Visual Analytics of Missing Data in Epidemiological Cohort Studies

S. Alemzadeh, U. Niemann, T. Ittermann, H. Völzke, D. Schneider, M. Spiliopoulou, B. Preim

Eurographics Workshop on Visual Computing for Biology and Medicine (2017)
S. Bruckner, A. Hennemuth, and B. Kainz (Editors)

Visual Analytics of Missing Data in Epidemiological Cohort Studies

S. Alemzadeh[1], U. Niemann[1], T. Ittermann[4], H. Völzke[4] , D. Schneider[3], M. Spiliopoulou[2] , B. Preim[1]

[1] Department of Simulation and Graphics, Otto-von-Guericke University Magdeburg, Germany
[2] Department of Technical and Business Information Systems, Otto-von-Guericke University Magdeburg, Germany
[3]Department of Statistics, Otto-von-Guericke University Magdeburg, Germany
[4]University Medicine Greifswald, Germany

Abstract

We introduce a visual analytics solution to analyze and treat missing values. Our solution is based on general approaches to handle missing values, but is fine-tuned to the problems in epidemiological cohort study data. The most severe missingness problem in these data is the considerable dropout rate in longitudinal studies that limits the power of statistical analysis and the validity of study findings. Our work is inspired by discussions with epidemiologists and tries to add visual components to their current statistics-based approaches. In this paper we provide a graphical user interface for exploration, imputation and checking the quality of imputations.

Categories and Subject Descriptors (according to ACM CCS): J.3 [Computer Applications]: Life and Medical Sciences—

1. Introduction

Epidemiologists gather data from medical examinations, physical conditions, treatments and questionnaire forms. To do investigations in longitudinal cohort studies, they often repeat these examinations several times. The classical cohort has a baseline examination. Thereafter, only information on newly occurring diseases and mortality is collected. Missing values are an unavoidable part of such collecting process. Generally, there are three types of missing data: when the data are missing completely at random (MCAR), missing at random (MAR), where the missing values depend on the other observed variable/s, and missing not at random (MNAR), where the probability of missingness cannot be explained by the other observed values. When the missing data are MAR, simple methods fail to predict missing values, bias occurs in the results and we loose a part of information [SWC*09]. However, more complicated methods like multiple imputation can handle MAR missingness.

Dropouts are the most severe problem in epidemiology longitudinal studies, where participants do not follow the examinations. Today, there are statistical tools and appropriate methods that treat missing values [BGO11].

Often, the patterns of missing data can be observed and are essential to assess whether missingness requires special treatment. Visual analytics of missing values is a rare topic [FG14], whereas the statistical treatment of missing values is an established topic. Visual analysis could have a strong role in the investigation of missing data and to get information on the quality of imputed data. Visualization of the data that are not observed can give insights to the analyst about the source of missingness and whether a particular individual denied to participate in the next follow-up examination or if some variables stayed unreported. The correlation between missing values is interesting to select an appropriate method for treating them. There is a lack of appropriate systems to analyze the missing data and to easily impute them and check the quality of imputations. In this paper we provide a user interface for the visualization and imputation of missing data on epidemiological data. Our contributions include:

- A system for the exploration and imputation of missing values,
- Suggestions for predictors to variables with missing values, and
- Methods to check the quality of imputations.

The rest of paper is organized as follows: Section 2 gives an overview of related works on the visual analysis of missing data and cohort study data. In Section 3, we discuss missing data and requirements to handle them. In Section 4, we discuss the presented application to address requirements by giving a use case scenario. We conclude with a summary and future work in Section 5.

2. Related Work

This section provides an overview of related works for imputation techniques of clinical and epidemiological data, the visual analysis of cohort studies and missing data visualizations.

There are many studies that compare different imputation methods for cross-sectional and longitudinal epidemiology and medical data [DvdHSM06, Twi13, SWC*09]. They mostly conclude that single imputations are most frequently used to fill the missing values, e.g. imputation by the mean value. However,

