EasyChair Preprint

№ 12591

Handling Missing Data in Longitudinal Anthropometric Data Using Multiple Imputation Method

Dhruv Varma, Chittaranjan S Yajnik, Aniket Thorave and Neha Sharma

March 19, 2024

Handling Missing Data in Longitudinal Anthropometric Data using Multiple Imputation Method

Dhruv Varma[1*], Chittaranjan S Yajnik[2], Aniket Thorave[3], Neha Sharma[4]
[1]Dwarkadas J. Sanghvi College of Engineering, Mumbai
[2]Diabetes Unit, KEM Hospital & Research Centre, Pune
[3,4]Data and Analytics Practice, Tata Consultancy Services
*Corresponding author
[1]dhruvvarma@36gmail,com, [2]csyajnik@gmail.com, [3]aniketthorave19@gmail.com,
[4]nvsharma@rediffmail.com

Abstract

Diabetes mellitus, a prevalent and an ever-growing metabolic problem, is a widespread global health challenge. Type 2 diabetes is traditionally attributed to genetic factors and unhealthy lifestyle that can lead to obesity. Recent research has shown intrauterine fetal programing as an additional risk factor. To investigate fetal programing of diabetes in Indians, Pune Maternal Nutrition Study (PMNS) was set up by the Diabetes Unit of KEM Hospital, Pune in 1993 in six villages near Pune. The objective was to investigate determinants of fetal growth and study the lifecourse evolution of phenotype of diabetes. The children born in the study and their parents have been serially followed-up for their growth, development and cardiometabolic risk factors. A large dataset over 5000 variables is created over 30 years, including demographics, anthropometry, socioeconomic status, nutrition intake, cardiometabolic risk factors, etc. investigation of the dataset revealed a substantial number of missing values which would create an impedance in performing analytics. Hence it was decided to impute the missing value and prepare the data for analysis in the first phase of the project. It was decided to focus initially on only 177 columns pertaining to anthropometry. To impute the missing values in the longitudinal dataset, popular algorithms like K-Nearest Neighbors and Multiple Imputation by Chained Equation (MICE) were applied. The imputed values generated from both the algorithm were compared and it was found that MICE excelled in maintaining the temporal coherence of the dataset.

In conclusion, this imputation exercise underscores the paramount significance of preserving temporal consistency in longitudinal research, specifically when reporting long-term health outcomes.

Keywords: K-Nearest Neighbors (KNN), Multiple Imputation by Chained Equations (MICE), Diabetes, Anthropometric Data, Missing data, Data Imputation

1. INTRODUCTION

Diabetes is one of the commonest chronic metabolic disorders that mainly occur due to defective insulin action on the tissues or defective insulin secretion by the pancreas, leading to raised blood

glucose levels [1]. Diabetes is primarily classified in two types - type 1 and type 2. Type 1 diabetes is an autoimmune condition which is less prevalent but with more devastating effects. Type 2 diabetes is the most common and accounts for the majority of diabetes cases worldwide [2]. According to the estimates of the International Diabetes Federation (IDF), ~ 537 million adults in the age of 20-79 had diabetes worldwide in 2021 [1], and is projected that this number will reach 700 million by 2045 [1]. India has one of the largest number of people with diabetes as estimated by INDIAB 2023 [2]. Uncontrolled diabetes may cause widespread complications in the body like kidney disease, retinopathy, neuropathy, cardiovascular disease, foot ulcers, etc. These complications can be prevented by good management of blood glucose levels and other risk factors [3]. Nevertheless, the cost associated with diabetes management, treatment, and their complications is substantial, and poses a significant economic burden on healthcare systems and individuals [4-8].

The conventional model of type 2 diabetes envisages predisposition by genetic factors and precipitation of metabolic disorders by 'unhealthy' lifestyle contributing to obesity. In last few decades, an additional predisposing factor has been described, called intrauterine 'fetal programming'. Fetal programming depends on epigenetic modification in the genome of the growing fetus due to a number of 'environmental' factors including: maternal under-nutrition or over-nutrition, diabetes, maternal stress, a number of environmental pollutants, etc. India is also known for its large number of small babies due to undernutrition of young mothers. In addition, a condition called gestational diabetes is also increasing in young Indian mothers. Thus, fetal under-nutrition and over-nutrition are common in India.

Technology has enhanced diagnostics, patient care, administrative tasks and medical research in recent times. One of the key technologies with transformative applications across various domains is machine learning, which is a subset of Artificial Intelligence (AI). The objective of AI and machine learning is to enable machine to perform like intelligent human beings. This can be achieved by developing algorithms and building models that can learn, predict or make decision based on data [9-11]. The application of machine learning is tremendous in all the verticals and healthcare is no exception. Machine learning is also contributing towards advancing diabetes management by enabling more accurate predictions, personalized treatment plans, and efficient monitoring of patients. Besides, it has been increasingly utilized to analyze complex datasets, including those related to fetal programming and type 2 diabetes risk.

The purpose of this research is to use machine learning to derive insights from the pre-conceptional observational birth cohort dataset and to compare it with the results of the traditional research. To commence the machine learning journey, first ingredient required is data that comes from a thirty-year journey of research under PMNS. This data qualifies as longitudinal data as it is obtained through a series of observations of the same entities over some extended time frame and rendering it valuable for measuring change. The original dataset spans a daunting 5800 columns and 800 data points regarding anthropometric measurements, socio-economic status and a number of biological variables. The dataset contains complex information which is collected as a manual process. Certain amount of data is missing, which may affect analysis. The authors recognize the enormous potential in the dataset and are committed to

participate in the on-going fetal programming research by utilizing machine learning tools and techniques. Data integrity, data quality and data completeness are paramount in medical research, and only when this challenge posed by missing or ambiguous data points is overcome, it is possible to uncover valuable information in the data and reveal the insights that will aid the physicians and policy makers in decision making. The process of estimating missing values based on available data plays an important role in the analysis. Null or obscure data points must be properly labeled. In this endeavour, the choices of techniques adopted to fill in the missing values will largely determine the outcome of our research and subsequent results.

