The Journal of Open Source Software

dtrackr: An R package for tracking the provenance of data

Robert Challen [1,2]

1 Engineering Mathematics, University of Bristol, Bristol, UK **2** College of Engineering, Mathematics and Physical Sciences, University of Exeter, Devon, UK

DOI: 10.21105/joss.04707

Software

- Review
- Repository
- Archive

Editor: Andrew Stewart

Reviewers:

- @debruine
- @craig-willis

Submitted: 28 June 2022
Published: 28 November 2022

Summary

An accurate statement of the provenance of data is essential in bio-medical research. Powerful data manipulation tools available in the `tidyverse` R package ecosystem (Wickham et al., 2019) provide the infrastructure to assemble, clean and filter data prior to statistical analysis. Manual documentation of the steps taken in the data pipeline and the provenance of data is a cumbersome and error prone task which may restrict reproducibility. `dtrackr` is a wrapper around a subset of the standard `tidyverse` data manipulation tools that allows automatic tracking of the processing steps applied to a data set, prior to statistical analysis. It allows early detection and reporting of data quality problems, and automatically documents a pipeline of data transformations as a flowchart in a format suitable for scientific publication, including, but not limited to CONSORT diagrams (Schulz et al., 2010).

`dtrackr` is first and foremost a utility to accelerate and improve research by facilitating documentation, supporting extraction of knowledge from data sets, and the execution of research by helping identify data quality issues. The general capability however fits into a broader context of other provenance or data pipeline research. This includes initiatives such as `C2Metadata` (Alter et al., 2021), which focus on a language independent representation of a data pipeline, and R packages such as `targets` (Landau, 2021) which focus on documenting pipeline code, and managing the execution of a pipeline, or `RDataTracker` which focusses on tracking the execution of a arbitrary R script (Lerner et al., 2018). `dtrackr` takes a more data oriented approach, which could be complementary, in which we remain agnostic to the detail of a data pipeline script or nature of its execution, but capture a subset of the transformations applied to data alongside the data itself, thereby documenting the data state as it is being manipulated. This is achieved by overriding the execution of `dplyr` pipeline functions and results in a retrospective record of provenance (Pimentel et al., 2019). `dtrackr` also has the ability to insert secondary analysis as annotations into the pipeline, and allows control over what information is collected, ultimately with a view to producing simple human readable output. The approach of `dtrackr` is analogous to a `git` commit history for dataframes, and there is potential synergy with emerging versioned databases such as `dolt` (*Dolt Is Git for Data!*, 2019/2022; Ross, 2022).

Statement of need

The collection of experimental or observational data for research is often an iterative endeavour, involving curation of complex data sets designed for multiple goals. Systematic data quality checking for such sets is a major challenge, particularly when they are assembled to identify emerging or rapidly evolving issues. Feedback from early data analysis can identify specific data quality issues, resolution of which can considerably improve data for the task at hand. However this requires a clear understanding of why and when individual data items are excluded, which is potentially tedious and may be seen as lower priority compared to statistical analysis.

Data analysis using `tidyverse` in R is a rapid means of transforming raw data into a format suitable for statistical analysis. The transformations involved can, however affect the results of statistical analysis, and meticulous care must be taken to ensure that any assumptions made during data processing are well documented. It is often too easy to inadvertently exclude data where filtering on missing items, or joining linked data sets with incomplete foreign key relationships.

In complex data analysis, the use of interactive programming environments such as Read-Eval-Print Loops (REPL) in R markdown documents, interim caching of results, or conditional branching data pipelines, can result in the current state of a processed data set becoming decoupled from the code that is designed to generate them.

To surface these issues bio-medical journal articles are usually required to report data manipulation to an agreed standard. For example, CONSORT diagrams are part of the requirements in reporting parallel group clinical trials. They are described in the updated 2010 CONSORT statement (Schulz et al., 2010), and clarify how patients were recruited, selected, randomized and followed up. For observational studies, such as case control designs, an equivalent requirement is the STROBE statement (von Elm et al., 2008). There are many other similar requirements for other types of study, such as the TRIPOD statement for multivariate models (Collins et al., 2015). Maintaining such CONSORT diagram over the course of a study when data sets are being actively collected and data quality issues being addressed is time-consuming.

