World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:18, No:4, 2024

Imputing Missing Data in Electronic Health Records: A Comparison of Linear and Non-Linear Imputation Models

Alireza Vafaei Sadr, Vida Abedi, Jiang Li, Ramin Zand

Abstract—Missing data is a common challenge in medical research and can lead to biased or incomplete results. When the data bias leaks into models, it further exacerbates health disparities; biased algorithms can lead to misclassification and reduced resource allocation and monitoring as part of prevention strategies for certain minorities and vulnerable segments of patient populations, which in turn further reduce data footprint from the same population – *thus, a vicious cycle*. This study compares the performance of six imputation techniques grouped into Linear and Non-Linear models, on two different real-world electronic health records (EHRs) datasets, representing 17864 patient records. The mean absolute percentage error (MAPE) and root mean squared error (RMSE) are used as performance metrics, and the results show that the Linear models outperformed the Non-Linear models in terms of both metrics. These results suggest that sometimes Linear models might be an optimal choice for imputation in laboratory variables in terms of imputation efficiency and uncertainty of predicted values.

Keywords—EHR, Machine Learning, imputation, laboratory variables, algorithmic bias.

I. Introduction

HEALTHCARE relies heavily on data for improvements and optimizations. A primary data source is Electronic Health Records (EHRs), which are comprehensive repositories of patient information. When effectively utilized, this information can significantly enhance healthcare delivery, resource optimization, and patient outcomes. EHRs can be used for improving healthcare delivery, resource optimization, health disparity, and access [1]–[6]. Clinical data, extracted from EHRs can be analyzed to identify patterns, predict outcomes, and develop treatment plans [7]–[13]. However, EHRs are often noisy and suffer from missing. EHR imputation is the process of using statistical techniques to fill in missing data since the accuracy and completeness of EHRs are crucial for their effective utility. The EHR data may be incomplete due to various reasons, such as missing data points, inconsistent coding, patients moving to different regions/health systems, or simply poor documentation [14]. This lack of information can make it challenging to derive meaningful and unbiased insights. EHR imputation plays a crucial role in clinical research and healthcare delivery [15]. More reliable and higher density EHR data can help to better identify high-risk patient groups for specific conditions, enabling early diagnosis, improved prognosis, development of more personalized treatment plans, and enhancing model-driven resource optimization. Given the importance and utility of EHR in improving healthcare efficiency, there is growing interest in developing novel machine learning-based methodologies in the field of EHR data imputation. Yet, often the newly developed methods are still based on the assumption of random missing patterns which leads one to assume that the missing values can be predicted based on the available data [16]–[18]. Mean imputation, mode imputation, and regression imputation are some of the most commonly used simple imputation techniques [19], [20].

Recent studies in EHR imputation have shown promising results with the use of machine learning techniques such as deep learning and graph models to predict missing values. These methods are capable of modeling complex relationships between variables and can handle high-dimensional data, making them ideal for a broad range of applications, especially when dealing with the challenges of handling big data. For instance, recent studies proposed various frameworks that use deep learning to address missing EHR data, achieving state-of-the-art performance compared to traditional imputation methods on the evaluated datasets [21]–[27]. These studies highlight the potential of deep learning-based imputation methods for improving the accuracy of EHR data imputation.

In our study, we strive to highlight the effectiveness of both Linear and Non-Linear machine learning models for iterative imputation models. Given the dataset's size and the potential risk of overfitting with more intricate models, it is crucial to acknowledge that more complexity does not necessarily guarantee superior performance compared to simpler models. Therefore, it is necessary to explore the performance of various models in different scenarios. By comparing the performance of Linear models, such as linear regression, with Non-Linear models, such as tree-based and deep learning models, on two real-world datasets, we aim to provide researchers with valuable insights that highlight the importance of considering both Non-Linear and Linear models, as Non-Linear models are not always the superior choice for complex medical datasets.

This paper is structured to present the data, imputation methodology and evaluation metrics, results, and discussion. The second section of this paper will introduce the data, imputation algorithms, and evaluation metrics. The third section will present the results of our experiments. The last section provides a summary and conclusion. Overall, our study

AVS and VA are with the Department of Public Health Sciences, College of Medicine, The Pennsylvania State University, Hershey, PA 17033 (e-mail: asadr@pennstatehealth.psu.edu).

JL is with the Department of Molecular and Functional Genomics, Weis Center for Research, Geisinger Health System, Danville, PA 17822.