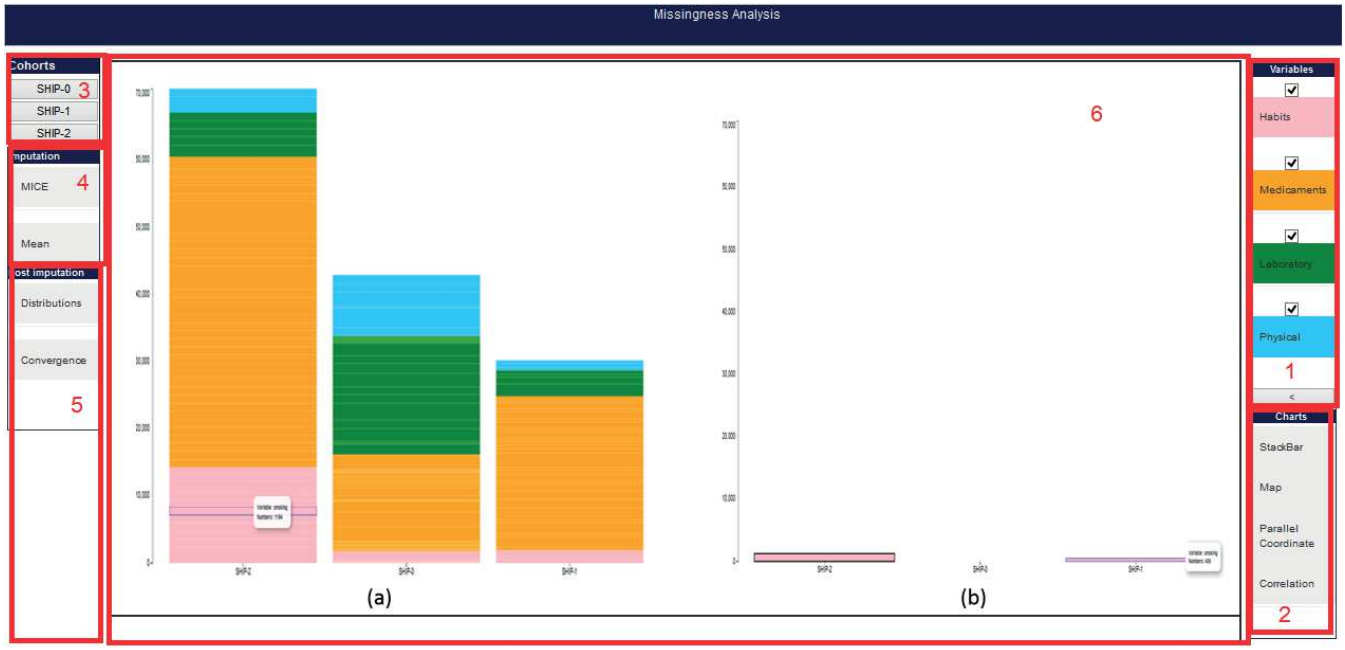

Figure 1: *User interface of our system. Panel 1 shows several categories of variables. Check boxes define which categories of variables are incorporated in the analysis. Panel 2 contains charts for the exploration of missing values. Panel 3 contains the datasets for the analysis. In panel 4, the user can set the parameters for missing value imputation. Panel 5 consists of plots to check the quality of imputations. All plots and results are displayed in panel 6.*

these heuristics lead to a biased estimation, since they reduce the variability of the true distribution. In contrast, multiple imputation methods yield a reasonable amount of accuracy and bias. Multiple imputation generates *m* multiple imputed datasets by considering the dependency between variables. In fact, predicting the exact values of missing data is impossible, but multiple imputation gives an appropriate amount of uncertainty by assessing multiple plausible values in *m* datasets. The differences between imputed values in *m* datasets is the amount of uncertainty.
Although the visual analysis of missing data is of potentially high benefit for researchers, there are very few attempts in the literature. It is obviously challenging to display something that does not exist. Cheng et al. [CCH*15] proposed a graphical user interface to show and impute the missing values in R. The plots let the user explore and compare different imputation methods.

Johansson, Fernstad and Glen [FG14] investigate different aspects of missing values by using a visual analysis tool. They displayed the relationship of missing data among different variables before and after a simple imputation by a combination of parallel coordinates and bar charts. Although it is a good start to visualize the missing values, many points such as appropriate visualizations for quality assessments of imputations were left unsolved.
Eaton et al. [EPD05] examined different visualization techniques to represent missing data. They reported the effect of missing data on visualization techniques in an empirical study. In the study, they encode missing values by missing spaces and encoded glyphs. The results showed that encoded values will give the user more perception of the missing data.
Klemm et al. [KLG*16] presented a regression-based technique for the discovery knowledge from cohort study data. They implemented 2D and 3D heat maps to show the correlations between variables. The analyst can interactively choose a model for getting more details and analyze risk factors for diseases. Zhang et al. [ZGP15] provide CAVA, a framework for the visual analysis of cohort study data. CAVA enables the analyst to build group of patients for further investigations by iterative search.
Alemzadeh et al. [AHN*17] presented a framework to explore the results of subspace clustering of epidemiological data. A technique is provided to let the analyst check the replication of a subpopulation in independent cohort study data. Preim et al. [PKH*16] give a review on the role of visual analytics on cohort studies (more specifically image-centering data). They investigated the combination of clustering analysis and visualization.

3. Missing Data

In this section, we discuss the source and types of missing values, especially the problems with missing values and why we need to impute them carefully to preserve the precision of the data.

3.1. Sources of Missing Data

Missing values of longitudinal epidemiological data may have several reasons. Epidemiologists gather the data by collecting information from questionnaire forms (paper-based or online), labora-

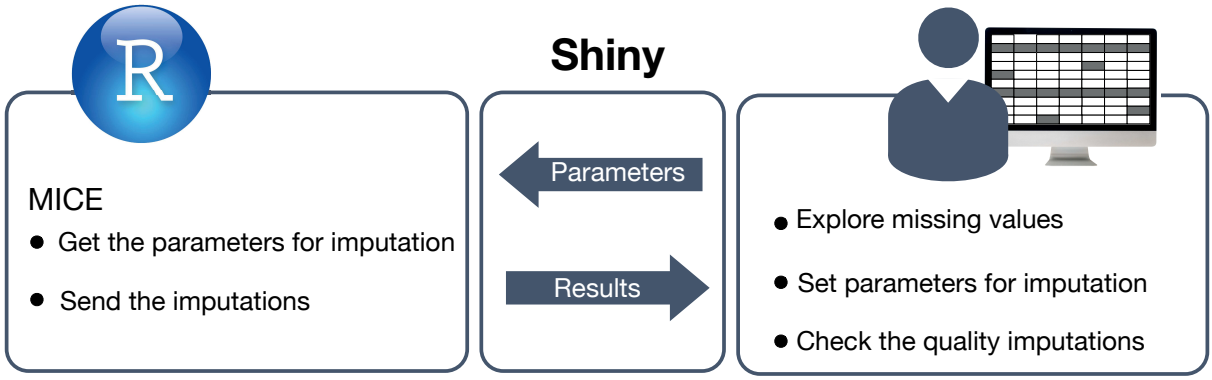

Figure 2: *The analyst explores the missing values and sets the parameters for multiple imputation to the R tool via Shiny package. The MICE library in R gets the parameters and runs the imputation process. After imputation the results will be sent to the analyst to check the quality of imputations via Shiny.*

