Vaar Project
Missingness in health data: evaluating imputation and complete case analysis on downstream models

- University of Sunderland (online)
- Computer Science with Data Science MSc
- Student: Amanda Harris

- Date: August 2023

Contents

Introduction

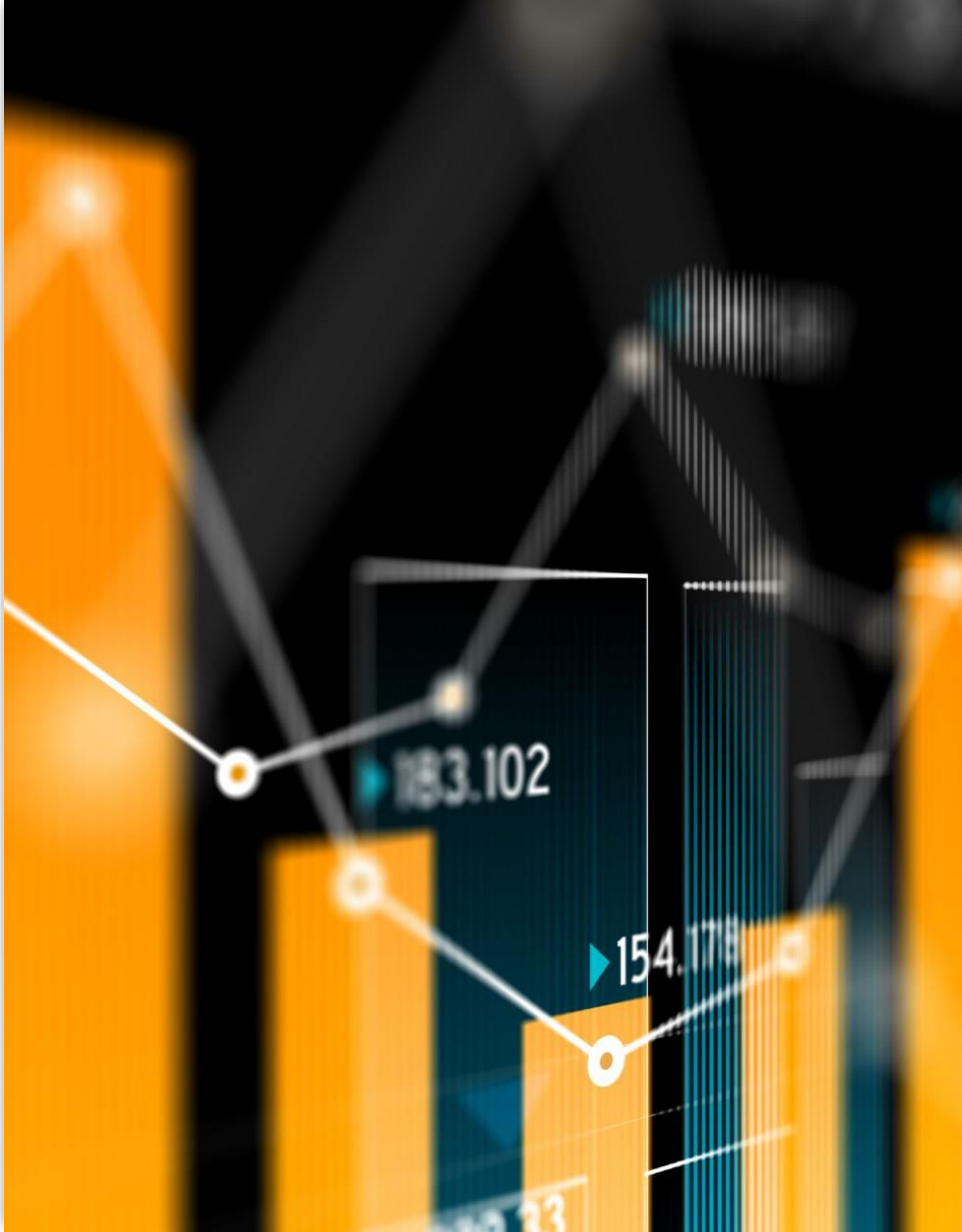
Purpose, relevance, objectives and importance

Purpose and Relevance

- Health data provides import insights but it’s typically not research-ready
- Missing data a common issue
- The aim is to develop a prototype system for evaluating data imputation efforts
- Vaar is an Orcadian sailing term meaning to guide or direct
- Can the Vaar project contribute to providing some clarity?

