

Hit and Miss

**An evaluation of imputation
techniques from Machine Learning**

Motivation

- Approval of a drug combination that patients were already taking as separate tablets.
- Was a separate trial necessary?
 - Pool evidence from two Randomized Clinical Trials and 2 registry studies

Problem statement

- *Registry studies typically have many missing assessments, this is particularly of concern on baseline data which is needed for pooling*
 - Is machine learning suited for this task?
 - Can we provide a robust imputation strategy?
 - How can we assess our predictive performance?

What's wrong with the current methods?

- Deletion
 - Loss of power, potential bias of result
- Imputation
 - “Simple” Imputation
 - mean/median
 - worst observation (WOOF), last observation (LOCF)

Current approach: Drop imputation

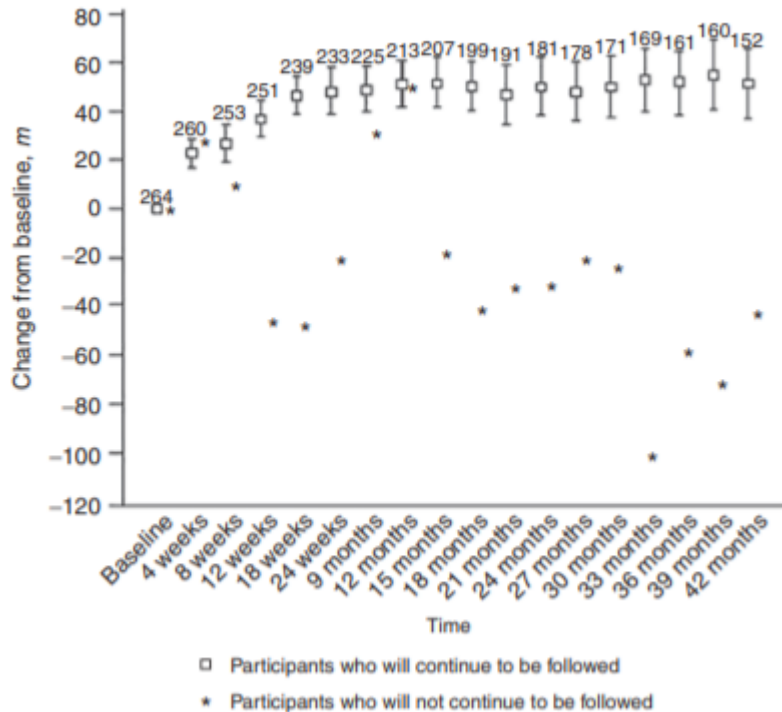
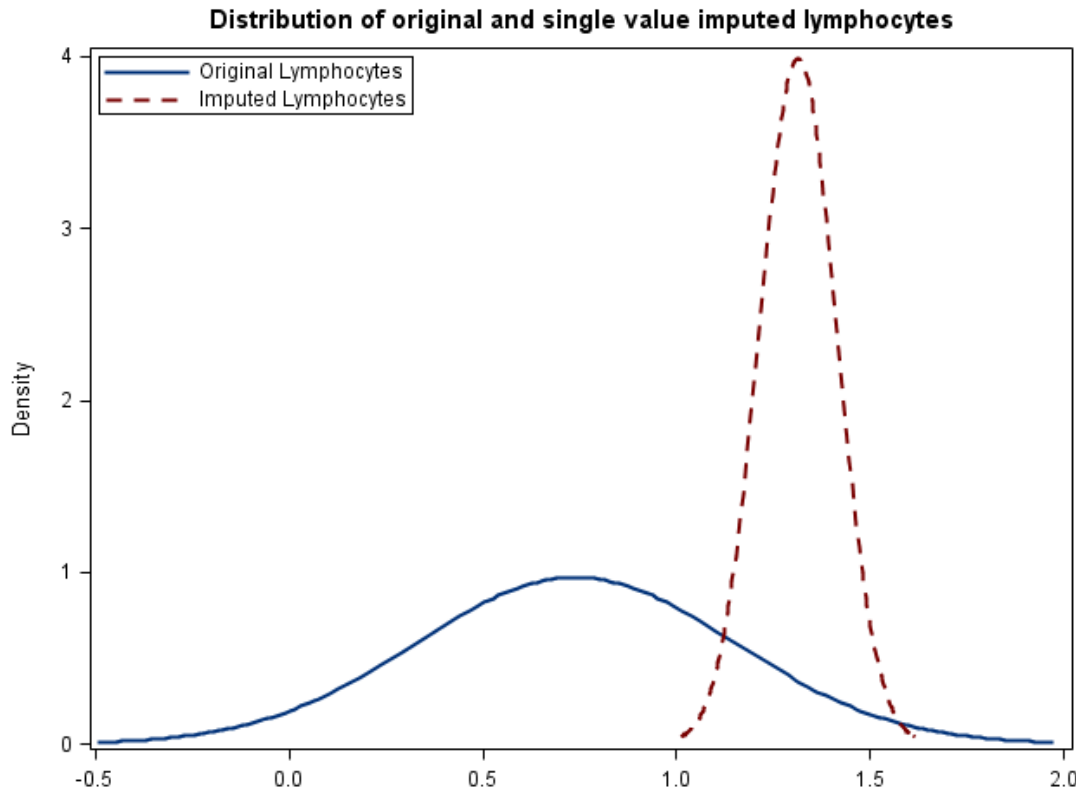


Figure 1 Change from baseline in 6-min walk distance over 48 months after initiation of treatment in patients with pulmonary arterial hypertension. At each scheduled visit, the average change from baseline in 6-min walk distance is plotted separately for the subgroups that will and will not remain under follow-up at the time of the next scheduled visit. The numbers above the squares represent, at each scheduled visit, the number of patients in the subgroup who will remain under follow-up at the time of the next scheduled visit. Hence, these are the numbers of patients whose data contribute to the calculations of the mean values and the 95% confidence intervals for that subgroup. Figure extracted from ref. 7.