tory tests, medical images and other examinations. In longitudinal studies, they usually repeat examinations in different time points to observe the changes. During these time points, some participants drop out from the study. This may have several reasons, e.g., inconvenient places of examinations, physical disabilities, forgetting appointments. Sometimes, variables remain unfilled for some participants. In this case, participants did not answer some questions of the questionnaire forms, biological samples are unavailable or there are problems with medical images and devices. Generally, having a large amount of missing values leads to notable bias, loss of accuracy and loosing a part of the original sample.

3.2. Types of Missing Data

There are three types of missing data [SWC*09]:

- MCAR: When one variable for a participant is missing that has no relevance to other variables of that participant. For example, accidentally the participant does not respond to a question.
- MAR: In many cases, the probability is high to miss information that depend on the other available information of that participant. For example, the blood pressure measure is more probably be missed for younger participants than for older ones.
- MNAR: If the probability that an observation is missing depends on an unobserved data, this case is called missing not at random. For example, people with a specific disease are more likely to not be present for examinations.

3.3. Handling Missing Data

Studies show that in the case of MCAR simple techniques like the complete case analysis yield unbiased results [GF95]. On the other hand, when the data are MAR, simple techniques like mean imputation lead to biased results. Complicated methods like multiple imputation showed unbiased results in several studies [VB98]. When just a small part of the data is missing, single imputation could be a suitable choice, because this method is less complex. Otherwise, multiple imputation is a better choice. The values for dropout participants can be predicted by the baseline values. The missing values will be filled with multiple predictions.

Multiple Imputation by Chained Equations. Multiple imputation is a frequently used technique for missing value imputation in epidemiological data [SWC*09]. Generally, multiple imputation uses the distribution of observed data to produce multiple plausible values for missing data. As an output, several datasets are generated, i.e. there are multiple estimations for each unobserved value. The variance of the estimated values reflects the amount of uncertainty in the prediction models. *Multiple imputation by chained equations (MICE)* [WRW11] is a common approach to generate imputations for datasets containing several variables with missing values. First, the missing values of each variable are filled by a simple imputation technique, e.g. applying the mean value or drawing a random sample. Then, for each variable v a regression model is learned only for the observed values of v on both observed and filled values of all other variables. Finally, the missing values of v are replaced by the predictions from the regression model. This procedure, also called *a cycle*, is repeated *varit* times yielding *varit* imputed datasets which are aggregated at the end. To ensure stability of the results, researchers should specify a sufficiently large value for a maximum number of iterations (*maxit*).

3.4. Requirements

Epidemiologists advised us that a combination of multiple imputation (their established technique) with interactive visual analysis would be beneficial. In particular, our discussions resulted in the following requirements:

R1 In longitudinal study data, the proportions of missing values in each time point of study should be displayed. It is also interesting to see whether all variables in the baseline examination are available in the follow-up examinations.

R2 To choose the right imputation method, the pattern of missing data should be identified. When the data are MCAR, simple imputations are provided to treat the missing values. When the missing data are MAR, multiple imputation should be applied.

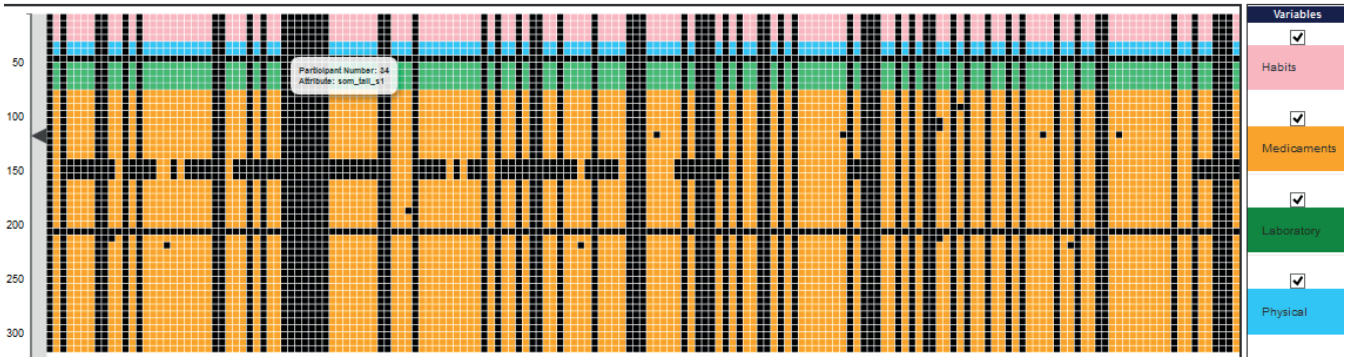

Figure 3: *The missingness map shows an overall overview of missing data based on the groups of variables. Rows stand for variables and columns show participants. Blacked cells represent missing values. One completely blacked column represents a drop-out participant. The zoom tool on the left side is used to adjust the degree of detail. Tooltips give information about the variable description and the participant number in the data base.*

R3 For predicting missing values of a variable, multiple imputation by default gets its regression on all other variables. Due to computational complexity, it is infeasible to do this when we have a large number of variables. Thus, predicting dropouts by the values of variables in the baseline examinations, which are correlated to the variable with missing values, helps to reduce the complexity.