This paper consists of the subsequent sections: Section 2 concentrates on related research work carried out by various researchers across the globe. Section 3 emphasizes on the dataset and algorithms to be used in the current research. Section 4 presents the experimental results and section 5 summarizes the research and concludes the paper. Finally, the last section lists out all the references cited in the paper.

2. LITERATURE REVIEW

The literature review section is a critical component of any academic or research paper. It serves several important purposes, contributing significantly to the overall quality and credibility of the work by contextualizing the study, identifying of gaps and forming the research questions. Given the stress of the current analysis on imputation of missing values for machine learning approach on a longitudinal large dataset, we have restricted literature review to this area and not on the science of fetal programming.

Machine Learning has proved to be a powerful tool in diabetes management, for it offers innovative solutions to improve patient care, early detection, and treatment. Various studies have explored the use of ML algorithms for predicting the onset of diabetes. For instance, Li et al. build a predictive model based on genetic factors and lifestyle using support vector machines (SVM) to identify individuals at high risk for diabetes (type 2) [12]. Rajpurkar, P., et al. build machine learning models that are used to analyze medical imaging data, such as retinal images, to detect diabetic retinopathy early [13]. Jian et al predicted the development of type 2 diabetes and the risk of complication using machine learning algorithms to analyze patient data [14]. Artificial pancreas systems which are also known as closed-loop control systems, have appeared as an encouraging approach for the management of diabetes. The study by Anderson et al. (2020) implemented a closed-loop control system using reinforcement learning, demonstrating improved glycemic control in individuals with type 1 diabetes [15]. ML algorithms have been employed to develop personalized treatment plans for individuals with diabetes. In their study, Smith et al. (2019) used machine learning to analyze patient data and recommend personalized insulin dosages, resulting in improved glycemic control [16].

In fields such as epidemiology, psychology, and the social sciences, managing missing data is a crucial part of data analysis. Researchers have devised statistical methods and strategies to tackle this issue. Handling missing data in longitudinal studies is crucial for reliable data analysis and informed decision-making in various fields [17]. Multiple Imputation (MI) methods

were explored by Huque et al. and others and offered a robust approach to imputing missing data in longitudinal studies [18]. These methods provided less biased estimates and better coverage for statistical models [19]. Ibrahim et al.'s research compares 14 different methods for imputing missing data in a longitudinal cohort. This comparative analysis sheds light on the effectiveness of various imputation techniques in specific longitudinal contexts [20]. MICE (Multivariate Imputation by Chained Equations) imputation is a powerful modern method to predict missing values. It involves building models and taking the mean of predictions, making it valuable for researchers and data scientists dealing with datasets containing missing values [21,22].

Objective: The research paper embarks on a mission to determine the most appropriate imputation method for our refined dataset, one that has been meticulously curated to encapsulate over three decades of research. The primary objective is to select a robust and effective strategy for handling missing data of a longitudinal medical study centered on a pre-conceptional observational birth cohort.

3. MATERIALS AND METHODS

This section provides a detailed account of how the study was conducted. This section is useful for reproducibility, transparency, validity and reliability, ethical considerations, experimental design, instruments and materials used, sampling methods, data collection procedures, statistical analyses.

3.1 Data Source

Diabetes Unit of the KEM Hospital, Pune is at the forefront of investigating fetal programming of diabetes, in collaboration with Prof David Barker and his team in Southampton, UK. They set up the Pune Maternal Nutrition Study (PMNS), which is a pre-conceptional observational birth cohort started in 1993 [23] in six villages near Pune. More than 800 pregnancies were studied in these rural women between 1993-96. Detailed measurements of socioeconomic status, maternal nutrition, physical activity, and metabolism were made, 2-3 times during pregnancy. Fetal growth was recorded by ultrasound and father's body size and metabolic measurements were also made. Anthropometric measurements were made in the newborn and serially over next two decades. Periodically, the parents and the offspring underwent detailed measurements of growth, development, and a range of cardiometabolic factors. Follow-up rates have been more than 90%. A huge data has been collected and stored on servers. A biobank of blood, urine, and other biological samples has been archived. The entire process adopted during PMNS along with the time-line is demonstrated in Figure 1.

