EG UK Computer Graphics & Visual Computing (2024)
A. Slingsby and D. Hunter (Editors)

Multi-level visualization for exploration of structures in missing data

Sarah Alsufyani[†][1,5], Matthew Forshaw[1], Silvia Del Din[2,3,4],
Alison Yarnall[2,3,4], Lynn Rochester[2,3,4] and Sara Johansson Fernstad[1]

[1]School of Computing, Newcastle University, Newcastle upon Tyne, UK
[2] Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK
[3] National Institute for Health and Care Research (NIHR) Newcastle Biomedical Research Centre (BRC), Newcastle University, UK
[4]The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK
[5] Department of Computer Science, College of Computers and Information Technology, Taif University, Saudi Arabia

Abstract

Missing data refers to the absence of a value in the dataset where it was expected to be present. This absence is common across various fields. It can be caused by a range of factors in the data collection process, and may severely impact analysis through unreliable or biased results. Missing data visualization provides an effective approach to exploring the missing data, recognizing the missingness patterns and structures, and determining optimal solutions through interactive visual interfaces. This paper presents a visualization prototype that incorporates two novel techniques, the MissVisG glyph and the MissVis plot, to support the exploration of missing values in data. The visualization provides an overview of missing values, and helps identify patterns in the data to guide users in selecting appropriate methods for dealing with the missingness. A multi-step evaluation process is utilized to assess and ensure the usability and effectiveness of the visualization.

CCS Concepts

• ***Human-centered computing*** → ***Visualization design and evaluation methods;***

1. Introduction

Missing data, often referred to as missing values or incomplete data, occurs when no data value is stored for a cell in the dataset. This issue is prevalent in various fields - including healthcare, social sciences, and finance - and can significantly affect the data analysis process and cause biased and unreliable results [JJ22]. Missing data can arise from various factors, such as data entry errors, equipment malfunctions, or survey non-responses. Visualization of missing data is crucial in identifying the nature of the missingness, understanding the patterns and mechanisms behind the absence of values, and exploring relationships between missing and observed values. Furthermore, visual analysis can support the analysts in gaining insights into the sources of missingness, e.g., if a particular individual opted out of participation at a certain time point, or if variables were not reported under certain circumstances. As such, visualization can be critical in diagnosing data quality. It supports making informed decisions about data collection, imputation, and analysis strategies, which can improve the quality and reliability of data-driven decisions. Despite its importance, the visual analysis of missing values is relatively unexplored and often overlooked. Many researchers have emphasized the importance of visualization methods for missing data analysis [SS19], but there is a notable lack of suitable techniques for analyzing and visualizing missing data [Joh19]. The need for robust and intuitive visualization methods for missing data has led to some development of novel techniques and tools, but they may not be scalable to the growing data sizes in areas such as medical sciences, and are often focused on representing imputed values rather than on exploration and understanding of missingness structures.

This paper presents a multi-level visualization prototype that allows the user to explore missingness in data using a combined glyph and plot design. The nine-stage framework for design studies [SMM12] was adopted to ensure usefulness and usability, alongside a multi-step evaluation process. The main contributions are:

- A novel visualization prototype that helps explore missingness structures in data at multiple levels of detail;
- MissVisG, a novel glyph style that supports understanding of data quality aspects such as missingness, outliers, and distribution; which can be used to enhance other visualization methods or can function as an independent visualization technique;
- An exemplar of a successful design study and multi-step evaluation process to ensure and evaluate the usability and effectiveness of the designed visualization.

The work was carried out in close collaboration with researchers in biomedical engineering, mobility and digital health, who often struggle with large amounts of missing data. The visualization approaches are, however, generically applicable across domains.

[†] Corresponding author: S.H.H.Alsufyani2@newcastle.ac.uk

DOI: 10.2312/cgvc.20241212

2. Related Work

Viewing missing values as informative signals could provide crucial insights and reveal potential issues in data collection, preprocessing, and analysis processes [JJ22]. Hence, visualizing missing data can serve many purposes; it helps in diagnosing the missingness mechanism, identifying the distribution of missing values, and could be used as a guide in choosing appropriate imputation methods. Fielding et al. [FFR09] and Djurcilov and Pang [DP00] highlighted the importance of visualization in understanding missingness in data and stated that often the absence of the value is more meaningful than replacing them with estimated values, and that users need to be aware of the incompleteness of data. There is, however, limited research introducing novel methods and techniques for missing data visualization. Several early and more recent publications use approaches where missing values are visually separated and represented by colour [Bed90; TCS94; ASMP17; CCH15]. MANET [UHHS96; THSU97] use visual representations such as bar charts, scatter plots, and histograms to visualize missing values. xGobi [SB98] and gGobi [SLBC03] provide techniques for the interactive exploration of missing values and their correlations between variables. Templ et al. [TAF12] introduce the R-package VIM, that supports visualization and imputation by highlighting and exploring the missingness in data. Amelia II [HKB11] provides graphical user interfaces for imputation methods that distinguish imputed missing values from recorded values using colour schemes in a missingness map. While providing important and interesting approaches, a large part of visualization methods designed for missing data analysis may not be able to deal with the growing data sizes in data-generating domains, such as medical sciences. Furthermore, many methods are mainly focused on representing imputed values, and less on exploration and understanding of missingness patterns.

