

International Journal of Population Data Science

Journal Website: www.ijpds.org



Handling of missing values in whole-population electronic health records: a simulation study

Alexia Sampri¹, Samantha Ip¹, Carmen Petitjean¹, Paul Madley-Dowd², Tim P Morris³, Kate Tilling², Jonathan A C Sterne², and Angela M Wood¹

¹University of Cambridge, Department of Public Health and Primary Care, Cambridge, United Kingdom

²University of Bristol, Bristol Medical School, Bristol, United Kingdom

³MRC Clinical Trials Unit at UCL, University College London, London, United Kingdom

Objectives

This study evaluates the scalability of multiple imputation methods, specifically Multivariate Imputation by Chained Equations (MICE), for addressing missing data in whole-population EHRs (Electronic Health Records). We investigate the impact of the number of imputations, subsampling strategies, missing data mechanisms, and missingness levels on the accuracy of the results.

Methods

A simulation study was conducted using whole-population NHS England EHRs (primary/secondary care, and COVID-19 related records) from January 2024. We simulated missing BMI (Body Mass Index) scenarios under MAR (Missing At Random) and MNAR (Missing Not At Random) conditions. We examined the effect of MICE combined with subsampling strategies (i.e., the imputation and analysis models derived in subsamples and/or the full population) on the accuracy and precision of estimates for the association between COVID-19 vaccination status and severe outcomes adjusted for BMI and other confounders, using logistic regression. We evaluated the estimand accuracy, computational efficiency, and the environmental impact of the imputations.

Results

Preliminary results indicate that MICE effectively manages missing BMI data within large EHR datasets, preserving the integrity and accuracy of statistical outputs through grouped logistic regression. However, analysis reveals that subsampling strategies (e.g., deriving imputation model in 20% random sample whilst executing analysis on full population) and number of imputations can substantially reduce runtimes, memory usage and environmental impact, but compromises the accuracy of the adjusted log odds ratios and their corresponding standard errors, underscoring the importance of carefully chosen imputation strategies. We plan to extend this method-

ology to explore other critical missing variables within EHRs, such as blood pressure, ethnicity, and cholesterol levels, to further validate the versatility of MICE in handling diverse data types.

Conclusion

Scalable imputation methods like MICE are promising for robust analysis of EHR datasets, ensuring accuracy and data completeness while optimising computational resources and minimising the carbon footprint of data-intensive analyses.



Temporary page!

$\LaTeX$ was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because $\LaTeX$ now knows how many pages to expect for this document.