CLEMI: A Proof-of-Concept Computational Framework for Clustering to Evaluate Multiple Imputation

Anthony Chapman*†, Steve Turner‡, Wei Pang*, Lorna Aucott†

*School of Natural and Computing Sciences, University of Aberdeen, UK, AB24 3UE

†Institute of Applied Health Science, University of Aberdeen, UK, AB25 2ZD

‡Child Health, University of Aberdeen, UK, AB25 2ZG

Abstract

Data with missing values are ubiquitous across the world Although computational and statistical approaches have been developed to deal with missing values, a vast majority of researchers choose to carry out complete case analysis instead of using software to predict missing values. Our approach evaluates the effect of imputation by using clustering. To our knowledge there has not been any ongoing evaluation of imputation.

This paper presents a framework for a system, the purpose of which, is to allow those researchers who do not have sufficient computational or statistical analysis skills to facilitate better understanding of different imputation techniques and how they affect and contribute to data analysis. In addition, it sets guidelines on how such a system can be used to choose an appropriate imputation methods. The system could also be used to optimise an imputation method by analysing the effect changing the parameters have on the dataset.

1. Introduction

Researchers from all disciplines have to analyse raw data presented to them at some stage of their research. There are common problems in data analysis, for instance, organizing the data, dealing with outliers, missing values, data manipulation; just to name a few. When dealing with routinely acquired data[1], one of the main problems a researcher faces is what to do about missing values. Multiple imputation[2] is a possible solution for missing values, although there is work on how to impute (replace missing value with a value) values, there is no way to evaluate it's effects or efficiency. We present a way to evaluate the effect of any given imputation method and provides guidelines to make such a system accessible to all researchers regardless of their background.

Firstly, we describe some of the problems a researcher faces when analysing data with missing values. Secondly, we discuss possible solutions for dealing with missing data including ways to evaluate such solutions, we then outline a framework which is not language specific.

This paper is the first attempt in an effort to create software which supports researchers in evaluating the effect of multiple imputation, regardless of their computing and statistical abilities. We finally provide some preliminary results from our implementation which is being written in R, a statistical programming language.

2. Background

In this section we introduce some key concepts of routinely acquired data, why missing values occurs, the current practises when dealing with missing data and some ways to analyse data with missing values.

2.1. Missing Data

Our project deals with routinely acquired maternity data, like many other scientists from all disciplines[3], we encounter datasets with missing values. The amount of missingness in one of our datasets is as high as 85% of all records having one or more values missing. Before any analysis can be carried out, we need to decide on the best way to analyse data with missing values. We could ignore any records with missing

values and carry out the analysis on the subset consisting of all records with no missing values (complete case analysis). Additionally, we could impute missing values with predicted values.

After searching relevant literatures, and current systematic reviews of cohort studies [4], [5], [6] which show how hundreds of different studies dealt with missing data. We noticed that the majority (circa 70%) use complete case analysis, and some (circa 25%) do not mention whether there were any missing values or what they did if there were.

(Theory about most data possible for better analysis)

In our case, complete cases analysis on the maternity dataset will only yield 15% of the data. This means only 15% of the available data could be used for analysis. Although ignoring missing values seems to be acceptable in some cases, the systematic reviews also show studies which impute missing values prior to any analysis. Imputation has been shown to be effective[7] but it should not be taken lightly, 9% and 6% (respectively) of the studies reviewed [4], [5] used imputation methods that are known to produce biased results[8].

and depend heavily on the assumptions about the relationship of the missing data and the outcome of interest taking us back to the problem that it only represents 15% and could be biased w.r.t. any specific outcome variable

2.2. Imputation

We decided to use the statistical computing software R [9] for the analysis as we are comfortable with the language and there is a large community for support. After deciding on the language, we chose a package called MICE [10]. While MICE seems to suit our needs, we are unsure whether this imputation method (or another) will enhance our data for better analytical results. Even if this method has been proven to work on a dataset, it does not indicate that it will work on all others.

Multiple imputation works by replacing each missing value with a set of plausible values that represent the uncertainty about the right value to impute. The imputed datasets are then analysed and the results are combined. Multiple imputation does not attempt to estimate each missing value through simulated values but rather to represent a random sample of the missing values. This process results in valid statistical inferences that properly reflect the uncertainty due to missing values; for example, valid confidence intervals for parameters[2].