Wong and Vargas [WV12] and Johansson Fernstad and Glen [JG14] emphasize the need for visualization techniques that facilitate the exploration and understanding of missingness in data. However, little research has focused on finding the best techniques to represent missing data. Eaton et al. [EPD05] provides one of the first user studies related to the visualization of missing data that aims to understand users' ability to interpret graphs including missing data. The study found that users may not notice that data is missing when it is replaced by a default value and may be compelled to make general conclusions with partial data even if the missing data is realized. Their results indicated that inadequate representation of missing values negatively impacts interpretation and recommended enhancing visualizations with dedicated visual attributes to highlight the presence of missing data. However, they did not specify which visual attributes would be most appropriate for missing data visualization. Andreasson and Riveiro [AR14] evaluated the impact on decision-making of three techniques for representing missing values visually: Emptiness, Fuzziness and Emptiness plus explanation. Their empirical study found that emptiness plus explanation was the most preferred technique with the highest degree of decision confidence. Song and Szafir [SS19] discusses how imputation and visualization methods of missing values can influence analysts' perceptions of data quality and their confidence in their conclusions. They use three categories to encode imputed values (highlight, downplay, and annotation), information removal, and two common visualizations (bar charts and line graphs). Their study found that highlighting the missing data resulted in higher perceived confidence and data quality, while downplaying and removing information are perceived as lower quality. Ruddle et al. [RAH22] provided interactive set visualisation techniques to identify missing data patterns and derive actionable insights. The visualizations include bar charts for sets, heatmaps to represent set intersections, and histograms to display the distributions within both sets and intersections. Johansson Fernstad [Joh19] evaluated scatter plot matrix, heat map and parallel coordinates (PC) from the VIM package, which are enhanced by using visual attributes to present missing values, to identify the three types of missingness patterns. They found that the heat map with missing values represented by colour performed best for Amount Missing (AM) and Joint Missingness (JM) tasks, while PC with missing values represented above the axis performed better for Conditional Missingness (CM) tasks. Following this, MissiG [JJ22], a glyph-based visualization technique, was designed to support the exploration of missingness structures. MissiG represents the amount missing in variables using a combination of histograms and bar charts, the joint missingness across variables, the distribution of recorded values and the relationship between missing and recorded across variables using the missingness patterns defined in [Joh19]. Unlike existing techniques that primarily focus on the amount of missing data and missing patterns, our paper introduces novel visualization methods that facilitate the exploration of missing values in data. These methods take into account variables' distributions and outliers, which can further aid in the selection of an appropriate imputation method.

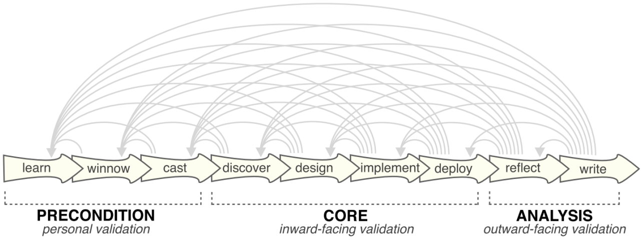


Figure 1: *Nine-stage design study methodology framework classified into categories. [SMM12]*

3. Methodology

This work utilises the methodology of Sedlmair et al. [SMM12], including three main phases (Figure 1). First, the *precondition* phase, within which we prepared for the work by conducting a comprehensive review of the existing missing data visualization techniques and establishing collaboration with experts in biomedical engineering and digital health in the Brain and Movement Research Group at Newcastle University, UK. They struggle with large amounts of missing data, with diverse causes, such as technical failures in data collection and study participant dropout. Second, the *core* phase, where collaboration with domain experts included meetings and brainstorming sessions to formulate design requirements, begin the iterative design process of the visualization solution, implement the prototype, and seek feedback from our collaborators (Section 6.1). Third, the *analysis* phase, where we refined the final design, proposed guidelines, and composed this paper. The reminder of this paper mainly focuses on describing the core phase.

4. The visualization design

This section will describe the MissVis method, which was designed to support exploration of missingness patterns in incomplete datasets which are considerably affected by large amounts of missing values. The design is separated into a glyph design (Section 4.1) building on glyph design criteria [BKC*13], and a multivariate plot design (Section 4.2) which integrates the glyph. The underlying design considerations are discussed in Section 4.3.

4.1. MissVisG Design

MissVisG was designed as a glyph style that can be used to enhance other visualization methods (Figure 4.2), as well as to function independently as a stand alone visualization technique. The overall goal of the glyph is to help the users gain insights into the distribution of missingness in the data, in relation to overall data distribution, explore missingness, and aid in the selection of an appropriate imputation method. To determine the best approach for handling missing values, it is essential not only to be aware of the proportion of missing values in the data but also to identify any underlying structure to the missingness and, if present, the type of structure. The MissVisG represents a variable in the data in a rectangle shape combining a stacked bar chart and box plot (Figure 2). Following the thorough review and consultations with domain experts, it was agreed that the designed glyph would encode the following elements:

- Percentage of missing values and recorded values were presented as a stacked bar chart, separated into the percentage of missing values in red colour and the percentage of recorded values in blue colour.
- The variable distribution was portrayed via a box plot. The variable distribution is an important feature to represent as it helps to choose the proper imputation method, with improper imputation often impacting the distribution, and can hence be used to evaluate the imputation method by comparing the distribution before and after imputation [SSH*18].
- Outliers were also plotted using a box plot. The visualization of outliers was included as it adds further aspect of data quality that may impact decisions on how to deal with missing data, as outliers can be addressed through removal or replacement methods [KK17] similar to those used for missing values.

4.2. Plot Design

This section will describe the MissVis plot (see Figure 3), a novel visualization technique designed to facilitate the exploration of missingness, which has been improved through the integration of the MissVisG glyph. Data analysts are interested in examining the amount of missing values, understanding their correlations, and identifying meaningful patterns among the recorded and missing values [ANI*17]. The goal of the MissVis plot is to help the users gain further insights into the missingness by supporting comparison of multiple variables in terms of missing data patterns, distributions, relationships of missing values and recorded values, and outliers.