Source: *The Prevention and Treatment of Missing Data in Clinical Trials: An FDA Perspective on the Importance of Dealing With It* (O'Neill, Temple 2012)

Current approach (single imputation)



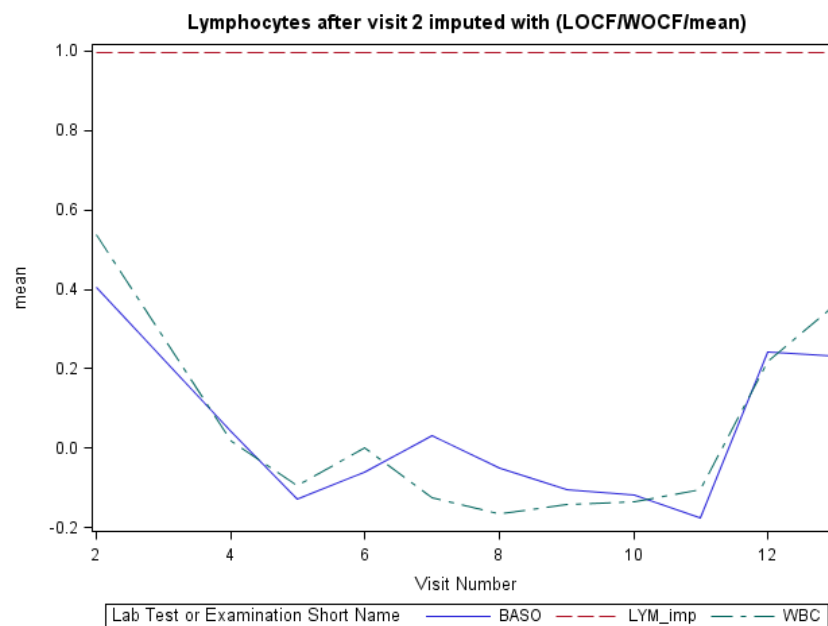
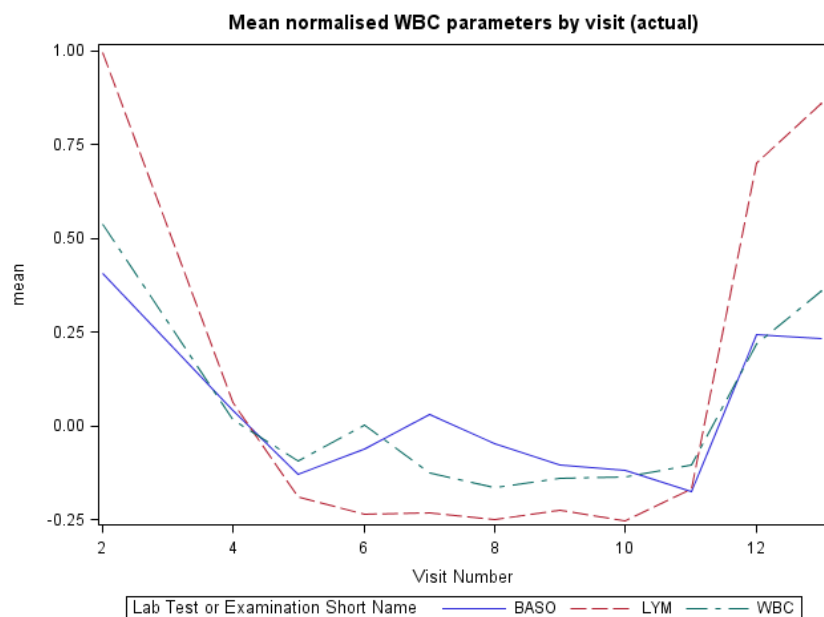
Single imputation (Mean/Median/WOCF/LOCF) assumes greater information is known than is available at the time of analysis due to the imputed values being assumed as known. **This can lead to narrow confidence intervals and biased p-values**

in a way that can be understood by clinicians.