RZ is with the Department of Neurology, College of Medicine, The Pennsylvania State University, Hershey, PA 17033, and the Geisinger Neuroscience Institute, Geisinger Health System, Danville, PA 17822

aims to provide insights into the performance of Linear and Non-Linear machine learning models for imputing laboratory variables extracted from EHR. Selecting the most appropriate imputation technique for a given set of variables can lead to the development of more accurate and less biased datasets for EHR-based model development.

II. Methodology: Dataset, Imputation Algorithms, and Evaluation Metrics

Dataset: We utilized two real-world clinical datasets to evaluate the performance of Linear versus Non-Linear machine learning-based models for iterative imputation. In this study, our focus was on laboratory variables due to their significant impact on disease diagnosis and prognosis. Imputation or lack thereof can introduce algorithmic bias, emphasizing the importance of careful design to mitigate bias stemming from the training dataset.

The first dataset is an EHR-based cohort of 9037 ischemic stroke patients from a large integrated healthcare system with multiple hospitals in the United States. The stroke dataset includes 45 most common variables with missingness less than 75 percent. Table IV provides a detailed summary of the laboratory variables used, their percentage missing, and the summary statistics of the observed values. The pattern of missing in these 45 variables varies, depending on the level of missingness. For instance, the most common laboratory variables with the lowest missing rate exhibit a more random missing pattern, while the variables with a higher missing pattern show a not completely random missing pattern.

The second dataset is a subset of the Medical Information Mart for Intensive Care (MIMIC) database version 1.4 [28], from which laboratory findings are extracted. More specifically, we selected the subset of the MIMIC dataset without missing for the 45 laboratories used in the stroke database; we simulated a missing pattern similar to the ischemic stroke dataset, to be able to assess model performance. The complete MIMIC dataset contains clinical data for over 38,000 ICU patients, including demographic information, vital signs, laboratory results, and medication orders. Our subset of the MIMIC dataset included 8827 patient records. Table V provides a detailed summary of the laboratory variables used, their percentage missing, and the summary statistics of the observed values. The MIMIC data missingness pattern is similar to the stroke data.

Additionally, we used the stroke patients' observed missingness pattern to also simulate missingness on the MIMIC laboratory values. Using this strategy, we avoided the assumption of missing at random and used observed missing patterns to generate the holdout values for model evaluation (the ground truth values are therefore available for comparison with the imputed values); for additional details please refer to the model evaluation section below.

Imputation algorithms: We compared Linear and Non-Linear imputation techniques. In total, we used three imputation algorithms for Linear and three imputation algorithms for Non-Linear-based models. For the Linear models, we used linear regression (LR) [29], ridge regression (Ridge) [30], and Lasso regression (Lasso) [31]. For the Non-linear models, we used random forest (RF) [32], extreme gradient boosting (XGB) [33], and multi-layer perceptron (MLP) [34].

Model evaluation: We used Python to implement and evaluate the performance of Linear and Non-Linear machine learning models employing an iterative imputation strategy on the two real-world datasets. To ensure the robustness of the results, we ran all the experiments 50 times and reported the mean and 68% confidence intervals. To evaluate the performance of our imputation models, we used two commonly used regression metrics, Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). The following two equations describe RMSE and MAPE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \tag{1}$$

where n is the number of observations, y_i is the actual value of the i^{th} observation, and $\hat{y_i}$ is the predicted value of the i^{th} observation.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y_i}}{y_i}\right| \times 100 \tag{2}$$

where n is the number of observations, y_i is the actual value of the i^{th} observation, and $\hat{y_i}$ is the predicted value of the i^{th} observation. The absolute value of the difference between the actual and predicted values is divided by the actual value and multiplied by 100% to obtain the percentage difference, which is then averaged across all observations. RMSE measures the average difference between the predicted and observed laboratory values, while MAPE measures the average percentage difference between the predicted and observed values. Since our data consists of continuous laboratory values, we selected these regression metrics to evaluate and compare the models.

To test the performance of the models, we randomly held out 50 values from each dataset as a test set and used the remaining data to train and validate our models. Overall, applying the regression metrics on these datasets allowed us to compare the performance of the Linear and Non-Linear machine learning models for iterative imputation and identification of the optimal approach for imputing missing laboratory variables from real-world clinical data. It has been previously shown that a hold-out strategy of 50 values can provide robust results [11].

III. Results

Fig. 1 and Table I present the results of evaluating six different models (three Non-Linear and three Linear) on the MIMIC dataset using MAPE and RMSE as performance metrics. The Non-Linear model achieved the highest MAPE value of 2.53 [2.05, 3.07] while the Linear model achieved a slightly lower value of 2.3 [1.81, 2.84]. In terms of RMSE, the Linear model had a lower value of 0.0027 [0.00136, 0.00396] in comparison with the Non-Linear model with a value of 0.00313 [0.00165, 0.00431].