R4 In the multiple imputation process, parameters like the number of datasets and the sufficient number of iterations should be selected carefully in order to get more accurate results.

R5 After the imputation process, the imputed datasets should be diagnosed carefully in order to make sure that the predictions are plausible values.

R6 The convergence of imputed curves should be followed to check whether the number of iterations is sufficient. Usually, this is achieved by observing the mean value and standard deviation curves of imputed values over different iterations [BGO11].

In the following, we are trying to handle these requirements.

3.4.1. Parameters

Multiple imputation chained equations (MICE) [BGO11] in R (recall Section 3.3) provides a platform to let the user flexibly change the parameters and get the imputation results. For setting the MICE procedure the following parameters are essential for the quality and complexity of imputations.

- **Maximum iterations:** The process of imputation is repeated with *maxit* numbers to generate an imputed dataset. Actually, imputations are repeated until the imputation curves over iterations reach a state where the mean and standard deviation values of a particular imputed variable separately are mixed together. This means that mean and standard deviation values of imputed data sets should not vary too much from each other. The imputation model should ideally achieve a convergence among imputations at the end.
- **Number of imputed datasets:** During the imputation process the values for missing values will change and for the observed variables they remain fix in all generated datasets. In many statistical tools (e.g. MICE) the number of imputed datasets is set to 5 by default. The difference between imputed values in datasets reflects the amount of uncertainty about the predicted values. Although, when we have a large number of missing values, it is suggested to have a large number of imputed datasets [GOG07]. Due to performance reasons this is not feasible with a large number of variables.
- **Predictors of target variable:** As discussed in Section 3.3, the prediction of missing values of one variable is by regression on other variables. The dropout to follow-up participants could be explained by the variables from the baseline examinations. By default, the missing values will be filled by the regression on all other variables. When we have many variables in our dataset, the prediction of one target variable is not feasible by all other variables, because of computational complexity. Thus, carefully selecting a set of variables as predictor in regression models that imposes the least risk of bias in estimations is necessary [CSK01].
  To achieve better imputation results with less complexity, the MICE package allows the use of a prediction matrix. This matrix gives information on predictors for the target variables. If n is the number of variables, then this matrix has an $n * n$ dimension which is filled by 1 and 0. The rows represent target variables, and the value of 1 is assigned if a variable is used for imputation and 0 if it is not. In the matrix of Eq. 1 , v_1 is imputed only by information gained from v_2 and ‘ v_2 is not imputed at all, because it has no missing values.
  This matrix can be set by default (use all variables as predictors), or be customized (use variables to predict one variable that has a correlation with a selected variable) to increase the performance and speed of imputation. The goal is to find variables that depend on each other and, thus, they can be used to predict each other. In general, there are two types of variables in our dataset: factor

and numeric. In this work, to address R3 and find good predictors for numerics a correlation matrix (Pearson correlation) is calculated. If the absolute value of a correlation is greater than an adjustable threshold (by default it is 0.2), the variables are used for imputation of each other, 1 (correlated) is written in the predictor matrix. Otherwise, 0 (not correlated) is written in the matrix.
The connections between factor variables are found by performing a Chi-squared Test. If the test finds a significant relation between two variables, 1 is written in the predictor matrix. The significance is defined by the p-value of the test. If it is smaller than the threshold alpha (default is 0.05), they are considered as non-correlated.

$$\begin{bmatrix} v_{11}{=}0 & v_{12}{=}1 & v_{13}{=}0 & \dots & v_{1n}{=}0 \\ v_{21}{=}0 & v_{22}{=}0 & v_{23}{=}0 & \dots & v_{2n}{=}0 \\ \dots & \dots & \dots & \dots & \dots \\ v_{n1}{=}0 & v_{n2}{=}0 & v_{n3}{=}1 & \dots & v_{nn}{=}0 \end{bmatrix} \quad (1)$$

There are special cases that are set as follows:

1. When a variable has no missing value, the whole row of the predictor matrix is set to 0, since imputation is not necessary.
2. It is not possible to perform a test for two variables. This could for example be the case when the standard deviation of a variable is zero. In this case, the correlation cannot be calculated. If there is an error, the cell in the predictor matrix is set to 0.

We only consider linear correlations, whereas non-linear correlations are also typical in epidemiology data, e.g., J-shaped and U-shaped distributions [PKH*16].

4. Visual Analysis of Missing Data

A number of statistical tools for the imputation of missing values are established. They provide functions to address the missing values by removing unobserved values, single imputations (overall mean and median imputation) or multiple imputation. However, the tools do not display the missing values and relationships between them and other variables. This would be essential to decide which imputation model is more appropriate. Additionally, it is necessary to check the quality of imputed values to ensure that they are sensible regarding the variables. Here, we provide a web-based graphical user interface that allows epidemiologists to explore missing values and to impute them by interactively changing the parameters for the imputation method (Fig. 1). Moreover, we provide measurements and different plots to check the quality of imputations. In the following, the components of the visual analysis framework are discussed.

