Title

Handling missing values in machine learning to predict patient-specific risk of adverse cardiac events: Insights from REFINE SPECT registry

Authors

Rios, Richard
Miller, Robert JH
Manral, Nipun
et al.

Publication Date

2022-06-01

DOI

10.1016/j.compbiomed.2022.105449

Peer reviewed

HHS Public Access

Author manuscript

Comput Biol Med. Author manuscript; available in PMC 2023 June 01.

Published in final edited form as:

Comput Biol Med. 2022 June ; 145: 105449. doi:10.1016/j.compbiomed.2022.105449.

Handling missing values in machine learning to predict patient-specific risk of adverse cardiac events: insights from REFINE SPECT registry

Richard Rios[1,2], **Robert J.H. Miller**[1,3], **Nipun Manral**[1], **Tali Sharir**[4,5], **Andrew J. Einstein**[6,7], **Mathews B. Fish**[8], **Terrence D. Ruddy**[9], **Philipp A. Kaufmann**[10], **Albert J. Sinusas**[11], **Edward J. Miller**[11], **Timothy M. Bateman**[12], **Sharmila Dorbala**[13], **Marcelo Di Carli**[13], **Serge D. Van Kriekinge**[1], **Paul B. Kavanagh**[1], **Tejas Parekh**[1], **Joanna X. Liang**[1], **Damini Dey**[1], **Daniel S. Berman**[1], **Piotr J. Slomka**[1]

[1]Departments of Medicine (Division of Artificial Intelligence in Medicine), Imaging, and Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA, USA

[2]Universidad Nacional de Colombia, Sede de La Paz, GAUNAL, La Paz, Colombia

[3]Department of Cardiac Sciences, University of Calgary, Calgary, AB, Canada

[4]Department of Nuclear Cardiology, Assuta Medical Center, Tel Aviv, Israel

[5]Faculty of Health Sciences, Ben Gurion University of the Negev, Beer Sheba, Israel

[6]Division of Cardiology, Department of Medicine, Columbia University Irving Medical Center and New York-Presbyterian Hospital, New York, NY, USA

[7]Department of Radiology, Columbia University Irving Medical Center and New York-Presbyterian Hospital, New York, NY, USA

[8]Department of Nuclear Medicine, Oregon Heart and Vascular Institute, Sacred Heart Medical Center, Springfield, OR, USA

[9]Division of Cardiology, University of Ottawa Heart Institute, Ottawa, ON, Canada.

Corresponding author: Piotr Slomka, PhD, Cedars-Sinai Medical Center, 8700 Beverly Boulevard, Metro 203, Los Angeles, California 90048, Phone: 310-423-4348, Fax: 310-423-0173, piotr.slomka@cshs.org.

[8]Conflict of Interest Statement

Drs. Berman, Van Kriekinge, and Slomka, and Mr. Kavanagh participate in software royalties for QPS software at Cedars-Sinai Medical Center. Dr. Slomka has received research grant support from Siemens Medical Systems. Drs. Berman, Dorbala, Einstein, and Edward Miller have served as consultants for GE Healthcare. Dr. Dorbala has served as a consultant to Bracco Diagnostics; her institution has received grant support from Astellas. Dr. Di Carli has received research grant support from Spectrum Dynamics and consulting honoraria from Sanofi and GE Healthcare. Dr. Ruddy has received research grant support from GE Healthcare and Advanced Accelerator Applications Dr. Einstein's has served as a consultant to W. L. Gore & Associates and his institution has received research support from Toshiba America Medical Systems, Roche Medical Systems, and W. L. Gore & Associates. Dr. Edward Miller has served as a consultant for Bracco Inc; and he and his institution has received grant support from Bracco Inc. Dr. Berman's institution has received grant support from HeartFlow. All other authors have reported that they have no relationships relevant to the contents of this paper to disclose.

[10]Department of Nuclear Medicine, Cardiac Imaging, University Hospital Zurich, Zurich, Switzerland

[11]Department of Internal Medicine, Section of Cardiovascular Medicine, Yale University School of Medicine, New Haven, CT, USA.

[12]Cardiovascular Imaging Technologies LLC, Kansas City, MO, USA.

[13]Division of Nuclear Medicine and Molecular Imaging, Department of Radiology, Brigham and Women's Hospital, Boston, MA, USA.

Abstract

Background: Machine learning (ML) models can improve prediction of major adverse cardiovascular events (MACE), but in clinical practice some variables may be missing. We evaluated the influence of missing values in ML models for patient-specific prediction of MACE risk.

Methods: We included 20,179 patients from the multicenter REFINE SPECT registry with MACE follow-up data. We evaluated seven methods for handling missing values: 1) removal of variables with missing values (ML-Remove), 2) imputation with median and unique category for continuous and categorical variables, respectively (ML-Traditional), 3) unique category for missing variables (ML-Unique), 4) cluster-based imputation (ML-Cluster), 5) regression-based imputation (ML-Regression), 6) Miss-Ranger imputation (ML-MR), and 7) multiple imputation (ML-MICE). We trained ML models with full data and simulated missing values in testing patients. Prediction performance was evaluated using area under the receiver-operating characteristic curve (AUC) and compared with a model without missing values (ML-All), expert visual diagnosis and total perfusion deficit (TPD).

Results: During mean follow-up of 4.7±1.5 years, 3,541 patients experienced at least one MACE (3.7% annualized risk). ML-All (reference model-no missing values) had AUC 0.799 for MACE risk prediction. All seven models with missing values had lower AUC (ML-Remove: 0.778, ML-MICE: 0.774, ML-Cluster: 0.771, ML-Traditional: 0.771, ML-Regression: 0.770, ML-MR: 0.766, and ML-Unique: 0.766; $p<0.01$ for ML-Remove vs remaining methods). Stress TPD (AUC 0.698) and visual diagnosis (0.681) had the lowest AUCs.

Conclusion: Missing values reduce the accuracy of ML models when predicting MACE risk. Removing variables with missing values and retraining the model may yield superior patient-level prediction performance.

Keywords

Machine learning; clinical implementation; missing values; prognosis; myocardial perfusion imaging

1. Introduction

Myocardial perfusion imaging (MPI) is frequently used for risk stratification in patients with known or suspected coronary artery disease (CAD) [1,2] Machine learning (ML) can be utilized to improve prediction of major adverse cardiovascular events (MACE)[3], predict

revascularization[4], or select patients for rest scan cancellation[5]. Guidelines suggest reporting the risk for adverse outcomes during clinical interpretation[6]. Recently it has been shown that machine learning (ML) models including fewer variables have higher prediction performance for MACE risk compared to standard interpretation methods and traditional clinical models[7,8]. These ML models have the potential of being implemented in clinical practice to provide patient-specific estimations of MACE risk.

However, in clinical practice, missing values in some clinical variables are unavoidable since patient questionnaires may be incomplete or stress tests may not yet be interpreted. This missing data could significantly reduce the accuracy of ML models, but the impact of missing values on prognostic accuracy has not been rigorously evaluated. Additionally, while dedicated methods exist for handling missing values in ML models [9-11], it remains unknown how these influence the accuracy of patient-specific risk estimations, especially after developing. The aim of this study was to evaluate the influence of missing values in variables on the accuracy of ML models to assess individual MACE risk using seven distinct methods for handling missing values.

2. Material and Methods

2.1 Study Population

The cohort included 20,414 consecutive patients referred for clinically indicated myocardial perfusion imaging from 2009 to 2014 at 5 centers. The institutional review boards at each center approved local data collection and the institutional review board at Cedars-Sinai Medical Center approved the overall registry. The study complies with the Declaration of Helsinki. To the extent allowed by data sharing agreements and IRB protocols, the data from this manuscript will be shared upon written request.

2.2. Data pre-processing

Patients with missing values (n = 235) in any study variable were excluded to avoid confounding factors during the training and testing process, leaving 20,179 patients with all values available for all variables. During data pre-processing no values were imputed (there were no missing values in the included dataset) and integer encoding was performed for categorical variables.

2.2 Clinical Data

Clinical data included: age, sex, body mass index, past medical history, symptoms, and family history of coronary artery disease (CAD). Past medical history included: diabetes mellitus, hypertension, dyslipidemia, smoking, previous myocardial infarction, previous percutaneous coronary intervention, and prior coronary artery bypass grafting [12].

