The Impact of Missing Value Imputation on the Interpretations of Predictive Models: A Case Study on One-year Mortality Prediction in ICU Patients with Acute Myocardial Infarction

Seyedeh Neelufar Payrovnaziri, MS[1], Aiwen Xing, BS[1], Salman Shaeke, MS[1], Xiuwen Liu, PhD[1], Jiang Bian, PhD[2], Zhe He, PhD[1,*]
[1]Florida State University, Tallahassee, Florida, USA;
[2]University of Florida, Gainesville, Florida, USA

Abstract

Acute Myocardial Infarction (AMI) is responsible for the death of millions of people annually around the world, which makes predictive analyses of AMI mortality risk necessary. Rich clinical data in electronic health records (EHR) makes such predictive modeling possible. However, missing values in EHR data is a major issue. Also, the interpretability of predictive models in medicine and healthcare is vital for medical professionals. Therefore, this study examines the impact of imputing missing values in EHR data on the performance and interpretations of predictive models. Our experiments showed a small standard deviation in root mean squared error of different runs of imputation under similar method does not necessarily imply small standard deviation in prediction models' performance and interpretation. Our findings reveal that the imputation method and the level of missingness impact not only the predictive models' performance but also the interpretation of the models in terms of feature importance.

Introduction

Cardiovascular diseases (CVDs) remain the leading cause of death worldwide.[1] Based on a recent report from the European Heart Network,[2] CVDs cause 3.9 million deaths in Europe every year. In 2012, there were 17.5 million deaths from CVDs worldwide with 7.4 million deaths because of coronary heart disease and 6.7 million deaths due to heart attack.[3] CVDs cause a heavy social and economic toll on the nations as well as governments.[4] Among various CVDs, acute myocardial infarction (AMI) is the most severe form of coronary artery disease and a fatal CVD responsible for the death of millions of people annually around the world.[5] Thus, predictive analyses for AMI mortality risk are important for early interventions and procedures.

The wide adoption and implementation of electronic health records (EHR) systems in the United States is the result of a government initiative[6] leading to a large amount of clinical data accumulated in digital forms.[7] These data are a rich source of patient information for predictive analytics in healthcare.[8] Predictive analytics in healthcare and clinical decision-support is not a new topic.[9] Nevertheless, in recent years, there is an increasing demand of using routinely collected real-world data (RWD) such as EHRs, administrative claims, and billing data to generate real-world evidence (RWE) that informs medical care.[10] On the other hand, the emergence and efficient implementation[14] of state-of-the-art machine learning[11] and deep learning methods,[12] as well as powerful computing infrastructure make predictive analytics using EHR data more possible than ever.

However, using EHR data for predictive analytics with machine learning and deep learning methods is still challenging. One major issue is the quality of EHR data due to incompleteness.[15] The existence of missing values in EHR data is due to various reasons including human error, lack of documentation (e.g., the medical expert did not document an evaluation result), or lack of collection (e.g., the medical expert did not perform an evaluation).[16] Thus, a significant body of studies has attempted to approach this challenge by imputing missing values, rather than eliminating records with missing data.[17] Mean imputation is a common approach for imputing missing values in EHR data mainly due to its ease of implementation.[18] Researchers have used multiple imputation by chained equations (MICE)[19] or its variations to deal with missing values in EHR data.[16,20] Other imputation methods based on predictive modeling using machine learning methods are also popular approaches such as MissForest[21] and K-nearest neighbors (KNN)-based imputation.[22] A few recent studies have also used deep learning methods such as generative adversarial networks (GANs)[23] and autoencoders for missing value imputation in EHR data.[22]

[*] Corresponding author: Zhe He. Email: zhe@fsu.edu

On the other hand, predictive analyses using complex machine learning methods (e.g., deep learning), which yield superior prediction accuracy, usually result in black-box models that are not easily interpretable by the end-users. Despite their promising performance, using such complex predictive models in healthcare and clinical decision-making process is quite challenging. Medical professionals need to understand the rationale behind the predictive models' prediction,[24] thus, prefer less complex models such as logistic regression for clinical decision-making.[12] Researchers in the field have taken different approaches to address the interpretability of machine learning models, for instance, feature interaction and importance,[25,26] attention mechanism,[27,28] data dimensionality reduction,[29,30] knowledge distillation and rule extraction.[31,32] However, there are still some fundamental issues that need to be addressed in this regards such as fidelity of the post-hoc interpretation methods to the reference model, evaluation of the interpretation methods, and design biases due to focusing on the intuition of researchers rather than real end-users' (medical professionals in this context) needs. In this study, we consider the interpretation of the predictive models as using feature coefficients and importance of intrinsically interpretable models. Further, although missingness has been recognized as a major kind of the data quality issues of EHR for secondary reuse,[33] how different imputation methods would affect interpretations of machine learning models based on EHR data has not been explored.

