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

Nijman SWJa, 1, †, Oberman HIb, 1,†, Brandjes Mc, Jacobs JJLc, Bots MLa, Asselbergs FWdef, Moons KGMa, Vink Gb, †, Debray TPAaf, †, Smeden van Ma, †

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

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

c Department of Health, Ortec B.V. Zoetermeer, The Netherlands;

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

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

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

1 First author

† Equal contribution

Corresponding author: Steven WJ Nijman, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX, Utrecht, the Netherlands. E-mail: s.w.j.nijman@umcutrecht.nl. Telephone: +31-(0)88-75 6801

# Abstract

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

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

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

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

# Introduction

Incompleteness of medical records is a ubiquitous problem when using healthcare data. Besides the well-documented issues that missing data can create in data analyses, incompleteness of medical records may also create practical issues in clinical practice (1,2). For instance, a prediction model that relies on historical but unrecorded data for a particular patient or prediction models that are used as early-warning systems for individual patients (3,4). Most prediction models are not designed to be used when predictors are not fully observed, and ad-hoc approaches such as replacing the missing value with the population average value (i.e., mean imputation) is generally not advised (1,5). As prediction models are increasingly being integrated in the electronic health record (EHR) via clinical decision support systems (CDSS), the issues concerning missing data and the need to deal with those missing values when applying prediction models in individual patients becomes more evident (6,7). The issue is further complicated as the common strategies to mend or circumvent missing data in research are not directly applicable for use when predicting an outcome for an individual patient in a clinical practice setting.

Various strategies to handle different manifestations of missing data have been studied thoroughly and can usually provide more plausible substitution values (e.g., via imputation) (2). Multiple imputation is often considered to be the gold standard for missing data problems and is known to provide valid estimates and correct standard errors in circumstances where the missingness does not depend on the unobserved values (8). Most imputation algorithms, however, require direct access to data from multiple instances (i.e., multiple patients or multiple measurements) and are therefore not directly suitable for use on a case-by-case basis. Further, when a prediction model is applied to a single patient in clinical practice via a CDSS there is usually no access to any data from other individuals due to computational and privacy constraints [ref].

An intuitive alternative to imputation is to solve for the missingness inside the prediction model instead of the data. Two promising methods of this type are the pattern submodel (PS) approach and surrogate splits (SS). PS are attractive to a variety of parameter-based modeling techniques (e.g., regression). The so-called submodels incorporate the nature of the missing data by developing a separate prediction model for all possible missing data patterns (9). Then, when applied to a new case or out-of-sample individual the corresponding prediction model that matches the individual’s missing data pattern is used. Whereas the PS approach lends itself to various kinds of prediction models, SS come naturally to tree-based methods, such as random forest models (10,11). Briefly, SS attempt to preserve the partitioning of the original split by finding the next most optimal split given other observed variables. When the model is applied, each original split for which the predictor is missing will be replaced by the best available ‘surrogate’ variable to decide the split direction (10,11).

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

# Missing data handling methods for prediction models

We consider the following three prediction modeling strategies for real-time handling of missing data: (i) prediction models that adopt joint modeling imputation, (ii) prediction models that adopt a pattern submodel approach (iii) prediction models that adopt random forests with surrogate splits (9,11–13).

## Joint Modeling Imputation (JMI)

JMI is an imputation method that involves estimating the multivariate (joint) density of the predictor data and is used to generate imputed values directly from the conditional distribution (14). An advantage of JMI is that it can be applied to a previously developed prediction model. Because distribution parameters cannot directly be estimated in incomplete data, JMI typically requires the implementation of a Gibbs sampler. Recently, an extension to JMI was proposed to allow for real-time imputation in individual patients (13,15). With the extension the development of a JMI model consists of two separate steps. In the first step, the means and covariance of all predictor variables are estimated in a complete training sample from the population to which the prediction model will be applied. Since JMI assumes that every predictor variable is normally distributed, the population characteristics (i.e., means and covariance) can directly be used to generate, or draw, imputations on an individual level. In clinical practice, when a prediction model now encounters missing values, the developed JMI model can be utilized to generate imputations for each missing value on each predictor variable. We implemented three variants of JMI to be evaluated: single draw (JMI-SD, where a single draw from the conditional distribution is the imputed value), multiple draw (JMI-MD, where the average of 50 draws from the conditional distribution is the imputed value) and the conditional mean (JMI-CM, where the expected value of the conditional distribution is the imputed value). See Figure 1 for a schematic depiction of JMI.

