

RESEARCH ARTICLE

Open Access



How are missing data in covariates handled in observational time-to-event studies in oncology? A systematic review

Orlagh U. Carroll^{1*} , Tim P. Morris^{1,2} and Ruth H. Keogh¹

Abstract

Background: Missing data in covariates can result in biased estimates and loss of power to detect associations. It can also lead to other challenges in time-to-event analyses including the handling of time-varying effects of covariates, selection of covariates and their flexible modelling. This review aims to describe how researchers approach time-to-event analyses with missing data.

Methods: Medline and Embase were searched for observational time-to-event studies in oncology published from January 2012 to January 2018. The review focused on proportional hazards models or extended Cox models. We investigated the extent and reporting of missing data and how it was addressed in the analysis. Covariate modelling and selection, and assessment of the proportional hazards assumption were also investigated, alongside the treatment of missing data in these procedures.

Results: 148 studies were included. The mean proportion of individuals with missingness in any covariate was 32%. 53% of studies used complete-case analysis, and 22% used multiple imputation. In total, 14% of studies stated an assumption concerning missing data and only 34% stated missingness as a limitation. The proportional hazards assumption was checked in 28% of studies, of which, 17% did not state the assessment method. 58% of 144 multivariable models stated their covariate selection procedure with use of a pre-selected set of covariates being the most popular followed by stepwise methods and univariable analyses. Of 69 studies that included continuous covariates, 81% did not assess the appropriateness of the functional form.

Conclusion: While guidelines for handling missing data in epidemiological studies are in place, this review indicates that few report implementing recommendations in practice. Although missing data are present in many studies, we found that few state clearly how they handled it or the assumptions they have made. Easy-to-implement but potentially biased approaches such as complete-case analysis are most commonly used despite these relying on strong assumptions and where often more appropriate methods should be employed. Authors should be encouraged to follow existing guidelines to address missing data, and increased levels of expectation from journals and editors could be used to improve practice.

Keywords: Missing data, Time-to-event, Observational studies, Survival, Epidemiology, Oncology, Multiple imputation

*Correspondence: Orlagh.Carroll@lshtm.ac.uk

¹Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street, London, UK

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Time-to-event or survival studies focus on the analysis of times to an outcome or event. Missing data in covariates is a problem in many such investigations. It can render estimators biased if applied to the complete-cases or using an ad hoc approach to handling missingness, and a loss of power to detect associations between explanatory variables and times-to-event. The presence of missing data can also lead to further challenges in a survival setting such as the handling of time-varying effects or dealing with time-dependent covariates when values are missing in the covariates in question. Additionally, it can lead to questions about how best to approach the checking of model assumptions, for example, the proportional hazards assumption when using a Cox model. Missing data also brings further challenges not specific to time-to-event scenarios, such as how to address the selection of covariates into a model or the flexible modelling of covariates. Another type of missingness concerning time-to-event scenarios is missing observations of the event time due to patients being censored, for example, due to administrative censoring or loss to follow-up. This is typically addressed in the analyses for right-censored data, making the assumption that the censoring is uninformative. Missingness in the outcome is not assessed in this review which instead focuses on missing data in the explanatory covariates only.

Complete-case analysis is both a simple and popular method for dealing with missing data, which involves restricting the analysis to individuals with no missing data. Other simple approaches involve replacing missing observations in a covariate with the mean, median or modal value or the use of a missing indicator category for categorical covariates. While popular, these methods can be biased, inefficient or underestimate the variance of estimates. Multiple imputation is an increasingly popular method for handling missing data which involves replicating the original dataset multiple times and in each replication replacing the missing values with plausible observations drawn from the posterior predictive distribution [1]. It is typically conducted using the 'missing at random' (MAR) assumption [2], which also subsumes 'missing completely at random' (MCAR). MCAR means that missingness does not depend on the observed or missing values while MAR means that missingness is conditionally independent of the missing values given those which have been observed. Further methodology has been developed to adapt the use of multiple imputation in a survival setting. White and Royston in 2009 [3] focused on the Cox model and recommend including the Nelson-Aalen estimate and event indicator in the imputation model. Bartlett et al in 2015 [4] described an alternative imputation approach suitable for several analysis models including the Cox model and Keogh and Morris (2018) [5]

adapted both approaches to handle time-varying covariate effects - that is, non-proportionality of hazards.

In addition to developed methodology, there have been several published guidelines focusing on how to conduct and report an observational study with some recommendations pertinent to reporting with incomplete covariate data, summarised in Table 1. Some guidelines, such as Sterne et al. [6] focus purely on the handling and reporting of missing data while using multiple imputation, whereas STROBE [7, 8] and ROBINS-I [9] focus more generally on reporting of observational studies. Examples of recommendations range from providing detail on eligibility criteria of patients to clearly stating the selection process for the final analysis model to reporting the amount of missingness in each covariate and which method was chosen to deal with the missing observations. Sensitivity analyses are also recommended to investigate plausibility of any assumptions assumed and the robustness of results. These published guidelines aim to introduce transparency as well as replicability of results if another analyst were to conduct the same investigation.

Time-to-event studies are commonly conducted in oncology with a search for time-to-event or survival studies on Web of Science indicating oncology to be the most popular category at approximately 30% of journal articles. As such, this review focused on studies conducted in any area of oncology. Common scenarios involve assessing the risk factors of patients developing a specific cancer or investigating factors associated with survival post-diagnosis. Proportional hazards models and Cox regression, in particular, continues to be the dominant analysis technique in time-to-event studies. As such, the review focuses on proportional hazards models while allowing for the extension of the Cox model to include time-varying effects.

Given the developed methodology in this field and the detailed recommendations in place, this review aims to:

- understand which methods researchers are using in time-to-event analyses when missing data are present
- assess if methods used are being carried out appropriately and the relevant assumptions stated
- assess how other challenges such as covariate selection, choice of functional forms (i.e. whether the covariate should be included as a linear term or be more flexibly modelled) for continuous covariates and checking of model assumptions are handled, particularly in the presence of missing data.

Methods

Databases, search strategy and screening

Medline and Embase databases were searched for studies published between January 2012 and January 2018 to allow time for developed methods and guidelines to

Table 1 Summary of recommendations or considerations from STROBE, ROBINS-I and Sterne et al. guidelines

Recommendation	Explanation	STROBE	ROBINS-I	Sterne et al.
Patient Selection				
State eligibility criteria	State inclusion and exclusion criteria of study participants, including criteria concerning missing data	✓		✓
Report the number of individuals at each stage of the study	Give reasons for exclusion at each stage	✓		
	Indicate the amount of individuals discarded due to missingness at each stage of the study	✓		✓
	Give consideration to selection bias introduced by exclusion criteria		✓	
	May use a flowchart to summarise	✓		
Modelling and Covariate Selection				
Covariates	Detail whether included as continuous or categorical and, if relevant, detail how the quantitative covariate was categorised	✓	✓	
	Consider departures from linearity for continuous covariates and state which transformation, if any, was used	✓	✓	
State analysis model	make it clear which method will be used to model the data	✓	✓	
Covariate Selection	describe the procedure used to reach the final model	✓	✓	
	this includes, but is not restricted to, missing data imputation, transformation of covariates, interactions between covariates or inclusion of covariates for a priori reasons	✓	✓	
Results	Provide unadjusted estimates and the final adjusted model	✓	✓	
	State the number of participants included in unadjusted and adjusted analyses	✓		
Missing Data				
Report the number of participants with missing data	Report this for each covariate of interest or the number of complete data for the important covariates	✓		✓
	Give reasons for missing values	✓	✓	✓
	Investigate if there are key differences between those observed and those with missing data - this may be compared across exposure/intervention groups.		✓	✓
<i>Missing data methods (general)</i>				
Which method was used to handle missing data?	State clearly the method used	✓	✓	✓
State any missing data assumptions that were made	Such as whether the data are MCAR, MAR or MNAR	✓	✓	✓
Sensitivity analysis	Should investigate robustness of findings	✓	✓	
	Compare method with a complete-case analysis		✓	
	If necessary, assess validity of methods if there are differences	✓	✓	
	Assess plausibility of missing data assumptions		✓	

Table 1 Summary of recommendations or considerations from STROBE, ROBINS-I and Sterne et al. guidelines (*Continued*)

Recommendation	Explanation	STROBE	ROBINS-I	Sterne et al.
<i>Multiple Imputation</i>				
Give details of the imputation model	State the software used and key settings for imputation model			✓
	State the number of imputations used			✓
	State variables included in imputation model			✓
	State how non-normal or binary covariates were handled			✓
	Were interactions in analysis model included in imputation model?			✓
If a large fraction of data are imputed, compare observed and imputed values				✓
Missing data assumptions	Discuss if variables included in the imputation model make MAR assumption plausible			✓
Sensitivity analyses	Compare MI results with CC results			✓
	Investigate departures from MAR assumption			✓
	If necessary, suggest explanations for why there are differences in results across sensitivity analyses			✓

be used in practice. The search strategy for observational studies consisted of three main components: oncology, missing data and time-to-event analyses; additional details can be found in Additional file 1.

