Statistical modelling of competing risks with incomplete data: with applications to allogeneic stem cell transplantation

Bonneville, E.F.

Chapter 2

Handling missing covariate data in clinical studies in haematology

Chapter based on: **Bonneville, E. F.**, Schetelig, J., Putter, H., et al. (2023) Handling missing covariate data in clinical studies in haematology. *Best Practice & Research Clinical Haematology*, 36, 101477. DOI: 10.1016/j.beha.2023.101477

Abstract

Missing data are frequently encountered across studies in clinical haematology. Failure to handle these missing values in an appropriate manner can complicate the interpretation of a study's findings, as estimates presented may be biased and/or imprecise. In the present work, we first provide an overview of current methods for handling missing covariate data, along with their advantages and disadvantages. Furthermore, a systematic review is presented, exploring both contemporary reporting of missing values in major haematological journals, and the methods used for handling them. A principle finding was that the method of handling missing data was explicitly specified in a minority of articles (in 76 out of 195 articles reporting missing values, 39%). Among these, complete case analysis and the missing indicator method were the most common approaches to dealing with missing values, with more complex methods such as multiple imputation being extremely rare (in 7 out of 195 articles). An example analysis (with associated code) is also provided using haematopoietic stem cell transplant data, illustrating the different approaches to handling missing values. We conclude with various recommendations regarding the reporting and handling of missing values for future studies in clinical haematology.

2.1 Introduction

Missing data are widely encountered in clinical research across both covariate and outcome information. For example, a patient known to have relapsed from a particular malignant disease may be lacking information on the precise timing of the relapse (missing outcome), but may also have incomplete data on important prognostic factors that may be time-consuming or expensive to collect, or were not considered relevant at the moment of diagnosis or start of treatment (missing covariate).

While missing values generally represent a nuisance to the researcher, failure to adequately account for them may lead to results that are potentially biased and/or imprecise (Sterne *et al.*, 2009). Insufficient reporting of missing values may also complicate the interpretation of a study's results. Both the reporting of missing values and the methods used for handling them have been the subject of multiple reviews across a multitude of study types and outcome models (Bell *et al.*, 2014; Burton and Altman, 2004; Carroll *et al.*, 2020; Sullivan *et al.*, 2017).

On the subject of clinical studies in haematology, there has to the best of our knowledge been no explicit investigation into contemporary reporting and handling of missing data. Similarly, field-specific guidelines seem to be restricted to a section from Delgado *et al.* (2014), which are limited in scope and arguably outdated. The present work seeks to fill this knowledge gap, focusing mainly on missing covariate data. We will first introduce the assumptions underlying missing data, as well as the common methods used to handle them and their associated pitfalls. We follow this up with a systematic review of missing data related practices across major haematological journals, and an applied data example. We then conclude with recommendations for further practice.

2.2 Assumptions underlying methods for handling missing data

Any discussion concerning missing values in the context of a particular study should always begin with one simple question: why are the values missing? This enquiry into the possible causes of the missing values is in fact an assessment of the missing data mechanism. Little and Rubin (2019) formalised the concept, assuming that every value a priori had some probability of becoming missing, which could depend on either observed or unobserved information.

Suppose that in the context of a retrospective study, performance status (measured for example by means of the Karnofsky Performance Score, KPS) is missing for a portion of the patient cohort. These values are missing completely at random (MCAR) when all patients are equally likely to have a missing value. That is, no one subgroup defined by the data (e.g. older patients) has any comparatively higher likelihood of

having a missing KPS value. The missingness is independent from both observed and unobserved information.

In the case where observed information fully explains whether the data are missing or not, we refer to the mechanism as missing at random (MAR). For example, younger patients may have their KPS score recorded less routinely than older patients. In other words, whether or not the KPS score is observed can be explained using the observed age information. If the missing data are related to unobserved information, we refer to them as missing not at random (MNAR). An example of this would be if patients with a higher KPS score would also have a higher likelihood of having a corresponding missing value, e.g., because fitter—but not necessarily younger—patients are monitored less intensively. The missingness in this example is therefore related to the unobserved KPS values or patient fitness not measured by any other variable.

Given that MNAR data by definition depends on unobserved information, it is impossible to distinguish from MAR data based on data alone. Since the most common missing data handling methods will rely on data being MAR, it is critical to discuss how plausible the assumption is in a given context. Such a discussion should take place not only between statisticians and researchers, but also together with data managers, and other persons who possess intimate knowledge of the data collection process.

2.3 Common approaches and associated pitfalls

In what follows, we outline common approaches for handling missing values in the context of a regression model—referred to as the analysis or substantive model. We assume that the outcome of the model is observed for all individuals, and that one or more of the predictors used have missing values.