Why is the Vaar Project important?

- Good practice for managing and reporting missing data not often followed
- Missing data is a challenging issue:
    - Cannot statistically differentiate between missing at random (MAR) and missing not at random (MNAR) data
    - Missingness mechanism important consideration for management method chosen
    - Time consuming to pre-process data
- Inappropriate methods can lead to:
    - Biased results
    - Fragile results
    - Invalid conclusions
    - Loss of information
    - Poor generalisation



Vaar System Objectives

- Evidence-based
- Open, transparent and reproducible
- Adaptable to variability of data issues and problems
- Ascertain MCAR or MAR/MNAR
- Recommended an approach that considers stability of test model results
- Evaluate imputation efforts against test model

Theoretical Underpinning

Literature Review

Areas of Consensus

- Reason for missingness is an important first step
- Complete case analysis (CCA) is usually only appropriate if data is missing completely at random (MCAR)
- Outliers often represent valuable information that must not be discarded
- Normalisation reduces bias from skewed data in downstream models
- Scaling allows models to compare relative relationships between data points more effectively
- Missing data patterns and information from auxiliary variables impact imputation effectiveness
- Sensitivity analyses should be conducted with different methods

Lack of Consensus

- The proportion of missing data at which imputation no longer boosts performance
- Relationship between reason for missingness and most effective imputation models, with particular challenges on handling non-ignorable missingness (MNAR – missing not at random)
- The effect of data distribution on pre-processing steps

Hypotheses tested

Imputation out-performs CCA with up to 50% missingness where data is MCAR or MAR.

Reasons for missingness impacts the effectiveness of different imputation approaches, especially where missingness levels are higher.

Missing data patterns in dependent, independent and confounder variables impact imputation effectiveness.

Research Methods

Practical and Experimental Work

Research Design

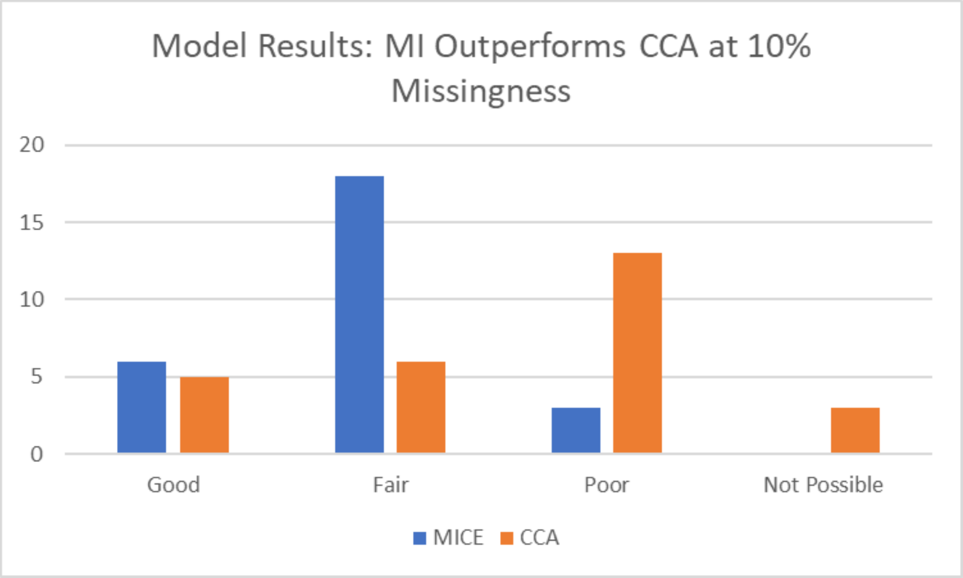
Method/ Activity	Purpose	Implementation
Quantitative survey.	To ensure prototype system supports common tools and requirements.	Microsoft Form shared with NHS-R and Python communities via Slack.
Quantitative experiments: compare missingness mechanisms and missingness levels managed with CCA or MI on downstream data models.	The **primary** analysis method to test hypotheses and provide evidence for prototype system.	• 9 complete UCI health datasets with 10%, 20%, 50%, 70% and 90% missingness simulated (MCAR, MAR and MNAR) using missMethods R package. (Rockel, 2022) • Complete data created for each using CCA and MICE R package for MI (van Buuren et al. 2023) as far as missingness levels and mechanisms would allow. • 9 baseline data models created with original complete data, compared to models based on CCA and imputed datasets. • All models evaluated against baseline test data.
Qualitative test and code review.	To peer-review code and test usability and effectiveness of prototype system.	Code review, usability task and qualitative interview.

