

mosaicQA – A General Approach to Facilitate Basic Data Quality Assurance for Epidemiological Research

Martin Bialke^{1*}; Henriette Rau^{1*}; Thea Schwaneberg¹; Rene Walk²; Thomas Bahls¹; Wolfgang Hoffmann¹

¹Institute for Community Medicine, Section Epidemiology of Health Care and Community Health, University Medicine Greifswald, Greifswald, Germany;

²Institute for Community Medicine, Section GANI_MED, University Medicine Greifswald, Greifswald, Germany

Keywords

Medical data management, data quality assurance

Summary

Background: Epidemiological studies are based on a considerable amount of personal, medical and socio-economic data. To answer research questions with reliable results, epidemiological research projects face the challenge of providing high quality data. Consequently, gathered data has to be reviewed continuously during the data collection period.

Objectives: This article describes the development of the mosaicQA-library for

non-statistical experts consisting of a set of reusable R functions to provide support for a basic data quality assurance for a wide range of application scenarios in epidemiological research.

Methods: To generate valid quality reports for various scenarios and data sets, a general and flexible development approach was needed. As a first step, a set of quality-related questions, targeting quality aspects on a more general level, was identified. The next step included the design of specific R-scripts to produce proper reports for metric and categorical data. For more flexibility, the third development step focussed on the generalization of the developed R-scripts, e.g. extracting charac-

teristics and parameters. As a last step the generic characteristics of the developed R functionalities and generated reports have been evaluated using different metric and categorical datasets.

Results: The developed mosaicQA-library generates basic data quality reports for multivariate input data. If needed, more detailed results for single-variable data, including definition of units, variables, descriptions, code lists and categories of qualified misings, can easily be produced.

Conclusions: The mosaicQA-library enables researchers to generate reports for various kinds of metric and categorical data without the need for computational or scripting knowledge. At the moment, the library focusses on the data structure quality and supports the assessment of several quality indicators, including frequency, distribution and plausibility of research variables as well as the occurrence of missing and extreme values. To simplify the installation process, mosaicQA has been released as an official R-package.

Correspondence to:

Martin Bialke
Institute for Community Medicine
Section Epidemiology of Health Care and
Community Health
University Medicine Greifswald
Ellernholzstr. 1–2
17487 Greifswald
Germany
E-mail: martin.bialke@uni-greifswald.de

Methods Inf Med 2017; 56(Open): e67–e73
<https://doi.org/10.3414/ME16-01-0123>

received: October 17, 2016

accepted: April 6, 2017

published: April 29, 2017

Funding

This research is funded by the German Research Foundation (DFG) as a part of the research grant programme „Information infrastructure for research data“ (grant number HO 1937/2–1).

* These authors contributed equally to this work

1. Introduction

Epidemiological research in the context of cohort studies and registries requires extensive personal, medical and socio-economic data often from multi-site acquisition. Consequently, research becomes more and more complex regarding information acquisition and exchange between partici-

pating institutions as well as data quality assurance.

According to Neugebauer et al. [1], a reliable rating of research data quality is only possible within the context of a specific research question. As an example, Schrappe et al. [2] highlight the importance of monitoring the completeness of research data in the context of registries. Therefore, relevant

indicators must be identified to be able to capture reliable information about data quality with respect to the research question at hand [2]. Stausberg et al. [3], have identified a considerable degree of heterogeneity: in their review they compiled 34 different concepts of data quality indicators.

According to the guideline for data quality [4], provided by the Technology, Methods and Infrastructure for Networked Medical Research e.V. (TMF), medical research needs to consider the (a) quality of structures, the (b) quality of processes as well as the (c) quality of results. Further to (a) selected variables, quality assessment needs to consider (b) specific data sets and (c) complex research data pools.

Following the recommendations of this TMF guideline, the monitoring of data quality should start during the data capture process, “to detect anomalies in research data efficiently” (see [4], p. 119ff). Hereby, the highlighted data quality indicators are of significant importance.

