2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Feature selection for unbiased imputation of missing values : A case study in healthcare

Chetanya Puri[1,2], Gerben Kooijman[2], Xi Long[2], Paul Hamelmann[2], Sima Asvadi[2], Bart Vanrumste[1*] and Stijn Luca[3*]

***Abstract*— Datasets in healthcare are plagued with incomplete information. Imputation is a common method to deal with missing data where the basic idea is to substitute some reasonable guess for each missing value and then continue with the analysis as if there were no missing data. However unbiased predictions based on imputed datasets can only be guaranteed when the missing mechanism is completely independent of the observed or missing data. Often, this promise is broken in healthcare dataset acquisition due to unintentional errors or response bias of the interviewees. We highlight this issue by studying extensively on an annual health survey dataset on infant mortality prediction and provide a systematic testing for such assumption. We identify such biased features using an empirical approach and show the impact of wrongful inclusion of these features on the predictive performance.**

***Clinical relevance*— We show that blind analysis along with plug and play imputation of healthcare data is a potential pitfall that clinicians and researchers want to avoid in finding important markers of disease.**

I. INTRODUCTION

Missing data is a ubiquitous problem in statistical analysis or data science irrespective of the domain, be it social sciences or health sciences. Most, if not all the machine learning algorithms presume that all the information is present for all the available features. Conventional techniques of handling missing data include performing complete case analysis which is deletion of missing cases but this strategy results in lesser informative subset of the dataset.

Health data are being massively generated due to the advancement of both data acquisition and analysis technologies, examples of which include time-series data from intensive care units (ICU), biomarker data, electronic health records (EHR), or health surveys. The global market for big data in health care has been projected to grow significantly from US$19.6 billion in 2018 to US$ 47.7 billion in 2022 [12]. Undoubtedly, this rise is due to the penetration of data analytics for better predictive clinical outcomes, analyzing disease, and tracking patterns thus increasing overall public health. Modelling such large scale data and predicting the health status for improvement of the patient is challenging. One such challenge is addressing missing values in data, that arise, for example, from unrecorded data from ICU machines due to lead detachment or respondents intentional/unintentional non-responsiveness to health surveys [6].

Datasets (particularly in healthcare) are often preprocessed by various imputation techniques that rely on the assumption of independence between the missing mechanism and the observed data. Statistical tests to verify this assumption often fail when missing data is abundant and a subset of reasonable size of complete data is absent [11].

In this article, we illustrate the common pitfall of blindly applying imputation techniques that can lead to biased results. To this end, we utilize a publicly available dataset from the annual health survey in India and show how state-of-the art imputation techniques fall short in reliable feature matrix completion for classification purposes. Furthermore, we propose an empirical approach to study the effect of including features that are strongly associated to the occurrence of missing data.

A large part of existing literature on missing data analysis that we discuss later studies one or more methods to impute data. In this article, we highlight the biased effect that imputation might have on the results of a predictive classification model in the presence of imbalanced missingness across different classes and we propose a method that can support in preventing careless imputation of missing data.

The remainder of the paper is structured as follows. In section II, we talk about the imputation techniques and types of missingness. Further, we describe the dataset in section III. Section IV elaborates upon the experiments performed in order to show the impact of the described challenges with unbiased missing values imputation. We conclude by giving final remarks to the reader in section V.

II. RELATED WORK

Several approaches exist that handle missing data by (a) *deletion* of the cases that have values missing for a single variable, simply excluding such cases can be used to build complete datasets [4] or (b) estimating a single set of missing values by *single imputation* using statistical moments, *k*-nearest neighbours or (c) a confidence interval imputation by much more complex *multiple imputation* [13]. A specific implementation of multiple imputation strategy known as the Multivariate Imputation by Chained Equations (MICE) involves multiple steps of imputation in which every variable is imputed conditionally on all other variables [5]. Deletion based imputation can lead to loss of statistical power and can

[1]eMedia Lab and STADIUS, Department of Electrical Engineering (ESAT), KU Leuven, Belgium. <firstname>.<lastname>@kuleuven.be
[2]Philips Research, Eindhoven, The Netherlands <firstname>.<lastname>@philips.com
[3]Department of Data Analysis and Mathematical Modelling, Ghent University, Ghent, Belgium stijn.luca@ugent.be
*These authors contributed equally to this work.

introduce bias when a smaller complete subset is selected from a non-complete dataset.