2.3.1 Complete case analysis

The simplest solution to deal with missing covariate data is to perform a complete case analysis (CCA). This excludes patients from an analysis as soon as they have an unknown value on at least one of the predictor variables. Therefore, in a multivariable analysis model with missingness spanning multiple predictors, the potential loss in statistical precision can be substantial—as reflected by wide confidence intervals. While a CCA is valid under the assumption of MCAR, it is also unbiased in various MAR and even MNAR situations (Hughes *et al.*, 2019). In turn, when missingness depends on the outcome, or observed variables explaining the missingness are not accounted for (e.g. by adjusting for them in a multivariable model), CCA will generally be biased. That is, the estimate(s) obtained using CCA will differ with respect to those that would have been obtained had the data been complete. Note also that CCA

is the (often implicit) default strategy for handling missing values in most software packages.

2.3.2 Missing indicator method

A straightforward way to exploit all available data is to use the missing indicator method (MIM). For categorical variables, this simply involves creating an additional level for the missing values. For continuous variables, the missing values are first replaced by a constant (commonly zero, or the observed average), and thereafter a binary variable indicating whether the value was missing in the first place or not is added to the model. Two major advantages of this method are that the patients with missing information are monitored in univariable analyses (e.g. checking whether those with missing data have much higher or lower overall survival compared to those with observed information), and that it retains all information for multivariable modelling. The MIM has historically been criticised, as it has been shown to potentially be biased, even in cases with data that are MCAR (Donders *et al.*, 2006). Nevertheless, it may prove to be a pragmatic solution under particular assumptions and mild confounding (Blake *et al.*, 2020).

2.3.3 Single imputation

One may also choose to replace all missing values in a variable (for example by the average among observed values) and proceed to analyse the data as if they were complete. This type of single imputation is generally discouraged as it does not take into account any of the uncertainty regarding the missing values (Sterne *et al.*, 2009). Data are analysed as if they were complete (i.e. failing to capture that each missing value could have taken a range of different values, other than for example the observed average), leading to confidence intervals that are too narrow. A less well known instance of single imputation occurs in categorical variables, where the analyst may choose to combine those individuals with missing values together with another factor level, usually the most frequent one. This may in fact also occur at a pre-processing stage, as may be the case for the calculation of the disease risk index (DRI), used in registry studies about patients undergoing allogeneic stem cell transplantation (alloSCT) (Armand *et al.*, 2014). In different implementations of the DRI, patients with myelodysplastic syndromes or acute myeloid leukemia lacking cytogenetic risk classification (a component of the DRI) are assumed to belong to (i.e. singly imputed as) the most commonly observed category, intermediate cytogenetic risk (Armand *et al.*, 2012; Saccardi *et al.*, 2023; Snowden *et al.*, 2020). In the original implementation of the DRI, this was justified on the basis of outcomes being similar between those unavailable and intermediate cytogenetics. Note also that the MIM is not a form of

single imputation: it assumes (in the case of a categorical variable), that missing values belong to a separate category, rather than to any one of the existing categories.

2.3.4 Multiple imputation

A more complex strategy to handle missing values is multiple imputation (MI). MI leverages the observed data to repeatedly replace the missing values by plausible (model-based) values. Doing so results in multiple 'complete' datasets that can separately be analysed, and whose results can be combined in a way that properly accounts for the uncertainty induced by the missing values (using so-called Rubin's rules). MI is one of multiple missing data methods that operate under the MAR assumption, and are capable of producing more precise estimates compared to CCA (White and Carlin, 2010).

There are different variants of MI, of which the best known is multivariate imputation by chained equations (MICE) (van Buuren *et al.*, 1999), implemented for example in the 'mice' R package (van Buuren and Groothuis-Oudshoorn, 2011). For each variable with missing data, an imputation model needs to be specified and fitted using the observed part of the variable, and can thereafter be used to generate imputed values. This approach allows to flexibly specify different models for different variable types, such as a logistic regression if the variable to be imputed is binary. Once these models are specified, there are two main settings to fix: the number of imputed datasets, and the number of iterations (also known as cycles). To generate a single imputed dataset, MICE will start with an initial guess for each missing value by choosing randomly from the observed values. Naturally, these initial guesses may be far from values we could have expected given a patient's other variables: for example, initial guesses for KPS scores among younger patients could be low due to the random choice, when we might have plausibly expected them to be high. In order to move in the direction of more plausible values given the observed data, we 'chain' the imputation models together: the most recently imputed values are used as predictors in the imputation model for the next variable to be imputed. One pass across all variables with missing data represents a single cycle or iteration. It is necessary to continue these cycles (i.e. imputed values at the end of the first cycle are used as starting values for the second cycle, and so forth) until 'convergence' is reached, i.e. the sequence of imputed values generated has stabilised. The imputed values at the end of the final cycle form a single imputed dataset. Independent (i.e. with different starting values) runs of these cycles give rise to multiple imputed datasets.