Thus, inspired by the importance of predictive modeling in medicine, the challenge of missing values in EHR, and the necessity for building interpretable models in medicine, in this study, we examine the impact of imputation methods on EHR data of AMI patients in ICU on the produced interpretations of several predictive models in terms of feature importance. We use a complete dataset without missing values as the baseline and introduce different levels of missingness (i.e., from 10% to 50%) through simulations. Then, we apply different statistical and machine learning-based imputation methods including mean (mode for two binary variables), MICE, MissForest, and a KNN-based method, as well as Generative Adversarial Imputation Networks (GAIN)[23] - a novel imputation method based on neural networks. We build less complex machine learning models that are intrinsically interpretable and preferred by medical experts, such as logistic regression, linear support vector machine (SVM), and decision tree. We compared these models' performances and interpretations. Further, we build predictive models based on DeepConsensus[34] to experiment if the consensus mechanism would reduce the variance in the performance of predictive models based on slightly different imputed datasets. Our goal is to investigate the impact of imputation methods on not only the performance of the resulting models, but also the interpretation results. To the best of our knowledge, this is the first study to investigate how imputation methods might impact derived interpretations of predictive models on EHR data. Figure 1 demonstrates the workflow of this study.

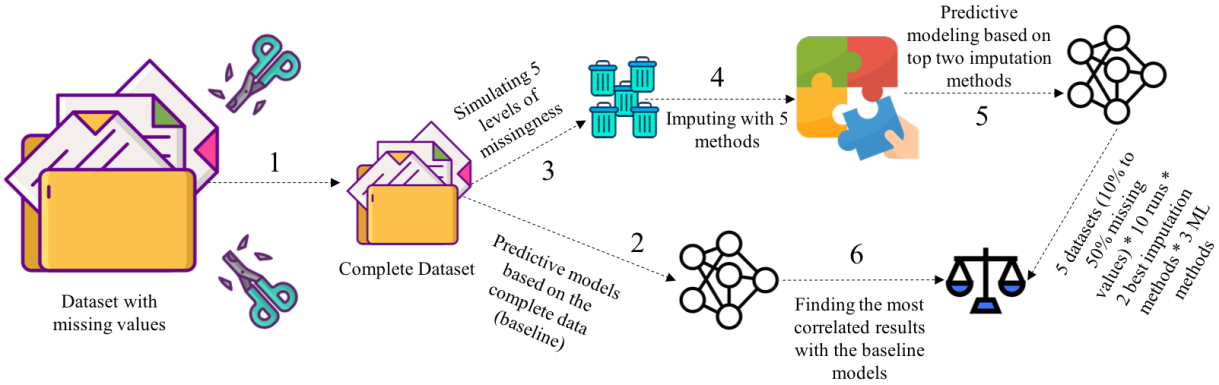

Figure 1. The workflow of this study. *(icons are downloaded from https://www.flaticon.com/)*

Methods

Missingness mechanisms

Since there might be various reasons for a value to be missing in a dataset, there are three main categories of missingness mechanism, including missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). MCAR means the missingness mechanism is not dependent on the outcome variable or any other variables of the dataset. MAR means the missingness did not happen at random. In other words, the missingness is dependent on one or more variables in the dataset. NMAR, the most difficult condition to model for, means the missingness is dependent on the actual value of the missing data.[35] Characterizing the missingness mechanism in EHR data can be an indicator of choosing an appropriate imputation method. This impact has been

explored previously.[22] In this study, we focused on exploring the impact of imputation methods on models' performance and derived interpretation. We simulated random missingness on the complete dataset. Any value in the data was as likely to be missing as any other value. Thus, the missingness mechanism in this study was MCAR.

Data

The Medical Information Mart for Intensive Care (MIMIC-III) database is an integration of de-identified and comprehensive EHR data of patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts.[36] This dataset is freely available and contains patient information spanning over more than a decade. MIMIC-III is widely used in different academic and industrial research. Considering the importance of studying the mortality risk in cardiovascular diseases, especially AMI, in this study, we focused on patients with AMI and post myocardial syndrome (PMS). Thus, the International Classification of Diseases, 9th revision, Clinical Modification (ICD-9-CM) codes considered for this study are 410.0 to 411.0. For each admission, we aggregated the laboratory and chart values and considered the average value of each of the 19 numerical features. We also considered two categorical features with few missing values including gender and initial emergency room diagnosis of AMI. Since our purpose for this study was to examine how the ranking of feature importance would change under different imputation methods and level of missingness, we did not consider longitudinal structure of features. Also, because we compared deep learning-based models with conventional machine learning-based models, we chose to include consistent features throughout the whole experiments. Further, because we should have had a reference to compare the ranking of features under different imputation methods and levels of missingness, we excluded features with more than 50% missing values and applied listwise deletion to the rest of the original dataset. The resulting complete dataset has 3054 observations and 21 features. The binary outcome was dead or not dead within one year after admission. Features description and summary statistics are reported in Table 1.

Table 1. Summary statistics of the variables in the reference complete dataset with 3054 instances.

