Comparison of Modern Imputation Methods for Missing Laboratory Data in Medicine

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2013-002847
Article Type:	Research
Date Submitted by the Author:	07-Mar-2013
Complete List of Authors:	Waljee, Akbar; VA Center for Clinical Management Research, Ann Arbor VA Medical Center; University of Michigan, Department of Internal Medicine Mukherjee, Ashin; University of Michigan, Department of Statistics Singal, Amit; UT Southwestern Medical Center, Department of Internal Medicine Zhang, Yiwei; University of Michigan, Department of Statistics Warren, Jeffrey; University of Michigan, Pathology Balis, Ulysses; University of Michigan, Pathology Marrero, Jorge; UT Southwestern Medical Center, Department of Internal Medicine Zhu, Ji; University of Michigan, Department of Statistics Higgins, Peter; University of Michigan, Department of Internal Medicine
Primary Subject Heading:	Gastroenterology and hepatology
Secondary Subject Heading:	Diagnostics, Gastroenterology and hepatology, Health informatics
Keywords:	Gastroduodenal disease < GASTROENTEROLOGY, Information technology < BIOTECHNOLOGY & BIOINFORMATICS, Inflammatory bowel disease < GASTROENTEROLOGY

Comparison of Modern Imputation Methods for Missing Laboratory Data in Medicine

Ashin Mukherjee MS[1], Akbar K. Waljee MD MS[2,3], Amit G. Singal MD MS[4,5], Yiwei Zhang MS[1], Jeffrey Warren MD[6], Ulysses Balis MD[6], Jorge Marrero[4], Ji Zhu PhD[1], Peter DR Higgins MD PhD[2]

[1]Department of Statistics, University of Michigan, Ann Arbor, MI

[2]Department of Internal Medicine, University of Michigan, Ann Arbor, MI

[3]Veterans Affairs Center for Clinical Management Research, Ann Arbor, MI

[4]Department of Internal Medicine, UT Southwestern Medical Center, Dallas, TX

[5]Department of Clinical Sciences, UT Southwestern, Dallas, TX

[6]Department of Pathology, University of Michigan, Ann Arbor, MI

Correspondence:

Akbar K. Waljee, M.D., M.Sc.
VA Center for Clinical Management Research,
Ann Arbor VA Medical Center,
2215 Fuller Road, IIID,
Ann Arbor, MI 48105
e-mail: awaljee@med.umich.edu

Financial Support: Dr. Waljee's research is funded by a VA HSR&D CDA-2 Career Development Award 1IK2HX000775. Dr. Singal's research is funded by an ACG Junior Faculty Development Award and grant number KL2 RR024983-05. Dr. Higgins' research is supported by NIH R01 GM097117. The content is solely the responsible of the authors and does not necessarily represent the official views of UT-STAR, the University of Texas Southwestern Medical Center at Dallas and its affiliated academic and health care centers, the National Center for Advancing Translational Sciences, or the National Institutes of Health.

Acknowledgements:

Ashin Mukherjee and Akbar Waljee contributed equally to this work and should be considered co-first authors.

The content is solely the responsibility of the authors and does not necessarily reflect the official views of the Department of Veterans Affairs, the National Center for Research Resources, or the National Institutes of Health.

Conflicts of Interest: The authors disclose no conflicts.

Abstract Word Count: 191

Manuscript Word Count: 2671

Article Summary:

1) Article Focus

- Multi-analyte Assays with Algorithmic Analyses (MAAAs) are a relatively new approach to leveraging value from laboratory data to predict clinical outcomes. It is not known how robust MAAA models are when individual laboratory data points are missing.

- Recent developments in machine learning have used laboratory data to build MAAA models to predict health care outcomes - These models can be sensitive to missing laboratory data. It s not known whether modern imputation methods can robustly address the problem of missing data, and whether predictive models will remain accurate when imputed values are used.

- Multiple methods have been developed in order to deal with missing data, including single imputation, multiple imputation, multiple imputation by chained equations (MICE) , nearest neighbor estimation , and missForest . Although these have been shown to be effective with other types of missing data, little data exists regarding the absolute or comparative effectiveness of these methods in accurately imputing missing laboratory data in predictive models. The aims of our study were: (1) to compare the accuracy of four different imputation methods for missing laboratory data in two large data sets; and (2) to compare the effect of imputed values from each method on the accuracy of predictive models based on these data sets.

2) Key Message

- We found that missForest methods consistently produced the lowest imputation error and had the smallest prediction difference when models used imputed laboratory values.

- The small absolute changes in predictions with these models, despite 10-30% missing laboratory data, speak to the robustness of these multi-analyte assays with algorithmic analysis (MAAAs).

- With increasing complexity of these models, and increasing numbers of analytes, the risk of missing values increases, and methods to cope with missing values and preserve the accuracy of the model are needed. missForest appears to be a robust and accurate approach to the issue of missing laboratory values when used in these two MAAAs.

3) Limitations and Strengths

- The main limitations of missForest as a solution to missing laboratory data for predictive modeling applications are: a requirement for skilled R programming for implementation, and slightly more demanding computational needs, compared to NN or MICE methods, with minimal increases in processing time (generally in the range of 10-20 sec).

- The strength of this is that missForest methods consistently produced the lowest imputation error and had the smallest prediction difference when models used imputed laboratory values and that it is a readily available freeware R package making it a very convenient solution for any practical missing value problems.

Abstract

Background: Missing laboratory data is a common issue, but the optimal method of imputation of missing values has not been determined. The aims of our study were to compare the accuracy of four imputation methods for missing laboratory data and to compare the effect of the imputed values on the accuracy of predictive models.

Methods: Non-missing laboratory data was randomly removed with varying frequencies from two large data sets, and we then compared the ability of four methods – missForest, mean imputation, nearest neighbor imputation, and multivariate imputation by chained equations (MICE) – to impute the simulated missing data. We characterized the accuracy of the imputation and the effect of the imputation on predictive ability.

Results: MissForest had the least imputation error for both continuous and categorical variables at each frequency of missingness, and it had the smallest prediction difference when models used imputed laboratory values. In both data sets, MICE had the second least imputation error and prediction difference, followed by nearest neighbor imputation and mean imputation.

Conclusion: MissForest is a highly accurate method of imputation for missing laboratory data and outperforms other common imputation techniques in terms of imputation error and maintaining predictive ability with imputed values.

Introduction

" You can have data without information, but you cannot have information without data"

- Daniel Keys Moran

Missing data is a nearly ubiquitous problem when conducting research, particularly when using large data sets. Missing data can occur in the form of random or non-random patterns. Non-random missing data can introduce systematic error and make the study population less representative of the general population. Although random missing data does not introduce systematic error, it leads to significant loss in statistical power and predictive ability. Missing data are rarely completely at random and must be carefully managed. Multi-analyte Assays with Algorithmic Analyses (MAAAs) are a relatively new approach to leveraging value from laboratory data to predict clinical outcomes. Several of these are now available with implemented CPT codes (i.e. FibroSure, Risk of Ovarian Malignancy (ROMA), PreDx Diabetes Risk Score), but it is not known how robust MAAA models are when individual laboratory data points are missing.

Recent developments in machine learning have used laboratory data to build MAAA models to predict health care outcomes [1-3] - These models can be sensitive to missing laboratory data, which may result from hemolyzed samples, clumped platelets, or other uncommon sample or processing problems. It is not known whether modern imputation methods can robustly address the problem of missing data, and whether predictive models will remain accurate when imputed values are used.

Multiple methods have been developed in order to deal with missing data, including single imputation, multiple imputation, multiple imputation by chained equations (MICE)[3] , nearest neighbor estimation [4], and missForest[5] . Although these have been shown to be effective with other types of missing data, little data exists regarding the absolute or comparative effectiveness of these methods in accurately imputing missing laboratory data in predictive models. The aims of our study were: (1) to compare the accuracy of four different imputation methods for missing laboratory data in two large

data sets; and (2) to compare the effect of imputed values from each method on the accuracy of predictive models based on these data sets.

Methods

The University of Michigan (UM) Cirrhosis Cohort and Predictive Model for Hepatocellular Carcinoma

Between January 2004 and September 2006, consecutive patients with cirrhosis but no detectable HCC were prospectively identified and entered into a surveillance program using ultrasound and alpha fetoprotein (AFP), as has been previously described in greater detail [6]. Patients were enrolled if they had Child-Pugh class A or B cirrhosis and absence of known HCC at the time of initial evaluation. Patients diagnosed with HCC within the first six months of enrollment (prevalent cases) were excluded. Other exclusion criteria included clinical evidence of significant hepatic decompensation (refractory ascites, grade 3-4 encephalopathy, active variceal bleeding, or hepatorenal syndrome), co-morbid medical conditions with a life expectancy of less than one year, prior solid organ transplant, and a known extrahepatic primary tumor. Patients were followed until the time of HCC diagnosis, liver transplantation, death, or until the study was terminated on July 31, 2010.

The following demographic and clinical data were collected at the time of enrollment: age, gender, race, body mass index (BMI), past medical history, lifetime alcohol use, and lifetime tobacco use. Data regarding their liver disease included the underlying etiology and presence of ascites, encephalopathy, or esophageal varices. Laboratory data of interest at the time of enrollment included: platelet count, aspartate aminotransferase (AST), alanine aminotransferase (ALT), alkaline phosphatase, bilirubin, albumin, international normalized ratio (INR), and AFP. This data set was used as the basis of a published predictive model to identify patients with hepatocellular carcinoma with a c statistic of 0.70 [2].

The UM Inflammatory Bowel Disease Cohort and Predictive Model for Thiopurine Clinical Response.

The study sample included all patients who had thiopurine metabolite analysis, CBC, and a comprehensive chemistry panel drawn within a 24-hour period at the University of Michigan between May 1, 2004, and August 31, 2006 and is described in greater detail in the manuscript [1]. This study was approved by the University of Michigan Medical Institutional Review Board with a waiver of explicit consent from the subjects. The patient sample included 774 cases, in a total of 346 individuals. For the analysis of the outcome of clinical response to thiopurines, 5 exclusion criteria were applied: exclusion of patients who did not have IBD, exclusion of patients who had not started on thiopurines at the time when metabolites were measured, exclusion of patients on biologic anti-tumor necrosis factor therapy, exclusion of patients without documentation of their clinical status at the time of laboratory measurement, and exclusion of patients who had an infection that confounded assessment of clinical response. This data set was used as the basis of a predictive model to identify patients with clinical response to thiopurine immune suppressant medication with a c statistic of 0.86[1].