In studies where missing data are abundant, it is crucial to pick a sufficiently large number of imputed datasets. When the proportion of missing data is high, the imputation models are fitted on a smaller portion of data, which will lead to a larger variability in the imputed values. Intuitively, we need to make sure that conclusions made on the basis of for example 10 imputed datasets, do not substantially differ from

those that would have been made under another set of 10 imputed datasets (i.e. had we repeated the entire multiple imputation procedure). A rule of thumb is to use as many imputed datasets as the percentage of incomplete observations (e.g. 40 imputed datasets if a CCA discards 40% of observations), although recent research suggests that the number of imputed datasets should be even larger (Mertens *et al.*, 2020; von Hippel, 2020). Generally, at least 10-20 iterations are recommended, and convergence should be assessed using visual diagnostic tools, such as those described in section 6.4.2 of the text by van Buuren (van Buuren, 2018).

Another key consideration in MI is the issue of compatibility between analysis and imputation models, that is, that the imputation models should make assumptions that are consistent with those made by the proposed analysis model. Practically speaking, this means that all variables in the analysis model, including the outcome, should also be included as predictors in the imputation model. It also means that special terms, such as interactions, should be adequately handled. A version of MICE which naturally ensures compatibility between the analysis and imputation models in such situations is substantive-model-compatible fully conditional specification (SMC-FCS) (Bartlett *et al.*, 2015), implemented in the 'smcfcs' R package (Bartlett *et al.*, 2022).

A closely related approach to SMC-FCS is fully Bayesian imputation, implemented in the 'JointAI' R package (Erler *et al.*, 2021). This performs analysis and imputation simultaneously, again ensuring compatibility between the imputation and analysis models. An overview of these and related methods are found in the work of Carpenter and Smuk (Carpenter and Smuk, 2021).

Research concerning the theory and performance of different MI approaches extends to the field of survival analysis, with particular focus on the widely-used Cox proportional hazards model. When the analysis model of interest is a Cox model, White and Royston (2009) suggested that the imputation model for a partially observed variable should include as predictors the remaining covariates from the analysis model, the event indicator (indicating whether an patient experienced an event or was censored) and the Nelson–Aalen estimate of the cumulative hazard. This approach is expected to work well in settings with a low cumulative incidence, and when the true effects of the covariates are small. When this does not hold, results are expected to be biased: the imputation and analysis model are only approximately compatible. That is, this imputation model produces imputed values which are not completely consistent with data where the key assumption made by a Cox model (multiplicative covariate effects on the hazard) is assumed to hold.

The SMC-FCS approach, which addresses this compatibility issue directly, has shown superior performance (when the analysis model is well specified) compared to the standard MICE approach described in the previous paragraph across multiple simulation studies, including settings with time-varying effects of covariates (Keogh and Morris, 2018), excess hazard models (Antunes *et al.*, 2021), and competing risks (Bartlett and

Taylor, 2016; Bonneville *et al.*, 2022). While to our knowledge there has been no systematic evaluation of the fully Bayesian approach in the context of the Cox model, the expectation is that it should perform at least as well as the SMC-FCS approach. This is because it not only ensures that the imputations are consistent with the analysis model, but also that the different imputation models (when there are multiple variables to be imputed) are consistent with each other. The latter point may be of concern when there are complex non-linear relationships between covariates.

2.4 Systematic review

2.4.1 Search strategy and data extraction

We performed a systematic review to obtain a broad picture of current practices in missing data reporting and handling across research articles published in major haematological journals. We used the Ovid platform to search the MEDLINE and Embase databases for articles written in English in 2021. We excluded articles that did not contain new data analyses (such as review articles), as well as letters to the editor, since their brevity would preclude full reporting on the issues we are interested in. Meta-analyses, methodological publications, and articles that were co-authored by authors of the present work were also excluded. A total of 16 journals were selected, based on two primary criteria: a 5-year Journal Impact Factor (2021) larger than 3 (data obtained via Journal Citation Reports Science Edition, Clarivate Analytics 2018), and a journal scope focused on clinical research in haematological malignancies. The 5-year Journal Impact Factor criteria was used in order to target articles with (on average) better quality of methodology and reporting, and a larger readership.

