PREDICTION MODELS, MACHINE LEARNING AND MISSING DATA

STEVEN NIJMAN

Prediction models, machine learning

and missing data

Steven Nijman

Provided by thesis specialist Ridderprint, ridderprint.nl
Printing: Ridderprint
Layout and Cover design: Jules Verkade, persoonlijkproefschrift.nl

ISBN/EAN: 978-90-393-7472-6

Financial support by the Dutch Heart Foundation for the publication of this thesis is gratefully acknowledged. The research described in this thesis was supported by a grant of the Dutch Heart Foundation (#2018B006). The publication of this thesis itself was made possible by the Heart Foundation.

Prediction models, machine learning and missing data

Predictie modellen, machine learning en missende waarden
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college van promoties
in het openbaar te verdedigen op

dinsdag 7 juni 2022 des middags te 12.15 uur

door

Steven Willem Joost Nijman

geboren op 22 mei 1992
te Utrecht

Promotoren:

Prof. dr. K.G.M. Moons
Prof. dr. F.W. Asselbergs

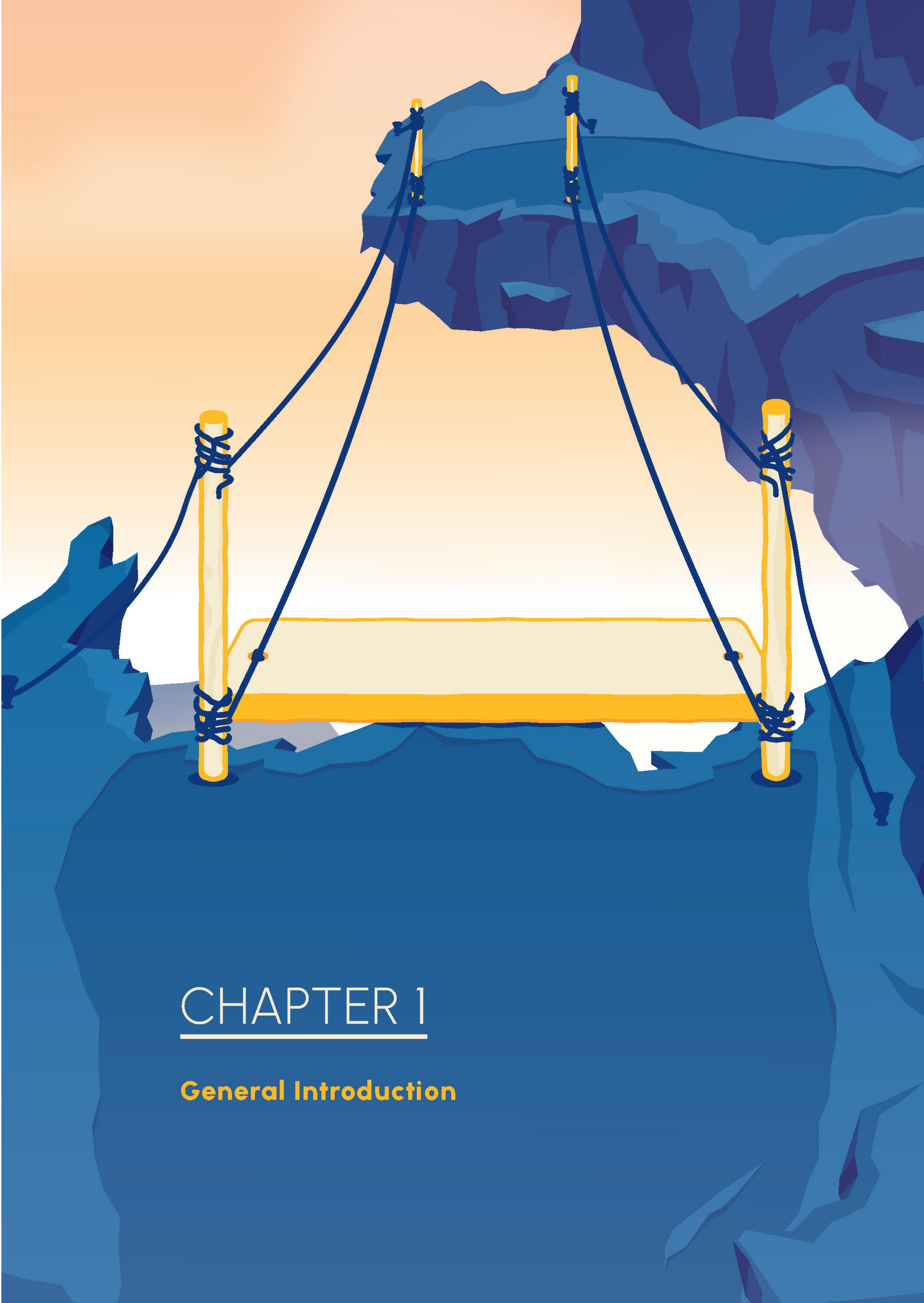
Copromotor:

Dr. T.P.A. Debray

Beoordelingscommissie:

Prof. dr. M.M. Rovers
Prof. dr. D.L. Oberski PhD
Prof. dr. F.E. Scheepers (voorzitter)
Prof. dr. F.L.J. Visseren
Prof. dr. M.E.E. Kretzschmar

Table of Content



CHAPTER 1

General Introduction

Missing data is one of the most universal dilemmas in research using clinical healthcare data [1–4]. A broken scale or an older patient's refusal to have a lab-value measured or answer any clinical questions (e.g., about income or weight) can already have tremendous impact on the ease and validity of any research making use of these healthcare data. As a result, Rubin, who initiated the inception of classical missing data theory, and many others studied and published thoroughly about the proper handling of missing data [5–8].

Historically, methods for handling missing data either (i) delete cases (e.g., patients) or columns (i.e., variables) with missing data, often referred to as complete-case analysis (CCA), (ii) include a missing indicator (i.e., a dummy variable denoting whether the variable is missing or not) in the prediction model or (3) estimate a plausible substitution value (i.e., imputation) for each missing predictor value [6–10]. Under most circumstances, It is recommended to generate multiple imputations for valid statistical inference [6,9,11,12].

A field which makes use of these clinical healthcare data and is well-known for its clashes with missing data is prediction research. Missing data frequently frustrates the development and validation of clinical – diagnostic and prognostic - prediction models [7,8,13–19]. Briefly, clinical prediction models combine patient and disease characteristics to estimate an individual's absolute risk of a pre-specified outcome (e.g., heart disease) with a fixed time window [13,20]. Examples of these predictions models can be found in the field of cardiovascular disease; models such as the Framingham heart score (FHS) and HEART-SCORE are widely applied in medical practice [21,22]. When developing or validating a risk prediction model, existing epidemiological reporting guidelines, congruent with the increasing amount of supportive literature, usually also recommend the use of multiple imputation [23–25].

When a prediction model is applied in daily practice to an individual, however, multiple imputation is mostly infeasible due to its computational time and required access to raw patient data. Hence, the actual, real-time, use of a risk prediction model, even when properly developed and validated, is limited in daily medical practice as risk prediction models usually have no direct, built-in problem-solving ability in case a predictor value of the individual is missing [13]. This is evident in the fields where prediction models are already being applied, such as in the cardiovascular domain, and studies have shown that the adoption of risk prediction models is severely hampered by missing predictor values [14,20,22,26,27]. This real-time aspect is unique to the application of risk prediction models in daily medical practice and seems to be underexposed in the literature, as evident by the variety of available solutions to deal with missing data when developing or validating a prediction model and the very limited guidance on dealing with missing data when applying them [6,9,12,28]. Clearly, simply ignoring the predictor from the prediction model which data is not observed is not a logical solution.

Lately, prediction research in medicine considers machine learning (ML) based risk prediction more often [29–35]. These ML-based methods allow for more flexibility, handle multi-dimensional complex data, and may circumvent the necessity for substituting missing values completely [30,35]. By changing their risk estimation to account for missing predictor values, these ML-based approaches do not require the previously explained imputation methods. Rather, they are risk prediction models capable of handling missing data as they occur in medical practice with built-in mechanisms that account for the missing predictors. An example are the so-called surrogate splits, which form an extension to the well-known decision trees [36–38]. Decision trees are one of the more common instances of ML based prediction that are used in clinical practice. As the name suggests, decision trees use a tree like structure to find the optimal cut-off point which partitions the data for optimal predictive performance. Based on the values of the pre-defined predictor variables, each branch in the tree represents a possible direction or decision [30,38]. Surrogate splits try and preserve these splits by learning from missing predictor values in the training data and adjusting the partitioning to resemble the original split in the tree as good as possible in the presence of missing predictor values [37–39].

In this thesis we evaluated traditional statistical and modern machine learning strategies for handling of missing predictor data when applying prediction models in real-time medical settings. We focus on strategies that do not require continuous access to raw patient data sets, that are computationally efficient and can, if desired, provide direct access to the imputed predictor value [40]. Since these imputation strategies allow existing prediction models to keep their current format and assigned weights, they provide an elegant and useful approach for enabling existing prediction models directly in medical practice.

Thesis outline

In **chapter 2** we provide a review evaluating the extent to which prediction model studies that use ML based techniques report on the presence and nature of missing data, which included the common methods used or handling missing data during model development, validation, and implementation.

Existing strategies to impute missing data are not applicable in implementation settings. In **Chapter 3** we expand on two well-known methods that make real-time imputation of missing predictor values possible and compare their imputation accuracy with mean imputation. In

Chapter 4 we further evaluate the impact of two of these imputation strategies on a prediction model's performance.

Intuitive alternatives exist to real-time imputation and are characterized by their ability to solve missingness inside the prediction model instead of in the data. In **Chapter 5** we compare various real-time missing data handling approaches other than imputation when implementing specific modeling techniques in clinical practice.

In addition to reporting on missing data, prediction model studies should report on many other aspects to ensure potential sources of bias have been handled appropriately. In **Chapter 6** we assess the methodological quality and risk of bias of supervised ML-based prediction model studies.

The use of imputation and risk prediction in clinical care will likely rely upon using large datasets from the electronic health record (EHR) or multi-centre studies, though complex strategies may be required to develop generalizable clinical prediction models. In **Chapter 7** we illustrate how advanced evidence synthesis methods can be used to evaluate this need in large population-level datasets.

The thesis ends with a general discussion.

Nijman SWJ[a], Leeuwenberg AM[a], Beekers I[b], Verkouter I[b], Jacobs JJL[b], Bots ML[a], Asselbergs FW[cde], Moons KGM[a], Debray TPA[ae]

[a] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands;

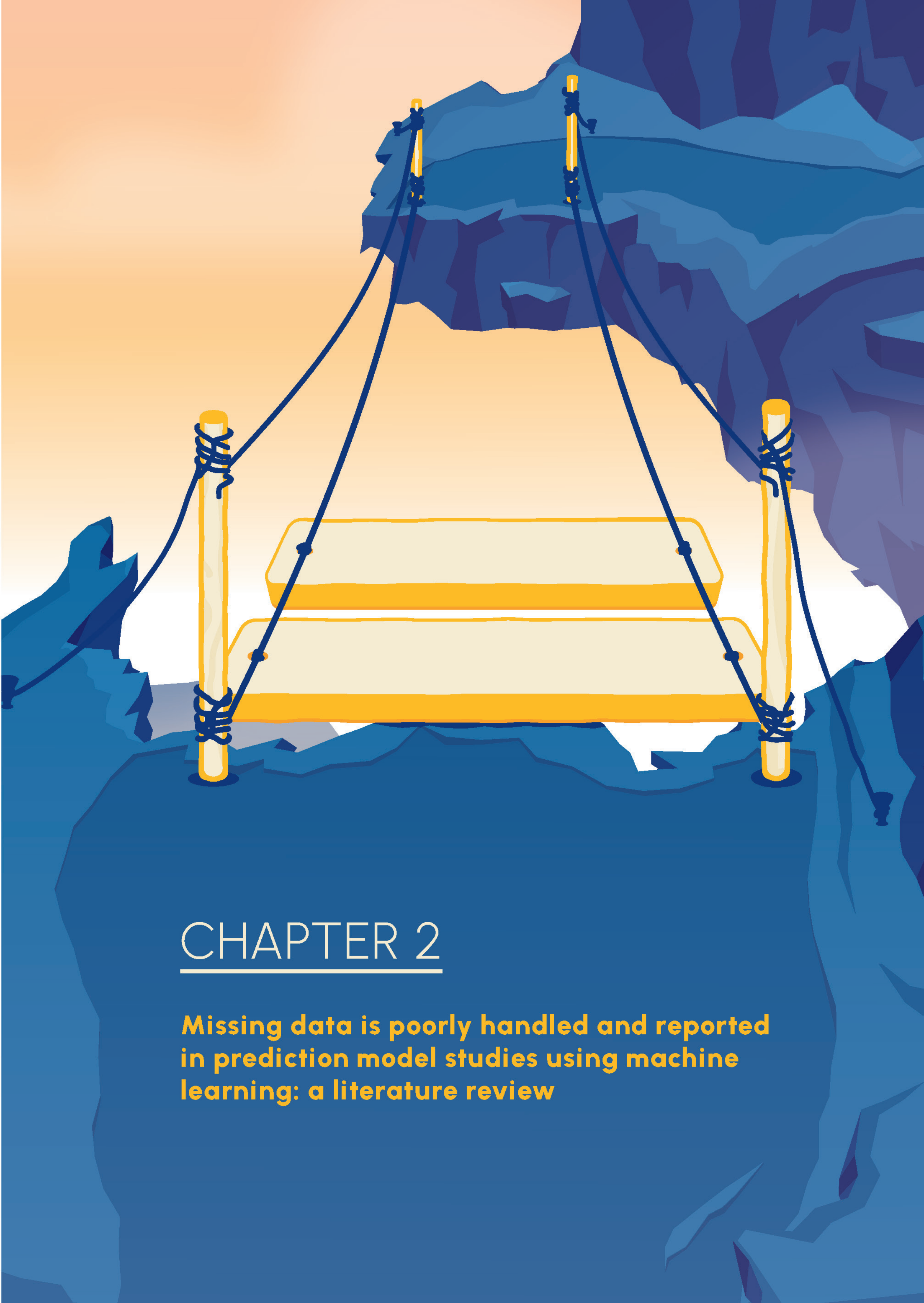
[b] Department of Health, Ortec B.V. Zoetermeer, The Netherlands;

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

[d] Institute of Cardiovascular Science, Faculty of Population Health Sciences, University College London, London, United Kingdom;

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

Journal of Clinical Epidemiology, 2022; 142 218–229. DOI: 10.1016/j.jclinepi.2021.11.023



CHAPTER 2

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

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.

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 was 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.

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 [1,23,24,41–43]. 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 [9,44]. 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 [6,12,41]. 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 [6,12,28,45,46]. 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 [6,45]. 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 [47,48]. This also makes it possible use multiple imputation techniques when actually implementing and applying prediction models in electronic healthcare software in daily clinical practice [28,40,47,48].

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) [4,37–39,49,50].

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 [4,9,61,65]. 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 [62,63,66]. Additionally, other issues may arise in the application of missing indicators as the manner of data collection between different practices is likely to vary [63].

Surrogate splits. Preserves the partitioning of each original split as good as possible in the presence of missing predictor values [37–39]. 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.

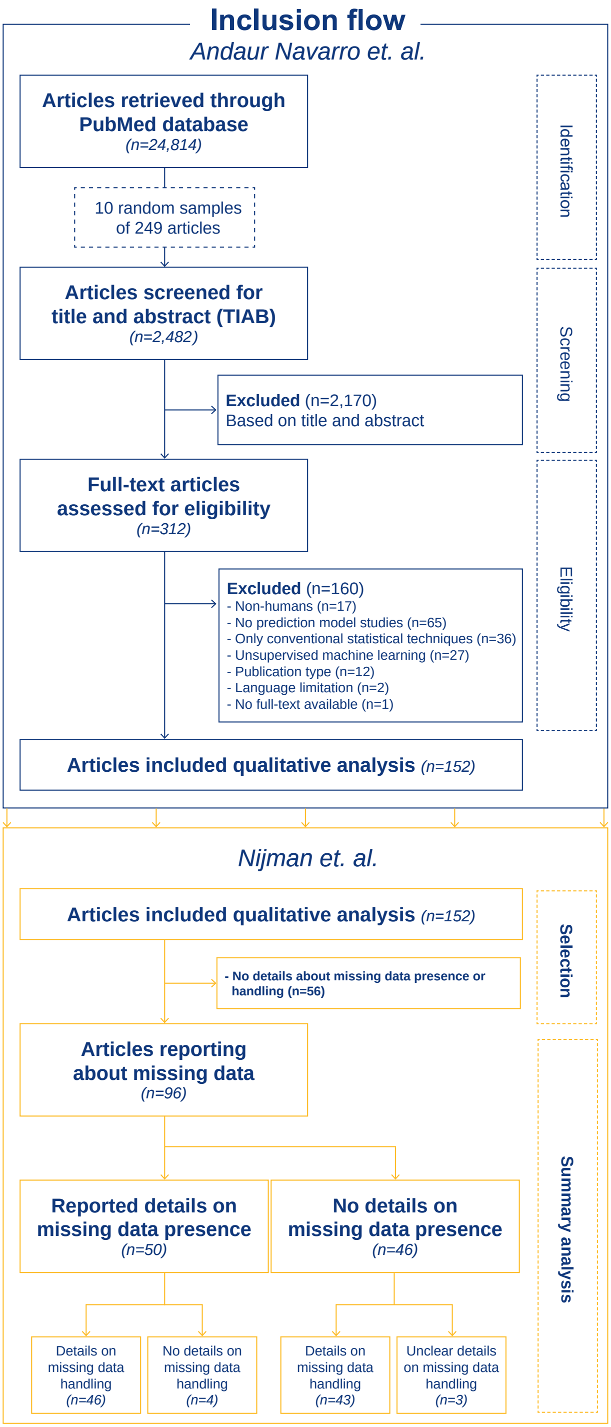
Sparsity aware splitting. A default direction is added for each tree node in a decision tree (e.g., XGBoost) [49]. Whenever a missing predictor value is encountered, the instance is classified into the pre-specified default direction. The optimal default direction, and thus best direction to handle missing 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 [50]. 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 use*d*.

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 [6,23–25,51]. 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 [3,52–56].

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.

Figure 1. Inclusion flow continuation after systematic review


Inclusion flow
Andaur Navarro et. al.
Articles retrieved through PubMed database (n=24,814)
10 random samples of 249 articles
Articles screened for title and abstract (TIAB) (n=2,482)
Excluded (n=2,170) Based on title and abstract
Full-text articles assessed for eligibility (n=312)
Excluded (n=160)
- Non-humans (n=17)
- No prediction model studies (n=65)
- Only conventional statistical techniques (n=36)
- Unsupervised machine learning (n=27)
- Publication type (n=12)
- Language limitation (n=2)
- No full-text available (n=1)
Articles included qualitative analysis (n=152)
Identification
Screening
Eligibility
Nijman et. al.
Articles included qualitative analysis (n=152)
- No details about missing data presence or handling (n=56)
Articles reporting about missing data (n=96)
Reported details on missing data presence (n=50)
No details on missing data presence (n=46)
Details on missing data handling (n=46)
No details on missing data handling (n=4)
Details on missing data handling (n=43)
Unclear details on missing data handling (n=3)
Selection
Summary analysis

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 [57,58]. The protocol of which was registered and published (PROSPERO, CRD42019161764) [59]. 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 (Figure 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 [6,23,24,41] 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 [35]. 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 [37–39,49]. 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 (SN, AL), 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 (TD). 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.

Results

After screening, 152 eligible articles were available for the present study (Figure 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.

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%]).

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 eight 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 (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.

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	[23]
	1.2 Potential reasons for the presence of missing data	[23,24,41]
	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	[24,41]
	› 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*	
	2.2 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)	[23,41]
	2.3 If complete case analysis was performed, a rationale for exclusion	[23,41]
	2.4 Comparison of overall patient characteristics of patients with and patients without missing values	[41]
	2.5 If possible, results of complete case analysis (to compare) and their interpretation	[41]
	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	[24,41]

Table 1. (continued)

	Details to be reported		Inspired by
3.0 Imputation-based approaches	3.1 Type of imputation › Single imputation (SI) (3.9) › Multiple imputation (MI) (3.10)		[24,41]
	3.2 Explicit mentioning of the assumptions that were made (e.g., MAR, MCAR or MNAR)		[23,41]
	3.3 Motivation for the assumptions made (3.3.1) or inclusion of sensitivity analyses for testing robustness (3.3.2)		[41]
	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		[24,41]
	3.5 Details on how continuous and non-continuous variables were imputed		[24,41]
	3.6 Details on what variables were included in the imputation procedure		[24,41]
	3.7 Inclusion of outcome as variable in the imputation procedure		[24]
	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)	[24]
		3.9.2 Details on if method takes into account noise or imputes a fixed value	[24]
	3.10 Multiple imputation details	3.10.1 Type of MI used (e.g., FCS or joint imputation)	[24,41]
		3.10.2 Number of imputed datasets	[24,41]
		3.10.3 Details on conditional models used (e.g., PMM, Random Forest, logistic regression, neural network, machine learning, etc.)	[24,41]
		3.10.4 Details on convergence of the imputation model	[24,41]

Table 1. (continued)

	Details to be reported		Inspired by
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
	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

Legend – 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

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.

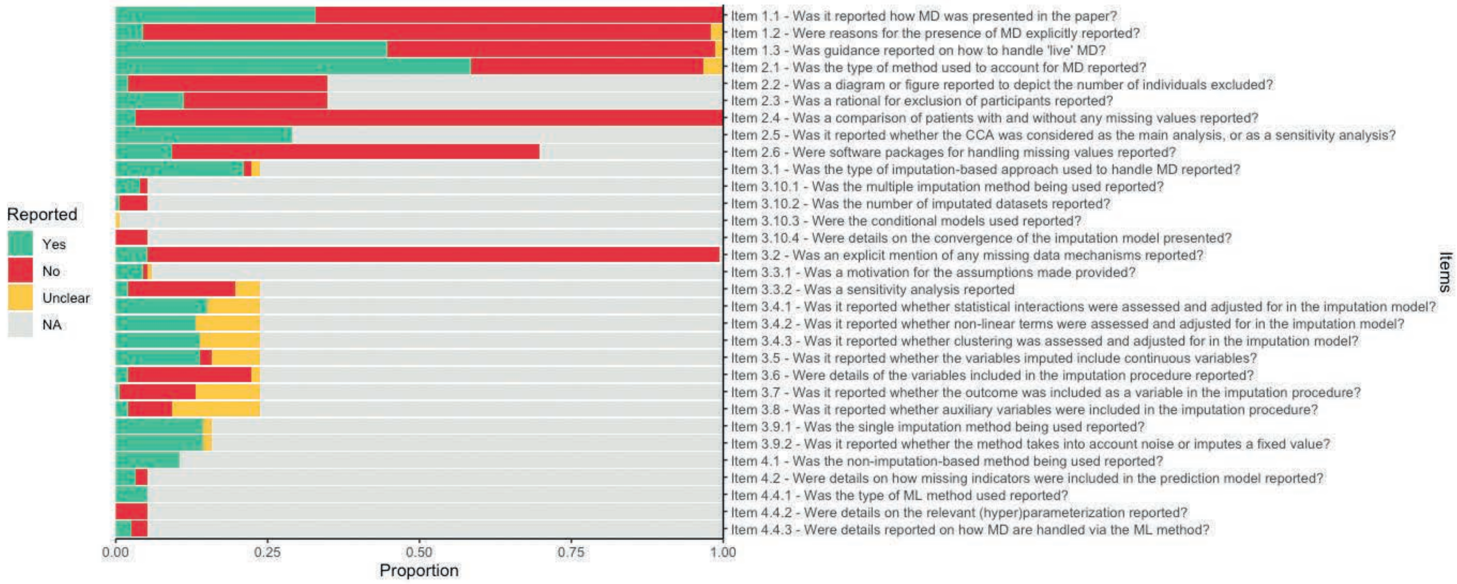
A complete overview of the extracted data can be found in the Appendix.

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 [3,60]. 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 (Figure 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 [9,28,45,61]. 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 [4,9,61,62]. 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 [63]. 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.

Figure 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

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 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 [36,37,39].

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 [48,64]. 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 [3,6,52]. 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 [37,49]. 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 [36,37,39].

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 [59]. 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 [23,42]. 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.

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).

All other 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.

Author contributions

Nijman SWJ: Conceptualization, Methodology, Investigation, Writing- Original draft preparation, Visualization. **Leeuwenberg AM:** Investigation, Writing- Reviewing and Editing. **Beekers I:** Writing- Reviewing and Editing, **Verkouter I:** Writing- Reviewing and Editing. **Jacobs JJL:** Writing-Reviewing and Editing. **Bots ML:** Conceptualization, Writing- Reviewing and Editing, Supervision, Funding acquisition. **Asselbergs FW:** Conceptualization, Writing- Reviewing and Editing. **Moons KGM:** Supervision, Conceptualization, Writing- Reviewing and Editing, Project administration. **Debray TPA:** Validation, Conceptualization, Methodology, Supervision, Writing- Reviewing and Editing.

Data availability statement

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

Acknowledgements

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.

Appendix

Appendix A. 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.

Appendix B. 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%)

Appendix B. (continued)

#	Item	Total (%)
	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%)

Appendix C. 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%)

Appendix D. Reported details on imputation (n=36)

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

Appendix D. (continued)

#	Item	Total (%)
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%)

Appendix E. 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%)

Appendix F. 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%)
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%)

Appendix G. 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%)

Andaur-Navarro CL[ab], Damen JAA[ab], Takada T[a], Nijman SWJ[a], Dhiman P[cd], Ma J[c], Collins GS[cd], Bajpai R[e], Riley RD[e], Moons KGM[ab], Hooft L[ab]

a Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands;

b Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands.

c Center for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom.

d NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom

e Centre for Prognosis Research, School of Medicine, Keele University, Keele, United Kingdom.

British Medical Journal, 2021; 375:n2281. DOI: 10.1136/bmj.n2281

CHAPTER 3

Risk of bias in studies on prediction models developed using supervised Machine Learning techniques: A systematic review

Abstract

Objective. To assess the methodological quality of machine learning (ML)-based prediction model studies across all medical fields.

Design. Systematic review.

Data sources. PubMed from 1 January 2018 to 31 December 2019.

Eligibility criteria. We included articles reporting on the development or development with external validation of a multivariable prediction model (either diagnostic or prognostic) developed using supervised ML for individualized predictions. No restrictions were made based on study design, data source, or predicted patient-related health outcomes.

Review methods. To determine the methodological quality of the ML-based prediction model studies, we evaluated the risk of bias (RoB) using the Prediction Risk Of Bias ASsessment Tool (PROBAST). We measured RoB per domain (participants, predictors, outcome, and analysis) and per study (overall).

Results. We included 152 studies, 58 (38.2%) diagnostic and 94 (61.8%) prognostic studies. We applied PROBAST to 152 developed models and 19 external validations. Out of these 171 analyses, 148 (86.5%, 95% confidence interval 80.6% to 90.9%) were rated at high RoB. The Analysis domain was the most frequently rated at high RoB. We observed 85/152 (55.9%, 48.0% to 63.6%) models developed with an inadequate number of events per candidate predictor, 62 with poor handling of missing data (40.8%, 33.3% to 48.7%) and 59 with unproper assessment of overfitting (38.8%, 31.4% to 46.7%). Most models used appropriate data sources to develop (73.0%, 65.5% to 79.4%) and externally validate their ML-based prediction models (73.7%, 51.2% to 88.2%). However, information about blinding of outcome and blinding of predictors was absent in 60/152 (39.5%, 32.1% to 47.4%) and 79/152 (52.0%, 44.1% to 59.8%) developed models, respectively.

Conclusion. Most ML-based prediction model studies show poor methodological quality and are at high risk of bias. Factors contributing to the risk of bias include small study size, poor handling of missing data, and failure to address overfitting. Efforts to improve the design, conduct, reporting, and validation of ML-based prediction model studies are necessary to boost its application in clinical practice.

Systematic review registration PROSPERO, CRD42019161764

What is already known on this topic?

- Several publications have highlighted the poor methodological quality of regression-based prediction models studies.
- The number of clinical prediction models developed using supervised machine learning is rapidly increasing, however, evidence about their methodological quality and risk of bias is scarce.

What this study adds?

- Prediction model studies developed using supervised machine learning have poor methodological quality. Limited sample size, poor handling of missing data, and inappropriate evaluation of overfitting contributed largely to the overall high risk of bias.
- Machine learning prediction models often claim superior accuracy compared to regression-based approaches. However, reported performance may be at high risk of bias based on the study design and modelling strategies used. Caution is needed when interpreting these findings.
- Future research should improve transparency when reporting and the study designs used to develop, validate, and compare prediction models to reduce methodological biases.

Introduction

A multivariable prediction model is defined as any combination of two or more predictors (i.e. variables, features) for estimating the probability or risk of an individual to have (diagnosis) or will develop (prognosis) a particular outcome.[67–70] Properly conducted and well reported prediction model studies are essential for a proper implementation in clinical practice. Even though prediction model studies are abundant in biomedical literature, a limited amount of them are used in clinical practice. As a result, many published studies contribute to research waste.[71] We anticipate that the rise of modern data-driven modelling techniques will boost the existing popularity of prediction model studies in the biomedical literature.[72,73]

Machine learning (ML), a subset of artificial intelligence (AI), has gained considerable popularity in recent years. Broadly, machine learning refers to computationally intensive methods that use data-driven approaches to develop models that require fewer modelling decisions by the modeler compared to traditional modelling techniques.[74–77] Within machine learning, there are two approaches: supervised and unsupervised learning. While supervised learning is defined as an algorithm that learn to predict using previously labelled outcomes, unsupervised learning learns to find unexpected patterns using unlabelled outcomes.[78] Traditional prediction models in healthcare usually resemble supervised learning: datasets used for development are labelled and the objective is to predict an outcome in new data. Supervised learning includes tree-based methods, such as random forests, naïve bayes, and gradient boosting machines, support vector machines, neural networks. Supervised ML-based prediction model studies have shown promising and even superior predictive performance compared to conventional statistical techniques, however, recent systematic reviews have shown otherwise. [79–82] Although several publications have raised concern about the methodological quality of prediction models developed with conventional statistical techniques[72,83,84], a formal methodological and risk of bias (RoB) assessment of supervised ML-based prediction model studies across all medical disciplines has not yet been carried out.

Shortcomings in study design, methods, conduct, and analysis may set the study at high RoB, which could lead to deviated estimates of models' predictive performance.[85,86] The Prediction model Risk Of Bias Assessment Tool (PROBAST) was developed to facilitate RoB assessment, and thus provides a methodological quality assessment of primary studies that report on development, validation, or update of prediction models, regardless of the clinical domain, predictors, outcomes, or modelling technique used. [85,86] Using a prediction model considered at high RoB, might lead to unnecessary or insufficient interventions, and thus affect patients' health and health systems. Rigorous RoB evaluation of prediction model studies is, therefore, essential to ensure reliability, fast, and valuable application of prediction models. Therefore, we conducted a systematic review to assess the methodological quality and RoB of supervised ML-based prediction model studies across all medical fields in a contemporary sample of recent literature.

Methods

Our systematic review was reported following the PRISMA statement.[87] The review protocol was registered (PROSPERO, CRD42019161764) and published.[88]

Identification of prediction model studies

We searched for eligible studies published in PubMed between 1 January 2018 and 31 December 2019. We restricted the search to obtain a contemporary sample of articles that would reflect the current practices in prediction modelling using machine learning to date. The search was performed on 19 December 2019 with a strategy that is provided in Supplemental File 1.

Eligible publications needed to describe the development or validation of at least one multivariable prediction model using any supervised ML technique aiming for individualized prediction of risk or patient-related health outcomes. Details about inclusion and exclusion criteria are stated in our protocol.[88] A publication was also eligible if it aimed to develop a prediction model based on model extension or incremental value of new predictors. No restrictions were made based on study design, data source, or types of patient-related health outcomes. We defined a publication to be an instance of ML when a non-regression statistical technique was used to develop or validate a prediction model. Hence, studies using only linear regression, logistic regression, lasso regression, ridge regression, or elastic net were excluded. Publications that report about the association of a single predictor, test, or biomarker, or its causality with an outcome were excluded. Publications that aimed to use ML to enhance the reading of images or signals or those where ML models only used genetic traits or molecular markers as predictors, were also excluded. We also excluded systematic reviews, methodological articles, conference abstracts, and publications for which full text was unavailable through our institution. The search was restricted to human subjects and English-written articles.

Screening process

Titles and abstracts were screened by two independent reviewers, from a group of seven (CLAN, TT, SWJN, PD, JM, RB, JAAD). A third reviewer was involved when required to resolve any disagreements (JAAD). After selection of potentially eligible studies, full-text articles were retrieved and two independent researchers reviewed them for eligibility; one researcher (CLAN) screened all articles and six researchers (TT, SWJN, PD, JM, RB, JAAD) collectively screened the same articles for agreement. In case of any disagreement, a third reviewer was asked to read the article in question and resolve (JAAD).

Data extraction

We developed a data extraction form based on the four-domain structure (participants, predictors, outcome, and analysis) and 20 signalling questions (SQ) as described in PROBAST.[85] [86] The Participants domain refers to the selection of the participants and data sources. The Predictors domain evaluates potential sources of bias by the definition and measurement of the candidate predictors. The Outcome domain assesses how and when the outcome was defined and determined. Finally, the Analysis domain examines the statistical methods that the authors have used to develop and validate the model, including study size, handling of continuous predictors and missing data, selection of predictors, and model performance measures.

Our extraction form contained 3 sections per domain: two to nine specific signalling questions, judgement of RoB, and rationale for the judgment. Signalling questions were formulated to be answered 'yes/probably yes', 'no/probably no', and 'no information'. All signalling questions were phrased so that 'yes/probably yes' indicated absence of bias. Likewise, judgement of RoB was defined as 'high RoB', 'low RoB', and 'unclear RoB'. Also, we requested reviewers to provide a rationale for judgment as free-text comments.

If a study included external validation, we applied the extraction form to both, the development and external validation of the model. Signalling question 4.5 –was selection of predictors based on univariable analysis avoided? –, 4.8 –Were model overfitting and optimism in model performance accounted for? –, and 4.9 –Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis? – did not apply to external validation. If a study reported more than one model, we applied PROBAST to the recommended model defined by the authors in the article. If the authors did not recommend a single model, the model with highest accuracy (in terms of discrimination) was selected as the recommended model. The PROBAST tool, its considerations, and related publications are available on the PROBAST website (www.probast.org). A summary table with the criteria to judge risk of bias is provided in Supplemental File 2.

Two reviewers independently extracted data from each article using the constructed form. To accomplish consistent data extraction, the form was piloted on five articles by all reviewers. During pilot, reviewers clarified differences in interpretation and standardise data extraction. After the pilot, articles used were randomly assigned and screened again in the main data extraction.

One researcher (CLAN) extracted data from all articles and six researchers (TT, SWJN, PD, JM, RB, JAAD) collectively extracted data from the same articles. Any disagreements in data extraction were settled by consensus among each pair of reviewers.

Data analysis

Prediction model studies were categorized as prognosis or diagnosis and into four types of prediction models studies: development (with internal validation), development with external validation (same model), development with external validation (different model), and external validation only. *Model development studies* aim to develop a prediction model to be used for individualized predictions where its predictive performance is directly evaluated using the same data, either by resampling participant data or random/non-random split sample (internal validation). *Model development studies with external validation (same model)* have the same aim as the previous type, but the development of the model is followed by quantifying the predictive performance of the model in a different dataset. *Model development studies with external validation (different model)* aim to update or adjust an existing model that performs poorly by recalibrating or extending the model. *External validation only* studies aim to assess only the predictive performance of existing prediction models using data external to the development sample. [86,89]

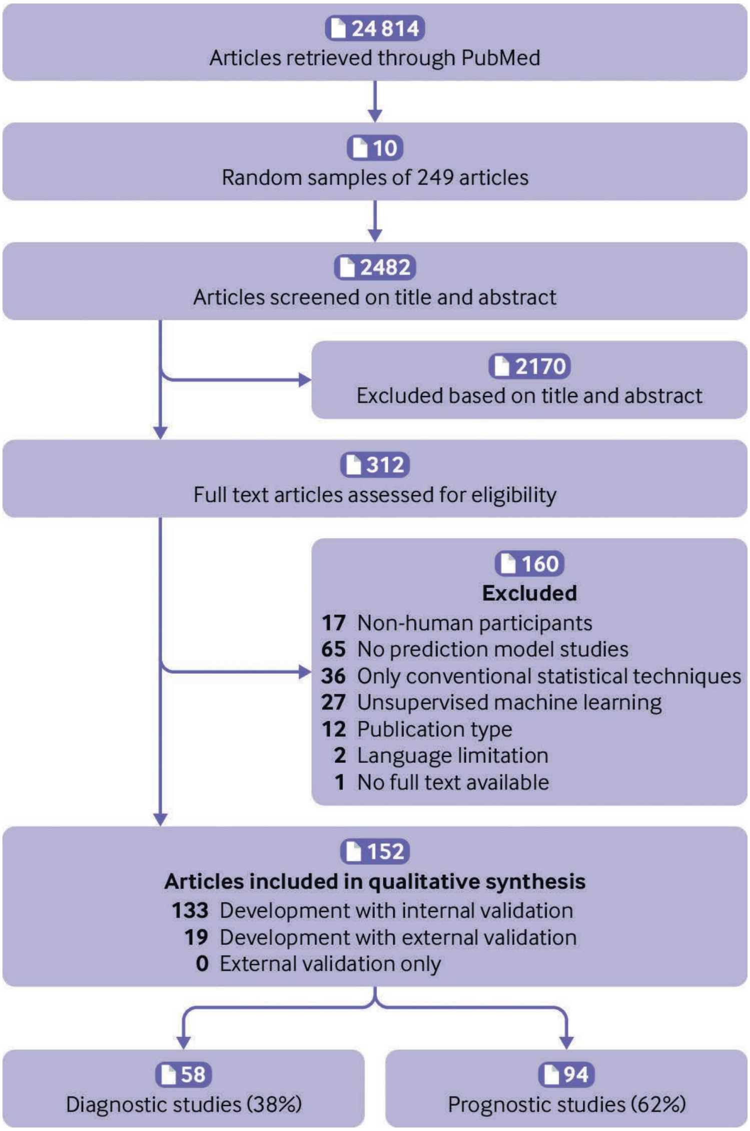
Two independent reviewers each assessed signalling question by the degree of compliance with the PROBAST recommendations. If there was any disagreement, it was discussed until consensus was reached. The RoB judgement per domain was based on the answers to the signalling questions. If the answer to all signalling questions was 'yes/probably yes', the RoB domain was judged as 'low RoB'. If reported information was insufficient to answer the signalling questions, these were judged as 'no information', and the RoB domain scored as 'unclear RoB'. If any signalling question was answered as 'no/probably no', reviewers applied their judgment to rate the domain as 'low RoB', 'high RoB', or 'unclear RoB'.

After judging all the domains, we performed an overall assessment per application of PROBAST. PROBAST recommends rating the study as 'low RoB' if all domains had 'low RoB'. If at least one domain had 'high RoB', overall judgment should be rated as 'high RoB'. 'Unclear RoB' was assigned if 'unclear RoB' was noted in at least one domain and all other domains had 'low RoB'. Judgement rationale was recorded to facilitate discussion among reviewers when solving discrepancies. We removed signalling question 4.9 –Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis? – because it is tailored for regression-based studies. Results were summarized as percentages with 95% confidence intervals and visual plots. Analyses were performed using R version 3.6.2 (R Core Team, 2020).

Patient and public involvement

We conducted a methodological appraisal; thus, no patients were involved in setting the research question, nor were they involved in the design or implementation of the study, or the interpretation or writing up of results.

Figure 1. Flowchart of included studies


24 814
Articles retrieved through PubMed
10
Random samples of 249 articles
2482
Articles screened on title and abstract
2170
Excluded based on title and abstract
312
Full text articles assessed for eligibility
160
Excluded
17 Non-human participants
65 No prediction model studies
36 Only conventional statistical techniques
27 Unsupervised machine learning
12 Publication type
2 Language limitation
1 No full text available
152
Articles included in qualitative synthesis
133 Development with internal validation
19 Development with external validation
0 External validation only
58
Diagnostic studies (38%)
94
Prognostic studies (62%)

Results

The search identified 24,814 publications, of which we sampled ten random sets of 249 publications each. Of the 2,482 screened publications, 152 were eligible: 94 (61.8%) prognostic and 58 (38.2%) diagnostic ML-based prediction model studies (Figure 1). Detailed description of the included studies is provided in Supplemental File 3. We classified publications according to their research aims: 132 (86.8%) articles were classified as development with internal validation, 19 (12.5%) as development with external validation of the same model, and 1 (0.6%) as development with external validation of another model (eventually included as development with internal validation). Across the 152 studies, a total of 1429 ML-based prediction models were developed and 219 validated. For our analyses, we selected only the recommended model by the authors for our RoB assessment. Hence, we applied PROBAST 171 times: in 152 developed models and 19 external validations. The most common ML techniques for the first model reported were Classification and Regression Tree (CART [10.1%]), Support Vector Machine (SVM [9.4%]), and Random Forest (RF [9.4%]). Detailed list of techniques assessed is provided in Supplemental File 3. The clinical fields with the most publications were oncology (21/152 [13.8%]), surgery (20/152 [13.5%]), and neurology (20/152 [13.5%]).