Description of Imputation Techniques

We compared missForest with three other commonly used imputation methods that can handle both continuous and categorical variables, namely, mean Imputation, Nearest Neighbor Imputation, and Multivariate Imputation by Chained Equations (MICE). We briefly introduce these methods below.

The recently proposed missForest method makes use of highly flexible and versatile random forest models [7, 8] to achieve missing value imputation. It creates a random forest model for each variable using the rest of the variables in the data set and uses that to predict the missing values for that variable. This is done in a cyclic fashion for all variables and the entire process is iteratively repeated until a stopping criterion is attained. The advantages of using the random forest model are that it can handle

both continuous and categorical responses, requires very little tuning, and provides an internally cross-validated error estimate. This was implemented via the 'missForest' package available in R.

Nearest Neighbor algorithms were originally proposed in the supervised pattern recognition literature. Troyanskaya et al. proposed an imputation method based on nearest neighbor search [4]. The basic idea is to compute a distance measure between each pair of observations based on the non-missing variables. Then the k-nearest observations that have non-missing values for that particular variable are used to impute a missing value via a weighted mean of the neighboring values. In order to accommodate both continuous and categorical variables the Gower distance is used [9]. For the categorical variables we imputed the missing values by weighted mode instead of a weighted mean as used for continuous variables. Cross-validation error measures are used to select the optimal number of nearest neighbors denoted by k. The function 'kNN' in R package 'VIM' was used to implement this method.

Mean imputation is one of the most naïve and easiest methods for imputing missing values. The mean (for continuous variables) or mode (for categorical variables) of the non-missing values of each variable were used to impute the missing values. This method does not take advantage of any correlation among the variables and therefore can perform rather poorly when such correlations are present.

MICE was proposed by Van Buuren et al. [3]. It requires the user to specify a conditional model for each variable, using the other variables as predictors. By default we used a linear regression model for continuous variables, a logistic regression model for binary variables, and a polytomous logistic regression for categorical variables with more than two levels. The algorithm works by iteratively imputing the missing values based on the fitted conditional models until a stopping criterion is satisfied. In that way it is very similar to the missForest algorithm; the main difference being that missForest uses more flexible decision trees for each conditional model. We implemented this in R using the package ‘mice’[10] .

Statistical Analysis

We used two separate studies to perform the comparison between the methods of imputation described in the previous section. We describe the studies, implementation details, and our results below. The structure of the statistical analysis is the same for both studies. We start with a published predictive model built with the training data set. The test set refers to observations that were not part of the training set; these were solely used for assessing the performance of the model. The test sets did not have any missing values, so we randomly removed a proportion of values to simulate data missing at random. We then imputed the missing values by the four previously discussed methods, and the imputed laboratory results were compared to the actual values that were removed from the data set. We then used the imputed data to make clinical outcome predictions with the published models, and the results were compared with the predictions made using the complete test data with no missing data. We use the average relative error (for continuous variables) and misclassification error (for categorical variables) to assess the imputation performance. To quantify the effect of the imputation on predictive models we compare the predicted classes from non-missing test data to predicted classes from imputed test data and compute the misclassification error. This is important because if a particular variable had very little influence on the predictive model, then larger imputation errors are tolerable, resulting in negligible loss of prediction accuracy. On the other hand, small imputation errors in very important variables might lead to significantly different predicted class, which is of greater clinical concern. We varied the frequency of missing values to change the difficulty of the imputation problem. We report the average results over multiple random runs. We found that the Nearest Neighbor results are quite robust to the choice of number of nearest neighbors (k) if k is moderately large, therefore we fixed the number of nearest neighbors at 5 in both studies.

Results

Cirrhosis Cohort and HCC model

This study evaluated the effect of imputation on a published predictive model for HCC based on 21 predictor variables that included demographic, clinical and laboratory values using random forest modeling [2]. The random forest model was developed on a data set of 446 patients collected at University of Michigan (UM cohort). It proved to be more accurate than traditional logistic regression models. Of the 21 variables, 10 of them were categorical in nature while 11 were continuous. We used the first 200 observations from the publicly available data set from the HALT-C trial as our test set and randomly replaced 10%, 20%, or 30% of the observations with missing values. The process was repeated for 30 replications and we report the average results.

The accuracy of the four imputation methods is compared in Figure 1. The vertical axis plots the percentage relative error for continuous variables and percentage misclassification error for continuous variables, while the horizontal axis groups the results according to the proportion of missing values. Each boxplot represent the error measure over 30 random replications. As expected, the imputation error increases on an average as we increase the proportion of missing values in the test data but the variation tends to reduce slightly which is due to averaging over many more missing observations. MissForest has the least imputation error for both continuous and categorical variables at each level of missing proportion, followed by MICE, NN, and mean imputation of continuous laboratory values. MICE and NN have similar imputation accuracy for categorical variables.

In Figure 2 the vertical axis plots the error measure for imputation on predictive model, obtained by comparing the predicted classes (Low Risk/High Risk) for each test observation with no-missing values against the predicted class after imputing the artificial missing values. Therefore an error measure of 5 on the vertical axis implies that 5% of the test observations had their predicted classes wrongly switched (either low risk => high risk or high risk => low risk) due to the imputation. As above, each

boxplot reflects the results of 30 random runs. It is clear from the figure that missForest performs the best with NN and MICE following closely. The gap increases as we increase the proportion of missing observations making the problem harder.

Inflammatory Bowel Disease Cohort and Thiopurine Clinical Response Model

Waljee et al. showed that random forest models using laboratory values outperform 6-thioguanine (6-TGN) metabolite tests as well as traditional logistic regression models in predicting clinical response to thiopurines [1]. The analysis was carried out on a data set collected at University of Michigan that included 395 patients. 26 variables, which included 25 laboratory values and age, were used to predict the clinical response of each patient. In this study all of the variables were continuous in nature. To create a separate test set, we split the data set randomly into a training set consisting of 250 observations and a test set of 145 observations, using stratified sampling to keep the ratio of clinical responder to non-responders fixed. We then introduced random missing values into the test set as before and performed the same comparative study of the four imputation methods. The whole process was replicated 30 times to obtain stable results. Below we summarize our findings via boxplots.

Again in this study, missForest beats its competitors in both imputation accuracy and the effect of imputed values on the accuracy of clinical predictions. The trends remain the same for imputation error with MICE coming out second best followed by NN and mean imputation. For predictive accuracy we find the relative order becomes missForest > MICE > mean imputation > NN. This also shows that best method with respect to imputation error need not be the best when we consider the effect of imputation on predictive models. The performance gap between missForest and MICE is considerably lower than in the previous study. This might be explained by the fact that in the thiopurine study, both the training and test sets came from the same cohort, as we generated the training and test sets by random splits, while in HCC study the training and test sets were completely different cohorts leading to an extra degree of variation.

Discussion

We have performed an extensive simulation study using two actual datasets and two published predictive models to compare the performance of four methods of missing value imputation. We compared four popular methods namely, missForest, Nearest Neighbor, MICE and mean imputation in two studies simulating data missing at random. We found that missForest methods consistently produced the lowest imputation error and had the smallest prediction difference when models used imputed laboratory values. In addition, the ready availability of the freeware R package makes missForest a very convenient solution for any practical missing value problems. The main limitations of missForest as a solution to missing laboratory data for predictive modeling applications are: a requirement for skilled R programming for implementation, and slightly more demanding computational needs, compared to NN or MICE methods, with minimal increases in processing time (generally in the range of 10-20 sec)

The small absolute changes in predictions with these models, despite 10-30% missing laboratory data, speak to the robustness of these multi-analyte assays with algorithmic analysis (MAAAs). MAAAs are currently a hot topic, and several have been released with CPT codes in 2012. One example is the HCV FibroSure (LabCorp, code 0001M) which uses ALT, alpha 2 macroglobulin, apolipoprotein A1, total bilirubin, GGT, and haptoglobin to estimate fibrosis and necroinflammatory activity in the liver in patients with hepatitis C. With increasing complexity of these models, and increasing numbers of analytes, the risk of missing values increases, and methods to cope with missing values and preserve the accuracy of the model are needed. missForest appears to be a robust and accurate approach to the issue of missing laboratory values when used in these two MAAAs.

Contributorship:

Ashin Mukherjee – study concept and design, statistical analysis and interpretation of the data, drafting of the manuscript, critical revision of the manuscript
Akbar Waljee - study concept and design, acquisition of the data, statistical analysis and interpretation of the data, drafting of the manuscript, critical revision of the manuscript and study supervision.
Amit Singal - acquisition of data, drafting of the manuscript and critical revision of the manuscript.
Yiwei Zhang - statistical analysis and interpretation of the data and critical revision of the manuscript.
Jeffrey Warren - study concept and design, critical revision of the manuscript.
Ulysses Balis - study concept and design, critical revision of the manuscript.
Jorge Marrero - study concept and design, critical revision of the manuscript.
Ji Zhu - statistical analysis and interpretation of the data and critical revision of the manuscript.
Peter Higgins - study concept and design, acquisition of the data, statistical analysis and interpretation of the data, critical revision of the manuscript and study supervision.

Competing Interests: None

Ethics approval: University of Michigan IRB

Figure Legends

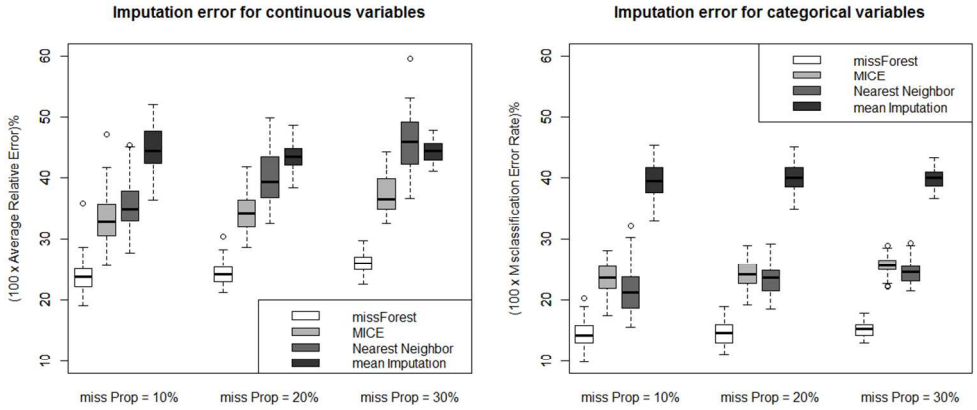
FIGURE 1. Imputation error comparison for categorical and continuous variables for four competing imputation methods at three levels of the proportion of missing values for the HCC study.

