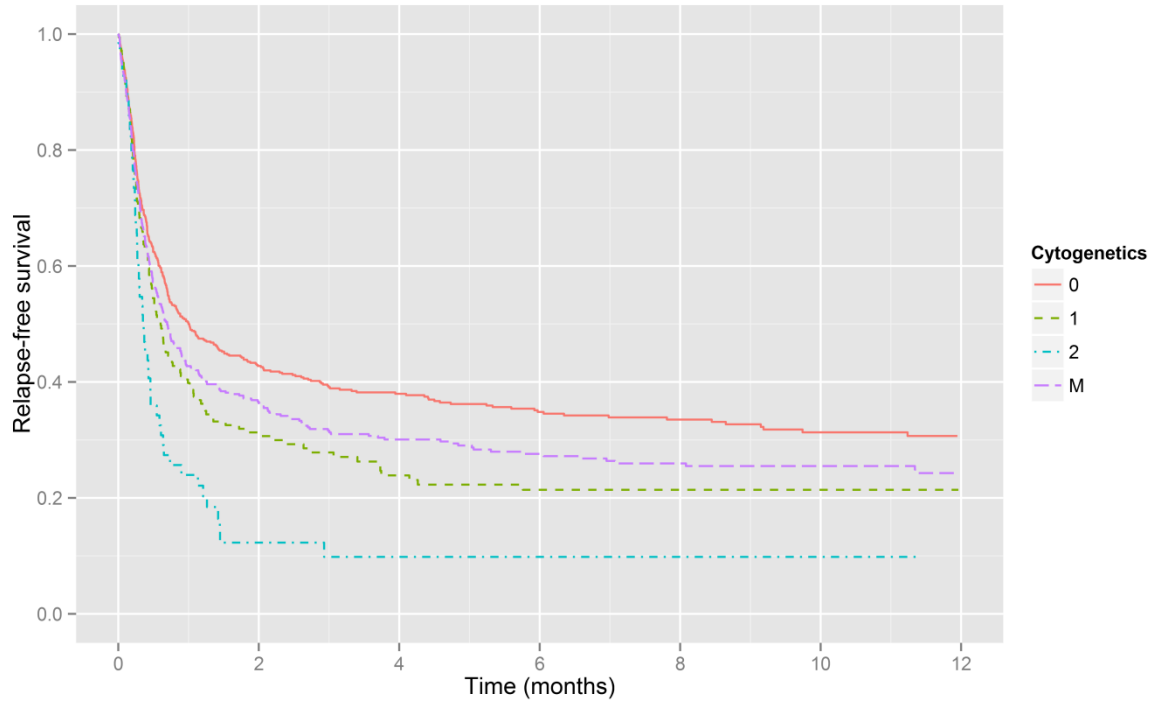


Figure 2.3 Kaplan-Meier plot of the survival times of *cytogenetics* with patients having: (0) a very good/ normal/ intermediate cytogenetic classification, (1) a poor cytogenetic classification, (2) a very poor cytogenetic classification and (M) missing cytogenetic classification.

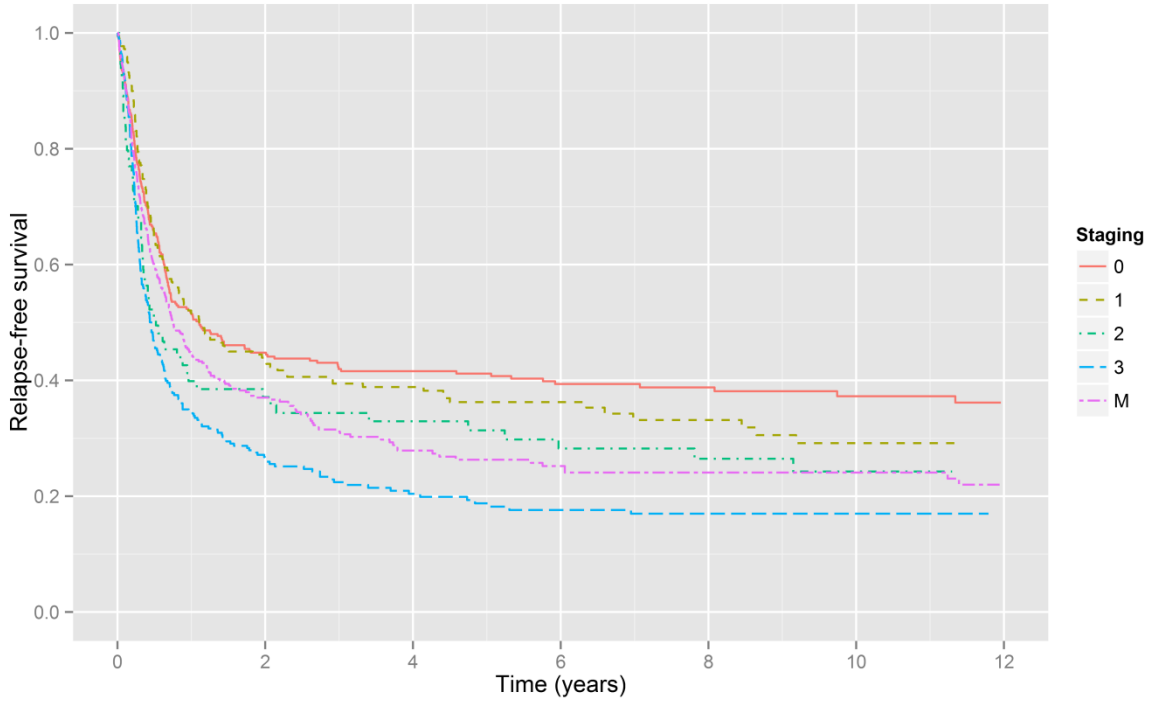


Figure 2.4 Kaplan-Meier plot of the survival times of *staging* with patients having: (0) RA/RARS/del5q/RC DM-RS untreated, (1) RAEB(t)/tAL/sMDS/CM ML in CR, (2) RAEB(t)/tAL/sMDS/CM ML untreated, (3) RAEB(t)/tAL/sMDS/CM ML not in CR and (M) missing staging.

Competing risks

Besides relapse-free-survival, information is available for the cumulative incidences of relapse and non-relapse mortality as well. The cumulative incidence of relapse at time t is defined as the probability of having experienced a relapse before time t and the cumulative incidence of non-relapse mortality is defined as the probability of having died without relapse before time t. These two can be analyzed in a competing risks framework, censoring patients that were reported alive and disease-free at the last time-point [9]. Even though no significant result was found between the complete case patients and incomplete case patients for the composite of relapse-free survival, a difference between these two groups might still be found for each specific cause.

In figure 2.5, the cumulative incidence of relapse curves of both the complete case patients and the incomplete case patients is presented. From the plot it is clear that the two groups have similar cumulative incidence curves. A log-rank test indicates the complete case patients and the incomplete case patients do not significantly differ in their rates of relapse (p-value = 0.788). The cumulative incidence curves of the non-relapse mortality for the complete case patients and incomplete case patients are shown in figure 2.6. From this plot we see the two curves start out quite similar, but after 6 months slowly diverge. A log-rank test indicates the complete case patients and the incomplete case patients do not significantly differ in their probability of non-relapse mortality (p-value = 0.073). Although not significant, a small difference could be seen between the two groups, where the incomplete case patients have a higher probability of non-relapse mortality than the complete case patients. This difference between complete case patients and incomplete case patients in non-relapse mortality might give a slight indication that the missing information is not MCAR.

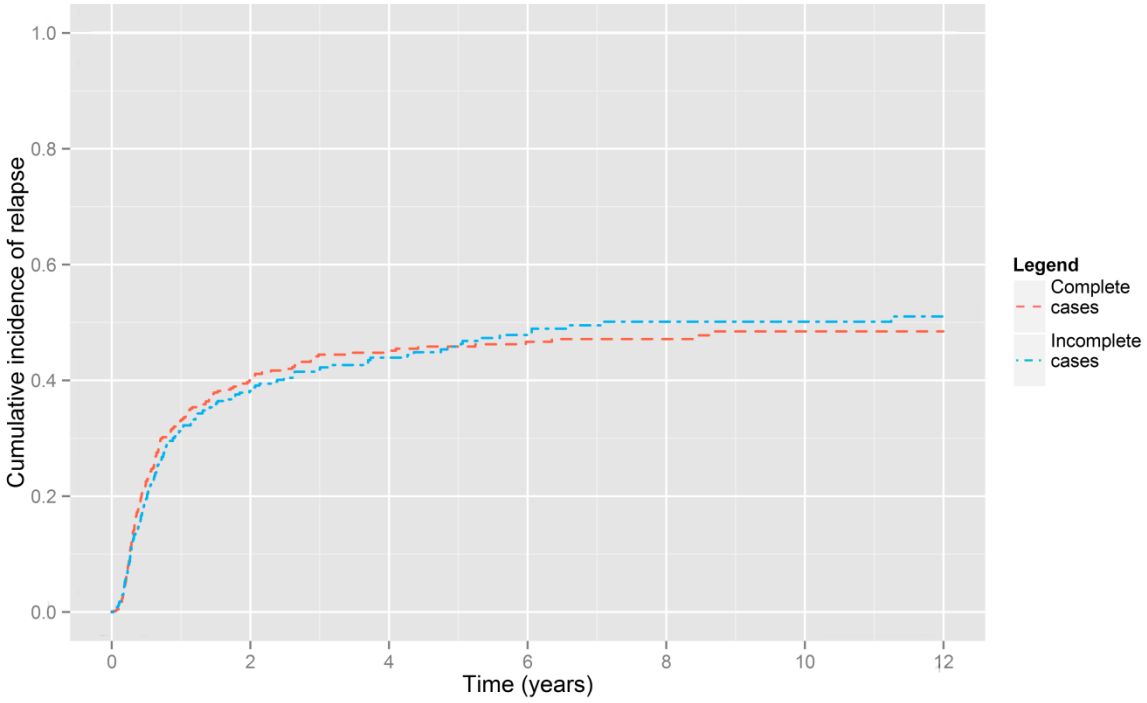


Figure 2.5 Cumulative incidence of relapse for the complete- and incomplete case individuals.

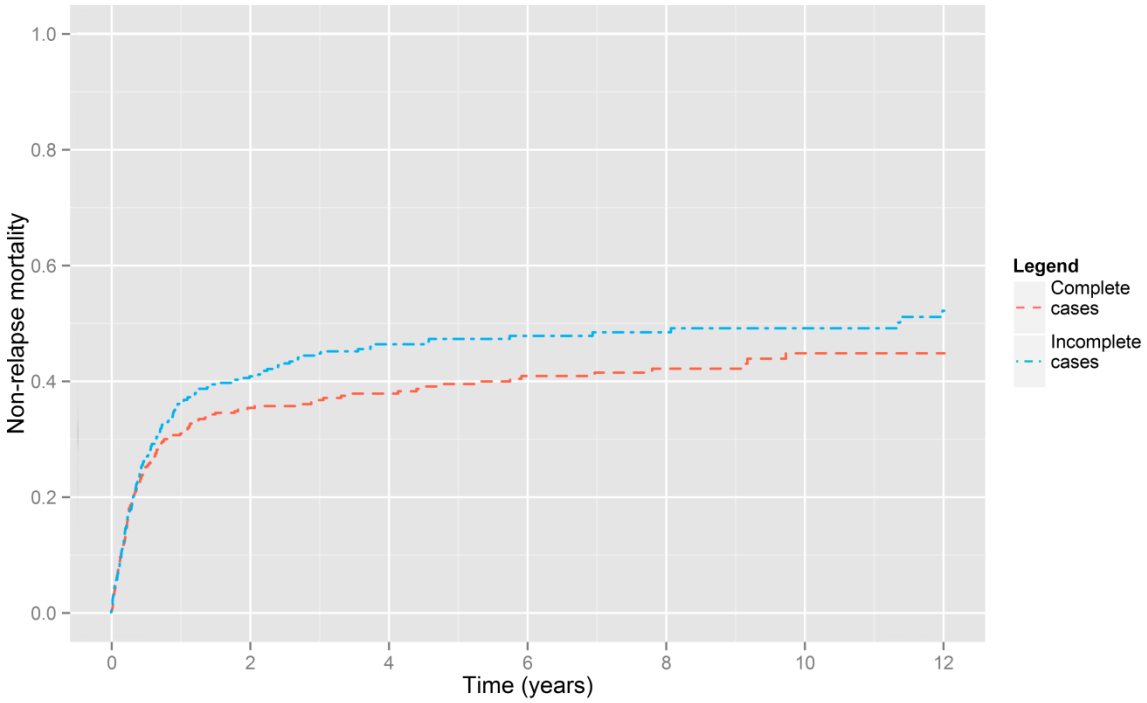


Figure 2.6 Cumulative incidence of non-relapse mortality for the complete- and incomplete case individuals.

For further analysis of the EMBT/data a prognostic model could be built with the selected variables *cytogenetics*, *staging*, *age* and *year* as covariates. Before building this model, we must choose a way to correctly handle the missing information that involves about half of the patients from the data. The above explorative analysis has given some indications that missing information from the EBMT data may not be missing completely at random (MCAR), but rather MAR/MNAR. From the results we have seen missing information in *staging* is significantly related to the year of transplantation and cytogenetic classification. We have also seen that missing information in *cytogenetics* is significantly related to the year of transplantation and the stage of the disease. Finally, in the context of competing risks, we have seen the incomplete case patients have a higher probability of non-relapse mortality than the incomplete case patients. However, this difference was just not significant. As the results of the explorative analysis suggest the data is either MAR or even MNAR (support for MNAR cannot be found from the data itself), this effects the choice of how to correctly handle the incomplete data, in order to make valid inference.

The EBMT data is the motivation of this work and will serve as a template for the simulation study on multiple imputation with chained equations and survival outcomes. After comparing several manners of inclusion of the survival outcome in the imputation model (in chapters 6 and 7), they are applied to the EBMT data for further analysis of this data. The results are presented in chapter 8.

3. Incomplete data

The problem of incomplete data and its corresponding difficulties occur frequently and in many fields of research. However, ignoring the process behind the missing data may bring along some statistical difficulties. By making a distinction between different kinds of missing data, Rubin suggested the well known clarification of the missing data mechanism [10]. The missing data mechanism helps understand the process behind missing data and which methods can be used for handling the incomplete data properly.