The search terms in Ovid format are reported in Table 2.1. After narrowing down by journals and year of publication, the remaining criteria focused on the malignant disease group, and the type of analysis model used. A wide selection of malignant diseases was included: acute and chronic myeloid leukemia, acute and chronic lymphocytic leukemia, myelodysplastic syndromes (MDS), and non-Hodgkin lymphomas (NHL) . We searched for articles that used a multivariable Cox proportional hazards model as part of their statistical analysis. Models were further classified into standard Cox, Fine–Gray (competing risks), frailty, multi-state and relative survival. The standard Cox category included standard outcomes such as overall or progression-free survival, but also cause-specific Cox models for outcomes such as relapse incidence and non-relapse mortality (since these are applied by treating competing risks as censored). Search terms for both analysis model and malignant disease group were based on detection of relevant strings in either title, abstract or keywords.

For each included article, the information extraction spanned three main areas: 1) exclusion of patients at 'population selection' phase based on missing data, 2) presence

Table 2.1: Search terms used in Ovid format, as entered into MEDLINE and Embase.

1.	(0006-4971 or 2352-3026 or 1756-8722 or 2044-5385 or 0887-6924 or 0390-6078 or 0361-8609 or 2473-9529 or 0007-1048 or 2666-6367 or 0268-3369 or 2040-6207 or 0278-0232 or 0939-5555 or 1545-5009 or 1042-8194).is.
2.	(cox or HR or aHR or (hazard adj1 ratio) or hazard or (proportional adj1 hazards) or multivaria*).ti,ab,kf.
3.	((acute adj1 myeloid adj1 leukemia) or (myelodysplastic adj1 syndrome*) or (chronic adj1 lymphocytic adj1 leukemia) or non-Hodgkin* or (chronic adj1 myeloid adj1 leukemia) or (acute adj1 lymphocytic adj1 leukemia) or (multiple adj1 myeloma) or leukem* or leukaem*).ti,ab,kf.
4.	1 and 2 and 3
5.	limit 4 to ('conference review' or 'review')
6.	limit 4 to (article or article in press or journal article)
7.	6 not 5
8.	limit 7 to english language
9.	limit 8 to yr='2021 - 2021'
10.	remove duplicates from 9

and explicit reporting of missing data in baseline covariates used in one or more of the reported analysis models, and 3) explicit reporting of methods used to handle the missing data. The first part of the extraction is based on the findings in Carroll *et al.* (2020), namely that the population is occasionally filtered on the basis of information/variable availability, leaving little or no missing data at the analysis phase. Two possible examples of this are, a) retrospective analysis of data from multiple trials, and a particular trial being excluded because information on a covariate of interest was not collected; b) in a retrospective study, including only those with sequencing data at a point in time (thereby implicitly excluding those with unavailable sequencing data).

We checked whether there was any missing data in any of the covariates making up the multivariable model, whether it be reported in the descriptive 'Table 1' (hereafter referred to as the 'descriptives table'), in a figure or in the main text. If there were multiple multivariable Cox models reported in the article, we recorded whether in at least one of them there was a covariate with missing data. Note that sometimes the missing values are only implicitly reported, as is for example the case when the numbers per level of categorical variables are reported, and these fail to add up to the total.

In terms of missing data handling, we paid particular attention to the use of CCA. Specifically, when missing values are reported, authors can be explicit about use of complete cases in two main ways: specifying the number of subjects used when reporting the multivariable model, or including a sentence in-text explicitly mentioning the use of CCA (the sentence 'missing values were not imputed' is also appropriate). Otherwise, the use of CCA was considered implicit—which can be problematic since the reader is unaware of the extent of the power loss. Likewise, we also recorded

whether other methods were used, such as: MIM, single imputation, MI or other. We also recorded the software used for the analysis. The full extraction sheet is available in the online supplement.

An initial investigation was carried out by EB, LdW and HP using 10 randomly selected papers. This was done in order to assess the consistency of data extraction, sharpen the data extraction checklist and agree on how to extract information when answers were ambiguous. Data extraction was then carried out by EB.

2.4.2 Results

A total of 398 research articles were identified after eliminating duplicate records obtained via MEDLINE (n = 86) and Embase (n = 391). From those, 99 were excluded due to either co-authors of the present manuscript being involved as co-author on the publication (n = 8), meta-analyses (n = 6), no Cox models reported (n = 48) and absence of multivariable Cox model (n = 36). One article was excluded as the supplementary material (which described the predictors used in the multivariable model) was not available on the publisher's website. A total of 299 articles were therefore included in the review. The journals where these articles most frequently featured in were Bone Marrow Transplantation (n = 46), Transplantation and Cellular Therapy (n = 40), Blood Advances (n = 35), Annals of Hematology (n = 33), and Leukemia and Lymphoma (n = 28).