#### Graphical user interface, diagram, timeline Description automatically generated with medium confidence **Figure 1.** Joint Modeling Imputation (JMI)

## Pattern Submodel (PS) approach

Another approach to address missing data without requiring imputation is to develop separate prediction models (so called pattern submodels, or briefly, PS) for each missing data pattern (9). Each PS is to be made specifically for one of the identified missing data patterns in the training data and the missing data patterns that are encountered in clinical practice. When applied to a new, out-of-sample, individual, PS approach uses the corresponding prediction model (i.e., matching the missing data pattern at hand). A recent study has shown that the use of PS for prediction performs similarly to multiple imputation and outperform multiple imputation in some cases when the data are missing not at random (MNAR, when missing data is dependent on unobserved values) (9,16,17). As such, PS may provide an elegant and intuitive to understand method for handling missing data when implementing prediction models. See figure 2 for a schematic depiction of the PS approach.

#### Graphical user interface, text, application Description automatically generated **Figure 2.** Pattern submodel approach

## Surrogate Splits (SS)

A well-known family of ML-based prediction models are the tree-based models, with as a simple case a (single) decision tree (18,19). Decision trees use a tree like structure to find the optimal cut-off point which partitions the data for optimal predictive performance. Based on the values of the pre-defined predictor variables, each branch in the tree represents a possible direction or decision. In essence, random forests combine multiple decision trees by using a combination of a random subspace method (i.e., random combinations of features) and bagging (i.e., random sample of observations). As an early extension to the well-known decision tree and random forest, SS were developed to circumvent the necessity for imputation (10,11,20). Briefly, SS try to preserve the partitioning of each original split in a tree as good as possible in the presence of missing predictor values. Whenever the model is applied to an individual and encounters a missing predictor value, it will use the pre-specified surrogate (i.e., replacement) variable, rather than the missing predictor variable, to decide upon the split direction. See figure 3 for a schematic depiction of SS in the context of a single decision tree. In this study we use SS in combination with a random forest prediction model.

#### Figure 3. Decision tree with surrogate splits

# Simulation design

## Aims

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

#### **Figure 4.** Simulation study

#### Diagram Description automatically generated

## Data-generating mechanism

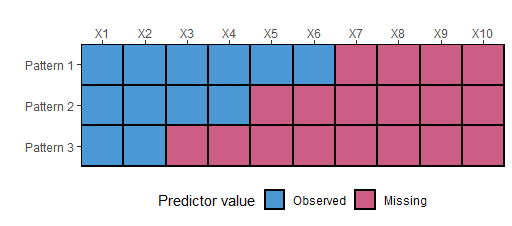
All data are generated from a single model-based population, consisting of ten continuous predictors and one dichotomous outcome. In each simulation iteration, we draw two samples from the population: a complete development set (*n* = 10.000), and a validation set in which we introduce missing values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

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

Where is the intercept, s are regression coefficients, and is the residual error term . We differentiate between two types of regression coefficients: is a vector of regression coefficients for the main effects of the predictors, ; is a vector of regression coefficients for the interactions with the first predictor, . This introduces a polynomial effect of the second degree, , and nine interaction effects. For additional non-linearity, we use a transformation in the effect of the second predictor, . All regression coefficients can be found in Supplementary materials X. The expected occurrence of the outcome is 15%.

The validation set is amputed (i.e., made incomplete) according to several missingness mechanisms and missingness rates. In this study, we focus primarily on the Missing At Random (MAR) missingness mechanism and additionally on the Missing Not At Random (MNAR) missing mechanism (21). We use a mixture of the four kinds of MAR missingness, as described by Schouten and others (22). The overall missingness rate is 60%, but the number of missing predictor entries differs between cases. The hypothetical patients in our validation set are missing either 40%, 60%, or 80% of the observations in the predictor space. The resulting missing data pattern is visualized in Figure 5.