2.3 Imaging Protocols

Studies were performed as previously reported[12]. Patients underwent same-day rest/stress (58.1%), stress/rest (27.3%), stress-only (13.5%), or 2-day stress/rest (1.1%) imaging protocols. Stress results included exercise duration, stress-induced symptoms, resting and

stress heart rate, systolic blood pressure (SBP), diastolic blood pressure (DBP), and electrocardiogram (ECG) changes.

2.4 Visual Analysis

Expert visual interpretation was performed with access to all data during clinical reporting, including clinical, stress test, and local quantitative imaging results. Studies were interpreted using either a 4-point visual scale as 0-normal, 1-probably normal, 2-equivocal, and 3-abnormal or summed stress scores (SSS) [12]. SSS were re-classified to reader diagnosis as normal (SSS = 0), probably normal (SSS = 1), equivocal (SSS = 2 to 3), or abnormal (SSS ≥4)[12].

2.5 Quantitative Analysis

Images were anonymized and transferred to the core laboratory at Cedars-Sinai Medical Center. Quantitative Perfusion/Gated SPECT software (QGS+QPS, Cedars- Sinai Medical Center, Los Angeles, CA) was used for automatic quantification of imaging variables including stress total perfusion deficit (TPD), left ventricular (LV) volumes, and phase analysis parameters [13,14]. Stress TPD was quantified in the default imaging position, as described previously[15]. Combined 2-position stress TPD was obtained as previously described [16]. A full list of variables is available in Table S1.

2.6 Primary Outcome

The primary outcome was MACE, which included all-cause mortality, non-fatal myocardial infarction, admission for unstable angina, or late coronary revascularization (>90 days after imaging). All-cause mortality was determined from the Social Security Death Index for US sites, Ministry of Health National Death Database for Israel, and through the Open Architecture Clinical Information System in Canada. Non-fatal myocardial infarction was defined based on hospital admission for chest pain, elevated cardiac enzyme levels, and typical ECG changes [17] Admission for unstable angina was defined as hospital admission for cardiac chest pain without elevated cardiac enzymes. All non-fatal events were adjudicated by experienced cardiologists after reviewing all available clinical, laboratory, and imaging information.

2.7 Machine Learning

Figure 1 presents an overview of the ML pathway followed in this study. Based on our previous work[3], we considered a reduced ML model for MACE risk prediction, including: 9 imaging, 8 clinical, and 6 stress test variables, see Table S1. Extreme gradient boosting (XGBoost) and random forest (RF) were used to build ML models and evaluate methods for handling missing values, seeking to determine whether the influence of missing values on patient-specific predictions holds disregarding the learning technique[11,18].

2.7.1 Method for handling missing values—Methods for handling missing values are detailed and depicted in Table 1 and Figures 2-4. The imputation models were developed using data from the training populations to avoid overfitting. The following methods for handling missing values were assessed: 1) removal of variables with missing values and

retraining the ML model (ML-Remove), 2) imputation of unique "unknown" category – e.g., NA – for all missing variables to introduce a sparsity pattern, seeking to use the default method for handling missing values of XGBoost or RF (ML-Unique), see supplemental material for specific details, 3) imputation of values with the population median for continuous variables and a distinct missing category for categorical variables (ML-Traditional), 4) clustering patients based on known variables and subsequently imputing values with cluster-based medians for continuous variables or with a distinct missing category for categorical variables (ML-Cluster), 5) estimation of continuous values based on variables with complete data using linear regression or with a distinct missing category for categorical variables (ML-Regression), 6) non-parametric estimation of missing values using the MissRanger method, which builds a random forest model for each variable with the observed values (ML-MR), see supplemental material for specific details, and 7) estimation of multiple values for missing values using multiple imputation by chained equations [19] (ML-MICE). The last four methods (ML-Cluster, ML-Regression, ML-MR, and ML-MICE) are described in detail in the supplemental material. Additional details for the cluster-based imputation are shown in fig S1-4.

2.7.2 Simulation of multiple missing values—We used 10-fold cross validation for training (90% of data) and testing (10% of data) ML models. In the stratified 10-fold cross-validation procedure, the population was randomly split into 10 equally sized folds (which included a similar proportion of patients with MACE and representation from all 5 sites). Each fold had a separate model trained using 90% of the data, which was not used in any way during model testing. ML models were initially built with XGBoost and RF using the training set and full data (no missing values), where corresponding hyperparameters were optimized with an internal 5-fold cross validation procedure following a grid search method. Later, missing values were simulated for all patients in the testing set (which were not used in any way during training). We considered clinical and stress test variables that could be missing for reasons like incomplete data questionnaires or advance Interpretation of stress test data, see Table 2. Imaging variables were not considered since these variables can automatically be obtained with quantitative software directly from images and therefore should always be available for ML prediction. We simulated missing values for all patients to assess the influence on the accuracy of patient-specific predictions, where values can either be present or missing. Imputation methods were then used to estimate the variables that were missing in the testing set. Imputation models and parameters were optimized or derived using instances in the training set (i.e., instances with complete data). ML-Remove models were retrained without the variables with missing values and tested with full patient-specific data. For ML-MICE, ML-MR, ML-Regression, ML-Traditional, ML-Cluster, and ML-Unique, simulated missing data were replaced with the imputed values, where population median values or model-based imputations were initially derived in the training set with full patient-specific data. These ML methods were compared to the model built and tested with all variables without missing values (ML-All).

2.7.3 Estimation performance for missing values of imputation methods—We evaluated the performance of the imputation methods to estimate the simulated missing

values in the testing sets. For all methods, we used the following metrics for continuous and categorical variables:

$$Mean\ absolute\ percentage\ error\ (MAPE) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y-\hat{y}}{y}\right|$$

$$Mean\ absolute\ error\ (MAE) = \frac{1}{n}\sum_{i=1}^{n}|\ y_i - \hat{y}_i\ |$$

$$NHD = \frac{1}{n}\sum_{i=1}^{n} 1\cdot f(y_i,\hat{y}_i),\quad f(y_i,\hat{y}_i) = |\ sign(y_i-\hat{y}_i)\ | = \begin{cases}1, & if\ y_i \neq \hat{y}_i\\ 0, & if\ y_i = \hat{y}_i\end{cases}$$

Where y and $\hat{y}$ were the actual and estimated output values, respectively. Meanwhile, n was the number of instances. We used MAPE to evaluate the imputation error for Resting HR and BMI. Similarly, to avoid division by zero, we used MAE to evaluate the estimation error for Magnitude of ST deviation. The metric NHD can be observed as a normalized version of the hamming distance for two vectors, which counts the number of values or components that are different between two vectors[20]. For the multiple imputation method (ML-MICE), we firstly aggregated the multiple imputations into a single estimation before evaluating the imputation error with MAPE, MAE, and NHD. For example, for continuous variables, $\hat{y}$ was the average of the multiple estimations, while for categorical variables, $\hat{y}$ was the most frequent estimated category (i.e., the majority category).

2.7.4 Simulation of Increasing Number of Variables with Missing Values—

We performed an additional analysis to determine the threshold at which the number of variables missing with missing values significantly impact prediction performance. For this analysis, we also compared the two best strategies. The number of variables with missing values were increased following the ranking of variable importance derived in our previous work[7]. This order was: indication for test, resting heart rate, body mass index, ECG response, Symptoms, ST deviation, and clinical response to stress. Several ML models were then built with an increasing number of variables with missing data.

2.8 Prediction Performance

Prediction performance was evaluated using area under the receiver-operating characteristic curve (AUC) and compared with the performance of 4-point scale visual diagnosis and stress total perfusion deficit (TPD) variables. All models and statistics were implemented in R language (version 4.0.3), using the following open-source packages: *xgboost* (version 1.2.0.1) [11], randomforest (4.6.14), survival (3.2.7), *PredictABEL (version 1.2.−4)*, and pROC (1.16.2).

3. Results

3.1 Study Population

The baseline clinical characteristics of the study population (n = 20,179) are shown in Table S2. During mean follow-up of 4.7 ± 1.5 years, 3495 patients experienced at least one MACE including: 1,617 deaths, 379 MI, 1,895 late revascularizations, and 300 admissions for unstable angina. The annual rate of MACE was 3.7%. In the study population, patients

who experienced MACE were older (mean age 68 vs 63, p<0.001) and more likely to have a history of diabetes (38% vs 23%, p<0.001) or previous percutaneous coronary intervention (34% vs 16%, p<0.001).

