Towards a Robust Imputation Evaluation Framework

Anthony Chapman, Wei Pang, George Coghill
Department of Computing Science
University of Aberdeen, Aberdeen, UK
Email: {r01ac14, pang.wei, g.coghill}@abdn.ac.uk

***Abstract*—Missing data research is hindered by a lack in imputation evaluation techniques. Imputation has the potential to increase the impact and validity of studies from different sectors (research, public and private). By creating robust evaluation software, more researchers may be willing to use and justify using imputation methods. This paper aims to encourage further research for robust imputation evaluation by defining a framework which could be used to optimise the way we impute datasets prior to data analysis. We propose a framework which uses a prototypical approach to create testing data and machine learning methods to create a new metric for evaluation. We introduce our implementation of such a framework and present some preliminary results. The results show how, for our dataset, records with less than 40% missingness could be used for analysis, which increases the amount of available data for future studies using that dataset.**

Keywords*—*Missing Data; Evaluating Imputation; Imputation; Clustering; Prototypical Testing.

I. Introduction

The number of papers evaluating imputation methods (methods that predict missing values) is so large we cannot fit all of them in this paper, yet there is no evaluation software. Although individually evaluating an imputation method on a dataset has it's place, there are many problems (discussed in the sections to follow) currently associated with it. Even though imputation research has experienced a surge in recent years, evaluating imputation has not advanced at the same pace. This is problematic given recent findings, namely the potentially negative effects imputation can have on the validity and reliability of data analysis [2], thus, more of our attention must be directed to the evaluation of such methods.

Recent reports have noted an increase in the number of studies using imputation methods [3], [4]. Although imputation is being used more, the preferred method is still complete case analysis (aka likewise deletion or masking), where records with missing values are omitted from analysis [5], [6]. Consequently, newer statistical techniques which have eclipsed complete case analysis, in terms of appropriateness, for most circumstances [7], [8], are not being widely used. A robust imputation evaluation method could lead to a rise of popularity in imputation by allowing users to (relatively) easily see the effects an imputation method has on their datasets.

Evaluating imputation must be at the forefront of missing data research in order for imputation to be more widely accepted and, ultimately, used. By enabling others to evaluate imputation, they may be more inclined to consider imputing a dataset and, when appropriate, to use the imputed dataset for an appropriate study. This, in turn, could enable more of the available/meaningful data to be used [9] and this could help decrease uncertainty in studies, as illustrated in Figure I.

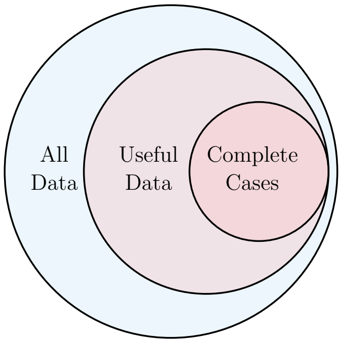

Figure. 1. Useful Data. There may be a larger amount of useful data than the subset of complete cases.

There are many challenges which need to be considered when evaluating imputation. One of which is that new imputation methods are constantly being developed and existing ones keep evolving [38]. Because of the evolving nature of imputation methods, any previous evaluations may become redundant/irrelevant and more up-to-date evaluations are constantly required. As there is currently no straightforward approach to evaluate imputation, we are constantly lagging behind new and improving imputation methods.

Another problem to consider is the complex structure of datasets: how will an imputation method behave with different datasets? Datasets could be very different in structure from one to another. They might consist of solely numerical values, solely non-numerical or could contain some mixture of the two types [10]. An evaluation method which could cope with such diversity could help overcome these issues.

Given the problems already stated, we will propose an imputation evaluation framework and the paper is structured as follows: Section II describes the motivations, implications and background related to this research. Section III describes the proposed framework and Section IV gives a breakdown of the benefits the framework could have on a system. Section V introduces, CLustering to Evaluate Multiple Imputation (CLEMI), our implementation of this framework and shows some preliminary results. The remaining sections consist of a discussion, which includes limitations, in Section VI, and

ISBN: 978-1-61208-646-0

finally, a conclusion and future work in Section VII.

II. Motivation & Background

Missing data is a common occurrence [11] and can negatively affect inferences on the conclusions that may be drawn from data analysis [12], [13].