4.1. Examined Data

The Study of Health in Pomerania (SHIP) is a population-based project in the Northeast of Germany, where the data are gathered in three different time points during an eleven year follow-up examination period [Vol12]. The participants are invited to the study and information are collected from questionnaires, laboratory tests, medical and dental images. SHIP-0 was a baseline examination of 4308 participants aged between 20 and 81, in the years between 1997 and 2001. The second wave, SHIP-1, was conducted between 2002 and 2006, where the participants of SHIP-0 were re-invited with 3301 participants. About 16.4 percent of the individuals dropped out in SHIP-1, more people dropped out in SHIP-2 (only 1879 participated again). Here, we used hepatic steatosis SHIP datasets of female participants for our analysis, where many values are missing because of an increasing dropout rate in each wave.

4.2. Application and Use Case Scenario

We implemented the graphical user interface based on the web technology and the screenshot of the system is shown in Figure 1. The application is implemented by HTML5 and JavaScript. As shown in Figure 2, to estimate the imputations, we used the MICE package in R [BGO11]. To visualize the results and for interaction with the analyst the D3.js library is used [BOH11]. To make the connection between R and JavaScript, the RShiny package is used. The proposed framework provides GUI to cover a range of functions for exploration and imputation of missing values in epidemiological data. Additionally, it enables the analyst to diagnose imputations. In this section, we demonstrate in a use case scenario how the framework supports the user to do investigations for imputation. We explain each step by screen shots of the analysis stages. In this use case, we follow an expert user who is familiar with the data (hepatic steatosis SHIP data).

4.2.1. Grouping of Variables

Overall, we have 266 variables (excluding time and date variables) in the baseline examination. To facilitate the exploration and interpreting the plots, we provide a categorization of variables in four groups:

- Habits (e.g. smoking behavior)
- Physical status (e.g. body mass index)
- Laboratory tests (e.g. cholesterol)
- Medicament (e.g. enalapril)

One unique color is assigned to each category and all plots are colored based on this categorization.

4.2.2. Exploration of Missing Data

The missingness map provides a compact overview on the pattern of missing and observed values. A basic missingness map is provided in R [TF08, HKB*11].

The analyst is interested to see the proportion of missing values, how they are correlated and find meaningful patterns of observed and missing values.
To start, the analyst should select one dataset from the cohort panel. To continue the use case, the SHIP-1 dataset is selected. The following section describes the further analysis regarding to the SHIP-1 dataset. A filtering option is also provided so that the user can exclude categories of variables from the analysis by unchecking the check box on top of each variable's category (i.e. variables related to habit group). To see the variables of each category, the user can click at the button of the corresponding category in the variables'

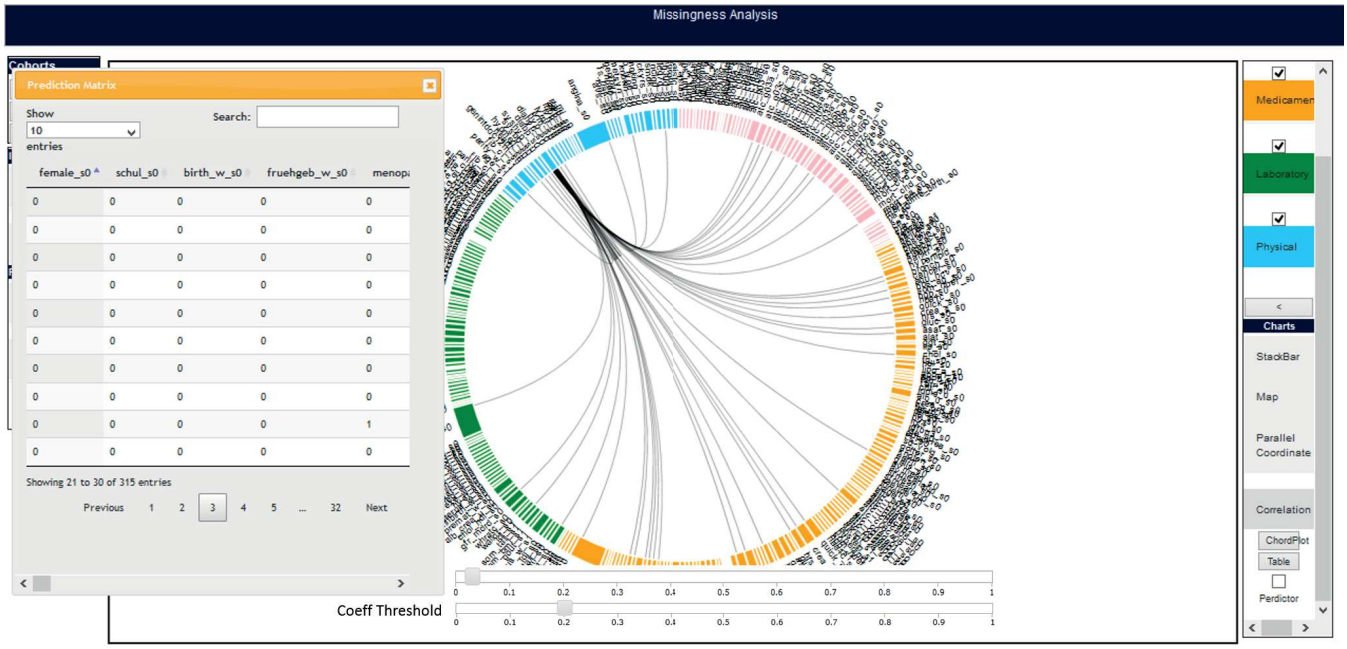

Figure 4: *The chord chart illustrates the correlation matrix. The variables are colored and sorted based on the variable's taxonomy. When there is a correlation between the two variables, then an arc shows this connection. A table shows the values of the prediction matrix. A slider can adjust the threshold values for generating the prediction matrix.*

panel. Then, variables will be shown in an accordion menu. In the following analysis, all groups are included.