-
- 10 Single imputation methods like last observation carried forward and baseline observation carried forward should not be used as the primary approach to the treatment of missing data unless the assumptions that underlie them are scientifically justified.
-

Current approach (correlation)

- Effect of single imputation not using correlation



Types of missingness

- Missing Completely At Random (MCAR)
 - The reason for data being missing does not depend on the observed or the unobserved missing data
- Missing At Random (MAR)
 - The reason for data being missing may depend on observed data (trajectory), but not on the unobserved missing data
- Missing Not At Random (MNAR)
 - The reason for data being missing depends on the unobserved missing data (or on observed data not taken into account in the model)

Missingness Intuition (images)

Cool Dog (Original)



Missingness Intuition (MCAR)

10% missing



20% missing



50% missing



Missingness Intuition (MCAR vs MAR)

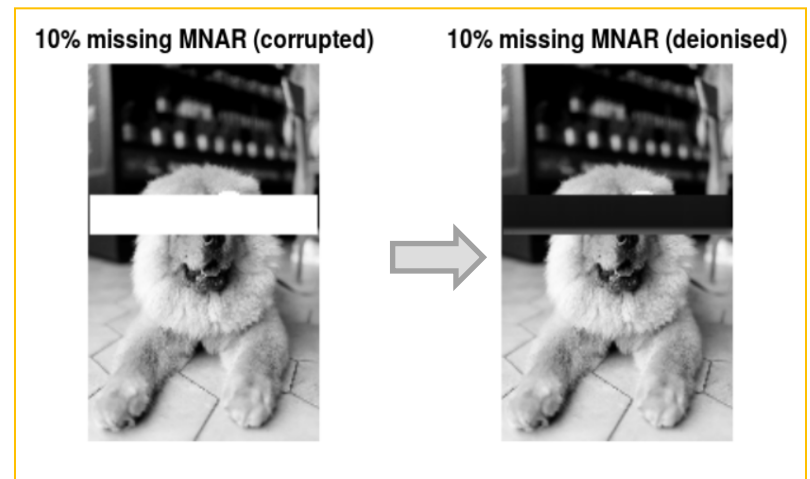
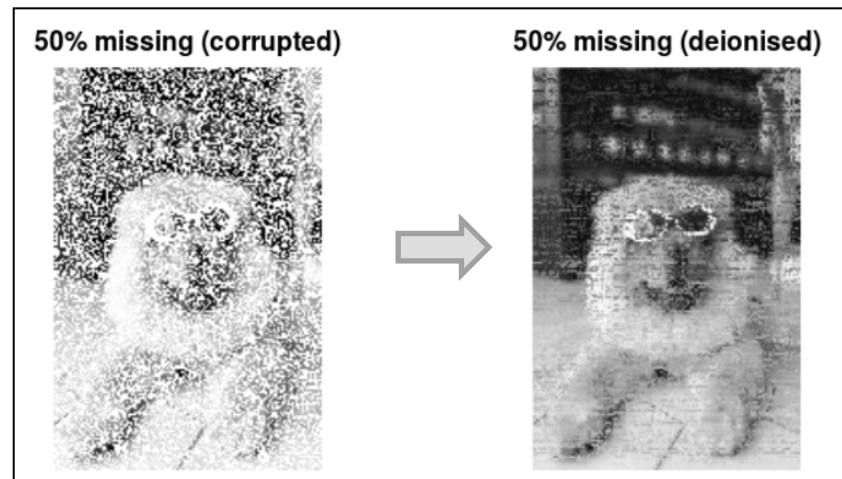
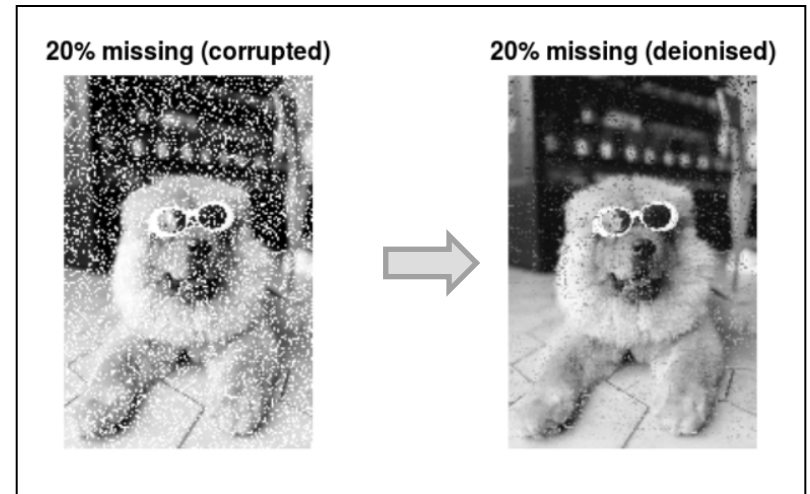
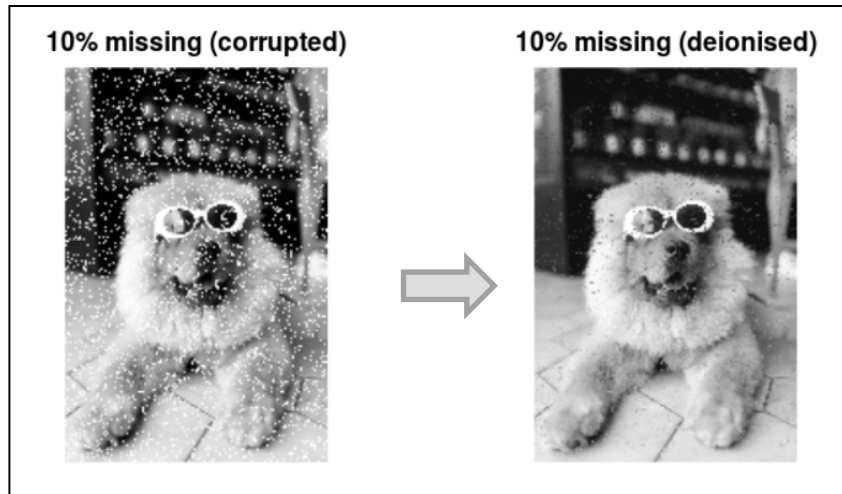
10% missing MCAR