Data Experiments

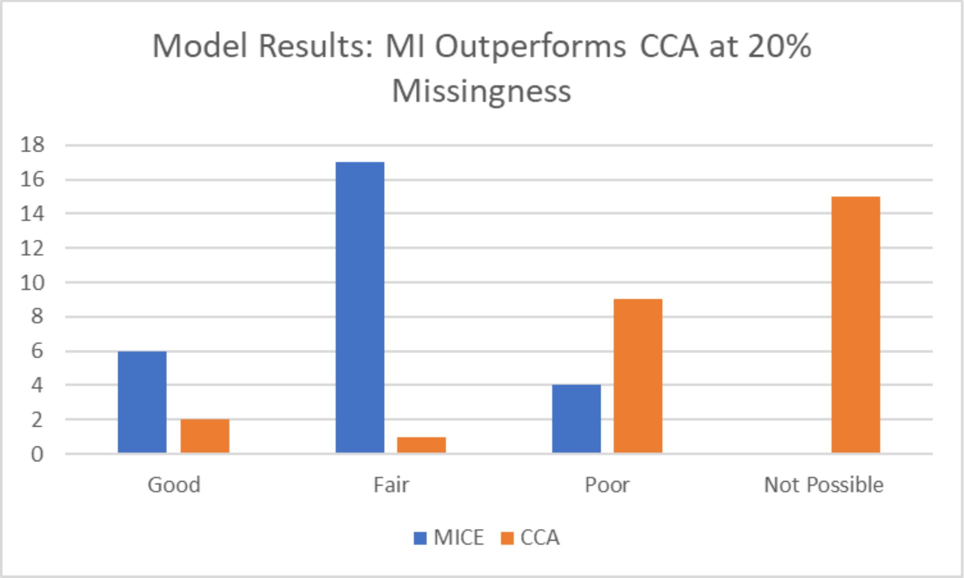
Outcomes, Results and Evaluation

MI Consistently Outperforms CCA

From 10% Missingness

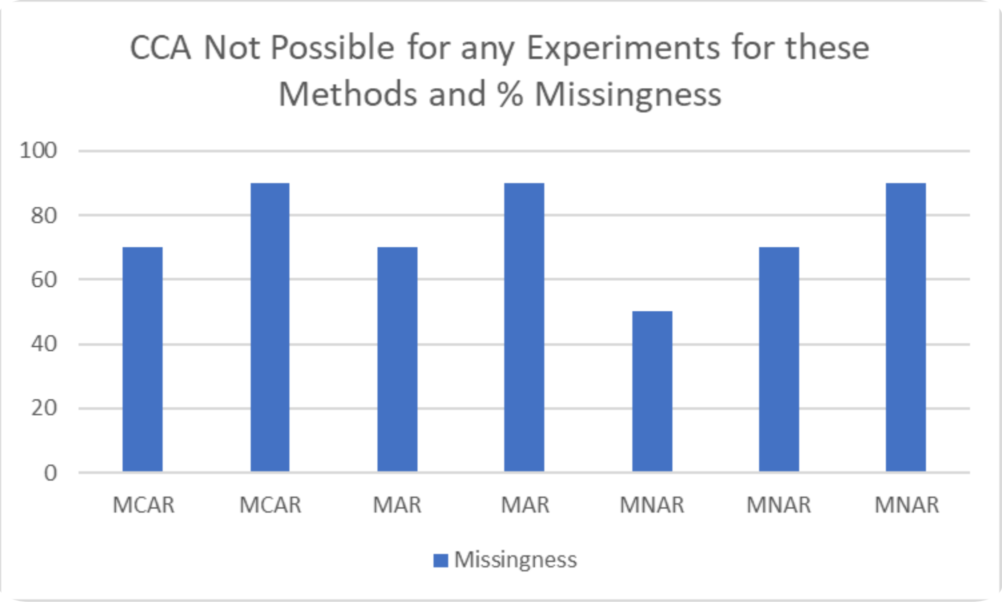


