Yazdani, J Biom Biostat 2018, 9:2
DOI: 10.4172/2155-6180.1000393

Journal of Biometrics & Biostatistics

Research Article Open Access

Using Statistical Techniques and Replication Samples for Missing Values Imputation with an Application on Metabolomics

Akram Yazdani* **and Azam Yazdani**

Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, USA

***Corresponding author:** Akram Yazdani, Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, USA, Tel: 2126598542; E-mail: Akram.Yazdani@mssm.edu

Received date: Dec 18, 2017; **Accepted date:** Feb 20, 2018; **Published date:** Feb 27, 2018

Abstract

Background: Data preparation, such as missing values imputation and transformation, is the first step in any data analysis and requires crucial attention. We take advantage of availability of replication samples to identify the empirical distribution of missing values through utilization of statistical techniques. We apply these techniques to metabolomics data for imputation.

Results: Using replication samples, we obtained the empirical distribution of missing values. After application of the techniques on metabolites, we observed that the rate of missing values is approximately distributed uniformly across metabolite range. Therefore, the missing values cannot be imputed with the lowest values. To have a realistic simulation, we designed a simulation study based on empirical distribution of missing values to find an optimal imputation approach. Our findings validated the optimal approach introduced previously for metabolomics.

Conclusions: Our analysis utilized replication samples as a new approach to metabolite imputation and found empirical distribution of missing values, designed a simulation study close to reality, and compared different approaches for selecting an optimal imputation approach. The result of this study validated the optimal approach for metabolite imputation through a different data set and different approach, and the aim was to encourage researchers to pay more attention to metabolite imputation since imputing metabolomic missing values with lowest value is going to be a common approach, for example in genomic-metabolomic data analysis.

Keywords: Statistical techniques; Missing value imputation; Empirical distribution; Optimal imputation; Metabolomics; Replication samples

Background

Extracting relevant and substantial biological information from large-scale datasets at different biological levels is one of the challenges in modern biomedical research and is highly dependent on data preparation, including imputation and data transformation to hold underlying assumptions. To provide a valid analysis, missing value imputation. Using inappropriate imputation approaches results in losing power and misleads conclusions.

Missing value imputation has been addressed through a wide range of approaches, including: disregarding all variables with missing values or using univariate, multi variate, Bayesian approaches, and algorithms for imputation, e.g. [1-4]. Some studies highlight the importance of imputation of missing values in data processing pipelines by demonstrating the major effect of imputation algorithms on the outcome of data analysis, e.g. [5-7]. Here we introduce a simple technique through utilizing replication samples instead of discarding them. Replication samples happen in many biomedical and biological studies and this approach facilitates to estimate the distribution of missing values.

Missing values widely occur in mass spectrometry metabolomics datasets due to a variety of reasons, such as values that exist below the detection limit of the mass spectrometer or technical issues unrelated to the metabolite processing [1,8,9]. The first step of pre-processing raw metabolomics data includes baseline correction, noise reduction, smoothing, peak detection, and alignment [2,5,10,11]. After this pre-processing, some other steps, such as imputation and transformation, are required to prepare the data for further analysis. It has been shown that K nearest neighborhood is an optimal approach for imputation of missing values in metabolomics [5,7]. However, imputing metabolomics missing values with lowest value of measured metabolites is going to be a common approach. Using the replication samples, we obtained the empirical distribution of missing data in a metabolomics data set and showed the missing values are not necessarily low. Since we had estimated the distribution of missing values, we designed a simulation study close to real data. We found that the KNN algorithm performs better than the other approaches. The result of this study provides a validation to other studies that made the same conclusion [5,7,12] through different approaches, and is an encouragement to pay more attention to missing value imputation in metabolomics data analysis, instead of easily imputation by the lowest values.

Methods and Results

Study sample

Our metabolomics data were collected on a subset of the Atherosclerosis Risk in Communities (ARIC) study [13]. Metabolite profiling was completed in June 2010 using fasting serum samples that had been stored at -80 degrees centigrade since collection at the

baseline examination in 1987-1989. For the discovery of African-American samples, detection and quantification of metabolites was completed over 2475 individuals, using an untargeted, gas chromatography-mass spectrometry, GC-MS, and liquid chromatography-mass spectrometry, LC-MS, based metabolomics quantification protocol [14,15]. Pre-processing of the raw data, including baseline correction, noise reduction, smoothing, peak detection and alignment, was carried out by Metabolon Inc.

