Journal of Clinical Epidemiology 142 (2022) 218–229

Journal of Clinical Epidemiology

REVIEW

Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review

SWJ Nijman[a,*], AM Leeuwenberg[a], I Beekers[b], I Verkouter[b], JJL Jacobs[b], ML Bots[a], FW Asselbergs[c,d,e], KGM Moons[a], TPA Debray[a,e]

[a] *Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Heidelberglaan 100, Utrecht, 3584 CX , The Netherlands*

[b] *Department of Health, Ortec B.V. Zoetermeer, The Netherlands*

[c] *Department of Cardiology, University Medical Center Utrecht, Utrecht University, The Netherlands*

[d] *Institute of Cardiovascular Science, Population Health Sciences, University College London, London, UK*

[e] *Health Data Research UK, Institute of Health Informatics, University College London, London, UK*

Received in revised form 1 November 2021; Accepted 10 November 2021; Available online 16 November 2021

Abstract

Objectives: Missing data is a common problem during the development, evaluation, and implementation of prediction models. Although machine learning (ML) methods are often said to be capable of circumventing missing data, it is unclear how these methods are used in medical research. We aim to find out if and how well prediction model studies using machine learning report on their handling of missing data.

Study design and setting: We systematically searched the literature on published papers between 2018 and 2019 about primary studies developing and/or validating clinical prediction models using any supervised ML methodology across medical fields**.** From the retrieved studies information about the amount and nature (e.g. missing completely at random, potential reasons for missingness) of missing data and the way they were handled were extracted.

Results: We identified 152 machine learning-based clinical prediction model studies. A substantial amount of these 152 papers did not report anything on missing data (n = 56/152). A majority (n = 96/152) reported details on the handling of missing data (e.g., methods used), though many of these (n = 46/96) did not report the amount of the missingness in the data. In these 96 papers the authors only sometimes reported possible reasons for missingness (n = 7/96) and information about missing data mechanisms (n = 8/96). The most common approach for handling missing data was deletion (n = 65/96), mostly via complete-case analysis (CCA) (n = 43/96). Very few studies used multiple imputation (n = 8/96) or built-in mechanisms such as surrogate splits (n = 7/96) that directly address missing data during the development, validation, or implementation of the prediction model.

Conclusion: Though missing values are highly common in any type of medical research and certainly in the research based on routine healthcare data, a majority of the prediction model studies using machine learning does not report sufficient information on the presence and handling of missing data. Strategies in which patient data are simply omitted are unfortunately the most often used methods, even though it is generally advised against and well known that it likely causes bias and loss of analytical power in prediction model development and in the predictive accuracy estimates. Prediction model researchers should be much more aware of alternative methodologies to address missing data.

Keywords: Missing data; Machine learning; prediction; reporting; literature review

Conflict of Interest: All authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

* Corresponding author: Telephone: +31-(0)88-75 6801

E-mail address: s.w.j.nijman@umcutrecht.nl (S. Nijman).

https://doi.org/10.1016/j.jclinepi.2021.11.023

What is new?

Key findings

- Prediction model studies that adopt machine learning (ML) methods rarely report the presence and handling of missing data.
- Although many types of machine learning methods offer built-in capabilities for handling missing values, these strategies are rarely used. Instead, most ML-based prediction model studies resort to complete case analysis or mean imputation.

What this adds to what is known?

- Missing data are often poorly handled and reported, even when adopting advanced machine learning methods for which advanced imputation procedures are available.

What is the implication, and what should change now?

- The handling and reporting of missing data in prediction model studies should be improved. A general recommendation to avoid bias is to use multiple imputation. It is also possible to consider machine learning methods with built-in capabilities for handling missing data (e.g., decision trees with surrogate splits, use of pattern submodels, or incorporation of autoencoders).
- Authors should take note of and appreciate the existing reporting guidelines (notably, TRIPOD and STROBE) when publishing ML-based prediction model studies. These guidelines offer a minimal set of reporting items that help to improve the interpretation and reproducibility of research findings.

1. Introduction

Thorough contemplation about the handling and reporting of missing data is an integral part of any research addressing and using clinical data, including clinical prediction model research [1–6]. Clinical prediction models use multiple input variables (i.e., covariates, predictors) to calculate the absolute risk of a specific outcome presence (diagnostic models) or incidence (prognostic models). In the medical literature, most diagnostic and prognostic prediction models are derived or validated using regression modelling strategies. When missing values are present in the model development or validation sample, additional efforts preparatory to model development are required.

The most common approach is to adopt a complete-case analysis (CCA), wherein individuals with missing data on any of the predictor or outcomes variables are (automatically) deleted from the analysis [7,8]. Although this strategy is (only) valid under very stringent circumstances, it is generally inefficient and can lead to severe bias in estimates of the estimated model parameters (e.g., regression coefficients) and thus in the model's predictive performance [3,9,10]. For example, removing incomplete cases could lead to loss of a significant number of informative observations.

For this reason, it is generally recommended to implement multivariable imputation models that generate multiple imputations conditionally on other (observed) patient characteristics [9–13]. When multiple imputation is used during prediction model development, multiple completed versions of the incomplete datasets are generated in which the prediction model coefficients are estimated separately. The model coefficients from each imputed dataset are then pooled using Rubin's rules, and subsequently used for calculating absolute risk probabilities in new patients [10,11]. Although multiple imputation strategies are consequently applied to an entire prediction model development or validation dataset, it is possible to generate imputations tailored to individual patients [14,15]. This also makes it possible use multiple imputation techniques when actually implementing and applying prediction models in electronic healthcare software in daily clinical practice [13–16].