10% missing MNAR



DAE Reconstruction



MNAR image completion

- Globally and Locally Consistent Image Completion (Iizuka et al, 2017)



Error Metrics

- How good is our prediction?
 - Normalised Root Means Square for continuous data (lower is better)
 - Proportion Falsely Classified for categorical data (lower is better)

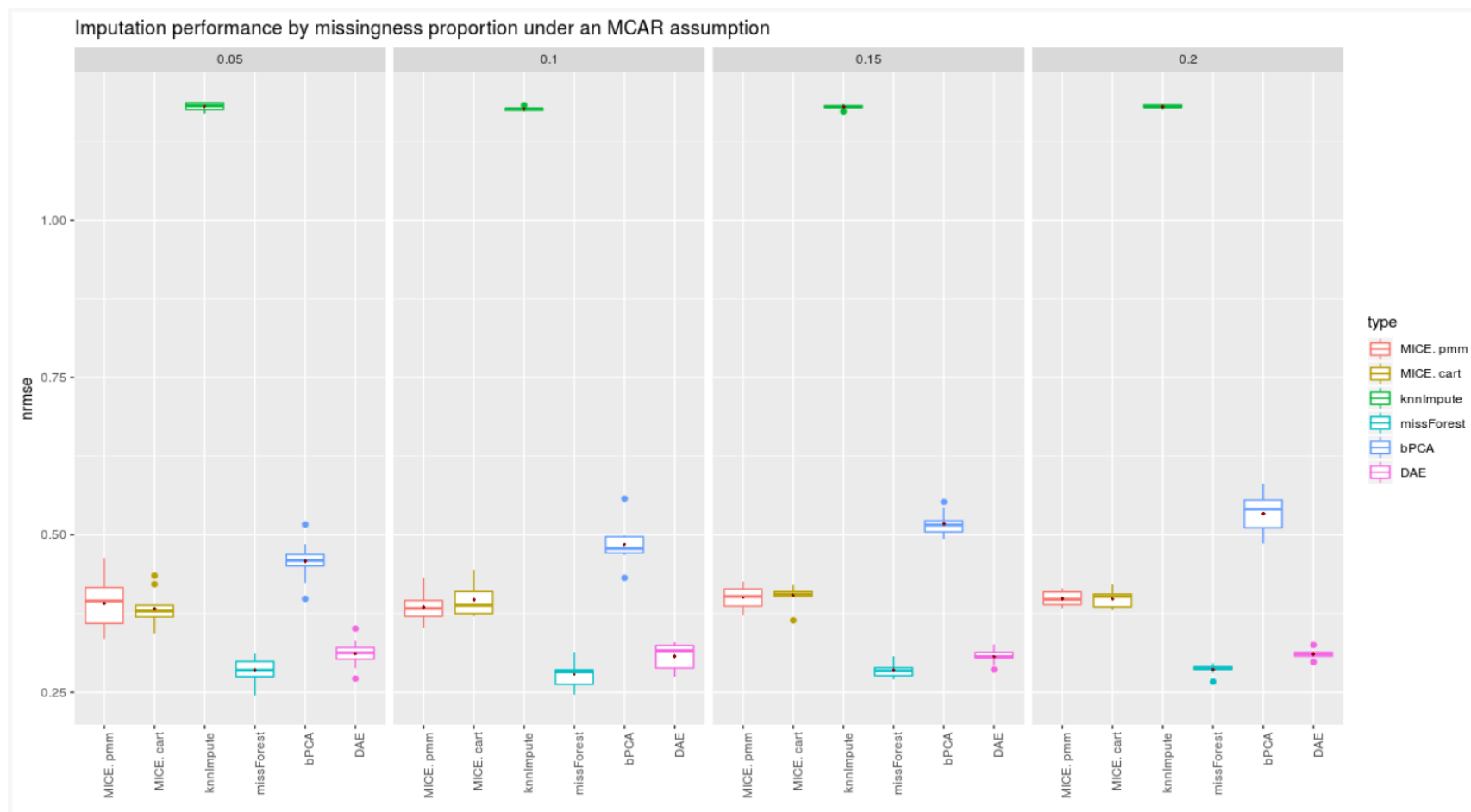
$$NRMSE = \sqrt{\frac{\text{mean}((\text{observed} - \text{imputed})^2)}{\text{var}(\text{observed})}}$$