For inclusion, studies had to use a proportional hazards or an extended Cox model (includes an interaction between a covariate and time) in a cancer setting. The study also had to have a reference to missing data (either ‘complete’ or ‘missing’) in the abstract or in the full-text. Studies involving only competing risks, frailty models, accelerated failure time models or excess hazards in the abstract or full-text were excluded from the review. If the abstract mentioned a time-to-event outcome but did not specify the analysis models used, the paper proceeded to a full-text review. Papers not written in English or which focused on methodology, meta-analyses, validations of previously created models, and primary or secondary trial outcomes were excluded. However, retrospective observational analyses of a trial cohort were included.

Data extraction

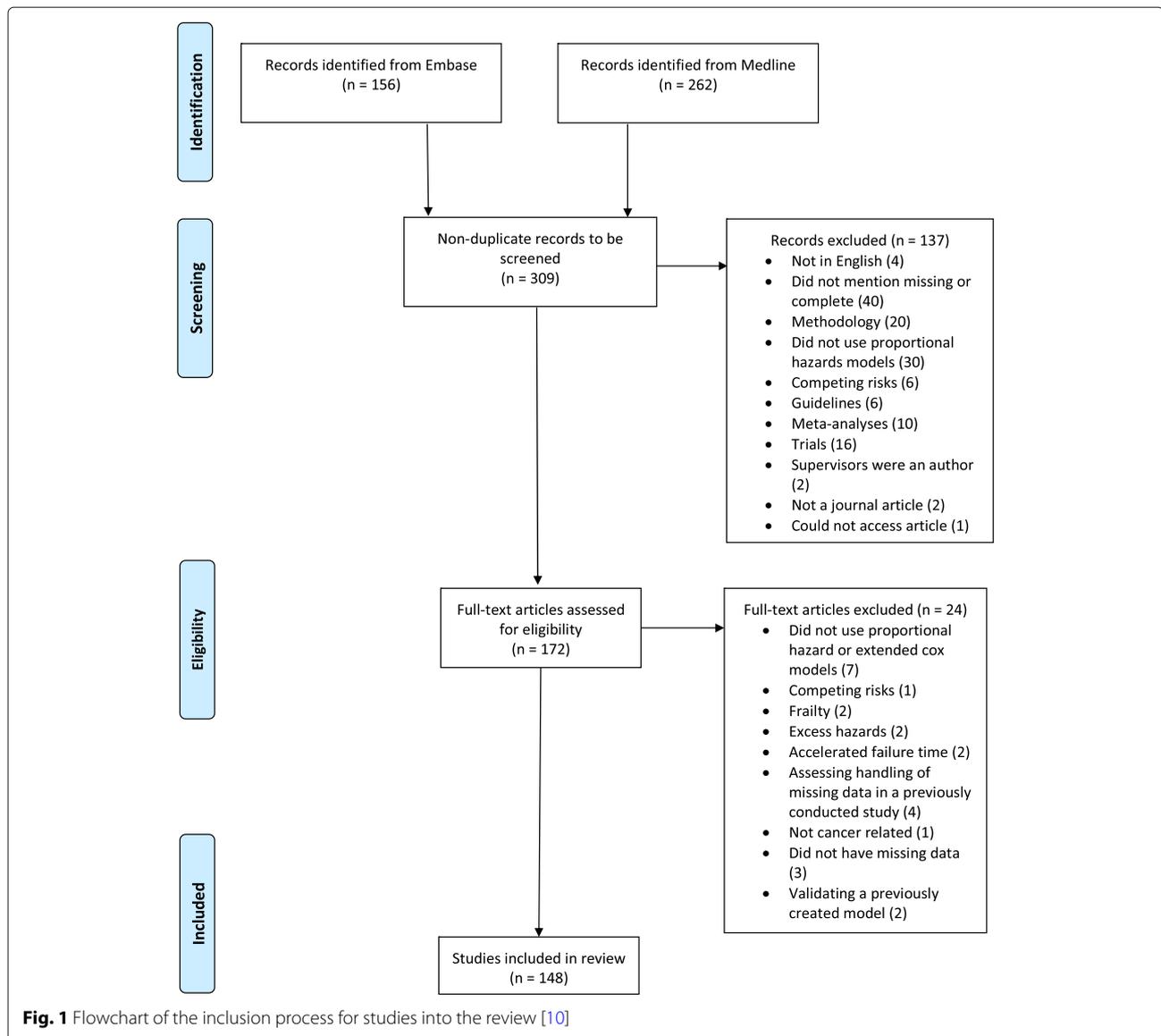
The information extracted focused on two key areas: missing data and features of the time-to-event analysis. The missing data component assessed the sample size used in the study, how much missing data had been discarded, if assumptions about the treatment of missing data in the analysis were stated and how any missing data were handled in the analysis: complete-case analysis, single imputation techniques or multiple imputation. Where multiple imputation was used, the choice of univariate or multivariate imputation was recorded, the number of imputations used and which covariates were

included in the imputation model. Online supplementary materials were accessed only when referenced with regards to the handling of missing data in the text. The features of the time-to-event analysis assessed were whether the proportional hazards assumption was investigated, how covariates were selected for model inclusion and the assessment of the functional form (if continuous covariates were included). We also assessed, where relevant, how missing data were treated in the context of these features. In addition, the software used for the analysis was also extracted by searching for ‘Stata’, ‘SAS’, ‘SPSS’, ‘R’ and ‘plus’ (for S-plus and Mplus). Papers which did not mention one of these six programs were then searched for the software used. A detailed list of the information extracted can be found in Additional file 1 which was motivated by the guideline recommendations found in Table 1 and are evaluated in the “Results” section.

A pilot investigation consisting of 10 randomly selected papers was carried out by OUC, TPM and RHK to assess the consistency of data extraction, refine the data extraction checklist and agree on how to extract information when answers were ambiguous. Data extraction was then carried out by OUC.

Results

The PRISM diagram [10] summarising the review inclusion process is shown in Fig. 1. Four hundred and eighteen papers were identified from Embase and Medline, of which 309 were non-duplicates and proceeded to the screening step. One hundred and thirty-seven studies did not meet the inclusion criteria during screening and were therefore excluded. After a full-text assessment, a



further 24 studies were excluded with a total of 148 studies included within the review. The studies included came from 110 journals, of which the most prominent were BMC Cancer (5), International Journal of Radiation Oncology, Biology, Physics (4) and Journal of the National Cancer Institute (4).

Missing data

Reporting extent of missing data

In the pilot stage we noticed that many studies excluded individuals due to missing data on key covariates in an initial phase (in which the study population was determined using inclusion or exclusion criteria). That is, having certain covariates observed was used as part of the inclusion criteria. One hundred and six (72%) studies excluded missing data while determining their study population

in this initial phase. Of the 106 studies which excluded observations, 66 (62%) reported the number of individuals excluded. On average, 14% of individuals were discarded during this stage. After inclusion criteria had been applied, 102 (69%) studies contained patients with missing data. Table 2 shows the breakdown of missing data during the initial phase and analysis stage of the study.