FIGURE 2. Percentage of wrongly predicted observations after missing value imputation by the four competing methods at three levels of missing value proportions in the test data.

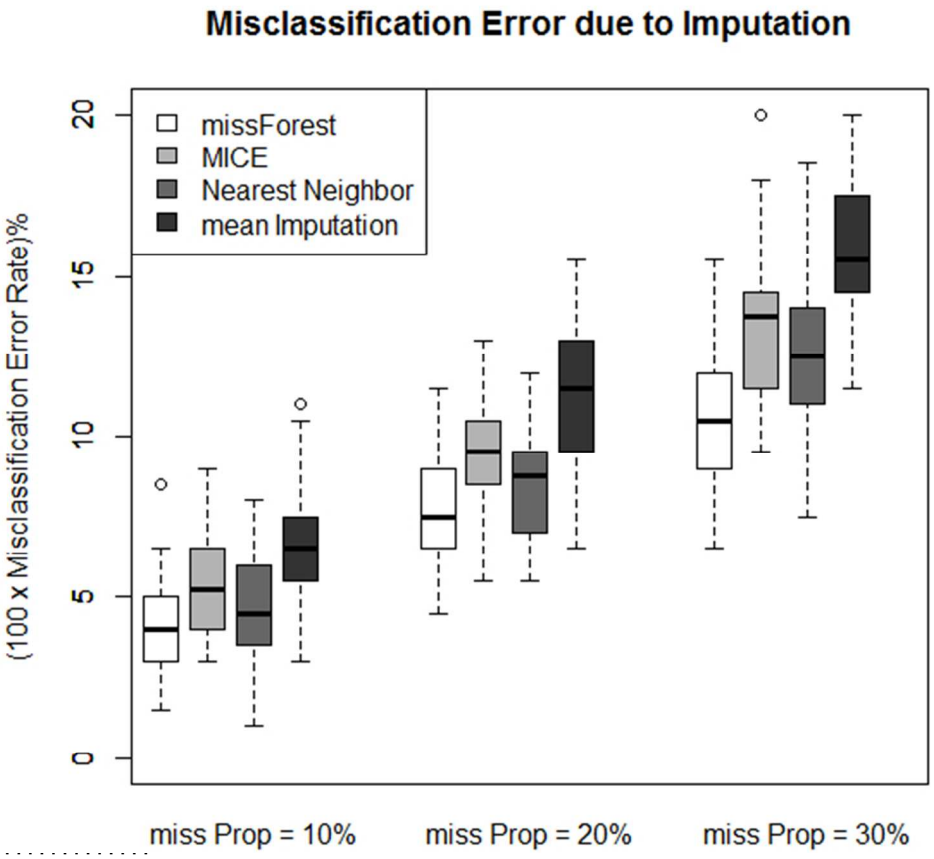
FIGURE 3. Left: Imputation error comparison, Right: Misclassification error in predicted classes due to imputation for four competing imputation methods at three levels of the proportion of missing values for the Thiopurine study.

References

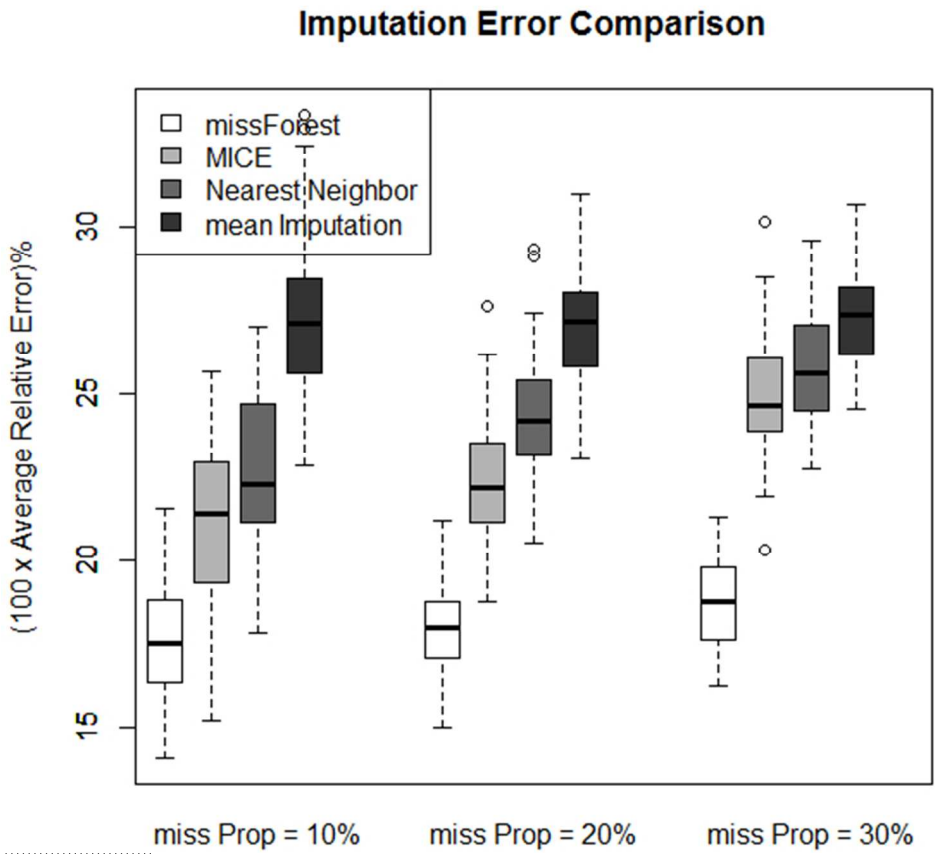
1 Waljee AK, Joyce JC, Wang S, et al. Algorithms outperform metabolite tests in predicting response of patients with inflammatory bowel disease to thiopurines. Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association. 2010 Feb;**8**(2):143-50.
2 Singal AG, Waljee AK, Mukherjee A, Higgins PD, Zhu J, Marrero JA. Machine Learning Algorithms Outperform Conventional Regression Models in Identifying Risk Factors for Hepatocellular Carcinoma in Patients With Cirrhosis. Gastroenterology. 2012 May;**142**(5):S984-S.
3 van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999 Mar 30;**18**(6):681-94.
4 Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001 Jun;**17**(6):520-5.
5 Stekhoven DJ, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012 Jan 1;**28**(1):112-8.
6 Singal AG, Conjeevaram HS, Volk ML, et al. Effectiveness of hepatocellular carcinoma surveillance in patients with cirrhosis. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2012 May;**21**(5):793-9.
7 Liaw A, Wiener M. Classification and Regression by randomForest. 2002;**2**(3):18-22.
8 Breiman L. Random forests. 2001;**45**(1):5-32.
9 Gower JC. A General Coefficient of Similarity and Some of Its Properties. Biometrics. 1971;**27**(4):857-71.
10 Buuren van S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;**45**(3).



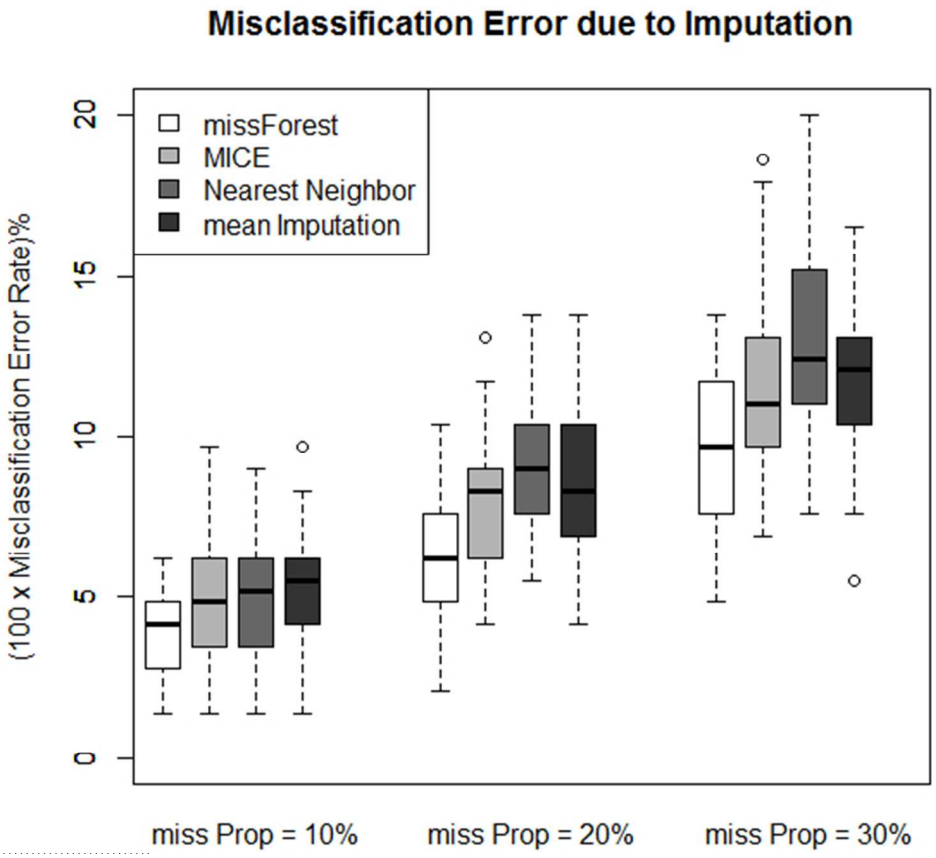
430x204mm (72 x 72 DPI)



204x204mm (72 x 72 DPI)



204x204mm (72 x 72 DPI)



204x204mm (72 x 72 DPI)

Comparison of Modern Imputation Methods for Missing Laboratory Data in Medicine

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2013-002847.R1
Article Type:	Research
Date Submitted by the Author:	15-May-2013
Complete List of Authors:	Waljee, Akbar; VA Center for Clinical Management Research, Ann Arbor VA Medical Center; University of Michigan, Department of Internal Medicine Mukherjee, Ashin; University of Michigan, Department of Statistics Singal, Amit; UT Southwestern Medical Center, Department of Internal Medicine Zhang, Yiwei; University of Michigan, Department of Statistics Warren, Jeffrey; University of Michigan, Pathology Balis, Ulysses; University of Michigan, Pathology Marrero, Jorge; UT Southwestern Medical Center, Department of Internal Medicine Zhu, Ji; University of Michigan, Department of Statistics Higgins, Peter; University of Michigan, Department of Internal Medicine
Primary Subject Heading:	Gastroenterology and hepatology
Secondary Subject Heading:	Diagnostics, Gastroenterology and hepatology, Health informatics
Keywords:	Gastroduodenal disease < GASTROENTEROLOGY, Information technology < BIOTECHNOLOGY & BIOINFORMATICS, Inflammatory bowel disease < GASTROENTEROLOGY

Comparison of Modern Imputation Methods for Missing Laboratory Data in Medicine

Akbar K. Waljee MD MS[2,3], Ashin Mukherjee MS1, Amit G. Singal MD MS[4,5], Yiwei Zhang MS[1], Jeffrey Warren MD[6], Ulysses Balis MD[6], Jorge Marrero[4], Ji Zhu PhD[1], Peter DR Higgins MD PhD[2]

[1]Department of Statistics, University of Michigan, Ann Arbor, MI

[2]Department of Internal Medicine, University of Michigan, Ann Arbor, MI

[3]Veterans Affairs Center for Clinical Management Research, Ann Arbor, MI

[4]Department of Internal Medicine, UT Southwestern Medical Center, Dallas, TX

[5]Department of Clinical Sciences, UT Southwestern, Dallas, TX

[6]Department of Pathology, University of Michigan, Ann Arbor, MI

Correspondence:

Akbar K. Waljee, M.D., M.Sc.