Blood samples were sent to laboratory for metabolomics measurement in 4-5 years apart. There was a set of 97 blood samples that were sent to laboratory twice. These 97 replication samples shared 159 metabolites. Instead of discarding the set of the replication samples, we used them to obtain the empirical distribution of missing values. Since the source of missing metabolomics data varies from biological to technical reasons, using statistical techniques here, we focused on imputation of metabolites that have 50 percent or less missing values. Since this study is based on data driven techniques, we do not bias the result and conclusions by combining large rate of missing values (>50%) that might have purely biological reasons. We introduced a technique that includes three main stages to identify an optimal approach for missing value imputation based on the data in our hand. For each stage of the analysis and the requirements, we used a particular subset of the data with specific features to efficiently use available data. These stages are:

Assessing the effect of freezing: Since these replication samples were acquired 4-5 years apart, we first determined the effect of 4-5 years longer freezing on serum metabolites. We note here that the metabolites were measured from frozen serum more than 20 years, which may have already resulted in the loss of some metabolites. However, we wanted to address whether 4-5 years longer freezing (after 20 years total freezing) has a significant impact on metabolite loss. The data set considered in this stage includes 39 metabolites measured for 97 individuals with no missing values.

Identifying the distribution of missing values: We considered 47 metabolites measured twice for 97 individuals (replication samples). Number of missing values of metabolites ranges from 1 to 55. At this stage, we pooled missing values in order to overcome the small set of missing values for each metabolite. The pooled set included 575 values to have sufficient sample size for estimating the distribution of missing values. To pool the missing data, we carried out some assessments described later.

Identifying the optimal imputation approach using identified missing values distribution: This stage of analysis includes 39 metabolites measured for 1977 individuals that are large enough for illustrating a simulation study.

Effect of freezing

By comparing the empirical distributions of 39 metabolites with no missing values in replication set, we evaluated whether the effect of longer freezing up to 5 years was significant. To see if the measured metabolites in two different time points were comparable at an individual level, we plotted the replication samples for each metabolite versus each other, Supplementary 1. The plots do not represent any significant evidence of differences or trends associated to the time points, although they might be slightly different. We employed parametric approaches for set of metabolites that were normally distributed, and nonparametric approaches for metabolites that were not normal. For the set of metabolites with a normal distribution, we applied t-test and F-test, and for the set of non-normal metabolites, we applied two-sample Kolmogorov-Smirnov Test and Wilcoxon signed-rank test to assess whether the distribution of a metabolite in replication samples was statistically the same. Using those approaches, the null hypothesis was not rejected at level 0.05, which means there is no significant difference between metabolites that are frozen for ~20 years and those that are frozen for 4-5 years longer (~25 years). Furthermore, we calculated the Kullback-Leibler divergence (KLD) presented for each metabolite in Supplementary 1 that ranges from 0.02 to 0.271. The calculated KLDs that are close to zero reveal the similar distribution of replication samples at the two time points that is in consistence with the other results and plots.

Empirical distribution of missing values

Since the missing values in each metabolite did not provide enough sample for making any conclusion, we pooled all the missing values that are observed through replication to assess how they were distributed across the range of metabolites. In order to provide sufficient conditions that allow us to pool missing values, we carried out some assessments:

The first assessment was related to distribution of metabolites. We noticed that different transformations were required to transform metabolites to normal distribution. Therefore, we selected 47 metabolites that were normally distributed using the same transformation. We then standardized the metabolites in order to pool their missing values.

The second assessment was related to the missing values that were not observed in any replication. We excluded those missing values from the analysis for this step. Table 1 shows the number of those missing values that are not observed in any replication.

Metabolite	Trehalose	Theophylline	Stearoylcarnitine	Phenyl acetate	Glycodeoxycholate
Non-observed missing values	2	4	1	3	4

Table 1: Number of missing values remained unobserved in replication samples.

We plotted empirical distribution of the metabolites for missing and observed values separately. Figure 1 demonstrates the distributions for three metabolites, while the distributions of the entire set are provided in Supplementary 2. The distributions showed that the missing values were distributed across the range of metabolites and were not only low values.

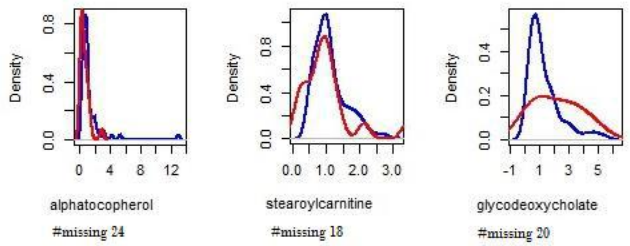

Figure 1: Blue: empirical distribution of observed metabolites. Red: empirical distribution of missing metabolites. The captions represent name of metabolites and number of missing values.