Focusing on single data variables, data quality monitoring might include the assessment of missing values (TMF-1012 – TMF-1016), the examination of disallowed values (TMF-1021 – TMF-1026), the distribution of values (TMF-1006), the validation of representability (TMF-1048), the observation of completeness (TMF-1046) as well as the investigation of extreme values (TMF-1018) [4].

According to recommendations of Müller et al. [5], at least a periodical surveillance of distribution characteristics should be applied. Especially the accuracy as well as the type and frequency of potential recurring data errors must be an essential part of the investigation. A frequent quality and integrity problem of research data is the proportion, and distribution of missing entries, which can substantially complicate the evaluation of research data and the interpretation of statistical results. Typically, incidental and systematic missings are distinguished, whereas only systematic missings need to be further categorized regarding quality aspects.

However, high data quality is connected to costs and knowledge. Especially in smaller studies, budget and personnel constraints are limiting parameters. Basic statistics and descriptive plots enable researchers to identify quality problems through visualization of main characteristics as well as relationships and, therefore, facilitate monitoring and reporting quality of research data [6].

To be able to assess data quality, a scientist needs to have an understanding of

(a) the data and its meaning, (b) basic statistical methods and approaches, (c) methods of visualization, for example graphs, and (d) at least one statistical software product. The more knowledge one has the more detailed and conclusive the data can be interpreted. However, the training period is one of the typical challenges of assuring data quality and generating statistics. To facilitate a reliable data quality assurance, scientists have to be able to apply quick checks on current datasets, e.g. regarding data distributions or the count of characteristics.

Canuel et al. [7] analyzed several existing data quality solutions provided by the scientific community. Five out of seven were open source assessment solutions based on R [7, 8], which suggests a widespread application of R in the scientific community. The authors point out that the developed in-house visualization tools were designed to fit project-individual needs and provide several highly detailed functionalities. Reusability of many solutions is therefore limited. The authors conclude that an “increased development of customizable and reusable tools and libraries would be a great help for the field” (refer to [7], p. 287).

Hence, the aims of the mosaicQA-library are a basic assessment of data quality and the generation of multiple different graphs without the need for in-depth experience with statistical methods or statistical software. Within the MOSAIC-project [9] (funded by the German Research Foundation (HO 1937/2–1)), a modular systematic approach to implement a centralized data management, R was used to address this issue because it is a free and open source powerful statistical software [8]. Due to the large user community, R is well-documented. Since the MOSAIC-target group consists primarily of researchers with limited or no IT-expertise and -resources [10], generating reports on data quality is a particular challenge. Therefore, the aim of the developed library is to facilitate the generation of quality reports of collected data reducing the thresholds for non-experts in statistical software. In its most generic layout mosaicQA focuses on basic statistics and the occurrence of missings.

2. Objectives

This paper focusses on the technical development and implementation of a free R-library (package) for basic monitoring and reporting of data quality in epidemiological studies. This library reduces the effort for implementation and maintenance of monitoring and reporting tools in cohort studies and registries. It is readily reusable for individual statistical scenarios. In this case, monitoring is considered as a continuous quality control of research data regarding selected aspects of data quality during the data collection period, whereas reporting means the visualization of data quality aspects based on the complete amount of research data at the time of report provision.

3. Methods

3.1 Developing Functions to Monitor and Report Data Quality without “Knowing the Data”

When the development of the mosaicQA-library was initiated, information about the research data or the respective research context was very limited. Therefore, a general and flexible approach was demanded. To generate quality reports for various scenarios, the mosaicQA-library has to be able to handle data, without further information on specific characteristics. The question on hand is how to evaluate data quality without knowing their content and meaning. To address this, all quality related aspects have to be assessed in their most general form before developing generic mechanisms and applying them to a specific data set. As a result, mosaicQA is designed to produce valid quality reports for various scenarios and data sets without the need to edit the underlying R code. Thus, this library enables epidemiological researchers without previous knowledge of R to gain insights in their data. The first step was to develop a set of quality-related questions, targeting quality aspects on a more general level:

1. How are the values distributed?
2. What are common statistical values of the distributions?

3. How high is the proportion of missing values compared to findings or other values?
4. Are there outliers in the data set?
5. Is my data model applicable to answer the scientific question?