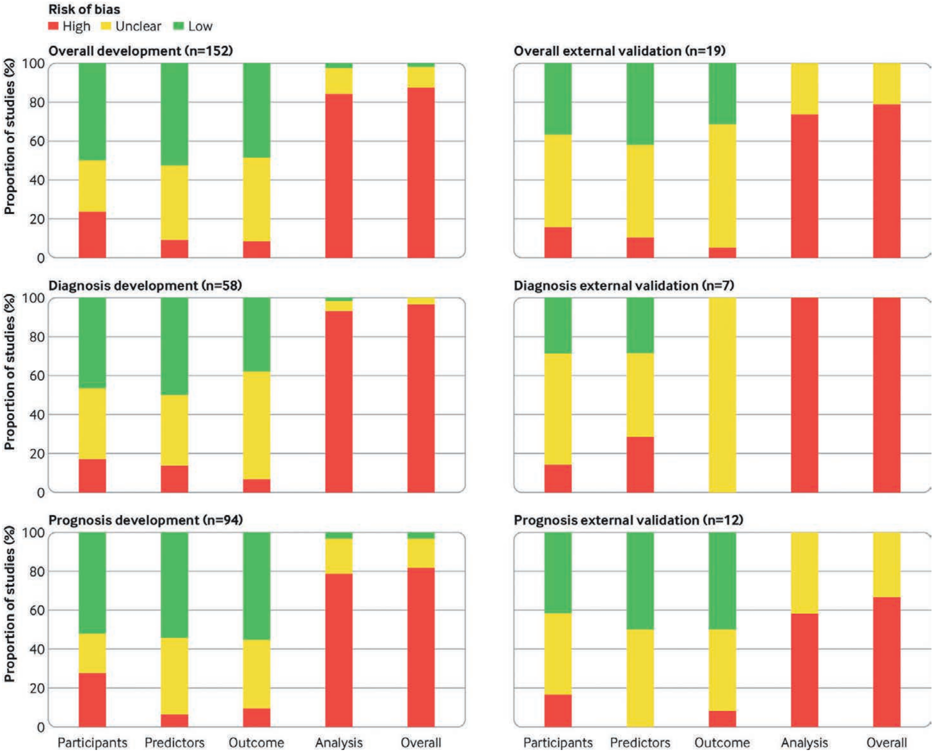
Domain 1: Participants

In total, 36/152 (23.7%) developed models and 3/19 (15.8%) external validations were scored as high RoB for the Participants domain (Figure 2). Prospective and longitudinal data sources (SQ1.1) were properly used for model development in 111/152 (73.0%) and to externally validate in 14/19 (73.7%). We were unable to evaluate whether the inclusion and exclusion of participants (SQ1.2) was representative of the target population in 47/152 (30.9%) developed models and in 12/19 (63.1%) external validations (Table 1).

Domain 2: Predictors

We rated 14/152 (9.2%) developed models and 2/19 (10.5%) external validations to be at high RoB for the Predictors domain (Figure 2). Candidate predictors were defined and assessed in a similar way for all included participants (SQ2.1) in 109/152 (71.7%) developed models and in 8/19 (42.1%) external validations. Information on blinding of predictor assessment to outcome data (SQ2.2) was missing in 60/152 (39.5%) developed models and in 7/19 (36.8%) external validations. All considered predictors should be available at the time the model is intended to be used (SQ2.3), which we found appropriate in 116/152 (76.9%) developed models and in 12/19 (63.1%) external validations (Table 1).

Figure 2. Risk of bias of included studies (n=152) and stratified by study type



Domain 3: Outcome

The domain Outcome was scored as unclear RoB in 65/152 (42.8%) and 12/19 (63.2%) of developed models and external validations, respectively (Figure 2). We missed information about the outcome being determined without knowledge of predictors' information (SQ3.5) in 79/152 (52.0%) developed models and in 14/19 (73.7%) external validations. Predictors were excluded from the outcome definition (SQ3.3) in 90/152 (59.2%) developed models and in 10/19 (52.6%) external validations. We considered the time interval between predictor measurement and outcome determination appropriate (SQ3.6) in 110/152 (72.4%) developed models and in 11/19 (57.9%) external validations. We observed in 114/152 (75%) developed models and in 12/19 (63.1%) external validations that the outcome was determined using appropriate methods, thus reducing risk of misclassification (SQ3.1). Similarly, 118/152 (77.6%) developed models and 13/19 (68.4%) external validations used prespecified, standard or consensus-based definitions to determine the outcome (SQ3.2). The outcome was defined and measured with the same categories or thresholds for all included participants (SQ3.4) in 118/152 (77.6%) developed models and 10/19 (52.6%) external validations (Table 1).

Domain 4: Analysis

We classified 128/152 (84.2%) developed models and 14/19 (73.7%) external validations as high RoB in the Analysis domain. We considered that the number of participants with the outcome (SQ4.1) was insufficient (i.e. event per predictor parameter <10) in 85/152 (55.9%) developed models and 8/19 (42.1%) external validations (i.e. number of events <100). Information about methods to handle continuous and categorical predictors (SQ4.2) was missed in 81/152 (53.3%) developed models and 18/19 (94.7%) external validations. We found that 84/152 (55.3%) developed models and 10/19 (52.6%) external validation included in their statistical analyses all enrolled participants (SQ4.3).

Handling of missing data (SQ4.4) was inappropriate (i.e. participants with missing data were omitted from the analysis or imputation method was flawed) in 62/152 (40.8%) developed models and in 7/19 (36.8) external validation. We observed that 28/152 (18.4%) developed models used univariable analyses to select predictors (SQ4.5). We were unable to assess if censoring, competing risks or sampling of control participants (SQ4.6) were considered in 54/152 (35.5%) developed models and in 7/19 (36.8%) external validations. Similarly, the reporting of relevant model performance measures (e.g., both discrimination and calibration) (SQ4.7) was missing in 91/152 (59.9%) developed models, while 13/19 (68.4%) external validations lacked this information too. 76/152 (50.0%) developed models accounted for model overfitting and optimism (SQ4.8).

Overall Risk of Bias

Finally, the overall RoB assessed using PROBAST let to 133/152 (87.5%) developed models, and 15/19 (78.9%) external validations being classified as high RoB (Figure 2). Further information about each signalling question answered as ‘Yes/probably yes’, ‘No/probably no’, and ‘No information’ is provided in Table 1.

Diagnostic versus prognostic models

Regarding diagnostic versus prognostic prediction models, the Analysis domain is the major contributor to an overall high RoB in both. We evaluated 56/58 (96.6%) developed models and 7/7 (100%) external validation as high RoB in diagnostic studies, and 77/94 (81.9%) developed models and 8/13 (66.7%) external validation in prognostic studies (Figure 2). External validations of both diagnostic and prognostic models suffer from unclear information to judge RoB. While in diagnostic models, signalling questions in domain Outcome were frequently answered with ‘no information’ (Table S2), in prognostic models this was the case for both Outcome and Analysis domains (Table S3). Further information about each signalling question is provided in Supplemental file 3.

Table 1. PROBAST Signaling questions for model development and validation analyses in 152 included studies

		Developed models (n=152)			External validations (n=19)		
		Yes, probably yes	No, probably no	No information	Yes, probably yes	No, probably no	No information
		n (%; 95% CI)	*n* (%; 95% CI)	*n* (%; 95% CI)	*n* (%; 95% CI)	*n* (%; 95% CI)	*n* (%; 95% CI)
Participants							
1.1	Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?	111 (73.0%, 65.5 to 79.4)	32 (21%, 15 to 28)	9 (6%, 3 to 11)	14 (74%, 51 to 88)	5 (26%, 12 to 49)	0
1.2	Were all inclusions and exclusions of participants appropriate?	89 (59%, 51 to 66)	16 (11%, 7 to 16)	47 (31%, 24 to 39)	7 (37%, 19 to 59)	0	12 (63%, 41 to 81)
Predictors							
2.1	Were predictors defined and assessed in a similar way for all participants?	109 (71.7, 64.1 to 78.3)	19 (13, 8 to 19)	24 (16, 11 to 22)	8 (42, 23 to 64)	1 (5, 0 to 25)	10 (53, 32 to 73)
2.2	Were predictor assessments made without knowledge of outcome data?	88 (58, 50 to 66)	4 (3, 1 to 7)	60 (40, 32 to 47)	10 (53, 32 to 73)	2 (11, 3 to 31)	7 (37, 19 to 59)
2.3	Are all predictors available at the time the model is intended to be used?	117 (77.0, 69.7 to 83.0)	4 (3, 1 to 7)	31 (20, 15 to 28)	12 (63, 41 to 81)	1 (5, 0 to 25)	6 (32, 15 to 54)
Outcome							
3.1	Was the outcome determined appropriately?	114 (75.0, 67.6 to 81.2)	6 (4, 2 to 8)	32 (21, 15 to 28)	12 (63, 41 to 81)	0	7 (37, 19 to 6)
3.2	Was a prespecified or standard outcome definition used?	118 (77.6, 70.4 to 83.5)	6 (4, 2 to 8)	28 (18, 13 to 25)	13 (68, 46 to 85)	0	6 (32, 15 to 54)
3.3	Were predictors excluded from the outcome definition?	90 (59, 51 to 67)	8 (5, 3 to 1)	54 (36, 28 to 43)	10 (53, 32 to 73)	0	9 (47, 27 to 69)
3.4	Was the outcome defined and determined in a similar way for all participants?	118 (77.6, 70.4 to 83.5)	11 (7, 4 to 13)	23 (15, 10 to 22)	10 (53, 32 to 73)	1 (5, 0 to 25)	8 (42, 23 to 64)
3.5	Was the outcome determined without knowledge of predictor information?	63 (41, 34 to 49)	10 (7, 4 to 12)	79 (52, 44 to 60)	4 (21, 9 to 43)	1 (5, 0 to 25)	14 (74, 51 to 88)
3.6	Was the time interval between predictor assessment and outcome determination?	110 (72.4, 64.8 to 78.9)	2 (1, 0 to 5)	40 (26, 20 to 34)	11 (60, 36 to 77)	1 (5, 0 to 25)	7 (37, 19 to 59)

		Developed models (n=152)			External validations (n=19)		
		Yes, probably yes	No, probably no	No information	Yes, probably yes	No, probably no	No information
		n (%; 95% CI)	*n* (%; 95% CI)	*n* (%; 95% CI)	*n* (%; 95% CI)	*n* (%; 95% CI)	*n* (%; 95% CI)
Analysis							
4.1	Were there a reasonable number of participants with the outcome?	52 (34, 27 to 42)	85 (56, 48 to 64)	15 (10, 6 to 16)	8 (42, 23 to 64)	8 (42, 23 to 64)	3 (16, 6 to 38)
4.2	Were continuous and categorical predictors handled appropriately?	37 (24, 18 to 32)	34 (22, 17 to 30)	81 (53, 45 to 61)	0	1 (5, 0 to 25)	18 (95, 75 to 100)
4.3	Were all enrolled participants included in the analysis?	84 (55, 47 to 63)	29 (19., 14 to 26)	39 (26, 19 to 33)	10 (53, 32 to 73)	3 (16, 6 to 38)	6 (32, 15 to 54)
4.4	Were participants with missing data handled appropriately?	20 (13, 9 to 20)	62 (41, 33 to 49)	70 (46, 38 to 54)	3 (16, 6 to 38)	7 (37, 19 to 59)	9 (47, 27 to 68)
4.5	Was selection of predictors based on univariable analysis avoided?	101 (66.4, 58.6 to 73.5)	28 (18, 13 to 25)	23 (15., 10 to 22)		NA	
4.6	Were complexities in the data (e.g., censoring, competing risks, sampling of control participants) accounted for appropriately?	63 (41, 34 to 49)	35 (23, 17 to 30)	54 (36, 28 to 43)	8 (42, 23 to 64)	4 (21, 9 to 43)	7 (37, 19 to 59)
4.7	Were relevant model performance measures evaluated appropriately?	15 (10, 6 to 16)	46 (30, 24 to 38)	91 (60, 52 to 67)	3 (16, 6 to 38)	3 (16, 6 to 38)	13 (68, 46 to 85)
4.8	Were model overfitting and optimism in model performance accounted for?	76 (50, 42 to 58)	59 (39, 31 to 47)	17 (11, 7 to 17)		NA	

(NA) Signaling question 4.5 and 4.8 are only applicable for model development

Discussion

Principal findings

We have conducted a detailed assessment of the methodological quality of supervised ML-based prediction model studies across all clinical fields. Overall, 133/152 (87.5%) developed models and 15/19 (78.9%) external validations showed high RoB. The Analysis domain was most commonly rated as high RoB in developed models and external validations, mainly due to a low number of participants with the outcome (relative to the number of candidate predictors), risk of overfitting, and inappropriate handling of participants with missing data. Although there are still no conclusive studies about sample size calculations for developing prediction models using ML techniques, these usually require (many) more participants and events than conventional statistical approaches.[90,91] One hundred studies failed to either provide the number of events or reported an event per candidate predictor (EPV) lower than 10, which historically is a marker of potentially low sample size. Furthermore, ML studies with a low number of participants with the outcome are likely to suffer from overfitting, that is the model is too much tailored to the development dataset. [90–93] Only half of the included studies examined potential overfitting of models either by using split data, bootstrapping or cross-validation. Random-split was often relied on to internally validate models (i.e. validation based on the same participants' data), whereas bootstrapping and cross-validation are generally considered more appropriate.[94]

Most studies carried out complete-case analyses or mean/median imputation. Multiple imputation is generally preferred as it prevents biased model performance due to deletion or single imputation of participants' missing data. Unfortunately, multiple imputation is still unpopular within models developed with ML techniques.[95,96] Some ML techniques have the power to incorporate this missingness by including a separate category of a predictor variable that has missing values.[97] Therefore, we urge algorithm developers to improve imputation methods and incorporate informative missingness in their models when possible.

Several signalling questions were scored as 'No information' making it impossible for us to judge potential biases. It was often unclear whether all enrolled participants were included in the analyses, how many participants had missing values, and how missing data were handled. ML are powerful and automated techniques that will learn from data, however, if there was selection bias in the dataset, predictions made using the trained ML algorithm will also be biased. Similarly, several signaling questions in PROBAST are tailored to identify lack of blinding (SQ 2.2, SQ 3.3, SQ 3.5); however, almost half of included articles failed to report any information for us to assess blinding. Furthermore, model calibration tables or plots were often not presented, whereas classification measures (i.e. confusion matrix) were commonly reported with an overreliance on accuracy.[98] Reporting and assessment of discrimination (i.e., ability to discriminate between cases and non-cases) and calibration (i.e., agreement between predictions and observed outcomes) is essential to assess a models' predictive accuracy.[98]

Comparison with other studies

A systematic review of 23 studies about ML for diagnostic and prognostic predictions in emergency departments shows that analysis was the most poorly rated domain with 20 studies at high RoB.[99] This study found deficiencies in how continuous variables and missing data were handled, and found that model calibration was rarely reported. Another publication about ML risk prediction models for triage of patients entering the emergency room also considered 22/25 studies considered at high RoB.[100] A study assessing the performance of diagnostic deep learning algorithms for medical imaging reported 58 of 81 studies being classified as overall high RoB.[73] Similar to our results, major deficiencies were found in the analysis domain including the number of events per variable, inclusion of enrolled participants in the analysis, reporting of relevant model performance measures, and overfitting. Recently, a living systematic review about COVID-19 prediction models indicated that all 57 studies that used ML were at high RoB due to insufficient sample size, unreported calibration, and internal validation based on training-test split.[101]

Strength and limitations of the study

We evaluated the risk of bias of supervised ML-based prediction model studies in a broad sample of articles which included prognostic and diagnostic development only and development with external validation studies. After using a validated search strategy, we retrieved nearly 25,000 publications which is similar to a previous study. We finally screened the tenth part of the whole sample; therefore, our results are presented using confidence intervals to extrapolate them to the whole sample. The present analyses considered results from studies that were published over one year ago; nevertheless, we expect these findings to be still applicable and relevant for the clinical prediction field. We adopted PROBAST as the benchmark to evaluate RoB enhancing the objectivity and consistency, however, this is not without certain limitations. While two signalling question in PROBAST might become less relevant within the ML context (i.e. selection of predictors based on univariable analysis and reporting of weighted estimates in the final model correspond to the results from the reported multivariable analysis), further signalling questions related to data generation, feature selection, and overfitting might be necessary.

Implication for researchers, editorial offices, and future research

The number of ML-based studies is increasing every year; thus, their identification, reporting and assessment become even more relevant. It will remain a challenge to determine the risk of bias if detailed information about data and modelling approach (including justifications to any decision made that may biases estimates) is not clearly reported in articles. To better judge studies, we recommend researchers to adhere to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement (35,36). Though TRIPOD was not exilicilty developed for machine learning prediction models, all items are applicable. Similarly, while there is yet no RoB assessment tool available specifically for supervised ML models, we suggest researchers to follow PROBAST recommendations to reduce potential biases when planning and modelling primary prediction model studies using either regression or non-regression models. For example, the adoption of multiple imputation to handle missing value and cross-validation or bootstrapping to internally validate the developed models.

Currently, extensions of TRIPOD and PROBAST for prediction models developed using machine learning are under development (TRIPOD-AI, PROBAST-AI).[102,103] As sample size contributed largely to the overall high RoB, future methodological research could focus on determine appropriate sample sizes for each supervised learning technique. Giving the rapid and constant evolution of machine learning, periodic systematic reviews of prediction model studies need to be conducted. Although high quality ML-based prediction model studies are scarce, those who stand out need to be validated, re-calibrated, and promptly implemented in clinical practice.[101] To avoid research waste, we suggest peer-reviewers and journal's editors to promote the adherence to reporting guidelines.[71,104,105] Facilitating the documentation of studies (i.e. supplemental material, data, and code) and setting unlimited word count may improve methodological quality assessment, as well as independent validation (i.e. replication). Likewise, requesting external validation of prediction models upon submission might help setting minimum standards to ensure generalizability of supervised ML-based prediction models studies.

Conclusion

Most supervised ML-based prediction model studies show poor methodological quality and are at high risk of bias. Factors contributing to the risk of bias include the exclusion of participants, small sample size, poor handling of missing data, and failure to address overfitting. Efforts to improve the design, conduct, reporting, and validation of supervised ML-based prediction model studies are necessary to boost its application in clinical practice and avoid research waste.

Disclosures

Authors' contributions

The study concept and design were conceived by CLAN, JAAD, PD, LH, RDR, GSC, and KGMM. CLAN, JAAD, TT, SN, PD, JM, and RB conducted article screening and data extraction. CLAN performed data analysis and wrote the first draft of this manuscript, which was revised by all authors who have provided the final approval of this version. CLAN, the corresponding author, is the guarantor of the review. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Acknowledgements

The authors would like to thank and acknowledge the support of René Spijker, information specialist. In addition, we thank the BMJ editors and peer reviewers for their comments and suggestions which have been incorporated into the manuscript.

Funding

There is no specific funding for this study. GSC is supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) and by Cancer Research UK program grant (C49297/A27294). PD is supported by the NIHR Oxford BRC. The views expressed are those of the authors and not necessarily those of the NHS, NIHR, or Department of Health. None of the funding sources had a role in the design, conduct, analyses, or reporting of the study or in the decision to submit the manuscript for publication.

Competing interests

GSC, RDR and KGMM are members of the PROBAST Steering Group. All authors have completed the Unified Competing Interest form and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years, no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval

We analysed only published data; therefore, ethics approval was not required.

Data sharing

The study protocol is available at *doi*: 10.1136/bmjopen-2020-038832 . Detailed extracted data on all included studies are available upon reasonable request.

Transparancy

The guarantor of this review affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

Dissemination plans

We plan to disseminate the findings and conclusions from this study through social media (such as Twitter), a plain-language summary on www.probast.org, and scientific conferences. In addition, the findings will provide insights to the development of PROBAST-AI.

Twitter: @GSCollins, @SWJNijman, @pauladhiman, @RamBajpai, @Richard_D_Riley, @CarlMoons

Additional files

- Supplemental file 1. Search Strategy
- Supplemental file 2. Summary table with criteria to judge risk of bias.
- Supplemental file 3. Table S1. Characteristics of included studies (n=152)
- Supplemental file 4. Table S2-S3. Signalling questions for diagnosis and prognosis model studies.

Nijman SWJ[a]*, Groenhof TKJ[a]*, Hoogland J[a], Bots ML[a], Brandjes M[b], Jacobs JJL[b], Asselbergs FW[cde], Moons KGM[a], Debray TPA[ad]

* contributed equally

a Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands;

b LogiqCare, Ortec B.V. Zoetermeer, The Netherlands;

c Department of Cardiology, Division Heart & Lungs, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands;

d Institute of Cardiovascular Science, Faculty of Population Health Sciences, University College London, London, United Kingdom;

e Health Data Research UK, Institute of Health Informatics, University College London, London, United Kingdom

Journal of Clinical Epidemiology, 2021; 134 22-34. DOI: 10.1016/j.jclinepi.2021.01.003

CHAPTER 4

Real-time imputation of missing predictor values improved the application of prediction models in daily practice

Abstract

Objectives – In clinical practice, many prediction models cannot be used when predictor values are missing. We therefore propose and evaluate methods for real-time imputation.

Study design and Setting – We describe (i) mean imputation (where missing values are replaced by the sample mean), (ii) joint modeling imputation (JMI, where we use a multivariate normal approximation to generate patient-specific imputations) and (iii) conditional modeling imputation (CMI, where a multivariable imputation model is derived for each predictor from a population). We compared these methods in a case study evaluating the root mean squared error (RMSE) and coverage of the 95% confidence intervals (i.e. the proportion of confidence intervals that contain the true predictor value) of imputed predictor values.

Results –RMSE was lowest when adopting JMI or CMI, although imputation of individual predictors did not always lead to substantial improvements as compared to mean imputation. JMI and CMI appeared particularly useful when the values of multiple predictors of the model were missing. Coverage reached the nominal level (i.e. 95%) for both CMI and JMI.

Conclusion – Multiple imputation using, either CMI or JMI, is recommended when dealing with missing predictor values in real time settings.

Keywords: missing data; multiple imputation; real-time imputation; prediction; computerized decision support system; electronic health records

Highlights

- Cardiovascular risk management guidelines advocate use of prediction models in routine clinical practice.
- The implementation of a prediction model in routine care typically requires complete information on all predictor values. If one or more predictor values are unknown, the model cannot provide a prediction.
- The implementation of a prediction model (e.g. in a decision support system) should always include a strategy for dealing with missing predictor values.
- Traditional (multiple) imputation methods require information from other patients and therefore cannot be used when patients present individually, as is the case in clinical practice.
- It is possible to adapt existing imputation strategies for real-time use. This requires to estimate the conditional distribution for each predictor variable in a training sample, and to make this summary information available to the implementation of a prediction model.
- In general, two approaches are possible to model the conditional distribution of the predictor variables in a training sample. One approach is to estimate each distribution separately using a flexible (e.g. regression) modeling strategy. Alternatively, it is possible to directly estimate the joint distribution of all predictor variables. When this joint distribution is normal, then the conditional distributions can directly be derived from the mean and covariance of the training sample.
- Simulations indicate that joint modelling imputation and conditional modelling imputation results in fewer inappropriate treatment decisions and has minimal impact on predicted risk, especially for high-risk patients.

Introduction

In present-day medical practice, characterized by an aging population, multimorbidity and high complexity of diseases, attention has grown towards personalized medicine aiming to administer the most applicable treatment to the individual patient given their risk profile [106–108]. In cardiovascular disease management, guidelines advocate the use of prediction models to assess the patients' risk of developing a certain cardiovascular disease to guide treatment decision making [106]. To integrate risk-guided care in daily practice, technological solutions such as computerized decision support systems (CDSS) are increasingly developed [109,110]. Using predictor values directly extracted from the electronic health record (EHR), CDSS can provide an immediate risk assessment of each encountered patient at a glance [17,19,111].

The use of prediction models in daily practice in individual patient requires real-time availability of the patient's values of the predictors in the model. Most prediction models cannot provide a risk estimate in the presence of missing predictor values, which hampers implementation and may ultimately limit guideline adherence [112]. Therefore, predictor values should be measured and registered (e.g. in the Electronic Health Record; EHR) in such a way that they are available in real-time. Yet, routine clinical care data is often incomplete because certain measurements are deemed unnecessary, time-consuming, or expensive, or because they cannot directly be extracted from the EHR (e.g., registered as free text) [113].

Missing data is a well-known challenge in (medical) research, for which several scalable solutions exist [114]. Multiple imputation by chained equations has often been recommended to handle missing data in a research setting where data from multiple patients are available for study analysis purposes [115,116]. This approach, however, is not directly applicable when applying a prediction model real-time to a single patient in the consulting room. In particular, the models used for imputation cannot be generated "live" in clinical practice, and therefore need to be derived elsewhere and beforehand [117].

One option is to replace missing predictor values by their respective mean/median, which in turn is estimated from another data set or training sample [118,119]. Whilst straightforward to implement, mean imputation may be insufficient when the predictor with missing values is a strong predictor or exhibits large variability such that assigning an overall mean may lead to less predictive accuracy of the prediction model and to misinformed treatment decisions. Mean imputation does not distinguish between patients and may therefore likely impute values that are unrealistic given the patient's observed predictor values. Also, mean imputation obfuscates any uncertainty about the imputed values.

To address these issues, we expand on two well-known methods that may also be used in real time imputation of missing predictor values [117]: joint modeling imputation (JMI) [120] and conditional modeling imputation (CMI, also known as multiple imputation by chained equations) [116]. As opposed to mean imputation, these methods are able to incorporate the relation between multiple patient characteristics, and therefore allow imputations to be adjusted for observed patient specific characteristics. Similar to mean imputation, these relations can be learned from training data and, in real time, applied on new patients that are not part of the training sample. Additionally, both methods allow for multiple imputations to be estimated, reflecting the uncertainty with respect to the imputed value.

Using a real-world example and empirical data set on cardiovascular risk prediction, we compared the accuracy and usability of three imputation methods (mean imputation, JMI, and CMI) to deal with missing values of predictors in the prediction model in real time. Though it is well known that mean imputation is problematic, it was chosen as a comparison due to its straightforward implementation when implementing a prediction model in routine clinical practice or in a decision support [121–124].

What is new?

Key findings

- Multiple imputation approaches can be adapted without much difficulty to allow for real-time imputation of missing predictor variables.
- Both conditional modelling imputation (CMI) and joint modelling imputation (JMI) give more accurate estimates of missing predictor values when compared to mean imputation.

What this adds to what was known?

- Imputation of missing predictor values does not require 'live' access to a source dataset. Simple population characteristics (such as the mean and covariance) can be used to generate imputations that are tailored to a specific individual.

What is the implication and what should change now?

- Real-time multiple imputation, using either CMI or JMI, should be made available in clinical practice (e.g. via a computerized decision support system) to support guideline recommended use of prediction models and to be more transparent about uncertainty
- When developing or validating a prediction model, researchers should report the mean and covariance of the study population, as this information can directly be used to impute missing values in routine care.

Methods

Imputation methods

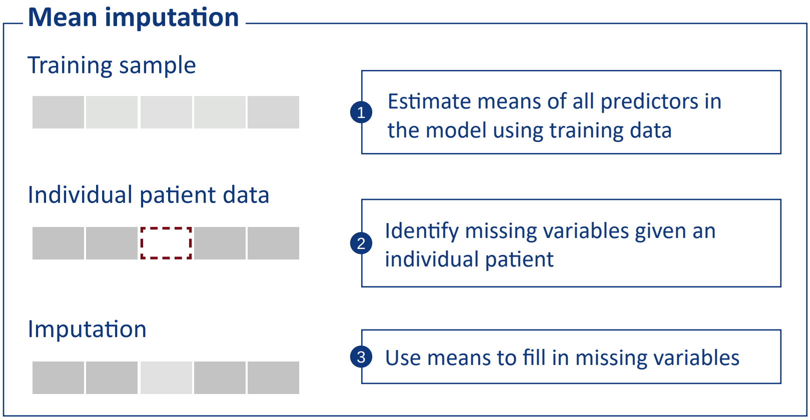
To facilitate live imputation of missing values in routine care, it is essential to obtain information on the distribution of the target population. This summary information can, for instance, be derived in an epidemiologic (e.g. cohort) study and then be utilized to train live imputation models. A key constraint given is that all methods, after being trained, are independent and stand-alone, which means that they can directly be used for live imputation in a new, single, patient without requiring the need for any additional procedures.

The three methods under evaluation are mean imputation, joint modeling imputation (JMI), and conditional modeling imputation (CMI) [116,117,120]. All methods were implemented in R and facilitate live imputation of missing values in individual patients. Source code is available from the supplementary information (Appendix D).

Mean imputation

The training sample is used to derive the means of all predictors in the model (Figure 1). Missing predictor values are then imputed by their respective mean (or proportion in case of binary variables). This method is relatively straightforward to implement, and can be extended to subgroup-specific means (i.e. creating subdivisions based on certain parameters of a population of which multiple means are respectively calculated).

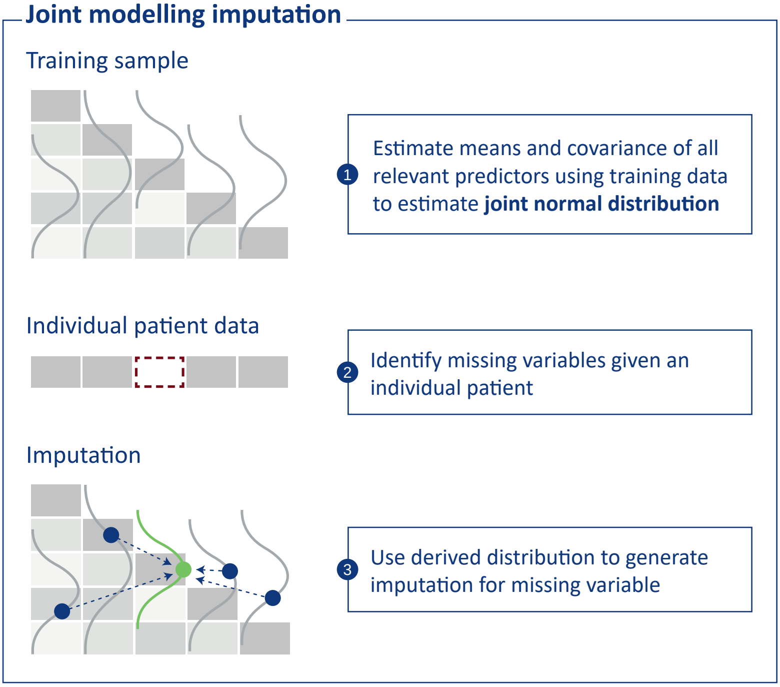
Figure 1. Mean imputation


Mean imputation
Training sample
Individual patient data
Imputation
1
Estimate means of all predictors in the model using training data
2
Identify missing variables given an individual patient
3
Use means to fill in missing variables

Joint modeling imputation

The training sample is used to derive the means and covariance of all predictor variables (Figure 2). It is assumed that all predictor variables of the training sample are normally distributed, such that imputations for an individual patient can directly be generated from the mean and covariance of the training sample and the observed predictor values [117,120]. In contrast to overall mean imputation, use of covariances between all predictors incorporates the relation between the predictors, and therefore allows imputations to be tailored to an individual's patient own characteristics. A more detailed description is provided in Appendix A [117].

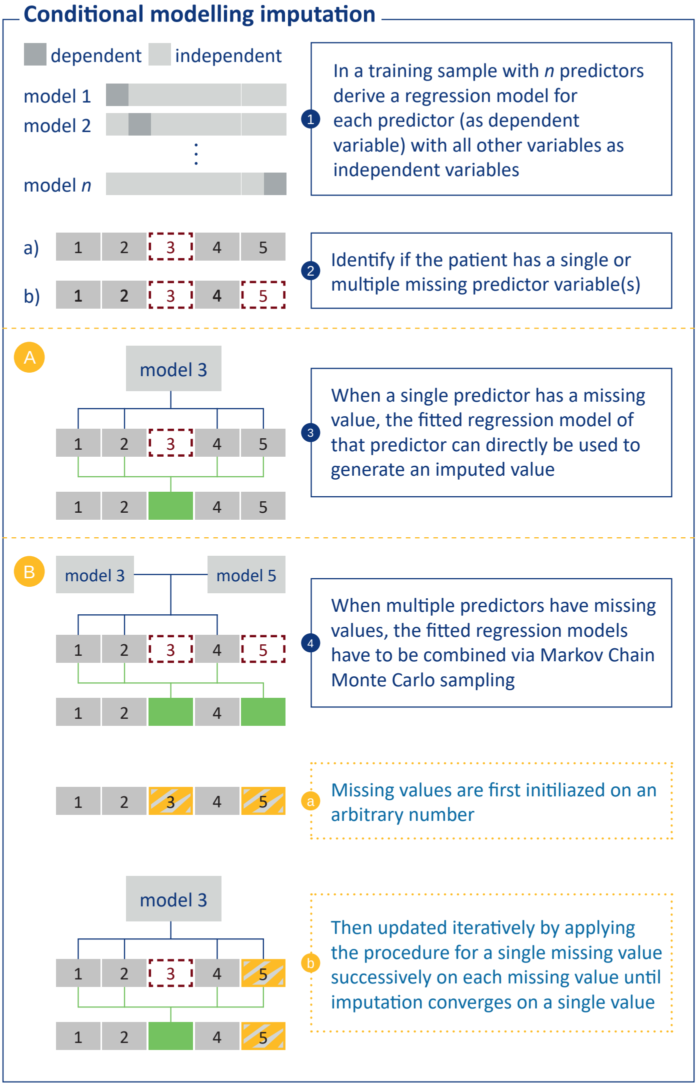
Figure 2. Joint modelling imputation


Joint modelling imputation
Training sample
1 Estimate means and covariance of all relevant predictors using training data to estimate **joint normal distribution**
Individual patient data
2 Identify missing variables given an individual patient
Imputation
3 Use derived distribution to generate imputation for missing variable

Conditional modeling imputation

The training sample is used to derive a flexible (e.g. regression) model for each predictor (as dependent variable) with all other predictor variables as independent variables (Figure 3). These models describe the conditional distribution of each predictor, and usually need to be estimated using a Gibbs sampling procedure (as predictor values may also be missing in the training sample). Due to the flexible nature of these conditional models, it is no longer assumed that predictor variables of the training sample are normally distributed (as does JMI). For instance, a logistic regression model can be used to estimate the conditional distribution of a binary predictor variable (e.g. current smoker). Subsequently, when the smoking status for a new patient is unknown, the logistic regression model can be used to generate a probability that they are a current smoker. This probability can directly be used as imputed value (in case only 1 imputation is needed). Alternatively, if multiple imputations are required, a Bernoulli distribution (with aforementioned probability) can be used to sample multiple (discrete) values for the patient's current smoking status. If multiple predictor values are missing, the conditional models need to be used successively using an iterative Monte Carlo procedure (Appendix A).

Figure 3. Conditional modelling imputation



1 In a training sample with *n* predictors derive a regression model for each predictor (as dependent variable) with all other variables as independent variables

2 Identify if the patient has a single or multiple missing predictor variable(s)

3 When a single predictor has a missing value, the fitted regression model of that predictor can directly be used to generate an imputed value

4 When multiple predictors have missing values, the fitted regression models have to be combined via Markov Chain Monte Carlo sampling

a Missing values are first initiliazed on an arbitrary number

b Then updated iteratively by applying the procedure for a single missing value successively on each missing value until imputation converges on a single value

Simulation study

Cardiovascular disease prevention is an example of a setting where risk-guided management of predictors – smoking, blood pressure, cholesterol - is common practice [125]. Numerous risk prediction models have been developed and the (international) guidelines advocate the use of risk classification to inform treatment decisions [126,127]. These models are typically implemented in a CDSS, where a patient's characteristics of the predictors can be entered manually or are automatically retrieved from the patient's EHR [19,109,128].

For this study we used a data set of the ongoing Utrecht Cardiovascular cohort initiative (UCC). This cohort includes all patients who come for a first-time visit the Center for Circulatory Health at the UMC Utrecht for the evaluation of a symptomatic vascular disease or an asymptomatic vascular condition. A minimum set of predictors, according to the Dutch Cardiovascular Risk Management Guidelines, is collected in all patients. No data on outcomes (i.e. time-to-event data) was recorded. UCC has been approved by the Institutional Review Board of the UMC Utrecht (Biobank Ethics committee). For the present analyses an anonymized dataset was used of the UCC cohort up to November 2018 [129,130].

The sample consisted of 3880 patients with information on 23 variables, measured during the patient's visit (Table 1 and Appendix B). To ensure full utilization of the observed data, we completed this dataset using all 23 variables in k-nearest neighbor imputation, which aggregates the values of the *k* nearest neighbors to an imputation [131].

To evaluate the quality of the three selected imputation methods in individual patients, a leave-one-out-cross-validation (LOOCV) procedure was used in the completed UCC dataset. In LOOCV, all but one patient are used as the training sample from which the overall mean or proportion (method 1), or imputation models (method 2 and 3) are derived (Figure 4). In the remaining hold-out patient, missing values are introduced for one or more predictor variables. As we apply each scenario to each patient exactly once, the missing data mechanism is essentially missing-completely-at-random (MCAR) [121]. The summary information from the training sample is then used to impute the missing predictor values in the hold-out patient. For CMI and JMI, we generated 50 imputations for each missing predictor value. This process is repeated until all patients have been taken from the dataset exactly once.

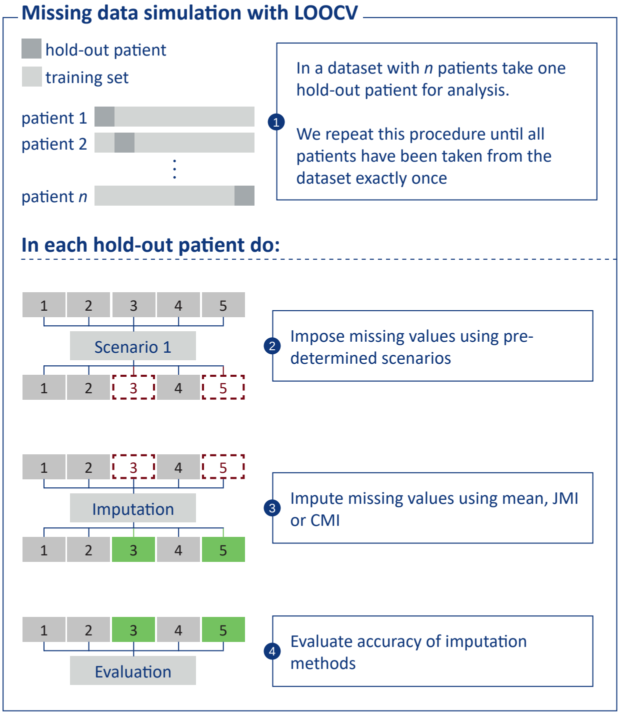
We consider 8 scenarios where missing values occur for one predictor variable, and 8 scenarios where multiple predictor variables are simultaneously missing (Figure 5). A detailed description of how the scenarios were selected and of the R code are listed in Appendix C and D respectively.

Table 1. Descriptive statistics (after imputation)

	Part of missing data scenarios	Mean (sd) or n/total (%)*	Original missing %
Age (years)	No	61.7 (18.2)	0.00
Sex (1=female; 0=male)	No	1987/3880 (51.2)	0.00
Smoking (1=yes; 0= no)	No	363/3880 (9.4)	24.07
SBP (mmHg)	Yes	142.8 (24.2)	10.54
TC (mmol/l)	Yes	5.1 (1.2)	24.54
LDL-c (mmol/l)	Yes	3.1 (1.3)	26.01
HDL-c (mmol/l)	No	1.4 (0.4)	25.39
eGFR (mL/min/1.73m2)	Yes	81.8 (24.6)	15.98
History of CVD (1=yes; 0= no)	Yes	1971/3880 (50.8)	23.45
History of PAD (1=yes; 0= no)	No	335/3880 (8.6)	23.45
History of CHD (1=yes; 0= no)	No	591/3880 (15.2)	23.45
History of CHF (1=yes; 0= no)	No	284/3880 (7.3)	23.45
History of CVA (1=yes; 0= no)	No	579/3880 (14.9)	23.45
History of DM (1=yes; 0= no)	No	607/3880 (15.6)	23.45
Polyvascular disease	No	0.6 (0.7)	23.45
# of medications	No	0.8 (1.7)	27.24
BP lowering medication (1=yes; 0= no)	No	705/3880 (18.2)	27.24
Statin (1=yes; 0= no)	No	415/3880 (10.7)	27.24
HbA1c (mmol/mol)	No	40 (10.7)	26.37
Years since first CVD (years)	Yes	4.6 (8.1)	26.21
Diabetes (1=yes; 0= no)	Yes	755/3880 (19.5)	8.12
Diabetes duration (years)	No	11.3 (7.3)	86.11
Pulse pressure (mmHg)	No	61.7 (18.9)	10.54

Legend – SBP: systolic blood pressure, TC: total cholesterol, LDL-c: low-density lipoprotein cholesterol, HDL-c: high-density lipoprotein cholesterol, eGFR: estimated glomerular filtration rate according to the CKD epi formula, CVD: cardiovascular disease, PAD: peripheral artery disease, CHD: coronary heart disease, CHF: chronic heart failure, CVA: cerebrovascular accident, DM: diabetes mellitus, BP: blood pressure, HbA1c: glycated hemoglobin. * after KNN-imputation

Figure 4. Missing data simulation procedure



Missing data simulation with LOOCV

In a dataset with *n* patients take one hold-out patient for analysis.

We repeat this procedure until all patients have been taken from the dataset exactly once

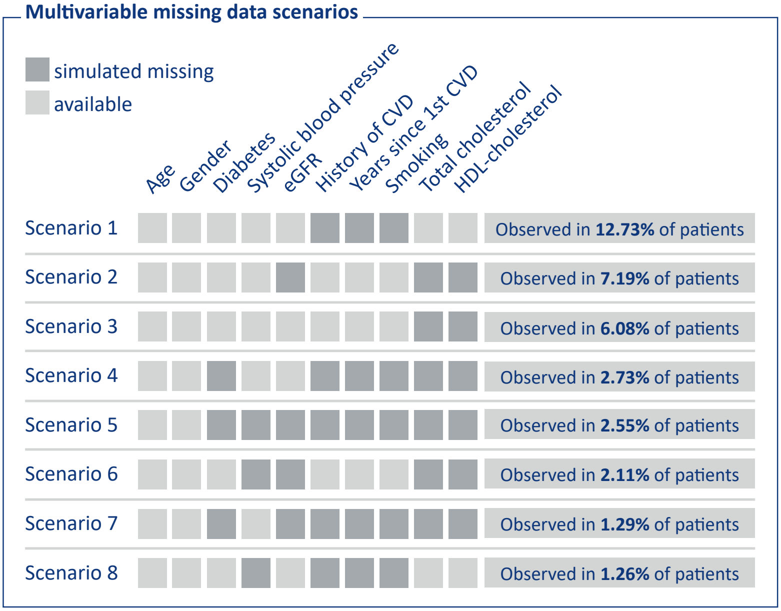
In each hold-out patient do:

Impose missing values using pre-determined scenarios

Impute missing values using mean, JMI or CMI

Evaluate accuracy of imputation methods

Figure 5. Multivariable missing data scenarios



Multivariable missing data scenarios

■ simulated missing

□ available

	Age	Gender	Diabetes	Systolic blood pressure	eGFR	History of CVD	Years since 1st CVD	Smoking	Total cholesterol	HDL-cholesterol	
Scenario 1	□	□	□	□	□	■	■	■	□	□	Observed in **12.73%** of patients
Scenario 2	□	□	□	□	■	□	□	□	■	■	Observed in **7.19%** of patients
Scenario 3	□	□	□	□	□	□	□	□	■	■	Observed in **6.08%** of patients
Scenario 4	□	□	■	□	□	■	■	■	■	■	Observed in **2.73%** of patients
Scenario 5	□	□	■	■	■	■	■	■	■	■	Observed in **2.55%** of patients
Scenario 6	□	□	□	■	■	□	□	□	■	■	Observed in **2.11%** of patients
Scenario 7	□	□	■	□	■	■	■	■	■	■	Observed in **1.29%** of patients
Scenario 8	□	□	□	■	□	■	■	■	□	□	Observed in **1.26%** of patients

Measures of performance

To evaluate the performance of the three imputation methods we used four performance metrics:

1. We calculated the root mean squared error (RMSE) between the average of the multiple imputed predictor values (i.e. 50 imputations) and the true, original (i.e. before the simulation of missing) predictor value to evaluate the accuracy of the imputations. The RMSE is a performance measure that aggregates error due to bias and variability. Generally, an RMSE of zero means perfect imputation and an increasing RMSE means decreasing performance of the imputation. Clinical relevance of an RMSE depends on the natural range of the predictor. For example, an RMSE of 0.5 is large for LDL-c (mean 3.0 SD 1.3 mmol/L) but not for SBP (mean 143 SD 24 mmHg).
2. For each hold-out patient, we assessed whether the original predictor value was in the 95% confidence interval around the imputed predictor value. Subsequently, we calculated the proportion of confidence intervals that consisted the original value (coverage). For a 95% CI, the coverage should ideally be equal to 95% [132]. A lower coverage translates to imputed predictor values that are too precise (which in turn may lead to estimates of predicted risk that are too precise), whereas a coverage above 95% indicates that imputed predictor values are too imprecise [116]. We assessed coverage only for continuous predictor variables.
3. We assessed the effect on treatment decision support for blood pressure in patients with manifest cardiovascular disease (n=1971) to evaluate the clinical implications of the imputed predictor values. Guidelines indicate that all patients with a history of CVD should receive blood pressure lowering treatment when their blood pressure is higher than 140/90mmHg [106,130]. We adopted the LOOCV approach, and set values for SBP missing in the hold-out patient. Subsequently, we imputed the missing value and compared the treatment decision for the true value with the treatment decision for the imputed value (SBP <> 140mmHg). Afterwards, we calculated the sensitivity, specificity, positive predictive value and negative predictive value. Also, we illustrated the importance of reporting confidence intervals based on imputed values to inform the discussion around treatment commencement.