To 20% Missingness

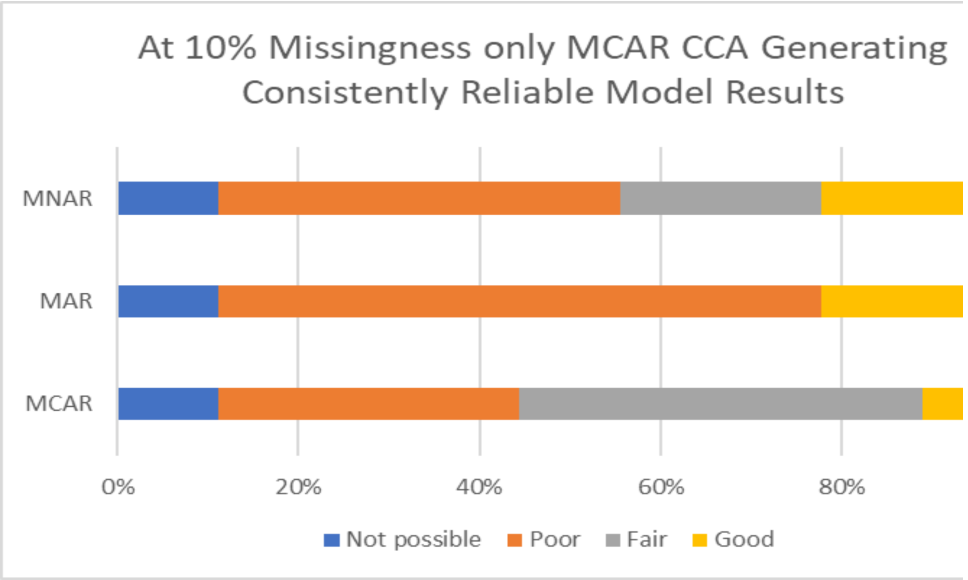


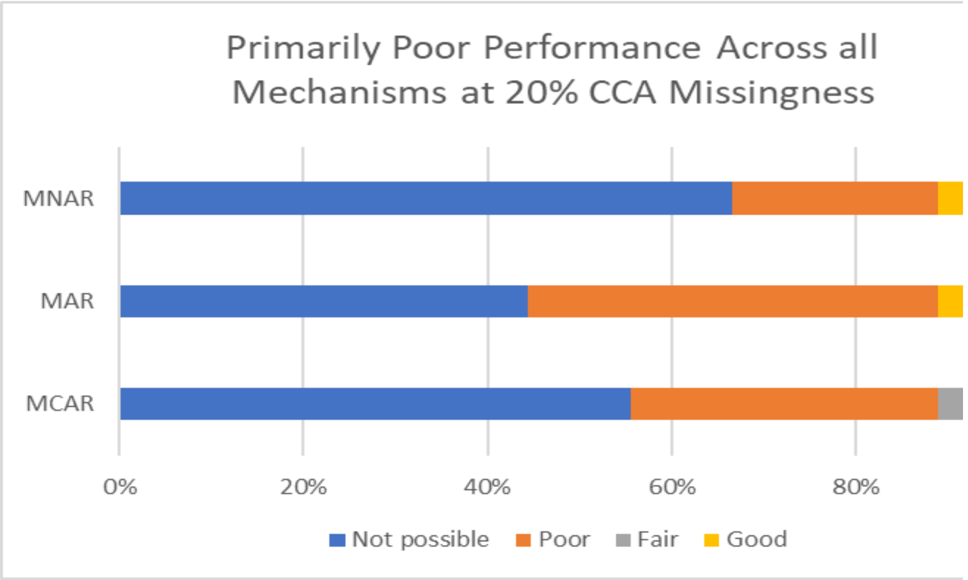
Comparisons Not Possible Beyond 20%

