Informatics in Medicine Unlocked 17 (2019) 100275

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: http://www.elsevier.com/locate/imu

MICE vs PPCA: Missing data imputation in healthcare

Harshad Hegde, Neel Shimpi, Aloksagar Panny, Ingrid Glurich, Pamela Christie, Amit Acharya *

Center for Oral and Systemic Health, Marshfield Clinic Research Institute, Marshfield, WI, USA

ARTICLE INFO

Keywords:
Imputation
Probabilistic principal component analysis
PPCA
Multiple imputations using chained equations
MICE
Medical dental data
Dental informatics

ABSTRACT

Retrospective analyses of real-world clinical data face challenges owing to the absence of some data elements. Historically, missing data was addressed by first classifying its presence into one of three categories: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Imputation techniques continue to be developed and tested to gauge their capacity to mitigate the negative impact of missing data types on analyses and their results. This study undertook a comparison of two techniques of data imputation: probabilistic principal component analysis (PPCA) and multiple imputation using chained equations (MICE).

Retrospective data from 41,543 unique patients including both medical and dental variables ($n = 116$) were mined from the institutional research data warehouse, which captures data through an integrated medical and dental electronic health record (iEHR). A subset with complete data on all variables of interest was sampled. “Missing data” were artificially created by randomly removing data elements to create the missing data problem. Applying PPCA and MICE, the capacity of the two techniques to create an accurate imputed dataset was tested. Comparisons were drawn between imputed dataset and sampled subset, to investigate which technique more closely simulated the true data.

PPCA outperformed MICE with an overall correct imputation percentage (accuracy) and root mean square error (RMSE) of approximately 65% and 0.29, respectively, compared to MICE, which yielded approximately 38% accuracy with a RMSE of 0.83.

Overall, this study concluded that PPCA demonstrated higher capacity to impute MCAR data than MICE.

1. Introduction

Secondary use of clinically captured data when used for research often presents the missing data challenge leading to the potential of introducing bias or negative impact on analytical outcomes [1–4]. Reasons for the existence of missing data are myriad. Examples might include 1) failure of staff to consistently document a value under the structured data element in an electronic health record, or 2) technical failures precluding data capture by a device designed to track specific data, or 3) capture of data mainly in unstructured formats not readily mined electronically without manual abstraction or preprocessing, making these data not readily available for analysis. Data not consistently recorded lead to missing data and thus limit the analysis and systematic interpretability surrounding relative impact of specific data elements [5]. Identifying the type of missing data is crucial to find solutions to address them. Missing data has been categorized into three different types namely: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) [1,2,6,7]. MCAR is defined as data whose absence occurs independent of the observed and/or unobserved values [1,2]. For example, the body weight of an individual was not recorded because the scale used for weight capture was temporarily decommissioned. Missing data decrease the sample size of the available study population and subsequently reduce statistical power, albeit without introduction of bias [8]. By contrast, MAR occurs when the reason for the missing data point can be deduced by the observed values only and are not related to unobserved values. For example, highly educated people are less likely to reveal their salary in surveys if it is in the higher range. A data point could be MNAR if there was an explanation for it to be missing.

One way of handling missing data is simply omitting that portion of the data from further analysis [9]. But excluding the data variable with missing value is not a good technique of handling missing data [10]. Deleting rows of data due to the absence of a few variables leads to a loss of valuable information observed that would have been beneficial for

* Corresponding author. Marshfield Clinic Research Institute, Research Scientist, Center for Oral and Systemic Health, 1000 North Oak Avenue, Marshfield, WI, 54449. USA.

E-mail address: acharya.amit@marshfieldresearch.org (A. Acharya).

https://doi.org/10.1016/j.imu.2019.100275
Received 13 September 2019; Received in revised form 21 November 2019; Accepted 23 November 2019
Available online 26 November 2019

analyses. This could lead to biased estimates. Missing data can also be supplemented by findings from clinical notes using Natural Language Processing techniques (NLP) [11]. Alternatively, trying to predict missing values by a close estimation based on data context is called data imputation [2,6]. Techniques have been developed to impute missing data elements. An example includes 'mean value imputation' (MVI) where the missing values for a variable are filled by calculating and substituting the mean for a missing value [2,6]. Another single-value technique involves carrying forward the last observation [6].This technique was used for studies that treated data collected from individuals over a span of time [6]. There is a possibility that these techniques may result in biased results and hence they are deemed as suboptimal [6,12]. Moreover, these techniques do not directly use information obtained from the observed values [6]. Additional techniques continue to be developed, and some have proven to be better at imputing data. These include: Maximum Likelihood (ML) based methods, Hot Deck Imputation and Multiple Imputation (MI) [6]. MI is recommended over single-value techniques to handle missing data [9,12]. MI handles missing data by gradual replacement after several iterations [1,2,6]. It employs statistical analyses on known information along with handling the uncertainty caused by the presence of missing data to generate an estimate [9]. 'Multiple Imputation Using Chained Equations' (MICE), is an example of a MI technique [13] that was adopted in our study. We compared MICE to an ML based technique called Probabilistic Principal Component Analysis (PPCA) which employs an Expectation-Maximization (EM) algorithm to estimate values of missing data points [4,14]. PPCA is a derivation of Principal Component Analysis (PCA), which is used for dimensionality reduction. PPCA estimates missing values when original data is recovered from its dimension-reduced form [14].