VA Center for Clinical Management Research,

Ann Arbor VA Medical Center,

2215 Fuller Road, IIID,

Ann Arbor, MI 48105

e-mail: awaljee@med.umich.edu

Financial Support: Dr. Waljee's research is funded by a VA HSR&D CDA-2 Career Development Award 1IK2HX000775. Dr. Singal's research is funded by an ACG Junior Faculty Development Award and grant number KL2 RR024983-05. Dr. Higgins' research is supported by NIH R01 GM097117. The content is solely the responsible of the authors and does not necessarily represent the official views of

UT-STAR, the University of Texas Southwestern Medical Center at Dallas and its affiliated academic and health care centers, the National Center for Advancing Translational Sciences, or the National Institutes of Health.

Acknowledgements:

Ashin Mukherjee and Akbar Waljee contributed equally to this work and should be considered co-first authors.

The content is solely the responsibility of the authors and does not necessarily reflect the official views of the Department of Veterans Affairs, the National Center for Research Resources, or the National Institutes of Health.

Conflicts of Interest: The authors disclose no conflicts.

Abstract Word Count: 211

Manuscript Word Count: 2843

Article Summary:

1) Article Focus

- Multi-analyte Assays with Algorithmic Analyses (MAAAs) are a relatively new approach to leveraging value from laboratory data to predict clinical outcomes. It is not known how robust MAAA models are when individual laboratory data points are missing.

- Recent developments in machine learning have used laboratory data to build MAAA models to predict health care outcomes - These models can be sensitive to missing laboratory data. It s not known whether modern imputation methods can robustly address the problem of missing data, and whether predictive models will remain accurate when imputed values are used.

- Multiple methods have been developed in order to deal with missing data, including single imputation, multiple imputation, multivariate imputation by chained equations (MICE), nearest neighbor estimation, and missForest. Although these have been shown to be effective with other types of missing data, little data exists regarding the absolute or comparative effectiveness of these methods in accurately imputing missing completely at random laboratory data in predictive models. The aims of our study were: (1) to compare the accuracy of four different imputation methods for missing laboratory data in two large data sets; and (2) to compare the effect of imputed values from each method on the accuracy of predictive models based on these data sets.

2) Key Message

- We found that missForest methods consistently produced the lowest imputation error and had the smallest prediction difference when models used imputed laboratory values.
- The small absolute changes in predictions with these models, despite 10-30% missing laboratory data, speak to the robustness of these multi-analyte assays with algorithmic analysis (MAAAs).
- With increasing complexity of these models, and increasing numbers of analytes, the risk of missing values increases, and methods to cope with missing values and preserve the accuracy of the model are needed. missForest appears to be a robust and accurate approach to the issue of missing laboratory values when used in these two MAAAs.

3) Limitations and Strengths

- The main limitations of missForest as a solution to missing laboratory data for predictive modeling applications are: a requirement for skilled R programming for implementation, and slightly more demanding computational needs, compared to NN or MICE methods.
- The simulations in this manuscript use data missing at random. The results presented here may not be generalizable to situations in which laboratory values are missing in a bias, non-random way.
- The strength of this is that missForest methods consistently produced the lowest imputation error and had the smallest prediction difference when models used imputed laboratory values and that it is a readily available freeware R package making it a very convenient solution for any practical missing value problems.

Abstract

Background: Missing laboratory data is a common issue, but the optimal method of imputation of

missing values has not been determined. The aims of our study were to compare the accuracy of four imputation methods for missing completely at random laboratory data and to compare the effect of the imputed values on the accuracy of two clinical predictive models.

Methods: Non-missing laboratory data was randomly removed with varying frequencies from two large data sets, and we then compared the ability of four methods – missForest, mean imputation, nearest neighbor imputation, and multivariate imputation by chained equations (MICE) – to impute the simulated missing data. We characterized the accuracy of the imputation and the effect of the imputation on predictive ability in two large datasets.

Results: MissForest had the least imputation error for both continuous and categorical variables at each frequency of missingness, and it had the smallest prediction difference when models used imputed laboratory values. In both data sets, MICE had the second least imputation error and prediction difference, followed by nearest neighbor imputation and mean imputation.

Conclusion: MissForest is a highly accurate method of imputation for missing laboratory data and outperforms other common imputation techniques in terms of imputation error and maintaining predictive ability with imputed values in two clinical predicative models.

Introduction

" You can have data without information, but you cannot have information without data"

- Daniel Keys Moran

Missing data is a nearly ubiquitous problem when conducting research, particularly when using large data sets. Missing data can occur in the form of random or non-random patterns. Non-random missing data can introduce systematic error and make the study population less representative of the general population. Although random missing data does not introduce systematic error, it leads to significant loss in statistical power and predictive ability. Missing data are rarely completely at random and must be carefully managed. Multi-analyte Assays with Algorithmic Analyses (MAAAs) are a relatively new approach to leveraging value from laboratory data to predict clinical outcomes. Several of these are now available with implemented CPT codes (i.e. FibroSure, Risk of Ovarian Malignancy (ROMA), PreDx Diabetes Risk Score), but it is not known how robust MAAA models are when individual laboratory data points are missing.

Recent developments in machine learning have used laboratory data to build MAAA models to predict health care outcomes [1-3] - These models can be sensitive to missing completely at random laboratory data, which may result from hemolyzed samples, clumped platelets, or other uncommon sample or processing problems. It is not known whether modern imputation methods can robustly address the problem of missing data, and whether predictive models will remain accurate when imputed values are used.

Multiple methods have been developed in order to deal with missing data, including single imputation, multiple imputation, multivariate imputation by chained equations (MICE)[3] , nearest neighbor estimation [4], and missForest[5] . Although these have been shown to be effective with other types of missing data, little data exists regarding the absolute or comparative effectiveness of these methods in accurately imputing missing laboratory data in predictive models. The aims of our study were: (1) to compare the accuracy of four different imputation methods for missing completely at random

laboratory data in two large data sets; and (2) to compare the effect of imputed values from each method on the accuracy of predictive models based on these data sets.

Methods

The University of Michigan (UM) Cirrhosis Cohort and Predictive Model for Hepatocellular Carcinoma

Between January 2004 and September 2006, consecutive patients with cirrhosis but no detectable HCC were prospectively identified and entered into a surveillance program using ultrasound and alpha fetoprotein (AFP), as has been previously described in greater detail [6]. Patients were enrolled if they had Child-Pugh class A or B cirrhosis and absence of known HCC at the time of initial evaluation. Patients diagnosed with HCC within the first six months of enrollment (prevalent cases) were excluded. Other exclusion criteria included clinical evidence of significant hepatic decompensation (refractory ascites, grade 3-4 encephalopathy, active variceal bleeding, or hepatorenal syndrome), co-morbid medical conditions with a life expectancy of less than one year, prior solid organ transplant, and a known extrahepatic primary tumor. Patients were followed until the time of HCC diagnosis, liver transplantation, death, or until the study was terminated on July 31, 2010.

The following demographic and clinical data were collected at the time of enrollment: age, gender, race, body mass index (BMI), past medical history, lifetime alcohol use, and lifetime tobacco use. Data regarding their liver disease included the underlying etiology and presence of ascites, encephalopathy, or esophageal varices. Laboratory data of interest at the time of enrollment included: platelet count, aspartate aminotransferase (AST), alanine aminotransferase (ALT), alkaline phosphatase, bilirubin, albumin, international normalized ratio (INR), and AFP. This data set was used as the basis of a published predictive model to identify patients with hepatocellular carcinoma with a c statistic of 0.70 [2].

The UM Inflammatory Bowel Disease Cohort and Predictive Model for Thiopurine Clinical Response. The study sample included all patients who had thiopurine metabolite analysis, CBC, and a comprehensive chemistry panel drawn within a 24-hour period at the University of Michigan between May 1, 2004, and August 31, 2006 and is described in greater detail in the manuscript [1]. This study was approved by the University of Michigan Medical Institutional Review Board with a waiver of explicit consent from the subjects. The patient sample included 774 cases, in a total of 346 individuals. For the analysis of the outcome of clinical response to thiopurines, 5 exclusion criteria were applied: exclusion of patients who did not have IBD, exclusion of patients who had not started on thiopurines at the time when metabolites were measured, exclusion of patients on biologic anti-tumor necrosis factor therapy, exclusion of patients without documentation of their clinical status at the time of laboratory measurement, and exclusion of patients who had an infection that confounded assessment of clinical response. This data set was used as the basis of a predictive model to identify patients with clinical response to thiopurine immune suppressant medication with a c statistic of 0.86[1].

Description of Imputation Techniques

We compared missForest with three other commonly used imputation methods that can handle both continuous and categorical variables, namely, mean Imputation, Nearest Neighbor Imputation, and Multivariate Imputation by Chained Equations (MICE). We briefly introduce these methods below.

The recently proposed missForest method makes use of highly flexible and versatile random forest models [7, 8] to achieve missing value imputation. It creates a random forest model for each variable using the rest of the variables in the data set and uses that to predict the missing values for that variable. This is done in a cyclic fashion for all variables and the entire process is iteratively repeated until a stopping criterion is attained. The advantages of using the random forest model are that it can handle both continuous and categorical responses, requires very little tuning, and provides an internally cross-validated error estimate. This was implemented via the 'missForest' package available in R.