Based on the type of missingness, three basic mechanisms are present [4], described as follows, Suppose we have missing data on a variable Y and we have some other variable X, then, one defines:

- Missing completely at random (MCAR) : If the probability of missing data on Y is unrelated to the value of Y itself or to the values of any other variables in the data set, the data is said to be MCAR.
- Missing at random (MAR): If the missingness depends only on the data that are observed but not on the missing components, the data are MAR. i.e., $P(Y\text{missing}|Y, X) = P(Y\text{missing}|X)$
- Not missing at random (NMAR): If the probability that Y is missing depends on the unobserved value of Y itself, then the mechanism is NMAR.

Most of the imputation strategies work under the assumption that the missingness is MCAR [11]. Statistical tests like Little's test [11] exist that can test whether the data is MCAR or not. However, in the absence of a small complete subset (when missing data is abundant), it is difficult to conduct such a test and existing imputation techniques tend to fail in reliably predicting the missing values. Authors in [8] and [7] discuss different imputation methods and compare the performance of imputation techniques with different amount of missingness on different datasets. They advise that different missing data mechanism needs different imputation strategy, however none of the previous works talk about the imbalance in missingness that can be present in different classes when considering a classification problem. Imputation without analysis of such an imbalance can lead to erroneous completion of the feature matrix which we will show later.

In this article, we illustrate the challenges of using imputation methods when the MCAR assumption is not met. For this purpose, we use a case study from healthcare and we propose an algorithm to study the effect of including features that are strongly associated to the occurrence of missing data.

III. DATA

We chose a publicly available healthcare survey dataset conducted over women that underwent pregnancy in several states in India [1]. Child mortality remains a major challenge in India and is responsible for approximately 39.1 deaths per 1,000 live births in 2017 [2]. Child mortality as a pregnancy outcome is considered a major attribute in building efforts to preventive antenatal care thus reducing infant mortality. Poor pregnancy outcome in India is not just attributed in defining the outcome but is also a consequence of substandard health information systems. The National Institute for Medical Statistics of the Indian Council of Medical Research (ICMR - NIMS) has launched the National Data Quality Forum (NQDF) in collaboration with the Population Council. The purpose of the NQDF is establishing protocols and good practices for betterment of data collection, storage and dissemination [9]. Major barriers to the data quality include (a) lack of comparability, (b) discordance between system and survey level estimates, (c) lengthy questionnaires, (d) questions related to socially restricted conversation topics, (e) age-reporting errors or non-response, (f) intentional skipping of questions, (g) under-reporting due to subjective question interpretation and incompleteness, and (h) paucity of data to generate reliable estimates on mortality [9]. We select data from the open government platform in India where the Indian government has provided open access to datasets, documents, etc. for public use. This dataset is also collected as part of a joint initiative between government of India and US government. Authors in [14] have shown the risks of using such open datasets from non-verified sources such as [3]. They identify that Woman Schedule Section 1 and Section 2 (called WPS dataset) is from a verified source [1]. A number of 355 features in the WPS dataset [1] are present in the form of questionnaire, with fields related to social, economic, health status or demographic indicators as well as the outcome of pregnancy (live or stillbirth).

Since the dataset consists of questions from surveys, some questions are explicitly on the child birth outcome thus making some of the features highly correlated with the fact whether the child birth resulted in a live or stillbirth. Hence, features such as baby weight taken or not, weight measurement, immunization card details, different vaccines, polio, hepatitis, vit. A, IFA tablet, feeding details, breastfed, animal dairy, solid food month, etc. were removed to maintain causality of the labels with respect to the feature set because these features can only be recorded if the pregnancy outcome is positive. We then selected 233 features out of 355 as the final feature list for further analysis.

IV. THE CASE STUDY

Given a feature matrix $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, $\mathbf{x} \in \mathbb{R}^d$ observed for N subjects, the objective is to learn a function $h : \mathbf{X} \rightarrow Y$, where $Y = \{0, 1\}$ corresponds to prediction of still or live birth respectively. The class of stillbirth also includes all cases of induced abortion and spontaneous abortion. The number of cases for live birth are much more than all the stillbirth cases. Hence, we look at the problem of learning a model for binary classification of live and stillbirth.

Imputation of the feature matrix occurs during preprocessing before training the model, as shown in Figure 1.

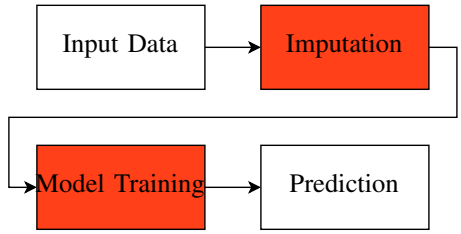

Fig. 1: Typical processing pipeline for learning with missing data.