Yet another approach is to address missing data directly during the prediction model development, validation, or application. This strategy can, for instance, be achieved by including missing indicator variables, by adopting pattern-mixture models, tree-based ensembles, or other machine learning (ML) methods that circumvent the use of missing data imputation (Box 1) [17–22].

Existing prediction model reporting guidelines (TRIPOD), congruent with the increasing amount of supportive literature, recommend to at least report whether prediction model development sets and validation sets indeed suffered the presence of missing data and to what extent, and how such missing data were addressed in the analysis [1,2,10,23,24]. So far, adherence to these reporting guidelines seems to be limited in applied prediction research. Even in prediction model studies that adopt more traditional (regression-based) methods, many reviews have found that missing data is often inadequately handled or completely ignored [25–30].

With the emergence of ML methods for prediction modeling, which may circumvent the need for imputation (e.g., random forests with surrogate splits), it becomes less evident whether and how missing data is handled during model development or validation. The question remains how often researchers adopting these ML methods make use of alternative and proper strategies and in what way. The objective of this study is, therefore, to investigate how well prediction model studies that used ML based techniques reported on the presence, nature and extent of missing data in the used data sets, and which methods were commonly used for handling missing data during prediction model development, validation, or (if done) implementation.

2. Methods

In a recent review by Andaur Navarro et al. we systematically searched the medical literature for primary studies developing and/or validating prediction models using any supervised ML methodology, published between January 2018 and December 2019 [31,32]. The protocol of which was registered and published (PROSPERO, CRD42019161764) [33]. The search initially yielded 24.814 results, from which 10 random sets of 249 articles were sampled. From the sampled 2.482 publications, 152 were included in the review. The present review uses the same data set of this review (Fig. 1). Similarly for the present review, articles were eligible for inclusion when a primary study described the development or validation of a multivariable prediction model using any kind of supervised ML methodology. We defined a study using supervised ML as the use of algorithmic approaches to develop or validate a prediction model (e.g., any tree-based methods, neural networks, or support vector machines). We excluded studies that adopted common statistical techniques such as linear regression, logistic regression, lasso regression, ridge regression, or elastic net. Also, studies were excluded when only a single variable was studied. All human medical fields, with the notable exception of medical imaging, were included. To address the aim of the present review, first, a list of key reporting items that may facilitate the interpretation of prediction model studies in the presence of missing data, were defined (Table 1). These items were based on prevailing reporting guidelines [1–3,10] and consider:

1) Information on the presence, amount, and distribution of missingness on the study variables, including reasons for the missing data and assumptions about the missing data mechanism;
2) Methods for missing data handling, including the type (e.g., imputation, missing indicator, surrogate splits);
3) Implementation details of the missing data method, including total number of imputed datasets and (auxiliary, i.e. not part of the prediction model) variables used in the imputation models (Table 1).

Existing machine learning reporting guidelines sparsely refer to the need to report on missing data details [34]. As a consequence, items specifically about the ML modeling techniques were based on key characteristics of known ML methods with built-in strategies to handle missing data [17–20]. Subsequently, we reviewed each eligible study and assessed whether missing data was present. For studies that reported the presence of missing data, we evaluated the level of reporting of the items listed in Table 1. If applicable, data extraction was done both for the prediction model development and validation. When a sensitivity analysis was utilized, applied methods for handling missing data in these sensitivity analyses were also assessed separately. Supplementary material was considered when available. Ten percent of the total set was reviewed first by two reviewers (S.N., A.L.), in which disagreements were resolved for mutual learning by discussing the found discrepancies. The two reviewers then independently reviewed fifty percent of all studies respectively. Unresolved disagreements were resolved through consensus with a third reviewer (T.D.). All items used in the data extraction can be found in the Appendix. For the data extraction some reporting items (e.g., Item 2.1) about identifying and handling missing data from Table 1 were split up into several separate data extraction items.

3. Results

After screening, 152 eligible articles were available for the present study (Fig. 1). A total of 56 (37%) prediction model studies did not report on missing data and could not be analyzed further. We included 96 (63%) studies which reported on the handling of missing data. Across the 96 studies, 46 (48%) did not include information on the amount or nature of the missing data.

3.1. Presence and mechanism of missing data

Papers that reported on the amount of missing data most often (n = 31/50 [62%]) reported the overall number or frequency of missingness (e.g., the total number of patients or variables with one or more missing values). For these papers, the overall median percentage of missingness was 4.7% (IQR 1.85-28). In most other cases it was unclear how many values were missing. It was often unclear which variables exactly were missing (n = 39/50 [78%]). In 7 papers it was explicitly stated that the outcome was missing [14%]. Only a small proportion of papers provided possible reasons for missingness of predictor values (n = 7/50 [14%]) or compared the characteristics of patients with and without any missing values (n = 5/50 [10%]). Additionally, a statement about the (potential) mechanism by which the data were missing was seldom reported (n = 8/50 [16%]).