- **Amount of missingness:** The user may be interested to see the proportion of missing values in each wave of the cohort study data, with regard to variables and the overall number of missing values (R1). The analyst clicks on the stack bar from the charts panel and an interactive stack bar shows the number of missing values in all waves of cohort data. The bars are sorted based on the variable categorization. By moving the mouse over stacks, the user can see a tool tip containing the caption of variable name and number of missingness. Stacks show the proportion of missing values for each variable in cohorts. To compare the number of missing values of a particular variable, by clicking on the corresponding variable's stack the user can see the proportion of missing values separately (see Fig. 1 panel 6).
- **Pattern:** As discussed in Section 4.2.2, the simplest way to plot the missing and observed values is by using a missingness map. The missingness map shows the place of missing values in each variable. It reveals which participants dropped out from the study or whether simultaneous missingness occurred in variables. If this is the case, the analyst may conclude that there is a relation with these variables. Here, we implemented a categorized missingness map (R2). As shown in Fig. 3, the columns stand for participants and the rows demonstrate the variables. The values that do not exist are colored in black. A zoom tool is embedded to let the user zoom in and out the plot to adjust the level of detail. When the user is moving the mouse across each cell in the missingness map, a tool tip provides some information about the participant and the variable. When the column (participant) is completely black, the corresponding participant dropped out of the study (unit non-response). In contrast, when only one cell is black, this means that it is a non-response case.
- **Correlations:** As the next step, the analyst interacts with the framework to see the correlation between variables (R3). This correlation may be used for the imputation of missing values. As discussed in Section 3.4.1, the Pearson correlation coefficient is used to characterize the relationship between numeric variables. The Chi-squared test explains the correlation between categorical variables. Variables are correlated to each other if they exceed a threshold value. As shown in Figure 4, a chord chart illustrates the correlations between SHIP-0 and SHIP-1 variables, where connected arcs show that the two variables are correlated. With sliders the analyst can adjust the threshold value and observe the correlations between variables. Here, the Pearson coefficient threshold (correff) is set to 0.2 and the Chi-squared test threshold (alpha) is set to 0.05. To see the matrix in a table schema, the analyst clicks on the table button. Then, a dialog box will show the prediction matrix. If the analyst wants to use these correlations as predictor matrix in the imputation model, he should set the "predictor" check box in the correlation panel.

4.2.3. Imputation

After investigating missing and observed data, the user decides to impute the missing values. In the imputation panel, techniques for treating missing values are suggested. Here, the analyst selects the MICE imputation. By clicking on MICE the according

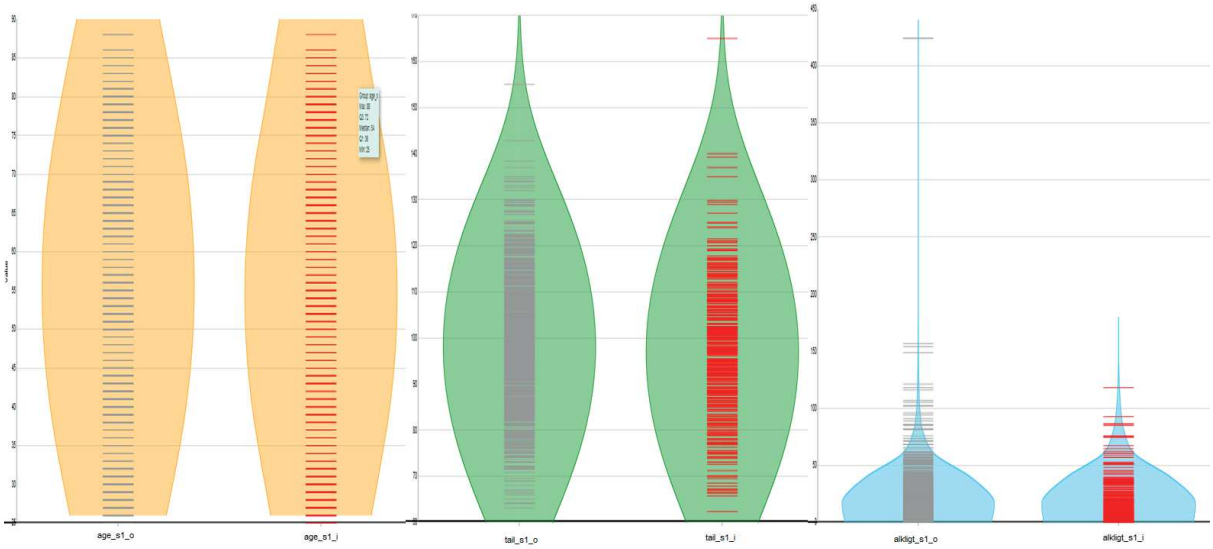

Figure 5: *Bean plots are used to compare the distributions of imputed and observed values in dense and sparse regions where the lines inside the bean plot represent the values. The red lines show imputed values and gray ones display observed values.*