At this stage metric and categorical data have to be distinguished. Measurements or other quantifiable data are considered metric, whereas answer options like “yes”, “no”, “not specified” or “not asked” are examples for categorical data.

As one example, a selected and anonymized sample dataset originating from the GANI_MED-project [11] was used in a next development step to design an R-script for metric and categorical data (N=1000). Metric and categorical data require different tools for visualization. For metric data, basic statistics are, for example, histogram, boxplots and QQ-plots.

An example for the visualization of metric data using a histogram is shown in ► Figure 1. Moreover mean values, standard deviation and other basic statistical items have been added.

A Box-Whisker-Plot presents quantiles and enables the scientist to identify outliers easily within the dataset. Additionally, the QQ-Plot (Quantile-Quantile-Plot) shows the deviation of the observed data in comparison with any given probability distribution, e.g. the normal distribution. The result of such a QQ-Plot is especially useful to verify whether the observed data fits the applied data model. The QQ-Plot shows the deviation of research data in relation to the expected, e.g. normal, distribution. If both graphs are aligned closely that would indicate that this is the case.

Categorical data require absolute and relative distribution frequencies as well as cumulative percent-ages. Since categorical data refers to different categories, for example “female” and “male” for gender, tabular representations (with absolute and relative frequencies) as well as distribution frequency plots can be utilized to reveal the relation of values (within an expected value range) and qualified missings (exceeding the expected value range, cf. ► Figure 2).

For more flexibility and in order to support individual scenarios a third step included the generalization of the developed

R-scripts. This was necessary since the available data for statistical analysis depend on individual characteristics and parameters of a study or registry. Therefore, dependencies were resolved, functions were parameterized and outsourced, and several supporting functions to comfortably individualize labels, value ranges, code lists and other plot parameters were added.

A matter of utmost importance is the handling of missings. Absent information should be categorized in terms of “qualified missings” by application of unique codes

explaining why an expected item is not available. Thus, researchers are enabled to determine in later analyses whether a missing value relates to information not collected, due to data collection errors during the study period, or is simply a result of:

- Missing information (the interviewee did not know the answer)
- Unwillingness to cooperate (the interviewee did not want to answer)
- Missing coverage by the survey or interview tool (question/response options were not applicable)

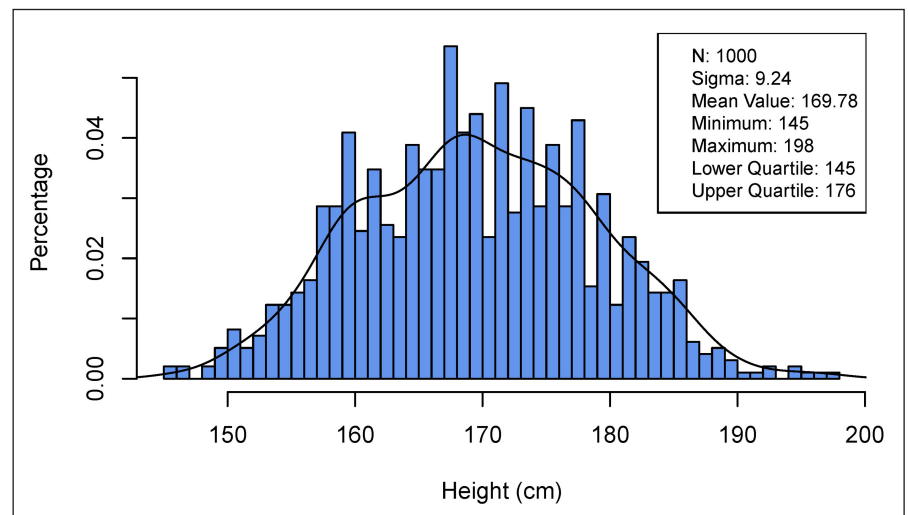


Figure 1 Distribution of a metric variable example (sample size limited to N = 1000; Sigma defined as standard deviation) from the GANI_MED-project [11] – visualized with the mosaicQA-package.