Variable Name	Description	Variable Type	Min	Max	Median	Mean	Standard Deviation
diastolic	Diastolic blood pressure	Numerical	18.31	134.69	51.2	52.2	11.28
systolic	Systolic blood pressure	Numerical	37.97	484.12	104.96	105.44	22.74
heartRate	Heart rate	Numerical	42	139.48	83.81	84.13	11.97
respRate	Respiratory rate	Numerical	9	42.7	19.33	19.62	3.28
bicarbonate	Bicarbonate	Numerical	8	41.88	24.92	24.61	3.65
calcium	Calcium	Numerical	5.6	13.95	8.43	8.44	0.59
chloride	Chloride	Numerical	80.42	125.61	104	104.01	4.55
potassium	Potassium	Numerical	1.87	6.9	4.14	4.18	0.36
sodium	Sodium	Numerical	118.18	158.5	138.72	138.67	3.47
glucose	Glucose	Numerical	65.67	543	131.76	141.25	39.93
hematocrit	Hematocrit	Numerical	21.11	50.61	30.97	31.53	3.6
hemoglobin	Hemoglobin	Numerical	6.4	16.27	10.43	10.63	1.34
wbc	White blood cell count	Numerical	0.45	107.68	10.93	11.75	5.19
alt	Alanine aminotransferase	Numerical	2	5509	30.33	89.71	270.59
ast	Aspartate aminotransferase	Numerical	2	13511.7	46	142.36	486.22
alp	Alkaline phosphatase	Numerical	19	1147.92	80	102.61	83.6
albumin	Albumin	Numerical	1.2	5	3.2	3.2	0.6
bilirubin	Bilirubin	Numerical	0.1	31.14	0.6	0.97	1.77
admitAge	Age at admission	Numerical	21.22	97.52	72.55	70.89	12.96
InitialERDiagnosisMI (0 = No, 1 = Yes)	Initial emergency room diagnosis was AMI or rule out AMI (#0s = 2050 , #1s = 1004)	Binary	Not applicable				
gender (0 = Female, 1= Male)	Gender (#0s = 1202, #1s = 1852)	Binary					

Imputation Methods

In this study, we evaluated five different imputation methods, including (1) mean value imputation, (2) MICE, (3) K-nearest neighbors (KNN)-based, (4) MissForest, and (5) GAIN. For implementation purposes, we used available packages in R to implement methods (1) to (4). The implementation code of GAIN in Python is made available by its authors on GitHub (https://github.com/jsyoon0823/GAIN). The performance of different imputation methods was compared using root mean squared error (RMSE). The details of these methods are described as follows.

Mean value imputation: The easiest to implement and most conventional approach to impute missing values in EHR data is mean value imputation. However, this simplicity might result in ignoring the underlying statistical information in data and introduce unintentional biases in the subsequent analyses.[16]

MICE: MICE is one of the most popular methods for imputing missing values in EHR data. The main reason resides in its ability to impute different types of variables that might be present in the EHR data. Using MICE, each variable with missing observations is regressed on all the remaining variables in the dataset. The missing values are replaced with the predicted value, and this imputation process is repeated sequentially until all missing values are imputed.

KNN-based: KNN is a machine learning method that can be used for imputing missing values in EHR data.[22] In this approach, missing values are replaced with the mean value of *k* most similar complete observations. A distance function (e.g., Euclidean) is used to measure this similarity.

MissForest: MissForest is a promising imputation method for missing values in EHR data.[21] In this method, first, mean imputation (or any other imputation method) is performed as an initial guess for the missing values. Then, variables in the dataset are sorted based on the amount of missing values they have with the one with fewest missing values ordered first. Further, for each variable *x*, a random forest [37] model is fitted on all other variables' observed values and the outcome variable being the observed values of variable *x*. Then, the trained model is used to predict the missing values of *x*. This process is repeated until a stopping criterion is met.

GAIN: Recently, GAIN, a neural network-based imputation method was introduced for missing value imputation. This imputation method is based on the generative adversarial networks (GAN)[23] framework. In the framework, corresponding to a minimax two-player game, two models are trained simultaneously, a generative model and a discriminative model. The generative model captures the data distribution while the discriminative model estimates the probability of a sample being from the training data or from the generative model. The objective of the generative model is to make the discriminative model make more mistakes. In GAN, the generative and discriminative models are defined based on multilayer perceptron (feedforward neural networks). GAIN is an imputing GAN framework in which the goal of the generative model is to accurately impute the missing values in data, while the goal of the discriminative model is to predict the probability of a value being from the original dataset or from the generative model (observed or imputed component). The objective of the discriminative model in GAIN is to minimize the error loss (on guessing if the elements in the generative model's output are produced by the generative model or from the original data) while the generative model's goal is to maximize discriminative model's mistakes. The authors of GAIN have reported superior imputation performance of GAIN in comparison to autoencoders and other statistical and conventional machine learning-based imputation methods.

