Journal of Clinical Epidemiology 145 (2022) 70–80

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Population median imputation was noninferior to complex approaches for imputing missing values in cardiovascular prediction models in clinical practice

Gijs F.N. Berkelmans[a,*], Stephanie H. Read[b,c], Soffia Gudbjörnsdottir[d], Sarah H. Wild[b], Stefan Franzen[d], Yolanda van der Graaf[e], Björn Eliasson[d,g], Frank L.J. Visseren[a], Nina P. Paynter[f,#], Jannick A.N. Dorresteijn[a,#]

[a] *Department of Vascular Medicine, University Medical Center Utrecht, the Netherlands*
[b] *Usher Institute, University of Edinburgh, Edinburgh, Scotland, UK and on behalf of the Scottish Diabetes Research Network epidemiology group*
[c] *Women's College Research Institute, Canada*
[d] *Swedish National Diabetes Register, Center of Registers in Region, Gothenburg, Sweden*
[e] *Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, the Netherlands*
[f] *Harvard Medical School, Brigham & Women's Hospital, Boston, USA*
[g] *Department of Molecular and Clinical Medicine, Institute of Medicine, University of Gothenburg, Gothenburg, Sweden*

Accepted 17 January 2022; Available online 21 January 2022

Abstract

Objectives: To compare the validity and robustness of five methods for handling missing characteristics when using cardiovascular disease risk prediction models for individual patients in a real-world clinical setting.

Study design and setting: The performance of the missing data methods was assessed using data from the Swedish National Diabetes Registry (n = 419,533) with external validation using the Scottish Care Information - diabetes database (n = 226,953). Five methods for handling missing data were compared. Two methods using submodels for each combination of available data, two imputation methods: conditional imputation and median imputation, and one alternative modeling method, called the naïve approach, based on hazard ratios and populations statistics of known risk factors only. The validity was compared using calibration plots and c-statistics.

Results: C-statistics were similar across methods in both development and validation data sets, that is, 0.82 (95% CI 0.82–0.83) in the Swedish National Diabetes Registry and 0.74 (95% CI 0.74–0.75) in Scottish Care Information-diabetes database. Differences were only observed after random introduction of missing data in the most important predictor variable (i.e., age).

Conclusion: Validity and robustness of median imputation was not dissimilar to more complex methods for handling missing values, provided that the most important predictor variables, such as age, are not missing.

Keywords: Missing patient characteristics; Epidemiology; Cardiovascular risk prediction; Real-world setting; clinical practise

What is new?

- The 2021 European guidelines on cardiovascular disease (CVD) prevention in clinical practice advise using individual CVD-risk prediction to optimize and support treatment decisions about intensified preventive treatment options. However, CVD-risk prediction models do not allow missing patient information, potentially limiting their use.

Conflicts of interest: The 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.

Funding: Partly funded by American Heart Association grant: Award Number 17IGMV33860009 (NP)
Partly funded by Hartstichting, grant number 2016T026 (JD)

[#] Authors contributed equally

[*] Corresponding author: Tel.: +31 (0)88 7555161; fax: +31 (0)30 2523741
E-mail address: f.l.j.visseren@umcutrecht.nl (Frank L.J. Visseren).

https://doi.org/10.1016/j.jclinepi.2022.01.011

- Five methods for dealing with missing patient information were validated in datasets with real-world missing information and in datasets with randomly introduced patterns of missing information.
- When important patient characteristics, such as age and history of CVD, are not known, the most accurate methods are conditional imputation and the reduced model method (i.e., creating submodels for all combinations of available data). In the real-world datasets studied, however, these important predictors were always available.
- In the datasets with real-world missing information, accuracy of all five methods was not dissimilar. Thus, in clinical practice, the simplest and most pragmatic method (i.e., imputing a population median value) performs similar to more complex methods.

1. Introduction

The use of cardiovascular disease (CVD) risk prediction models to assist clinical decision-making regarding preventive medications is recommended by clinical guidelines [1,2], and is expected to increase in the future [3–5]. The 2021 European guidelines on CVD prevention in clinical practice advise using individual CVD-risk prediction to optimize and support treatment decisions about intensified preventive treatment options [6]. At time of risk prediction, some patient characteristics necessary for risk prediction may not be available. For example, the ADVANCE risk engine and DIAL model use albuminuria as a predictor of risk which is not routinely measured in all diabetic patients by clinicians [7]. This may cause healthcare providers to opt against using the risk prediction tool, potentially leading to suboptimal care [8]. In a report from the ESC prevention of CVD program, seven considerations for the selection of the best risk prediction tool for an individual patient (i.e., patients in primary care, type 2 diabetes mellitus (T2DM) patients, or patients with a history of CVD) have been stated. One of seven is the ability of the prediction model to offer features that enable dealing with missing or unavailable values such replacing the missing value with the median value for a given population [9].

While extensive research has been undertaken into the handling of missing data in the development and validation of risk prediction models, there is limited evidence on methods to deal with missing patient characteristics in the implementation stage [10–12]. Of the 23 prediction models discussed by Tsvetvanova et al, less then halve provided methods and guidance for approaches to estimating patient CVD risk in the absence of patient characteristics included in the model. Of the nine methods that allow absence of patient characteristics eight impute a value representing a healthy person and one model imputes the population mean. None of the models use more sophisticated methods such as conditional imputation or reduced model methods [13].

Several methods dealing with missing patient characteristics have been described for the implementation stage. However, most studies were simulation studies (not a real-world clinical setting), and were not extended to a Cox proportional hazard model. Examples are the reduced model methods [14], hybrid model method [14], conditional single imputation, mean/median imputation [15], and the naïve approach [16].

The objective of the present study is to compare the validity and robustness of individual CVD risk prediction using the before mentioned five methods for handling missing patient characteristics in a real-world clinical setting, using the development of the Swedish National Diabetes Registry (NDR) risk prediction model and its validation in Scottish data for patients with T2DM as an example [17].