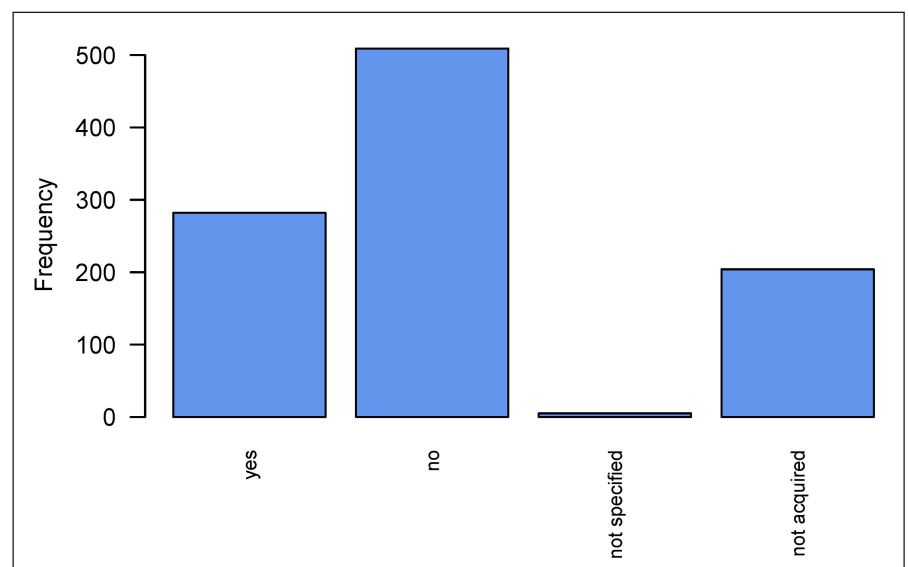


Figure 2 Example plot for a distribution frequency of categorical data displaying valid values (yes, no) and qualified missings (not specified, not acquired).

Table 1 Overview of implemented functionality within the mosaicQA-library (version 1.2.0). Functions called by the user (User Functions) invoke calls of internal functions.

Function	Description	Call Type
mosaic.setGlobalMissingThreshold (threshold)	Set Global Threshold for Missings, e.g. 99000	User
mosaic.setGlobalUnit(unit)	Set Global Unit Label to be used in graphs, e.g. "(cm)"	User
mosaic.setGlobalDescription (description)	Set Global Description for variable data, e.g. "waist circumference"	User
mosaic.loadCsvData(filename)	Load CSV Data from file, e.g. "c:\data.csv"	User
mosaic.setGlobalCodelist (codelist)	Set and parse a global code list for categorical data, e.g. c("1=yes", "2=no", "99996=no information")	User
mosaic.createSimple PdfCategorical(inputfile, outputfolder)	Create simple PDF-file for categorical CSV data using the functions listed above	User
mosaic.createSimplePdfCategorical Dataframe(dataframe, outputfolder)	Create simple PDF-file for categorical data frame with n columns using the functions listed above	User
mosaic.createSimple PdfMetric(inputfile, outputfolder)	Create simple PDF-file for metric CSV data with n columns using the functions listed above	User
mosaic.createSimplePdfMetric Dataframe(dataframe, outputfolder)	Create simple PDF-file for metric data frame with n columns using the functions listed above	User
mosaic.countValue(searchvalue, datacolumn)	Count occurrence of search value in data column	Internal
mosaic.preProcessMetricData(data)	Pre-process metric data to allow missing-ratio table	Internal
mosaic.preProcessCategoricalData(data)	Identify unique values in data column, get absolute, percentage and cumulative statistics	Internal
mosaic.generateMetricTablePlot (data, num of columns, column index, varname)	Generate missing-ratio table for metric data	Internal
mosaic.generateMetricPlots (data snippet, varname)	Generate graphs for metric data	Internal
mosaic.beginPlot(varname, outputfolder)	Begin plotting, generate PDF-file with given variable name	Internal
mosaic.addFootnote()	Add a footnote with timestamp and MOSAIC text	Internal
mosaic.finishPlot()	Finish plotting, close PDF-file	Internal
mosaic.generateCategoricalPlot (dataframe, varname)	Create plots for categorical data	Internal
mosaic.getTimestamp()	Get formatted timestamp, e.g. 2015_09_16_235811	Internal