$$PFC = \sqrt{\frac{\sum(\text{observed} \neq \text{imputed})}{N}}$$

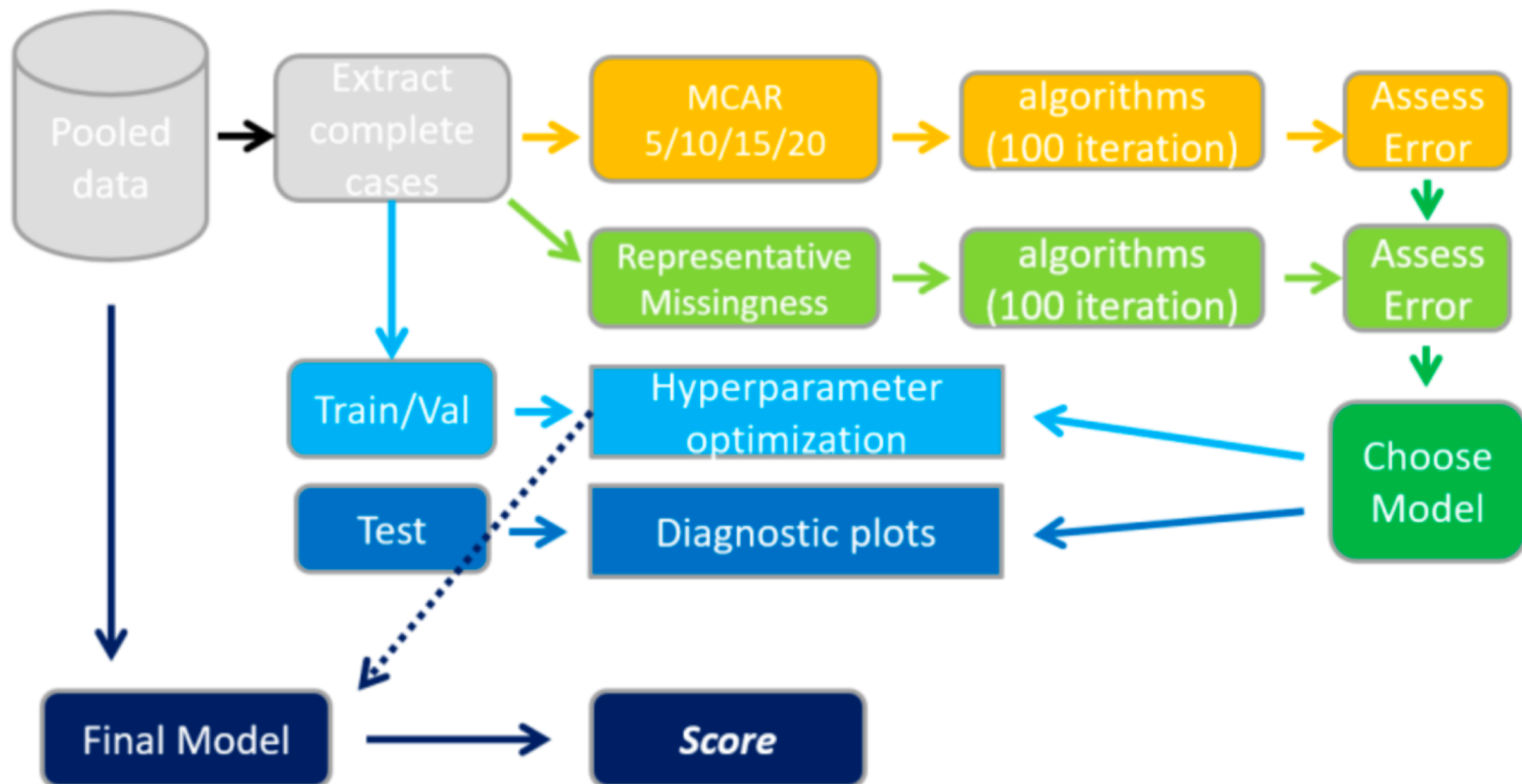
Introduction to Algorithms

- 6 algorithms assessed
 - Multiple Imputation Chained Equations
 - missForest - Random Forest
 - Classification and Regression Trees
 - kNN / missXGB
 - Denoising Autoencoder
 - Bayesian Principal Component Analysis