A. Implications

Missing data prevention mechanisms may not ensure all data is recored or stored [14]. This may be due to a number of reasons ranging from study design to computer error. Ensuring all data is recorded is usually unfeasible in real life, and comping mechanisms for missing data, such as imputation, are paramount in maximising the use of available information [15].

Complete case analysis is still the most common mechanism used to cope with missing data, but records with missing values could also yield valuable information [16]. Although complete case analysis may be appropriate in some cases, ignoring records with missing values could lead to overestimation of results [8], depending on how the standard error is affected [17]. Furthermore, the analysis of information from a subset rather than the entire period of interest is also likely to alter the results of a study [18].

Incorrect imputation has the potential to produce drastically incorrect results [12] and some of the limitations of imputation (such as assuming regressions capture all necessary data for imputation) are discussed in Shih's paper [8]. Imputation could also lead to underestimation or overestimation of test statistics, depending on how standard errors are affected [19], so further analysis will be required by the users. These issues will require analysts to have in-depth knowledge of the data.

Evaluation methods could help with these problems by allowing the analysts to visualise the effects different imputation methods have on datasets of interest. Optimisation is another challenge met by those using imputation methods. Without evaluating imputation, how can we be sure that we have, not only the more appropriate imputation method, but also, optimised the chosen method to perform at its best capability.

B. Missing Data

Although the effects different types of missing data have on imputation have been studied, we are still a long way from truly understanding the effects they have on imputation. An evaluation system could greatly advance current understanding regarding how imputation is affected by different types of missing data.

Missing data occurs for to a wide variety of reasons. The most common reason for missing data is participants dropping out of studies [20]. Other reasons include having too few participants, not reporting data, or the data not being applicable to the study [21]. Computer based reasons include computer error, from the mismatch of variables between datasets to improper merging [22]. These reasons could be minimised by improving study design, though it is unlikely that missing data can be prevented altogether.

Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing not at Random (MNAR) are missing data mechanisms introduced by Rubin [23]. MCAR is when data are missing randomly throughout the dataset without any dependence between the variables. MAR is when missing data may be dependent on other observed data but the not dependent on other missing values. Finally, MNAR is when data are not MAR and the missing values are related to other missing values, with MNAR missing values could be linked to other missing values as well as the observed ones

These different mechanisms will make imputation behave differently. For example, datasets which have MCAR data are the least likely to produce bias [24] when complete case analysis is used. MNAR is seen as the most susceptible to bias but there are ways to minimise this. There are many assumptions and studies which look at the relationship between the different mechanisms and imputation, more research is required for us to understand the effects such mechanisms might have. A robust imputation evaluation methods will make testing the effects more accessible for anyone wishing to do so. Such tests could be done by controlling the type of mechanism used and analysing the effects imputation has, by doing this, we may be able to prove or disprove current notions of the behaviour of missing data mechanisms.

C. Imputation

Imputation research has received a lot of interest in recent years as researchers and industry alike are trying to use as much of the available data as possible [25]. Although there are many methods at our disposal, which vary in appropriateness and complexity, imputation is still not being considered (as a method to deal with missing data) as much as it should be.

Imputation methods widely range in complexity and appropriateness. One of the simplest methods is default value imputation (where all missing values are replaced with a value, such as 0 or Female). Default value is generally regarded as taboo since they could potentially create bias in data. Mean imputation is slightly more intricate, it replaces any missing value with the mean of the corresponding column. Studies have found that mean imputation can overestimate results when there is more than 5%-10% missing values [42]. More complex methods also exist, such as multiple imputation, where regression models are created from the recorded data and missing values are imputed according to these regression models. Multiple imputation generates a number of datasets to account for uncertainty [26].

Imputation is still not being used adequately. One reason why it is not widely used could be due to the potential to underestimate values by imputing incorrectly [27]. Another reason could be a general lack of understanding on how imputation methods perform with different quantities and types of missing data [25]. Finally, imputation methods have, until recently, not been readily accessible to researchers [28].

One of the biggest factors against imputation is the lack of understanding in the effects they have on dataset with missing values. This understanding deficiency leads to a lack of trust in imputation methods, which leads to people not using as much of the available data, or the misuse of imputation methods, both of which change the data's underlying information and leads to a negative impact on studies.

Many imputation methods have a certain amount of flexibility which, potentially, allows users to impute datasets in a more efficient way. Without evaluation methods, it is sometimes difficult to choose the best parameters when imputing; this leads to sub-optimal imputation. Many studies either fail to look for optimal results or simply do not report how they have optimised imputation [29].