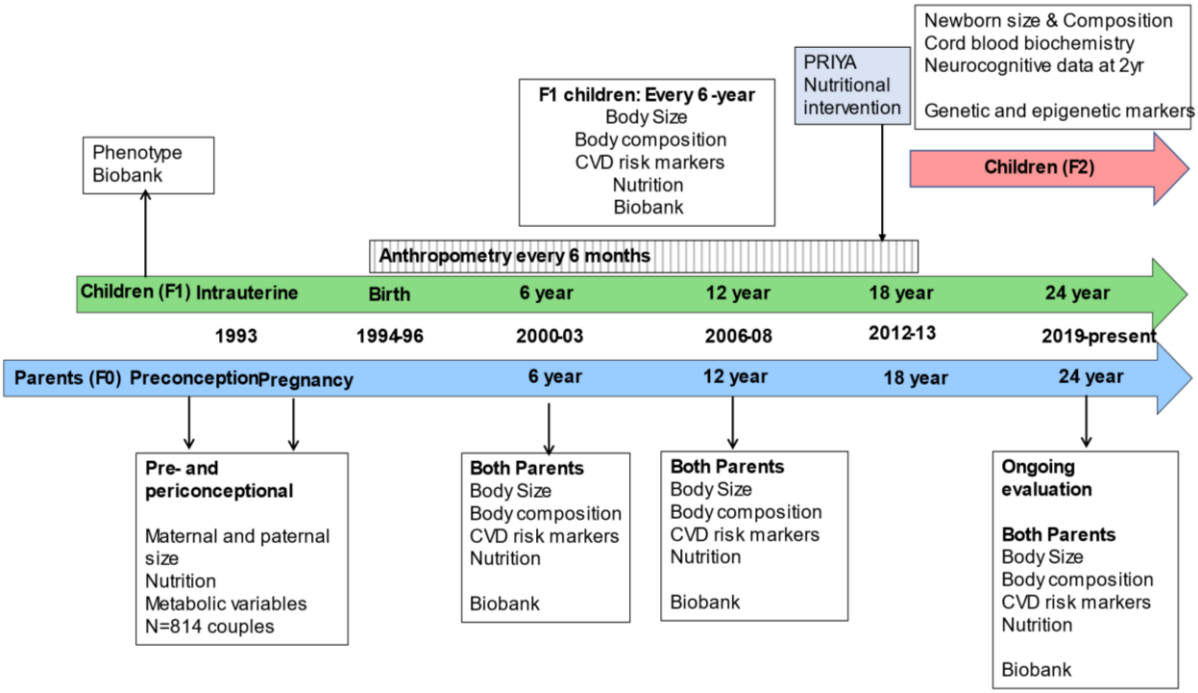


Figure 1. Pune Maternal Nutrition Study Process

The PMNS data provides rich information on: 1) determinants of fetal growth in India, 2) lifecourse evolution of cardiometabolic risk (including diabetes) in a Developmental Origins of Health and Disease (DOHaD) model, 3) birth of third generation allows a transgenerational rather than intergenerational model, and 4) findings in the mothers in the cohort led to a preconceptional 'primordial' micronutrient intervention, probably the first in India. A number of papers have been published using conventional statistics to test various hypothesis which initiated the research. It is increasingly recognized that modern data science may help significantly transform the understanding of various biological, social, economic, and other environmental issues which affect health of the population.

The initial dataset encompasses overwhelming 5800 features for 800 patients that can be broadly classified under anthropometric measurements, socio-economic status, obstetrics history, medication, nutrition intake and other vitals. However, as a first phase of the project, the focus would be to process the anthropometric data, which are around 177 columns. Table 1 below presents the snapshot of a few of the columns pertaining to the anthropometric data from the dataset and also the explanation of the columns which comes from the data dictionary.

Table 1. First 10 variables of the dataset along with their purpose

Variable	Variable information
mas3_id_no	Identity Number
f0_m_age_eld_child	Age (yrs) of eldest Childs Of F0 Gen Mothers
f0_m_age	Mothers Derived Age - F0 Gen

f0_socio_eco_sc	Socioeconomic score - Trivedi Pareek (F0 Gen)
f0_dt_prepreg_visit	Mothers Date of prepregnancy visit - F0 Gen
f0_m_ht	Mothers Height (cm) - F0 Gen
f0_m_wt_prepreg	Mothers Weight (kg) at prepregnancy visit - F0 Gen
f0_m_bmi_prepreg	Mothers BMI (kg/m^2) at Prepregnancy Visit - F0 Gen
f0_m_tr_prepreg	Mothers Triceps (cm) at prepregnancy visit - F0 Gen

3.2 Imputation Methods

The techniques used to handle missing data by estimating the absent values based on the available / observed data, is known as imputation methods. Missing or absent data is a prevalent problem with most of the datasets. Imputation helps analysts and researchers maintain the statistical power of the analyses. The commonly used imputation methods are Mean / Median imputation, Mode Imputation, Regression Imputation, Multiple Imputation, Hot-Deck Imputation, Cold-Deck Imputation, K-Nearest Neighbors (KNN) Imputation, Expectation-Maximization (EM) Imputation, and many more. The choice of imputation method depends on the nature of the data, the missing data mechanism, and the assumptions that can reasonably be made about the missing data. Each method has its strengths and limitations, and the appropriateness of the method should be carefully considered in the context of the specific dataset and research question. For the present research, to tackle the problem of missing data, thorough testing was conducted on two well-known imputation methods: K-Nearest Neighbors (KNN) and Multiple Imputation by Chained Equations (MICE).

3.2.1 K-Nearest Neighbors (KNN) Imputation

To address missing data in the longitudinal dataset, the K-Nearest Neighbors (KNN) imputation method was used. This technique estimates missing values by examining the values of their nearest neighbors within the dataset, offering a practical and effective way to fill in missing data [24]. The choice to use KNN imputation was mainly driven by the fact that a significant portion of the dataset consisted of categorical columns. In order to make these categorical variables suitable for analysis, one-hot encoding was performed to convert them into numeric variables. Subsequently, KNN imputation was applied to fill in missing values, which allowed to gain a better understanding of the data's behavior and correlations.

Systematic steps followed to carry out KNN imputation algorithm are as follows:

Initial requirements:

a. PMNS data containing missing values (Longitudinal dataset)
b. Decide the number of nearest neighbors to consider
c. Decide the metric used to calculate distances between data points (e.g., Euclidean distance).

