*MEDINFO 2023 — The Future Is Accessible*
*J. Bichel-Findlay et al. (Eds.)*

*doi:10.3233/SHTI231252*

# Compatibility in Missing Data Handling Across the Prediction Model Pipeline: A Simulation Study

Antonia TSVETANOVA[a,1], Matthew SPERRIN[a], David JENKINS[a], Niels PEEK[a,b], Iain BUCHAN[c], Stephanie HYLAND[d], Glen MARTIN[a]
[a] *Centre for Health Informatics, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, England, UK*
[b] *The Christabel Pankhurst Institute for Health Technology Research and Innovation, University of Manchester, Manchester, England, UK*
[c] *Institute of Population Health, University of Liverpool, Liverpool, England, UK*
[d] *Microsoft Research Cambridge, Cambridge, England, UK*
ORCiD ID: Antonia Tsvetanova https://orcid.org/0000-0003-1875-1921

**Abstract.** Careful handling of missing data is crucial to ensure that clinical prediction models are developed, validated, and implemented in a robust manner. We determined the bias in estimating predictive performance of different combinations of approaches for handling missing data across validation and implementation. We found four strategies that are compatible across the model pipeline and have provided recommendations for handling missing data between model validation and implementation under different missingness mechanisms.

**Keywords.** Statistical models, missing data, imputation, simulation study

## 1. Introduction

Proper handling of missing data is critical for ensuring optimal accuracy of clinical prediction models (CPM). Current guidelines for handling missing data are inadequate for risk prediction modelling, where missing data can occur in model development, validation, or implementation. Emerging recommendations suggest that the approach to handle missingness should be compatible throughout CPM production. We determined which combinations of approaches estimate the performance of a CPM without bias, and which are appropriate when the model does not allow missingness at deployment.

## 2. Methods

To determine which missing data handling approaches are compatible across model validation and implementation, we performed an extensive simulation study, focusing on

[1] Corresponding Author: Antonia Tsvetanova, Centre for Health Informatics, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, England, UK; email: antonia.tsvetanova@manchester.ac.uk.

a logistic model for binary outcome, where data were generated under various missingness mechanisms for validation and implementation. We imputed data using the following approaches: multiple imputation (MI) including or omitting the outcome in the imputation model, mean imputation, and complete case analysis (CCA).

We calculated the bias in the predictive performance by subtracting the 'ground truth', i.e. implementation performance metrics estimates, from performance metrics estimates of each validation dataset for every combination of missing data handling approach. We determined which imputation-method-model-pipeline stage combinations are compatible and which ones led to bias in the C-statistic, Brier Score, Calibration-in-the-large and Calibration slope. Data generating mechanisms are presented in Figure 1.

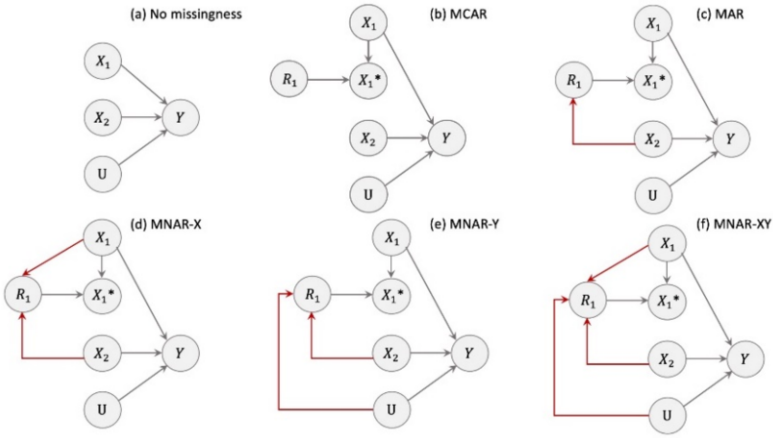


**Figure 1.** Directed acyclic graphs for all missingness structures. In a) no missingness is used to generate the development data, to which the model is fit; (b) – (f) are used for validation and implementation data generation. $X_1$ is continuous or binary; $X_1$* is the observed part of $X_1$. $X_2$ is always observed, U is an unmeasured variable which potentially induces the relationship between the missingness indicator ($R_1$) and the outcome Y. MCAR = missing completely at random, MAR = missing at random, MNAR = missing not at random.

## 3. Results

When the model allowed for missing data at deployment, we found that using the same method to handle missing data between validation and deployment resulted in unbiased estimates of the model performance. This was the case when the missingness mechanism remained constant across the two stages or changed from MCAR/MAR and MAR/MCAR. When missing data was prohibited at deployment scenarios, using either MI with Y or CCA at validation resulted in unbiased predictive performance estimates under MCAR, MAR and some special cases of MNAR.

## 4. Conclusions

We provide evidence that commonly used combinations of approaches for handling missing data across CPM validation and deployment may not be compatible and may result in bias of the CPM's predictive performance. Choice of method for handling missing data should be based on whether a shift in the missingness mechanism is expected in deployment, and/or whether the model allows for missingness at deployment.