We pooled missing values across 47 metabolites normally distributed with the same transformation and estimated the distribution of missing values using Kolmogorov-Smirnov Goodness-of-fit. The p-value 0.026 [16] reveals that the missing values are approximately normally distributed, Figure 2. While the metabolites were normally distributed, we concluded that rate of missing values was approximately uniform, Figure 3, over the range of metabolite values. Although, the rate of missing values in the first quartile is slightly higher than the others.

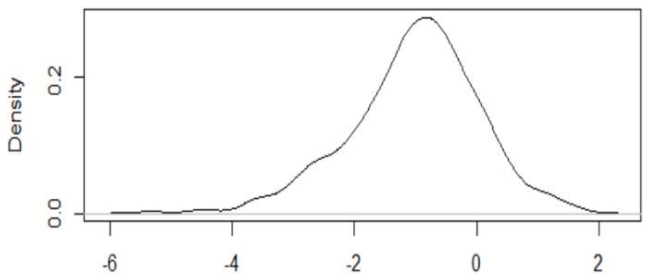

Figure 2: Distribution of missing metabolites observed through replication. Null hypothesis of normal distribution of the missing values is not rejected using Kolmogorov-Smirnov Goodness-of-fit at level 0.01 [16].

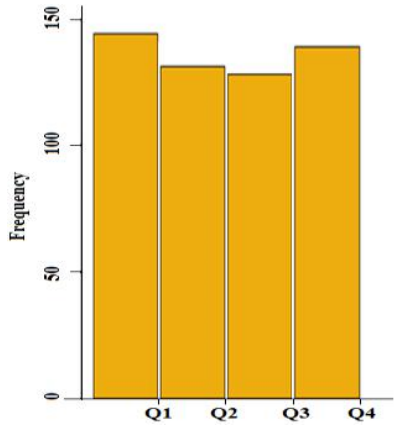

Figure3: Rate of missing value across range of metabolites. Qith in x-axis stands for ith quartile.

Hence, the assumption that metabolites are missing because they are low is not supported by this analysis. Therefore, replacing missing values with the lowest values can severely distort the distribution of metabolites and result in misleading and inaccurate conclusions.

ptimal approach for imputation

To obtain an optimal approach for metabolomics imputation, we conducted a cross validation based analysis. We used 39 metabolites measured for 1977 individuals who had no missing values. While the distribution of rate of missing values is approximately uniform, for each metabolite, we considered 14 equal intervals across the range of the metabolite and randomly selected 10% of measured metabolites in each interval (14 values) and set them as missing values. To impute those missing values, we utilized five approaches that are widely applied: Iterative Robust Model-based Imputation (IR-MI) [17], which each iteration uses one variable as an outcome and the remaining variables as predictors. Multiple Imputation (MU-IM) [18], which includes multiple imputations of incomplete multivariate data values in place of missing values by running a bootstrapped EM (expected maximization) algorithm. Maximum Likelihood estimation for multivariate normal data (ML-ES) [19], which is focused on a complete variance-covariance matrix based on maximum likelihood. K nearest neighbor (KNN) [20], which assumes data are missing at random and missing data depends on the observed data. Finally, Random Forest approach [21], which is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. For imputation using these approaches, we used the R packages irmi [22], AMELIA II [23], Mvnmle [23], SeqKnn [24], and missForest [25], respectively. The performance of these methods was evaluated in terms of mean square imputed errors (MSIE) of 40 times repeated simulation scenario. Among those methods, the KNN algorithm, which uses sequential imputation, performed better than the other methods. Although it was slightly better than RF, it significantly performed better than the other approaches. Figure 4 shows the performance of the models for K=5 as parameter of KNN, while K in (6 to 12) did not shown significant difference in our analysis.

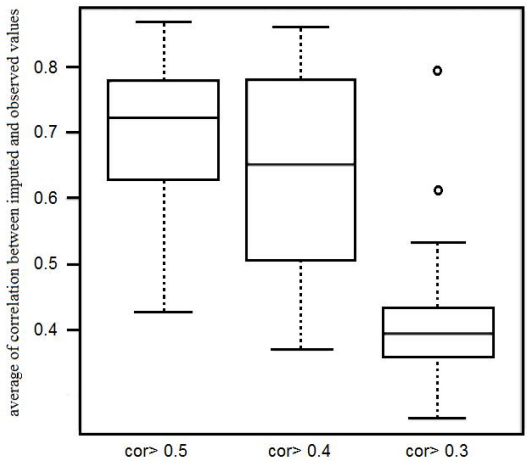

Figure 4: Performance of imputation methods as a function of MSIE. Y-axes: average of MSIE over 40 sets of simulation.