In order to check whether the imputation method works on a given dataset, one can first apply the method to a benchmark dataset and analyse the effects of such an action. For this process to imply the effects of the method on the given dataset, the benchmark must follow the same pattern of missingness as our dataset. Unfortunately, it is often difficult and time consuming to find such a benchmark. It would be almost impossible to find a benchmark which could also be used for a wide range of datasets. Thus, we felt that benchmark datasets should also be specific to the given dataset. One could create a benchmark by analysing the way values are missing in the dataset and applying it to a dataset which is complete. The best complete dataset to mimic the original behaviour could be a complete cases dataset extracted from the original dataset.

2.3. Clustering

The next challenge will be to interpret the effect of imputing a dataset with any given imputation method. Measure theory [11] and cluster analysis [12] provide good starting points on achieving this since measure theory will help us compare multidimensional datasets to each other and clustering can create comparable efficiency outputs from imputed datasets. Measure theory provides systematic ways in which to assign numbers to suitable members of a group, this way one can give a sense of distance in a less conventional manner (for example, what is the distance between record 5 in a dataset and the average record in the same dataset, note the dataset could include categorical variables).

Clustering is an unsupervised computation method which takes a group of items and puts them into smaller groups according to their characteristics. The idea being, all items that behave similarly will be put into a group

all other items will be in other groups. Each of these groups is called a cluster and a group of clusters is called a clustering, some cluster and clustering characteristics that might be of use are the cluster sizes (amount of objects in each cluster), maximum dissimilarity (maximum distance between an object in a cluster and the cluster's centre object), average dissimilarity (the average distance between all points in a cluster and it's centre object), the average width of each cluster in a clustering and the cluster isolation values (how separated each cluster is from the other clusters)[13].

3. The Challenges

When working with raw routinely acquired data, one of the first challenges a researcher will have to overcome is how to deal with missing values [14]. With large amounts data being collected daily [3], data corruption is inevitable and can happen at any point during data acquisition. These data corruptions may happen due to human error, computational inefficiency or other unforeseen circumstances [15]. Missing values in a dataset create an array of problems that need to be solved. Firstly, missing values could affect the results from any analysis. Secondly, a way to replace missing values needs to be chosen. Thirdly, one needs to confirm whether the chosen method will work on any dataset. Finally, we need to evaluate the chosen method to know whether it is the best way. We now discuss some of these challenges and provide possible solutions.

3.1. Incompleteness / Missing Values

Missing values are more likely the closer you access the dataset at it's source, thus raw data will have the most missing values. Although there are ways to combat missing data, such as mean-value imputation or multiple imputation [16], [17], [18], it is still quite normal for researchers to perform complete cases analysis [19], [4], [5], [6]. As an example, in [19], the authors use only 2,758 records for analysis out of the possible 44,261 mainly due to missing data, and this is a mere 6.2% out of the records available.

3.1.1. Possible Solution: Imputation. Imputation is the process of replacing missing values with substituted values. One has to be careful when imputing data as there are many techniques (default value, mean value and multiple imputation just to name a few [20]) and using them without care will lead to erroneous data analysis [21]. By creating a user-friendly system with clear guidelines on how to impute data and some explanation on how it works, we believe that researchers that would normally ignore data with missing values will be more likely to use more of the available data through imputation, thus improving the quality and credibly of their analysis.

3.2. Applicability of Imputation

After deciding that imputation is beneficial to the study, the next step will be to find an imputation method for the dataset. Muliple Imputation by Chained Equations (MICE [10]) using the computational language R [9] or the Impute Missing Values function in the statistical software SPSS [22] are two example of imputation which are accompanied by documentation to support them. The problem is, how does one know if the imputed values are representative to the truth, how does one know whether the imputed values are the correct values after applying the imputation method.

Even if the imputation method has been proven to work on a specific dataset such as [23], it is no indication it will work for any other dataset. This is due to the complex nature in which missing data is created, for example there might be a underlying reasons for the variables to be missing[24], [25].

In order to test whether an imputation method works on other datasets, one needs something to compare the results to; a benchmark. We could the analyse the effects of imputation by comparing the effects on the benchmark. Unfortunately, it is difficult to find a complete dataset to use as a benchmark which would reflect the effects imputation will have on the original dataset.

3.2.1. Possible Solution: Testing one's own data. The proposed solution is to create clone datasets by analysing the missingness characteristics from a given dataset and applying them to a new dataset created by the complete records. We can therefore create benchmark datasets which have the same characteristics as the original dataset. We now have an original dataset, a subset consisting of only the complete records (the benchmark) and artificially incomplete datasets created from the complete records with the missingness characteristics from the original.