We compare the performance of the imputation approach by keeping the processing pipeline fixed i.e the training data and the classifier and its parameters are fixed and only the imputation approach is varied. For our experiments, we perform a 10-fold cross validation with minority class as the positive class (stillbirth) and plot the average receiver operating characteristic. Figure 2 shows that a random forest classifier with single imputation methods like constant based filling for imputation achieves the best performance. This motivated us to look closely into the features and the missingness in relation to the class label.

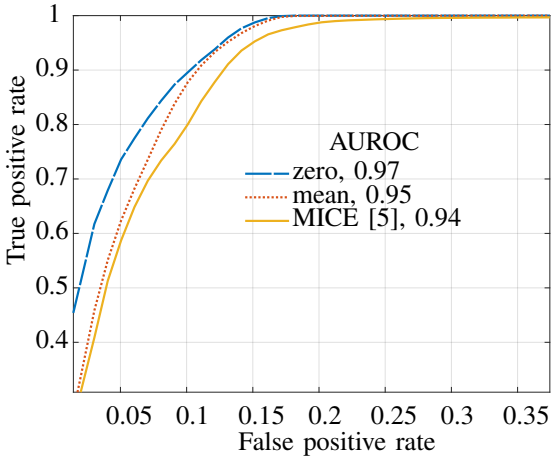

Fig. 2: A simple zero based constant filling appears to have the most predictive power (keeping classifier and its parameters fixed) when the imputation methods are applied blindly without understanding the type of missingnes.[1]

We take two exemplary features that are discrete-valued categorical features namely "*source_of_anc*" and "*maternity_financial_assistance*". In the annual health survey, "*source_of_anc*" refers to the institution offering antenatal care (ANC). 12 different government or private institutions operating at different governance level are assigned real numbers. For example, women receiving antenatal care at government operated rural center called *anganwadi* are assigned the real number '1'. Similarly, women receiving ANC from private hospitals are assigned the value '9'. The complete description of the domain space is mentioned in [1] and is mapped to $\mathbb{R}$ in $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 99\}$. For the feature "*maternity_financial_assistance*", women who took financial assistance under the government scheme *Janani Suraksha Yojana (JSY)* are assigned the value '1' for this feature. If they avail any other government scheme, real number '2' is assigned, '3' for any other non-government scheme and '4' in case no financial assistance was availed. The domain space for this feature is mapped in $\mathbb{R}$ to $\{1, 2, 3, 4\}$. Figure. 3(a) and (b) represent the feature "*source_of_anc*" and "*maternity_financial_assistance*" respectively. These two features are representative for multiple features which have a lot of missing values or are filled with zero in the questionnaire, possibly due to errors in the interview. For the sake of discrimination, we do not combine the missing values and zero-entries even if they mean the same thing. As can be observed from Figure. 3(a) 9.7% data is missing in class "0" and 77.48% data is missing for class "1" for the feature "*source_of_anc*". Similarly from Figure. 3(b), "*maternity_financial_assistance*" feature has around 0.187% data missing for class "0" and 75.82% data is missing for class "1". This percentage imbalance in missing data will be further irritated if we consider the occurrence of zero in the data as '0' is not in the domain space of most of the features and was recorded maybe as a missing value. Suppose we fill the missing data with a simple single imputation approach, for example, a constant '$c \in \mathbb{R}$' or mean, for feature "*source_of_anc*", then for 77.48% of the data in class '1' the feature value will be c and the remaining 22.52% will take values somewhere in the domain of the feature $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 99\}$. On the other hand, only 9.7% of the data is missing for class '0' and will be assigned the value c. Rest of the class '0' (90.263%) will take values from the domain of the feature. Since class "1" has more missing data than class "0", constant-filling based imputation methods will provide a false sense of discriminatory power to the feature.

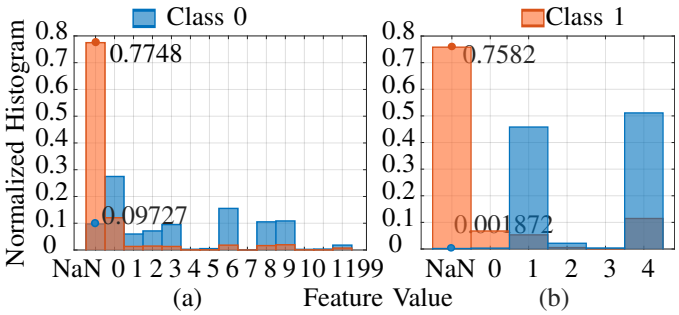

Fig. 3: Exempelary features (a)"*source_of_anc*" and (b) "*maternity_financial_assistance*" with different classwise imbalance in terms of availability of the data, Class 1 = live birth, Class 0 = stillbirth