Data imputation using MICE in healthcare have been reported in various studies [15–19] in the past. However, data imputation using PPCA is more prevalent in studies working with traffic and transportation data [4,20–22]. The objective of our work was to compare the efficiency of data imputation by MICE and PPCA, beginning with a complete baseline dataset including medical and dental variables, and then simulating missing data and its imputation. Data for this study was derived from a larger study that involved systematic capture and modeling of data to achieve development of a predictive model estimating a patient's relative risk of being diagnosed with diabetes mellitus (DM) [23].

2. Material and methods

2.1. Retrieval of baseline data

Baseline data ($D_{baseline}$) from 1979 to 2018 was retrieved retrospectively from the Marshfield Clinic Health System (MCHS) enterprise data warehouse. The study was reviewed by the Marshfield Clinic Research Institute's Institutional Review Board (IRB) and was classified as "exempt" under 45 CFR 46.101(b) (2). The integrated Electronic Health Record (*iEHR*) [24] at the MCHS facilitated retrieval of both medical and dental variables of 41,543 patients. A systematic review was conducted by the research team members to obtain the list of candidate medical and dental variables that could contribute to building a prediction model for assessing dysglycemic risk among patients in dental settings with undiagnosed dysglycemia. The dataset consisted of 116 variables including 18 demographic, 12 medical and 86 dental variables of patients diagnosed with DM (cases = 5,286) (based on ICD-9/10 diagnostic codes) compared to observations surrounding these variables among patients with no diabetes (controls = 36,257). Fig. 1 shows the data imputation process.

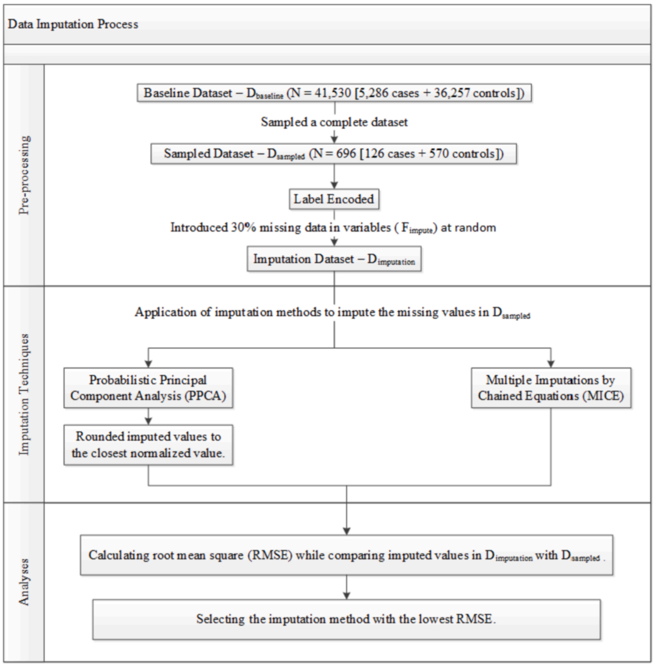

Data Imputation Process
Pre-processing
Baseline Dataset – D_baseline (N = 41,530 [5,286 cases + 36,257 controls])
Sampled a complete dataset
Sampled Dataset – D_sampled (N = 696 [126 cases + 570 controls])
Label Encoded
Introduced 30% missing data in variables (F_impute) at random
Imputation Dataset – D_imputation
Imputation Techniques
Application of imputation methods to impute the missing values in D_sampled
Probabilistic Principal Component Analysis (PPCA)
Rounded imputed values to the closest normalized value.
Multiple Imputations by Chained Equations (MICE)
Analyses
Calculating root mean square (RMSE) while comparing imputed values in D_imputation with D_sampled .
Selecting the imputation method with the lowest RMSE.

Fig. 1. Data imputation process.

The average proportion of missing data amongst all features combined in $D_{baseline}$ was around 30%. Probability of missing data was not influenced by any observed or unobserved variables and hence could be categorized as MCAR. Out of the 116 variables, 46 variables were always collected in the iEHR and hence had no missing data ($F_{complete}$) while the remaining 70 variables had missing values (F_{impute}). Table 1 shows the variable list along with types and imputation status and percentage of missing records per variable.

2.2. Sampling of baseline data

$D_{baseline}$ was sampled such that there were no missing data points across all of the 116 variables. The sampled data subset ($D_{sampled}$) consisted of 696 patients (126 cases and 570 controls) with 116 variables.

2.3. Label encoding and feature scaling for categorical variables

$D_{sampled}$ included 87 categorical variables which were converted to numerical values and 29 continuous variables as shown in Table 1. In order to avoid biased weight distribution of variables while developing the prediction model, feature scaling was performed, wherein categorical variables were scaled between the range of 0 and 1 using the following formula.

$$n = \frac{\text{Original value} - \min(\text{possible value for the variable})}{\max(\text{possible value for the variable}) - \min(\text{possible value for the variable}))} \tag{1}$$

Table 1
Variable list along with types and imputation status and percentage of missing records per variable.