Step 1: Preprocess to normalize or scale the data to ensure that features are on a similar scale (if needed).
Step 2: Identify the locations of missing values in the dataset.
Step 3: Repeat the following steps for all missing values:

a. Iterative imputation process

For each row with missing values:
Calculate the distances between this row and all other rows in the dataset.
Identify the k-nearest neighbors with the smallest distances.
Impute the missing values by aggregating the values from the k-nearest neighbors.

b. Calculate distance between two data points by using the chosen distance metric (for example, Euclidean distance)
c. Imputation Strategy
    i. Impute numeric missing values by using the mean / median of the corresponding feature values from the k-nearest neighbors.
    ii. Impute categorical missing values by using the mode (most frequent category) from the k-nearest neighbors.

Final Output: The dataset with missing values imputed using the KNN algorithm.

3.2.2 Multiple Imputation by Chained Equations (MICE)

Three different mechanisms for generating missing data are defined: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [21]. These mechanisms can be distinguished by the definition of the outcome variable Y, which consists of both the observed data (Yobserved) and the missing data (Ymissing), as well as the missing indicator R (1 for observed and 0 for missing). Each missing data mechanism suggests a specific approach for imputation. If the probability of missingness is not related to either the observed or unobserved data, then the data are considered MCAR. According to this mechanism, the distribution of missing data is represented by $P(R|Y) = P(R)$. While this assumption may not hold in practice, the listwise deletion method is an unbiased approach for handling missing data when this assumption is met. On the other hand, multiple imputation (MI) approaches are developed under the MAR assumption, which states that the probability of missingness is related to the observed data and not dependent on the unobserved data. In this case, the distribution of missing data is defined as $P(R|Y) = P(R|Yobserved)$, and MI methods can generate unbiased and efficient results. Recognizing that the dataset exhibited missing at random (MAR), the selection of Multiple Imputation by Chained Equations (MICE) as another imputation method was done.

MICE is sophisticated imputation method used to handle missing data by iteratively filling in missing values to create multiple datasets that are complete. These completed datasets are then analyzed individually, and the results from each one of them are combined to deliver more precise and robust estimates.

Steps followed are as follows:

Initial requirements:

a. PMNS data containing missing values (Longitudinal dataset)
b. Decide the desired number of imputed datasets to be generated
c. For each attribute/variable with missing data, decide on the set of imputation models

Step 1: Initialize imputed datasets to represent different imputation by creating n copies of the original dataset
Step 2: Identify the locations of missing values in the dataset.

Step 3: Repeat the following steps for all n imputations

a. Iterative imputation process

For each imputation (i = 1 to n):

For each variable with missing values (j = 1 to p, where p is the number of variables with missing data):

Replace missing values in variable j using imputation model Mj based on observed values from other variables.

Repeat until convergence or a specified number of iterations.

b. Check for convergence

Monitor the convergence of imputed values. Convergence can be assessed using various criteria, such as small changes in imputed values across iterations.

Step 4: Combine the n imputed datasets to create a single dataset with imputed values.

Step 5: Use the desired statistical analysis, like - regression, classification, to analyze each imputed dataset separately. Use appropriate rules, like - Rubin's rules for parameter estimates, combining p-values, to combine the results from each imputed dataset.

Final Output: The final imputed dataset and the results of the statistical analysis.

4. PRE-PROCESSING AND ANALYSIS

Before diving into the principle analysis, it is a good idea to carry out a thorough exploratory statistics analysis (EDA) to gain a better know-how of the dataset evolved over time, especially in terms of distribution and behavior. Various descriptive statistics like mean, median, deviations, percentiles, etc. for every column was calculated. This helped tremendously to understand and summarize the changes in key variables over time. In addition, facts reduction was conducted to streamline the initial dataset, which contained over 5000 variables. Through a meticulous choice system, only anthropometric data was considered and the number columns were brought down to a hundred and seventy relevant variables. A visual snapshot of the dataset prior to the imputation process is shown in Figure 2.

	mas3_id_no	f0_m_age_eld_child	f0_m_age	f0_socio_eco_sc	f0_dt_prepreg_visit	f0_m_ht	f0_m_wt_prepreg	f0_m_bmi_prepreg	f0_m_tr_prepreg
0	1	6.0	22.0	15.0	1994-07-21	154.2	46.1	19.387963	9.9
1	2	1.0	22.0	19.0	1994-10-27	158.5	39.0	15.524087	7.3
2	3	NaN	20.0	28.0	1994-07-09	145.4	34.6	16.366179	9.7
3	4	13.0	35.0	20.0	1994-07-21	149.9	69.0	30.707596	27.0
4	5	NaN	16.0	19.0	1995-09-05	149.1	38.7	17.408273	10.5

5 rows × 177 columns

Figure 2. First 5 rows of the data frame before imputing

The PMNS data is the longitudinal dataset where certain variables were categorical and needed encoding to enable temporal analysis. Label encoding was used to encode the categorical variables. Over the course of four decades, missing data became a major concern. Before proceeding with imputation, it was essential to identify the extent and patterns of missing data

within the dataset. Missing data can arise due to various reasons, including non-response, data entry errors, or changes in data collection practices over time. A systematic approach was undertaken to identify variables and data points with missing values. To maintain the completeness of the dataset and ensure robust analysis, two prominent imputation methods like: KNN and MICE Imputation Method were applied. The snapshot of the dataset after imputation is shown in Figure 3.