Figure. 4 represents a compact view of all the features plotted with respect to availability of data in each class. All the features that exhibit classwise-imbalance in availability of feature data are shown in $*$. The line $y = x$ in Figure 4 represents the features that have equal amounts of missing data in each class (marked in $\circ$). The margins along the line $y = x$ represent the tolerance level (e.g. = 10% tolerance) for visualising whether the feature is useful or not in the absence of actual feature value. One way of finding out if the features are missing completely at random is by performing Little's test [10]. We found on performing Little's test that the data is not missing completely at random.

We develop an empirical approach to evaluate the features that exhibit such behaviour and use the algorithm provided in Algorithm 1. The algorithm first, calculates the percentage missing data in each class. If the difference in percentage of the missing data calculated in the previous step differs by a

[1]Notice the difference in x and y coordinates as this is a zoomed-in snippet of the AUROC curve to improve the visibility of the curves.

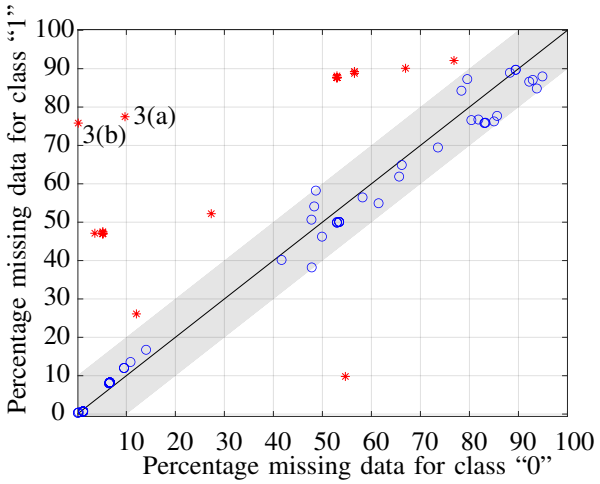

Fig. 4: Each data point ($*$, $\circ$) represents a feature with x and y coordinates being the missing percentage in class 0 and 1 respectively. Each feature outside the tolerance margins (marked as $*$) have high absolute percentage difference between the available class "0" and class "1". As depicted, features from Figure. 3(a and b) are also apparently intolerable features

Algorithm 1 Finding features inside tolerable range

procedure FIND TOLERABLE FEATURES
  $Input : \mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, $\mathbf{x} \in \mathbb{R}^d$, $Y = \{0, 1\}$
  $Parameter : perThresh \in [0, 100]$ ▷ tolerance (in %)
  $mis0 = 0$ ▷ Initiate missing count for class 0
  $mis1 = 0$ ▷ Initiate missing count for class 1
  $tolerableFeatInd = []$
  for $i = 1 : d$ **do**
    $f0 = \mathbf{x}_i(Y == 0)$
    $f1 = \mathbf{x}_i(Y == 1)$
    for $j = 1 : length(f0)$ **do**
      if $isnan(f0(j))$ **then**
        $mis0 = mis0 + 1$
    for $j = 1 : length(f1)$ **do**
      if $isnan(f1(j))$ **then**
        $mis1 = mis1 + 1$
    $misPer0 = 100 * mis0/length(f0)$
    $missPer1 = 100 * mis1/length(f1)$
    $absDiffMiss = abs(missPer0 - missPer1)$
    if $(absDiffMiss < perThresh)$ **then**
      $tolerableFeatInd = [tolerableFeatInd, i]$

pre-decided tolerable limit, then we say that the feature is a tolerable feature with respect to the imbalance in missing data, otherwise, it is an intolerable feature. For example, as can be observed from Figure 3a and b, both the features have an absolute difference of > 60 which is greater than a pre-decided tolerance limit of 10, decided empirically. Hence, both the features are intolerable and have false discriminatory power for model-learning if used with imputation.