Performance under MCAR assumptions

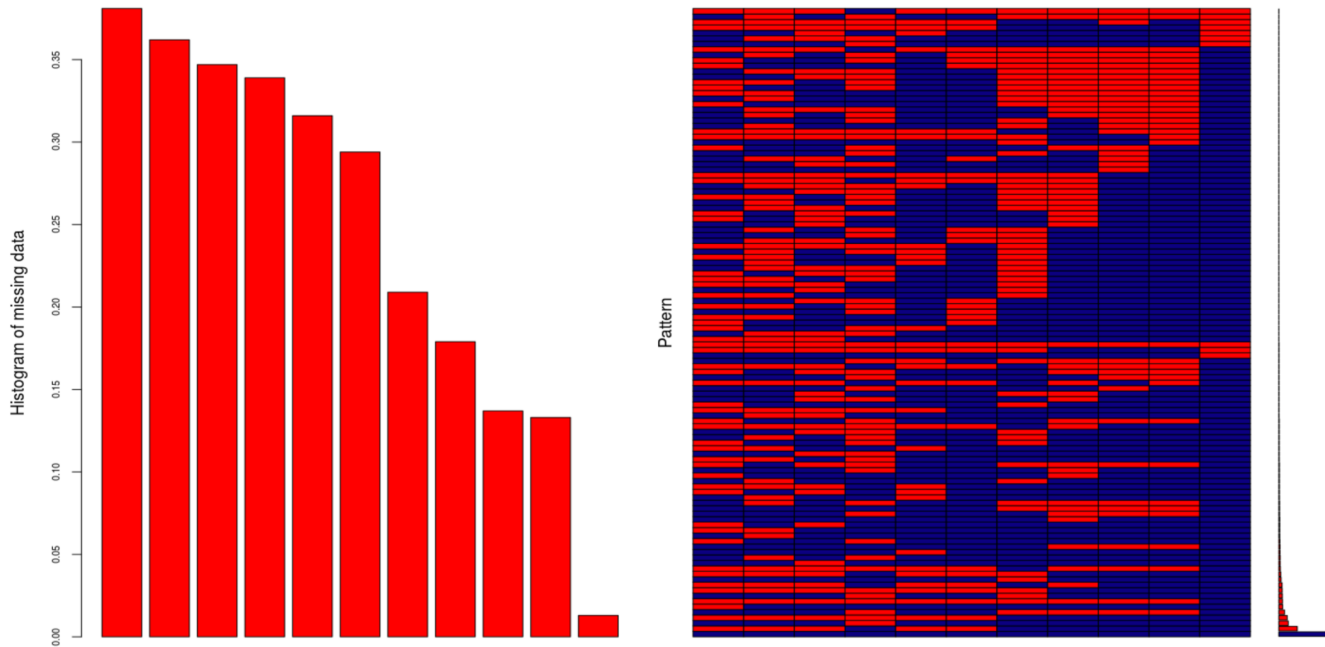


Machine Learning Pipeline for Multiple Imputation methods

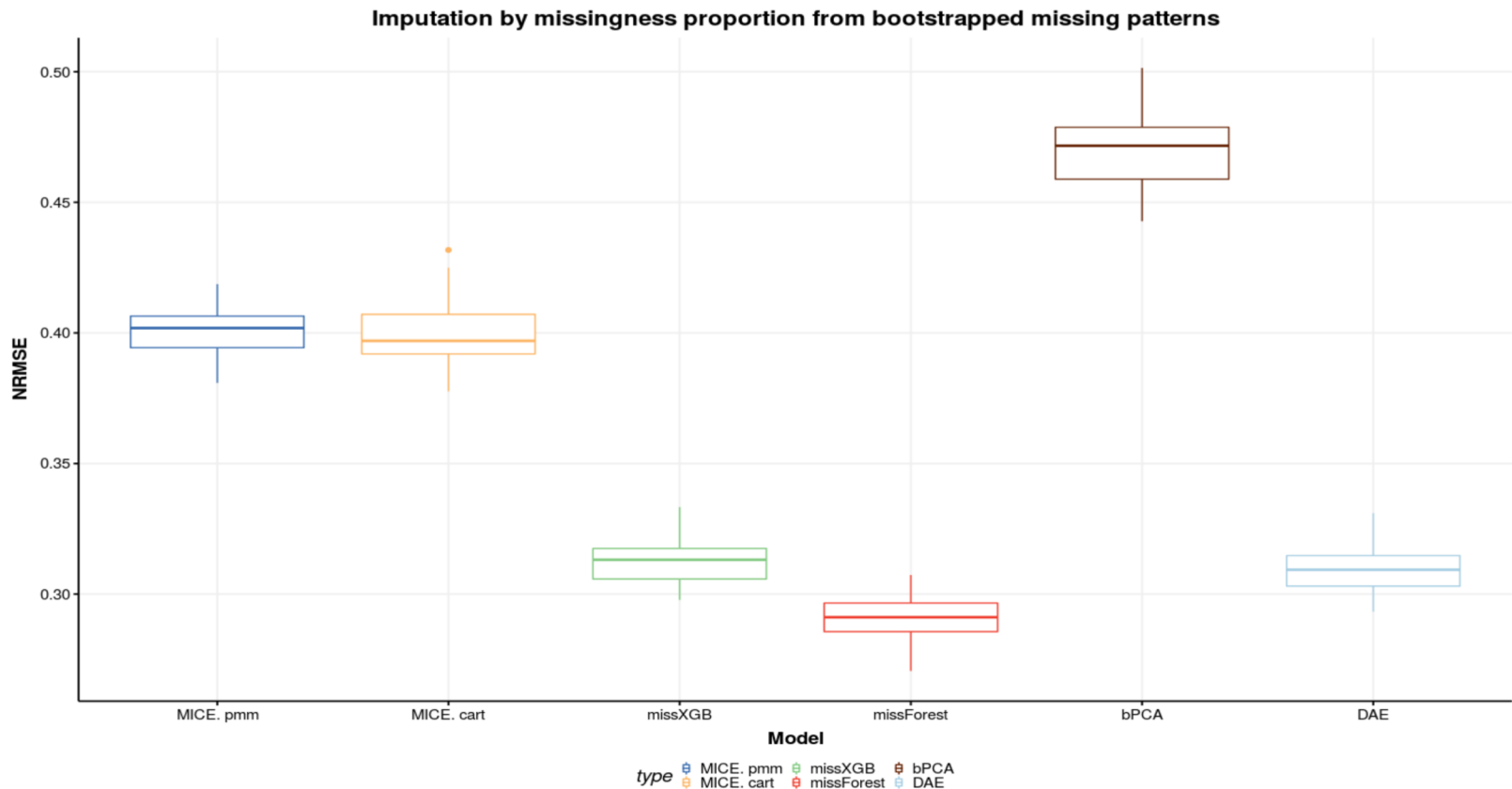


Missingness Pattern of baseline parameters

- Can we create a representative train/test set?