In the demographics table (often considered to be ‘Table 1’ in publications), 87 (59%) studies summarised the missing data in covariates, 47 (32%) reported the breakdown of missingness in incomplete covariates and two (1%) used missing data pattern plots. Thirty-four (23%) used both the text and a table to report the extent of missingness. For the 48 (32%) who did not use a plot, use a table or explicitly break down the missing values in each covariate a general statement was typically made stating which

Table 2 Breakdown of the number of individuals with missing data

Description	Number	(%)
Excluded missing data in initial phase (N=106)		
Excluded individuals with missing data in any covariate ¹	44	(42)
Excluded individuals with missing data in a subset of covariates	62	(58)
Reported the number of individuals excluded	66	(62)
<i>Percentage (%) of individuals excluded (n = 66)</i>		
Mean (SD)	14.14	(12.40)
Median (IQR)	10.22	(4.73, 18.34)
Min, Max	0.11	47.38
Missing data present for the analysis stage (N=102)		
Reported missing data in baseline table for incomplete covariates	82	(80)
Used a complete-case analysis ²	35	(34)
Used other missing data methods	36	(35)
Quantified the complete-case sample size	25	(25)
<i>Percentage (%) of individuals excluded (n = 25)</i>		
Mean (SD)	31.65	(21.90)
Median (IQR)	31.34	(13.67, 37.76)
Min, Max	1.77	94.16

The initial phase is the stage when defining the study population using inclusion and exclusion criteria.

¹potentially used a complete-case in initial phase but did not clearly state their methods

²A further 31 were not clear on whether they used a complete-case during the analysis

variables were incomplete or that variables or patients were excluded due to having incomplete data.

Analyses performed

Table 3 summarises the methods used for the analysis in the presence of missing data. Complete-case analysis was the most popular and was used in 79 (53%) studies either in the initial phase or at the analysis stage (either as the primary method used to deal with missing data or as a sensitivity analysis). This was followed in popularity by removing individuals with missing values in certain key covariates (62, 42%) and multiple imputation (33, 22%). Some studies used multiple methods for handling missing data with 18 (12%) using both complete-case and multiple imputation.

68 (50%) of all studies used a complete-case analysis as their primary analysis method and 24 (16%) reported multiple imputation as their main analysis. Of those using

Table 3 Methods used in studies for the handling of missing data

Missing data methods	Count	(%)*
Complete-case	79	(53)
Removed individuals with incomplete data for a subset of covariates	67	(45)
Multiple Imputation	33	(22)
Missing indicator	10	(7)
Worst or best case scenario ¹	2	(1)
Stochastic imputation	1	(1)
Mean value imputation	1	(1)
Mode value imputation	1	(1)
Growth models	1	(1)
Bayesian model incorporating handling of missing data	1	(1)
Full-information maximum likelihood estimation ²	1	(1)
Selection procedure ³	1	(1)
Unclear	33	(22)

*Percentages do not sum to 100 as there is overlap with some studies using more than one method.

¹[11, 12]

²[11]

³A selection model to account for missing data and time-varying covariates [13]

complete-case analysis as the main analysis, nine (13%) also used MI or other methods. Of those using MI as the main analysis, 12 (50%) used complete-case analysis or another method as a secondary analysis.

Missing data assumptions

Of the 148 studies, 128 (86%) did not state the assumptions that their chosen analysis made regarding the missing data. Eighteen (12%) stated the MAR assumption, of which 16 (89%) gave a general statement such as 'MAR was assumed', with no further explanation. One (0.7%) study stated MCAR and another stated 'missing not at random'.

Sensitivity analyses and stating missing data as a limitation

Ninety-eight (66%) studies did not mention the presence of missing data as a limitation to their analysis. Twenty-six (18%) used sensitivity analyses to check the robustness of their final results to either different assumptions concerning the missingness or comparing results with other techniques to handle missing data.

Description of complete-case analysis

Thirty-five (34%) used a complete-case analysis and a further 31 (30%) were suspected to have used complete-case during the analysis stage based on the information provided but did not state this clearly in their paper. On average, 32% of individuals were discarded by applying a

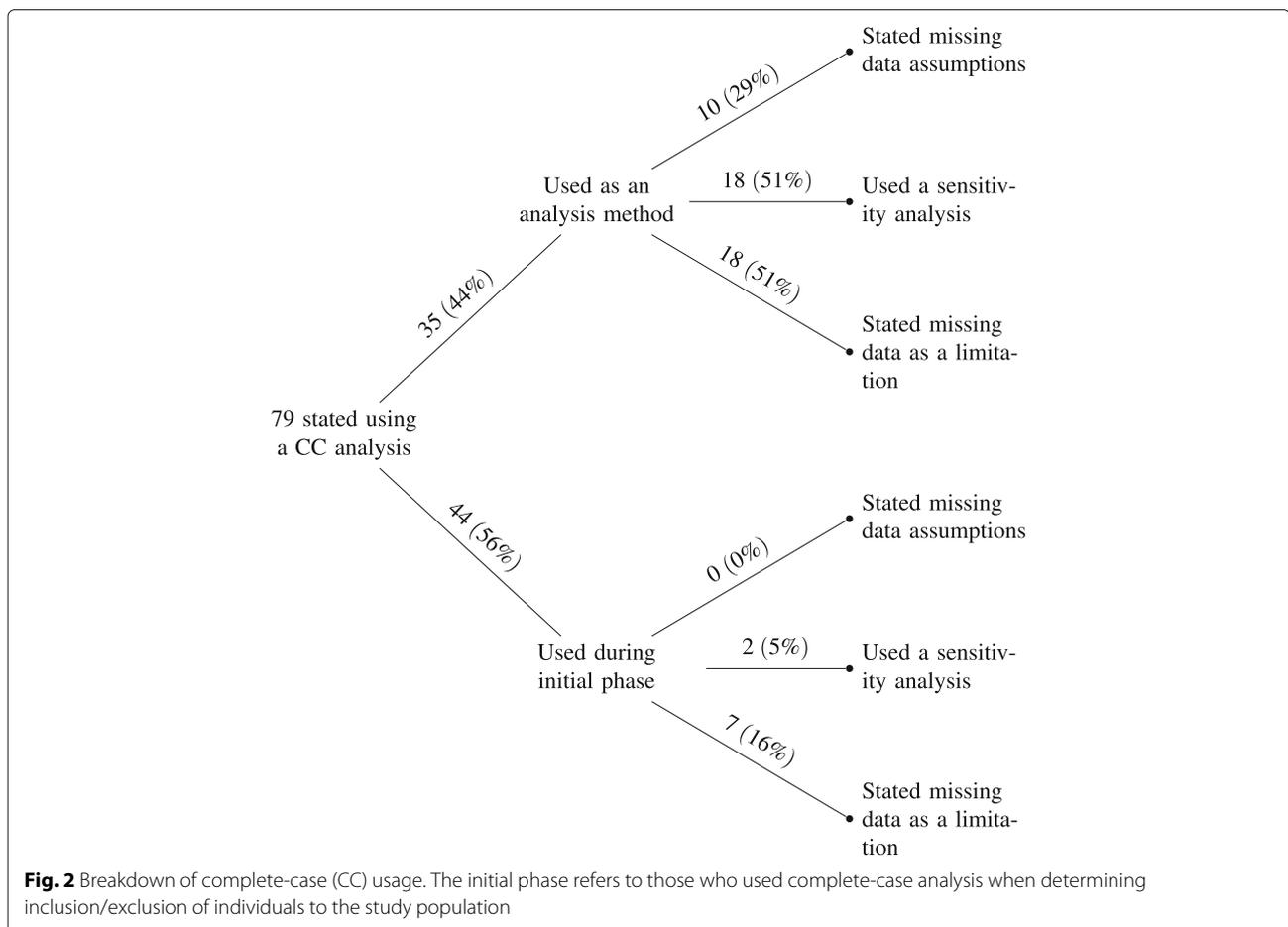
complete-case analysis, the maximum being 94% where complete-case was used as a sensitivity analysis for comparison with the main analysis using multiple imputation. Figure 2 summarises the reporting of missing data in the 79 studies that used a complete-case analysis (either during the initial phase or analysis stage). Seven (16%) of the 44 (56%) studies using the initial phase complete-case stated missing data as a limitation. Of the two (5%) studies using a sensitivity analysis, one compared with multiple imputation and the other compared the initial complete-case results pre and post propensity score matching and therefore with different sample sizes. Thirty-five (44%) studies used complete-case during the analysis stage, of which 18 (51%) stated missing data as a limitation. In addition, we presumed based on the information provided that a further 33 studies used a complete-case in the initial phase or analysis stage but did not clearly state this as their method to handle missing data.