Suppose a *d*-dimensional space $\boldsymbol{\chi} = \boldsymbol{\chi_1} \times \ldots \times \boldsymbol{\chi_d}$ and X as a random variable $\boldsymbol{X} = (\boldsymbol{X_1}, \ldots, \boldsymbol{X_d})$ taking values in $\boldsymbol{\chi}$ and a mask vector $\boldsymbol{M} = (\boldsymbol{M_1}, \ldots, \boldsymbol{M_d})$ as another random variable that takes 0 if the value in X is missing and 1 if it is not missing. A new space is defined for each $\boldsymbol{i} \in \{\boldsymbol{1}, \ldots \boldsymbol{d}\}$ as $\widetilde{\boldsymbol{\chi}}_{\boldsymbol{i}} \cup \{*\}$, where * represents an unobserved value. $\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{X}}_{\boldsymbol{1}}, \ldots, \widetilde{\boldsymbol{X}}_{\boldsymbol{d}}) \in \widetilde{\boldsymbol{\chi}}$ is the partial observation of X that takes the corresponding value of X if M is 1 and * otherwise. We inputted the missing data to GAIN and got the imputed datasets as output using Equation 1 and 2,

$$\overline{\boldsymbol{X}} = \boldsymbol{G}(\widetilde{\boldsymbol{X}}, \boldsymbol{M}, (\boldsymbol{1} - \boldsymbol{M}) \odot \boldsymbol{Z}) \quad (1)$$

$$\widehat{\boldsymbol{X}} = \boldsymbol{M} \odot \widetilde{\boldsymbol{X}} + (\boldsymbol{1} - \boldsymbol{M}) \odot \overline{\boldsymbol{X}} \quad (2)$$

where $\overline{\boldsymbol{X}} \in \boldsymbol{\chi}$ is the vector of imputed values, *G* is the generator function, $\widehat{\boldsymbol{X}} \in \boldsymbol{\chi}$ is the vector obtained by partial observations $\widetilde{\boldsymbol{X}}$ and replacing missing values with the values of $\overline{\boldsymbol{X}}$, Z is a *d*-dimensional noise which is independent

of all the other variables, and $\odot$ is element-wise multiplication. In (1) G function takes $\widetilde{\boldsymbol{X}}$, M, and Z as input and outputs $\overline{\boldsymbol{X}}$. Then in (2), $\widehat{\boldsymbol{X}}$ is obtained by replacing each * with the corresponding value from $\overline{\boldsymbol{X}}$.

Predictive Modeling

To compare the prediction performance and feature importance ranking of different prediction models with different level of missingness, we performed predictive analyses using three popular machine learning methods in predictive modeling with EHR data,[38] namely logistic regression, SVM, and decision tree. Further, we captured feature coefficients and importance in each model to compare to the same in the reference model of its own kind. For comparison, we used Pearson correlation coefficients. Higher correlations mean closer results (in terms of feature importance) to the reference model based on the complete dataset. Also, we built a deep learning model (i.e., DeepConsensus) to investigate if it can reduce variance. The binary prediction task was patient mortality within one year after admission. The dataset is divided to separate training and testing sets at the ratio of 0.9 to 0.1 respectively. The dataset is imbalanced with 65 (negative class) to 35 (positive class) ratio. For implementation purposes we used Python programming language with Tensorflow, NumPy, Pandas, and Sklearn packages. We give a brief description of DeepConsensus in the following.

DeepConsensus: The main idea behind DeepConsensus is that since different deep neural networks tend to classify training samples accurately, they generate similar linear regions. Thus, these models should behave similarly on classifying training samples. Such behavior enables multiple models to agree with each other on classifying valid inputs and filtering out adversarial examples, while individual models are sensitive to those examples. Using consensus among different models helps to capture the underlying structure of data. It is shown that consensus helps to differentiate extrinsically classified samples (i.e., classified under extrinsic factors such as randomness of weight initialization) from consistently classified samples (i.e., samples that are classified in the same class with high probability by multiple models). Thus, such consensus mechanism among multiple models can reduce the variance caused by extrinsic factors. The effectiveness of this method is demonstrated in the reference paper.[34]

Results

Imputation Performance

We compared five different imputation methods including MICE, MissForest, KNN-based, Mean (mode for binary variables), and GAIN, on 10% to 50% missing datasets. We ran each experiment 10 times and computed the average RMSE along with its standard deviation. The results are reported in Table 2.

Table 2. Average RMSE from different imputation methods on 10% to 50% missing data (10 runs).

Missingness level / Imputation Method	10%	20%	30%	40%	50%
MICE	0.2254±0.0025	0.2249±0.0020	0.2292±0.0013	0.2343±0.0014	0.2325±0.0014
MissForest	0.1935±0.0017	0.1950±0.0013	0.1997±0.0018	0.2051±0.0018	0.2062±0.0011
KNN-based	0.21447	0.2219	0.2241	0.2267	0.2228
Mean\Mode	0.2038	0.2046	0.2052	0.2066	**0.2059**
GAIN	**0.1757±0.0049**	**0.1763±0.0064**	**0.1838±0.0039**	**0.1963±0.0114**	0.2088±0.0100

Averaging the RMSE of different imputation methods across all different levels of missing values, GAIN showed the best performance. MissForest was the second-best performing imputation method following GAIN showing only 0.01276 difference in RMSE on average. Mean came next, showing quite monotonic behavior across all datasets with different amount of missing values. KNN-based and MICE showed similar performance reporting the highest RSME (worst performance).