3.1. The missing data mechanism

The missing data mechanism governs the likelihood of being missing for each data point. To this extent, missing data are grouped into three categories: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

Paragraph 2.2.4 in the book of van Buuren *Flexible Imputation of missing data* [1] serves as a good introduction to the missing data mechanism and is depicted here in this section. Let Y be a n x p matrix of partially observed sample data and let R be a vector indicating the missing data pattern in data Y, where R=0 indicates missing values. The missing data model depicts the level of dependence of the distribution of R on Y = (Y_{obs}, Y_{mis}). The general expression of the missing data model is $P(R|Y_{obs}, Y_{mis}, \psi)$, where ψ denotes a vector containing the parameters of the missing data model. With incomplete data classified as MCAR, the likelihood of missingness depends only on the overall probability of being missing, ψ:

$$P(R = 0|Y_{obs}, Y_{mis}, \psi) = P(R = 0|\,\psi)$$

Under MCAR, complete case analysis gives consistent results. The next category of missing data is MAR. Data is said to be MAR when the probability of being missing differs between groups. These groups are required to be defined by observed data and the probability of being missing within the groups must be the same, i.e. the probability of being missing within the groups must be MCAR. Therefore we can say, with incomplete data classified as MAR, the likelihood of missingness depends on both the observed data and the parameters ψ:

$$P(R = 0|Y_{obs}, Y_{mis}, \psi) = P(R = 0|Y_{obs}, \psi)$$

Under MAR, likelihood based methods give consistent results. When data is said to be MNAR, the likelihood of missingness differs for groups or individual data points for reasons that depend on unobserved information. This means the general expression of the missing data model does not simplify and the probability of being missing depends on unobserved information as well:

$$P(R = 0|Y_{obs}, Y_{mis}, \psi)$$

The distinctions made by the missing data mechanism provide a better insight and understanding on the process behind the missing data and how to proceed with the analyses. Most simple missing data methods only provide valid statistical inferences under the MCAR assumption. However, the MCAR assumption is restrictive and most often does not hold [1].

4. Multiple Imputation

4.1. General concept

Multiple imputation is a statistical technique for handling missing data, developed by Rubin [11, 12] in the 1970's. The technique consists of three main steps: generating multiply imputed datasets, analyzing the multiply imputed datasets and pooling the estimates from the multiply imputed datasets. Figure 4.1 schematically represents these three steps as also depicted in the book of van Buuren [1, p.17]. For an observed incomplete dataset, the multiple imputation technique replaces each missing value with several, *m*, plausible values. These *m* values are drawn from a distribution modelled for each individual incomplete variable, using the observed data. This way, the multiple imputation technique creates *m* complete datasets that are identical for the observed values and may only differ from each other by the imputed values (step 1). Each complete dataset is then analyzed using standard complete data procedures (step 2). Finally, the obtained set of *m* parameter estimates is pooled into an overall estimate. The overall variance is estimated as well, by combining the conventional sampling variance (within-imputation variance) with the extra variance caused by the missing data (between-imputation variance) (step 3) [1, 2]. Multiple imputation is now considered to be the best method to deal with incomplete data in many fields [1].

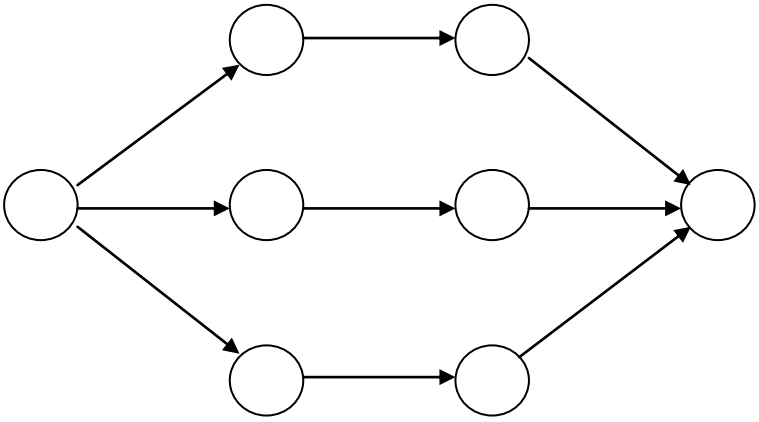


Figure 4.1 Schematic representation of main steps in multiple imputation [1, p.17]

4.2. Detailed explanation

The three main steps of multiple imputation briefly described in section 4.1, are explained in more detail in this section.

Step 1: generating the multiply imputed datasets

The multiply imputed datasets are generated by a procedure including three tasks: a modelling task, estimation task and imputation task [2]. First m multiple imputations are generated for the missing values of the target variable in data Y, denoted by Y_{mis}. The m values are drawn from the posterior distribution of Y_{mis} under the chosen model. This posterior distribution is the conditional distribution of Y_{mis} given the observed values of the complete covariate(s) in Y, denoted by Z, and the observed values of the target variable in Y, denoted by Y_{obs}. The posterior distribution to draw from is $P(Y_{mis}|\, Z, Y_{obs})$. The modelling task consists of choosing a specific model for imputation of the missing data, with θ as model parameter. For example, the model could be specified as a general normal linear model with continuous p-variate Y_i given the q-variate predictor Z, or as a multivariate logistic regression model with discrete p-variate Y_i given the q-variate predictor Z. Many other models are possible as well.

The estimation task includes calculating the posterior distribution of θ, so that θ can be randomly drawn from its posterior distribution. However, this is easier said than done. Calculating the posterior distribution of θ is most often analytically hard and computationally demanding. Therefore, Monte Carlo techniques may be used to draw θ more easily from an approximate posterior distribution. In some cases, drawing θ from its posterior distribution is not too difficult though; when a standard normal linear model is used, for example [2].

The imputation task includes drawing m imputation values for Y_{mis} from its posterior distribution under the specified model:

$$P(Y_{mis}|Z, Y_{obs}) = \int P(Y_{mis}|Z, Y_{obs}, \theta)\, P(\theta|Z, Y_{obs})\, d\theta$$

It is best to first look at the second part of the right-hand side of the equation and start by drawing a value of θ from its posterior distribution. The obtained drawn value θ^* can then be used to replace θ, when we draw Y_{mis} from its posterior distribution. This then becomes $P(Y_{mis}|Z, Y_{obs}, \theta^*)$. By repeating this process m times, there are m values drawn for imputation. Filling in the drawn values into the incomplete dataset, creates m imputed datasets [2].

Step 2: analyzing the multiply imputed datasets

Once the m imputed datasets are created, each one of them is analysed separately using standard complete data methods [2]. Let Q be a k x 1 vector with k parameters, that could only be calculated when the entire population is observed. The actual value of Q is unknown if the population data is incomplete. Multiple imputation focuses on finding an unbiased and confidence valid estimate of Q. Unbiased in the sense that the obtained estimates $\hat{Q}$ over all possible samples Y from the population should on average be equal to Q. Confidence valid in the sense that the average of the estimated variance-covariance matrix, U, over all possible samples Y from the population is equal or larger than

the variance of $\hat{Q}$. In case of incomplete population data, the actual value of Q can only be calculated with certainty when the missing values of the data are perfectly regenerated. Since this is hardly ever possible, the distribution of Q needs to be represented under varying imputation values [1]. Therefore, $\hat{Q}$ is obtained from each imputed dataset, along with the corresponding variance-covariance matrix U. Since the missing values have been replaced by different imputations in step 1, the results of these m analyses will differ from each other as well.

Step 3: pooling the estimates from the multiply imputed datasets

In the third step, the m estimates $\hat{Q}$ with corresponding variance-covariance matrices U are combined into an overall estimate and a total variance-covariance matrix. Suppose $\hat{Q}_l$ is an estimate obtained from the lth imputed dataset l, then the combined overall estimate $\bar{Q}$ is equal to the average of the individual estimates:

$$\bar{Q} = \frac{1}{m}\sum_{l=1}^{m}\hat{Q}_l$$

The total variance of $\bar{Q}$ is composed of a within-imputation variance component and a between-imputation variance component. The within-imputation variance is the conventional measure of variability, i.e. the variance due to sampling [1, 2]. The within-imputation variance is equal to the average of the repeated complete data variances of $\hat{Q}$. Suppose U_l is the estimated variance of $\hat{Q}_l$ from the lth imputed dataset l, then the combined within-imputation variance $\bar{U}$ is equal to:

$$\bar{U} = \frac{1}{m}\sum_{l=1}^{m}U_l$$

The variance between the m complete data estimates is called the between-imputation variance, denoted by B. The unbiased estimate of the between-imputation is given by:

$$B = \frac{1}{m-1}\sum_{l=1}^{m}(\hat{Q}_l - \bar{Q})(\hat{Q}_l - \bar{Q})'$$

Combining the within-imputation variance with the between-imputation variance leads to the total variance of $\bar{Q}$, denoted by V, as follows:

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

Here, the factor $(1 + m^{-1})$ multiplying B is an adjustment to correct for the extra variance caused by using a finite number of imputations m to estimate $\bar{Q}$. This adjustment is needed to make valid inferences with low m. Otherwise, the analysis would result in too low p-values or too short confidence intervals. The procedure described above to pool the repeated-imputation results is referred to as Rubin's rules [1, 2, 13].

4.3. Advantages of Multiple Imputation

Certain advantages of multiple imputation apply to imputation in general. Rubin [2] mentions imputation has the advantages of being able to allow standard complete data methods of analysis on the imputed data and the ability of incorporating the data collector's knowledge. The ability to allow standard complete data methods of analysis on the imputed data is a great advantage for the user, especially for a statistically less experienced user. It is much easier to use and draw inferences from the general standard methods of analysis on the imputed data, because scientists are familiar with these methods and they are broadly understood [2].

The ability of incorporating the data collector's knowledge is an advantage when the imputations are created by the data collector who has more information and a better understanding of the data and the non-response of it. For instance, when confidential information may be used for imputation which is not publicly available. This advantage is also enhanced by multiple imputations because it allows 'data collectors to use their knowledge to reflect uncertainty about which values to impute' [2, p.16]. The amount of certainty about an imputed value may differ between cases. When one is quite sure about the likeliness of an imputed value, there may be little variation between them across the *m* datasets. On the other hand, when there is little information about the likeliness of a value to fill in, the imputations will vary much more throughout the *m* datasets. So the variation between the imputed values indicates the level of confidence in them [1].

Furthermore, advantages of multiple imputation over single imputation include increased efficiency and the ability of making valid inferences. The imputations are randomly drawn from the updated represented distribution of the data. Hence, the efficiency of estimation is increased by multiple imputation. By combining complete data inferences according to Rubin's rules, multiple imputation is also able to make valid inferences [2]. Multiple imputation has the ability to incorporate all sources of variability and uncertainty, both within-imputation variance and between-imputation variance. By capturing the between-imputation variance, it solves the problem of too small standard errors (a typical problem of single imputation) [1, 13].

4.4. Imputation of multivariate missing data

Missing values may occur in many kinds of data, including multivariate data. Several practical problems can occur with imputation of multivariate missing data: predictors Y_{-j} (all covariates but the jth covariate) for imputing multivariate Y_j (the jth covariate) can contain missing values as well; circular dependence can occur, when the missing values of two incomplete variables depend on each other, because these variables are correlated; variables are of different types (binary, categorical, continuous etc.); collinearity or empty cells may occur when the data has large p and small n. Various other problems can occur with multivariate missing data as well [14]. Several special strategies for imputation of multivariate missing data exist, such as monotone data imputation, joint modelling and fully conditional specification [1]. This section focuses on the latter.

Fully conditional specification (FCS), also known as chained equations, is a method that imputes multivariate missing data in a variable-by-variable fashion. For each incomplete variable an imputation model needs to be specified. For incomplete multivariate data Y, the challenge lies in obtaining the multivariate distribution of θ, either explicitly or implicitly. By iteratively sampling from conditional distributions:

$$P(Y_1|Y_{-1}, \theta_1)$$
$$\vdots$$
$$P(Y_k|Y_{-k}, \theta_k),$$

the FCS-strategy acquires a posterior distribution of θ [14]. Instead of creating imputations by means of the modelling task, estimation task and imputation task (see section 4.2), imputation under FCS begins by initializing starting values for the incomplete variables. Starting values are obtained by simple random sampling from the corresponding marginal distributions. Then, imputations are created by drawing from the directly specified conditional distributions, while iterating over the conditionally specified imputation models. Imputation under FCS has the important advantage of no longer needing to specify a multivariate model for the data. The conditional distributions are specified by the user and therefore the existence of a multivariate joint distribution is assumed, but may not necessarily exist [1].

4.4.1. The MICE algorithm

One way of implementing multiple imputation under FCS is Multivariate Imputation by Chained Equations (MICE). The MICE algorithm is a Markov Chain Monte Carlo (MCMC) method and is briefly described in Algorithm 4.1 [1, p.110]. The MICE algorithm starts by specifying an imputation model $P\left(Y_k^{mis}\middle|Y_k^{obs}, Y_{-k}\right)$ for incomplete variable Y_k with k = 1, …, p. The missing values of each incomplete variable are filled in with starting imputations $Y_k^{*(0)}$ which are generated by simple random sampling from the observed values of Y_k^{obs}. After initialisation of the starting values, the iterations begin by imputing the first incomplete variable at the first iteration. The first incomplete variable Y_1, is regressed on all other variables $Y_2, \ldots, Y_p$ under the specified model for imputation of Y_1. For each incomplete variable Y_k conditional on the values of all other variables of Y, there is a univariate solution exploited where the MICE algorithm iteratively generates imputations, on a variable-by-variable basis. Multiple imputations are generated by executing the MICE algorithm in parallel m times [1].