3.2 MACE Prediction with Missing Values

With XGboost as the ML technique, the best MACE risk prediction was obtained when no variables were missing (ML-All AUC: 0.799, 95% confidence interval [CI]: 0.792 to 0.807, p<0.01 vs all), see Figure 5. The second-best performance was obtained by removing variables with missing values during model training (ML-Remove AUC: 0.778, 95% CI: 0.770 to 0.786), followed by multiple imputation by chained equations, traditional, cluster-based, and regression imputation methods, see Table 1. ML-MR (AUC: 0.766, 95% CI: 0.758 to 0.774) and ML-Unique (AUC: 0.766, 95% CI: 0.758 to 0.774) obtained the lowest prognostic accuracy of the ML models, but this was still higher compared with stress TPD (AUC: 0.698, 95% CI: 0.688 to 0.708) and visual diagnosis (AUC: 0.681, 95% CI: 0.671 to 0.691; p < 0.01 for ML-MR vs stress TPD and Diagnosis).

Similarly, with RF as ML technique, the best MACE risk prediction was obtained when no variables were missing (ML-All AUC: 0.792, 95% confidence interval [CI]: 0.784 to 0.800, p<0.01 vs all), see Figure 6. Also, ML-MR (AUC: 0.755, 95% CI: 0.746 to 0.763) obtained the lowest prognostic accuracy of the ML models, but this was still higher compared with stress TPD (AUC: 0.698, 95% CI: 0.688 to 0.708) and visual diagnosis (AUC: 0.681, 95% CI: 0.671 to 0.691; p < 0.01 for ML-MR vs stress TPD and Diagnosis). Unlike with XGBoost, the second-best performance was obtained by removing variables with missing values during model training and multiple imputation (ML-Remove AUC: 0.77, 95% CI: 0.762 to 0.778 vs ML-MICE AUC: 0.77, 95% CI: 0.762 to 0.779; $p = 0.732$), followed by regression, traditional, and cluster-based imputation methods.

Table 3 provides the prediction performance of each imputation methods for estimating missing values. For continuous variables, ML-Traditional, ML-Cluster, and ML-Regression obtained in overall the best prediction performances. Meanwhile, for categorical variables, ML-MR obtained the best prediction performance.

3.3 Robustness of ML models with Increasing Number of Variables with Missing Values

We compared the prediction performance of ML-Remove and ML-MICE as a function of the number of variables with missing values, see Figure 7. The prediction performance of ML-Remove decreased significantly (decreased AUC of more than 1%) after three variables with missing values were present. Meanwhile, the prediction performance of ML-MICE decreased significantly when two or more variables with missing values were present.

4. Discussion

In this study we evaluated the impact of missing values on the prognostic accuracy of ML models which incorporate clinical, stress-test, and imaging variables when performing patient-specific evaluations of cardiovascular risk. We compared six different methods for imputing missing variables as well as a model retrained without the variables with missing variables. We show that the best strategy for handling missing values was to rebuild the ML

model, removing variables with missing values, to avoid bias introduced by imputation. The second-best strategy was to estimate multiple possible values (ML-MICE) for the missing values to account uncertainties in the imputation around the true value. Of the remaining imputation methods, ML-Regression, ML-cluster, and ML-traditional achieved similarly high accuracy followed by ML-Unique and ML-MR. Importantly, all of the ML models had higher prognostic accuracy compared to standard interpretation methods. In prospective studies assessing clinical implementations of ML for risk prediction, the effect of missing data in ML models should be considered to avoid possible model biases[21-23].

ML models have been increasingly applied to many aspects of cardiovascular imaging[24-27]. The strength of ML is the ability to objectively integrate a variety of clinical, stress, and imaging variables to predict the outcome of interest. However, in clinical practice as the number of required variables increases some of this information will invariably be missing. These missing values could adversely impact the accuracy of predictions -but this issue has not been explored in depth. Previously, we demonstrated that a ML model incorporating only 9 of 32 manually collected variables could achieve >99.5% of the prediction performance of the full model [3]. In the current work we demonstrate that missing values reduce the accuracy of ML models in prediction of MACE risk. We show that imputing values for missing variables introduced biases into ML models that decrease prediction performance. The most accurate solution for handling missing values was to develop a ML model without the variables (ML-Remove).

To date, the most accurate method to impute missing values has not been well established, therefore we evaluated several frequently used techniques [3-5,7,28]. Mishra et al.[29] previously demonstrated that multiple imputation using either regression or propensity-based methods significantly reduced mean square error compared to last observation carried forward imputation. Rusdah et al. demonstrated that the XGBoost default imputation method demonstrated higher accuracy compared to imputation with either k-nearest neighbor imputation or population mean[30]. However, in our dataset the default imputation method for XGBoost and RF (ML-Unique) obtained the second lowest accuracy of the ML models, while ML-Remove had the highest accuracy[10]. The XGBoost default method can be considered as a greedy algorithm that does not necessarily lead to the best solution (in this case MACE prediction), i.e., the sparsity-aware split algorithm only ensures that on average a default direction is the best solution given the already traversed decision nodes. The default method for RF based on proximity takes characteristics of ML-Cluster and ML-traditional as it takes averages of non-missing values weighted by proximities (i.e. average of most similar instances). Jerez et al. demonstrated that imputing missing values with k-nearest neighbor achieved higher model accuracy compared to list-wise deletion (removing observations with missing values)[31]. Similarly, Wohlrab et al. demonstrated that several methods for imputation generally outperformed list-wise deletion[32]. However, these imputation methods may eliminate individual characteristics that make a specific patient differ from the overall population (i.e. each missing value was assigned to the same single value – median/mean-or unique category), leading to biased assessments and treatments [30]. In our study, multiple imputation resulted in the highest accuracy of the imputation methods tested. The variability between studies suggests that the optimal method depends on the number of variables available, the extent of bias caused by imputation and potentially other

factors. For cardiovascular applications, the bias related to value imputation is unpredictable but the number of variables potentially available for ML prediction is large. Therefore, generating new ML models without the variables with missing values would be expected to more consistently demonstrate high accuracy.

While our study suggests that ML-Remove is the most accurate method to handle missing values, it may not be clinically feasible to have ML models available for every possible data combination. Therefore, a compromise between feasibility and accuracy is necessary. The best solution may be to develop a limited set of ML models with different configurations [23] For stress-only imaging, for instance, a ML model with only imaging-derived variables can be built to provide fully automatic MACE risk calculation, or semi-automatic by adding a few demographic variables with high importance which are rarely missing (e.g. age, sex, and diabetes). A full ML model can be built to support assessments with all patient-specific values when full data is available to ensure the highest possible accuracy since we also demonstrated that ML-All had significantly higher AUC compared to all other methods. If variables required for the desired ML-model could not be obtained, a population-based imputation method could be utilized such as ML-MICE which achieved only marginally lower accuracy. Based on our results, assigning a unique “unknown” category, which is commonly used and the default method for XGBoost, should not be incorporated in MACE risk estimation. Importantly, if data collection was incomplete, it would be important to warn the user about key missing features with high prediction importance.

We also evaluated the prognostic accuracy of ML models while the number of missing variables were progressively increased. We found that the prognostic accuracy was less affected when variables with missing values were removed (ML-Remove) compared to imputing values (ML-MICE). With ML-Remove, 3 variables with missing values were needed to reduce accuracy by 1%, while with ML-MICE only 2 variables with missing values were required. These results suggest that imputation of a single missing value should not meaningfully impact predictions in clinical practice.

Our study has a few important limitations. We simulated the impact of missing values using variables with the highest missing rate in our registry which is still relatively infrequent. However, missing values in other variables with high importance (such as previous cardiac intervention or age) could have greater impact on patient-specific risk estimation. Nevertheless, based on our results, the best strategy may be to remove variables with missing values rather than imputing data. Additionally, we tested the various models using stratified 10-fold cross validation and separate validation in data from new sites may yield different results.

5. Conclusions

Missing values significantly reduce the accuracy of ML when predicting MACE risk. ML models generated without variables with missing values are the most accurate method to address missing values. Meanwhile, imputing missing values based on population databases is a more clinically feasible alternative, which has clinically similar prediction performance and still higher prediction performance compared to standard interpretation methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

This research was supported in part by grant R01HL089765 from the National Heart, Lung, and Blood Institute/ National Institutes of Health (NHLBI/NIH) (PI: Piotr Slomka). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The work was also supported in part by the Miriam and Sheldon Adelson Medical Research Foundation.