At population selection, 80 articles (27%) explicitly reported having excluded observations on the basis of missing information. At this stage of the extraction, the focus was not yet solely on covariates that would make part of the multivariable model(s) in a particular article. These 80 articles could thus for example comprise exclusions based on missing outcome data. Given the retrospective nature of many studies, many of these exclusions were based on lack of cytogenetic information or no minimal residual disease (MRD) assessments, as was the case (exclusion of patients without cytogenetic information) for example in Hansen *et al.* (2021). It is important to note that while this approach can seem natural, it could come at the cost of ending up with a slightly different population than the one originally targeted. A possible sanity test would be to compare the univariable outcomes (e.g. Kaplan–Meier based overall survival) between those excluded and those remaining.

The vast majority of articles (287 out of 299, 96%) included at least one standard Cox model. The Fine–Gray model for competing risks was used in 69 articles (23%), and frailty models were employed in 14 articles. Multi-state models (n = 2) and relative survival approaches (n = 3) were rarely employed. Furthermore, the software used for the analyses was R (n = 144), SPSS (n = 94), SAS (n = 55), Stata (n = 35), unknown/ not specified (n = 40), and other (n = 28). These numbers do not add up to the total as 86 articles used two or more of these software packages in combination.

A total of 195 articles (65%) reported missing data in at least one of the covariates that formed part of one or more of the presented multivariable models. In most cases, these were explicitly reported in the descriptives table (n = 124), in both the descriptives table and in text (n = 24) and in the main text only (n = 19). In some instances, these were reported as part of a figure/flowchart (n = 3). The missing values (for at least one of the variables) were implicit in 20 articles, deducted based on subcategories in the descriptives table not adding up to totals.

In 39% (n = 76) of the 195 articles reporting missing values, the method (or at least one of the methods, if multiple were used) for handling missing data was explicitly specified. The most common methods for dealing with missing values among these were CCA (n = 34), MIM (n = 29), MI (n = 6) and single imputation (n = 5). One article used both MIM and CCA together, while another used both MI and CCA. Han *et al.* (2021) provide a clear example of explicit method reporting in-text: they mention the use of MIM for a particular mutation status indicator, and the use of CCA on the remaining variables with missing values. Furthermore, in that same study, treatment data was not available for all patients: they proceeded to compare survival outcomes of patients with and without available treatment data, and checked that there were no differences. Occasionally, variables were explicitly excluded from the multivariable model if their proportion of missing values were deemed too large. This was the case in Sharma *et al.* (2021), where multiple variables (such as KPS and cytomegalovirus serostatus) were excluded from analyses on the basis on having a proportion of missing values larger than 35%.

In 67% (n = 131) of cases, all or part of reported missing data was handled implicitly, which was assumed to correspond to implicit CCA. Indeed, this meant that one or more predictors included in the multivariable model(s) had missing values, and that results were reported as if the data were complete: with no explicit mention of running a CCA in text, or no information provided on the reduction in sample size. Inoue *et al.* (2021) provide an example of both explicit and implicit reporting: the MIM was used for the HCT-CI comorbidity index, however the handling method for the other incomplete adjustment variables (such as performance score) was not stated.

From the few articles using MI, three explicitly mention using MICE, with the remaining simply referring to their approach to handling the missing values as general 'multiple imputation'. Information on the details of the imputation procedure was lacking across all these articles, with only three mentioning the number of imputed datasets (10, 15 and 30 imputed datasets used), and none of the articles outlining the contents of the imputation models or the number of cycles (also not in the supplementary materials).

2.5 Illustrative example

We make use of data published by Schetelig *et al.* (2019), describing long-term outcomes of patients with myelodysplastic syndromes (MDS) and secondary acute myeloid leukemia (sAML) following an alloSCT to demonstrate selected options how to deal with missing data. The dataset contained both outcome information (timing of relapse or non-relapse mortality, if either occurred) and variables measured at baseline for 6434 patients registered with the EBMT, transplanted between 2000 and 2012. Several of these recorded baseline predictors presented a substantial amount of missing data: IPSS-R cytogenetic classification (62.2%), HCT-CI comorbidity index (59.9%), donor age (49.5%), KPS (32.8%), and cytomegalovirus (CMV) status in both donor and patient (17.8%). For full description of the variables, we refer to Table 2.2. In our previous publication, we illustrated several approaches for dealing with these missing values in competing risks analyses (Bonneville *et al.*, 2022).

Table 2.2: Data dictionary with predictor variables and their descriptions, levels and proportion missing data for the illustrative example, adapted from Bonneville *et al.* (2022). The 'Summary' column reports median and interquartile range for continuous variables, as well as counts and proportion per level of categorical variables. Abbreviations: CMV = cytomegalovirus, CR = complete remission, IPSS-R = International Prognostic Scoring System, V. = very, interm. = intermediate, HLA = Human leukocyte antigen, HCT-CI = Hematopoietic stem cell transplantation-comorbidity index, M = male, F = female, MDS = myelodysplastic syndromes, sAML = secondary acute myeloid leukemia, w/= with, w/o = without.