In special cases when the conditionals are compatible and the joint distribution exists, the MICE algorithm is a Gibbs sampler. In short, a Gibbs sampler is a Bayesian simulation technique that samples from the conditional distributions, so that samples from the joint distribution are obtained. In such a case, the full conditional distributions are derived from the joint probability distribution. For the start of the first iteration of chained equations a Gibbs sampler draws for the first incomplete variable Y_1 [15]:

$$\theta_1^{*(1)} \sim P\left(\theta_1\middle|Y_1^{obs}, Y_2^{(0)}, \ldots, Y_p^{(0)}\right)$$

$$Y_1^{*(1)} \sim P\left(Y_1\middle|Y_1^{obs}, Y_2^{(0)}, \ldots, Y_p^{(0)}, \theta_1^{*(1)}\right)$$

where $Y_k^{(0)} = (Y_k^{obs}, Y_k^{*(0)})$ is the kth completed variable after initializing starting values before the iterative procedure starts. For notational convenience, the notation for the originally completely observed variables is suppressed here, so all distributions are implicitly conditional on the fully observed variables. Next, the subsequent variable with missing values Y_2 is regressed on all other

variables under the specified model, including the imputed variable Y_1. This process of imputing the incomplete variables one at a time continues up until the last incomplete variable Y_p, when the first cycle is completed. In general notation, the jth iteration of chained equations is a Gibbs sampler that successively draws [15]:

$$\theta_1^{*(j)} \sim P\left(\theta_1 \middle| Y_1^{obs}, Y_2^{(j-1)}, \dots, Y_p^{(j-1)}\right)$$

$$Y_1^{*(j)} \sim P\left(Y_1 \middle| Y_1^{obs}, Y_2^{(j-1)}, \dots, Y_p^{(j-1)}, \theta_1^{*(j)}\right)$$

$$\vdots$$

$$\theta_p^{*(j)} \sim P\left(\theta_p \middle| Y_p^{obs}, Y_1^{(j)}, \dots, Y_{p-1}^{(j)}\right)$$

$$Y_p^{*(j)} \sim P\left(Y_p \middle| Y_p^{obs}, Y_1^{(j)}, \dots, Y_{p-1}^{(j)}, \theta_p^{*(j)}\right)$$

where $Y_k^{(j)} = (Y_k^{obs}, Y_k^{*(j)})$ is the kth imputed variable at iteration j. The imputed values, and therefore the incomplete variables itself, are updated at each iteration.

In the MICE algorithm, the conditional distributions are specified by the user. The joint distribution is therefore only implicitly known, but it does not necessarily exist. After a certain number of iterations, the Markov chain should converge to a stationary distribution. At that point the chain must be irreducible, aperiodic and recurrent [1]. The Gibbs sampler aims for convergence in distribution, but without deriving the conditional distributions from the joint distribution convergence may not always be self-evident. Monitoring convergence is quite easily done by plotting statistics of interest, such as the mean and standard deviation for instance, in each stream against the number of iterations. The number of iterations needed for the chain to converge differs per data, but is usually quite small with about 5-10 iterations.

After the m imputed datasets are created according to the MICE algorithm, the regular multiple imputation steps follow (step 2 and 3): analyzing the m complete datasets, followed by pooling the results of the m complete datasets. An important advantage of MICE is that it can handle a set of mixed measurement levels (continuous, binary, categorical), due to the ability of using different imputation models for each incomplete variable [1].

Algorithm 4.1: the MICE algorithm [1, p.110]

1. Specify an imputation model $P\left(Y_k^{mis}\middle|Y_k^{obs},\ Y_{-k}, R\right)$ for incomplete variable Y_k, with k = 1, …, p.
2. For each incomplete variable, initialize starting imputations $Y_k^{*(0)}$ by random draws from Y_k^{obs}.
3. Repeat for iterations j = 1, … J:
4. Repeat for number of incomplete variables k = 1, … p:
5. Define the imputed data $Y_{-k}^{(j)}$ = $\left(Y_1^{(j)}, \ldots, Y_{k-1}^{(j)}, Y_{k+1}^{(j-1)}, \ldots, Y_p^{(j-1)}\right)$ as the currently complete data at iteration t except Y_k.
6. Draw $\theta_k^{*(j)} \sim P\left(\theta_k\middle|Y_k^{obs},\ Y_{-k}^{(j)},\ R\right)$.
7. Draw imputations $Y_k^{*(j)} \sim P\left(Y_k^{mis}\middle|Y_k^{obs},\ Y_{-k}^{(j)},\ R,\ \theta_k^{*(j)}\right)$.
8. End repeat j.
9. End repeat k.

5. Handling missing covariate values with censored data

Incomplete data can be handled in many different ways. One of the options is multiple imputation of covariates with chained equations. For analyzing survival data with missing values in one or more covariates, the survival outcome needs to be included in the imputation model [3]. The main purpose of this simulation study is to compare different manners in which survival data with missing values in one or more covariates can be analyzed. This chapter describes nine different manners of handling incomplete survival data.

5.1. Complete case analysis

Complete case analysis is probably the simplest way of handling missing data. By this procedure all cases with one or more missing values are omitted from the analysis, leaving only the complete cases. The major advantage of complete case analysis is the ease of implementation, since standard methods for computation can be applied after omitting incomplete cases. When the missing data are MCAR, complete case analysis produces unbiased estimates of means, variances and regression weights. The standard errors and significance levels are correct for the reduced subset of data, but may be larger than they would have been without missing information [1].

The main drawback of complete case analysis is that it is wasteful, especially for data with a large amount of missing values. Many missing values often occur when there are many predictors in the dataset, however not necessarily. Losing (more than) half of the cases, when omitting the incomplete ones is quite common. As a result, it becomes harder to detect the effects of interest [1]. Besides, with a small subset of data used for estimation, the estimates may be less precise than initially planned. In case the incomplete data are MAR or MNAR, complete case analysis gives biased estimates and invalid inferences [16]. From now on we will refer to complete case analysis as CC.

5.2. Creating an extra category

For incomplete unordered categorical variables, a simple and often useful approach of handling the missing values is to add an extra category for the variable indicating missing values. This method treats the missing information as valid data by adding the extra category and has as advantage over complete cases that it retains all data. However, by adding an extra category for the categorical variable, dissimilar classes may be put together into one group and potentially cause severe bias. The method as described here is meant for categorical covariates only. The impact of this approach on the inferences depends on the missing data mechanism and on how the missing values are divided among the real categories [www.missingdata.org.uk]. From this point on the method of creating an extra category is referred to as XCAT.

5.3. Multiple Imputation without inclusion of the survival outcome

Using multiple imputation (with chained equations) for analyzing incomplete survival data, first implies specifying an imputation model. According to Moons et al., the imputation model must not

only consist of the covariates that have any association with the incomplete data assessed in the analysis model [4], but also the outcome [3]. Since in this method the missing covariate values are imputed without taking the survival outcome into account ('under the null-model'), it is likely that after multiple imputation, the regression coefficients of the missing covariates are attenuated towards zero.

The imputation is based on a unique regression model for each of the incomplete covariates, depending on the characteristics of these covariates and using the other covariates as predictors. For imputation of the incomplete covariate Z_{inc}, the imputation model consists of all covariates designated as predictors, but not the survival outcome. Imputation values for incomplete covariate Z_{inc}, are generated by:

$$Z_{inc} \sim \beta_0 + \sum \beta_j Z_j,$$

which serves as the notation of the regression model. The type of the regression depends on outcome Z_{inc}, with a logistic regression in case of a binary Z_{inc} for instance. On the right-hand-side, β_0 represents the intercept and the remaining covariates Z_j serve as predictors. The observed event or censoring time T and status indicator δ are left out of the regression that generated imputation values for Z_{inc}, as they are not included in the imputation model. The method of multiple imputation without survival analysis is later on referred to as NO-T.

5.4. Multiple Imputation including the survival outcome

The first and most obvious way in which the survival outcome may be included in the imputation model, would be to simply add both the observed event or censoring time T and the status indicator δ into the imputation model. This way, both T and δ are included as predictors next to some other covariate(s) for imputing missing covariate values. This approach is used by many authors [8] and could be considered common practice with incomplete survival data. Imputation values for incomplete covariate Z_{inc}, are generated by:

$$Z_{inc} \sim \beta_0 + \sum \beta_j Z_j + \delta + T,$$

where the sum is over the remaining covariates. Multiple imputation including the survival outcome is later on referred to as T.

5.5. Multiple Imputation including the log of the survival outcome

Some authors suggest including the survival outcome in the imputation model, while taking the logarithm of the observed event or censoring time, $\log(T)$ [6]. This way, both $\log(T)$ and δ are included in the imputation model as predictors next to some other covariate(s) for imputing missing covariate values. Imputation values for incomplete covariate Z_{inc}, are generated by:

$$Z_{inc} \sim \beta_0 + \sum \beta_j Z_j + \delta + \log(T),$$

where the sum is over the remaining covariates. From now on the method of multiple imputation including the log of the survival outcome is referred to as LOGT.

5.6. Multiple Imputation including the cumulative baseline hazard

A simulation study conducted by White & Royston [5] has shown that including the cumulative baseline hazard in the imputation model is preferred over inclusion of (the log of) the survival outcome. They prove a number of exact and approximate results that serve as evidence to support their suggestion of including the cumulative baseline hazard in the imputation model. Their line of reasoning is depicted below.

Let X and Z denote the incomplete covariate of interest and the other covariates (for simplicity of notation we consider only one). Let $L(T, \delta|X, Z)$ denote the likelihood of the time-to-event outcome, given the covariates. It is given by

$$L(T, \delta|X, Z) = h(T|X, Z)^{\delta} \cdot \exp\big(-H(T|X, Z)\big), \tag{5.1}$$

where, under the proportional hazards model,

$$h(t|X, Z) = h_0(t) \cdot \exp(\beta_X X + \beta_Z Z), \tag{5.2}$$

and similarly for the cumulative hazard $H(t|X, Z)$. Taking the log likelihood results in

$$\begin{aligned}\log\big(L(T, \delta|X, Z)\big) &= \delta \log\big(h(T|X, Z)\big) - H(T|X, Z) \\ &= \delta\big(log\big(h_0(T)\big) + \beta_X X + \beta_Z Z\big) - H_0(T) \exp(\beta_X X + \beta_Z Z).\end{aligned}$$

By Bayes theorem, we have

$$P(X = x|T, \delta, Z) = \frac{L(T, \delta|X = x, Z) \cdot P(X = x|Z)}{\sum_x L\,(T, \delta|X = x, Z) \cdot P(X = x|Z)}.$$

It follows that

$$\frac{P(X = 1|T, \delta, Z)}{P(X = 0|T, \delta, Z)} = \frac{L(T, \delta|X = 1) \cdot P(X = 1|Z)}{L(T, \delta|X = 0) \cdot P(X = 0|Z)},$$

and as a result (with $logit(p) = \log\left(\frac{p}{1-p}\right)$), that

$$logit(P(X = 1|T, \delta, Z)) = \log\left(\frac{L(T, \delta|X = 1, Z)}{L(T, \delta|X = 0, Z)}\right) + logit\big(P(X = 1|Z)\big). \tag{5.3}$$

If we denote $logit\big(P(X = 1|Z)\big)$ by ζ_Z and combine equations (5.1), (5.2) and (5.3), we obtain

$$
\begin{aligned}
logit(P(X=1|T,\delta,Z)) &= \log\big(L(T,\delta|X=1,Z)\big) - \log\big(L(T,\delta|X=0,Z)\big) + \zeta_Z \\
&= \delta\big(\log\big(h(T|X=1,Z)\big) - \log\big(h(T|X=0,Z)\big)\big) \\
&\quad - \big(H(T|X=1,Z) - H(T|X=0,Z)\big) + \zeta_Z \\
&= \delta\big(log\big(h_0(T)\big) + \beta_X + \beta_Z Z - log\big(h_0(T)\big) - \beta_Z Z\big) \\
&\quad -(H_0(T)\exp(\beta_X + \beta_Z Z) - H_0(T)\exp(\beta_Z Z)) + \zeta_Z \\
&= \delta\beta_X - H_0(T)\cdot\exp(\beta_Z Z)\cdot(\exp(\beta_X) - 1) + \zeta_Z
\end{aligned}
$$

Without extra covariate(s) Z, this is a logistic regression with δ and $H_0(T)$ as covariates in the imputation model for X. With one categorical covariate Z, it is a logistic regression model with δ, $H_0(T)$ and the interaction between Z and $H_0(T)$.

5.6.1. Cumulative baseline hazard approximated by the Nelson-Aalen estimator

The cumulative baseline hazard is usually unknown and must be estimated. When the covariate effects $\beta_{Z_{inc}}$ and β_Z are small, the non-parametric estimator of the cumulative hazard rate function, Nelson-Aalen estimator, may be used for estimating the cumulative baseline hazard; $H_0(T) \approx H(T)$ [5]. The estimation of the cumulative baseline hazard takes place before imputation, so that $\widehat{H}(T)$ is included in the imputation model together with δ and some other covariate(s). Now that $\widehat{H}(T)$ is included in the imputation model, the ordinary observed event or censoring time T is no longer included. Imputation values for incomplete covariate Z_{inc}, are generated by:

$$
Z_{inc} \sim \beta_0 + \sum \beta_j Z_j + \delta + \widehat{H}(T),
$$

where the sum is over the remaining covariates. The method of multiple imputation including the cumulative baseline hazard approximated by the Nelson-Aalen estimator is later on referred to as NA.

5.6.2. Cumulative baseline hazard estimated by a Cox proportional hazards model

The cumulative baseline hazard may also be estimated iteratively by means of a Cox proportional hazards model, during the imputation process in MICE [5]. In practice, this can be done by adjusting the beginning of the imputation method used for imputation of the missing covariate values. The default method specified by MICE can be used as imputation method for regression or one could choose another appropriate imputation method. Either way, iteratively estimating $H_0(T)$ is implemented in the imputation method. The method will be adjusted, such that it starts by fitting a Cox proportional hazards model to the data using the current (iteratively updated) values of the covariates Z_{inc}, Z and extracting the updated estimated cumulative baseline hazard. Then, the updated estimated cumulative baseline hazard replaces the observed event or censoring time T as predictor for imputation of Z_{inc}. Imputation values for incomplete covariate Z_{inc}, are generated by:

$$Z_{inc} \sim \beta_0 + \sum \beta_j Z_j + \delta + \hat{H}_0(T),$$

where the sum is over the remaining covariates. From now on the method of multiple imputation including the cumulative baseline hazard estimated by the Cox proportional hazards model is referred to as COX.

5.7. Multiple Imputation including pseudo-observations

Another way to consider the survival outcome for imputation of covariates with chained equations, is by using pseudo-observations [17]. Pseudo-observations can be used to make the observed event or censoring time T complete. The pseudo-observations create a transformation of the observed survival time by the change in the Kaplan-Meier estimate when leaving out a single observation from the survival data. Subsequently, the survival time variable is completed, such that T now only consists of observed event time points. The obtained pseudo observations can be included into the imputation model together with some covariate(s) as predictors for imputation of missing covariate values. The survival outcome is then included in the imputation model in the form of pseudo-observations. In paragraph 5.7.1. the definition of pseudo-observations are explained and in paragraph 5.7.2. it is described how pseudo-observation may be used in analysis.