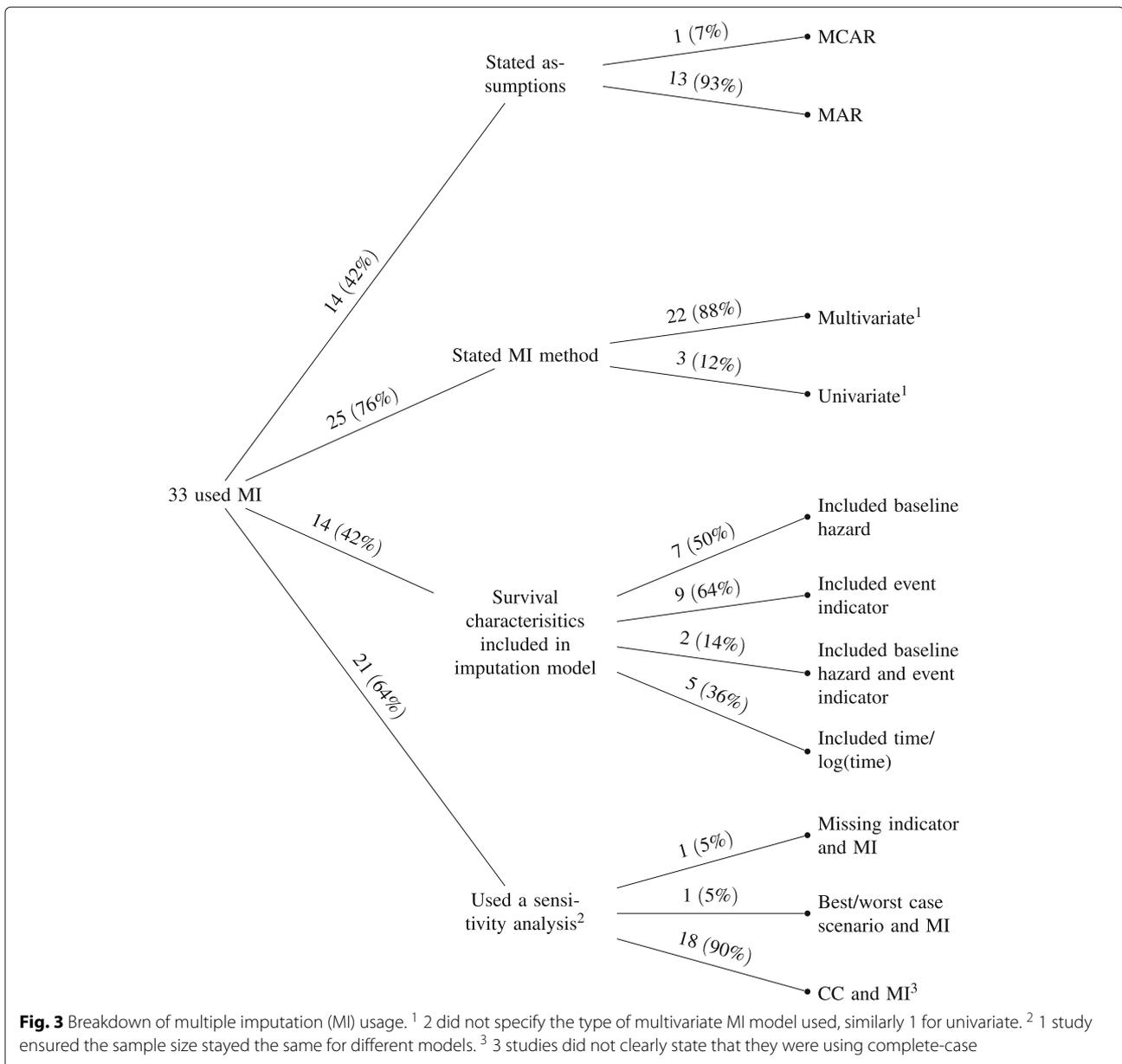
Sensitivity analyses with complete-case: Eighteen (51%) studies used a sensitivity analysis, of which 14 (78%) involved multiple imputation versus complete-case

analysis, two (11%) used complete-case analysis where individuals with missing data in any covariate were excluded versus excluded if there were missing data in a specific subset of covariates (known as available case analysis), one (6%) tested various missing data assumptions and one (6%) did not specify.

Description of multiple imputation

The breakdown of multiple imputation usage can be seen in Fig. 3. Thirty-three (22%) studies used multiple imputation, of which 24 (73%) reported the multiple imputation estimates as their main study results. Fourteen (42%) stated a missing data assumption and 25 (76%) described whether a multivariate or univariate approach was taken. For those using a multivariate imputation approach (22, 88%), multivariate imputation by chained equations (MICE) was the most popular method (19, 86%). In total, 14 (42%) studies included a component of the time-to-event outcome in their imputation model. These included the baseline hazard (7, 50%), the event indicator (9, 14%) or both (2, 14%). Twenty-six (79%) studies using multiple imputation stated the number of imputations. One (3%) used a single imputation, five (15%)





used five, six (18%) used 10, seven (21%) used 20, two (6%) used 25 and five (15%) used 50. Some studies (example: [14]) cited the White, Royston and Wood paper [15] which suggests that the rule of thumb for choosing the number of imputations should be at a minimum the percentage of cases that are incomplete while other studies (example: [16]) stated the number of imputations with no justification.

Sensitivity analyses with multiple imputation: Of the 21 (64%) studies that conducted a sensitivity analysis, 18 (90%) compared complete-case and multiple imputation (three of which did not explicitly state complete-case) and

10 (56%) used multiple imputation as the main analysis method and one (6%) was unclear on the main strategy while reporting both multiple imputation and complete-case results.

Missing data assumptions with methods

Of the 18 studies which stated the MAR assumption, 11 (61%) used multiple imputation, two (11%) used complete-case and one (16%) was not clear on whether they used complete-case or multiple imputation, two (11%) were suspected to have used complete-case but did not clearly state, one (16%) used a stochastic single regression imputation model and one (16%) used a fully Bayesian model. The one study stating MCAR used complete-case

analysis and the other stating 'missing not at random' performed an analysis using a selection model for the joint distribution of the missing covariates, the outcome and the probability that covariate data are missing [13].

None of the 44 (56%) studies using the initial phase complete-case stated a missing data assumption and for the 35 using complete-case analysis during the analysis stage 10 (29%) stated a missing data assumption. This consists of one (10%) study stating MCAR and nine (90%) MAR, of which seven (78%) used multiple imputation and complete-case analysis together for sensitivity analyses (six (86%) of these used multiple imputation as the main method for handling missing data and complete-case analysis used as a comparison). For the 14 studies using multiple imputation having stated a missing data assumption, 13 (93%) used MAR and one (7%) used MCAR.

Features of the analysis

Selection of covariates into the model

One hundred and forty-four (97%) studies used a multivariable model and therefore used some selection method or criteria to select which covariates should be included. Of these, 85 (59%) stated a clear selection procedure. The use of a predefined set of covariates (33, 39%), stepwise methods (31, 37%) and univariable analyses (32, 38%) were most commonly used, with eight (10%) studies using both a predefined set and univariable analyses. Of the 31 studies using stepwise methods, backwards elimination was used in 18 (58%), six (19%) used forwards selection and seven (23%) did not state which type of stepwise method was used.

Eleven (35%) used complete-case as the main method to handle missing data. Five (45%) out of the 31 studies using stepwise methods stated excluding individuals with missing data on key covariates in the initial phase, and were left with no additional missing data at the analysis phase. Six (55%) studies that used a stepwise procedure did so in a complete-case analysis. We suspect that an additional six (19%) studies used complete-case analysis but this was not clearly stated.

