Evaluation of Missing Data Imputation Methods for Electronic Medical Records Considering Imputation Performance and Applicability

Asami Tada[1], Nesrine Berjab[2], and Hieu Hanh Le[1]

[1] Ochanomizu University, Tokyo, Japan
g2640817@edu.cc.ocha.ac.jp, le@is.ocha.ac.jp
[2] Institute of Science Tokyo, Tokyo, Japan
berjab@c.titech.ac.jp

Abstract. Imputation methods are required for analysis of datasets containing missing values. Although electronic medical records (EMRs) may include missing values, it remains a challenge to identify an effective imputation method from a practical perspective. In EMR analysis, whether a laboratory value is abnormal can be more clinically relevant than its exact numerical value. Furthermore, practical analyses often require methods that can be applied to a large proportion of missing values. This paper evaluates imputation methods from two practical perspectives: imputation performance in predicting abnormality and practical applicability. Compared with linear interpolation alone, combining linear interpolation with last observation carried forward and next observation carried backward resulted in an approximately 24.15% relative increase in applicability. In addition, incorporating time intervals improved recall for abnormality prediction.

Keywords: Missing data imputation methods · Statistical imputation methods · Electronic medical records

1 Introduction

Large volumes of time series data are generated in various contexts. As a result, extensive analysis has been conducted to extract useful knowledge from such big data. However, numerical variables in these datasets may contain missing data. Removal of records with missing data may introduce bias and potentially affect the reliability of analytical results. Therefore, imputation methods are used.

In recent years, the secondary use of EMRs has attracted increasing attention [7]. As an example of EMR analysis, Le et al. [5] proposed a method to understand variant factors in medical-order sequence variants, which are typical sequences of medical orders, by logistic regression analysis using explanatory variables such as patient's age, hospitalization period, treatment history, and laboratory test results. Their study demonstrated that these types of patient information are important for understanding variant factors. In this way, useful knowledge can be extracted from EMRs; however, EMRs may contain missing

Publication Notice. This version of the contribution has been accepted for publication, after peer review (when applicable) but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/DOI (TBD). Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use: https://www.springernature.com/gp/open-science/policies/accepted-manuscript-terms.

data, which can hinder analysis. For example, laboratory test results are important for clinical decision-making, yet in laboratory test results of MIMIC-IV [3, 4, 1], real-world EMR data, approximately 10.14% of data were missing. To conduct analyses using such data, missing data imputation methods are required.

This paper evaluates imputation methods from two perspectives: imputation performance in predicting abnormality and practical applicability. We focus on statistical imputation methods, specifically four basic approaches—last observation carried forward (LOCF), next observation carried backward (NOCB), linear interpolation (LI), and mean imputation—and propose their combinations and methods considering time intervals. Imputation methods are applied to blood test results from MIMIC-IV. We evaluate the applicability, defined as the proportion of missing data to which the imputation method could be applied relative to the total number of missing data, as well as the imputation performance in predicting abnormality determined from the imputed values. Compared with LI alone, combining LI with LOCF and NOCB resulted in an approximately 24.15% relative increase in applicability. In addition, incorporating time intervals improved recall for abnormality prediction.

The main contributions of this study are as follows:

- We evaluate imputation methods from two perspectives: imputation performance in predicting abnormality and applicability.
- We propose a combination strategy to improve applicability and methods considering time intervals to improve imputation performance.
- We validate the evaluated methods using real-world EMR data from MIMIC-IV.

2 Related Work

Imputation methods can be classified into three categories: conventional statistical methods, machine learning methods, and deep learning methods [8]. Many studies comparing statistical, machine learning, and deep learning methods have reported that machine learning and deep learning methods achieve higher imputation performance [6]. However, one of the characteristics of EMRs is their sparsity [7]. Laboratory test results also tend to be sparse because the intervals and types of tests vary among patients. Considering that, a substantial number of missing values may need to be pre-imputed to construct training data, which may prevent machine learning and deep learning methods from achieving sufficient performance. Xu et al. [9] noted that the imputation performance of their proposed method, which is machine-learning-based, may deteriorate when variables with such high missing rates are included as features. Therefore, in this study, a statistical approach was adopted.