2. Methods

2.1. Study population

Using data from nationwide health registers from the Swedish NDR and the Scottish Care Information-Diabetes (SCI-Diabetes) database linked to hospital admission and death records. Patients were aged >18 years with a diagnosis of T2DM registered in the Swedish NDR [18] between 2002 and 2012 or in the SCI-Diabetes register between 2004 and 2016. The definition of T2DM in the Swedish NDR was treatment with 1) diet only, 2) oral hypoglycaemic agents only, or 3) insulin only or combined with oral agents, and onset age of diabetes ≥40 years. In the SCI – diabetes database, T2DM was defined using an algorithm which uses information on diabetes type recorded by the clinician, prescription data (use of and timing of sulphonylureas and insulin) and age at diagnosis. People with a previous diagnosis of cancer (ICD-10 codes C00-C97) were excluded due to their increased risk of (short-term) mortality [19]. Use of each register's data was approved by institutional review boards.

Clinical characteristics at baseline for patients registered in the Swedish NDR and SCI –diabetes database were identified in the first year after registration. The Swedish NDR included people with prevalent and incident diabetes. At baseline, also the age at onset of diabetes was registered. The SCI-diabetes database only registered people with incident diabetes mellitus. Clinical characteristics included in the updated Swedish NDR risk score were age (years), sex (female/male), age at onset of T2DM (years), smoking status (yes/no), body-mass index (BMI in kg/m²), systolic blood pressure (SBP in mm Hg), hemoglobin A1c (HbA1c in mmol/mol), non-high-density lipoprotein cholesterol (non-HDLc in mmol/L), albuminuria (no/micro/macro), estimated glomerular filtration rate (eGFR in mL/min/1.73 m²), retinopathy (yes/no), and a his-

tory of CVD (yes/no) and atrial fibrillation (yes/no). Microalbuminuria was defined as an albumin/creatinine ratio 3 to 30 mg/mmol or urine-albumin 20 to 300 mg/L, and macro-albuminuria was defined as an albumin/creatinine ratio >30 mg/mmol or urine-albumin >300 mg/L. All baseline characteristics had missing data, excluding age, sex, history of CVD, and atrial fibrillation.

2.2. Five methods for handling missing data

An arbitrary 25% of patients picked at random from the Swedish NDR (development dataset) was used to re-estimate the Swedish NDR risk equation [17] (supplemental methods, including calibration plots of internal and external validation) and generate the necessary algorithms and statistics for each method for handling missing data. These five developed methods were:

1) Reduced model method (also described as complete case submodels) [20]: development of a comprehensive set of models for each possible combination of available characteristics. This means that in addition to the full model including all variables, all other possible models with a combination of fewer variables are developed. In the current example, the full model consists of 13 variables. Therefore, for the reduced model method, 2^13 = 8,192 models were developed within the development dataset.
2) Hybrid model method: development of a set of models that allow one missing value at a time and averaging the predicted risks if more than one value is missing. This means that in addition to the full model including all variables, models with one variable missing are developed. In the current example (full model of 13 predictors), the hybrid model method generates 14 models (i.e., one full model plus 13 single variable missing models). When more than one variable is missing, multiple risk predictions are calculated using each applicable single variable missing model and averaging the result of those models. Other missing variables in these models are imputed using the median value for continuous variables or mean value for categorical variables.
3) Conditional single imputation: iterative algorithm-based imputation of missing values. This means that in addition to the risk score, imputation models are developed to impute missing variables based on the available characteristics. In the current example (full model of 13 predictors), this results in one risk score and 13 imputation models. All missing values were estimated with a linear or logistic regression model for continuous and categorical variables, respectively. The imputation model estimates the missing variable using all other variables available. In case of multiple missing variables, the different imputation models estimating the missing variable with a median value for other missing variables. After the first estimates of missing variables were calculated, these estimates were included in the different imputation models at a new second time of estimating variables. This was repeated for 30 times to ensure accurate estimates for missing variables also when more than one variable was missing.
4) Median imputation method: Imputation of missing values using the population median for continuous predictors or the population mean proportion for categorical predictors derived from the data in which the risk score was originally developed. In the current example (full model of 13 predictors), this results in one risk score algorithm and 13 imputation values.
5) The naïve approach: not using the original risk score, but estimating individual risk based on hazard ratios and population statistics of known risk factors only. This means that no statistical model is developed from the development dataset, but only the following population statistics: The baseline population survival, the prevalence of categorical predictors, the mean values of continuous predictors, and the independent hazard ratios for all predictors. These population statistics enable calculation of individual risk based on a stepwise process that starts with the assumption that the baseline population survival is the best estimate of individual risk if nothing is known about that individual. This first baseline individual risk estimate can be updated in a stepwise process with each variable that is known using the following formula: $Baseline\ population\ survival^{\left(\frac{Hazard\ ratio}{population\ relative\ risk}\right)}$, where the population relative risk is equal to $(Prevalence\ of\ risk\ factor \times HR\ of\ the\ factor) + ((1 - prevalence) \times 1.0)$ for categorical variables. For continuous variables, the $\frac{Hazard\ ratio}{Population\ relative\ risk}$ is equal to the $\frac{Hazard\ ratio \times individual\ continuous\ value}{Hazard\ ratio \times median\ value\ of\ population}$ This formula, thus, is used for each known risk factor value, whereby the predicted survival of one step (e.g., average population risk updated for patient's age) is used as the new updated baseline survival for the next step (e.g., age-specific risk updated for patient's sex). For individual patients, this method calculates individual risk based on available risk factor information only. Missing data, thus, are no issue when calculating individual risk using the naïve approach method.

2.3. Internal validation in the Swedish NDR

In the remaining arbitrary 75% of patients (internal test dataset) from the Swedish NDR, real-world missing data were available to compare each method for handling missing data. Each method for handling missing data was applied in the estimation of 5-year risk for all patients using the Swedish NDR risk score estimated from the other 25% of the population. The effect of missing data on the predictive accuracy of risk predictions was quantified by comparing C-statistics and calibration plots, stratified by the number of missing variables. To further test the robustness of the five methods, missing patient characteristics were

introduced in the subset of the Swedish NDR with complete data in three different ways. First, for all patients with complete data, one variable was removed completely for all patients. This was performed subsequently for all variables in the risk score. For example, in the complete data, SBP was removed and assumed unknown for all patients with all other variables available.