We compared the risk predictions that were obtained in the absence of missing values (i.e. in the original data) with the risk predictions that are based on imputations to evaluate the impact of the imputed values on the precision of predicted risk. Ideally, the predictions that are based on imputed values should have a similar distribution as the predictions that are derived from the complete original data. To explore any deviation, we assessed the interquartile range of predicted risk for a single missing predictor scenario and for a multiple missing predictor scenario. Rather than developing a prediction model ourselves, we used the previously developed SMART prediction model for risk of 10 year recurrent vascular disease as an example [133]. The prediction model includes 11 variables: age, sex, current smoker, SBP, diabetes, history of cerebrovascular disease, aortic aneurysm or peripheral vascular disease, polyvascular disease, HDL-cholesterol and total cholesterol.

Results

Root Mean squared error

With the exception of smoking, all predictor variables in single missing predictor scenarios had a lower RMSE when using JMI or CMI as compared to mean imputation (Table 2). For most multiple missing predictor scenarios, the RMSE is consistently lower when using JMI or CMI as compared to mean imputation. The exceptions being history of CVD and smoking. Performance diminished as more variables were missing. For example, the RMSE of years since 1st CVD event are 6.30 and 6.26 for JMI and CMI respectively when univariately missing, whilst mean imputation has a RMSE of 8.06. When additional variables (e.g., SBP, history of CVD and smoking) are missing, the RMSE for years since 1st CVD event for both JMI and CMI increases to 7.58 and 7.84 respectively.

Coverage rate

For JMI, the coverage reached nominal levels for all single missing predictor scenarios and multiple missing predictor scenarios (Table 3). For CMI, the coverage reached nominal levels for all single missing predictor scenarios and multiple missing predictor scenarios. For mean imputation, coverage was 0% by definition for all imputed predictors because no uncertainty is taken into account.

Clinical decision accuracy

When assessing the treatment decision for blood pressure management according to the prevailing clinical guidelines (see above), we selected 1971 out of the total 3880 patients with manifest cardiovascular disease. We found that 1134 patients (57.53%) should be treated. However, when blood pressure values were set to missing, the overall mean imputed value was 142 mmHg (Table 1), which is just above the treatment threshold of 140 mmHg. As a result, everyone would have been treated when adopting overall mean imputation, such that 837 patients (42.47%) would have been treated unnecessarily. When adopting JMI or CMI, only 16.08% or, respectively, 15.98% of patients would have been treated unnecessarily (Table 4). Hence, imputation of missing blood pressure values using CMI or JMI was more adequate than mean imputation in terms of decision making.

Table 2. RMSE values for each combination of, individual or multiple, missing predictor values

Single missing variable scenarios									
		Diabetes	SBP	eGFR	History of CVD	Years since 1st CVD	Smoking	Total cholesterol	HDL-cholesterol
Mean imputation*		0.40	24.24	24.56	0.50	8.06	0.29	1.24	0.36
JMI		0.17	22.31	19.60	0.39	6.30	0.30	1.19	0.34
CMI		0.21	22.29	19.69	0.39	6.26	0.29	1.19	0.34

Multiple missing variables scenarios									
Method	Scenario	Diabetes	SBP	eGFR	History of CVD	Years since 1st CVD	Smoking	Total cholesterol	HDL-cholesterol
JMI	1				0.46	7.59	0.30		
CMI	1				0.51	7.78	0.29		
JMI	2			19.68				1.19	0.35
CMI	2			19.69				1.20	0.35
JMI	3							1.19	0.33
CMI	3							1.19	0.35
JMI	4	0.17			0.48	7.65	0.30	1.22	0.35
CMI	4	0.20			0.50	7.83	0.28	1.21	0.35
JMI	5	0.17	22.62	19.86	0.47	7.66	0.30	1.23	0.35
CMI	5	0.21	22.48	19.87	0.51	7.86	0.29	1.22	0.35
JMI	6		22.45	19.61				1.19	0.34
CMI	6		22.50	19.59				1.20	0.34
JMI	7	0.17		19.83	0.48	7.69	0.30	1.22	0.35
CMI	7	0.21		19.75	0.50	7.84	0.29	1.23	0.35
JMI	8		22.36		0.46	7.58	0.30		
CMI	8		22.35		0.51	7.84	0.29		

Legend – JMI: joint modelling imputation, CMI: conditional modelling imputation, SBP: systolic blood pressure, eGFR: estimated glomerular filtration rate according to the CKD epi formula, CVD: cardiovascular disease

Multiple missing predictor scenarios: (1) history of CVD, years since 1st CVD event & smoking, (2) eGFR, total cholesterol & HDL-cholesterol, (3) total cholesterol & HDL-cholesterol, (4) all variables but SBP & eGFR, (5) all variables, (6) SBP, eGFR, total cholesterol & HDL-cholesterol, (7) all variables but SBP and (8) SBP, history of CVD, years since 1st CVD event and smoking.

* Mean imputation is only included in the single missing variable scenarios as the performance of the model, when multiple variables are missing, is equivalent.

Table 3. Coverage values for each combination of, individual or multiple, imputations

Single missing variable scenarios						
		SBP	eGFR	Years since 1st CVD	Total cholesterol	HDL-cholesterol
	JMI	0.945	0.948	0.952	0.952	0.950
	CMI	0.945	0.948	0.954	0.953	0.948

Multiple missing variable scenarios						
Method	Scenario	SBP	eGFR	Years since 1st CVD	Total cholesterol	HDL-cholesterol
JMI	1			0.951		
CMI	1			0.948		
JMI	2		0.947		0.951	0.951
CMI	2		0.946		0.955	0.949
JMI	3				0.949	0.951
CMI	3				0.950	0.949
JMI	4			0.951	0.950	0.951
CMI	4			0.949	0.952	0.952
JMI	5	0.944	0.947	0.951	0.952	0.951
CMI	5	0.948	0.948	0.946	0.953	0.953
JMI	6	0.945	0.950		0.951	0.948
CMI	6	0.948	0.948		0.949	0.949
JMI	7		0.950	0.951	0.951	0.951
CMI	7		0.947	0.950	0.948	0.951
JMI	8	0.945		0.952		
CMI	8	0.945		0.950		

Legend – JMI: joint modelling imputation, CMI: conditional modelling imputation, SBP: systolic blood pressure, eGFR: estimated glomerular filtration rate according to the CKD epi formula, CVD: cardiovascular disease

Multiple missing predictor scenarios: (1) history of CVD, years since 1st CVD event & smoking, (2) eGFR, total cholesterol & HDL-cholesterol, (3) total cholesterol & HDL-cholesterol, (4) all variables but SBP & eGFR, (5) all variables, (6) SBP, eGFR, total cholesterol & HDL-cholesterol, (7) all variables but SBP and (8) SBP, history of CVD, years since 1st CVD event and smoking.

Table 4. 2x2 tables of guideline adherence to treatment threshold given the point estimate of each method

		True value		
Mean imputation		**Treatment advised (≥ 140mmHg)**	**Treatment not advised (< 140mmHg)**	**Totals**
Point estimate	Treatment advised (> 140mmHg)	1134	837	1971
	Treatment not advised (< 140mmHg)	0	0	0
Totals		1134	837	1971

Sensitivity 100%, specificity 0%, Positive Predictive Value 58%, Negative Predictive Value (cannot be calculated) %

		True value		
Joint modeling imputation		**Treatment advised (≥ 140mmHg)**	**Treatment not advised (< 140mmHg)**	**Totals**
Point estimate	Treatment advised (> 140mmHg)	946	317	1263
	Treatment not advised (<140 mmHg)	188	520	708
Totals		1134	837	1971

Sensitivity 83%, specificity 62%, Positive Predictive Value 75%, Negative Predictive Value 73%

		True value		
Conditional modeling imputation		**Treatment advised (≥ 140mmHg)**	**Treatment not advised (< 140mmHg)**	**Totals**
Point estimate	Treatment advised (> 140mmHg)	960	315	1275
	Treatment not advised (< 140mmHg)	174	522	696
Totals		1134	837	1971

Sensitivity 85%, specificity 62%, Positive Predictive Value 75%, Negative Predictive Value 75%

To illustrate the importance of measuring uncertainty we provided an example in which we compare the use of imputation in a real-life situation (table 5). In the example a patient with an imputed SBP of 144mmHg was given an indication for blood pressure lowering treatment according to the guidelines [106]. However, given that the uncertainty around the imputed predictor value crosses the treatment line of 140mmHG (scenario A), there is reasonable doubt this imputation is too uncertain to be used for treatment decision making.

Table 5. Clinical interpretation of imputed SBP values and 95% confidence intervals from a patient with a history of CVD

	True	Scenario A	Scenario B
SBP (95%CI)	144	144 (138-150)	144 (142-146)
Treatment based on point estimate	>140mmHg, Start treatment	>140mmHg, Start treatment	>140mmHg, Start treatment
Treatment based on 95% CI	NA	Uncertain	>140mmHg, Start treatment

* estimated. Legend: SBP = systolic blood pressure, 95% CI = 95% confidence interval A= hypothetical situation where imputed value interval contains treatment threshold B=hypothetical situation where imputed value interval does not contain treatment threshold

Effect on risk predictions

The predicted risks, given each method, did not seem to deviate much from the originally predicted risk, given the complete data (table 6). When assessing the single missing predictor scenario there was a difference between overall mean imputation (median difference of -1.713% to the originally predicted risk) and the combination of JMI and CMI (median difference of respectively 0.301% and 0.399% to the originally predicted risk). Further, we found that predicted risks for mean imputation were more similar when compared to the complete data (standard deviation = 15.12 versus the reference of 18.91). In contrast standard deviations of JMI and CMI were 17.87 and 17.86 respectively.

In the multiple missing predictor scenario, there was a similar difference between mean imputation (median difference of -2.064% to the originally predicted risk) and JMI and CMI (median difference of respectively 0.375% and 0.390% to the originally predicted risk). With multiple missing predictors the predicted risks for mean imputation were again more similar than the predicted risk given the complete data (standard deviation = 14.42 versus the reference of 18.91). The standard deviations of JMI and CMI were 17.67 and 17.68 respectively.

The difference between mean imputation and both JMI and CMI is especially apparent in higher risk patients (i.e., 75% IQR) where mean imputation, as expected, underestimates the risk. This is because mean imputation pulls the risk predictions of patients with missing values towards the prediction for an “average” patient. As such JMI and CMI perform much better with regards to their impact on prediction in higher risk patients, when compared to mean imputation.

Table 6. Differences in predicted 10-year risk of CVD for both a single missing predictor scenario and a multiple missing predictor scenario

Single missing predictor: eGFR	**25% IQR**	Absolute risk difference to completed data	**Median**	Absolute risk difference to completed data	**75% IQR**	Absolute risk difference to completed data
Predicted risk complete data	8.382%	-	13.711%	-	28.170%	-
Predicted risk (mean)	7.287%	-1.095%	11.997%	-1.713%	23.035%	-5.135 %
Predicted risk (joint)	8.767%	0.385%	14.012%	0.301%	27.734%	0.435%
Predicted risk (conditional)	8.786%	0.404%	14.110%	0.399%	27.783%	0.387%

Multiple missing predictors: SBP, TC, LDL-c and eGFR	**25% IQR**	Absolute risk difference to completed data	**Median**	Absolute risk difference to completed data	**75% IQR**	Absolute risk difference to completed data
Predicted risk complete data	8.382%	-	13.711%	-	28.170%	-
Predicted risk (mean)	7.473%	-0.909%	11.647%	-2.064 %	22.692%	-5.478%
Predicted risk (joint)	8.809%	0.427%	14.085%	0.375%	28.410%	0.240%
Predicted risk (conditional)	8.786%	0.404%	14.100%	0.390%	28.267%	0.097%

Legend: eGFR = estimated glomerular filtration rate, SBP = systolic blood pressure, TC = total cholesterol, LDL-c = low density lipoprotein cholesterol, IQR = inter quartile range.

Discussion

This project described the development and performance of three imputation methods to handle missing data on an individual patient level in real-life clinical decision making. As expected, both JMI – using draws from a normal distribution constructed from means and covariance in the training sample and observed values in the patient – and CMI – using a conditional distribution of each variable based on regression models fitted on all other variables, – were more accurate and showed better coverage as compared to mean imputation, resulting in fewer inappropriate treatment decisions and lower impact on predicted risk.

The accuracy measures – RMSE, coverage and clinical decision accuracy – were comparable for JMI and CMI. Hence, both methods can be used for generating live imputations in routine care. Based on usability, we recommend JMI, as its implementation in decision support systems is fairly straightforward and only requires information on the mean and covariance of the target population. Although its assumption of multivariate normality may be unrealistic for real life clinical data, simulation studies have demonstrated that this rarely affects the performance of imputation [134–136].

Previous studies on imputation methods to handle missing data on an individual patient level have focused on the impact of missing values on the performance of a prediction model and evaluated the use of mean imputation as well as the (re)development of a simplified prediction model [118,119]. Mean imputation was recommended due to its applicability in practice and relatively good performance compared to other models, but was considered insufficient when strong predictors were missing. For this reason, our proposed multiple imputation models appear particularly relevant when strong or multiple predictors are missing. This was confirmed in our simulation study: RMSE and coverage did not much deteriorate with increasing number of predictor values that were simultaneously missing for the individual patient. It is noted that our simulations, due to the way missing data was introduced, were not able to distinguish between various mechanisms by which data can be missing, e.g. data that is missing at random (MAR) versus data that is missing completely at random (MCAR) [121].

Furthermore, because the described imputation methods can accommodate for numerous patient characteristics that are not necessarily disease-specific, they are highly scalable to other settings and populations. However, it is likely that some local tailoring is necessary when imputation models are derived from specific studies or settings that do not fully match the intended target population. For JMI, the means and covariances could for instance simply be replaced by their respective values in a local ”training” sample. For CMI, the regression coefficients can be revised using recently described updating methods [137]. When the (local) training data are affected by missing predictor values, advanced methods exist to estimate the mean and the covariance [138]**.** All methods can be potentially incorporated within an EHR based computerized decision support system and generate imputations based on observed data from individual patients extracted from the EHR. Evidently, before implementing imputation models in clinical practice, it is of the utmost importance to assess their validity, likely impact on treatment decisions, patient outcomes, as well as any practical, security and ethical constraints.

Although multiple imputation offers a computational framework to account for missing values, we recommend to always first optimize data collection, and to avoid having missing values: a clinical decision making should never be based solely on imputed values. However, imputed values can serve as a proxy for prior risk, setting an indication for more (advanced) diagnostic tests. This is especially useful for expensive tests, tests associated with complications or when tests are unavailable. Additional diagnostic testing should preferably only be performed when it is expected to change treatment and the potential clinical benefit outweighs the tests risk). Note that in this study we do not take into account the (un)certainty around imputed values when assessing treatment decision support.

In cardiovascular risk management, the decision to start treatment of a risk factor is based on i) the predicted risk for a cardiovascular disease or patient characteristics that are per definition associated with a high risk for cardiovascular disease and ii) the absolute value of the risk factor itself. We focused on imputation models to recover the missing value and to quantify its uncertainty. We demonstrated that the choice of imputation method may impact risk predictions and decision making. Whilst the magnitude of this effect was not always substantial, it may vary according to the number of missing predictors and their weight in the decision-making process and should therefore be evaluated when applying these models in different settings and populations.

Lastly, traditional (e.g. regression-based) prediction models assume complete input data, which is often not realistic in routine clinical practice. Although we developed models for imputing the missing values, which can subsequently be used to generate predictions, it is also possible to develop prediction models that do not require complete information on the predictors. Well-known examples are the use of decision trees with surrogate or sparsity-aware splits [37,39,139], the use of submodels [140], or the use of missing indicator variables [4]. More research is warranted to evaluate whether these methods may offer any improvement in model predictions, as well facilitate their implementation in routine care.

In summary, this study describes three imputation methods to handle missing values in the context of computerized decision support systems in clinical practice. We found that JMI and CMI provide imputations that are closer to the original value (as compared to mean imputation) and able to reflect uncertainty due to missing data. We therefore recommend their implementation in situations where information on relevant predictors is often incomplete due to practical constraints.

Disclosures

Funding sources

This work was supported by the Netherlands Heart foundation (public-private study grant, grant number: #2018B006); and the Top Sector Life Sciences & health (PPP allowance made available to Netherlands Heart Foundation to stimulate public-private partnerships).

Competing interests

The authors declare no competing interests.

Author contributions

TD, KM, KG and SN conceived of the presented idea, in correspondence with earlier work by FA, MB and KG. SN, TD, JH and KG derived the models and analyzed the data. TD and JH verified the analytical methods. TD supervised the findings of this work. SN, KG, TD and JH contributed to the interpretation of the results. KG and SN wrote the manuscript. All authors provided critical feedback and helped shape the research, analysis and manuscript.

Data availability statement

The data that support the findings of this study are available from the UCC upon reasonable request (https://www.umcutrecht.nl/en/Research/Strategic-themes/Circulatory-Health/Facilities/UCC).

Acknowledgements

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

Appendix

Appendix A. Explanation Imputation Methods

In this supplementary material we will explain which values are required to be calculated in the training data and which R packages are used per implemented method. We will also explain step by step what we do for each method. We will focus specifically on JMI and CMI as mean imputation is relatively straightforward. In addition, we will shortly cover the requirements and step-by-step instructions for each evaluation method used. All code is added in appendix C.

Joint modelling imputation

As stated JMI allows tailored imputations, making use of covariances between all predictors. More specifically, imputations are randomly sampled from a (multivariate) normal distribution that is conditioned on the observed predictor values. For binary variables, a logistic regression model is used to transform the drawn continuous values into discrete imputations.

To implement JMI we first have to calculate the expectation (mean) of all variables included in the data and save this in a single vector. Additionally, a covariance matrix of the data has to be saved in a separate object. We also save the class of each variable included in the data. On a patient-by-patient basis we extract which variables are missing and which are not missing. From the variables that are not missing we save the observed values in a separate vector. Then, using the *rcmvnorm* function in R, we estimate the conditional multivariate normal distribution using the provided expectations (mu), covariance matrix (sigma), dependent variables (i.e. names of the missing variable), the given non-missing variables and all observed values [138].

For example, consider a situation where we have two variables of interest x_1 (e.g. blood pressure) and x_2 (e.g. Body Mass Index). These variables have been fully observed in a previous cohort study, where we calculated their respective means μ_1 and μ_2, their respective variance σ_1^2 and σ_2^2, and their correlation $\rho_{1,2}$. Consider now the encounter with a single new patient for which the Body Mass Index has been measured (i.e. x_2 is known), but for which the blood pressure cannot be retrieved (and therefore is missing). Assuming that BMI and blood pressure follow a bivariate Normal distribution, likely values for x_1 (given that x_2 is known) can be described by a Normal distribution with mean $\mu_{1|2}$ and variance $\sigma_{1|2}^2$ where:

$$\mu_{1|2} = \mu_1 + \rho_{1,2}\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2)$$

and

$$\sigma_{1|2}^2 = (1 - \rho_{1,2}^2)\sigma_2^2$$

Hence, imputations for x_1 can simply be generated by drawing random samples from the distribution N($\mu_{1|2}, \sigma^2_{1|2}$). If only a single imputation is desired, the most likely value for x_1 is simply given by $\mu_{1|2}$.

Consider now that x_1 is a binary variable (e.g. smoking) instead of a continuous variable. In this case, samples from N($\mu_{1|2}, \sigma^2_{1|2}$), denoted as $x^*_{1|2}$, may not be appealing to be used as imputation because they are unlikely to be discrete (e.g. 0 or 1) and may even take negative values. For this reason, imputations for x_1 are generated according to $Bernoulli\ (\frac{1}{1+\exp\ (-x^*_{1|2})})$. Note that for each imputation, a new value of $x^*_{1|2}$ need to be sampled.

The code that was used can be found in appendix C (function: *joint.MI()*). Note that the amount of imputed values is specified beforehand (i.e. n.imp). Also note that the mean vector *mu* and covariance matrix *sigma* of the training data can simply be obtained by applying the R functions *colMeans()* and *cov()* to the corresponding data frame. In case the training data are affected by missing values, R packages such as *mvnmle* can be used to obtain maximum likelihood estimates for the original mean and covariance.

Conditional modelling imputation

To implement CMI we, before calculating other separate values, estimate each conditional model based on the training data. This entails iterating over all columns in the training data, specifying a conditional model (e.g. logistic) based on the type of dependent variable (e.g. binary). We save the conditional models in a list to be used in our imputation.

Note that we use the function *estimice* instead of *glm* when modeling continuous variables. This approach is analogous to the imputation of missing continuous variables in the R package *mice* when adopting the *mice.impute.norm* function. The *estimice* function is a least squares implementation of ridge regression, and can therefore better handle situations where training samples are relatively small.

The fitted regression models (one for each variable of interest) can then be used to generate imputations in new patients. In similar fashion to JMI, our implementation of CMI requires the means, covariance and data classes of the training data. The method first checks, on a patient-by-patient basis, how many variables are missing. We start with this distinction as single and multiple missing variable require a different approach.

In short: when a single predictor has a missing value, the fitted regression model of that predictor can directly be used to generate an imputed value. When multiple predictors have jointly missing values, the conditional models need to be combined through Markov Chain Monte Carlo sampling [141]. Missing values are then first initialized to an arbitrary value, and updated iteratively by applying the procedure for a single missing value successively on each missing value.

More specifically: when the patient has a single missing variable, we specify the variables on which each model should be based, thus excluding the missing variable. If the missing variable is binary we use the regression coefficients of the relevant imputation model (i.e. as estimated in *conditional.estimation* function) and its corresponding covariance matrix to draw a random sample of imputation coefficients. Hereto, we use a multivariate T-distribution as implemented in the R function *rmvt* [142]. The imputation coefficients are then used to calculate a probability, which is then used with a Bernoulli distribution to draw an imputation for the missing value. This process of drawing the betas, calculating a probability and drawing a value from the Bernoulli distribution is done the amount of times we specify (i.e. n.imp).

When the missing variable is continuous, we use the Bayesian multiple imputation approach described by van Buuren and implemented in the R function *mice.impute.norm* [116]. This approach generates imputation coefficients by sampling from a posterior distribution that is based on the regression coefficients of the relevant imputation model (i.e. as estimated in *conditional.estimation* function) and standard non-informative priors. This adaptation was necessary to ensure that estimation uncertainty for the residual error variance is also taken into account when generating imputations.

When two or more variables are missing for a single patient, the conditional imputation models need to be used in conjunction to generate reliable imputations. Because each imputation model requires complete data on all but one variable, we first initialize each missing variable with a random starting value. To this purpose, we use the means and covariance of the training sample. Then, on a variable-by-variable basis, the starting values are updated by imputing them using all other (original or initiated) values. This process of updating randomly initiated values iterates over each missing variable and is then repeated for a specified number of times to also replace the updated values numerous times. This cyclic generation of updated values is necessary to ensure that the imputed variables depend on each other and the observed data, but no longer on their initial values. Updated values from the last iteration are then extracted and used as the imputed values. The process of initializing starting values, updating these values and extracting them is repeated for a prior specified amount (i.e. n.imp).

The code in appendix C (functions: *conditional.estimation()* and *conditional.MI()*) was used to implement conditional modelling imputation. Note that the object *model_estimation* is a list containing the conditional imputation models for each variable, and can be obtained using the function *conditional.estimation()*.

Evaluation measures

Each method provided an three-dimensional array of the data, where the third dimension consisted of the prior specified amount of imputations (i.e. 50 in our analysis) for each of the missing variables. When calculating the RMSE we square the difference between the mean of those imputations and the true value, which gives us a vector of squared deviations. The root of the mean of that vector is the RMSE reported in this study. We calculated the coverage rate by first calculating a 95% confidence interval for each imputed predictor in the hold-out patient according to

$$mean(x_i^*) \pm t_{50}^{0.05/2} \times sd(x_i^*)$$

Where xi is the ith imputed value (out of a total of 50), and t is a value from a two-sided t-distribution with 50 degrees of freedom. We then specify a binary indicator showing if the confidence interval included the true value. Taking the mean of the binary indicator gives us the percentage of confidence intervals containing the true value.

The code in appendix C was used to calculate both evaluation measures for a single missing predictor (function: *test_single_missing()*). Note that the object *knn1* is the exemplar dataset where all predictors are fully observed. To accommodate deriving the necessary population characteristics from the training data we completed any missing values in the UCC data using K-nearest neighbor imputation (KNN) [131]. In addition the character *test_var* specifies the variable for which" missing values are introduced in the Jack-knife procedure.

Appendix B – Descriptive statistics before imputation

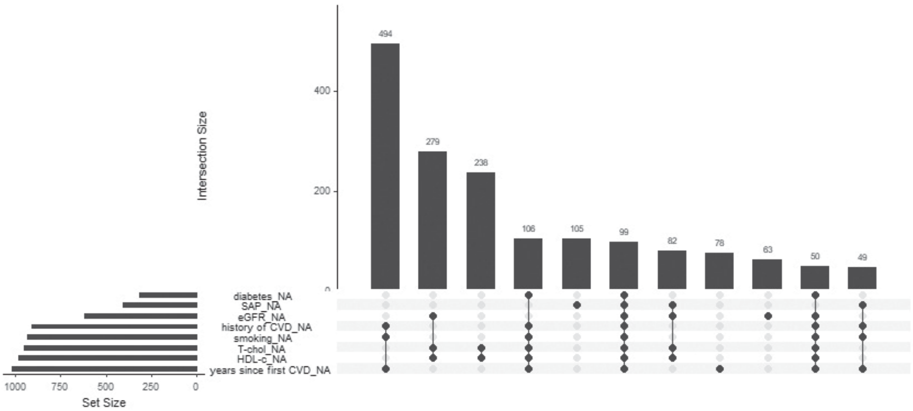
	Mean (sd) or n/total (%)*	% Missing
Age (years)	61.7 (18.2)	0.00
Sex (1=female; 0=male)	1987/3880 (51.2)	0.00
Smoking (1=yes; 0= no)	363/2583 (14.05)	24.07
SBP (mmHg)	141.49 (24.2)	10.54
TC (mmol/l)	5.2 (1.4)	24.54
LDL-c (mmol/l)	3.0 (1.2)	26.01
HDL-c (mmol/l)	1.4 (0.4)	25.39
eGFR (mL/min/1.73m2)	80.7 (25.6)	15.98
History of CVD (1=yes; 0= no)	1063/1907 (55.7)	23.45
History of PAD (1=yes; 0= no)	271/2699 (10.0)	23.45
History of CHD (1=yes; 0= no)	472/2498 (18.9)	23.45
History of CHF (1=yes; 0= no)	283/2687 (10.5)	23.45
History of CVA (1=yes; 0= no)	449/2521 (17.8)	23.45
History of DM (1=yes; 0= no)	607/2363 (25.6)	23.45
Polyvascular disease	0.5 (0.8)	23.45
# of medications	1.0 (1.9)	27.24
BP lowering medication (1=yes; 0= no)	599/2224 (26.9)	27.24
Statin (1=yes; 0= no)	395/2428 (16.3)	27.24
HbA1c (mmol/mol)	40.7 (11.8)	26.37
Years since first CVD (years)	3.8 (8.5)	26.21
Diabetes (1=yes; 0= no)	755/2810 (26.9)	8.12
Diabetes duration (years)	14.9 (12.0)	86.11
Pulse pressure (mmHg)	61.7 (19.5)	10.54

Legend – SBP: systolic blood pressure, TC: total cholesterol, LDL-c: low-density lipoprotein cholesterol, HDL-c: high-density lipoprotein cholesterol, eGFR: estimated glomerular filtration rate according to the CKD epi formula, CVD: cardiovascular disease, PAD: peripheral artery disease, CHD: coronary heart disease, CHF: chronic heart failure, CVA: cerebrovascular accident, DM: diabetes mellitus, BP: blood pressure, HbA1c: glycated hemoglobin. * after KNN-imputation

Appendix C – Selection of variables

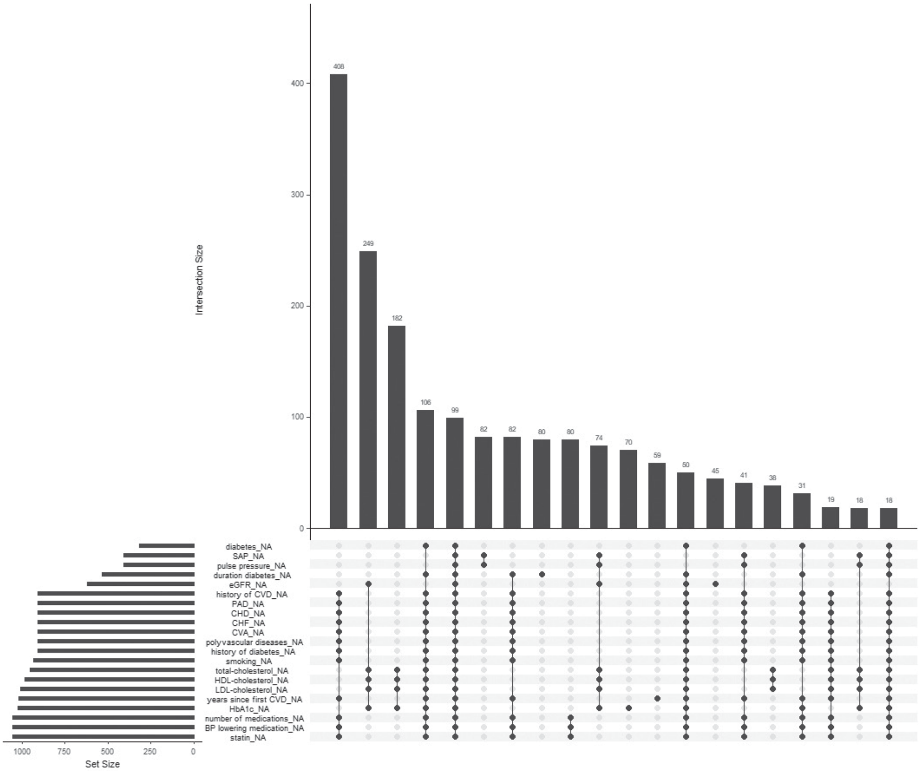
Given that the interest of the study is to provide a method with which a prediction model is able to be used whilst missing predictor values are present, we looked at combinations of missing predictor values that are observed in real data (see below). This figure describes the most common missing intersections of predictor variables. No distinction concerning variable importance is made. All single missing predictor scenarios are included, regardless of their occurrence in real data, as such the apparent single scenarios in the figure below can be ignored.

Each of these intersections is used in the study as a possible scenario for which the imputation methods should realistically work well. For a combination of missing predictor values to be included in the study it should at least be apparent in >1% of patients. This resulted in the inclusion of eight distinct multiple missing predictor scenarios.



The next part of variable selection was identifying the auxiliary variables that are inextricably linked to any of the predictor variables. Using these variables in an attempt to impute their respective predictor value via JMI or CMI would overestimate their performance as they are highly reliant on the relationship between available variables and the missing predictor value to be imputed. As such it is important that these auxiliary variables are not available for information extraction when their respective predictor values are missing. The variables were identified using the clinical experience of the authors as well as by using visualizations of the various combinations of missing value scenarios in the complete data (see next figure). For example, it was noticed that pulse pressure, or SAP, were never exclusively missing.

The combinations identified are: (1) SAP and pulse pressure, (2) diabetes and diabetes duration, (3) history of CVD and history of PAD, CHD, CHF, CVA and polyvascular disease and (4) total cholesterol, HDL-cholesterol and LDL-cholesterol.


Intersection Size
400
300
200
100
0
408
249
182
106
99
82
82
80
80
74
70
59
50
45
41
38
31
19
18
18
diabetes_NA
SAP_NA
pulse pressure_NA
duration diabetes_NA
eGFR_NA
history of CVD_NA
PAD_NA
CHD_NA
CHF_NA
CVA_NA
polyvascular diseases_NA
history of diabetes_NA
smoking_NA
total-cholesterol_NA
HDL-cholesterol_NA
LDL-cholesterol_NA
years since first CVD_NA
HbA1c_NA
number of medications_NA
BP lowering medication_NA
statin_NA
1000
750
500
250
0
Set Size

Appendix D – R code

Code available upon reasonable request.

Nijman SWJ[a], Hoogland J[a], Groenhof TKJ[a], Brandjes M[b], Jacobs JJL[b], Bots ML[a], Asselbergs FW[cde], Moons KGM[a], Debray TPA[ae]

On behalf of the UCC-CVRM and UCC-SMART study groups.

[a] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, 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, Faculty of Population Health Sciences, University College London, London, United Kingdom;
[e] Health Data Research UK, Institute of Health Informatics, University College London, London, United Kingdom

European Heart Journal - Digital Health, 2021; 2 154-164. DOI: 10.1093/ehjdh/ztaa016

CHAPTER 5

Real-time imputation of missing predictor values in clinical practice

Abstract

Introduction – Use of prediction models is widely recommended by clinical guidelines, but usually requires complete information on all predictors that is not always available in daily practice.

Methods – We describe two methods for real-time handling of missing predictor values when using prediction models in practice. We compare the widely used method of mean imputation (M-imp) to a method that personalizes the imputations by taking advantage of the observed patient characteristics. These characteristics may include both prediction model variables and other characteristics (auxiliary variables). The method was implemented using imputation from a joint multivariate normal model of the patient characteristics (joint modeling imputation; JMI). Data from two different cardiovascular cohorts with cardiovascular predictors and outcome were used to evaluate the real-time imputation methods. We quantified the prediction model's overall performance (mean squared error (MSE) of linear predictor), discrimination (c-index), calibration (intercept and slope) and net benefit (decision curve analysis).

Results – When compared with mean imputation, JMI substantially improved the MSE (0.10 vs. 0.13), c-index (0.70 vs 0.68) and calibration (calibration-in-the-large: 0.04 vs. 0.06; calibration slope: 1.01 vs. 0.92), especially when incorporating auxiliary variables. When the imputation method was based on an external cohort, calibration deteriorated, but discrimination remained similar.

Conclusions – We recommend JMI with auxiliary variables for real-time imputation of missing values, and to update imputation models when implementing them in new settings or (sub) populations.

Keywords: missing data; joint modeling imputation; real-time imputation; prediction; computerized decision support system; electronic health records

Introduction

The identification and treatment of patients at increased risk for disease is a cornerstone of personalized and stratified medicine [14,15,143]. Often, identification of high-risk patients involves the use of multivariable risk prediction models. These models combine patient and disease characteristics to provide estimates of absolute risk of a disease in an individual [20,26,144–146]. For example, prediction models for cardiovascular disease such as Framingham heart score (FHS) [27], HEART score [143], ADVANCE [147], Elderly [148] and SMART [126] are well known examples [19]. Additionally, cardiovascular guidelines recommend use of prediction models integrated in computerized decision support systems (CDSS), to support guideline adherent, risk-informed decision making [19,143].

When applying a risk prediction model in real-time, which constitutes its application in individual patients in routine clinical practice, one needs to have the individual's information (values) on all predictors in the model. Otherwise no absolute risk prediction by the model can be generated, restricting its use in situations when a physician is unable to acquire certain patient measurements. For example, for cardiovascular risk assessment, prediction models require complete information typically on age, sex, smoking, co-morbidities, blood pressure and lipid levels [127]. With the increased availability of large databases with information from electronic health care records, automated implementation and use of risk prediction models within CDSS using routine care (EHR) data has gained much interest [109,149–152]. However, the use of EHR databases faces many challenges, notably the incompleteness of data in the records [152–155]. The usability of a prediction model may thus still be limited in clinical practice if its implementation cannot standardly handle missing predictor values in real time. A detailed example is given in Box 1.

A variety of strategies have been developed for daily practice to handle missing predictor values in real-time[40,156]. Imputation strategies are of interest since they allow for direct use of well-known prediction models in their original form. In short, imputation substitutes a missing predictor value with one or more plausible values (imputations). In its simplest form, these imputations solely rely on the estimated averages of the missing variables in the targeted population. Therefore, they reflect what is known about the average patient. These simple methods can be applied directly in real-time clinical practice, provided that summary information (e.g. mean predictor values) about the targeted population is directly available. Additionally, imputations can account for the individual patient's *observed* predictor values by making use of the estimated associations between the patient characteristics in other patients. In that case, the imputations reflect all what is known about the specific individual at hand. Usually, the implementation of more complex imputation strategies requires direct access to the raw data from multiple individuals, which is typically problematic in clinical practice (e.g. due to operational or privacy constraints). As such, alternative strategies are required to make the imputation model applicable in real-time clinical practice.

Although real-time imputation of missing predictor values in clinical practice offers an elegant solution to generate predictions in the presence of incomplete data, the accuracy of these predictions may be severely limited if imputed values are a poor representation of the unobserved predictor values. In particular, problems may arise when (i) the imputation procedure does not adequately leverage information from the observed patient data, and (ii) if the estimated population characteristics used to generate the imputation(s) poorly represent the population to which the individual patient belongs. It is currently unclear how these novel real-time imputation methods influence the accuracy of available prediction models.

In this paper we explicitly focus on the relatively new area of real-time imputation, which has not been studied often before in similar literature. Most similar studies that address missing data consider and attempt to halt the onset of missing data in a particular dataset with missing values in study individuals, rather than a missing predictor in a single individual that is encountered in real-time clinical practice. Briefly, we investigate the performance of these two real-time imputation methods to handle missing predictor values when using a prediction model in daily practice. We evaluated both the accuracy of imputation and the impact of imputation on the prediction model's performance. Furthermore, transportability of the imputation procedures across different populations was empirically examined in two cardiovascular cohorts.

Box 1. An example of real-time imputation in an individual patient

Example. A patient visits their physician for a regular check-up. The patient and physician have access to a clinical decision support system that provides information on previously ordered test results (automatically retrieved from a registry). The physician would like to know the 10-year risk for the patient to suffer from a cardiovascular event, in order to determine whether any lifestyle changes or preventative therapies are needed. A calculator to determine this risk (e.g. the pooled cohort equations) is incorporated in the clinical decision support system, but requires complete information on several patient characteristics, including their BMI, cholesterol levels, and blood pressure. Many of these predictors are directly available (e.g. age, gender) at the visit. However, for some patients, important lab results (e.g. LDL cholesterol) are yet unknown or outdated (e.g. when retrieved from the registry). It is then not possible to determine the absolute risk of CVD for these patients. Our algorithm provides a substitute value for the missing LDL-cholesterol in real-time, enabling the calculation of a risk estimate 'on the spot'.

Methods

Short description

We conducted a simulation study to evaluate the impact of real-time imputation of missing predictor values on the absolute risk predictions in routine care. Hereto, we considered 2 large datasets and two real-time imputation methods. The datasets considered were the ongoing Utrecht Cardiovascular cohort - Cardiovascular risk management (UCC-CVRM) and the Utrecht Cardiovascular cohort - Secondary Manifestation of ARTerial disease (UCC-SMART) study [129,157]. Both studies focused on cardiovascular disease prevention and included newly referred patients visiting the University Medical Center (UMC) Utrecht for evaluation of cardiovascular disease [129,157]. Baseline examinations (i.e. predictors) for the UCC-CVRM included only the minimum set as suggested by the Dutch Cardiovascular Risk Management Guidelines [125].

Imputation methods

We considered mean imputation (M-Imp) and joint modelling imputation (JMI) [120,134]. Mean imputation was chosen as a comparison due to its straightforward implementation and extensive use during prediction model development and validation [9,158–160]. A major advantage of mean imputation is that it does not require information on individual patient characteristics and can be implemented without much difficulty in daily clinical practice. Using mean imputation, missing predictor values are simply imputed by their respective mean, usually from a representative sample (e.g., observational study). JMI was chosen because it allows to personalize imputations by adjusting for observed characteristics. To this purpose, JMI implements multivariate methods that have extensively been studied in the literature [6,120,134,161]. Some modifications are required to implement JMI for real-time imputation, these have been discussed previously [40]. In JMI, missing predictor values are imputed by taking the expected value from a multivariate distribution that is conditioned on the observed patient data. Implementations of JMI commonly assume that all variables are normally distributed, as this greatly simplifies the necessary calculations. This method then minimally requires mean and covariance estimates for all variables that are included as predictors in the prediction model from a representative sample (e.g., observational study). As an extension to JMI, we also consider that additional patient data (auxiliary variables) are available and can be used to inform the imputation of missing values (denoted as JMI^{aux}) [162].

All imputation methods can be directly applied to individuals and only require access to estimated population characteristics (i.e., mean and covariance estimates of the predictors) to account for missing predictor values. For both imputation methods the required population characteristics are easily stored and accessible in ‘live’ clinical practice within any accompanying

CDSS. The outcome is excluded from the imputation procedure as this information is not available when imputing the missing predictor values, and is the target of the prediction model. The corresponding source code is available from the supplementary information (Appendix E).