For the 31 studies using stepwise methods, 10 (32%) used them in combination with multiple imputation. Of this, eight (80%) used multiple imputation as the main method to handle missing data, one (10%) used a missing indicator as the main method with multiple imputation as a sensitivity analysis and one (10%) used a sensitivity analysis but did not state whether multiple imputation or complete-case was the main method. For those who used multiple imputation as the primary method to handle missing data, seven (88%) did not state how they combined it with the stepwise methods and the other is suspected to have applied the stepwise procedure in a complete-case analysis to determine the set of covariates

to be included, before using this set in the model fitted in each imputed dataset, however this was not clearly stated.

For the 32 studies using univariable analyses, four (13%) studies used multiple imputation as the main analysis method, 14 (44%) stated that they used a complete-case as the main method for handling missing data, 12 (38%) were presumed to have used a complete-case analysis based on the information available, one (3%) used stochastic single regression imputation model and one (3%) did not include incomplete covariates in the analysis model (available case analysis).

Functional form of continuous covariates

Sixty-nine (47%) studies included continuous covariates in their model, of which 57 (83%) did not report considering whether any form other than linear was required, i.e. its appropriate functional form. For those that did consider it, splines were the most popular way of transforming covariates in the model (8, 12%), followed by fractional polynomials (2, 3%) and Martingale residuals (2, 3%). Including a quadratic term or a 'flexible non-linear model' were each used once. One study used Martingale residuals, cubic splines and fractional polynomials to investigate evidence of non-linear associations [17]. For a further 11 (7%) studies, it was not clear whether included covariates were continuous or categorical. For the 12 studies which reported assessing the functional form of covariates, three (25%) used multiple imputation as the main method for handling missing data, three (25%) used initial phase complete-case, three (25%) presumably used complete-case analysis but this was not clearly stated, one (8%) used stochastic single regression imputation, one (8%) used a study-specific model to impute missing values and one (8%) used available case analysis by restricting to individuals with complete data in the covariates to be included in the analysis model.

Proportional hazards assumption and time-varying effects of covariates

The primary analysis method in 142 (96%) studies was the Cox model and the remaining six (4%) stated the use of a proportional hazards model. When investigating the proportional hazards assumption, the covariates included within the analysis model should be assessed. Forty-one (28%) studies stated that the proportional hazards assumption was assessed either using a general statement (example:[18]) or specifically detailing how to handle the covariates which violated it (example:[19]). Of those who checked, seven (17%) did not state the method used to assess the assumption. Schoenfeld residuals were most frequently used (18, 44%), followed by visual inspection of plots of Kaplan-Meier estimates of survivor curves, or functions thereof (12, 29%). Of these two methods, seven (17%) studies used both. Ten (24%) studies tested the

assumption by including an interaction between covariates and follow-up time in the model.

For the studies that checked the assumption, 13 (32%) used multiple imputation as the main method to handle missing data, three (7%) used a missing indicator, 11 (27%) used complete-case analysis, seven (17%) presumably used complete-case analysis but did not clearly state, two (5%) had no missing data in covariates chosen for inclusion in the analysis model, one (2%) used both multiple imputation and complete-case but did not state which was the main method, one (2%) excluded incomplete covariates from the analysis model and one (2%) removed individuals with missing data in specific covariates. For the 18 studies using Schoenfeld residuals, six (33%) used multiple imputation as the main method for handling missing data. For the 12 studies using visual inspection of survivor curve plots four (33%) used multiple imputation and for the 10 including an interaction with time, three (30%) used multiple imputation as the main method.

Five studies discovered evidence for time-varying effects, of which three (60%) had incomplete covariates associated with time-varying effects. Two (67%) of these used multiple imputation to impute the missing values in the covariate, of which one took into account the time-varying effect using methods developed by Keogh and Morris [5] and the other stated using MICE while the third study was unclear on how they handled the missing data.

Software

Forty-four (30%) studies used SPSS, 41 (28%) used SAS, 36 (24%) used Stata, 11 (7%) used R, two (1%) used winbugs, one (1%) used XL-stat life and 17 (11%) did not state. Of these, three (2%) used both SAS and SPSS, one (1%) used SAS and Stata together, one (1%) used SAS and S-plus and one (1%) used SAS and Mplus. Of the 11 studies using R, four (36%) used multiple imputation with three using the MICE package and one using Hmisc. Examples of other potential packages that could have been used are Amelia [20], jomo [21] or smcfcs [22].

Discussion

Missing data is a pervasive problem in observational time-to-event studies. However, this review has found that few studies appropriately report this issue. Whether this is due to a lack of appreciation of the potential implications of missing data from the researcher, or to the handling of missing data not being deemed of high enough importance to be described in the “Methods” section is unclear. There are general guidelines in place such as STROBE [7, 8] and Sterne’s specific multiple imputation recommendations [6] from 2007 and 2009, respectively, but it appears that many of their recommendations are still not being implemented. By considering literature from 2012

onwards, all papers we reviewed came after the publication of these guidelines. Over half of papers considered (53%) were from 2016 onwards. A surprising finding was that in many studies it was not clear how the study population was selected and what the extent of missing data was. We recommend that authors provide clear and comprehensive information on these aspects including detailing the finalisation of the study population, and stating the sample size used in each model when missing data are present. These recommendations would aid in the transparency of research findings.

Methods for handling missing data such as the multiple imputation approach of White and Royston [3] were implemented by two studies in 2014 [23] and 2016 [24], five and seven years respectively after the method was published. Although valid methods have been developed to handle missing data, the easier-to-implement approach of complete-case analysis is still the most popular method used. However, the studies suggest that little or no consideration is being given to the missing data assumptions needed for this method and whether they are introducing bias to their results. It is plausible that some authors had not noticed the missing data, since software by default runs complete-case analysis without flagging that some individuals were dropped from the analysis. Also of note are the studies which have a ‘fully’ observed dataset and therefore had no need to consider any missing data assumptions or methods. However, this ‘fully’ observed dataset originated from using a complete-case inclusion/exclusion criteria for individuals entering their study. These studies gave no consideration to missing data assumptions and only seven (16%) considered the missing data excluded to be a limitation.

Several systematic reviews have been conducted to assess the handling of missing data in studies, most of which have focus on randomized trials. Wood et al. [25] reviewed the handling of missing outcome data in randomized control trials published in 2001. They found that missing data are typically handled inadequately and that there was almost no use of modern data methods with complete-case used in 46% of studies. Similar findings were made in other reviews covering trials published between 2005 and 2014 [26–28]. Karahalios et al. [29] focused on missing data in cohort studies published between 2000 and 2009 and found inconsistent reporting of missing data and inappropriate methods used with 66% of studies using complete-case analysis. With regards to missing data, these reviews collectively looked at studies published between 2001 and 2014. They, along with our own review looking at papers from 2012 to 2018, highlight the lack of progress that has been made in appropriate handling of missing data in both trials and observational studies.

Our review revealed a lack of rigour in other aspects of a study investigation. 42% of studies did not state how the covariates were selected for their final model. When conducting a covariate selection procedure, thought should also be given to continuous covariates and whether categorising is worth a loss of power to detect associations or the occurrence of residual confounding [30]. Clarity should also be required for how the selection procedure is combined when using multiple imputation. For example, [31] states using multiple imputation for multivariable analyses and goes on to detail that univariable models and backward selection were used. However, no discussion is given as to if this process was repeated across the multiple imputed datasets and, if so, what happened when there were disagreements across them regarding the selection process? In 2008, Wood et al. [32] discusses methods to handle covariate selection with multiple imputation. Studies included in the review tended to be exploratory or predictive in nature and consideration should be given to the selection procedure for including covariates into these models. Stepwise methods were used in 37% of studies which stated a covariate selection procedure despite the disadvantages being well-known [33, p. 68]. These include underestimating standard errors of parameter estimates, narrow confidence intervals, low p-values and parameter estimates biased away from zero. VanderWeele also discusses the use of stepwise methods and their drawbacks in a causal setting [34].