Second, we simulated multivariate missing data using the ampute function of the mice package in R. [21] Missing data were generated as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). The missing data patterns necessary for the ampute function were specified to create a similar pattern of missing data as in the original dataset. In the original dataset, 15% of the cases had no missing data. These were excluded before patterns were specified to avoid a pattern of cases without any missing characteristics. For all other cases in the original dataset with missing data between 1 and 9 missing variables, patterns were specified and the proportions of each pattern was calculated. Proportions were the number of appearances of a certain pattern divided by the total number of cases with missing characteristics. By specifying the proportion in the ampute function for all different patterns, the constructed dataset with missing variables (from the selected complete cases) had a proportion and number of cases with missing between 1 and 9 that was similar to the to the original datasets. The R statistical code of this part of the statistical analyses is included in the supplemental materials.

Similar to the original dataset, values for age, sex, history of CVD, and atrial fibrillation were never missing. Therefore, as a third way of testing robustness of the methods, missing data was also introduced for age, sex, history of CVD, and atrial fibrillation. Missing data for these characteristics were added to the constructed data set using the ampute function with all possible patterns (so addition of 1 to 4 missing characteristics with all possible permutations as patterns with equal proportions for any of the patterns).

2.4. External validation in the SCI-diabetes database

For external validation of the methods, data from the SCI-diabetes database with real-world missing data were used. First, 25% of the SCI-diabetes database was used to recalibrate the Swedish NDR 5-year risk equation for differences in baseline hazards (supplemental methods). Only the risk equation was recalibrated, not the algorithms and population statistics developed for imputation. For feasibility reasons, the reduced model method could not be externally validated as this would require recalibration of 8,192 developed (reduced) models. The remaining 75% of the SCI-diabetes database was used for external validation. All statistical analyses were conducted using R version 3.4.1.

3. Results

3.1. Study population

The baseline characteristics (including percentage of missing characteristics) of patients in the Swedish NDR development dataset (n = 104,883), the Swedish NDR test dataset (n = 314,650) and the external validation SCI-diabetes database (n = 170,215) are shown in Table 1. Notably, in the Swedish NDR and the SCI-diabetes database, age, sex, history of CVD, and history of atrial fibrillation were always available (0% missing), and also age at onset of T2DM and retinopathy was never missing in the SCI-diabetes database. The remaining missing data were not missing completely at random. Patients without missing data were in general younger (median age of patients: 65 years without missing data vs. 66 years with missing data). Also, patients without missing data had a longer duration of diabetes, with a difference in median duration of 2 years.

3.2. Internal validation in the Swedish NDR

An example of how the models would be applied to a single participant with a common pattern of missing data in the test dataset of the Swedish NDR is given in Supplemental Table 1. Overall, the predicted 5-year risks using any of the methods for dealing with missing data showed good agreement with 5-year observed risks (Fig. 1). There was no difference in discriminative power as evaluated by c-statistics of 0.82 (95% CI 0.82–0.83) for all methods. Also, when stratified for the number of missing characteristics, no differences in c-statistics between the methods were observed (Fig. 2). Even with 9 missing patient characteristics (only age, sex, history of CVD, and atrial fibrillation available), c-statistics remained high (0.81; 95% CI 0.78–0.83).

3.3. Removal of one variable in patients with complete data

The results were different after missing data were introduced in the subset of patients of the Swedish NDR with complete data (n = 46,971; 15%). When age, the most important variable (Supplemental Table 2) was missing for all patients, the single imputation method, reduced model method, and hybrid model method resulted in c-statistics of 0.80 (95% CI 0.80–0.81) compared to c-statistics of only 0.75 (95% CI 0.74–0.75) using median imputation or the naïve approach (Fig. 3). Missing data of variables with the highest chi-squares in the model (i.e., age or history of CVD; Supplemental Table 2) resulted in 5% underestimation of predicted risk in the highest quintile of observed risk when median imputation or the naïve approach was applied (Supplemental Fig. 1). The observed and predicted risks showed adequate agreement when applying the other

Table 1. Baseline characteristics of people aged >18 years with a diagnosis of T2DM registered in the Swedish NDR between 2002 and 2012 foe development and internal test cohorts or in the 75% of the people in the SCI-diabetes database between 2004 and 2016 included in the external validation

	Development dataset Swedish NDR (n = 104,883)		Internal Test dataset Swedish NDR (n = 314,650)		External validation dataset SCI-diabetes database (n = 170,215)	
	Median/frequency	Missing	Median/frequency	Missing	Median/frequency	Missing
Age (y)	66 (57–75)	0 (0%)	66 (58–75)	0 (0%)	61 (52–71)	0 (0%)
Male sex	58,557 (56%)	0 (0%)	176,051 (56%)	0 (0%)	95,869 (56%)	0 (0%)
Current smoking	13,111 (16%)	23,030 (22%)	39,591 (16%)	68,839 (22%)	29,634 (17%)	41,892 (25%)
Age at onset of T2DM (y)	60 (52–69)	12,794 (12%)	61 (52-69)	38,054 (12%)	61 (52–71)	0 (0%)
Systolic blood pressure (mm Hg)	140 (128–150)	15,473 (15%)	140 (128-150)	46,283 (15%)	135 (125–145)	21,003 (12%)
Body mass index (kg/m2)	29 (26–33)	26,824 (26%)	29 (26–33)	80,221 (25%)	31 (28–36)	65,145 (38%)
HbA1c (mmol/mol)	50 (44–59)	11,842 (11%)	50 (44–59)	35,611 (11%)	53 (45–65)	23,207 (14%)
Non-HDL-c (mmol/l)	3.7 (3.0–4.4)	42,240 (40%)	3.7 (3.0–4.4)	127,107 (40%)	3.1 (2.7–4.2)	60,971 (36%)
eGFR (mL/min/1.73m^2)	82 (67–95)	22,059 (21%)	83 (66–95)	66,912 (21%)	81 (66–95)	29,052 (17%)
Micro-albuminuria	9,106 (15%)	43,862 (42%)	27,499 (15%)	132,091 (42%)	14637 (9%)	77,099 (45%)
Macro-albuminuria	4,712 (8%)		14,090 (8%)		1,553 (1%)	
Retinopathy	7,249 (21%)	70,718 (67%)	21,800 (21%)	212,485 (68%)	20,958 (12%)	0 (0%)
Atrial fibrillation	7,582 (7%)	0 (0%)	23,208 (7%)	0 (0%)	8,311 (5%)	0 (0%)
History of CVD	13,952 (13%)	0 (0%)	41,117 (13%)	0 (0%)	27,919 (16%)	0 (0%)

All data are shown as median (inter quartile range) or frequency (%). NDR: National Diabetes Registry. SCI: Scottish Care Information. Micro-albuminuria was defined as an albumin/creatinine ratio 3 to 30 mg/mmol or urine-albumin 20 to 300mg/L. Macro-albuminuria was defined as an albumin/creatinine ratio >30 mg/mmol or urine-albumin >300 mg/L.

methods with 0% to 2% underestimation of predicted risk in the highest quantile of observed risk.