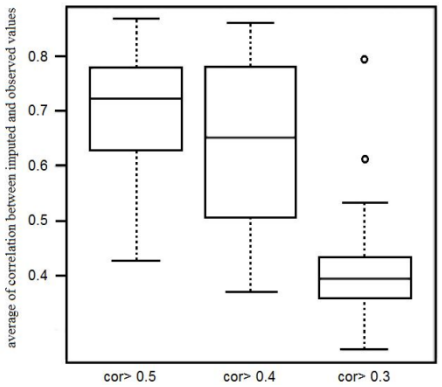

Figure 5: Performance of the KNN algorithm for different correlation cutoff (x-axes) among metabolites. Y-axes: average of correlation among imputed and observed values over 40 sets of simulation.

KNN imputation is based on Euclidean and correlation based distance metrics and is a simple but efficient approach that attempts to preserve original data structure and avoids distorting the distribution of imputed variables. The KNN approach is able to take advantage of multivariate relationships in the complete data. The algorithm starts from a complete subset of the data set, Xc, and sequentially estimates the missing values for an incomplete observation, x*, by minimizing the determinant of the covariance of the augmented data matrix, X* = [Xc; x']. Then the observation x* is added to the complete data matrix, and the algorithm continues with the next observation [26].

As mentioned, to preserve characteristics and relationship between different metabolites, KNN takes into account correlation among metabolites as similarity criteria. To assess how the correlation affects the imputation performance, we illustrated a simulation analysis for a set of metabolites that were selected with different correlation cutoff (0.5, 0.4, 0.3). Figure 5 demonstrates the average of correlation between imputed values and observed values of metabolites over 40 sets of simulation for each cutoff. Using this result, we imputed metabolites that showed at least 30% correlation with other metabolites and discarded the others from the analysis to avoid inaccurate imputation.

Discussion

Imputation of missing values is the major tasks in data pre-processing since no presentation of analysis of metabolomics data is complete without careful consideration of missing data. The aim of this manuscript was to introduce techniques for missing value imputation by taking advantage of dozens-to-hundreds of replication samples. The introduced techniques can be generalized to a variety of studies, although we applied the techniques to metabolomics data.

Missing values widely occur in mass spectrometry metabolomics datasets due to a variety of reasons, from biology to totally technical reasons [1,8,9]. After pre-processing of metabolomics data including baseline correction, noise reduction, smoothing, peak detection, and alignment [2,5,10,11] some other steps, such as imputation and transformation, are required to prepare the data for analysis. Missing value imputation has been addressed through a wide range of approaches. However, most attempts for metabolomics imputation involve easy approaches, such as using the mean, median, or lowest value of measured metabolites.

Each data set is unique, and imputation of missing values needs to be carried out carefully. Our metabolomics data set, included replication samples. Thus, we estimated the distribution of missing values in a set of 97 individuals with replication samples and noticed that rate of missing values across range of metabolites was approximately scattered as uniform. The results contradict the common belief that missing metabolite values are low values. Therefore, replacing missing data with the lowest value imposes biases to data analyses. Furthermore, based on our assessments, we observed that different transformations were required for different metabolites to be transformed to normal, and log transformation did not normalize all metabolites.

Using the empirical distribution of missing values, we conducted a simulation study close to real data to compare the performance of different imputation approaches and identify an optimal imputation method. We compared the performance of five commonly applied imputation approaches, IR-MI, MU-IM, ML-ES, KNN, and RF. Among those five approaches, the KNN algorithm showed the best performance for metabolite imputation. The KNN algorithm which

assumes data are missing at random was employed to impute the missing values in our metabolomics data set.

The KNN algorithm has already been suggested by some other studies as an optimal approach to metabolite imputation. However, the common approach utilized by metabolomics data analyzers is replacing metabolite missing values by the lowest value. Therefore, the purpose of this study was to approach metabolomics imputation with new techniques. Our findings provide another validation for the optimal approach to metabolite imputation through a different metabolite data set, ARIC metabolites, and a new approach, using replication samples.

Conclusion

Using replication samples, we could identify the empirical distribution of missing values which did not support the assumption that metabolites are missing because they are low. Therefore, replacing missing values with the lowest values can mislead the analysis and result in inaccurate conclusions. We could observe that the rate of missing values are uniformly distributed across the range of metabolites. Based on this fact, we conducted a simulation study to select the best approach for imputation. We could see the KNN algorithm performs better than some other approaches for metabolomics imputation which was a validation to some studies with the same conclusion e.g. [5,12].