9. References

1. Fihn SD, Gardin JM, Abrams J, et al. 2012 ACCF/AHA/ACP/AATS/PCNA/SCAI/STS Guideline for the diagnosis and management of patients with stable ischemic heart disease: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines, and the American College of Physicians, American Association for Thoracic Surgery, Preventive Cardiovascular Nurses Association, Society for Cardiovascular Angiography and Interventions, and Society of Thoracic Surgeons. J Am Coll Cardiol. 2012;60(24):e44–e164. doi: 10.1016/j.jacc.2012.07.013 [PubMed: 23182125]
2. Knuuti J, Wijns W, Saraste A, et al. 2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromesThe Task Force for the diagnosis and management of chronic coronary syndromes of the European Society of Cardiology (ESC). Eur Heart J. 2020;41(3):407–477. doi:10.1093/eurheartj/ehz425 [PubMed: 31504439]
3. Betancur J, Otaki Y, Motwani M, et al. Prognostic value of combined clinical and myocardial perfusion imaging data using machine learning. JACC Cardiovasc Imaging. 2018;11(7):1000–1009. doi: 10.1016/j.jcmg.2017.07.024 [PubMed: 29055639]
4. Hu LH, Betancur J, Sharir T, et al. Machine learning predicts per-vessel early coronary revascularization after fast myocardial perfusion SPECT: results from multicentre REFINE SPECT registry. Eur Heart J Cardiovasc Imaging. 2020;21(5):549–559. doi: 10.1093/ehjci/jez177 [PubMed: 31317178]
5. Hu LH, Miller RJH, Sharir T, et al. Prognostically safe stress-only single-photon emission computed tomography myocardial perfusion imaging guided by machine learning: report from REFINE SPECT. Eur Heart J Cardiovasc Imaging. 2021;22(6):705–714. doi: 10.1093/ehjci/jeaa134 [PubMed: 32533137]
6. Tilkemeier PL, Bourque J, Doukky R, Sanghani R, Weinberg RL. ASNC imaging guidelines for nuclear cardiology procedures : Standardized reporting of nuclear cardiology procedures. J Nucl Cardiol. 2017;24(6):2064–2128. doi: 10.1007/s12350-017-1057-y [PubMed: 28916938]
7. Rios R, Miller RJH, Hu LH, et al. Determining a Minimum Set of Variables for Machine Learning Cardiovascular Event Prediction: results from REFINE SPECT registry. Cardiovasc Res. Published online Accepted 2020.
8. Haro Alonso D, Wernick MN, Yang Y, Germano G, Berman DS, Slomka P. Prediction of cardiac death after adenosine myocardial perfusion SPECT based on machine learning. J Nucl Cardiol. 2019;26(5):1746–1754. doi: 10.1007/s12350-018-1250-7 [PubMed: 29542015]
9. Little RJA, Rubin DB Statistical Analysis with Missing Data. 2nd ed. John Wiley & Sons, Ltd; 2002.
10. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2d edition. Springer Series in statistics; 2017.
11. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. Published online August 13, 2016:785–794. doi: 10.1145/2939672.2939785
12. Slomka PJ, Betancur J, Liang JX, et al. Rationale and design of the REgistry of Fast Myocardial Perfusion Imaging with NExt generation SPECT (REFINE SPECT). J Nucl Cardiol. Published online June 19, 2018. doi:10.1007/s12350-018-1326-4

13. Nakazato R, Berman DS, Gransar H, et al. Prognostic value of quantitative high-speed myocardial perfusion imaging. J Nucl Cardiol. 2012;19(6):1113–1123. doi:10.1007/s12350-012-9619-5 [PubMed: 23065414]
14. Slomka PJ, Nishina H, Berman DS, et al. Automated quantification of myocardial perfusion SPECT using simplified normal limits. J Nucl Cardiol. 2005;12(1):66–77. [PubMed: 15682367]
15. Otaki Y, Betancur J, Sharir T, et al. 5-Year Prognostic Value of Quantitative Versus Visual MPI in Subtle Perfusion Defects: Results From REFINE SPECT. JACC Cardiovasc Imaging. Published online June 8, 2019. doi:10.1016/j.jcmg.2019.02.028
16. Nishina H, Slomka PJ, Abidov A, et al. Combined supine and prone quantitative myocardial perfusion SPECT: method development and clinical validation in patients with no known coronary artery disease. J Nucl Med. 2006;47(1):51–58. [PubMed: 16391187]
17. Thygesen Kristian, Alpert Joseph S., White Harvey D., et al. Universal Definition of Myocardial Infarction. Circulation. 2007;116(22):2634–2653. doi:10.1161/CIRCULATIONAHA.107.187397 [PubMed: 17951284]
18. Breiman L Random Forests. Mach Learn. 2001;45(1):5–32. doi:10.1023/A:1010933404324
19. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 2011;20(1):40–49. doi:10.1002/mpr.329 [PubMed: 21499542]
20. Deza MM, Deza E. Encyclopedia of Distances. 3rd ed. 2014 edition. Springer; 2014.
21. Chan AW, Tetzlaff JM, Altman DG, et al. SPIRIT 2013 Statement: Defining Standard Protocol Items for Clinical Trials. Ann Intern Med. 2013;158(3):200–207. doi:10.7326/0003-4819-158-3-201302050-00583 [PubMed: 23295957]
22. Chan AW, Tetzlaff JM, Gøtzsche PC, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. BMJ. 2013;346:e7586. doi:10.1136/bmj.e7586 [PubMed: 23303884]
23. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020;26(9):1320–1324. doi:10.1038/s41591-020-1041-y [PubMed: 32908275]
24. Lin A, Kolossváry M, Motwani M, et al. Artificial Intelligence in Cardiovascular Imaging for Risk Stratification in Coronary Artery Disease. Radiol Cardiothorac Imaging. Published online February 25, 2021. doi:10.1148/ryct.2021200512
25. Dey D, Slomka PJ, Leeson P, et al. Artificial Intelligence in Cardiovascular Imaging: JACC State-of-the-Art Review. J Am Coll Cardiol. 2019;73(11):1317–1335. doi:10.1016/j.jacc.2018.12.054 [PubMed: 30898208]
26. Slomka PJ, Miller RJ, Isgum I, Dey D. Application and Translation of Artificial Intelligence to Cardiovascular Imaging in Nuclear Medicine and Noncontrast CT. Semin Nucl Med. 2020;50(4):357–366. doi:10.1053/j.semnuclmed.2020.03.004 [PubMed: 32540032]
27. Petersen SE, Abdulkareem M, Leiner T. Artificial Intelligence Will Transform Cardiac Imaging-Opportunities and Challenges. Front Cardiovasc Med. 2019;6:133. doi:10.3389/fcvm.2019.00133 [PubMed: 31552275]
28. Nakanishi R, Slomka PJ, Rios R, et al. Machine Learning Adds to Clinical and CAC Assessments in Predicting 10-Year CHD and CVD Deaths. JACC Cardiovasc Imaging. 2021;14(3):615–625. doi:10.1016/j.jcmg.2020.08.024 [PubMed: 33129741]
29. Mishra S, Khare D. On comparative performance of multiple imputation methods for moderate to large proportions of missing data in clinical trials: a simulation study. J Med Stat Inform. 2014;2(1):9. doi:10.7243/2053-7662-2-9
30. Rusdah DA, Murfi H. XGBoost in handling missing values for life insurance risk prediction. SN Appl Sci. 2020;2(8):1336. doi:10.1007/s42452-020-3128-y
31. Jerez JM, Molina I, García-Laencina PJ, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med. 2010;50(2):105–115. doi:10.1016/j.artmed.2010.05.002 [PubMed: 20638252]
32. Wohlrab L, Fürnkranz J. A review and comparison of strategies for handling missing values in separate-and-conquer rule learning. J Intell Inf Syst. 2011;36(1):73–98. doi:10.1007/s10844-010-0121-8

HIGHLIGHTS

- Removing variables with missing values and retraining was the most accurate method
- Multiple imputation had the highest accuracy of the imputation methods
- Having a selection of reduced ML models may be a practical clinical solution

