# On the fly handling of missing data in the application of prediction models: a comparison of methods

Nijman SWJa\*, Oberman HIb\*, 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

\* 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 occurrence of real-time missing predictor values is unique to the application of prediction models and seems to be 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. We aim to evaluate the influence of built-in missing data handling mechanisms on prediction accuracy and compare it with existing real-time imputation methods (e.g., joint modelling imputation). We evaluate the effect of various missing data handling methods under specific missing data circumstances as would occur in medical practice.

**Methods –**

**Results –**

**Discussion –**

# 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 substantial impediment of missing data on the direct use of prediction models in individual patients becomes more evident (6,7). The issue is further compounded as the (gold) standard strategies to mend or circumvent missing data are not suited for use in individual patient data in live clinical practice.

Various strategies to handle challenging 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 and can provide valid estimates and correct standard errors when the missing data are missing completely at random (MCAR, when missing data is unrelated to the data) or missing at random (MAR, when missing data is dependent on observed 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 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 the missingness inside the prediction model instead of the data. Two promising methods of this type are surrogate splitting and pattern mixture modeling. The surrogate split approach is natural to random forest models (9,10). Briefly, these surrogate splits attempt to preserve the partitioning of the original split by finding the second most optimal split given other observed variables. When the model is applied, each original split for