For the 41 studies that checked the PH assumption, it was not clear how the 13 studies using multiple imputation incorporated the use of Schoenfeld residuals or inspection of survivor curves as these details were not provided. For those using a time-interaction and multiple imputation, only one did not make it clear how they were incorporating the two methods. Using again the example of [31] it is possible that they checked the assumption using scaled Schoenfeld residuals over time in a complete-case scenario or individually in each imputed dataset but without specification it is difficult to say whether the assumption diagnostics were carried out appropriately. It is important to note that when considering compatibility between the analysis and imputation model thought should also be given to allowing for time-varying effects in the imputation process, in order to allow for valid tests of the proportional hazards assumption. Further thought should also be provided on whether there is sufficient statistical power to detect violations of the proportional hazards assumption [35].

This review demonstrates poor adherence to guidelines already in place and further drives the need for clear reporting. Ideally, an external analyst should be able to rerun the study analysis from the information published which is currently not possible in many studies. Finally,

Table 4 provides some related references for consideration of different aspects of missing data and time-to-event features in a study.

Limitations of review

A large number of search terms were used to extract the relevant studies. However, it is possible that some time-to-event studies did not mention how they handled missing data in the title, abstract or keywords and therefore were not included in the review. The search also focused solely on oncology, it is possible that in other medical setting studies could be performed differently in terms of reporting or methods used. A further limitation stems from only one reviewer identifying, screening and extracting information from the studies which may have introduced bias from the selection and interpretation of papers. An agreement check was conducted with RHK and TPM and initially found poor agreement in the collection of sample size of studies and the amount of missing data. The data collection check-list was reviewed and amended to improve discrepancies.

Many journals have a page or word limit which restricts the study analysts from fully detailing methods conducted and results. It is possible that studies were unable to detail information such as checking the PH assessment or conducting a sensitivity analysis. However, most journals also allow for online supplementary materials which could have been used.

For this review we focused on methods used in the oncology field. It is possible that the handling of missing data may be better or worse in other medical fields or study designs.

Recommendations for multiple imputation in time-to-event analyses

While it is difficult to recommend a gold standard method as it can depend on the context of the study, for time-to-event studies involving the Cox model we would recommend using the substantive model compatible fully conditional specification (SMC-FCS) of Bartlett et al. [4] as the gold standard method for multiple imputation. It allows for compatibility between the study analysis model and the imputation model. This method is available in both Stata and R software. Keogh and Morris [5] have adapted SMC-FCS to allow for the presence of time-varying effects and proposed an algorithm to allow for model selection with time-varying effects.

White and Royston [3] recommend the inclusion of the event indicator and the Nelson-Aalen estimator in the imputation model for an approximately compatible model. While this is simpler and more straightforward using widely available MI software, the approximation can perform badly in 'extreme' scenarios such as strong covariate effects and a high event rate. The approximation also

Table 4 Selected papers describing methods for addressing common issues arising in the analysis of time-to-event data when there is missing covariate data

Consideration	Some recommended references	
Missing data (general)		
General recommendations	[6]	<i>Sterne et al.</i> : Recommendations for missing data and multiple imputation
Simple imputation	[36]	<i>Zhang</i> : Mean, median, mode, regression imputations
Complete-case bias considerations	[37]	<i>Bartlett et al.</i> : When CC is valid
	[38]	<i>Carpenter & Kenward</i> : When CC is valid
Multiple imputation		
Number of imputations to use	[15]	<i>White et al.</i> : at least the percentage of incomplete cases
	[39]	<i>von Hippel</i> : two-stage quadratic rule
Covariate selection procedures	[32]	<i>Wood et al.</i> : Repeated use of Rubin's rules or stacking approach
	[40]	<i>Morris et al.</i> : Adapted for MFP including selection procedure and functional form
Non-linear effects	[40]	<i>Morris et al.</i> : Adapted for MFP including selection procedure and functional form
	[41]	<i>Seaman et al.</i> : recommend just another variable (JAV) approach
Using a Cox model	[3]	<i>White & Royston</i> : inclusion of Nelson-Aalen estimate and event indicator in imputation model
	[4]	<i>Bartlett & Seaman</i> : full conditional specification adjusting for the analysis model of choice
Testing the Proportional hazards assumption and modelling time-varying effects of covariates	[5]	<i>Keogh & Morris</i> : adapting White & Royston and Bartlett & Seaman approaches for time-varying effects
Time-dependent covariates	[42]	<i>De Silva et al.</i> : Investigating performance of two-fold fully conditional specification for time-dependent covariates
	[43]	<i>Moreno-Betancur et al.</i> : Use of joint modelling for time-dependent covariates
Time-to-event features not concerning missing data		
Functional form	[44]	<i>Sauerbrei et al.</i> : multivariable fractional polynomial time i.e. MFP in survival setting accounting for time-varying effects
	[45]	<i>Buchholz & Sauerbrei</i> : comparison of procedures for assessing time-varying effects and functional form
	[46]	<i>Heinzel & Kaider</i> : Using cubic spline functions to assess functional form
	[47]	<i>Wynant & Abrahamowicz</i> : Importance of assessing time-varying effects and functional form
Covariate selection procedures	[48]	<i>Abrahamowicz & MacKenzie</i> : Joint estimation of time-varying effects and functional form using splines
	[44]	See above
Testing the Proportional hazards assumption	[49]	<i>Yan & Huang</i> : Assessing time-varying effects using an adaptive lasso method
	[35]	<i>Austin</i> : Assessing power of tests to assess proportional hazards assumption
Time-varying effects	[50]	<i>Bellera et al.</i> : Recommend assessing proportional hazards assumption and inclusion of time-varying effects where necessary
	[51]	<i>Abrahamowicz et al.</i> : use of regression splines to model time-varying effects
	[52]	<i>Hess</i> : use of cubic splines to model time-varying effects
Time-varying effects	[44]	See above
	[45]	See above
	[46]	See above
	[47]	See above

Table 4 Selected papers describing methods for addressing common issues arising in the analysis of time-to-event data when there is missing covariate data (*Continued*)

Consideration	Some recommended references	
	[48]	See above
	[49]	See above
	[50]	See above
	[52]	See above
General study considerations		
Categorising of covariates	[53]	<i>MacCallum et al.</i> : Discussion on dichotomising continuous covariates
Non-linear effects	[54]	<i>Royston & Sauerbrei</i> : Text book providing overview of model selection with a focus on MFP procedures
	[33]	<i>Harrell</i> : Text book providing overview of strategies for regression modelling
Covariate selection procedures	[54]	See above
	[55]	<i>Heinze et al.</i> : Review of methods for covariate selection

MFP: Multivariable fractional polynomials

has weaker statistical properties (estimators will generally be inconsistent) than SMC-FCS due to semi-compatibility of the imputation and analysis model. Keogh and Morris have also adapted White and Royston's method to handle time-varying effects.

Conclusions

More consideration is required for observational time-to-event analyses with missing data, including clear reporting of how the missing data were handled and how any selection procedures or assumption checks were conducted in conjunction with the missing data method implemented. Wider thought should be given to the limitations the missing data introduces to the observational study, such as bias of parameter estimates, and which methods can be used to help deal with this. While methods such as complete-case analysis are well ingrained in the community there are more modern methods which should also be considered when conducting a study. There appears to be a delay between methodology publication and uptake into the applied research field [56] or, rather, a delay in departing from simpler favoured methods of the field. There are many published guidelines readily available to help researchers conduct and report their study and these should be consulted, alongside a statistician. All recommendations that came from conducting the review were found to have already been emphasised in the published guidance discussed in the Introduction section of this paper. Finally, we recommend that journal editors have requirements for appropriate reporting in the presence of missing data to ensure high quality studies are published and that their results are robust.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12874-020-01018-7>.

Additional file 1: This contains additional information on the database search terms used, the full data extraction checklist and a list of all studies included within the review. This information is stored as a PDF file.

Additional file 2: This contains the data extraction spreadsheet, stored as a .xls file.

Abbreviations

CC: Complete-case; MCAR: Missing completely at random; MAR: Missing at random; MFP: Multivariable fractional polynomials; MI: Multiple Imputation; MICE: Multivariate imputation by chained equations

Acknowledgements

We thank Ian White for providing helpful comments on the initial results of this review and the Economic and Social Research Council Doctoral Training Partnership for funding OUC.

