Benchmarking missing-values approaches for predictive models on health databases.

CIMD presentation

Alexandre Perez-Lebel [1,2] Gaël Varoquaux [1,2] Marine Le Morvan [2]
Julie Josse [2] Jean-Baptiste Poline [1]

[1]McGill University, Canada

[2]Inria, France

October 11, 2021

Content

1. Overview
2. Introduction
3. Benchmark
   - Methods benchmarked
   - Datasets
   - Protocol
4. Results
   - Prediction performance
   - Computational time
   - Significance
5. Discussion
   - Interpretation
   - Limitations
   - Conclusion

Overview

- Benchmark on real-world data.
- 4 health databases with missing values.
- Arbitrarily defined prediction tasks.
- Methods to handle missing values:
  Missing Incorporated in Attribute (MIA) vs imputation.
- Evaluate prediction score and computational time.

Overview

Results:

- MIA performed better at little cost, but not always significantly.
- Conditional imputation is on a par with constant imputation.
- Complex imputation can be untractable at large scale.

Introduction

Introduction: scope of the study

Focus on supervised learning with missing values.
Different tradeoffs: risk minimization instead of parameters estimation.

In supervised learning, most statistical models and machine learning algorithms are not designed for incomplete data.

How to deal with missing values in this framework?

- Delete samples having missing values → to avoid.
- Use imputation.
  - Constant imputation (mean, median)
  - Conditional imputation (KNN, MICE)
- Adapt or create predictive models to handle missing values natively.
  - Boosted-trees with *Missing Incorporated in Attribute* (MIA) adaptation [Twala et al., 2008].
  - NeuMiss networks in the regression setting [Le Morvan et al., 2020].

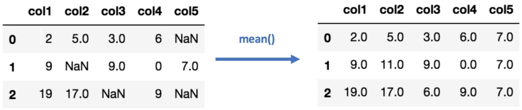
Introduction: problem

- How does MIA experimentally compare to imputation?
- Constant imputation vs conditional imputation.

Imputation

Replace missing values by values.

- Constant imputation:
  Mean or Median.



	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

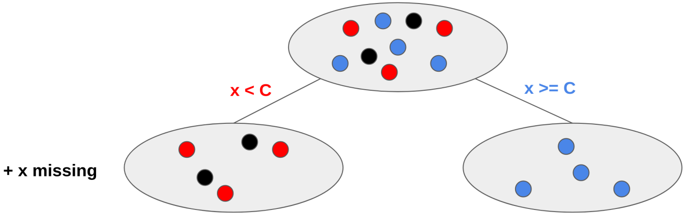
- Conditional imputation:
  $X_{mis} \leftarrow \mathbb{E}\left[X_{mis} \mid X_{obs}\right]$.
  MICE [Buuren and Groothuis-Oudshoorn, 2010] or KNN.

Add a binary mask to keep track of imputed values.

Missing Incorporated in Attribute (MIA)

Adaptation of boosted-trees to account for missing values.

Idea:
For each split on a variable, all samples with a missing value in this variable are either sent to the left or to the right child node depending on which option leads to the lowest risk.



Methods benchmarked

Table: **Methods compared in the main experiment.**

In-article name	Imputer	Mask	Predictive model
MIA	-	-	Gradient-boosted trees
Mean	Mean	No	Gradient-boosted trees
Mean+mask	Mean	Yes	Gradient-boosted trees
Median	Median	No	Gradient-boosted trees
Median+mask	Median	Yes	Gradient-boosted trees
Iterative	MICE	No	Gradient-boosted trees
Iterative+mask	MICE	Yes	Gradient-boosted trees
KNN	KNN	No	Gradient-boosted trees
KNN+mask	KNN	Yes	Gradient-boosted trees

Datasets

- Traumabase [The Traumabase Group,], 20 000 samples.
- UK BioBank [Sudlow et al.,], 500 000 samples.
- MIMIC-III [Johnson et al.,], 60 000 samples.
- NHIS [National Center for Health Statistics, 2017], 88 000 samples.