No.	Variable	Label	Possible values	Imputation Status	Percentage of missing records per variable
1	Age[a]	1	21–30 years	Not-imputed	0%
		2	31–40 years		
		3	41–50 years		
		4	51–60 years		
		5	61–70 years		
		6	71–80 years		
		7	80–89 years		
2	BMI[a] [c]	1	Underweight: < 18.5	Imputed	0.93%
		2	Normal: 18.5–24.99		
		3	Overweight: 25.0–29.99		
		4	Obese: > 30.0		
3–30	Bleeding on probing[a] Each tooth was probed at six sites. Bleeding aligned to the corresponding tooth surface that showed the deepest PPD among the six probing surfaces of a tooth (except third molars).	1	Present	Imputed	55.22%
		2	Absent		
31	Prescription of corticosteroids medications[a]	1	Yes	Imputed	0.007%
		2	No		
32	Creatinine levels[a] [c]	1	Low: Females: < 0.6 mg/dl; Males: < 0.7 mg/dl	Imputed	15.63%
		2	Normal: Females:0.6–1.1 mg/dl; Males:0.7–1.3 mg/dl		
		3	High: Females:1.1 mg/dl <; Males:1.3 mg/dl <		
33	Prescription of diabetic medications[a]	1	Yes	Imputed	0.007%
		2	No		
34	Ethnicity[a] [d]	1	Declined	Imputed	1.23%
		2	Hispanic or Latino		
		3	Not Hispanic or Latino		
		4	Patient Does Not Know		
35	Family history of diabetes[a]	1	Yes	Not-imputed	0%
		2	No		
36	Gender[a]	1	Male	Not-imputed	0%
		2	Female		
37	HDL cholesterol[a] [c] (High density lipids levels)	1	Poor: <40 mg/dl	Imputed	29.55%
		2	Better: 40–59 mg/dl		
		3	Best: 60 mg/dl ≤		
38	Hypertension[a] [c]	1	Normal: <120 mm Hg and <80 mm Hg	Imputed	12.83%
		2	Prehypertension: 120–129 mm Hg and <80 mm Hg		
		3	Stage 1 hypertension: 130–139 mm Hg or 80–89 mm Hg		
		4	Stage 2 hypertension: ≥140 mm Hg or ≥90 mm Hg		
		5	Hypertensive crisis: ≥140 mm Hg or ≥120 mm Hg		
39	Hypertensive medications[a]	1	Hypertensive medication prescribed	Imputed	0.007%
		2	Hypertensive medication not prescribed		
40–45	Health Insurance[a]	1	Present	Not-imputed	0%
		2	Absent		
46	LDL cholesterol[a] [c] (Low density lipids levels)	1	Optimal: < 100 mg/dl	Imputed	31.08%
		2	Near optimal: 100 mg/dl to 129 mg/dl		
		3	Borderline high: 130 mg/dl to 159 mg/dl		
		4	High: 160 mg/dl to 189 mg/dl		
		5	Very High: > 190 mg/dl		
47	Number of dental visits + (Total number of unique dental visits of a patient in the given measurement year)	N/A	Minimum value = 1 Maximum value = 24	Not-imputed	0%
48	Periodontal disease (PD)[a] [c]	1	Healthy	Imputed	27.5%
		2	Type 1		
		3	Type 2		
		4	Type 3		
		5	Type 4		
		6	Type 5		
49–76	Periodontal Pocket Depth (PPD) [b] Each tooth is probed at six sites and maximum PPD value is assigned as the PPD for each tooth.	N/A	Minimum vale = 0 Maximum value = 16	Imputed	58.28%
77–84	Race[a] (American Indian or Alaska Native; Asian; Black or African American; Native Hawaiian or Other Pacific Islander; White; Patient Does Not Know; Declined and Unknown)	1	Race (Yes)	Not-imputed	0%
		2	Race (No)		

(continued on next page)

Table 1 (*continued*)

No.	Variable	Label	Possible values	Imputation Status	Percentage of missing records per variable
85	Prescription for statins (Use of Statins)[a]	1	Yes	Imputed	0.007%
		2	No		
86	Tobacco use status[a] [d] (History of tobacco use)	1	Current	Imputed	25.39%
		2	Former		
		3	Never		
		4	Missing		
87–114	Tooth present/absent[a]	1	Tooth present	Not-imputed	0%
		2	Tooth absent		
115	Total Triglycerides levels[a] [c]	1	Normal: < 150 mg/dl	Imputed	29.72%
		2	Borderline high: 150 mg/dl to 199 mg/dl		
		3	High: 200 mg/dl to 499 mg/dl		
		4	Very High: > 500 mg/dl		
116	WBC[a] [c] (White blood cell count)	1	Leukopenia: Less than $4.0*10^9$/L	Imputed	25.08%
		2	Normal: $4.0*10^9$/L to $11.0*10^9$/L		
		3	Leukocytosis: More than $11.0*10^9$/L		

[a] Categorical.
[b] Continuous.
[c] Ordinal Variable.
[d] Nominal Variable.

where n stands for normalized value [25,26]. The max and min values for the data variables were obtained from our data warehouse.

2.4. *Introducing missing data in F_{impute}*

We introduced 30% missing data into F_{impute} (as shown in Fig. 1) using the MCAR mechanism to mimic the missing data pattern of $D_{baseline}$. The missing data was imposed. This formed the imputation dataset $D_{imputation}$.

2.5. *Imputation methods*

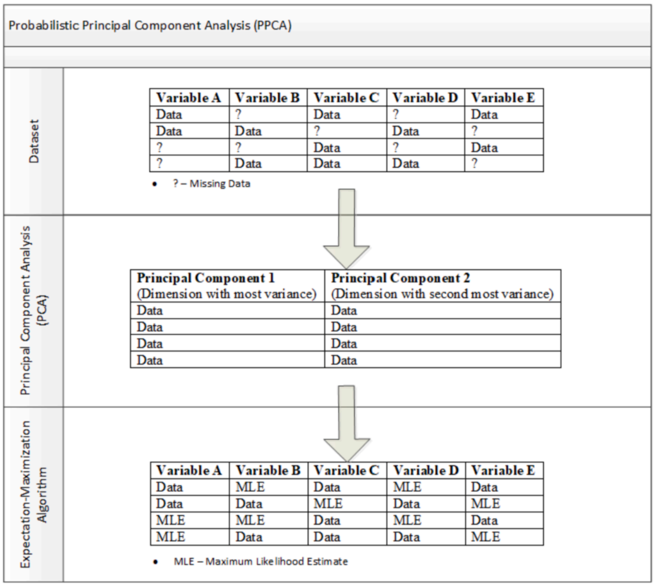
We employed PPCA and MICE techniques to impute the missing data points in $D_{imputation}$ and compare the imputed values with the original values in $D_{sampled}$. We used open sourced R packages (pcaMethods for PPCA and mice for MICE) to implement these imputation methods [27].