Having a robust and versatile imputation evaluation method could lead to better understanding how datasets are affected by different imputation methods.This could then enable and improve optimisation of imputation methods.

D. *Evaluating Imputation*

Current imputation evaluation methods are mostly based on statistical regressions [11], [26], [30]–[35], some machine learning approaches have also been proposed [36]–[38]. Regression based evaluation performs well for individual cases but the results from such evaluations are not generalisable, the outcomes may not be applicable to others and are manually intensive to obtain.

Standards have been proposed to streamline imputation which, in theory, could lead to widespread evaluation to be carried out when imputing datasets [39]–[42]. These papers provide guidelines to handle missing data and suggest "good practices" for imputing datasets. They also provide useful information such as how some imputation methods might behave when applied to different types and/or quantities of missing data. Although these standards provide useful information, they do not advance on the problems posed for evaluating imputation. Some of the problems include:

- Results from evaluations may be unsuitable for other datasets or imputation methods
- It is not straightforward to evaluate the prediction of something which is truly missing
- Does the evaluation show whether the imputation method predict "true" values
- There are no standards to evaluate imputation methods
- Regression based evaluation is manually intensive.

Recent studies evaluating various imputation methods applied to a selection of datasets may be advantageous to resolve individual problems [31], [33], [38], [43]. However, due to inherent complexities of datasets, their results cannot be generalised for others to use. For example, [32] reported that for their particular dataset, multilevel imputation gave the best results in their study. This result may not be the same given a different dataset.

Similarly, three imputation methods were evaluated in [33] and four imputation methods were evaluated in [38]. Due to the different approaches used to evaluate the methods, these results are not comparable to each other; as one study may yield more appropriate results for their specified problem whereas an alternative study may find contradicting results. From this, we suggest that an evaluation method should be able to be compared to other methods in order for any results to be used by others.

By being able to compare different evaluation outcomes, may enable us to not only find the most appropriate method to impute a specific dataset, but also help us optimise an imputation method. By evaluating the same imputation method multiple times and changing any parameters every time and comparing the results, we may be able to optimise the method for a given dataset.

III. Proposed Framework

Now that we have identified a lack of research on evaluating imputation, we propose a framework which could help with most of the current problems we face when evaluating imputation. The framework can be split into several stages. The first stage involves creating a benchmark dataset to evaluate imputation. The benchmark must be similar to the original dataset in order to preserve the relationship between them. In the second stage prototype datasets are created which can represent the original dataset for testing purposes. We will use these datasets to find the effect imputation has by comparing the imputed datasets to the benchmark.

In the third stage, the imputation methods are applied to the prototype datasets. It will be applied on all datasets in the same manner, specified by the user, in order to reduce uncertainty in the results. The forth stage will evaluate the imputed datasets by comparing them to the benchmark. In theory, a suitable imputation method will create values which are similar to the benchmark, conversely, an unsuitable imputation method will creates values which differ from the benchmark.

A. *Benchmark*

In order to evaluate an imputation, a benchmark could be used. As different datasets are not guaranteed to behave the same, individual benchmarks must be used for every dataset.

We propose using the subset of the dataset consisting of the complete cases as the benchmark. Doing so, we reduce the variance between the dataset in question and the benchmark. This will decrease the evaluation uncertainty by maintaining a close link between the benchmark and the original dataset. This process is shown in Algorithm 1, line 3.

B. *Prototypes*

To evaluate an imputation method, we could apply imputation to testing datasets and quantify the results. We will create prototypes from the benchmark, to act as our testing datasets. Then, impute the prototypes and compare the results to determine if the imputation method created realistic results, namely, whether the results have a small dissimilarity to the benchmark.

