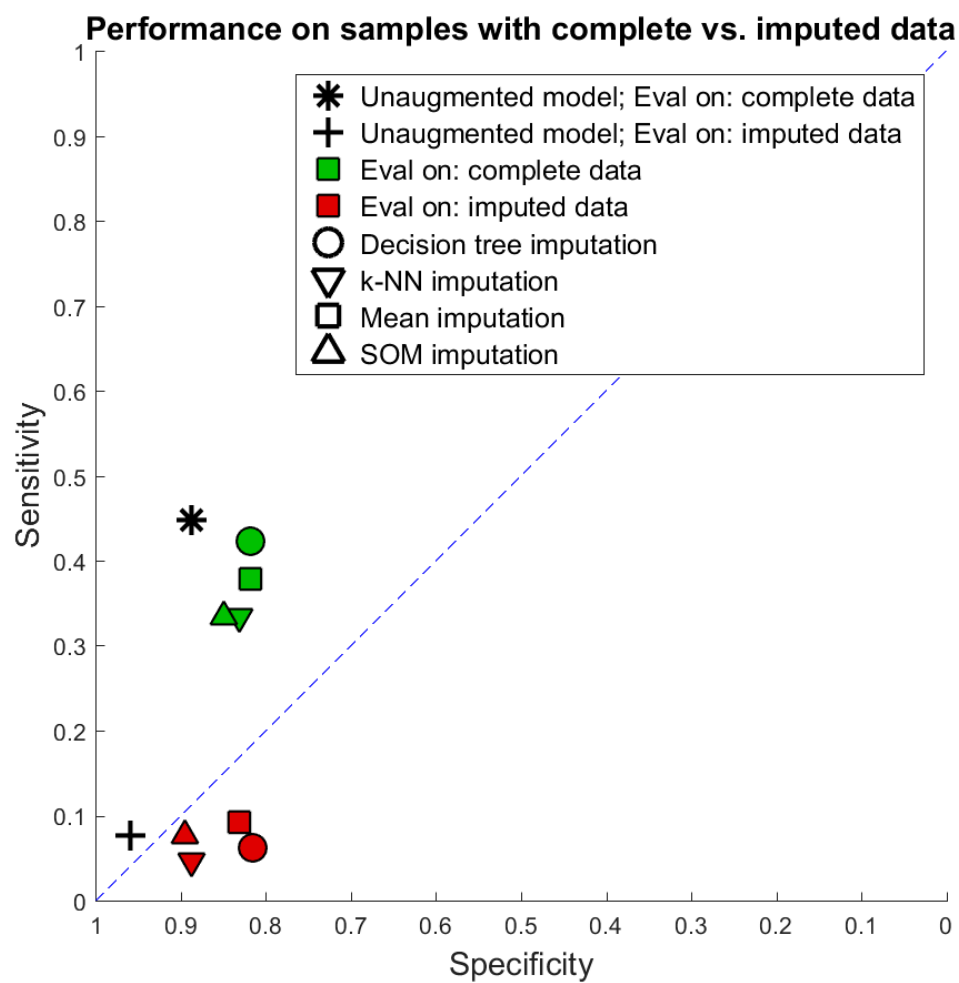
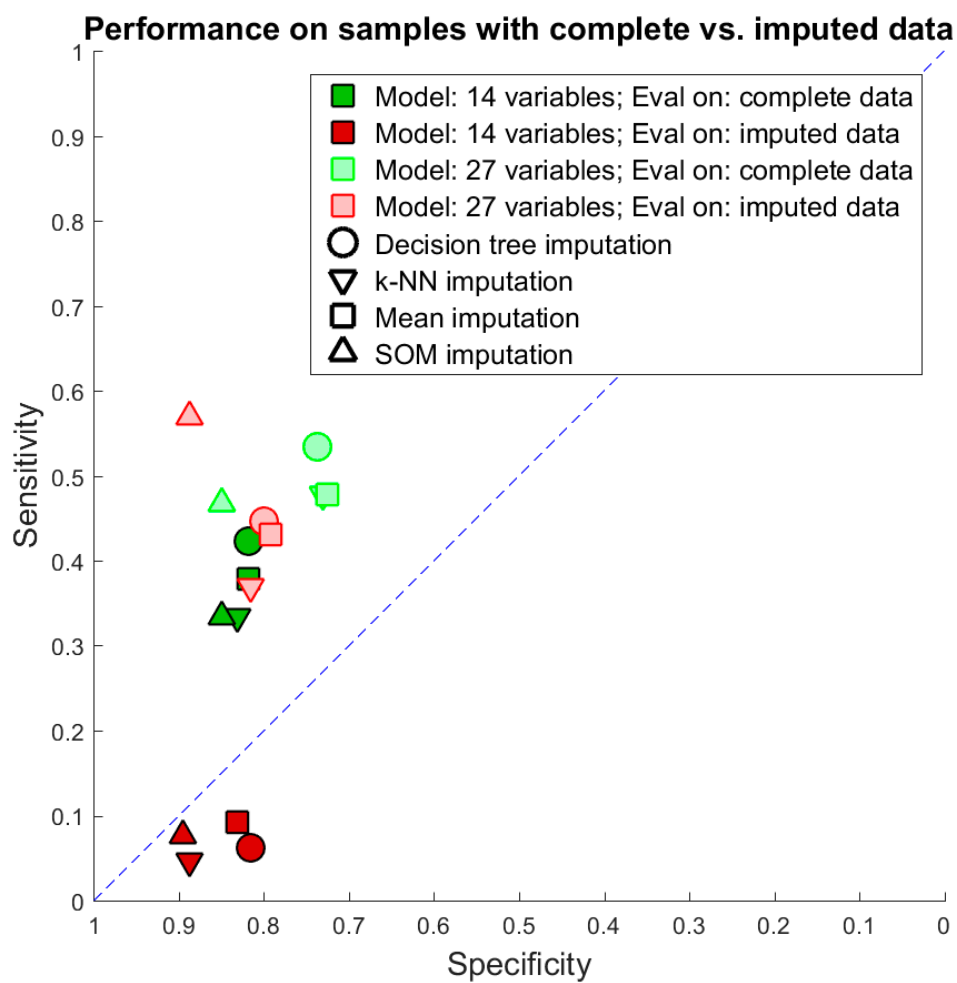