Prediction Performance

Next, we based our experiments on the datasets imputed with the best performing imputation methods GAIN and MissForest. First, we built the reference models that served as the benchmark for our comparisons (models based on the complete dataset). Then, we built models based on datasets with varying percentage of missingness that are imputed using GAIN and MissForest (through 300 experiments = 5 levels of missingness * 2 imputation methods* 3 ML methods * 10 runs each). The performance of these models based on GAIN imputed datasets and MissForest imputed datasets in terms of area under the receiver operating characteristic curve (AUROC) are depicted in Figure

2. AUROC is a classification performance measure that illustrates how models perform in terms of discriminating between the classes. The performance of SVM and logistic regression models based on MissForest imputed datasets (Figure 2) showed lower variance in comparison to the same models based on GAIN imputed datasets. The variance in decision tree models based on the datasets imputed by both GAIN and MissForest is relatively high, making those models' performance unstable even under low levels of missingness.

By increasing the number of missing values in the datasets from 10% to 50%, models based on the datasets imputed by MissForest showed a more stable behavior in comparison to GAIN across all models. A closer look at SVM performance (based on AUROC) from 10% to 50% missing values imputed by MissForest showed an increase of standard deviation from 0.00375523 to 0.01062501 with an average of 0.00674935. This measure was 0.00277818 to

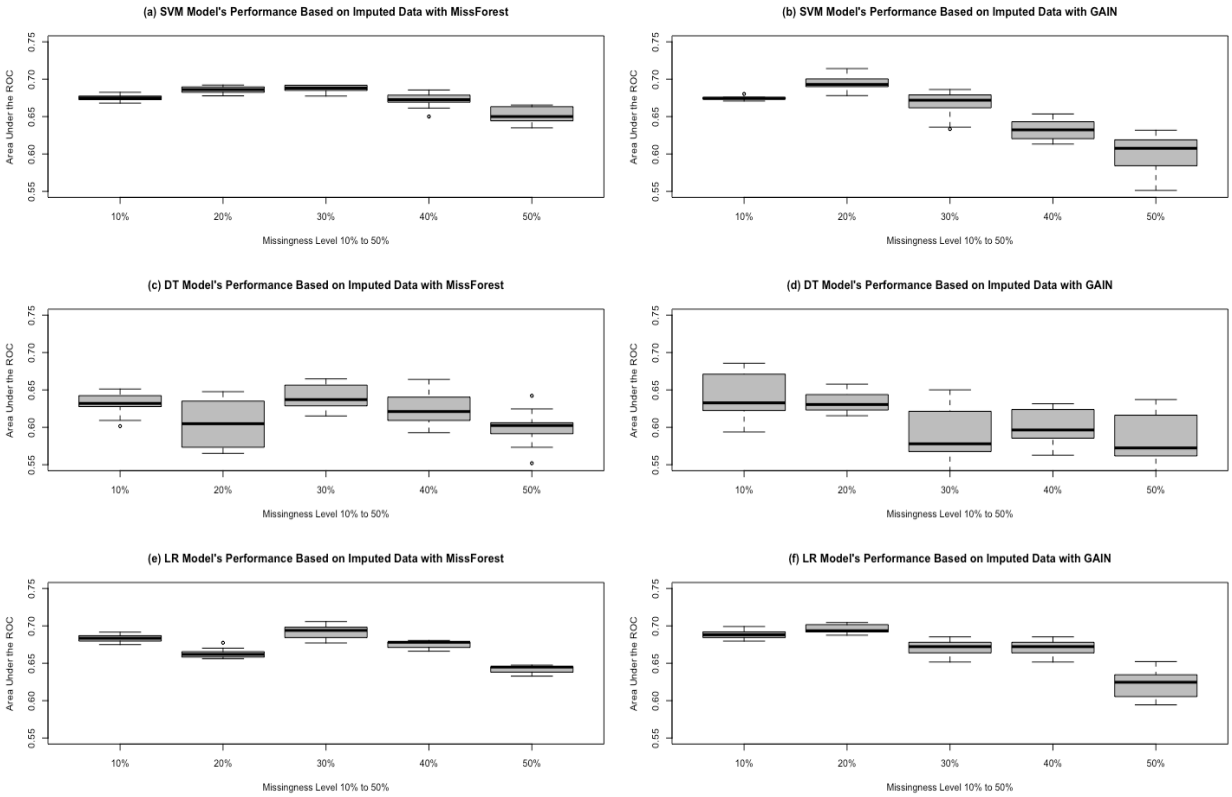

Figure 2. Comparing the performance of models that are built using (a) support vector machine (SVM) based on the imputed datasets with MissForest, (b) SVM based on the imputed datasets with GAIN, (c) decision tree (DT) based on the imputed datasets with MissForest, (d) DT based on the imputed datasets with GAIN, (e) logistic regression (LR) based on the imputed datasets with MissForest, (f) LR based on the imputed datasets with GAIN, under gradually increasing missingness level. The performance is reported on area under the receiver operating characteristic curve (AUROC).