5.7.1. Definition of pseudo-observations

Survival data is characterized by its incomplete survival time X. Would X have been observed for all individuals, standard methods for quantitative data could be applied directly for the survival time. One way of achieving this with censored survival data is to replace X by the pseudo-observations [17]. This procedure will be explained with an illustrative example.

Let X be the survival time of the patients and let ψ be a parameter of the form $\psi = E(X)$. The data is assumed to be independently right-censored data. The counting process of the data given the number of observed events in [0, t] is as follows:

$$N(t) = \sum_i I(X_i \leq t, D = 1).$$

The number of individuals for which an event was not (yet) observed and who are therefore still at risk at time t is:

$$Y(t) = \sum_i I(X_i \geq t).$$

With the number of observed events and the number of individuals at risk, the survival function can be estimated using the Product-Limit estimator [18]. The Product-Limit estimator provides efficient and unbiased means of estimating the survival function for independent right-censored data. The survival function can be estimated using the Product-Limit estimator:

$$\hat{S}(t) = \prod_{u \leq t} \left(1 - \frac{dN(u)}{Y(u)}\right),$$

after which the pseudo-observations can be calculated by taking the integral of the Product-Limit estimator over all observed event or censoring time points:

$$\hat{\psi} = \int_0^\infty \hat{S}(t)\, dt.$$

To avoid infinite integrals, we restrict the integral from zero to the last observed time point. The i-th pseudo-observation of X is defined as:

$$\hat{\psi}_i = n \cdot \hat{\psi} - (n-1)\, \hat{\psi}^{-i},$$

where $\hat{\psi}^{-i}$ is the estimator applied to the data with sample size $n-1$ obtained by leaving out the i-th individual. Intuitively the i-th pseudo-observation $\hat{\psi}_i$ can be seen as the contribution of the individual i to the $E(X)$ estimate on the sample of size n. Usually, pseudo-observations are used in regression models for censored outcomes. In the absence of censoring, it can be seen that $\hat{\psi} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and that we have $\hat{\psi}_i = X_i$. In the presence of censoring, the $\hat{\psi}_i$'s will be close to X_i for the uncensored cases and we hope that they represent some expected value of X_i conditional on being greater than the observed time of subject i. We follow the general recommendations in using pseudo-observations for all individuals, not only for those where X_i was unobserved [17].

5.7.2. Including pseudo-observations in the imputation model

As mentioned before, pseudo-observations can be used to make survival data complete. The pseudo-observations replace the observed event or censoring time T, after which various kinds of models can be used for analysis. When pseudo-observations are included in the imputation model as predictor next to the other covariate(s) for imputation of the missing covariate values, the original survival outcome T and δ are substituted and no longer needed. Imputation values for incomplete covariate Z_{inc}, are generated by:

$$Z_{inc} \sim \beta_0 + \sum \beta_j Z_j + \psi,$$

where the sum is over the remaining covariates and ψ denotes the pseudo-observations. The method of multiple imputation including pseudo-observations in the imputation model is later on referred to as PSEUDO.

5.8. Multiple Imputation including imputed censored observations

The observed survival data consists of n observations with a time variable T, a status indicator δ and some (incomplete) covariates Z. As before, the survival outcome may be included as predictor for imputing the missing values of the covariates Z. However, for right censored patients it is unknown when the event will take place. Therefore, right censored observations can be considered missing values as well. While including the time variable T and specifying the right censored observations as missing values in MICE, the time variable T will be imputed itself. Consequently, the imputed time variable $T_{complete}$ will serve as predictor for imputing the covariates Z, instead of the 'incomplete' right censored time variable T.

For imputation of the censored time points in T, a new imputation method is developed. An existing imputation method cannot be used since imputing the censored observations needs to take into account certain assumptions. For instance, the imputed survival times must always be larger than the original censored time points. The imputation method for imputing the censored observations starts by using a Cox proportional hazards model including the covariates Z. With this model an estimate can be made of the regression coefficients $\hat{\beta}$ and the baseline hazard rate $\hat{h}_0(t)$. In the data, there are n_0 censored observations $(t_i, \delta_i = 0)$ for which the actual event times are now considered missing values. Consequently, for each patient the hazard rate is calculated:

$$\hat{h}_i(t, Z_i) = \hat{h}_0(t) \exp\left(\sum_{j=1}^{p} \hat{\beta}_j Z_{ij}\right)$$

Then a survival curve for that patient can be derived as follows:

$$\hat{S}_i(t,\ Z_i) = \ \exp\ (-\int_0^t \hat{h}_i(s, Z_i)\ ds),$$

where t denotes the survival time (with observed event time points and censored time points). For imputation we draw one value from all survival times of the survival curve that are higher than the observed censored time with certain probabilities, $\hat{P}(T > t \mid T > t_i,\ Z_i\ )$. The value used for imputation must be higher than the censored value, since the patient was event-free until this point in time. The probability of drawing a value of the survival times higher than the observed censored time, is the survival probability of the data divided by the survival probability of the censored data:

$$\hat{P}(T > t \mid T > t_i, Z_i\ ) = \frac{\hat{S}_i(t, Z_i)}{\hat{S}_i(t_i, Z_i)} = \ \exp\ (-\int_{t_i}^t \hat{h}_i(\mathrm{s},\ Z_i)\ ds).$$

The drawn value is used for imputation. However, it might occur that a certain censored time point is the highest observed time point, such that there is no higher event time point to choose from. In such case, the imputed value becomes equal to the censored time point. This newly proposed method is a special case of risk set imputation as described by van Buuren [1, p.80].

Summarized, before imputation of the incomplete covariate(s), the censored observations will be imputed with MICE according to the above described imputation method. Then, imputation values for incomplete covariate Z_{inc}, are generated by:

$$Z_{inc} \sim \beta_0 + \sum \beta_j Z_j + T_{complete},$$

where the sum is over the remaining covariates. The missing covariate values can still be imputed by using any of the appropriate imputation methods already available (including the default methods), as long as the censored observations are imputed according to a specialized imputation method, such as the one described above. Multiple imputation including imputed censored observations is from now on referred to as CENS.

6. Simulation study

The main purpose of this study is to compare different manners in which survival data with missing values in one or more covariates can be analyzed. Now that nine different manners of handling incomplete survival data are described in the previous chapter, they are compared by means of a simulation study to see which one performs best on the generated datasets.

6.1. Design of the simulation study

6.1.1. Generating data

Datasets were randomly generated resembling the EBMT data. The generated data consists of the survival time T, the status indicator δ and five covariates; *year* (Z_1)*, age* (Z_2)*, country* (Z_3), *staging* (Z_4) and *cytogenetics* (Z_5). Both *year* and *age* are continuous covariates, *country* is a binary covariate and *staging* and *cytogenetics* are categorical covariates with four and three categories, respectively. The first covariate, *year*, was generated by randomly drawing from a uniform distribution on [1982, 2010] - 2000. The covariate *age* was centered around the age 50 and generated by drawing from a normal distribution with μ = 0 and σ = 6. The third covariate, *country*, was generated by drawing from a binomial distribution with size 1 and probability 0.75. The covariates *cytogenetics* and *staging* were generated with marginal probabilities resembling the EBMT data, but with maximum correlation between the two covariates. To this extent a matrix, $M_{ij} = P(Z_4 = i, Z_5 = j)$, was generated with the shared probabilities of both covariates. To the upper left and bottom right values of the matrix, a value Δ = 0.035 was added, while in the opposite corners of the matrix this value was subtracted. Randomly sampling numbers 1-12 with replacement and shared probabilities from the defined matrix

$$M = \begin{bmatrix} 0.245 + \Delta & 0.070 & 0.035 - \Delta \\ 0.210 & 0.060 & 0.030 \\ 0.035 & 0.010 & 0.005 \\ 0.210 - \Delta & 0.060 & 0.030 + \Delta \end{bmatrix},$$

generates the shared categories for both covariates *cytogenetics* and *staging*. The resulting correlation between the two categorical covariates is about 0.3.

The survival times were drawn from a Weibull distribution with hazard $h_T(t) = \alpha \lambda_T t^{\alpha-1} \exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_{41} \cdot I(Z_4 = 1) + \beta_{42} \cdot I(Z_4 = 2) + \beta_{43} \cdot I(Z_4 = 3) + \beta_{51} \cdot I(Z_5 = 1) + \beta_{52} \cdot I(Z_5 = 2)$ and with scale and shape parameters (λ_T, α) = (0.005, 1) or (0.1, 0.5), so that the mean survival is 200 in both cases, roughly equal to the mean survival (in months) of the EBMT data. Random censoring times were drawn from a Weibull distribution with scale and shape parameters (λ_C, α) = (0.005, 1) or (0.1, 0.5), $h_C(t) = \alpha \lambda_C t^{\alpha-1}$. Values of the β's are mimicking the regression coefficients obtained by a Cox proportional hazards model on the EBMT data:

β_1=-0.02, β_2=0.01, β_3=0.8, β_{41}=0.09, β_{42}=0.4, β_{43}=0.6, β_{51}=0.2, β_{52}=0.7. This parameterization corresponds to 35% censoring. Taking the minimum between the generated survival times and the generated censoring times, results in the observed event or censoring time T for each individual.

Status indicator δ was set to be 0 in case of censoring time and 1 in case of an event time point. The sample size of the data, n, varies per simulation and is set to n = {500, 1000}.

6.1.2. Missing values

After generating each dataset, a part of the data is set to be missing. Incorporating missing values occurred under the assumption of the three groups of the missing data mechanism of Rubin [10]: MCAR, MAR and MNAR (see section 2.1). Similar to the EBMT dataset, missing values were only created in the two categorical covariates; *cytogenetics* and *staging*. In each generated dataset, missing values are incorporated in *cytogenetics* and *staging* according to MCAR, MAR and MNAR, leaving three types of datasets available for analysis. Data can be classified as MCAR when the probability of being missing is the same for each data point. Therefore, creating missing values according to MCAR, involves calculating the random chance of being missing. This was achieved by drawing n times from a binomial distribution with size 1 and probability p_1, where p_1 indicates the chosen fraction of missing values in covariate *cytogenetics*. Parameter p_1 may vary per simulation in order to change the fraction of missing values in *cytogenetics*. Parameter p_1 is set to p_1 = {0.4, 0.6}, ranging from a large to severe fraction of missing values in *cytogenetics*. The corresponding observations for *cytogenetics* from the binomial distribution are set to NA. After creating missing values for *cytogenetics*, missing values are created in a similar manner for the other categorical covariate, *staging*. There was n times randomly drawn from a binomial distribution with size 1 and probability p_2, where p_2 indicates the chosen portion of missing values in *staging*. Parameter p_2 is set to p_2 = {0.4, 0.6}, ranging from a large to severe fraction of missing values in covariate *staging*. The corresponding observations for *staging* from the binomial distribution are set to NA.

Data is said to be MAR when the probability of being missing differs between the observed groups. Creating missing values according to MAR, involves calculating the random chance of being missing given some variable of the data that is also included in the analysis. The dependent variable is chosen to be the binary covariate *country*. The data is divided into two groups: about 25% of the patients are treated in one country (group 1) and about 75% is treated in the other country (group 2). The chance of a missing value is set to be lower in group 1 than it is in group 2. Creating missing values in *cytogenetics* for group 1 according to MAR is achieved by randomly drawing from a binomial distribution with size 1 and probability 0.1. For group 2 this is achieved by drawing randomly from a binomial distribution with size one and probability ($4 \cdot p_1/3$ - 0.1/3) for group 2. Note that the overall probability of missing values in *cytogenetics* equals p_1. The parameter p_1 is set to p_1 = {0.4, 0.6}, ranging from a large to severe fraction of missing values in covariate *cytogenetics*. The corresponding observations for *cytogenetics* from the binomial distribution are set to NA. After creating missing values for *cytogenetics*, missing values are created in a similar manner for the other categorical covariate, *staging*.

Data is classified as MNAR, when the probability of being missing differs for groups or individual data points for reasons that depend on unobserved information. Therefore, creating missing values according to MNAR, involves calculating the random chance of being missing given some unobserved variable. The data under MNAR in the simulation is the exact same data as the data under MAR. This time, however, *country* will be excluded from the analysis. By excluding *country* from the analysis, the different probabilities of being missing between the two groups now depend on a variable that is unobserved and therefore unknown to us.

6.1.3. Imputation

After the missing values are incorporated in the generated datasets, the nine different procedures are applied to the incomplete datasets under the assumption of all three groups of the missing data mechanism of Rubin [10]. This implies that each procedure is applied to all three types of generated datasets (MCAR, MAR and MNAR). The first two procedures for handling survival data with missing covariate values described in the first two sections of chapter 5, *Complete case analysis* and *Creating an extra category*, do not use multiple imputation and are therefore left out of consideration in this section. The remaining seven procedures described, all impute the missing covariate values with MICE. The survival outcome plays a different role in each of these procedures. For comparison, the simulated datasets are also analyzed before introducing missing values (later referred to as PERFECT).

Generated covariates *cytogenetics* and *staging* are both categorical with more than two categories and have missing values. Categorical incomplete covariates are imputed by the polytomous regression model, which is the default imputation method for categorical variables in R. The method starts by fitting the categorical response as a multinomial model, then computed the predicted categories and, finally, adds appropriate noise to the predictions. Bias due to perfect prediction is avoided by augmenting the data. Brand [19] provides a more detailed explanation of the method in his dissertation. Polytomous regression is generalization of logistic regression, used for imputation of an incomplete binary variable and modeled as:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p;$$

with $\pi = P(y = 1|x_1, \ldots, x_p)$. Polytomous regression for s categories can be modelled as a series of separate logistic regression models of the categories $1, \ldots, s-1$ against a baseline category 0 according to:

$$\ln\left(\frac{P(y = j|x)}{P(y = 0|x)}\right) = \beta_{j0} + \beta_{j1} x_1 + \ldots + \beta_{jp} x_p;$$

with $j = 1, \ldots, s-1$ [19]. Each time the datasets are multiply imputed, there are m = 10 imputed datasets constructed, with a burn-in of 15 iterations. Table 6.1 briefly summarizes the imputation procedures, where Z_{inc} denotes the incomplete covariate and Z denotes the covariates included as predictors for imputation of Z_{inc}.