which the predictor is missing is replaced by the best available ‘surrogate’ variable, or split, to decide the split direction (9,10). The pattern sub models, which can be used for a variety of modelling techniques (e.g., regression), approach the missing data by developing, for each (possible) pattern of missing data, a separate prediction model (11). Then, when applied to a new (out-of-sample) individual the corresponding (i.e., matching the missing data pattern in the individual) prediction model is used.

In this article we compare various real-time missing data handling approaches when implementing specific modelling techniques in live 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 clinical practice, without necessitating the use of data from other individuals at the point of care. We present an extensive 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.

# Methods

**Brief description**

We evaluate and compare various modelling strategies for real-time handling of missing data at the implementation level via a simulation study: (i) prediction models that adopt separate imputation models (i.e., joint modelling imputation or JMI), (ii) prediction models that adopt random forests with surrogate splits and (iii) prediction models that adopt a submodel approach (10–13).

**Missing data handling methods**

Real-time imputation via Joined Modelling Imputation (JMI)

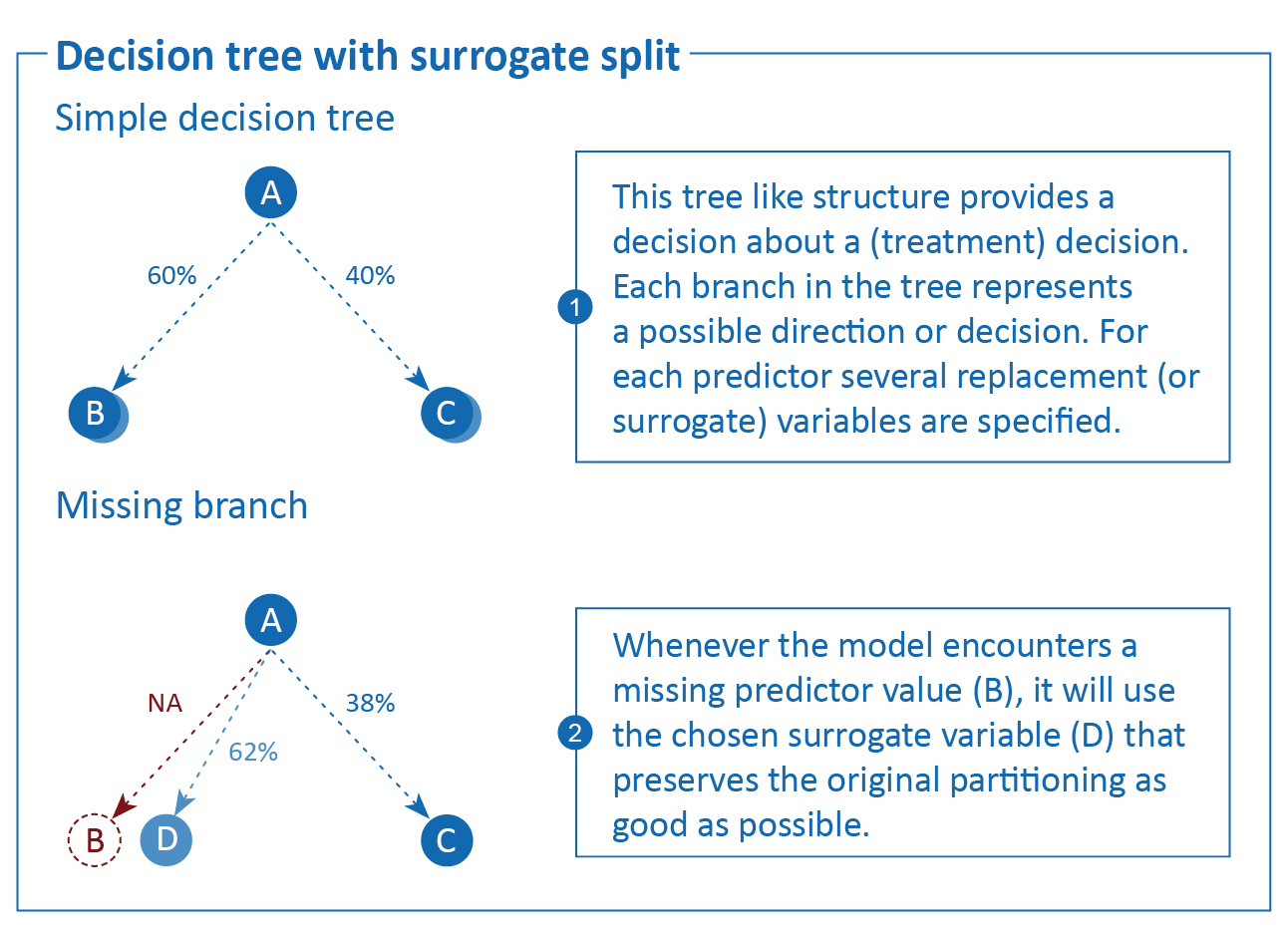
JMI is an imputation method that involves estimating the multivariate (joint) density of the data (14). JMI typically involves Monte Carlo sampling to estimate the distribution parameters and impute the missing values. Recently, an extension to JMI was proposed to allow for real-time imputation in individual 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 development sample. 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 of the missing variables. An advantage of JMI is that it can be applied to a previously developed prediction model. See Figure x for a schematic depiction of JMI.