3.2. Handling of missing data

From the 96 papers reporting on missing data handling, the most common approach was deletion (n = 65/96 [68%]), with the majority using complete case analysis (CCA) (n = 43/65 [66%]). About a third of papers reporting on missing data handling, used imputation (n = 36/96 [38%]), most often single imputation (23/36 [61%]) with the mean (12/23 [52%]). Only a handful used the recommended multiple imputation (n = 8/36 [22%]). Of these 8 papers, important details such as the number of imputed datasets, whether predictor and outcome variables were included in the imputation models, exact imputation method applied, or whether auxiliary variables were used, was only rarely reported (1–3 papers). Missing indicators were used by some authors (n = 8/96 [8%]), most often in combination with any deletion or imputation method

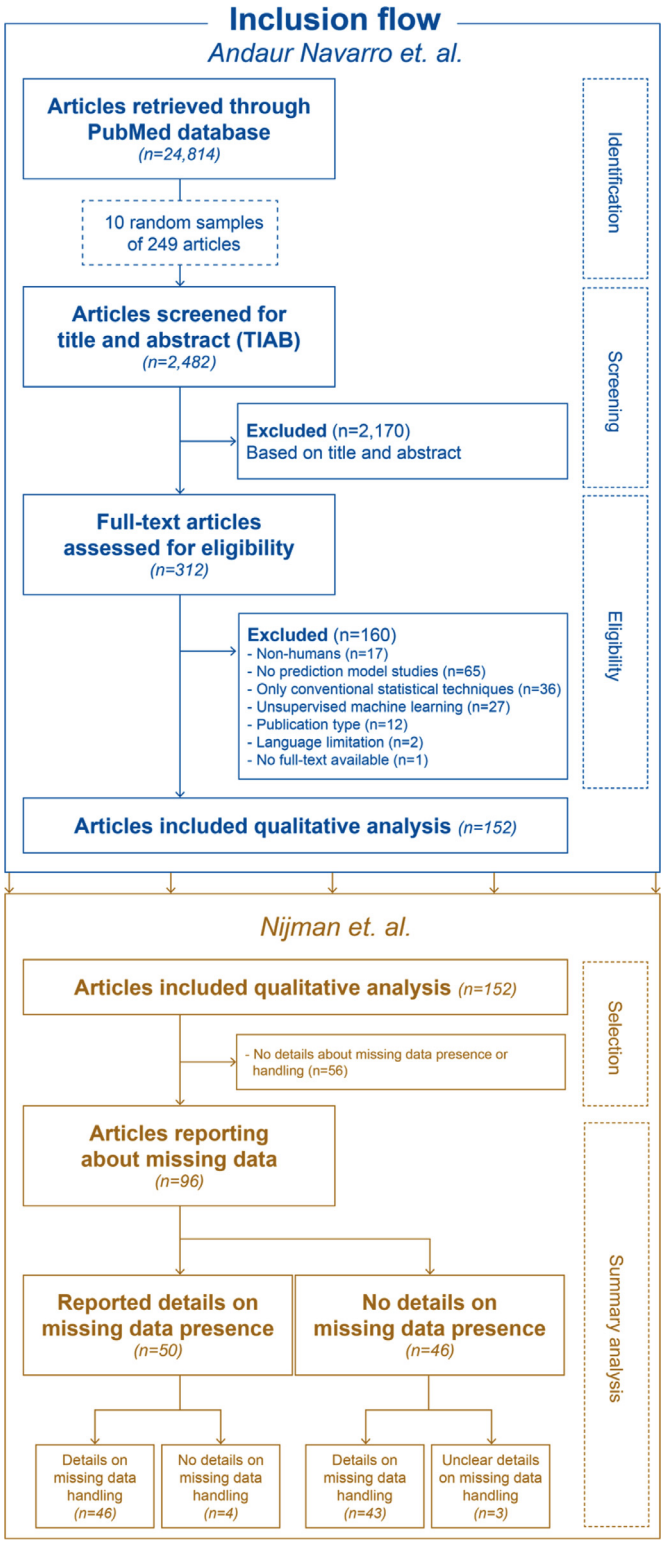

Fig. 1. Inclusion flow continuation after systematic review. For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.

Table 1. Missing data details recommended to be reported in prediction model studies

	Details to be reported		Inspired by
1.0 Missing data	1.1 For each variable of interest (e.g., candidate predictor, outcome): The amount of missing data or the number of cases with (in)complete data		[1]
	1.2 Potential reasons for the presence of missing data		[1–3]
	1.3 Guidance on how the prediction model should be implemented in new patients (i.e., how to deal with 'live' missing values)		Expert opinion
2.0 Missing data handling details	2.1 The type of method used to account for missing data • Deletion (e.g., case-wise deletion, complete-case analysis) *Methods which omit part of the data to allow for analysis* • Imputation-based approach (e.g., single or multiple imputation) *Methods which fill-in plausible estimates for missing data* • Non-imputation-based approach (e.g., missing indicator, surrogate splits) *Methods which provide predictions without imputing missing data by taking note of missing data in various ways*		[2,3]
	2.2 If complete case analysis was performed, the number of individuals excluded, e.g., in a diagram depicting the participant flow (e.g., 'CONSORT' participant flow diagram)		[1,3]
	2.3 If complete case analysis was performed, a rationale for exclusion		[1,3]
	2.4 Comparison of overall patient characteristics of patients with and patients without missing values		[3]
	2.5 If possible, results of complete case analysis (to compare) and their interpretation		[3]
	2.6 If software was applied (e.g., for imputation-based or non-imputation-based approaches), provide details on software and key settings of the approach (e.g., packages used), supplementary material allowed		[2,3]
3.0 Imputation-based approaches	3.1 Type of imputation • Single imputation (SI) (3.9) • Multiple imputation (MI) (3.10)		[2,3]
	3.2 Explicit mentioning of the assumptions that were made (e.g., MAR, MCAR or MNAR)		[1,3]
	3.3 Motivation for the assumptions made (3.3.1) or inclusion of sensitivity analyses for testing robustness (3.3.2)		[3]
	3.4 Details of the adjustment for statistical interactions (3.4.1), non-linear terms (3.4.2), and clustering (3.4.3) in the imputation model		[2,3]
	3.5 Details on how continuous and non-continuous variables were imputed		[2,3]
	3.6 Details on what variables were included in the imputation procedure		[2,3]
	3.7 Inclusion of outcome as variable in the imputation procedure		[2]
	3.8 Inclusion and details of auxiliary variables in the imputation procedure		Expert opinion
	3.9 Single imputation details	3.9.1 Type of SI used (e.g., mean imputation)	[2]
		3.9.2 Details on if method takes into account noise or imputes a fixed value	[2]
	3.10 Multiple imputation details	3.10.1 Type of MI used (e.g., FCS or joint imputation)	[2,3]
		3.10.2 Number of imputed datasets	[2,3]
		3.10.3 Details on conditional models used (e.g., PMM, Random Forest, logistic regression, neural network, machine learning, etc.)	[2,3]
		3.10.4 Details on convergence of the imputation model	[2,3]
4.0 Non-imputation-based approaches	4.1 Type of non-imputation-based method • Missing indicator method • Likelihood-based methods (e.g., using expectation-maximization) • Use of submodels (4.3) • ML method (e.g., decision trees with surrogate splits) (4.4) • Other		Expert opinion