Table 6.1 Imputation of missing covariate values

Abbreviation	Description
NO-T	Polytomous regression of Z_{inc} on Z
T	Polytomous regression of Z_{inc} on Z, T and δ
LOGT	Polytomous regression of Z_{inc} on Z, $\log(T)$ and δ
NA	Polytomous Regression of Z_{inc} on Z, $\widehat{H}(T)$ and δ
COX	Polytomous Regression of Z_{inc} on Z, $\widehat{H}_0(T)$ and δ
PSEUDO	Polytomous Regression of Z_{inc} on Z and ψ
CENS	Polytomous Regression of Z_{inc} on Z and $T_{complete}$

6.1.4. Analysis

After the first step of multiple imputation is completed, the 10 imputed datasets are analyzed separately by means of a Cox regression on the five complete covariates in case of MCAR and MAR. For datasets with missing values generated MNAR, the 10 imputed datasets are analyzed separately by means of a Cox regression on the four complete covariates, since *country* is left out of the analysis. The results from the analyses are then combined using Rubin's rules [1].

After completing the simulation, the results are explored by calculating the bias, root mean square error (RMSE) and the per cent coverage for each procedure. The RMSE is a commonly used measure that represents the error between the estimated values and the observed values. It is used as a measure of accuracy between the procedures, with a smaller value of the RMSE as higher accuracy. The per cent coverage is calculated as the proportion of time that the 95% confidence interval contains the theoretical values of the regression coefficients. The coverage of a procedure is good when it is (close to) 95%. Coverage lower than 95% is called undercoverage and could either be attributed to highly biased results or too small standard errors. Coverage higher than 95% is called overcoverage and might be caused by too large standard errors.

6.2. Results simulation study

Results of the simulation study concerning the RMSE and the per cent coverage are presented in tables 6.2 and 6.3. These results are only presented for the third category of *cytogenetics* (the very poor risk patients), as this category has the strongest effect on relapse free survival in the EBMT data. Since the simulation data was generated resembling the EBMT data, the corresponding parameter for the very poor risk patients was given an effect of β_{52}=0.7, similar to the effect the very poor risk patients have on relapse free survival in the EBMT data after complete cases analysis. Results for other parameters are found to be quite similar, but less outspoken. Parameters with a weaker effect on relapse free survival show a smaller RMSE (results not shown), but the proportion between the nine methods is similar to the proportion between the nine methods for parameter β_{52}.

Root mean square error

The RMSE for the different methods in estimating the parameter β_{52} is presented in table 6.2 below. The table shows that with similar sample size and percentage of missing values no appreciable differences in RMSE were found between the generated survival time following an Exponential- or a Weibull distribution. Furthermore, we see the RMSE is usually smaller for data with a larger sample size (n = 1000), than it is with a smaller sample size (n = 500). This result is not surprising. Differences in RMSE are clearly found between data with a large percentage of missing values in both *staging* and *cytogenetics* (40%) and data with a severe percentage of missing values (60%). Data with a large percentage of missing values show a smaller RMSE than data with a severe percentage of missing values in both covariates. This is what we expect, since there is less certainty on the real values with severe missingness than when there is a smaller percentage of missingness in the data.

Comparing the procedures all together, we see procedures T, LOGT and CENS perform best under all three conditions (MCAR, MAR and MNAR), with respect to the RMSE. Under MNAR, however, procedure XCAT performs quite good as well. Procedure NO-T seems to perform worst of all. The high values of the RMSE for NO-T are mainly caused by the strong bias towards the null of the

procedure. Since in this procedure the missing covariate values are imputed without taking the survival outcome into account ('under the null-model'), we already expected that after multiple imputation, the regression coefficients of the missing covariates would be attenuated towards zero. Contrary to expectation, the bias of CC and XCAT was quite modest, also under MAR and MNAR. From the table we see procedure CC does not perform well with respect to the RMSE, which is actually as we would expect. For CC the high values of the RMSE are most likely caused by the smaller effective sample size (incomplete cases are left out). Surprisingly, all procedures seem to perform better under MAR/MNAR data than under MCAR data, as the RMSE seems smaller under MAR/MNAR data than under MCAR data. This was probably caused by the percentage of complete cases that was left over in the data after creating missing values. The percentage of complete cases was 36% under MCAR and 39% under MAR/MNAR when creating 40% missing values in two covariates and it was 16% under MCAR and 26% under MAR/MNAR when creating 60% missing values in two covariates. Furthermore, with sample size n = 500, we see all procedures show quite similar results under MNAR, except for procedure CC, which performs worse than the others. With sample size n = 1000, differences between procedures under MNAR are clearer and as discussed above.

Per cent coverage

In table 6.3, the percentage of inclusion of the true value of parameter β_{52} in the 95% confidence interval is presented (per cent coverage). No appreciable differences in the per cent coverage were found between the generated survival time following an Exponential- or a Weibull distribution. Differences between per cent coverage under the three conditions (MCAR, MAR and MNAR) seem small, although most per cent coverage of approximately 95 are found under MAR. With a few exceptions, we see all of the procedures perform quite well with respect to the coverage. Procedures NO-T and COX perform worst here, as they show some heavy undercoverage in some cases. Any overcoverage is not observed. Procedure LOGT shows most often approximately 95 per cent coverage. Procedures CC, XCAT and CENS perform quite good as well.

It is important to mention that imputation of missing covariate values according to the models NA and PSEUDO, is not considered Bayesianly proper [2], since these models do not take into account the uncertainty in estimating the cumulative baseline hazard and the pseudo observations, respectively. Consequently, this may result in too small standard errors and too narrow confidence intervals. Important undercoverage was not expected, as parameter uncertainty was taken into account in the imputation process by making use of the Bayesian polytomous regression model. Still, procedures NA and PSEUDO did not perform well, which may be caused by improper imputation. Overall, all procedures seem quite adequate, except for NO-T, which has shown strongly biased results. Procedure LOGT seems to perform best in this simulation study.

Table 6.2: Root mean square error of the simulation results for parameter β_{52}

	Settings				Analyses									
	α	λ	n	mis (%)	PERFECT	CC	XCAT	NO-T	T	LOGT	NA	COX	PSEUDO	CENS
Root mean square error														
MCAR	1	0.005	500	40	0.191	0.340	0.258	0.283	0.245	0.239	0.262	0.283	0.233	0.245
	1	0.005	500	60	0.191	1.011	0.323	0.424	0.319	0.327	0.377	0.410	0.325	0.312
	0.5	0.1	500	40	0.191	0.340	0.258	0.283	0.236	0.239	0.263	0.283	0.263	0.235
	0.5	0.1	500	60	0.191	1.011	0.323	0.424	0.311	0.327	0.378	0.410	0.394	0.312
	1	0.005	1000	40	0.133	0.224	0.181	0.274	0.172	0.175	0.236	0.255	0.188	0.172
	1	0.005	1000	60	0.133	0.385	0.235	0.412	0.220	0.234	0.344	0.368	0.259	0.221
	0.5	0.1	1000	40	0.133	0.224	0.181	0.274	0.174	0.175	0.236	0.255	0.247	0.185
	0.5	0.1	1000	60	0.133	0.385	0.235	0.412	0.228	0.234	0.344	0.368	0.362	0.253
MAR	1	0.005	500	40		0.243	0.220	0.225	0.212	0.211	0.217	0.229	0.209	0.213
	1	0.005	500	60		0.257	0.230	0.277	0.225	0.228	0.261	0.296	0.237	0.227
	0.5	0.1	500	40		0.243	0.220	0.225	0.210	0.211	0.217	0.229	0.216	0.212
	0.5	0.1	500	60		0.257	0.230	0.277	0.229	0.228	0.261	0.296	0.264	0.239
	1	0.005	1000	40		0.164	0.150	0.185	0.148	0.151	0.171	0.184	0.156	0.148
	1	0.005	1000	60		0.169	0.153	0.243	0.158	0.161	0.218	0.247	0.185	0.161
	0.5	0.1	1000	40		0.164	0.150	0.185	0.151	0.151	0.171	0.184	0.177	0.156
	0.5	0.1	1000	60		0.169	0.153	0.243	0.169	0.161	0.217	0.247	0.23	0.186
MNAR	1	0.005	500	40		0.245	0.219	0.221	0.218	0.217	0.217	0.214	0.212	0.217
	1	0.005	500	60		0.257	0.227	0.249	0.224	0.226	0.234	0.229	0.225	0.226
	0.5	0.1	500	40		0.245	0.219	0.221	0.216	0.217	0.217	0.214	0.216	0.215
	0.5	0.1	500	60		0.257	0.227	0.249	0.224	0.226	0.234	0.229	0.240	0.227
	1	0.005	1000	40		0.164	0.146	0.186	0.155	0.155	0.173	0.173	0.161	0.155
	1	0.005	1000	60		0.169	0.152	0.217	0.158	0.159	0.194	0.191	0.172	0.158
	0.5	0.1	1000	40		0.164	0.146	0.186	0.158	0.155	0.173	0.173	0.180	0.161
	0.5	0.1	1000	60		0.169	0.152	0.217	0.163	0.159	0.194	0.191	0.204	0.171

Table 6.3: Per cent coverage of the simulation results for parameter β_{52}

	Settings				Analyses									
	α	λ	n	mis (%)	PERFECT	CC	XCAT	NO-T	T	LOGT	NA	COX	PSEUDO	CENS
Per cent coverage														
MCAR	1	0.005	500	40	95.0	93.5	94.0	93.6	93.1	95.0	94.0	91.3	94.7	94.4
	1	0.005	500	60	95.0	92.7	92.3	79.6	90.1	89.6	83.6	72.8	88.4	91.8
	0.5	0.1	500	40	95.0	93.5	94.0	93.6	95.0	95.2	94.2	91.3	94.0	94.9
	0.5	0.1	500	60	95.0	92.7	92.3	79.6	89.7	89.4	83.2	72.8	80.8	92.4
	1	0.005	1000	40	94.8	95.0	92.4	79.1	94.0	95.0	84.5	80.1	92.1	94.9
	1	0.005	1000	60	94.8	92.8	90.6	32.7	88.7	87.5	64.3	47.3	82.2	90.8
	0.5	0.1	1000	40	94.8	95.0	92.4	79.1	93.0	95.0	84.5	80.1	82.7	93.2
	0.5	0.1	1000	60	94.8	92.8	90.6	32.7	87.5	87.6	64.5	47.3	53.1	86.6
MAR	1	0.005	500	40		94.5	93.5	96.1	95.9	95.8	95.8	94.3	95.7	95.4
	1	0.005	500	60		94.2	94.4	91.8	94.2	95.1	93.0	88.5	93.8	95.2
	0.5	0.1	500	40		94.5	93.5	96.1	95.6	95.8	95.7	94.3	95.7	95.4
	0.5	0.1	500	60		94.2	94.4	91.8	94.5	95.1	93.2	88.5	93.5	94.7
	1	0.005	1000	40		94.8	94.9	92.4	95.2	95.9	93.6	91.1	94.8	95.8
	1	0.005	1000	60		96.1	95.4	83.4	95.3	95.9	87.4	79.2	91.7	95.4
	0.5	0.1	1000	40		94.8	94.9	92.4	94.6	95.9	93.4	91.1	93.3	95.7
	0.5	0.1	1000	60		96.1	95.4	83.4	92.8	95.9	87.0	79.2	84.8	91.8
MNAR	1	0.005	500	40		93.8	93.6	94.3	93.7	94.7	94.4	94.8	94.5	94.2
	1	0.005	500	60		94.3	94.2	92.9	92.9	93.8	93.2	93.1	94.1	93.3
	0.5	0.1	500	40		93.8	93.6	94.3	93.4	94.7	94.4	94.8	94.6	94.1
	0.5	0.1	500	60		94.3	94.2	92.9	92.8	93.8	93.2	93.1	92.8	93.6
	1	0.005	1000	40		94.0	95.3	88.9	93.3	92.8	91.4	90.8	92.3	93.2
	1	0.005	1000	60		95.5	95.6	84.6	93.4	94.1	88.8	89.0	91.4	93.3
	0.5	0.1	1000	40		94.0	95.3	88.9	92.5	92.8	91.5	90.8	89.7	91.6
	0.5	0.1	1000	60		95.5	95.6	84.6	92.0	94.1	88.7	89.0	86.8	91.2

7. Competing risks

Throughout this study, we have spoken of the observed event or censoring time T as time to relapse or death (whichever occurs first). In other datasets, T may be the observed time to another type of event, such as time to death or time to infection. Another possibility is competing risks, where each individual may fail due to one of the K ($K \geq 2$) causes. The problem of competing risks occurs in the EBMT data as well. In the EBMT data the two competing risks for treatment failure are cumulative incidence of relapse and non-relapse mortality. Whenever an individual experiences one of the events, this precludes us from observing the other event on this person [20].

Analyzing survival data with competing risks requires a different approach. Taking the EBMT data as an example, there are two competing risks. Let X_k, with $k = 1, 2$, be the survival time to occurrence of the ith competing risk. The observed event time for each patient is the time at which the patient fails from any cause $T = Min(X_1, \ldots, X_p)$. The survival status indicator δ shows from which of the two causes the patient has failed; $\delta = k$ if $T = X_k$. The most essential competing risks parameter is the cause-specific hazard rate $h_k(t)$, which indicates the rate at which patients who are still at risk are experiencing the kth competing cause of failure. The overall hazard rate $h_T(t)$ of the observed time to event T is simply the sum of the two cause specific hazard rates [20].

This chapter presents the procedures to be used for handling missing covariate values in case of competing risks and discusses the results of a simulation study where the presented procedures are applied to the generated simulation data.

7.1 Handling missing covariate values in the context of competing risks

In Chapter 5, a complete overview is given of several manners of handling missing covariate values with censored data. Some of these procedures are still useful in the context of competing risks. Such procedures are complete cases analysis (CC), creating an extra category (XCAT) and multiple imputation without inclusion of the survival outcome (NO-T) and are described in sections 5.1-5.3. In the context of competing risks, the described procedures of multiple imputation including the survival outcome (T) and multiple imputation including the log of the survival outcome (LOGT) only differ from before with respect to the survival outcome. The survival outcome now consists of δ and T and is included in the imputation model as such. As the status indicator δ is no longer binary ($K \geq 2$), it is important to treat this covariate as a factor in the imputation model.