Nearest Neighbor algorithms were originally proposed in the supervised pattern recognition literature. Troyanskaya et al. proposed an imputation method based on nearest neighbor search [4]. The basic idea is to compute a distance measure between each pair of observations based on the non-missing variables. Then the k-nearest observations that have non-missing values for that particular variable are used to impute a missing value via a weighted mean of the neighboring values. In order to accommodate both continuous and categorical variables the Gower distance is used [9]. For the categorical variables we imputed the missing values by weighted mode instead of a weighted mean as used for continuous variables. Cross-validation error measures are used to select the optimal number of nearest neighbors denoted by k. The function 'kNN' in R package 'VIM' was used to implement this method.

Mean imputation is one of the most naïve and easiest methods for imputing missing values. The mean (for continuous variables) or mode (for categorical variables) of the non-missing values of each variable were used to impute the missing values. This method does not take advantage of any correlation among the variables and therefore can perform rather poorly when such correlations are present.

MICE was proposed by Van Buuren et al. [3]. It requires the user to specify a conditional model for each variable, using the other variables as predictors. By default we used a linear regression model for continuous variables, a logistic regression model for binary variables, and a polytomous logistic regression for categorical variables with more than two levels. The algorithm works by iteratively imputing the missing values based on the fitted conditional models until a stopping criterion is satisfied. In that way it is very similar to the missForest algorithm; the main difference being that missForest uses more flexible decision trees for each conditional model. We implemented this in R using the package ‘mice’[10] .

Statistical Analysis

We used two separate studies to perform the comparison between the methods of imputation described in the previous section. We describe the studies, implementation details, and our results below. The structure of the statistical analysis is the same for both studies. We start with a published predictive model built with the training data set. The test set refers to observations that were not part of the training set; these were solely used for assessing the performance of the model. The test sets did not have any missing values, so we randomly removed a proportion of values to simulate data missing completely at random. We then imputed the missing values by the four previously discussed methods, and the imputed laboratory results were compared to the actual values that were removed from the data set. We then used the imputed data to make clinical outcome predictions with the published models, and the results were compared with the predictions made using the complete test data with no missing data. The prediction models were created using both logistic models and random forest models as some might argue that using random forest imputation methods might favor the random forest prediction models. We use the average relative error (for continuous variables) and misclassification error (for categorical variables) to assess the imputation performance for both the logistic and random forest models. To quantify the effect of the imputation on predictive models we compare the predicted classes from non-missing test data to predicted classes from imputed test data and compute the misclassification error for both the logistic and random forest models. This is important because if a particular variable had very little influence on the predictive model, then larger imputation errors are tolerable, resulting in negligible loss of prediction accuracy. On the other hand, small imputation errors in very important variables might lead to significantly different predicted class, which is of greater clinical concern. We varied the frequency of missing values to change the difficulty of the imputation problem. We report the average results over multiple random runs. We found that the Nearest Neighbor results are quite robust to the choice of number of nearest neighbors (k) if k is moderately large, therefore we fixed the number of nearest neighbors at 5 in both studies.

Results

Cirrhosis Cohort and HCC model

This study evaluated the effect of imputation on a published predictive model for HCC based on 21 predictor variables that included demographic, clinical and laboratory values using random forest modeling [2]. The random forest model was developed on a data set of 446 patients collected at University of Michigan (UM cohort). It proved to be more accurate than traditional logistic regression models. Of the 21 variables, 10 of them were categorical in nature while 11 were continuous. We used the first 200 observations from the publicly available data set from the HALT-C trial as our test set and randomly replaced 10%, 20%, or 30% of the observations with missing values. The process was repeated for 1000 replications and we report the average results.

The accuracy of the four imputation methods for both continuous and categorical variables are compared in Figure 1 for the cirrhosis cohort and HCC model. Figure 1A and 1B represents the logistic model and Figure 1C and 1D reflect the random forest prediction model. The vertical axis plots the percentage relative error for continuous variables and percentage misclassification error for categorical variables, while the horizontal axis groups the results according to the proportion of missing values. Each boxplot represent the error measure over 1000 random replications. As expected, the imputation error increases on an average as we increase the proportion of missing values in the test data but the variation tends to reduce slightly which is due to averaging over many more missing observations. MissForest has the least imputation error for both continuous and categorical variables at each level of missing proportion, followed by MICE, NN, and mean imputation of continuous laboratory values. MICE and NN have similar imputation accuracy for categorical variables. MissForest works well using both logistic and random forest prediction models.

In Figure 2A and 2B, which represents the logistic and random forest prediction models respectively, the vertical axis plots the error measure for imputation on predictive model, obtained by comparing the predicted classes (Low Risk/High Risk) for each test observation with no-missing values against the

predicted class after imputing the artificial missing values. Therefore an error measure of 5 on the vertical axis implies that 5% of the test observations had their predicted classes wrongly switched (either low risk => high risk or high risk => low risk) due to the imputation. As above, each boxplot reflects the results of 1000 random runs. It is clear from the figure that missForest performs the best with NN and MICE following closely. The gap increases as we increase the proportion of missing observations making the problem harder.

Inflammatory Bowel Disease Cohort and Thiopurine Clinical Response Model

Waljee et al. showed that random forest models using laboratory values outperform 6-thioguanine (6-TGN) metabolite tests as well as traditional logistic regression models in predicting clinical response to thiopurines [1]. The analysis was carried out on a data set collected at University of Michigan that included 395 patients. 26 variables, which included 25 laboratory values and age, were used to predict the clinical response of each patient. In this study all of the variables were continuous in nature. To create a separate test set, we split the data set randomly into a training set consisting of 250 observations and a test set of 145 observations, using stratified sampling to keep the ratio of clinical responder to non-responders fixed. We then introduced random missing values into the test set as before and performed the same comparative study of the four imputation methods. The whole process was replicated 1000 times to obtain stable results. Below we summarize our findings via boxplots.

Again in this study, missForest beats its competitors in both imputation accuracy and the effect of imputed values on the accuracy of clinical predictions based on the logistic model (Figure 3A) and the random forest model (Figure 3B). The trends remain the same for imputation error, Figure 4A and B representing logistic models and random forest models respectively, with MICE coming out second best followed by NN and mean imputation. For predictive accuracy we find the relative order becomes missForest > MICE > mean imputation > NN. This also shows that best method with respect to imputation error need not be the best when we consider the effect of imputation on predictive models.

The performance gap between missForest and MICE is considerably lower than in the previous study. This might be explained by the fact that in the thiopurine study, both the training and test sets came from the same cohort, as we generated the training and test sets by random splits, while in HCC study the training and test sets were completely different cohorts leading to an extra degree of variation.

Discussion

We have performed an extensive simulation study using two clinical datasets and two published predictive models to compare the performance of four methods of missing value imputation for missing data completely at random. We included both local (randomForest) and global (logistic) modeling approaches to avoid bias that might favor a local (MissForest) imputation approach. While the superiority of MissForest for imputation of missing lab values this will not be generalizable to all predictive models or datasets, this manuscript highlights the value of missForest to impute missing data. We compared four popular methods namely, missForest, Nearest Neighbor, MICE and mean imputation, in two studies simulating data missing completely at random. We found that these simulation methods consistently produced the lowest imputation error and had the smallest prediction difference when models used imputed laboratory values. In addition, the ready availability of the freeware R package makes missForest and its simulations a very convenient solution for any practical missing value problems. The main limitations of these simulations as a solution to missing laboratory data for predictive modeling applications are: a requirement for skilled R programming for implementation, and slightly more demanding computational needs, compared to NN or MICE methods. An additional limitation in this study is that these simulations did not address the issue of data missing for non-random reasons. There could be an association between the clinical outcome of interest and the missingness of certain predictors. At this point, we cannot generalize these results to situation in which data is missing for non-random reasons.

The small absolute changes in predictions with these models, despite 10-30% missing laboratory data, speak to the robustness of these multi-analyte assays with algorithmic analysis (MAAAs). MAAAs are

currently a hot topic, and several have been released with CPT codes in 2012. One example is the HCV FibroSure (LabCorp, code 0001M) which uses ALT, alpha 2 macroglobulin, apolipoprotein A1, total bilirubin, GGT, and haptoglobin to estimate fibrosis and necroinflammatory activity in the liver in patients with hepatitis C. With the increasing complexity of these models, and increasing numbers of analytes, the risk of missing completely at random values increases, and methods to cope with missing values and preserve the accuracy of the model are needed. missForest appears to be a robust and accurate approach to the issue of missing laboratory values when used in these two MAAAs and may be applicable to other datasets with missing completely at random datasets.

Contributions:

Ashin Mukherjee – study concept and design, statistical analysis and interpretation of the data, drafting of the manuscript, critical revision of the manuscript

Akbar Waljee - study concept and design, acquisition of the data, statistical analysis and interpretation of the data, drafting of the manuscript, critical revision of the manuscript and study supervision.

Amit Singal - acquisition of data, drafting of the manuscript and critical revision of the manuscript.

Yiwei Zhang - statistical analysis and interpretation of the data and critical revision of the manuscript.

Jeffrey Warren - study concept and design, critical revision of the manuscript.

Ulysses Balis - study concept and design, critical revision of the manuscript.

Jorge Marrero - study concept and design, critical revision of the manuscript.

Ji Zhu - statistical analysis and interpretation of the data and critical revision of the manuscript.

Peter Higgins - study concept and design, acquisition of the data, statistical analysis and interpretation of the data, critical revision of the manuscript and study supervision.

Competing Interests: None

Ethics approval: University of Michigan, IRB

Data Sharing: Agreement: A prediction model using the cross-sectional data from the hepatocellular carcimona cohort had currently been submitted to the American Journal of Gastroenterology and we are have a revise and resubmit

Figure Legends

FIGURE 1. Imputation error comparison for categorical and continuous variables for four competing imputation methods at three levels of the proportion of missing values for the logistic prediction model (Figure 1A and 1B) and random forest prediction model (Figure 1C and 1D) in the HCC study

FIGURE 2. Percentage of wrongly predicted observations after missing value imputation by the four competing methods at three levels of missing value proportions in the test data for the logistic prediction model (Figure 2A) and the random forest prediction model (Figure 2B) in the HCC study.

FIGURE 3. Imputation error for four competing imputation methods at three levels of the proportion of missing values for the logistic prediction model (Figure 3A) and random forest prediction model (Figure 3B) in the Thiopurine response model

Figure 4. Percentage of wrongly predicted observations after missing value imputation by the four competing methods at three levels of missing value proportions in the test data for the logistic prediction model (Figure 4A) and the random forest prediction model (Figure 4B) in the Thiopurine response model.

References

1 Waljee AK, Joyce JC, Wang S, et al. Algorithms outperform metabolite tests in predicting response of patients with inflammatory bowel disease to thiopurines. Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association. 2010 Feb;**8**(2):143-50.