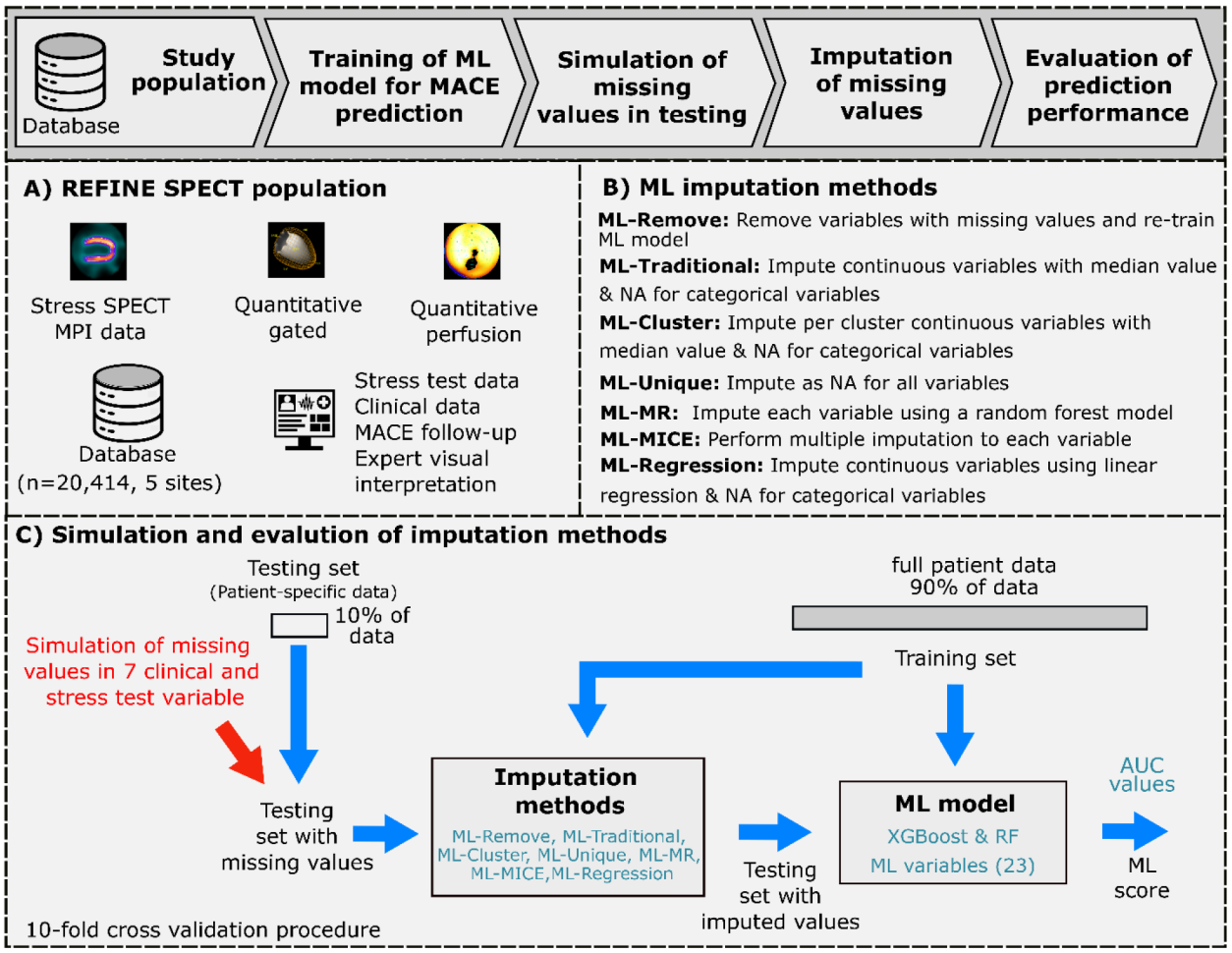
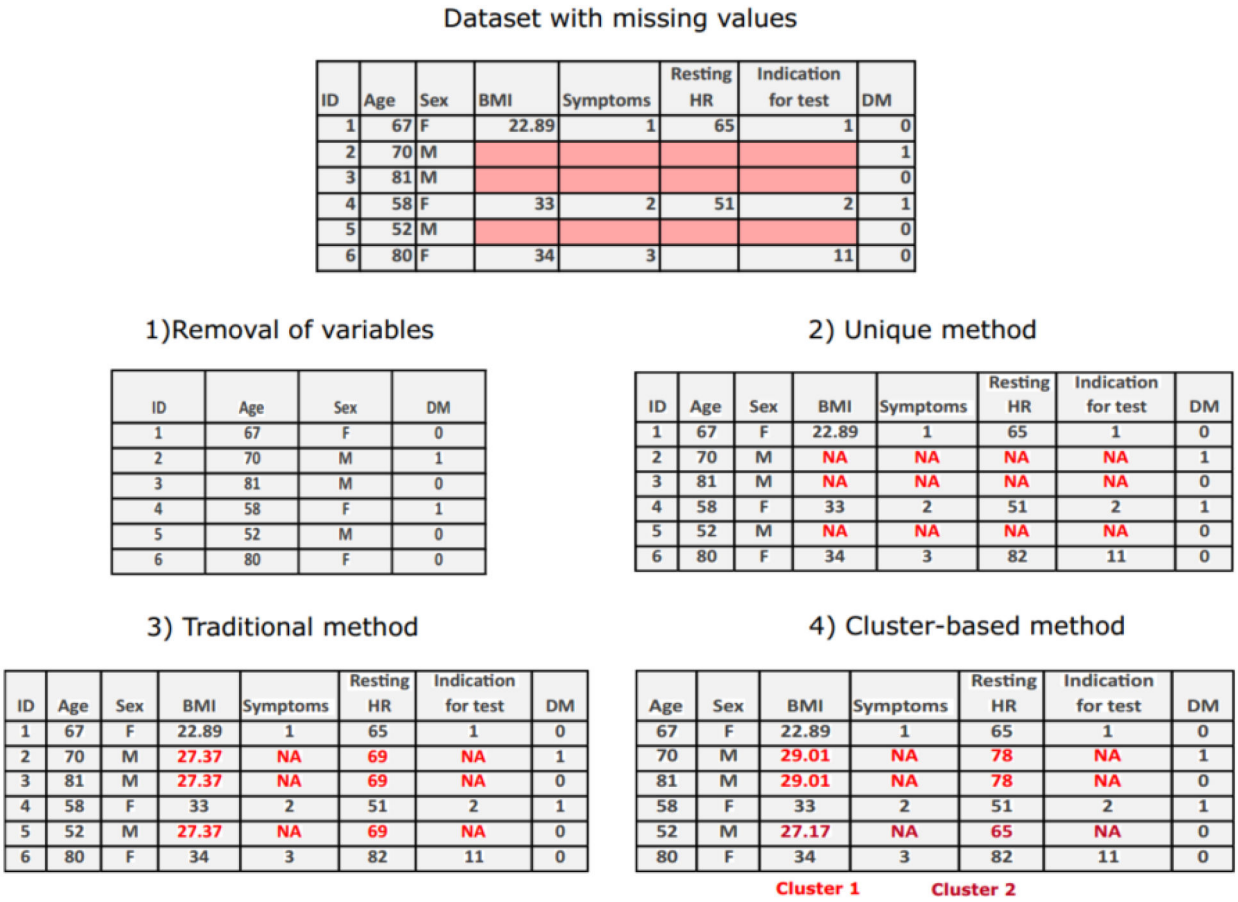


Figure 1: Machine learning workflow.

The machine learning (ML) pathway and validation procedures used in this study. **A:** The REFINE SPECT study population was used to build a ML model for MACE risk prediction, including quantitative gated and perfusion variables, stress and clinical data**. B**) Seven methods were evaluated for handling missing values**. C**) Models were trained and tested using stratified 10-fold cross validation. In the training sets (90% of data with full patient information), we trained ML models with XGBoost and Random Forest (RF) for MACE prediction as well as the imputation models used for estimating missing values. Imputation models were also developed using data from the training set. Subsequently, for some clinical and stress test variables in the testing sets (10% of data), we simulated missing values for all patients and imputed values to evaluate their impact on the prediction performance for patient-specific estimations of MACE risk. ML models were compared to traditional risk models, expert visual interpretation, and stress total perfusion deficit. The method was repeated 10 times, with a separate fold (10% of data) used for testing in each repetition. Abbreviations: AUC – area under the receiver operating characteristic curve, MACE – major adverse cardiovascular events, NA – unique category for missing values, MICE – multiple imputation by chained equations, MR – MissRanger.