	mas3_id_no	f0_m_age_eld_child	f0_socio_eco_sc	f0_dt_prepreg_visit	f0_m_ht	f0_m_wt_prepreg	f0_m_bmi_prepreg	f0_m_tr_prepreg	f0_m_bi_prepreg
0	1	6.000000	15.0	1994-07-21	154.2	46.1	19.387963	9.9	4.3
1	2	1.000000	19.0	1994-10-27	158.5	39.0	15.524087	7.3	3.3
2	3	3.643989	28.0	1994-07-09	145.4	34.6	16.366179	9.7	5.1
3	4	13.000000	20.0	1994-07-21	149.9	69.0	30.707596	27.0	18.5
4	5	3.228441	19.0	1995-09-05	149.1	38.7	17.408273	10.5	6.1

Figure 3. First 5 rows of the data frame after imputing

In this research study, the prime focus is on the analysis of longitudinal data, emphasizing several crucial aspects:

Longitudinal Data Analysis: The core of the research revolved around the exploration and analysis of longitudinal data. This type of data offers valuable insights into how variables change over time. A significant challenge in the dataset was the presence of numerous columns with 30% - 50% null values. To streamline the process of analysis, it was needed to decompose the dataset and scale down the volume of the dataset, hence anthropometric data for three generation were considered and rest were dropped. This led to a substantial reduction in the dataset, transforming it from over 5,000 variables to a more manageable 177 variables. Figure 4 presents the percentage of column-wise missing value.

Column Name	Null Value %	Column Name	Null Value %	Column Name	Null Value %	Column Name	Null Value %
mas3_id_no	0	f0_m_hip_circ_v1	0	f0_m_bmi_6yr	13.054499	f0_f_wt_ini	4.309252
f0_m_age_eld_child	39.163498	f0_m_glu_f_v1	3.675539	f0_m_head_circ_6y	12.547529	f0_f_bmi_ini	5.449937
f0_socio_eco_sc	0.126743	f0_m_b12_v1	9.252218	f0_m_waist_circ_6y	12.547529	f0_f_waist_circ_ini	4.309252
f0_dt_prepreg_visit	0.126743	f0_m_rcf_v1	11.533587	f0_m_hip_circ_6y	12.547529	f0_f_hip_cir_ini	4.309252
f0_m_ht	0	f0_m_fer_v1	5.576679	f0_m_biceps_6y	12.547529	f0_f_head_cir_ini	39.416984
f0_m_wt_prepreg	0.126743	f0_m_sys_bp_r1_v2	8.238276	f0_m_triceps_6y	12.547529	f0_f_glu_f_ini	18.504436
f0_m_bmi_prepreg	0.887199	f0_m_dia_bp_r1_v2	8.238276	f0_m_subscap_6y	12.547529	f0_f_vitB12_ini	24.207858
f0_m_tr_prepreg	0.126743	f0_m_pulse_r1_v2	24.461343	f0_m_suprai_6y	12.547529	f0_f_pfol_ini	24.334601
f0_m_bi_prepreg	0.126743	f0_m_sys_bp_r2_v2	8.365019	f0_m_syst_bp_r1_6y	12.420786	f0_f_ferr_ini	22.560203
f0_m_ss_prepreg	0.126743	f0_m_dia_bp_r2_v2	8.365019	f0_m_syst_bp_r2_6y	12.420786	f0_f_wt_6y	16.983523
f0_m_su_prepreg	0.126743	f0_m_pulse_r2_v2	24.714829	f0_m_dia_bp_r1_6y	12.420786	f0_f_ht_6y	16.983523
f0_m_ma_prepreg	0.126743	f0_m_wt_v2	7.984791	f0_m_dia_bp_r2_6y	12.420786	f0_f_bmi_6yrs	16.983523
f0_m_waist_prepreg	0.126743	f0_m_bmi_v2	7.984791	f0_m_pulse_r1_6y	12.420786	f0_f_head_circ_6y	16.983523
f0_m_hip_prepreg	0.126743	f0_m_waist_circ_v2	7.984791	f0_m_pulse_r2_6y	12.420786	f0_f_waist_circ_6y	16.983523
f0_m_diab	0.633714	f0_m_hip_circ_v2	7.984791	f0_m_ht_12y	16.730038	f0_f_hip_circ_6y	16.983523
f0_m_sys_bp_r1_v1	0.633714	f0_m_glu_f_v2	13.434728	f0_m_wt_12y	16.730038	f0_f_biceps_6y	16.983523
f0_m_dia_bp_r1_v1	0.633714	f0_m_b12_v2	16.856781	f0_m_bmi_12y	16.730038	f0_f_triceps_6y	16.983523
f0_m_pulse_r1_v1	35.868188	f0_m_rcf_v2	21.799747	f0_m_syst_bp_r1_12y	29.024081	f0_f_subscap_6y	16.983523
f0_m_sys_bp_r2_v1	0.633714	f0_m_fer_v2	17.490494	f0_m_syst_bp_r2_12y	29.024081	f0_f_suprai_6y	16.983523
f0_m_dia_bp_r2_v1	0.633714	f1_c_sex	1.140684	f0_m_dia_bp_r1_12y	29.024081	f0_f_syst_bp_r1_6y	17.237009
f0_m_pulse_r2_v1	35.99493	f0_m_calfat_v1	1.267427	f0_m_dia_bp_r2_12y	29.024081	f0_f_syst_bp_r2_6y	17.237009
f0_m_wt_v1	0	f0_m_calfat_v2	5.449937	f0_m_pulse_r1_12y	29.024081	f0_f_dia_bp_r1_6y	17.237009
f0_m_bmi_v1	0	f0_m_wt_6y	12.547529	f0_m_pulse_r2_12y	29.024081	f0_f_dia_bp_r2_6y	17.237009
f0_m_waist_circ_v1	0	f0_m_ht_6y	12.547529	f0_f_ht_ini	5.323194	f0_f_pulse_r1_6y	17.237009