Graphical user interface, diagram, timeline

Description automatically generated with medium confidence  
*Figure x. Joint Modelling Imputation (JMI)*

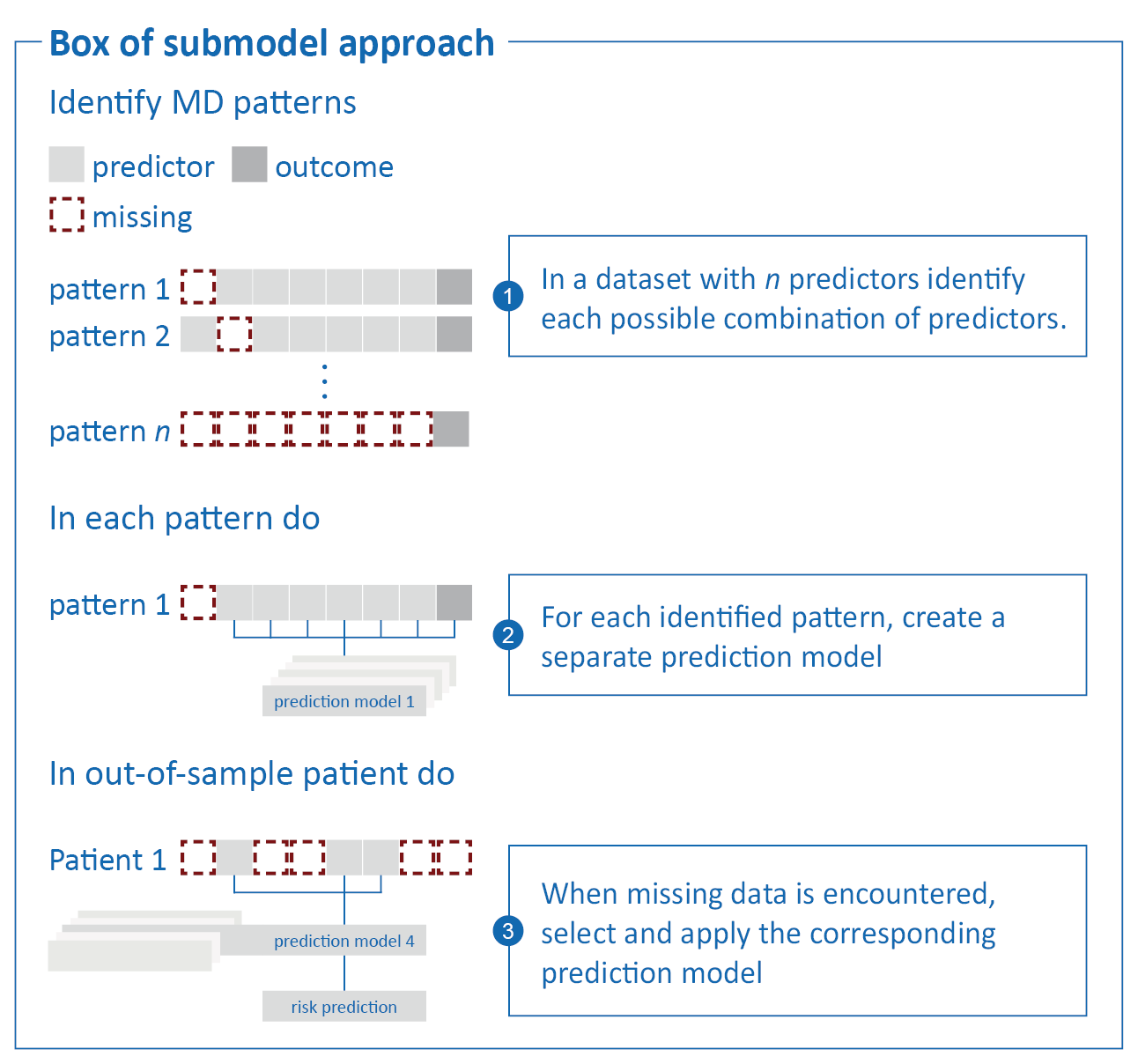
Surrogate splits

As an early extension to the well-known decision tree, surrogate splits were developed to circumvent the necessity for imputation (9,10,16). Decision trees use, as the name suggests, 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 to be merged for improved prediction accuracy. Briefly, surrogate splits 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. We use surrogate splits in the context of random forests. See figure x for a schematic depiction of surrogate splits.

*  
Figure x. Decision tree with surrogate splits*

Missing data pattern submodels

Another approach to address missing data without requiring imputation is to develop separate submodels for each missing data pattern (11). Each submodel 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 real-time clinical practice. When applied to a new, out-of-sample, individual, the collection of submodels uses the corresponding (i.e., matching the missing data pattern at hand) prediction model. A recent study has shown that the use of submodels for prediction performs similarly to multiple imputation and can be used when the data are missing not at random (MNAR, when missing data is dependent on unobserved values) (11). As such, submodels may provide an elegant and intuitive to understand method for handling missing data when implementing prediction models. See figure x for a schematic depiction of pattern submodels

  
*Figure x. Pattern submodels*

**Motivating example**

We conducted a motivational example showing the effect of the built-in methods and real-time imputation models when used in real patients. Hereto, we considered the large Medical Information Mart for Intensive Care (MIMIC)-III dataset (17). MIMIC-III provides a large database which contains information about patients staying in critical care units of the single tertiary care Beth Israel Deaconess Medical Center.

Similarly, to the simulation study, the prediction models of interest were the flexible logistic regression and random forest models. We derived both models in MIMIC-III using predictors from existing relevant prediction models using mortality as the primary outcome. For the logistic regression, we considered the Sequential Organ Failure Assessment (SOFA) prediction score and for the random forest model we considered the … (18). The SOFA score estimates the number and severity of failed organs, with in-hospital mortality as the primary outcome.

**Simulation study**

*Aims*

The aim of the simulation study is to evaluate the performance of several on-the-fly missing data handling approaches in the context of single patient application of prediction models as would occur in clinical practice (Figure x). To that end, we compare the performance of these different approaches on their calibration and discriminative performance.

**Diagram

Description automatically generated**

*Figure x. Simulation study*

*Data-generating mechanism*

All data are generated from a single model-based population, consisting of ten continuous predictors and one dichotomous outcome. In each simulation iteration, we draw two samples from the population: a complete development set (*n* = 10.000), and a validation set in which we introduce missingness 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. All 10 predictors have a mean of zero, . The covariance matrix

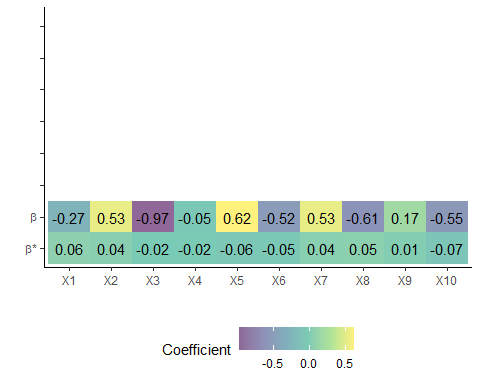
is visualized in Figure XYZ.

|  |
| --- |
| Figure XYZ. Correlation coefficients between predictors |

From the predictor space, we define the binary outcome Y. Y is a function of through the logit link function,

where s are regression coefficients, and residual error . We differentiate between three types of regression coefficients: 1) the intercept, ; 2) a vector of regression coefficients for the main effects of the predictors, ; and 3) an additional vector of regression coefficients for the interactions with the first predictor, . This introduces a polynomial effect of the second degree, , and nine moderation effects. For additional non-linearity, we use a transformation in the effect of the second predictor, . The regression coefficient vectors

are visualized in Figure XYZ. With an intercept of , the population prevalence of is 15%.

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

Before any strategies were applied, missing data was introduced in the validation set, with hypothetical patients, via varying mechanisms. We choose to focus on a mixture of MAR missing data (“MAR left”, “MAR right”, “MAR mid”, and “MAR tail”) (8). We also used a mixture of three missing data patterns, with 40%, 60%, and 80% of variables being missing per case (see figure x below). The set of hypothetical patients was generated and evaluated 1.000 times.

Chart, bar chart, histogram

Description automatically generated

***Figure x.*** *missing data patterns*

|  |
| --- |
|  |
| **Figure X.** Data generating mechanism correlation matrix. |

**Figure X.** Model coefficients.

*Missing data handling strategies*

For handling missing data in real-time, we consider JMI, surrogate splits and pattern submodels. We use JMI via three different approaches. We impute with (i) the conditional mean, (ii) a random draw from the conditional multivariate distribution and (iii) the mean of multiple (50) draws from the conditional multivariate distribution.

*Prediction models*

Each prediction model is applied to the hypothetical patients to estimate the absolute risk of the outcome for each individual observation (i.e., patient). To accommodate the use of these real-time missing data handling strategies, several types of prediction models need to be considered. We use a combination of random forests and flexible logistic regression models. We use both the random forest and logistic regression to accommodate JMI and pattern submodels. Only random forests are used to accommodate the use of surrogate splits.

*Generation of risk predictions*

The target in each risk prediction is for each prediction model to estimate the absolute risk of the outcome in each hypothetical individual observation.We evaluated the calculated risks after using five different approaches to handling missing data (Table 1). Whilst the JMI approaches and submodel approach are specifically related to the real-time imputation of missing predictor values, they are also applied to the random forest to compare between the two prediction models. The remaining strategy (i.e., surrogate splits) is specific to the prediction method used (i.e., random forest).

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Flexible logistic regression | Random forest |
| JMI | Conditional mean1 | X | X |
| Random draw2 | X | X |
| Average of multiple draws3 | X | X |
| Surrogate splits4 | |  | X |
| Pattern submodels5 | | X | X |

**Table 1.** summary of missing data methods and prediction models  
  
1. Missing values are imputed by their mean conditional on the observed predictor variables; 2. Missing values are imputed by a random draw from their conditional multivariate distribution (i.e., non-deterministic imputation); 3. Missing values are imputed 50 times by a random draw from their multivariate normal distribution, The resulting 50 absolute risk predictions are then averaged to obtain the final prediction; 4. Missing values are circumvented by selecting the appropriate pattern submodel for calculating absolute risks. 5. Missing values are handled using surrogate splits.

*Performance measures*

We evaluated the approaches outlined above and compared the following five performance metrics to evaluate the predicted risks under the 5 strategies: (i) root mean squared error (RMSE) of the predicted risk, (ii) brier score, (iii) concordance (C-) statistic, (iv) calibration-in-the-large (CITL) and (v) the calibration slope (Table 2).