2.5.1. *Probabilistic principal component analysis (PPCA)*

PPCA employs the Expectation–Maximization (EM) algorithm to iteratively calculate the Maximum Likelihood Estimates (MLE) of an incomplete data set [14]. Principal Component Analysis (PCA) is a method used for dimensionality reduction first described by Pearson [28] and developed by Hotelling [29]. PCA reduces the data dimensionality by linear presentation of the original data variables in a lower dimensional space [30]. It is regarded as one of the most reliable techniques for dimensionality reduction as it minimizes the reconstruction error/loss on variance during the data compression by reducing the Euclidean distance between the original data points and the projected/estimated data points (latent data points) [31].

This property of PCA can be utilized for imputing the missing data points by first estimating the distribution of the compressed information based on the non-missing data and then reconstructing the missing data from the compressed information as estimated/projected data points [32]. Several PCA algorithms for handling missing data were proposed [33,34] which differ in the assumption on the relationships between the original data points and the latent data points. MLE is one of the many methods to estimate the missing/unknown parameters of the data [35]. Each iteration of the EM algorithm during the data imputation process consists of two steps including: the expectation (E) step and the maximization (M) step [36]. The expectation step imputes all the missing values from the observed values in the dataset [36,37]. The maximization step then updates the imputed parameters using the MLE utilizing the complete dataset (Observed values + imputed values) created in the expectation step [36,37]. This iterative process consisting of both "E&M" steps continues until no further improvements are possible in the likelihood estimates [37]. This process results in imputed datasets with MLE suggesting higher accuracy of imputed values. Fig. 2 illustrates how PCA is used to reduce the dimensionality of a multivariate dataset by 'compressing' it into a lower dimensional space. In the next steps the EM algorithm is applied to the compressed dataset to impute missing data points of the multivariate dataset using MLE. Thus, the combination of PCA and EM constitutes PPCA which imputes missing data using MLE as shown in Fig. 2. We used an R function (RPackage: 'pcaMethods', v1.64.0) [38,39] to implement PPCA.

2.5.1.1. Dummy variable generation. All nominal variables (from Table 1) were converted into dummy variables before running through the imputation process. The process of dummy variable generation was illustrated in a study by Gujarati et al. [40].


Probabilistic Principal Component Analysis (PPCA)
Dataset
Principal Component Analysis (PCA)
Expectation-Maximization Algorithm

Variable A	Variable B	Variable C	Variable D	Variable E
Data	?	Data	?	Data
Data	Data	?	Data	?
?	?	Data	?	Data
?	Data	Data	Data	?

- ? – Missing Data

Principal Component 1 (Dimension with most variance)	Principal Component 2 (Dimension with second most variance)
Data	Data
Data	Data
Data	Data
Data	Data

Variable A	Variable B	Variable C	Variable D	Variable E
Data	MLE	Data	MLE	Data
Data	Data	MLE	Data	MLE
MLE	MLE	Data	MLE	Data
MLE	Data	Data	Data	MLE

- MLE – Maximum Likelihood Estimate

Fig. 2. Data imputation process using PPCA.

2.5.1.2. Rounding to normalized values. Using R, we first prepared lists of all variables along with the normalized values. The imputed values for categorical variables were then rounded to the closest normalized value for the corresponding variable. For example, the variable "blood glucose level" had three categories/values and were denoted as 1(low), 2 (normal) and 3(high). These values were feature scaled to 0.3333, 0.6666, and 1.000, respectively, in D_{sample}. An imputed value of 0.68 in $D_{imputation}$ was rounded to 0.6666, since it was the closest normalized value in D_{sample}. This process was applied to define all imputed values in the dataset $D_{imputation}$.

2.5.2. Multiple imputation using chained equations (MICE)

In this technique of imputation, numerous regression models are run in such a way that the variable with missing data is modeled depending on other variables in the dataset [13]. Each variable is modeled taking into account the type of the variable. For example, logistic regression is used to model binary variables, whereas predictive mean matching is used for continuous variables [13]. Per Melissa et al., the chained equation process is broken down into four core steps which are repeated until optimal results are achieved [13]. The first step involves replacing every missing data with the mean of the observed values for the variable, which acts as a placeholder. The second step involves setting these mean imputations back to 'missing'. In the third step, the observed values of a variable (e.g., 'x') are regressed on the other variables such that 'x' is the dependent variable and the rest are independent variables. In the current study, we used logistic regression for binary variables (2 levels), polytomous logistic regression for nominal categorical variables (>2 levels), proportional odds model for ordinal categorical variables (>2 levels) and predictive mean matching for the continuous variables. The fourth step involves replacing the missing values with the predictions derived from the regression model. This imputed value would then be part of the independent variables along with observed values for other variables. Steps '2' to '4' are then repeated for each variable that has missing values, constituting one 'iteration'. After one iteration, all missing values are replaced by the regressed predictions related to the observed data. The imputed values are replaced after every iteration, and the number of iterations may vary [13]. In the present study, we investigated results for 10, 20, and 30 iterations.