#### **Figure 5.** Missing data pattern.



## Methods

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

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

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

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

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

## Performance measures

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

**Table 2.** Performance measures

|  |  |
| --- | --- |
| Measure | Performance metric |
| Overall prediction accuracy | Root mean square error (RMSE). The RMSE of the predictions reflects the difference between the estimated probability of Y and the true underlying probability of the outcome before amputation. Like the estimand and estimates, the RMSE lies on the probability scale. Lower values indicate better performance (23). |
| Brier score. The brier score is defined as the squared difference between the predicted risk and the observed outcome value. A brier score of 0 would represent a perfect model, whilst the maximum brier score is determined by the incidence of the outcome (24). |
| Discrimination | Concordance (C-)statistic. The C-statistic is a rank-order statistic, which is used to describe how well a classification model can discriminate between those with an event and those without. The C-statistic shows the probability of taking two random subjects (one with and one without the outcome) and correctly attributing the one with the outcome with a high risk. A C-statistic of 0.5 describes a model with no discriminative performance and a C-statistic 1 describes a model with perfect discriminative performance. |
| Calibration | Calibration-in-the-large (CITL). The CITL represents the overall calibration of a model. In other words, the extent of agreement between the average predicted risk and the original predicted risk (25). The metric ultimately describes the amount of systematic over- or under-estimation of the predicted risk. A value of 0 is ideal and represents perfect agreement. |
| The calibration slope. In contrast with the CITL, the calibration slope does not evaluate the average predicted, or original, risk. Rather, it quantifies the extent by which the predicted risks vary too much (i.e., slope <1) or too little (i.e., slope >1). Ideally, the slope is 1. |

Simulation Results

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

## Root mean squared error

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

## Brier score

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

## C-statistic

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

## Calibration-in-the-large

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

## Calibration slope

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

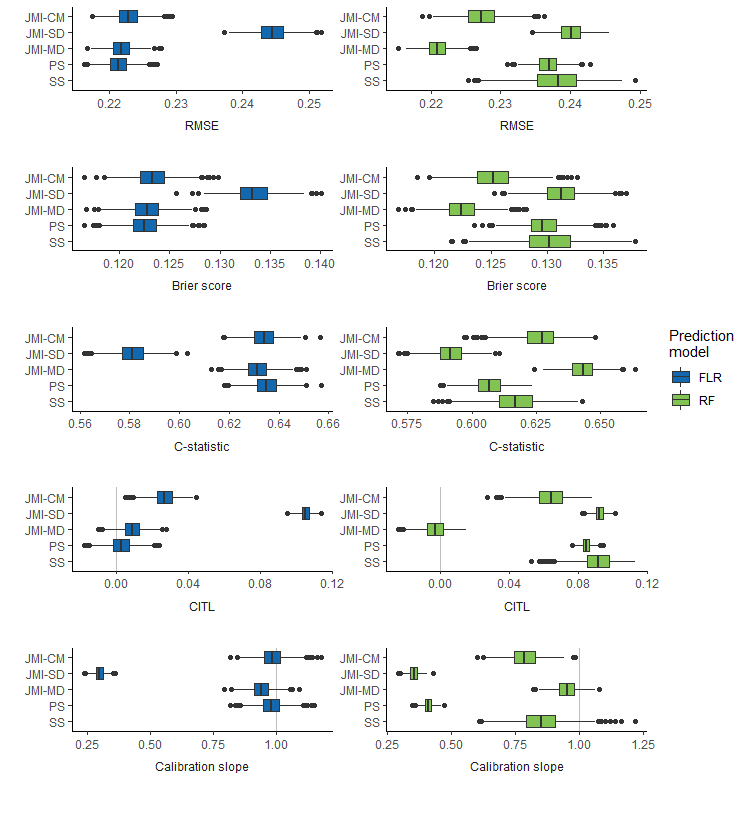
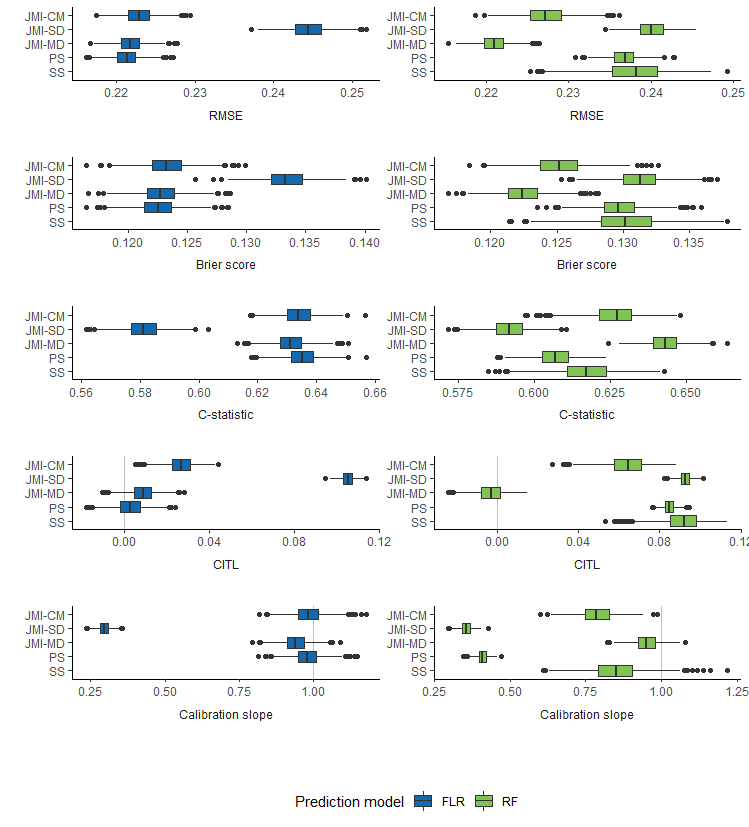
## Calibration plots