Dataset with missing values

ID	Age	Sex	BMI	Symptoms	Resting HR	Indication for test	DM
1	67	F	22.89	1	65	1	0
2	70	M					1
3	81	M					0
4	58	F	33	2	51	2	1
5	52	M					0
6	80	F	34	3		11	0

1)Removal of variables

ID	Age	Sex	DM
1	67	F	0
2	70	M	1
3	81	M	0
4	58	F	1
5	52	M	0
6	80	F	0

2) Unique method

ID	Age	Sex	BMI	Symptoms	Resting HR	Indication for test	DM
1	67	F	22.89	1	65	1	0
2	70	M	NA	NA	NA	NA	1
3	81	M	NA	NA	NA	NA	0
4	58	F	33	2	51	2	1
5	52	M	NA	NA	NA	NA	0
6	80	F	34	3	82	11	0

3) Traditional method

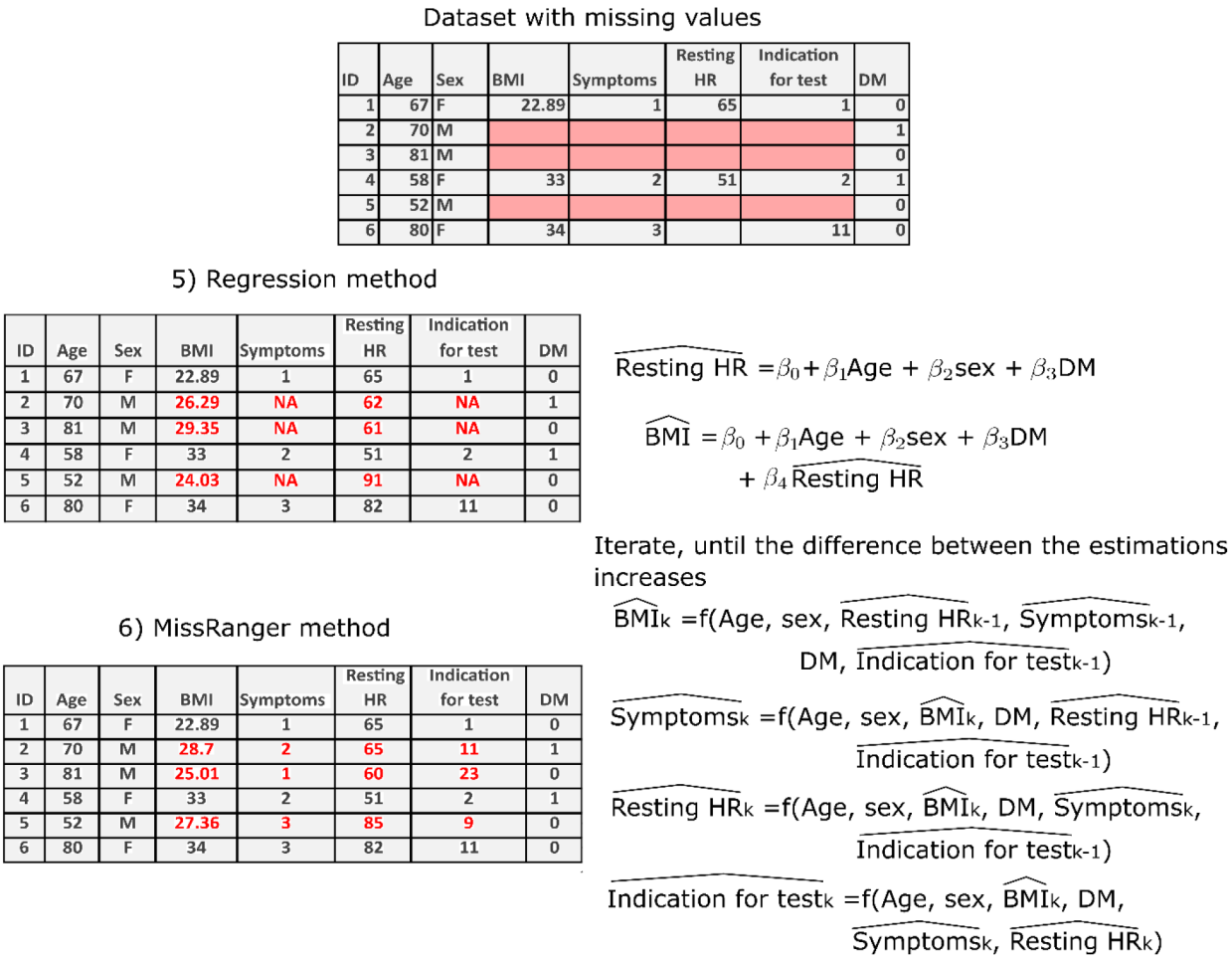
ID	Age	Sex	BMI	Symptoms	Resting HR	Indication for test	DM
1	67	F	22.89	1	65	1	0
2	70	M	27.37	NA	69	NA	1
3	81	M	27.37	NA	69	NA	0
4	58	F	33	2	51	2	1
5	52	M	27.37	NA	69	NA	0
6	80	F	34	3	82	11	0

4) Cluster-based method

Age	Sex	BMI	Symptoms	Resting HR	Indication for test	DM
67	F	22.89	1	65	1	0
70	M	29.01	NA	78	NA	1
81	M	29.01	NA	78	NA	0
58	F	33	2	51	2	1
52	M	27.17	NA	65	NA	0
80	F	34	3	82	11	0

Figure 2. Visual example for imputing missing values with the unique, traditional, and cluster-based methods.

Missing values highlighted with red color on a given dataset were imputed using the following methods: 1) Removal of variables with missing values, 2) unique “unknown (NA)” category imputed for all missing variables (XGBoost default method for handling the missing variables), 3) traditional approach to impute variables with the population median for continuous variables and a distinct missing category for categorical variables, and 4) clustering patients based on known variables and subsequently imputing cluster-based medians for continuous variables or with a distinct missing category for categorical variables. Cluster’s centroids and medians were derived with instances with complete data.



ID	Age	Sex	BMI	Symptoms	Resting HR	Indication for test	DM
1	67	F	22.89	1	65	1	0
2	70	M					1
3	81	M					0
4	58	F	33	2	51	2	1
5	52	M					0
6	80	F	34	3		11	0

ID	Age	Sex	BMI	Symptoms	Resting HR	Indication for test	DM
1	67	F	22.89	1	65	1	0
2	70	M	26.29	NA	62	NA	1
3	81	M	29.35	NA	61	NA	0
4	58	F	33	2	51	2	1
5	52	M	24.03	NA	91	NA	0
6	80	F	34	3	82	11	0

ID	Age	Sex	BMI	Symptoms	Resting HR	Indication for test	DM
1	67	F	22.89	1	65	1	0
2	70	M	28.7	2	65	11	1
3	81	M	25.01	1	60	23	0
4	58	F	33	2	51	2	1
5	52	M	27.36	3	85	9	0
6	80	F	34	3	82	11	0

$$\widehat{\text{Resting HR}} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{sex} + \beta_3 \text{DM}$$

$$\widehat{\text{BMI}} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{sex} + \beta_3 \text{DM} + \beta_4 \widehat{\text{Resting HR}}$$

$$\widehat{\text{BMI}}_k = f(\text{Age, sex, } \widehat{\text{Resting HR}}_{k-1}, \widehat{\text{Symptoms}}_{k-1}, \text{DM, } \widehat{\text{Indication for test}}_{k-1})$$

$$\widehat{\text{Symptoms}}_k = f(\text{Age, sex, } \widehat{\text{BMI}}_k, \text{DM, } \widehat{\text{Resting HR}}_{k-1}, \widehat{\text{Indication for test}}_{k-1})$$

$$\widehat{\text{Resting HR}}_k = f(\text{Age, sex, } \widehat{\text{BMI}}_k, \text{DM, } \widehat{\text{Symptoms}}_k, \widehat{\text{Indication for test}}_{k-1})$$

$$\widehat{\text{Indication for test}}_k = f(\text{Age, sex, } \widehat{\text{BMI}}_k, \text{DM, } \widehat{\text{Symptoms}}_k, \widehat{\text{Resting HR}}_k)$$

Figure 3. Visual example for imputing missing values with the regression and MissRanger method.

