The Journal of Open Source Software

dataquieR 2: An updated R package for FAIR data quality assessments in observational studies and electronic health record data

Stephan Struckmann [1], **Joany Mariño** [1], **Elisa Kasbohm** [1], **Elena Salogni** [1], **and Carsten Oliver Schmidt** [1]

1 Institute for Community Medicine, University Medicine Greifswald

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 22 February 2024
Published: unpublished

Summary

dataquieR version 2 is a major update to the dataquieR package (Richter et al., 2021), which enables extensive, highly standardized, and accessible data quality assessments related to data integrity (e.g. data type errors, duplicates), completeness (e.g. missing values), consistency (e.g. range violations, contradictions), and accuracy (e.g. outliers, time trends, examiner effects) on tabular form data. This update extends the coverage of data quality indicators from the underlying framework (Schmidt et al., 2021) based on a substantially improved information model, comprises performance enhancements, interactive output, and versatile options to grade data quality issues. The broader framework coverage is achieved by integrating dataquieR 2 with a differentiated spreadsheet-type metadata schema. This schema includes descriptions, expectations and requirements about the following data objects in a machine-readable way: (1) single variables and data fields, (2) combinations of variables, (3) segments of the data, (4) data tables, (5) representation and classification of missing data, and (6) identifier codes. The new metadata schema makes additional assumptions underlying data quality assessments explicit, thereby improving reproducible data quality reporting in compliance with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles (Wilkinson et al., 2016). While dataquieR has primarily been designed for and is currently used in observational health studies, such as population-based cohort studies (Peters et al., 2022; Schmidt et al., 2023), it can also be applied to all sorts of tabular data from other sources, for instance, electronic health record (EHR) data, or registries.

Statement of Need

The need to thoroughly assess data before using them for any substantive scientific purpose is undisputed, and much work has been done in this regard. On the one hand, several data quality frameworks have been developed in the health sciences to guide related assessments. Examples are frameworks with a focus on medical registries (Lee et al., 2017; Nonnemacher et al., 2014), EHR data (Kahn et al., 2016; Liaw et al., 2021; Weiskopf & Weng, 2013), or observational health research data collections (Nonnemacher et al., 2014; Schmidt et al., 2021). Other conceptual work focuses on checking data properties as part of an initial data analysis (Huebner et al., 2018). On the other hand, many tools have been made available to assess data quality in different programming languages (Ehrlinger & Woss, 2022; Mariño et al., 2022). Still, substantial challenges remain; a recent review of 27 R packages to assess data quality revealed important issues (Mariño et al., 2022). The most important drawbacks in most existing R packages for quality evaluation are that these tools are not based on formal data quality frameworks, and the rules underlying the assessments need to be provided in a program-specific syntax rather than being available in an interoperable, reusable format. Exceptions are dataquieR (Richter et al., 2021), and DQAstats (Kapsner et al., 2021). The latter is used for inpatient EHR data, while dataquieR targets designed observational research studies. Despite their different underlying concepts, both define quality-related metadata separately from the programming code. Stand-alone metadata files comprising data descriptions, expectations and requirements are, however, a major precondition to making reproducible data quality assessments (Mariño et al., 2022). Regarding dataquieR, three main shortcomings existed: first, the functions and previous metadata concept only allowed calculating 18 out of 34 data quality indicators from the underlying data quality framework (Schmidt et al., 2021); second, performance issues in handling larger numbers of variables in a single report limited its use; third, even complex output was static, thus limiting the possibility of users to interactively focus on details of concern.

New functionalities

Metadata is now organized in six tables (formerly two tables) that specify overall expectations at the levels of: (1) single variables and data fields (e.g. range violations of individual data values, expected range of a mean), (2) combinations of variables (e.g. contradictions between values of different variables), (3) segments of the data (e.g. different examinations), (4) data tables (e.g. data from the entire study), (5) representation and classification of missing data (e.g. linking a study specific use of missing value codes with a generic approach (AAPOR, 2023)), and (6) valid identifier codes (e.g. pseudo ids). Each metadata table can be handled as a spreadsheet in a workbook, allowing metadata input directly in the spreadsheet or by specifying the source file for a specific item (e.g. another spreadsheet or a URL). For an easier and more accessible way of annotating complex contradiction rules, dataquieR 2 supports a syntax language inspired by REDCap's syntax (Harris et al., 2009). Examples for contradictions in the new syntax are: (1) "[sbp1] < [dbp1]" to indicate that systolic blood pressure cannot be lower than diastolic blood pressure; or (2) "[DIABETES_KNOWN_0] = 'yes' and [DIAB_AGE_ONSET_0] = ''" to denote that for study participants with a known diabetes diagnosis, the age of onset of diabetes should also be specified (i.e. cannot be empty). Whereas the earlier approach could only handle the relation of two variables, now there is no formal limit to rule complexity.