The multiple iterations ideally result in convergence of the regression coefficients. This constitutes one 'imputation' [13]. Several imputations are performed by keeping the observed values of all variables constant, with only the missing values changing to their respective imputation prediction. This results in the formation of several imputed datasets depending on the number of imputations (n = 30, in this study). The number of imputations depend on the missing values [13]. We chose 30 imputations since the proportion of missing data was roughly 30%, and was further based on a prior publication by White et al. [41]. We used the R package 'MICE', v 3.3.0 [42] for imputing missing values. Normalization of the imputed values is not required for MICE as it imputes the data variable-wise.

2.6. Comparing imputed values ($D_{imputation}$) with original values (D_{sample})

We calculated the percentage of values that were correctly imputed for each variable and across $D_{imputation}$. We also calculated the root mean square error (RMSE) using R function 'RMSE' (RPackage: 'pcaMethods', v1.64.0) [39] to compare the performances between PPCA and MICE.

Table 2
Comparison of imputation technique performances.

Measure	PPCA	MICE (10 iterations)	MICE (20 iterations)	MICE (30 iterations)
Correct Imputation	64.51% ± 0.26%	37.82% ± 0.273%	37.80% ± 0.27%	38.11% ± 0.28%
RMSE	0.29	0.83	0.83	0.83

3. Results

The results of the comparison that executed MICE with 30 imputations (combinations of 10, 20 and 30 iterations) and PPCA is shown in Table 2. Approximately 65% of data variables were correctly imputed by PPCA and 38% by MICE. In terms of RMSE, PPCA outperformed all MICE iterations with the lowest value of 0.29. Table 2 shows the comparison of imputation technique performances.

Fig. 3 illustrates the variables that were imputed most accurately by the four imputation techniques along with mean and standard deviation of all variables combined.

Table 3 shows the imputation accuracies for all features. For PPCA, the imputation accuracy for hypertensive status was the lowest (14.96%) while for MICE, the lowest imputation accuracy were diabetic medications for MICE 10 iterations (7.41%) and 30 iterations (7.69%) whereas BOP (8.61 ± 2.5%) for MICE 20 iterations. .

4. Discussion

This study demonstrated the accuracies and RMSE of missing data imputation done by PPCA and MICE. Overall, PPCA outperformed MICE in terms of handling of missing data for medical and dental variables. MICE generally assumes the underlying data to be MAR [43] whereas in this study, introducing missing data via MCAR may have potentially contributed to the underperformance of MICE. Notably, Ambler et al. compared MICE imputation performed over MAR and MCAR data and both approaches yielded a similar RMSE [18]. Similarly, Baneshi et al. also applied MICE imputation to MCAR data and reported better results following imputation compared to complete case analysis [19]. Future studies may involve investigating methodologies for introducing missing data via MAR before using MICE for imputation. PPCA can be performed on data that is both MAR and MCAR [22].

Multiple approaches for imputing data have been explored in various studies. For example, deep learning methodologies have been employed for vehicle transportation data by Duan et al. [44]. Studies have reported RMSE, which indicates the sample standard deviation of the difference between imputed and original values [45]. RMSE reported in the current

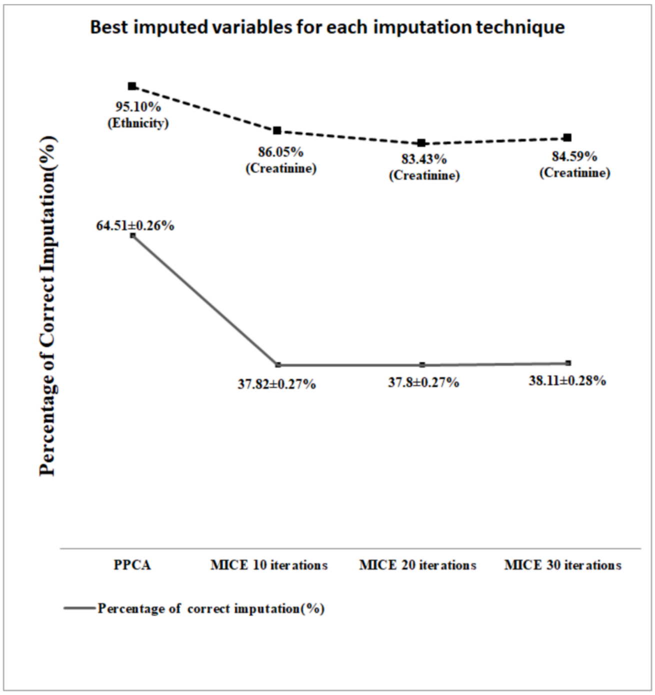


Fig. 3. Best imputed variables for each imputation technique along with mean and standard deviation of accuracy of all variables combined.

Table 3
Imputation accuracies by variables.