We conducted an independent t-test to assess the statistical significance of performance differences among the models.

TABLE I
MEAN AND 68%-CI FOR METRICS ON *MIMIC*

Model	MAPE	RMSE
Non-Linear	$2.53^{3.07}_{2.05}$	$0.00313^{0.00431}_{0.00165}$
Linear	$2.30^{2.84}_{1.81}$	$0.0027^{0.00396}_{0.00136}$

Specifically, the p-value for comparing the RMSE values of Non-Linear versus the Linear models is lower than 0.05, indicating that the differences in RMSE are statistically significant. However, the p-value for comparing the MAPE values of Non-Linear versus the Linear models is 0.105, indicating that the difference in MAPE between these two models is not statistically significant. Overall, these results suggest that the Linear models can outperform the Non-linear models in terms of RMSE, when tested on 45 different laboratory variables from the MIMIC dataset.

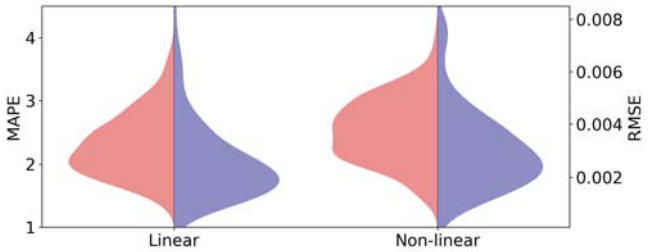

Fig. 1 Compared predicted and actual values for missing lab data in **MIMIC** dataset using linear and non-linear models; error bars show standard deviation

Fig. 2 and Table II show the mean MAPE and RMSE values for the Linear and Non-Linear imputation models on the holdout values in the MIMIC dataset. The reported values are the mean and the confidence interval.

The reported p-values indicate that the difference in RMSE between the Non-Linear and Linear models is insignificant, with a p-value of 0.2283. Similarly, the results for MAPE show no significant difference in the imputation method for missing data in the laboratory variables from the MIMIC dataset.

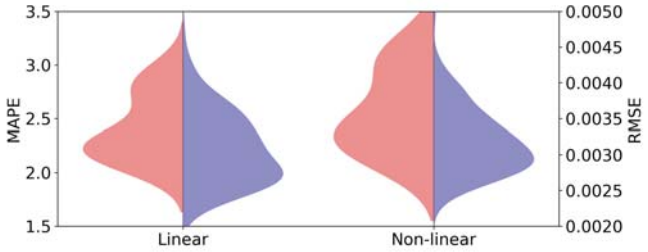

Fig. 2 Performance of Linear, Non-linear models on 50 **holdout** values from **MIMIC** dataset compared; X-axis: model category, Y-axis: RMSE, mean MAPE

Fig. 3 and Table IV present the mean MAPE and RMSE for Linear and Non-Linear imputation models on the EHR-based cohort of stroke patients. The Linear models outperformed the Non-Linear models in terms of both metrics. The MAPE for the Non-Linear and Linear models were 19.03% and 18.04%, respectively. And the RMSE for Non-Linear and Linear

TABLE II
MEAN AND 68%-CI FOR METRICS ON *holdout-MIMIC*

model	$MAPE_h$	$RMSE_h$
Non-Linear	$2.56^{3.06}_{2.19}$	$0.00317^{0.00360}_{0.00275}$
Linear	$2.42^{2.81}_{2.17}$	$0.00310^{0.00361}_{0.00259}$

TABLE III
MEAN AND 68%-CI FOR METRICS ON *Ischemic Stroke* COHORT

model	$MAPE_h$	$RMSE_h$
Non-Linear	$19.03^{20.4}_{17.91}$	$0.01232^{0.01322}_{0.01125}$
Linear	$18.04^{19.89}_{16.11}$	$0.01185^{0.01269}_{0.01105}$

models were 0.01232 and 0.01185, respectively. The p-values suggest that the differences between the models are statistically significant. These results corroborate that Linear models are preferred for imputing missing values of the laboratory variables from the stroke cohort, extracted from real-world EHRs.