The metadata extensions enable new indicator functions to discover data quality issues. They foremost target the Integrity and Accuracy dimensions of the guiding framework (Schmidt et al., 2021), comprising new options to detect unexpected data elements or data records, duplicates, as well as the detection of unexpected proportion and unexpected location parameters (e.g. mean outside a defined range). Thus, dataquieR 2 allows for the evaluation of 24 data quality indicators out of the 34 in the concept, where the extended metadata scheme already lays the basis for the coverage of the remaining indicators. Another development is classifying data quality issues into up to five categories according to their severity. This grading enables users to obtain a comprehensive overview and identify potentially problematic variables more quickly.

Numerous output improvements have been implemented. Examples include Figure 1: (1) a pie chart summarizing the number of variables in each data quality category, (2) a summary matrix displaying all potential data quality issues or problems related to missing/deficient metadata, (3) easier navigation thanks to the inclusion of breadcrumbs that show the current location in the report, (4) more detailed output for specific data quality checks combining tables and interactive plots, (5) the use of hover text in plots to better interpret the results, (6) the possibility to download plots and tables directly from the report as well as to obtain the source code to reproduce plots individually, (7) the inclusion of descriptive statistics and metadata information in the report.

dataquieR 2 reports are produced using a two-stage approach: computation and rendering. First, the computation stage creates a list with all possible and reasonable function calls according to the metadata. These independent functions are executed in parallel. We eliminated idle times and sequential parts in the report computation through different parallelization strategies, most prominently, a job-queue-based approach (Bengtsson, 2021, 2023; Csárdi & Chang, 2022).

In the rendering stage, reports are no longer produced using R Markdown (Allaire et al., 2023). dataquieR 2 directly renders HTML files using htmltools (Cheng et al., 2023), which speeds up the reporting and improves the use of HTML-specific capabilities (e.g. interactive figures using plotly (Sievert, 2020)). Importantly, dataquieR 2 does not write single large files containing all required resources but creates a complete mini-website for the data quality report, which can be displayed by web browsers using fewer resources of the client computer (i.e. in terms of memory and CPU usage).

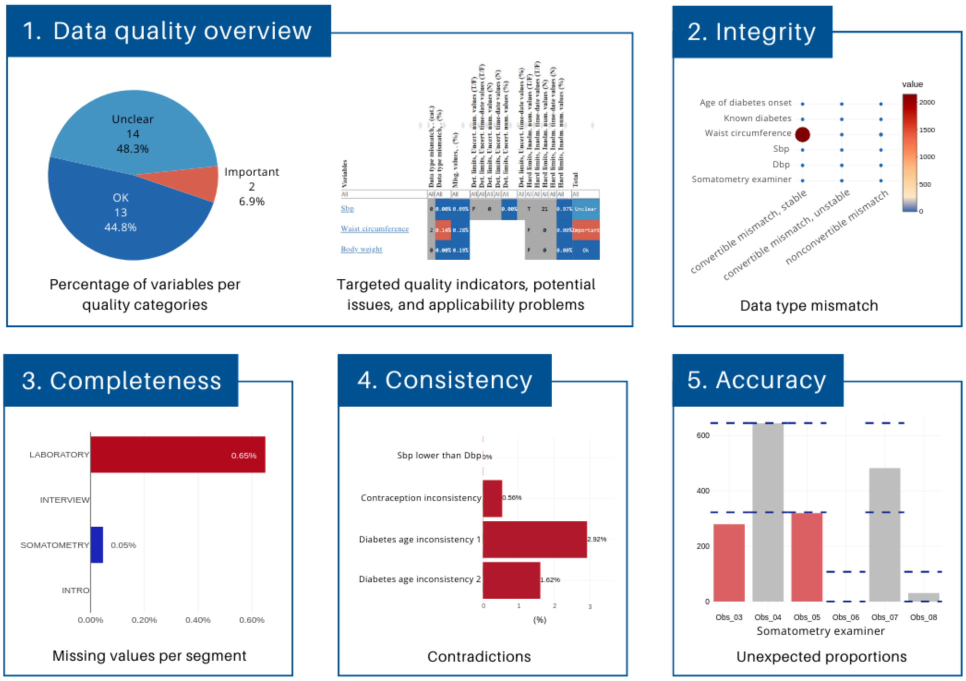

Figure 1: Outline of dataquieR 2's report, showing the main sections with selected outputs for each case.