Variable	Description	Levels	% Missing	Summary
Age (Donor)	Donor age at alloHCT (years)		49.49	42.1 (31.2, 52.5)
Age (Patient)	Patient age at alloHCT (years)		0	56 (46.9, 61.9)
CMV Patient/Donor	CMV status in patient and donor	Patient -/Donor -	17.8	1439 (27%)
		Patient -/Donor +		544 (10%)
		Patient +/Donor -		1281 (24%)
		Patient +/Donor +		2024 (38%)
Comorbidity score	HCT-CI score	Low risk (0)	59.93	1322 (51%)
		Interm. risk (1 – 2)		657 (25%)
		High risk (≥ 3)		599 (23%)
Cytogenetics	Cytogenetics categories used for IPSS-R	V. good/good/interm.	62.23	1784 (73%)
		Poor		287 (12%)
		V. poor		359 (15%)
HLA match patient/donor	HLA match between patient and donor	HLA-identical sibling	0	2666 (41%)
		Other		3767 (59%)

Karnofsky	Karnofsky performance status	≥ 90	32.8	3130 (72%)
		80		898 (21%)
		≤ 70		295 (7%)
MDS class	MDS groups based on subclassification at alloSCT	MDS w/o excess blasts	0	1355 (21%)
		MDS w/ excess blasts		2716 (42%)
		sAML		2362 (37%)
Patient/Donor sex match	Sex match patient and donor	M/M	1.68	2545 (40%)
		M/F		1196 (19%)
		F/M		1474 (23%)
		F/F		1110 (18%)
Stage	Stage at alloHCT	CR	3.17	2119 (34%)
		no CR		2156 (35%)
		Untreated		1954 (31%)

In the current work, using this dataset we present a multivariable Cox model for relapse-free survival (RFS), using the same covariates as in the cause-specific Cox models in Bonneville *et al.* (2022): the baseline variables with missing values described above, together with the completely observed variables patient age, stage at alloSCT, patient-donor human leukocyte antigen (HLA) match, patient-donor sex match and MDS classification. We compared various methods for dealing with the missing values: CCA, MIM, MICE, SMC-FCS, and fully Bayesian imputation using the 'JointAI' R package. The MI analyses were performed using 100 imputed datasets with 15 iterations, and the analysis using 'JointAI' used 2000 iterations following 200 adaptation iterations. Hazard ratios obtained using each method are presented in Figure 2.1 along with their corresponding 95% confidence interval. Full code to reproduce the analysis is available at https://github.com/survival-lumc/ReviewHaemaMissing.

Regarding results, we first note that the loss in efficiency when using a CCA is apparent in this application, with confidence intervals that are considerably larger than in any of the other methods. Indeed, the CCA made use of only 17.5% of the available observations (5309 patients were omitted from the analysis). CCA also presents point estimates that differ considerably compared to the remaining methods, as is the case for example with patient-donor sex match categories, the CMV status in patient and donor, and cytogenetic risk classification. Such differences can raise skepticism as to the validity of the MAR assumption made: assuming the specified model is 'correct' and contains all variables explaining the missingness mechanism (covariate dependent missingness), the point estimates obtained with CCA and MI should generally be in alignment. However, note that when missingness depends on the observed outcome (often unlikely in survival data), MI should theoretically outperform CCA, while the reverse holds true in the MNAR case where the missingness depends on the variable itself (Carpenter and Smuk, 2021). Of course in reality, missingness will often be a mixture of MAR and MNAR, and results should therefore be interpreted with care.