Variable No. based on Table 1	Variable	Missing values introduced	PPCA	MICE (30-10)	MICE (30-20)	MICE (30-30)
2	BMI	50.86%	29.80%	45.19%	45.48%	45.77%
3–30	Bleeding on probing (BOP)	50.29 ± 0.02%	84.92 ± 0.03%	8.35 ± 0.025%	8.61 ± 0.025%	8.45 ± 0.024%
31	Prescription of corticosteroids medications	50.58%	39.77%	23.82%	23.53%	22.94%
32	Creatinine levels	51.15%	91.38%	86.05%	83.43%	84.59%
33	Prescription of diabetic medication	49.57%	18.56%	7.41%	9.40%	7.69%
34	Ethnicity	51.15%	95.10 ± 0.05%	80.85%	80.28%	84.23%
37	HDL Cholesterol	50.43%	52.17%	39.88%	45.16%	37.83%
38	Hypertension	50.00%	14.96%	29.91%	34.31%	35.19%
39	Hypertensive medication	50.72%	34.67%	17.89%	17.07%	17.07%
46	LDL Cholesterol	54.17%	36.66%	35.88%	37.85%	33.33%
48	Periodontal Disease (PD) Types	51.00%	65.27%	46.13%	47.28%	48.71%
49–76	Periodontal Pocket Depth (PPD)	50.18 ± 0.02%	48.08 ± 0.05%	63.97% ± 0.033%	63.59% ± 0.04%	64.51 ± 0.044%
85	Prescription of statin medication	53.16%	17.75%	9.62%	11.81%	10.99%
86	Tobacco use status	50.58%	65.48 ± 0.12%	39.46%	31.89%	36.76%
115	Total triglyceride levels	51.58%	49.71%	53.71%	56.08%	59.94%
116	WBC (White blood cell count)	51.29%	90.13%	77.46%	79.15%	78.03%

study (0.05–0.96 for PPCA) as compared to that reported by Schmitt et al. using Random Forest under MICE for data imputation (RMSE = 0.05 to 0.45) [45]. In their study that compared imputation techniques for handling missing predictor values in risk models with binary outcomes, Ambler et al. reported an average RMSE of 0.66 for MICE [18] in comparison to 0.83 in this study. Similarly, in a non-medically-related context, Qu et al. reported RMSE for PPCA to be around 0.8 for a missing ratio of 0.3 [22] which is comparable to 0.29 for the same missing ratio in the current study. In our approach, we only compared the original values to the imputed values and calculated the RMSE to show the performance of PPCA and MICE while studies have reported different techniques for imputation method assessment [46–48].

Few studies have defined imputation techniques for dental variables. Pahel et al. [49] employed a zero-inflated Poisson (ZIP) regression model to impute missing dental caries data. A study by White et al. [50] used a monotone multiple imputation technique for imputing missing data for dental pain after third molar extractions. This study used PPCA and MICE on a combination of medical and dental data to impute missing data values with an accuracy of about 65%. To the best of our knowledge, no other study has performed data imputation on a combined medical-dental dataset using MICE or PPCA. In this study, 30% of the overall dataset was replaced with ‘blanks’ to simulate missing data. As shown in Table 3, approximately 50% of data were missing for each variable. Despite this high proportion of data missing, PPCA correctly imputed 31 out of 70 variables (44.28%) with an accuracy of greater than 80%. MICE on the other hand correctly imputed only two variables (Creatinine and Ethnicity) with an accuracy of greater than 80%. Studies have shown that MICE performs better when missing rates of variables are between 2.5 and 30% [51,52]. However in the present study, all the variables had around 50% missing data and thus may have contributed to the poor performance of MICE.

MI is recognized as an efficient approach to addressing missing data in clinical and epidemiological research. Several studies have reported on the importance of applying an adequate number of imputations to avoid a large Monte Carlo error [53–55] A review of 99 articles using MI by Mackinnon stated that less than half of the articles reported on the number of imputations performed [56]. Stating the number of imputations applied can help readers make a better judgment about the imputation methods and is important to defining reproducibility of results. We have thus provided information on number of imputations and iterations performed during the implementation of MICE and also have tabulated the percentage of correctly imputed data of each variable for various numbers of iterations (i.e. 10, 20, 30 iterations with 30 imputations as shown in Fig. 3).

Despite its popularity, not many clinical and epidemiological studies have reported details on implementation of imputation methods [57]. A systemic review of 103 articles on MI by Hayati et al. stated that only one third of studies had mentioned the use of imputation methods, and often these details were not explicit [57]. Further, the authors noted that only one third of the articles clearly specified the variables used in the imputation process [57]. Providing detailed information about what variables were used and how the imputation was carried out is important for reproducibility of results. In line with previously published guidelines regarding the importance of documenting the imputation methods [56,58], we have provided detailed information regarding data preprocessing, imputation techniques used, and methods to validate the imputation models we employed. Notably we have tabulated details of both the medical and dental variables used in our imputation process in Tables 1 and 2 Further we have presented detailed percentage of correctly imputed values for each of the variables using various imputation techniques in Table 3.

The study acknowledges a few limitations. The data used for the study was focused on factors that were associated with type 2 diabetes mellitus and hence lacks generalization to other variables associated with other health conditions. The data used belonged to a single healthcare system having an integrated electronic medical dental record. Translation of these techniques to external datasets needs further validation. The dataset $D_{sampled}$ (n = 696) generated represented complete data points that were retrieved from the data warehouse. Since simulation of missing data was MCAR, this dataset could not be assumed to be representative of the missing data in $D_{baseline}$. There was no true way to determine the exact reasons for the presence of missing data; hence MCAR was the methodology preferred.

5. Conclusion

Our study explored imputation techniques that could be used for a combined medical-dental dataset to efficiently predict missing values. Outcomes of our analyses identified PPCA to be a more efficient method to impute data as compared to MICE. Furthermore, this technique holds potential for extension to other healthcare-related research studies which consist of integrated medical-dental datasets.