3.4. Simulated multivariate missing data using ampute function

Second, with the introduction of multiple missing characteristics patterns similar to the real-world missing data in the Swedish NDR, the predicted 5-year risks showed good agreement with 5-year observed risks for all methods dealing with missing characteristics (Supplemental Fig. 2). Also, there was no difference in discriminative power as evaluated by the c-statistic of 0.81 (95% CI 0.81–0.82) for all methods. No difference between methods was observed when stratified for number of missing characteristics. These findings were similar for different missing data mechanisms (MCAR, MAR, and MNAR; Fig. 4).

3.5. Introduction of missing age, sex, history of CVD, and atrial fibrillation

Third, when age, sex, history of CVD, and atrial fibrillation (i.e., variables that were never missing in the real-world data) were included in the missing data patterns, adequate agreement between predicted 5-year risk and 5-year observed risk was still observed using the reduced model method only. However, when using the single imputation method, hybrid model method, median imputation, and naïve approach, respectively, increasing levels of underestimation of risk were observed (Supplemental Fig. 3). C-statistics were lower overall compared to the real-world missing data patterns (c-statistics between 0.72 and 0.77), with the highest c-statistic for the reduced model method, followed by the single imputation method, hybrid model method, median imputation, and naïve approach. These results were similar for the different missing data mechanisms, however the overall discriminative power was lower for the MNAR compared to the MCAR and MAR patterns (Supplemental Fig. 4).

3.6. External validation in the SCI-diabetes database

After recalibration of the Swedish NDR risk equation (supplemental methods), in 75% the SCI-diabetes database (n = 170,215), there was no difference in discriminative ability between the hybrid method, single imputation, median imputation, or naïve approach with c-statistics of 0.74 (95% CI 0.74–0.75; Fig. 5). An example of how the models would be applied to a single patient with common patterns of missing data in the SCI-diabetes database is given in Supplemental Table 2. Predicted and observed 5-year risks were similar in patients with <30% observed risk using each of the five methods for handling missing data. In patients with an observed risk >30%, all methods over-

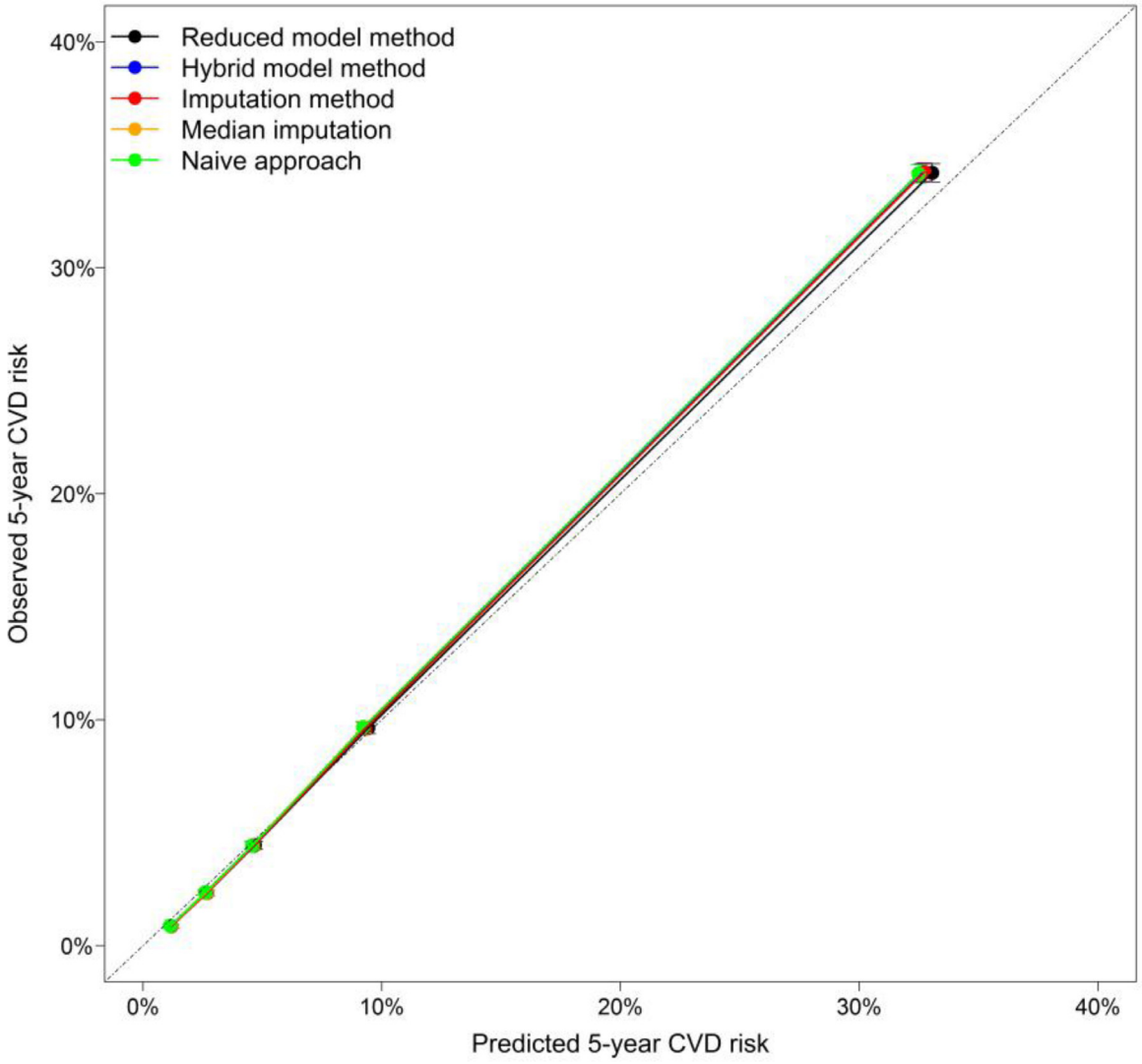

Fig. 1. Calibration plot of observed vs. predicted risk among patients in the Swedish National Diabetes Register (n = 314,650) with real-world missing patient characteristics using five methods of dealing with missing characteristics. Dots represent mean risks with 95% confidence intervals of people grouped by quintiles of predicted risk. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

estimated risk as expected based on the recalibration curve (Supplemental Fig. 5).

