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** [1,2], **Elisa Kasbohm** [1], **Elena Salogni** [1], **and Carsten Oliver Schmidt** [1]

1 Institute for Community Medicine, University Medicine Greifswald, Germany **2** Dept. of Internal Medicine B – Cardiology, University Medicine Greifswald, Germany

DOI: 10.21105/joss.06581

Software

- Review
- Repository
- Archive

Editor: Julia Romanowska

Reviewers:

- @EstherPlomp
- @anandhi-iyappan

Submitted: 22 February 2024
Published: 28 June 2024

Summary

dataquieR version 2 is a major update to the dataquieR package (Richter et al., 2021). It enables extensive, highly standardized, and accessible data quality assessments related to data integrity (such as data type errors or duplicates), completeness (for example, missing values), consistency (for instance, range violations or contradictions), and accuracy (such as time trends or 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. It furthermore 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 metadata schema. This spreadsheet-type 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 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 data quality 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 frameworks have been developed in the context of the European Health Data Space (TEHDAS, 2022), the European Open Science Cloud (Lacagnina et al., 2023), in the context of regulatory decision making (Data Analytics and Methods Task Force, 2023) or with focus 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 findings (Mariño et

al., 2022). 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 mainly used for inpatient EHR data, while dataquieR targets designed observational research studies. Both define quality-related metadata separately from the programming code. Such stand-alone metadata files comprise data descriptions, expectations and requirements about the data. They are a major precondition to conduct 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 reference 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 (for instance, range violations of individual data values, expected range of a mean), (2) combinations of variables (such as contradictions between values of different variables), (3) segments of the data (for example, different examinations), (4) data tables (for instance, the expected sample size), (5) representation and classification of missing data (for example, linking a study specific use of missing value codes with a generic approach (AAPOR, 2023)), and (6) identifier codes (for example, 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 (for example, another spreadsheet or a URL). To ease the annotation of complex contradiction rules, dataquieR 2 now supports a syntax language inspired by REDCap (Harris et al., 2009). Examples for contradictions in the new syntax are: (1) "[sbp1] $<$ [dbp1]" to express 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 and cannot be empty. 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 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 (for example, mean outside a defined range). Thus, dataquieR 2 allows for the evaluation of 24 data quality indicators out of the 34 in the concept. However, the extended metadata scheme already lays the basis for the coverage of the remaining indicators. Another development is grading data quality issues into up to five categories according to their severity. This grading enables users to identify potentially problematic variables more quickly. The grading can be adapted to study specific needs.

In addition, numerous output improvements have been implemented. Examples include: (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, 2024a; Csárdi & Chang, 2024a).

In the rendering stage, reports are no longer produced using R Markdown (Allaire et al., 2024). dataquieR 2 directly renders HTML files using htmltools (Cheng et al., 2024). This speeds up the reporting and improves the use of HTML-specific capabilities (for example, 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 then be displayed by web browsers using fewer resources of the client computer (that is, in terms of memory and CPU usage).