Existing studies often evaluate imputation methods using numerical proximity measures such as RMSE [6], and evaluation from the perspectives of imputation performance in predicting abnormality and practical applicability has not been sufficiently conducted. In EMR analysis, it is sometimes not the numerical values themselves but rather the abnormality of those values—important

to clinical decision-making—that is of greater importance. Therefore, this study adopted imputation performance with respect to abnormality as the evaluation metric. In addition, depending on the imputation method, there may be missing values to which the method cannot be applied. For example, Jazayeri et al. [2] proposed an imputation method that takes inter-patient similarity and used LI in combination when enough similar patients were not available in the dataset. In practice, it is necessary to address all missing values; however, the proportion of missing values to which a method can be applied relative to the total number of missing values has not been sufficiently evaluated. Therefore, we also used applicability as an evaluation metric.

3 Imputation Methods and Their Applicability

3.1 Basic Methods

Last Observation Carried Forward (LOCF) For data grouped by patient, the value preceding a missing point is imputed into the missing point.

Next Observation Carried Backward (NOCB) For data grouped by patient, the value following a missing point is imputed into the missing point.

Linear Interpolation (LI) For data grouped by patient, a straight line between the two points before and after a missing point is considered. The missing point is calculated assuming that it lies on the line segment defined by these two points and imputed into the missing point. The missing point and the two surrounding points are assumed to be equally spaced along the line segment.

Mean Imputation In mean imputation, the value to be imputed varies depending on how the mean is calculated. We examined the following three approaches.

Mean Imputation Using Values from All Patients (MIA) The mean value is calculated using data from all patients.

Mean Imputation Using All Values for Each Patient (MIP) For data grouped by patient, the mean is calculated using all values within each patient group.

Mean Imputation Using Only Preceding Values for Each Patient (pre-MIP) For data grouped by patient, the mean is calculated using all values preceding the missing point for each patient.

3.2 Time-Considering Methods

Linear Interpolation Considering Time Intervals (t-LI) For data grouped by patient, a straight line between the two points before and after a missing point, $(time_1, y_1)$ and $(time_2, y_2)$, is considered. The missing point $(time_m, y_m)$ is assumed to lie on the line segment defined by these two points, and the value y_m is calculated accordingly (1) and imputed into the missing point. In this method, the x-coordinates of the two surrounding points and the missing point correspond to their respective time information, $time_1$,$time_2$, and $time_m$, and the value y_m is determined by taking the time intervals into account.

$$y_m = \frac{y_2 - y_1}{time_2 - time_1}(time_m - time_1) + y_1 \tag{1}$$

Time-Weighted Mean Imputation Using Only Preceding Values (t-pre-MIP) The data were grouped by patient. Time-weights were calculated for the observed values preceding the missing point. For the missing point at row m, the time difference between the timestamp of the row including missing point, $time_m$, and that of row $(m-i)$, $time_{m-i}$, was computed. This difference was divided by a parameter t, rounded up to the nearest integer, and the reciprocal was taken to obtain the time-weight w_{m-i} (2). The parameter t was set to $t = 60 \times 60$, and weights were calculated on an hourly basis. Finally, the weighted mean was computed using all available preceding values and their corresponding weights (3). The value at row m is defined as $value_m$, and the value at row $(m-i)$ is defined as $value_{m-i}$. We assume that available values exist up to the $(m-n)$ row.

$$w_{m-i} = \frac{1}{\lceil \frac{time_m - time_{m-i}}{t} \rceil} \tag{2}$$

$$value_m = \frac{\sum_{i=1}^{n}(value_{m-i} \times w_{m-i})}{\sum_{i=1}^{n} w_{m-i}} \tag{3}$$

3.3 Combination of Methods

In this study, LI is applied only when interpolation is feasible. Therefore, this method cannot be applied when missing points occur at the end points of a sequence of data. With practical applicability in mind, we combined LI with LOCF or NOCB to impute missing points for which LI could not be applied. The combinations of methods are as follows: i) LI+LOCF, ii) LI+NOCB, iii) LI + LOCF +NOCB, iv) t-LI+LOCF, v) t-LI+NOCB, vi) t-LI + LOCF +NOCB.

3.4 Applicability Study

Using hypothetical data, Table 1 summarizes the correspondence between the missing data patterns and indicates whether each method can be applied.