- Question was not asked or exam was not carried out (e. g. because the participant is missing the needed body part)
- Missing documentation (result of a measurement was not documented)

Code lists for qualified missings can be used in the mosaicQA-library to customize the labels of graphs and tables. Further information on how to handle and code qualified missings can be found in the "reference manual to describe a data dictionary" available at the MOSAIC homepage [12].

3.2 Functions of the mosaicQA-library

Altogether the developed library provides a set of reusable functions to enable re-

searchers to simply evaluate research data based on preformatted quality reports. Therefore, the implemented functions contain the required functionality to generate the previously described statistical visualizations of research data in a report format. ► Table 1 presents an overview of the currently implemented functions within the mosaicQA-library (version 1.2.0).

4. Results

The aims of mosaicQA were to develop a library to provide a basic assessment of data quality and to generate a set of informative graphs. Especially, there should be no demand for the potential researcher to master R or any other statistical soft-

ware. Based on simple samples the researcher should be enabled to reach the desired goal with a readily understandable set of commands.

The developed library enables researchers to generate reports for various kinds of metric and categorical data. Additionally, general reports for multivariate input data and, if needed, detailed results for single-variable data can be easily produced.

CSV-files as well as data frames can be used as input format to create a report. The results are instantly saved in an automatically generated PDF-file. For each study variable within the data input file a separate PDF-file with standard or, if applicable, customized plots and tables is produced. These standard reports enable the user to

monitor and report the data integrity and completeness. However, for more specific reports the knowledge of metadata is necessary, including definition of units, variables, descriptions, code lists and categories of qualified missings.

► Figure 3 provides a code snippet of a sample script to depict the generation of PDF-reports. Only four easy-to-use steps have to be performed from data import to report generation: The researcher has to (a) load the mosaicQA-library, (b) define the data source, (c) the output folder and (d) has to decide whether metric or categorical reports should be created. Necessary additional libraries (e.g. gplots, psych, etc.) are installed automatically. Even users with no experience with statistical software are encouraged to customize thresholds for qualified missings, units and the description of variables by calling three further respective library functions (*mosaic.setGlobalMissingTreshold*, *mosaic.setGlobalUnit*, *mosaic.setGlobalDescription*). The necessary data preparation, the customizing of utilized plots and tables and the required formatting is performed automatically by simply calling the function *mosaic.createSimplePdfmetric* (in case of metric data). As a result a PDF-file is created; containing essential visualizations to monitor and report the specified data (cf. ► Figure 4).

The mosaicQA-package was implemented using R (version 3.3.2) and RStudio. The developed R-package can easily be installed using the official CRAN repository [13]. Additional R libraries will be automatically installed if necessary. Additional sample scripts as well as sample data are available at the MOSAIC- project website [14]. Thus, the provided mosaicQA-library is a ready-to-use implementation, which enables even non-experts to quickly visualize their data and create own functions using working examples as a starting point. To simplify the adaption of the script examples every customizable method is completely documented including an invitation to “specify”, “set” or “adjust” the given standard values (cf. ► Figure 3).

```
# specify the csv import file with metric data, use one column per variable
metric_datafile='c:/mosaic/sample_data/metric_single_var.csv'

#specify output folder
outputFolder='c:/mosaic/output/'

# load mosaicQA package
library('mosaicQA')

#set missing threshold, optional, default is 99900
mosaic.setGlobalMissingTreshold(99900)

#set variable unit, optional
mosaic.setGlobalUnit('cm')

#set variable description, optional
mosaic.setGlobalDescription('Height')

#create PDF-report
mosaic.createSimplePdfmetric(metric_datafile, outputFolder)
```

Figure 3 Code snippet to create a report for a metric dataset sample using the R-package “mosaicQA”.