0.0253859 for GAIN with an average of 0.01441592. Also, a big jump in standard deviation was not observed in case of MissForest until 40% of missingness. This jump occurred at 20% of missingness in case of GAIN. The standard deviation of AUROC of multiple runs for logistic regression models based on GAIN showed an increasing trend from 0.00561881to 0.01919611 with the average of 0.0101355. However, in case of MissForest, the trend is increasing from 10% to 30% and then decreasing from 30% to 50% with an average of 0.00619782. We hypothesized that the variance among models on the same missing level of missing data that is imputed with the same method 10 times can be reduced by using the consensus mechanism among multiple models. Applying DeepConsensus on the complete dataset (baseline) showed a significant increase of performance as expected: from 0.7212 AUROC in logistic regression and 0.7137 in SVM to 0.8095 in DeepConsensus. Further, we narrowed down our experiments to 10 datasets that were the result of imputing 10% missingness (lowest missing rate) 10 times with MissForest (imputation method with less variance on machine learning models). The average AUROC across 10 experiments showed an

increase of performance to 0.7908 (from 0.67505 in SVM and 0.68307 in logistic regression). However, the variance in performance still persists at 0.034460858.

Feature Importance Ranking Comparison

Pearson correlation analyses (feature importance of each model with feature importance of corresponding baseline model) are reported in Table 3. These results showed a generally lower performance for decision tree models across all datasets. Focusing on SVM and logistic regression with higher performance, averaging coefficients as a result of 10 runs for each level of missingness showed a correlation coefficient of more than 0.99 with statistically significant results on the imputed datasets of 10% missing value with GAIN and MissForest. However, this trend on models based on an increasing level of missingness on average showed a decreasing correlation with the baseline model across all machine learning methods and imputation methods. A further principal component analysis (PCA) on the complete dataset showed that a linear model based on 30% missing values could capture between 90.130% (14 features) and 92.824% (15 features) of the statistical information in this dataset.

Table 3. Pearson correlation coefficients and p-values of feature importance comparison between the models based on complete dataset (baselines) and the models based on the imputed datasets.

Machine Learning method	Imputation method	Missingness %	Pearson correlation coefficient	p-value	Imputation method	Missingness %	Pearson correlation coefficient	p-value
Decision Tree	GAIN	10%	0.9661	1.24E-12	MissForest	10%	0.9445	1.23E-10
		20%	0.9591	7.09E-12		20%	0.9146	6.56E-09
		30%	0.9174	4.84E-09		30%	0.9066	1.49E-08
		40%	0.8426	1.64E-06		40%	0.8741	2.24E-07
		50%	0.9033	2.05E-08		50%	0.7707	4.34E-05
SVM		10%	0.9924	8.57E-19		10%	0.9941	7.71E-20
		20%	0.9867	1.80E-16		20%	0.9839	1.12E-15
		30%	0.9491	5.54E-11		30%	0.9851	5.31E-16
		40%	0.8348	2.51E-06		40%	0.9758	5.09E-14
		50%	0.7939	1.73E-05		50%	0.9118	8.83E-09
Logistic Regression		10%	0.9954	6.80E-21		10%	0.9968	2.56E-22
		20%	0.9883	5.14E-17		20%	0.9872	1.28E-16
		30%	0.9620	3.62E-12		30%	0.9854	4.24E-16
		40%	0.8959	4.00E-08		40%	0.9787	1.53E-14
		50%	0.8541	8.39E-07		50%	0.9589	7.44E-12

Discussion

Comparing the imputation methods' RMSE performance reported in Table 2 might imply that (1) GAIN performs better than MissForest, and (2) the standard deviation between different runs of the same method on the same dataset with missing values is small. However, our experiments showed that choosing the best imputation method might not always be a straightforward process. Although GAIN surpassed all imputation methods in terms of RMSE on all datasets, MissForest imputation yielded more stable results (smaller standard deviation on average) at the presence of gradually increasing missing values. Also, comparing the performance of models based on datasets with different percentage of missingness reveals the fact that higher performance does not necessarily indicate more similar interpretations to the reference model. We observed that, on average a relatively small standard deviation of RMSE across all levels of missingness yielded a bigger standard deviation in models' performance and a lower correlation

of feature importance between baseline models and the models based on imputed datasets. Also, the dilemma of bias/variance is well understood regarding neural networks which are hyperparameterized.[39] Training neural networks requires a larger number of training samples to achieve acceptable performance and less variance. Thus, although using consensus of deep models did not resolve the issue of variance in this study, we hypothesize that using a bigger dataset (with more samples and more features) could potentially yield a more stable consensus of deep learning models and result in less variance.