Figure 4. Column-wise Null Value percentage

5. EXPERIMETAL RESULTS AND EVALUATION

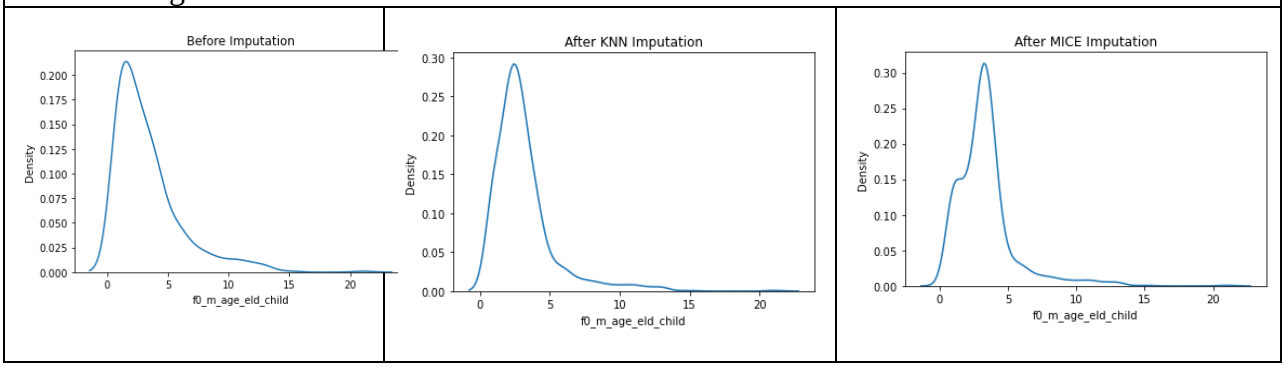
This particular section of the paper holds significant importance due to its provision of a comprehensive depiction of the results derived from the conducted study as well as the efficacy of the proposed techniques or methodologies. It is of utmost importance for the purpose of demonstrating the effectiveness of the employed techniques, algorithms, or approaches. The current findings are juxtaposed with those that already exist, thereby enabling a direct comparison between the experimental outcomes. Within this section, there are quantitative measures presented which include performance metrics, accuracy rates, error rates, or any other pertinent measures. This numerical data enhances the precision of the findings and allows for an objective evaluation. It forms the empirical basis that substantiates your assertions, contributes to the advancement of knowledge, and facilitates communication within the scientific community.

5.1 Comparative Analysis of Imputation Methods

In the investigation of the imputation strategies—K-Nearest Neighbors (KNN) and Multiple Imputation by using Chained Equations (MICE)—in context of a longitudinal dataset spanning four decades, the focus was on comparing their effectiveness in preserving temporal integrity.

This preference was bolstered by means of observations in the KDEplots, wherein MICE constantly yielded extra normalized outcomes as compared to KNN. These meticulous steps in handling missing values and deciding on appropriate imputation method were critical in ensuring the correctness and integrity of the evaluation of longitudinal information. Combination of data reduction, imputation, and visualization techniques allowed to extract meaningful insights from the extensive dataset. Figure 5 depicts the distribution of some columns before imputation and after imputation (for both KNN and MICE Imputation methods). It is observed that the right-skewed original data, after imputation gets normalized in the case of MICE Imputation than KNN Imputation