2 Singal AG, Waljee AK, Mukherjee A, et al. Machine Learning Algorithms Outperform Conventional Regression Models in Identifying Risk Factors for Hepatocellular Carcinoma in Patients With Cirrhosis. Gastroenterology. 2012 May;**142**(5):S984-S.

3 van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999 Mar 30;**18**(6):681-94.

4 Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001 Jun;**17**(6):520-5.

5 Stekhoven DJ, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012 Jan 1;**28**(1):112-8.

6 Singal AG, Conjeevaram HS, Volk ML, et al. Effectiveness of hepatocellular carcinoma surveillance in patients with cirrhosis. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2012 May;**21**(5):793-9.

7 Liaw A, Wiener M. Classification and Regression by randomForest. 2002;**2**(3):18-22.

8 Breiman L. Random forests. 2001;**45**(1):5-32.

9 Gower JC. A General Coefficient of Similarity and Some of Its Properties. Biometrics. 1971;**27**(4):857-71.

10 Buuren van S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;**45**(3).

Comparison of Modern Imputation Methods for Missing Laboratory Data in Medicine

Ashin Mukherjee MS[1], Akbar K. Waljee MD MS[2,3], Amit G. Singal MD MS[4,5], Yiwei Zhang MS[1], Jeffrey Warren MD[6], Ulysses Balis MD[6], Jorge Marrero[4], Ji Zhu PhD[1], Peter DR Higgins MD PhD[2]

[1]Department of Statistics, University of Michigan, Ann Arbor, MI

[2]Department of Internal Medicine, University of Michigan, Ann Arbor, MI

[3]Veterans Affairs Center for Clinical Management Research, Ann Arbor, MI

[4]Department of Internal Medicine, UT Southwestern Medical Center, Dallas, TX

[5]Department of Clinical Sciences, UT Southwestern, Dallas, TX

[6]Department of Pathology, University of Michigan, Ann Arbor, MI

Correspondence:

Akbar K. Waljee, M.D., M.Sc.

VA Center for Clinical Management Research,

Ann Arbor VA Medical Center,

2215 Fuller Road, IIID,

Ann Arbor, MI 48105

e-mail: awaljee@med.umich.edu

Financial Support: Dr. Waljee's research is funded by a VA HSR&D CDA-2 Career Development Award 1IK2HX000775. Dr. Singal's research is funded by an ACG Junior Faculty Development Award and grant number KL2 RR024983-05. Dr. Higgins' research is supported by NIH R01 GM097117. The content is solely the responsible of the authors and does not necessarily represent the official views of

UT-STAR, the University of Texas Southwestern Medical Center at Dallas and its affiliated academic and health care centers, the National Center for Advancing Translational Sciences, or the National Institutes of Health.

Acknowledgements:

Ashin Mukherjee and Akbar Waljee contributed equally to this work and should be considered co-first authors.

The content is solely the responsibility of the authors and does not necessarily reflect the official views of the Department of Veterans Affairs, the National Center for Research Resources, or the National Institutes of Health.

Conflicts of Interest: The authors disclose no conflicts.

Abstract Word Count: 211

Manuscript Word Count: 2843

Article Summary:

1) Article Focus

- Multi-analyte Assays with Algorithmic Analyses (MAAAs) are a relatively new approach to leveraging value from laboratory data to predict clinical outcomes. It is not known how robust MAAA models are when individual laboratory data points are missing.

- Recent developments in machine learning have used laboratory data to build MAAA models to predict health care outcomes - These models can be sensitive to missing laboratory data. It s not known whether modern imputation methods can robustly address the problem of missing data, and whether predictive models will remain accurate when imputed values are used.

- Multiple methods have been developed in order to deal with missing data, including single imputation, multiple imputation, multivariate imputation by chained equations (MICE), nearest neighbor estimation, and missForest. Although these have been shown to be effective with other types of missing data, little data exists regarding the absolute or comparative effectiveness of these methods in accurately imputing missing completely at random laboratory data in predictive models. The aims of our study were: (1) to compare the accuracy of four different imputation methods for missing laboratory data in two large data sets; and (2) to compare the effect of imputed values from each method on the accuracy of predictive models based on these data sets.

2) Key Message

- We found that missForest methods consistently produced the lowest imputation error and had the smallest prediction difference when models used imputed laboratory values.
- The small absolute changes in predictions with these models, despite 10-30% missing laboratory data, speak to the robustness of these multi-analyte assays with algorithmic analysis (MAAAs).
- With increasing complexity of these models, and increasing numbers of analytes, the risk of missing values increases, and methods to cope with missing values and preserve the accuracy of the model are needed. missForest appears to be a robust and accurate approach to the issue of missing laboratory values when used in these two MAAAs.

3) Limitations and Strengths

- The main limitations of missForest as a solution to missing laboratory data for predictive modeling applications are: a requirement for skilled R programming for implementation, and slightly more demanding computational needs, compared to NN or MICE methods.
- The simulations in this manuscript use data missing at random. The results presented here may not be generalizable to situations in which laboratory values are missing in a bias, non-random way.
- The strength of this is that missForest methods consistently produced the lowest imputation error and had the smallest prediction difference when models used imputed laboratory values and that it is a readily available freeware R package making it a very convenient solution for any practical missing value problems.

Abstract

Background: Missing laboratory data is a common issue, but the optimal method of imputation of missing values has not been determined. The aims of our study were to compare the accuracy of four imputation methods for missing completely at random laboratory data and to compare the effect of the imputed values on the accuracy of two clinical predictive models.

Methods: Non-missing laboratory data was randomly removed with varying frequencies from two large data sets, and we then compared the ability of four methods – missForest, mean imputation, nearest neighbor imputation, and multivariate imputation by chained equations (MICE) – to impute the simulated missing data. We characterized the accuracy of the imputation and the effect of the imputation on predictive ability in two large datasets.

Results: MissForest had the least imputation error for both continuous and categorical variables at each frequency of missingness, and it had the smallest prediction difference when models used imputed laboratory values. In both data sets, MICE had the second least imputation error and prediction difference, followed by nearest neighbor imputation and mean imputation.

Conclusion: MissForest is a highly accurate method of imputation for missing laboratory data and outperforms other common imputation techniques in terms of imputation error and maintaining predictive ability with imputed values in two clinical predicative models.

Introduction

" You can have data without information, but you cannot have information without data"

- Daniel Keys Moran

Missing data is a nearly ubiquitous problem when conducting research, particularly when using large data sets. Missing data can occur in the form of random or non-random patterns. Non-random missing data can introduce systematic error and make the study population less representative of the general population. Although random missing data does not introduce systematic error, it leads to significant loss in statistical power and predictive ability. Missing data are rarely completely at random and must be carefully managed. Multi-analyte Assays with Algorithmic Analyses (MAAAs) are a relatively new approach to leveraging value from laboratory data to predict clinical outcomes. Several of these are now available with implemented CPT codes (i.e. FibroSure, Risk of Ovarian Malignancy (ROMA), PreDx Diabetes Risk Score), but it is not known how robust MAAA models are when individual laboratory data points are missing.

Recent developments in machine learning have used laboratory data to build MAAA models to predict health care outcomes [1-3] - These models can be sensitive to missing completely at random laboratory data, which may result from hemolyzed samples, clumped platelets, or other uncommon sample or processing problems. It is not known whether modern imputation methods can robustly address the problem of missing data, and whether predictive models will remain accurate when imputed values are used.

Multiple methods have been developed in order to deal with missing data, including single imputation, multiple imputation, multivariate imputation by chained equations (MICE)[3] , nearest neighbor estimation [4], and missForest[5] . Although these have been shown to be effective with other types of missing data, little data exists regarding the absolute or comparative effectiveness of these methods in accurately imputing missing laboratory data in predictive models. The aims of our study were: (1) to compare the accuracy of four different imputation methods for missing completely at random

laboratory data in two large data sets; and (2) to compare the effect of imputed values from each method on the accuracy of predictive models based on these data sets.

Methods

The University of Michigan (UM) Cirrhosis Cohort and Predictive Model for Hepatocellular Carcinoma

Between January 2004 and September 2006, consecutive patients with cirrhosis but no detectable HCC were prospectively identified and entered into a surveillance program using ultrasound and alpha fetoprotein (AFP), as has been previously described in greater detail [6]. Patients were enrolled if they had Child-Pugh class A or B cirrhosis and absence of known HCC at the time of initial evaluation. Patients diagnosed with HCC within the first six months of enrollment (prevalent cases) were excluded. Other exclusion criteria included clinical evidence of significant hepatic decompensation (refractory ascites, grade 3-4 encephalopathy, active variceal bleeding, or hepatorenal syndrome), co-morbid medical conditions with a life expectancy of less than one year, prior solid organ transplant, and a known extrahepatic primary tumor. Patients were followed until the time of HCC diagnosis, liver transplantation, death, or until the study was terminated on July 31, 2010.

The following demographic and clinical data were collected at the time of enrollment: age, gender, race, body mass index (BMI), past medical history, lifetime alcohol use, and lifetime tobacco use. Data regarding their liver disease included the underlying etiology and presence of ascites, encephalopathy, or esophageal varices. Laboratory data of interest at the time of enrollment included: platelet count, aspartate aminotransferase (AST), alanine aminotransferase (ALT), alkaline phosphatase, bilirubin, albumin, international normalized ratio (INR), and AFP. This data set was used as the basis of a published predictive model to identify patients with hepatocellular carcinoma with a c statistic of 0.70 [2].

The UM Inflammatory Bowel Disease Cohort and Predictive Model for Thiopurine Clinical Response.

The study sample included all patients who had thiopurine metabolite analysis, CBC, and a comprehensive chemistry panel drawn within a 24-hour period at the University of Michigan between May 1, 2004, and August 31, 2006 and is described in greater detail in the manuscript [1]. This study was approved by the University of Michigan Medical Institutional Review Board with a waiver of explicit consent from the subjects. The patient sample included 774 cases, in a total of 346 individuals. For the analysis of the outcome of clinical response to thiopurines, 5 exclusion criteria were applied: exclusion of patients who did not have IBD, exclusion of patients who had not started on thiopurines at the time when metabolites were measured, exclusion of patients on biologic anti-tumor necrosis factor therapy, exclusion of patients without documentation of their clinical status at the time of laboratory measurement, and exclusion of patients who had an infection that confounded assessment of clinical response. This data set was used as the basis of a predictive model to identify patients with clinical response to thiopurine immune suppressant medication with a c statistic of 0.86[1].