These observations might not be generalizable to other datasets or imputation methods. There is no universally optimal approach for missing data imputation or predictive modeling using EHR data. However, these experiments showed that the way we approach missing values in EHR data impacts not only the model performance but also the interpretations of the models' predictions. In real-world predictive analyses of EHR data, it is usually not possible to obtain a dataset with no missing values. However, in cases where the interpretations of predictive models matter, in order to choose the best imputation method, just relying on RMSE or model performance measures might not be sufficient. In these cases, we suggest to run extensive experiments on a smaller complete-case version of the dataset first, evaluate the impact of different imputation methods on the interpretations in comparison to the complete-case, and then apply the best performing method on the original dataset with missing values. In cases where it is not possible to have the complete-case dataset, researchers should be aware of this potential impact, use different imputation methods for predictive modeling, and discuss the resulting interpretations with medical experts or compare to the medical knowledge when choosing the imputation method that yields the most reasonable interpretations. Also, more in-depth analyses of data with methods such as PCA can be used to investigate the redundancy in datasets and determine the maximal allowed missing value rate.

Limitations and Future Opportunities

A potential limitation of this study was the relatively small and imbalanced dataset (65:35). Although the findings in this study are robust, future studies could be done on datasets with more balance and more samples to investigate how these results would change. Another limitation was using snapshot of features instead of longitudinal structure. The focus of this study was to examine how different imputation methods can potentially impact the resulting performance and interpretations of different predictive models. Thus, we had to be consistent in terms of the data representation to models and the number of features we used across all experiments. Future studies could be conducted on longitudinal features and investigating the impact of imputation methods on the resulting interpretations. However, modeling and interpreting longitudinal EHR data is inherently challenging due to different granularity of different variables. Another limitation to this study was the fact that the missingness mechanism of data was MCAR. However, in real-world EHR data this might not be the case. Investigating the impact of imputation methods, especially on the models' interpretations, under MAR and NMAR in comparison to MCAR can provide a broader view and understanding of the underlying challenges with regards to EHR data imputation. Another limitation of this study is that it is an empirical study rather than a theoretical study. We encourage future theoretical studies on investigating the possible impacts of missing value imputation on the models' interpretations.

Conclusions

In this study, we simulated 5 levels of missingness (10% to 50%) on a complete EHR dataset of 21 features for 3054 patients with AMI from MIMIC-III database. We examined different statistical and machine learning-based imputation methods such as mean, MICE, MissForest, and KNN-based, as well as GAIN-a novel imputation method based on GAN. Our experiments showed that GAIN and MissForest yielded best performance in terms of RMSE and small standard deviations across all levels of missingness. However, further predictive modeling (using machine learning and deep learning methods) based on each of these datasets revealed the fact that the variance in their performance (in terms of AUROC) gradually grows with more missingness. Also, Pearson correlation analyses showed that the similarity of feature importance of models based on the imputed datasets to the feature importance of baseline models gradually decreases, a trend that could not initially be inferred by just looking at the performance of imputation and predictive modeling in terms of RMSE and AUROC respectively.

Acknowledgments

This study was supported in part by the National Institute on Aging (NIA) of the National Institutes of Health (NIH) under Award Number R21AG061431; and the University of Florida Clinical and Translational Science Institute, which is supported in part by the NIH National Center for Advancing Translational Sciences under award number

UL1TR001427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. al-Aiad A, Duwairi R, Fraihat M. Survey: Deep Learning Concepts and Techniques for Electronic Health Record. In: 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA). 2018. p. 1–5.
2. Wilkins E, Wilson L, Wickramasinghe K, Bhatnagar P, Leal J, Luengo-Fernandez R, et al. European Cardiovascular Disease Statistics 2017. 2017 Feb 15 [cited 2019 Jul 23]; Available from: https://researchportal.bath.ac.uk/en/publications/european-cardiovascular-disease-statistics-2017
3. Organisation mondiale de la santé. Global status report on noncommunicable diseases 2014: attaining the nine global noncommunicable diseases targets; a shared responsibility. Geneva: World Health Organization; 2014.
4. Ford ES, Capewell S. Coronary Heart Disease Mortality Among Young Adults in the U.S. From 1980 Through 2002: Concealed Leveling of Mortality Rates. Journal of the American College of Cardiology. 2007 Nov 27;50(22):2128–32.
5. Reed GW, Rossi JE, Cannon CP. Acute myocardial infarction. The Lancet. 2017 Jan 14;389(10065):197–210.
6. Blumenthal D. Implementation of the Federal Health Information Technology Initiative. New England Journal of Medicine. 2011 Dec 22;365(25):2426–31.
7. ANSI I. ISO/DTR 20514: Health informatics–electronic health record—definition, scope and context. ISO; 2005.
8. Milenkovic MJ, Vukmirovic A, Milenkovic D. Big data analytics in the health sector: challenges and potentials. Management: Journal of Sustainable Business and Management Solutions in Emerging Economies. 2019;24(1):23–33.
9. Parikh RB, Kakad M, Bates DW. Integrating Predictive Analytics Into High-Value Care: The Dawn of Precision Delivery. JAMA. 2016 Feb 16;315(7):651–2.
10. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-World Evidence — What Is It and What Can It Tell Us? New England Journal of Medicine. 2016 Dec 8;375(23):2293–7.
11. Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. PLoS One [Internet]. 2018 Aug 31 [cited 2020 Feb 20];13(8). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6118376/
12. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE Journal of Biomedical and Health Informatics. 2018 Sep;22(5):1589–604.
13. Obermeyer Z, Emanuel EJ. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016 Sep 29;375(13):1216–9.
14. Erickson BJ, Korfiatis P, Akkus Z, Kline T, Philbrick K. Toolkits and Libraries for Deep Learning. J Digit Imaging. 2017 Aug 1;30(4):400–5.
15. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. Summit on Translat Bioinforma. 2010 Mar 1;2010:1–5.
16. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. EGEMS (Wash DC). 2013;1(3):1035.
17. Little RJ, Rubin DB. Statistical analysis with missing data. Vol. 793. John Wiley & Sons; 2019.
18. Salgado CM, Azevedo C, Proença H, Vieira SM. Missing Data. In: MIT Critical Data, editor. Secondary Analysis of Electronic Health Records [Internet]. Cham: Springer International Publishing; 2016 [cited 2020 Feb 20]. p. 143–62. Available from: https://doi.org/10.1007/978-3-319-43742-2_13
19. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of statistical software. 2010;1–68.
20. Sun P. MICE-DA: A MICE method with Data Augmentation for missing data imputation in IEEE ICHI 2019 DACMI Challenge. In: 2019 IEEE International Conference on Healthcare Informatics (ICHI). 2019. p. 1–3.
21. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012 Jan 1;28(1):112–8.
22. Beaulieu-Jones BK, Lavage DR, Snyder JW, Moore JH, Pendergrass SA, Bauer CR. Characterizing and managing missing structured data in electronic health records: data analysis. JMIR medical informatics. 2018;6(1):e11.