(*continued on next page*)

Table 1 (*continued*)

	Details to be reported		Inspired by
	4.2 If missing indicator method was used, details on how missing indicators were included in the prediction model		Expert opinion
	4.3 Submodels details	4.3.1 Type of submodel used (e.g., pattern mixture kernel submodels)	Expert opinion
		4.3.2 The total number of developed submodels	Expert opinion
		4.3.3 Details on how each submodel is derived (e.g., in completed data or in a missing data pattern specific subset of data)	Expert opinion
	4.4 ML method details	4.4.1 Type of ML method used	Expert opinion
		4.4.2 Details on the relevant (hyper)parameterization (e.g., range, selection method for configuration, specification of parameters)	Expert opinion
		4.4.3 Details on how missing data are handled	Expert opinion

SI, single imputation; MI, multiple imputation; ML, machine learning; MAR, missing at random; MCAR, Missing completely at random; MNAR, Missing not at random; PMM, predictive mean matching; FCS, full conditional specification.

(n = 6/8 [75%]). Many studies used a type of prediction model development or validation (e.g., random forest) capable of handling missing data via built-in mechanisms (n = 77/152 [51%]). Few articles explicitly stated that the machine learning method could handle missing data via built-in mechanisms (n = 13/77 [17%]), this concerned almost exclusively tree-based models.

There were many studies (n = 23/96 [24%]) where a combination of missing data handling methods was used, most often combining deletion practices with imputation methods (n = 15/23 [65%]). Only sometimes were these reported as sensitivity analyses (n = 3/23 [13%]). There were no studies in which a submodel approach was used.

A complete overview of the extracted data can be found in the Appendix.

4. Discussion

This work comprised a comprehensive review of 152 ML-based clinical prediction model development or validation studies, to evaluate the reporting and methodological quality with regards to the presence, amount, and handling of missing data in such studies. Consistent with similar reviews on the reporting of prediction models or missing data, the quality of reporting in ML-based prediction model studies with regards to missing data was generally poor. This makes the judgement of the validity of the reported prediction models or their predictive accuracy difficult or even impossible [25,35]. Examples of common pitfalls in the handling of missing data largely match that of similar reviews which analyzed studies reporting on prevailing statistical models: the exclusion of study participants with any missing data and a lack of primary details on the amount or nature of the missing data, and the imputation methods used, if done (Fig. 2).

Methods such as CCA and single imputation, often via mean imputation (52%), were highly common in the ML studies included in this review. It can seem efficient to apply methods such as mean imputation or CCA, but it is generally expected that these ad-hoc methods are unfit for working with healthcare data [7,11,13,36]. Only under stringent circumstances to which healthcare data, and certainly not routine healthcare data, usually do not abide, mean imputation and CCA could provide unbiased estimates. Similarly, there are strong recommendations to avoid the use of missing indicators, for example because it may alter the way clinicians approach the use of a predictive model, given that the model suggests missing data may also be informative [7,22,36,37]. Likewise, missing indicators require continued monitoring and dynamic revision for the various different missing data circumstances upon which they may be used, which is incredibly convoluted when applied in a medical decision-making context [38]. Surprisingly, this method is often used by studies using a non-imputation-based approach (53%). This tendency in combination with frequent absence of explicit motivations for choosing certain missing data handling strategies and sparse reference to missing data in existing machine learning reporting guidelines, illustrate an overall lack of appreciation about the severe consequences of improper handling of missing data in prediction model studies and also in clinical decision making based on prediction models.

Overall, there is clearly room for improvement in the strategies for handling missing values of the prediction model studies adopting state-of-the-art ML methods. Although multiple imputation is currently considered the gold standard, it is only rarely implemented in these published studies (8/152 [5%]). In addition, several alternative strategies (e.g., pattern-mixture models, surrogate splits, etc.) are available that circumvent the need for imputation. These