Authors' contributions

OUC performed data extraction and analysis, and wrote the manuscript. RHK and TPM contributed to the design of the data extraction checklist, conducted a pilot check of data extraction involving 10 papers and provided feedback on all versions of the paper. All authors read and approved the final manuscript.

Funding

OUC was funded by the Economic and Social Research Council Doctoral Training Partnership (ES/P000592/1). Tim Morris was supported by the Medical Research Council (grant numbers MC_UU_12023/21 and MC_UU_12023/29). RHK was supported by a MRC fellowship (MR/M014827/1) and a UKRI Future Leaders Fellowship (MR/S017968/1).

Availability of data and materials

The dataset supporting the conclusions of this article is included within the Additional files 1 and 2.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street, London, UK. ²MRC Clinical Trials Unit at UCL, 90 High Holborn, London, UK.

Received: 30 September 2019 Accepted: 18 May 2020

Published online: 29 May 2020

References

- Rubin DB. Multiple Imputation for Nonresponse in Surveys. United States of America: Wiley; 1987, p. 258.
- Little RJA, Rubin DB. Statistical Analysis with Missing Data, 2nd edn. United States of America: Wiley; 2002. <https://onlinelibrary.wiley.com/doi/book/10.1002/9781119013563>.
- White IR, Royston P. Imputing missing covariate values for the Cox model. *Stat Med*. 2009;28(15):1982–98. <https://doi.org/10.1002/sim.3618>.
- Bartlett JW, Seaman SR, White IR, Carpenter JR. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Stat Methods Med Res*. 2015;24(4):462–87. <https://doi.org/10.1177/0962280214521348>.
- Keogh RH, Morris TP. Multiple imputation in Cox regression when there are time-varying effects of covariates. *Stat Med*. 2018. <https://doi.org/10.1002/sim.7842>.
- Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ (Clin Res Ed)*. 2009;338:2393. <https://doi.org/10.1136/BMJ.B2393>.
- Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening of Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies for the STROBE initiative. *Lancet*. 2007;370(9596):1453–57.
- Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M, Initiative fTS. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. *PLoS Med*. 2007;4(10):297. <https://doi.org/10.1371/journal.pmed.0040297>.
- Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I, Carpenter JR, Chan A-W, Churchill R, Deeks JJ, Hróbjartsson A, Kirkham J, Jüni P, Loke YK, Pigott TD, Ramsay CR, Regidor D, Rothstein HR, Sandhu L, Santaguida PL, Schünemann HJ, Shea B, Shrier I, Tugwell P, Turner L, Valentine JC, Waddington H, Waters E, Wells GA, Whiting PF, Higgins JP. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ (Clin Res Ed)*. 2016;355. <https://doi.org/10.1136/BMJ.I4919>.
- Moher D, Liberati A, Tetzlaff J, Altman DG, Group TP. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med*. 2009;6(7):1000097. <https://doi.org/10.1371/journal.pmed.1000097>.
- Elaidi R, Harbaoui A, Beuselinc B, Eymard J-C, Bamias A, De Guillebon E, Porta C, Vano Y, Linossier C, Debruyne PR, Gross-Goupil M, Ravaud A, Aitelhaj M, Marret G, Oudard S. Outcomes from second-line therapy in long-term responders to first-line tyrosine kinase inhibitor in clear-cell metastatic renal cell carcinoma. *Ann Oncol Off J Eur Soc Med Oncol*. 2015;26(2):378–85. <https://doi.org/10.1093/annonc/mdl552>.
- Clive AO, Kahan BC, Hooper CE, Bhatnagar R, Morley AJ, Zahan-Evans N, Bintcliffe OJ, Boshuizen RC, Fysh ETH, Tobin CL, Medford ARL, Harvey JE, Van Den Heuvel MM, Lee YCG. Predicting survival in malignant pleural effusion: Development and validation of the LENT prognostic score. *Thorax*. 2014;69(12):1098–104. <https://doi.org/10.1136/thoraxjnl-2014-205285>.
- Bradshaw PT, Ibrahim JG, Stevens J, Cleveland R, Abrahamson PE, Satia JA, Teitelbaum SL, Neugut AI, Gammon MD. Postdiagnosis change in bodyweight and survival after breast cancer diagnosis. *Epidemiology*. 2012;23(2):320–7. <https://doi.org/10.1097/EDE.0b013e31824596a1>.
- Lukic M, Licaj I, Lund E, Skeie G, Weiderpass E, Braaten T. Coffee consumption and the risk of cancer in the Norwegian Women and Cancer (NOWAC) Study. *Eur J Epidemiol*. 2016;31(9):905–16. <https://doi.org/10.1007/s10654-016-0142-x>.
- White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*. 2010;30(4):377–99. <https://doi.org/10.1002/sim.4067>.
- Vogiatzoglou A, Mulligan AA, Bhaniani A, Lentjes MAH, McTaggart A, Luben RN, Heiss C, Kelm M, Merx MW, Spencer JPE, Schroeter H, Khaw K-T, Kuhnle GGC. Associations between flavan-3-ol intake and CVD risk in the Norfolk cohort of the European Prospective Investigation into Cancer (EPIC-Norfolk). *Free Radic Biol Med*. 2015;84:1–10. <https://doi.org/10.1016/j.freeradbiomed.2015.03.005>.
- Thompson EM, Hielscher T, Bouffet E, Remke M, Luu B, Gururangan S, McLendon RE, Bigner DD, Lipp ES, Perreault S, Cho Y-J, Grant G, Kim S-K, Lee JY, Rao AAN, Giannini C, Li KKW, Ng H-K, Yao Y, Kumabe T, Tominaga T, Grajkowska WA, Perek-Polnik M, Low DCY, Seow WT, Chang KTE, Mora J, Pollack IF, Hamilton RL, Leary S, Moore AS, Ingram WJ, Hallahan AR, Jouvett A, Fevre-Montange M, Vasiljevic A, Faure-Contier C, Shofuda T, Kagawa N, Hashimoto N, Jabado N, Weil AG, Gayden T, Wataya T, Shalaby T, Grotzer M, Zitterbart K, Sterba J, Kren L, Hortobagyi T, Klekner A, Laszlo B, Pocza T, Hauser P, Schuller U, Jung S, Jang W-Y, French PJ, Kros JM, van Veelen M-LC, Massimi L, Leonard JR, Rubin JB, Vibhakhar R, Chambless LB, Cooper MK, Thompson RC, Faria CC, Carvalho A, Nunes S, Pimentel J, Fan X, Muraszko KM, Lopez-Aguilar E, Lyden D, Garzia L, Shih DJH, Kijima N, Schneider C, Adamski J, Northcott PA, Kool M, Jones DTW, Chan JA, Nikolic A, Garre ML, Van Meir EG, Osuka S, Olson JJ, Jahangiri A, Castro BA, Gupta N, Weiss WA, Moxon-Emre I, Mabbott DJ, Lassaletta A, Hawkins CE, Tabori U, Drake J, Kulkarni A, Dirks P, Rutka JT, Korshunov A, Pfister SM, Packer RJ, Ramaswamy V. Prognostic value of medulloblastoma extent of resection after accounting for molecular subgroup: a retrospective integrated clinical and molecular analysis. *Lancet Oncol*. 2016;17(4):484–95. <https://doi.org/10.1016/S1470-2045%2815%2900581-1>.
- Renfro LA, Grothey A, Xue Y, Saltz LB, Andre T, Twelves C, Labianca R, Allegra CJ, Alberts SR, Loprinzi CL, Yothers G, Sargent DJ, Group ACCCA. ACCENT-based web calculators to predict recurrence and overall survival in stage III colon cancer. *J Natl Cancer Inst*. 2014;106(12). <https://doi.org/10.1093/jnci/dju333>.
- Ali HR, Dawson S-J, Blows FM, Provenzano E, Leung S, Nielsen T, Pharoah PD, Caldas C. A Ki67/BCL2 index based on immunohistochemistry is highly prognostic in ER-positive breast cancer. *J Pathol*. 2012;226(1):97–107. <https://doi.org/10.1002/path.2976>.
- Honaker J, King G, Blackwell M. Amelia II: A program for missing data. *J Stat Softw*. 2011;45(7):1–47.
- Quartagno M, Carpenter J. jomo: A Package for Multilevel Joint Modelling Multiple Imputation. 2019. <https://CRAN.R-project.org/package=jomo>. Accessed 17 Feb 2020.
- Bartlett J, Keogh R. smcfcs: Multiple Imputation of Covariates by Substantive Model Compatible Fully Conditional Specification. 2019. R package version 1.4.0. <https://CRAN.R-project.org/package=smcfcs>. Accessed 17 Feb 2020.
- Ali HR, Provenzano E, Dawson S-J, Blows FM, Liu B, Shah M, Earl HM, Poole CJ, Hiller L, Dunn JA, Bowden SJ, Twelves C, Bartlett JMS, Mahmoud SMA, Rakha E, Ellis IO, Liu S, Gao D, Nielsen TO, Pharoah PDP. Association between CD8+ T-cell infiltration and breast cancer survival in 12 439 patients. *Ann Oncol*. 2014;25(8):1536–43. <https://doi.org/10.1093/annonc/mdl191>.
- McCabe EL, Larson MG, Lunetta KL, Newman AB, Cheng S, McCabe EL, Larson MG, Lunetta KL, Newman AB, Cheng S, Murabito JM. Association of an Index of Healthy Aging With Incident Cardiovascular Disease and Mortality in a Community-Based Sample of Older Adults. *J Gerontol Ser A Biol Sci Med Sci*. 2016;71(12):1695–701.
- Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials*. 2004;1:368–76.
- Fiero MH, Huang S, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials*. 2016;17(72). <https://doi.org/10.1186/s13063-016-1201-z>.