23. Yoon J, Jordon J, Schaar M. GAIN: Missing Data Imputation using Generative Adversarial Nets. In: International Conference on Machine Learning [Internet]. 2018 [cited 2020 Feb 27]. p. 5689–98. Available from: http://proceedings.mlr.press/v80/yoon18a.html
24. Ahmad MA, Eckert C, Teredesai A. Interpretable Machine Learning in Healthcare. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics [Internet]. Washington, DC, USA: Association for Computing Machinery; 2018 [cited 2020 Feb 21]. p. 559–560. (BCB '18). Available from: https://doi.org/10.1145/3233547.3233667
25. Eck A, Zintgraf LM, de Groot EFJ, de Meij TGJ, Cohen TS, Savelkoul PHM, et al. Interpretation of microbiota-based diagnostics by explaining individual classifier decisions. BMC Bioinformatics. 2017 Oct 4;18(1):441.
26. Pan L, Liu G, Mao X, Li H, Zhang J, Liang H, et al. Development of Prediction Models Using Machine Learning Algorithms for Girls with Suspected Central Precocious Puberty: Retrospective Study. JMIR Med Inform. 2019 Feb 12;7(1):e11728.
27. Kaji DA, Zech JR, Kim JS, Cho SK, Dangayach NS, Costa AB, et al. An attention based deep learning model of clinical events in the intensive care unit. PLoS ONE. 2019;14(2):e0211057.
28. Park S, Kim YJ, Kim JW, Park JJ, Ryu B, Ha J-W. [Regular Paper] Interpretable Prediction of Vascular Diseases from Electronic Health Records via Deep Attention Networks. In: 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE). 2018. p. 110–7.
29. Bernardini M, Romeo L, Misericordia P, Frontoni E. Discovering the Type 2 Diabetes in Electronic Health Records using the Sparse Balanced Support Vector Machine. IEEE Journal of Biomedical and Health Informatics. 2019;1–1.
30. Zhao LP, Bolouri H. Object-oriented regression for building predictive models with high dimensional omics data from translational studies. J Biomed Inform. 2016 Apr;60:431–45.
31. Che Z, Purushotham S, Khemani R, Liu Y. Interpretable Deep Models for ICU Outcome Prediction. AMIA Annu Symp Proc. 2017 Feb 10;2016:371–80.
32. Luo G. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. Health Inf Sci Syst. 2016 Dec;4(1):2.
33. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013;20(1):144–51.
34. Salman S, Payrovnaziri SN, Liu X, Rengifo-Moreno P, He Z. Consensus-based Interpretable Deep Neural Networks with Application to Mortality Prediction. arXiv:190505849 [cs, stat] [Internet]. 2019 Sep 11 [cited 2020 Mar 11]; Available from: http://arxiv.org/abs/1905.05849
35. Scheffer J. Dealing with missing data. 2002 [cited 2020 Feb 28]; Available from: https://mro.massey.ac.nz/handle/10179/4355
36. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific Data. 2016 May 24;3(1):1–9.
37. Breiman L. Random Forests. Machine Learning. 2001 Oct 1;45(1):5–32.
38. Taslimitehrani V, Dong G, Pereira NL, Panahiazar M, Pathak J. Developing EHR-driven heart failure risk prediction models using CPXR(Log) with the probabilistic loss function. Journal of Biomedical Informatics. 2016 Apr 1;60:260–9.
39. Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. Neural computation. 1992;4(1):1–58.