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

#### **Figure 6.** Performance measures per method

Diagram, schematic

Description automatically generated

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

#### **Table 3.** Average performance across simulations.

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

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

#### **Figure 7.** Calibration plots

Chart, diagram, line chart

Description automatically generated

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

# Case Study

**Data Source**

To illustrate the methods, we used the MIMIC-III dataset. This is a freely available database that comprises data of 62722patients admitted to the ICU of the Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2001 and 2012. The data consists of demographics, vital sign measurements made every hour, laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality.

**Data of Interest**

For the current case study, a few specific variables are of interested, based on the Sepsis-related Organ Failure Assessment (SOFA). The SOFA is a measure that is designed to describe a pattern of complications in critically ill patients that are related to organ failure and death. For computation of the SOFA score, the variables FiO2, PaO2, the number of platelets, the bilirubin level, the Glasgow coma score, MAP, creatinine level, and the urine output are of importance. Of specific interest are the ‘worst’ scores of these variables in the last 24 hours, i.e., the level in the last 24 hours that would yield the worst SOFA score.

**Data Extraction**

We extracted the data using the ‘ricu’ package in R. For each of the relevant variables, the worst value in 24 hours before discharge from the ICU was extracted, as well as the outcome variable indicating the patient had deceased. The outcome had an incidence of 0.12, with 6589 of the patients having the outcome and 54303 not. The training set presented 115 different missing data patterns in total, out of which the 20 most frequent were selected to match the test set. Explorations showed that 1830 patients had missingness on all predictor variables and were thus excluded from the sample. Table […] shows descriptives of the relevant continuous variables of the 60892 remaining patients.

**Table […].** *Descriptives of Worst Values in the Last 24 Hours of Predictors from MIMIC Dataset.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictor** | ***M*** | ***SD*** | ***NAs*** | ***Complete Rate*** |
| FiO2 | 79.71 | 25.58 | 33162 | .46 |
| PaO2 | 91.55 | 384.86 | 48030 | .21 |
| Platelets | 190.04 | 102.59 | 1160 | .98 |
| Bilirubin | 3.23 | 5.58 | 20666 | .66 |
| Glasgow Coma Score (non-sedated) | 12.65 | 3.47 | 8537 | .86 |
| Glasgow Coma Score (total) | 9.31 | 5.03 | 31853 | .48 |
| MAP | 109.31 | 26.09 | 8459 | .86 |
| Creatinine | 1.81 | 1.90 | 7153 | .88 |
| Urine Output | 1360.93 | 1168.65 | 11502 | .81 |

**Figure 8.** Missing data patterns (Note. 115 patterns total, 20 most frequent selected, then 17 left after feasibility check).