Table 1. Applicability of each method

Pattern	Value	Applicability of each method													
		(1)LOCF	(2)NOCB	(3)LI	(4)MIA	(5)MIP	(6)pre-MIP	(7)t-LI	(8)t-pre-MIP	(9)LI +LOCF	(10)LI +NOCB	(11)LI +LOCF+NOCB	(12)t-LI +LOCF	(13)t-LI +NOCB	(14)t-LI +LOCF+NOCB
I	missing missing	-	-	-	✓	-	-	-	-	-	-	-	-	-	-
II	missing 0.36 0.33 0.41 0.43	-	✓	-	✓	✓	-	-	-	-	✓	✓	-	✓	✓
III	0.30 0.32 missing	✓	-	-	✓	✓	✓	-	✓	✓	-	✓	✓	-	✓
IV	0.31 missing 0.30 0.38	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

4 Experimental Evaluation

4.1 Dataset

The experiments used data extracted from the labevents table in MIMIC-IV. The valuenum column, which records laboratory test results, contains missing values. These were imputed in this experiment.

Construction of the Ground Truth and Preprocessing In the flag column, the string "abnormal" is recorded when a laboratory test result indicates an abnormality. We confirmed that although no numerical test result was available and the valuenum column was missing, the comments sometimes contained indications of abnormal findings. Furthermore, it was inferred that some of these cases were associated with comments that corresponded to the string "abnormal" in the flag column. Based on this observation, we considered that, after imputing the values in the valuenum column, the abnormality of the imputed values could be determined by comparing them with the reference range provided in the labevents table and evaluated against the flag column. Therefore, the flag column was used as the ground truth.

As preprocessing related to the construction of the ground truth, the following two procedures were performed.

1. In some cases, the contents of the comments column suggested that the result indicated an abnormality, although nothing was recorded in the flag column. In such cases, the string "abnormal" was added to the flag column.
2. Because abnormality could not be determined from the contents of the comments column in certain cases, rows in which the valuenum column was missing and the comments column was either NULL or completely anonymized were excluded from the dataset.

Laboratory Test Items Used 21 test items were initially extracted from the labevents table, and preprocessing was applied. As a result of the preprocessing, some test items could not be used for the experiment. For example, those in

Table 2. Characteristics of the Dataset

Item name	Number of data	Number of missing values	Rate of abnormal in missing values
Albumin	1,032,387	51	100.00%
Platelet count	4,199,386	1,499	100.00%
Creatinine	4,334,840	938	100.00%
Urea nitrogen	4,203,472	1,544	100.00%
Uric acid	192,219	98	100.00%
Alkaline phosphatase	1,602,488	33	100.00%
Potassium	4,481,959	384	100.00%
C-reactive protein	177,842	3,573	94.26%
Glucose	3,903,505	860	100.00%
Creatine kinase	334,663	72	100.00%

which all rows with missing values in the valuenum column were excluded. After excluding such items, 10 laboratory test items were used in the experiments. Characteristics of the dataset are shown in Table 2.

4.2 Evaluation Method

Applicability As summarized in Table 1, the applicability of each method differs depending on the pattern of missingness. We defined applicability as the proportion of missing values to which the imputation method could be applied relative to the total number of missing values. For each laboratory test item, applicability was calculated.

Imputation Performance In EMR analysis, whether a laboratory value is abnormal can be clinically relevant. Precision and recall with respect to abnormality determined from the imputed values were calculated.

4.3 Experimental Results

Applicability, precision, and recall were calculated for each laboratory test item, and their distributions are summarized in Fig. 1 and 2. Regarding (4)MIA in Fig. 2(a), both true positives and false positives were zero for albumin, platelet count, uric acid, and potassium, so precision could not be calculated. Therefore, we present the results for the six test items, excluding these four items.

Regarding applicability, methods that combined both LOCF and NOCB when LI could not be applied—such as (11)LI + LOCF + NOCB and (14)t-LI + LOCF + NOCB—achieved a relatively high applicability. Compared with using (3)LI or (7)t-LI alone, both (11)LI + LOCF + NOCB and (14)t-LI + LOCF + NOCB improved the applicability ratio by approximately 24.15%. In practice, the ability to handle a larger number of missing values increases the amount of data that can be used for analysis. Therefore, improving the applicability rate is of practical significance.

As shown in Fig. 2(a), the overall precision was approximately 1.00 on average. In the dataset, almost all the missing values were considered abnormal data; when the imputed values were predicted as abnormal, they were always correct. Therefore, we discuss imputation performance using recall shown in Fig. 2(b).