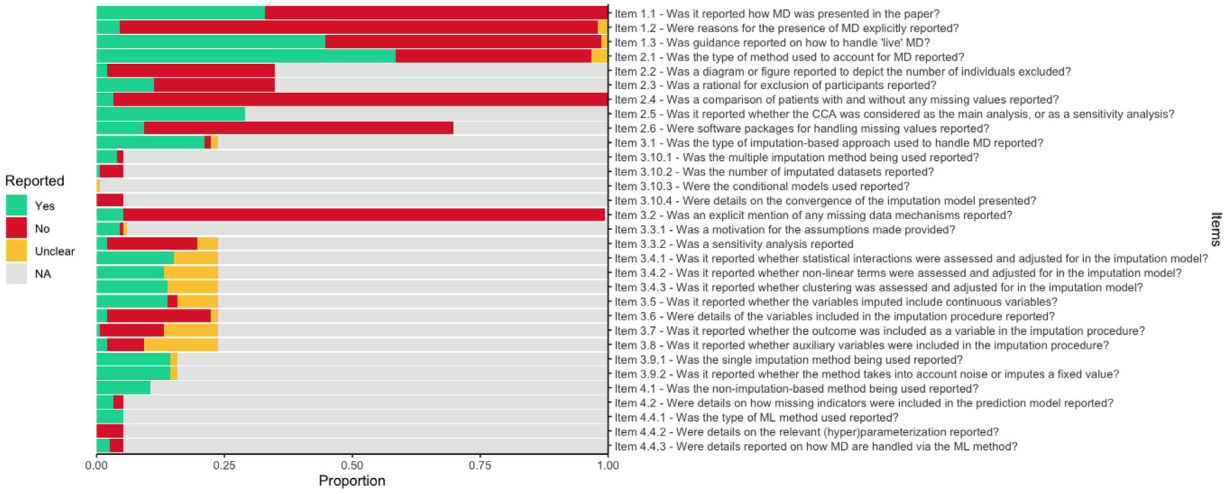

Fig. 2. Overview of missing data details reported on
Item 4.3.1., 4.3.2 and 4.3.3 are not shown as no study included the use of submodels. For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.

strategies may be particularly appealing to enhance the development, validation and implementation of developed prediction models, as they offer a unified approach to generate predictions in the presence of missing data. Still, among these approaches, it is yet unclear which is to be preferred, and consensus about their effectivity when compared with, more classical, missing data handling methods is lacking; more research on this is warranted [18,19,39].

The level of reporting is arguably just as important as the quality of an imputation model. Sufficient detail to be able to replicate the study is a key obligation of scientific research and reporting. Almost all studies that used multiple imputation lacked sufficient detail on which variables were included, the conditional imputation models used, and the number of multiple imputed datasets. Also, the limited utilization of sensitivity analyses suggests that authors did not consider the potential consequences of handling missing data much. Further, the lack of detail on which variables were included in the imputation model suggests that known extensions that can improve the accuracy of the imputation model (e.g., use of auxiliary variables) are unexploited [15,40]. To promote good missing-data-handling-practice, we echo previous recommendations to acknowledge sufficient reporting on missing data and any applied missing data handling method, to allow others to interpret the quality of the results, to allow for their replication and to enhance the application of the prediction model [10,25,26]. Furthermore, journals are encouraged to ask for these details to be published in the original text or as supplementary files.

Many included papers used prediction models based on decision trees or random forests, for which built-in capabilities exist for handling missing data during its development, validation and implementation [17,18]. Most authors, however, did not clarify whether and how these were used. It is possible that many authors used the default way of handling missing data as programmed for these models, i.e, usually CCA. However, due to the limited inclusion of programming details (i.e., code, libraries and packages) it remains largely uncertain how often these methods were used. The implementation of automated or built-in missing data handling methods is rare in software packages, which may explain their underreported use. Another possibility is that these built-in methods are taken for granted, which again suggests that there may be an overall lack of knowledge about the consequences of improper missing data handling. There is generally no consensus on how well these built-in methods work with regards to clinical prediction model development, validation or implementation, which warrants additional research and caution when using them in the presence of missing data [18,19,39].

A limitation of our review may be related to the restricted search strategy from the original review, as only articles published in PubMed over a time span of two years (between January 2018 and December 2019) were considered and only a subsample (n = 2.482) from the initial search results (n = 24.814) was screened [33]. However, we believe that even with these restrictions the final study sample remains representative of the current status in the field, since no recent reporting or methods guideline were likely issues that may have caused any improvements since then.

To our knowledge, this is the first comprehensive review evaluating the level of reporting and handling of missing data in ML-based clinical prediction model studies. We believe this review of a representative sample of model development and validation prediction model studies in healthcare has highlighted severe issues with the general

conduct and reporting of missing data in ML-based prediction model studies. It is well known that inappropriate handling of missing data can greatly reduce the validity and generalizability of predictions and corresponding estimates of prediction model performance [1,5]. An improved understanding about the negative consequences of inappropriate handling of missing data and effective ways to remedy these issues through improved conduct and reporting is warranted. We recommend authors to take note of and appreciate the existing reporting guidelines (notably, TRIPOD and STROBE) when publishing ML-based prediction model studies. These guidelines include a minimal set of reporting items that help to improve the interpretation and reproducibility of research findings.

Box 1 Prediction with built-in missing data handling

Missing indicator. For each variable in the model a dichotomous dummy variable (0/1) is added to indicate whether that variable is missing or not [7,22,36,41]. These dummy variables are then included in the statistical (i.e., risk prediction) model as separate predictors. The original, missing, predictor variable is usually set to 0. Missing indicators may contain relevant information for predictions, but are susceptible to so-called feedback loops; as soon as a clinician is aware of the informative missingness in certain predictors, their predictive value changes [37,38,42]. Additionally, other issues may arise in the application of missing indicators as the manner of data collection between different practices is likely to vary [38].

Surrogate splits. Preserves the partitioning of each original split as good as possible in the presence of missing predictor values [18–20]. Accordingly, the model, whenever it encounters a missing predictor value, will use the surrogate variable (rather than the missing predictor variable) to decide upon the split direction.