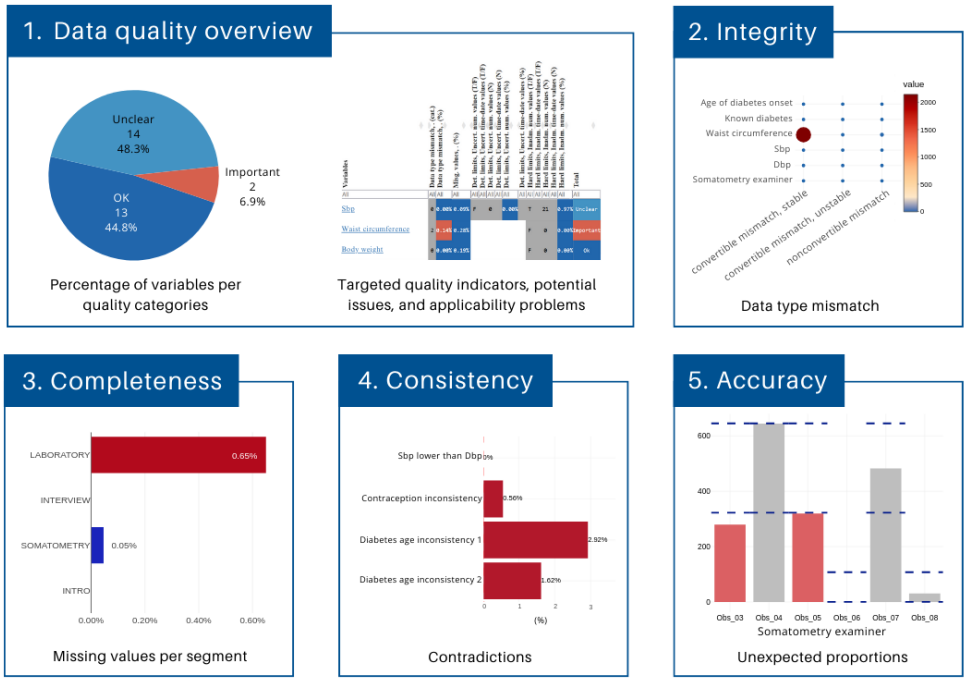

Figure 1: Outline of dataquieR 2's report, showing the main sections with selected outputs for each case.

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

```
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/, and see also Figure 1.

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
```

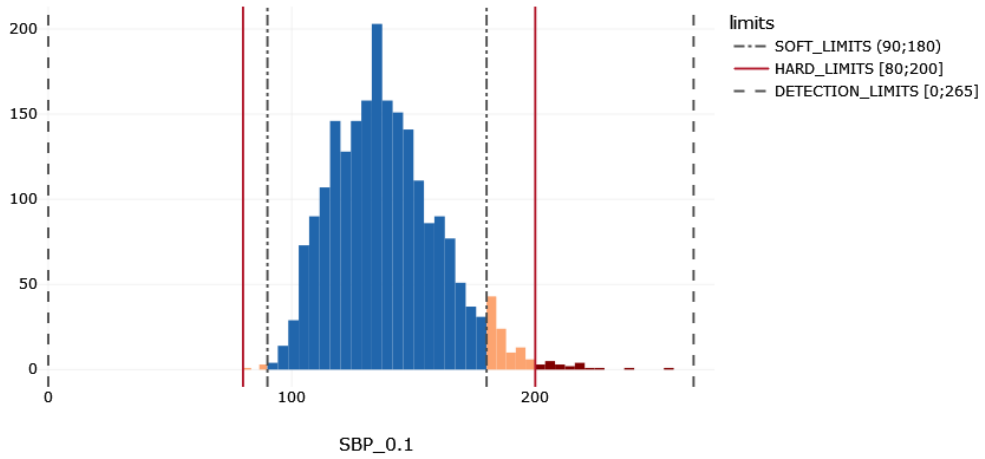

Figure 2: Improved output of the con_limit_deviations indicator function showing the three possible limits for a variable in a single plot (from Example 2).

Dependencies

For data loading, shaping and analysis, dataquieR makes use of the R packages dplyr (Wickham, François, et al., 2023), rlang (Henry & Wickham, 2024), withr (Hester et al., 2024), lifecycle (Henry & Wickham, 2023), emmeans (Lenth, 2024), MASS (Venables & Ripley, 2002), lme4 (Bates et al., 2015), MultinomialCI (Villacorta, 2021), robustbase (Maechler et al., 2024; Todorov & Filzmoser, 2009), ggplot2 (Wickham, 2016), patchwork (Pedersen, 2024), R.devices (Bengtsson, 2024b), scales (Wickham, Pedersen, et al., 2023), lubridate (Grolemund & Wickham, 2011), parallelMap (Bischl et al., 2021), qmrparser (Rosat & Coscollà, 2022), rio (Chan et al., 2023), readr (Wickham et al., 2024), and units (Pebesma et al., 2016).

Optionally, the following packages can be employed for extended features, if installed. These features are: (1) Interactive HTML reports: htmltools (Cheng et al., 2024), htmlwidgets (Vaidyanathan et al., 2023), markdown (Xie et al., 2023), rmarkdown (Allaire et al., 2024; Xie et al., 2018, 2020), and Rdpack (Boshnakov, 2023), additionally, DT (Xie et al., 2024) and plotly (Sievert, 2020) for interactive tables and interactive figures in an HTML report respectively. (2) Parallel computing: processx (Csárdi & Chang, 2024b), R6 (Chang, 2021), and callr (Csárdi & Chang, 2024a) for fast, queue-based parallel computing, parallelly (Bengtsson, 2024a) or rJava (Urbanek, 2024) for correct detection of compute cores in containers, and future (Bengtsson, 2021) as an alternative parallel compute strategy. (3) Improved metadata handling: rvest (Wickham, 2024) for reading metadata from web servers, xml2 (Wickham, Hester, et al., 2023) for full support of the `UNIT` column in item-level metadata, textutils (Schumann, 2023) for supporting `VALUE_LABELS` with special characters, whoami (Csárdi, 2019) to auto-detect a report author's name. (4) User interface: cli (Csárdi, 2023) for nicer console output, rstudioapi (Ushey et al., 2024) for RStudio progress bars, shiny (Chang et al., 2024) for improved integration with Shiny user interfaces. (5) Specific figures: GGally (Schloerke et al., 2024) for descriptive pairs-plots, colorspace (Stauffer et al., 2009; Zeileis et al., 2009, 2020) for nicer colors in some plots, mgcv (Wood, 2003, 2004, 2011, 2017; Wood et al., 2016) for more efficient LOESS plot computation.