Defined 13 prediction tasks (10 classifications, 3 regressions).
Outcomes chosen arbitrarily.
Feature selection:

- ANOVA (11 tasks)
- Expert knowledge (2 tasks).

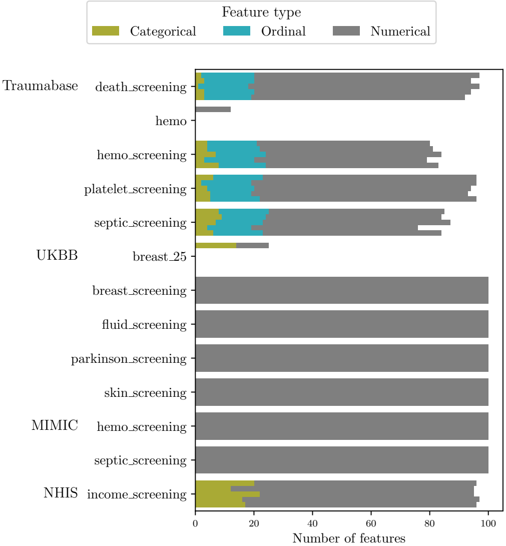


Figure: **Types of features** *before encoding*.

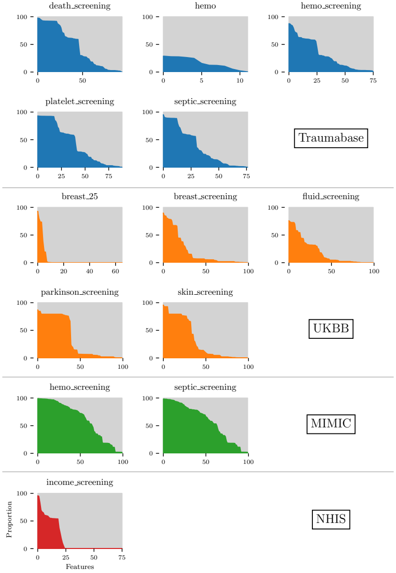


Figure: **Missing values distribution.**

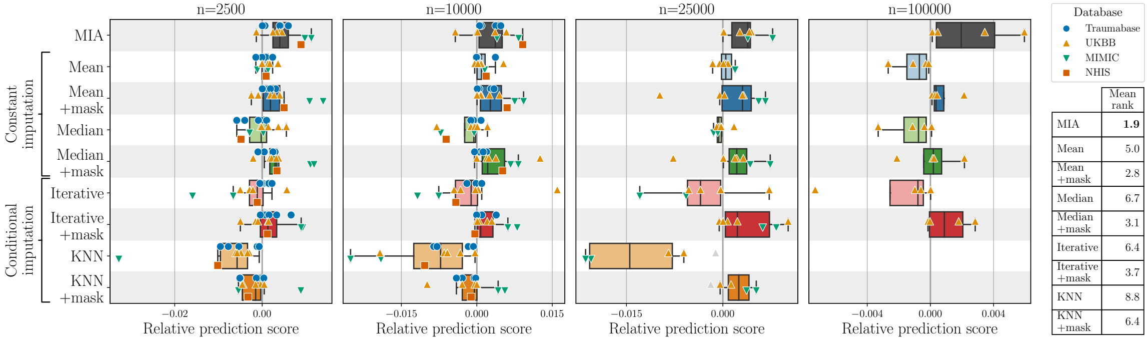
Table: **Correlation between features.**

			Threshold		
			0.1	0.2	0.3
Database	Task	# features			
Traumabase	death_screening	92	68%	41%	22%
	hemo	12	50%	23%	12%
	hemo_screening	76	65%	36%	20%
	platelet_screening	90	67%	40%	22%
	septic_screening	76	68%	37%	18%
UKBB	breast_25	11	40%	20%	19%
	breast_screening	100	26%	12%	8%
	fluid_screening	100	21%	10%	6%
	parkinson_screening	100	28%	16%	11%
	skin_screening	100	24%	11%	8%
MIMIC	hemo_screening	100	22%	6%	3%
	septic_screening	100	21%	6%	2%
NHIS	income_screening	78	15%	6%	4%
Average		79	40%	20%	12%

Experimental protocol

- 9 methods, 13 prediction tasks.
- One-hot encode categorical features.
- Feature selection trained on $1/3$ of the samples: 5 trials, 100 features.
- Sub-sampled the tasks: 2 500, 10 000, 25 000 and 100 000 samples.
- Cross-validation.
- Tuned the hyper-parameters of the predictive model.
- Evaluate prediction with accuracy or r2.