The prototypes are created by copying the benchmark and then analysing the missing data structure of the original dataset and imposing the same levels of missingness onto the copy, as illustrated in Figure 2. This is randomised, by creating different (but similar) prototypes, to increase the variability of the datasets. We randomise to reduce uncertainly when

Algorithm 1: Pseudo code for imputation evaluation framework. ©2018 Chapman, Pang & Coghill

```
   input : A dataset with missing values and parameters (if
           any) for the imputation methods.
   output: Evaluation Score: Difference between the
           imputed prototypes and the benchmark.
 1 data ← original dataset with missing values ;
 2 param ← Imputation parameters ;
 3 bench ← complete cases from data ;
 4 n ← amount of prototypes;
 5 missingDist ← missingness distribution from data ;
 6 for i ← 1 to n do
 7 |   p(i) ← bench.delete(missingDist) ;
 8 end
 9 for i ← 1 to n do
10 |   pImp(i) ← impute(p(i), method=param) ;
11 |   pMEAN(i) ← impute(p(i), method=mean) ;
12 end
13 for i ← 1 to n do
14 |   pClustImp(i) ← cluster(pImp(i)) ;
15 |   pClustMean(i) ← cluster(pMean(i)) ;
16 end
17 benchClust ← cluster(bench) ;
18 for i ← 1 to n do
19 |   disImp(i) ← dissimilarity(pClustImpE(i),benchClust) ;
20 |   disMean(i) ←
   |     dissimilarity(pClustMean(i),benchClust) ;
21 end
```

analysing multiple imputed datasets. This process is shown in Algorithm 1, lines 4-8.

One simple to impose missingness onto the prototypes in a way that mimics the original, is to calculate the proportion of missing values per column in the original and then delete the same proportion from the prototype. Although this is a simple method, it does rely on some assumptions. One is that it assumes no relationship between the variables. Another is that any missing data mechanisms are not analysed. To have a strong relationship between the original dataset and the benchmark, these assumptions must be analysed further and maybe extended so any underlying relationships are accounted for.

C. Imputation

By this stage, we should have an original dataset, a benchmark and multiple prototypes. The framework will now impute the prototypes, with any parameters specified by the user, independently. It is important to apply the imputation, with the same parameters, to each prototype in order to obtain reliable results. This process is shown in Algorithm 1, lines 9-12.

D. Evaluation

The final stage will involve comparing the imputed prototypes to the benchmark. This will allow us to evaluate how well imputation has performed based on how similar the imputed datasets are to the benchmark. The underlying theory is that the better the imputation, the smaller the difference between the imputed prototypes and the benchmark will be, conversely, a bigger difference implies a worse imputation

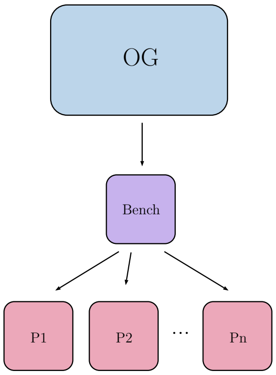

Figure. 2. Creation of Prototypes. The benchmark (Bench) is set as the subset of complete cases from the original (OG). Prototypes (P1 to Pn) are created by imposing missing values onto the bench.

One of the biggest challenges in comparing imputed datasets and the benchmark, is the problem posed by complex dataset dissimilarity measures, ie. how to define the distance between mixed data datasets. By clustering the datasets, we may be able to compare the clustering meta-data (such as cluster widths, density, size etc..) and compare the meta-data instead of the datasets. This may be possible due to the deterministic nature of clustering. If two datasets are similar, then their clustering meta-data will be similar. This process is shown in Algorithm 1, lines 13-21.

By comparing the meta-data, we mean that the (dis)similarity between clusterings can be represented by their structure. We would, for instance, create a similarity metric solely on the amount of data points per cluster (assuming the same amount of clusters) or on the density of their clusters. Thus, we could say two clusterings are similar, if their cluster sizes are similar. We could then move a step further and create a metric based on six clustering meta-data (cluster size, max dissimilarity, avg dissimilarity, cluster isolation score, individual silhouette widths and the avg silhouette width, as shown in Figure 3). We can then say two clustering are similar if their collective meta-data is similar.

IV. Framework Benefits

Our framework expands the field in a number of ways. An underlying objective throughout our work has been to strive towards the provision of labour reducing imputation evaluation software. To do so, it is necessary to establish means to automatise processes, such as creating custom benchmarks, tailored ad-hoc prototypes and using a dissimilarity measure (created from clustering meta-data) which can be applicable to a variety of data types.

We have achieved this by using the subset of complete cases as a benchmark, creating a complete dataset with similar structure as the original. Using the missingness structure from the original to create prototypes, again, ensures these datasets follow a similar structure to the original. Finally, clustering techniques are used to define a dissimilarity measure between datasets with (possibly) mixed data.