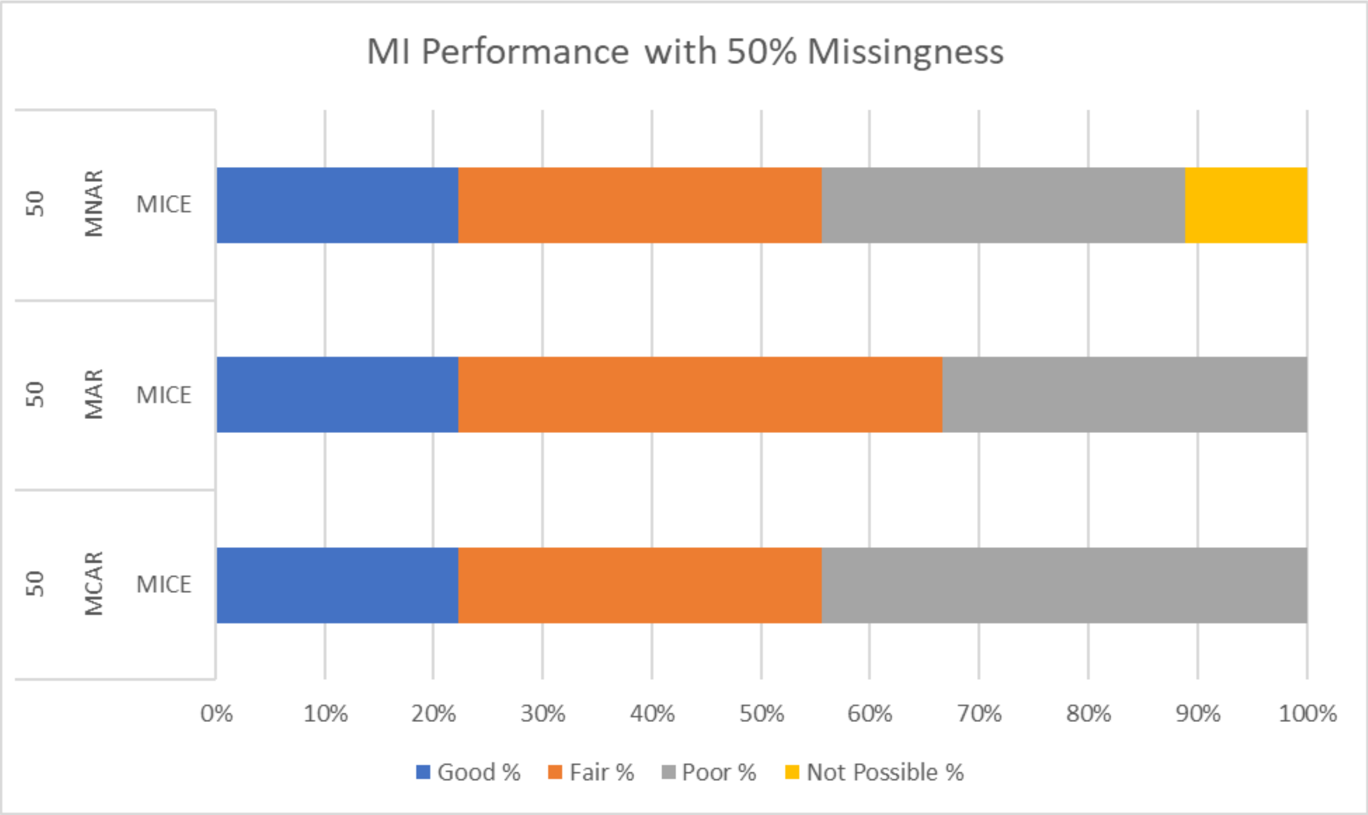
- Lost Learning Opportunities
  - No workable experiments beyond 20% missingness
  - Complete case analysis quickly becomes unworkable