Installation and examples

dataquieR 2 is available on the Comprehensive R Archive Network (CRAN) and can be installed using install.packages("dataquieR").

Example 1: Computing a data quality report

```r
library(dataquieR)

report_ship_data <- dq_report2(
            study_data = "ship",
            meta_data_v2 = "ship_meta_v2",
            dimensions = NULL,
            label_col = LONG_LABEL,
            title = "SHIP data quality report"
            )

print(report_ship_data, dir = "~/data_quality_reports/report_ship_data")
```

The resulting report is available at https://dataquality.qihs.uni-greifswald.de/report_2024q1/

Example 2: Calculating a data quality indicator (uncertain and inadmissible numerical values, Figure 2)

```
ship_data <- prep_get_data_frame("ship")

prep_load_workbook_like_file("ship_meta_v2")

sbp1_limits <- con_limit_deviations(
                 resp_vars = "sbp1",
                 study_data = ship_data
                 )

sbp1_limits$SummaryPlotList

sbp1_limits$ReportSummaryTable

sbp1_limits$SummaryData
```

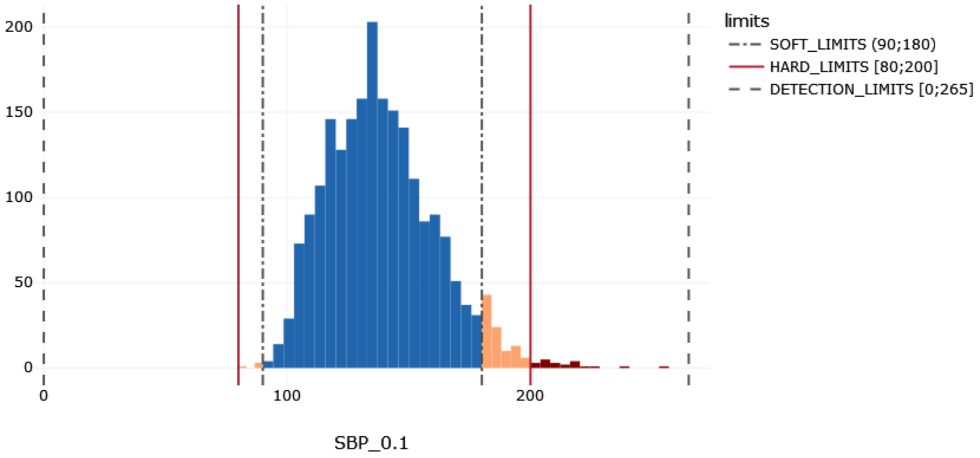

Figure 2: Improved output of the con_limit_deviations indicator function showing the three possible limits for a variable in a single plot (from Example 2).

Acknowledgements

This work was supported by the German Research Foundation (DFG: SCHM 2744/3-4, and NFDI 13/1), and the German National Cohort (NAKO) as funded by the Federal Ministry of Education and Research (BMBF: 01ER1301A and 01ER1801A).

References

AAPOR. (2023). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (10th ed.). The American Association for Public Opinion Research.

Allaire, J. J., Xie, Y., Dervieux, C., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2023). *Rmarkdown: Dynamic documents for r. R package version 2.25*. https://github.com/rstudio/rmarkdown

Bengtsson, H. (2021). A unifying framework for parallel and distributed processing in r using futures. *The R Journal*, *13*(2), 273–291. https://doi.org/10.32614/RJ-2021-048

Bengtsson, H. (2023). *Parallelly: Enhancing the 'parallel' package. R package version 1.36.0*. https://CRAN.R-project.org/package=parallelly

Cheng, J., Sievert, C., Schloerke, B., Chang, W., Xie, Y., & Allen, J. (2023). *Htmltools: Tools for HTML. R package version 0.5.7*. https://CRAN.R-project.org/package=htmltools

Csárdi, G., & Chang, W. (2022). *Callr: Call r from r. R package version 3.7.3*. https://CRAN.R-project.org/package=callr

