The Journal of Open Source Software

dataquieR: assessment of data quality in epidemiological research

Adrian Richter[1], **Carsten Oliver Schmidt**[1], **Markus Krüger**[1], **and Stephan Struckmann**[1]

1 Institute for Community Medicine, University Medicine Greifswald

DOI: 10.21105/joss.03093

Software

- Review
- Repository
- Archive

Editor: Charlotte Soneson

Reviewers:

- @borishejblum
- @cmirzayi

Submitted: 01 March 2021
Published: 19 May 2021

Summary

dataquieR is an R package to conduct data quality assessments in data collections designed for research. It makes strong use of metadata that specify the requirements of the study data. Spreadsheet tables can be used to collect this information in a standardized manner. dataquieR starts with checking the formal compliance of study data with expectations defined in the metadata, such as the data type, during *integrity* analyses. Depending on available metadata, further data quality assessments cover the dimensions *completeness*, *consistency*, and *accuracy* as proposed by the framework of Schmidt et al. (2020). Three dataquieR functions investigate the completeness of data within and across observational units. Consistency-related analysis comprises two aspects. First, depending on the data type, the compliance of data elements with either user-defined limits or the adherence to expected value lists is investigated. Second, contradictions between data values of two data elements can be identified by using one of eleven logical comparisons, e.g., if systolic blood pressure is lower than diastolic blood pressure whereas the opposite is expected. Eight dataquieR functions support accuracy-related analyses by aiming at unexpected distributions of single or multiple data elements. Particular focus is placed on the influence of observers, examiners, and devices on the measurement process.

Statement of Need

Various data quality concepts have been proposed to evaluate data's "fitness for use" including different definitions of terms and focus areas (Cai & Zhu, 2015). To comprehend differences underlying these approaches, Keller et al. (2017) stressed the importance to differentiate between (a) designed data collections, (b) administrative data, and (c) opportunity data. Kahn et al. (2016) had already proposed a concept of data quality tailored for electronic health records (EHR) data. Schmidt et al. (2020) have recently introduced a framework addressing specifically the requirements of designed research data collections. Data collected for research purposes differs substantially from EHR data as the researchers are involved in the design, the conduct and the control of the measurement process. Further, enriched metadata, describing the collected data elements beyond datatypes and labels, is commonly available, as well as process information, i.e. the circumstances under which data have been generated (Richter et al., 2019). dataquieR was developed to make specific use of metadata and process information for data quality assessments in designed data collections, and to complement a data quality framework for research data collections.

dataquieR package

The R package dataquieR is currently available on the comprehensive R archive network (CRAN) (R Development Core Team, 2020) and can be installed using:

```
install.packages("dataquieR")
```

For data shaping and analysis dataquieR makes use of the R packages dplyr (Wickham et al., 2021), emmeans (Lenth, 2020), lme4 (Bates et al., 2015), lubridate (Grolemund & Wickham, 2011), MASS (Venables & Ripley, 2002), MultinomialCI (Villacorta, 2019), parallelMap (Bischl et al., 2020), reshape (Wickham, 2007), rlang (Henry & Wickham, 2020), robustbase (Maechler et al., 2021; Todorov & Filzmoser, 2009), and utils (R Development Core Team, 2020). Graphical illustration is realized using the R packages cowplot (Wilke, 2020), ggplot2 (Wickham, 2016), ggpubr (Kassambara, 2020), and R.devices (Bengtsson, 2021). All features of the package have been documented using vignettes created with R Markdown (Allaire et al., 2020; Xie et al., 2018, 2020) and are available on the companion website under software.

