

REDCapTidieR: Extracting complex REDCap databases into tidy tables

Richard Hanna¹, Ezra Porter¹, Stephany Romero¹, Paul Wildenhain⁶, William Beasley⁷, and Stephan Kadauke^{2,3,4,5}

¹ Division of Oncology, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania ² Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania ³ Department of Pathology and Laboratory Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania ⁴ Division of Transfusion Medicine, Children's Hospital of Philadelphia, Pennsylvania ⁵ Division of Pathology Informatics, Children's Hospital of Philadelphia, Pennsylvania ⁶ Division of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania ⁷ Department of Pediatrics, The University of Oklahoma Health Sciences Center, College of Medicine, Oklahoma City, Oklahoma, USA

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

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

Editor: [↗](#)

Submitted: 05 December 2023

Published: unpublished

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

Summary

Capturing and storing electronic data is integral in the research world. REDCap (Harris et al., 2009, 2019) offers a secure web application that lets users build databases and surveys with a robust front-end interface that can support data of any type, including data requiring compliance with standards for protected information.

Many REDCap users use the R programming language (R Core Team, 2020) to extract and analyze their data. The REDCapR (Beasley, 2023) and redcapAPI (Nutter & Lane, 2023) packages allow R users to extract data directly into their programming environment. While this works well for simple REDCap databases, it becomes cumbersome for complex databases, because the REDCap API outputs a “block matrix”—a single table with varied granularity levels, which conflicts with the “tidy data” framework (Wickham, 2014) that advocates for standardized data organization.

To address this, we introduce REDCapTidieR, an open-source package that streamlines data extraction and restructures it into an intuitive format compatible with the tidy data principles. This facilitates seamless data analysis in R, especially for complex longitudinal studies.

While there are several tools available for REDCap data management, REDCapTidieR introduces a unique solution by transforming the challenging block matrix into a standardized tidy data structure that we term the “supertibble”. This approach not only aligns with good data science practice but also caters to databases of any complexity. By providing a suite of utility functions to work with the supertibble, REDCapTidieR provides a complete framework for extracting REDCap data designed with user-friendliness at its core.

Statement of Need

As of 2023, the REDCap Consortium boasts nearly 3 million users across over 150 countries. REDCap databases range from single-instrument projects to complex builds that use both repeating instruments and repeating events. These data structures are needed to capture multiple items related to a specific visit, such as concomitant medications, or events that cannot be planned ahead of time, such as adverse events.

REDCap databases that contain repeating events and instruments require significant manual

pre-processing, a major pain point for researchers and analysts. This is because the REDCap API returns a single table (Figure 1) that includes data from instruments that record data at different levels of granularity.

While there are a few existing REDCap tools (Table 1), REDCapTidieR occupies a unique space by providing analysts with a framework returns a tidy data structure regardless of the size or complexity of the extracted database. Although some of these tools also offer functions for data processing, such as the `tidyREDCap` (Balise et al., 2023) and `REDCapDM` (Carmezim et al., 2023) packages, only REDCapTidieR restructures the block matrix into an easy to use format.

REDCapTidieR is built with production readiness in mind. In addition to an extensive 98% coverage test suite, REDCapTidieR execution is evaluated against 15 test databases that cover many complex configuration scenarios. It also provides ample documentation through a pkgdown site (Hanna et al., 2023). It is also built on top of REDCapR, which contains its own extensive test suite, and evaluated against an additional 26 test databases. REDCapTidieR meets the rigorous requirements of the [OpenSSF Best Practices Badge](#) (“OpenOpen Source Security Foundation_2023,” 2023), which certifies open-source projects that adhere to criteria for delivering high-quality, robust, and secure software.

Package	Exports from REDCap	Imports into REDCap	Tidy Reformatting	Extensive Test Suite
redcapAPI	x	x		x
REDCapR	x	x		x
tidyRED-Cap	x			
RED-CapDM	x			
REDCapTi-dieR	x		x	x

Table 1: Comparative breakdown of the landscape for REDCap tools in R.

Design

The `REDCapTidieR::read_redcap()` function leverages REDCapR to make API calls to query the data and metadata of a REDCap project and returns the supertibble (Figure 1). The supertibble, named after the `tibble` package (Müller & Wickham, 2023), is an alternative presentation of the data in which multiple tables are linked together in a single object in a fashion consistent with tidy data principles.