MCAR CCA More Consistent Performance to 10% Missingness

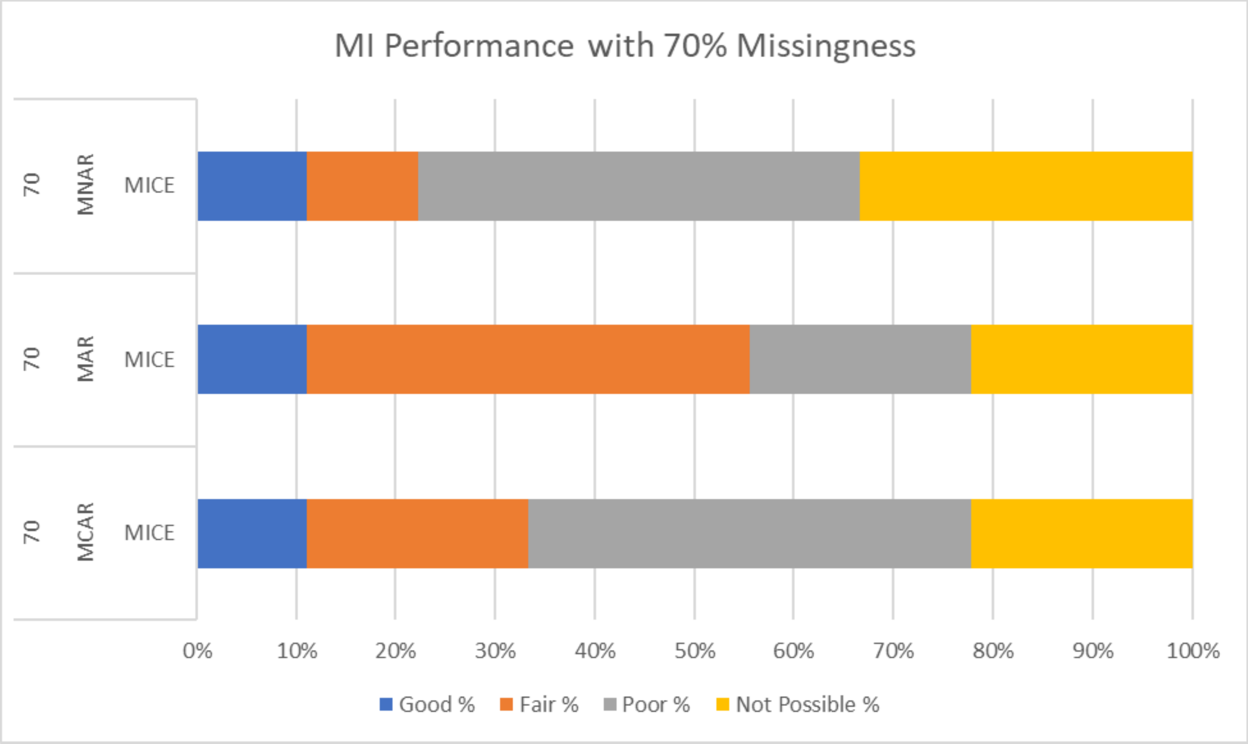






MI Returning Fair Results at 50% Missingness

Only MAR MI Performance Consistent at 70% Missingness



QUANTITATIVE SURVEY KEY POINTS

RESPONSE RATE

Too few responses for requirements decision making.

Requirements determined by secondary research; survey illustrative and indicated other areas for desk research.

DESK RESEARCH

Additional Python packages noted to explore further.

OPINION REFLECTIVE OF LITERATURE

Consensus and differences in opinion, reflective of areas where there is lack of consensus in literature.

VARIABLITY

Variability, adaptability and flexibility themes noted in comments throughout.

Vaar Notebook

System Requirements, Design and Demonstration

Vaar Requirements

Objectives	Solution
Open, transparent and reproducible	Markdown Notebooks created in R Studio - R and Python Versions - Templates and exemplars created - Open-source packages - Transparent code
Adaptable to data variability and problems	Key steps included in Notebook, users can add additional code as required.
Whether data is MCAR or MAR/ MNAR	Little’s MCAR Test (results sometimes not possible due to data singularity) and data missingness pattern.
Recommend missing data approach based on stability of results	Sensitivity analysis using delta adjustment that also considers missingness mechanism, level of missingness and missing data pattern.
Evaluate imputation efforts against test model	CCA compared to MI results for test model, alongside recommended approach.