Sparsity aware splitting. A default direction is added for each tree node in a decision tree (e.g., XGBoost) [17]. Whenever a missing predictor value is encountered, the instance is classified into the prespecified default direction. The optimal default direction, and thus best direction to handle missing data, is learnt from the data.

Pattern-mixture models. For each pattern of missing data, a separate risk prediction model is made and included in the pattern-mixture model [21]. Then, when applied to a new (out-of-sample) individual the corresponding (i.e., matching the missing data pattern in the individual) prediction model is used.

CRediT authorship contribution statement

SWJ Nijman: Conceptualization, Methodology, Investigation, Writing – original draft, Visualization. **AM Leeuwenberg:** Investigation, Writing – review & editing. **I Beekers:** Writing – review & editing. **I Verkouter:** Writing – review & editing. **JJL Jacobs:** Writing – review & editing. **ML Bots:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition. **FW Asselbergs:** Conceptualization, Writing – review & editing. **KGM Moons:** Supervision, Conceptualization, Writing – review & editing, Project administration. **TPA Debray:** Validation, Conceptualization, Methodology, Supervision, Writing – review & editing.

Disclosures

The UCC is primarily financed by the UMC Utrecht. A grant from the Netherlands Organization for Health Research and Development (#8480-34001) was obtained to develop feedback procedures. UCC website: www.umuctrecht.nl/ucc (in Dutch). Contact information UCC: ucc@umcutrecht.nl.

SWJN is supported by a Public-Private Study grant of the Netherlands Heart foundation for the CVRM-IMPROVE project (#2018B006). This Research Project is financed by the PPP Allowance made available by Top Sector Life Sciences & Health to Netherlands Heart Foundation to stimulate public-private partnerships. TPAD is supported by the Netherlands Organisation for Health Research and Development (#91617050).

Data availability statement

The data that support the findings of this study are available from upon reasonable request.

Acknowledgments

This study was conducted on behalf of the Utrecht Cardiovascular Cohort- CardioVascular Risk Management (UCC- CVRM) study group. Members of the UCC-CVRM Study group: F.W. Asselbergs, Department of Cardiology; G.J. de Borst, Department of Vascular Surgery; M.L. Bots (chair),Julius Center for Health Sciences and Primary Care; S. Dieleman, Division of Vital Functions (anesthesiology and intensive care); M.H. Emmelot, Department of Geriatrics; P.A. de Jong, Department of Radiology; A.T. Lely, Department of Obstetrics/Gynecology; I.E. Hoefer, Laboratory of Clinical Chemistry and Hematology; N.P. van der Kaaij, Department of Cardiothoracic Surgery; Y.M. Ruigrok, Department of Neurology; M.C. Verhaar, Department of Nephrology & Hypertension, F.L.J. Visseren, Department of Vascular Medicine, University Medical Center Utrecht and Utrecht University

We are grateful to the authors of the original review for the search conducted.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jclinepi.2021.11.023.

Appendix

Details of missingness (n = 152)

#	Item	Total (%)
1.1	How was missing data presented in the paper?	
	Not summarized	102 (67%)
	Overall	31 (20%)
	By all candidate predictors	8 (5%)
	By all final predictors	3 (2%)
	Other	8 (5%)
1.2	Were reasons for the presence of missing data explicitly reported?	
	Yes	7 (5%)
	No	142 (93%)
	Unclear	3 (2%)
1.3	Was guidance provided on how to handle 'live' MD? (i.e., how to apply the prediction models in new patients with MD)	
	Yes, explicitly	7 (5%)
	Yes, implicitly (e.g., mean imputation)	61 (40%)
	No	82 (5%)
	Unclear	2 (1%)
1.4	Was a comparison of patient characteristics for patients without any missing values, and patients with one or more missing values made?	
	Yes	5 (3%)
	No	147 (97%)

Legend: MD: missing data, CCA: complete-case-analysis.

Details of missing data handling (n = 152)

#	Item	Total (%)
2.1	Was the type of method used to account for MD reported?	
	Yes	89 (59%)
2.2	*If yes, what was the method being used?*	
	Deletion (i.e., CCA)	44 (47%)
	Imputation-based	16 (17%)
	Non-imputation-based	7 (7%)
	A combination of the above	23 (25%)
	A combination of deletion and imputation	15 (65%)
	A combination of deletion and non-imputation	3 (13%)
	A combination of imputation and non-imputation	2 (9%)
	A combination of all three methods	3 (13%)
	Unclear	4 (4%)
	No	58 (38%)
	Unclear	5 (3%)
2.3	Is there evidence to suggest the developed prediction model can handle the presence of missing data?	
	Yes / probably yes	13 (9%)
	No / probably no	75 (49%)
	Unclear	64 (42%)
2.4	Was an explicit mention of any missing data mechanisms given?	
	Yes	8 (5%)
2.5	*Was a motivation for the assumptions made provided? (i.e., missing data mechanisms)*	
	Yes	7 (88%)
	Unclear	1 (13%)
	No	144 (95%)

Reported details on deletion (n = 65)

#	Item	Total (%)
3.1	Were results of a CCA presented?	
	Yes	44 (68%)
3.2	*Was the CCA considered as the main analysis, or as a sensitivity analysis?*	
	Main analysis	42 (96%)
	Sensitivity analysis	2 (5%)
	No	18 (28%)
	Unclear	3 (5%)
3.3	Was a diagram or figure used to depict the number of individuals excluded (e.g., participant flow diagram)?	
	Yes	3 (5%)
	No	62 (95%)
3.4	Was an explicit rationale for exclusion of participants reported?	
	Yes	17 (26%)
	No	48 (74%)

Reported details on imputation (n = 36)