We then have a benchmark and a testing dataset which behaves similar to the original dataset in terms of missing values. The idea being that if the imputation method works on the testing datasets, it will work on the original, and we can test whether imputation is successful by comparing to the benchmark.

3.3. Choosing the best imputation technique

The following problem applies to researchers, even those computationally competent, who wish to know whether one imputation method is better than any others. There is nothing to easily compare results from different imputation methods or the same imputation methods with slightly different parameters. The main problem arises when one tries to compare the outcomes from one method to another, here an adequate analogy would be to compare imputation method A with

method B would be like comparing chocolate with a bicycle; the outcomes might not be comparable.

A framework which can take different imputation methods and output scores in order to compare the effects of imputing a dataset is needed. It is essential that the framework can be implemented into a system accessible by all researchers.

3.3.1. Possible Solution: Comparing Imputation. In order for a researcher to be able to compare different imputation techniques on their own datasets, the outcomes of the techniques need to be compatible. A system that imputes a dataset and outputs a standardised efficiency classification which will make it easier to compare different imputation methods on the same dataset using this standardised efficiency classification.

A program which can be used by every researcher, regardless of their computing or statistical ability, to deal with missing data would be beneficial.Thus every researcher will be able to compare different imputation methods without having to understand the individual imputation technique outputs.

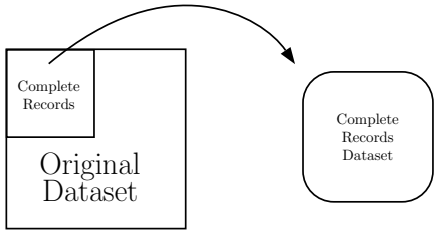
4. CLEMI: The Framework

The idea: The underlying concept is to first create a benchmark dataset from the complete records, then create artificially missing datasets which represent the original dataset out of the benchmark by mimicking the original dataset. We achieve this by analysing the pattern of data missing for a specific dataset and create testing datasets by imposing the same missingness into the benchmark. We would then impute the artificially missing datasets and analyse how far they have deviated from the benchmark. We can check how far imputation has taken the datasets from the benchmark by clustering the benchmark and the testing datasets. By doing so we will be able to see the effects of imputing a dataset.

Pre-requisite: In order to evaluate the effect of applying an imputation method on a dataset, one has to first have a dataset with missing values and choose an imputation method. At this stage, any imputation method can be used to accommodate the ability to compare different imputation methods on the same dataset to evaluate the best one. Similarly, one can apply the same imputation methods to the dataset multiple times by changing the imputation parameters, thus finding the optimal imputation. We will call the dataset O and imputation method $imp(x)$, where x is any incomplete dataset.

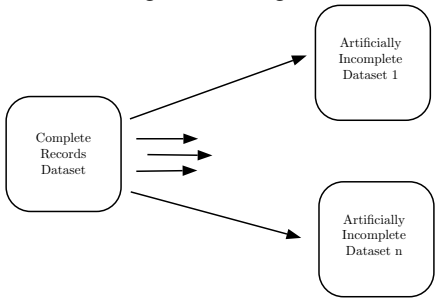
Stage 1: Extracting a Benchmark Firstly, a benchmark needs to be created, this can be done by extracting the complete records from O; call this benchmark dataset CC for Complete Cases. O can then be analysed to find the missingness characteristics, this will be used to create replicas later in the process. Notice that $CC \subset O$

Figure 1. Stage 1



Stage 2: Create Dummy Datasets Next, artificially incomplete datasets are created, called $artMiss.i$ where i is a number from 1 to n by applying the missingness characteristics from O to CC n times. It is important to apply the missingness in a manner that treats each $artMiss.i$ separately; doing so will aid in a more robust test. Thus we now have a benchmark dataset CC, and n artificial datasets with missing data which follow the same structure as the original dataset.

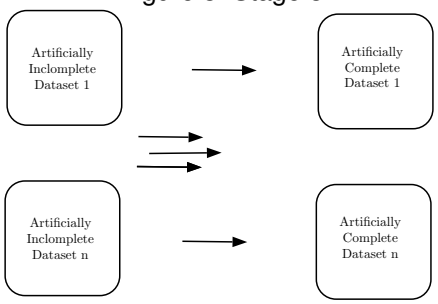
Figure 2. Stage 2



Stage 3: Impute Dummy Datasets The next step will be to impute all $artMiss.i$ with the chosen imputation method, it is important to apply exactly the same procedure (same imputation with the same parameters) to all datasets in order to have reliable results. This will create n artificially complete (imputed) datasets, called $artComp.i$ where i ranges from 1 to n.