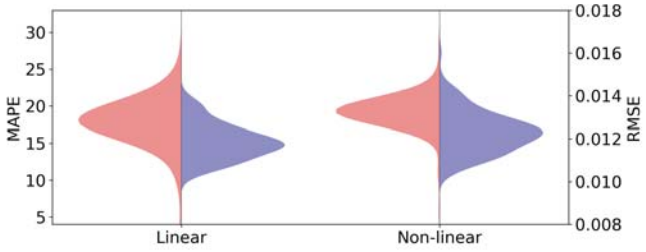

Fig. 3 Performance of Linear and Non-Linear models on 50 **holdout** values, from the **ischemic stroke** cohort; the x-axis is the category of models, and the y-axis shows RMSE and MAPE

IV. DISCUSSION

In this study, we compared the performance of six machine learning-based models used in iterative imputation in terms of Linear and Non-Linear groups. We employed laboratory-based variables from two real-world clinical datasets. The results are aggregated for three Linear and three Non-Linear imputation models for a more robust comparison. Our results are based on 45 different laboratory variables from two datasets representing a total of 8827 ICU patients and 9037 ischemic stroke patients. Our findings demonstrated that the Linear models can outperform the Non-Linear models in terms of MAPE and RMSE.

The improved performance exhibited by Linear models can be attributed to their lower susceptibility to overfitting and their ability to effectively capture linear relationships among variables. Moreover, it is important to acknowledge that these findings align with the principles of the "no free lunch" theorem [35], which states that there is no universally superior model for all tasks, reinforcing the validity of our results. On the other hand, Non-Linear models may have difficulty capturing relevant features in the tabular datasets, leading to overfitting and less optimal performance. Further investigation is needed to compare deep learning models such as GAIN [36] and

decipher when and where more complex imputation techniques can be valuable for laboratory-based variables. Moreover, the absence of a standardized implementation for recently developed methods poses challenges in terms of reproducibility and robustness of findings, despite some indications of their superior performance. The datasets tested in this study are based on laboratory variables from EHRs; however, even common laboratory variables can have high levels of missingness and the missingness may not be completely at random. The mechanism, pattern, and level of missingness are some of the important factors when selecting the optimal imputation techniques.

The results of this study have practical implications for data scientists and practitioners who are interested in using EHR data for understanding care utilization, improving the quality of care, and resource optimization, all of which require the highest quality of real-world data. Our findings suggest that, in certain circumstances, Linear models are a reliable and efficient option for imputing missing data in laboratory variables extracted from clinical datasets. However, it is important to note that the choice of imputation method should depend on the mechanism and pattern of missingness in target datasets and the research question at hand. It is also crucial to assess the mechanism of missing or the reason behind the missingness, which may aid in determining how to design the imputation procedure. In some instances, the inclusion of auxiliary variables has been shown to improve imputation for laboratory-based variables with a high level of missingness [3], [11].

In conclusion, this study contributes to the expanding body of literature on imputation methods for handling missing data. By focusing specifically on laboratory variables derived from two independent and distinct real-world EHR datasets, valuable insights have been gained regarding the performance of Linear and Non-Linear models. The Non-Linear models may prioritize "local complexity," leading to overfitting. This overemphasis on capturing intricate patterns within the data can hinder the generalizability of the model, thereby impacting its performance. Furthermore, the training process of Non-Linear models may be time-consuming, potentially limiting their practicality in large-scale applications. Moreover, when imputing missing values for prediction variables that exhibit a higher linear association with the response variable, the utilization of Linear models can be more efficient. Linear models exhibit greater resilience to the presence of outliers within the data. By demonstrating less sensitivity to outliers, Linear models can avoid unnecessary complexity introduced by Non-Linear models. The latter models, in their attempt to account for outliers, may introduce additional intricacies into the imputation process, resulting in increased uncertainty of imputed values. The findings underscore the need to consider the nature of the variables, the potential for overfitting, computational efficiency, and sensitivity to outliers when selecting an appropriate imputation method. These insights contribute to the existing knowledge and understanding of imputation techniques, further enhancing the applicability and effectiveness of missing data handling in various domains.

As a future direction, we are assessing the performance of other imputation methods, such as multiple imputation and Bayesian imputation, and exploring how imputation techniques are impacted by different missing patterns - e.g., missing not at random (MNAR), missing at random (MAR) - and missing level. We are also working towards the creation of synthetic datasets that can more accurately mimic real-world datasets with higher diversity and broader synthetic patient representation for a more comprehensive model evaluation and systematic assessment of imputation techniques.

This study had also some limitations, including using only a selected number of imputation algorithms and using only two distinct cohorts. One of the cohorts was from the ICU, which represents a very specific type of patients, and one was from ischemic stroke patients, which may not be a representative cohort for other chronic conditions or for younger patients as stroke affects primary the aging population with an average age of 65-75 years old. Finally, this study highlights the importance of careful consideration and evaluation of imputation methods for healthcare applications as this data processing step can lead to biased data for model training.