5. Conclusions

The basic idea of a set of reusable functions to simplify the data quality assurance process is not new. R is an open source solution and supports the implementation of automated processes in terms of reusable statistical analyses. R has been “blamed for being a bit rough” (see [15], p. 187). Individual R-packages, RStudio and cheat sheets [16] are intended to simplify the usage of R especially for non-statistical experts.

The developed library mosaicQA enables users to more easily visualize collected research data and reduces the need for specific implementations for each epidemiological study or registry. Computational or scripting knowledge are not required. For mosaicQA different metric and categorical datasets were used in order to generate appropriate reports as well as to evaluate the generic characteristics of the applied R methods. The library is not designed to deal with complex data types like images or biological samples.

Currently, mosaicQA focusses on the data structure quality and supports the assessment of several quality indicators recommended by the TMF including frequency, distribution and plausibility of research

variables and the occurrence of missing values (TMF-1013 – TMF-1016) and extreme values (TMF-1018) [4]. However, in terms of the categories adopted by the TMF guidelines the provided library only addresses the examination of category (c) selected variables. The present implementation focusses on the visualization of absolute values. Further implementations will also support the formula-based, relative TMF quality indicators.

mosaicQA was released as an official R-package for basic monitoring and reporting of data quality. As a consequence, the active developer community and users in the translational field are now encouraged to contribute (see [15], p. 288). The “library and sample script approach” supports the user to modify existing functionalities and add new features with minimal effort.

In summary, this article described how the mosaicQA-library was developed and how the presented functionality can be used to create essential statistics to review data integrity and completeness. The implemented R-package provides the necessary flexibility, portability and reusability for usage in various epidemiological research projects.