The procedures of multiple imputation including the cumulative baseline hazard (NA; COX), slightly differ from before. In the case of two competing risks, the cumulative baseline hazard is approximated for $\delta = 1$ and for $\delta = 2$ separately, so that two cumulative baseline hazards, $\widehat{H}_{10}(t)$ and $\widehat{H}_{20}(t)$ respectively, are included in the imputation model. Extending the proof from White & Royston (section 5.6), motivates the inclusion of two cumulative baseline hazards in the imputation model in case of competing risks. In the context of competing risks, equation (5.3) remains valid, with

$$L(T, \delta | X = 1, Z) = h_1(T|X,Z)^{\delta_1} \cdot h_2(T|X,Z)^{\delta_2} \cdot \exp\big(-H(T|X,Z)\big), \qquad (8.1)$$

where δ_1 = 1 { δ =1} and δ_2 = 1 {δ =2}, and

$$H(T|X,Z) = H_1(T|X,Z) + H_2(T|X,Z). \tag{8.2}$$

Under the proportional hazards model for the cause-specific hazards we have, for k=1,2

$$h_k(t|X,Z) = h_{k0}(t) \cdot \exp(\beta_{Xk}X + \beta_{Zk}Z). \tag{8.3}$$

Now applying (5.3) with (8.1) - (8.3), yields

$$\begin{aligned} logit(P(X=1|T,\delta,Z)) &= \log\big(L(T,\delta|X=1,Z)\big) - \log\big(L(T,\delta|X=0,Z)\big) + \zeta_Z \\ &= \delta_1\big(\log\big(h_1(T|X=1,Z)\big) - \log\big(h_1(T|X=0,Z)\big)\big) \\ &\quad + \delta_2\big(\log\big(h_2(T|X=1,Z)\big) - \log\big(h_2(T|X=0,Z)\big)\big) \\ &\quad - \big(H(T|X=1,Z) - H(T|X=0,Z)\big) + \zeta_Z \\ &= \delta_1\big(log\big(h_{10}(T)\big) + \beta_{X1} + \beta_{Z1}Z - log\big(h_{10}(T)\big) - \beta_{Z1}Z\big) \\ &\quad + \delta_2\big(log\big(h_{20}(T)\big) + \beta_{X2} + \beta_{Z2}Z - log\big(h_{20}(T)\big) - \beta_{Z2}Z\big) \\ &\quad - (H_{10}(T)\exp(\beta_{X1} + \beta_{Z1}Z) - H_{10}(T)\exp(\beta_{Z1}Z)) \\ &\quad - (H_{20}(T)\exp(\beta_{X2} + \beta_{Z2}Z) - H_{20}(T)\exp(\beta_{Z2}Z)) + \zeta_Z \\ &= \delta_1\beta_{X1} + \delta_2\beta_{X2} - H_{10}(T)\cdot\exp(\beta_{Z1}Z)\cdot(\exp(\beta_{X1}) - 1) \\ &\quad - H_{20}(T)\cdot\exp(\beta_{Z2}Z)\cdot(\exp(\beta_{X2}) - 1) + \zeta_Z \end{aligned}$$

As in the ordinary survival case, without additional covariates Z, this is a logistic regression with δ_1, δ_2, $H_{10}(T)$ and $H_{20}(T)$ as covariates. With one categorical covariate Z, this is a logistic regression with δ_1, D_2, $H_{10}(T)$ and $H_{20}(T)$ and the interactions between $H_{10}(T)$ and Z and between $H_{20}(T)$ and Z, as covariates. The two procedures related to the above motivation (sections 5.6.1-5.6.2) stay intact for the greater part. The main difference lies in the fact that there are now two cumulative baseline hazards to be estimated. The imputation model of the procedure of approximating the cumulative baseline hazard by the Nelson-Aalen estimator (NA) changes into

$$Z_{inc} \sim \beta_0 + \sum \beta_j Z_j + \delta + \hat{H}_1(T) + \hat{H}_2(T),$$

where the sum is over the remaining covariates. The imputation model of the procedure of estimating the cumulative baseline hazard by a Cox proportional hazards model (COX) changes into

$$Z_{inc} \sim \beta_0 + \sum \beta_j Z_j + \delta + \hat{H}_{10}(T) + \hat{H}_{20}(T),$$

where the sum is over the remaining covariates.

7.2 Simulation study

7.2.1 Generating data

Datasets were randomly generated resembling the EBMT data in the context of competing risks. The generated data still consists of the survival time T, the status indicator δ (now with δ = 0,1,2) and five covariates; *year* (Z_1)*, age* (Z_2)*, country* (Z_3), *staging* (Z_4) and *cytogenetics* (Z_5). The five covariates are generated in the exact same way as earlier described in section 6.1.1. The cause specific hazard of relapse equals $h_1(t) = \alpha\, \lambda_1 t^{\alpha-1}$, with

$$\lambda_1 = \lambda_{01} \exp(\beta_{11}Z_1 + \beta_{12}Z_2 + \beta_{13}Z_3 + \beta_{141}\cdot I(Z_4=1) + \beta_{142}\cdot I(Z_4=2) + \beta_{143}\cdot I(Z_4=3) + \beta_{151}\cdot I(Z_5=1) + \beta_{152}\cdot I(Z_5=2),$$

and with (λ_{01}, α) = (0.00125, 1) or (0.025, 0.5). The cause specific hazard of non-relapse mortality equals $h_2(t) = \alpha\, \lambda_2 t^{\alpha-1}$, with

$$\lambda_2 = \lambda_{02} \exp(\beta_{21}Z_1 + \beta_{22}Z_2 + \beta_{23}Z_3 + \beta_{241}\cdot I(Z_4=1) + \beta_{242}\cdot I(Z_4=2) + \beta_{243}\cdot I(Z_4=3) + \beta_{251}\cdot I(Z_5=1) + \beta_{252}\cdot I(Z_5=2),$$

and with (λ_{02}, α) = (0.00375, 1) or (0.075, 0.5). Since the two hazards have the same shape parameter α, the observed failure time T was drawn from a Weibull distribution with hazard $h_T(t) = h_1(t) + h_2(t) = \alpha\,(\lambda_1 + \lambda_2)\, T^{\alpha-1}$. Random censoring times were drawn from a Weibull distribution with scale and shape parameters (λ_C, α) = (0.005, 1) or (0.1, 0.5) and hazard $h_C(t) = \alpha\, \lambda_C\, t^{\alpha-1}$. Values of the β_1's and β_2's are mimicking the regression coefficients obtained by the cause specific Cox proportional hazards on the EBMT data. For failure cause 1 (relapse): β_{11}=-0.02, β_{12}=0.02, β_{13}=0.8, β_{141}=0.5, β_{142}=0.6, β_{143}=1.1, β_{151}=0.3, β_{152}=1. For failure cause 2 (non-relapse mortality): β_{21}=-0.03, β_{22}=-0.001, β_{23}=0.8, β_{241}=-0.4, β_{242}=0.3, β_{243}=0.2, β_{251}=0.1, β_{252}=0.4. This parameterization corresponds to 35% censoring. Taking the minimum between the generated survival times and the generated censoring times, results in the observed event or censoring time T for each individual. For each non-censored individual with observed time t^*, the probability of experiencing failure cause 1 (relapse) is $\frac{h_1(t^*)}{h_1(t^*)+ h_2(t^*)}$. The status indicator δ = 0 in case of censoring time, δ = 1 in case of relapse and δ = 2 in case of non-relapse mortality. The sample size of the data is set to n = 500. Missing values are incorporated under the assumptions of MCAR, MAR and MNAR, just as described in section 6.1.2.

7.2.2 Imputation of missing covariate values in the context of competing risks

After the missing values are incorporated in the generated datasets, the seven procedures described in section 7.1 are applied to the incomplete datasets with competing risks under the assumption of all three groups of the missing data mechanism of Rubin [10]. Similar to before, this implies that each procedure is applied to all three types of generated datasets (MCAR, MAR and MNAR). The first two procedures for handling survival data with missing covariate values described in the first two sections of Chapter 5, *Complete case analysis* (CC) and *Creating an extra category* (XCAT), do not use multiple

imputation and are therefore left out of consideration in this section. The remaining five procedures described, all impute the missing covariate values with MICE. The survival outcome plays a different role in each of these procedures. For comparison, the simulated datasets are also analyzed before introducing missing values (later on referred to as PERFECT).

Generated covariates *cytogenetics* and *staging* are both categorical with more than two categories and have missing values. Polytomous regression is used by default here in the simulation study for imputation of *cytogenetics* and *staging*. Each time the datasets are multiply imputed, there are m = 10 imputed datasets constructed, with a burn-in of 15 iterations. After the first step of multiple imputation is completed, the 10 imputed datasets are analyzed separately by means of a cause specific Cox regression model. The results are combined using Rubin's rules [1]. After completing the simulation, the results are explored by calculating the RMSE and the per cent coverage for each procedure and both competing risks, as described in paragraph 6.1.4. An overview of the five imputation procedures is given in table 7.1 briefly below, where Z_{inc} denotes the incomplete covariate and Z denotes the covariates included as predictors in the imputation model for imputation of Z_{inc}.

The last two procedures described in chapter five, *Multiple imputation including pseudo observations* (PSEUDO) and *Multiple imputation including imputed censored observations* (CENS) are not included in our simulation study in the context of competing risks. Similar to procedures T and LOGT, these procedures would perform best when the two observed failure times T_1 and T_2 are available in the context of two competing risks. However, in the motivating EBMT data only one T is observed, which is defined as the time from transplantation to relapse (cause 1) or non-relapse-mortality (cause 2), whichever occurred first. As there is only one observed failure time variable T available, the relatively new proposed procedures PSEUDO and CENS are not expected to perform very well and since they are both very computationally demanding, they are therefore omitted from the simulation study. We also expect procedures T and LOGT to perform worse than when both observed failure times T_1 and T_2 would be observed. However, these two procedures are less complicated, less computationally demanding and, most importantly, they are widely used for imputation of missing covariates already. Therefore, we see added value for including T and LOGT in the simulation study and comparing their performance with the other procedures.

Table 7.1 Imputation of missing covariate values in the context of competing risks

Abbreviation	Description
NO-T	Polytomous regression of Z_{inc} on Z
T	Polytomous regression of Z_{inc} on Z, T and δ *
LOGT	Polytomous regression of Z_{inc} on Z, $\log(T)$ and δ *
NA	Polytomous Regression of Z_{inc} on Z, $\widehat{H}_1(T)$, $\widehat{H}_2(T)$ and δ *
COX	Polytomous Regression of Z_{inc} on Z, $\widehat{H}_{01}(T)$, $\widehat{H}_{02}(T)$ and δ *

* Status indicator δ is always categorical, so two dummies δ_1 and δ_2 are included in the imputation model

7.3 Results simulation study

Results of the simulation study in the context of competing risks concerning the RMSE and the per cent coverage are presented in table 7.2. These results are only presented for the third category of *cytogenetics* (the very poor risk patients) on failure cause 1 (relapse), as this category has one of the strongest effect on the cumulative incidence of relapse in the EBMT data, and since relapse is

considered the most important cause of failure by the medical researchers. Since the simulation data was generated resembling the EBMT data, the corresponding parameter for the very poor risk patients was given an effect of β_{152}=1, similar to the effect the very poor risk patients have on the cumulative incidence of relapse in the EBMT data after complete cases analysis. For parameter β_{152} on failure cause 2 (non-relapse mortality), similar results were found as on failure cause 1. Results for other parameters are also found to be quite similar, but less outspoken. Parameters with a weaker effect on relapse free survival show a smaller RMSE (results not shown), but the proportion between the seven methods is similar to the proportion between the seven methods for parameter β_{152}.

Root mean square error

The RMSE for the different methods in estimating the parameter β_{152} is presented in table 7.2 below. The table shows that with similar sample size and percentage of missing values no appreciable differences in RMSE were found between the generated survival time following an Exponential- or a Weibull distribution. As expected, differences in RMSE are clearly found between data with a large percentage of missing values in both *staging* and *cytogenetics* (40%) and data with a severe percentage of missing values (60%). Data with a large percentage of missing values show a smaller RMSE than data with a severe percentage of missing values in both covariates. This result is due to less certainty on the real values with severe missingness than with a smaller percentage of missing values in the data.

Comparing the procedures all together, we see procedures XCAT, T and LOGT perform best under all three conditions (MCAR, MAR and MNAR), with respect to the RMSE. Procedures CC, NO-T and COX seem to perform poorly. The high values of the RMSE for CC are most likely caused by the smaller effective sample size (incomplete cases are left out). For NO-T and COX the high values of the RMSE are mainly caused by the strong bias towards the null of the procedures. As mentioned before, in procedure NO-T the missing covariate values are imputed without taking the survival outcome into account. We therefore already expected that after multiple imputation, the regression coefficients of the missing covariates would be attenuated towards zero. Contrary to expectation, the bias of CC and XCAT was quite modest, also under MAR and MNAR. Surprisingly, all procedures seem to perform better under MAR/MNAR data than under MCAR data, as the RMSE seems smaller under MAR/MNAR data than under MCAR data. This was probably caused by the percentage of complete cases that was left over in the data after creating missing values. The percentage of complete cases was 36 percent under MCAR and 39 percent under MAR/MNAR when creating 40 percent missing values in two covariates and it was 16 percent under MCAR and 26 percent under MAR/MNAR when creating 60 percent missing values in two covariates.

Per cent coverage

In the lower half of table 7.2, the percentage of inclusion of the true value of parameter β_{152} in the 95% confidence interval is presented (per cent coverage). No appreciable differences in the per cent coverage were found between the generated survival time following an Exponential- or a Weibull distribution. Differences between per cent coverage under the three conditions (MCAR, MAR and MNAR) seem small, although most per cent coverage of approximately 95 are found under MAR and MNAR. We see all of the procedures perform quite well with respect to the coverage. Procedures CC, NO-T and LOGT perform best, as they show most often approximately 95 per cent coverage. Any overcoverage or heavy undercoverage was not observed.

As discussed in the end of paragraph 6.3, the imputation of missing covariate values according to the model NA is not considered Bayesianly proper [2], since the model does not take into account the uncertainty in estimating the cumulative baseline hazard. Consequently, this may result in too small standard errors and too narrow confidence intervals. As parameter uncertainty was taken into account in the imputation process by making use of the Bayesian polytomous regression model, important undercoverage was not expected and also not found in the results. Overall, procedure LOGT performed best in this simulation study in the context of competing risks. Procedures XCAT and T both performed good as well.