Supplementary Figure 1. Imputed values for the variables in the 27-variable dataset.





Supplementary Figure 2. Performance of 14 and 27-variable imputation-augmented rulesets when values were discretized pre-imputation using a weighted frequency binning method.

Supplementary Table 1. Review of papers comparing performance of imputation methods.

Paper	Imputation Methods	Dataset Types
Kang (2013) [1]	Locally linear reconstruction (LLR) vs. mean imputation, hot deck, k-NN, expectation conditional maximization, k-means, and mixture of Gaussians	13 real world datasets with simulated missingness 1%-50% (MCAR)
	Evaluation Metric	Performance Finding
	Accuracy (%) improvement compared to listwise deletion when training several classifiers	Neural network classifier: At 1% missing values: Worst performing was +0.39% and best was +0.86%. At 50% missing values: Worst performing was +36.77% and best was +44.09%. Other classifiers similar. LLR performed best in majority of tests, but not by much.
Batista and Monard (2003) [2]	k-NN, mean/mode imputation, and no imputation	4 real world datasets with simulated missingness 0%-50% (MCAR)
	Error rates when training decision tree classifiers	Pima dataset, C4.5, 1 attribute missing: At 10% missing values: 26.17% with no imputation, 26.42% with mean imputation, 24.86% with k-NN imputation. Performance varied with dataset. No imputation performed better than k-NN imputation in 11/108 comparisons. Mean imputation performed better than k-NN imputation in 20/108 comparisons.
Troyanskaya et al. (2001) [3]	k-NN vs. mean imputation and SVD imputation	Gene expression datasets with real missing values removed and simulated missing values added 1%-20% (MCAR)
	Normalized root mean squared error of imputed values versus original values	> 0.40 for mean imputation, < 0.33 for SVD imputation, and < 0.25 for k-NN imputation. SVD better than k-NN in cases where expression data is dominated by strong patterns. Both relatively insensitive to percentage of data missing.
Suyundikov et al. (2015) [4]	Weighted k-NN vs. MCMC and EM multiple imputation	Simulated gene expression datasets with simulated missing values added 10%-50% (MCAR)
	Root mean squared error of imputed values versus original values	MCMC and EM multiple imputation had RMSE around 1.5-2.0 for various missing percentages and sizes of datasets, while weighted k-NN had RMSE around 1.3.