The current state of the released R-package is a starting point for more ad-

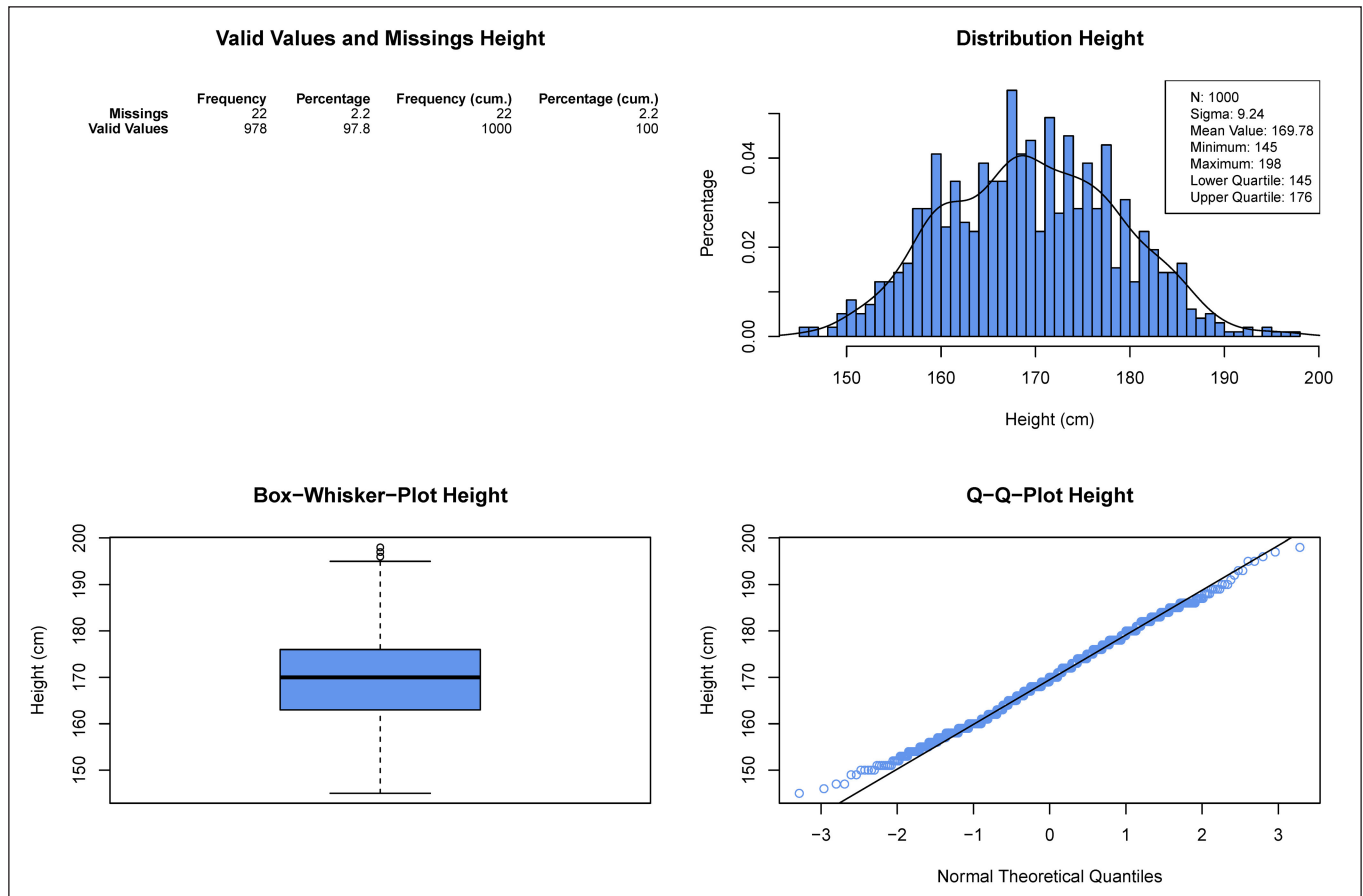


Figure 4 Applying the code snippet for a metric data variable (cf. ► Figure 3) generates a full set of visualizations to facilitate monitoring and reporting of data integrity and completeness.

vanced automated processes. Zalatel et al. (see [15], p. 191) recommend a continuous controlling of selected data quality indicators based on routinely implemented processes as a possible strategy to reduce data quality problems. Thus, the integration in more complex application scenarios is conceivable, such as the MOSAIC Toolbox for Research [17] or transSMART [18]. Consequently, data quality reports could easily be provided automatically and web-based to scientific partners on a daily basis.

The mosaicQA-library will be optimized continuously to improve the applicability in the translational field. At the moment, mosaicQA focusses on CSV-based and data frame-based imports and visualizations for single data variables. To be a basis for a data quality framework in order to support data exploration as well as the generation of hypotheses [7], the support for complex plots (2-n data variables)

is an essential next step. In addition, further development of mosaicQA shall involve the support for respective TMF quality indicators to investigate the number of values below and/or above of specified ranges (TMF-1019) and to identify the number of disallowed values (TMF-1022 – TMF-1025).

The active field of developing data quality monitoring tools can support better processes in research. The interpretation and management of necessary consequences, however, remains the obligation of the responsible scientist [4].

Conflict of Interest

There is no existing conflict of interest regarding this work.

References

1. Neugebauer EAM, Icks A, Schrappe M. Memorandum III: Methods for Health Services Research (Part 2). *Das Gesundheitswesen* 2010; 72(10): 739–748. doi: 10.1055/s-0030-1262858.
2. Schrappe M, Glaeske G, Gottwik M, Kilian R, Papadimitriou K, Scheidt-Nave C, et al. Memorandum II for Health Services Research “Conceptual, methodical and structural requirements for Health Service Research” (Memorandum II zur Versorgungsforschung „Konzeptionelle, methodische und strukturelle Voraussetzungen der Versorgungsforschung“). *Z ärztl Fortbild Qual Gesundheitswes.* 2005; 99(10): 648–51.
3. Stausberg J, Nasseh D, Nonnemacher M. Measuring Data Quality: A Review of the Literature between 2005 and 2013. *Stud Health Technol Inform* 2015; 210: 712–716. doi: 10.3233/978-1-61499-512-8-712.
4. Nonnemacher M, Nasseh D, Stausberg J. Data quality in medical research – Guideline to adaptive management of data quality in cohort studies and registries (Datenqualität in der medizinischen Forschung – Leitlinie zum adaptiven Management von Datenqualität in Kohortenstudien und Regis-