Table 7.2: Simulation results of the cumulative incidence of relapse for parameter β_{152}

	Settings				Analyses							
	α	λ	n	mis (%)	PERFECT	CC	XCAT	NO-T	T	LOGT	NA	COX
Root mean square error												
MCAR	1	0.005	500	40	0.266	0.494	0.374	0.395	0.361	0.358	0.380	0.443
	1	0.005	500	60	0.266	2.698	0.468	0.562	0.466	0.471	0.498	0.636
	0.5	0.1	500	40	0.265	0.490	0.370	0.396	0.353	0.357	0.381	0.439
	0.5	0.1	500	60	0.265	2.698	0.468	0.562	0.463	0.471	0.501	0.636
MAR	1	0.005	500	40		0.345	0.301	0.304	0.306	0.302	0.304	0.325
	1	0.005	500	60		0.361	0.320	0.372	0.327	0.326	0.347	0.443
	0.5	0.1	500	40		0.344	0.301	0.300	0.304	0.300	0.301	0.323
	0.5	0.1	500	60		0.361	0.320	0.372	0.329	0.326	0.347	0.443
MNAR	1	0.005	500	40		0.344	0.299	0.305	0.309	0.310	0.309	0.309
	1	0.005	500	60		0.358	0.315	0.340	0.320	0.320	0.326	0.348
	0.5	0.1	500	40		0.343	0.298	0.303	0.306	0.307	0.306	0.306
	0.5	0.1	500	60		0.358	0.315	0.340	0.319	0.320	0.326	0.348
Per cent coverage												
MCAR	1	0.005	500	40	95.8	94.6	91.8	95.7	93.4	95.0	92.8	92.2
	1	0.005	500	60	95.8	94.2	90.4	89.7	90.8	90.8	89.6	71.7
	0.5	0.1	500	40	96.1	95.1	92.3	95.8	94.0	95.3	92.5	92.3
	0.5	0.1	500	60	96.1	94.2	90.4	89.7	90.9	90.8	88.2	71.7
MAR	1	0.005	500	40		95.1	95.4	96.3	95.7	96.0	95.4	95.4
	1	0.005	500	60		95.6	94.6	95.0	95.0	94.3	94.4	89.0
	0.5	0.1	500	40		95.2	95.4	96.5	95.6	96.2	95.5	95.8
	0.5	0.1	500	60		95.6	94.6	95.0	94.6	94.3	94.6	89.0
MNAR	1	0.005	500	40		94.4	95.2	95.6	93.8	94.5	94.6	95.9
	1	0.005	500	60		95.7	94.9	95.4	94.2	94.5	95.3	95.2
	0.5	0.1	500	40		94.6	95.5	95.7	94.8	94.9	94.8	96.0
	0.5	0.1	500	60		95.7	94.9	95.4	94.5	94.5	95.1	95.2

8. Application to the EBMT data

The simulation study described in chapters 6 and 7, compared 9 and 7 procedures for handling missing covariate values in survival and competing risks data, respectively. The different imputation procedures, complete cases analysis and the added extra category method are now applied to the motivating EBMT data. A short description of this dataset was given in chapter 3. The analysis model of interest for the EBMT data is a multivariate Cox proportional hazards model, including the covariates s*taging*, y*ear*, a*ge* and c*ytogenetics*. Imputation was performed on the incomplete categorical covariates s*taging* and c*ytogenetics* and the survival outcomes were included in the imputation models appropriate to each of the procedures. For imputation, m = 10 imputed datasets were created with MICE, with a burn-in of 15 iterations. In table 8.1, the pooled results of the Cox proportional hazards models on the data with each of the procedures are presented.

Starting with the differences between the complete cases analysis and the multiple imputation procedures, we see the estimated coefficients $\hat{\beta}$ of the complete cases are usually higher or equal to the estimated coefficients of the imputation procedures (except for poor risk cytogenetic patients, where it is lower). For patients having RAEB(t)/tAL/sMDS/CM ML untreated, the coefficient is much higher for the complete cases analysis than for the multiple imputation procedures. The standard errors of the complete cases are always larger or equal to the ones corresponding to the imputation procedures, which is as expected. The estimated coefficients for XCAT are also somewhat high compared to the estimated coefficients of the multiple imputation procedures, while there are no distinct differences between the standard errors seen. Remarkable are the estimated coefficients and standards errors for both variables y*ear* and a*ge*, which are equal for all nine methods. Turning to comparisons between imputation procedures, the main differences are seen for NO-T compared to the other multiple imputation procedures. Especially for patients having RAEB(t)/tAL/sMDS/CM ML not in CR, poor risk patients and very poor risk patients we see the estimated coefficients for NO-T are much lower than the other estimated coefficients.

From the simulation study we have seen that the multiple imputation procedure including the log of the survival outcome, LOGT, was the best procedure for handling missing covariate values on data resembling the EBMT data, where the missing covariate values are MAR/MNAR. Therefore, we might conclude that this is the best procedure for handling the missing covariate values for the EBMT data. We continue our analysis of the data, after the data is imputed according to the multiple imputation procedure including the log of the survival outcome. The hazard ratios and corresponding 95% confidence intervals and p-values are given in table 8.2.

From the table we see that very poor risk cytogenetic patients (HR=1.85) have a significant higher risk of relapse or death than very good/normal/intermediate risk cytogenetic patients. The stage of the disease is a significant predictor when comparing patients with RAEB(t)/tAL/sMDS/CM ML untreated (HR=1.37) and patients with RAEB(t)/tAL/sMDS/CM ML not in CR (HR=1.66) with RA/RARS/del5q/RC DM-RS untreated. The first two groups of patients mentioned have a higher risk of relapse or death than the latter group of patients. Older patient age at transplantation was a negative significant predictor for relapse free survival (HR=1.01) and more recent year of transplantation was a positive significant predictor for relapse free survival (HR=0.98).

Table 8.1 Results of application to the EBMT data. Tabulated values are $\hat{\beta}$ *(standard error)*

	(n = 437)	(n = 1354)	Imputation procedures (n = 1354)						
Variables	CC	XCAT	NO-T	T	LOGT	NA	COX	PSEUDO	CENS
Cytogenetics									
Very good/ normal/ intermediate									
Poor	0.20	0.31	0.20	0.27	0.27	0.27	0.30	0.30	0.29
	(0.12)	*(0.10)*	*(0.09)*	*(0.09)*	*(0.10)*	*(0.11)*	*(0.09)*	*(0.10)*	*(0.09)*
Very poor	0.74	0.69	0.46	0.65	0.61	0.67	0.62	0.58	0.65
	(0.16)	*(0.15)*	*(0.15)*	*(0.13)*	*(0.15)*	*(0.14)*	*(0.17)*	*(0.15)*	*(0.14)*
Staging									
RA/RARS/del5q/RC DM-RS untreated									
RAEB(t)/tAL/sMDS/CM ML in CR	0.06	0.03	0.02	0.03	0.01	0.01	0.01	0.05	0.03
	(0.13)	*(0.11)*	*(0.10)*	*(0.11)*	*(0.10)*	*(0.10)*	*(0.12)*	*(0.11)*	*(0.13)*
RAEB(t)/tAL/sMDS/CM ML untreated	0.44	0.28	0.23	0.30	0.31	0.32	0.31	0.29	0.24
	(0.18)	*(0.15)*	*(0.15)*	*(0.15)*	*(0.16)*	*(0.15)*	*(0.14)*	*(0.15)*	*(0.15)*
RAEB(t)/tAL/sMDS/CM ML not in CR	0.62	0.53	0.38	0.50	0.51	0.51	0.50	0.49	0.47
	(0.12)	*(0.10)*	*(0.10)*	*(0.10)*	*(0.09)*	*(0.10)*	*(0.10)*	*(0.09)*	*(0.09)*
Age *(per 10 years)*	0.10	0.11	0.10	0.10	0.10	0.11	0.11	0.11	0.11
	(0.04)	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*
Year *(per 10 years)*	-0.21	-0.19	-0.17	-0.18	-0.17	-0.18	-0.17	-0.17	-0.19
	(0.09)	*(0.06)*	*(0.06)*	*(0.06)*	*(0.06)*	*(0.06)*	*(0.06)*	*(0.06)*	*(0.06)*

Table 8.2 Cox regression model for relapse-free survival with procedure T (n = 1354)

	Relapse-free survival		
	HR	95% CI	P-value
Cytogenetics*			
Very good/ normal/ intermediate	1.00		
Poor	1.31	1.11-1.50	0.011
Very poor	1.85	1.55-2.14	<0.001
Staging*			
RA/RARS/del5q/RC DM-RS untreated	1.00		
RAEB(t)/tAL/sMDS/CM ML in CR	1.01	0.81-1.20	0.948
RAEB(t)/tAL/sMDS/CM ML untreated	1.37	1.05-1.68	0.054
RAEB(t)/tAL/sMDS/CM ML not in CR	1.66	1.49-1.84	<0.001
Age *(per 10 years)*	1.10	1.05-1.17	0.001
Year *(per 10 years)*	0. 85	0.73-0.96	0.008

* Compared to the first group (reference category)

8.1 Competing risks

The simulation study described in chapter 7, compared five imputation procedures of imputing missing covariate values in survival data in the context of competing risks. These different imputation procedures, together with complete cases and the added extra category method are now applied to the motivating data in a competing risks context. In this dataset, patients may either experience relapse (cause 1) or non-relapse mortality (cause 2). A short description of this dataset and the competing risks was given in chapter 2. The analysis model of interest for the EBMT data with competing risks is a multivariate cause-specific Cox proportional hazards model, including the covariates s*taging*, y*ear*, *age* and *cytogenetics*. Imputation was performed on the incomplete categorical covariates *staging* and *cytogenetics* and the survival outcomes were included in the imputation model appropriate to each of the procedures. For imputation, m = 10 imputed datasets were created with MICE, with a burn-in of 15 iterations. In table 8.3, the pooled results of the cause-specific Cox model are presented.

What stands out from the results is that in case of cumulative incidence of relapse, the estimated coefficients $\hat{\beta}$ are usually closer to zero when applying imputation procedure NO-T than any other procedure, but this difference seems less abundant in case of non-relapse mortality. Remarkable are the estimated coefficients and standards errors for both variables *year* and *age* at both causes, as they are almost equal for all seven procedures, except for procedure CC and in some cases XCAT. Furthermore, the coefficients of the complete case analysis and the extra category are often somewhat high with respect to the other procedures. The standard errors of the complete cases are always larger or equal to the ones corresponding to the imputation procedures, which is as expected. Remarkable is the fact that in case of non-relapse mortality, the estimated coefficients are negative for patients with stage RAEB(t)/tAL/sMDS/CM ML in CR, which is contradictory to the cumulative incidence of relapse and the relapse free survival.

From the simulation study we have seen that the multiple imputation procedure including the log of the survival outcome, LOGT, was the best procedure for handling missing covariate values on MAR/MNAR data resembling the EBMT data in context of competing risks. Therefore, for analyzing the EBMT data with competing risks, we further only consider this procedure. We continue our

analysis after imputation according to the multiple imputation procedure including the log of the survival outcome. The hazard rates and corresponding 95% confidence intervals and p-values for both competing risks are given in table 8.4.

From the table we see significant predictors for the cumulative incidence of relapse are poor risk cytogenetic patients (HR=1.48) and very poor risk cytogenetic patients (HR=2.20) when comparing with very good/normal/intermediate patients. The first two groups of patients mentioned have a higher risk of relapse than the latter group of patients. The stage of the disease is a significant predictor when comparing patients with RAEB(t)/tAL/sMDS/CM ML in CR (HR=1.83), patients with RAEB(t)/tAL/sMDS/CM ML untreated (HR=1.73) and patients with RAEB(t)/tAL/sMDS/CM ML not in CR (HR=2.63) with RA/RARS/del5q/RC DM-RS untreated. All three groups of patients have a higher risk of relapse than the patients with RA/RARS/del5q/RC DM-RS untreated. Patients who are older at the time of transplantation have a higher risk of relapse than patients who are younger at the time of transplantation (HR=1.15). Contrary to relapse free survival in the previous section, the year of transplantation is not a significant predictor for cumulative incidence of relapse.

For non-relapse mortality, cytogenetics is not a significant predictor. The stage of the disease is a significant predictor when comparing patients with RAEB(t)/tAL/sMDS/CM ML in CR (HR=0.60) with RA/RARS/del5q/RC DM-RS untreated. Meaning that, patients with RAEB(t)/tAL/sMDS/CM ML in CR have a higher non-relapse death risk, compared to patients with stage RA/RARS/del5q/RC DM-RS untreated. A more recent year of transplantation is a positive significant predictor for non-relapse mortality (HR=0.74), meaning that patients who underwent transplantation in a more recent year have lower non-relapse death risk than patients who underwent transplantation in an earlier calendar year.

Table 8.3 Results of application to the EBMT data. Tabulated values are $\hat{\beta}$ *(standard error)*.