Column: Age of the elder child



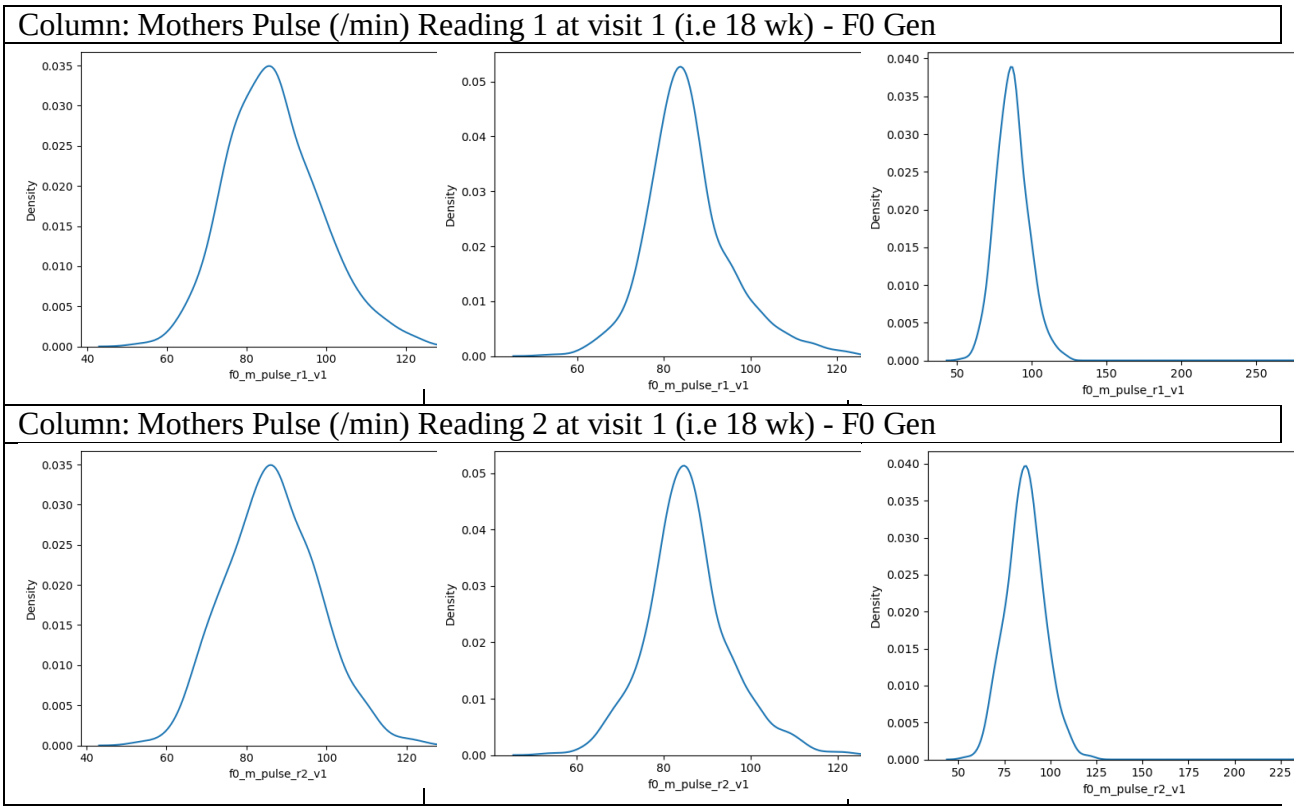


Figure 5. Data distribution of a few columns before imputation and after imputation
(for both KNN and MICE Imputation methods)

- **K-Nearest Neighbors (KNN) Imputation:**

The KNN imputation approach was used, which estimated missing values by examining the values of their k-nearest neighbors in the dataset. After imputation, an in depth statistical evaluation was carried out to evaluate its overall performance.

- **Statistical Analysis after KNN Imputation:**

Temporal Distortion Assessment: Variables with sizeable temporal shifts had been recognized, and the volume of distortion was measured in the temporal distribution. This involved comparing the distribution of those variables prior and after imputation.

- **Multiple Imputation by Chained Equations (MICE):**

The MICE imputation method was also tested, which iteratively imputed missing data while respecting temporal dependencies, thus producing multiple imputed datasets. Following MICE imputation, rigorous statistical analysis was performed to evaluate its effectiveness.

1. **Statistical Analysis after MICE Imputation:**

Temporal Integrity Preservation: The preservation of temporal integrity was assessed by comparing the temporal distributions of variables before and after imputation. Specifically the

correlation coefficients were computed and hypothesis tests were conducted to determine if the temporal characteristics were retained.

Comparative Analysis of Mean and Standard Deviation

Our analysis of the mean and standard deviation for each variable, both before and after imputation, yielded valuable insights:

a. KNN Imputation: KNN imputation, while initially addressing missing data, exhibited limitations in preserving the temporal distribution of variables. Notably, variables with substantial temporal shifts often displayed distortions.
b. MICE Imputation: In contrast, MICE excelled in maintaining the temporal coherence of the dataset. Means and standard deviations computed after MICE imputation closely mirrored those of the original dataset, affirming its ability to preserve the temporal evolution of data.

Figure 6 depicts the statistical comparison between the data of a few variables before and after imputing (by both KNN and MICE Imputation methods).

Column Name	**Before Imputation Mean**	**After KNN Imputation Mean**	**After MICE Imputation Mean**
f0_m_age_eld_child	3.372916667	3.140684411	3.36019396
f0_m_pulse_r1_v1	86.85573122529644	85.75893536121676	87.08614855242524
f0_m_pulse_r2_v1	86.42178217821782	85.57870722433462	86.54064103009914

Column Name	**Before Imputation Std**	**After KNN Imputation Std**	**After MICE Imputation Std**
f0_m_age_eld_child	2.789125886	2.2331217	2.193351357
f0_m_pulse_r1_v1	11.723925640950977	9.820688660611285	12.732705214604545
f0_m_pulse_r2_v1	11.405856925911545	9.518790012455863	11.667464849012498

Column Name	**Before Imputation Median**	**After KNN Imputation Median**	**After MICE Imputation Median**
f0_m_age_eld_child	3	2.6	3.123262414
f0_m_pulse_r1_v1	86.0	84.6	86.65036410651682
f0_m_pulse_r2_v1	86.0	85.0	86.0

Figure 6. Statistical Comparative Analysis between KNN and MICE Imputations (both before and after imputing)

6. CONCLUSION

This study, centered on missing data imputation within a longitudinal dataset spanning three decades, underscores the critical importance of retaining temporal integrity in longitudinal research. The findings clearly support the superiority of MICE as the preferable choice for addressing missing data within a longitudinal dataset. Not only did it surpass KNN imputation in terms of mean and standard deviation preservation, but the rigorous statistical analysis additionally confirmed its brilliant potential to maintain the nuanced temporal traits of the records. Beyond the specifics of imputation strategies, this study emphasizes the pivotal function of methodological alternatives in the context of long-term datasets. Researchers engaged in longitudinal research ought to not forget MICE as a reliable tool to cope with missing data while maintaining the dataset's temporal integrity.

As a conclusion to this study, it acknowledges the paramount importance of methodological rigor in longitudinal research. The selection of an appropriate imputation approach, as exemplified by means of MICE in this study, can profoundly affect the validity and reliability of conclusions derived from longitudinal records, contributing to the robustness of such research endeavors.

Acknowledgement

The authors thank the Diabetes Unit, KEM Hospital & Research Centre, Pune for the extensive help with the dataset and expertise in the domain of the research. The research team extends a gesture of gratitude to Society for Data Science, Pune for their guidance on the usage of data science algorithms.

REFERENCES

[1] International Diabetes Federation (IDF). (2019). IDF Diabetes Atlas, 9th edition.