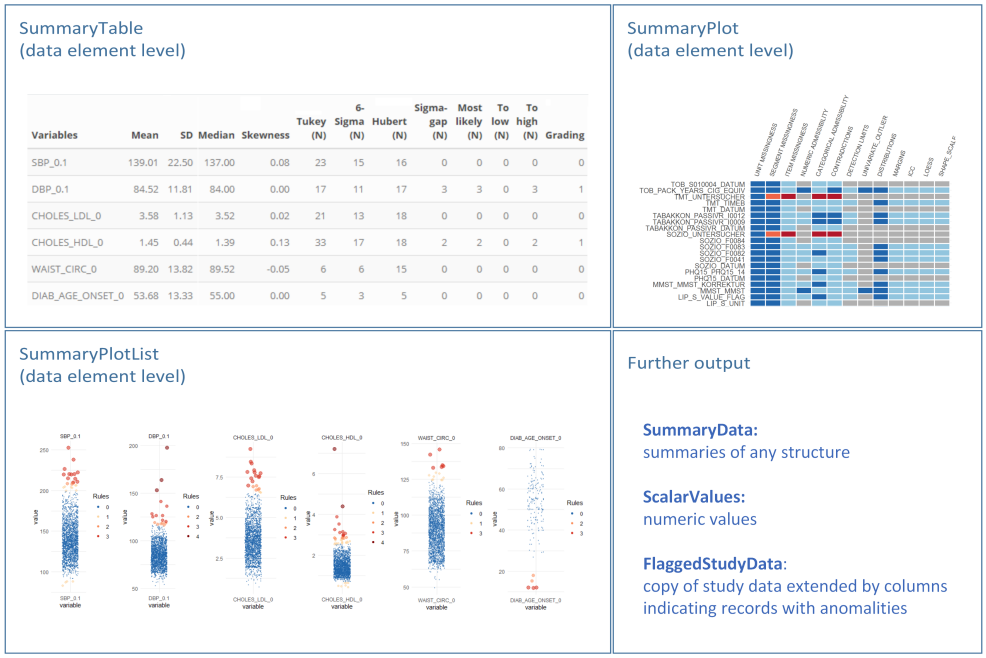
dataquieR can be used in two separate ways. First, the following command:

```
dataquieR::dq_report(study_data = studydata,
                     meta_data  = metadata)
```

will generate a flexdashboard (Iannone et al., 2020) summarizing results of integrity, completeness, and consistency analyses by default. Further arguments of this function can be used to reference the source of additional quality requirements in the metadata. The generated report will then comprise also accuracy-related analyses.

Second, all exported functions of dataquieR may be applied individually to create customized reports. Besides potential modifications of the output, this approach allows for inclusion of transformed or new data elements created during the quality assessment.

Sample output of both approaches are shown in Figure 1. SummaryTable(s), returned as dataframes, and ggplot2 (Wickham, 2016) objects (SummaryPlot, SummaryPlotList) are the most frequently used outputs of dataquieR.



Variables	Mean	SD	Median	Skewness	Tukey (N)	6-Sigma (N)	Hubert (N)	Sigma-gap (N)	Most likely (N)	To low (N)	To high (N)	Grading
SBP_0.1	139.01	22.50	137.00	0.08	23	15	16	0	0	0	0	0
DBP_0.1	84.52	11.81	84.00	0.00	17	11	17	3	3	0	3	1
CHOLES_LDL_0	3.58	1.13	3.52	0.02	21	13	18	0	0	0	0	0
CHOLES_HDL_0	1.45	0.44	1.39	0.13	33	17	18	2	2	0	2	1
WAIST_CIRC_0	89.20	13.82	89.52	-0.05	6	6	15	0	0	0	0	0
DIAB_AGE_ONSET_0	53.68	13.33	55.00	0.00	5	3	5	0	0	0	0	0

Figure 1: Figure 1: Output types of dataquieR.

dataquieR adds to versatile R packages assessing data quality such as validate (Loo & Jonge, 2019), smartEDA (Putatunda et al., 2019), DataExplorer (Cui, 2019), and dataMaid (Petersen & Ekstrøm, 2019) in enabling R users to create extensive data quality reports. The full functionality of dataquieR rests on the existence of well-defined metadata. Therein, one row of the metadata corresponds to one data element of the study data (Richter et al., 2019); currently up to 20 attributes can be used by dataquieR. Such attributes comprise, e.g., the data type, missing codes, different types of limits in interval notation (e.g. "[0; Inf)" for float-type data), value codes (e.g. "1=female | 2=male" for nominal data), distributional assumptions, and the keys to process variables describing the measurement process. While such information can be set up without programming knowledge, the efforts to create such metadata for large numbers of data elements are considerable. Yet, appropriate metadata increase research data FAIRness (Wilkinson et al., 2018) and transparency of research.

For further details regarding the concept and metadata requirements please visit the companion website.

Acknowledgements

This work was supported by the German Research Foundation (DFG: SCHM 2744/3-1), and by the Innovation Programme under grant agreement No. 825903 (euCanSHare project). In addition, we are grateful for the feedback on the functionality and documentation of dataquieR to Dr. Cornelia Enzenbach (University of Leipzig, Institute for Medical Informatics, Statistics and Epidemiology (IMISE)) and Dr. Matthias Ernst.