References

1. Shah JS, Rai SN, DeFilippis AP, Hill BG, Bhatnagar A, et al. (2017) Distribution based nearest neighbor imputation for truncated high dimensional data with applications to pre-clinical and clinical metabolomics studies. BMC Bioinformatics 18: 114.
2. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-Mcintyre S, et al. (2011) Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat Protoc 6: 1060-1083.
3. Xia J, Psychogios N, Young N, Wishart DS (2009) MetaboAnalyst: A web server for metabolomic data analysis and interpretation. Nucleic Acids Res 37: W652-W660.
4. Bijlsma S, Bobeldijk I, Verheij ER, Ramaker R, Kochhar S, et al. (2006) Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation. Anal Chem 78: 567-574.
5. Hrydziuszko O, Viant MR (2012) Missing values in mass spectrometry based metabolomics: An undervalued step in the data processing pipeline. Metabolomics 8: 161-174.
6. Böttcher C, Lahaye ER, Willscher E, Scheel D, Clemens S (2007) Evaluation of matrix effects in metabolite profiling based on capillary liquid chromatography electrospray ionization quadrupole time of flight mass spectrometry. Anal Chem 79: 1507-1513.
7. Armitage EG, Godzien J, Alonso-Herranz V, Lopez-Gonzalvez A (2015) Missing value imputation strategies for metabolomics data. Electrophoresis 36: 3050-3060.
8. Hughes G, Cruickshank-Quinn C, Reisdorph R, Lutz S, Petrache I, et al. (2014) MSPrep-Summarization, normalization and diagnostics for processing of mass spectrometry-based metabolomic data. Bioinformatics 30: 133-134.
9. Di Guida R, Engel J, Allwood JW, Weber RJM, Jones MR, et al. (2016) Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. Metabolomics 12: 93.
10. Silva LP, Lorenzi PL, Purwaha P, Yong V, Hawke DH, et al. (2013) Measurement of DNA concentration as a normalization strategy for metabolomic data from adherent cell lines. Anal Chem 85: 9536-9542.
11. Cole RF, Mills GA, Bakir A, Townsend I, Gravell A, et al. (2016) A simple, low cost GC/MS method for the sub-nanogram per litre measurement of organotins in coastal water. MethodsX 3: 490-496.
12. Kumar N, Hoque MA, Shahjaman M, Islam SMS, Mollah MNH (2017) Metabolomic Biomarker Identification in Presence of Outliers and Missing Values. Biomed Res Int 2017: 1-11.
13. The ARIC Investigators (1989) The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. Am J Epidemiol 129: 687-702.
14. Evans AM, DeHaven CD, Barrett T, Mitchell M, Milgram E (2009) Integrated, Nontargeted Ultrahigh Performance Liquid Chromatography/ Electrospray Ionization Tandem Mass Spectrometry Platform for the Identification and Relative Quantification of the Small-Molecule Complement of Biological Systems. Anal Chem 81: 6656-6667.
15. Ohta T, Masutomi N, Tsutsui N, Sakairi T, Mitchell M, et al. (2009) Untargeted Metabolomic Profiling as an Evaluative Tool of Fenofibrate-Induced Toxicology in Fischer 344 Male Rats. Toxicol Pathol 37: 521-535.
16. Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V (2006) Compositional Data Analysis in the Geosciences: From Theory to Practice. Geological Society Special Publication. SP 264.
17. Templ M, Kowarik A, Filzmoser P (2011) Iterative stepwise regression imputation using standard and robust methods. Comput Stat Data Anal 55: 2793-2806.
18. Rubin DB (1996) Multiple Imputation after 18+ Years. J Am Stat Assoc. 91: 473-489.
19. Pinheiro JC, Bates DM (1996) Unconstrained parametrizations for variance-covariance matrices. Stat Comput 6: 289-296.
20. Fix E, Hodges JL (1951) Discriminatory Analysis - Nonparametric discrimination consistency properties. International Statistical Review/ Revue Internationale de Statistique 57: 238-247.
21. Breiman L (2001) Random Forests. Mach Learn 45: 5-32.
22. Lumley T, Miller A (2009) leaps: regression subset selection.
23. Honaker J, King G, Blackwell M (2011) AMELIA II: A Program for Missing Data. J Stat Softw 45: 1-54.
24. Horton NJ, Lipsitz SR (2001) Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables. Am Stat 55: 244-254.
25. Liaw A, Wiener M (2002) Classification and Regression by random Forest. R news 2: 18-22.
26. Verboven S, Branden K Vanden, Goos P (2007) Sequential imputation for missing values. Comput Biol Chem 31: 320-327.