- tern). 2nd ed. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft; 2014.
5. Müller D, Augustin M, Banik N, Baumann W, Bestehorn K, Kieschke J, et al. Memorandum Registry for Health Services Research. *Das Gesundheitswesen* 2010; 72(11): 824–839. doi: 10.1055/s-0030-1263132.
 6. Kowarik A, Meindl B, Templ M. sparkTable: Generating graphical tables for websites and documents with R. *The R Journal* 2015; 7(1): 24–37. doi: <https://journal.r-project.org/archive/2015-1/templ-kowarik-meindl.pdf>.
 7. Canuel V, Rance B, Avillach P, Degoulet P, Burgun A. Translational research platforms integrating clinical and omics data: a review of publicly available solutions. *Brief Bioinform* 2015; 16(2): 280–290. doi: 10.1093/bib/bbu006.
 8. The R Foundation. The R Project for Statistical Computing. [Online]. 2015 [cited 2015 Feb 24]. Available from: <http://www.r-project.org/>.
 9. The MOSAIC Project. MOSAIC Homepage. [Online]. 2016 [cited 2016 Oct 15]. Available from: <https://mosaic-greifswald.de>.
 10. Bialke M, Bahls T, Havemann C, Piegsa J, Weitmann K, Wegner T, et al. MOSAIC. A modular approach to data management in epidemiological studies. *Methods Inf Med* 2015; 54(4): 364–371. doi: 10.3414/ME14-01-0133.
 11. Grabe HJ, Assel H, Bahls T, Dörr M, Endlich K, Endlich N, et al. Cohort profile: Greifswald approach to individualized medicine (GANI_MED). *Journal of Translational Medicine* 2014; 12: 144. doi: 10.1186/1479-5876-12-144.
 12. The MOSAIC Project. Guideline for describing a data dictionary. [Online]. 2017 [cited 2017 Feb 2]. Available from: https://mosaic-greifswald.de/fileadmin/Produkte/Leitfaden_DataDictionary/2017_02_01-GuidelineDataDictionary_v1.1_eng.pdf.
 13. The MOSAIC Project. CRAN-Repository: mosaicQA. [Online]. 2016 [cited 2016 Sep 17]. Available from: <https://cran.r-project.org/package=mosaicQA>.
 14. The MOSAIC Project. R-script library for basic data quality assurance. [Online]. 2015 [cited 2016 Oct 8]. Available from: <https://mosaic-greifswald.de/werkzeuge-und-vorlagen/mosaicqa>.
 15. Zalatel M, Kralj M, editors. *Methodological Guidelines and Recommendations for Efficient and Rational Governance of Patient Registries*. Ljubljana: National Institute of Public Health; 2015.
 16. RStudio.com. R Markdown Cheat Sheet. [Online]. 2014 [cited 2017 Feb 03]. Available from: <https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf>.
 17. Bialke M, Schuld R, Blumentritt A. TMF Dockerbank Workshop – An example for orchestrating docker containers – “MOSAIC Toolbox for Research” (TMF Dockerbank – Container-Orchestrierung am Beispiel der „MOSAIC Toolbox for Research“). [Online]. 2016 [cited 2016 Nov 10]. Available from: http://www.tmf-ev.de/DesktopModules/Bring2mind/DMX/Download.aspx?Method=attachment&Command=Core_Download&EntryId=28919&PortalId=0.
 18. Athey B, Braxenthaler M, Haas M, Guo Y. transMART: An Open Source and Community-Driven Informatics and Data Sharing Platform for Clinical and Translational Research. *AMIA Summits on Translational Science Proceedings*. 2013; p. 6–8.