Ethical statement

Authors are in compliance with ethical authorship guidelines. All authors of this manuscript have directly participated in the planning, development, and writing of the article. All authors affirm that the final manuscript has been seen and approved by them for submission to the *Informatics in Medicine Unlocked* journal. This article has not been published in another publication of any type, nor is it under consideration by any other journal. Neither will it be submitted elsewhere unless and until it is declared unacceptable for publication by the journal.

Declaration of competing interest

The authors declare no real or potential conflicts of interest.

Acknowledgement

This study was supported in part by funds from Delta Dental of Wisconsin, Family Health Center of Marshfield Inc., and Marshfield Clinic Research Institute.

References

[1] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338. https://doi.org/10.1136/BMJ.B2393. b2393.

[2] Li P, Stuart EA, Allison DB. Multiple imputation. J Am Med Assoc 2015;314:1966. https://doi.org/10.1001/jama.2015.15281.

[3] Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using mice: a caliber study. Am J Epidemiol 2014;179:764–74. https://doi.org/10.1093/aje/kwt312.

[4] Ke J, Zhang S, Yang H, Chen X. PCA-based missing information imputation for real-time crash likelihood prediction under imbalanced data. 2018.

[5] Li T, Hutfless S, Scharfstein DO, Daniels MJ, Hogan JW, Little RJA, et al. Standards should be applied in the prevention and handling of missing data for patient-centered outcomes research: a systematic review and expert consensus. J Clin Epidemiol 2014;67:15–32. https://doi.org/10.1016/j.jclinepi.2013.08.013.

[6] Newgard CD, Lewis RJ. Missing data. J Am Med Assoc 2015;314:940. https://doi.org/10.1001/jama.2015.10516.

[7] Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. Annu Rev Public Health 2000;21:121–45. https://doi.org/10.1146/annurev.publhealth.21.1.121.

[8] Mack C, Su Z, Westreich D. Managing missing data in patient registries. Agency for Healthcare Research and Quality (US); 2018.

[9] Manly CA, Wells RS. Reporting the use of multiple imputation for missing data in higher education research. Res High Educ 2015;56:397–409. https://doi.org/10.1007/s11162-014-9344-9.

[10] Masconi KL, Matsha TE, Erasmus RT, Kengne AP. Effects of different missing data imputation techniques on the performance of undiagnosed diabetes risk prediction models in a mixed-ancestry population of South Africa. PLoS One 2015;10:e0139210. https://doi.org/10.1371/journal.pone.0139210.

[11] Hegde H, Shimpi N, Glurich I, Acharya A. Tobacco use status from clinical notes using Natural Language Processing and rule based algorithm. Technol Health Care 2018;1–12. https://doi.org/10.3233/THC-171127.

[12] Eekhout I, de Vet HCW, Twisk JWR, Brand JPL, de Boer MR, Heymans MW. Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. J Clin Epidemiol 2014;67:335–42. https://doi.org/10.1016/J.JCLINEPI.2013.09.009.

[13] Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple Imputation by Chained Equations: what is it and how does it work?. n.d, https://doi.org/10.1002/mpr.329.

[14] Tipping ME, Bishop CM. Probabilistic principal component analysis. J R Stat Soc Ser B Stat Methodol 1999;61:611–22. https://doi.org/10.1111/1467-9868.00196.

[15] Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med 2010;50:105–15. https://doi.org/10.1016/j.artmed.2010.05.002.

[16] Jolani S, Debray TPA, Koffijberg H, van Buuren S, Moons KGM. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. Stat Med 2015;34:1841–63. https://doi.org/10.1002/sim.6451.

[17] Chowdhury MH, Islam MK, Khan SI. Imputation of missing healthcare data. In: 20th Int. Conf. Comput. Inf. Technol. IEEE; 2017. p. 1–6. https://doi.org/10.1109/ICCITECHN.2017.8281805. 2017.

[18] Ambler G, Omar RZ, Royston P. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. Stat Methods Med Res 2007;16:277–98. https://doi.org/10.1177/0962280206074466.

[19] Baneshi MR, Talei AR. Does the missing data imputation method affect the composition and performance of prognostic models? Iran Red Crescent Med J 2012;14:31–6.

[20] (Michael) Ke J, Zhang S, Yang H, Chen X. PCA-based missing information imputation for real-time crash likelihood prediction under imbalanced data. Transp A Transp Sci 2019;15:872–95. https://doi.org/10.1080/23249935.2018.1542414.

[21] Li L, Li Y, Li Z. Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. Transp Res C Emerg Technol 2013;34:108–20. https://doi.org/10.1016/J.TRC.2013.05.008.

[22] Qu Li, Hu Jianming, Li Li, Zhang Yi. PPCA-based missing data imputation for traffic flow volume: a systematical approach. IEEE Trans Intell Transp Syst 2009;10:512–22. https://doi.org/10.1109/TITS.2009.2026312.

[23] Hegde H, Shimpi N, Panny A, Glurich I, Christie P, Acharya A. Development of non-invasive diabetes risk prediction models as decision support tools designed for application in the dental clinical environment. Informatics Med Unlocked 2019. https://doi.org/10.1016/J.IMU.2019.100254. 100254.

[24] Shimpi N, Glurich I, Acharya A. Integrated care case study: Marshfield clinic health system. 2019. p. 315–26. https://doi.org/10.1007/978-3-319-98298-4_17.

[25] Aksoy S, Haralick RM. Feature normalization and likelihood-based similarity measures for image retrieval. Pattern Recognit Lett 2001;22:563–82. https://doi.org/10.1016/S0167-8655(00)00112-4.

[26] Kumar Jain Y, Kumar Bhandare S. Min max normalization based data perturbation method for privacy protection, vol. VIII; 2011.