References

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2020). *rmarkdown: Dynamic Documents for R*. https://github.com/rstudio/rmarkdown

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bengtsson, H. (2021). *R.devices: Unified Handling of Graphics Devices*. https://CRAN.R-project.org/package=R.devices

Bischl, B., Lang, M., & Schratz, P. (2020). *parallelMap: Unified Interface to Parallelization Back-Ends*. https://CRAN.R-project.org/package=parallelMap

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, *14*. https://doi.org/10.5334/dsj-2015-002

Cui, B. (2019). *DataExplorer: Automate Data Exploration and Treatment*. https://CRAN.R-project.org/package=DataExplorer

Grolemund, G., & Wickham, H. (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25. https://doi.org/10.18637/jss.v040.i03

Henry, L., & Wickham, H. (2020). *rlang: Functions for Base Types and Core R and 'Tidyverse' Features*. https://CRAN.R-project.org/package=rlang

Iannone, R., Allaire, J., & Borges, B. (2020). *flexdashboard: R Markdown Format for Flexible Dashboards*. https://CRAN.R-project.org/package=flexdashboard

Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., Estiri, H., Goerg, C., Holve, E., & Johnson, S. G. (2016). A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *eGEMs*, *4*(1). https://doi.org/10.13063/2327-9214.1244

Kassambara, A. (2020). *gpubr: 'ggplot2' Based Publication Ready Plots*. https://CRAN.R-project.org/package=ggpubr

Keller, S., Korkmaz, G., Orr, M., Schroeder, A., & Shipp, S. (2017). The evolution of data quality: Understanding the transdisciplinary origins of data quality concepts and approaches. *Annual Review of Statistics and Its Application*, *4*(1). https://doi.org/10.1146/annurev-statistics-060116-054114

Lenth, R. V. (2020). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://CRAN.R-project.org/package=emmeans

Loo, M. P. van der, & Jonge, E. de. (2019). Data validation infrastructure for R. *arXiv Preprint arXiv:1912.09759*. https://arxiv.org/abs/1912.09759

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. L. T., & Anna di Palma, M. (2021). *robustbase: Basic Robust Statistics*. http://robustbase.r-forge.r-project.org/

Petersen, A. H., & Ekstrøm, C. T. (2019). dataMaid: Your Assistant for Documenting Supervised Data Quality Screening in R. *Journal of Statistical Software*, *90*(1), 1–38. https://doi.org/10.18637/jss.v090.i06

Putatunda, S., Rama, K., Ubrangala, D., & Kondapalli, R. (2019). SmartEDA: An R Package for Automated Exploratory Data Analysis. *Journal of Open Source Software*, *4*(41), 1509. https://doi.org/10.21105/joss.01509

R Development Core Team. (2020). *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/

Richter, A., Schössow, J., Werner, A., Schauer, B., Radke, D., Henke, J., Struckmann, S., & Schmidt, C. (2019). Data quality monitoring in clinical and observational epidemiologic studies: the role of metadata and process information. *GMS Medizinische Informatik, Biometrie Und Epidemiologie*, *15*(1). https://doi.org/doi:%2010.3205/mibe000202

Schmidt, C., Struckmann, S., Enzenbach, C., Reineke, A., Stausberg, J., Damerow, S., Huebner, M., Schmitt, B., Sauerbrei, W., & Richter, A. (2020). *Facilitating Harmonized Data Quality Assessments. A Data Quality Framework for Observational Health Research Data Collections With Software Implementations in R*. Research Square. https://doi.org/10.21203/rs.3.rs-119457/v1

Todorov, V., & Filzmoser, P. (2009). An Object-Oriented Framework for Robust Multivariate Analysis. *Journal of Statistical Software*, *32*(3), 1–47. https://doi.org/10.18637/jss.v032.i03

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth Edition). Springer-Verlag New York. ISBN: 0-387-95457-0

Villacorta, P. J. (2019). *MultinomialCI: Simultaneous Confidence Intervals for Multinomial Proportions According to the Method by Sison and Glaz*. https://CRAN.R-project.org/package=MultinomialCI

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12). https://doi.org/10.18637/jss.v021.i12

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr

Wilke, C. O. (2020). *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. https://CRAN.R-project.org/package=cowplot

Wilkinson, M. D., Sansone, S.-A., Schultes, E., Doorn, P., Bonino da Silva Santos, L. O., & Dumontier, M. (2018). A design framework and exemplar metrics for FAIRness. *Scientific Data*, *5*(1), 180118. https://doi.org/10.1038/sdata.2018.118

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R Markdown: The Definitive Guide*. Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown

Xie, Y., Dervieux, C., & Riederer, E. (2020). *R Markdown Cookbook*. Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook