Missing data in clinical prediction modelling

Primer and current projects within Prediction Modelling Group, under the NIHR Maudsley BRC theme Trials, Prediction and Genomics.

Harper Clees-Baron[1], Diana Shamsutdinova[1,2], Daniel Stahl[1,2]
[1] King's College London, IoPPN, Biostatistics and Health Informatics Department
[2] NIHR Maudsley Biomedical Research Centre

Missing data in medical research

Statistical analyses of medical data are fundamental to modern clinical research methods and increasing use of electronic data collection and storage have opened new avenues with which to leverage statistics and data science to improve patient outcomes. One such application is clinical prediction modelling, which aims to predict patient outcomes of previously unseen cases using specific criteria. These models are the foundation of precision medicine—a promising clinical treatment pathway in which individual patient data are leveraged to predict which treatments will be most efficacious via prediction modelling (Luo *et al.*, 2020). Given the significant potential of clinical prediction modelling to improve medical system efficiency and patient outcomes, managing the problem of missing data for prediction models is paramount.

Missing observations appear almost ubiquitously across all types of datasets, and for the statistician, missing data are as troublesome as they are common. Analyses which ignore or exclude observations with missing values can lead to biased results (Rubin, 1976). Methods for handling this missingness vary depending on the missingness 'mechanism' which resulted in the missed observations. However, the missingness mechanism is often not knowable, and therefore must be assumed (Austin *et al.*, 2021). Therefore, outside of the limited cases in which the mechanism is known for certain to be the most innocuous form, missing data should be considered "not ignorable" and require a complicated fix (Austin *et al.*, 2021).

Conceptualizing missing data

The set of solutions which are appropriate for handling missing data depend on the missingness 'mechanism' of that data. These missingness mechanisms have been classified into three distinct groups (Rubin, 1976). The first, and simplest, is known as "missing

completely at random" or MCAR. Data are MCAR if the probability of missingness is entirely unrelated to any variables in the data. This is the easiest form of missingness to handle and can be addressed by listwise deletion of any observations which contain missing values. The MCAR assumption is quite strong, however, and is rarely reasonable in most contexts (Rubin, 1976). It is defined mathematically as $P(MISSING|y_o, y_m) = P(MISSING)$ for y_o = observed data, and y_m = missing data. An example of missing data which may be MCAR could be if there is a technical issue with the database, and certain random observations are not saved—in this case, the missing values are not related to any other observed or unobserved data.

For three variables in which values are ordered from 0 to 100, MCAR missing data patterns are reflected below. Notice that the missingness status of any given observation appears not to be correlated with the value of either the variable it occurs in, or in any other observed variable in the simulated dataset below.

Figure 1: MCAR data example

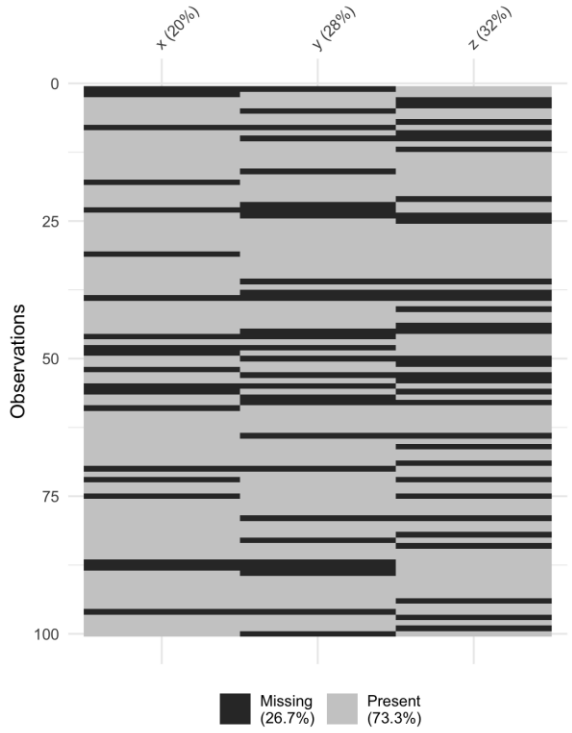
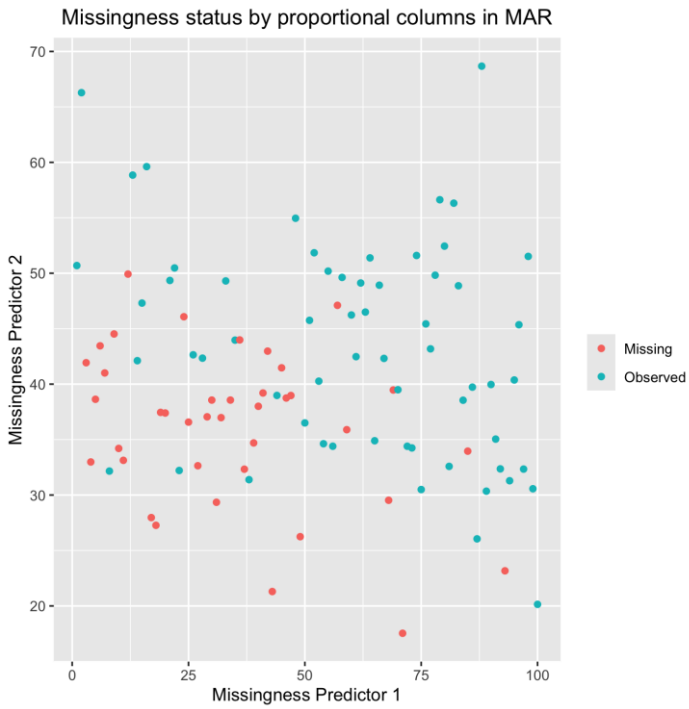


Figure 2: MAR Data Example



Data, for which the missingness is associated with values of other variables in the data set *but not itself,* is known as 'missing at random' (MAR). However, as Scheffer observes, this is a misnomer (Scheffer, 2002). The MAR assumption simply states that the values of the missing data are related to other predictors in the data set. Therefore, if the statistician accurately models the correlation structure between the missing values and the predictors which influence it, this missingness can be imputed. MAR is defined as $P(MISSING|y_o, y_m) = P(MISSING|y_o)$. An example of MAR data could be cholesterol levels across patient age groups. Because clinicians may be more likely to measure cholesterol levels in older patients, younger patients will have higher rates of missing values. Therefore, the missing values (cholesterol readings) are related to a fully observed variable (age), and the missingness can be considered random *within* a given age group.

Finally, there are missing values for which the probability of a value being missing is directly related *to that value*. This construction is known as Missing Not At Random (MNAR), or informative missingness. An illustrative example of data which is commonly MNAR is self-reported annual income (Scheffer, 2002). Those with very low incomes may be uncomfortable sharing their income information and are therefore less likely to answer. The resulting data will therefore be biased towards higher income values. To make matters worse, the probability of missingness depends on the missing value itself, and therefore other variables in the data frame may provide little information for the statistician to accurately impute the missing value. As is shown in this example, MNAR is exceedingly difficult to manage correctly. It is defined mathematically as follows $P(MISSING|y_o, y_m) = P(MISSING|y_m)$.

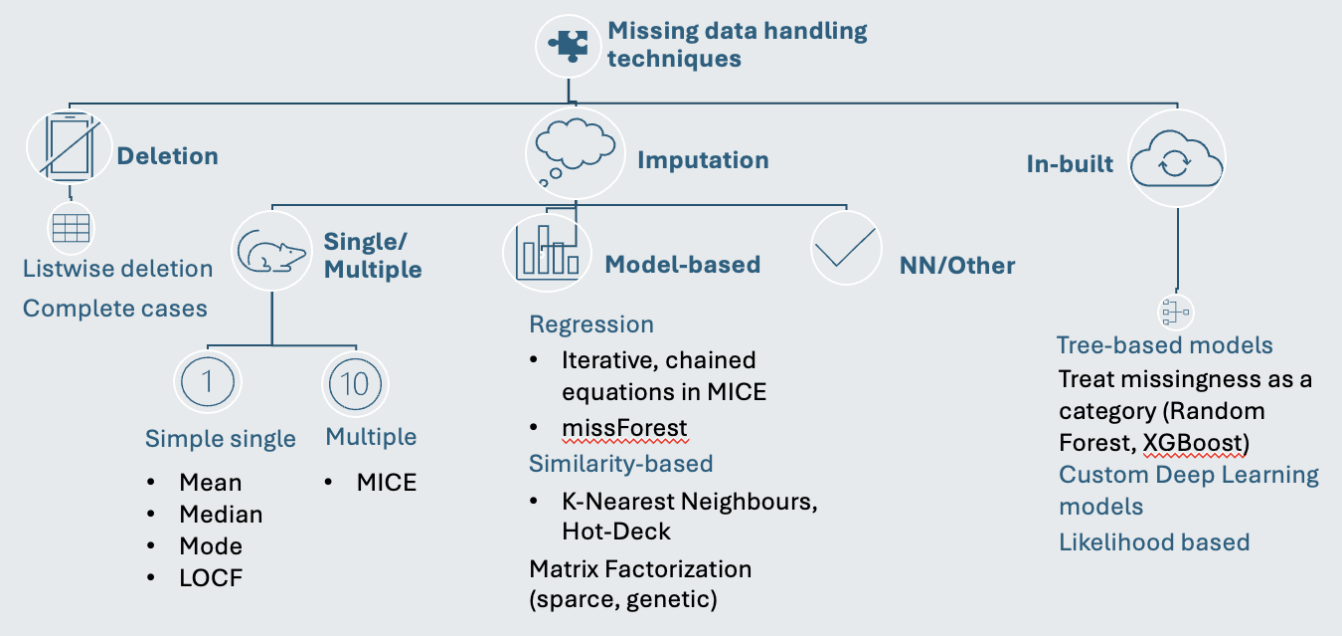
When conducting an analysis of data with missing variables, the statistician must assume one of these missingness mechanisms based on their understanding of the way that the data was generated. Often a statistician will assume that data are MAR simply because it is the most realistic assumption that still allows for some forms of imputation. Although these assumptions are not testable, Steyerberg (2019) observes that "MAR assumption becomes more reasonable with imputation models that include a wide range of characteristics, including predictors, the outcome, and auxiliary variables."

Handling missing data

Generally, missing data can be handled by deletion, imputation, or being embedded in prediction model (Figure 3). Deletion means deleting features with high level of missingness or removing some participants completely. Participants removal in particularly can lead to the loss of representativeness and reduced statistical power, and hence not recommended.

Another way to deal with missingness is to use a prediction model that internally handles missingness and does not require prior imputation. For example, tree-based models such as Random Forest and XGBoost, can make predictions from incomplete data. They treat missingness as a special data category and learn at each tree node whether the missing data is best split to the right or to the left.

Figure 3. Overview of missing data handling techniques



Handling missing data by imputation

In most cases, missing data is addressed via “imputation”, which is an umbrella term referring to any method used to guess the value of a missing datum. Below we briefly introduce some of the most popular and recommended imputation models.

- Mean imputation is quite common because of its simplicity and ease-of-use in popular statistical software (Lin *et al.*, 2019). However, it is not valid in MAR and MNAR, and, regardless of missingness type, it can have the undesirable effect of artificially decreasing the variance of the imputed variable (Lin *et al.*, 2019; Rubin, 1976). Therefore, it is most useful as a baseline comparator for other, more sophisticated imputation techniques.
- K-Nearest Neighbors (KNN) is a sophisticated imputation technique which works by creating a hyperplane of all variables in the dataset. It then uses a distance metric to find K observations with observed values that lie closest in the hyperplane to the observation with the missing value. The observations of the variable with the missing value in these neighbours are pooled and the pooled value is imputed (Lin *et al.*, 2019). This technique is comparatively easy to implement and only contains one parameter: number of neighbours considered. However, in non-Euclidean spaces, a different distance metric must be used.
- MICE can use any imputation strategy which leverages other variables in the dataset to predict the missing values of a particular variable. Instead of making a single guess, however, MICE creates a series of datasets, each with different imputations. Traditionally, the chosen model is run on all datasets independently, and the parameter estimates from each dataset are pooled together. The resulting pooled parameter estimates are the final parameter estimates of the study (Rubin, 1976). MICE can be difficult to implement and computationally expensive because of the many datasets it creates, but it is a common method for handling missing data, particularly when the MAR mechanism is suspected.

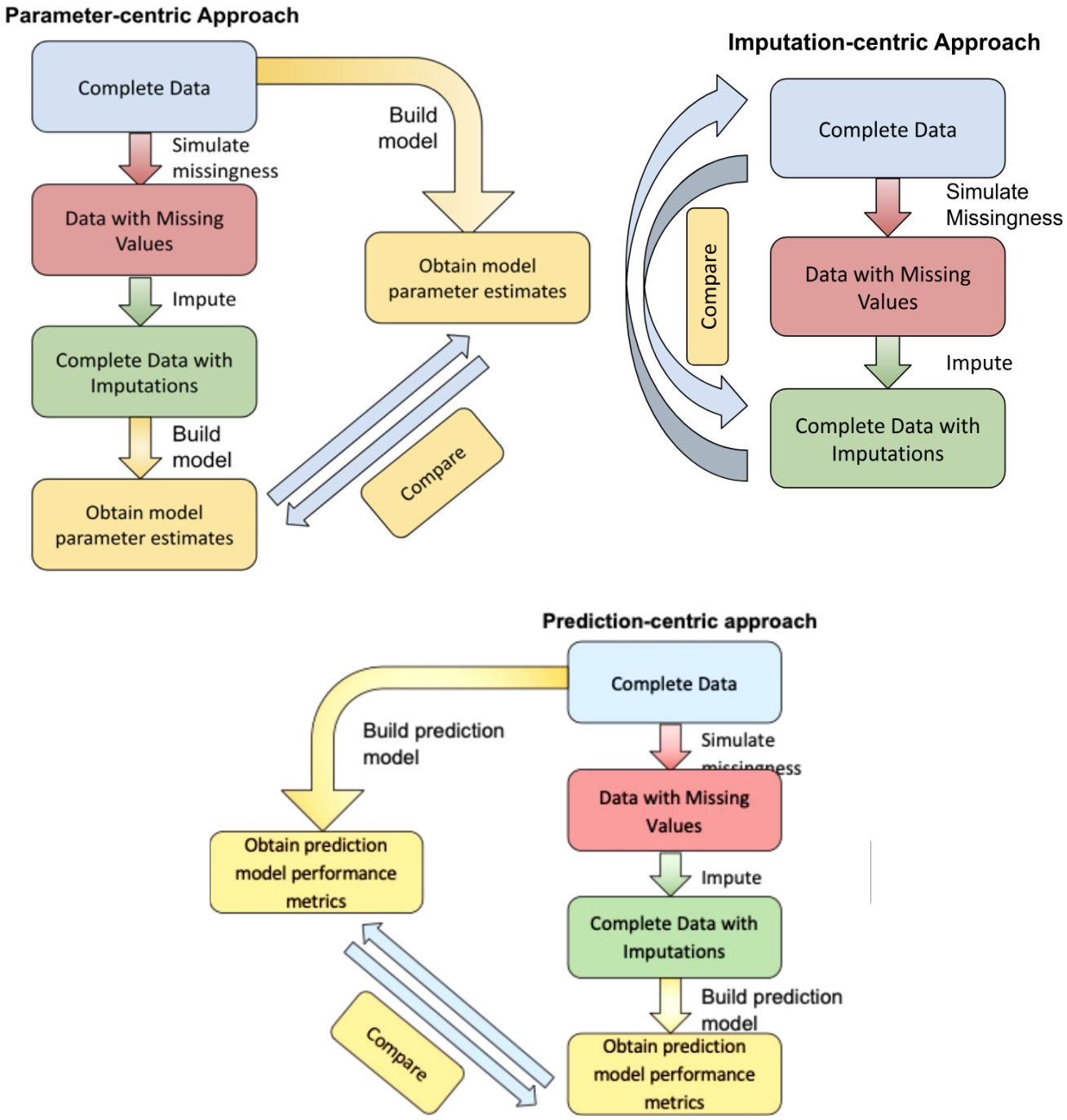
Choosing imputation model for clinical prediction

Imputation methods of all sorts have been studied in prediction modelling contexts (Lin *et al.*, 2019). However, metrics for the quality of the imputation vary widely. Some studies consider the accuracy of the imputation—how similar a given imputation is to the true value of the missing data—while others consider the accuracy of the imputations based on the values of the parameters of the models built on the imputed dataset (Lin *et al.*, 2019). A new and exciting imputation comparator assesses the validity of a series of imputation methods across

missingness mechanisms using the quality of the predictions *themselves* as the metric of imputation performance. To best understand the differences between imputation-centric, parameter-centric, and prediction-centric imputation evaluation metrics, the processes of each technique are diagrammed in Figure 3.

Imputation technique comparison via prediction metrics is not novel. For instance, Garciarena et al. (2017) consider imputation technique performance for a series of prediction models. While the predictions models, they consider are quite complex (Regularized Logistic Regression, Gradient Boosting, Deep Neural Networks, Support Vector Machines, to name a few), the imputation techniques they consider are relatively naïve, and include Mean imputation, Last Value Carried Forward, and Hot Deck imputation.

Figure 3: Imputation Evaluation Techniques using Simulated Missingness



Prediction modelling group projects: imputation for predicting survival outcomes.

In a project by our group members, we utilized a predictive approach to imputation technique evaluation and applied the method to less studied survival data. We compared naïve and sophisticated imputation techniques with the aim to find out which imputation technique would lead to a better predictive performance. The tested imputation models included 1) Multiple Imputations by Chained Equations (MICE), 2) Random Forest imputation, 3) K-Nearest Neighbors imputation, and 4) mean imputation as a baseline comparator.

Our experiments were based on two real-life clinical datasets, both with survival, and time-to-event, outcomes. First, we use the English Longitudinal Study of Aging (ELSA) dataset, which tracks nearly 6000 aging Britons over multiple decades. Here, we predicted the outcome of type 2 diabetes diagnosis, which was observed in 7.5% of participants. The second dataset was the German Breast Cancer Study Group (GBSG), which included 720 participants. The outcome was time to cancer recurrence or death, sometimes known as progression-free survival, and was observed for 42% of participants by the end of follow-up. We started from the complete data with no missing values, and created missingness according to MCAR, MAR, or MNAR, with a varied percentage of missing data (20%/40%/60%). Then, we trained and internally validated Cox-Lasso model and assessed predictive performance by Concordance index (C-index). C-index varies between 0.5 and 1.0, with 0.5 being a chance model and >0.9 implying excellent prediction. The experiments were repeated 50 times to get the mean and standard deviation for the estimated model performances. The results of these analyses are presented in Figure 4a and 4b.

Analyses for both datasets showed striking similarities: in nearly all combinations of missingness percentage, mechanism, and dataset, MICE underperforms other imputation techniques. In some contexts, as with high percentages of missing data in the GBSG set, MICE underperforms by a significant margin. Interestingly, KNN performs very similarly to other imputation techniques, except in the case of MNAR missingness in the GBSG dataset. This may be due to nonlinearities in the GBSG data which KNN is able to capture more effectively than other imputation techniques. Finally, it is notable that mean imputation, which is intended as a naive baseline, does not perform significantly worse than other imputation

techniques. One counterintuitive conclusion of this unexpected result is simply that these imputation tasks are quite difficult to do successfully, and therefore outperforming a naive baseline technique is nontrivial. A second and perhaps more hopeful possibility is that this analysis is somewhat limited in the range of hyperparameters it tests when tuning the imputation techniques. For example, KNN is only ever tested with k = 25 neighbors, while MICE imputation only considers one imputed dataset and only tests one imputation technique within it (predictive mean matching). Given these limitations, which are the result of simple computing capacity constraints, it is plausible that the more advanced imputation methods may exhibit better performance with more extensive hyperparameter tuning.

Summing up, we confirmed that predictive performance can significantly deteriorate in high missingness scenarios, and there is a variability in performances between the imputation techniques. Secondly, we found less differences between MNAR and MAR/MCAR scenarios despite theory suggesting this to be a much more difficult environment. By using this bespoke comparison technique, we hope our analysis provides a roadmap for future imputation research and ideally streamlines the handling of missingness in prediction models such that the true potential of precision medicine for patient benefit can be realized.

Prediction modelling group projects: PyPOTS (Python for Partially Observed Time Series).

Time series is another type of medical data, where missingness should be handled. Such data could include longitudinal information from medical devices, tracking of vital signs in intensive care units (heart rate, blood pressure, respiratory rate, dosage and administration of pharmacological treatments), electrocardiogram or electroencephalogram data, metabolic readings.

PyPOTS (Python for Partially Observed Time Series, https://pypots.com/) is a project by prediction modelling group members which addresses handling incomplete time series data. PyPOTS is an open-source machine learning library, which provides a collection of state-of-the-art algorithms for imputation, classification, clustering, and anomaly detection, specifically tailored for time series with missing values. Built on PyTorch, PyPOTS enables efficient model training and inference, supporting both research and practical applications in fields such as healthcare, finance, and IoT. The library offers an easy-to-use API, making it accessible for data scientists and researchers to experiment with advanced time series analysis techniques while ensuring scalability and performance.

In this article, we highlighted the complexities involved in choosing the most effective missing data handling technique while developing a predictive model. As demonstrated through our studies on survival data and time-series applications, different imputation strategies can significantly impact model outcomes. Moving forward, advancements in machine learning and statistical modeling will continue to refine imputation techniques, ultimately enhancing the reliability of predictive models in healthcare.

Figure 4a: GSBG Analysis Results

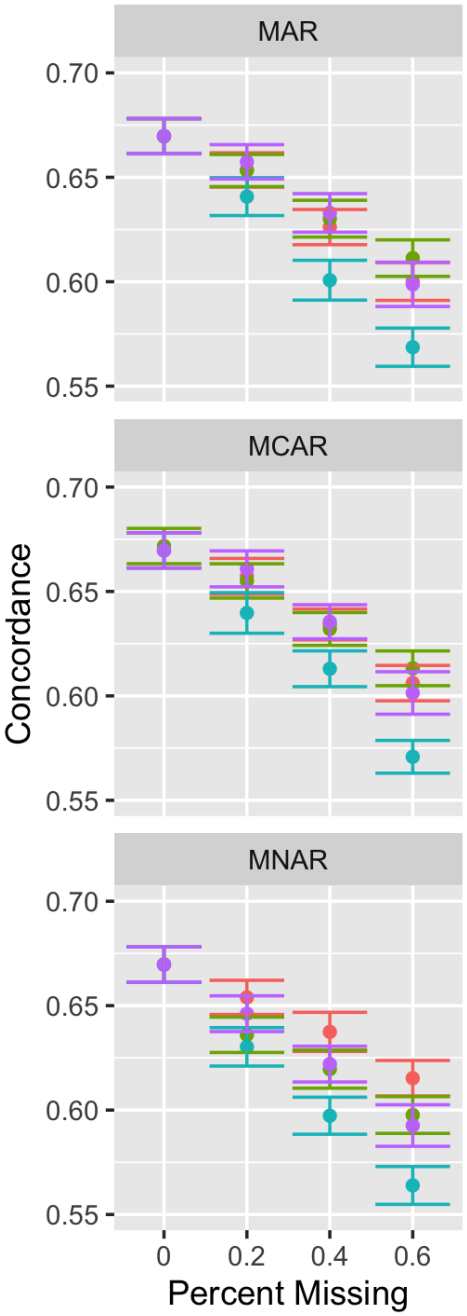
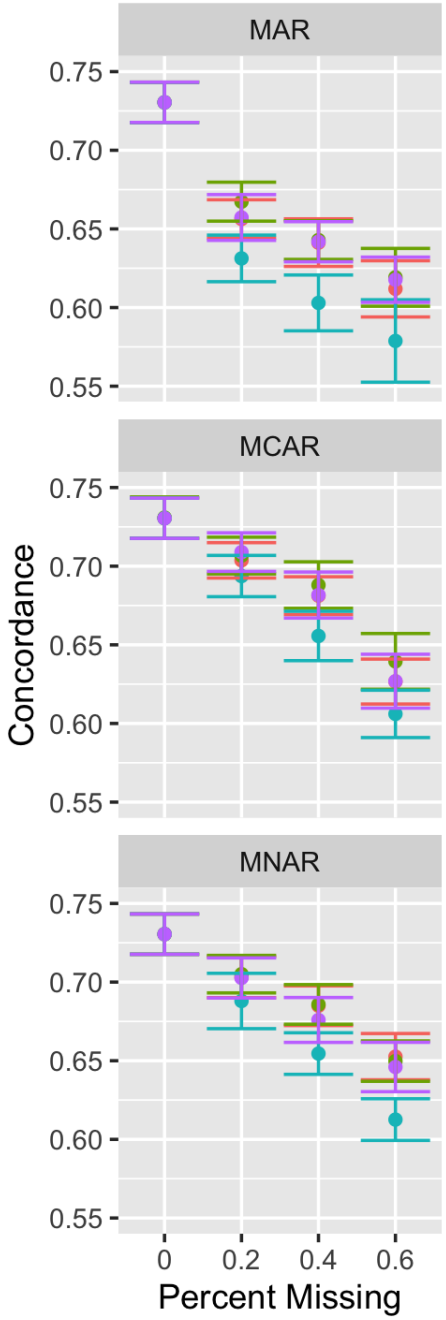
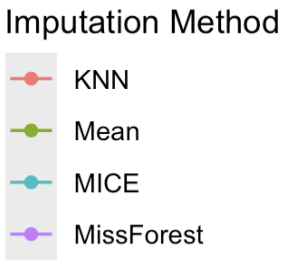


Figure 4b: ELSA Analysis Results





References

Austin, P. C., Stuart, E. A., & Small, D. S. (2021). Missing data in clinical research: A tutorial on multiple imputation. *Canadian Journal of Cardiology*, *37*(9), 1322-1331. https://doi.org/10.1016/j.cjca.2020.11.010

Garciarena, U., & Santana, R. (2017). An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. *Expert Systems with Applications*, *89*, 52-65. https://doi.org/10.1016/j.eswa.2017.07.026

Lin, W.-C., & Tsai, C.-F. (2020). Missing value imputation: A review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, *53*(2), 1487-1509. https://doi.org/10.1007/s10462-019-09709-4

Luo, J. C., Zhao, Q. Y., & Tu, G. W. (2020). Clinical prediction models in the precision medicine era: old and new algorithms. *Annals of translational medicine*, *8*(6), 274. https://doi.org/10.21037/atm.2020.02.63

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, *63*(3), 581. https://doi.org/10.2307/2335739

Scheffer, J. (2002). Dealing with Missing Data. *Res. Lett. Inf. Math. Sci*, *3*, 153–160. https://mro.massey.ac.nz/server/api/core/bitstreams/3b2f5383-da88-4ab9-b426-66f05b01339a/content

Steyerberg, E. W. (2019). *Clinical prediction models: A practical approach to development, validation, and updating* (2nd ed.). Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-16399-0