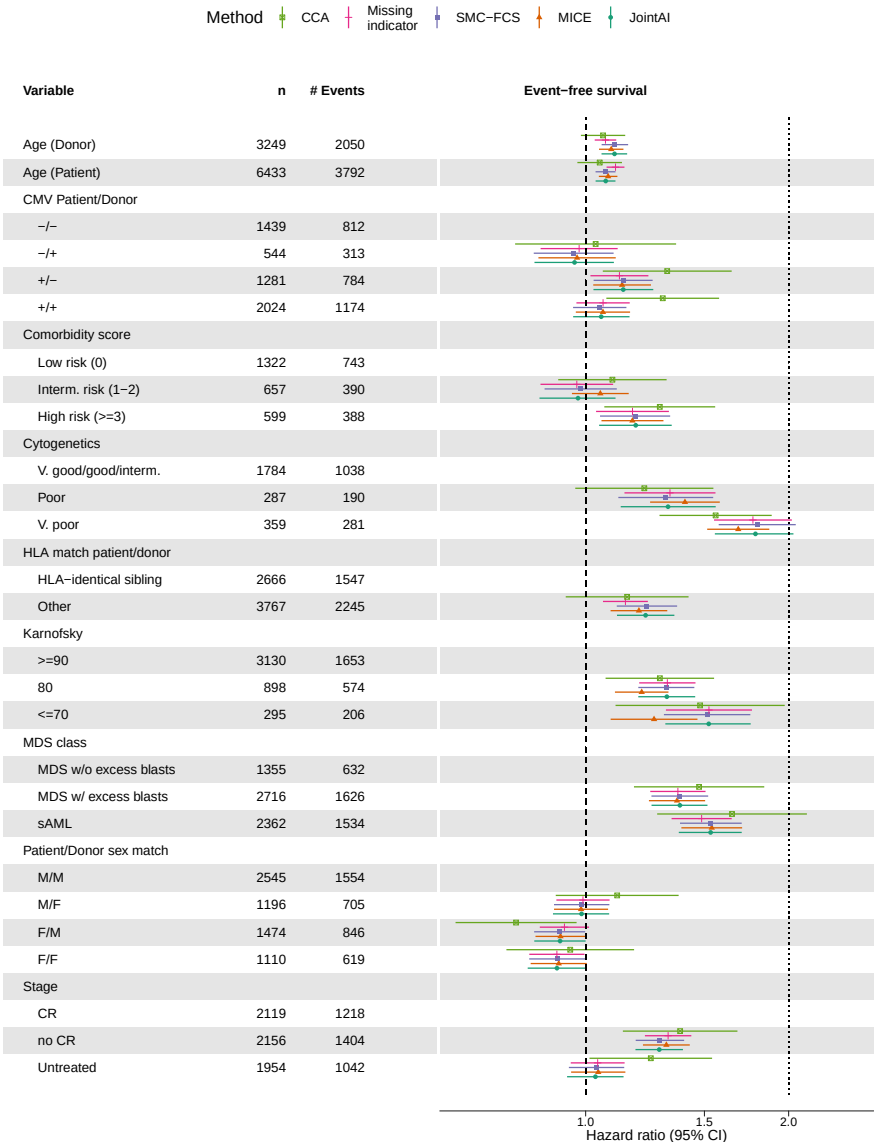

Figure 2.1: Point estimates and associated 95% confidence intervals for the Cox model for relapse-free survival, according to missing data handling method. Variables and their descriptions can be found in Table 2.2. Per level of factor and for continuous variables, we show the observed counts (n) and the number of events (# Events, which is the sum of relapse and non-relapse mortality events) in the full dataset.

Moreover, CCA point estimates will be more uncertain compared to their MI based counterparts, and one should therefore not expect to them to be equal (White *et al.*, 2022). One could indeed make the case that in the present example, given that the point estimates from the imputation methods fall within the confidence intervals of the CCA estimates (that is, within two standard errors), we should not be concerned.

Concerning the imputation methods, SMC-FCS and JointAI were in consistent agreement across all coefficients. The MICE approach differed noticeably from the previous two approaches for both the KPS and cytogenetic risk coefficients, which can perhaps be attributed to its theoretical limitations in the survival context: it does not ensure full compatibility between analysis and imputation model, in contrast to the other two MI approaches. The MIM was also consistently in agreement with estimates obtained with both SMC-FCS and JointAI. Given that the assumptions in Blake *et al.* (2020) concern missing data in adjustment variables (i.e. not in the outcome or in the exposure/treatment variable), it is difficult to reason about the validity of the MIM in the current context where the multivariable model is presented as prognostic.

2.6 Discussion and recommendations

Our systematic review demonstrated that missing data feature prominently across studies from major journals in the field of clinical haematology. Studies are often observational in nature, sometimes making use of large registries where whether or not a variable is recorded can depend on a multitude of reasons (e.g. variables collected only from a certain date, in particular centers). While presence of missing values was seemingly consistently reported (usually in the descriptives table), the method used for dealing with them was not. CCA was the dominant approach to handling missing values, and the associated loss of efficiency was generally poorly documented. Importantly, we also note that articles with seemingly complete data (particularly in observational studies) may have for simplicity chosen to filter out the missing values at a pre-processing step, without reporting it in the main manuscript—implying that the true prevalence of missing data is likely higher than reported in our review.