`dtrackr` addresses these issues by instrumenting a commonly used subset of standard `tidyverse` data manipulation pipeline functions from `dplyr` and `tidyr`. It can automatically record the steps taken, records excluded and a summary of the result of each data processing step, as part of the data set itself in a "history graph". In this way data sets retain an accurate history of their own provenance regardless of the actual route taken to assemble them. This history includes a complete record of any data quality issues that lead to excluded records. The history is a directed graph which can be expressed in the commonly used `GraphViz` language (Gansner & North, 2000) and may be visualised as a flowchart such as in Figure 1; this uses the Chronic Granulomatous Disease dataset from the `survival` package (Terry M. Therneau & Patricia M. Grambsch, 2000; Therneau, 2022) as an example of a parallel group study and produces a STROBE like flowchart.

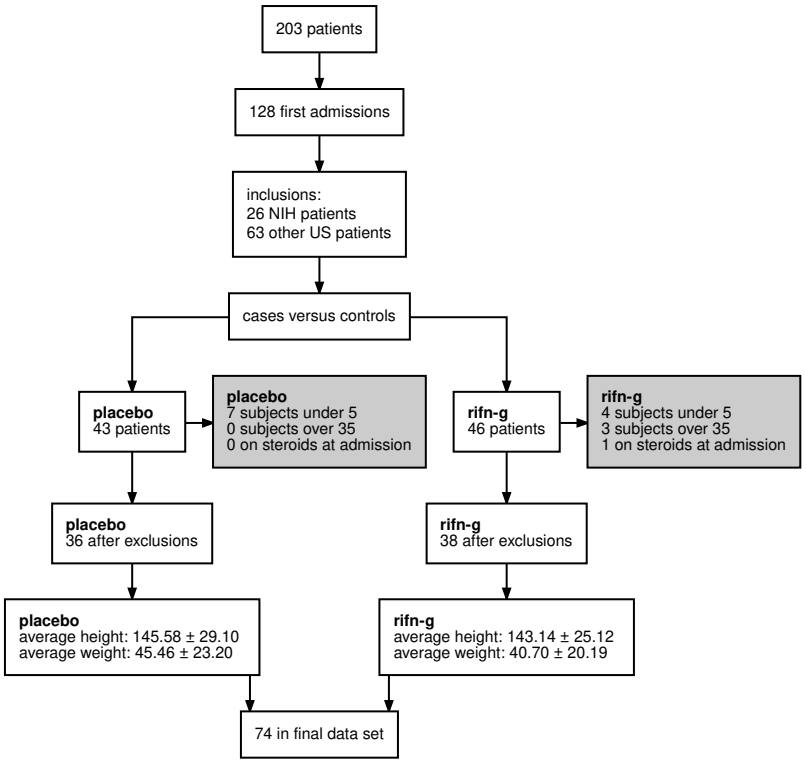

Figure 1: An example flowchart derived directly from a simple analysis of the Chronic Granulomatous Disease dataset demonstrating use of `dtrackr` to generate the key parts of a STROBE or CONSORT diagram.

`dtrackr` was originally conceptualized during an analysis I undertook of the severity of the Alpha variant of SARS-CoV-2 (Challen et al., 2021), and has since been used for other epidemiological studies including an analysis of the incidence of hospitalization of acute lower respiratory tract disease in Bristol (Hyams, Challen, Begier, et al., 2022), and a comparative analysis of the severity of the SARS-CoV-2 Omicron variant, versus the Delta variant against a range of hospital outcomes (Hyams, Challen, Marlow, et al., 2022).

Although the specific example presented here is in the bio-medical domain, tracking the provenance of data is a much broader issue, and we anticipate there are many other applications for `dtrackr`.

