PCN432

# Evaluation of Missing Data Imputation Strategies in Clinical Trial and EMR Data Using Standardized Data Models

McLean C[1], Ransom J[1], Galaznik A[1]

[1]Medidata Solutions, Boston, MA, USA

## Background

- Data missingness is a major challenge and a source of bias in evidence-based medicine.[1]
- Missing data is common in both randomized controlled trials (RCT) and observational real-world data (RWD) studies.[2]
- Records that contain incomplete data are often excluded from studies, but this strategy can introduce bias. Biases tend to differ between RCT and observational RWD datasets.[3]
- In this study, we conducted an assessment for methods for imputing missing data in both RCT and RWD acute myeloid leukemia (AML) datasets.
- A challenge of this type of analysis is the presentation of data in variable formats.[1] To address this issue, an evaluation of the utility of standardizing data format was also performed.

## Objectives

- Evaluate the utility of different techniques for imputing missing data
- Evaluate the utility of standardizing data format

## Methods

### Data Source

- The clinical trial data cohort was derived from a pooled dataset of 7 clinical trials (n=719) for relapsed/refractory AML, conducted from March 2008 - Nov 2017, from the Medidata archive of > 3,000 trials.[4]
  - Pooling was accomplished through harmonization to Clinical Data Interchange Standards Consortium Study Data Tabulation Model version 1.4.[5]
- De-identified Oncology Electronic Medical Record (EMR) data was obtained from the Guardian Research Network™ (GRN) of integrated delivery systems from Jan 1990 – July 2018.
  - GRN is a nationwide consortium that aggregates hundreds of thousands of cancer patients' electronic medical records from multiple integrated community health systems into a single searchable database.[6]

### Variables for Imputation

- Data containing the same five variables for each data source were created as sample data (Table 1).
- Metrics were selected to represent a number of potential variables of interest in both clinical and health outcomes study contexts.
- Imputation was evaluated using categorical, continuous numerical, and binary values.
- Only non-demographic observations were imputed; demographic variables were used as covariates in imputation processes.

### Data Transformation and Analysis

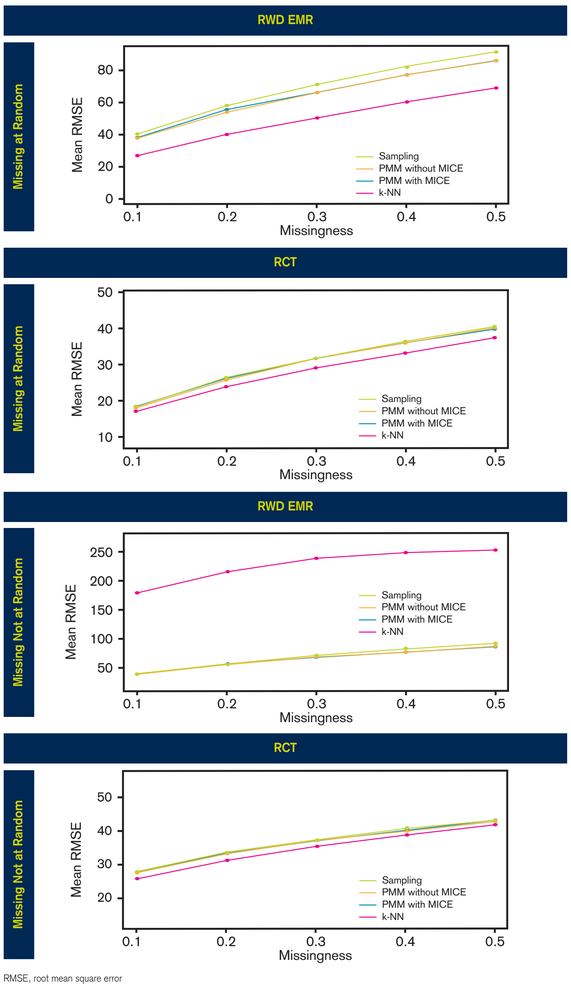
- Both RCT and RWD datasets were converted to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), version 5.[7]
- Analyses were generated using SHYFT Quantum version 6.7.0 and R v 3.2.5.
- The standardized code sets present in the OMOP CDM and the SHYFT STRATA and QUANTUM platforms were used to guarantee that variables were derived identically for each dataset (Table 2).