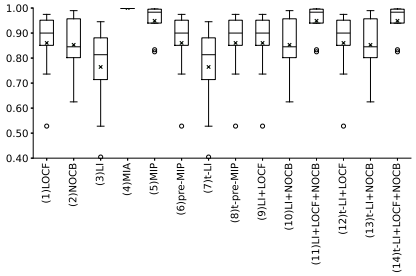
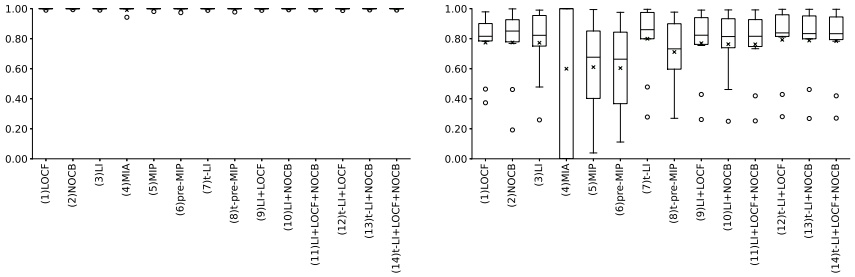


Fig. 1. Applicability Result



(a) Precision (b) Recall

Fig. 2. Precision and Recall Results

The recall of (1)LOCF, (2)NOCB, and (3)LI was approximately 0.77 on average, while that of (7)t-LI was around 0.80, which was relatively high. Conversely, no clear differences were observed among combinations of methods. Examination of the actual data suggested that abnormality did not change substantially before and after missing values in many cases. Since (1)LOCF, (2)NOCB, and (3)LI impute missing values using adjacent values, combining these methods tends to produce similar imputations, which may explain the lack of notable differences among method combinations. Comparing the results of (3),(9),(10),(11)LI with those of (7),(12),(13),(14)t-LI, as well as the results of (6)pre-MIP and (8)t-pre-MIP, it is apparent that recall increased when time intervals were considered: by approximately 3.12% for LI and by about 17.77% for mean imputation methods. In the dataset, the time intervals between tests were not constant. As a characteristic of EMRs, it is common for the intervals between tests to be uneven. The increase in recall can be attributed to taking time intervals into account.

5 Conclusion

We evaluate imputation methods from two perspectives: imputation performance in predicting abnormality and practical applicability. We focus on statistical imputation methods and propose their combinations and methods considering time

intervals. Compared with linear interpolation alone, combining linear interpolation with last observation carried forward and next observation carried backward increased applicability by approximately 24.15%. In addition, incorporating time intervals improved recall for abnormality prediction. In future work, we will explore combinations of methods that incorporate correlations among laboratory tests or similarities between patients. In addition, a limitation of the dataset is that the abnormality rate at the missing values was nearly 100.00%. To address this limitation, we will artificially mask observed values to construct an evaluation dataset containing ground-truth values for both normal and abnormal cases.

Acknowledgments. ChatGPT and Copilot were used to support English translation and grammar improvement of the manuscript.

References

1. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000)
2. Jazayeri, A., Liang, O.S., Yang, C.C.: Imputation of missing data in electronic health records based on patients' similarities. Journal of Healthcare Informatics Research **4**, 295–307 (2020)
3. Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L.A., Mark, R.: MIMIC-IV (version 3.1). PhysioNet (2024)
4. Johnson, A.E.W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., Lehman, L.H., Celi, L.A., Mark, R.G.: MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data **10**, 1 (2023)
5. Le, H.H., Yamada, T., Honda, Y., Sakamoto, T., Matsuo, R., Yamazaki, T., Araki, K., Yokota, H.: Methods for analyzing medical-order sequence variants in sequential pattern mining for electronic medical record systems. ACM Transactions on Computing for Healthcare **4**(1), 3:1–28 (2023)
6. Ren, W., Liu, Z., Wu, Y., Zhang, Z., Hong, S., Liu, H., on behalf of the Missing Data in Electronic health Records (MINDER) Group: Moving beyond medical statistics: A systematic review on missing data handling in electronic health records. Health Data Science **4**, 0176 (2024)
7. Sarwar, T., Seifollahi, S., Chan, J., Zhang, X., Aksakalli, V., Hudson, I., Verspoor, K., Cavedon, L.: The secondary use of electronic health records for data mining: Data characteristics and challenges. ACM Computing Surveys **55**(2), 33 (2022)
8. Sun, Y., Li, J., Xu, Y., Zhang, T., Wang, X.: Deep learning versus conventional methods for missing data imputation: A review and comparative study. Expert Systems with Applications **227**, 120201 (2023)
9. Xu, X., Liu, X., Kang, Y., Xu, X., Wang, J., Sun, Y., Chen, Q., Jia, X., Ma, X., Meng, X., Li, X., Xie, G.: A multi-directional approach for missing value estimation in multivariate time series clinical data. Journal of Healthcare Informatics Research **4**, 365–382 (2020)