Description of Imputation Techniques

We compared missForest with three other commonly used imputation methods that can handle both continuous and categorical variables, namely, mean Imputation, Nearest Neighbor Imputation, and Multivariate Imputation by Chained Equations (MICE). We briefly introduce these methods below.

The recently proposed missForest method makes use of highly flexible and versatile random forest models [7, 8] to achieve missing value imputation. It creates a random forest model for each variable using the rest of the variables in the data set and uses that to predict the missing values for that variable. This is done in a cyclic fashion for all variables and the entire process is iteratively repeated until a stopping criterion is attained. The advantages of using the random forest model are that it can handle both continuous and categorical responses, requires very little tuning, and provides an internally cross-validated error estimate. This was implemented via the 'missForest' package available in R.

Nearest Neighbor algorithms were originally proposed in the supervised pattern recognition literature. Troyanskaya et al. proposed an imputation method based on nearest neighbor search [4]. The basic idea is to compute a distance measure between each pair of observations based on the non-missing variables. Then the k-nearest observations that have non-missing values for that particular variable are used to impute a missing value via a weighted mean of the neighboring values. In order to accommodate both continuous and categorical variables the Gower distance is used [9]. For the categorical variables we imputed the missing values by weighted mode instead of a weighted mean as used for continuous variables. Cross-validation error measures are used to select the optimal number of nearest neighbors denoted by k. The function 'kNN' in R package 'VIM' was used to implement this method.

Mean imputation is one of the most naïve and easiest methods for imputing missing values. The mean (for continuous variables) or mode (for categorical variables) of the non-missing values of each variable were used to impute the missing values. This method does not take advantage of any correlation among the variables and therefore can perform rather poorly when such correlations are present.

MICE was proposed by Van Buuren et al. [3]. It requires the user to specify a conditional model for each variable, using the other variables as predictors. By default we used a linear regression model for continuous variables, a logistic regression model for binary variables, and a polytomous logistic regression for categorical variables with more than two levels. The algorithm works by iteratively imputing the missing values based on the fitted conditional models until a stopping criterion is satisfied. In that way it is very similar to the missForest algorithm; the main difference being that missForest uses more flexible decision trees for each conditional model. We implemented this in R using the package ‘mice’[10] .

Statistical Analysis

We used two separate studies to perform the comparison between the methods of imputation described in the previous section. We describe the studies, implementation details, and our results below. The structure of the statistical analysis is the same for both studies. We start with a published predictive model built with the training data set. The test set refers to observations that were not part of the training set; these were solely used for assessing the performance of the model. The test sets did not have any missing values, so we randomly removed a proportion of values to simulate data missing completely at random. We then imputed the missing values by the four previously discussed methods, and the imputed laboratory results were compared to the actual values that were removed from the data set. We then used the imputed data to make clinical outcome predictions with the published models, and the results were compared with the predictions made using the complete test data with no missing data. The prediction models were created using both logistic models and random forest models as some might argue that using random forest imputation methods might favor the random forest prediction models. We use the average relative error (for continuous variables) and misclassification error (for categorical variables) to assess the imputation performance for both the logistic and random forest models. To quantify the effect of the imputation on predictive models we compare the predicted classes from non-missing test data to predicted classes from imputed test data and compute the misclassification error for both the logistic and random forest models. This is important because if a particular variable had very little influence on the predictive model, then larger imputation errors are tolerable, resulting in negligible loss of prediction accuracy. On the other hand, small imputation errors in very important variables might lead to significantly different predicted class, which is of greater clinical concern. We varied the frequency of missing values to change the difficulty of the imputation problem. We report the average results over multiple random runs. We found that the Nearest Neighbor results are quite robust to the choice of number of nearest neighbors (k) if k is moderately large, therefore we fixed the number of nearest neighbors at 5 in both studies.

Results

Cirrhosis Cohort and HCC model

This study evaluated the effect of imputation on a published predictive model for HCC based on 21 predictor variables that included demographic, clinical and laboratory values using random forest modeling [2]. The random forest model was developed on a data set of 446 patients collected at University of Michigan (UM cohort). It proved to be more accurate than traditional logistic regression models. Of the 21 variables, 10 of them were categorical in nature while 11 were continuous. We used the first 200 observations from the publicly available data set from the HALT-C trial as our test set and randomly replaced 10%, 20%, or 30% of the observations with missing values. The process was repeated for 1000 replications and we report the average results.

The accuracy of the four imputation methods for both continuous and categorical variables are compared in Figure 1 for the cirrhosis cohort and HCC model. Figure 1A and 1B represents the logistic model and Figure 1C and 1D reflect the random forest prediction model. The vertical axis plots the percentage relative error for continuous variables and percentage misclassification error for categorical variables, while the horizontal axis groups the results according to the proportion of missing values. Each boxplot represent the error measure over 1000 random replications. As expected, the imputation error increases on an average as we increase the proportion of missing values in the test data but the variation tends to reduce slightly which is due to averaging over many more missing observations. MissForest has the least imputation error for both continuous and categorical variables at each level of missing proportion, followed by MICE, NN, and mean imputation of continuous laboratory values. MICE and NN have similar imputation accuracy for categorical variables. MissForest works well using both logistic and random forest prediction models.

In Figure 2A and 2B, which represents the logistic and random forest prediction models respectively, the vertical axis plots the error measure for imputation on predictive model, obtained by comparing the predicted classes (Low Risk/High Risk) for each test observation with no-missing values against the

predicted class after imputing the artificial missing values. Therefore an error measure of 5 on the vertical axis implies that 5% of the test observations had their predicted classes wrongly switched (either low risk => high risk or high risk => low risk) due to the imputation. As above, each boxplot reflects the results of 1000 random runs. It is clear from the figure that missForest performs the best with NN and MICE following closely. The gap increases as we increase the proportion of missing observations making the problem harder.

Inflammatory Bowel Disease Cohort and Thiopurine Clinical Response Model

Waljee et al. showed that random forest models using laboratory values outperform 6-thioguanine (6-TGN) metabolite tests as well as traditional logistic regression models in predicting clinical response to thiopurines [1]. The analysis was carried out on a data set collected at University of Michigan that included 395 patients. 26 variables, which included 25 laboratory values and age, were used to predict the clinical response of each patient. In this study all of the variables were continuous in nature. To create a separate test set, we split the data set randomly into a training set consisting of 250 observations and a test set of 145 observations, using stratified sampling to keep the ratio of clinical responder to non-responders fixed. We then introduced random missing values into the test set as before and performed the same comparative study of the four imputation methods. The whole process was replicated 1000 times to obtain stable results. Below we summarize our findings via boxplots.

Again in this study, missForest beats its competitors in both imputation accuracy and the effect of imputed values on the accuracy of clinical predictions based on the logistic model (Figure 3A) and the random forest model (Figure 3B). The trends remain the same for imputation error, Figure 4A and B representing logistic models and random forest models respectively, with MICE coming out second best followed by NN and mean imputation. For predictive accuracy we find the relative order becomes missForest > MICE > mean imputation > NN. This also shows that best method with respect to imputation error need not be the best when we consider the effect of imputation on predictive models.

The performance gap between missForest and MICE is considerably lower than in the previous study. This might be explained by the fact that in the thiopurine study, both the training and test sets came from the same cohort, as we generated the training and test sets by random splits, while in HCC study the training and test sets were completely different cohorts leading to an extra degree of variation.

Discussion

We have performed an extensive simulation study using two clinical datasets and two published predictive models to compare the performance of four methods of missing value imputation for missing data completely at random. We included both local (randomForest) and global (logistic) modeling approaches to avoid bias that might favor a local (MissForest) imputation approach. While the superiority of MissForest for imputation of missing lab values this will not be generalizable to all predictive models or datasets, this manuscript highlights the value of missForest to impute missing data. We compared four popular methods namely, missForest, Nearest Neighbor, MICE and mean imputation, in two studies simulating data missing completely at random. We found that these simulation methods consistently produced the lowest imputation error and had the smallest prediction difference when models used imputed laboratory values. In addition, the ready availability of the freeware R package makes missForest and its simulations a very convenient solution for any practical missing value problems. The main limitations of these simulations as a solution to missing laboratory data for predictive modeling applications are: a requirement for skilled R programming for implementation, and slightly more demanding computational needs, compared to NN or MICE methods. An additional limitation in this study is that these simulations did not address the issue of data missing for non-random reasons. There could be an association between the clinical outcome of interest and the missingness of certain predictors. At this point, we cannot generalize these results to situation in which data is missing for non-random reasons.

The small absolute changes in predictions with these models, despite 10-30% missing laboratory data, speak to the robustness of these multi-analyte assays with algorithmic analysis (MAAAs). MAAAs are

currently a hot topic, and several have been released with CPT codes in 2012. One example is the HCV FibroSure (LabCorp, code 0001M) which uses ALT, alpha 2 macroglobulin, apolipoprotein A1, total bilirubin, GGT, and haptoglobin to estimate fibrosis and necroinflammatory activity in the liver in patients with hepatitis C. With the increasing complexity of these models, and increasing numbers of analytes, the risk of missing completely at random values increases, and methods to cope with missing values and preserve the accuracy of the model are needed. missForest appears to be a robust and accurate approach to the issue of missing laboratory values when used in these two MAAAs and may be applicable to other datasets with missing completely at random datasets.

Contributions:

Ashin Mukherjee – study concept and design, statistical analysis and interpretation of the data, drafting of the manuscript, critical revision of the manuscript

Akbar Waljee - study concept and design, acquisition of the data, statistical analysis and interpretation of the data, drafting of the manuscript, critical revision of the manuscript and study supervision.

Amit Singal - acquisition of data, drafting of the manuscript and critical revision of the manuscript.

Yiwei Zhang - statistical analysis and interpretation of the data and critical revision of the manuscript.

Jeffrey Warren - study concept and design, critical revision of the manuscript.

Ulysses Balis - study concept and design, critical revision of the manuscript.

Jorge Marrero - study concept and design, critical revision of the manuscript.

Ji Zhu - statistical analysis and interpretation of the data and critical revision of the manuscript.

Peter Higgins - study concept and design, acquisition of the data, statistical analysis and interpretation of the data, critical revision of the manuscript and study supervision.

Competing Interests: None

Ethics approval: University of Michigan, IRB

Data Sharing: Agreement: There is no additional data available.

Figure Legends

FIGURE 1. Imputation error comparison for categorical and continuous variables for four competing imputation methods at three levels of the proportion of missing values for the logistic prediction model (Figure 1A and 1B) and random forest prediction model (Figure 1C and 1D) in the HCC study

FIGURE 2. Percentage of wrongly predicted observations after missing value imputation by the four competing methods at three levels of missing value proportions in the test data for the logistic prediction model (Figure 2A) and the random forest prediction model (Figure 2B) in the HCC study.