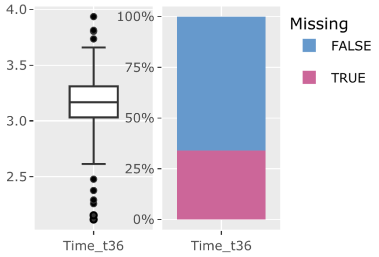


Figure 2: *The MissVisG glyph structure shows the amount of missing values in red colour, the amount of recorded values in blue colour, the distribution of the variable, and its outliers.*

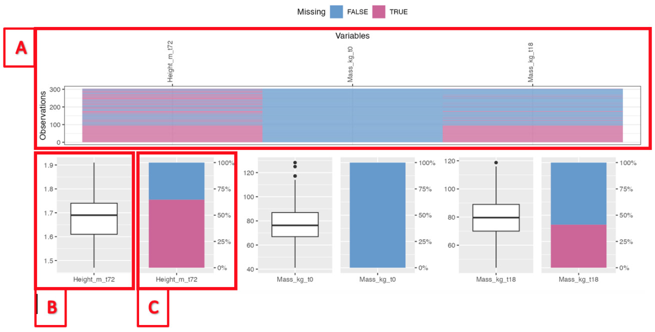


Figure 3: *MissVis plot: A) The heat map represents the missing data pattern. B) The box plot represents the variable distribution and outliers. C) The bar chart represents the percentage of missing values and recorded values in the variable.*

Following literature review and consultations with domain experts, it was agreed that the plot would encode the missing data pattern. Since a straightforward method to visualize missing and observed values is through a missingness map, a heat map, which plots whether the observation is missing or recorded, was used (Figure 3A). This technique shows the exact location of the missing values in each variable, it facilitates the comparison among multiple variables and explores the relation of missingness among the selected variables including the missing patterns. The values that do not exist are coloured in red and the recorded values are blue. The plot was enhanced using MissVisG under each variable to represent the percentage of missing values and recorded values (Figure 3C), the distribution, and outliers (Figure 3B).

4.3. Design Considerations

Typically, a glyph is designed using a number of visual channels, where each visual channel encodes a specific attribute of a data entity. The MissVisG and the MissVis plot were designed based on the principles and guidelines provided by [CLP*15] and [BKC*13]. MissVisG uses three main visual channels: colour hue to distinguish between missing values and the recorded values, size/length to represent magnitude and shape for representation of distribution and outliers (Figure 2). The MissVis plot integrates a heat map with the MissVisG glyph (Figure 3), using four main visual channels:

position to distinguish the location and relationship of missing and recorded values within and across variables, colour hue to distinguish between missing and recorded values, size/length to represent magnitude, and shape for representation of distribution and outliers. The visualization design was based on these components and several well-established glyph design principles [CLP*15] and and [BKC*13], such as: learnability, typedness, separability, searchability, and pop-out effect.

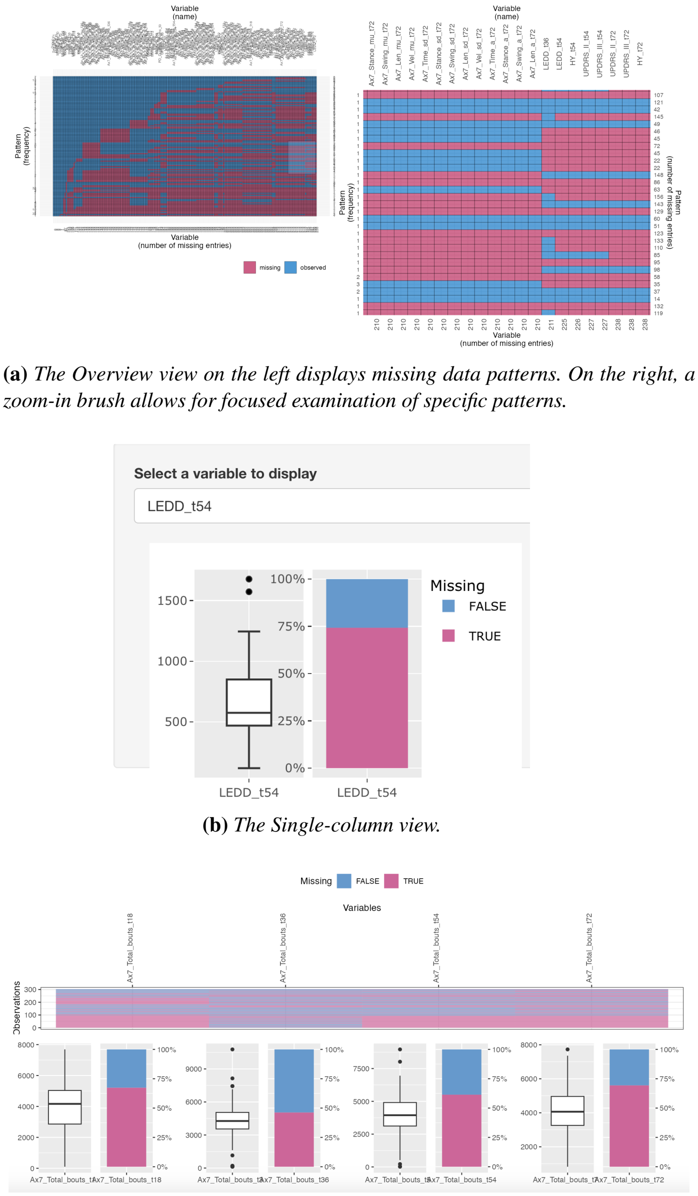
5. Implementation

The interface of the MissVis prototype was implemented using R Shiny and utilizes a combination of MissVisG and MissVis plot to support the exploration of missing data. The prototype adapted Shneiderman's information-seeking mantra [Shn03], which outlines the essential components of interaction in visualization, including overview first, zoom and filter, then details-on-demand. Building on this approach, the system includes three tabs (Figure 4): overview view, single-column view, and multi-column view.