Stage 4: Cluster all Datasets In order to evaluate the effect of an imputation method, by using clustering techniques one can evaluate the distance between all $artComp.i$ and CC. By finding the distance between sets of cluster; one can see how close (or far) an imputation method has taken each $artMiss.i$ from the true values (the benchmark CC). It will now be

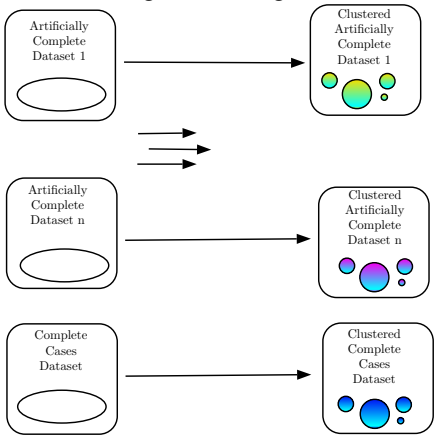
Figure 3. Stage 3



possible to compare the effect of different imputation methods.

Thus it is needed to cluster our benchmark dataset CC and all imputed datasets $artComp.i$, $clustCC$ will be the clustering of CC and $clust.i$ will be all clustered $artComp.i$. Note that as with imputing all datasets one needs to make sure the same clustering method with the same parameters are being used on all datasets. This way one can accurately calculate the distances between the clusterings.

Figure 4. Stage 4



Stage 5: Distance Between Artificial Datasets and Benchmark Finally, calculate the distance between all $clust.i$ and $clustCC$. This will indicate the effect of imputing an incomplete dataset by finding the distance between the clustering of the imputed datasets ($clust.i$) and the clustering of our benchmark ($clustCC$). The system should output an efficiency indicator to show how far away all $clust.i$ are from $clustCC$; this will be how the user judges whether an imputation method gives correct values or not. Whether the final output describes a successful or efficient imputation method will be subjective. The output should be a normalised result in order to make comparison with other imputation outcomes clearer. The aim is to make it easier for all researchers to see the effect of imputation and feel more comfortable in using data with missing values for analysis.

It will be up to the researcher to decide whether the imputed values are close enough to represent the truth or whether they are too far to provide fair results from any type of analysis.

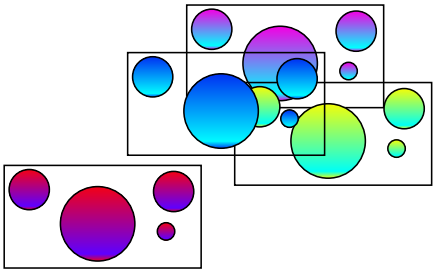


Figure 5. Stage 5

Figure 5 is a visual representation of what comparing different clustered datasets looks like. Each box contains represents a clustering and contain clusters. Using clustering characteristics, one can compare clustering and evaluate whether clustering A is more similar to clustering B than clustering C is.

5. Preliminary Results

Our current implementation, called CLEMI (CLustering to Evaluate Multiple Imputation), implemented in R will be published for the public shortly.

One of the biggest challenges in this project has been to create a normalised output which can be used to compare different methods. Such outputs can then be used to create a graph of optimal inclusion (how much data should be included for imputation; ignore any records with ¿90% missing values). We would like to find out how much missingness can be tolerated without negatively affecting our analysis.

The following preliminary results partially show how one can use the clustering characteristics to compare the effect of multiple imputation, we use mean imputation as a reference point I created the mean imputation method in R, not sure how to reference that. The system created multiple numbers of imputed datasets and after clustering them, we create a table of the averaged cluster characteristics.

The characteristics are: the amount of objects in each cluster (clust size), the maximum dissimilarity for each cluster (max-diss), the average dissimilarity for each cluster (avg-diss), the average width for each cluster

(avd-width) and the average clusters width for each clustering (avg-widths).

For the following tests, the system has created 6 artificial missing datasets by deleting values from the benchmark, it them imputed 1 with mean imputation and 5 with MICE. The system then clusters the benchmark dataset, the mean imputed dataset and all 5 MICE imputed datasets. It then combines the 5 MICE imputed clusterings and outputs the characteristics.

Benchmark: Table 1 shows the clustering characteristics for the benchmark dataset, that is the complete cases dataset.