### Study Design

- Missingness was artificially introduced into the variables to be imputed using the ampute function of the R Package Multiple Chained Imputation Equations (MICE).
- Missing at random and missing not at random patterns of missingness were evaluated.
- The missing not at random missingness pattern assumed that the missingness was most dependent on treatment response, somewhat dependent on stratification and treatment variables, and least dependent on demographic variables.
- Missingness was introduced at different levels from 10% to 50% of the data. It was assumed that any number of variables to be imputed could be missing at any given time.
- Imputation was conducted using the R Packages MICE and Data Mining with R (DMwR).
- To represent a baseline for comparison to all other imputation strategies, the data was first imputed by randomly sampling non-missing values. This method was then compared with predictive means matching (PMM) without MICE, PMM with MICE, and K Nearest Neighbors (k-NN) imputation methods.
- For each missingness pattern, dataset, and imputation method, 500 unique amputed datasets for imputation were generated.
- These unique amputed datasets were imputed and the imputation results were compared to the original dataset with no missingness.

## Results

Figure 1: RMSE

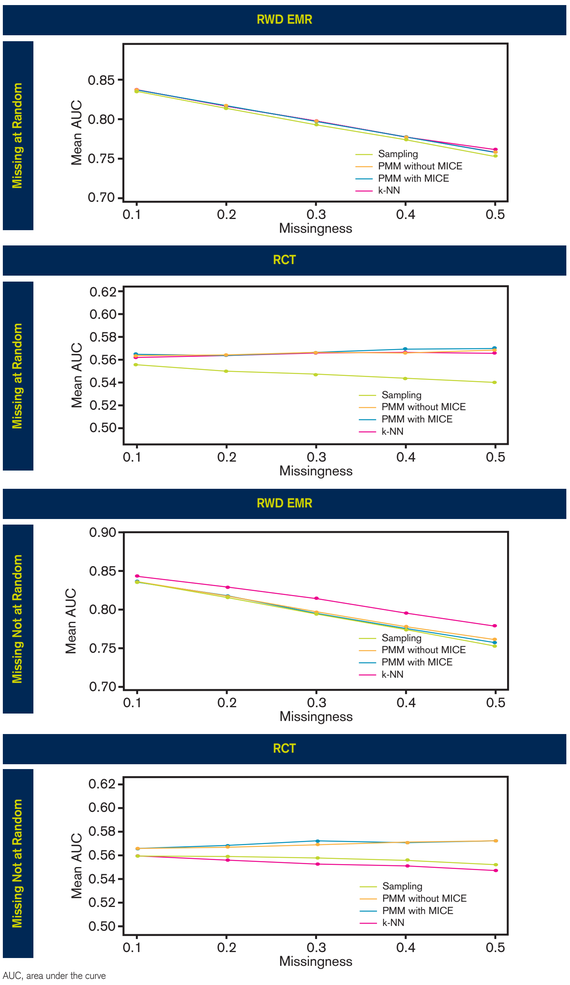



RMSE, root mean square error

Table 1: Variables for Imputation

| Variable | Research Usage | Type | Imputed |
|---|---|---|---|
| Gender | Demographic | Categorical | No |
| Age at Index | Demographic | Continuous | No |
| Count of Blood Transfusions | Patient Stratification | Continuous | Yes |
| Presence of Azacitizine Exposure | Treatment | Binary | Yes |
| Time to Death | Response | Continuous | Yes |