Results - Prediction performance



	Mean rank
MIA	**1.9**
Mean	5.0
Mean +mask	2.8
Median	6.7
Median +mask	3.1
Iterative	6.4
Iterative +mask	3.7
KNN	8.8
KNN +mask	6.4

Figure: **Prediction performance.**

For a specific task:

Relative prediction score = prediction score - average prediction score of the 9 methods.

Iterative = MICE

Results - Computational time

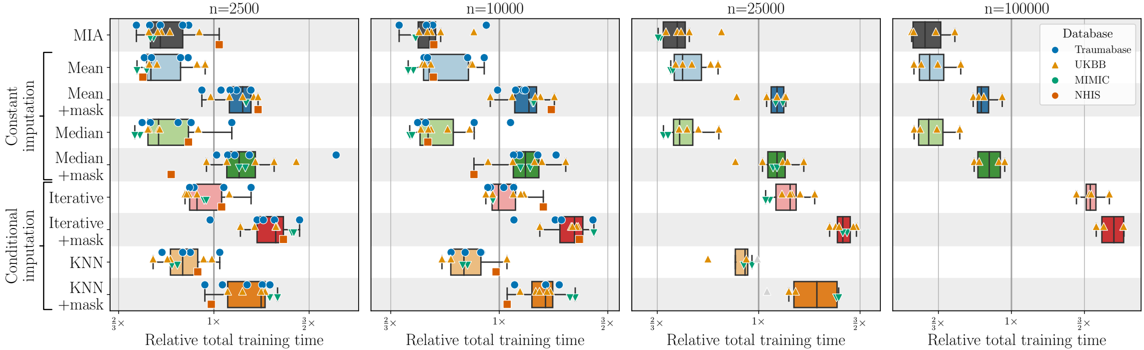


Figure: **Computational time.**

For a specific task:

Relative prediction time = prediction time - average prediction time of the 9 methods.

Results - Significance

Friedman test. [Friedman, 1937]
Null hypothesis: “methods are equivalent”.

	p-value
Size	
2500	1.6e-10
10000	2.6e-10
25000	2.8e-04
100000	8.5e-03

Nemenyi test. [Nemenyi, 1963]
Once the Friedman test is rejected, the Nemenyi test can be applied. It provides a critical difference CD which is the minimal difference between the average ranks of two algorithms for them to be significantly different. (N: number of datasets, k: number of algorithms)

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

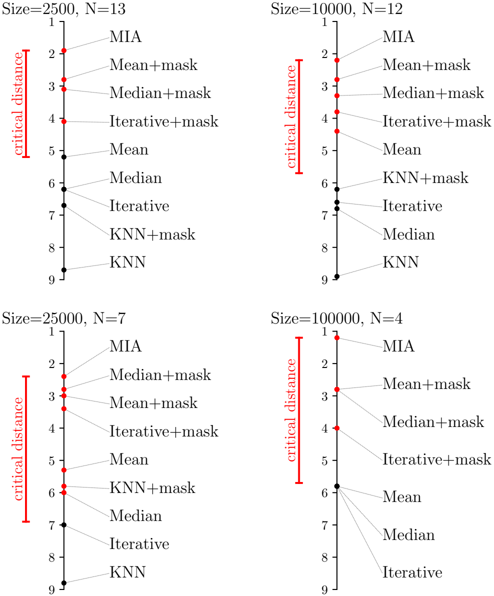


Figure: **Mean ranks by method and by size of dataset.**

Results - Significance

Same conclusions with the Wilcoxon one-sided signed rank test. [Wilcoxon, 1945]

Findings and interpretation

Findings:

- MIA takes the lead at little cost, although not significantly.
- Adding the mask improves prediction.