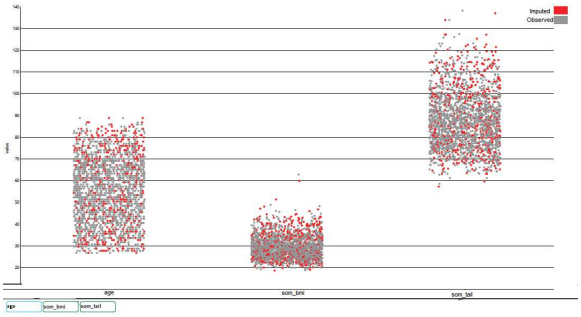

Figure 6: *The strip plot displays the distribution of imputed values over observed values in a combined way. The user can drag and drop important variables to see distributions. The gray points represent observed values and the red ones represent imputed values.*

panel will be expanded and the user can set parameters to continue imputations (R4). These parameters include the number of imputed datasets (*m*) and the maximum number of iterations for multiple imputation (*maxit*). If the analyst wants to pass the predictor matrix defined in Section 3.4.1, he should set its corresponding check box to 'checked'. By clicking on the impute button, the parameters and the data will be sent to the R tool via the Rshiny library for calculations.
After calculations, the results including the *m* imputed dataset and the mean and standard deviation of the variables in different iterations will be sent to the client from the R server.

4.2.4. Quality Control

After imputation, checking the prediction to assess whether the imputations are plausible or not is very critical. Impossible imputations (e.g., negative values for cholesterol) will corrupt the predictions.

Plausibility of imputed values. It is necessary to compare the distribution of imputed values over observed values.
As the next step, the analyst checks the quality of imputations in *m* datasets separately to see whether the imputed values are plausible to address R5. Thus, in the post imputation panel items to assess the quality of imputations are listed. So, the user clicks on the distribution button, then the corresponding menu contains the plots extended to show the distribution of imputed values over missing values. At first, from the dropdown list in the distribution section, the user selects the number corresponding to the imputed dataset. This value is between 1 and *m*. Then, from the variable panel he can drag and drop a variable to the plot panel to check the distributions.
Next, the analyst clicks on the bean plot (see Fig. 5) to compare the distribution of observed and imputed values. In the bean plot, the density is shown by beans and individuals are shown inside the beans in a strip chart [K*08]. Bean plots are chosen since they enables the analyst to compare the density of observed data over imputed data. The bean plot with gray strips shows the distributions of observed data and the red strips represents the imputed data. As shown in Fig. 5, the imputed data relating to age are distributed in a similar way to the observed data. The minimum and maximum values are approximately the same and the imputed values seems plausible.
Next, the analyst wants to compare this distribution in a more compact way. By clicking on the strip plot, it illustrates the missing and imputed values of the selected variable in a combined way. As shown in Figure 6, the imputation covers the gaps between observed points for all selected variables.

Convergence. Healthy imputations usually occur when the mean

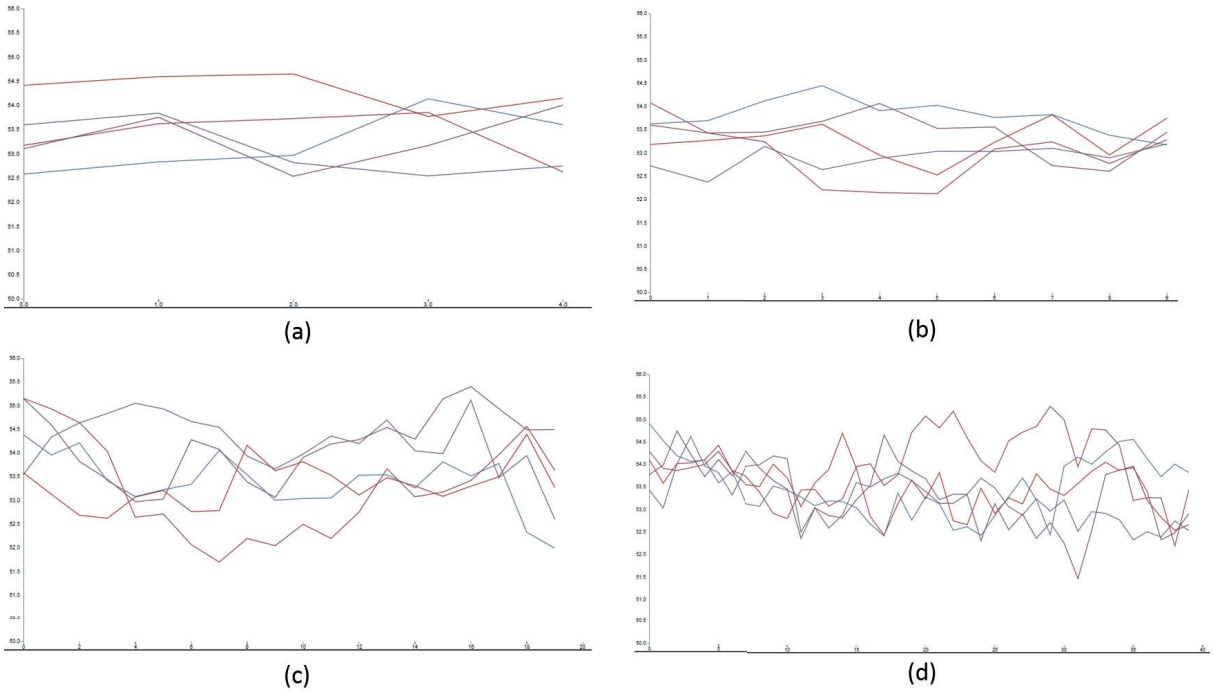

Figure 7: *(a)-(d) show the means of the imputed values of the age variable for the SHIP-1 dataset over different iterations (5, 10, 20 and 40, respectively). As shown in (a), 5 iterations are very low and the curves did not mix together.*

and standard deviation curves are twisted to each other. To monitor the convergence, the mean and standard deviation can be plotted against the number of iterations in the imputation model [BGO11]. Checking the convergence of imputed data helps to understand the imputed dataset to reach a sufficient number of iterations or to understand if it needs an extra iteration. Thus, as the next step to address R6, to expand the menu the user clicks on the convergence corresponding button from the post imputation panel (recall Section 4.2.4). In the following step, the user selects the information to plot (mean or standard deviation), then drags and drops the desired variable from the variable panel to the plot panel. Then, the line chart shows the curves of imputation in different iterations 7.