Figure 2: AUC




AUC, area under the curve

Table 2: Data Set Summary Statistics

| | RWD | RCT |
|---|---|---|
| Total (N) | 2,002 | 719 |
| Gender (%) | | |
| Male | 52 | 58 |
| Female | 48 | 42 |
| Age at Index – Mean years (SD) | 63.8 (17.2) | 68.9 (39.2) |
| Blood Transfusion – Mean years (SD) | 142.0 (362.0) | 14.0 (20.0) |
| Azacitidine Exposure (%) | 13 | 5 |
| Time to Death – Mean years (SD) | 0.4 (3.9) | 0.9 (137.7) |

SD, standard deviation

## Summary

- The performance of our imputation models was in many cases consistent with other investigations,[8,9] though it would seem in some cases MICE performs less well than we might expect.
- k-NN imputation generally performs best compared to other methods, but the degree to which it outperforms even random sampling varies by data type and missingness.
- Imputation was generally more successful in missing at random datasets.
- While models for EMR data appear to be more predictive based on the AUC metric, they are more error-prone.
  - This is likely because even in a relatively complete dataset, EMR provides more patients with fewer records each, and thus there may be relatively few events of interest per patient.
  - For example, in the missing not at random EMR imputations, the skewed distribution of blood transfusion events introduces very high levels of error into the k-NN imputation.
  - Because EMR may more often be missing at random, k-NN imputation may remain a viable strategy.
  - It is worthwhile to consider using a subset of exceptionally well-captured individuals as a training dataset for k-NN imputation in any EMR dataset.
- In general, the varying performance of each imputation strategy suggests significant value for repeated evaluation of these strategies whenever an imputed dataset is to be used for analysis.
- During this evaluation, a key enabler of the repeat analysis was the usage of the OMOP CDM standard model.
  - Consistent cohorts and variables between datasets could be reliably identified.
  - Analysis could be readily streamlined due to standardized data model.

## Limitations

- This study represents a preliminary investigation into these methods, and not all components of imputation methodology could be considered.
  - While the variables selected were chosen to be representative of variables of interest in health economics and outcomes research, they do not represent a comprehensive or large set of variables. The number of variables to be imputed can significantly impact algorithm performance.
  - Similarly, identical parameterization was used for all datasets and missingness types. In a true application of data imputation, these would ideally be tuned to the data itself.
  - No additional training dataset was used in any method, and we would expect this to impact the performance of different methods.
- The disease cohort selected for this research may be somewhat idiosyncratic: AML is a rapidly progressing disease and thus measures related to treatment and response may behave atypically compared to what we would see in other therapeutic areas of interest.

## Conclusions

- Imputation techniques can significantly improve the informativeness of health economics and outcomes research when appropriate methods are tested and applied.
- Clinical data standards such as the OMOP CDM are well suited to enable rigorous and repeatable methodological evaluations, which should be a key consideration when imputing a dataset.

### References

1. Bell ML et al. Differential dropout and bias in randomised controlled trials: When it matters and when it may not. BMJ 2013; 346: e8668.
2. Berger M et al. Opportunities and challenges in leveraging electronic health record data in oncology. Future Oncology 12:10.
3. Gunsoy N et al. How to tackle the estimation of treatment impact in the presence of differential withdrawal and missing data among study arms? ISPOR Conference Workshop, May 21 2019.
4. https://www.cdisc.org/standards/foundational/sdtm.
5. https://www.guardianresearch.org/.
6. https://www.ohdsi.org/data-standardization/the-common-data-model/.
7. Grinblatt DL et al. Transfusion independence in patients with hematologic disorders receiving Azacitidine who are enrolled in AVIDA, a longitudinal patient registry. Blood, 2008, 112 (11): 2683.
8. Schmitt P et al. A comparison of six methods for missing data imputation. J Biom Biostat 6:224.
9. Jadhav, A et al. Comparison of Performance of Data Imputation Methods for Numeric Dataset. Applied Artificial Intelligence. 33:10.

### Disclosures:

- CM, JR, and AG are employees of Medidata Solutions.

*Presented at ISPOR Europe 2019, 2-6 November 2019*
*Copenhagen, Denmark*

QR CODE