The primary goal is to either use more of the available data (by showing it responds well to imputation) or justify not using imputation (by showing that it does not respond well). Whether the amount of extra available data justifies the means, ie whether it is worth it, is subjective. Even a dataset with relatively small amounts of missing data, may benefit from such methods, alternatively, some may think that this framework is not worth the effort required when there is only a small amount of missing data. Either way, we will not know whether the partially missing data is useful until it has been evaluated.

Creating an evaluation score enables results from different evaluations to be easily compared. By comparing scores, we may be able to reinforce post-imputation analysis and potentially discover more about the relationship between missing data and imputation.

Clustering techniques could be used to create a dissimilarity measure. This makes us able to not only quantify the difference between non-numerical data, we may also be able to create a metric which can be used to compare different evaluation scores to each other.

Using an evaluation score, we are able to run the evaluation system multiple times, and can change the imputation parameters every time. From this, we may be able to optimise the imputation parameters for a given dataset by comparing the scores produced by imputing with different parameters. Although this was possible before, through regression comparisons, our framework makes it more straightforward for someone who wants to optimise their imputation methods, since the tests will be carried out autonomously.

V. Preliminary Results

We are currently implementing the framework proposed in this paper, called CLustering to Evaluate Multiple Imputation (CLEMI). CLEMI is being implemented in R, the statistical language, and we hope to make it a publicly available package/library once it is completed.

CLEMI is currently being validated using controlled tests by varing degrees of missing data and analysing the outcomes. We are also currently testing our metric, which uses clustering meta-data to find the dissimilarity between datasets. We hope to have enough results to publish shortly after the summer.

CLEMI uses MICE and Mean imputation (freely available on R), and compared the difference between 1. MICE imputed datasets and the benchmark and 2. Mean imputed datasets and the benchmark. We have used MICE as it is one of the more widely used imputation methods and we have used Mean imputation as a comparison as Mean imputation has shown to produced biased results in a lots of cases. The ultimate goal will be to find the smallest difference between MICE imputation and the benchmark (showing imputation yields similar results to the benchmark), and having results which are better than Mean imputation (if MICE produces similar results to Mean, then they are likely to produce biased results).

Figure 3 shows some preliminary user case results for CLEMI. Each of the six charts represents one part of a metric (one clustering meta-data value) used and all should be considered together for the final decision. When analysing the results, we look for the MICE box plot, blue box on the left, to be as low as possible whilst being lower than the Mean box plot, red box on the right.

This small user case uses a partially complete dataset with 10 variables which are mixed (both numerical and non-numerical data). We created 9 datasets which range from the amount of missing data we allow to remain, for example the first only have complete records and records with 9 recorded values, the second datasets contains all records with 8 recorded values or more, and so on until you have almost the original records (without records which are fully missing). From Figure 3, we can see that MICE is consistently lower than Mean and at its lowest point between 40–60% missing data within the records. So for this particular dataset, we should remove records with more than circa 60% missing values but we can impute the others; allowing us to decide how much of the available data should be used.

VI. Discussion & Limitations

The proposed framework focuses on using prototypes and clustering meta-data to evaluate imputation. Research regarding imputation evaluation is crucial to informatics and having methods to cope with missing data, when missing data is present, will only strengthen data analysis. Evaluation methods could be used to optimise imputation and improve the credibility of studies using imputation. Such methods could also be used to improve our knowledge on the effect different types and/or quantities of missing data have on imputation.

A number of limitations with the framework were identified. One limitation is that the framework assumes that the prototypes represent the original dataset and the evaluation of the prototypes will reflect the imputation of the original. Some work will be needed to show whether this assumption is acceptable or not.

A more technical limitation lies at the heart of modeling theory. Although regressions have great modeling power, they also include a degree of uncertainty. Most multiple imputation methods rely on regressions to predict the missing values, this is done multiple times to reduce uncertainty but we cannot guarantee that the regressions created for the prototypes will be exactly those which were created for the original dataset. A good imputation method relies on a good regression model, but a good regression model is not guaranteed in every run.

Using clustering, we are able to create a dissimilarity measure between the prototypes and the benchmark. However, this is not easy in practice and there are many things to consider, for example, clustering creates meta-data, which we can use to create the metric for evaluation but it may not be easily interpreted, as shown in the preliminary user case.