|  |  |
| --- | --- |
| Measure | Performance metric |
| Prediction accuracy | Root mean square error (RMSE). The RMSE of the predictions reflects the error between the original predicted risk and the outcome risk which was calculated under various conditions with missing data (19). The error is presented, like the original predictions, on the probability scale. Lower values indicate better performance. |
| Brier score. The brier score calculates the squared differences between the predicted risk and the original outcome. A brier score of 0 should always reflect a perfect model, whilst the incidence of the outcome ultimately indicates what the maximum brier score is (19). |
| Discrimination | Concordance (C-)statistic. The C-statistic is a rank-order statistic, which is used to describe how well the model can discriminate between those with the 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 extend of agreement between the average predicted risk and the original predicted risk (20). The statistic ultimately explains the amount by which the predictions systematically over or underestimate 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 are too extreme (i.e., <1) or too narrow (i.e., >1) when compared with the original risk over the whole range of predicted risks. Ideally the slope is 1. |

**Table 2.** performance measures

Results

**Reference performance**

**Root mean squared error**

**Brier score**

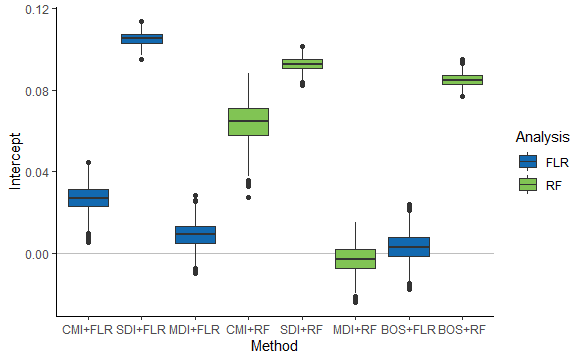
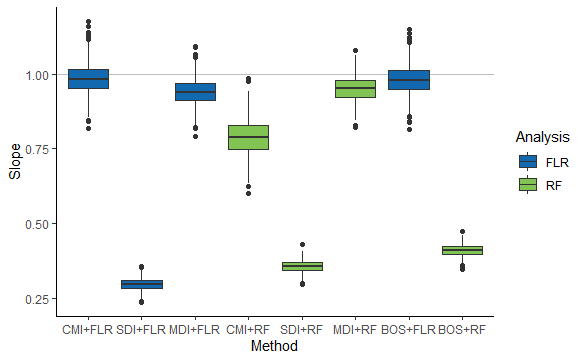
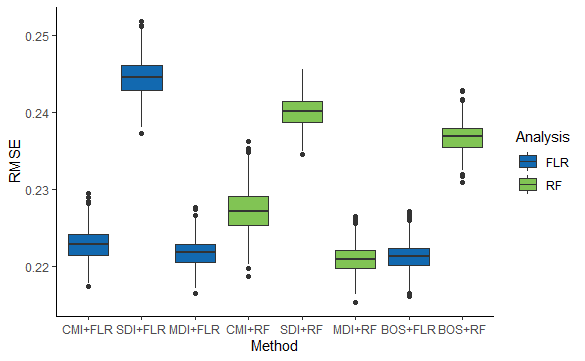
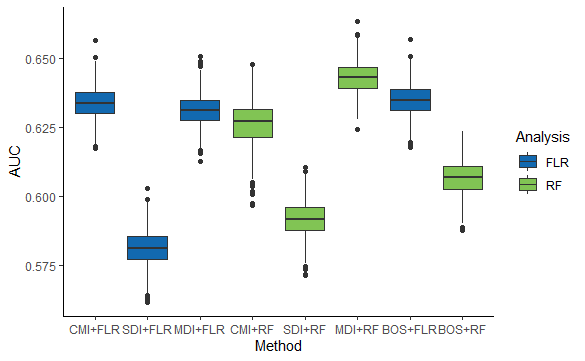
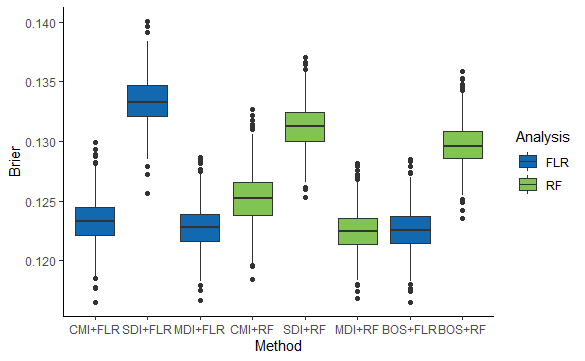
**C-index**

**Calibration-in-the-large**

**Calibration slope**

Both JMI with the conditional mean and multiple draws results in a well calibrated slope when applied for the logistic regression (…). Single imputation, via JMI, has sever miscalibration. The performance diminishes using the conditional mean when applied for the random forest. In contrast, the use of multiple draws does not diminish in its performance. Similarly, the use of submodels was consistently well calibrated for both the logistic regression and the random forest models. The use of surrogate splits causes miscalibration in the lower and higher risk predictions. All calibration plots can be found in figure x.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Label | Method | Intercept | Slope | AUC | RMSE | Brier |
| FLR | CMI | 0.027 | 0.985 | 0.634 | 0.223 | 0.123 |
| SDI | 0.105 | 0.297 | 0.581 | 0.244 | 0.133 |
| MDI | 0.009 | 0.941 | 0.631 | 0.222 | 0.123 |
| Submodels | 0.003 | 0.981 | 0.635 | 0.221 | 0.123 |
| Surrogate splits |  |  |  |  |  |
| RF | CMI | 0.064 | 0.789 | 0.627 | 0.227 | 0.125 |
| SDI | 0.093 | 0.355 | 0.592 | 0.240 | 0.131 |
| MDI | -0.003 | 0.952 | 0.643 | 0.221 | 0.122 |
| Submodels | 0.085 | 0.410 | 0.607 | 0.237 | 0.130 |
| Surrogate splits |  |  |  |  |  |

# Discussion

This simulation study aimed to evaluate the effectivity of using real-time missing data handling strategies to handle missing predictor values in individual patients. We considered JMI, submodels or surrogate splits for the real-time handling of missing data when using either a flexible logistic regression or random forest model. Our results suggest that built-in mechanisms such as surrogate splits, when compared with the other missing data handling approaches in this simulation study, show severe miscalibration for the low end and high end of predicted risks.

The performance of the imputation approach for each of the modelling techniques depends on the method of implementation. The conditional mean and multiple imputation (i.e., average over multiple draws) variant both performed, in terms of calibration and discrimination, much better than the single imputation variant when a flexible logistic regression or random forest is used. The use of multiple imputation performed more consistently, in terms of calibration and discrimination, than the conditional mean when a random forest is used. The difference in performance between the multiple imputation and conditional mean variant, when used for a random forest, may be explained by the congeniality, or compatibility, of the imputation model. Briefly, it means that the random forest, when compared to the flexible logistic regression, may be better at surmising the information provided to it as (non-linear) input from the completed data. When compared to the conditional mean, multiple imputations are ultimately less (parametrically) restrictive and allow for more variability and as such play more to the strength of a random forest.

Previous work has shown that the performance of JMI is also associated with the correlations between predictor variables, and that low correlations were associated with limited performance, in terms of calibration and discrimination (12). Since highly correlated variables are unlikely to all be used in the same prediction model it is likely that, when only predictor values are used, the accuracy of JMI remains limited. However, the amount of information able to be leveraged from a patient can generally be much higher when other, auxiliary, variables (i.e., not part of the prediction model) are included (12). To prevent a similar limitation in this study, correlations were kept sufficiently high for JMI. The generated correlations may be limiting the propagation of the imputation methods to more realistic settings as they are slightly higher than usually observed in clinical settings. Under these circumstances the incorporation of auxiliary variables may be necessary for suitable imputation performance.

Similarly, the deficient performance of surrogate splits may be explained by the high dependency on the correlation between the missing predictor value and the surrogate replacement value (21). However, in contrast with JMI, surrogate splits depend on only a singular surrogate attribute to base its performance. As a result, it may be that the correlations in this simulation study, however high, are not high enough to guarantee satisfactory surrogate splitting.

The use of submodels seem to work well for both modelling techniques and results in optimal calibration in the presence of mining predictor values. Given that a submodel approach does not in any way depend on the interrelationship between predictor variables is compelling and seems to be an advantage over the other methods evaluated in this simulation study.

**May be added, depending on the results:**

*Table X. Description of the two data generation mechanisms*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DGM | Included predictor variables | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |  |
| 1 | L | L | L | L | L | L | L | L | L | L |
| 2 | L | Ln\* | L\* | L\* | L\* | L\* | L\* | L\* | L\* | L\* |

IT=interaction terms. L=Linear effects. Ln=Natural log transformation. \*=With interaction.

The first DGM includes all predictor variables as linear effects. The second DGM uses RCS with three knots for variable 4, uses a log transformation for variable 2, and interaction terms between the first variable and all other variables.

# 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. 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

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

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

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. 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

17. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 Dec 20;3(1):160035.

18. Jones AE, Trzeciak S, Kline JA. The Sequential Organ Failure Assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation\*: Critical Care Medicine. 2009 May;37(5):1649–54.

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

20. 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.

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