Table 1. general characteristics of the study populations

	UCC-SMART Mean (sd) or n/total (%)*	Role	UCC-CVRM Mean (sd) or n/total (%)**	Role
Age (years)	56.28 (12.45)	Predictor	61.7 (18.18)	Predictor
Gender (1=male)	8258 (65.50)	Predictor	1987 (51.21)	Predictor
Smoking (1=yes)	3560 (28.24)	Predictor	363 (9.36)	Predictor
SBP (mmHg)	144.67 (21.58)	Predictor	142.75 (24.24)	Predictor
TC (mmol/l)	5.11 (1.37)	Predictor	5.07 (1.24)	Predictor
HDL-c (mmol/l)	1.27 (0.38)	Predictor	1.36 (0.36)	Predictor
DM (1=yes)	2299 (18.23)	Predictor	755 (19.46)	Predictor
AD (1=yes)	8332 (66.09)	Predictor	705 (18.17)	Predictor
LDL-c (mmol/l)	3.15 (1.22)	auxiliary	3.08 (1.27)	auxiliary
HbA1c (mmol/mol)	3.69 (0.20)	auxiliary	3.66 (0.22)	auxiliary
MDRD (ml/min/1.73m2)	79.90 (19.54)	auxiliary	81.79 (24.56)	auxiliary
History of CVD (1=yes)	8134 (64.51)	auxiliary	1971 (50.80)	auxiliary
Time since 1st CVD event (years)	2.37 (5.93)	auxiliary	4.642 (8.06)	auxiliary
MPKR (mg/mmol)	4.10 (13.71)	auxiliary	NA	None
CRP (mg/L)	0.71 (1.13)	auxiliary	NA	None
AF (1=yes)	164 (1.30)	auxiliary	NA	None
LLD (1=yes)	6836 (54.22%)	auxiliary	NA	None
PAI (1=yes)	6805 (53.97%)	auxiliary	NA	None

Legend – SBP: systolic blood pressure, TC: total cholesterol, HDL-c: high-density lipoprotein cholesterol, , DM: diabetes mellitus, AD: antihypertensive drugs, LDL-c: low-density lipoprotein cholesterol, HbA1c: glycated hemoglobin, MDRD: modification of diet in renal diseases, MPKR: micro-protein/creatinine ratio, AF: atrial fibrillation, lipid-lowering drugs, PAI: platelet aggregation inhibitors. * after multiple imputation by chained equations.

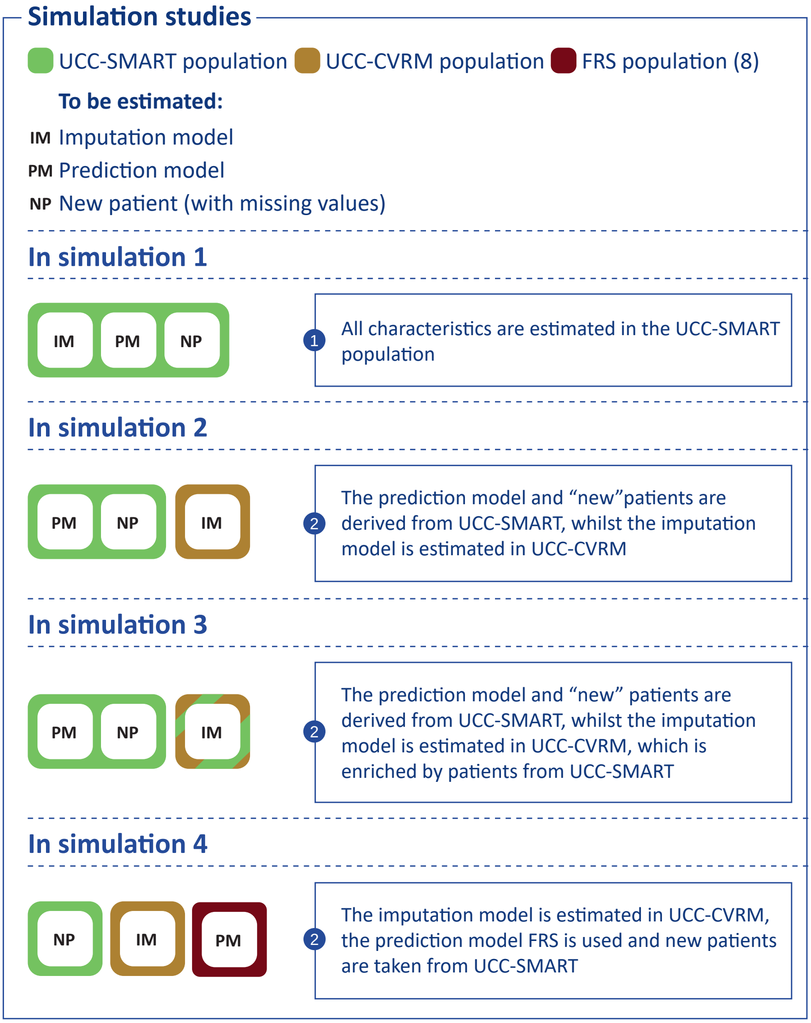
Study population

The UCC-CVRM sample consisted of 3.880 patients with 23 variables and the UCC-SMART study consisted of 12.616 patients with 155 variables. Some patient values were missing in UCC-CVRM (for 1057/3880 patients) and in UCC-SMART (for 2028/12616 patients). For the purpose of our methodological study, we had to have complete control over the patterns of missing predictor data and the true underlying predictor values, and needed to start with a fully observed data set that could be considered as the reference situation. To that end, for each dataset separately, we imputed all missing data once using Multiple Imputation by Chained Equations (for UCC-SMART) and nearest neighbor imputation (for UCC-CVRM) [6]. These then completed data sets formed the reference situation after which missing predictor values were generated according to various patterns (see below). Table 1 provides an overview of the completed variables in both cohorts, and how they were subsequently used in our simulation study. To assess the relatedness between UCC-CVRM and UCC-SMART, we calculated the membership c-statistic [163], which ranges between 0.5 (both samples have a similar case-mix) and 1 (the case-mix between both samples does not have any overlap). We found a membership c of 0.86, which indicates that the population characteristics of UCC-CVRM and UCC-SMART differ greatly.

Simulation study

We performed 4 simulation studies to investigate the impact of real-time predictor imputation on absolute risk predictions (Figure 1). In the first simulation, we considered the ideal situation where a (new) patient stems from the same population (i.e. UCC-SMART) as the one that is used to develop the prediction model, to derive the population characteristics, and to test the accuracy of individual risk predictions after the real-time imputations. In the second simulation, we considered a less ideal situation where imputations are based on the characteristics from a different, but related, population (i.e. UCC-CVRM). This simulation mimics the situation where development data are unavailable (or otherwise insufficient) to inform the imputation procedure, and thus assesses the transportability of the imputation model. In the third simulation, we investigated the situation where the estimated population characteristics underlying the imputations are derived from an external cohort (UCC-CVRM) and subsequently updated using local data (from UCC-SMART). This resembles a situation in which a small amount of local data is available, though insufficient to entirely inform the real-time imputation procedure. In the final simulation, we considered the most extreme scenario where 3 different populations are used to derive a prediction model (Framingham Risk Score [27]), the imputation model (UCC-CVRM), and to test the accuracy of the real-time imputations on the individuals' absolute risk predictions (UCC-SMART). This simulation mimics a more common predicament in which local data is insufficient to inform the imputation procedure and there is no access to the data from which the prediction model had been developed.

Figure 1. the simulation studies illustrated



All characteristics are estimated in the UCC-SMART population

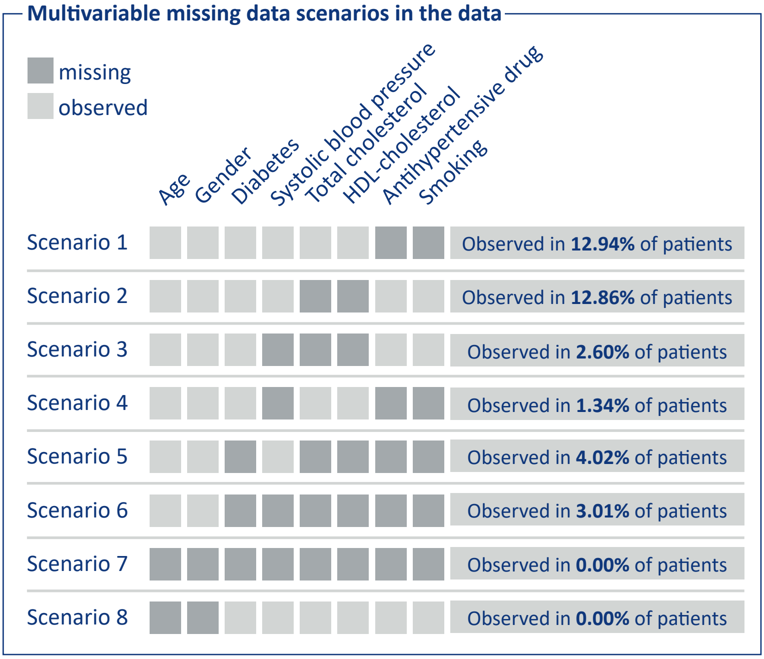
The prediction model and “new”patients are derived from UCC-SMART, whilst the imputation model is estimated in UCC-CVRM

The prediction model and “new” patients are derived from UCC-SMART, whilst the imputation model is estimated in UCC-CVRM, which is enriched by patients from UCC-SMART

The imputation model is estimated in UCC-CVRM, the prediction model FRS is used and new patients are taken from UCC-SMART

In all simulation studies, we considered UCC-SMART as the target population. For simulations 1-3, we adopted a leave-one-out-cross-validation (LOOCV) approach to develop the prediction model, to derive the population characteristics, and to evaluate the accuracy of risk predictions. This procedure ensures that independent data are used for the evaluation of risk predictions. In the LOOCV approach both the prediction model imputation model were derived from all but one patient (leave-one-out) of UCC-SMART. In the remaining hold-out patient, one or more predictor variables were then set to missing (see Figure 2 for an overview of which sets of predictor values were set to missing). The leave-one-out procedure was repeated until all patients had been removed from UCC-SMART exactly once (Figure 3).

Figure 2. Multivariate scenarios of missing predictor values observed in UCC-CVRM



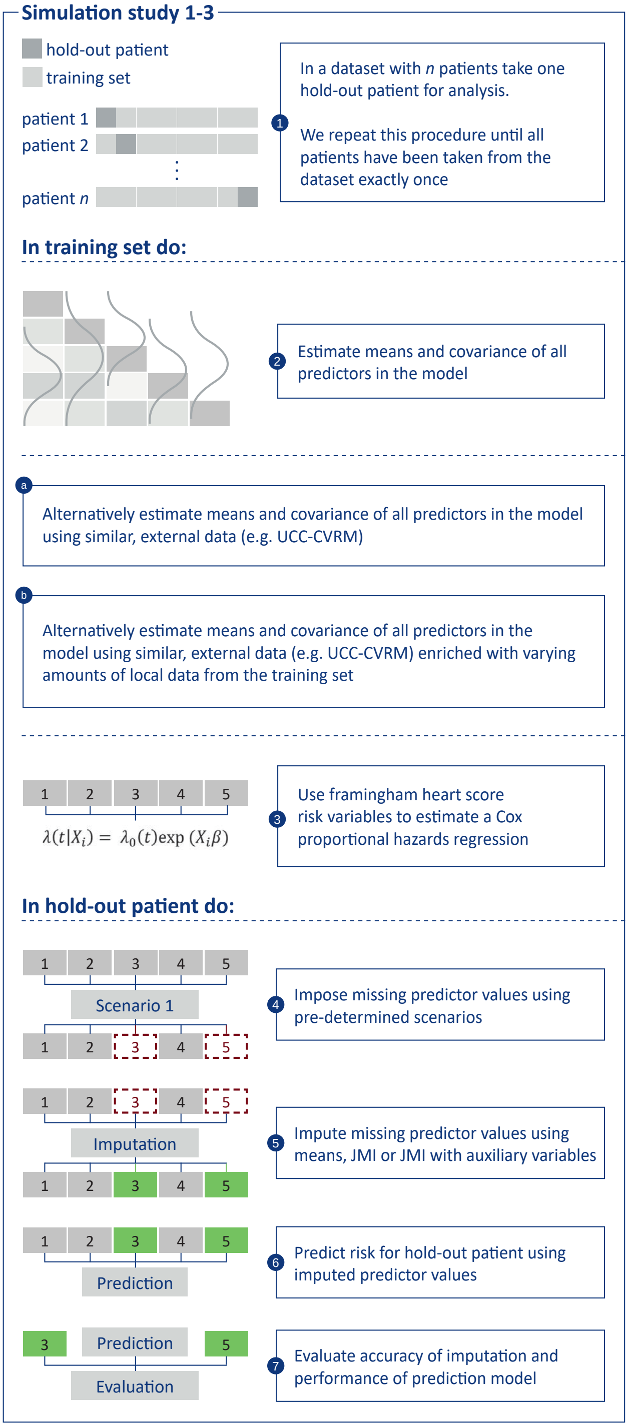
Multivariable missing data scenarios in the data

■ missing
□ observed

	Age	Gender	Diabetes	Systolic blood pressure	Total cholesterol	HDL-cholesterol	Antihypertensive drug	Smoking	
Scenario 1	□	□	□	□	□	□	■	■	Observed in **12.94%** of patients
Scenario 2	□	□	□	□	■	■	□	□	Observed in **12.86%** of patients
Scenario 3	□	□	□	■	■	■	□	□	Observed in **2.60%** of patients
Scenario 4	□	□	□	■	□	□	■	■	Observed in **1.34%** of patients
Scenario 5	□	□	■	□	■	■	■	■	Observed in **4.02%** of patients
Scenario 6	□	□	■	■	■	■	■	■	Observed in **3.01%** of patients
Scenario 7	■	■	■	■	■	■	■	■	Observed in **0.00%** of patients
Scenario 8	■	■	□	□	□	□	□	□	Observed in **0.00%** of patients

LOOCV was not needed for the 4th simulation as each task (prediction model development, derivation of population characteristics, and evaluation of risk predictions) involved a different dataset (Figure 4).

Figure 3. Simulation study 1-3 in detail



1 In a dataset with *n* patients take one hold-out patient for analysis.

We repeat this procedure until all patients have been taken from the dataset exactly once

2 Estimate means and covariance of all predictors in the model

a Alternatively estimate means and covariance of all predictors in the model using similar, external data (e.g. UCC-CVRM)

b Alternatively estimate means and covariance of all predictors in the model using similar, external data (e.g. UCC-CVRM) enriched with varying amounts of local data from the training set

3 Use framingham heart score risk variables to estimate a Cox proportional hazards regression

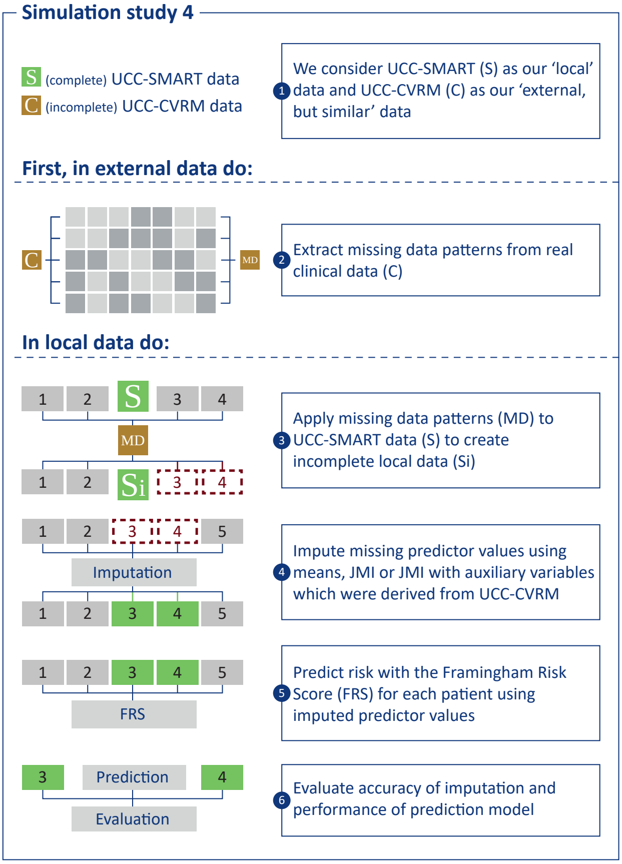
4 Impose missing predictor values using pre-determined scenarios

5 Impute missing predictor values using means, JMI or JMI with auxiliary variables

6 Predict risk for hold-out patient using imputed predictor values

7 Evaluate accuracy of imputation and performance of prediction model

Figure 4. Simulation study 4 in detail



1. We consider UCC-SMART (S) as our ‘local’ data and UCC-CVRM (C) as our ‘external, but similar’ data

2. Extract missing data patterns from real clinical data (C)

3. Apply missing data patterns (MD) to UCC-SMART data (S) to create incomplete local data (Si)

4. Impute missing predictor values using means, JMI or JMI with auxiliary variables which were derived from UCC-CVRM

5. Predict risk with the Framingham Risk Score (FRS) for each patient using imputed predictor values

6. Evaluate accuracy of imputation and performance of prediction model

Step 1. Estimation of the prediction model

For all simulation studies, the prediction model of interest was a Cox proportional hazards model predicting the onset of cardiovascular disease or coronary death. This model was derived in the LOO (leave-one-out) subset of UCC-SMART using predictors from the original FRS (simulation 1-3), or retrieved from the literature (simulation 4). A detailed description of how the prediction models were fit and the R code is listed in Appendix E. As a sensitivity analysis, we fitted a Cox regression model with only age and gender as predictors and included a scenario where, though unrealistic, age and gender were missing.

Step 2: Estimation of the population characteristics

We estimated the population characteristics necessary for the real-time missing data methods (i.e. the imputation models) in the following data (Figure 1):

- in the entire LOO subset of UCC-SMART (simulation 1),
- in the entire dataset of UCC-CVRM (simulation 2 and 4)
- in the entire dataset of UCC-CVRM, plus a random sample of the LOO subset of UCC-SMART, which were simply stacked. (simulation 3)

Step 3: Introduction and imputation of missing values

For simulation 1-3, we set one or more predictor variables to missing in each hold-out patient of UCC-SMART (scenarios illustrated in Figure 2). To match the introduction of missing values with real life occurrences of missingness, we included scenarios based on observed patterns of missingness in UCC-CVRM. For simulation 4, missing values were generated for the entire UCC-SMART dataset, rather than for individual patients. We subsequently impute the missing values once using the following strategies:

1. Mean imputation. Any missing predictor value was imputed with their respective mean as estimated in step 2.
2. JMI with observed predictors only. Each missing predictor value is replaced by its expected value conditional on the individual's observed *predictors*. The expected value is derived using the estimated population means and covariances from step 2.
3. JMI with observed predictors and auxiliary variables. Each missing predictor value is replaced by its expected values conditional on *all* the observed patient data. Note that this includes additional patient data that are not included as predictors in the prediction model (Table 1).

Step 4 – Risk prediction and validation of model performance

The imputed missing predictor values were then used together with the observed predictor values to calculate the linear predictor η_i (where $\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots$) and the 10-year predicted absolute risk. The predictions from all UCC-SMART patients were then used to assess the following performance measures: 1) Mean Squared Error (MSE) of the prediction model's linear predictor, (2) concordance (C-)statistic, 3) calibration-in-the-large, 4) the calibration slope and 5) the decision curve [144,149,164,165].

4. *The MSE of the linear predictor of the prediction model* can be described as the average squared difference between the linear predictor after imputation and the true, original linear predictor (i.e. before introducing missing values) [13]. The linear predictor can be described as the weighted sum of the predictors of a given patient, where the weights consist of the model coefficients [164]. Lower values for the MSE are preferred.
5. *The C-statistic* can be described as the ability of the model to discern those who have experienced an event and those who haven't [13,145,166]. It is represented by the probability of correctly discerning who, between two random subjects, has the higher predicted probability of survival. The C-statistic is ideally close to 1.
6. *Calibration-in-the-large (CITL)* can be described as the overall calibration of the model (i.e. agreement between average predicted risk and average observed risk) [144,145,166,167]. It is interpreted as an indication of the extent to which the predictions systematically over- or underestimate the risk; the ideal value is 0.
7. *The calibration slope* can be described as a quantification of the extent that predicted risks vary too much (slope <1) or too little (slope > 1), and is often used as an indication of overfitting or lack of transportability [13,145,149,166,167]. The ideal value is 1.
8. *The decision curve* can be described as a way of identifying the potential impact of leveraging individual risk predictions for decision making [144,165,168]. It considers a range of thresholds (e.g. 10%) to classify patients into high risk (indication of treatment) or low risk (no treatment required) and calculates the net benefit (NB) for each cut-off value. A decision curve is then constructed for 3 different treatment strategies: treat all, treat none, or treat according to risk predictions. Ideally, the decision curve of the latter strategy depicts consistently better NB over the complete range of thresholds.

Results

Prediction model performance in the absence of missing values

Based on internal validation by means of LOOCV, the optimism corrected c-statistic for our newly derived prediction model in UCC-SMART was 0.705. As expected, the CITL and calibration slope were near 0 (-0.0005) and 1 (0.9999) respectively. Therefore, there were no signs of miscalibrations and/or over/underfitting of the developed CVD risk prediction model. The prediction model that was based on age and gender yielded an optimism corrected c-statistic of 0.679, with a slope of 0.9999 and an intercept of -0.00005. Finally, the refitted FRS model (as derived from the literature) yielded a c-statistic of 0.6280 and a slope of 0.8205 in UCC-SMART.

Prediction model performance in presence of missing data

Mean squared error

The MSE of the linear predictor was consistently lower when adopting JMI, as compared to M-Imp. The implementation of JMI was particularly advantageous when adjusting for auxiliary variables that were not part of the prediction model (see table 2 for the results of scenario 1 and 5). For instance, when total cholesterol (TC), HDL-cholesterol (HDL-c), use of Antihypertensive Drugs (AD), smoking and Diabetes Mellitus (DM) were missing (i.e. scenario 5), M-Imp yielded an MSE of 0.130, whereas the MSE for JMI was 0.126 or even 0.101 when utilizing auxiliary variables. As expected, differences in MSE were lower, when imputing other predictors that did not have a strong contribution in the prediction model, or much more pronounced when imputing important predictors (see table 3 for the results of the sensitivity analysis with age and gender missing). This expected discrepancy results from the fact that the linear predictor is a weighted average of the predictors and the important variables simply have larger weights. When imputation was based on the characteristics of a different, but related, cohort to UCC-SMART, all imputation strategies yielded a substantially larger MSE. For instance, when TC, HDL-c, AD, smoking and DM were missing (i.e. scenario 5), the MSE increased from 0.130 to 0.193 for M-Imp, and from 0.1014 to 0.159 for JMI^{aux}. Again, JMI^{aux} was superior to M-Imp and JMI based on predictor variables only. As expected, the MSE for all imputation methods improved when the imputation model was based on a mixture of patients from both the UCC-CVRM (different but related) and the UCC-SMART (the target cohort for predictions). However, the lowest MSE's were obtained when imputations were based on UCC-SMART data only.

C-statistic

The c-statistic was higher for both implementations of JMI, when compared to M-Imp (Table 2). Using JMI^{aux} further increased the c-statistic substantially, especially when important predictors (i.e. age and gender) were missing (Table 3). In this scenario, M-Imp yielded a c-statistic of 0.61, whereas JMI yielded a c-statistic of 0.62 or even 0.67 if auxiliary variables were used. Discrimination performance did not much deteriorate when imputation was based on the characteristics from a different but related population. Again, JMI^{aux} was superior to M-Imp and JMI based on predictor variables only. The c-statistic, for all imputation methods, improved when the population characteristics from UCC-CVRM were augmented with data from UCC-SMART. However, when an external prediction model was used in combination with external population characteristics (simulation 4), the utilization of auxiliary variables did not seem to improve on the discriminatory ability of risk predictions (Table 4). The highest c-statistics were obtained when imputations were based on UCC-SMART data only and a locally derived prediction model was used.

Calibration-in-the-large

The CITL was consistently closer to the ideal value (i.e. 0) for all scenarios when using both implementations of JMI, when compared to M-Imp. Using JMI^{aux} improved the CITLs further towards their ideal value (Table 2). When imputation used estimated population characteristics from UCC-CVRM, all imputation strategies had a substantially worse CITL. The performance drop was most notable as more predictors in the model were missing. Again, JMI^{aux} was superior to M-Imp and JMI based on predictor variables only. The CITL, for all imputation methods, improved when the population characteristics from UCC-CVRM were augmented with data from UCC-SMART. When an external prediction model was used, M-Imp yielded the “best” CITL (-0.167 as opposed to -0.2030 for JMI and -0.2256 for JMI^{aux}; Table 4). The CITLs were closest to 0 when imputations were based on UCC-SMART data only.

Calibration slope

The use of JMI^{aux} improved the calibration slope as compared to M-Imp or JMI using predictor variables only (Table 2). When imputation used population characteristics from UCC-CVRM, the variability of predicted risks generally became too large (slope < 1 for all imputation methods). The performance drop was most notable as more predictors were missing. When an external prediction model was used, both JMI and JMI^{aux} yielded better calibration as compared to M-Imp (Table 4), although JMI^{aux} performed worse than JMI. The best calibration slopes were found for imputations based on UCC-SMART data only.

Figure 5 visualizes calibration plots for scenarios 1, 5 and 8. It shows that when important predictors (i.e. age and gender in scenario 8) are missing there is a notable impact on the calibration of 10-year risk predictions, especially when using external data for generating imputations. When less important predictors are missing (scenario 1 and 5) the differences between the imputation methods are much less pronounced in the calibration plots.

Decision curve

When important variables were missing, imputation through JMI with auxiliary variables yielded an improved net benefit over the whole range of thresholds when compared to M-Imp and JMI (Figure 6), and was substantially better than treat-all or treat-none strategies. The observed net benefit did not much deteriorate when imputation was based on a different, but related, dataset.

A complete detailed overview of all results (e.g. all scenarios) can be found in the supplementary material.

Figure 5. Calibration plots for scenario 1, 5 and 8

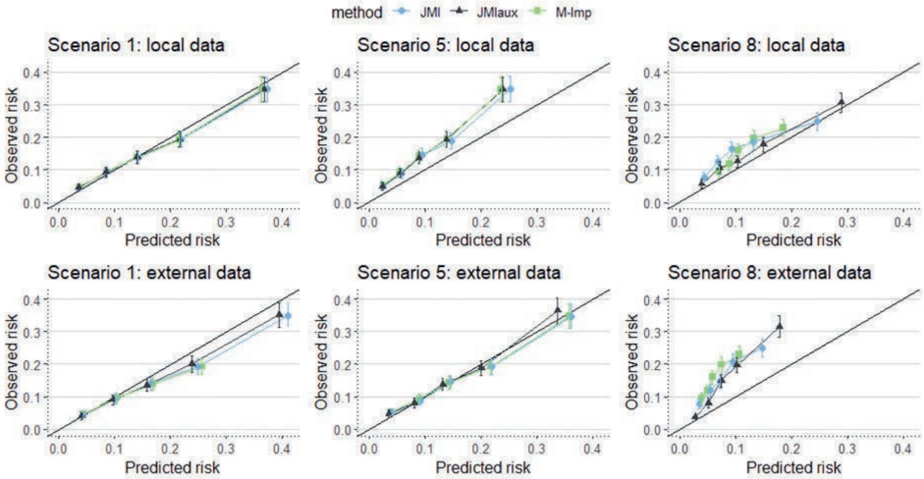


Table 2. Results of simulating scenarios with small and large amounts of missing data

Scenario 1 (small amount of missing data):	**Imputation**	**MSE of the LP**			
(1) SBP, (2) smoking	**methods**	**(% difference to M-Imp)**	**C-index**	**CITL**	**Calibration slope**
Apparent performance (reference)			0.7051	-0.0001	0.9999
Simulation 1 Local data (for informing imputation)	M-Imp	0.0702	0.6908	0.0228	0.9415
	JMI	0.0685 (-2.35%)	0.6913	0.0242	0.9552
	JMI^{aux}	0.0649 (-7.50%)	0.6975	0.0221	0.9928
Simulation 2 External data (for informing imputation)	M-Imp	0.0802	0.6908	0.1227	0.9415
	JMI	0.0782 (-2.56%)	0.6911	0.1018	0.9269
	JMI^{aux}	0.0801 (0.001%)	0.6902	0.1123	0.9251
Simulation 3 External data with 1.500 local patients	M-Imp	0.0746	0.6909	0.0845	0.9393
	JMI	0.0718 (-3.90%)	0.6913	0.0510	0.9278
	JMI^{aux}	0.0708 (-5.37%)	0.6911	0.0485	0.9315

Scenario 5 (large amount of missing data):	**Imputation**	**MSE of the LP**			
(1) TC, (2) HDL-c, (3) AD (4) smoking, (5) DM missing	**methods**	**(% difference to M-Imp)**	**C-index**	**CITL**	**Calibration slope**
Apparent performance (reference)			0.7051	-0.0001	0.9999
Simulation 1 Local data (for informing imputation)	M-Imp	0.1300	0.6797	0.0581	0.9199
	JMI	0.1262 (-2.98%)	0.6803	0.0549	0.9211
	JMI^{aux}	0.1014 (-21.98%)	0.6960	0.0369	1.0052
Simulation 2 External data (for informing imputation)	M-Imp	0.1930	0.6797	0.3090	0.9199
	JMI	0.1806 (-6.42%)	0.6803	0.2797	0.9067
	JMI^{aux}	0.1591 (-17.57%)	0.6844	0.2595	0.9475
Simulation 3 External data with 1.500 local patients	M-Imp	0.1683	0.6790	0.2418	0.9387
	JMI	0.1603 (-4.78%)	0.6792	0.2078	0.9095
	JMI^{aux}	0.1334 (-20.72%)	0.6851	0.1677	0.9573

Legend – SBP: systolic blood pressure, TC: total cholesterol, HDL-c: HDL-cholesterol, AD: antihypertensive drug, SBP: systolic blood pressure, DM: diabetes mellitus, MSE: mean squared error, LP: linear predictor, CITL: calibration in the large, M-Imp: mean imputation, JMI: joint modelling imputation, JMIaux: joint modelling imputation with auxiliary variables.

Table 3. Results sensitivity analysis

Scenario 8: (1) Age, (2) gender missing	**Imputation methods**	**MSE of the LP (% difference to M-Imp)**	**C-index**	**CITL**	**Calibration slope**
Apparent performance (reference)			0.7051	-0.0001	0.9999
Simulation 1 Local data (for informing imputation)	M-Imp	0.7438	0.6063	0.1958	0.8225
	JMI	0.6373 (-14.32%)	0.6223	0.1616	0.8052
	JMI^{aux}	0.4517 (-39.26%)	0.6931	0.0794	1.0828
Simulation 2 External data (for informing imputation)	M-Imp	0.8334	0.6064	-0.1037	0.8230
	JMI	0.7963 (-4.45%)	0.6116	-0.2221	0.5769
	JMI^{aux}	0.7018 (-15.79%)	0.6721	-0.3649	0.8453
Simulation 3 External data with 1.500 local patients	M-Imp	0.792383	0.6107	-0.0205	0.8429
	JMI	0.7252996 (-9.25%)	0.6131	-0.0659	0.6480
	JMI^{aux}	0.5739753 (-38.05%	0.6856	-0.1451	0.9654

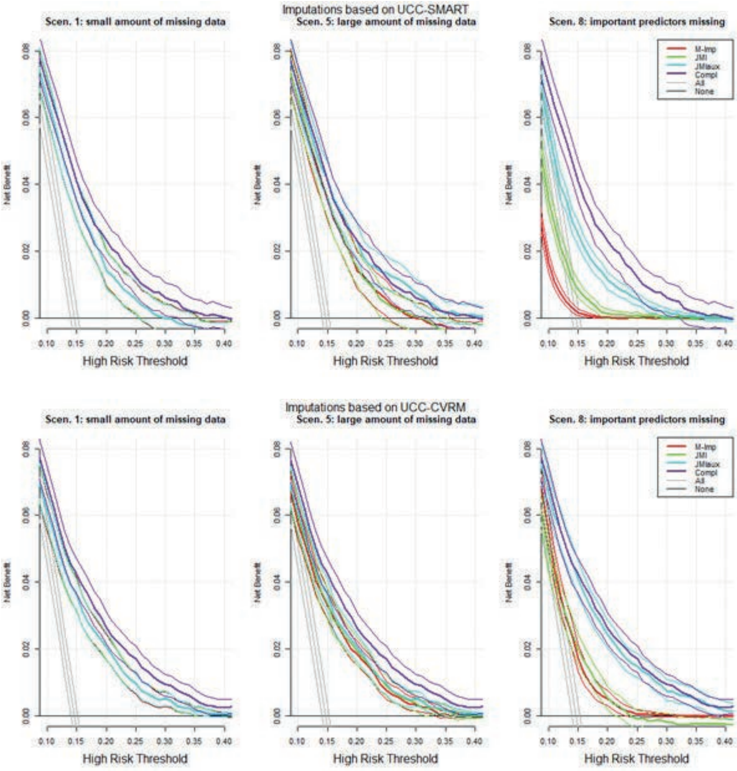
Legend – MSE: mean squared error, LP: linear predictor, CITL: calibration in the large, M-Imp: mean imputation, JMI: joint modelling imputation, JMI+: joint modelling imputation with auxiliary variables.

Table 4. Multivariable missing data imputation (simulation 4): the use of an external prediction and imputation model

Combination of all missing data scenarios	**Imputation methods**	**MSE of the LP (% difference to M-Imp)**	**C-index**	**CITL**	**Calibration slope**
Reference when no variables are missing			0.6280	-0.0888	0.8205
Simulation 4 External prediction and imputation model	M-Imp	0.1689	0.6095	-0.1674	0.7424
	JMI	0.1585 (-6.56%)	0.6145	-0.2030	0.7549
	JMI^{aux}	0.1334 (-26.61%)	0.6135	-0.2257	0.7495

Legend – CITL: calibration in the large, M-Imp: mean imputation, JMI: joint modelling imputation, JMI^{aux}: joint modelling imputation with auxiliary variable

Figure 6: Decision curve analysis simulation 1



Discussion

Our aim was to evaluate the impact of using real-time imputation of missing predictor values on the performance of cardiovascular risk prediction models in individual patients. We considered mean imputation and joint modeling imputation to provide automated real-time imputations. Our results demonstrate that in all scenarios and for all parameters studied (c-index, calibration and decision curve analysis) JMI leads to more accurate risk predictions than M-Imp, especially when used to impute a higher number of missing predictors (e.g. scenario 5 for prediction of cardiovascular events). The performance of JMI greatly improved when imputations were based on all observed patient data, and not restricted to only the predictors that were in the prediction model. Finally, we found that real-time missing predictor imputations were most accurate when the imputation method relied on characteristics that were directly estimated a sample from the target population (i.e. the population for which predictions are required), rather than from an external though related dataset. In the latter case, while discriminative performance was stable, calibration clearly deteriorated (in terms of both CITL and calibration slope). This implies that the need for local updating, as is well known in clinical prediction modeling, may extend to imputation models. In practice, a prediction model is ideally developed together with an appropriate missing data method for real-time imputation. When high quality local data are available, performance gains can be expected for that setting by local updating of both the prediction model and the imputation model.

Our findings suggest that JMI should be preferred over M-Imp for real-time imputation of missing predictor values in routine care, ideally making use of additional patient data (variables) that are not part of the prediction model. The underlying rationale, is that some variables that are highly correlated are unlikely to both end up in a prediction model (due to little added value), but are quite valuable for imputation purposes when one or the other is missing. The implementation of JMI is very straightforward, and only requires estimating the mean and covariance of all relevant patient variables in a representative sample. Imputations are then generated using a set of mathematical equations that are well established in the statistical literature [40]. As JMI does not rely on disease-specific patient characteristic and lends itself excellently for local tailoring [169], it is considered highly scalable to a multitude of clinical settings and populations. Routine reporting of population characteristics (i.e. means and covariance) would greatly facilitate the implementation of risk prediction models in the presence of missing predictor data in daily practice, and has previously been recommended to improve the interpretation of validation study results [163].

A limitation we observed in the data was that most of the explained variability in risk of cardiovascular disease, as defined in our study, could be inferred based on age and gender.

Although additional predictors (e.g. blood pressure, cholesterol levels) somewhat improved the model's discrimination and calibration performance, their individual added value appears small. A further limitation of the data was the lack of strong correlations between predictors other than age and gender (appendix D). Consequently, the information available for JMI to leverage observed patient characteristics was limited. These findings are in line with earlier research, suggesting that M-Imp performs similarly to more advanced imputation methods when considering commonly encountered missing data patterns in cardiovascular routine care [170]. However, our study reveals that JMI had the advantage even under these typical but difficult settings. Gains are expected to be larger when the interrelation of predictors is stronger and especially when key auxiliary variables can be identified. Moreover, for many disease areas, risk prediction relies more strongly on a multitude patient characteristic that are more likely to be missing (e.g. certain imaging characteristics, biomarkers or genetic profiles), and JMI offers a larger advantage.

Various other aspects need to be addressed to fully appreciate these results. First, we restricted our comparison to M-Imp and JMI. Considering M-Imp was picked as a comparator, we choose JMI as it was well established in the statistical literature and permitted relatively straightforward adjustments to be applied in clinical practice via the EHR [6,120]. Other, more flexible, imputation strategies exist, and have been discussed at length [40]. These strategies generally require more complex descriptions of the population characteristics and adopt more advanced procedures to generate imputations. For this reason, their implementation appears less straightforward in routine care. A more detailed overview of the impact of using other strategies for handling real-time missing predictor value imputation is warranted. Also, the use of multiple imputation may be preferable with respect to prediction accuracy in case of models with a non-linear link function such as the Cox or logistic model, the reason is multiple imputation can correctly convey the influence of imputation uncertainty on the expected prediction. The available R code already provides in this, though in this study we explicitly choose to use single imputation. We choose single imputation due to its convenience in real-time clinical practice. The imputation process is quick, in contrast to the usually computationally expensive multiple imputation, and it presents an individual's imputed predictor value which may be informative to the clinician. Additionally, rather than imputing a random draw, we impute the most likely value in order to be able to easily reproduce model predictions from the imputed data. Ideally, the predictions would be based on multiple imputation from the conditional distribution of the missing predictors rather than representing their conditional means. Further extensions, for example multilevel multiple imputation, may also be recommended in specific situations where the prediction model and accompanying imputation models are derived from large datasets with clustering [171]. Lastly, whilst there are many clinical settings and populations the study only considered cardiovascular risk prediction. The performance of JMI, when compared to M-Imp, might have been further emphasized had other clinical settings been considered.

In summary, this study evaluates the use of two imputation methods for handling missing predictor values when applying risk prediction models in daily practice. We recommend JMI over mean imputation, preferably based on estimated from local data and with the use of available auxiliary variables. The added value of JMI is most evident when missing predictors are associated with either observed predictor values or auxiliary variables.

Disclosures

Funding sources

This work was supported by the Netherlands Heart foundation (public-private study grant, grant number: #2018B006); and the Top Sector Life Sciences & health (PPP allowance made available to Netherlands Heart Foundation to stimulate public-private partnerships).

Conflicts of interest

None declared.

Author contributions

TD, KM, KG, and SN conceived of the presented idea, in correspondence with earlier work by FA, FV, MB and KG. SN, TD and JH derived the models and analyzed the data. TD and JH verified the analytical methods. TD supervised the findings of this work. SN, KG, TD and JH contributed to the interpretation of the results. SN wrote the manuscript. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

Data availability statement

The data that support the findings of this study are available from the UCC upon reasonable request (https://www.umcutrecht.nl/en/Research/Strategic-themes/Circulatory-Health/Facilities/UCC).

Acknowledgements

We gratefully acknowledge the contribution of the research nurses; R. van Petersen (data-manager); B. van Dinther (study manager) and the members of the Utrecht Cardiovascular Cohort-Second Manifestations of ARTerial disease-Studygroup (UCC-SMART-Studygroup): F.W. Asselbergs and H.M. Nathoe, Department of Cardiology; G.J. de Borst, Department of Vascular Surgery; M.L. Bots and M.I. Geerlings, Julius Center for health Sciences and Primary Care; M.H. Emmelot, Department of Geriatrics; P.A. de Jong and T. Leiner, Department of Radiology; A.T. Lely, Department of Obstetrics & Gynecology; N.P. van der Kaaij, Department of Cardiothoracic Surgery; L.J. Kappelle and Y.M. Ruigrok, Department of Neurology; M.C. Verhaar, Department of Nephrology, F.L.J. Visseren (chair) and J. Westerink, Department of Vascular Medicine, University Medical Center Utrecht and Utrecht University.

Appendix A. Simulation 1: imputing multiple missing predictor values scenarios using local data

#	Variables missing	Methods	MSE of the LP (% difference to M-Imp)	C-index	CITL	Calibration slope
	Apparent performance (reference)			0.7051	-0.0001	0.9999
1	(1) SBP, (2) smoking	M-Imp	0.0702	0.6908	0.0228	0.9415
		JMI	0.0685 (-2.35%)	0.6913	0.0242	0.9552
		JMI^{aux}	0.0649 (-7.50%)	0.6975	0.0221	0.9928
2	(1) TC, (2) HDL-c	M-Imp	0.0265	0.7005	0.0230	0.9766
		JMI	0.0241(-8.97%)	0.7013	0.0192	0.9751
		JMI^{aux}	0.0220 (-16.84%)	0.7046	0.0153	0.9901
3	(1) TC, (2) HDL-c, (3) SBP	M-Imp	0.0333	0.6994	0.0406	1.0003
		JMI	0.0314 (-5.74%)	0.7000	0.0268	0.9787
		JMI^{aux}	0.0276 (-17.26%)	0.7040	0.0212	0.9935
4	(1) TC, (2) HDL-c, (3) SBP, (4) AD	M-Imp	0.0459	0.6981	0.0546	1.0217
		JMI	0.0425 (-7.41%)	0.6983	0.0255	0.9702
		JMI^{aux}	0.0394 (-14.13%)	0.7044	0.0231	1.0010
5	(1) TC, (2) HDL-c, (3) AD (4) smoking, (5) DM	M-Imp	0.1300	0.6797	0.0581	0.9199
		JMI	0.1262 (-2.98%)	0.6803	0.0549	0.9211
		JMI^{aux}	0.1014 (-21.98%)	0.6960	0.0369	1.0052
6	(1) TC, (2) HDL-c, (3) AD, (4) smoking, (5) DM, (6) SBP	M-Imp	0.1383	0.6785	0.0758	0.9430
		JMI	0.1351 (-2.31%)	0.6788	0.0637	0.9249
		JMI^{aux}	0.1087 (-21.45%)	0.6955	0.0441	1.0128
7	(1) Age, (2) gender, (3) TC, (4) HDL-c, (5) AD, (6) smoking, (7) DM, (8) SBP	M-Imp	0.9137	0.5112	0.2897	77.819
		JMI	0.9137 (0.00%)	0.5112	0.2897	77.819
		JMI^{aux}	0.5990 (-34.44%)	0.6892	0.1544	1.3754
8	(1) Age, (2) gender	M-Imp	0.7438	0.6063	0.1958	0.8225
		JMI	0.6373 (-14.32%)	0.6223	0.1616	0.8052
		JMI^{aux}	0.4517 (-39.26%)	0.6931	0.0794	1.0828

Legend – MSE: mean squared error, LP: linear predictor, CITL: calibration in the large, M-Imp: mean imputation, JMI: joint modelling imputation, JMI^{aux}: joint modelling imputation with auxiliary variables, SBP: systolic blood pressure, TC: total cholesterol, HDL-c: HDL-cholesterol, AD: antihypertensive drug, DM: diabetes mellitus.