Performance under representative data

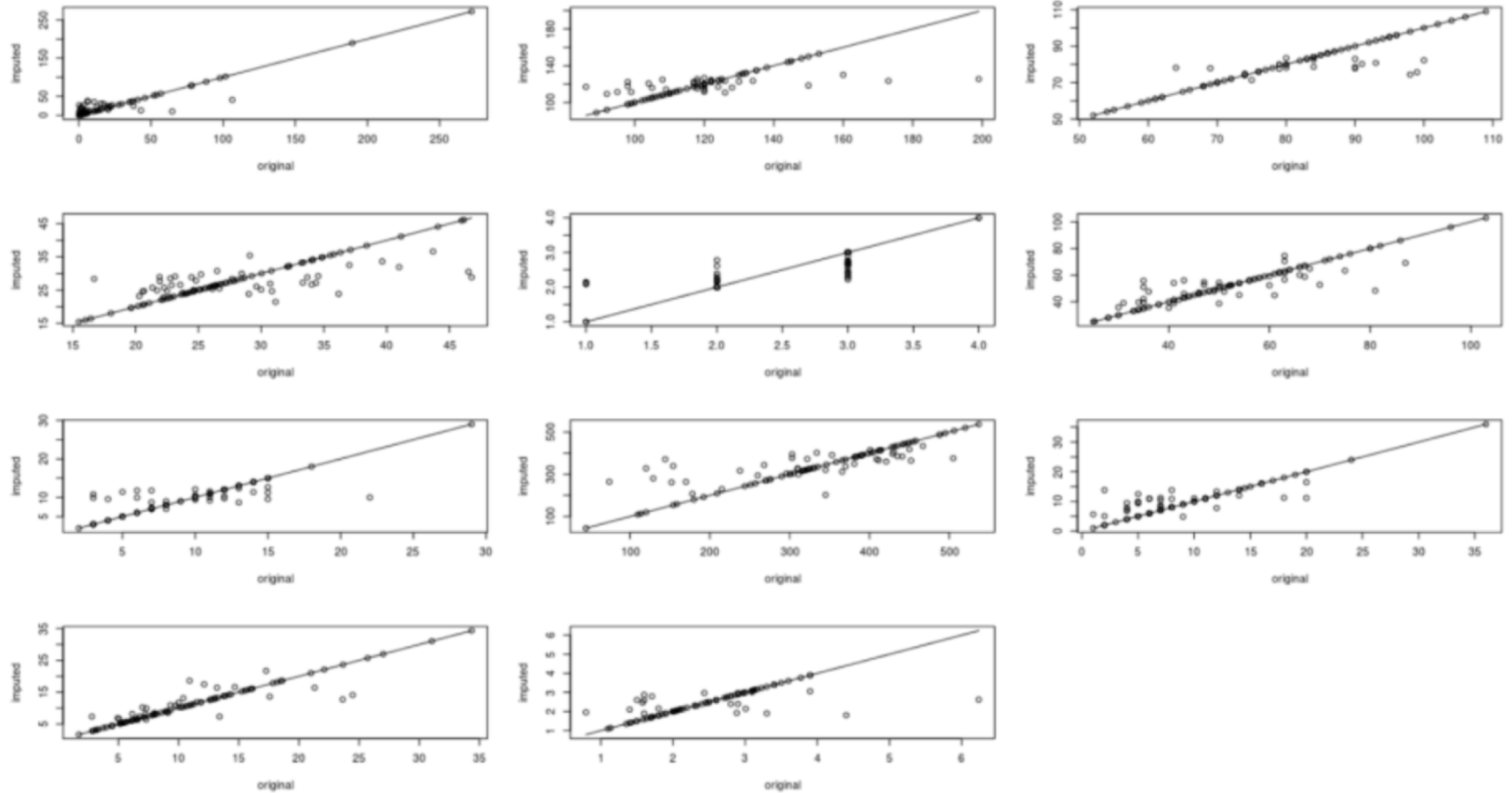


Variable level performance

Mean (95% CI) NRMSE of bootstrap sampled imputation

MICE, pmm	1.10±0.01	1.25±0.02	1.39±0.01	1.38±0.01	1.25±0.01	1.38±0.01	1.34±0.01	1.45±0.04	1.22±0.01	1.21±0.01	1.28±0.01	1.30±0.01
MICE, cart	1.12±0.01	1.24±0.02	1.35±0.01	1.37±0.01	1.33±0.01	1.38±0.01	1.33±0.01	1.42±0.04	1.29±0.01	1.20±0.01	1.30±0.01	1.32±0.01
missXGB	0.84±0.01	0.94±0.01	1.02±0.00	1.07±0.01	1.00±0.00	1.08±0.00	1.03±0.01	1.05±0.01	0.95±0.00	0.95±0.00	0.99±0.01	1.04±0.00
missForest	0.79±0.01	0.88±0.00	0.96±0.00	0.99±0.00	0.91±0.00	0.99±0.00	0.95±0.00	0.96±0.01	0.89±0.00	0.88±0.00	0.93±0.01	0.96±0.00
bPCA	0.94±0.00	0.97±0.00	1.00±0.00	1.00±0.00	0.99±0.00	0.99±0.00	0.99±0.00	1.11±0.01	1.00±0.00	1.46±0.01	0.97±0.00	0.98±0.00
DAE	0.84±0.00	0.93±0.00	0.99±0.00	1.02±0.00	0.97±0.00	1.02±0.00	0.98±0.00	1.02±0.01	0.96±0.00	0.94±0.00	0.97±0.01	0.98±0.00

