

secuTrialR: Seamless interaction with clinical trial databases in R

Patrick R. Wright^{1, 3}, Alan G. Haynes^{2, 4}, and Milica Markovic^{1, 3}

1 University Hospital Basel, Clinical Trial Unit, Basel, Switzerland 2 CTU Bern, University of Bern 3 Data Management Platform of the Swiss Clinical Trial Organisation (SCTO) 4 Statistics and Methodology Platform of the Swiss Clinical Trial Organisation (SCTO)

DOI: [10.21105/joss.02816](https://doi.org/10.21105/joss.02816)

Software

- [Review](#) ↗
- [Repository](#) ↗
- [Archive](#) ↗

Editor: [Charlotte Soneson](#) ↗

Reviewers:

- [@pacoramon](#)
- [@sachsmc](#)

Submitted: 26 October 2020

Published: 20 November 2020

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Summary

Elementary clinical trials have been conducted for hundreds of years ([Meinert & Tonascia, 1986](#)). The most famous early example is the proof that sailors' scurvy can be cured by the consumption of citrus fruit ([Lind, 2014](#)) performed by James Lind in the 18th century. Since those initial days of clinical research, trials have significantly evolved methodically, ethically, and technologically. While it was viable and legitimate to collect clinical trials data in unversioned spreadsheets in the past, this is no longer true and digital clinical data management systems (CDMS) have taken over. CDMS allow constraint-based and version-controlled data entry into a clinical trial database, which ensures traceability, integrity, and quality of study data.

There is a vast market of heterogeneous CDMS solutions, each of which has individual advantages and limitations ([Kuchinke et al., 2010](#)). One limitation can be the interaction with the data after it has been collected. Specifically, a CDMS may be tailored for optimal data capture while, at least to some extent, disregarding ease-of-use of study data after the conclusion of data entry. It is, however, vital that the interaction between data sources and data analysts is fast and seamless in order to avoid loss of valuable time due to technical overhead. This point has been prominently highlighted by the currently ongoing coronavirus pandemic ([Callaway et al., 2020](#)) in which issues have been reported regarding the timely and complete transfer of information for the preparation of up-to-date infection counts ([Kelion, 2020](#); [Merlot & Pauly, 2020](#)). These issues led to confusion and may have ultimately delayed important actions. While this is a stark example, it still serves to show how severe the influence of technical friction between digital systems can be.

To this end we have developed the open source R statistics ([R Core Team, 2020](#)) software package `secuTrialR`, which enables seamless interaction with data collected in the commercially available CDMS `secuTrial` (vendor [interActive Systems Berlin](#)). In addition to parsing and reading the data, it performs data transformation for dates, date times, and categorical data to reduce the data preparation overhead and to allow a swift transition into the analytical phase. Furthermore, `secuTrialR` includes standard functionalities to show descriptive statistics such as study recruitment or completeness of entered data per case report form for `secuTrial` data exports.

Statement of need

Due to the size and complexity of clinical trial and registry databases, technical friction during the initial interaction with data exported from `secuTrial` can be expected. Our own first-hand experience revealed that this overhead can sometimes significantly redirect scarce time and

energy away from analysis and towards data management. The amount of time spent on data management should be as small as possible. The use of `secuTrialR` leads to a pronounced reduction of time necessary for data management, enables swift quantitative analyses through preimplemented functionalities, and most importantly, standardizes the interaction with data exports from `secuTrial`, thus allowing robust and reproducible science.

While some CDMS provide APIs (e.g., REDCap ([Harris et al., 2009, 2019](#))) or Open Database Connectivity (ODBC) connections (e.g., [2mt's WebSpirit](#)) to download data easily, using `secuTrial`'s SOAP API involves querying individual datapoints. This results in an extraordinarily high number of queries even to download a relatively small database, and high demand on servers. As such, approaches such as those for REDCap (e.g., the [REDCapR](#) package, which can interface to REDCap's REST API and download all data in a single query, but does no data preparation) are not suitable for `secuTrial`. Another approach is to parse data exported manually from websites (e.g., the [ox](#) package for importing [OpenClinica](#) exports into R). This approach is used in `secuTrialR`.

Design

All `secuTrial` data exports share a certain common technical structure independent of the specific database at hand. In `secuTrialR` we make use of this information to build an S3 object of class `secuTrialdata`, which is a list, while the data is being read into R. All downstream functions implemented in `secuTrialR` expect a `secuTrialdata` object as input but custom analyses with other components of R statistics are also an option (see Figure 1). While editing the `secuTrialdata` object is technically possible, this is not advisable. Instead, it should be treated as raw data archive from which data can be extracted for analysis. However, if necessary, it is possible to extract subsets of `secuTrialdata` objects with the `subset_secuTrial()` function and return intact `secuTrialdata` objects. The individual elements of the `secuTrialdata` object can be accessed via regular list access operations or the `as.data.frame()` method, which assigns all objects to an environment of choice.