Appendix B. Simulation 2: imputing multiple missing predictor values scenarios using external data

#	Variables missing	Methods	MSE of the LP (% difference to M-Imp)	C-index	CITL	Calibration slope
	Apparent performance (reference)			0.7051	-0.0001	0.9999
1	(1) SBP, (2) smoking	M-Imp	0.0802	0.6908	0.1227	0.9415
		JMI	0.0782 (-2.56%)	0.6911	0.1018	0.9269
		JMI^{aux}	0.0801 (0.001%)	0.6902	0.1123	0.9251
2	(1) TC, (2) HDL-c	M-Imp	0.0280	0.7005	0.0618	0.9766
		JMI	0.0251 (-10.13%)	0.7010	0.0244	0.9639
		JMI^{aux}	0.0248 (-11.35%)	0.7017	0.0155	0.9724
3	(1) TC, (2) HDL-c, (3) SBP	M-Imp	0.0343	0.6994	0.0715	1.0003
		JMI	0.0325 (-5.14%)	0.6995	0.0308	0.9629
		JMI^{aux}	0.0324 (-5.29%)	0.7003	0.0192	0.9703
4	(1) TC, (2) HDL-c, (3) SBP, (4) AD	M-Imp	0.0654	0.6982	0.1943	1.0217
		JMI	0.0615 (-5.95%)	0.6978	0.1661	0.9756
		JMI^{aux}	0.0583 (-10.91%)	0.6998	0.1530	0.9881
5	(1) TC, (2) HDL-c, (3) AD (4) smoking, (5) DM	M-Imp	0.1930	0.6797	0.3090	0.9199
		JMI	0.1806 (-6.42%)	0.6803	0.2797	0.9067
		JMI^{aux}	0.1591 (-17.57%)	0.6844	0.2595	0.9475
6	(1) TC, (2) HDL-c, (3) AD, (4) smoking, (5) DM, (6) SBP	M-Imp	0.1974	0.6785	0.3189	0.9431
		JMI	0.1914 (-3.01%)	0.6788	0.2889	0.9045
		JMI^{aux}	0.1664 (-15.70%)	0.6831	0.2663	0.9483
7	(1) Age, (2) gender, (3) TC, (4) HDL-c, (5) AD, (6) smoking, (7) DM, (8) SBP	M-Imp	0.9167	0.5157	0.2332	86.984
		JMI	0.9167 (0.00%)	0.5157	0.2332	86.984
		JMI^{aux}	0.7269 (-20.70%)	0.6589	-0.0783	0.9687
8	(1) Age, (2) gender	M-Imp	0.8334	0.6064	-0.1037	0.8230
		JMI	0.7963 (-4.45%)	0.6116	-0.2221	0.5769
		JMI^{aux}	0.7018 (-15.79%)	0.6721	-0.3649	0.8453

Legend – MSE: mean squared error, LP: linear predictor, CITL: calibration in the large, M-Imp: mean imputation, JMI: joint modelling imputation, JMI^{aux}: joint modelling imputation with auxiliary variables, SBP: systolic blood pressure, TC: total cholesterol, HDL-c: HDL-cholesterol, AD: antihypertensive drug, DM: diabetes mellitus.

Appendix C imputing multiple missing predictor values scenarios using enriched external data; simulation 3: scenario 1

#	Variables missing	Methods	MSE of the LP (% difference to M-Imp)	C-index	CITL	Calibration slope
	Apparent performance			0.7051	-0.0001	0.9999
1	MD scenario 1* (i.e. without local data, reference)	M-Imp	0.0802	0.6908	0.1227	0.9415
		JMI	0.0782 (-2.56%)	0.6911	0.1018	0.9269
		JMI+	0.0801 (0.001%)	0.6902	0.1123	0.9251
2	+100 local patients	M-Imp	0.0794	0.6908	0.1169	0.9402
		JMI	0.0769 (-3.25%)	0.6912	0.0915	0.9252
		JMI+	0.0780 (-1.79%)	0.6904	0.0987	0.9246
3	+300 local patients	M-Imp	0.0792	0.6902	0.1190	0.9393
		JMI	0.0763 (-3.80%)	0.6904	0.0910	0.9242
		JMI+	0.0771 (-2.72%)	0.6897	0.0966	0.9229
4	+750 local patients	M-Imp	0.0765	0.6902	0.1091	0.9396
		JMI	0.0733 (-4.37%)	0.6904	0.0710	0.9233
		JMI+	0.0725 (-5.52%)	0.6902	0.0693	0.9268
5	+1500 local patients	M-Imp	0.0746	0.6909	0.0845	0.9393
		JMI	0.0718 (-3.90%)	0.6913	0.0510	0.9278
		JMI+	0.0708 (-5.37%)	0.6911	0.0485	0.9315
6	+5000 local patients	M-Imp	0.0713	0.6869	0.0779	0.9420
		JMI	0.0699 (-2.00%)	0.6874	0.0545	0.9389
		JMI+	0.0683 (-4.39%)	0.6875	0.0482	0.9464
7	+10000 local patients	M-Imp	0.0704	0.6732	0.0863	0.8778
		JMI	0.0686 (-2.65%)	0.6733	0.0743	0.8797
		JMI+	0.0671 (-4.92%)	0.6739	0.0681	0.8887

Legend – MSE: mean squared error, LP: linear predictor, CITL: calibration in the large, M-Imp: mean imputation, JMI: joint modelling imputation, JMIaux: joint modelling imputation with auxiliary variables, *(1) systolic blood pressure, (2) smoking.

Appendix C (cont.) simulation 3: scenario 5

#	Variables missing	Methods	MSE of the LP (% difference to M-Imp)	C-index	CITL	Calibration slope
	Apparent performance			0.7051	-0.0001	0.9999
1	MD scenario 5* (i.e. without local data, reference)	M-Imp	0.1974	0.6785	0.3189	0.9431
		JMI	0.1914 (-3.01%)	0.6788	0.2889	0.9045
		JMI+	0.1664 (-15.70%)	0.6831	0.2663	0.9483
2	+100 local patients	M-Imp	0.1929	0.6785	0.3081	0.9418
		JMI	0.1865 (-3.33%)	0.6788	0.2775	0.9046
		JMI+	0.1605 (-16.81%)	0.6836	0.2509	0.9502
3	+300 local patients	M-Imp	0.1899	0.6785	0.3054	0.9464
		JMI	0.1827 (-3.78%)	0.6788	0.2749	0.9114
		JMI+	0.1557 (-18.01%)	0.6834	0.2445	0.9557
4	+750 local patients	M-Imp	0.1794	0.6787	0.2851	0.9506
		JMI	0.1714 (-4.42%)	0.6790	0.2521	0.9167
		JMI+	0.1425 (-20.56%)	0.6844	0.2105	0.9607
5	+1500 local patients	M-Imp	0.1683	0.6790	0.2418	0.9387
		JMI	0.1603 (-4.78%)	0.6792	0.2078	0.9095
		JMI+	0.1334 (-20.72%)	0.6851	0.1677	0.9573
6	+5000 local patients	M-Imp	0.1472	0.6736	0.1960	0.9377
		JMI	0.1424 (-3.31%)	0.6737	0.1696	0.9166
		JMI+	0.1206 (-18.11%)	0.6789	0.1292	0.9585
7	+10000 local patients	M-Imp	0.1454	0.6630	0.1832	0.8820
		JMI	0.1410 (-3.06%)	0.6629	0.1603	0.8668
		JMI+	0.1199 (-17.54%)	0.6703	0.1291	0.9201

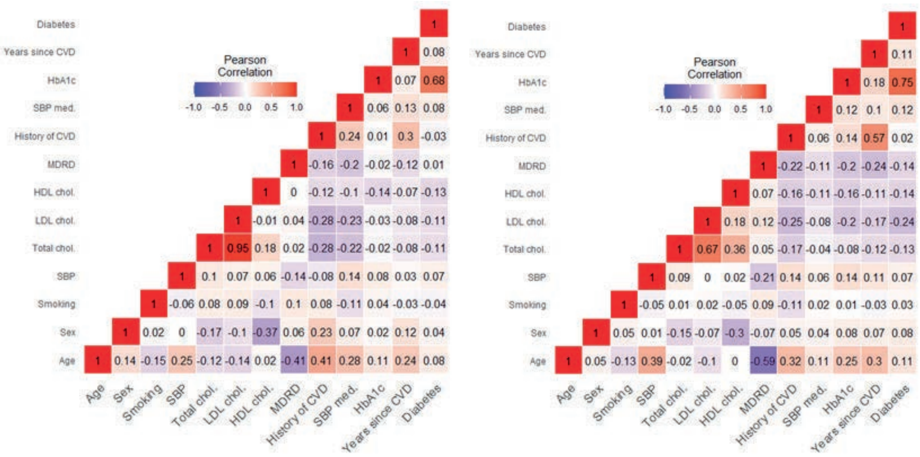
Legend – MSE: mean squared error, LP: linear predictor, CITL: calibration in the large, M-Imp: mean imputation, JMI: joint modelling imputation, JMI^{aux}: joint modelling imputation with auxiliary variables, *(1) systolic blood pressure, (2) total cholesterol, (3) HDL-cholesterol, (4) smoking, (5) antihypertensive drugs, (6) Diabetes mellitus.

Appendix C (cont.) simulation 3: scenario 8

#	Variables missing	Methods	MSE of the LP (% difference to M-Imp)	C-index	CITL	Calibration slope
	Apparent performance			0.7051	-0.0001	0.9999
1	MD scenario 8* (i.e. without local data, reference)	M-Imp	0.8334	0.6064	-0.1037	0.8230
		JMI	0.7963 (-4.45%)	0.6116	-0.2221	0.5769
		JMI+	0.7018 (-15.79%)	0.6721	-0.3649	0.8453
2	+100 local patients	M-Imp	0.8281	0.6064	-0.0965	0.8178
		JMI	0.7787 (-6.34%)	0.6122	-0.1910	0.5893
		JMI+	0.6764 (-22.43%)	0.6733	-0.3225	0.8608
3	+300 local patients	M-Imp	0.8173	0.6044	-0.0803	0.8118
		JMI	0.7677 (-6.46%)	0.6103	-0.1662	0.5841
		JMI+	0.6510 (-25.55%)	0.6734	-0.2786	0.8665
4	+750 local patients	M-Imp	0.8089	0.6050	-0.0540	0.8005
		JMI	0.7455 (-8.50%)	0.6108	-0.1166	0.6132
		JMI+	0.6117 (-32.24%)	0.6778	-0.2138	0.9232
5	+1500 local patients	M-Imp	0.7923	0.6107	-0.0205	0.8429
		JMI	0.7253 (-9.24%)	0.6131	-0.0659	0.6480
		JMI+	0.5740 (-38.03%)	0.6856	-0.1451	0.9654
6	+5000 local patients	M-Imp	0.7461	0.6144	0.0905	0.9096
		JMI	0.6695 (-11.44%)	0.6202	0.0387	0.7332
		JMI+	0.5094 (-46.47%)	0.6914	-0.0503	1.0142
7	+10000 local patients	M-Imp	0.7560	0.6024	0.1577	0.7556
		JMI	0.6734 (-12.27%)	0.6065	0.1049	0.6707
		JMI+	0.4999 (-51.23%	0.6973	0.0328	1.0664

Legend – MSE: mean squared error, LP: linear predictor, CITL: calibration in the large, M-Imp: mean imputation, JMI: joint modelling imputation, JMI^{aux}: joint modelling imputation with auxiliary variables, *(1) age, (2) gender.

Appendix D: Correlation matrix (with additional patient variables) – left: local data (SMART), right: external data (UCC)



Appendix E – R code

Code available upon reasonable request.

Nijman SWJ[a, 1], Oberman HI[b, 1], Bots ML[a], Asselbergs FW[cde], Moons KGM[a], Vink G[b], Debray TPA[ae], Smeden van M[a]

a Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands;

b Department of Methodology and Statistics, Utrecht University, Utrecht, the Netherlands

c Department of Cardiology, University Medical Center Utrecht, Utrecht University, The Netherlands;

d Institute of Cardiovascular Science, Faculty of Population Health Sciences, University College London, London, United Kingdom;

e Health Data Research UK, Institute of Health Informatics, University College London, London, United Kingdom

1 First author

Statistics in Medicine, 2022. Submitted.

CHAPTER 6

Real-time handling of missing data in the application of prediction models: a comparison of methods

Abstract

Introduction – The need to account for missing values in real time is unique to the application of prediction models but is underrepresented in the literature. In this study, we aim to evaluate various real-time strategies to handle the pervasive problem of missing data when using clinical data to make predictions on patients for whom part of the data is missing. We assess the influence of built-in missing data handling mechanisms on prediction accuracy and compare it with existing real-time imputation methods (e.g., joint modeling imputation).

Methods – We evaluate the effect of various missing data handling methods under specific missing data circumstances as would occur in medical practice in a simulation study. Hereto, we consider three types of missing data handling strategies: Joint Modelling Imputation (JMI), Pattern Submodels (PS), and Surrogate Splits (SS). The predicted risks are evaluated in terms of overall prediction accuracy (i.e., root mean squared error of the predicted risk and brier score), and in terms of discrimination (C-statistic) and calibration (i.e., calibration-in-the-large and the calibration slope).

Results – Simulation results suggests that both PS and JMI work reasonably well, provided JMI generated multiple imputations for each missing value. In comparison, when a RF was used, the performance of PS diminished.

Discussion – We recommend JMI-MD as it yielded good performance for both FLR and RF. When the goal is to use a RF, the use of JMI-CM and SS are not recommended.

Introduction

Incompleteness of medical records is a ubiquitous problem when using healthcare data. Besides the well-documented issues that missing data can create in data analyses, incompleteness of medical records may also create practical issues in clinical practice [9,28]. For instance, a prediction model that relies on historical but unrecorded data for a particular patient or prediction models that are used as early-warning systems for individual patients [27,126]. Most prediction models are not designed to be used when predictors are not fully observed, and ad-hoc approaches such as replacing the missing value with the population average value (i.e., mean imputation) is generally not advised [9,143]. As prediction models are increasingly being integrated in the electronic health record (EHR) via clinical decision support systems (CDSS), the issues concerning missing data and the need to deal with those missing values when applying prediction models in individual patients becomes more evident [19,110]. The issue is further complicated as the common strategies to mend or circumvent missing data in research are not directly applicable for use when predicting an outcome for an individual patient in a clinical practice setting.

Various strategies to handle different manifestations of missing data have been studied thoroughly and can usually provide more plausible substitution values (e.g., via imputation) [28]. Multiple imputation is often considered to be the gold standard for missing data problems and is known to provide valid estimates and correct standard errors in circumstances where the missingness does not depend on the unobserved values [6]. Most imputation algorithms, however, require direct access to data from multiple instances (i.e., multiple patients or multiple measurements) and are therefore not directly suitable for use on a case-by-case basis. Further, when a prediction model is applied to a single patient in clinical practice via a CDSS there is usually no access to any data from other individuals due to computational and privacy constraints [19].

An intuitive alternative to imputation is to solve for the missingness inside the prediction model instead of the data. Two promising methods of this type are the pattern submodel (PS) approach and surrogate splits (SS). PS are attractive to a variety of parameter-based modeling techniques (e.g., regression). The so-called submodels incorporate the nature of the missing data by developing a separate prediction model for all possible missing data patterns [50,172]. Then, when applied to a new case or out-of-sample individual the corresponding prediction model that matches the individual's missing data pattern is used. Whereas the PS approach lends itself to various kinds of prediction models, SS come naturally to tree-based methods, such as random forest models [36,37,173,174]. Briefly, SS attempt to preserve the partitioning of the original split by finding the next most optimal split given other observed variables. When the model is applied, each original split for which the predictor is missing will be replaced by the best available 'surrogate' variable to decide the split direction.

In this article we compare various real-time missing data handling approaches when implementing specific modeling techniques in clinical practice. We use the term 'real-time' to refer to methods that can be applied to data from a single individual as would occur in a clinical practice setting, possibly without the availability of data from other individuals. We present a simulation study and a motivating example to compare the different missing data handling strategies that can be used at the implementation level. The aim is to identify strengths and weaknesses of these approaches on the ability to estimate individualized risk, as quantified by the discrimination and calibration of the predictions.

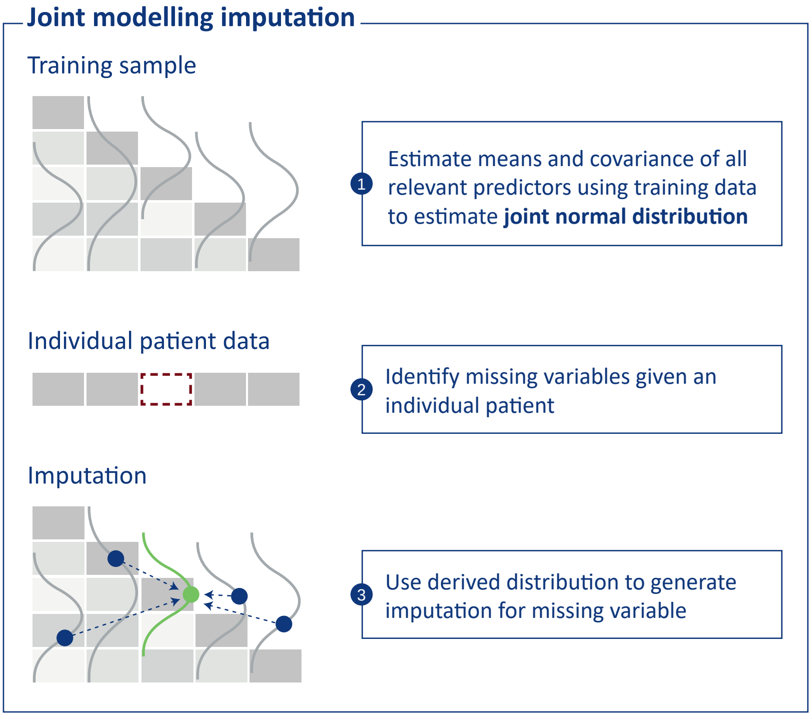
Missing data handling methods for prediction models

We consider the following three prediction modeling strategies for real-time handling of missing data: (i) prediction models that adopt joint modeling imputation, (ii) prediction models that adopt a pattern submodel approach (iii) prediction models that adopt random forests with surrogate splits [36,48,50,175].

Joint Modeling Imputation (JMI)

JMI is an imputation method that involves estimating the multivariate (joint) density of the predictor data and is used to generate imputed values directly from the conditional distribution [47]. An advantage of JMI is that it can be applied to a previously developed prediction model. Because distribution parameters cannot directly be estimated in incomplete data, JMI typically requires the implementation of a Gibbs sampler. Recently, an extension to JMI was proposed to allow for real-time imputation in individual patients [40,175]. With the extension the development of a JMI model consists of two separate steps. In the first step, the means and covariance of all predictor variables are estimated in a complete training sample from the population to which the prediction model will be applied. Since JMI assumes that every predictor variable is normally distributed, the population characteristics (i.e., means and covariance) can directly be used to generate, or draw, imputations on an individual level. In clinical practice, when a prediction model now encounters missing values, the developed JMI model can be utilized to generate imputations for each missing value on each predictor variable. We implemented three variants of JMI to be evaluated: single draw (JMI-SD, where a single draw from the conditional distribution is the imputed value), multiple draw (JMI-MD, where the average of 50 draws from the conditional distribution is the imputed value) and the conditional mean (JMI-CM, where the expected value of the conditional distribution is the imputed value). See Figure 1 for a schematic depiction of JMI.

Figure 1. Joint Modeling Imputation (JMI)



1 Estimate means and covariance of all relevant predictors using training data to estimate **joint normal distribution**

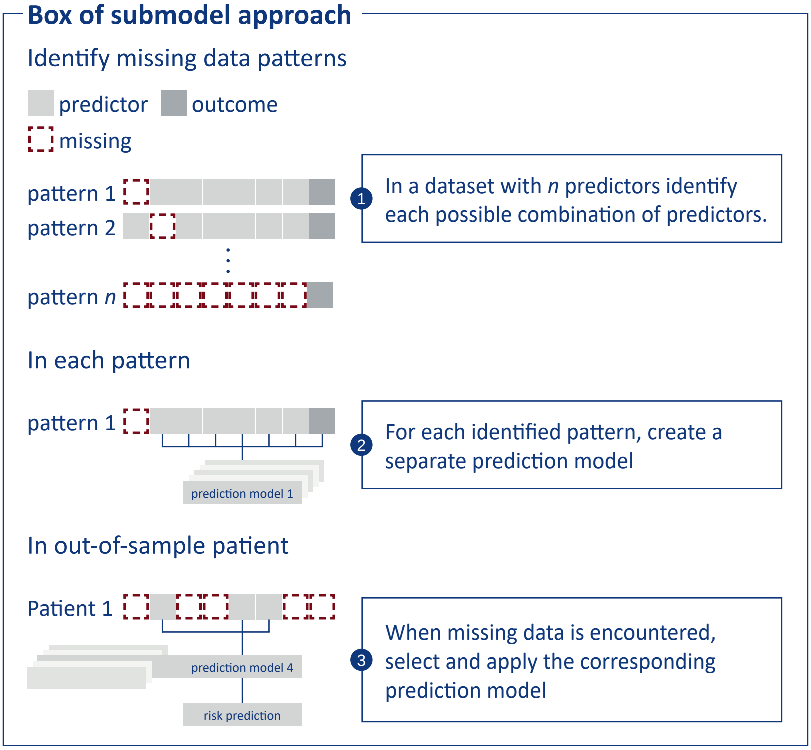
2 Identify missing variables given an individual patient

3 Use derived distribution to generate imputation for missing variable

Pattern Submodel (PS) approach

Another approach to address missing data without requiring imputation is to develop separate prediction models (so called pattern submodels, or briefly, PS) for each missing data pattern [50]. Each PS is to be made specifically for one of the identified missing data patterns in the training data and the missing data patterns that are encountered in clinical practice. When applied to a new, out-of-sample, individual, PS approach uses the corresponding prediction model (i.e., matching the missing data pattern at hand). A recent study has shown that the use of PS for prediction performs similarly to multiple imputation and outperform multiple imputation in some cases when the data are missing not at random (MNAR, when missing data is dependent on unobserved values) [50,172,176]. As such, PS may provide an elegant and intuitive to understand method for handling missing data when implementing prediction models. See figure 2 for a schematic depiction of the PS approach.

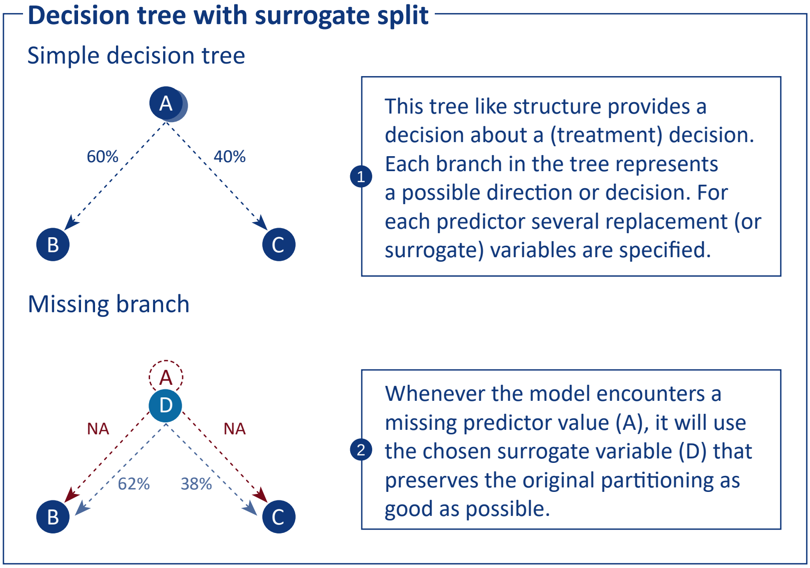
Figure 2. Pattern submodel approach


Box of submodel approach
Identify missing data patterns
predictor
outcome
missing
pattern 1
pattern 2
pattern *n*
1
In a dataset with *n* predictors identify each possible combination of predictors.
In each pattern
pattern 1
prediction model 1
2
For each identified pattern, create a separate prediction model
In out-of-sample patient
Patient 1
prediction model 4
risk prediction
3
When missing data is encountered, select and apply the corresponding prediction model

Surrogate Splits (SS)

A well-known family of ML-based prediction models are the tree-based models, with as a simple case a (single) decision tree [30,38]. Decision trees use a tree like structure to find the optimal cut-off point which partitions the data for optimal predictive performance. Based on the values of the pre-defined predictor variables, each branch in the tree represents a possible direction or decision. In essence, random forests combine multiple decision trees by using a combination of a random subspace method (i.e., random combinations of features) and bagging (i.e., random sample of observations). As an early extension to the well-known decision tree and random forest, SS were developed to circumvent the necessity for imputation [36,37,39]. Briefly, SS try to preserve the partitioning of each original split in a tree as good as possible in the presence of missing predictor values. Whenever the model is applied to an individual and encounters a missing predictor value, it will use the pre-specified surrogate (i.e., replacement) variable, rather than the missing predictor variable, to decide upon the split direction. See figure 3 for a schematic depiction of SS in the context of a single decision tree. In this study we use SS in combination with a random forest prediction model.

Figure 3. Decision tree with surrogate splits



1 This tree like structure provides a decision about a (treatment) decision. Each branch in the tree represents a possible direction or decision. For each predictor several replacement (or surrogate) variables are specified.

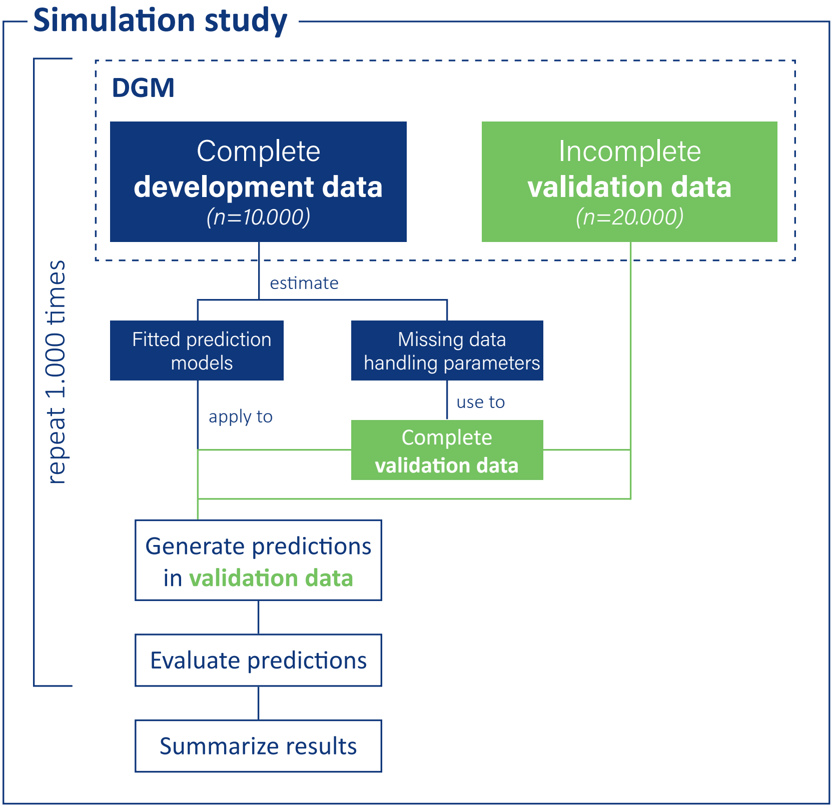
2 Whenever the model encounters a missing predictor value (A), it will use the chosen surrogate variable (D) that preserves the original partitioning as good as possible.

Simulation design

Aims

The aim of the simulation study is to emulate how a single patient would present themselves in clinical practice, with incomplete prediction model data, and to evaluate the performance of several real-time missing data handling approaches. We compare the performance of these missing data approaches on their ability to generate accurate risk predictions. We consider the situation in which a complete dataset is available for prediction model development, and that the resulting model is then applied to individual patients with missing observations for one or more variables. For an overview of the simulation, see Figure 4; for the full script and technical details, see github.com/hanneoberman/real-time-missing.

Figure 4. Simulation study



Data-generating mechanism

All data are generated from a single model-based population, consisting of ten continuous predictors and one dichotomous outcome. In each simulation iteration, we draw two samples from the population: a complete development set (*n* = 10.000), and a validation set in which we introduce missing values to mimic how patients would present themselves in clinical practice (*n* = 20.000).

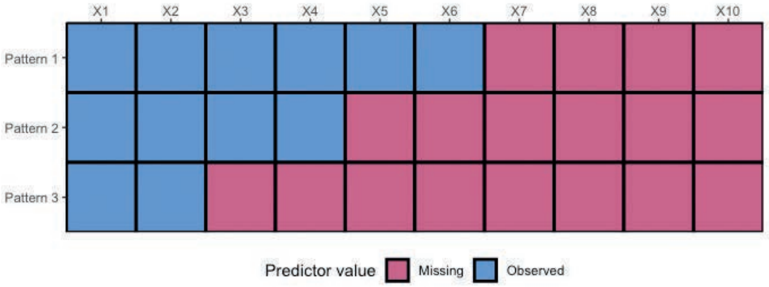
The data generating mechanism of the predictor space is a multivariate normal distribution, $X \sim \mathcal{N}(\mu,\Sigma)$, with mean vector Y and covariance matrix Σ (Supplementary materials A). Correlations between the ten predictors range from *r* = -.37 to *r* = .36. From the predictor space, we define the binary outcome vector Y. is a function of X through the logit link function,

$$\text{logit}(Pr(Y = 1)) = \alpha + \beta \times X + \beta^{*} \times x_1 \times X + \varepsilon,$$

Where α is the intercept, βs are regression coefficients, and ε is the residual error term $\varepsilon \sim \mathcal{N}(0{,}2)$. We differentiate between two types of regression coefficients: β is a vector of regression coefficients for the main effects of the predictors, $\beta = [\beta_1, \beta_2, \ldots, \beta_{10}]$; β^* is a vector of regression coefficients for the interactions with the first predictor, $\beta^* = [\beta_1^*, \beta_2^*, \ldots, \beta_{10}^*]$. This introduces a polynomial effect of the second degree, $\beta_1^* \times x_1^2$, and nine interaction effects. For additional non-linearity, we use a transformation in the effect of the second predictor, $\beta_2 \times \log(|x_2|)$. All regression coefficients can be found in Supplementary materials B. The expected occurrence of the outcome is 15%.

The validation set is amputed (i.e., made incomplete) according to several missingness mechanisms and missingness rates. In this study, we focus primarily on the Missing At Random (MAR) missingness mechanism and additionally on the Missing Not At Random (MNAR) missing mechanism [5]. We use a mixture of the four kinds of MAR missingness, as described by Schouten and others [177]. The overall missingness rate is 60%, but the number of missing predictor entries differs between cases. The hypothetical patients in our validation set are missing either 40%, 60%, or 80% of the observations in the predictor space. The resulting missing data pattern is visualized in Figure 5.

Figure 5. Missing data pattern.



Methods

Our methods consist of nine pairs of missing data methods and prediction models to predict the absolute risk of the outcome in real-time. For an overview of all methods, see Table 1.

To accommodate for missing predictor values in real-time, we consider three types of missing data handling strategies: JMI, PS, and SS. Since JMI can have different implementations, we further subdivide this strategy into (i) imputing the conditional mean (JMI-CM), (ii) single imputation with a random draw from the conditional multivariate distribution (JMI-SD), and (iii) multiple imputation with 50 draws from the conditional multivariate distribution and pooling (i.e., taking the average of) the predictions of the outcome (JMI-MD).

We obtain predictions of the outcome by applying two models on the incomplete (imputed) predictor space. The first prediction model is flexible logistic regression (FLR) with a natural cubic spline. The second prediction model is a random forest (RF). Both prediction models are compatible with the JMI and PS. The SS missing data strategy is only available for tree-based prediction models, such as a random forest. Technical details such as model tuning can be found in Supplementary Materials C and on github.com/hanneoberman/real-time-missing.

Table 1. Overview of missing data methods and prediction models.

		Prediction model	
	Missing data technique	**FLR**	**RF**
JMI-CM	**Conditional mean imputation.** Missing values are imputed by the predictor mean, conditional on the observed values of the other predictors.	x	x
JMI-SD	**Single draw imputation.** Missing values are imputed by a random draw from the conditional multivariate distribution of the predictor.	x	x
JMI-MD	**Multiple draw imputation.** Missing values are imputed 50 times by a random draw from the multivariate normal distribution, and subsequently used to obtain 50 predictions of the outcome, which are then averaged to obtain one pooled prediction.	x	x
PS	**Pattern submodels.** Missing values are circumvented by selecting the appropriate pattern submodel for predicting the outcome.	x	x
SS	**Surrogate splits.** Missing values are accommodated using surrogate splits.		x

Performance measures

We evaluate the estimates (the predicted risk of the outcome for each of the hypothetical patients) in terms of overall prediction accuracy at the individual patient-level, and in terms of discrimination and calibration. Subsequently, all metrics are averaged across simulation iterations. Table 2 provides an overview of the performance measures: root mean squared error

(RMSE) of the predicted risk, brier score, concordance (C-) statistic, calibration-in-the-large (CITL), and the calibration slope.

Table 2. Performance measures

Measure	Performance metric
Overall prediction accuracy	Root mean square error (RMSE). The RMSE of the predictions reflects the difference between the estimated probability of Y and the true underlying probability of the outcome before amputation. Like the estimand and estimates, the RMSE lies on the probability scale. Lower values indicate better performance [178].
	Brier score. The brier score is defined as the squared difference between the predicted risk and the observed outcome value. A brier score of 0 would represent a perfect model, whilst the maximum brier score is determined by the incidence of the outcome [13].
Discrimination	Concordance (C-)statistic. The C-statistic is a rank-order statistic, which is used to describe how well a classification model can discriminate between those with an event and those without. The C-statistic shows the probability of taking two random subjects (one with and one without the outcome) and correctly attributing the one with the outcome with a high risk. A C-statistic of 0.5 describes a model with no discriminative performance and a C-statistic 1 describes a model with perfect discriminative performance.
Calibration	Calibration-in-the-large (CITL). The CITL represents the overall calibration of a model. In other words, the extent of agreement between the average predicted risk and the original predicted risk [144]. The metric ultimately describes the amount of systematic over- or under-estimation of the predicted risk. A value of 0 is ideal and represents perfect agreement.
	The calibration slope. In contrast with the CITL, the calibration slope does not evaluate the average predicted, or original, risk. Rather, it quantifies the extent by which the predicted risks vary too much (i.e., slope <1) or too little (i.e., slope >1). Ideally, the slope is 1.

Results

Figure 6 displays the performance of the real-time missing data approaches across simulations. Table 3 presents the average performance across simulations. The additional simulation under a MNAR missingness mechanism showed equivalent results, and can be found in Supplement D. For reasons of brevity, we exclude the severely under-performing missing data approach JMI-SD from any further reported results.

Root mean squared error

Overall, imputation and non-imputation missing data handling methods were very similar in their ability to recover the original probability of the outcome. When implemented with a FLR, PS performed best. A very similar performance was obtained when adopting a FLR model after imputation with JMI-CM or JMI-MD. For the random forest prediction model, JMI-MD outperformed all other missing data approaches. RF with SS and PS showed relatively low accuracy.

Brier score

When paired with a FLR, both imputation (JMI-MD and JMI-CM) and non-imputation (PS) missing data handling methods had an equivalent performance. When a random forest prediction model was used, JMI-MD appeared to be slightly better at approximating the binary realization of the outcome than JMI-CM, with SS and PS again showing relatively poor performance.

C-statistic

The use of JMI-MD paired with RF marginally exceeded the performance of other techniques, now in terms of discriminating between cases and non-cases. The discriminatory ability of JMI-CM and JMI-MD with FLR are mostly equivalent. The performances of JMI-CM and PS are diminished when comparing the random forest prediction model to FLR. And, although slightly better than PS, the performance of SS is below par.

Calibration-in-the-large

Both PS and JMI-MD showed near perfect overall calibration when paired with a FLR. With JMI-CM showing an only marginally worse performance. Whilst all missing data handling techniques had very similar performances when paired with a RF, JMI-MD remained the favourite with near perfect calibration.

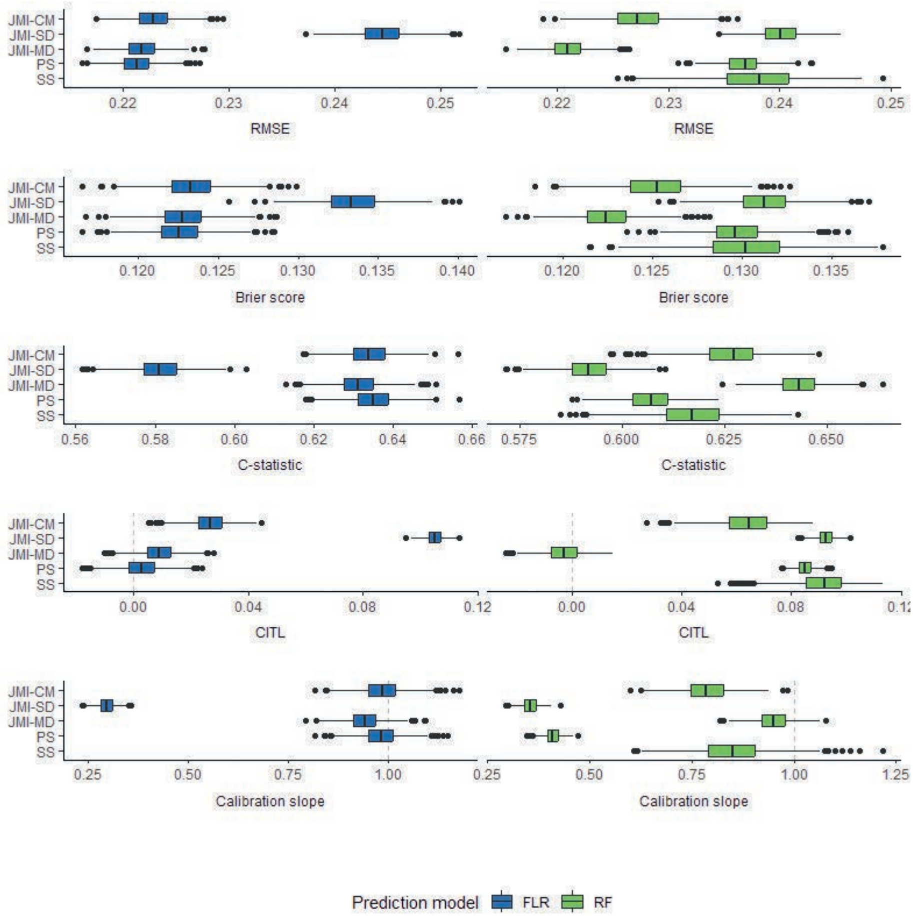
Calibration slope

In contrast with other performance metrics, the best performance is observed with JMI-CM paired with FLR, which could best quantify the extremeness of predicted risks across the whole range. Both JMI-MD and PS had similar performance. Apart from JMI-MD, all missing data handling techniques showed miscalibration when a random forest prediction model is used.

Calibration plots

Figure 7 presents calibration plots for the methods of interest, taken from a single iteration in the simulation. The missing data approaches can be found in the row-wise panels; the prediction models in the columns (left = FLR, right = RF). Within each plot, dashed lines show optimal calibration (i.e., perfect match between predicted and actual probabilities), colored lines (blue for FLR, green for RF) are Loess lines with standard errors through the calibration, and the shaded grey area represents the density of the predicted probabilities.

Figure 6. Performance measures per method



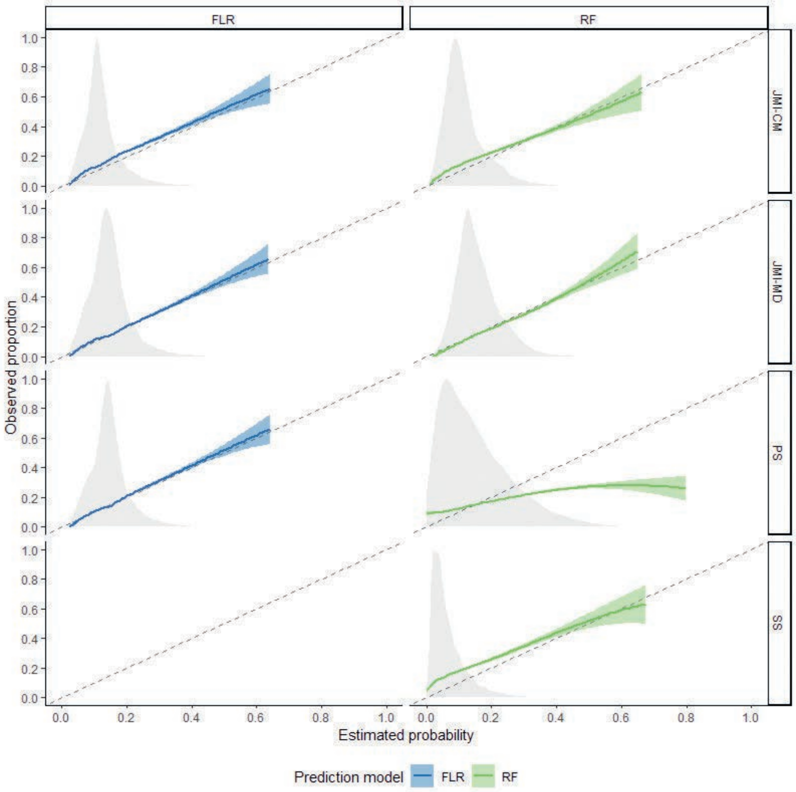
Legend – JMI-CM: conditional mean imputation; JMI-SD: single draw imputation; JMI-MD: multiple draw imputation; PS: pattern submodels; SS: surrogate splits; AUC: area under the curve; RMSE: root mean squared error; FLR: flexible logistic regression; RF: random forest

Table 3. Average performance across simulations.