Overview view: On the left, the `ggmice` package [**ggmice**] is used to plot the missing data patterns as rows, and variables as columns. On the top x-axis, the variables are ordered by number of missing patterns (the rightmost variable is involved in the largest number of the missingness patterns). On the bottom x-axis, the number of missing entries per variable is represented. The number of missing entries per pattern is represented on the right y-axis, while the pattern frequency in the whole dataset is represented on the left y-axis. On the right, a zoom-in brush (visible as a lighter area in the main plot) is embedded to adjust the level of details to improve the readability and focus on certain patterns (Figure 4a). The overview helps in understanding overall missingness patterns as well as the identification of variables of interest for further examination in the single-column and multi-column views.

Single-column view: This view includes MissVisG as an independent visualization technique (Figure 4b). The user can select any variable to explore, and gain more insight about it through the glyph. Tooltips are used to provide detail on variable name, number of missing values, and recorded values. The user can also inspect the distribution within the variable and its outliers.

Multi-column view: This view (Figure 4c) uses the MissVis plot and allows the user to select multiple variables to compare. The missingness map, in the top part of the plot, offers a compact overview of the missing pattern including missing values and recorded values. This map illustrates the locations of missing values for each variable, enabling identification of multivariate patterns. As described in section 4.2, MissVisG is included to enhance the plot and support further understanding of distribution and outliers. Columns are sorted based on the user selection and can be manually modified.



(a) *The Overview view on the left displays missing data patterns. On the right, a zoom-in brush allows for focused examination of specific patterns.*

Select a variable to display

LEDD_t54

(b) *The Single-column view.*

(c) *The Multi-column view*

Figure 4: *The three tabs of the MissVis prototype.*

6. Evaluation of the visualization

To ensure the usefulness and usability of the visualization, a multi-step validation and evaluation process was used, including collaboration with experts, heuristics evaluation, and user testing. For the evaluation, a real-life walking monitoring dataset from the Incidence of Cognitive Impairment in Cohorts with Longitudinal Evaluation (ICICLE) study [YBD*14] was used. This dataset contains data recorded on Parkinson's disease patients, and includes about 30 attributes related to gait characteristics and walking behaviour, which was collected longitudinally across six years.

6.1. Collaboration with experts

Følstad [Føl07] found that engaging domain experts as evaluators in a usability assessment often leads to highly significant outcomes. Consequently, we adopted this strategy by following four steps:

1. **Identify the right experts**: In the early design phase, collaboration with domain experts was established to ensure the right design choices to meet their actual needs.
2. **Meetings with experts**: Meetings took place online and in person before and during the design process. During these, the visualization and the prototype were discussed using techniques such as think-aloud and brainstorming. We took notes and occasionally recorded meetings to maintain focus.
3. **Seek feedback and iteration**: Early versions of the prototype were presented to identify any drawbacks in the design and implementation. For this, think-aloud protocol was used which allowed the domain experts to describe their thoughts and requirements. The experts' feedback was then used to guide the design process and evaluate the end product.
4. **Review findings**: Key insights and recommendations were identified from these meetings. We then carefully analyzed the derived data to extract meanings and drawbacks.

6.2. Heuristics evaluation

6.2.1. Procedure

To evaluate the designed visualization, a heuristics evaluation, with visualization and digital health experts, was conducted in the format of one-to-one semi-structured interviews. The goals of this step were to assess user acceptance of the designed visualization and prototype, to ensure their usefulness and usability, and to collect feedback for updates and refinements. Wall et al. [WAM*19] provided a heuristic technique for quantifying the potential benefit of visualisation in terms of data comprehension. Their value evaluation hierarchy framework includes four components (ICE-T): Insight, Confidence, Essence, and Time. The heuristics evaluation and analysis presented in this paper is inspired by Wall et al.'s technique.

The interviews were conducted separately for each participant, starting with a brief introduction going through each view of the prototype, highlighting its key features and benefits, followed by a live demonstration of its various functions and customization options. Following this, the participants took over control to interact with the visualization and the prototype by themselves. They were encouraged to explore the missingness in the data, find any patterns, and trying to gain overall understanding of the missingness in the dataset. Participants were allowed to ask questions and further interact with the study lead at any time during the session. The sessions took on average 40 minutes and were recorded to allow full focus on the interview rather than note-taking. After the interview, the participants were asked to fill out and return a form with 16 heuristic items (Table 1), and respond using a 5-point Likert scale (1: Strongly disagree, 2: Disagree, 3: Neither agree nor disagree, 4: Agree, and 5: Strongly agree). Additionally, a comments space was provided with each heuristic to allow participants to include more detailed comments. The heuristics were grouped based on their relevance to specific views and functionality (component). Each heuristic was accompanied by a description and screenshot of the visualization to facilitate the evaluation process.

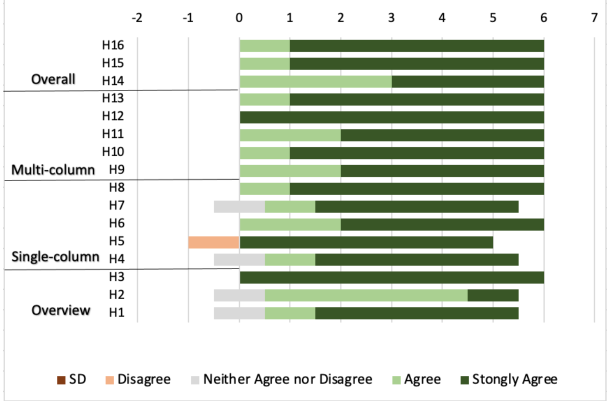


Figure 5: *Number of positive and negative Likert scale responses for each heuristic.*

6.2.2. Participants

This evaluation was run as a qualitative study with 6 participants, of which two had expertise in bio-engineering and digital health, and four were visualization experts.