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., Xie, Y., Dervieux, C., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2024). *Rmarkdown: Dynamic documents for R*. https://doi.org/10.32614/CRAN.package.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). 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. (2024a). *Parallelly: Enhancing the 'parallel' package*. https://doi.org/10.32614/CRAN.package.parallelly

Bengtsson, H. (2024b). *R.devices: Unified handling of graphics devices*. https://doi.org/10.32614/CRAN.package.R.devices

Bischl, B., Lang, M., & Schratz, P. (2021). *parallelMap: Unified interface to parallelization back-ends*. https://doi.org/10.32614/CRAN.package.parallelMap

Boshnakov, G. N. (2023). *Rdpack: Update and manipulate Rd documentation objects*. https://doi.org/10.5281/zenodo.3925612

Chan, C., Leeper, T. J., Becker, J., & Schoch, D. (2023). *Rio: A swiss-army knife for data file I/O*. https://doi.org/10.32614/CRAN.package.rio

Chang, W. (2021). *R6: Encapsulated classes with reference semantics*. https://doi.org/10.32614/cran.package.r6

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2024). *Shiny: Web application framework for R*. https://doi.org/10.32614/cran.package.shiny

Cheng, J., Sievert, C., Schloerke, B., Chang, W., Xie, Y., & Allen, J. (2024). *Htmltools: Tools for HTML*. https://doi.org/10.32614/CRAN.package.htmltools

Csárdi, G. (2019). *Whoami: Username, full name, email address, 'GitHub' username of the current user*. https://doi.org/10.32614/CRAN.package.whoami

Csárdi, G. (2023). *Cli: Helpers for developing command line interfaces*. https://doi.org/10.32614/CRAN.package.cli

Csárdi, G., & Chang, W. (2024a). *Callr: Call R from R*. https://doi.org/10.32614/cran.package.callr

Csárdi, G., & Chang, W. (2024b). *Processx: Execute and control system processes*. https://doi.org/10.32614/cran.package.processx

Data Analytics and Methods Task Force. (2023). *Data quality framework for EU medicines regulation*. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/data-quality-framework-eu-medicines-regulation_en.pdf

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

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

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

Henry, L., & Wickham, H. (2023). *Lifecycle: Manage the life cycle of your package functions*. https://doi.org/10.32614/CRAN.package.lifecycle

Henry, L., & Wickham, H. (2024). *Rlang: Functions for base types and core R and 'tidyverse' features*. https://doi.org/10.32614/CRAN.package.rlang

Hester, J., Henry, L., Müller, K., Ushey, K., Wickham, H., & Chang, W. (2024). *Withr: Run code 'with' temporarily modified global state*. https://doi.org/10.32614/CRAN.package.withr

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

Lacagnina, C., David, R., Nikiforova, A., Kuusniemi, M.-E., Cappiello, C., Biehlmaier, O., Wright, L., Schubert, C., Bertino, A., Thiemann, H., & Dennis, R. (2023). Towards a data quality framework for EOSC (1.0.0). *Zenodo*. https://doi.org/10.5281/zenodo.7515816

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

Lenth, R. V. (2024). *Emmeans: Estimated marginal means, aka least-squares means*. https://doi.org/10.32614/CRAN.package.emmeans

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

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. (2024). *Robustbase: Basic robust statistics*. https://doi.org/10.32614/CRAN.package.robustbase

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

Pebesma, E., Mailund, T., & Hiebert, J. (2016). Measurement units in R. *R Journal*, *8*(2), 486–494. https://doi.org/10.32614/RJ-2016-061

Pedersen, T. L. (2024). *Patchwork: The composer of plots*. https://doi.org/10.32614/CRAN.package.patchwork

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

Rosat, J. G., & Coscollà, R. M. (2022). *Qmrparser: Parser combinator in R*. https://doi.org/10.32614/CRAN.package.qmrparser

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., & Crowley, J. (2024). *GGally: Extension to 'ggplot2'*. https://doi.org/10.32614/CRAN.package.GGally

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

Schumann, E. (2023). *Textutils: Utilities for handling strings and text*. https://doi.org/10.