Appendix A
Description of the Stroke Cohort

Table IV presents a summary of variables for the stroke cohort data. Each row corresponds to a specific laboratory measurement, identified by a unique Lab code. The variables include count, mean, standard deviation (std), minimum (min), maximum (max), and the percentage of missing values for each laboratory measurement.

The Lab column provides the names of the laboratory tests, such as sodium (NA+), potassium (K+), chloride (CL-), total carbon dioxide (TCO2), calcium (CA), blood urea nitrogen (BUN), glucose (GLU), creatinine, hemoglobin, hematocrit, and various other clinical markers.

The count column indicates the number of observations available for each laboratory test, while the mean column represents the average value of the measurements. The std column shows the standard deviation, providing a measure of the dispersion of the data. The min and max columns indicate the minimum and maximum values observed, respectively. Lastly, the Missing (%) column displays the percentage of missing values for each laboratory test.

This table provides a comprehensive overview of the variables in the stroke cohort data, enabling researchers and practitioners to understand the distribution and characteristics of the laboratory measurements.

Appendix B
Variables for MIMIC Data - Laboratory Measurements

Table V provides an overview of various laboratory measurements obtained from the MIMIC dataset. MIMIC is a publicly available database that contains de-identified health records of patients admitted to intensive care units (ICUs). These measurements play a crucial role in assessing patients' health status and monitoring their progress during ICU stays.

The table includes essential information for each laboratory variable, such as the variable code, the name of the measurement (e.g., Potassium, Chloride, Albumin), the count of

TABLE IV
SUMMARY OF LABORATORY VARIABLES FOR THE STROKE COHORT

LOINC	Laboratory name	count	mean value	std	min value	max value	Missing (%)
X2951.2	NA+	5748	139.210000	3.390000	122.000000	152.000000	36
X2823.3	K+	5798	4.260000	0.480000	2.500000	6.600000	36
X2075.0	CL-	5770	101.970000	3.940000	85.000000	118.000000	36
X2028.9	TCO2	5774	26.840000	3.030000	14.000000	40.000000	36
X17861.6	CA	5776	9.270000	0.550000	7.000000	11.600000	36
X3094.0	BUN	5725	19.500000	8.550000	3.000000	54.000000	37
X2345.7	GLU	5569	120.700000	39.970000	46.000000	275.000000	38
X2160.0	CREATININE	5566	1.010000	0.330000	0.100000	2.300000	38
X718.7	HEMOGLOBIN	5489	13.080000	1.940000	6.300000	20.300000	39
X4544.3	HEMATOCRIT	5445	39.050000	5.380000	19.300000	58.800000	40
X786.4	MCHC	5429	33.420000	1.240000	27.600000	37.900000	40
X785.6	MCH	5407	30.290000	2.110000	21.600000	38.800000	40
X6690.2	WBC	5392	8.070000	2.640000	0.650000	18.750000	40
X32623.1	PLATELET MEAN VOL	5413	9.850000	1.400000	5.800000	14.600000	40
X789.8	RBC	5432	4.330000	0.640000	1.910000	6.690000	40
X787.2	MCV	5410	90.560000	5.430000	67.800000	113.500000	40
X777.3	PLATELET COUNT	5371	233.880000	77.080000	5.000000	553.000000	41
X788.0	RDW	5304	13.950000	1.320000	11.000000	19.100000	41
X10466.1	ANION GAP	5364	10.520000	3.220000	0.000000	25.000000	41
X61151.7	ALBUMIN	4798	3.950000	0.490000	2.000000	5.300000	47
X1743.4	ALT	4822	21.580000	10.620000	4.000000	66.000000	47
X2885.2	TP	4728	6.880000	0.640000	4.200000	9.500000	48
X5905.5	MONOCYTE%	4729	8.490000	2.840000	0.500000	19.000000	48
X742.7	MONOTYPE ABS	4691	0.680000	0.270000	0.010000	1.810000	48
X731.0	LYMPHOCYTE ABS	4688	1.720000	0.780000	0.090000	5.140000	48
X706.2	BASEPHIL%	4656	-7.680000	10.010000	-21.000000	0.850000	48
X736.9	LYMPHOCYTE%	4743	22.230000	9.590000	1.000000	69.000000	48
X1975.2	TBIL	4614	0.530000	0.290000	0.100000	1.900000	49
X6768.6	AP	4636	81.780000	27.660000	20.000000	198.000000	49
X30239.8	AST	4650	24.330000	8.540000	6.000000	59.000000	49
X770.8	NEUTRPHIL%	4631	65.690000	11.450000	12.000000	97.000000	49
X711.2	EOSINOPHILS ABS	4401	-0.830000	0.390000	-2.000000	0.750000	51
X713.8	EOSINOPHILS%	4397	0.320000	0.350000	-1.000000	1.530000	51
X704.7	BASEPHILS ABS	4239	-1.450000	0.300000	-2.000000	-0.210000	53
X2093.3	CHOLESTEROL	4191	178.580000	46.360000	58.000000	376.000000	54
X2085.9	HDL	4160	48.090000	14.870000	6.000000	110.000000	54
X2571.8	TRIGLYCERIDES	4068	144.880000	74.670000	24.000000	460.000000	55
X13457.7	LDL	4054	99.850000	38.610000	1.000000	261.000000	55
X751.8	NEUTROPHIL ABS	3938	5.410000	2.350000	0.210000	15.020000	56
X9830.1	CHOLESTEROL:HDL RATIO	4021	3.950000	1.380000	1.400000	9.800000	56
X50560.2	URINE PH	3468	6.010000	0.790000	5.000000	9.000000	62
X3016.3	Thyrotropin	3363	2.240000	1.470000	0.010000	8.590000	63
X5902.2	PT	3002	14.490000	2.300000	9.400000	23.200000	67
X6301.6	INR	2965	1.140000	0.230000	0.740000	2.050000	67
X17856.6	HgA1C	2633	7.070000	1.720000	3.900000	14.200000	71

