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

**Stephen P Fortin, Jenna Reps**

**Background**

In real world evidence research, the use of continuous measurements is often avoided due to incomplete recording and lack of standardization. The current study explores the degree to which continuous measurement concepts are recorded across differrent databases mapped to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) among patients newly diagnosed with heart failure. Furthermore, continuous measurement concepts with sufficient records are used to assess the performance of various imputation methods to estimate the values of missing data.

**Methods**

We identified 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 in 6 large healthcare databases mapped to the OMOP CDM (index = first diagnosis). For each database, we measured the number of unique continuous measurement concepts observed within 365 days prior to index and their respective prevalence. To investigate imputation strategies, we focused on the Optum® de-identified Electronic Health Record Dataset (Optum EHR), which is known to contain more continuous measurements as compared to administrative claims databases. Continuous measurement values were mapped to a uniform scale for each measurement concept with a prevalence ≥30% in the year prior to index. We then used the missCompare R package to compare the performance of 13 imputation methods (e.g., mean, median, multiple imputation, random forest, k-nearest neighbor and principal component analysis [PCA] imputation) in terms of compute time and root-mean squared error (RMSE) on the aforementioned uniformly scaled continuous measurement concepts[2]. Specifically, we measured patient age, sex and clinical characteristics (i.e., components of the Charlson comorbidity index). The missCompare package was used to simulate these data in addition to each continuous measurement concept, but without missing values, while preserving correlations between covariates; and 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.

**Results**

A total of 3,166 to 261,824 patients meeting the study criteria were identified across each database, including 61,345 patients in Optum EHR. The number of unique continuous measurement concepts and their respective prevalence in each database is summarized in Table 1. The prevalence of of continuous measurement concepts was highest in Optum EHR.

| | CCAE | Optum EHR | MDCR | JMDC | Optum DOD | Optum SES |
|---|---|---|---|---|---|---|
| Cohort patient count (N) | 49108 | 61345 | 30189 | 7181 | 261824 | 234919 |
| Number of unique continuous measurement concepts observed in prior 365 days | 885 | 1023 | 478 | 23 | 9012 | 8742 |
| with prevalence ≥1% | 47 | 190 | 49 | 23 | 250 | 239 |
| with prevalence ≥5% | 0 | 121 | 0 | 22 | 109 | 108 |
| with prevalence ≥10% | 0 | 89 | 0 | 22 | 75 | 71 |
| with prevalence ≥20% | 0 | 58 | 0 | 20 | 49 | 46 |
| with prevalence ≥30% | 0 | 48 | 0 | 19 | 41 | 35 |
| with prevalence ≥40% | 0 | 45 | 0 | 17 | 8 | 5 |
| with prevalence ≥50% | 0 | 41 | 0 | 14 | 0 | 0 |
| with prevalence ≥60% | 0 | 40 | 0 | 8 | 0 | 0 |
| with prevalence ≥70% | 0 | 32 | 0 | 0 | 0 | 0 |
| with prevalence ≥80% | 0 | 9 | 0 | 0 | 0 | 0 |
| with prevalence ≥90% | 0 | 0 | 0 | 0 | 0 | 0 |

**Table 1.** Number of continuous measurement concepts and their respective prevalence within 6 healthcare databases

CCAE: IBM MarketScan® Commercial Database; Optum EHR: Optum® de-identified Electronic Health Record Dataset; MDCR: IBM MarketScan® Medicare Supplemental Database; JMDC: Japan Medical Data Center; Optum DOD: Optum® De-Identified Clinformatics Data Mart Database – Date of Death; Optum SES: Optum® De-Identified Clinformatics Data Mart Database – Socioeconomic Status

The average computation time associated with each imputation method is shown in Figure 1. Mean, median and k-nearest neighbor imputation were associated with computation times over 25 times faster than PCA imputation.