Links to Templates and Vaar Exemplars

- Templates in github:
  - [bi23le/Vaar (github.com)]
- **View Python Exemplar HTML Page**
- [Breast Cancer Wisconsin Original] (https://rpubs.com/bi23le/1070978)
- **View R Exemplar HTML Pagess**
- [Breast Cancer Wisconsin Original] (https://rpubs.com/bi23le/1070975)
- [Cervical Cancer Risk Factors] (https://rpubs.com/bi23le/1070977)

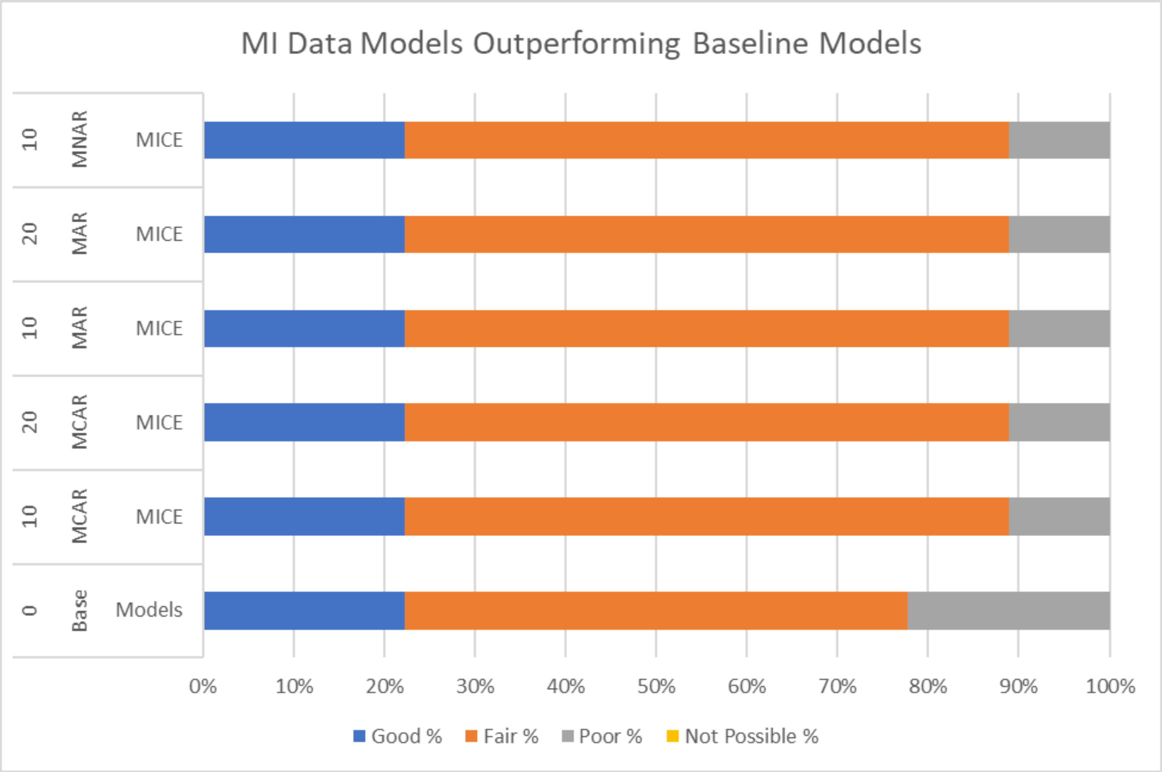
Discussion

Limitations and Potential Extensions

Limitations

- Low survey numbers meant Vaar requirements based on secondary research
- More effective missing data packages in R than Python (e.g. MICE for multiple imputation) (van Buuren et al. 2023)
- Python version calls on naniar R package for Little’s MCAR Test and add_prop_miss function (Tierney, N, 2023)
- More thorough testing required for different datasets and problems
  - Flexibility built into Notebook approach
  - Open code to allow for adaptability
  - Logic rules and variable setting are the core Vaar elements

Surprising Result: MI Outperforming Baseline Models



References

- van Buuren, *et al.* (2023) ‘mice: Multivariate Imputation by Chained Equations’. CRAN. Available at: https://CRAN.R-project.org/package=mice (Accessed: 10 August 2023).
- Mandreoli, F. *et al.* (2022) ‘Real-world data mining meets clinical practice: Research challenges and perspective’, *Frontiers in Big Data*, 5. Available at: https://doi.org/10.3389/fdata.2022.1021621.
- Rockel, T. (2022) ‘missMethods: Methods for Missing Data’. R CRAN. Available at: https://CRAN.R-project.org/package=missMethods (Accessed: 11 August 2023).
- Tierney, N. (2023) ‘naniar: Data Structures, Summaries, and Visualisations for Missing Data’. cran.r-project.org. Available at: https://CRAN.R-project.org/package=naniar (Accessed: 11 August 2023).

Thank you