VII. Conclusions & Future Work

This paper has identified weaknesses in existing imputation evaluation research, which, if not addressed, could lead to

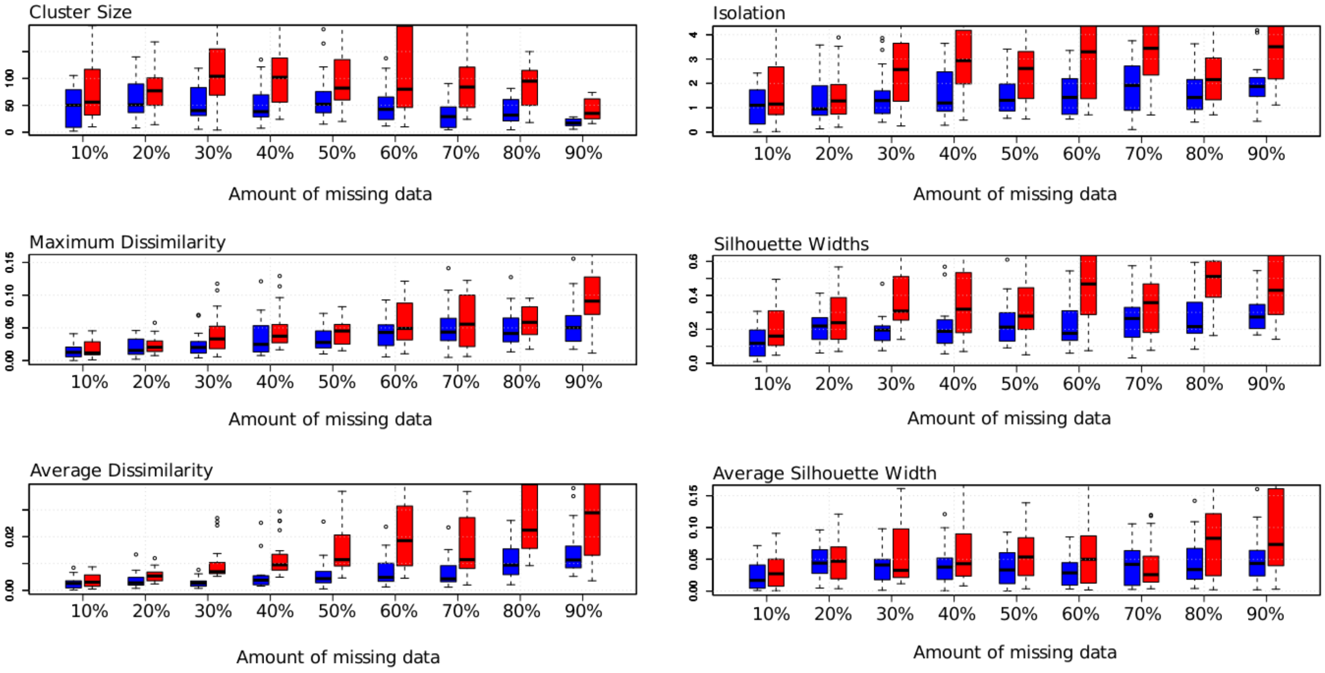

Figure. 3. Removing data (10% to 90%) to see how MICE and Mean are affected by the different amounts of missing data.

studies having a wost impact than they otherwise would have. The first problem identified is that complete case analysis seems to be the norm when faced with an incomplete dataset, this method does not use as much of the available data as possible, as illustrated in Figure 1.

Another problem identified, in current research, is the lack of imputation evaluation software, although we notice that there are many imputation methods. Finally, we identified a gap in literature for efficiently optimising imputation methods without having to create a new system for every dataset which needs to be imputed.

The proposed imputation evaluation framework may be used on a large variety of datasets, without having to manually create different methods for every evaluation. The framework includes a method for comparing the dissimilarity of datasets, efficiently, by using clustering to define a dissimilarity measure, this measure may work on both numerical and non-numerical data. Using such an evaluation method, we may be able to use more of the available data, and consequently, increase the impact from a given study.

We introduced CLEMI which will output dataset specific evaluation scores. Users will then have to decide whether the scores imply a satisfactory imputation method, for use in their studies, or not. Using an imputed dataset with a low evaluation score may lead to unreliable or biased results. The proposed framework will enable users to not only use more of the available data but even possibly strengthen the validity of their conclusions. This is especially important as we live in a world where the quantity of data being gathered may increase at a faster pace than data mining techniques and mechanisms.