27. Bell ML, Fiero M, Horton NJ, Hsu C-H. Handling missing data in RCTs; a review of the top medical journals. *BMC Med Res Methodol*. 2014;14(1):118. <https://doi.org/10.1186/1471-2288-14-118>.
28. Powney M, Williamson P, Kirkham J, Kolamunnage-Dona R. A review of the handling of missing longitudinal outcome data in clinical trials. BioMed Central Ltd. 2014. <https://doi.org/10.1186/1745-6215-15-237>.
29. Karahalios A, Baglietto L, Carlin JB, English DR, Simpson JA. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. *BMC Med Res Methodol*. 2012;12(1):96. <https://doi.org/10.1186/1471-2288-12-96>.
30. Altman DG, Royston P. Statistics Notes The cost of dichotomising continuous variables. *BMJ*. 2006;332(7549):. <https://doi.org/10.1136/bmj.332.7549.1080>.
31. van Maaren MC, de Munck L, Jobsen JJ, Poortmans P, de Bock GH, Siesling S, Strobbe LJA. Breast-conserving therapy versus mastectomy in T1-T2 stage breast cancer: a population-based study on 10-year overall, relative, and distant metastasis-free survival in 3071 patients. *Breast Cancer Res Treat*. 2016;160(3):511–21. <https://doi.org/10.1007/s10549-016-4012-8>.
32. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? *Stat Med*. 2008;27:3227–46. <https://doi.org/10.1002/sim.3177>.
33. HARRELL F. *Regression Modeling Strategies: with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Switzerland: Springer International Publishing; 2016.
34. VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol*. 2019;34(3):211–9. <https://doi.org/10.1007/s10654-019-00494-6>.
35. Austin PC. Statistical power to detect violation of the proportional hazards assumption when using the Cox regression model. *J Stat Comput Simul*. 2018;88(3):533–52. <https://doi.org/10.1080/00949655.2017.1397151>.
36. Zhang Z. Missing data imputation: focusing on single imputation. *Ann Trans Med*. 2016;4(1):9. <https://doi.org/10.3978/j.issn.2305-5839.2015.12.38>.
37. Bartlett JW, Harel O, Carpenter JR. Asymptotically Unbiased Estimation of Exposure Odds Ratios in Complete Records Logistic Regression. *Am J Epidemiol*. 2015;182(8):730–6. <https://doi.org/10.1093/aje/kwv114>.
38. Carpenter JR, Kenward MG. *Multiple Imputation and Its Application*: Wiley; 2013. <https://www.wiley.com/en-gb/Multiple+Imputation+and+its+Application+-+p-9780470740521>.
39. von Hippel PT. How Many Imputations Do You Need? A Two-stage Calculation Using a Quadratic Rule. *Sociol Methods Res*. 2018;004912411774730. <https://doi.org/10.1177/0049124117747303>.
40. Morris TP, White IR, Carpenter JR, Stanworth SJ, Royston P. Combining fractional polynomial model building with multiple imputation. *Stat Med*. 2015;34(25):3298–317. <https://doi.org/10.1002/sim.6553>.
41. Seaman SR, Bartlett JW, White IR. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC Med Res Methodol*. 2012;12(1):46. <https://doi.org/10.1186/1471-2288-12-46>.
42. De Silva AP, Moreno-Betancur M, De Livera AM, Lee KJ, Simpson JA. A comparison of multiple imputation methods for handling missing values in longitudinal data in the presence of a time-varying covariate with a non-linear association with time: a simulation study. *BMC Med Res Methodol*. 2017;17(1):114. <https://doi.org/10.1186/s12874-017-0372-y>.
43. Moreno-Betancur M, Carlin JB, Brilleman SL, Tanamas SK, Peeters A, Wolfe R. Survival analysis with time-dependent covariates subject to missing data or measurement error: Multiple Imputation for Joint Modeling (MIJM). *Biostatistics*. 2018;19(4):479–96. <https://doi.org/10.1093/biostatistics/kxx046>.
44. Sauerbrei W, Royston P, Look M. A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. *Biom J*. 2007;49(3):453–73.
45. Buchholz A, Sauerbrei W. Comparison of procedures to assess non-linear and time-varying effects in multivariable models for survival data. *Biom J*. 2011;53(2):308–31.
46. Heinzl H, Kaidler A. Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions. *Comput Methods Prog Biomed*. 1997;54(3):201–8. [https://doi.org/10.1016/S0169-2607\(97\)00043-6](https://doi.org/10.1016/S0169-2607(97)00043-6).
47. Wynant W, Abrahamowicz M. Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis. *Stat Med*. 2014;33(19):3318–37. <https://doi.org/10.1002/sim.6178>.
48. Abrahamowicz M, MacKenzie TA. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Stat Med*. 2007;26(2):392–408. <https://doi.org/10.1002/sim.2519>.
49. Yan J, Huang J. Model Selection for Cox Models with Time-Varying Coefficients. *Biometrics*. 2012;68(2):419–28. <https://doi.org/10.1111/j.1541-0420.2011.01692.x>.
50. Bellera CA, MacGrogan G, Debléd M, de Lara CT, Brouste V, Mathoulin-Pélissier S. Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC Med Res Methodol*. 2010;10:20. <https://doi.org/10.1186/1471-2288-10-20>.
51. Abrahamowicz M, Mackenzie T, Esdaile JM. Time-Dependent Hazard Ratio: Modeling and Hypothesis Testing With Application in Lupus Nephritis. *J Am Stat Assoc*. 1996;91(436):1432–39.
52. Hess KR. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Stat Med*. 1994;13(10):1045–62. <https://doi.org/10.1002/sim.4780131007>.
53. MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the Practice of Dichotomization of Quantitative Variables. *Psychol Methods*. 2002;7(1):19–40. <https://doi.org/10.1037/1082-989X.7.1.19>.
54. Royston P, Sauerbrei W. *Multivariable Model-building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables*; 2008, pp. 1–299. <https://doi.org/10.1002/sim.3499>.
55. Heinze G, Wallisch C, Dunkler D. Variable selection-A review and recommendations for the practicing statistician. *Biom J Biom Z*. 2018;60(3):431–49. <https://doi.org/10.1002/bimj.201700067>.
56. Binder N. The gap between development of time-to-event methods and their application in epidemiology. In: *Survival Analysis for Junior Researchers*; 2019. <https://publicisv.sund.ku.dk/~safjr2019/NadineBinderAbstract.pdf>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