Acknowledgements

Thanks for contributions from TJ McKinley. I gratefully acknowledge the financial support of the EPSRC via grants EP/N014391/1, EP/T017856/1, the MRC (MC/PC/19067), and from the Somerset NHS Foundation Trust, Global Digital Exemplar programme.

References

Alter, G. C., Gager, J., Heus, P., Hunter, C., Ionescu, S., Iverson, J., Jagadish, H. V., Lyle, J., Mueller, A., Nordgaard, S., Risnes, O., Smith, D., & Song, J. (2021). Capturing Data Provenance from Statistical Software. *International Journal of Digital Curation*, *16*(1, 1), 14–14. https://doi.org/10.2218/ijdc.v16i1.763

Challen, R., Brooks-Pollock, E., Read, J. M., Dyson, L., Tsaneva-Atanasova, K., & Danon, L. (2021). Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: Matched cohort study. *BMJ*, *372*, n579. https://doi.org/10.1136/bmj.n579

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *BMC Medicine*, *13*(1), 1. https://doi.org/10.1186/s12916-014-0241-z

Dolt is Git for Data! (2022). [Computer software]. DoltHub. https://github.com/dolthub/dolt (Original work published 2019)

Gansner, E. R., & North, S. C. (2000). An open graph visualization system and its applications to software engineering. *Software - Practice and Experience*, *30*(11), 1203–1233. https://doi.org/10.1002/1097-024X(200009)30:11%3C1203::AID-SPE338%3E3.0.CO;2-N

Hyams, C., Challen, R., Begier, E., Southern, J., King, J., Morley, A., Szasz-Benczur, Z., Garcia Gonzalez, M., Kinney, J., Campling, J., Gray, S., Oliver, J., Hubler, R., Valluri, S. R., Vyse, A., McLaughlin, J. M., Ellsbury, G., Maskell, N., Gessner, B., … Finn, A. (2022). *Incidence of Community Acquired Lower Respiratory Tract Disease in Bristol, UK During the COVID-19 Pandemic* [SSRN Scholarly Paper]. https://doi.org/10.2139/ssrn.4087373

Hyams, C., Challen, R., Marlow, R., Nguyen, J., Begier, E., Southern, J., King, J., Morley, A., Kinney, J., Clout, M., Oliver, J., Ellsbury, G., Maskell, N., Jodar, L., Gessner, B., McLaughlin, J., Danon, L., Finn, A., & Group, T. A. C. R. (2022). *Severity of Omicron (B.1.1.529) and Delta (B.1.1.617.2) SARS-CoV-2 infection among hospitalised adults: A prospective cohort study* (p. 2022.06.29.22277044). medRxiv. https://doi.org/10.1101/2022.06.29.22277044

Landau, W. M. (2021). The targets R package: A dynamic Make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software*, *6*(57), 2959. https://doi.org/10.21105/joss.02959

Lerner, B., Boose, E., & Perez, L. (2018). Using Introspection to Collect Provenance in R. *Informatics*, *5*(1), 12. https://doi.org/10.3390/informatics5010012

Pimentel, J. F., Freire, J., Murta, L., & Braganholo, V. (2019). A Survey on Collecting, Managing, and Analyzing Provenance from Scripts. *ACM Computing Surveys*, *52*(3), 47:1–47:38. https://doi.org/10.1145/3311955

Ross, N. (2022). *Doltr: A client for the dolt database* [Manual].

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *BMJ*, *340*, c332. https://doi.org/10.1136/bmj.c332

Terry M. Therneau, & Patricia M. Grambsch. (2000). *Modeling survival data: Extending the Cox model*. Springer. ISBN: 0-387-98784-3

Therneau, T. M. (2022). *A package for survival analysis in r*. https://CRAN.R-project.org/package=survival

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., Vandenbroucke, J. P., & STROBE Initiative. (2008). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Journal of Clinical Epidemiology*, *61*(4), 344–349.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, 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