Missing values highlighted with red color on a given dataset were imputed using the following methods: 5), sequential estimation of continuous variables using linear models with variables that have complete data, either with observed or estimated values derived from previous steps (e.g. $\widehat{Resting\ HR}$), or with a distinct missing category for categorical variables , 6) non-parametric estimation of missing variables by building a random forest model for each variable based on instances with complete data, either with observed or estimated values derived from previous steps. The iterative procedure is stopped when the difference between the new and old matrix of imputed values increases for the first time with respect to both continuous and categorical variables. The function f represents the random forest model. For both imputation methods, model parameters are optimized using instances with complete data.

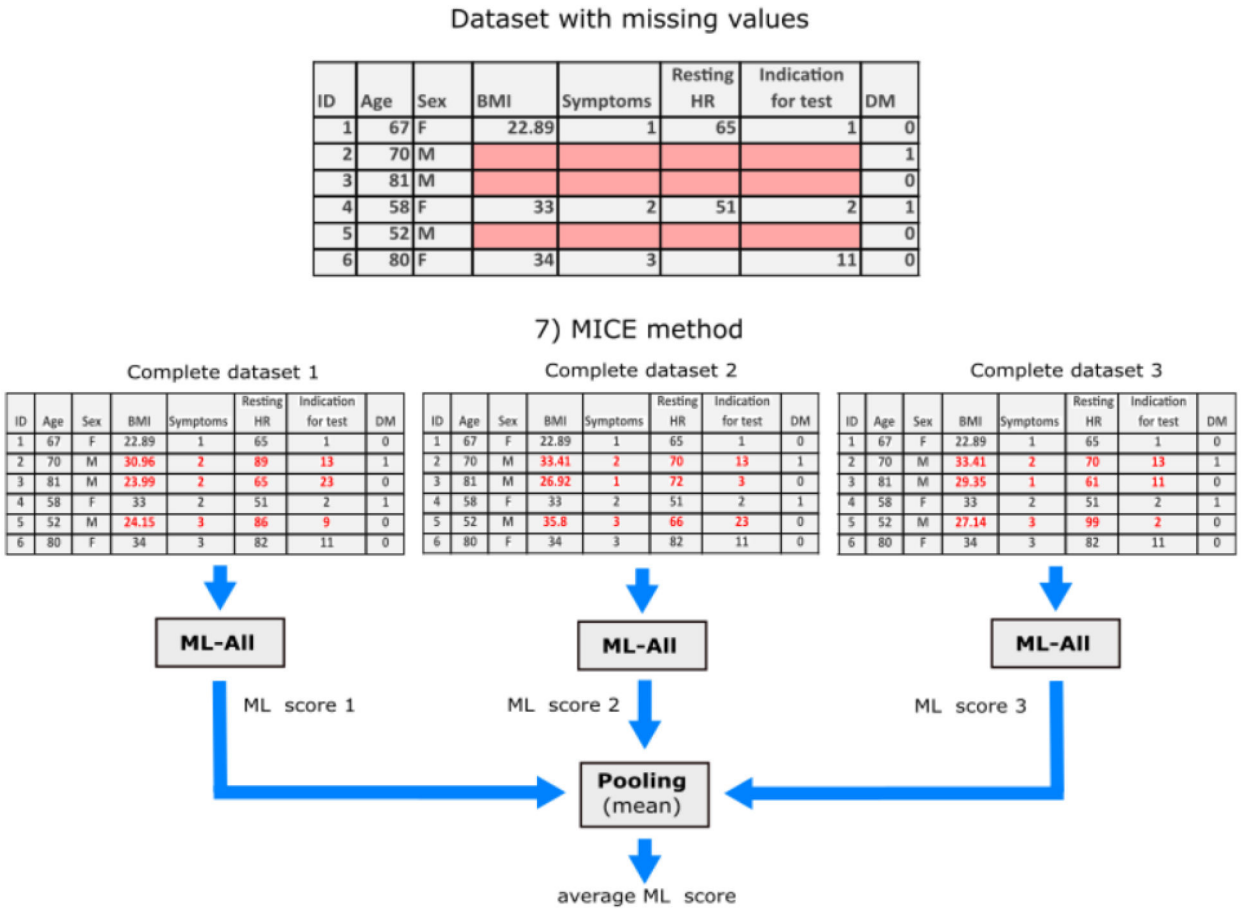


Figure 4. Visual example for imputing missing values using multiple imputation by chained equations (MICE).

Missing values highlighted with red color on a given dataset were imputed with multiple imputation. It was created $m = 3$ "complete" datasets to predict the risk of MACE with the ML-All for each patient in the testing set of the 10-fold cross validation procedure (i.e. the desired analysis). Later, the ML scores for MACE risk were pooled into a single prediction by averaging the ML scores.

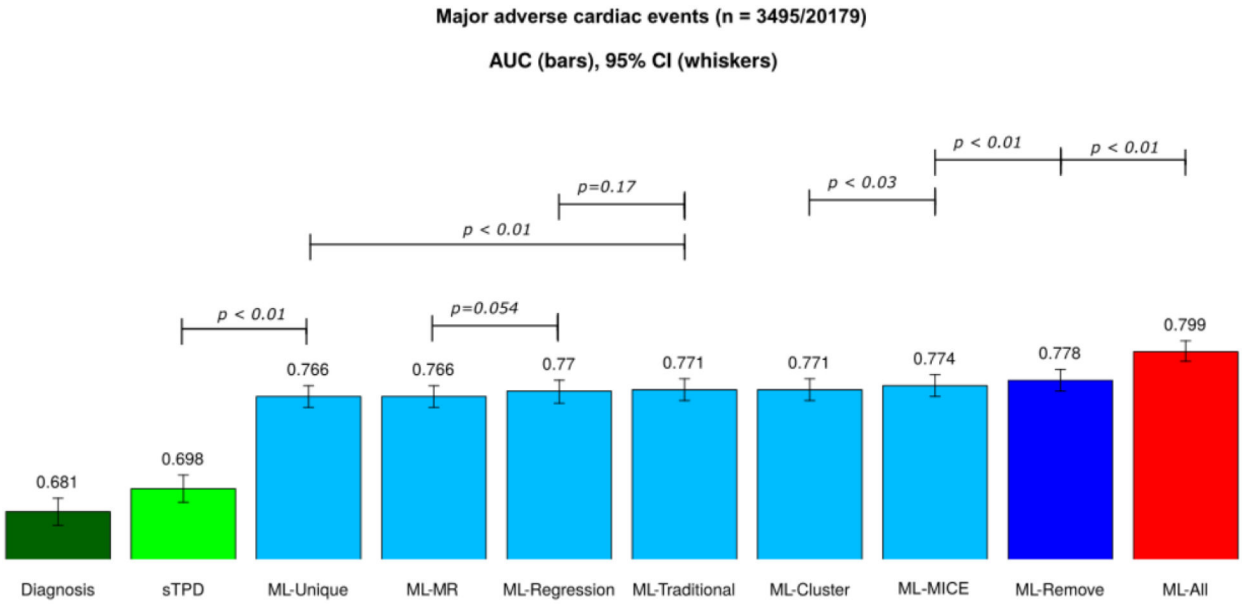


Figure 5. Performance prediction for MACE risk with multiple missing values.

Receiver-operating characteristic (ROC) curves for the machine learning (ML) and standard interpretation methods. ML-All included all variables with patient specific values. ML-Remove was developed with variables having missing values removed. For ML-MICE, ML-MR, ML-Regression, ML-Traditional, ML-Unique, and ML-cluster all missing values were imputed to assess the impact on risk-estimation for an individual patient. Abbreviations: MACE – major adverse cardiovascular events, AUC – area under the ROC curve, CI – confidence interval, sTPD – stress total perfusion deficit.