#	Item	Total (%)
4.1	Was the type of imputation-based approach reported?	
	Yes	32 (89%)

(continued on next page)

#	Item	Total (%)
4.2	*What was the imputation method being used?*	
	Single imputation	23 (72%)
	Multiple imputation	8 (25%)
	Unclear	1 (3%)
	No	2 (6%)
	Unclear	2 (6%)
4.3	Was a sensitivity analysis performed?	
	Yes	3 (8%)
	No	27 (75%)
	Unclear	6 (17%)
4.4	Were statistical interactions assessed and adjusted for in the imputation model?	
	Yes	2 (6%)
	No	21 (58%)
	Unclear	13 (36%)
4.5	Were non-linear terms assessed and adjusted for in the imputation model?	
	Yes	1 (3%)
	No (non-linear terms were assessed in the main analysis, but not adjusted for during imputation)	2 (6%)
	No (non-linear terms were not assessed in the main analysis and not adjusted for during imputation)	17 (47%)
	Unclear	16 (44%)
4.6	Was clustering assessed and adjusted for in the imputation model?	
	Yes	1 (3%)
	No	20 (56%)
	Unclear	15 (42%)
4.7	Did the variables imputed include continuous variables?	
	Yes / probably yes	21 (58%)
4.8	*Was it described how these were modelled?*	
	Linear	1 (5%)
	Non-linear	3 (14%)
	Categorized	2 (10%)
	Not reported	15 (71%)
	No	3 (8%)
	Unclear	12 (33%)
4.9	Was any other preprocessing performed?	
	Standardization / normalization	10 (28%)
	Outlier removal	2 (6%)
	Not reported	2 (6%)
	Unclear	16 (44%)
	No	6 (17%)
4.10	Were details of the variables included in the imputation procedure presented?	
	Yes	3 (8%)
4.11	*Was a motivation for the inclusion of variables in the imputation procedure provided?*	
	No	3 (100%)
	No	31 (86%)
	Unclear	2 (6%)
4.12	Was the outcome included as a variable in the imputation procedure?	
	Yes	1 (3%)
	No / probably no	19 (53%)
	Unclear	16 (44%)
4.13	Were auxiliary variables included in the imputation procedure?	
	Yes	3 (8%)
4.14	*Were any details on auxiliary variables used presented?*	
	No	3 (100%)
	No / probably no	11 (31%)
	Unclear	22 (61%)

Reported details on single imputation (n = 23)

#	Item	Total (%)
5.1	What is the single imputation method being used?	
	Mean / median imputation	12 (52%)
	K-nearest neighbor imputation	3 (13%)
	Combination of imputation methods	2 (9%)
	Regression method	1 (4%)
	Random forest imputation	1 (4%)
	Last observation carried forward	1 (4%)
	Unclear	2 (9%)
5.2	Does the method take into account noise or impute a fixed value?	
	Fixed value	21 (91%)
	Unclear	2 (9%)

Reported details on multiple imputation (n = 8)

6.0	Multiple imputation details	
6.1	What is the multiple imputation method being used?	
	Predictive mean matching	2 (25%)
	MissForest	2 (25%)
	Full conditional specification	1 (13%)
	Which conditional models were used?	
	Unclear	1 (100%)
	Bayesian ridge regression	1 (13%)
	Unclear	2 (25%)

(*continued on next page*)

6.0	Multiple imputation details	
6.2	Was the number of imputed datasets reported?	
	Yes	1 (13%)
	No	7 (88%)
6.3	Were details on the convergence of the imputation model presented?	
	No	8 (100%)

Reported details on non-imputation-based approaches (n=15)

#	Item	Total (%)
7.1	Was the non-imputation-based method implicitly or explicitly reported as capable of handling MD?	
	Explicit	11 (73%)
	Implicit	4 (27%)
7.2	What is the non-imputation-based method being used?	
	Missing indicator method	8 (53%)
7.3	*Were details on how missing indicators were included in the prediction model reported?*	
	Yes	5 (63%)
	No	3 (38%)
	Machine learning method	7 (47%)
7.4	*What was the type of ML method used?*	
	Tree-based (e.g., random forest)	6 (86%)
	Bayesian network	1 (14%)
7.10	*Are details provided on how MD are handled via the ML method? (e.g., Imputation)*	
	Yes	3 (43%)
	No	4 (57%)

References

[1] Vandenbroucke JP, Poole C, Schlesselman JJ, Egger M. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration. PLoS Med 2007;4(10):27.

[2] Collins GS, Reitsma JB, Altman DG, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med 2015;13(1):1.

[3] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338(jun29 1):b2393.

[4] Groenwold RHH, Moons KGM, Vandenbroucke JP. Randomized trials with missing outcome data: how to analyze and what to report. CMAJ 2014;186(15):1153–7.

[5] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med 2015;162(1):W1.

[6] Little RJ, Emerson SS, Hogan JW, Molenberghs G, Neaton JD, Shih WJ. The Prevention and Treatment of Missing Data in Clinical Trials. N Engl J Med 2012;6.

[7] Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. J Clin Epidemiol 2006;59(10):1087–91.

[8] Little RJA. Rubin:Statistical Analysis with Missing Data, XIV+278. New York: Wiley; 1987.

[9] Harel O, Mitchell EM, Perkins NJ, Cole SR, Tchetgen Tchetgen EJ, Sun B, et al. Multiple imputation for incomplete data in epidemiologic studies. Am J Epidemiol 2018;187(3):576–84.

[10] Van Buuren S. Flexible Imputation of Missing Data. 2nd ed. CRC Press; 2018.