	(n = 437)	(n = 1354)	Imputation procedures (n = 1354)				
Variable	CC	XCAT	NO-T	T	LOGT	NA	COX
Cumulative incidence of relapse							
Cytogenetics							
Very good/ normal/ intermediate							
Poor	0.35 *(0.17)*	0.44 *(0.14)*	0.29 *(0.13)*	0.39 *(0.13)*	0.43 *(0.15)*	0.42 *(0.15)*	0.33 *(0.12)*
Very poor	1.02 *(0.21)*	0.89 *(0.19)*	0.63 *(0.19)*	0.79 *(0.20)*	0.81 *(0.19)*	0.85 *(0.18)*	0.73 *(0.22)*
Staging							
RA/RARS/del5q/RC DM-RS untreated							
RAEB(t)/tAL/sMDS/CM ML in CR	0.55 *(0.19)*	0.56 *(0.16)*	0.36 *(0.15)*	0.60 *(0.17)*	0.53 *(0.15)*	0.56 *(0.15)*	0.36 *(0.16)*
RAEB(t)/tAL/sMDS/CM ML untreated	0.62 *(0.28)*	0.51 *(0.25)*	0.34 *(0.23)*	0.55 *(0.22)*	0.54 *(0.23)*	0.56 *(0.23)*	0.44 *(0.23)*
RAEB(t)/tAL/sMDS/CM ML not in CR	1.13 *(0.18)*	1.03 *(0.15)*	0.63 *(0.15)*	0.97 *(0.16)*	0.96 *(0.16)*	0.99 *(0.14)*	0.80 *(0.14)*
Age *(per 10 years)*	0.22 *(0.06)*	0.16 *(0.05)*	0.13 *(0.05)*	0.14 *(0.05)*	0.14 *(0.05)*	0.14 *(0.05)*	0.14 *(0.05)*
Year *(per 10 years)*	-0.16 *(0.13)*	-0.07 *(0.09)*	-0.01 *(0.09)*	-0.03 *(0.09)*	-0.02 *(0.10)*	-0.03 *(0.10)*	-0.01 *(0.09)*
Non-relapse mortality							
Cytogenetics							
Very good/ normal/ intermediate							
Poor	0.07 *(0.17)*	0.19 *(0.15)*	0.12 *(0.12)*	0.15 *(0.14)*	0.15 *(0.15)*	0.15 *(0.13)*	0.22 *(0.14)*
Very poor	0.40 *(0.27)*	0.45 *(0.23)*	0.27 *(0.24)*	0.49 *(0.22)*	0.41 *(0.23)*	0.45 *(0.20)*	0.51 *(0.22)*
Staging							
RA/RARS/del5q/RC DM-RS untreated							
RAEB(t)/tAL/sMDS/CM ML in CR	-0.41 *(0.19)*	-0.45 *(0.16)*	-0.29 *(0.15)*	-0.50 *(0.15)*	-0.46 *(0.17)*	-0.48 *(0.16)*	-0.37 *(0.17)*
RAEB(t)/tAL/sMDS/CM ML untreated	0.29 *(0.24)*	0.14 *(0.20)*	0.14 *(0.19)*	0.12 *(0.19)*	0.16 *(0.22)*	0.10 *(0.19)*	0.18 *(0.19)*
RAEB(t)/tAL/sMDS/CM ML not in CR	0.16 *(0.17)*	0.15 *(0.13)*	0.17 *(0.13)*	0.13 *(0.12)*	0.16 *(0.12)*	0.13 *(0.12)*	0.19 *(0.13)*
Age *(per 10 years)*	-0.01 *(0.06)*	0.07 *(0.04)*	0.08 *(0.04)*	0.08 *(0.04)*	0.08 *(0.04)*	0.08 *(0.04)*	0.08 *(0.04)*
Year *(per 10 years)*	-0.26 *(0.13)*	-0.30 *(0.09)*	-0.31 *(0.09)*	-0.31 *(0.09)*	-0.30 *(0.09)*	-0.31 *(0.09)*	-0.31 *(0.09)*

Table 8.4 Cox regression models for cause specific hazards with procedure T (n = 1354)

	Cumulative incidence of relapse			Non-relapse mortality		
	HR	95% CI	P-value	HR	95% CI	P-value
Cytogenetics*						
Very good/ normal/ intermediate	1.00			1.00		
Poor	1.48	1.19-1.78	0.005	1.17	0.87-1.46	0.297
Very poor	2.20	1.83-2.58	<0.001	1.63	1.18-2.08	0.077
Staging*						
RA/RARS/del5q/RC DM-RS untreated	1.00			1.00		
RAEB(t)/tAL/sMDS/CM ML in CR	1.83	1.54-2.12	0.001	0.60	0.27-0.94	0.011
RAEB(t)/tAL/sMDS/CM ML untreated	1.73	1.28-2.18	0.021	1.12	0.69-1.55	0. 472
RAEB(t)/tAL/sMDS/CM ML not in CR	2.63	2.32-2.94	<0.001	1.14	0.90-1.37	0.174
Age *(per 10 years)*	1.15	1.05-1.24	0.003	1.08	1.00-1.16	0.061
Year *(per 10 years)*	0.97	0.77-1.62	0.813	0.74	0.56-0.91	0.001

* Compared to the first group (reference category)

9. Discussion

In this study we have evaluated different manners of handling survival data with missing covariate values by means of a simulation study. The primary objective of this research was to find out how the survival outcome should be included in the regression model for imputation of missing covariate values. In our simulation study we compared different combinations of inclusion of the survival outcome together with complete cases analysis and the addition of an extra category. An imputation model without survival outcome and two new proposals for inclusion of the survival outcome in the imputation model were also included in the simulation. The two new proposals were pseudo observations and imputing observed event time. From the simulation study we found all procedures to perform quite adequately, except for the imputation model not including the survival outcome, due to strong biased results. The procedure of multiple imputation with chained equations and an imputation model including both the status indicator and the log of the survival time, performed best in our simulation study.

From the simulation study we have hardly found any evidence for including the cumulative baseline hazard into the imputation model, as was suggested by White & Royston [5]. They show that a suitable model for imputing a binary or normal incomplete covariate is a logistic or linear regression model on the status indicator, the cumulative baseline hazard, the other covariate and the interaction term between other covariate and the cumulative baseline hazard. This result is exact in the case of an incomplete binary covariate. In our simulation study, we imputed two incomplete categorical covariates with more than two categories with a polytomous regression model, which is actually a series of separate logistic regression models of all categories -1 against a baseline category. In their simulation study White & Royston found that including the cumulative baseline hazard performed better than including the log of the survival outcome in a univariate imputation model under MCAR. In a bivariate imputation model, however, this difference was very minimal under MCAR, while under MAR the two procedures show approximately equal performance. The different result in our simulation study and that of White & Royston on including the cumulative baseline hazard in the imputation model, may be caused by the number of extra predictors included and the inclusion of interaction terms, or the way of constructing MAR models.

A number of surprising results were found in our simulation study as well. Complete cases analysis and the procedure of adding an extra category have shown merely modestly biased results, even under MAR and MNAR. A similar result was found in the simulation study of White & Royston, where complete cases analysis has shown much less biased results under MAR than the multiple imputation procedures [5]. Furthermore, for each procedure we found a slight improvement of performance under MAR and MNAR over the performance under MCAR. This was probably caused by the percentage of complete cases that was left over in the data after creating missing values. The percentage of complete cases was 36 percent under MCAR and 39 percent under MAR/MNAR when creating 40 percent missing values in two covariates and it was 16 percent under MCAR and 26 percent under MAR/MNAR when creating 60 percent missing values in two covariates. When creating 30 percent missing values or less in two covariates, the percentage of complete cases left over in the data is more or less equal under MCAR, MAR and MNAR. In such a case, an improvement of performance under MAR and MNAR over the performance under MCAR is not expected. Furthermore, the generated missing data under MAR and MNAR may not have been specific enough to trigger important differences and the failure of complete cases analysis and the procedure of

adding an extra category, while these procedures were found to be inferior in other simulation studies.

9.1 General recommendations

Starting with non-imputation procedures, complete case analysis and adding an extra category did not perform best in our simulation study and the general idea is that these procedures are inefficient. Our results support this view both for ordinary survival data and in the context of competing risks. The procedure of multiple imputation without including the survival outcome has shown poor performance, mainly caused by a strong attenuation of the results towards the null. One of the newly proposed procedures in case of ordinary survival data, multiple imputation including pseudo observations, did not perform well when we consider the RMSE and the per cent coverage. Adjustments may be made to improve this procedure, but in the current form, it is not recommended for usage in practice.

Procedures of multiple imputation including the survival outcome and multiple imputation including the imputed observed event time have shown to perform quite adequately. However, the latter procedure is very computationally demanding and as it does not outperform the much simpler procedures, applying this procedure to your data may not be worth the computational burden. From our results we have seen that multiple imputation including the log of the survival outcome performs best in both ordinary survival data and in the context of competing risks. This procedure was found to perform well in the literature and even though more recent improvements have been suggested, this procedure is widely used by many authors. Basing our thoughts both on our own simulation results and on other studies, we come to the conclusion that the procedure of multiple imputation including the log of the survival outcome performs well and could be recommended for imputing missing covariate values with survival data and in the context of competing risks.

9.2 Suggestions for further research

In our simulation study we have purely focussed on the regression coefficients for comparison between the different procedures for handling missing covariate values, but other scientific quantities of interest could be evaluated as well. Quantities such as the baseline hazards or the estimated survival probabilities. An estimate of the baseline hazard function could be desired in several applications, for instance for model-based prediction of survival probabilities in independent data. It might be interesting to evaluate the estimation of these two quantities (as well as others), with different procedures for handling missing covariate values under the different missing data mechanisms. Thereby, differences in performance of the procedures for estimation of several quantities may be detected as well.

As mentioned in the beginning of this chapter, the different result in our simulation study and that of White & Royston on including the cumulative baseline hazard in the imputation model, may be caused by the number of extra predictors included and the inclusion of interaction terms. In our simulation study three or four extra predictors were included in the imputation model. Interaction terms between the extra predictors and the cumulative baseline hazard were not considered, as this would have made the imputation model much more complicated and since White & Royston found

in their simulation study that including an interaction term in the imputation model, performed similar to not including an interaction term (both in case of only one extra predictor). However, according to the exact result proven by White & Royston, the interaction term needs to be included in the model. Not taking the interactions into consideration in our simulation study may have lead to an incorrect imputation model and consequently to less performance of the suggested procedures. Further research on this topic could be performed as it would be interesting to see whether including interaction terms in the imputation model in case of more than one extra predictor, improves the method's performance.

One of the advantages of multiple imputation is that even with a low number of imputed datasets (m=2), unbiased estimates with correct confidence intervals may be produced. Multiple imputation should be able to work with such a low number of imputed datasets as the between-imputation variance B is enlarged by a factor $1/m$ before calculating the total variance. Even though the classic advice was to use 3-5 imputed datasets for moderate missing data, several authors have made more recent suggestions that it could be beneficial to use 20-100 imputed datasets. As multiple imputation is a simulation technique, the estimated $\bar{Q}$ and the corresponding variance estimate are subject to simulation error [1]. One of the more recent suggestions came from White, Royston & Wood [13]. They suggest a rule of thumb that 'the number of imputations should be similar to the percentage of cases that are incomplete' [13, p.388], at least with fractions of missing information up to 0.5. Applying this rule of thumb would result in a much smaller Monte Carlo error and will therefore also provide an adequate level of reproducibility. According to this rule of thumb, our simulation study would have needed a number of m=50 imputed datasets, instead of m=10. Such a large number of imputed datasets involves much more computation and storage, which was beyond the scope of our study. However, further research might be desired, where similar procedures are compared using a much higher number of imputed datasets or at least a number according to the rule of thumb suggested by White, Royston & Wood.

When using multiple imputation, parameter uncertainty should be included in the imputation process for proper imputation, either by the Bayesian method or the Bootstrap method. The Bayesian methods draw the parameters directly from their posterior distributions and Bootstrap methods resample the observed data and re-estimate the parameters from the resampled data [1]. In our simulation study, we imputed the missing covariate values by polytomous regression, the default method for imputing categorical variables, which does not include parameter uncertainty. The easiest way of accomplishing parameter uncertainty would be to include bootstrapping at the beginning of this method. However, large differences are not expected. Besides the imputation method, two of the multiple imputation procedures used in our simulation study are not considered Bayesian proper; multiple imputation including the cumulative baseline hazard approximated by the Nelson-Aalen estimator and multiple imputation including pseudo observations. These procedures are not considered Bayesian proper since they do not take into account the uncertainty in estimating the cumulative baseline hazard [5] and the pseudo observations, respectively. Consequently, this may result in too small standard errors and too narrow confidence intervals. The procedure of multiple imputation including pseudo observations may be improved by iteratively calculating the pseudo observations during imputation. This could be done by adjusting the chosen imputation method or by writing a new one for imputation of missing covariate values. Naturally, calculating the pseudo observations iteratively requires more calculation time, but it might be worth the wait.

Bibliography

[1] S. Van Buuren, Flexible imputation of missing data, Boca Raton, Florida: Chapman & Hall/CRC, 2012.

[2] D. Rubin, Multiple imputation for nonresponse in surveys, New York: John Wiley & Sons, 1987.

[3] K. Moons, R. Donders, T. Stijnen and F. Harrell Jr, "Using the outcome for imputation of missing predictor values was preferred.," *Journal of Clinical Epidemiology,* vol. 59, pp. 1092-1101, 2006.

[4] J. Schafer, Analysis of Incomplete Multivariate Data, London: Chapman & Hall, 1997.

[5] I. White and P. Royston, "Imputing missing covariate values for the Cox model," *Statistics in Medicine,* no. 28, pp. 1982-1998, 2009.

[6] S. van Buuren, H. Boshuizen and D. Knook, "Multiple imputation of missing blood pressure covariates in survival analysis.," *Statistics in Medicine,* no. 18, pp. 681-694, 1999.

[7] T. Clark and D. Altman, "Developing a prognostic model in the presence of missing data:an ovarian cancer case study.," *Journal of Clinical Epidemiology,* no. 56, pp. 28-37, 2003.

[8] F. Barzi and M. Woodward, "Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies," *American Journal of Epidemiology,* no. 160, pp. 34-45, 2004.

[9] C. Koenecke et al, "Impact of IPSS-R cytogenetics on outcome after allogeneic stem cell transplantation for MDS and secondary AML evolving from MDS: A retrospective multicenter stufy of the EBMT," Upublished manuscript.

[10] D. Rubin, "Inference and Missing Data," *Biometrika,* vol. 63, no. 3, pp. 581-592, 1976.

[11] D. Rubin, "The design of a general flexible way system for handling non-response in sample surveys. Manuscript prepared for the U.S. Social Security Administration, July 1, 1977," *Later published in The American Statistician,* vol. 58, no. 4, pp. 298-302, 1977.

[12] D. Rubin, "Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse," *Proceedings of the Survey Research Methods Section of the American Statistical Association,* pp. 20-34, 1978.

[13] I. White, P. Royston and A. Wood, "Multiple imputation using chained equations: issues and guidance for practise," *Statistics in Medicine,* vol. 30, pp. 377-399, 2011.

[14] S. van Buuren, J. Brands, C. Groothuis-Oudshoorn and D. Rubin, "Fully conditional specification in multivariate imputation," *Journal of Statistical Computation and Simulation,* vol. 76, no. 12,

pp. 1049-1064, 2006.

[15] S. van Buuren and K. Groothuis-Oudshoorn, “mice: Multivariate Imputation by Chained Equations in R,” *Journal of Statistical Software,* vol. 45, no. 3, pp. 1-67, 2011.

[16] R. Little and D. Rubin, Statistical Analysis with Missing Data, New York: John Wiley & Sons, 2002.

[17] P. Andersen and M. Perme, “Pseudo-observations in survival analysis,” *Statistical Methods in Medical Research,* vol. 19, pp. 71-99, 2010.

[18] E. Kaplan and P. Meier, “Nonparametric estimation from incomplete observations,” *Journal of the American Statistical Association,* vol. 53, no. 282, pp. 457-481, 1958.

[19] J. Brand, Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets, Rotterdam: Erasmus University, 1999.

[20] J. Klein and M. Moeschberger, Survival Analysis. Techniques for censored and truncated data, New York: Springer, 1997.