6.2.3. Interviews

These interviews were run online using Microsoft Teams that allowed the participants to take control and interact with the prototype and the visualizations. Prior to interviews, ethical approval was received from Newcastle University. Initial pilot interviews were run with two PhD students researching Data Visualization at Newcastle University, to ensure the performance of the prototype over Microsoft Teams and to refine the study. An email was sent to the participant before the interviews, including an overview of the evaluation, a consent form, and a series of heuristics based question (Table 1) used to evaluate the visualization prototype.

The heuristics results are reported in Table 2 and Figure 5, showing clearly positive responses across all heuristics. Table 2 displays the summary ratings of the participants on each heuristic with the average, standard deviation, maximum and minimum. The score aggregation metric in Wall et al. [WAM*19] was used in order to obtain an overall value rating for the visualization. Heuristics were grouped into components based on their similarity, structured as overview: H1-H3, single-column: H4-H8, multi-column: H9-H13, and overall: H14-H16 (Table 1). The Likert score of each heuristic is defined as s_h rating from 1-5. Each component score, s_c, is computed by averaging the values of its associated heuristics, where j represents the number of associated heuristics: $s_c = \frac{1}{j}\sum_{i=1}^{j} s_{h,i}$. Finally, the overall visualization score is computed by averaging all top-level components $s = \frac{1}{4}(s_{overview} + s_{single-column} + s_{multi-column} + s_{overall})$ (Table 3).

The result (Table 3) was generally positive and most participants favored the prototype. The lowest average group score was 3.6 (with a score of 3 indicating a neutral rating). Figure 5 provides an overview of the number of positive and negative scores for each heuristic. The multi-column component (H9-H13) was the highest scored, followed by overall (H14-H16) and single-column

Proceedings published by Eurographics - The European Association for Computer Graphics.

Component	H ID	Heuristic
Overview	H1	The overview is clear and intuitive
	H2	The overview clearly shows the number of missing entries.
	H3	The overview shows the pattern frequency.
Single-column	H4	The combined chart clearly shows the amount of missingness (%) of each selected variable.
	H5	The combined chart clearly identifies the outliers of each selected variable.
	H6	The combined chart clearly visualizes the distribution of each selected variable.
	H7	The combined chart clearly shows the max and min values of each selected column.
	H8	It is easy and intuitive to interactively select a variable to gain more information about it.
Multi-column	H9	The combined chart clearly shows the amount of missingness of all selected variables.
	H10	The combined chart clearly identifies the outliers of all selected variables.
	H11	The combined chart clearly visualizes the distribution of all selected variables.
	H12	Visualizing multiple columns together supports the ease of comparing them and provides more insights about missingness.
	H13	It is easy and intuitive to interactively select multiple variables to gain more information about them.
Overall	H14	The visualizations could provide a good guide to which variables are interesting to take a closer look and deal with missing values.
	H15	The visualizations help finding missing patterns more easily than other tools.
	H16	The visualization helps generate data-driven questions.

Table 1: *Heuristics set used within evaluation interviews.*

Heuristics	P1	P2	P3	P4	P5	P6	Avg	Std	Max	Min
H1	5	5	3	5	5	4	4.50	0.84	5	3
H2	4	4	4	4	5	3	4.00	0.63	5	3
H3	5	5	5	5	5	5	5.00	0.00	5	5
H4	5	5	3	5	5	4	4.50	0.84	5	3
H5	5	5	2	5	5	5	4.50	1.22	5	2
H6	5	4	5	5	5	4	4.67	0.52	5	4
H7	5	5	3	5	5	4	4.50	0.84	5	3
H8	4	5	5	5	5	5	4.83	0.41	5	4
H9	5	4	4	5	5	5	4.67	0.52	5	4
H10	5	5	4	5	5	5	4.83	0.41	5	4
H11	5	4	5	5	5	4	4.67	0.52	5	4
H12	5	5	5	5	5	5	5.00	0.00	5	5
H13	4	5	5	5	5	5	4.83	0.41	5	4
H14	4	5	4	5	5	4	4.50	0.55	5	4
H15	4	5	5	5	5	5	4.83	0.41	5	4
H16	5	4	5	5	5	5	4.83	0.41	5	4

Table 2: *Scores and statistics of the six participants (columns) for each heuristic (rows).*

Participant	Overview	Single-column	Multi-column	Overall
P1	4.7	4.8	4.8	4.3
P2	4.7	4.8	4.6	4.7
P3	4	3.6	4.6	4.7
P4	4.7	5	5	4.7
P5	5	5	5	5
P6	4	4.4	4.8	4.7
Component Avg	4.5	4.6	4.8	4.7

Table 3: *Score aggregation of the six participants.*

(H4-H8) components. Scores for each group were relatively consistent among participants. These results confirm the usability of the prototype and indicates the potential of the visualization as a useful tool to explore and gain insight into missingness patterns in data.

6.2.4. Feedback and Discussion

The overall reaction to the prototype was positive. Most of the provided feedback was minor, requesting small adjustments or additions.