		RMSE	EmpSE	Brier	EmpSE	C-index	EmpSE	CITL	EmpSE	Slope	EmpSE
FLR	JMI-CM	0.223	(0.002)	0.123	0.002	0.634	0.006	0.027	0.006	0.985	0.05
	JMI-SD	0.244	0.002	0.133	0.002	0.581	0.006	0.105	0.003	0.297	0.02
	JMI-MD	0.222	0.002	0.123	0.002	0.631	0.006	0.009	0.006	0.941	0.044
	PS	0.221	0.002	0.123	0.002	0.635	0.006	0.003	0.007	0.981	0.047
RF	JMI-CM	0.227	0.003	0.125	0.002	0.627	0.008	0.064	0.01	0.789	0.058
	JMI-SD	0.240	0.002	0.131	0.002	0.592	0.006	0.093	0.003	0.355	0.02
	JMI-MD	0.221	0.002	0.122	0.002	0.643	0.006	-0.003	0.007	0.952	0.041
	PS	0.237	0.002	0.130	0.002	0.607	0.006	0.085	0.003	0.410	0.018
	SS	0.238	0.004	0.130	0.003	0.617	0.01	0.091	0.01	0.851	0.087

Legend – RMSE: root mean squared error; EmpSE: empirical standard errors; C-index: concordance-index; CITL: calibration-in-the-large; FLR: flexible logistic regression; RF: random forest; JMI-CM: conditional mean imputation; JMI-SD: single draw imputation; JMI-MD: multiple draw imputation; PS: pattern submodels; SS: surrogate splits.

Figure 7. Calibration plots



Legend – FLR: flexible logistic regression; RF: random forest; JMI-CM: conditional mean imputation; JMI-MD: multiple draw imputation; PS: pattern submodels; SS: surrogate splits.

Discussion

This simulation study evaluated real-time missing data handling strategies to handle missing predictor values in individual patients. We considered JMI, PS and SS for the real-time handling of missing data when using either a FLR or RF. Our simulation study showed that the optimal choice of missing data handling technique may be dependent on the preferred prediction modeling approach. Overall, simulation results suggests that PS (when paired with FLR) and JMI (provided multiple imputations are generated) work reasonably well. Multiple imputation was found to be more consistent than imputing a conditional mean. In contrast, SS performed relatively poor. Likewise, imputing single draws severely underperformed on all metrics.

Generally, we found that missing data handling techniques yielded better performance when paired with FLR rather than RF. Possibly, this is because our dataset included mostly continuous predictors and the DGM was a logistic regression model. RF have been reported to perform particularly well when dealing with a very large number of discrete variables, especially in the presence of interactions [38,179]. Possibly, RF is also more prone to overfitting when estimated in smaller (sub)samples as compared to FLR. However, it is likely that due to the larger sample size in our simulation study, this is not the case. Due to the choice of DGM, comparisons between FLR and RF may be skewed in favour of FLR; consequently, any comparisons between the two modeling techniques may be irrelevant.

The good performance of JMI in our simulations may partly be driven by the choice of predictor correlation structure and missing data pattern in our simulations. Low correlations have previously been associated with limited performance of JMI [175]. Likewise, SS very heavily rely upon the correlation between the missing predictor value and the surrogate replacement value [174]. With the low to moderate correlations imposed in our DGM, it may be expected that multivariable approaches such as JMI perform better when compared with SS, which relies only on the single surrogate variable. For example, in the most extreme missing data scenario, when only and are observed, it is likely that optimal surrogate variables are not available. It may be evident that PS, which uses only the observed predictor variables, is also limited in circumstances such as these. In the end, when using clinical data, correlations between predictor variables need to be considered.

Additionally, to avoid overfitting, prediction models are typically designed as simple as possible and usually include predictors that do not intercorrelate much. Likewise, in our simulation study, we only generated 10 covariates, all of which were used for development of the prediction model and imputation strategies. In practice, however, many more additional variables may be available.

These auxiliary variables (i.e., not part of the prediction model) have previously shown to improve JMI performance [48]. If made available, it is likely that auxiliary variables, if not for prediction, may improve the accuracy of any missing data handling strategy which relies upon correlations between available variables.

Generally, PS has adequate prediction model performance in the presence of missing data. A major advantage for PS is that it does not require MAR assumptions. In real-world datasets PS, therefore, offer an appealing solution. When PS is paired with RF, however, problems arise. These problems may be explained by the fact that less predictors ultimately restrict how much a random forest may vary between each tree [179]. In other words, if there are less features available, as is the case for PS, the variability between trees is limited. Similarly, surrogate splits perform relatively poor, which can be explained by the strong dependence on high correlations between the surrogate variable and the missing predictor variable.

In summary, the best missing data handling technique depends on the prediction modeling technique. JMI-MD is considered the safest choice for handling missing data as it yielded good performance for both FLR and RF, whilst PS only obtained good performance when paired with FLR. The use of JMI-CM and surrogate splits are not recommended when using RF. Similarly, JMI-SD should be avoided.

Disclosures

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). HO is supported by a Special Interest Group grant for Applied Data Science from the University Utrecht.

All other 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.

Supplementary Materials

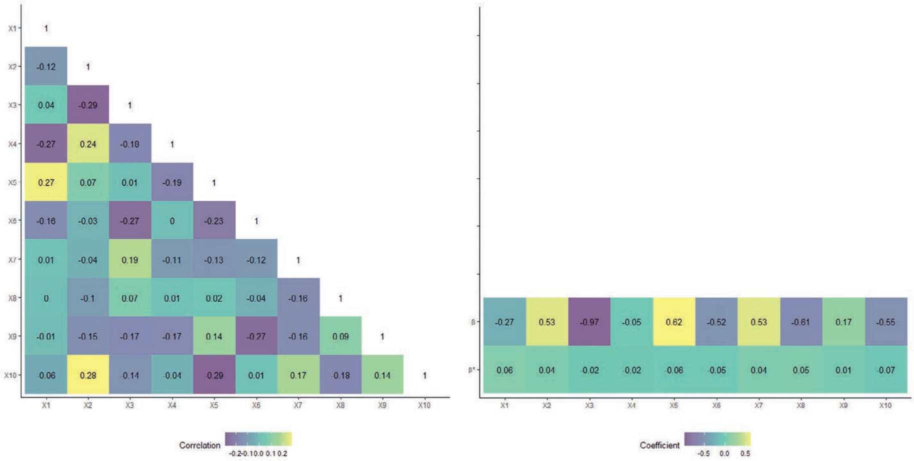
A. DGM

Means vector: All 10 predictors have a mean of zero, $\mu = [0{,}0{,}\ldots{,}0]$.

Covariance matrix:

$$\Sigma = \begin{bmatrix} 1.05 & -0.12 & 0.04 & -0.29 & 0.29 & -0.17 & 0.01 & 0.00 & -0.01 & -0.07 \\ -0.12 & 1.08 & -0.31 & 0.26 & 0.08 & -0.03 & -0.04 & -0.11 & -0.17 & 0.30 \\ 0.04 & -0.31 & 1.08 & -0.19 & 0.01 & -0.29 & 0.20 & 0.07 & -0.18 & -0.15 \\ -0.29 & 0.26 & -0.19 & 1.07 & -0.20 & 0.00 & -0.12 & 0.01 & -0.19 & -0.04 \\ 0.29 & 0.08 & 0.01 & -0.20 & 1.08 & -0.25 & -0.14 & 0.02 & 0.15 & -0.32 \\ -0.17 & -0.03 & -0.29 & 0.00 & -0.25 & 1.08 & -0.13 & -0.04 & -0.29 & 0.01 \\ 0.01 & -0.04 & 0.20 & -0.12 & -0.14 & -0.13 & 1.04 & -0.16 & -0.17 & 0.18 \\ 0.00 & -0.11 & 0.07 & 0.01 & 0.02 & -0.04 & -0.16 & 1.02 & 0.10 & -0.19 \\ -0.01 & -0.17 & -0.18 & -0.19 & 0.15 & -0.29 & -0.17 & 0.10 & 1.08 & 0.15 \\ -0.07 & 0.30 & -0.15 & -0.04 & -0.32 & 0.01 & 0.18 & -0.19 & 0.15 & 1.08 \end{bmatrix}$$

Correlations:



B. Regression coefficients:

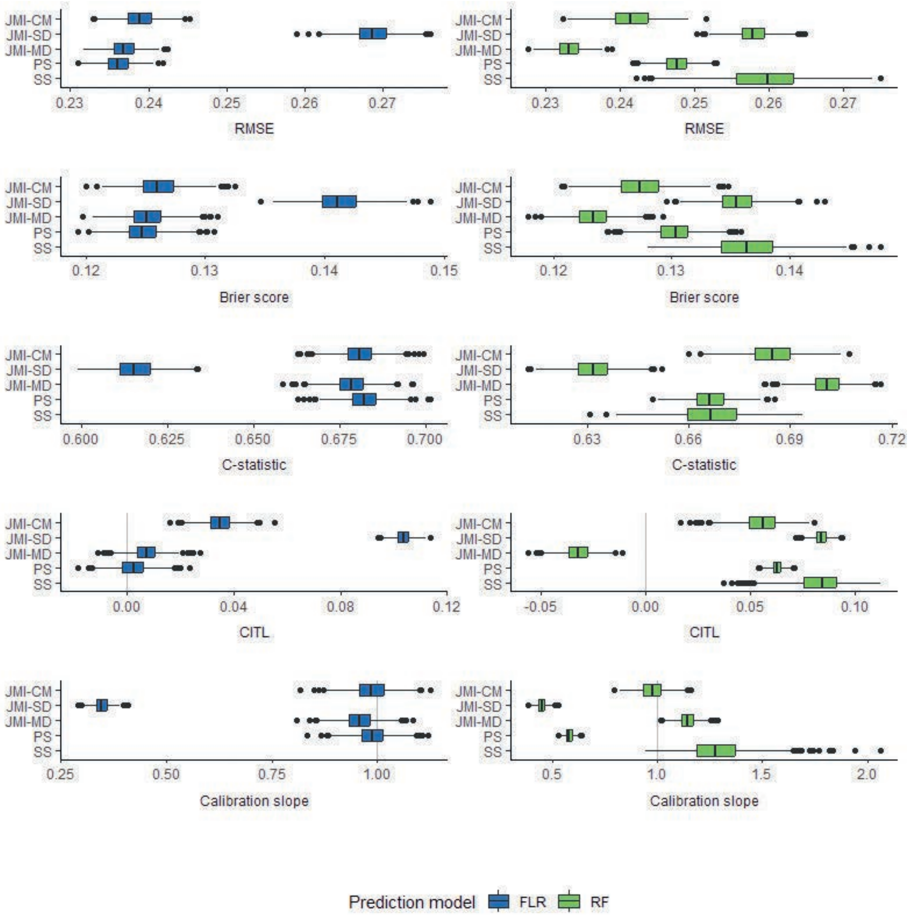
$\beta_0 = -3$

$$\begin{aligned} \beta &= [\quad -0.27 \quad 0.53 \quad -0.97 \quad -0.05 \quad 0.62 \quad -0.52 \quad 0.53 \quad -0.61 \quad 0.17 \quad -0.55 \quad] \\ \beta^* &= [\quad 0.06 \quad 0.04 \quad -0.02 \quad -0.02 \quad -0.06 \quad -0.05 \quad 0.04 \quad 0.05 \quad 0.01 \quad -0.07 \quad] \end{aligned}$$

C. Model tuning

- FLR: glm() with natural spline with 3 degrees of freedom.
- RF: ranger::ranger() with defaults (500 trees and 3 predictors considered for each split), party::cforest() with defaults (500 trees, 5 predictors considered for each split, and 3 surrogate variables considered for each split with missingness).

D. Performance under MNAR



Legend – JMI-CM: conditional mean imputation; JMI-SD: single draw imputation; JMI-MD: multiple draw imputation; PS: pattern submodels; SS: surrogate splits; AUC: area under the curve; RMSE: root mean squared error; FLR: flexible logistic regression; RF: random forest

Average performance under MNAR

		RMSE	Brier	C-index	CITL	Slope
FLR	JMI-CM	0.239	0.126	0.681	0.035	0.985
	JMI-SD	0.269	0.141	0.616	0.104	0.347
	JMI-MD	0.237	0.125	0.679	0.007	0.957
	PS	0.236	0.125	0.682	0.002	0.988
RF	JMI-CM	0.242	0.127	0.685	0.055	0.978
	JMI-SD	0.258	0.136	0.632	0.083	0.45
	JMI-MD	0.233	0.123	0.701	-0.032	1.144
	PS	0.248	0.13	0.666	0.062	0.581
	SS	0.259	0.136	0.667	0.083	1.287

Toshihiko Takada, Ph.D.[1], Steven Nijman, M.Sc.[1], Spiros Denaxas, Ph.D.[2,3,4,5],
Kym I.E. Snell, Ph.D.[6], Alicia Uijl, Ph.D.[1,7,8], Tri-Long Nguyen, Ph.D.[1,9],
Folkert W. Asselbergs, Ph.D.[2,8,10], Thomas P.A. Debray, Ph.D.[1,2]

1 Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Universiteitsweg 100, 3584 CG, Utrecht, The Netherlands
2 Health Data Research UK and Institute of Health Informatics, University College London, Gibbs Building, 215 Euston Road, London, NW1 2BE, United Kingdom
3 The Alan Turing Institute, British Library, 96 Euston Road, London, NW1 2DB, United Kingdom
4 The National Institute for Health Research University College London Hospitals Biomedical Research Centre, University College London, Suite A, 1st floor, Maple House, 149 Tottenham Court Road, London, W1T 7DN, United Kingdom
5 British Heart Foundation Research Accelerator, University College London, Gower Street, London, WC1E 6BT, United Kingdom
6 Centre for Prognosis Research, School of Medicine, Keele University, Keele, Staffordshire, ST5 5BG, United Kingdom
7 Division of Cardiology, Department of Medicine, Karolinska Institute, 171 77 Stockholm, Sweden
8 Department of Cardiology, Division Heart & Lungs, University Medical Center Utrecht, Utrecht University, Heidelberglaan 100, PO Box 85500, 3508GA, Utrecht, The Netherlands
9 Section of Epidemiology, Department of Public Health, University of Copenhagen, CSS, Øster Farimagsgade 5, DK-1353 Copenhagen K, Denmark
10 Institute of Cardiovascular Science, Faculty of Population Health Sciences, University College London, Gower Street, London, WC1E 6BT, United Kingdom

Journal of Clinical Epidemiology, 2021. 137 83–91. DOI: 10.1016/j.jclinepi.2021.03.025

CHAPTER 7

Internal-external cross-validation helped to evaluate the generalizability of prediction models in large, clustered datasets

Abstract

Objective – To illustrate how to evaluate the need of complex strategies for developing generalizable prediction models in large, clustered datasets.

Methods – We developed eight Cox regression models to estimate the risk of heart failure using a large population-level dataset. These models differed in the number of predictors, the functional form of the predictor effects (non-linear effects and interaction) and the estimation method (maximum likelihood and penalization). Internal-external cross-validation was used to evaluate the models' generalizability across the included general practices.

Results – Among 871,687 individuals from 225 general practices, 43,987 (5.5%) developed heart failure during a median follow-up time of 5.8 years. For discrimination, the simplest prediction model yielded a good concordance statistic, which was not much improved by adopting complex strategies. Between-practice heterogeneity in discrimination was similar in all models. For calibration, the simplest model performed satisfactorily. Although accounting for non-linear effects and interaction slightly improved the calibration slope, it also led to more heterogeneity in the observed/expected ratio. Similar results were found in a second case study involving patients with stroke.

Conclusion – In large, clustered datasets, prediction model studies may adopt internal-external cross-validation to evaluate the generalizability of competing models, and to identify promising modelling strategies.

What is new?

Key findings

- Flexible modelling strategies did not improve prediction model performance across different settings and populations.
- Although the inclusion of additional predictors marginally improved the model's discriminative performance, it also increased between-practice heterogeneity (thereby impairing model generalizability).

What this adds to what was known

- In contrast to traditional internal validation methods, internal-external cross-validation (IECV) can quantify the generalizability of a prediction model across different settings and populations.

What is the implication and what should change now?

- When developing prediction models using large, clustered datasets, both their internal and external validity should be studied.
- IECV can be used to compare the practical benefits of different modelling strategies, and to simplify model complexity.

Introduction

In medicine, there are an increasing number of clinical prediction models [180]. These models aim to predict a risk of having a certain condition or experiencing a health event in the future. Prediction models are often developed using a single and small dataset. This leads to prediction models that are more prone to overfitting with the dataset used for its development, which leads to poor accuracy and less generalizability of risk predictions when the model is validated or used in new individuals.

For this reason, there has been a growing interest in prediction model studies using large datasets from electronic health records (EHRs), multi-center studies or individual participant data [149,181–183]. An advantage of such large datasets is that parameters of the prediction model can accurately be estimated, thereby facilitating the development of complex models with many predictors, interaction terms and/or non-linear effects. Furthermore, a common feature of these large datasets is that individuals are often clustered within hospitals, primary care practices, or even within countries. Clusters may differ with respect to included participants, variable definitions, and measurement methods, all of which may affect the generalizability of developed prediction models. The presence of clustering, however, also offers an important opportunity, as the performance of a prediction model can be examined on multiple occasions and thus be used to explore its generalizability across different settings and populations. Recently, various strategies for such analyses using large, clustered data have been proposed [149,181,184–186].

The aim of this study was to illustrate how advanced methods can be used to evaluate the need of complex strategies for developing generalizable clinical prediction models in large, clustered datasets.

Methods

For illustration purpose, we used two case studies.

Case study 1

We compared various modelling strategies using an example of a prediction model for the incidence of heart failure (HF). In the field of cardiovascular diseases (CVD), HF is one of the most relevant outcomes due to its high morbidity and mortality [127,187–189].

Source of the data

We used an existing large population-level dataset which links three sources of EHRs in England: primary care records from the Clinical Practice Research Datalink (CPRD), secondary care diagnoses and procedures recorded during admissions in Hospital Episodes Statistics (HES), and the cause-specific death registration information sourced from the Office for National Statistics (ONS) registry. This study was carried out as part of the CALIBER © resource (https://www.ucl.ac.uk/health-informatics/caliber and https://www.caliberresearch.org/) [190,191]. CALIBER, led from the UCL Institute of Health Informatics, is a research resource providing validated EHR phenotyping algorithms and tools for national structured data sources. Data were recorded in five controlled clinical terminologies: Read version 2 (CPRD diagnoses), International classification of diseases (ICD)-9 and ICD-10 (HES diagnoses, ONS causes of death), the Office of Population Censuses and Surveys (OPCS)-4 (HES procedures) and British National Formulary (BNF) (CPRD medication prescriptions). The study was approved by the MHRA (UK) Independent Scientific Advisory Committee (14_246RMnA2), under Section 251 (NHS Social Care Act 2006).

Population

The construction of this cohort has been described by Uijl et al [192]. Briefly, we selected all individuals that were 55 years or older between 1st January 2000 and 25th March 2010, and had at least one year of follow-up by a general practitioner, in a practice that had at least one year of up-to-standard data recording in CPRD. The last date of the follow-up between the period above was considered cohort entry date (index date). Individuals with a history of HF before their index date were excluded. The study flow diagram is shown in Appendix A.

Predictors

We identified predictors that are commonly measured in CPRD or HES, and commonly used for prediction of HF [192,193]: Age, sex, current smoking, ethnicity (CE, Caucasian ethnicity), index of multiple deprivation (IMD), body mass index (BMI), creatinine level (CL), and total cholesterol (TC). IMD is a measure of multiple deprivation at the small area level, consisting of seven domains [194]. Within this set, we selected those predictors which were least affected by missing data. The closest measurement to index date between three years before and one year after the index date was used. Detailed information about the definition of each predictor is available on the CALIBER website [195].

Outcomes

The primary outcome was incidence of HF, based on the first record of HF from CPRD or HES after the index date. In CPRD, HF was defined by a diagnosis of HF or chronic left ventricular dysfunction on echocardiogram with READ codes. In HES, it was defined by a diagnosis of HF during a hospitalization using all positions of ICD-10. If no diagnosis of HF was made, censoring was defined as the first event among the following: death, de-registration from a practice, last practice data collection, or at the study end date.

Statistical analysis

Multilevel imputation

Multiple multilevel imputation which accounts for potential heterogeneity between the included clusters is recommended in the recent methodological guidelines [196], however, due to limited hardware processing capacity, we applied single multilevel imputation. The detail of the imputation process is described in Appendix B.

Derivation and validation of prediction models

We considered eight modelling strategies to predict the risk of developing HF using Cox regression. These models differed with respect to the number of predictors, the functional form of the predictor effects and the method of estimation. Each model and their estimation method are summarized in Table 1.

Model 1 included four predictors (age, sex, current smoking, and CE) as linear effects. Model 2 was an extension of Model 1 that included non-linear effect for age and for all possible two-way interactions between the four predictors. Model 3 and 4 included the same predictors as Model 1 and 2, respectively, but were estimated using a ridge penalty. Model 5 was an extension of Model 1 that also included IMD, BMI, CL and TC as linear effects. Model 6 – 8 were extended from Model 5 as similar to Model 2 – 4 from Model 1. In models with a ridge penalty (Model 3, 4, 7 and 8), all regression coefficients were shrunk towards zero by penalizing the partial log-likelihood for the magnitude of the squared coefficients (L2-norm) [197]. This strategy has been recommended to avoid overfitting, and to improve prediction model performance, particularly when it is applied in new population. We used the degree of penalty (lambda) which minimized the mean square error in ten-fold cross validation. The proportional hazards assumption of all models was checked using the Schoenfeld residuals.

Table 1. Description of the eight prediction models

Model	Included predictor variables								2-way IT	# RC	Estimation method
	Age	Male Sex	Smoking	CE	IMD	BMI	CL	TC			
1	L	L	L	L	-	-	-	-	no	4	Cox regression
2	RCS	L	L	L	-	-	-	-	yes	14	Cox regression
3	L	L	L	L	-	-	-	-	no	4	Ridge penalized Cox
4	RCS	L	L	L	-	-	-	-	yes	14	Ridge penalized Cox
5	L	L	L	L	L	L	L	L	no	8	Cox regression
6	RCS	L	L	L	RCS	RCS	RCS	RCS	yes	66	Cox regression
7	L	L	L	L	L	L	L	L	no	8	Ridge penalized Cox
8	RCS	L	L	L	RCS	RCS	RCS	RCS	yes	66	Ridge penalized Cox

IT=interaction terms. #RC=the total number of regression coefficients. CE=Caucasian ethnicity. IMD=index of multiple deprivation. BMI=body mass index. CL=creatinine level. TC=total cholesterol. L=Linear effects. RCS=restricted cubic splines Models 1, 3, 5 and 7 include all predictor variables as linear effects. Models 2, 4, 6 and 8 use RCS with three knots for all continuous predictor variables, and interaction terms between all possible combinations of two variables. For all models, the total number of regression coefficients is displayed.

We performed internal-external cross-validation (IECV) to compare the performance of the aforementioned eight prediction models at multiple occasions [181,184]. In contrast to traditional internal validation methods (e.g., bootstrapping, cross-validation) which evaluate the model's performance in new individuals from the same population (i.e., reproducibility), IECV assesses model performance in new individuals from different but related practices as compared to the original development sample. These practices (i.e., taken as cluster) may differ with respect to case-mix, variable definitions and measurement methods, and thus allow to investigate the model's generalizability [163,182]. Using IECV, the data from all but one practice are used for estimating the prediction model, after which its performance is evaluated in the remaining practice. The procedure is repeated by rotating the omitted practice, resulting in multiple estimates of prediction model performance. For each prediction model, we assessed the model's discrimination performance using Harrell's concordance (c-) statistic. For calibration, we constructed calibration plots in the overall population. We also estimated the calibration slope and the ratio of observed versus expected events (O:E ratio) at five years of follow-up [145]. Interpretation of each performance measure is described in Appendix C.

The performance measures resulting from IECV were pooled using random-effect meta-analysis [181,198,199]. This approach not only accounts for the precision of practice-specific performance estimates, but also quantifies the between-practice variability (heterogeneity) of model performance. Heterogeneity is quantified by the between-practice standard deviation of model performance (τ) [185]. Meta-analysis results were reported as point estimates with 95% confidence intervals (CI) and 95% prediction intervals (PI). The CI indicates the precision of the model's average performance across all practices. Conversely, the PI accounts for heterogeneity between practices and therefore indicates what performance can be expected when the model is applied within a specific practice.

Case study 2

In this case study, we used patient-level data from a large international, multi-center, randomized controlled trial [200]. Because the missingness proportion was very low (6.0%), we performed a complete case analysis. Eight modelling strategies using ridge penalized Cox regression model were considered to predict the risk of mortality from CVD in patients with acute ischemic stroke. These models differed with respect to the number of predictors, the functional form of the predictor effects (non-linear effects and/or interaction terms). We illustrated the advantage of IECV by comparing it with bootstrap internal validation. More detailed information is available in Appendix D.

All analyses were performed using R version 3.6.1.

Results

Case study 1

The cohort included 871,687 individuals from 225 general practices. Among these, 43,987 (5.5%) developed HF during a median follow-up time of 5.8 years (interquartile range [IQR] 2.7 – 9.9), with a median time-to-event of 3.7 years (IQR 1.8 – 6.4). Baseline characteristics are shown in Table 2.

Table 2. Baseline characteristics of the cohort

Predictor variable	Individuals with incident HF	Individuals without HF	Proportion of missing
Total number of patients	43,987	823,700	
Age, years, median (IQR)	75·5 (68·5 – 81·5)	60·6 (55·0 - 70·5)	0·0%
Male sex, n (%)	22,618 (51·4)	442,409 (53·7)	0·0%
Caucasian ethnicity, n (%)	42,065 (95·6)	754,756 (91·6)	39·2%
Current Smoking, n (%)	10,843 (24·7)	190,851 (23·2)	66·2%
IMD, median (IQR)	16·2 (9·4 - 27·1)	13·7 (8·3 - 23·4)	0·3%
BMI, kg/m2, median (IQR)	27·4 (23·9 - 31·0)	26·9 (23·6 - 30·4)	60·2%
Creatinine, µmol/L, median (IQR)	102·4 (85·0 - 122·4)	88·7 (73·1 - 105·6)	66·5%
Total cholesterol, mmol/L, median (IQR)	5·3 (4·6 - 6·1)	5·5 (4·8 - 6·3)	72·3%

HF=heart failure. IQR=interquartile range. IMD=index of multiple deprivation. BMI=body mass index.

The number of patients with HF in each general practice was a median of 197 (IQR 128 – 282, range 3 – 622). We explored heterogeneity of case-mix across the included general practices by comparing their distribution of predicted risk according to Model 5. Results in Appendix E indicate that the standard deviation (SD) of the linear predictor (LP) in each general practice ranges between 1.09 and 1.41, and that the mean LP in each general practice ranges between -0.51 and 0.61.

The estimated regression coefficients of the eight prediction models, as obtained from the entire dataset, are presented in Appendix F. These results indicate that all included predictors are significantly associated with HF, and that interactions are present between various predictors. The performance of the estimated models, as evaluated using IECV, is summarized in Table 3.

Discrimination performance

The c-statistic across the general practices is shown in Appendix G. All models showed similar discrimination, although models that included more predictors yielded somewhat larger values for the c-statistic (0.79 in Model 1 – 4 vs. 0.81 in Model 5 – 8). For all models, there was notable between-practice heterogeneity in discrimination performance. For instance, the 95% PI for a Cox regression model including eight predictors as main effects (model 5) ranged from 0.756 to 0.852. Estimates for the between-study standard deviation (τ) were similar for all models, but slightly larger for prediction models that included eight predictors and allowed for non-linear effects and interactions.

Calibration performance

Calibration plot

Calibration plots in Figure 1 indicate that predicted and observed risks were almost in perfect agreement for the unpenalized Cox regression model that included non-linear effects and interactions between predictors (Model 2 and 6).

Predicted and observed risks are almost in perfect agreement for the unpenalized Cox regression models that included non-linear effects and interactions between predictors (Model 2 and 6). Some under-prediction for risk estimates around 10% is observed in the remaining models.

Table 3. Meta-analysis results of the model performance

Model	# RC	Model Estimation	Summary Estimate	95% CI		95% PI		SE	τ	Probability of good performance
Discrimination performance (c-statistic)										
1	4	Cox regression	0·792	0·788	0·796	0·741	0·835	0·012	0·145	95·1
2	14	Cox regression	0·793	0·789	0·797	0·742	0·836	0·012	0·144	95·2
3	4	Ridge penalized Cox	0·793	0·789	0·796	0·742	0·835	0·012	0·144	95·2
4	14	Ridge penalized Cox	0·793	0·789	0·796	0·742	0·836	0·012	0·144	95·2
5	8	Cox regression	0·808	0·804	0·812	0·756	0·852	0·012	0·156	98·4
6	66	Cox regression	0·806	0·802	0·810	0·744	0·856	0·014	0·180	96·4
7	8	Ridge penalized Cox	0·808	0·804	0·812	0·757	0·851	0·012	0·153	98·6
8	66	Ridge penalized Cox	0·809	0·805	0·813	0·754	0·854	0·013	0·163	98·2
Calibration performance (O:E ratio at 5 years)										
1	4	Cox regression	0·957	0·926	0·990	0·598	1·532	0·017	0·239	32·0
2	14	Cox regression	0·963	0·926	1·001	0·557	1·665	0·020	0·279	27·8
3	4	Ridge penalized Cox	0·959	0·928	0·991	0·609	1·511	0·017	0·231	33·0
4	14	Ridge penalized Cox	0·958	0·927	0·990	0·609	1·508	0·017	0·231	33·1
5	8	Cox regression	0·950	0·922	0·977	0·640	1·408	0·015	0·200	37·3
6	66	Cox regression	0·935	0·903	0·969	0·572	1·530	0·018	0·251	30·1
7	8	Ridge penalized Cox	0·947	0·921	0·974	0·648	1·385	0·014	0·193	38·3
8	66	Ridge penalized Cox	0·954	0·928	0·981	0·655	1·389	0·014	0·191	38·3
Calibration performance (calibration slope)										
1	4	Cox regression	1·021	1·005	1·036	0·835	1·206	0·008	0·094	69·7
2	14	Cox regression	1·010	0·992	1·028	0·789	1·231	0·009	0·112	62·3
3	4	Ridge penalized Cox	1·126	1·108	1·143	0·923	1·328	0·009	0·103	38·7
4	14	Ridge penalized Cox	1·088	1·071	1·105	0·888	1·287	0·009	0·101	51·5

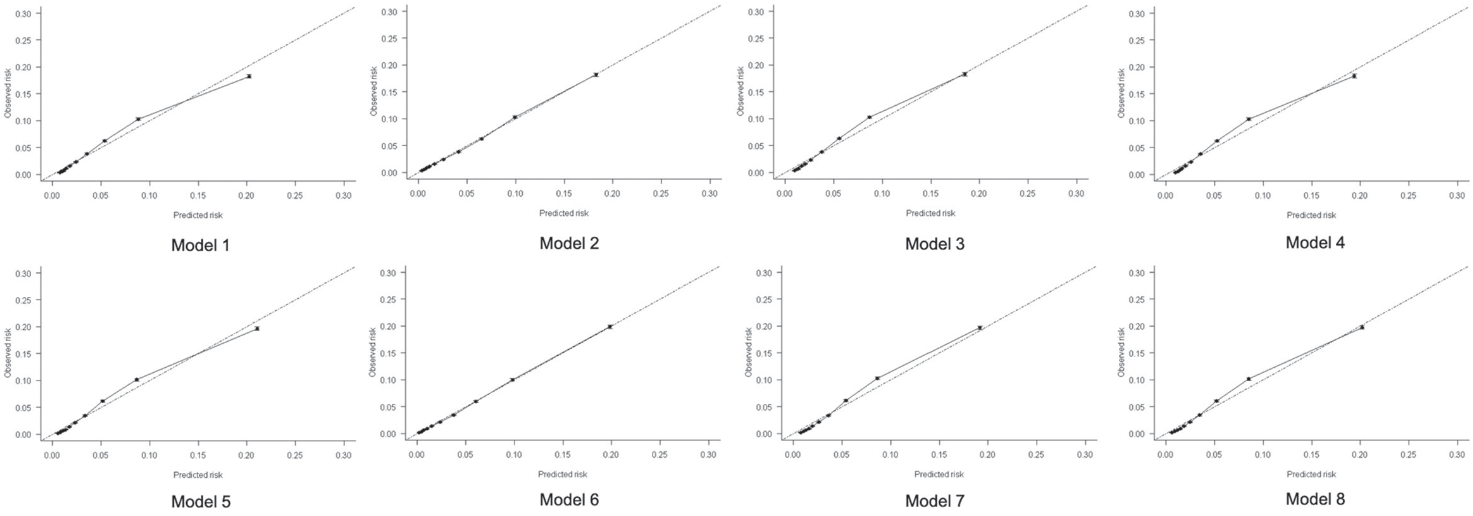
Table 3. (continued)

Model	# RC	Model Estimation	Summary Estimate	95% CI		95% PI		SE	τ	Probability of good performance
5	8	Cox regression	1·023	1·007	1·039	0·833	1·214	0·008	0·097	68·1
6	66	Cox regression	0·992	0·975	1·008	0·792	1·191	0·008	0·101	67·1
7	8	Ridge penalized Cox	1·138	1·120	1·156	0·917	1·358	0·009	0·112	35·1
8	66	Ridge penalized Cox	1·077	1·061	1·092	0·892	1·261	0·008	0·094	35·1

#RC=the total number of estimated regression coefficients. CI=confidence interval. PI=prediction interval. SE=standard error

For all models, SE and between-study heterogeneity (τ) are given on the scale of the meta-analysis (that is, the logit of c-statistic, the log of the O:E ratio and identity for the calibration slope).

Figure 1. Calibration plots of the eight prediction models



O:E ratio

The O:E ratio across the included general practices is shown in Appendix H. All models yielded summary O:E ratios at 5 years below one, especially those models that included eight predictors (Model 5 – 8). In addition, PIs indicate that all prediction models may substantially over- or under-predict the risk of HF when applied to individual patients from a new practice.

Calibration slope

Calibration slope across the included general practices is shown in Appendix I. Unpenalized prediction models yielded pooled calibration slopes most close to one (Model 1, 2, 5, and 6). Prediction models that adopted a ridge penalty yielded calibration slopes that were slightly larger than one, indicating that predicted risks did not vary enough and thus that too much shrinkage may have been applied in the development sample. For all models, the calibration slope was prone to a limited amount of between-practice heterogeneity. For instance, the prediction model that included eight predictors as main effects (model 5) yielded a 95% PI from 0.833 to 1.214. Estimates of between-study variance of the calibration slope were similar for all models.

Case study 2

The detailed results are shown in Appendix D. In short, among 16,280 patients from 14 countries, 2,745 (16.9%) died due to any CVD related conditions. Using bootstrap validation and IECV, we found that the c-statistic ranged from 0.65 to 0.71, and that models with more predictors discriminated better. Results of IECV also indicate that inclusion of non-linear terms and/or interaction effects) did not improve discrimination performance when the model is applied to new patients (from the original to new populations). In calibration performance, the effect of complex modelling strategies was small in both summary estimates of O:E ratio and calibration slope and their generalizability.

Discussion

We illustrated how evidence synthesis methods can be used to evaluate the need of complex strategies for developing generalizable clinical prediction models in large, clustered datasets. To this end, we applied IECV and quantified the model's average performance as well as its variability between clusters. In contrast to traditional internal validation methods, a major advantage of using IECV in large, clustered data is that the external validity of prediction models can be assessed on multiple occasions, thereby allowing researchers to explore the generalizability of different modelling strategies directly during the development process.

In the case study 1, we found that adopting complex modelling strategies did not much improve the external validity of developed prediction models for HF. In particular, prediction models that were based on four commonly available variables yielded a c-statistic of 0.79, which is comparable to existing models for HF using even more than 10 predictors including laboratory tests [187,188]. Although the inclusion of additional predictors marginally improved the discriminative performance, it also slightly increased the between-practice heterogeneity. When investigating model calibration, we found that all prediction modelling strategies yielded adequate calibration performance on average. However, because of between-practice heterogeneity, local revisions were often deemed necessary. In the case study 2, we also found that complex modelling did not meaningfully improve the generalizability of the prediction models, although the inclusion of additional predictors moderately improved their discrimination performance.

As we found in the case study 1, the incremental value of candidate predictors is often small in prediction model studies for the incidence of CVD [201,202]. For instance, systematic reviews have demonstrated a lack of incremental value for cholesterol level [202], BMI [202], and even biomarkers (e.g., triglycerides, C-reactive protein) for predicting CVD [201]. For this reason, it may sometimes be more advantageous to consider the inclusion of non-linear effects or interaction terms, rather than adding more predictors. This strategy is common in machine learning, where methods no longer assume additive linear effects and adopt penalization to avoid overfitting. We mimicked the use of flexible modelling strategies by including non-linear effects and non-linear interaction terms. However, this strategy also failed to improve model discrimination. Similar findings also have been reported in prediction model studies for the prognosis of patients with CVD [203,204]. For instance, a recent study adopting advanced machine learning algorithms failed to outperform traditional prediction models for readmissions in patients with HF, and yielded c-statistics around 0.60 [203]. In another study, discrimination performance to predict all-cause mortality in patients with coronary artery disease marginally increased from 0.793 (Cox regression model with 27 predictors) to 0.797 (random survival forests with 98 predictors) and to 0.801 (elastic net Cox regression model with 586 predictors) [204]. More generally, there is limited evidence that machine learning models can outperform simple prediction models involving additive linear terms, especially when predictions are only based on structured epidemiological data [205].

The following limitations need to be considered. In the first case study, the substantial presence of missing data is an important concern. Although we focused on the inclusion of variables with relatively few missing values, some were missing for more than 70% of participants. Multiple imputation is generally recommended to obtain reliable standard errors of the performance measures but only single imputation was pursued due to limited hardware processing capacity. There is still limited guidance on how to implement multiple imputation when developing and validating a prediction model in large, clustered datasets. Key issues that remain unclear are (i) how to combine multiple imputation with sampling procedures (e.g., IECV) [206,207], (ii) the order of pooling estimates (across imputations or across clusters) [208], (iii) how to ensure congeniality between the imputation model and the prediction model development procedure [209]. Another limitation was that we were not able to include non-linear and interaction terms in the imputation model due to non-convergence issues. Therefore, continuous variables were imputed as a linear term and no interaction term was included in imputation models. This strategy may have favored simpler modelling strategies in IECV. For this reason, we implemented those modelling strategies in the case study 2 where the presence of missing data was much less a concern. And we found similar findings to those in the case study 1.

Second, eligible individuals in both case studies were enrolled more than ten years ago. It is possible that population characteristics have substantially changed over time, and that complex associations (e.g., non-linear predictor effects or interaction terms) have become more common.

Third, we focused on regression-based methods and did not evaluate other flexible modelling strategies such as neural networks or random forests. It is possible that these strategies could yield more promising results, especially if (interaction between) predictor effects cannot adequately be described using the regression-based methods considered here.

Conclusion

We recommend the use of IECV in large, clustered datasets to assess the generalizability of prediction models during their development, and to identify whether complex modelling strategies may offer any advantages. In contrast to traditional internal validation methods, IECV allows to evaluate model performance in non-random hold-out samples with individuals from different settings or populations. In our case studies, we found that accurate prediction does not necessarily require complex modelling strategies, and that the need for local updating may be inevitable regardless of how much data are at hand during the model's development.

CRediT authorship contribution statement

Toshihiko Takada: Conceptualization, Formal analysis, Writing – original draft, Writing – review & editing. Steven Nijman: Formal analysis, Writing – original draft, Writing – review & editing. Spiros Denaxas: Data curation, Project administration, Resources, Writing – review & editing. Kym I.E. Snell: Methodology, Writing – review & editing. Alicia Uijl: Data curation, Writing – review & editing. Tri-Long Nguyen: Methodology, Writing – review & editing. Folkert W. Asselbergs: Project administration, Resources, Writing – review & editing. Thomas P.A. Debray: Conceptualization, Formal analysis, Supervision, Writing – original draft, Writing – review & editing.

Data sharing

Due to privacy laws and the data user agreement between the University College London and Clinical Practice Research Datalink, authors are not authorised to share individual patient data from these electronic health records. Requests to access data provided by Clinical Practice Research Datalink (CPRD) should be sent to the Independent Scientific Advisory Committee (ISAC) (https://www.cprd. com/ISAC/). The CALIBER portal (https://www. caliberresearch.org/portal/) does offer open sharing of phenotypic and analytic algorithms for use by other researchers. Aggregate data (e.g., model formulas, performance estimates) are available on request.

Funding

This work has received support from the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking BigData@Heart grant no. 116074. This project has also received funding from the European Union’s Horizon 2020 research and innovation programme under ReCoDID grant agreement no. 825746.

Acknowledgements

This study is based in part on data from the Clinical Practice Research Datalink obtained under licence from the UK Medicines and Healthcare products Regulatory Agency. The data is provided by patients and collected by the NHS as part of their care and support. The interpretation and conclusions contained in this study are those of the author/s alone. Copyright © (2020), re-used with the permission of The Health & Social Care Information Centre. All rights reserved.

Adapted from Steven W. J. Nijman, Saskia Haitjema, Inge Verkouter, John J.L. Jacobs, Michiel L. Bots, Folkert W. Asselbergs, Karel G.M. Moons, Ines Beekers, Thomas P.A. Debray.

Towards a knowledge base for the use of imputation models in live clinical practice: a pilot study. (Under review).

CHAPTER 8

General discussion

In this thesis, we investigated traditional statistical and modern machine learning (ML) methods for handling of missing predictor data when applying prediction models in real-time medical settings and evaluated how well ML-based prediction model studies follow recommendations from existing reporting guidelines on missing data. The main findings in this thesis are:

- Chapter 2 shows how a majority of the clinical prediction model studies using ML techniques does not report sufficient information on the presence and handling of missing data, despite missing values are highly common in routine healthcare data that often form the basis in ML prediction models studies. Consistent with similar reviews, strategies in which patient records with some missing variables are simply omitted are most often used, even though it is well known this likely causes bias and certainly loss of analytical power.

- Chapter 3 shows that ML-based prediction model studies adhered poorly to the current guideline Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD). Most items considered essential (e.g., about titles and abstract) were not completely addressed in prediction modelling studies. Some items and sub-items of TRIPOD may be less suitable for ML-based models; thus, the TRIPOD guideline requires tailored extensions for ML-based prediction model studies.