Interpretation:

- Good imputation does not imply good prediction.
  - Low correlation between features.
  - Strong non-linear mechanisms.
  - Constant imputation provides a simple structure that can be extracted by the learner.
- The missingness is informative (MNAR or outcome depends on missingness) $\rightarrow$ imputation is not applicable.

Strengths and limitations

Limitations:

- Not every difference is significant.
- Would benefit having more datasets and having more datasets with large number of samples.

Strengths of the benchmark:

- 12 000 CPU hours.
- Lots of datasets (only 6% of empirical NeurIPS articles build upon more than 10 datasets [Bouthillier and Varoquaux, 2020].)
- Real data.

Conclusion

Conclusion

- Using MIA provides small but systematic improvement over imputation.
- Complex imputation untractable at large scale.
- Experiments suggests that missingness is informative: imputation not grounded.
- Directly handling missing values in the predictive model is to be considered.
- Change habits in practice: better choices than imputation.

Reviewers' feedbacks

Manuscript submitted in GigaScience.

Some comments of the reviewers:

- What about multiple imputation?
- Break boxplots by task.
- Relative prediction score difficult to interpret.

Acknowledgments

Thank you for you attention.

Appendix

Introduction: the problem of missing values

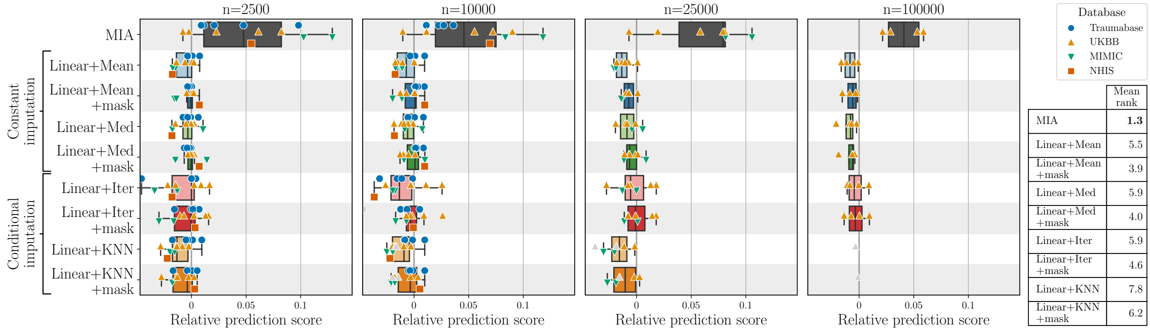
- Missing values are omnipresent in real world problems
- Have long been studied in the statistical literature within the inferential framework

[Rubin, 1976] defined several missing values mechanisms:

- *Missing At Random* (MAR): the probability of a value to be missing only depends on the observed variables.
- *Missing Not At Random* (MNAR): the missingness can depend on both the observed and unobserved values.

Most missing values methods in inference rely on the MAR hypothesis since theoretical results show that the mechanism can be ignored. In practice, real data is often MNAR.

Supplementary experiment



	Mean rank
MIA	**1.3**
Linear+Mean	5.5
Linear+Mean +mask	3.9
Linear+Med	5.9
Linear+Med +mask	4.0
Linear+Iter	5.9
Linear+Iter +mask	4.6
Linear+KNN	7.8
Linear+KNN +mask	6.2