Lakshminarayan et al. (1996) [5]	C4.5 decision tree imputation vs. Autoclass	Real world dataset with missing values (assumed MCAR)
	Error rate of classification task	C4.5 decision trees achieved an error rate of 22.6% compared to Autoclass, which achieved an error rate of 48.7%.
Twala (2009) [6]	Decision trees, mean/mode imputation, EM single and multiple imputation (EMSI and EMMI), listwise deletion (LD)	Real world datasets with simulated missing values (MCAR, MAR, and MNAR)
	Excess error % compared to performance on complete dataset using decision tree classification	On average, around 5% excess error difference between worst performing (LD) and best performing (EMMI), with slightly higher rates and differences in MNAR compared to MAR compared to MCAR. DT performed within 3% points of LD and EMMI.
Fessant and Midenet (2002) [7]	Self-organizing map (SOM) vs. mean imputation, hot deck, and multi-layered perceptron (MLP)	Real world dataset with simulated missing values (MCAR)
	Mean squared error of continuous variables, error percentages of categorical variables	MLP performed the best in continuous variables, and SOM performed best in categorical variables. Performance of SOM, MLP, and hot deck were very similar for all variables, and mean imputation performed worse for 3 of 4 variables imputed.
Vatanen et al. (2015) [8]	SOM, Imputation SOM, generative topographic mapping (GTM), Variational Bayesian principal components analysis (VBPCA), and mean imputation	Real world datasets with simulated missing values 10%-50% (MCAR)
	Root mean squared error of imputed values to original values	Both SOM/GTM methods and VBPCA beat mean imputation. SOM/GTM methods slightly underperformed compared to VBPCA on 10-40% missing data, and performed equally to VBPCA on 50% missing data. Mean imputation achieved around 1.01 RMSE, VBPCA from 0.71 to 0.82, SOM from 0.74 to 0.83, Imputation SOM from 0.75 to 0.82, and GTM from 0.75 to 0.82.
Di Zio et al. (2004) [9]	Bayesian networks, hot deck	Real world dataset with simulated missing values 5%-10% (MCAR and MAR)
	Δ , a sum of differences in frequency of each category of a variable in a dataset	Bayesian networks established a Δ of ~0.25 in MAR and ~0.11 in MCAR. Hot deck methods achieved a Δ of 0.24-0.36 in MAR and 0.90-0.13 in MCAR.

		On average, the Δ for Bayesian networks was about 73% that of hot decking in MAR and about 85-92% in MCAR.
Hruschka et al. (2007) [10]	Bayesian networks, EM, MCMC data augmentation, decision trees, and mean/mode imputation	Real world datasets with simulated missing values (MCAR)
	Distances to original of continuous variables and error percentages of categorical variables; average correct classification rate of imputed data in classification tasks	Bayesian networks performed similarly to decision trees in imputation error. Performance on the classification task was highly variable between datasets for each method.
Jerez et al. (2010) [11]	Listwise deletion, mean imputation, hot deck, multiple imputation (using SAS, Amelia, and MICE), MLP, k-NN, and SOM	Real world dataset with real missing values of around 5.61% (assumed MAR)
	AUC of neural network classification task	LD: 0.715 Mean: 0.723 Hot deck: 0.711 SAS: 0.722 Amelia: 0.717 MICE: 0.725 MLP: 0.734 k-NN: 0.735 SOM: 0.733 Standard deviations of AUC were around 0.03 for all. Machine learning methods (MLP, k-NN, SOM) produced statistically significant differences compared to LD.
Rahman and Davis (2013) [12]	Fuzzy unordered rule induction algorithm (FURIA) vs. decision tree, SVM, k-NN, and mean imputation	Real world dataset with real missing values, with variables containing 1-30% missing values and cases containing 4-56% missing values (assumed MAR)
	Accuracy of subsequent classification task	With decision tree classification: Decision tree imputation: 0.8 k-NN: 0.8 FURIA: 0.8 SVM: 0.78 Mean: 0.8 With k-NN classification: Decision tree: 0.71 k-NN: 0.81 FURIA: 0.79 SVM: 0.71