4. Discussion

Using nationwide registers in two countries, five methods for dealing with missing patient characteristics were developed and validated in real-world datasets with missing characteristics and in data with randomly introduced patterns of missing data. The hybrid model method, single imputation, median imputation, and naïve approach all showed similar discrimination and good calibration compared to the reduced model method. When important predictor variables were missing, such as age and history of CVD, optimal accuracy was achieved by single imputation or the reduced model method. However, when age, history of CVD, atrial fibrillation, and sex were available and up to 9 out of 13 variables were imputed using any of the five methods tested acceptable and comparable results for individual predicted risk were produced.

In the model development stage, multiple imputation is advocated as the preferred imputation method and also leads to more appropriate standard errors and *P* values; in single imputation, these are under-estimated [22]. However, it is not necessary that missing data is handled the same way at development, validation, or implementation stage [13]. This study only focused on the risk prediction at the implementation stage in the presence of missing predictors for individual patients. Point estimates of missing predictors using single imputation or multiple imputation were very similar [23]. The simplicity of single imputation without losing accuracy of the point estimate made us choose the single imputation as our only imputation method.

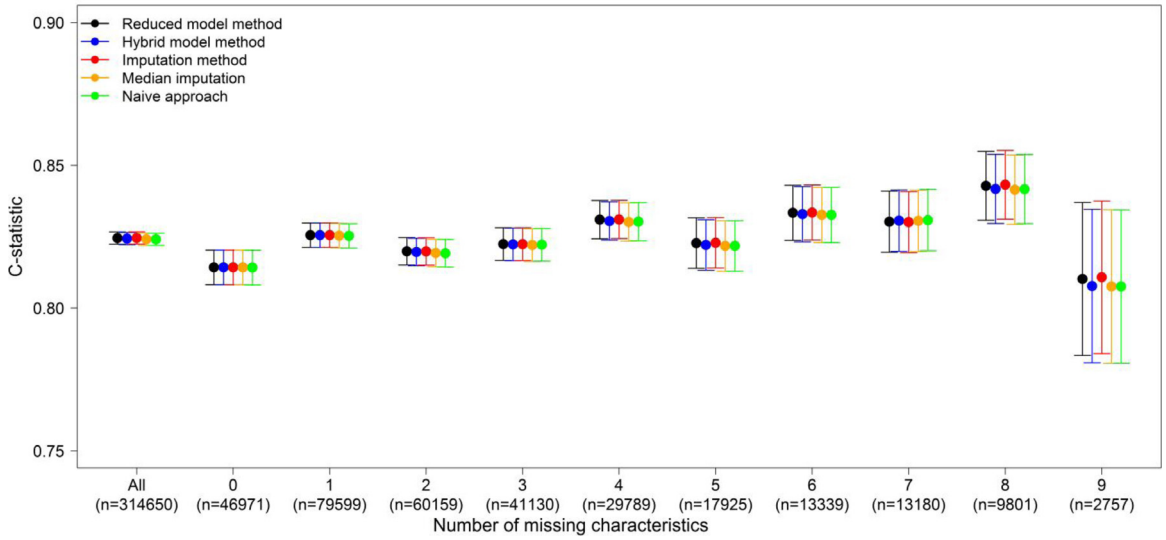

Fig. 2. C-statistics for each method of handling missing patient characteristics in the Swedish National Diabetes Register (n = 314,650) by number of real-world missing patient characteristics. Y-axis are scaled from 0.75 to 0.90. Groups with 0 to 9 missing characteristics are mutually exclusive. The different number of missing characteristics on the x-axis, represent 10 different subsets of all (n = 314,650) patients. Therefore, only c-statistics within each subset can be compared. The "all" patients group represents the total population. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

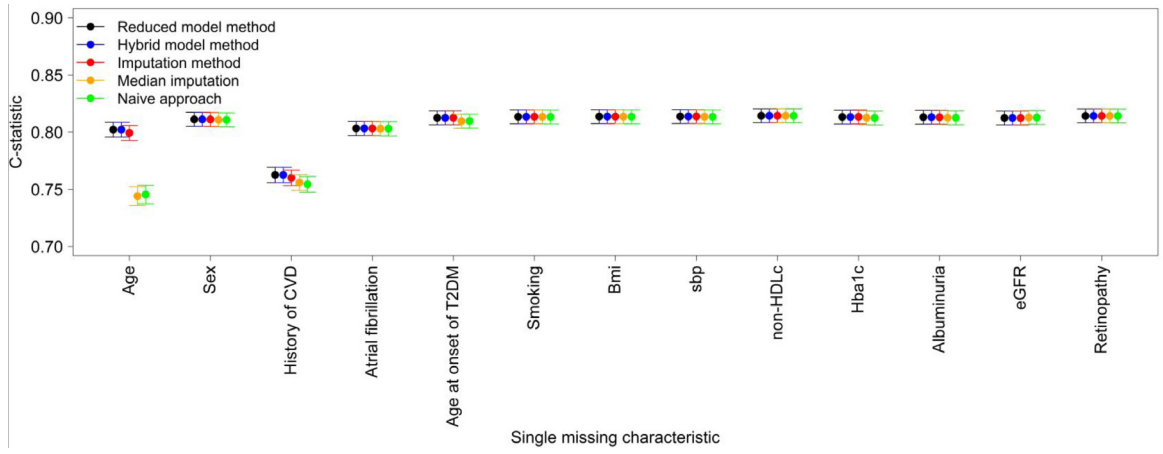

Fig. 3. C-statistics for each method of handling missing patient data among a subset of people with complete data and in whom missing data were introduced for each missing characteristic separately in the Swedish national diabetes registry (n = 46,971). Y-axis are scaled from 0.70 to 0.90. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

The results of this study are in line with previous studies in a diagnostic setting [15]. In a study with a diagnostic prediction model for deep venous thrombosis (DVT), the authors compared multiple imputation to other strategies for handling missing data. In the absence of a D-dimer test (the strongest predictor for the diagnosis of DVT), multiple imputation was the best way to deal with missing characteristics. In the absence of calf circumference, which is a weak predictor for the diagnosis of DVT, all strategies had similar results in terms of calibration and c-statistics. However, it must be emphasized that in the clinical setting for diagnosing DVT, only a few variables are needed in the model that are usually available. CVD prediction models usually contain 6 to 16 variables and therefore the chance of missing variables is higher [24].