recorded values, the mean value, standard deviation, minimum and maximum values observed, and the percentage of missing values using the simulated missing pattern. The count represents the number of data points available for each variable, giving an indication of the data's completeness.

The dataset encompasses a wide range of laboratory measurements, covering important aspects of patients' blood chemistry, such as electrolytes (e.g., Potassium, Chloride, Sodium), liver function markers (e.g., Albumin, Bilirubin), renal function markers (e.g., Creatinine, Urea Nitrogen), blood cell counts (e.g., Red Blood Cells, White Blood Cells, Platelet Count), and other parameters relevant to patient health.

Researchers and healthcare professionals can utilize this table to gain insights into the distribution and characteristics of the laboratory measurements in the MIMIC dataset.

TABLE V
SUMMARY OF LABORATORY VARIABLES FOUND IN THE MIMIC DATASET

LOINC	Laboratory name	count	mean value	std	min value	max value	Missing (%)
50822	Potassium, Whole Blood	5628	4.190000	0.630000	1.200000	20.400000	36
50902	Chloride	5650	103.860000	4.280000	81.000000	127.500000	36
50862	Albumin	5627	3.140000	0.630000	1.370000	5.400000	36
50863	Alkaline Phosphatase	5624	124.560000	108.570000	18.670000	2040.750000	36
50971	Potassium	5578	4.160000	0.330000	2.870000	6.330000	37
50912	Creatinine	5601	1.520000	1.270000	0.120000	15.240000	37
50885	Bilirubin, Total	5429	1.460000	3.130000	0.090000	53.680000	38
50882	Bicarbonate	5430	25.240000	3.650000	8.200000	46.900000	38
51250	MCV	5349	89.900000	5.720000	62.800000	124.190000	39
51301	White Blood Cells	5290	11.350000	9.010000	0.400000	404.200000	40
51279	Red Blood Cells	5267	3.500000	0.450000	1.970000	5.860000	40
50983	Sodium	5274	138.750000	3.250000	118.730000	157.780000	40
51491	pH	5270	5.870000	0.700000	5.000000	9.000000	40
51006	Urea Nitrogen	5305	30.050000	18.320000	3.420000	142.880000	40
51248	MCH	5252	30.050000	2.160000	18.940000	41.170000	41
51498	Specific Gravity	5167	1.020000	0.010000	1.000000	1.050000	41
50813	Lactate	5228	2.360000	1.720000	0.500000	24.800000	41
51277	RDW	5232	15.670000	1.870000	11.830000	27.560000	41
50820	pH	4692	7.380000	0.060000	6.900000	7.580000	47
51237	INR(PT)	4676	1.510000	0.570000	0.870000	8.060000	47
51256	Neutrophils	4626	75.560000	11.080000	0.230000	98.000000	48
50910	Creatine Kinase (CK)	4608	584.960000	2686.640000	7.000000	68132.000000	48
51265	Platelet Count	4575	248.960000	111.620000	15.720000	1142.370000	48
51254	Monocytes	4571	5.010000	2.190000	0.000000	43.950000	48
51222	Hemoglobin	4612	10.460000	1.250000	5.560000	17.570000	48
51249	MCHC	4517	33.470000	1.170000	27.170000	37.820000	49
51275	PTT	4518	40.250000	13.480000	18.800000	150.000000	49
50970	Phosphate	4529	3.630000	0.860000	1.400000	10.770000	49
50861	Alanine Aminotransferase (ALT)	4498	100.700000	337.930000	1.000000	7181.330000	49
50804	Calculated Total CO2	4540	25.360000	4.670000	4.000000	57.130000	49
50808	Free Calcium	4292	1.130000	0.080000	0.530000	2.300000	51
50809	Glucose	4289	146.770000	59.010000	23.000000	858.500000	51
50802	Base Excess	4132	-0.480000	4.070000	-25.000000	22.710000	53
50878	Asparate Aminotransferase (AST)	4066	145.270000	607.780000	6.000000	14929.500000	54
50868	Anion Gap	4036	13.920000	2.820000	6.600000	48.000000	54
50893	Calcium, Total	3945	8.450000	0.570000	5.920000	12.350000	55
50818	pCO2	3930	41.400000	8.380000	14.000000	121.000000	55
51274	PT	3843	15.960000	4.420000	9.800000	76.880000	56
51200	Eosinophils	3375	1.760000	1.740000	0.000000	23.910000	62
50960	Magnesium	3270	2.050000	0.220000	1.320000	4.490000	63
50931	Glucose	3217	133.890000	31.790000	51.000000	456.380000	64
51244	Lymphocytes	2881	14.690000	8.610000	0.000000	90.750000	67
51221	Hematocrit	2918	31.040000	3.470000	17.150000	50.930000	67
50821	pO2	2556	145.340000	57.300000	25.000000	529.000000	71
51146	Basophils	2264	0.350000	0.260000	0.000000	2.590000	74