Clust size	max-diss	avg-diss
160	5.857	2.027
128	3.947	1.897
46	3.976	2.375
Avg-width for each cluster		
0.146	0.265	0.347
Avg-widths for all clusters		
0.245		

Table 1. Benchmark Clustering Characteristics

Mean Imputed: Table 2 shows the clustering characteristics for the mean imputed dataset.

Clust size	max-diss	avg-diss
129	3.949	1.58
136	5.158	1.61
69	3.168	1.91
Avg-width for each cluster		
0.0976	0.1286	0.2294
Avg-widths for all clusters		
0.139,		

Table 2. Mean Clustering Characteristics

MICE Imputed Average: Table 3 shows the average cluster characteristics for all the imputed artificially missing datasets, these are created by applying the amount of missingness to the benchmark.

Clust size	max-diss	avg-diss
177	5.058	2.002
108	3.49	2.124
49	3.32	2.185
Avg-width for each cluster		
0.1815	0.2035	0.3265
Avg-widths for all clusters		
0.217		

Table 3. MICE Clustering Characteristics

From Tables 1, 2, and 3, mean imputation created the best clustering out of the three, this is due to the way mean imputation works. It replaces missing values by the average values for that variable, thus replacing all missing values for values which are more likely to be there, this forces all imputed values to be around the centre of every cluster, making the widths smaller and the distance between object also.

From these tables we notice that the characteristics from the MICE imputed datasets are closer to the benchmark's than the mean imputed ones are to the benchmarks. The difference between the benchmark clustering avg-clust and MICE clustering avg-clust is only 0.028, compared to the difference between the benchmark clustering avg-clust and mean clustering avg-clust which is 0.106, around 4 times bigger.

From these preliminary tests, we are able to see that imputing a dataset with MICE results in a more representative dataset than imputing with the mean.

6. Discussion

The purpose of this paper is to set a guideline for the implementation of a system for researchers to use as much of their data as possible for analysis. The aim is to motivate others to solve problems which arise from missing data as well as to provide solutions for all researchers who have similar problems.

Some work that need to be considered for the implementation are detailed in Sections 6.1 - 4.4:

6.1. Outcomes

The outcome from this evaluation process should be a number which can be used to compare different imputation methods without any further need to be manipulated. By this we mean that the output should be some normalised number, percentages are used throughout the country, so they could be a good choice. This will make it easier to compare different imputation methods which would normally have incomparable outputs.

6.2. Distance Between Clusterings

In order to evaluate the effect of imputation, we need to see "how far" the imputed dataset has travelled from the benchmark. Using clustering, one can view the differences between the clusterings (not individual clusters), the difference between the benchmark clusterings and the imputed data clusterings.

Some of the main clustering characteristics which might be useful for this could be, cluster sizes, cluster

medoids, average dissimilarities and isolation values to name a few. These are safe choices as they can be used to calculate distance measures between clusterings.

6.3. Reference Point

Another thing to consider is the use of a reference to evaluate the imputation method. A reference point is needed to determine whether a given imputed dataset can be classified as "successfully imputed dataset" or not. One could use a method which is know to be bad, such as mean imputation which can severely distort the distribution for this variable [8] and use this to see whether the imputation method in question is adequate.

6.4. Extra validation

Another possible way to evaluate the effect of imputing a dataset is to create regression models for each dataset (the benchmark, the imputed dataset and the reference dataset) and compare the models. If one created a model for one of the datasets and applies the same technique to the other datasets, the models should reflect any differences between the datasets. By examining how close the model from the imputed dataset is to the benchmark, the user can decide whether the imputation method can be classified as successful or not.

7. Conclusion and Future Work

Analysis of data with missing values has plagued researchers since the beginning of data collection. Most researchers, especially the less computing inclined, carry out complete case analysis which do not take advantage of all the data available. Doing so may lead to biased analysis or incorrect conclusions.

This paper discusses a framework which uses current imputation techniques on data with missing values and evaluate the effect of such imputation. As a result, anyone is be able to create a program which chooses the best imputation method for a given dataset as well as the optimal parameters for a given method. It could also be used to find the limit of missing data which should be allowed to remain (records with a certain amount of missing values will be omitted from analysis).

This paper can be used as a guideline for anyone who wishes to create a system to evaluate the effect of imputation. This paper also mentions key factors (ability to have any imputation method, change method parameters and output a standardised result) for such a system but has not exhausted them. This paper will be followed by an implementation of this framework, written in R, called CLEMI (Clustering to Evaluate Multiple Imputation). The following paper will explain our implementation choices as well as the testing and evaluation.