A more recent published study by Hoogland et al. handling missing data for a new individual in the context of prediction evaluated seven methods, including sub model methods (reduced model method) and stacked imputation which resembles the single imputation method used in this study. In this study the authors conclude that the reduced model method and the imputation method perform best dealing with missing data in new individuals. This is similar to our results when introducing missing values in all variables (including age, sex, history of CVD, and atrial fibrillation). It must be noted that the study of Hoogland

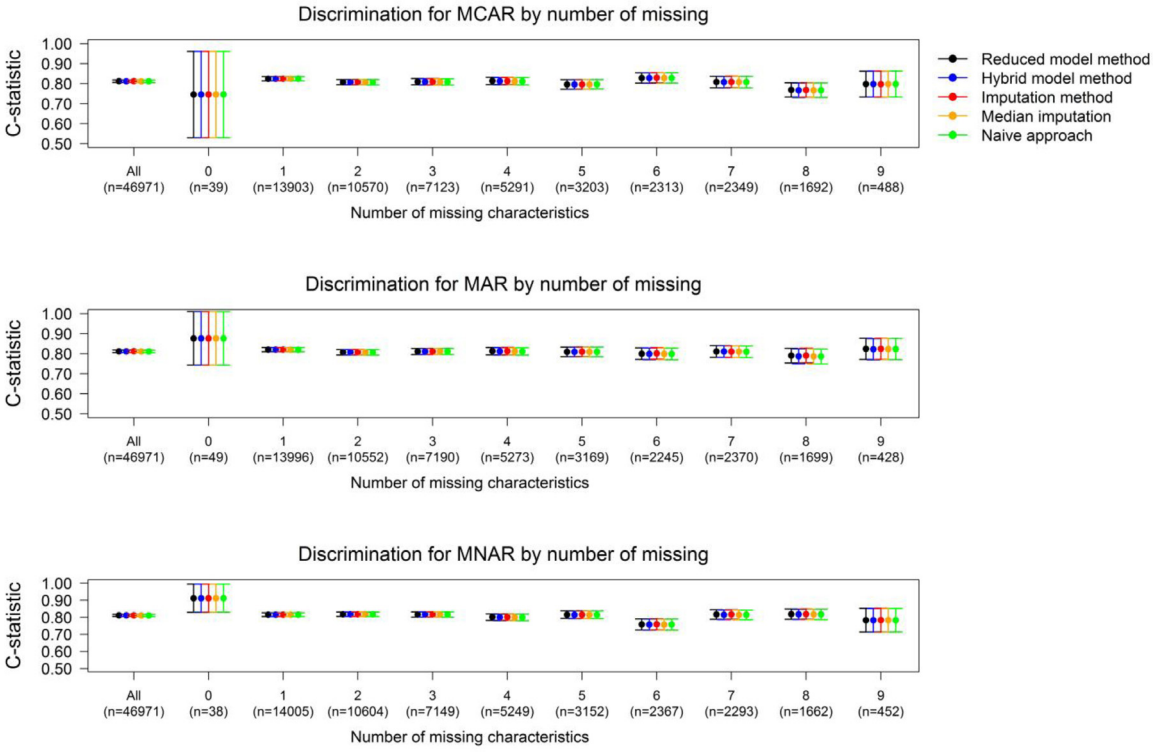

Fig. 4. C-statistics for each method of handling missing patient data among a subset of people with complete data from Swedish National Diabetes Registry (n = 46,971) and where missing data were introduced in a MCAR, MAR, and MNAR manner. The different number of missing characteristics on the x-axis, represent 10 different subsets of all (n = 46,971) patients. Therefore, only c-statistics within each subset can be compared. Y-axis are scaled from 0.50 to 1.00. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

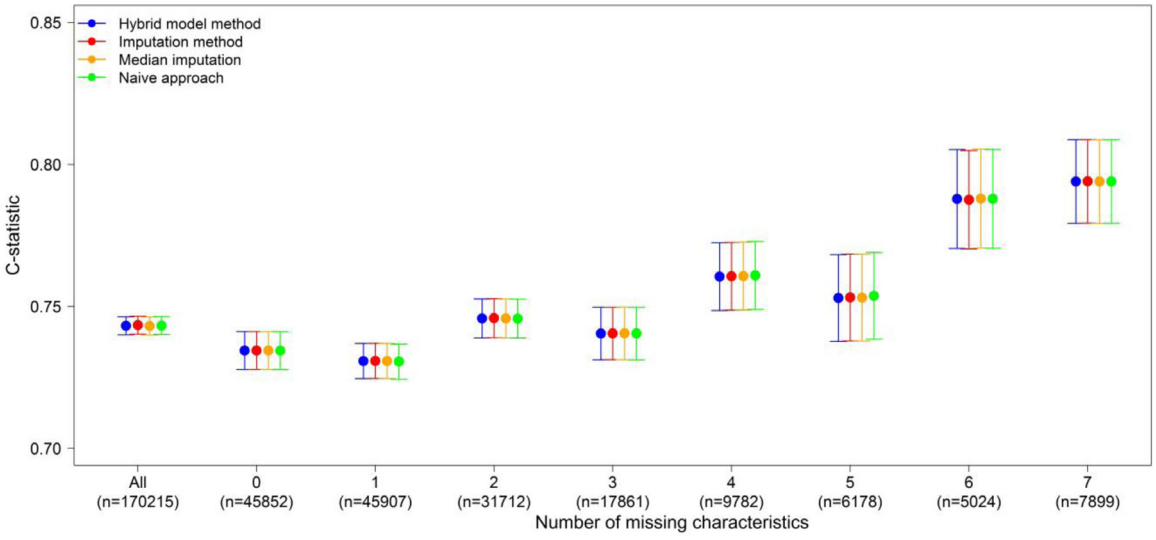

Fig. 5. C-statistics for each method of handling missing data among people in the Scottish Care Information - diabetes database (n = 170,250) according to number of missing characteristics. Y-axis are scaled from 0.70 to 0.85. Groups with 0 to 7 missing characteristics are mutually exclusive. The different number of missing characteristics on the x-axis, represent 10 different subsets of all (n = 170,215) patients. Therefore, only c-statistics within each subset can be compared. The "all" patients group represents the total population. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

et al. simulates missing data without taking into account the variables that are always available in clinical practice. No simulations were performed with values of (important) variables available such as age of a patient. However, using any method dealing with missing values in a clinical setting with small numbers of missingness showed no difference [25].