[2] American Diabetes Association. (2021). Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes.

[3] American Diabetes Association. (2018). Microvascular Complications and Foot Care: Standards of Medical Care in Diabetes.

[4] Bommer, C., Heesemann, E., Sagalova, V., et al. (2018). The global economic burden of diabetes in adults aged 20-79 years: a cost-of-illness study. The Lancet Diabetes & Endocrinology.

[5] Inzucchi, S. E., Bergenstal, R. M., Buse, J. B., et al. (2015). Management of hyperglycemia in Type 2 diabetes: a patient-centered approach. Diabetes Care.

[6] Tuomilehto, J., Lindström, J., Eriksson, J. G., et al. (2001). Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. New England Journal of Medicine.

[7] Powers, M. A., Bardsley, J., Cypress, M., et al. (2015). Diabetes self-management education and support in type 2 diabetes: A joint position statement of the American Diabetes Association, the American Association of Diabetes Educators, and the Academy of Nutrition and Dietetics.

[8] World Health Organization. (2021). Global report on diabetes.

[9] Sharma N., Ghosh S., Saha M. (2021) Open Data for Sustainable Community. Advances in Sustainability Science and Technology. Springer, Singapore. https://doi.org/10.1007/978-981-33-4312-2_10

[10] Sharma, N., De, P.K. (2023). Towards Net-Zero Targets: Usage of Data Science for Long-Term Sustainability Pathways. Advances in Sustainability Science and Technology. Springer, Singapore. https://doi.org/10.1007/978-981-19-5244-9

[11] Nair V., Joshi S., Patil M., Sharma N. (2021) Supply Chain Management During the Time of Pandemic. In: Sharma N., Chakrabarti A., Balas V.E., Bruckstein A.M. (eds) Data Management, Analytics and Innovation. Lecture Notes on Data Engineering and Communications Technologies, vol 70. Springer, Singapore. https://doi.org/10.1007/978-981-16-2934-1_20

[12] Li, X., Li, C., Liu, F., Wang, L., & Wang, J. (2018). Application of machine learning methods to predict type 2 diabetes mellitus incidence in a Chinese rural population: A prospective cohort study. Scientific Reports, 8(1), 1-8.

[13] Rajpurkar, Pranav & Irvin, Jeremy & Zhu, Kaylie & Yang, Brandon & Mehta, Hershel & Duan, Tony & Ding, Daisy & Bagul, Aarti & Langlotz, Curtis & Shpanskaya, Katie & Lungren, Matthew & Ng, Andrew. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv:1711.05225. https://doi.org/10.48550/arXiv.1711.05225

[14] Jian Y, Pasquier M, Sagahyroon A, Aloul F. A Machine Learning Approach to Predicting Diabetes Complications. Healthcare (Basel). 2021 Dec 9;9(12):1712. doi: 10.3390/healthcare9121712. PMID: 34946438; PMCID: PMC8702133.

[15] Anderson, S. M., & Peters, A. (2020). Reinforcement learning produces robust control policies for glucose regulation in type 1 diabetes. Journal of Diabetes Science and Technology, 14(1), 17-25.

[16] Smith, R. C., Domenico, D., & Roland, B. (2019). Personalized insulin dosage recommendations using machine learning for individuals with type 2 diabetes. Journal of Diabetes Science and Technology, 13(3), 493-501.

[17] Jahangiri, M., Kazemnejad, A., Goldfeld, K.S. (2023) A wide range of missing imputation approaches in longitudinal data: a simulation study and real data analysis. BMC Med Res Methodol 23, 161. https://doi.org/10.1186/s12874-023-01968-8

[18] Huque, M.H., Carlin, J.B., Simpson, J.A. et al. A comparison of multiple imputation methods for missing data in longitudinal studies. BMC Med Res Methodol 18, 168 (2018). https://doi.org/10.1186/s12874-018-0615-6

[19] Spratt M, Carpenter J, Sterne JA, Carlin JB, Heron J, Henderson J, Tilling K. Strategies for multiple imputation in longitudinal studies. Am J Epidemiol. 2010 Aug 15;172(4):478-87. doi: 10.1093/aje/kwq137. Epub 2010 Jul 8. PMID: 20616200.

[20] Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: a review. Test (Madr). 2009 May 1;18(1):1-43. doi: 10.1007/s11749-009-0138-x. PMID: 21218187; PMCID: PMC3016756.

[21] Engels JM, Diehr P. Imputation of missing longitudinal data: a comparison of methods. J Clin Epidemiol. 2003 Oct;56(10):968-76. doi: 10.1016/s0895-4356(03)00170-7. PMID: 14568628.

[22] Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 2011 Mar;20(1):40-9. doi: 10.1002/mpr.329. PMID: 21499542; PMCID: PMC3074241.

[23] D'souza N, Behere RV, Patni B, Deshpande M, Bhat D, Bhalerao A, Sonawane S, Shah R, Ladkat R, Yajnik P, Bandyopadhyay SK, Kumaran K, Fall C, Yajnik CS. Pre-conceptional Maternal Vitamin B12 Supplementation Improves Offspring Neurodevelopment at 2 Years of Age: PRIYA Trial. Front Pediatr. 2021 Dec 7;9:755977. doi: 10.3389/fped.2021.755977. Erratum in: Front Pediatr. 2022 Feb 21;10:860732. PMID: 34956975; PMCID: PMC8697851.

[24] Batista, Gustavo & Monard, Maria-Carolina. (2002). A Study of K-Nearest Neighbour as an Imputation Method.. Hybrid Intelligent Systems, ser Front Artificial Intelligence Applications. 30. 251-260.