References

[1] K. Y. E. Leung, F. van der Lijn, H. A. Vrooman, W. Niessen, and M. C. J. M. Sturkenboom, *IT infrastructure to support secondary use of routinely acquired clinical imaging data for research*. EPOS, 2013.

[2] Y. C. Yuan, "Multiple imputation for missing data: Concepts and new development (version 9.0)," *SAS Institute Inc, Rockville, MD*, vol. 49, 2010.

[3] ScienceDaily, "Big data, for better or worse." 2013 (accessed: January 18, 2016), http://www.sciencedaily.com/releases/2013/05/130522085217.htm.

[4] A. Karahalios, L. Baglietto, J. B. Carlin, D. R. English, and J. A. Simpson, "A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures," *BMC Medical Research Methodology*, 2012.

[5] K. L. Masconi, T. E. Matsha, J. B. Echouffo-Tcheugui, R. T. Erasmus, and A. P. Kengnecorresponding, "Reporting and handling of missing data in predictive research for prevalent undiagnosed type 2 diabetes mellitus: a systematic review," *The EPMA Journal*, 2015.

[6] L. Tooth, R. Ware, C. Bain, D. M. Purdie, and A. Dobson, "Quality of reporting of observational longitudinal research," *PRACTICE OF EPIDEMIOLOGY*, 2005.

[7] A. Karahalios, L. Baglietto, K. J. Lee, D. R. English, J. B. Simpson, Carlin, and J. A, "The impact of missing data on analyses of a time-dependent exposure in a longitudinal cohort: a simulation study," *Emerging Themes in Epidemiology*, 2013.

[8] N. Mittag, "Imputations: Benefits, risks and a method for missing data," http://home.cerge-ei.cz/mittag/papers/Imputations.pdf, 2013.

[9] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0 http://www.R-project.org.

[10] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011, http://www.jstatsoft.org/v45/i03/.

[11] P. R. Halmos, *Measure theory*. Springer, 2013, vol. 18.

[12] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009, vol. 344.

[13] P. I. Priya and D. Ghosh, "K-means clustering algorithm characteristics differences based on distance measurement," *International journal of computer applications*, vol. 59, no. 14, 2012.

[14] A. S. Fernandes, I. H. Jarman, T. A. Etchells, J. M. Fonseca, E. Biganzoli, C. Bajdik, and P. J. G. Lisboa, "Stratification methodologies for neural networks models of survival," in *Bio-Inspired Systems: Computational and Ambient Intelligence, 10th International Work-Conference on Artificial Neural Networks, IWANN 2009, Salamanca, Spain, June 10-12, 2009. Proceedings, Part I*, 2009, pp. 989–996.

[15] T. A. Factor, "Missing data mechanisms: A primer." 2013 (accessed: February 22, 2016), http://www.theanalysisfactor.com/causes-of-missing-data/.

[16] T. D. Pigott, "A review of methods for missing data," *Educational Research and Evaluation*, vol. 7, no. 4, pp. . 353–383, 2001.

[17] D. B. Rubin, "An overview of multiple imputation."

[18] A. C. Cock, "Working with missing values," *Journal of Marriage and Family*, vol. 174, no. 67, p. 10121028, 2005.

[19] A. M. Branum, J. D. Parker, K. S. A., and S. A. H., "Prepregnancy body mass index and gestational weight gain in relation to child body mass index among siblings," *American Journal of Epidemiology*, vol. 174, no. 10, pp. 1159–1165, 2011.

[20] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

[21] M. A. Ghazanfar and A. Prugel-Bennett, "The advantage of careful imputation sources in sparse data-environment of recommender systems: Generating improved svd-based recommendations," *Informatica*, vol. 37, no. 37, pp. 61–92, (2013).

[22] S. Inc, *SPSS Statistics for Windows, Version 17.0.*, Chicago: SPSS Inc, 2008, http://www-01.ibm.com/software/uk/analytics/spss.

[23] A. D. Shah and J. W. Bartlett, "Comparison of parametric and random forest mice in imputation of missing data in survival analysis," 2014.

[24] J. A. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter, "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls," *Bmj*, vol. 338, p. b2393, 2009.

[25] L. M. Collins, J. L. Schafer, and C.-M. Kam, "A comparison of inclusive and restrictive strategies in modern missing data procedures." *Psychological methods*, vol. 6, no. 4, p. 330, 2001.