REFERENCES

[1] P. Coorevits, M. Sundgren, G. O. Klein, A. Bahr, B. Claerhout, C. Daniel, M. Dugas, D. Dupont, A. Schmidt, P. Singleton, *et al.*, "Electronic health records: new opportunities for clinical research," *Journal of internal medicine*, vol. 274, no. 6, pp. 547–560, 2013.

[2] S. R. Raman, L. H. Curtis, R. Temple, T. Andersson, J. Ezekowitz, I. Ford, S. James, K. Marsolo, P. Mirhaji, M. Rocca, *et al.*, "Leveraging electronic health records for clinical research," *American heart journal*, vol. 202, pp. 13–19, 2018.

[3] V. Abedi, J. Li, M. K. Shivakumar, V. Avula, D. P. Chaudhary, M. J. Shellenberger, H. S. Khara, Y. Zhang, M. T. M. Lee, D. M. Wolk, *et al.*, "Increasing the density of laboratory measures for machine learning applications," *Journal of Clinical Medicine*, vol. 10, no. 1, p. 103, 2020.

[4] S. Khurshid, C. Reeder, L. X. Harrington, P. Singh, G. Sarma, S. F. Friedman, P. Di Achille, N. Diamant, J. W. Cunningham, A. C. Turner, *et al.*, "Cohort design and natural language processing to reduce bias in electronic health records research," *NPJ Digital Medicine*, vol. 5, no. 1, p. 47, 2022.

[5] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol, "Multimodal biomedical ai," *Nature Medicine*, vol. 28, no. 9, pp. 1773–1784, 2022.

[6] R. S. Vanguri and S. P. Shah, "Multimodal data integration improves immunotherapy response prediction," 2022.

[7] A. S. O'Malley, K. Draper, R. Gourevitch, D. A. Cross, and S. H. Scholle, "Electronic health records and support for primary care teamwork," *Journal of the American Medical Informatics Association*, vol. 22, no. 2, pp. 426–434, 2015.

[8] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records," *Scientific reports*, vol. 6, no. 1, pp. 1–10, 2016.

[9] J. J. Gong, T. Naumann, P. Szolovits, and J. V. Guttag, "Predicting clinical outcomes across changing electronic health record systems," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1497–1505, 2017.

[10] E. Kim, S. M. Rubinstein, K. T. Nead, A. P. Wojcieszynski, P. E. Gabriel, and J. L. Warner, "The evolving use of electronic health records (ehr) for research," in *Seminars in radiation oncology*, vol. 29, pp. 354–361, Elsevier, 2019.

[11] J. Li, X. S. Yan, D. Chaudhary, V. Avula, S. Mudiganti, H. Husby, S. Shahjouei, A. Afshar, W. F. Stewart, M. Yeasin, *et al.*, "Imputation of missing values for electronic health record laboratory data," *NPJ digital medicine*, vol. 4, no. 1, p. 147, 2021.