Finally, future investigation could be carried out to address the already discussed limitations. To justify the prototypical nature of our framework, we might test the appropriateness of using the prototypes as a representative of the original dataset. We could do this by externally validating the similarities between the prototype datasets and the original dataset.

Using machine learning techniques, such as metric learning or feature ranking we may be able to create a standardised evaluation score, using the clustering meta-data, which is both reliable and user friendly (easy to interpret). Our initial idea is that some cluster information is more relevant than others for an evaluation score and, using machine learning techniques, we may be able to combine the information to create a score.

REFERENCES

[1] D. B. Rubin, “Inference and missing data,” *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[2] P. E. McKnight, K. M. McKnight, S. Sidani, and A. J. Figueredo, *Missing data: A gentle introduction*. Guilford Press, 2007.

[3] E. A. Akl, K. Shawwa, L. A. Kahale, T. Agoritsas, R. Brignardello-Petersen, J. W. Busse, A. Carrasco-Labra, S. Ebrahim, B. C. Johnston, I. Neumann *et al.*, “Reporting missing participant data in randomised trials: systematic survey of the methodological literature and a proposed guide,” *BMJ open*, vol. 5, no. 12, p. e008431, 2015.

[4] J. A. Hussain, M. Bland, D. Langan, M. J. Johnson, D. C. Currow, and I. R. White, “Quality of missing data reporting and handling in palliative care trials demonstrates that further development of the consort missing data reporting guidance is required: a systematic review,” *Journal of Clinical Epidemiology*, 2017.

[5] C. A. Manly and R. S. Wells, “Reporting the use of multiple imputation for missing data in higher education research,” *Research in Higher Education*, vol. 56, no. 4, pp. 397–409, 2015.

[6] I. Eekhout et al., “Missing data in a multi-item instrument were best handled by multiple imputation at the item score level,” *Journal of Clinical Epidemiology*, vol. 67, no. 3, pp. 335–342, 2014.

[7] J. L. Schafer and J. W. Graham, “Missing data: our view of the state of the art.” *Psychological methods*, vol. 7, no. 2, p. 147, 2002.

[8] W. J. Shih, “Current Controlled Trials in Problems in dealing with missing data and informative censoring in clinical trials,” vol. 7, pp. 1–7, 2002.

[9] R. Somasundaram and R. Nedunchezhian, "Evaluation of Three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values," *International Journal of Computer Applications*, vol. 21, no. 10, pp. 14–19, 2011.

[10] C. Kuchler and M. Spiess, "The data quality concept of accuracy in the context of publicly shared data sets," *AStA Wirtschafts-und Sozialstatistisches Archiv*, vol. 3, no. 1, pp. 67–80, 2009.

[11] S. P. Mandel J, "A Comparison of Six Methods for Missing Data Imputation," *Journal of Biometrics & Biostatistics*, vol. 06, no. 01, pp. 1–6, 2015.

[12] J. W. Graham, "Missing data analysis: Making it work in the real world," *Annual review of psychology*, vol. 60, pp. 549–576, 2009.

[13] R. J. Little et al., *et al.*, "The prevention and treatment of missing data in clinical trials," *New England Journal of Medicine*, vol. 367, pp. 1355–1360, 2012.

[14] R. O'neill and R. Temple, "The prevention and treatment of missing data in clinical trials: an fda perspective on the importance of dealing with it," *Clinical Pharmacology & Therapeutics*, vol. 91, pp. 550–554, 2012.

[15] V. Tresp, R. Neuneier, and S. Ahmad, "Efficient methods for dealing with missing data in supervised learning," in *Advances in neural information processing systems*, 1995, pp. 689–696.

[16] J. Osborne, *Best Practices in Data Cleaning*. Sage, 2013.

[17] M. Soley-Bori, "Dealing with missing data: Key assumptions and methods for applied analysis," *Boston University*, 2013.

[18] S. F. Messner, "Exploring the consequences of erratic data reporting for cross-national research on homicide," *Journal of Quantitative Criminology*, vol. 8, no. 2, pp. 155–173, 1992.

[19] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: Issues and guidance for practice," *Statistics in Medicine*, vol. 30, no. 4, pp. 377–399, 2011.

[20] W. J. Shih, "Problems in dealing with missing data and informative censoring in clinical trials," *Current Controlled Trials in Cardiovascular Medicine*, vol. 3, no. 1, p. 4, 2002.