[11] Janssen KJM, Donders ART, Harrell FE, Vergouwe Y, Chen Q, Grobbee DE, et al. Missing covariate data in medical research: To impute is better than to ignore. J Clin Epidemiol 2010;63(7):721–7.

[12] Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. Int J Epidemiol 2019;48(4):1294–304.

[13] Janssen KJM, Vergouwe Y, Donders ART, Harrell FE, Chen Q, Grobbee DE, et al. Dealing with missing predictor values when applying clinical prediction models. Clin Chem. 2009;55(5):994–1001.

[14] Nijman S, Groenhof T, Hoogland J, Bots M, Brandjes M, Jacobs J, et al. Real-time handling of missing predictor values when implementing and using prediction models in daily practice. JCE 2021 Article in press.

[15] Nijman SWJ, Hoogland J, Groenhof TKJ, Brandjes M, Jacobs JJL, Bots ML, et al. Real-time imputation of missing predictor values in clinical practice. Eur Heart J - Digit Health 2021;2(1):154–64.

[16] Hoogland J, Barreveld M, Debray TPA, Reitsma JB, Verstraelen TE, Dijkgraaf MGW, et al. Handling missing predictor values when validating and applying a prediction model to new patients. Stats Med 2020 sim.8682.

[17] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13. p. 785–94.

[18] Feelders A. Handling Missing Data in Trees: Surrogate Splits or Statistical Imputation?. In: Żytkow JM, Rauch J, editors. Principles of Data Mining and Knowledge Discovery [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 1999. p. 329–34. Accessed 27 July 2021.

[19] Hapfelmeier A. Analysis of missing data with random forests [Internet]. 2012 [cited 2019 Sep 4]. 6–7 p. Available from: https://edoc.ub.uni-muenchen.de/15058/1/Hapfelmeier_Alexander.pdf, Accessed 27 July 2021

[20] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees [Internet]. Taylor & Francis; 1984. Accessed 27 July 2021.

[21] Fletcher Mercaldo S, Blume JD. Missing data and prediction: the pattern submodel. Biostatistics 2020;21(2):236–52.

[22] Groenwold RHH, White IR, Donders ART, Carpenter JR, Altman DG, Moons KGM. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. Can Med Assoc J 2012 Aug 7;184(11):1265–9.

[23] Lee KJ, Tilling K, Cornish RP, Little RJ, Bell ML, Goetghebeur E, et al. Framework for the treatment and reporting of missing data in observational studies: The TARMOS framework. arXiv:200414066 [stat] [Internet]. 2020 [cited 2020 Oct 6]; Available from: http://arxiv.org/abs/2004.14066. Accessed 27 July 2021.

[24] Kerr KF, Meisner A, Thiessen-Philbrook H, Coca SG, Parikh CR. RiGoR: reporting guidelines to address common sources of bias in risk model development. Biomark Res 2015;3(1):2.

[25] Collins GS, Omar O, Shanyinde M, Yu L-M. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. J Clin Epidemiol 2013;66(3):268–77.

[26] Tsvetanova A, Sperrin M, Peek N, Buchan I, Hyland S, Martin GP. Missing data was handled inconsistently in UK prediction models: a review of method used. J Clin Epidemiol 2021 S0895435621002882.

[27] Galbete A, Tamayo I, Librero J, Enguita-Germán M, Cambra K,

Ibáñez-Beroiz B. Cardiovascular risk in patients with type 2 diabetes: A systematic review of prediction models. Diabetes Res Clin Pract 2021:109089.

[28] Dhiman P, Ma J, Navarro CA, Speich B, Bullock G, Damen JA, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. J Clin Epidemiol 2021;138:60–72.

[29] Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. BMC Med Res Methodol 2015;15(1):30.

[30] 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.

[31] Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: A systematic review [Internet]. 2021 Jul [cited 2021 Sep 9]. Available from: http://medrxiv.org/lookup/doi/10.1101/2021.06.28.21259089 Accessed 27 July 2021.

[32] Andaur Navarro CL, Damen JAAG, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised Machine Learning techniques: A systematic review and critical appraisal. BMJ Open. In press.

[33] Andaur Navarro CL, Damen JAAG, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Protocol for a systematic review on the methodological and reporting quality of prediction model studies using machine learning techniques. BMJ Open 2020;10(11):e038832.

[34] Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016;18(12):e323.

[35] Mackinnon A. The use and reporting of multiple imputation in medical research - a review: The use and reporting of multiple imputation in medical research. J Int Med 2010;268(6):586–93.

[36] Knol MJ, Janssen KJM, Donders ART, Egberts ACG, Heerdink ER, Grobbee DE, et al. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. J Clin Epidemiol 2010;63(7):728–36.

[37] Groenwold RHH. Informative missingness in electronic health record systems: the curse of knowing. Diagn Progn Res 2020;4(1):8.

[38] van Smeden M, Groenwold RHH, Moons KGM. A cautionary note on the use of the missing indicator method for handling missing data in prediction research. J Clin Epidemiol 2020;125:188–90.

[39] Cevallos Valdiviezo H, Van Aelst S. Tree-based prediction on incomplete data using imputation or surrogate decisions. Information Sci 2015;311:163–81.

[40] Kappen TH, Vergouwe Y. Adaptation of clinical prediction models for application in local settings.:10.

[41] Sperrin M, Martin GP. Multiple imputation with missing indicators as proxies for unmeasured variables: simulation study. BMC Med Res Methodol 2020;20(1):185.

[42] Sperrin M, Martin GP, Sisk R, Peek N. Missing data should be handled differently for prediction than for description or causal explanation. J Clin Epidemiol 2020;125:183–7.