In our study, the results did not differ by missing data method when handling incomplete data on weak predictors since these weak predictors have limited effect on predictive accuracy. Therefore, each of the proposed methods were able to adequately deal with missing data for weak predictors [26,27]. The opposite is true when data for strong predictors are missing. In the case of missing strong predictors, the method resulting in the closest estimate of the true value is more likely to have the highest predictive accuracy compared to other methods. Thus, when dealing with missing strong predictors, median imputation and the naïve approach are insufficient.

Importantly, the present study focuses on the accuracy of risk prediction, since this is the main purpose of using cardiovascular risk prediction models in clinical practice. Other studies have investigated methods for estimating the actual missing value itself. For example, Nijman et al. showed that more sophisticated methods such as joint modelling imputation and conditional modeling imputation outperformed median imputation when estimating missing predictor values. However, they also conclude in their study that this hardly has any effect on the differences in risk prediction [28].

These findings should encourage the addition of imputation models within apps or web-based calculators to handle incomplete data to enable physicians to reliably use risk prediction models in the presence of missing patient characteristics. Although imputation may reduce the negative effects of missing characteristics in clinical practice, it should be emphasized that it is still preferable to have complete data available. Any method for dealing with missing data when using prediction models results in a small loss of predictive accuracy.

While the reduced model method, hybrid model method, single and median imputation methods, provide actual numbers for missing data, the naïve approach uses the population baseline hazard and the hazard ratios from the risk equation to estimate individual risks [16]. Interestingly, this fundamental difference in methods did not lead to differences in c-statistics or calibration. Thus, with all characteristics available, the naïve approach was as accurate as predictions using Cox proportional hazard models. This could provide further opportunities when other important patient information is available in addition to predictors in a risk model, such as coronary calcium score [29] or family history [30]. Both are mentioned in the ESC guidelines to downgrade or upgrade the risk in intermediate risk categories [31]. With the naïve approach, this information could be added to an existing model if the hazard ratio from large studies, ideally adjusted for all predictors in the model, and prevalence in the population is known.

Some strengths and limitations of the present study should be considered. Using nationwide registers to include data from a large number of patients seen in routine clinical practice, the observational nature, and the methods to gather patient data in the Swedish NDR and the SCI-diabetes database allowed for analyses in clinical data with real-world missing characteristics that can be generalized to clinical practice in similar settings. Whether the findings of this study could also be generalized to other fields in medicine is uncertain. However, the use of imputation as the most accurate way to deal with missing characteristics in the DVT example suggests that this method applies for prediction models in general. Although five methods were developed, only four methods were externally validated in the SCI-diabetes database. The reduced model method was not externally validated in the SCI-diabetes database because recalibration of the 8,192 models was computationally infeasible even when using a high- performance cloud server. Therefore, the reduced model method, despite being one of the most accurate methods for handling missing data, may not be suitable for clinical use. In non-randomized studies and clinical practice, a missing patient characteristic itself could be of value predicting risk of disease. For instance, a general practitioner measures albuminuria less often in patients that are less prone to illness, and therefore missing this predictor could mean a lower risk for disease. However, caution is required using this missing indicator method because there is a potential risk for bias [32].

In conclusion, pragmatic imputation of real-world missing values by median values resulted in valid and robust predictions in a clinical setting. Only when the model's most important characteristics, such as age and history of CVD in this example, are simulated as being missing, was median imputation outperformed by more sophisticated methods such as imputation or reduced model method. We conclude that the clinical use of CVD prediction tools in practice could be facilitated by automatic imputation of missing patient characteristics by median values assuming important characteristics to be available.

Disclosures

G.F.N. Berkelmans declares no conflict of interest.
S.H. Read declares no conflict of interest
S. Gudbjörnsdottir declares no conflict of interest.
S.H. Wild declares no conflict of interest.
S. Franzen declares no conflict of interest.
Y. van der Graaf declares no conflict of interest.
B. Eliasson reports personal fees from Amgen, personal fees from AstraZeneca, personal fees from Boehringer Ingelheim, personal fees from Eli Lilly, personal fees from Merck Sharp & Dohme, personal fees from Mundipharma, personal fees from Navamedic, personal

fees from NovoNordisk, personal fees from RLS Global, grants and personal fees from Sanofi, outside the submitted work.

F.L.J. Visseren declares no conflict of interest.

N.P. Paynter declares no conflict of interest.

J.A.N. Dorresteijn declares no conflict of interest.

Acknowledgments

Ann-Marie Svensson, R.N., Ph.D. and associate professor (deceased 2021), was involved in the planning and implementation of the study that forms the basis of this manuscript.

For the Swedish national diabetes registry, we thank all of the clinicians who were involved in the care of patients with diabetes for collecting data, and staff at the Swedish National Diabetes Registry.

We acknowledge with gratitude the contributions of people and organizations involved in providing data, setting up, maintaining and overseeing SCI-diabetes, including the Scottish Diabetes Research Network that is supported by National Health Service (NHS) Research Scotland, a partnership involving Scottish NHS Boards and the Chief Scientist Office of the Scottish Government.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jclinepi.2022.01.011.

References

[1] Goff DC Jr, Lloyd-Jones DM, Bennett G, et al. ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation 2013:S49–73 2014;129(25 Suppl 2). doi:10.1161/01.cir.0000437741.48606.98.

[2] Ryden L, Grant PJ, Anker SD, et al. ESC Guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD: the Task Force on diabetes, pre-diabetes, and cardiovascular diseases of the European Society of Cardiology (ESC) and developed in collaboration with the European Association for the Study of Diabetes (EASD). Eur heart j 2013;34:3035–87 doi:[published Online First: 2013/09/03]. doi:10.1093/eurheartj/eht108.

[3] Damen JA, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ 2016;353:i2416. doi:10.1136/bmj.i2416.