[21] A. I. for Research, "Three reasons for missing data: Engaging consumers in quality information," *Robert Wood Johnson Foundation*, 2012.

[22] SPSS, "Missing Data : The Hidden Problem," *Ibm Spss*, pp. 1–8, 2009.

[23] R. Little and D. Rubin, *Statistical Analysis with Missing Data*, ser. Wiley Series in Probability and Statistics. Wiley, 2014.

[24] B. K. Vaughn, "Data analysis using regression and multilevel/hierarchical models, by gelman, a., & hill, j." *Journal of Educational Measurement*, vol. 45, no. 1, pp. 94–97, 2008.

[25] J. Scheffer, "Dealing with missing data," 2002.

[26] R. W. Wiggins, M. Ely, and K. Lynch, "A comparative evaluation of currently available software remedies to handle missing data in the context of longitudinal design and analysis," *NCDS User Support Group Working Paper 51*, pp. 1–25, 2000.

[27] N. K. Malhotra, "Analyzing marketing research data with incomplete information on the dependent variable," *Journal of Marketing Research*, pp. 74–84, 1987.

[28] J. A. C. Sterne, I. R. White, and J. B. Carlin, "Multiple imputation for missing data in epidemiological and clinical research : potential and pitfalls," pp. 1–11, 2017.

[29] K. Swarts et al., *et al.*, "Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants," *The Plant Genome*, vol. 7, no. 3, 2014.

[30] C. Westermeier and M. M. Grabka, "Longitudinal Wealth Data and Multiple Imputation: An Evaluation Study," *Survey Research Methods*, vol. 10, no. 3, pp. 237–252, 2016.

[31] O. Akande, F. Li, and J. Reiter, "An Empirical Comparison of Multiple Imputation Methods for Categorical Data," vol. 27708, pp. 1–30, 2015.

[32] J. R. van Ginkel and P. M. Kroonenberg, "Evaluation of multiple-imputation procedures for three-mode component models," *Journal of Statistical Computation and Simulation*, vol. 87, no. 16, pp. 3059–3081, 2017.

[33] J. Baker, N. White, and K. Mengersen, "Missing in space: An evaluation of imputation methods for missing data in spatial analysis of risk factors for type II diabetes," *International Journal of Health Geographics*, vol. 13, no. 1, pp. 1–13, 2014.

[34] A. K. Waljee et al., "Comparison of imputation methods for missing laboratory data in medicine," *BMJ open*, vol. 3, no. 8, p. e002847, 2013.

[35] S. Seaman, J. Bartlett, and I. White, "Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods." *BMC medical research methodology*, vol. 12, no. Mi, p. 46, 2012.

[36] N. A. Samat, M. Najib, and M. Salleh, "A Study of Data Imputation Using Fuzzy C-Means with Particle Swarm Optimization," vol. 549, no. 2, 2017.

[37] R. Veroneze, F. O. De França, and F. J. Von Zuben, "Assessing the performance of a swarm-based biclustering technique for data imputation," *2011 IEEE Congress of Evolutionary Computation, CEC 2011*, pp. 386–393, 2011.

[38] Y. Liu and V. Gopalakrishnan, "An Overview and Evaluation of Recent Machine Learning Imputation Methods Using Cardiac Imaging Data," *Data*, vol. 2, no. 1, p. 8, 2017.

[39] D. Salfr, P. Jordan, and M. Spiess, "Missing data : On criteria to evaluate imputation methods," no. 4, 2016.

[40] G. Vink, "Towards a standardized evaluation of multiple imputation routines," pp. 1–16.

[41] Y. He, A. M. Zaslavsky, M. Landrum, D. Harrington, and P. Catalano, "Multiple imputation in a large-scale complex survey: a practical guide," *Statistical methods in medical research*, vol. 19, pp. 653–670, 2010.

[42] T. Li 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," *Journal of Clinical Epidemiology*, vol. 67, no. 1, pp. 15–32, 2014.

[43] A. D. Shah et al., "Comparison of random forest and parametric imputation models for imputing missing data using mice: A caliber study," *American Journal of Epidemiology*, vol. 179, no. 7, pp. 764–774, 2014.

[44] N. Mittag, "Imputations: Benefits, Risks and a Method for Missing Data," 2013.