Figure: **Prediction performance.**

Supplementary experiment

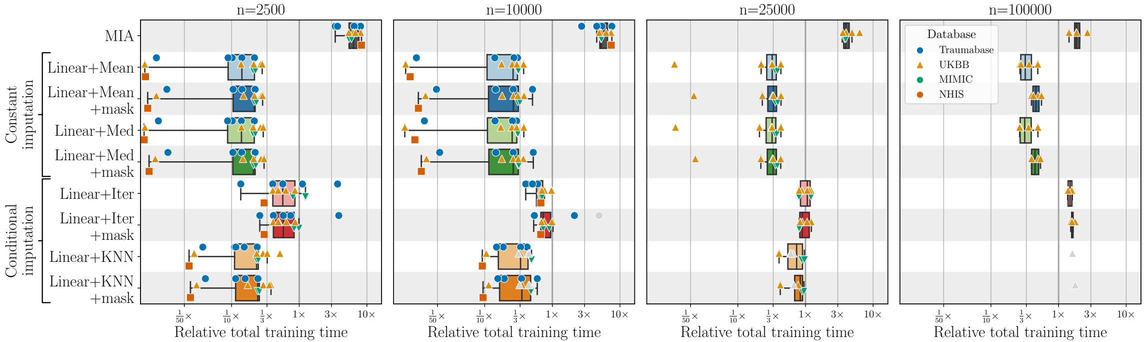


Figure: **Computational time.**

Results - Significance

Table: **Wilcoxon one-sided signed rank test.**

Size	2500	10000	25000	100000
Method				
Mean	1.2e-03**	4.6e-02*	2.3e-02*	6.2e-02
Mean+mask	4.0e-02*	2.3e-01	1.5e-01	6.2e-02
Median	5.2e-03*	1.7e-03**	2.3e-02*	6.2e-02
Median+mask	4.0e-02*	2.1e-01	1.5e-01	1.2e-01
Iterative	5.2e-03*	3.2e-02*	3.9e-02*	6.2e-02
Iterative+mask	2.4e-02*	2.1e-01	4.7e-01	6.2e-02
KNN	1.2e-04**	2.4e-04**	3.1e-02*	
KNN+mask	1.2e-04**	7.3e-04**	3.1e-02*	
Linear+Mean	6.1e-04**	4.9e-04**	7.8e-03*	6.2e-02
Linear+Mean+mask	8.5e-04**	7.3e-04**	1.6e-02*	6.2e-02
Linear+Med	6.1e-04**	4.9e-04**	7.8e-03*	6.2e-02
Linear+Med+mask	6.1e-04**	4.9e-04**	1.6e-02*	6.2e-02
Linear+Iter	3.1e-03**	1.2e-03**	1.6e-02*	6.2e-02
Linear+Iter+mask	2.3e-03**	1.2e-03**	1.6e-02*	6.2e-02
Linear+KNN	1.2e-04**	2.4e-04**	1.6e-02*	5.0e-01
Linear+KNN+mask	1.2e-04**	2.4e-04**	3.1e-02*	5.0e-01

References I

Bouthillier, X. and Varoquaux, G. (2020).
Survey of machine-learning experimental methods at NeurIPS2019 and ICLR2020.
Research report, Inria Saclay Ile de France.

Buuren, S. v. and Groothuis-Oudshoorn, K. (2010).
mice: Multivariate imputation by chained equations in r.
Journal of statistical software, pages 1–68.

Friedman, M. (1937).
The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance.
Journal of the American Statistical Association, 32(200):675–701.

References II

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G.
MIMIC-III, a freely accessible critical care database.
3(1):160035.

Le Morvan, M., Josse, J., Moreau, T., Scornet, E., and Varoquaux, G. (2020).
NeuMiss networks: differentiable programming for supervised learning with missing values.
Advances in Neural Information Processing Systems, 33:5980–5990.

National Center for Health Statistics (2017).
National Health Interview Survey (NHIS).

Nemenyi, P. (1963).
Distribution-free Multiple Comparisons.
Princeton University.

References III

- Rubin, D. B. (1976).
  Inference and missing data.
  Biometrika, 63(3):581–592.

- Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R.
  UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age.
  12(3):e1001779.

- The Traumabase Group.
  Traumabase.

References IV

Twala, B. E. T. H., Jones, M. C., and Hand, D. J. (2008).
Good methods for coping with missing data in decision trees.
Pattern Recogn. Lett., 29:950–956.

Wilcoxon, F. (1945).
Individual Comparisons by Ranking Methods.
Biometrics Bulletin, 1(6):80–83.