5. Conclusion and Future Work

In this paper, we presented a web-based system for the exploration and imputation of missing in epidemiological data. The system is designed to help the analyst to inspect the pattern of missingness in longitudinal cohort studies, where the drop-outs are the most common issue for the data being incomplete. The system makes suggestions to predict missing data by finding the correlations the values of baseline variables. The analyst can set parameters to impute the missing values and check the quality of imputations by plots. A categorization of variables is provided to facilitate the interpretation of plots. The basic idea as well as the specific requirements and the usage scenarios are derived from a number of discussions with epidemiologists.

For future work, we are looking for methods to expand the functionality of the system. For example, for finding the correlations of variables via classification rules. Additionally, we plan to add components to the framework for the comparison of different imputation models and consider quadratic or other polynomial regressions in the model. Finally, an evaluation is necessary to study the strengths and limitations of our approach in more detail.

References

[AHN*17] ALEMZADEH S., HIELSCHER T., NIEMANN U., CIBULSKI L., ITTERMANN T., VÖLZKE H., SPILIOPOULOU M., PREIM B.: Subpopulation discovery and validation in epidemiological data. In *Proceedings of the EuroVis Workshop on Visual Analytics* (2017), Eurographics Association.

[BGO11] BUUREN S., GROOTHUIS-OUDSHOORN K.: mice: Multivariate imputation by chained equations in r. *Journal of statistical software 45*, 3 (2011).

[BOH11] BOSTOCK M., OGIEVETSKY V., HEER J.: D^3 data-driven documents. *IEEE transactions on visualization and computer graphics 17*, 12 (2011), 2301–2309.

[CCH*15] CHENG X., COOK D., HOFMANN H., ET AL.: Visually exploring missing values in multivariable data using a graphical user interface. *Journal of Statistical Software 68*, 1 (2015), 1–23.

[CSK01] COLLINS L. M., SCHAFER J. L., KAM C.-M.: A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods 6*, 4 (2001), 330.

[DvdHSM06] DONDERS A. R. T., VAN DER HEIJDEN G. J., STIJNEN T., MOONS K. G.: Review: a gentle introduction to imputation of missing values. *Journal of clinical epidemiology 59*, 10 (2006), 1087–1091.

[EPD05] EATON C., PLAISANT C., DRIZD T.: Visualizing missing data: Graph interpretation user study. *Human-Computer Interaction-INTERACT 2005* (2005), 861–872.

[FG14] FERNSTAD S. J., GLEN R. C.: Visual analysis of missing data - to see what isn't there. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on* (2014), IEEE, pp. 249–250.

[GF95] GREENLAND S., FINKLE W. D.: A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American journal of epidemiology 142*, 12 (1995), 1255–1264.

[GOG07] GRAHAM J. W., OLCHOWSKI A. E., GILREATH T. D.: How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention Science 8*, 3 (2007), 206–213.

[HKB*11] HONAKER J., KING G., BLACKWELL M., ET AL.: Amelia ii: A program for missing data. *Journal of statistical software 45*, 7 (2011), 1–47.

[K*08] KAMPSTRA P., ET AL.: Beanplot: A boxplot alternative for visual comparison of distributions. *Journal of statistical software 28*, 1 (2008), 1–9.

[KLG*16] KLEMM P., LAWONN K., GLASSER S., NIEMANN U., HEGENSCHEID K., VÖLZKE H., PREIM B.: 3d regression heat map analysis of population study data. *IEEE transactions on visualization and computer graphics 22*, 1 (2016), 81–90.

[PKH*16] PREIM B., KLEMM P., HAUSER H., HEGENSCHEID K., OELTZE S., TOENNIES K., VÖLZKE H.: Visual analytics of image-centric cohort studies in epidemiology. In *Visualization in Medicine and Life Sciences III*. Springer, 2016, pp. 221–248.

[SWC*09] STERNE J. A., WHITE I. R., CARLIN J. B., SPRATT M., ROYSTON P., KENWARD M. G., WOOD A. M., CARPENTER J. R.: Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj 338* (2009), b2393.

[TF08] TEMPL M., FILZMOSER P.: Visualization of missing values using the r-package vim. *Reserach report cs-2008-1, Department of Statistics and Probability Therory, Vienna University of Technology* (2008).

[Twi13] TWISK J. W.: *Applied longitudinal data analysis for epidemiology: a practical guide*. Cambridge University Press, 2013.

[VB98] VACH W., BLETTNER M.: Missing data in epidemiologic studies. *Encyclopedia of biostatistics* (1998).

[Vol12] VOLZKE H.: Study of health in pomerania (ship). concept, design and selected results. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 55*, 6-7 (2012), 790–4.

[WRW11] WHITE I. R., ROYSTON P., WOOD A. M.: Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine 30*, 4 (2011), 377–399.

[ZGP15] ZHANG Z., GOTZ D., PERER A.: Iterative cohort analysis and exploration. *Information Visualization 14*, 4 (2015), 289–307.