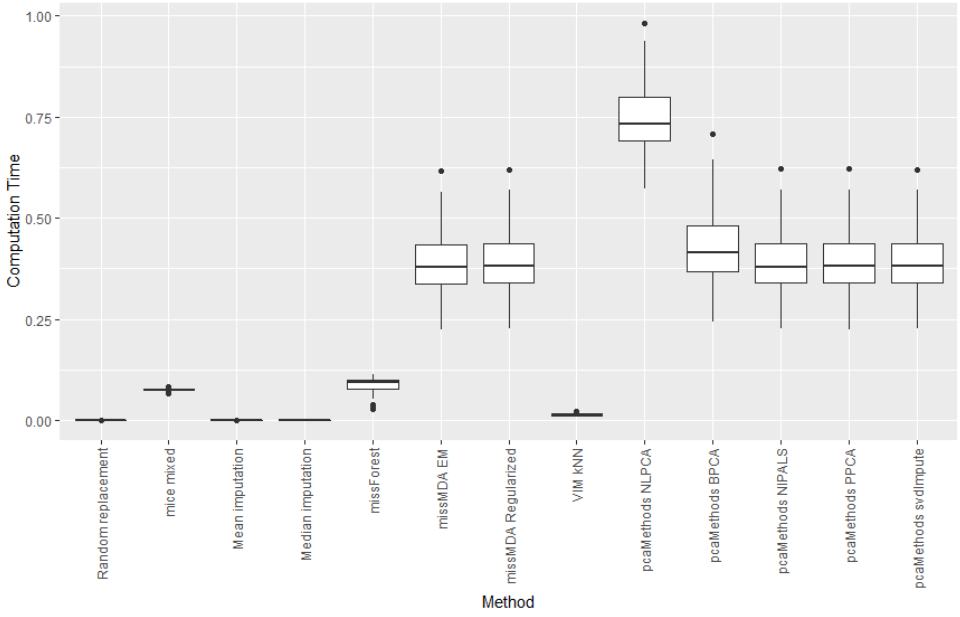


**Figure 1.** Average computation time associated with imputation methods tested in the missCompare R package

The RMSE associated with imputation methods under each missingness assumption is shown in Figure 2. Higher RMSE was observed with random replacement and multiple imputation (i.e., mice mixed) under all missingness assumptions. Conversely, a slight reduction in RMSE was observed with PCA imputation. Overall, MNAR was associated with increased RMSE across all imputation methods with one notable exception: PCA imputation was associated with lower RMSE among covariates with a lower fraction of populated values. In PCA imputation in MNAR, RMSE was positively correlated with the fraction of populated values for a given covariate.

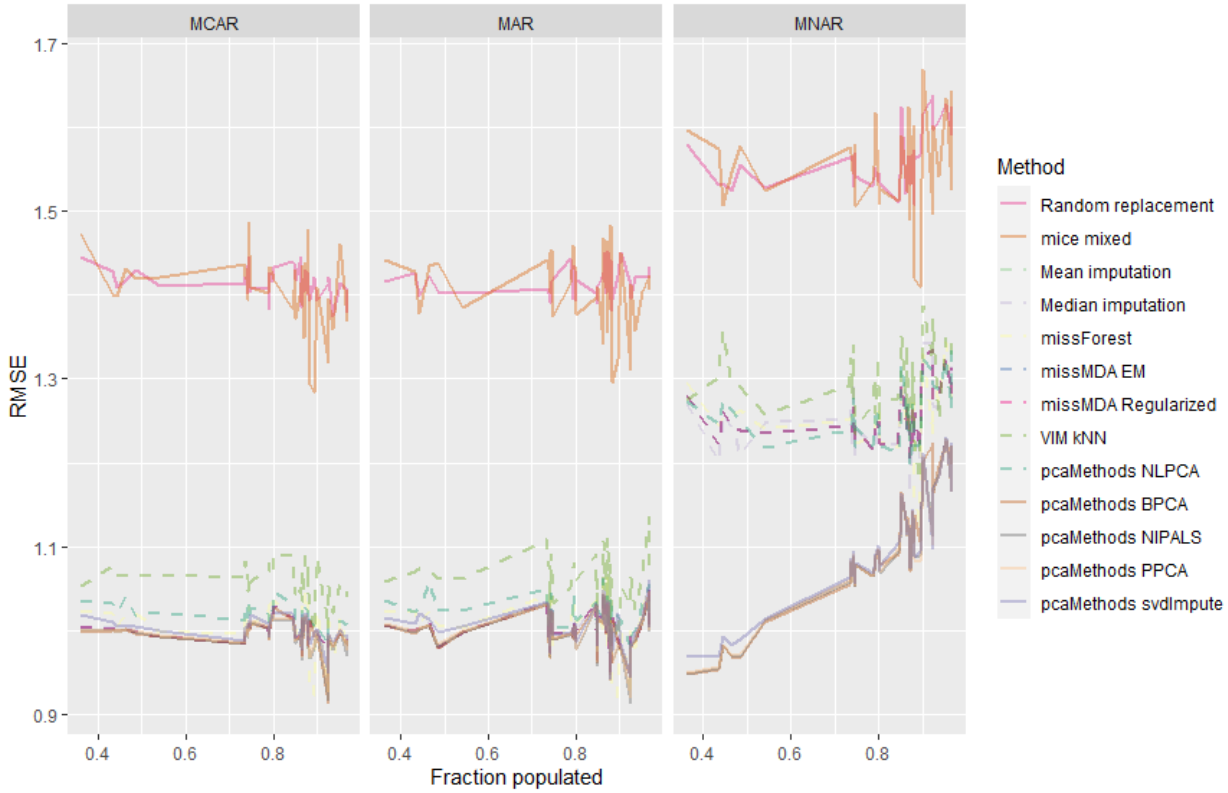


**Figure 2.** Root mean squared error (RMSE) of imputation methods based on fraction of populated values for a given continuous measurement concept under varied missingness assumptions

## Conclusion

The current study found the majority of observed continuous measurement concepts may be unsuitable for imputation due to low prevalence (<30%) within the data. In fact, continuous measurement concepts 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.

**References**


1. Kang H. The prevention and handling of the missing data. *Korean J Anesthesiol*. 2013;64(5):402-406. doi:10.4097/kjae.2013.64.5.402
2. Varga TV, Westergaard D. missCompare: Intuitive Missing Data Imputation Framework. R package version 1.0.3. 2020. https://CRAN.R-project.org/package=missCompare