		Mean: 0.77
de Souto et al. (2015) [13]	Mean imputation, median imputation, k-NN, Bayesian PCA, linear least squares	Real world dataset with real missing values (from 0.76% to 3.34% missing after filtering out variables with > 10% missing values) (assumed MCAR)
	Classification error	No statistically significant difference found between imputation methods over a range of classification methods.

- [1] P. Kang, "Locally linear reconstruction based missing value imputation for supervised learning," *Neurocomputing*, vol. 118, pp. 65–78, Oct. 2013.
- [2] G. E. A. P. A. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Appl. Artif. Intell.*, vol. 17, no. 5–6, pp. 519–533, May 2003.
- [3] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–5, Jun. 2001.
- [4] A. Suyundikov, J. R. Stevens, C. Corcoran, J. Herrick, R. K. Wolff, and M. L. Slattery, "Accounting for dependence induced by weighted KNN imputation in paired samples, motivated by a colorectal cancer study," *PLoS One*, vol. 10, no. 4, p. e0119876, Jan. 2015.
- [5] K. Lakshminarayan, S. A. Harp, R. P. Goldman, and T. Samad, "Imputation of Missing Data Using Machine Learning Techniques," in *The Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 1996, pp. 140–145.
- [6] B. Twala, "An Empirical Comparison Of Techniques For Handling Incomplete Data Using Decision Trees," *Appl. Artif. Intell.*, vol. 23, no. 5, pp. 373–405, May 2009.
- [7] F. Fessant and S. Midenet, "Self-Organising Map for Data Imputation and Correction in Surveys," *Neural Comput. Appl.*, vol. 10, no. 4, pp. 300–310, Apr. 2002.
- [8] T. Vatanen, M. Osmala, T. Raiko, K. Lagus, M. Sysi-Aho, M. Orešič, T. Honkela, and H. Lähdesmäki, "Self-organization and missing values in SOM and GTM," *Neurocomputing*, vol. 147, pp. 60–70, Jan. 2015.
- [9] M. Di Zio, M. Scanu, L. Coppola, O. Luzi, and A. Ponti, "Bayesian networks for imputation," *J. R. Stat. Soc. Ser. A (Statistics Soc.)*, vol. 167, no. 2, pp. 309–322, May 2004.
- [10] E. R. Hruschka, E. R. Hruschka, and N. F. F. Ebecken, "Bayesian networks for imputation in classification problems," *J. Intell. Inf. Syst.*, vol. 29, no. 3, pp. 231–252, Jan. 2007.
- [11] J. M. Jerez, I. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martín, and L. Franco, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artif. Intell. Med.*, vol. 50, no. 2, pp. 105–15, Oct. 2010.
- [12] M. M. Rahman and D. N. Davis, "Machine Learning-Based Missing Value Imputation Method for Clinical Datasets," in *IAENG Transactions on Engineering Technologies: Special Volume of the World Congress on Engineering 2012*, G.-C. Yang, S. Ao, and L. Gelman, Eds. Dordrecht: Springer Netherlands, 2013, pp. 245–257.
- [13] M. C. P. de Souto, P. A. Jaskowiak, and I. G. Costa, "Impact of missing data imputation methods on gene expression clustering and classification," *BMC Bioinformatics*, vol. 16, no. 1, p. 64, Jan. 2015.