- Chapter 4 shows the development of real time imputation methods for missing predictor values using either conditional modelling imputation (CMI, where a multivariable imputation model is derived for each predictor from a population) or joint modelling imputation (JMI, where we use a multivariate normal approximation to generate patient-specific imputations). These newly developed methods were compared with mean imputation (where missing values are replaced by the sample mean) in a case study evaluating the accuracy of the imputed missing predictor values, where we found that JMI and CMI were more accurate.

- Chapter 5 shows how the use of JMI, especially with auxiliary variables (i.e., variables not part of the prediction model), for real-time imputation of missing predictor values is preferred over JMI without auxiliary variables and mean imputation, in terms of the discrimination and calibration of the model predictions.

- Chapter 6 compares various ML modelling techniques that deal with missing predictor values. The use of surrogate splits were found to perform poorly, whilst pattern submodels showed good performance only when paired with a specific modelling technique. Overall, JMI is still to be preferred for both modelling techniques in terms of calibration and discrimination, provided multiple imputations are used.

- Chapter 7 describes how the adoption of internal-external cross-validation (IECV) is preferred to assess the generalizability of prediction models during their development, and to identify whether complex modelling strategies may offer any advantages. Briefly, IECV allows to evaluate model performance in non-random hold-out samples with individuals from different settings or populations. In our case studies, we found that accurate prediction does not necessarily require complex modelling strategies, and that the need for local updating may be inevitable regardless of how much data are at hand during the model's development.

In this final chapter we bring all these findings about current practice, reporting and advancements in the handling of missing predictor data in prediction modelling together, and explore how real-time imputation of missing predictor when using a prediction model in real time practice is perceived by healthcare professionals. We focus on how users of individualized prediction models in daily medical practice feel about imputing missing predictor values as we investigated via a vignette case study. Before that we briefly summarize the principles of risk prediction in daily care and the issue of missing predictor values. We will end this chapter by summarizing our future perspective on using missing data handling strategies for enabling risk prediction in daily medical care.

Risk prediction in daily medical care

Prediction models in routine clinical practice are able to provide actionable information to potentially improve shared clinical decision making in individual patients [22,210–213]. By combining patient, test result and disease characteristics these multivariable risk prediction models provide absolute risk estimates for diagnostic or prognostic purposes to guide further patient management [15,20,26,144–146]. Examples are the SMART risk score and the Framingham risk score [27,126].

Increasingly, with the introduction of build in prediction models in electronic health record (EHR) newly developed and carefully validated clinical prediction models can directly extract any individual's observed predictor value from the EHR and may provide risk-guided recommendations [17,19,109–111]. The actual use of such fully in EHR integrated clinical prediction models is however limited and often frustrated by missing predictor data in the EHR and the inability to real time handle these missing predictor data [19].

Unfortunately, missing predictor data are a hallmark of routine care datasets that are increasingly used for the development, validation, and implementation of prediction models, notably by prediction models based on ML. Consequently approaches for handling missing data (e.g. multiple imputation) in research that aims to develop or validate prediction models, have been developed and are now recommended by multiple reporting and methodological conduct guidelines

[9,24,28,45,214,215]. As it stands, there is limited adherence to these reporting guidelines in prediction model studies (Chapter 2 and Chapter 3) [3,52]. These approaches for handling missing data are, however, not directly suitable for real-time use to impute missing predictor values when using a prediction model in daily clinical practice [6]. Common imputation strategies are notably developed for valid statistical inference of prediction model research, for example, on estimated prediction model coefficients and not for application in single patients. Moreover, these methods typically require access to raw data from multiple individuals, which is unlikely to achieve in daily clinical care given privacy and computational constraints.

Still, imputation of missing predictor values is important to provide for the use of a prediction model in daily care and to provide an individual's prediction. Mean imputation of a missing predictor value has been recommended as a real-time missing predictor data handling strategy, due to its simple applicability in practice and relatively good performance, although it was also found to be insufficient when strong predictors were missing [48,170]. As a result, additional real-time missing data imputation developments, such as joint modelling imputation (JMI), have been made as also evaluated in Chapter 4 and 5. JMI alleviates the issues found with mean imputation as it estimates all associations between the relevant patient characteristics [48]. Briefly, JMI uses a two-step approach: first population characteristics (i.e., means and covariance) are estimated from raw individual patient data and stored in the EHR-system; second the prediction model handles any missing predictor data by drawing imputations using the stored population characteristics. As a consequence JMI is suitable for individual predictions by an EHR built-in prediction model, does not need large amounts of raw data, and can achieve near-real time handling of missing predictor data which makes it attractive for use in real-time model predictions [40,48,175].

Alternatives to imputation of missing predictor values exist and may be more intuitive as they solve the issue of missing data from within the prediction model, rather than via a separate imputation step as described above. In Chapter 6 we evaluated two of such approaches: so-called pattern submodels (PS) in which separate prediction models are developed for each possible missing predictor data pattern, and surrogate splits (SS), in which the original split direction of a tree-based method is preserved as good as possible by means of a surrogate variable [36,37,50]. Compared to PS, surrogate splits seem to perform poorly and are very dependent on the correlations between the surrogate variable and the missing predictor variable. Ultimately, JMI appeared in our research to still be preferred.

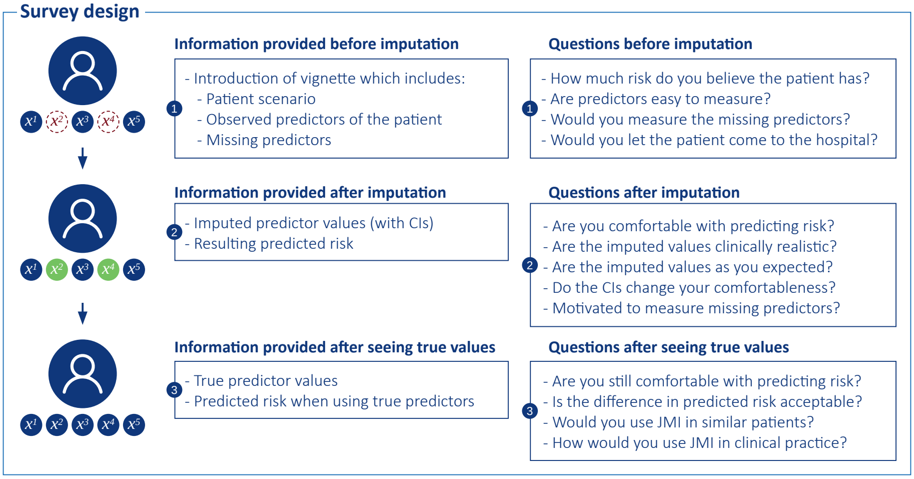
Still, the use of JMI requires careful interpretation by prediction models users such as the healthcare professionals, as imputations may be (very) uncertain. Furthermore, missing predictor imputation in daily care is not widely adopted yet and a valid and reliable assessment of the acceptance of using a prediction model combined with real-time imputation in practice is warranted. To do so, we assessed acceptance of a real-time imputation method by means of a vignette study.

A vignette study to acquire attitude towards using real-time imputation in daily clinical practice

A vignette describes a potential scenario as would occur in real life (e.g., patient scheduled for a consult) [216]. When simulating real-time use of missing predictor data handling, the information provided in each vignette needs to resemble an existing patient, with realistic missing patient characteristics or predictor values (Table 1). The participants – in this case healthcare professionals that use prediction models to guide their patient management - for a vignette study usually consist of those that may experience the potential scenario described.

To simulate the use of an existing prediction model with real-time missing predictor value imputation, we used the SMART risk score and paired it with the available U-Prevent prediction and decision model (figure 2) [21,126]. We approached potential study participants that may use both prediction models (i.e., clinicians) from the departments of cardiology, vascular medicine, or internal medicine at the UMC Utrecht. Our vignettes resembled real world patients as mimicked from the large scale Utrecht Cardiovascular Cohort (UCC) [129]. Further, we presented the participating clinicians with vignettes in fixed order and included separate questionnaires at different points in time of the mimicked clinical process: (i) before missing predictor imputation, (ii) after missing predictor imputation, and (iii) after unveiling the true values of the missing predictors (Figure 1). The questionnaires ultimately asked whether imputed predictor values were **clinically realistic**, the users were **comfortable** with using the imputed values to predict the patient's risk, and whether the imputation model provided **added value** for the clinician (Figure 1).

Figure 1. Overview of survey structure and design



Information provided before imputation

- Introduction of vignette which includes:
  - Patient scenario
  - Observed predictors of the patient
  - Missing predictors

Questions before imputation

- How much risk do you believe the patient has?
- Are predictors easy to measure?
- Would you measure the missing predictors?
- Would you let the patient come to the hospital?

Information provided after imputation

- Imputed predictor values (with CIs)
- Resulting predicted risk

Questions after imputation

- Are you comfortable with predicting risk?
- Are the imputed values clinically realistic?
- Are the imputed values as you expected?
- Do the CIs change your comfortableness?
- Motivated to measure missing predictors?

Information provided after seeing true values

- True predictor values
- Predicted risk when using true predictors

Questions after seeing true values

- Are you still comfortable with predicting risk?
- Is the difference in predicted risk acceptable?
- Would you use JMI in similar patients?
- How would you use JMI in clinical practice?

Legend – CI: confidence interval; JMI: joint modelling imputation.

Between the provided three vignettes (see Table 1) different combinations of predictors or patient characteristics were made missing, based on a combination of *variable types* (i.e., binary and continuous), *burden to retrieve the missing predictor value,* and *expected ease of interpretation of the imputed predictor values* (Table 1). We defined three categories for burden: low (i.e., when a phone call to the patient would suffice to retrieve the missing predictor value), medium (i.e., if the clinician can easily measure the variable with the patient during the physical examination), and high (i.e., when the missing predictor would require some additional, e.g. lab or imaging, test). In short, scenario 1 (table 1) with missing predictor values was the most prevalent and easiest to interpret, scenario 2 the most extreme and scenario 3 the easiest to fix with additional measurements.

With 17 clinicians , of which 13 completed all 9 questionnaires, the vignette study provided an exploratory look at how real world imputation of missing predictor values in clinical practice is perceived.

Overall, the imputed values themselves were perceived as very realistic (Table 2). The type of missing predictor did not influence this perception as both continuous, such as *SBP* (100%), and binary, such as *diabetes* or *anti-thrombotic treatment* (both 77%), predictors were rated similar. Except for *SBP* (46%) and *years since first CVD-event* (29%), the imputed values matched clinical expectations across variable types and levels of clinical burden. When many predictor variables were missing, the difference in predicted risk was perceived as unacceptable (23%) (reflected by scenario 2).

Table 1. Summary of vignettes and type of missing predictor values and their imputed and real values

	Missing predictor values	Variable type	Burden to retrieve the missing data	Imputed values	Real values
Scenario 1	SBP	Continuous	Medium	136	163
	Hs-CRP	Continuous	High	3.2	1.6
Scenario 2	Hs-CRP	Continuous	High	3	2.5
	Years since first CVD-event	Continuous	Low	14.7	4
	Total cholesterol	Continuous	High	5.2	4.1
	HDL-cholesterol	Continuous	High	1.2	1.2
	LDL-cholesterol	Continuous	High	3.3	2.3
Scenario 3	SBP	Continuous	Medium	138	138
	Diabetes	Binary	Low	11.9%	No
	Anti-thrombotic treatment	Binary	Low	86.9%	Yes

Legend – SBP: systolic blood pressure; hs-CRP: high sensitivity C-reactive protein; CVD: cardiovascular disease; LDL: lower-density lipoprotein; HDL: high-density lipoprotein.

The level of comfortableness was, altogether, low. Solely when few, exclusively continuous predictors, were missing, participants were comfortable with imputation of missing predictor values (67% in scenario 1). With too many predictors missing, independent of the burden to retrieve the missing predictor values, few participants were comfortable (29%). Only when predictors were mostly binary (as reflected in scenario 3), the level of comfortableness changed substantially after revealing the true predictor values (from 18% to 54%).

Participants seemed motivated to measure any missing predictor value, regardless of variable type or burden to retrieve the missing predictor value. The one exception was hs-CRP, for which participants were consistently not motivated to measure the missing values (35% and 13% for scenarios 1 and 2 respectively).

The view on comfortableness in predicting a patient's risk after having imputed a missing predictor value, seemed dependent on the type of the missing predictor. Possibly this is because binary predictors are imputed with percentages (e.g., 85% instead of yes/no), rather than a dichotomized imputed value, making the interpretation more difficult. Likewise, CIs for imputed predictor values were found to deteriorate the interpretability of the imputed value and comfortableness in the predicted risk (after imputation) more unrealistic.

These results indicate that the implementation of real time imputation seems better perceived as useful when it is used to impute continuous variables and not too many predictors are missing.

Also, the acceptance of real time imputation of missing predictor values is dependent on the importance of the predictor. Cholesterol levels, for example, were noted as important predictors and participants specifically stated that imputation could not be relied upon as it were part of the minimum set of predictors to be measured in cardiovascular risk prediction [22]. In comparison, hs-CRP was not considered important for deciding on treatment options and thus participants were not concerned when it was imputed. This vignette study thus indicates that there is acceptance by users of prediction models to apply missing predictor value imputation in real time, if not only to justify additional measurements.

One of the questions that remains is whether the use of confidence intervals around the imputed predictor values is helpful. Also, we note that this pilot of course addressed only a limited number of scenarios and clinical domains, which stresses the importance of further study on professionals' acceptance and use by of imputation of missing predictor values in real time.

Future perspectives regarding missing predictor data and their imputation in real time

The use of real-time missing predictor value imputation was found to be acceptable by potential users. Developments in terms of how to implement real time imputation models, which variables or information is to be used by the imputation models and how to present the imputed values and the correspondingly predicted risks, are required to ensure continued acceptableness of real-time imputation in daily practice. For example, an improved way to communicate the uncertainty around the missing predictor value imputations and the subsequently predicted risks by the model is warranted.

Inherently, there is always uncertainty when imputing missing predictor values in real time practice. When more predictor values are missing, the uncertainty around imputed values is evidently higher. Still, it is difficult for users to exactly interpret when an imputed predictor value in real time is too uncertain. As is, the use of confidence intervals seems to primarily cause doubt rather than providing confidence among prediction model users. Generating a rule of thumb for when a confidence interval is too wide or the uncertainty of an imputed predictor value is too high, may be possible but remains complex. Instead, we recommend to present the difference in the resulting model's predicted risk based on the confidence limits of the imputed predictor values.

Similarly, inclusion of auxiliary variables and information (extracted from the EHR) to be used in the real time imputation models has not yet been fully evaluated. There is tremendous opportunity in the typically rich EHRs for improved real-time missing predictor value imputation, which might result in more accurate imputations. Similarly, though unproven, the use of auxiliary variables may improve other missing data handling approaches such as surrogate splits, as discussed in our chapter 5. Still, privacy regulations may limit the use of other available patient data to be used in real time imputation models. Ultimately, to ensure the use of auxiliary variables in real time imputation models is feasible but should be researched further.

Table 2. summary of acceptance measures (shows percentage of participants that said yes)

			Clinical realism			Comfortableness		Added value	
Scenarios	**Missing variables**	Clinical burden	Are the imputed values clinically realistic?	Do the clinically realistic values fall in line with expectations?	Is the difference in predicted risks, between true and imputed predictor values, acceptable?	Are you comfortable with predicting risk with these missing predictor values?	Are you comfortable with predicting risk after seeing true values?	Would you use JMI in similar patients?	Are you motivated to measure missing variables?
Before (i), after (ii), or after seeing true values (iii)			ii	ii	iii	ii	iii	iii	i
Scenario 1	SBP	+/-	77%	46%	80%	67%	60%	87%	71%
	Hs-CRP	-	100%	67%					35%
Scenario 2	Time since first CVD event	+	47%	29%	23%	29%	23%	77%	80%
	Hs-CRP	-	93%	93%					13%
	Total cholesterol	-	93%	86%					93%
	HDL-cholesterol	-	93%	93%					93%
	LDL-cholesterol	-	93%	60%					93%
Scenario 3	Anti-thrombotic medication	+	77%	80%	92%	15%	54%	85%	69%
	Diabetes	+	77%	100%					69%
	SBP	+/-	100%	54%					77%

Legend – SBP: systolic blood pressure; hs-CRP: high-sensitivity C-reactive protein; CVD: cardiovascular; +: low; +/-: medium; -: high.

Likewise, though implied in our vignette study above, it has not yet been fully evaluated whether the intended users of prediction models and underlying real time imputation models, will actually improve the measurement of patient information to reduce the amount of missing data in future patients. The impact that the use of real-time imputation may have on the overall missingness rates in EHRs is yet unknown and more hands-on research in existing risk management systems may show whether this is the case.

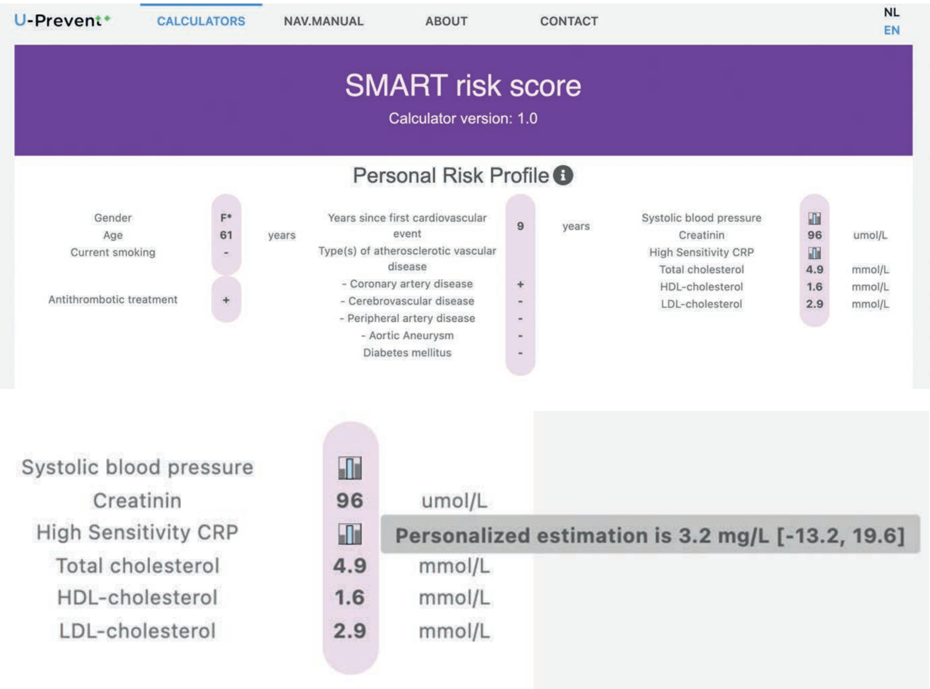
The framework and protocol for adopting real-time imputation in daily practice of healthcare professionals may also be a hurdle. Real-time imputation models need to be developed in or tailored (i.e., calibrated) to a suitable sample of patients from the targeted local population to which the prediction models will be applied. Fortunately, the development of the proposed real-time imputation models is relatively simple, and only requires estimating the means and covariance matrix of a targeted population. Consequently, it would be possible that this data needed for real time imputation, is directly provided from the research that led the development or validation of the prediction model itself. This would make the adoption of real time missing predictor value imputation easier.

The important question remains whether all these suggestions and developments on the use and implementation of real time imputation in daily practice will have a positive impact on clinical decision making and health outcomes in individual patients [212,217–223]. This is the ultimate aim of subsequent research in this area.

Concluding remarks

With the existence of extensive reporting guidelines and missing data handling theory, it remains surprising that missing data in clinical healthcare data continuous to be a persistent problem when developing, validating, or applying clinical prediction models. Omitting or ignoring missing predictor data seem the prevailing situation. Overall, this indicates an overall lack of appreciation about the severe consequences of improper handling of missing data in prediction model research and practice. Whilst improvements in clinical healthcare, such as improved clinical care pathways that minimize missing data, and the use of real time missing data imputation may provide a suitable solution, researchers and users of prediction models must first become more aware of the consequences of ignoring missing data. Otherwise, all improvements and solutions will not follow their implementation in future research or practice. We do believe that the research in this thesis will contribute to this acknowledgment.

Figure 2. Example of a hypothetical risk profile as presented to the clinician



SMART risk score

Calculator version: 1.0

Personal Risk Profile

Gender	F*	
Age	61	years
Current smoking	-	
Antithrombotic treatment	+	
Years since first cardiovascular event	9	years
Type(s) of atherosclerotic vascular disease		
- Coronary artery disease	+	
- Cerebrovascular disease	-	
- Peripheral artery disease	-	
- Aortic Aneurysm	-	
Diabetes mellitus	-	
Systolic blood pressure	📊	
Creatinin	96	umol/L
High Sensitivity CRP	📊	
Total cholesterol	4.9	mmol/L
HDL-cholesterol	1.6	mmol/L
LDL-cholesterol	2.9	mmol/L

Systolic blood pressure	📊	
Creatinin	96	umol/L
High Sensitivity CRP	📊	Personalized estimation is 3.2 mg/L [-13.2, 19.6]
Total cholesterol	4.9	mmol/L
HDL-cholesterol	1.6	mmol/L
LDL-cholesterol	2.9	mmol/L

Bibliography

1. Groenwold RHH, Moons KGM, Vandenbroucke JP. Randomized trials with missing outcome data: how to analyze and what to report. CMAJ. 2014;186(15):1153-1157. doi:10.1503/cmaj.131353
2. Masconi KL, Matsha TE, Echouffo-Tcheugui JB, Erasmus RT, Kengne AP. Reporting and handling of missing data in predictive research for prevalent undiagnosed type 2 diabetes mellitus: a systematic review. *EPMA Journal*. 2015;6(1):7. doi:10.1186/s13167-015-0028-0
3. Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *Journal of Clinical Epidemiology*. 2013;66(3):268-277. doi:10.1016/j.jclinepi.2012.06.020
4. 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. *Canadian Medical Association Journal*. 2012;184(11):1265-1269. doi:10.1503/cmaj.110977
5. Rubin DB. Inference and Missing Data. *Biometrika*. 1976;63(3):581-592. doi:10.2307/2335739
6. Van Buuren S. *Flexible Imputation of Missing Data*. 2nd ed. CRC Press; 2018.
7. Enders CK. *Applied Missing Data Analysis*. Guilford Press; 2010.
8. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*.; 2019. Accessed September 26, 2019. http://public.eblib.com/choice/publicfullrecord.aspx?p=5741221
9. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*. 2006;59(10):1087-1091. doi:10.1016/j.jclinepi.2006.01.014
10. Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *International Journal of Epidemiology*. 2019;48(4):1294-1304. doi:10.1093/ije/dyz032
11. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Med Res Methodol*. 2017;17(1):162. doi:10.1186/s12874-017-0442-1
12. Harel O, Mitchell EM, Perkins NJ, et al. Multiple Imputation for Incomplete Data in Epidemiologic Studies. *American Journal of Epidemiology*. 2018;187(3):576-584. doi:10.1093/aje/kwx349
13. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer; 2009.
14. Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Med Genomics*. 2015;8(1):33. doi:10.1186/s12920-015-0108-y
15. Fröhlich H, Balling R, Beerenwinkel N, et al. From hype to reality: data science enabling personalized medicine. *BMC Med*. 2018;16(1):150. doi:10.1186/s12916-018-1122-7
16. Hoffman MA, Williams MS. Electronic medical records and personalized medicine. *Hum Genet*. 2011;130(1):33-39. doi:10.1007/s00439-011-0992-y
17. Sitapati A, Kim H, Berkovich B, et al. Integrated precision medicine: the role of electronic health records in delivering personalized treatment: Integrated precision medicine. *WIREs Syst Biol Med*. 2017;9(3):e1378. doi:10.1002/wsbm.1378
18. Jiang X, Wells A, Brufsky A, Neapolitan R. A clinical decision support system learned from data to personalize treatment recommendations towards preventing breast cancer metastasis. Jeong J, ed. *PLoS ONE*. 2019;14(3):e0213292. doi:10.1371/journal.pone.0213292
19. Groenhof TKJ, Bots ML, Brandjes M, et al. A computerised decision support system for cardiovascular risk management 'live' in the electronic health record environment: development, validation and implementation—the Utrecht Cardiovascular Cohort Initiative. *Neth Heart J*. 2019;27(9):435-442. doi:10.1007/s12471-019-01308-w

20. Moons KGM, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart*. 2012;98(9):683-690. doi:10.1136/heartjnl-2011-301246

21. Visseren FLJ, Dorresteijn JAN, van der Graaf Y. U-prevent u bent “in control.” Published online 2018. Accessed October 7, 2019. https://www.u-prevent.nl/nl-NL

22. Visseren FLJ, Mach F, Smulders YM, et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. *European Heart Journal*. 2021;42(34):3227-3337. doi:10.1093/eurheartj/ehab484

23. Vandenbroucke JP, Poole C, Schlesselman JJ, Egger M. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. *PLoS Medicine*. 2007;4(10):27.

24. 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. doi:10.1186/s12916-014-0241-z

25. Lee KJ, Tilling K, Cornish RP, et al. Framework for the Treatment And Reporting of Missing data in Observational Studies: The TARMOS framework. *arXiv:200414066 [stat]*. Published online April 29, 2020. Accessed October 6, 2020. http://arxiv.org/abs/2004.14066

26. Grant SW, Collins GS, Nashef SAM. Statistical Primer: developing and validating a risk prediction model†. *European Journal of Cardio-Thoracic Surgery*. 2018;54(2):203-208. doi:10.1093/ejcts/ezy180

27. D’Agostino RB, Vasan RS, Pencina MJ, et al. General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. *Circulation*. 2008;117(6):743-753. doi:10.1161/CIRCULATIONAHA.107.699579

28. Janssen KJM, Vergouwe Y, Donders ART, et al. Dealing with Missing Predictor Values When Applying Clinical Prediction Models. *Clinical Chemistry*. 2009;55(5):994-1001. doi:10.1373/clinchem.2008.115345

29. Awan SE, Sohel F, Sanfilippo FM, Bennamoun M, Dwivedi G. Machine learning in heart failure: ready for prime time. *Current Opinion in Cardiology*. 2018;33(2):190-195. doi:10.1097/HCO.0000000000000491

30. Dorado-Díaz PI, Sampedro-Gómez J, Vicente-Palacios V, Sánchez PL. Applications of Artificial Intelligence in Cardiology. The Future is Already Here. *Revista Española de Cardiología (English Edition)*. 2019;72(12):1065-1075. doi:10.1016/j.rec.2019.05.014

31. Westcott RJ, Tcheng JE. Artificial Intelligence and Machine Learning in Cardiology. *JACC: Cardiovascular Interventions*. 2019;12(14):1312-1314. doi:10.1016/j.jcin.2019.03.026

32. Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *N Engl J Med*. 2018;378(11):981-983. doi:10.1056/NEJMp1714229

33. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthcare Journal*. 2019;(Vol 6, No 2):94-98.

34. Ghassemi M, Naumann T, Schulam P, Chen IY, Ranganath R. A Review of Challenges and Opportunities in Machine Learning for Health. :10.

35. Luo W, Phung D, Tran T, 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. doi:10.2196/jmir.5870

36. Cevallos Valdiviezo H, Van Aelst S. Tree-based prediction on incomplete data using imputation or surrogate decisions. *Information Sciences*. 2015;311:163-181. doi:10.1016/j.ins.2015.03.018

37. Feelders A. Handling Missing Data in Trees: Surrogate Splits or Statistical Imputation? In: Żytkow JM, Rauch J, eds. *Principles of Data Mining and Knowledge Discovery*. Vol 1704. Springer Berlin Heidelberg; 1999:329-334. doi:10.1007/978-3-540-48247-5_38

38. Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and Regression Trees*. Taylor & Francis; 1984. https://books.google.nl/books?id=JwQx-WOmSyQC

39. Hapfelmeier A. *Analysis of Missing Data with Random Forests*.; 2012. Accessed September 4, 2019. https://edoc.ub.uni-muenchen.de/15058/1/Hapfelmeier_Alexander.pdf

40. Hoogland J, Barreveld M, Debray TPA, et al. Handling missing predictor values when validating and applying a prediction model to new patients. *Statistics in Medicine*. Published online July 20, 2020:sim.8682. doi:10.1002/sim.8682

41. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338(jun29 1):b2393-b2393. doi:10.1136/bmj.b2393

42. Moons KGM, Altman DG, Reitsma JB, 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. doi:10.7326/M14-0698

43. 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*. Published online 2012:6.

44. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*.; 2019. Accessed September 26, 2019. http://public.eblib.com/choice/publicfullrecord.aspx?p=5741221

45. Janssen KJM, Donders ART, Harrell FE, et al. Missing covariate data in medical research: To impute is better than to ignore. *Journal of Clinical Epidemiology*. 2010;63(7):721-727. doi:10.1016/j.jclinepi.2009.12.008

46. Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *International Journal of Epidemiology*. 2019;48(4):1294-1304. doi:10.1093/ije/dyz032

47. Nijman S, Groenhof T, Hoogland J, et al. Real-time handling of missing predictor values when implementing and using prediction models in daily practice. *JCE*. 2021;Article in press. doi:https://doi.org/10.1016/j.jclinepi.2021.01.003

48. Nijman SWJ, Hoogland J, Groenhof TKJ, et al. Real-time imputation of missing predictor values in clinical practice. *European Heart Journal - Digital Health*. 2021;2(1):154-164. doi:10.1093/ehjdh/ztaa016

49. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Published online August 13, 2016:785-794. doi:10.1145/2939672.2939785

50. Fletcher Mercaldo S, Blume JD. Missing data and prediction: the pattern submodel. *Biostatistics*. 2020;21(2):236-252. doi:10.1093/biostatistics/kxy040

51. 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. doi:10.1186/s40364-014-0027-7

52. 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. *Journal of Clinical Epidemiology*. Published online September 2021:S0895435621002882. doi:10.1016/j.jclinepi.2021.09.008

53. 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 Research and Clinical Practice*. Published online October 2021:109089. doi:10.1016/j.diabres.2021.109089

54. Dhiman P, Ma J, Navarro CA, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. *Journal of Clinical Epidemiology*. 2021;138:60-72. doi:10.1016/j.jclinepi.2021.06.024

55. 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. doi:10.1186/s12874-015-0022-1

56. 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. doi:10.1186/1471-2288-12-96

57. Andaur Navarro CL, Damen JAA, Takada T, et al. *Completeness of Reporting of Clinical Prediction Models Developed Using Supervised Machine Learning: A Systematic Review*.; 2021. doi:10.1101/2021.06.28.21259089

58. Andaur Navarro CL, Damen JAAG, Takada T, 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.

59. Andaur Navarro CL, Damen JAAG, Takada T, 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. doi:10.1136/bmjopen-2020-038832

60. Mackinnon A. The use and reporting of multiple imputation in medical research - a review: The use and reporting of multiple imputation in medical research. *Journal of Internal Medicine*. 2010;268(6):586-593. doi:10.1111/j.1365-2796.2010.02274.x

61. Knol MJ, Janssen KJM, Donders ART, et al. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *Journal of Clinical Epidemiology*. 2010;63(7):728-736. doi:10.1016/j.jclinepi.2009.08.028

62. Groenwold RHH. Informative missingness in electronic health record systems: the curse of knowing. *Diagn Progn Res*. 2020;4(1):8. doi:10.1186/s41512-020-00077-0

63. 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. *Journal of Clinical Epidemiology*. 2020;125:188-190. doi:10.1016/j.jclinepi.2020.06.007

64. Kappen TH, Vergouwe Y. Adaptation of Clinical Prediction Models for Application in Local Settings. :10.

65. Sperrin M, Martin GP. Multiple imputation with missing indicators as proxies for unmeasured variables: simulation study. *BMC Med Res Methodol*. 2020;20(1):185. doi:10.1186/s12874-020-01068-x

66. Sperrin M, Martin GP, Sisk R, Peek N. Missing data should be handled differently for prediction than for description or causal explanation. *Journal of Clinical Epidemiology*. 2020;125:183-187. doi:10.1016/j.jclinepi.2020.03.028

67. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: What, why, and how? *BMJ (Online)*. 2009;338(7706):1317-1320. doi:10.1136/bmj.b375

68. Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Medicine*. 2013;10(2). doi:10.1371/journal.pmed.1001381

69. Riley, Richard D; van der Windt, Danielle; Croft, Peter; Moons KGM. *Prognosis Research in Health Care: Concepts, Methods, and Impact*. Oxford University Press; 2019. doi:10.1093/med/9780198796619.001.0001

70. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Second. Springer; 2019. doi:10.1007/978-3-030-16399-0

71. Glasziou P, Altman DG, Bossuyt P, et al. Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet*. 2014;383(9913):267-276. doi:10.1016/S0140-6736(13)62228-X

72. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ (Online)*. 2016;353. doi:10.1136/bmj.i2416

73. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. *The BMJ*. 2020;368:1-12. doi:10.1136/bmj.m689

74. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. *American Journal of Epidemiology*. 2019;188(12):2222-2239. doi:10.1093/aje/kwz189

75. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*. 2019;19(1):1-18. doi:10.1186/s12874-019-0681-4

76. Mitchell T. *Machine Learning*. McGraw Hill; 1997.

77. Obermeyer, Ziad MD, Emanuel, Ezekiel J., M.D. PhD. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine*. 2016;375(13):1212-1216. doi:10.1056/NEJMp1606181.Predicting

78. Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning and health systems. *Journal of Global Health*. 2018;8(2). doi:10.7189/jogh.08.020303

79. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Medicine*. 2018;1(1). doi:10.1038/s41746-018-0040-6

80. Shin S, Austin PC, Ross HJ, et al. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. *ESC Heart Failure*. 2021;8(1):106-115. doi:10.1002/ehf2.13073

81. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*. 2019;110:12-22. doi:10.1016/j.jclinepi.2019.02.004

82. Cho SM, Austin PC, Ross HJ, et al. MACHINE LEARNING COMPARED TO CONVENTIONAL STATISTICAL MODELS FOR PREDICTING MYOCARDIAL INFARCTION READMISSION AND MORTALITY: A SYSTEMATIC REVIEW. *Canadian Journal of Cardiology*. Published online March 5, 2021. doi:10.1016/j.cjca.2021.02.020

83. Collins GS, De Groot JA, Dutton S, et al. External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. *BMC Medical Research Methodology*. 2014;14(1):1-11. doi:10.1186/1471-2288-14-40

84. Bouwmeester W, Zuithoff NPA, Mallett S, et al. Reporting and methods in clinical prediction research: A systematic review. *PLoS Medicine*. 2012;9(5). doi:10.1371/journal.pmed.1001221

85. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*. 2019;170(1):51-58. doi:10.7326/M18-1376

86. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Annals of Internal Medicine*. 2019;170(1):W1-W33. doi:10.7326/M18-1377

87. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*. 2009;6(7). doi:10.1371/journal.pmed.1000097

88. Andaur Navarro CL, Damen JAAG, Takada T, 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):1-6. doi:10.1136/bmjopen-2020-038832

89. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*. 2019;170(1):51-58. doi:10.7326/M18-1376

90. Ploeg T Van Der, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry : a simulation study for predicting dichotomous endpoints. Published online 2014:1-13.

91. Riley RD, Ensor J, Snell KIE. Calculating the sample size required for developing a clinical prediction model. 2020;441(March):1-12. doi:10.1136/bmj.m441

92. Courvoisier DS, Combescure C, Agoritsas T. Performance of logistic regression modeling : beyond the number of events per variable , the role of data structure. 2021;64(2011):993-1000. doi:10.1016/j.jclinepi.2010.11.012

93. Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. *Journal of Clinical Epidemiology*. 2016;76:175-182. doi:10.1016/j.jclinepi.2016.02.031

94. Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Statistical Methods in Medical Research*. 2017;26(2):796-808. doi:10.1177/0962280214558972

95. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ (Online)*. 2009;339(7713):157-160. doi:10.1136/bmj.b2393

96. Vergouwe Y, Royston P, Moons KGM, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. *Journal of Clinical Epidemiology*. 2010;63(2):205-214. doi:10.1016/j.jclinepi.2009.03.017

97. Groenwold RHH. Informative missingness in electronic health record systems: the curse of knowing. *Diagnostic and Prognostic Research*. 2020;4(1):4-9. doi:10.1186/s41512-020-00077-0

98. Van Calster B, McLernon DJ, Van Smeden M, et al. Calibration: The Achilles heel of predictive analytics. *BMC Medicine*. 2019;17(1):1-7. doi:10.1186/s12916-019-1466-7

99. Kareemi H, Vaillancourt C, Rosenberg H, Fournier K, Yadav K. Machine Learning Versus Usual Care for Diagnostic and Prognostic Prediction in the Emergency Department: A Systematic Review. *Academic Emergency Medicine*. Published online 2020:1-13. doi:10.1111/acem.14190

100. Miles J, Turner J, Jacques R, Williams J, Mason S. Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review. *Diagnostic and Prognostic Research*. 2020;4(1):1-12. doi:10.1186/s41512-020-00084-1

101. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *The BMJ*. 2020;369. doi:10.1136/bmj.m1328

102. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *The Lancet*. 2019;393(10181):1577-1579. doi:10.1016/S0140-6736(19)30037-6

103. GS C, P D, CL AN, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ open*. 2021;11(7):e048008. doi:10.1136/BMJOPEN-2020-048008

104. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Annals of Internal Medicine*. 2015;162(1):W1-W73. doi:10.7326/M14-0698

105. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Annals of Internal Medicine*. 2015;162(1):55. doi:10.7326/M14-0697

106. Piepoli MF, Hoes AW, Agewall S, et al. 2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts)Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). *Eur Heart J*. 2016;37(29):2315-2381. doi:10.1093/eurheartj/ehw106

107. Hoffman MA, Williams MS. Electronic medical records and personalized medicine. *Hum Genet*. 2011;130(1):33-39. doi:10.1007/s00439-011-0992-y

108. Ginsburg G. Personalized medicine: revolutionizing drug discovery and patient care. *Trends in Biotechnology*. 2001;19(12):491-496. doi:10.1016/S0167-7799(01)01814-5

109. Groenhof TKJ, Groenwold RHH, Grobbee DE, Visseren FLJ, Bots ML, on behalf of the UCC-SMART study group. The effect of computerized decision support systems on cardiovascular risk factors: a systematic review and meta-analysis. *BMC Med Inform Decis Mak*. 2019;19(1):108. doi:10.1186/s12911-019-0824-x

110. Bezemer T, de Groot MC, Blasse E, et al. A Human(e) Factor in Clinical Decision Support Systems. *J Med Internet Res*. 2019;21(3):e11732. doi:10.2196/11732

111. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ*. 2005;330(7494):765. doi:10.1136/bmj.38398.500764.8F

112. Kotseva K, Wood D, De Bacquer D, et al. EUROASPIRE IV: A European Society of Cardiology survey on the lifestyle, risk factor and therapeutic management of coronary patients from 24 European countries. *Eur J Prev Cardiolog*. 2016;23(6):636-648. doi:10.1177/2047487315569401

113. Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. *Nat Rev Cardiol*. 2016;13(6):350-359. doi:10.1038/nrcardio.2016.42

114. Perkins NJ, Cole SR, Harel O, et al. Principled Approaches to Missing Data in Epidemiologic Studies. *American Journal of Epidemiology*. 2018;187(3):568-575. doi:10.1093/aje/kwx348

115. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work?: Multiple imputation by chained equations. *Int J Methods Psychiatr Res*. 2011;20(1):40-49. doi:10.1002/mpr.329

116. Van Buuren S. *Flexible Imputation of Missing Data*. 2nd ed. CRC Press; 2018.

117. Hoogland J, van Barreveld M, Debray TPA, et al. Handling missing predictor values when validating and applying a prediction model to new patients. *Statistics in Medicine*. Published online 2020:17. doi:10.1002/sim.8682

118. Berkelmans G. *Dealing with Missing Patient Characteristics When Using Cardiovascular Prediction Models in Clinical Practice.*; 2018.

119. Janssen KJM, Vergouwe Y, Donders ART, et al. Dealing with Missing Predictor Values When Applying Clinical Prediction Models. :8.

120. Carpenter JR, Kenward MG. *Multiple Imputation and Its Application*. Wiley; 2013.

121. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*. 2006;59(10):1087-1091. doi:10.1016/j.jclinepi.2006.01.014

122. Gökçay D, Eken A, Baltacı S. Binary Classification Using Neural and Clinical Features: An Application in Fibromyalgia with Likelihood based Decision Level Fusion. :10.

123. Debédat J, Sokolovska N, Coupaye M, et al. Long-term Relapse of Type 2 Diabetes After Roux-en-Y Gastric Bypass: Prediction and Clinical Relevance. *Dia Care*. 2018;41(10):2086-2095. doi:10.2337/dc18-0567

124. Chen R, Stewart WF, Sun J, Ng K, Yan X. Recurrent Neural Networks for Early Detection of Heart Failure From Longitudinal Electronic Health Record Data. *Circulation: Cardiovascular Quality and Outcomes*. Published online 2019:15.

125. Nederlands Huisartsen Genootschap. *Multidisciplinaire Richtlijn Cardiovasculair Risicomanagement*. Bohn Stafleu van Loghum; 2011.

126. Dorresteijn JAN, Visseren FLJ, Wassink AMJ, et al. Development and validation of a prediction rule for recurrent vascular events based on a cohort study of patients with arterial disease: the SMART risk score. *Heart*. 2013;99(12):866-872. doi:10.1136/heartjnl-2013-303640

127. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*. Published online May 16, 2016:i2416. doi:10.1136/bmj.i2416

128. Visseren FLJ, Dorresteijn JAN, van der Graaf Y. U-prevent u bent "in control." Published online 2018. Accessed October 7, 2019. https://www.u-prevent.nl/nl-NL

129. Asselbergs FW, Visseren FL, Bots ML, et al. Uniform data collection in routine clinical practice in cardiovascular patients for optimal care, quality control and research: The Utrecht Cardiovascular Cohort. *Eur J Prev Cardiolog*. 2017;24(8):840-847. doi:10.1177/2047487317690284

130. Nederlands Huisartsen Genootschap. *Multidisciplinaire Richtlijn Cardiovasculair Risicomanagement*. Bohn Stafleu van Loghum; 2011.