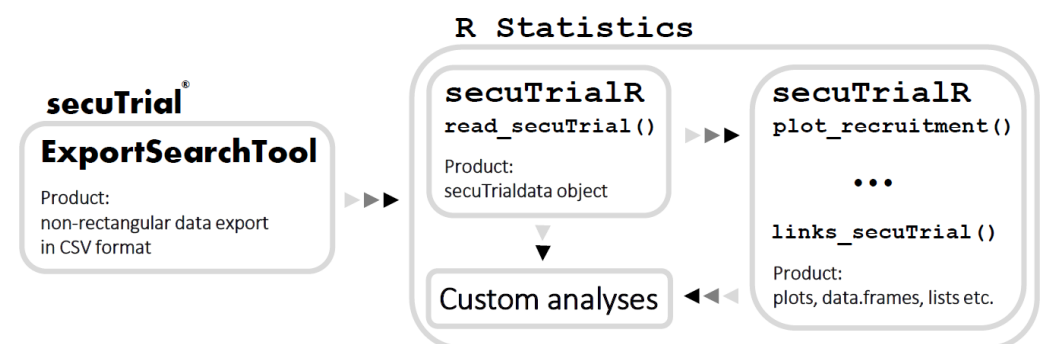


Figure 1: Information flow from `secuTrial` to R statistics and within R. Arrows indicate the direction from gray towards black. “...” indicates further functions working with `secuTrialdata` objects.

Availability

`secuTrialR` is available on [GitHub](#), [CRAN](#), [Anaconda Cloud](#), and should be functional on all major operating systems.

Dependencies

The development of `secuTrialR` made extensive use of the tidyverse (Wickham, Averick, et al., 2019) and greatly benefited from the `devtools` package (Wickham, Hester, et al., 2019) and RStudio (RStudio Team, 2015). Furthermore, `tcltk` and `igraph` (Csardi & Nepusz, 2006) are incorporated.

interActive Systems statement

InterActive Systems (iAS) has given permission for the open source development of this software package but accepts no responsibility for the correctness of any functionalities within.

iAS has read and approved this manuscript.

Acknowledgements

The authors thank Pascal Benkert, Nicole Bruni, Gilles Dutilh, Olivia Ebner, Stefanie von Felten, Thomas Fabbro, Inessa Kraft, Arnaud Künzi, Daniel Lengwiler, Armando Lenz, Pia Neuschwander, Henry Owusu, Hans Rock, Claudia Rokitta, Marie Roumet, Constantin Sluka, Klaus Steigmiller, Suvitha Subramaniam, Miriam Wegmann, Laura Werlen, and Thomas Zumbunn for ideas, testing, and constructive feedback on the `secuTrialR` package. We also thank Michael Sachs and Francisco Estupiñán-Romero for kindly reviewing this manuscript and the R package and making additional recommendations, and Charlotte Soneson for acting as editor. Furthermore, the authors thank the State Secretariat of Education, Research and Innovation and the Swiss National Science Foundation for the funding of this project and the Swiss Clinical Trial Organisation for its ongoing support.

Conflict of interest

The authors are not employees but customers of interActive Systems (iAS). The authors therefore declare no conflict of interest.

References

- Callaway, E., Cyranoski, D., Mallapaty, S., Stoye, E., & Tollefson, J. (2020). The coronavirus pandemic in five powerful charts. *Nature*, *579*, 482–483. <https://doi.org/10.1038/d41586-020-00758-2>
- Csardi, G., & Nepusz, T. (2006). The `igraph` software package for complex network research. *InterJournal, Complex Systems*, 1695. <http://igraph.org>
- Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., Kirby, J., & Duda, S. N. (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics*, *95*, 103208. <https://doi.org/10.1016/j.jbi.2019.103208>
- 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>

- Kelion, L. (2020). *Excel: Why using Microsoft's tool caused COVID-19 results to be lost*. <https://www.bbc.com/news/technology-54423988>.
- Kuchinke, W., Ohmann, C., Yang, Q., Salas, N., Lauritsen, J., Gueyffier, F., Leizorovicz, A., Schade-Brittinger, C., Wittenberg, M., Voko, Z., Gaynor, S., Cooney, M., Doran, P., Maggioni, A., Lorimer, A., Torres, F., Mcpherson, G., Charwill, J., Hellström, M., & Lejeune, S. (2010). Heterogeneity prevails: The state of clinical trial data management in Europe - results of a survey of ECRIN centres. *Trials*, 11, 79. <https://doi.org/10.1186/1745-6215-11-79>
- Lind, J. (2014). *A treatise of the scurvy, in three parts: Containing an inquiry into the nature, causes, and cure, of that disease*. Cambridge University Press. <https://doi.org/10.1017/CBO9781107256644>
- Meinert, C. L., & Tonascia, S. (1986). *Clinical trials: Design, conduct, and analysis*. Oxford University Press. <https://books.google.de/books?id=K92bnQEACAAJ>
- Merlot, J., & Pauly, M. (2020). *Statistikprobleme beim coronavirus - die große meldelücke*. <https://www.spiegel.de/a-13bd06d7-22a1-4b3d-af23-ff43e5e8abd6>.
- R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- RStudio Team. (2015). *RStudio: Integrated development environment for R*. RStudio, Inc. <http://www.rstudio.com/>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Golemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>
- Wickham, H., Hester, J., & Chang, W. (2019). *devtools: Tools to make developing R packages easier*. <https://CRAN.R-project.org/package=devtools>