32614/cran.package.textutils

Sievert, C. (2020). *Interactive web-based data visualization with R, plotly, and shiny*. Chapman; Hall/CRC. https://doi.org/10.1201/9780429447273

Stauffer, R., Mayr, G. J., Dabernig, M., & Zeileis, A. (2009). Somewhere over the rainbow: How to make effective use of colors in meteorological visualizations. *Bulletin of the American Meteorological Society*, *96*(2), 203–216. https://doi.org/10.1175/BAMS-D-13-00155.1

TEHDAS. (2022). *European health data space data quality framework. Deliverable 6.1*. https://tehdas.eu/tehdas1/results/tehdas-develops-data-quality-recommendations/

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

Urbanek, S. (2024). *rJava: Low-level R to Java interface*. https://doi.org/10.32614/CRAN.package.rJava

Ushey, K., Allaire, J., Wickham, H., & Ritchie, G. (2024). *Rstudioapi: Safely access the RStudio API*. https://doi.org/10.32614/cran.package.rstudioapi

Vaidyanathan, R., Xie, Y., Allaire, J., Cheng, J., Sievert, C., & Russell, K. (2023). *Htmlwidgets: HTML widgets for R*. https://doi.org/10.32614/cran.package.htmlwidgets

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth). Springer. https://doi.org/10.1007/978-0-387-21706-2

Villacorta, P. J. (2021). *MultinomialCI: Simultaneous confidence intervals for multinomial proportions according to the method by Sison and Glaz*. https://doi.org/10.32614/CRAN.package.MultinomialCI

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

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://doi.org/10.1007/978-3-319-24277-4

Wickham, H. (2024). *Rvest: Easily harvest (scrape) web pages*. https://doi.org/10.32614/cran.package.rvest

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation*. https://doi.org/10.32614/CRAN.package.dplyr

Wickham, H., Hester, J., & Bryan, J. (2024). *Readr: Read rectangular text data*. https://doi.org/10.32614/CRAN.package.readr

Wickham, H., Hester, J., & Ooms, J. (2023). *xml2: Parse XML*. https://doi.org/10.32614/cran.package.xml2

Wickham, H., Pedersen, T. L., & Seidel, D. (2023). *Scales: Scale functions for visualization*. https://doi.org/10.32614/CRAN.package.scales

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

Wood, S. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, *65*(1), 95–114. https://doi.org/10.1111/1467-9868.00374

Wood, S. (2004). Stable and efficient multiple smoothing parameter estimation for generalized

additive models. *Journal of the American Statistical Association*, *99*(467), 673–686. https://doi.org/10.1198/016214504000000980

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Wood, S. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/9781315370279

Wood, S., N., Pya, & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models (with discussion). *Journal of the American Statistical Association*, *111*, 1548–1575. https://doi.org/10.1080/01621459.2016.1180986

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R markdown: The definitive guide*. Chapman; Hall/CRC. https://doi.org/10.1201/9781138359444

Xie, Y., Allaire, J., & Horner, J. (2023). *Markdown: Render markdown with 'commonmark'*. https://doi.org/10.32614/CRAN.package.markdown

Xie, Y., Cheng, J., & Tan, X. (2024). *DT: A wrapper of the JavaScript library 'DataTables'*. https://doi.org/10.32614/CRAN.package.DT

Xie, Y., Dervieux, C., & Riederer, E. (2020). *R Markdown cookbook*. Chapman; Hall/CRC. https://doi.org/10.1201/9781003097471

Zeileis, A., Fisher, J. C., Hornik, K., Ihaka, R., McWhite, C. D., Murrell, P., Stauffer, R., & Wilke, C. O. (2020). colorspace: A toolbox for manipulating and assessing colors and palettes. *Journal of Statistical Software*, *96*(1), 1–49. https://doi.org/10.18637/jss.v096.i01

Zeileis, A., Hornik, K., & Murrell, P. (2009). Escaping RGBland: Selecting colors for statistical graphics. *Computational Statistics & Data Analysis*, *53*(9), 3259–3270. https://doi.org/10.1016/j.csda.2008.11.033