Ehrlinger, L., & Woss, W. (2022). A survey of data quality measurement and monitoring tools. *Front Big Data*, *5*(5), 850611. https://doi.org/10.3389/fdata.2022.850611

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Huebner, M., Cessie, S. le, Schmidt, C. O., & Vach, W. (2018). A contemporary conceptual framework for initial data analysis. *Observational Studies*, *4*(1), 171–192. https://doi.org/10.1353/obs.2018.0014

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., Liaw, S. T., Hamilton-Lopez, M., Meeker, D., Ong, T. C., Ryan, P., Shang, N., Weiskopf, N. G., Weng, C., Zozus, M. N., & Schilling, L. (2016). A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Wash DC)*, *4*(1), 1244. https://doi.org/10.13063/2327-9214.1244

Kapsner, L. A., Mang, J. M., Mate, S., Seuchter, S. A., Vengadeswaran, A., Bathelt, F., Deppenwiese, N., Kadioglu, D., Kraska, D., & Prokosch, H. U. (2021). Linking a consortium-wide data quality assessment tool with the MIRACUM metadata repository. *Appl Clin Inform*, *12*(4), 826–835. https://doi.org/10.1055/s-0041-1733847

Lee, K., Weiskopf, N., & Pathak, J. (2017). A framework for data quality assessment in clinical research datasets. *AMIA Annu Symp Proc*, *2017*, 1080–1089. https://www.ncbi.nlm.nih.gov/pubmed/29854176

Liaw, S. T., Guo, J. G. N., Ansari, S., Jonnagaddala, J., Godinho, M. A., Borelli, A. J., Lusignan, S. de, Capurro, D., Liyanage, H., Bhattal, N., Bennett, V., Chan, J., & Kahn, M. G. (2021). Quality assessment of real-world data repositories across the data life cycle: A literature review. *J Am Med Inform Assoc*, *28*(7), 1591–1599. https://doi.org/10.1093/jamia/ocaa340

Mariño, J., Kasbohm, E., Struckmann, S., Kapsner, L. A., & Schmidt, C. O. (2022). R packages for data quality assessments and data monitoring: A software scoping review with recommendations for future developments. *Applied Sciences*, *12*(9), 4238. https://doi.org/10.3390/app12094238

Nonnemacher, M., Nasseh, D., & Stausberg, J. (2014). *Datenqualität in der medizinischen forschung: Leitlinie zum adaptiven management von datenqualität in kohortenstudien und registern*. TMF e.V. https://doi.org/10.32745/9783954663743

Peters, A., German National Cohort, C., Peters, A., Greiser, K. H., Gottlicher, S., Ahrens, W., Albrecht, M., Bamberg, F., Barnighausen, T., Becher, H., Berger, K., Beule, A., Boeing, H., Bohn, B., Bohnert, K., Braun, B., Brenner, H., Bulow, R., Castell, S., … others. (2022). Framework and baseline examination of the german national cohort (NAKO). *Eur J Epidemiol*, *37*(10), 1107–1124. https://doi.org/10.1007/s10654-022-00890-5

Richter, A., Schmidt, C. O., Krüger, M., & Struckmann, S. (2021). dataquieR: Assessment of data quality in epidemiological research. *Journal of Open Source Software*, *6*(61), 3039. https://doi.org/10.21105/joss.03093

Schmidt, C. O., Struckmann, S., Enzenbach, C., Reineke, A., Stausberg, J., Damerow, S., Huebner, M., Schmidt, B., Sauerbrei, W., & Richter, A. (2021). Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in r. *BMC Med Res Methodol*, *21*(1), 63. https://doi.org/10.1186/s12874-021-01252-7

Schmidt, C. O., Struckmann, S., Scholz, M., Schossow, J., Radke, D., Richter, A., Reineke, A., Kasbohm, E., Coronado, J. M., Schauer, B., Henselin, K., Westphal, S., Balke, D., Leddig, T., Volzke, H., & Henke, J. (2023). Conducting an epidemiologic study and making it FAIR: Reusable tools and procedures from a population-based cohort study. *Stud Health Technol Inform*, *302*, 871–875. https://doi.org/10.3233/SHTI230292

Sievert, C. (2020). *Interactive web-based data visualization with r, plotly, and shiny*. Chapman; Hall/CRC.

Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *J Am Med Inform Assoc*, *20*(1), 144–151. https://doi.org/10.1136/amiajnl-2011-000681

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., Silva Santos, L. B. da, Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18