FIGURE 3. Imputation error for four competing imputation methods at three levels of the proportion of missing values for the logistic prediction model (Figure 3A) and random forest prediction model (Figure 3B) in the Thiopurine response model

Figure 4. Percentage of wrongly predicted observations after missing value imputation by the four competing methods at three levels of missing value proportions in the test data for the logistic prediction model (Figure 4A) and the random forest prediction model (Figure 4B) in the Thiopurine response model.

References

1 Waljee AK, Joyce JC, Wang S, et al. Algorithms outperform metabolite tests in predicting response of patients with inflammatory bowel disease to thiopurines. Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association. 2010 Feb;**8**(2):143-50.

2 Singal AG, Waljee AK, Mukherjee A, Higgins PD, Zhu J, Marrero JA. Machine Learning Algorithms Outperform Conventional Regression Models in Identifying Risk Factors for Hepatocellular Carcinoma in Patients With Cirrhosis. Gastroenterology. 2012 May;**142**(5):S984-S.

3 van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999 Mar 30;**18**(6):681-94.

4 Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001 Jun;**17**(6):520-5.

5 Stekhoven DJ, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012 Jan 1;**28**(1):112-8.

6 Singal AG, Conjeevaram HS, Volk ML, et al. Effectiveness of hepatocellular carcinoma surveillance in patients with cirrhosis. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2012 May;**21**(5):793-9.

7 Liaw A, Wiener M. Classification and Regression by randomForest. 2002;**2**(3):18-22.

8 Breiman L. Random forests. 2001;**45**(1):5-32.

9 Gower JC. A General Coefficient of Similarity and Some of Its Properties. Biometrics. 1971;**27**(4):857-71.

10 Buuren van S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;**45**(3).

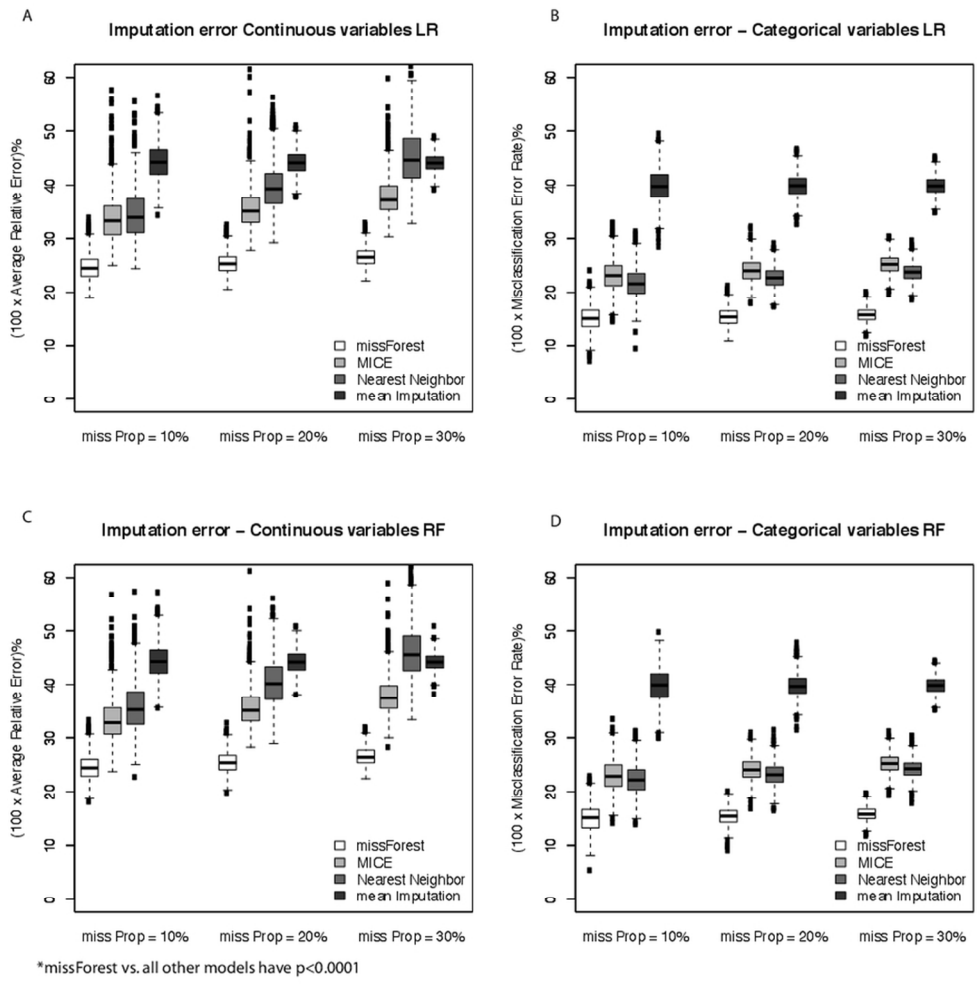


FIGURE 1. Imputation error comparison for categorical and continuous variables for four competing imputation methods at three levels of the proportion of missing values for the logistic prediction model (Figure 1A and 1B) and random forest prediction model (Figure 1C and 1D) in the HCC study. 90x90mm (300 x 300 DPI)

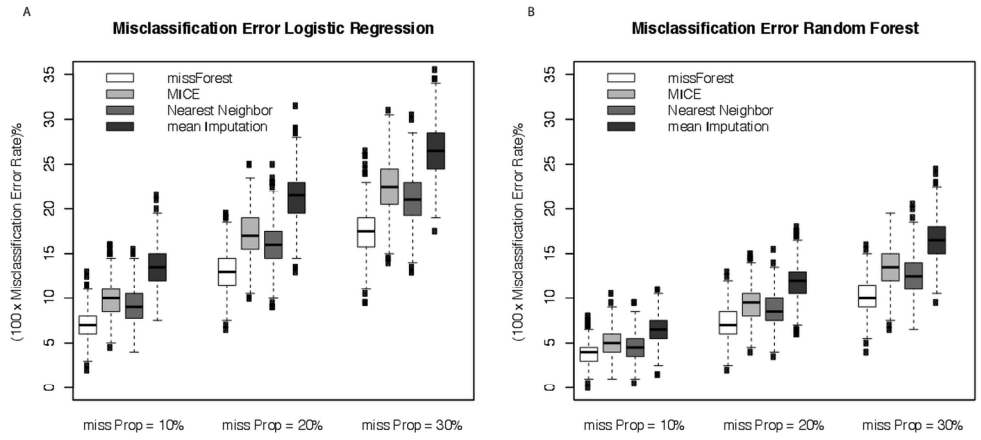


FIGURE 2. Percentage of wrongly predicted observations after missing value imputation by the four competing methods at three levels of missing value proportions in the test data for the logistic prediction model (Figure 2A) and the random forest prediction model (Figure 2B) in the HCC study.
185x90mm (300 x 300 DPI)

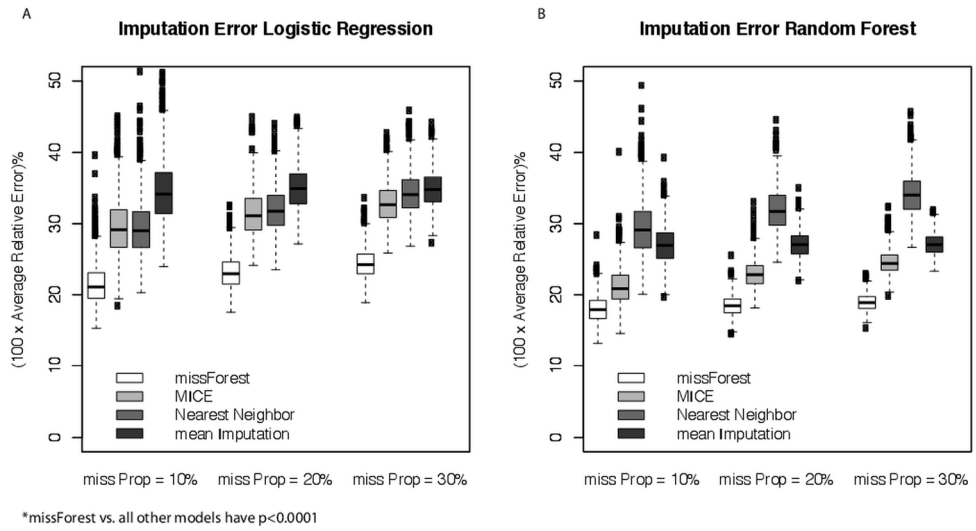


FIGURE 3. Imputation error for four competing imputation methods at three levels of the proportion of missing values for the logistic prediction model (Figure 3A) and random forest prediction model (Figure 3B) in the Thiopurine response model.
162x90mm (300 x 300 DPI)

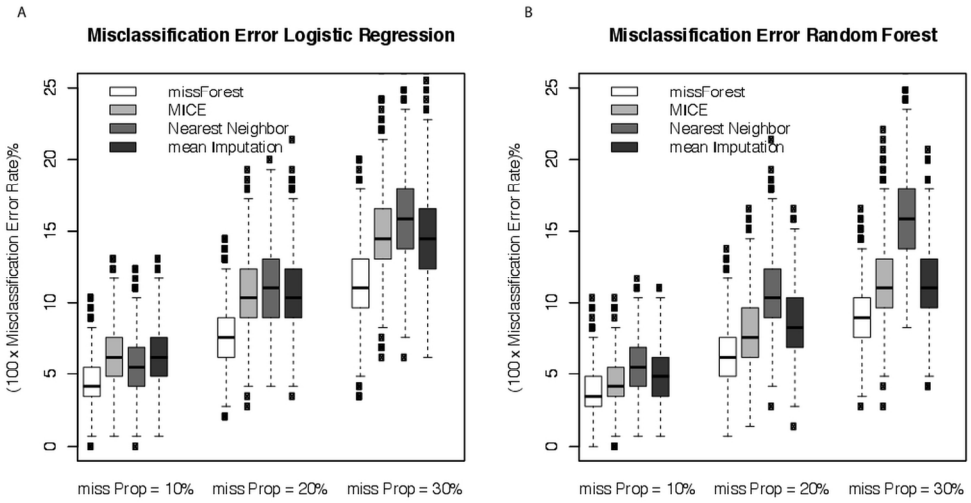


Figure 4. Percentage of wrongly predicted observations after missing value imputation by the four competing methods at three levels of missing value proportions in the test data for the logistic prediction model (Figure 4A) and the random forest prediction model (Figure 4B) in the Thiopurine response model.
161x90mm (300 x 300 DPI)