Results: Imputation



Results: On the final data

- We can estimate the error of the entire dataset based on the Out-Of-Bag error (OOB)
- Final model NRMSE: 0.27
 - Comparable from what was observed during testing

MissXGB

Algorithm 1 Impute missing values with missXGB

Require: X an $n \times p$ matrix, stopping criterion γ

- 1: Impute missing to respective column means ;
 - 2: $k \leftarrow$ vector of sorted indices of column X
w.r.t increasing amount of missing values ;
 - 3: **while** not γ **do**
 - 4: $X_{old} \leftarrow$ store previously imputed matrix ;
 - 5: **for** s in k **do**
 - 6: Fit XGBoost: $y_{obs} \sim x_{obs}$;
 - 7: Predict y_{miss} using x_{miss} ;
 - 8: $X \leftarrow$ update imputed matrix from y_{miss} ;
 - 9: **end for**
 - 10: update γ ;
 - 11: **end while**
 - 12: **return** Matrix X ;
-

MissXGB vs MissForest

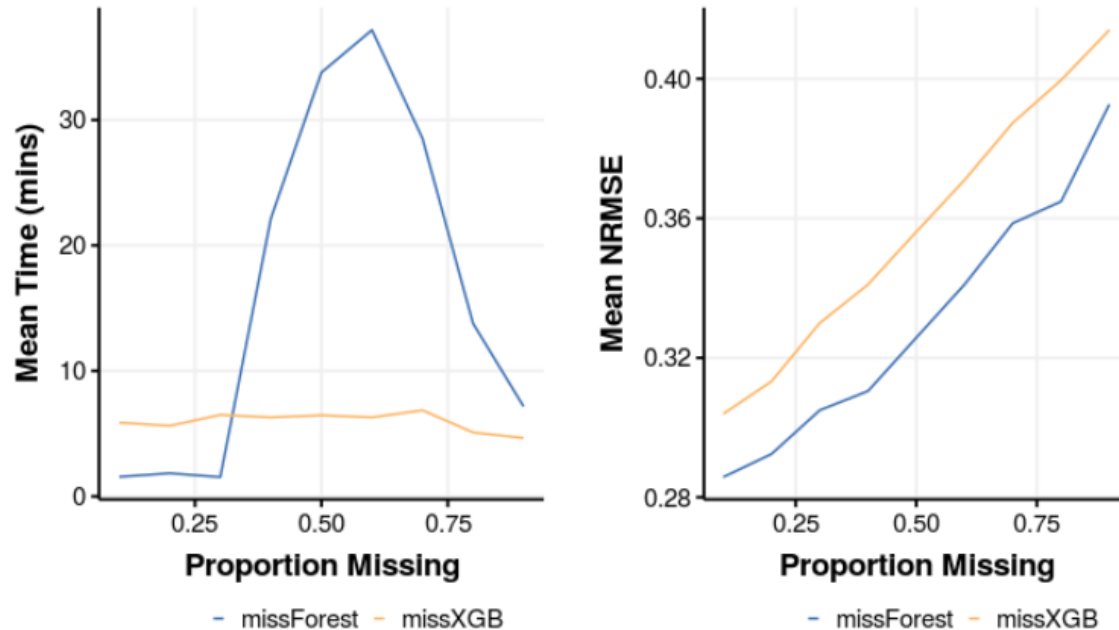


Figure 4: Comparison in speed and accuracy between missForest and missXGB

Questions

References

- ISPOR Special Interest Group: Statistical Methods in HEOR (Forum Presentation | ISPOR 2018 May 21, 2018 | Baltimore, MD, USA)
(https://www.ispor.org/docs/default-source/presentations/1455.pdf?sfvrsn=40350bca_1)
- The Prevention and Treatment of Missing Data in Clinical Trials: An FDA Perspective on the Importance of Dealing With It
(<https://website.aub.edu.lb/sharp/Publications/RCT-missingdata2.pdf>)
- Methods
 - MIDA (<https://arxiv.org/pdf/1705.02737.pdf>)
 - Bayesian PCA (<http://ishiilab.jp/member/oba/tools/BPCAFill.html>)
 - missForest (<https://academic.oup.com/bioinformatics/article/28/1/112/219101>)
 - MICE (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/>)
 - KNN (<http://www.bioconductor.org/packages/release/bioc/manuals/impute/man/impute.pdf>)
 - CART(http://civil.colorado.edu/~balajir/CVEN6833/lectures/cluster_lecture-2.pdf)