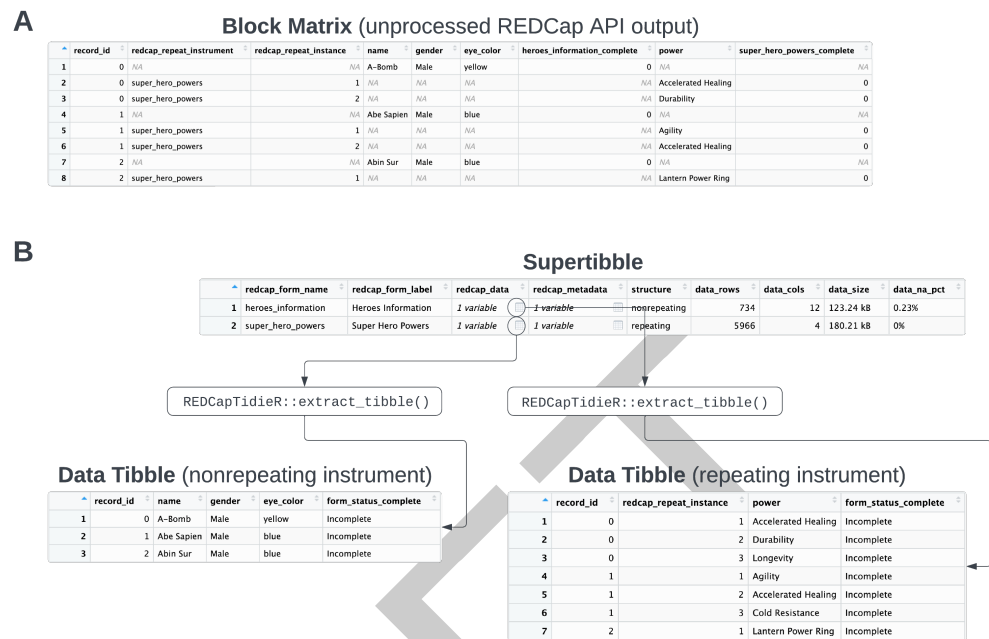


Figure 1: The REDCapTidieR Supertibble

Figure 1: The REDCapTidieR supertibble shown in the Data Viewer of the RStudio IDE. The “Superhero database” (Lingen, 2023) contains two instruments, one nonrepeating and one repeating. A. The REDCap API outputs a “Block Matrix”. Note an abundance of NA values, which do not represent missing values but rather fields that do not apply due to the data structure. B. The read_redcap() function returns a “Supertibble”. Note that each row represents one instrument, identified by the redcap_form_name column. The redcap_data column is a list column that links to tibbles containing the data from a specific instrument. The Data Viewer allows drilling down into individual tibbles by clicking on the table icon, allowing for rapid and intuitive data exploration without any preprocessing. Since each instrument has a consistent granularity, these tibbles can be tidy. Two data tibbles are shown, one from a nonrepeating and one from a repeating instrument. Note the differences in granularity between the instruments.

REDCapTidieR provides utility functions to work with the supertibble, all designed to work with the R pipe operator |>. The extract_tibble() function takes a supertibble object and returns a specific data tibble. The make_labelled() function leverages the labelled package (Larmarange, 2023) to apply variable labels to the supertibble. The add_skimr_metadata() function uses the skimr package (Waring et al., 2023) to add summary statistics. Using the write_redcap_xlsx() function, which leverages the openxlsx2 (Openxlsx2, 2023) package, users can easily export an the supertibble into a collaborator-friendly Excel document, in which each Excel sheet contains the data for an instrument.

REDCapTidieR cannot be used to write data to a REDCap project. We refer the reader to an excellent guide of how to accomplish this using REDCapR (Beasley & Balise, 2023).

Installation

REDCapTidieR is available on [GitHub](#) and [CRAN](#) and works on all major operating systems.

Acknowledgements

We would like to thank Will Beasley, Paul Wildenhain, and Jan Marvin for their feedback and support in development.

This package was developed by the [Cell and Gene Therapy Informatics Team](#) of the [Children's Hospital of Philadelphia](#).

Conflict of interest

The authors declare no financial conflicts of interest.

References

- Balise, R., Odom, G., Calderon, A., Bouzoubaa, L., DeFreitas, W., & Grealis, K. (2023). *tidyREDCap: Helper functions for working with 'REDCap' data*. <https://raymondbalise.github.io/tidyREDCap/index.html>
- Beasley, W. (2023). *REDCapR: Interaction between r and REDCap*. <https://ouhscbbmc.github.io/REDCapR/>
- Beasley, W., & Balise, R. (2023). *Writing to a REDCap project*. <https://ouhscbbmc.github.io/REDCapR/articles/workflow-write.html>
- Carmezim, J., Peñafiel, J., Satorra, P., García, E., Pallarés, N., & Tebé, C. (2023). *REDCapDM: 'REDCap' data management*. <https://ubidi.github.io/REDCapDM/>
- Hanna, R., Porter, E., & Kadauke, S. (2023). *REDCapTidieR*. <https://chop-cgtinformatics.github.io/REDCapTidieR/index.html>
- 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>
- Larmarange, J. (2023). *Labelled: Manipulating labelled data*. <https://larmarange.github.io/labelled/>
- Lingen, J. ter. (2023). *Superhero database*. <https://www.superheroddb.com/>
- Müller, K., & Wickham, H. (2023). *Tibble: Simple data frames*.
- Nutter, B., & Lane, S. (2023). *redcapAPI: Accessing data from REDCap projects using the API*. <https://doi.org/10.5281/zenodo.11826>
- Open source security foundation_2023. (2023). In *Open Source Security Foundation*. The Linux Foundation. <https://openssf.org/>
- openxlsx2: Read, write and edit 'xlsx' files. (2023). [Computer software]. <https://janmarvin.github.io/openxlsx2/>
- R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Waring, E., Quinn, M., McNamara, A., Arino de la Rubia, E., Zhu, H., & Ellis, S. (2023). *Skimr: Compact and flexible summaries of data*. <https://docs.ropensci.org/skimr/> (website)

¹²⁹ Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 59(10), 1–23. [https:](https://doi.org/10.18637/jss.v059.i10)
¹³⁰ [//doi.org/10.18637/jss.v059.i10](https://doi.org/10.18637/jss.v059.i10)

DRAFT