Various articles have attempted to raise the bar when it comes to reporting and handling of missing values more generally across observational studies. A convenient summary of guidelines found in Sterne *et al.* (2009), Sterne *et al.* (2016), and Vandenbroucke *et al.* (2007), can be found in Table 1 from Carroll *et al.* (2020), and should be considered alongside further guidelines given by Lee *et al.* (2021). The aforementioned guidelines put findings concerning missing values reporting in our systematic review into perspective: while the reporting may have been consistent across articles, it was by no means thorough. Paz *et al.* (2021) were perhaps the only standout article from the corpus in this respect. In their supplemental data, they provided a plot showing the missing data patterns (i.e. not only frequencies of missing values per variable,

but also frequencies for two or more variables being simultaneously missing), and presented a table showing the distribution of variables among complete cases versus cases with at least one missing value.

The review also showed that while multiple imputation remains an active topic of methodological research, its application in clinical haematological studies appears to be very limited. This does seem to contrast with findings regarding the increasing use of MI in the *Lancet* and *New England Journal of Medicine* over a period of 5 years (Hayati Rezvan *et al.*, 2015).

The systematic review had several strengths. The articles included covered the main haematological malignancies, as well as a (impact factor based) large range of journals. In order to obtain as representative a corpus as possible, we made sure to keep our initial criteria rather broad: filtering by journals, and by words (in abstract, title or keywords) directly referring to particular malignancies or multivariable Cox models. Importantly, we made sure not to include words such as imputation, missing or incomplete data in this initial search, as that could severely bias the sample towards articles which were exceptionally diligent with their missing data reporting.

One of the limitations of the present work is that the articles spanned a single year, and thereby may not be fully representative of trends in the literature. Furthermore, the extraction was limited in scope: we did not record many factors that may be of interest to stratify the results by, such as study type or goals (e.g. prediction), as was thoroughly done for example in Carroll *et al.* (2020). We effectively accepted the trade-off of having a larger sample of articles, in exchange for limited granularity on the information extraction.

As mentioned previously, extensive guidelines concerning the reporting and handling of missing values already exist, and are clearly not being adhered to. Among the few papers using multiple imputation, very little is reported except the number of imputed datasets: whether it be the number of iterations, the contents of the imputation model, or the discussion on the plausibility of the MAR assumption. We argue that at bare minimum, authors must (in the main article) a) state whether there were missing values, and if so, across which variables and in what frequency; b) make explicit whether and how their initial study population choice has been influenced by missing values; c) explicitly state the method used for handling the missing values. In the case of a CCA, one should then clearly report the sample size per analysis model. Given tight word count limitations for main text of many of the journals, authors should be encouraged to make use of supplemental materials to report additional information concerning the missing values. This may include discussion concerning the reasons why values were missing, comparison of outcomes between those with and without missing values, and full details behind the imputation procedure if an imputation method was used.

The choice of method for handling missing values in covariates will inevitably be highly context-dependent. For example, while the MIM is generally discouraged for observational studies (Groenwold *et al.*, 2012), its use has been approved for randomised trials (Sullivan *et al.*, 2018; White and Thompson, 2005). For observational studies, the flowchart shown in Figure 3 in Lee *et al.* (2021) provides a solid guide to motivating the choice of method for handling missing data. We would add that, if the MAR assumption is deemed plausible, results from a MI procedure (ideally SMC-FCS or fully Bayesian imputation, at least in the survival analysis context) should always be presented alongside the results of a CCA. Note that in cases where there are relatively few missing values, the imputation step could be skipped entirely: a CCA will likely be efficient enough, and will be unbiased under (covariate-dependent) MAR. In other cases, particularly when (auxiliary) variables related to either the variables with missing values or the mechanism are at the researcher's disposal, MI should be considered in order to mitigate a potentially important loss of statistical power. Nevertheless, the proportion of incomplete cases should not be the main criterion in deciding which method to use: when the imputation model is well specified and data are MAR, MI can yield unbiased results even with a large proportion of incomplete cases (Madley-Dowd *et al.*, 2019). Conversely, when either condition is not met (well specified imputation model and MAR), bias in the results will likely increase with the proportion of incomplete cases. Additionally, we caution that there is a lack of research concerning the use of MI for missing covariates in the presence of multiple outcomes. This is clearly relevant to articles in clinical haematology, which often present models for outcomes such as overall survival, relapse of a disease, and occurrence of graft-versus-host disease (GvHD) as part of a single study.

In conclusion, missing values are a prominent issue across studies in clinical haematology, and increased attention should be given in particular to the methods used to handle them, and the assumptions they make. In particular, we hope to stimulate a more open discussion about missingness mechanisms (and their implications on the validity of analyses), which can encourage better and more complete data collection in future studies. Nevertheless, researchers should remain prepared to discuss how the missing values in their particular context could affect the results of their study, both in terms of bias and statistical power.

Supplementary materials

The systematic review extraction sheet and associated reference list are available in the online supplement at https://doi.org/10.1016/j.beha.2023.101477.