Overview: Overall, the view that provides an overview of the data was well received. The view was generally interpreted well by all interviewees. The general amount of missing data and the frequency were recognised well. P3 stated *"Great to have an overview of the amount of missing data"*. The colours were well perceived: *“The red and blue contrast well”*. P5 stated *"The rearranging of variables based on missing values helped to recognize the pattern"*. Three interviewees out of five commented on the limitations of visualizing the large number of variables. One of them said *“Not able to actually quantify the precise "numbers" (Y axis on the right quite small)”*. Another interviewee commented on not being able to read the variables’ names. Two of the interviewees suggested adding some statistical data such as the percentage of missing values and the number of missing and observed data in general.

Single-column: The general response to the single-column view was positive. P5 said *"The interface is clear and easy to use."* However, potential areas of improvement were also indicated. P2 stated *"For showing the distribution could be nice to also have an actual distribution plot, on the right side of the column (like half of a violin plot)"*. Another request was to add the percentage of missing values and recorded values inside the bar char, P3 commented *"Would be useful to include the number of missing/observed data points as labels inside their respective box"*. Finally, P6 stated *"Would be great to include the legend"*.

Multi-column: All participants liked this view and found it very useful. The visualization was well-received and easily comprehended by all. P3 commented *"This is particularly useful when dealing with longitudinal data from ICICLE"*. Most of the comments were the same as in the single-column view. A potential scalability issue if comparing a large number of variables was highlighted by P1 *"Maybe it would be an issue if you needed to compare 20+ charts side by side"*.

Overall: All participants agreed on the usefulness and effectiveness of the prototype and its visualizations. P1 stated *"I think this would be an excellent tool"*, and P3 commented *"Some fantastic visualizations, that could be very useful with a few tweaks!"*

We have treated all identified issues seriously and have made the following changes to the prototype in response to this feedback:

- Focusing on a smaller subset of data to improve the readability, so we added interactive features in the overview tab to zoom in, which helped to improve the readability and reduce the clutter.
- Updating the MissVisG to show the amount of missing values

and recorded values inside the stacked bar chart, which improves readability.

- Adding a legend in each view, not only the first view, so the user does not need to recall any information from the previous views.

Our immediate priorities for development based on this feedback include adding basic statistical breakdowns on the number of missing/observed data in the overview view, and including more interactive features.

6.3. User testing

The last step in the multi-step evaluation process was the user testing. The primary goals from this step were to determine whether domain experts are indeed helped by the visualization prototype, and to get an additional level of feedback for improving the visualizations.

6.3.1. Participants

The prototype was presented to one expert in bio-engineering and digital health, and one data visualization expert, to enable capturing of a broader set of feedback. Both participants were affiliated with Newcastle University and also took part in the heuristics evaluation. A consent form and a short overview of the evaluation process were sent to the participants prior to the test.

6.3.2. Interviews

The user testing was conducted in person as individual interviews, which took approximately one hour for each participant. For consistency, the same interview structure was used for both participants, starting with a short introduction to the prototype, its functionalities, and the structure of the evaluation. After the introduction, participants were given a chance to explore the data and interact with the prototype by themselves. Participants were assigned tasks involving the use of the prototype, such as exploring the missingness in the dataset in general, the percentage of missing values and recorded values in some variables, identifying missing patterns, the distribution, and the outliers of some selected variables. During the interviews, participants were allowed to ask any question at any time. To ensure the reliability and validity of the results, the triangulation technique [Pri21] was used. Meaning that more than one method was used to collect data (taking notes, thinking aloud technique, and recording the screen and the audio). After the interaction phase, semi-structured interviews were conducted using 10 open questions:

1. Do you think that this prototype can help to investigate data quality, particularly missing data?
2. Is this prototype suitable for presenting all the information necessary to investigate data quality, particularly missing (as a variable and subset)?
3. Do you think this prototype can help identify missingness patterns?
4. Do you think that this prototype can help to more easily understand the missingness pattern?
5. Do you think that this prototype can help assess whether the outliers should be treated as missing values or recorded values?
6. Do you think this prototype can help determine which imputation methods are preferred?
7. What are the prototype's strengths? Where do you see the potential for improvement?
8. Does the prototype help make new findings comparing other tools?
9. Have you seen any tool similar to this prototype?
10. Other notes?

Subsequently, both interviews were transcribed, reviewed by participants, and then analyzed.

6.3.3. Results

First interview: The interviewee carried out the given tasks with confidence and navigated between views easily. As they had limited experience of the ICICLE dataset, considerable time was initially spent exploring the data. This interviewee used the overview's interactive zoom-in feature extensively to check the missingness patterns in more detail, and suggested adding further instructions on how to use the zoom-in feature. Following the initial examination, each task was easily and quickly carried out, using the overview, single-column, and multi-column views. Tasks related to identifying variables with no-missing values and the most missing values were answered quickly, with the interviewee realizing the missingness patterns in the data at a glance using the colours. The multi-column view was used to investigate distribution and outliers for multiple variables, and the interviewee stated that they liked the visualization and how it made it easy to compare variables. **Second interview:** The interviewee started by familiarizing themselves with the system through the views. They then continued to investigate the dataset using the overview. This interviewee was familiar with the ICICLE dataset and, hence, more confident in selecting variables and finding patterns. Following the initial exploration, all views and zoom in features were used effectively and confidently to carry out the tasks, without needing to ask any questions. They highlighted that it was useful to check the outliers in the data, to investigate the overall quality, and that the combined representation of distribution helped reviewing how to treat the missing and outlier values. This interviewee also spotted patterns and variables with missing values quickly through the colours. Moreover, they started to draw conclusions by using a combination of views and found reasons for the missing values in some variables (sensors and non-responses issues). For example, in Figure 4c, it was clearly visible from patterns in the visualization that the missingness pattern and distribution for the second column, recorded at month 36, differs considerably from the patterns and distribution for months 18, 54 and 72. Calling for further investigation into the cause.