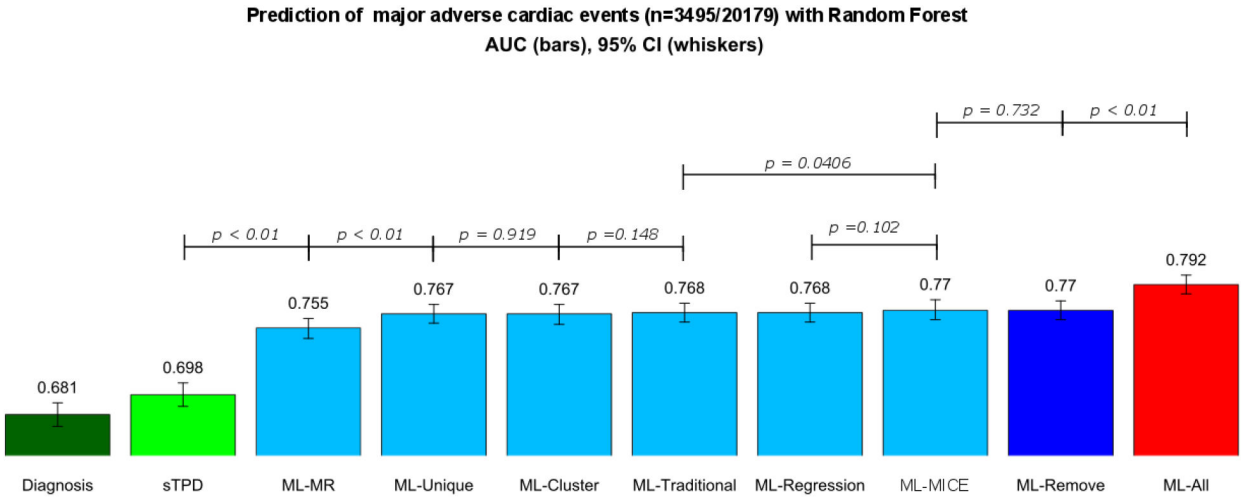


Figure 6. Performance prediction for MACE risk with random forest (RF) and multiple missing values.

Receiver-operating characteristic (ROC) curves for the machine learning (ML) and standard interpretation methods. ML-All included all variables with patient specific values. ML-Remove was developed with variables having missing values removed. For ML-MICE, ML-MR, ML-Regression, ML-Traditional, ML-Unique, and ML-cluster all missing values were imputed to assess the impact on risk-estimation for an individual patient. Abbreviations: MACE – major adverse cardiovascular events, AUC – area under the ROC curve, CI – confidence interval, sTPD – stress total perfusion deficit.

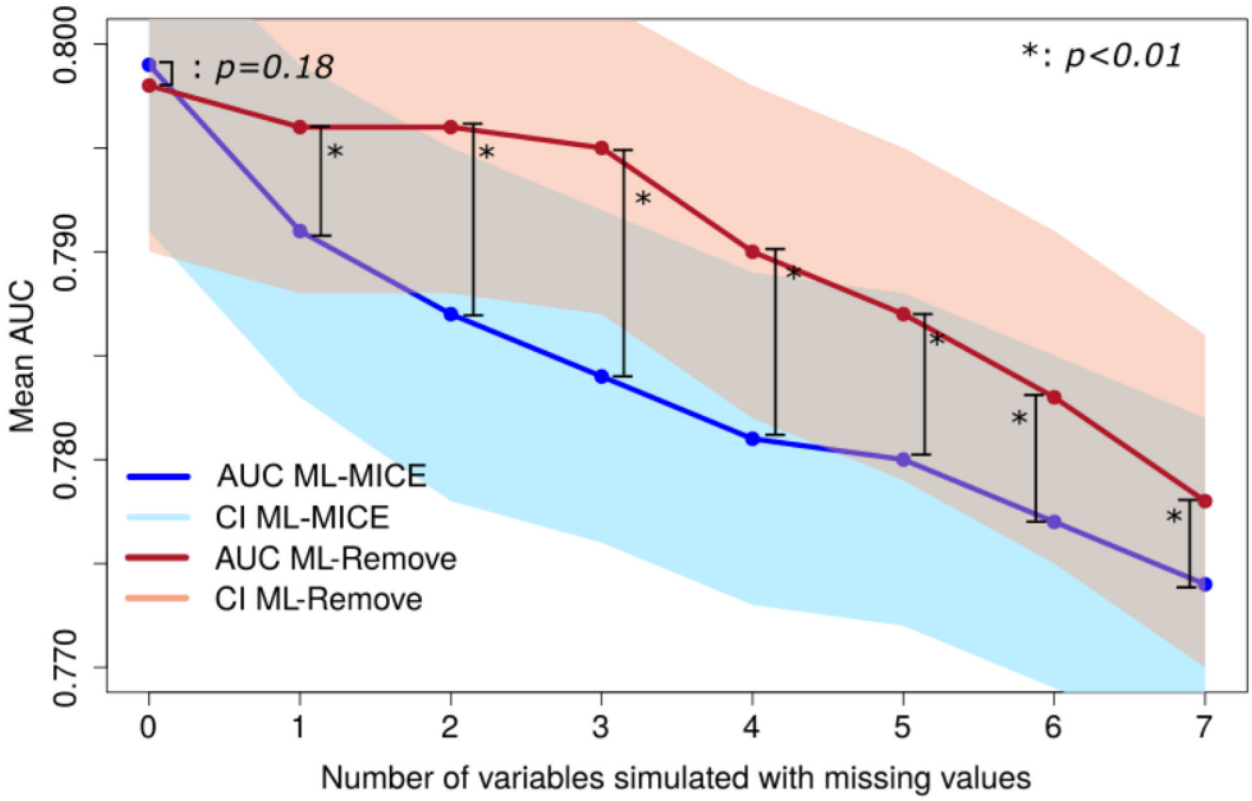


Figure 7. Performance prediction for MACE risk with increasing number of variables with missing values.

Area under the receiver-operating characteristic curve (AUC) and 95% confidence intervals (CI) for machine learning (ML) models as the number of simulated missing values increases. Missing values were imputed using ML-Remove and ML-MICE methods –the two best strategies for handling missing values. Abbreviations: MACE – major adverse cardiovascular events, AUC – area under the ROC, CI – confidence interval.

Table 1:

Methods for handling missing values.

Abbreviations: MI – multiple imputation by chained equations, ML – machine learning, MR – MissRanger.

Retraining Methods			
	Continuous and categorical Variables		**AUC (95% CI)**
ML-Remove	Variables with missing values were removed and a new ML model was trained using the training data set.		0.778 (0.770 to 0.786)
Imputation Methods			
	Continuous Variables	**Categorical Variables**	**AUC (95% CI)**
ML-Unique	Distinct missing category	Distinct missing category	0.766 (0.758 to 0.774)
ML-Traditional	Population median	Distinct missing category	0.771 (0.763 to 0.779)
ML-Cluster	Median value from population cluster	Distinct missing category	0.771 (0.763 to 0.779)
ML-Regression	Linear regression-based estimate using variables with complete values	Distinct missing category	0.770 (0.761 to 0.778)
ML-MR	Random Forest-based estimate with complete data		0.766 (0.758 to 0.774)
ML-MICE	Multivariate imputation using chained equations with complete data		0.774 (0.766 to 0.782)

Table 2.

List of variables with simulated missing values.

The reasons for missing values could potentially be related to incomplete data questionnaires or advance interpretation of stress test data.

Name	Type of variable	Potential Reason for missing value
Resting heart rate	Clinical variable	Incomplete patient questionnaires
Body mass index	Clinical variable	Incomplete patient questionnaires
Symptoms	Clinical variable	Incomplete patient questionnaires
Stress peak heart rate (HR)	Stress-test variable	Advanced interpretation of stress test data
ECG response to stress	Stress-test variable	Advanced interpretation of stress test data
ST deviation	Stress-test variable	Advanced interpretation of stress test data
Clinical Response to stress	Stress-test variable	Advanced interpretation of stress test data

Table 3.

Estimation performance for missing values of each imputation methods.

Error metrics to evaluate the estimation performance for missing values of each imputation method. For categorical variables, NHD describes a normalized version of the hamming distance for two vectors, which counts the number of values or components that are different between two vectors. Meanwhile, for continuous variables, mean absolute error (MAE) describes on average how far the imputations for missing values were off from their actual values. Similarly, mean absolute percentage error (MAPE) describes on average the percentage of how far the imputations for missing values were off from their actual values.

	ML-Traditional	ML-Cluster	ML-Regression	ML-MR	ML-MICE
$MAPE_{Resting\ HR}$	0.151	0.149	0.144	0.179	0.168
$MAPE_{BMI}$	0.143	0.143	0.147	0.173	0.169
$MAE_{Magnitude\ ST\ deviation}$	0.284	0.284	0.357	0.416	0.313
$NHD_{symptoms}$	----	----	----	0.604	0.95
$NHD_{ECG\ response\ to\ stress}$	----	----	----	0.508	0.403
$NHD_{Clinical\ Response\ to\ stress}$	----	----	----	0.393	0.404
$NHD_{Indication\ for\ test}$	----	----	----	0.978	0.998