131. Kowarik A, Templ M. Imputation with the R Package VIM. *J Stat Soft*. 2016;74(7). doi:10.18637/jss.v074.i07

132. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statist Med*. 2006;25(24):4279-4292. doi:10.1002/sim.2673

133. Dorresteijn JAN, Visseren FLJ, Wassink AMJ, et al. Development and validation of a prediction rule for recurrent vascular events based on a cohort study of patients with arterial disease: the SMART risk score. *Heart*. 2013;99(12):866-872. doi:10.1136/heartjnl-2013-303640

134. Hughes RA, White IR, Seaman SR, Carpenter JR, Tilling K, Sterne JA. Joint modelling rationale for chained equations. *BMC Med Res Methodol*. 2014;14(1):28. doi:10.1186/1471-2288-14-28

135. Demirtas H, Hedeker D. An imputation strategy for incomplete longitudinal ordinal data. *Statist Med*. 2008;27(20):4086-4093. doi:10.1002/sim.3239

136. Murray JS. Multiple Imputation: A Review of Practical and Theoretical Findings. *arXiv:180104058 [stat]*. Published online January 12, 2018. Accessed September 26, 2019. http://arxiv.org/abs/1801.04058

137. Vergouwe Y, Nieboer D, Oostenbrink R, et al. A closed testing procedure to select an appropriate method for updating prediction models: Method selection to update a prediction model. *Statist Med*. 2017;36(28):4529-4539. doi:10.1002/sim.7179

138. Varadhan R. condMVNorm: Conditional Multivariate Normal Distribution. 2015;R package version 2015.2-1. http://CRAN.R-project.org/package=condMVNorm

139. Burgette LF, Reiter JP. Multiple Imputation for Missing Data via Sequential Regression Trees. *American Journal of Epidemiology*. 2010;172(9):1070-1076. doi:10.1093/aje/kwq260

140. Hoogland J, van Barreveld M, Debray T, et al. Handling Missing Predictor Values When Validating and Applying a Prediction Model to New Patients. *Statistics in Medicine*. 2019;(Under review).

141. Chen MH. *Monte Carlo Methods in Bayesian Computation*. Springer; 2013.

142. Genz A, Bretz F, Miwa T, et al. mvtnorm: Multivariate Normal and t Distributions. *Journal of Statistical Software*. Published online 2018. http://CRAN.R-project.org/package=mvtnorm

143. Piepoli MF, Hoes AW, Agewall S, et al. 2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts)Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). *Eur Heart J*. 2016;37(29):2315-2381. doi:10.1093/eurheartj/ehw106

144. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European Heart Journal*. 2014;35(29):1925-1931. doi:10.1093/eurheartj/ehu207

145. Riley RD, van der Wind D, Croft P, Moons KGM. *Prognosis Research in Health Care: Concepts, Methods, and Impact*. 1st ed. Oxford University Press; 2019.

146. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ*. 2009;338. doi:10.1136/bmj.b375

147. Kengne AP. The ADVANCE cardiovascular risk model and current strategies for cardiovascular disease risk evaluation in people with diabetes. *CVJA*. 2013;24(9):376-381. doi:10.5830/CVJA-2013-078

148. Stam-Slob MC, Visseren FLJ, Wouter Jukema J, et al. Personalized absolute benefit of statin treatment for primary or secondary prevention of vascular disease in individual elderly patients. *Clin Res Cardiol*. 2017;106(1):58-68. doi:10.1007/s00392-016-1023-8

149. Riley RD, Ensor J, Snell KIE, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ*. Published online June 22, 2016:i3140. doi:10.1136/bmj.i3140

150. Hulsen T, Jamuar SS, Moody AR, et al. From Big Data to Precision Medicine. *Front Med*. 2019;6:34. doi:10.3389/fmed.2019.00034

151. Cook JA, Collins GS. The rise of big clinical databases: Big clinical databases. *Br J Surg*. 2015;102(2):e93-e101. doi:10.1002/bjs.9723

152. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc*. 2017;24(1):198-208. doi:10.1093/jamia/ocw042

153. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research: *Medical Care*. 2013;51:S30-S37. doi:10.1097/MLR.0b013e31829b1dbd

154. Wells BJ, Nowacki AS, Chagin K, Kattan MW. Strategies for Handling Missing Data in Electronic Health Record Derived Data. *eGEMs*. 2013;1(3):7. doi:10.13063/2327-9214.1035

155. Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. *Nat Rev Cardiol*. 2016;13(6):350-359. doi:10.1038/nrcardio.2016.42

156. Berkelmans GFN, Visseren F, van der Graaf Y, Dorresteijn J, Utrecht U. Lifetime predictions for individualized vascular disease prevention : Whom and when to treat? Published online 2018. https://dspace.library.uu.nl/handle/1874/372514

157. Simons PCG, Algra A. Second Manifestations of ARTerial disease (SMART) study: Rationale and design. *Eur J Clin Epi*.:9.

158. Gokcay D, Eken A, Baltaci S. Binary Classification Using Neural and Clinical Features: An Application in Fibromyalgia With Likelihood-Based Decision Level Fusion. *IEEE J Biomed Health Inform*. 2019;23(4):1490-1498. doi:10.1109/JBHI.2018.2844300

159. Debédat J, Sokolovska N, Coupaye M, et al. Long-term Relapse of Type 2 Diabetes After Roux-en-Y Gastric Bypass: Prediction and Clinical Relevance. *Dia Care*. 2018;41(10):2086-2095. doi:10.2337/dc18-0567

160. Chen R, Stewart WF, Sun J, Ng K, Yan X. Recurrent Neural Networks for Early Detection of Heart Failure From Longitudinal Electronic Health Record Data: Implications for Temporal Modeling With Respect to Time Before Diagnosis, Data Density, Data Quantity, and Data Type. *Circ: Cardiovascular Quality and Outcomes*. 2019;12(10). doi:10.1161/CIRCOUTCOMES.118.005114

161. Quartagno M, Carpenter JR. Multiple imputation for discrete data: Evaluation of the joint latent normal model. *Biometrical Journal*. Published online March 14, 2019:bimj.201800222. doi:10.1002/bimj.201800222

162. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*. 2001;6(4):330-351. doi:10.1037/1082-989X.6.4.330

163. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of Clinical Epidemiology*. 2015;68(3):279-289. doi:10.1016/j.jclinepi.2014.06.018

164. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol*. 2013;13(1):33. doi:10.1186/1471-2288-13-33

165. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Med Decis Making*. 2006;26(6):565-574. doi:10.1177/0272989X06295361

166. Harrell , FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer International Publishing; 2015. doi:10.1007/978-3-319-19425-7

167. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. *Epidemiology*. 2010;21(1):128-138. doi:10.1097/EDE.0b013e3181c30fb2

168. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res*. 2019;3(1):18. doi:10.1186/s41512-019-0064-7

169. Hansen N. The CMA Evolution Strategy: A Tutorial. *arXiv:160400772 [cs, stat]*. Published online April 4, 2016. Accessed May 7, 2020. http://arxiv.org/abs/1604.00772

170. Berkelmans GFN, Read SH, Gudbjörnsdottir S, et al. Dealing with missing patient characteristics when using cardiovascular prediction models in clinical practice. 2018;(Under review).

171. Jolani S, Debray TPA, Koffijberg H, van Buuren S, Moons KGM. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE: S. JOLANI *ET AL* . *Statist Med*. 2015;34(11):1841-1863. doi:10.1002/sim.6451

172. Pattern-Mixture Models for Multivariate Incomplete Data. Published online 2021:11.

173. Vries BBLP de, Smeden M van, Groenwold RHH. Propensity Score Estimation Using Classification and Regression Trees in the Presence of Missing Covariate Data. *Epidemiologic Methods*. 2018;7(1):20170020. doi:doi:10.1515/em-2017-0020

174. Twala B. AN EMPIRICAL COMPARISON OF TECHNIQUES FOR HANDLING INCOMPLETE DATA USING DECISION TREES. *Applied Artificial Intelligence*. 2009;23(5):373-405. doi:10.1080/08839510902872223

175. Nijman SWJ, Groenhof TKJ, Hoogland J, et al. Real-time imputation of missing predictor values improved the application of prediction models in daily practice. *Journal of Clinical Epidemiology*. 2021;134:22-34. doi:10.1016/j.jclinepi.2021.01.003

176. Glynn RJ, Laird NM, Rubin DB. Selection Modeling Versus Mixture Modeling with Nonignorable Nonresponse. In: Wainer H, ed. *Drawing Inferences from Self-Selected Samples*. Springer New York; 1986:115-142. doi:10.1007/978-1-4612-4976-4_10

177. Schouten RM, Lugtig P, Vink G. Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*. 2018;88(15):2909-2930. doi:10.1080/00949655.2018.1491577

178. van Smeden M, Moons KG, de Groot JA, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. *Stat Methods Med Res*. 2019;28(8):2455-2474. doi:10.1177/0962280218784726

179. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak*. 2019;19(1):281. doi:10.1186/s12911-019-1004-8

180. REILLY BM, EVANS AT. Translating clinical research into clinical practice : Impact of using prediction rules to make decisions. *Annals of internal medicine*. 2006;144(3):201-209.

181. Debray TPA, Moons KGM, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Statistics in medicine*. 2013;32(18):3158-3180.

182. Debray TP, Riley RD, Rovers MM, Reitsma JB, Moons KG. Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use. *PLoS medicine*. 2015;12(10):e1001886-e1001886.

183. Debray TPA, de Jong VMT, Moons KGM, Riley RD. Evidence synthesis in prognosis research. *Diagnostic and prognostic research*. 2019;3(1):13-13.

184. Ahmed I, Debray TPA, Moons KGM, Riley RD. Developing and validating risk prediction models in an individual participant data meta-analysis. *BMC medical research methodology*. 2014;14(1):3-3.

185. Debray TP, Damen JA, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Statistical methods in medical research*. 2019;28(9):2768-2786.

186. Snell KIE, Hua H, Debray TPA, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *Journal of clinical epidemiology*. 2016;69:40-50.

187. Agarwal SK, Chambless LE, Ballantyne CM, et al. Prediction of incident heart failure in general practice: the Atherosclerosis Risk in Communities (ARIC) Study. *Circulation Heart failure*. 2012;5(4):422-429.

188. Hippisley-Cox J, Coupland C. Development and validation of risk prediction equations to estimate future risk of heart failure in patients with diabetes: a prospective cohort study. *BMJ open*. 2015;5(9):e008503-e008503.

189. Smith M J Gustav, Newton-Cheh M Christopher, MD, Almgren Ms Peter, et al. Assessment of Conventional Cardiovascular Risk Factors and Multiple Biomarkers for the Prediction of Incident Heart Failure and Atrial Fibrillation. *Journal of the American College of Cardiology*. 2010;56(21):1712-1719.

190. Denaxas S, Gonzalez-Izquierdo A, Direk K, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. *Journal of the American Medical Informatics Association : JAMIA*. 2019;26(12):1545-1559.

191. DENAXAS SC, GEORGE J, HERRETT E, et al. Data Resource Profile: Cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). *International journal of epidemiology*. 2012;41(6):1625-1638.

192. Uijl A, Koudstaal S, Direk K, et al. Risk factors for incident heart failure in age- and sex-specific strata : a population-based cohort using linked electronic health records. *European journal of heart failure*. 2019;21(10):1197-1206.

193. Yang H, Negishi K, Otahal P, Marwick TH. Clinical prediction of incident heart failure risk: a systematic review and meta-analysis. *Open heart*. 2015;2(1):e000222-e000222.

194. The English Indices of Deprivation 2019. Accessed October 19, 2020. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019

195. CALIBER. https://www.ucl.ac.uk/health-informatics/caliber

196. Resche-Rigon M, White IR. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical methods in medical research*. 2018;27(6):1634-1649.

197. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of statistical software*. 2011;39(5):1-13.

198. Debray TPA, Damen JAAG, Snell KIE, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ*. 2017;356:i6460-i6460.

199. Snell KI, Ensor J, Debray TP, Moons KG, Riley RD. Meta-analysis of prediction model performance across multiple studies : Which scale helps ensure between-study normality for the C-statistic and calibration measures? *Statistical methods in medical research*. 2018;27(11):3505-3522.

200. Sandercock PAG, Niewada M, Członkowska A. The International Stroke Trial database. *Trials*. 2011;12:101-101.

201. Ioannidis JPA, Tzoulaki I. Minimal and Null Predictive Effects for the Most Popular Blood Biomarkers of Cardiovascular Disease. *Circulation research*. 2012;110(5):658-662.

202. van Bussel EF, Hoevenaar-Blom MP, Poortvliet RKE, et al. Predictive value of traditional risk factors for cardiovascular disease in older people: A systematic review. *Preventive medicine*. 2020;132:105986-105986.

203. Frizzell JD, Liang L, Schulte PJ, et al. Prediction of 30-Day All-Cause Readmissions in Patients Hospitalized for Heart Failure: Comparison of Machine Learning and Other Statistical Approaches. *JAMA cardiology*. 2017;2(2):204-209.

204. Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PloS one*. 2018;13(8):e0202344-e0202344.

205. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology*. 2019;110:12-22.

206. Wood AM, Royston P, White IR. The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. *Biometrical journal*. 2015;57(4):614-632.

207. Mertens BJA, Banzato E, de Wreede LC. Construction and assessment of prediction rules for binary outcome in the presence of missing predictor data using multiple imputation and cross-validation: Methodological approach and data-based evaluation. *Biometrical journal*. 2020;62(3):724-741.

208. Burgess S, White IR, Resche-Rigon M, Wood AM. Combining multiple imputation and meta-analysis with individual participant data. *Statistics in medicine*. 2013;32(26):4499-4514.

209. Morris TP, White IR, Carpenter JR, Stanworth SJ, Royston P. Combining fractional polynomial model building with multiple imputation. *Stat Med*. 2015;34(25):3298-3317. doi:10.1002/sim.6553

210. Fraccaro P, van der Veer S, Brown B, et al. An external validation of models to predict the onset of chronic kidney disease using population-based electronic health records from Salford, UK. *BMC Med*. 2016;14(1):104. doi:10.1186/s12916-016-0650-2

211. van Rijn MHC, van de Luijtgaarden M, van Zuilen AD, et al. Prognostic models for chronic kidney disease: a systematic review and external validation. *Nephrology Dialysis Transplantation*. 2021;36(10):1837-1850. doi:10.1093/ndt/gfaa155

212. Kappen TH, van Klei WA, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, Moons KGM. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagn Progn Res*. 2018;2(1):11. doi:10.1186/s41512-018-0033-6

213. Ratna MB, Bhattacharya S, Abdulrahim B, McLernon DJ. A systematic review of the quality of clinical prediction models in in vitro fertilisation. *Human Reproduction*. 2020;35(1):100-116. doi:10.1093/humrep/dez258

214. Allotey PA, Harel O. Multiple Imputation for Incomplete Data in Environmental Epidemiology Research. *Curr Envir Health Rpt*. 2019;6(2):62-71. doi:10.1007/s40572-019-00230-y

215. Carpenter JR, Kenward MG. *Multiple Imputation and Its Application*. 1st ed. John Wiley & Sons; 2013.

216. Evans SC, Roberts MC, Keeley JW, et al. Vignette methodologies for studying clinicians' decision-making: Validity, utility, and application in ICD-11 field studies. *International Journal of Clinical and Health Psychology*. 2015;15(2):160-170. doi:10.1016/j.ijchp.2014.12.001

217. Goehler A, Moore C, Manne-Goehler JM, et al. Clinical Decision Support for Ordering CTA-PE Studies in the Emergency Department—A Pilot on Feasibility and Clinical Impact in a Tertiary Medical Center. *Academic Radiology*. 2019;26(8):1077-1083. doi:10.1016/j.acra.2018.09.009

218. Gavrielides MA, Miller M, Hagemann IS, et al. Clinical Decision Support for Ovarian Carcinoma Subtype Classification: A Pilot Observer Study With Pathology Trainees. *Archives of Pathology & Laboratory Medicine*. 2020;144(7):869-877. doi:10.5858/arpa.2019-0390-OA

219. Sim LLW, Ban KHK, Tan TW, Sethi SK, Loh TP. Development of a clinical decision support system for diabetes care: A pilot study. Cras-Méneur C, ed. *PLoS ONE*. 2017;12(2):e0173021. doi:10.1371/journal.pone.0173021

220. Keenan GM, Lopez KD, Yao Y, et al. Toward Meaningful Care Plan Clinical Decision Support: Feasibility and Effects of a Simulated Pilot Study. *Nursing Research*. 2017;66(5):388-398. doi:10.1097/NNR.0000000000000234

221. Gudmundsson HT, Hansen KE, Halldorsson BV, Ludviksson BR, Gudbjornsson B. Clinical decision support system for the management of osteoporosis compared to NOGG guidelines and an osteology specialist: a validation pilot study. *BMC Med Inform Decis Mak*. 2019;19(1):27. doi:10.1186/s12911-019-0749-4

222. Menon S, Tarrago R, Carlin K, Wu H, Yonekawa K. Impact of integrated clinical decision support systems in the management of pediatric acute kidney injury: a pilot study. *Pediatr Res*. 2021;89(5):1164-1170. doi:10.1038/s41390-020-1046-8

223. Topaz M, Trifilio M, Maloney D, Bar-Bachar O, Bowles KH. Improving patient prioritization during hospital-homecare transition: A pilot study of a clinical decision support tool. *Res Nurs Health*. 2018;41(5):440-447. doi:10.1002/nur.21907

APPENDIX

Summary

Samenvatting

Dankwoord

Curriculum Vitae

Summary

The identification of individual patients at risk of disease has become an integral part of recent trends towards a more personalized healthcare system. A healthcare system that is personalized allows us to administer the most applicable treatment to an individual patient given their risk profile and, in turn, make our healthcare much more efficient. To that end, clinical prediction models are situated as prime candidates to assist clinicians with accurate risk estimates. By harnessing the information captured in various patient or disease related properties, these risk prediction models are able to chart a likely path that a disease might take (i.e., prognosis) or identify whether a specific disease is likely present in individual patients (i.e., diagnosis).

Recent efforts to computerize the use of various clinical prediction models in clinical practice have provided clinical decision support systems (CDSS) that are already usable in clinical practice. These CDSS already allow clinicians to potentially inform their clinical decision making by providing individual risk probabilities. However, because currently available risk prediction models require complete information to generate predictions, these models are severely hampered whenever any patient or disease properties are missing. Luckily, the ample guidance that exists on the handling of missing data provides useful stepping stones to develop flexible or missing data handling techniques usable in real-time clinical practice.

In **Chapter 2** we show that, so far, the majority of clinical prediction model studies that make use of machine learning (ML) techniques are not reporting enough information on the presence or handling of missing data when developing or validating a prediction model. Though ill-advised, the removal of patient records with missing variables is also used most often. These results were retrieved by evaluating whether a systematically searched subset of published papers included information on predefined features to be reported about missing data.

In addition to poor reporting on missing data, we show that the adherence of ML prediction model studies to current recommended reporting guidelines is also poor (**Chapter 3**). Several of the items deemed essential were reported incomplete, resulting in a heightened risk of bias for these studies. In addition, methodological quality was generally poor.

It is clear efforts are required to improve the design and consecutive reporting of prediction model studies (using ML or not). To that end, **Chapter 4** presents the development of several imputation methods for missing predictor values in real-time. In a case-study with a real-world empirical data set for cardiovascular risk prediction, we compared the accuracy of two common imputation methods which were adjusted for use in real time clinical practice: conditional

modeling imputation (CMI, where for each predictor a separate multivariable imputation model is derived) and joint modeling imputation (JMI, where we assume all predictors are normally distributed and use the observed patient information to generate imputations for each missing predictor). We then compare these methodologies with a method which is often used in practice: mean imputation (where missing values are replaced by the sample mean). Congruent with our expectations, simulations found that both JMI and CMI are generally to be recommended in terms of imputation accuracy. As JMI was generally faster and less complex, it was deemed more promising.

In Chapter 4 we evaluated novel imputation methods strictly on their imputation accuracy in terms of their root mean squared error. In **Chapter 5** we continue with the more promising imputation method (i.e., JMI) and evaluate it using common evaluation methods for prediction models (i.e., discrimination and calibration of the model predictions). We specifically focus on the use of auxiliary variables (i.e., variables not part of the prediction model), elaborate further on the idea of imputation model updating and make a comparison with the often-used method mean imputation. In summary, the use of JMI is found to be most beneficial when estimated in local data and with the use of these auxiliary variables. Its added value is most prominent whenever the missing predictors are correlated with other observed (auxiliary) variables.

The solution to missing data in clinical practice is not solely solvable by estimating substitute predictor values based on what we know of the individual patient. Multiple techniques exist which can handle missing values with a built-in design. In **Chapter 6** we evaluated multiple missing data handling methods and compared them with JMI. Specifically, we evaluated pattern submodels (PS, where for each pattern, by which variables are missing, a separate prediction model is developed) and surrogate splits (SS, where an optimal replacement value is found among the available patient information which can serve as a replacement for the missing predictor). Provided multiple imputations are used, JMI is still to be preferred over PS and SS.

We are hopeful that large, local datasets may become more available to inform proper imputation procedures which will enable real-time handling of missing data. Still, prediction models need to be generalizable to such data. In **Chapter 7** we show that internal-external cross-validation (IECV) is to be preferred, when the data is clustered, for assessing the generalizability of a prediction model during development. In short, IECV evaluated model performance in every hold-out sample which includes individuals from a different setting or population (e.g., a different hospital). Additionally, it can be adopted to evaluate whether complex modeling strategies (e.g., the use of penalization, interactions or non-linear effects) offer any benefits. We found that the accuracy of prediction models does not necessarily benefit from more complex modeling strategies, which shows that IECV is potentially useful for simplifying model complexity.

As of yet, it is largely uncertain whether personalized medicine, in the form of CDSS, will offer the benefits it gives the impression of providing. Clinicians are certainly inclined to believe so, but concrete evidence of positive impact on health outcomes is, as of yet, missing. First and foremost, and for fair comparison, the severe consequences of improper missing data handling must be appreciated and handled the right way.

Samenvatting

De identificatie van individuele patiënten met een risico op ziekte is een integraal onderdeel geworden van recente trends in de richting van een meer gepersonaliseerd gezondheidszorgsysteem. Een zorgsysteem dat gepersonaliseerd is stelt ons in staat om de meest toepasselijke behandeling te vinden voor een individuele patiënt gegeven diens risicoprofiel. Die persoonlijke benadering zou onze gezondheidszorg veel efficiënter kunnen maken. Daartoe zijn klinische voorspelmodellen de voornaamste kandidaten om clinici te helpen met nauwkeurige risicoschattingen. Door gebruik te maken van de informatie die is vastgelegd in verschillende patiënt- of ziektegerelateerde eigenschappen, kunnen deze risico-voorspelmodellen een waarschijnlijk pad in kaart brengen dat een ziekte zou kunnen nemen (d.w.z.,, prognose) of identificeren of een specifieke ziekte waarschijnlijk aanwezig is bij individuele patiënten (d.w.z.,, diagnose).

Recente inspanningen om het gebruik van verschillende klinische voorspelmodellen in de klinische praktijk te automatiseren, hebben geleid tot klinische beslissingsondersteunende systemen die al bruikbaar zijn in de klinische praktijk. Deze systemen stellen clinici in staat om hun klinische besluitvorming verder te verbeteren door het gebruik van individuele risico inschattingen. Omdat de momenteel beschikbare modellen voor risico-voorspelling echter volledige informatie vereisen om voorspellingen te genereren, worden deze modellen ernstig belemmerd wanneer eigenschappen van een patiënt of ziekte ontbreken. Gelukkig is de uitgebreide kennis over het omgaan met missende waardes in staat om nuttige opstapjes te geven om flexibele of verwerkingstechnieken te ontwikkelen die bruikbaar zijn in de ‘live’ klinische praktijk.

In **Hoofdstuk 2** laten we zien dat, tot dusverre, de meeste klinische voorspelmodel studies die gebruik maken van machine learning (ML) technieken, onvoldoende informatie rapporteren over de aanwezigheid of verwerking van ontbrekende data bij het ontwikkelen of valideren van een voorspelmodel. Hoewel het onverstandig is, wordt het verwijderen van patiëntendossiers of variabelen met missende waardes het vaakst gebruikt. Dit konden we vaststellen door na te gaan of een systematisch doorzochte subset van gepubliceerde artikelen informatie bevat over vooraf gedefinieerde kenmerken die moeten worden gerapporteerd over ontbrekende gegevens.

Naast ondermaatse rapportage over missende waardes, laten we zien dat de naleving van huidige aanbevolen rapportage richtlijnen voor ML-voorspelmodel studies ook slecht is (**hoofdstuk 3**). Verschillende van de items die als essentieel worden beschouwd, zijn onvolledig gerapporteerd, wat resulteerde in een verhoogd risico op vertekening voor deze onderzoeken. Bovendien was de methodologische kwaliteit over het algemeen slecht.

Het is duidelijk dat er inspanningen nodig zijn om het ontwerp en de opeenvolgende rapportage van voorspelmodel studies te verbeteren (al dan niet met ML). Met dat doel laat **Hoofdstuk 4** de ontwikkeling zien van verschillende imputatie methoden voor missende waardes in realtime. In een case-study met een real-world empirische dataset voor cardiovasculaire risico voorspelling, vergeleken we de nauwkeurigheid van twee veelgebruikte imputatie methoden die waren aangepast voor gebruik in de realtime klinische praktijk: conditional modelling imputation (CMI, waarbij voor elke voorspeller een apart multivariabel imputatie model is afgeleid) en joint modeling imputation (JMI, waarbij we uitgaan van enkel normaal verdeelde voorspellers en de beschikbare patiëntgegevens gebruiken om imputaties te genereren voor elke missende voorspeller). Vervolgens vergelijken we deze methodes met een veelgebruikte methode in de praktijk: gemiddelde imputatie (waarbij missende waardes worden vervangen door het steekproefgemiddelde). In overeenstemming met onze verwachtingen, vonden simulaties dat zowel JMI als CMI over het algemeen kunnen worden aanbevolen in termen van imputatie nauwkeurigheid. Omdat JMI over het algemeen sneller en minder complex was, werd het als de methode met meer potentie beschouwd.

In Hoofdstuk 4 evalueren we nieuwe imputatie methoden strikt op hun nauwkeurigheid in termen van hun gemiddelde kwadratische fout. In **Hoofdstuk 5** gaan we verder met de meer veelbelovende imputatie methode (d.w.z., JMI) en evalueren deze door middel van veel gebruikte evaluatie maten voor voorspelmodellen (d.w.z., discriminatie en kalibratie). We richten ons specifiek op het gebruik van auxiliaire variabelen (d.w.z., variabelen die geen deel uitmaken van het voorspellingsmodel), werken het idee van het flexibel bijwerken van het imputatie model verder uit en maken een vergelijking met de veel gebruikte methode gemiddelde imputatie. Samengevat blijkt het gebruik van JMI het voordeligst te zijn wanneer het wordt geschat in lokale gegevens en met behulp van auxiliaire variabelen. De toegevoegde waarde is het meest prominent wanneer de missende voorspellers correleren met andere waargenomen (auxiliaire) variabelen.

De oplossing voor missende waardes in de klinische praktijk is niet alleen op te lossen door vervangende voorspeller waarden te schatten op basis van wat we weten van de individuele patiënt. Er bestaan meerdere (ML) technieken die kunnen omgaan met ontbrekende waarden met een ingebouwd ontwerp. In **Hoofdstuk 6** evalueren we meerdere methoden voor het verwerken van missende waardes en vergeleken deze met JMI. Specifiek evalueren we patroon submodellen (PS, waar voor elk bestaand patroon waarmee voorspellers missend gevonden zijn een apart voorspelmodel ontwikkeld wordt) en surrogate splits (SS, waar een optimale vervangende waarde wordt gevonden onder de beschikbare patiëntgegevens die gebruikt kan worden in plaats van de missende voorspeller). Mits meerdere imputaties worden gebruikt, heeft JMI nog steeds de voorkeur boven PS en SS.

We hebben goede hoop dat er meer grote, lokale datasets beschikbaar zullen worden gemaakt om imputatie procedures uit te kunnen voeren die realtime verwerking van missende waardes mogelijk maken. Wel moeten voorspelmodellen generaliseerbaar zijn naar dergelijke gegevens. In **Hoofdstuk 7** laten we zien dat interne-externe kruisvalidatie (IECV) de voorkeur verdient, wanneer de gegevens worden geclusterd, voor het beoordelen van de generaliseerbaarheid van een voorspellingsmodel tijdens de ontwikkeling. In het kort, IECV evalueert de model prestaties in elke hold-out-steekproef die individuen omvat uit een andere setting of populatie (bijvoorbeeld een ander ziekenhuis). Bovendien kan het worden gebruikt om te evalueren of complexe modellering strategieën (bijvoorbeeld het gebruik van interacties of niet-lineaire effecten) voordelen bieden. We ontdekten dat de nauwkeurigheid van voorspelmodellen niet noodzakelijkerwijs baat heeft bij complexere modellering strategieën, wat aantoont dat IECV potentieel nuttig is voor het vereenvoudigen van algemene model complexiteit.

Vooralsnog is het grotendeels onzeker of gepersonaliseerde geneeskunde, in de vorm van CDSS, de potentie die het laat zien zal waarmaken. Clinici zijn zeker geneigd om van wel te geloven, maar concreet bewijs van een positieve impact op de gezondheidsuitkomsten ontbreekt tot nu toe. Eerst en vooral, en voor een eerlijke vergelijking, moeten de ernstige gevolgen van onjuiste verwerking van missende waardes worden erkend en op de juiste manier worden behandeld.

Dankwoord

Vanaf het allereerste begin, en zeker toen Nederland door COVID-19 getroffen werd, was het duidelijk dat het tot stand laten komen van dit proefschrift tot het moeilijkste zou behoren dat ik ooit gedaan heb. Nu het tot een einde komt kan ik niet anders dan dankbaar zijn voor alle mensen in mijn directe en indirecte omgeving die het voor mij mogelijk hebben gemaakt om door te gaan. De voldoening en de eer komt jullie allen toe.

Allereerst zou graag mijn promotoren **prof. dr. Moons**, **prof. dr. Asselbergs**, mijn copromotor **dr. Debray** en mentor **prof. dr. Bots** willen bedanken. Jullie stonden vanaf het allereerste begin in mijn hoek en hebben mij vol begrip gesteund op de momenten dat het moeilijk werd. Het is een genoegen en voorrecht geweest dit te mogen doen onder jullie begeleiding. Beste **Carl**, ik wil je heel graag bedanken voor je onbreekbare positiviteit gedurende de promotie. Je was begripvol, inspirerend en vooral geïnteresseerd in waar ik heen wilde. Jouw inspanningen en advies zorgde ervoor dat ik een goede grip op mijn onzekerheid kreeg gedurende mijn promotie en daar ben ik je nog altijd zeer dankbaar voor. Beste **Folkert**, dank voor je constante bereidheid om te helpen en in te springen waar nodig en je zeer plezierige aanwezigheid gedurende de hele promotie. Door de belangrijkste componenten van het onderzoek regelmatig te benadrukken en de hoofd en bijzaken goed te scheiden was je bijdrage van essentiële waarde. Het staat me nog goed bij dat we op een gegeven moment belde over een pijnpunt in ons onderzoek en je mij wees op de juiste invalshoek. Beste **Thomas**, met absolute zekerheid kan ik zeggen dat dit proefschrift er niet was geweest zonder jou. Onze wekelijkse gesprekken waren inspirerend, aanstekelijk, enthousiast, kritisch en bovenal nuttig. Je nam altijd de tijd voor mij en je streefde altijd naar een leuke teamsfeer (bijv. middels etentjes). Je was bereid naar me te luisteren en dacht mee als ik weer een gek idee had (die meestal niet werkte) en dit heb ik altijd zeer gewaardeerd. Dank voor je begeleiding en dat ik zoveel van je heb mogen leren. Beste **Michiel**, mijn dank gaat uit naar je betrokkenheid, je enthousiasme en je prettige benaderbare houding. Je aanwezigheid en communicatieve vaardigheden maakte elke reguliere overleg gemakkelijker, ondanks dat ik van tevoren nerveus kon zijn. Alhoewel je naar eigen zeggen geen technische expertise hebt, wist je altijd essentiële bijdrages te leveren door het kerndoel duidelijk te maken en herkende je zonder meer waar potentiële pijnpunten zouden kunnen liggen.

Aan mijn beoordelingscommissie, **prof. dr. Rovers**, **prof. dr. Oberski**, **prof. dr. Scheepers**, **prof. dr. Visseren** en **prof. dr. Kretzschmar**, dank voor uw bereidheid mijn proefschrift te lezen en beoordelen. Ook veel dank aan **dr. van Smeden** voor het plaatsnemen in de oppositie bij mijn verdediging.

Mijn dank gaat ook uit naar alle betrokkenen bij **ORTEC**, John, Ines, Inge, Menno en het team van Logiqcare. Jullie inbreng is van onschatbare waarde geweest. Bij Ortec heb ik me altijd welkom gevoeld en het is jammer dat het beperkt is gebleven door het vele thuiswerken. Beste **John**, ook jij bent sinds het begin betrokken geweest bij mijn onderzoek, dankjewel daarvoor. Je was plezierig en leek altijd het vertrouwen erin te houden, wat heel bemoedigend werkte. Dank dat je, ook toen het minder ging, met mij bleef praten. Onze wandeling door Utrecht zal ik altijd blijven waarderen. Beste **Ines**, alhoewel je later aansloot, was je inbreng direct zichtbaar en van meerwaarde. Heel veel dank voor al je inzet en je plezierige aanwezigheid bij de vele overleggen die we gehad hebben.

Aan alle **co-auteurs** die meegewerkt hebben aan de hoofdstukken in dit proefschrift, ontzettend veel dank voor jullie harde werk en vertrouwen. Beste **Jeroen**, zonder jouw bereidheid om mee te kijken en te schrijven aan mijn eerste onderzoeken was ik nooit enigszins beslagen ten ijs gekomen. Dank voor alle moeite die je erin gestopt hebt. Je humor en creatieve schrijfvaardigheid hebben altijd veel plezier gegeven. Beste **Constanza**, het immense werk van jouw review heeft de grondslag aan twee hoofdstukken in mijn proefschrift geleverd. Ontzettend bedankt dat ik hieraan mee heb mogen werken. Onze online koffiemomenten waren altijd plezierig en iets om naar uit te kijken, dank voor je vrolijke aanwezigheid. Beste **Tuur**, dank voor de gezelligheid tijdens onze samenwerking. Het is altijd met veel plezier geweest dat ik met je discussieerde over de verschillende componenten van ons onderzoek. Dear **Toshi**, it has been a pleasure to work with you. Thank you for providing the opportunity to do so. Beste **Saskia**, veel dank voor je betrokkenheid en inzet bij mijn laatste onderzoek. Je absolute positiviteit en enthousiasme werkte aanstekelijk en het was inspirerend te zien hoe je te werk gaat. Dit zijn vaardigheden die ik hoop eigen te kunnen maken.

Veel dank aan alle **collega's die meegewerkt hebben aan het Special Interest Group project**, Gerko, Hanne en Maarten. Nooit eerder waren de reguliere overleggen zo informeel en gezellig, alsook nuttig. Dat was precies waar ik op hoopte aan het eind van mijn proefschrift, veel dank daarvoor. Beste **Gerko**, alhoewel ik je soms nog steeds niet helemaal volg als je bepaalde missing data vakjargon gebruikt heb ik je intelligente en kritische blik enorm gewaardeerd. Het heeft er absoluut tot een beter resultaat geleid. Beste **Hanne**, het was heel erg fijn om samen te werken aan ons onderzoek. Je kennis en vaardigheden zijn bewonderenswaardig en maakte het project heel goed te doen. Het is ook niet zo gek dan dat je in de toekomst al een positie als promovendus hebt vastgesteld, veel succes daarmee. Verder heb ik erg genoten van onze discussies. Heel erg bedankt voor je betrokkenheid en harde werken. Beste **Maarten**, direct bij onze eerste kennismaking, waarbij ik je vertelde van mijn voornemen om ML toe te gaan passen voor missing data, was je (gekscherend) sceptisch voor het nut daarvan. Die kritische, en humoristische, blik heb je nooit laten varen en dat heb ik erg gewaardeerd gedurende de tijd dat we samenwerkte, dankjewel.

Aan alle **collega's in het methode team**, dank voor jullie inzicht en gezelligheid. Jullie presentaties zijn inspirerend geweest en ik keek altijd uit naar onze weekstart op maandag. De teamuitjes zijn helaas pas recent weer opgestart, maar ik heb veel genoten van de momenten die ik ondanks het vele thuiswerken nog met jullie gehad heb. We zullen maar moeten aannemen dat Michiel de hypothetische pubquiz gewonnen zou hebben. Lieve **Lotty**, ondanks dat we aan weinig projecten samen gewerkt hebben, heb je ongetwijfeld een positieve invloed gehad op mijn promotie. Je vrolijke, begripvolle en lieve aanwezigheid, en bereidheid te praten wanneer ik dat nodig had, heb ik enorm gewaardeerd. Het is één van de dingen die ik absoluut mis nu ik niet meer op het Julius werk. Beste **Valentijn**, alhoewel ik niet verwacht dat ik je ooit zal verslaan met AoE heb ik altijd erg genoten van je gezelschap deze jaren. We hebben veel gelachen en het was ontzettend leuk om samen deel uit te maken van Thomas z'n team (met de extra lunches en uitjes). Dankjewel voor al je positiviteit. Beste **Hans**, ondanks dat we niet samen hebben gewerkt wilde ik je graag bedanken voor het feit dat je de reden bent geweest dat ik deze kans heb gekregen. Heel erg bedankt dat je mijn naam destijds hebt doorgegeven aan Carl.

Lieve **kamergenoten van 6.118**, Noor, Marit, Zujie, Said, Antonis, Sieta en Katrin, dank voor jullie gezelligheid. Door het thuiswerken heb ik jullie gezelschap helaas veel moeten missen, maar het heeft het eerste jaar veel goed gedaan om met jullie op de kamer te zitten.

Lieve **Katrien**, in het eerste jaar heb jij mij de kneepjes van het vak geleerd. Je aanwezigheid was van onmiddellijke en onschatbare meerwaarde en maakte dat ik met goed vertrouwen direct aan de bak kon. Je laagdrempelige strategie middels zoete stroopwafels is mij nog lang bijgebleven, alhoewel ze meestal op waren voor ik ze kon gebruiken zoals je had voorgesteld. Dank voor al je vertrouwenwekkende aanwezigheid.

Anneke en **Reinoud**, wat fijn dat jullie mijn paranimfen willen zijn en achter mij zullen staan op het grote moment. Lieve **Anneke**, in mijn laatste jaar waren we regelmatig kamergenoten en daar heb ik zo enorm van genoten. Met veel plezier heb ik met je geluncht, waarbij het soms leek alsof wij de enige waren die niet over werk wilde praten, gewandeld, gepraat en nunchaku beoefend (bij de sportdag). Leuk om te horen dat je in Utrecht blijft. **Reinoud**, buddy, soms vraag ik me af of je wel door hebt hoeveel onze vriendschap voor mij betekent. Je vermogen om ongenadig te veroordelen als ik iets fout doe en tegelijkertijd lief en begripvol zijn als ik het nodig heb is prijzenswaardig. Het heeft op zoveel momenten geholpen om de juiste keuzes te maken en ik hoop dan ook dat ik daar nog heel erg lang van kan genieten. Dankjewel dat je er al die jaren al voor mij bent.

Al mijn lieve **vrienden, (schoon)familie en mede-nunchaku-ka's**, dank voor jullie interesse in mijn onderzoek en altruïstische steun deze afgelopen jaren. Het is zo fijn geweest om mijn ei kwijt te kunnen bij jullie als dat nodig was. Jullie hebben voor de nodige ontspanning gezorgd door met mij te filosoferen over de toekomst, bordspellen te spelen, kata's te lopen en te klagen over politiek. Jullie hebben het in deze tijd een stuk makkelijker gemaakt om te kunnen genieten, dank jullie allen.

Lieve **papa en mama**, het was heel fijn toen we uiteindelijk dichterbij jullie gingen wonen en we zo veel vaker bij jullie over de vloer konden komen. Jullie zijn altijd bereid geweest te luisteren als ik ergens mee zat en geïnteresseerd in hoe het met mij ging en wat ik deed aan onderzoek. Dit is zo belangrijk geweest voor mij in deze jaren. Jullie stonden altijd aan mijn kant, hielden het vertrouwen in mijn kunnen en gaven dat zonder verwachtingen ook dikwijls aan. Daar zal ik jullie altijd dankbaar voor zijn.

Lieve **Tom**, met veel plezier denk ik terug aan onze wandeling in Kaapverdië. De afgelopen jaren voelt het alsof we meer naar elkaar zijn toegegroeid en dat vind ik heel erg fijn. Jij bent al heel lang mijn grote voorbeeld geweest. Dank voor al je ondersteunende woorden de afgelopen tijd en je vertrouwen in mij.

Hannah, lieve Hannah, zonder jouw vertrouwen en steun had ik dit allemaal nooit kunnen doen. Jij bent mijn steun en toeverlaat. Mijn happy place. Alhoewel je af en toe genadeloos eerlijk was wat je van het onderwerp van mijn thesis vond heb je altijd je best gedaan deze te begrijpen en mee te denken wanneer ik dat nodig had. Jij was degene die mij ertoe zetten mijn gezondheid tijdelijk boven mijn werk te zetten en dat is wat ervoor gezorgd heeft dat ik het af heb kunnen maken. Door jou heb ik durven dromen over dingen die ik eerder niet voor mogelijk heb gehouden. We maken een super goed team samen en ik kijk vol verwachting en geluk uit naar onze toekomst samen.

En dan tot slot, **Mels**. Kleine, lieve Mels, je bent nu nog minder dan drie weken oud. Met grote ogen kijk je echter al om je heen en ik kan alleen maar hopen dat ik een goede vader voor je zal zijn. Gelukkig is nu de ene uitdaging afgerond en is het tijd voor een ander. Ik heb er heel veel zin in.

Steven Nijman
April 2022

Curriculum Vitae

Steven Nijman was born on May 22, 1992 in Utrecht, The Netherlands. He grew up in Groenekan and, in 2009, graduated from the Montessori Lyceum Herman Jordan in Zeist. In 2009, he entered the University of Applied Sciences Utrecht to pursue a bachelor's degree in Communication and Media Design. To chase more academic interests, he moved to Amsterdam and consecutively followed the masters Information Studies: Game Studies and Medical Informatics at the University of Amsterdam. As part of the Medical Informatics masters, Steven did internships at the Medical Informatics department on several topics, primarily focused on data science.

Following this internship, in 2018, Steven began working as a PhD student at the Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, supervised by Prof. Carl Moons, Prof. Folkert Asselbergs, Prof. Michiel Bots and Dr. Thomas Debray. His thesis focused on the handling of missing data in real-time clinical practice. He combined his PhD project with the postgraduate master Clinical Epidemiology. In 2020, he acquired an additional grant from the Special Interest Group grant for Applied Data Science from the University Utrecht.

Steven currently works as a data scientist in the Data Analytics, Research and Automated Reporting (DARA) team at the National institute for Public Health and the Environment (RIVM).