[12] F. Amrollahi, S. P. Shashikumar, A. L. Holder, and S. Nemati, "Leveraging clinical data across healthcare institutions for continual learning of predictive risk models," *Scientific Reports*, vol. 12, no. 1, p. 8380, 2022.

[13] R. Garriga, J. Mas, S. Abraha, J. Nolan, O. Harrison, G. Tadros, and A. Matic, "Machine learning model to predict mental health crises from electronic health records," *Nature medicine*, vol. 28, no. 6, pp. 1240–1248, 2022.

[14] T. Botsis, G. Hartvigsen, F. Chen, and C. Weng, "Secondary use of ehr: data quality issues and informatics opportunities," *Summit on translational bioinformatics*, vol. 2010, p. 1, 2010.

[15] A. Sharma, R. A. Harrington, M. B. McClellan, M. P. Turakhia, Z. J. Eapen, S. Steinhubl, J. R. Mault, M. D. Majmudar, L. Roessig, K. J. Chandross, *et al.*, "Using digital health technology to better generate evidence and deliver evidence-based care," *Journal of the American College of Cardiology*, vol. 71, no. 23, pp. 2680–2690, 2018.

[16] S. Van Buuren, H. C. Boshuizen, and D. L. Knook, "Multiple imputation of missing blood pressure covariates in survival analysis," *Statistics in medicine*, vol. 18, no. 6, pp. 681–694, 1999.

[17] C. M. Musil, C. B. Warner, P. K. Yobas, and S. L. Jones, "A comparison of imputation techniques for handling missing data," *Western journal of nursing research*, vol. 24, no. 7, pp. 815–829, 2002.

[18] A. Mackinnon, "The use and reporting of multiple imputation in medical research–a review," *Journal of internal medicine*, vol. 268, no. 6, pp. 586–593, 2010.

[19] B. Suthar, H. Patel, and A. Goswami, "A survey: classification of imputation methods in data mining," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 1, pp. 309–12, 2012.

[20] Z. Zhang, "Missing data imputation: focusing on single imputation," *Annals of translational medicine*, vol. 4, no. 1, 2016.

[21] B. K. Beaulieu-Jones, J. H. Moore, and P. R. O.-A. A. C. T. CONSORTIUM, "Missing data imputation in the electronic health record using deeply learned autoencoders," in *Pacific symposium on biocomputing 2017*, pp. 207–218, World Scientific, 2017.

[22] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ digital medicine*, vol. 1, no. 1, p. 18, 2018.

[23] C. Sun, S. Hong, M. Song, and H. Li, "A review of deep learning methods for irregularly sampled medical time series data," *arXiv preprint arXiv:2010.12493*, 2020.

[24] D. Xu, P. J.-H. Hu, T.-S. Huang, X. Fang, and C.-C. Hsu, "A deep learning–based, unsupervised method to impute missing values in electronic health records for improved patient management," *Journal of Biomedical Informatics*, vol. 111, p. 103576, 2020.

[25] Y.-H. Zhou and E. Saghapour, "Imputehr: a visualization tool of imputation for the prediction of biomedical data," *Frontiers in Genetics*, vol. 12, p. 691274, 2021.

[26] Y. Zou, A. Pesaranghader, Z. Song, A. Verma, D. L. Buckeridge, and Y. Li, "Modeling electronic health record data using an end-to-end knowledge-graph-informed topic model," *Scientific Reports*, vol. 12, no. 1, p. 17868, 2022.

[27] K. Psychogyios, L. Ilias, C. Ntanos, and D. Askounis, "Missing value imputation methods for electronic health records," *IEEE Access*, vol. 11, pp. 21562–21574, 2023.

[28] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[29] X. Su, X. Yan, and C.-L. Tsai, "Linear regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 3, pp. 275–294, 2012.

[30] G. C. McDonald, "Ridge regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 1, pp. 93–100, 2009.

[31] J. Ranstam and J. Cook, "Lasso regression," *Journal of British Surgery*, vol. 105, no. 10, pp. 1348–1348, 2018.

[32] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, pp. 197–227, 2016.

[33] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[34] M. Riedmiller and A. Lernen, "Multi layer perceptron," *Machine Learning Lab Special Lecture, University of Freiburg*, pp. 7–24, 2014.

[35] S. P. Adam, S.-A. N. Alexandropoulos, P. M. Pardalos, and M. N. Vrahatis, "No free lunch theorem: A review," *Approximation and Optimization: Algorithms, Complexity and Applications*, pp. 57–82, 2019.

[36] J. Yoon, J. Jordon, and M. Schaar, "Gain: Missing data imputation using generative adversarial nets," in *International conference on machine learning*, pp. 5689–5698, PMLR, 2018.