[27] R Core Team. R: a language and environment for statistical computing. 2018.

[28] Pearson K. On lines and plan. London, Edinburgh. Dublin Philos Mag J Sci 1901;2:559–72. https://doi.org/10.1080/14786440109462720.

[29] Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psychol 1933;24:417–41. https://doi.org/10.1037/h0071325.

[30] Kambhatla N, Leen TK. Dimension reduction by local principal component analysis. Neural Comput 1997;9:1493–516. https://doi.org/10.1162/neco.1997.9.7.1493.

[31] Mosci S, Rosasco L, Verri A. Dimensionality reduction and generalization. 2007. Corvallis.

[32] Ke J, Zhang S, Yang H, Chen X. PCA-based missing information imputation for real-time crash likelihood prediction under imbalanced data. 2018.

[33] Kiers HAL. Weighted least squares fitting using ordinary least squares algorithms. Psychometrika 1997;62:251–66. https://doi.org/10.1007/BF02295279.

[34] Grung B, Manne R. Missing values in principal component analysis. Chemometr Intell Lab Syst 1998;42:125–39. https://doi.org/10.1016/S0169-7439(98)00031-8.

[35] Anderson TW. Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. J Am Stat Assoc 1957;52:200–3. https://doi.org/10.1080/01621459.1957.10501379.

[36] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 1977;39:1–38. https://doi.org/10.2307/2984875.

[37] Josse J, Pagès J, Husson F. Multiple imputation in principal component analysis. Adv Data Anal Classif 2011;5:231–46. https://doi.org/10.1007/s11634-011-0086-7.

[38] Stacklies Wolfram, Redestig H, Wright K. A collection of PCA methods. 2018.

[39] Redestig Henning. pcaMethods package | R Documentation. n.d.;1.640, https://www.rdocumentation.org/packages/pcaMethods/versions/1.64.0. accessed August 23, 2018.

[40] Gujarati D. Use of dummy variables in testing for equality between sets of coefficients in two linear regressions: a note. Am Stat 1970;24:50. https://doi.org/10.2307/2682300.

[41] White IR, Royston P, Wood AM, Simoneau G. Multiple imputation using chained equations: issues and guidance for practice. n.d, https://doi.org/10.1002/sim.4067/full.

[42] Buuren S van, Groothuis-Oudshoorn K. Multivariate imputation by chained equations in R. J Stat Softw 2011;45:1–67. https://doi.org/10.18637/jss.v045.i03.

[43] Mongin D, Lauper K, Turesson C, Hetland ML, Klami Kristianslund E, Kvien TK, et al. Imputing missing data of function and disease activity in rheumatoid arthritis registers: what is the best technique? RMD Open 2019;5:e000994. https://doi.org/10.1136/rmdopen-2019-000994.

[44] Duan Y, Lv Y, Liu YL, Wang FY. An efficient realization of deep learning for traffic data imputation. Transp Res C Emerg Technol 2016. https://doi.org/10.1016/j.trc.2016.09.015.

[45] M, J el, M G Mandel JSP. A comparison of six methods for missing data imputation. J Biometrics Biostat 2015;06. https://doi.org/10.4172/2155-6180.1000224.

[46] Gelman A, Van Mechelen I, Verbeke G, Heitjan DF, Meulders M. Multiple imputation for model checking: completed-data plots with missing and latent data. Biometrics 2005;61:74–85. https://doi.org/10.1111/j.0006-341X.2005.031010.x.

[47] Nguyen CD, Carlin JB, Lee KJ. Diagnosing problems with imputation models using the Kolmogorov-Smirnov test: a simulation study. BMC Med Res Methodol 2013;13:144. https://doi.org/10.1186/1471-2288-13-144.

[48] He Y, Zaslavsky AM. Diagnosing imputation models by applying target analyses to posterior replicates of completed data. Stat Med 2012;31:1–18. https://doi.org/10.1002/sim.4413.

[49] Pahel BT, Preisser JS, Stearns SC, Rozier RG. Multiple imputation of dental caries data using a zero-inflated Poisson regression model. J Public Health Dent 2011;71:71–8. https://doi.org/10.1111/j.1752-7325.2010.00197.x.

[50] White IR, Daniel R, Royston P. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. Comput Stat Data Anal 2010;54:2267–75. https://doi.org/10.1016/J.CSDA.2010.04.005.

[51] Haji-Maghsoudi S, Haghdoost A-A, Rastegari A, Baneshi MR. Influence of pattern of missing data on performance of imputation methods: an example using national data on drug injection in prisons. Int J Health Policy Manag 2013;1:69–77. https://doi.org/10.15171/ijhpm.2013.11.

[52] Knol MJ, Janssen KJM, Donders ART, Egberts ACG, Heerdink ER, Grobbee DE, et al. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. J Clin Epidemiol 2010;63:728–36. https://doi.org/10.1016/j.jclinepi.2009.08.028.

[53] van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 2007;16:219–42. https://doi.org/10.1177/0962280206074463.

[54] White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med 2011;30:377–99. https://doi.org/10.1002/sim.4067.

[55] Newton EHJ, Cox NJ, Baum C, College B, Bellocco R, Institutet K, et al. The stata journal. [n.d].

[56] Mackinnon A. The use and reporting of multiple imputation in medical research - a review. J Intern Med 2010;268:586–93. https://doi.org/10.1111/j.1365-2796.2010.02274.x.

[57] Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. BMC Med Res Methodol 2015;15:30. https://doi.org/10.1186/s12874-015-0022-1.

[58] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338. https://doi.org/10.1136/bmj.b2393. b2393.