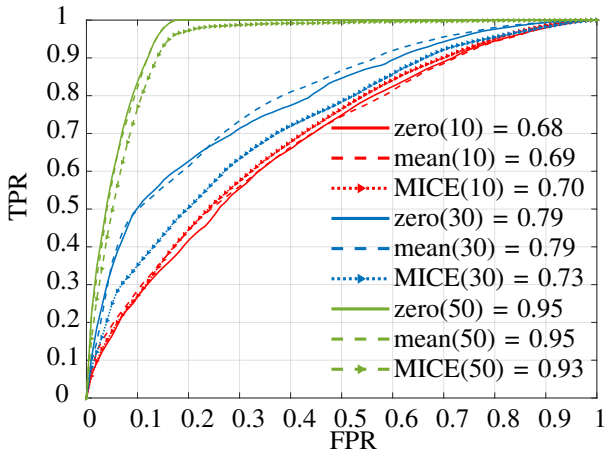

Fig. 5: Classification performance with zero, mean-filling and MICE based imputation when tolerance threshold varies from [10, 30, 50] and the area under the ROC curve represented upto two decimal places.

We test with different values of tolerance thresholds to test the variation of performance if such erroneous features are included in model-building blindly. Figure. 5 represents the classification of live-stillbirth prediction performance with different imputation strategies and different tolerance thresholds (margin as depicted in Figure. 3) as described in algorithm 1. A number of 86, 90 and 117 features were selected based on tolerance thresholds 10, 30 and 50 respectively. Figure 5 shows that we get much higher performance when the tolerance threshold is set high. This is due to the fact that at high tolerance threshold we include more features that are biased because of the imbalance in missingness in different classes. However, when the tolerance threshold is as low as 10, we include less biased features (depicted as $\circ$ in Figure 4). Here, the final performance achieved with tolerance level 10 is around 0.68. We also observe that at the threshold of 10, where minimum number of biased features are included, the state-of-the-art MICE approach performs better than the constant-filling approaches.

In this work in progress, we try to showcase the effect of selecting different thresholds for tolerance selection on a dataset from healthcare. In the future, we would like to work on finding an optimal strategy to find this threshold and test our approach on different datasets with more than two classes.

V. CONCLUSION

This paper reflects on the need for caution when imputing missing values for classification. The assumptions such as MCAR or MAR are not always easy to verify. Most of the state-of-the-art imputation techniques work well when data is MCAR and a subset of complete data is present for guiding the imputation process. We showed the effect of imputation on the performance by studying a case in healthcare. It was evident from our experiments that attention was needed when

features were used with missing values that are strongly associated with the class label and including these in a predictive model can lead to a false sense of discriminatory power. In the future, we would like to develop methods to find the tolerance threshold and fill the missing data in an unbiased manner.

ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 766139. This publication reflects only the authors' view and the REA is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] "Census of india : Annual health survey 2010 - 11 fact sheet," https://www.censusindia.gov.in/vital_statistics/AHSBulletins/Factsheets.html.

[2] "India demographics profile," https://www.indexmundi.com/india/demographics_profile.html.

[3] "Predict outcome of pregnancy," https://kaggle.com/rajanand/ahs-woman-1.

[4] P. D. Allison, *Missing data*. Sage publications, 2001, vol. 136.

[5] S. v. Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *Journal of statistical software*, pp. 1–68, 2010.

[6] N. R. Council *et al.*, *Nonresponse in social science surveys: A research agenda*. National Academies Press, 2013.

[7] A. Farhangfar, L. Kurgan, and J. Dy, "Impact of imputation of missing values on classification error for discrete data," *Pattern Recognition*, vol. 41, no. 12, pp. 3692–3705, 2008.

[8] L. A. Hunt, "Missing data imputation and its effect on the accuracy of classification," in *Data Science*. Springer, 2017, pp. 3–14.

[9] InsightsIAS, "National data quality forum(ndqf)," July 2019. [Online]. Available: https://www.insightsonindia.com/2019/07/26/national-data-quality-forum-ndqf/

[10] R. J. Little, "A test of missing completely at random for multivariate data with missing values," *Journal of the American statistical Association*, vol. 83, no. 404, pp. 1198–1202, 1988.

[11] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2019, vol. 793.

[12] R. ltd and Markets, "Healthcare big data analytics market: Global industry trends, share, size, growth, opportunity and forecast 2019-2024," https://www.researchandmarkets.com/reports/4856240/healthcare-big-data-analytics-market-global.

[13] J. S. Murray *et al.*, "Multiple imputation: A review of practical and theoretical findings," *Statistical Science*, vol. 33, no. 2, pp. 142–159, 2018.

[14] A. Trivedi, S. Mukherjee, E. Tse, A. Ewing, and J. L. Ferres, "Risks of using non-verified open data: A case study on using machine learning techniques for predicting pregnancy outcomes in india," *Proceedings of NeurIPS 2019 Workshop on Machine Learning for the Developing World: Challenges and Risks of ML4D, arXiv preprint arXiv:1910.02136*, 2020.