6.3.4. Discussions

The interviewees were positive to the visualization prototype, as evidenced by their interview responses. Both quickly understood how to interact with the data and were able to navigate between all views within a few minutes. Both participants agreed that the prototype presents important information and can help to investigate data quality, particularly missing data both as individual variables and subsets. The interviews also demonstrated that the prototype could be successfully used to identify where specific data is missing and its relationship to other missing data. They affirmed that the overview clearly shows the missing data, while the other views provide the tools to analyse them in detail. One participant said: *"Just*

Proceedings published by Eurographics - The European Association for Computer Graphics.

from looking at the overview, it was clear that there were large patterns of missing data without having to look too hard, and the other views had the tools needed to further explore patterns discovered in the overview". The other participant stated that it was easy to take a closer look at each variable and gain more information about its distribution, outliers, and the percentage of missing values. The interviewees liked the interactive features, particularly the zoom-in brush that improves readability, and mentioned that details on demand was beneficial and made it easy to get all information when needed. For strengths, they stated that the prototype does a good job of exploring missingness in the data. They also said that while there is a bit of a learning curve, the help section and intuitive design helps to understand it quickly. The general findings of the study infer that the prototype facilitate identification of missingness and outliers in the data, and could be used as a guide into how to treat them effectively. A few improvement suggestions were also made:

1. Include more ways to filter the overview, though the zoom-in helps a lot in this regard.
2. Hovering near variable axes could have a fisheye lens effect to display detail clearer.
3. Interactively auto-load a variable in the other views, by selection in the overview.
4. Possibility to compare multiple datasets side by side
5. Displaying information input from data collection forms could further help to identify specific reasons why data is missing.

6.4. Summary

The multi-step evaluation process, including collaboration with experts, heuristics evaluation and qualitative user testing, resulted in the design of a visualization prototype that overall was perceived as useful and usable, as indicated by feedback and heuristic scores. Most requests for improvements changes were minor and has subsequently been implemented, while some will be part of future work. These include: enabling comparison and exploration of multiple datasets side by side, further coordination across views by allowing selection in the overview to automatically load variables to the single-column and multi-column views, and implementing additional methods for filtering of the overview. The final findings of the evaluation concluded that the prototype provides effective ways of getting an overview of the missing values, identifying patterns of missing data, and serving as a valuable guide to help understand the causes of missingness and how to deal with it.

7. Conclusion and future directions

This paper presented a visualization prototype for exploration of missing values, including the novel MissVisG glyph style visualization and MissVis plot. These visualizations offer a comprehensive overview of missing values and help identify outliers and patterns in the data, which can provide valuable guidance on how to best address them. The MissVisG, which integrates a stacked bar chart and box plot, can enhance other visualizations or function independently as a standalone visualization technique. The MissVis plot, a missingness map that is enhanced by MissVisG, facilitates the comparison among variables in terms of exploring the missingness in the data. The multi-step evaluation demonstrated that the MissVisG and the MissVis plot provide an effective way to get an overview of the missing values, identify patterns of missing data, and that they could serve as a guide to choosing the appropriate method for dealing with data quality issues relating to missing values. There were several limitations in the evaluation process. Conducting the heuristics evaluation online posed a significant limitation, albeit we aimed to ensure participants had a similar experience to what they would have had in person. The follow on user testing was conducted in person to ensure direct interaction with the prototype by both interviewees. To limit the risks of relying too much input from a single interviewee, which may not suffice to justify modifications, we reviewed that feedback aligned with the main objectives of the prototype design, and weighed in the number of people providing similar feedback. Finally, the user testing only included two participants, due to limitations in experts availability. In the future, we aim to enhance the prototype by enabling users to upload multiple datasets, improving interaction features, and refining the visualizations to minimize clutter when comparing a large number of variables.

Acknowledgements

The authors would like to thank the participants and assessors of the ICICLE study. The ICICLE study was supported by Parkinson's UK (J-0802, G-1301) and by the NIHR Newcastle Biomedical Research Centre at Newcastle Upon Tyne Hospital NHS Foundation Trust and Newcastle University. The work was also supported by the NIHR/Wellcome Trust Clinical Research Facility infrastructure at Newcastle upon Tyne Hospitals NHS Foundation Trust. All opinions are those of the authors and not the funders.

Ethical approval

The study has been approved by Newcastle University.

References

[ANI*17] Alemzadeh, S., Niemann, U., Ittermann, T., et al. "Visual analytics of missing data in epidemiological cohort studies". *VCBM 2017 - Eurographics Workshop on Visual Computing for Biology and Medicine* (2017), 43–51 3.

[AR14] Andreasson, Rebecca and Riveiro, Maria. "Effects of Visualizing Missing Data: An Empirical Evaluation". *2014 18th International Conference on Information Visualisation*. IEEE, 2014, 132–138 2.

[ASMP17] Arbesser, Clemens, Spechtenhauser, Florian, Muhlbacher, Thomas, and Piringer, Harald. "Visplause: Visual Data Quality Assessment of Many Time Series Using Plausibility Checks". *IEEE Transactions on Visualization and Computer Graphics* 23 (2017) 2.

[Bed90] Beddow, Jeff. "Shape coding of multidimensional data on a microcomputer display". *Proceedings of the First IEEE Conference on Visualization: Visualization '90*. IEEE Comput. Soc. Press, 1990 2.

[BKC*13] Borgo, Rita, Kehrer, Johannes, Chung, David H S, et al. "Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications". *Eurographics State of the Art Reports* July 2014 (2013), 39–63 3, 4.