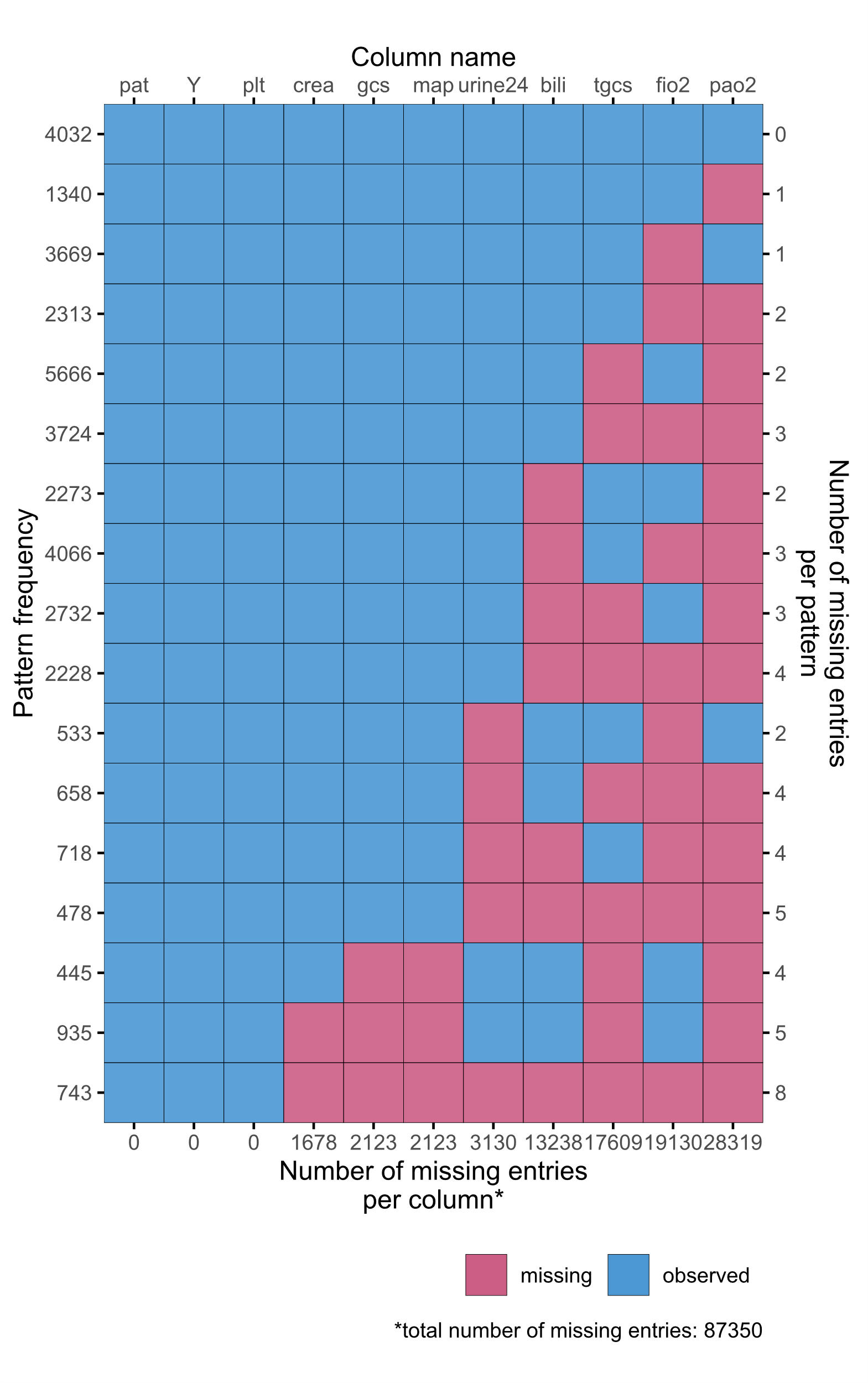


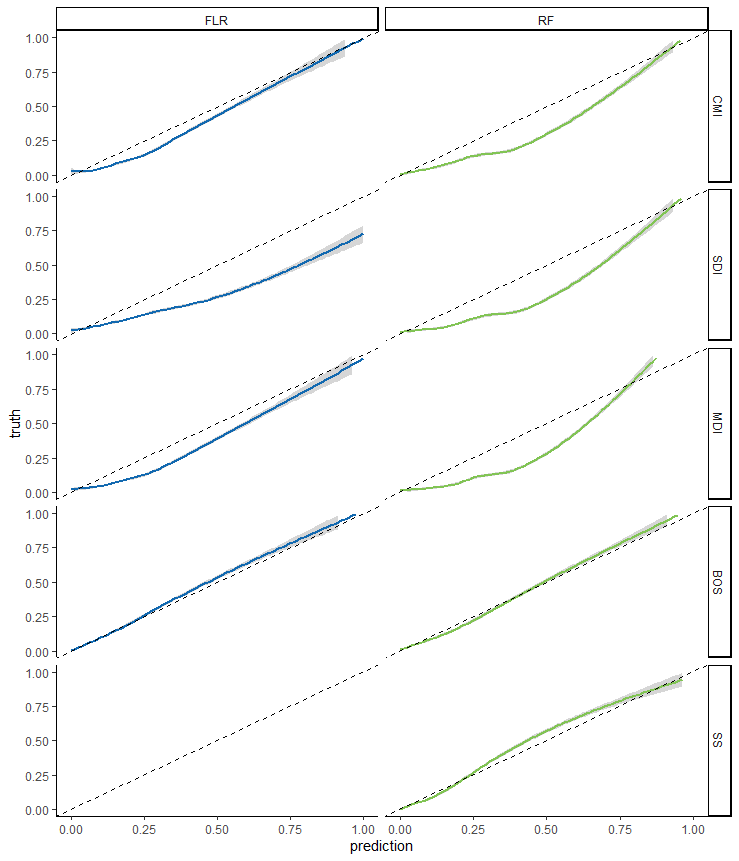
Table […] shows the results of the simulations. When considering FLR, the RMSE, Brier score, and C-index all indicate that single draw imputation performs the worst of the four methods, and the pattern submodel approach performs best. Regarding random forests, these measures also indicate single draw imputation to have the worst performance, and they indicate pattern submodels and surrogate splits to have a similarly high performance. Calibration curves are shown in Figure 9.

Table […]. Performance of Methods on MIMIC Data.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | RMSE | Brier | C-index | CITL | Slope |
| FLR | JMI-CM | 0.308 | 0.095 | 0.803 | 0.142 | 0.192 |
| JMI-SD | 0.321 | 0.103 | 0.774 | 0.155 | 0.189 |
| JMI-MD | 0.312 | 0.098 | 0.791 | 0.155 | 0.187 |
| PS | 0.300 | 0.090 | 0.826 | 0.102 | 0.210 |
| RF | JMI-CM | 0.302 | 0.091 | 0.837 | 0.151 | 0.278 |
| JMI-SD | 0.313 | 0.098 | 0.825 | 0.183 | 0.254 |
| JMI-MD | 0.309 | 0.095 | 0.830 | 0.184 | 0.255 |
| PS | 0.286 | 0.082 | 0.862 | 0.096 | 0.290 |
| SS | 0.287 | 0.082 | 0.865 | 0.093 | 0.270 |

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

**Figure 9.** Calibration curves case study.



# Discussion

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

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

The good performance of JMI in our simulations may partly be driven by the choice of predictor correlation structure and missing data pattern in our simulations. Low correlations have previously been associated with limited performance of JMI (13). Likewise, SS very heavily rely upon the correlation between the missing predictor value and the surrogate replacement value (25). With the low to moderate correlations imposed in our DGM, it may be expected that multivariable approaches such as JMI perform better when compared with SS, which relies only on the single surrogate variable. For example, in the most extreme missing data scenario, when only and are observed, it is likely that optimal surrogate variables are not available. It may be evident that PS, which uses only the observed predictor variables, is also limited in circumstances such as these. In the end, when using clinical data, correlations between predictor variables need to be considered.

Additionally, to avoid overfitting, prediction models are typically designed as simple as possible and usually include predictors that do not intercorrelate much [refs]. Likewise, in our simulation study, we only generated 10 covariates, all of which were used for development of the prediction model and imputation strategies. In practice, however, many more additional variables may be available. These auxiliary variables (i.e., not part of the prediction model) have previously shown to improve JMI performance (12). If made available, it is likely that auxiliary variables, if not for prediction, may improve the accuracy of any missing data handling strategy which relies upon correlations between available variables.