[4] Kleinrouweler CE, Cheong-See FM, Collins GS, et al. Prognostic models in obstetrics: available, but far from applicable. Am J Obstet Gynecol 2016;214:79–90 e36. doi:10.1016/j.ajog.2015.06.013.

[5] Noble D, Mathur R, Dent T, et al. Risk models and scores for type 2 diabetes: systematic review. BMJ 2011;343:d7163. doi:10.1136/bmj.d7163.

[6] Visseren FLJ, Mach F, Smulders YM, et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. European heart journal 2021;42:3227–337 [published Online First: 2021/08/31]. doi:10.1093/eurheartj/ehab484.

[7] Shin JI, Chang AR, Grams ME, et al. Albuminuria testing in hypertension and diabetes: an individual-participant data meta-analysis in a global consortium. Hypertension 2021;78:1042–52 doi:[published Online First: 2021/08/10]. doi:10.1161/HYPERTENSIONAHA.121.17323.

[8] Groenhof TKJ, Rittersma ZH, Bots ML, 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;27:435–42 [published Online First: 2019/08/03]. doi:10.1007/s12471-019-01308-w.

[9] Rossello X, Dorresteijn JA, Janssen A, et al. Risk prediction tools in cardiovascular disease prevention: a report from the ESC Prevention of CVD Programme led by the European Association of Preventive Cardiology (EAPC) in collaboration with the Acute Cardiovascular Care Association (ACCA) and the Association of Cardiovascular Nursing and Allied Professions (ACNAP). European journal of preventive cardiology 2019;26:1534–44. doi:10.1177/2047487319846715.

[10] Engels JM, Diehr P. Imputation of missing longitudinal data: a comparison of methods. J Clin Epidemiol 2003;56:968–76.

[11] Gorelick MH. Bias arising from missing data in predictive models. J Clin Epidemiol 2006;59:1115–23. doi:10.1016/j.jclinepi.2004.11.029.

[12] Shrive FM, Stuart H, Quan H, et al. Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. BMC Med Res Methodol 2006;6:57. doi:10.1186/1471-2288-6-57.

[13] Tsvetanova A, Sperrin M, Peek N, et al. Missing data was handled inconsistently in UK prediction models: a review of method used. J Clin Epidemiol 2021;140:149–58 [published Online First: 2021/09/15]. doi:10.1016/j.jclinepi.2021.09.008.

[14] Saar-Tsechansky M, Provost F. Handling missing values when applying classification models. J machine learning res [1532-4435] 2007;8:1623.

[15] Janssen KJ, Vergouwe Y, Donders AR, et al. Dealing with missing predictor values when applying clinical prediction models. Clin Chem 2009;55:994–1001. doi:10.1373/clinchem.2008.115345.

[16] Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Statistics in med 2004;23:1111–30. doi:10.1002/sim.1668.

[17] Zethelius B, Eliasson B, Eeg-Olofsson K, et al. A new model for 5-year risk of cardiovascular disease in type 2 diabetes, from the Swedish National Diabetes Register (NDR). Diabetes Res Clin Pract 2011;93:276–84. doi:10.1016/j.diabres.2011.05.037.

[18] Gudbjornsdottir S, Cederholm J, Nilsson PM, et al. The National Diabetes Register in Sweden: an implementation of the St. Vincent Declaration for Quality Improvement in Diabetes Care. Diabetes care 2003;26:1270–6.

[19] Mertens AC, Yasui Y, Neglia JP, et al. Late mortality experience in five-year survivors of childhood and adolescent cancer: the Childhood Cancer Survivor Study. J Clin Oncol 2001;19:3163–72. doi:10.1200/JCO.2001.19.13.3163.

[20] Fletcher Mercaldo S, Blume JD. Missing data and prediction: the pattern submodel. Biostatistics 2020;21:236–52. doi:10.1093/biostatistics/kxy040.

[21] Schouten RM, Lugtig P, Vink G. Generating missing values for simulation purposes: a multivariate amputation procedure. J Stat Comput Sim 2018;88:2909–30. doi:10.1080/00949655.2018.1491577.

[22] Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann internal medi 2015;162:W1–73 [published Online First: 2015/01/07]. doi:10.7326/M14-0698.

[23] Donders AR, van der Heijden GJ, Stijnen T, et al. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol

2006;59:1087–91 [published Online First: 2006/09/19]. doi:10.1016/j.jclinepi.2006.01.014.

[24] Wessler BS, Lai Yh L, Kramer W, et al. Clinical prediction models for cardiovascular disease: tufts predictive analytics and comparative effectiveness clinical prediction model database. Circ Cardiovasc Qual Outcomes 2015;8:368–75. doi:10.1161/CIRCOUTCOMES.115.001693.

[25] 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 med 2020;39:3591–607. doi:10.1002/sim.8682.

[26] Austin PC, Pencinca MJ, Steyerberg EW. Predictive accuracy of novel risk factors and markers: a simulation study of the sensitivity of different performance measures for the Cox proportional hazards regression model. Stat Methods Med Res 2017;26:1053–77. doi:10.1177/0962280214567141.

[27] Austin PC, Steyerberg EW. Predictive accuracy of risk factors and markers: a simulation study of the effect of novel markers on different performance measures for logistic regression models. Statistics in med 2013;32:661–72. doi:10.1002/sim.5598.

[28] Nijman SWJ, Groenhof TKJ, Hoogland J, et al. Real-time imputation of missing predictor values improved the application of prediction models in daily practice. J Clin Epidemiol 2021;134:22–34. doi:10.1016/j.jclinepi.2021.01.003.

[29] Polonsky TS, McClelland RL, Jorgensen NW, et al. Coronary artery calcium score and risk classification for coronary heart disease prediction. Jama 2010;303:1610–16 [published Online First: 2010/04/29]. doi:10.1001/jama.2010.461.

[30] Superko HR, Roberts R, Garrett B, et al. Family coronary heart disease: a call to action. Clin Cardiol 2010;33:E1–6. doi:10.1002/clc.20684.

[31] Piepoli MF, Hoes AW, Agewall S, et al. 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). *European* heart journal 2016 2016;37:2315–81. doi:10.1093/eurheartj/ehw106.

[32] van Smeden M, Groenwold RHH, Moons KG. A cautionary note on the use of the missing indicator method for handling missing data in prediction research. J Clin Epidemiol 2020;125:188–90. doi:10.1016/j.jclinepi.2020.06.007.