Imputation of Continuous Measurements in Large Healthcare Databases: Comparing the Performance of Imputation Algorithms

Stephen P Fortin[1], Jenna Reps[1]
[1]Janssen R&D, LLC, Raritan, NJ, USA

Background

- In real world evidence research, the use of continuous measurements is often avoided due to incomplete recording and lack of standardization
- Nevertheless, continuous measurements may contain valuable information and such practices may lead to significant information loss during model development

Study Objectives: To assess the following among patients newly diagnosed with heart failure:

1. The extent to which continuous measurement values are recorded across databases mapped to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)
2. The performance of various imputation methods to estimate values for continuous measurements

Methods

Study Design: Descriptive study
Data Sources: Data were from the 6 large healthcare databases:

1. IBM MarketScan® Commercial Database (CCAE)
2. Optum® de-identified Electronic Health Record Dataset (Optum EHR)
3. IBM MarketScan® Medicare Supplemental Database (MDCR)
4. Japan Medical Data Center (JMDC)
5. Optum® De-Identified Clinformatics Data Mart Database – Date of Death (Optum DOD)
6. Optum® De-Identified Clinformatics Data Mart Database – Socioeconomic Status (Optum SES)

Study Population: Hospitalized patients with a first-time diagnosis of heart failure occurring on or after 01-01-2017 and at least 365 days of prior continuous observation

Covariates

- Continuous measurements included all measurement concepts with at least 1 observed value in the "value_as_number" field at or within 365 days of index (value=last recorded at or prior to index)
- Patient characteristics included age, sex and clinical characteristics (i.e., components of the Charlson comorbidity index)

Statistical Analyses

- For each database, we measured the prevalence of all observed continuous measurements occurring at or within 365 days of index

Imputation of Continuous Measurements

- Conducted in Optum EHR, which is known to contain more continuous measurements as compared to administrative claims databases
- Continuous measurement values for continuous measurements with a prevalence ≥30% in the year prior to index were mapped to a uniform scale
- The missCompare R package used to compare performance of 13 imputation methods, including: mean, median, multiple, random forest, k-nearest neighbor and principal component analysis (PCA) imputation
  - For each patient, missCompare simulated patient characteristics and continuous measurements, but without missing values, while preserving correlations between covariates
  - Missingness was then simulated by removing measurement values under different missingness assumptions (i.e., missing completely at random [MCAR], missing at random [MAR], and missing not at random [MNAR]).
  - Each imputation method was tested across 10 randomly selected subsamples of 1,000 patients drawn from the simulated data.
- Performance assessed using computing time and root-mean squared error (RMSE)

Results

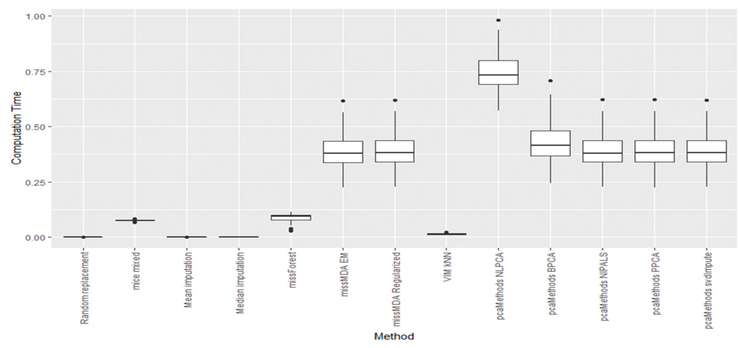
- A total of 7,181 to 261,824 patients met the study criteria across each database; including 61,345 in Optum EHR
- **Table 1** summarizes the prevalence of continuous measurements across databases. The highest prevalence of continuous measurements occurred in Optum EHR

Results

Table 1. Number of unique continuous measurements and their respective prevalence in each database

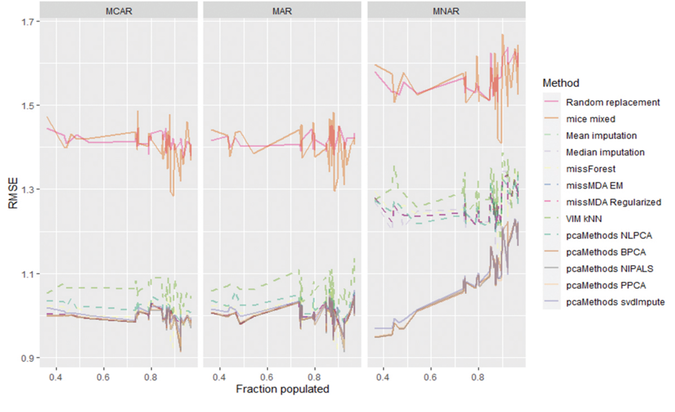
Prevalence	CCAE (N=49,108)	Optum EHR (N=61,345)	MDCR (N=30,189)	JMDC (N=7,181)	Optum DOD (N=261,824)	Optum SES (N=234,919)
≥0	885	1023	478	23	9012	8742
≥10	0	89	0	22	75	71
≥20	0	58	0	20	49	46
≥30	0	48	0	19	41	35
≥50	0	41	0	14	0	0
≥80	0	9	0	0	0	0

Figure 1. Average computation time (minutes) of imputation methods



- Mean, median and k-nearest neighbor imputation were associated with computation times 25 times faster than PCA imputation

Figure 2. Root mean squared error (RMSE) of imputation methods based on fraction of populated values for a given continuous measurement concept under varied missingness assumptions



- A higher RMSE was observed with random replacement and multiple imputation (i.e., mice mixed) under all missingness assumptions
- PCA imputation was associated with the lowest RMSE
- MNAR assumption was associated with increased RMSE across all imputation methods
- In PCA imputation with MNAR assumption, RMSE was positively correlated with the fraction of populated values for a given covariate

Conclusions

The current study found the majority of observed continuous measurement may be unsuitable for imputation due to low prevalence (<30%). In fact, continuous measurements occurring with prevalence ≥50% within 365 days of index were only observed within Optum EHR and JMDC. PCA imputation was associated with longer computation times but improved performance, especially under the missingness assumption of MNAR. Additional research is necessary to explore the positive correlation between RMSE and fraction of populated values for a given covariate achieved by PCA imputation under MNAR assumptions. Furthermore, it is important to note the study population included newly diagnosed heart failure patients, and, therefore, the generalizability of study findings may be limited.