Generally, PS has adequate prediction model performance in the presence of missing data. A major advantage for PS is that it does not require MAR assumptions. In real-world datasets PS, therefore, offer an appealing solution. When PS is paired with RF, however, problems arise. These problems may be explained by the fact that less predictors ultimately restrict how much a random forest may vary between each tree (27). In other words, if there are less features available, as is the case for PS, the variability between trees is limited. Similarly, surrogate splits perform relatively poor, which can be explained by the strong dependence on high correlations between the surrogate variable and the missing predictor variable.

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

# Disclosures

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

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

# References

1. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology. 2006 Oct;59(10):1087–91.

2. Janssen KJM, Vergouwe Y, Donders ART, Harrell FE, Chen Q, Grobbee DE, et al. Dealing with Missing Predictor Values When Applying Clinical Prediction Models. Clinical Chemistry. 2009 May 1;55(5):994–1001.

3. D’Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. Circulation. 2008 Feb 12;117(6):743–53.

4. Dorresteijn JAN, Visseren FLJ, Wassink AMJ, Gondrie MJA, Steyerberg EW, Ridker PM, et al. Development and validation of a prediction rule for recurrent vascular events based on a cohort study of patients with arterial disease: the SMART risk score. Heart. 2013 Jun 15;99(12):866–72.

5. Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C, Catapano AL, et al. 2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts)Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). Eur Heart J. 2016 Aug 1;37(29):2315–81.

6. Groenhof TKJ, Bots ML, Brandjes M, Jacobs JJL, Grobbee DE, van Solinge WW, 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 Sep;27(9):435–42.

7. Bezemer T, de Groot MC, Blasse E, ten Berg MJ, Kappen TH, Bredenoord AL, et al. A Human(e) Factor in Clinical Decision Support Systems. J Med Internet Res. 2019 Mar 19;21(3):e11732.

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

9. Fletcher Mercaldo S, Blume JD. Missing data and prediction: the pattern submodel. Biostatistics. 2020 Apr 1;21(2):236–52.

10. Feelders A. Handling Missing Data in Trees: Surrogate Splits or Statistical Imputation? In: Żytkow JM, Rauch J, editors. Principles of Data Mining and Knowledge Discovery [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 1999 [cited 2019 Oct 2]. p. 329–34. Available from: http://link.springer.com/10.1007/978-3-540-48247-5\_38

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

12. Nijman SWJ, Hoogland J, Groenhof TKJ, Brandjes M, Jacobs JJL, Bots ML, et al. Real-time imputation of missing predictor values in clinical practice. European Heart Journal - Digital Health. 2021 May 4;2(1):154–64.

13. Nijman SWJ, Groenhof TKJ, Hoogland J, Bots ML, Brandjes M, Jacobs JJL, et al. Real-time imputation of missing predictor values improved the application of prediction models in daily practice. Journal of Clinical Epidemiology. 2021 Jun;134:22–34.

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

15. Hoogland J, Barreveld M, Debray TPA, Reitsma JB, Verstraelen TE, Dijkgraaf MGW, et al. Handling missing predictor values when validating and applying a prediction model to new patients. Statistics in Medicine. 2020 Jul 20;sim.8682.

16. Glynn RJ, Laird NM, Rubin DB. Selection Modeling Versus Mixture Modeling with Nonignorable Nonresponse. In: Wainer H, editor. Drawing Inferences from Self-Selected Samples [Internet]. New York, NY: Springer New York; 1986. p. 115–42. Available from: https://doi.org/10.1007/978-1-4612-4976-4\_10

17. Pattern-Mixture Models for Multivariate Incomplete Data. 2021;11.

18. Dorado-Díaz PI, Sampedro-Gómez J, Vicente-Palacios V, Sánchez PL. Applications of Artificial Intelligence in Cardiology. The Future is Already Here. Revista Española de Cardiología (English Edition). 2019 Dec;72(12):1065–75.

19. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees [Internet]. Taylor & Francis; 1984. Available from: https://books.google.nl/books?id=JwQx-WOmSyQC

20. Hapfelmeier A. Analysis of Missing Data with Random Forests [Internet]. 2012 [cited 2019 Sep 4]. 6–7 p. Available from: https://edoc.ub.uni-muenchen.de/15058/1/Hapfelmeier\_Alexander.pdf

21. Rubin DB. Inference and Missing Data. Biometrika. 1976;63(3):581–92.

22. Schouten RM, Lugtig P, Vink G. Generating missing values for simulation purposes: a multivariate amputation procedure. Journal of Statistical Computation and Simulation. 2018 Oct 13;88(15):2909–30.

23. van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. Stat Methods Med Res. 2019 Aug;28(8):2455–74.

24. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. Berlin: Springer; 2009. 497 p. (Statistics for biology and health).

25. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. European Heart Journal. 2014 Aug 1;35(29):1925–31.

26. Twala B. AN EMPIRICAL COMPARISON OF TECHNIQUES FOR HANDLING INCOMPLETE DATA USING DECISION TREES. Applied Artificial Intelligence. 2009 May 4;23(5):373–405.

27. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak. 2019 Dec;19(1):281.

# Supplementary Materials

## DGM

Means vector: All 10 predictors have a mean of zero, .

Covariance matrix:

Correlations:

Chart, histogram

Description automatically generated

Figure XYZ. Correlation coefficients between predictors

Regression coefficients:

A picture containing Excel

Description automatically generated

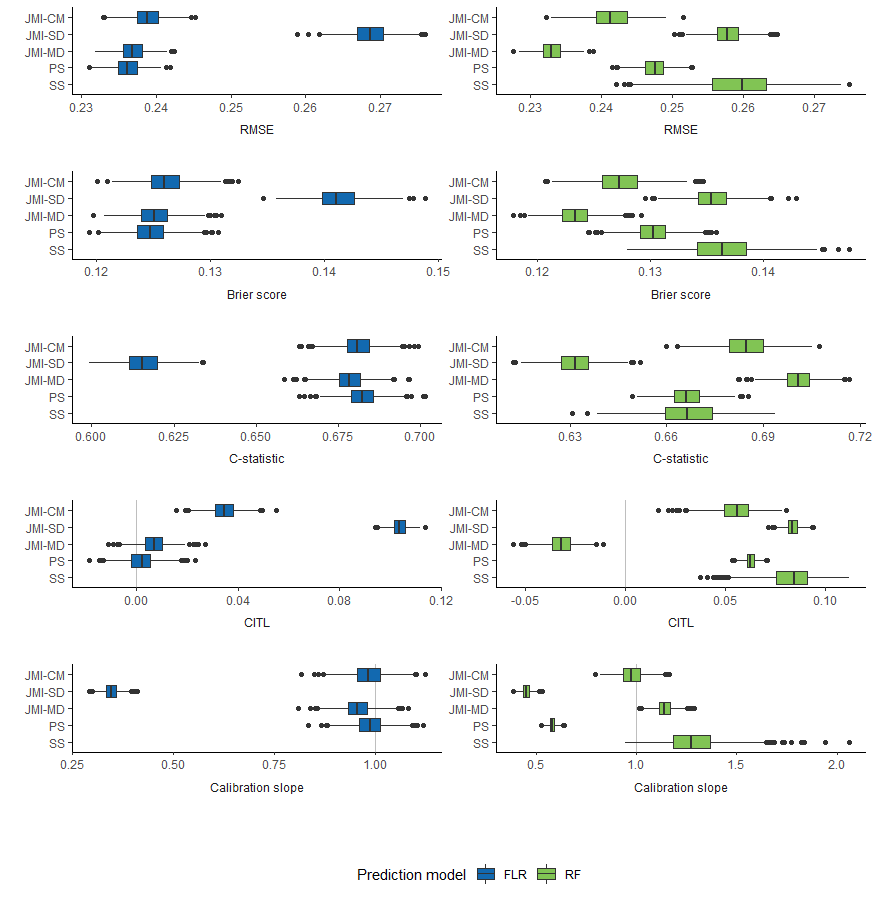
Figure XYZ. Regression coefficients of the main and interaction effects of the predictors

Model tuning

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

## Results

Performance under MNAR



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

#### Average performance under MNAR

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