[CCH15] Cheng, Xiaoyue, Cook, Dianne, and Hofmann, Heike. "Visually Exploring Missing Values in Multivariable Data Using a Graphical User Interface". *Journal of Statistical Software* 6 (2015) 2.

[CLP*15] Chung, David H.S., Legg, Philip A., Parry, Matthew L., et al. "Glyph sorting: Interactive visualization for multi-dimensional data". *Information Visualization* 14 (2015), 76–90 3, 4.

[DP00] Djurcilov, Suzana and Pang, Alex. "Visualizing sparse gridded data sets". *IEEE Computer Graphics and Applications* 20.5 (2000), 52–57 2.

[EPD05] Eaton, Cyntrica, Plaisant, Catherine, and Drizd, Terence. "Visualizing Missing Data: Graph Interpretation User Study". Vol. 3585 LNCS. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, 861–872 2.

[FFR09] Fielding, Shona, Fayers, Peter M., and Ramsay, Craig R. "Investigating the missing data mechanism in quality of life outcomes: A comparison of approaches". *Health and Quality of Life Outcomes* 7 (2009) 2.

[Føl07] Følstad, AsbjøRn. "Work-Domain Experts as Evaluators: Usability Inspection of Domain-Specific Work-Support Systems". *International Journal of Human-Computer Interaction* 22 (2007) 4.

[HKB11] Honaker, J., King, G., and Blackwell, M. "Amelia II: A program for missing dataJournal of Statistical Software". *Journal of Statistical Software* (2011), 1–47 2.

[JG14] Johansson Fernstad, Sara and Glen, Robert C. "Visual analysis of missing data - To see what isn't there". *2014 IEEE Conference on Visual Analytics Science and Technology, VAST 2014 - Proceedings* (2014) 2.

[JJ22] Johansson Fernstad, Sara and Johansson Westberg, Jimmy. "To Explore What Isn't There—Glyph-Based Visualization for Analysis of Missing Values". *IEEE Transactions on Visualization and Computer Graphics* 28 (2022) 1, 2.

[Joh19] Johansson Fernstad, Sara. "To identify what is not there: A definition of missingness patterns and evaluation of missing value visualization". *Information Visualization* 18.2 (2019), 230–250 1, 2.

[KK17] Kwak, Sang Kyu and Kim, Jong Hae. "Statistical data preparation: management of missing values and outliers". *Korean Journal of Anesthesiology* 70 (2017), 407 3.

[Pri21] Priya, Arya. "Case Study Methodology of Qualitative Research: Key Attributes and Navigating the Conundrums in Its Application". *Sociological Bulletin* 70 (2021), 94–110 7.

[RAH22] Ruddle, Roy A., Adnan, Muhammad, and Hall, Marlous. "Using set visualisation to find and explain patterns of missing values: a case study with NHS hospital episode statistics data". *BMJ Open* 12 (2022) 2.

[SB98] Swayne, Deborah F. and Buja, Andreas. "Missing data in interactive high-dimensional data visualization". *Computational Statistics* 13.1 (1998), 15–26 2.

[Shn03] Shneiderman, Ben. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations". *The Craft of Information Visualization*. Elsevier, 2003, 364–371 4.

[SLBC03] Swayne, Deborah F, Lang, Duncan Temple, Buja, Andreas, and Cook, Dianne. "GGobi: evolving from XGobi into an extensible framework for interactive data visualization". *Computational Statistics Data Analysis* 43 (2003), 423–444. issn: 01679473 2.

[SMM12] Sedlmair, Michael, Meyer, Miriah, and Munzner, Tamara. "Design study methodology: Reflections from the trenches and the stacks". *IEEE Transactions on Visualization and Computer Graphics* 18 (2012) 1, 2.

[SS19] Song, Hayeong and Szafir, Danielle Albers. "Where's My Data? Evaluating Visualizations with Missing Data". *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), 914–924 1, 2.

[SSH*18] Soares, Jastin, Santos, Miriam, Henriques Abreu, Pedro, et al. "Exploring the Effects of Data Distribution in Missing Data Imputation". 2018, 251–263 3.

[TAF12] Templ, Matthias, Alfons, Andreas, and Filzmoser, Peter. "Exploring incomplete data using visualization techniques". *Advances in Data Analysis and Classification* 1 (2012), 29–47 2.

[TCS94] Twiddy, Ray, Cavallo, John, and Shiri, S.M. "Restorer: a visualization technique for handling missing data". *Proceedings Visualization '94*. IEEE Comput. Soc. Press, 1994, 212–216 2.

[THSU97] Theus, Martin, Hofmann, Heike, Siegl, Bernd, and Unwin, Antony. "MANET Extensions to Interactive Statistical Graphics for Missing Values". 1997 2.

[UHHS96] Unwin, Antony, Hawkins, George Van Duzer, Hofmann, Heike, and Siegl, Bernd. "Interactive Graphics for Data Sets with Missing Values—MANET". *Journal of Computational and Graphical Statistics* (1996), 113–122 2.

[WAM*19] Wall, Emily, Agnihotri, Meeshu, Matzen, Laura, et al. "A Heuristic Approach to Value-Driven Evaluation of Visualizations". *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), 491–500 5.

[WV12] Wong, B.L. William and Varga, Margaret. "Black Holes, Keyholes And Brown Worms: Challenges In Sense Making". *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 56 (2012) 2.

[YBD*14] Yarnall, Alison J, Breen, David P, Duncan, Gordon W, et al. "Characterizing mild cognitive impairment in incident Parkinson disease: the ICICLE-PD study". *Neurology* 82.4 (2014), 308–316 4.

Proceedings published by Eurographics - The European Association for Computer Graphics.