DataRepExp: a R shiny Application that makes Data FAIR for Data Repositories

2024-02-22

Rory Chen1, Vibeke S Catts1, Ashleigh Vella1, Juan Carlo San Jose2, Sarah Bauermeister3,4, Josh Bauermeister3,4, Emma Squires3,5, Simon Thompson3,5, John Gallacher3,4, Perminder S. Sachdev1

1Centre for Healthy Brain Ageing, Discipline of Psychiatry & Mental Health, School of Clinical Medicine, University of New South Wales, Sydney, Australia

2Research Technology Services, University of New South Wales, Sydney, Australia

3Dementias Platform UK, Oxford, United Kingdom

4Department of Psychiatry, University of Oxford, Oxford, United Kingdom

5Population Data Science, Swansea University, Swansea, United Kingdom

## Summary

The Data Repository Explorer, DataRepExp, is an open-source R Shiny application developed to improve the findability, accessibility, interoperability, and reusability (FAIR) (ref) of research data hosted on a data repository.

The application displays standardised metadata across multiple studies including data availability by categories (such as demographics, medical history, imaging data and genomic data) to allow high-level comparison. It enables users to explore and run preliminary analysis from participants that match certain criteria. In addition, it provides features to export reports and aggregated results for data access application purposes.

The application was initially developed for a discipline-specific data-sharing platform, the Dementias Platform Australia (DPAU) [1]. Envisioning this work could be utilized by other data repositories, this demo application was created using simulated health-related data for demonstrating purposes, can be populated to other data repositories in diverse disciplines.

Source Code: <https://github.com/RoryChenXY/DataRepExp_public>

Web Application: <https://rorychenxy.shinyapps.io/DataRepExp/>

Contact: [xinyue.chen1@unsw.edu.au](mailto:xinyue.chen1@unsw.edu.au)

## Statement of need

Data repositories have become increasingly important in recent years as more emphasis has been placed on open science practices and data sharing. By making data publicly available through repositories, researchers can ensure data persistence and support data preservation, as well as facilitate the reuse of their data, thereby increasing the potential for new scientific discoveries. However, challenges exist for data findability, accessibility, interoperability, and reusability (FAIR) [2].

Even though most data repositories have adopted various metadata schemas to describe the dataset [3], it is increasingly a challenge for researchers to find relevant data that meet research interests or needs [4]. For multi-study research, applying to access different datasets usually comes with diverse and complicated data-sharing requirements and workflows, extensive administrative workloads and waiting periods. Upon approval, substantial efforts of data harmonization are usually required due to inconsistent data structures and labelling conventions, and harmonised dataset are hardly reused. We found that many data repositories do not provide comprehensive metadata, nor centralised tables for comparison. With repositories that provide data visualisation, Power BI and Tableau were commonly used but cost occurs. R-shiny could provide more flexibility and functions with a fraction of the cost.

Designed to enable easier access to research data hosted on data repositories, DPAU [1] seeks to address these challenges with R-Shiny. The application designed for DPAU includes rich metadata and a set of commonly used variables [5], identified as being of broad interest to dementia research, harmonised using the C-Surv data model [6], which has been developed by Dementias Platform UK (DPUK) [7], and adopted by Alzheimer’s Disease Data Initiative (ADDI) [8] and DPAU [1]. Researchers can identify datapoints from participants that match certain criteria, using filters at study and/or participant levels, then explore and conduct preliminary analysis on the filtered dataset. It allows users to export reports and aggregated results. The exported reports can then be used when submitting a single centralised data access application form for accessing data from multiple studies through the DPAU Data Portal [9].

DataRepExp was created with simulated data and a list of generalized health-related variables. This work can be modified and utilized by other data repositories by adopting the discipline-specific metadata schema and common variables. With rich metadata for findability, interactive visualization dashboard for accessibility, standardization and harmonization for data interoperability and reusability, this tool can improve the FAIR of research data hosted on a data repository. R programming skill is required for reproducibility, detailed documentation and syntax is open-source and publicly available.

## Methods

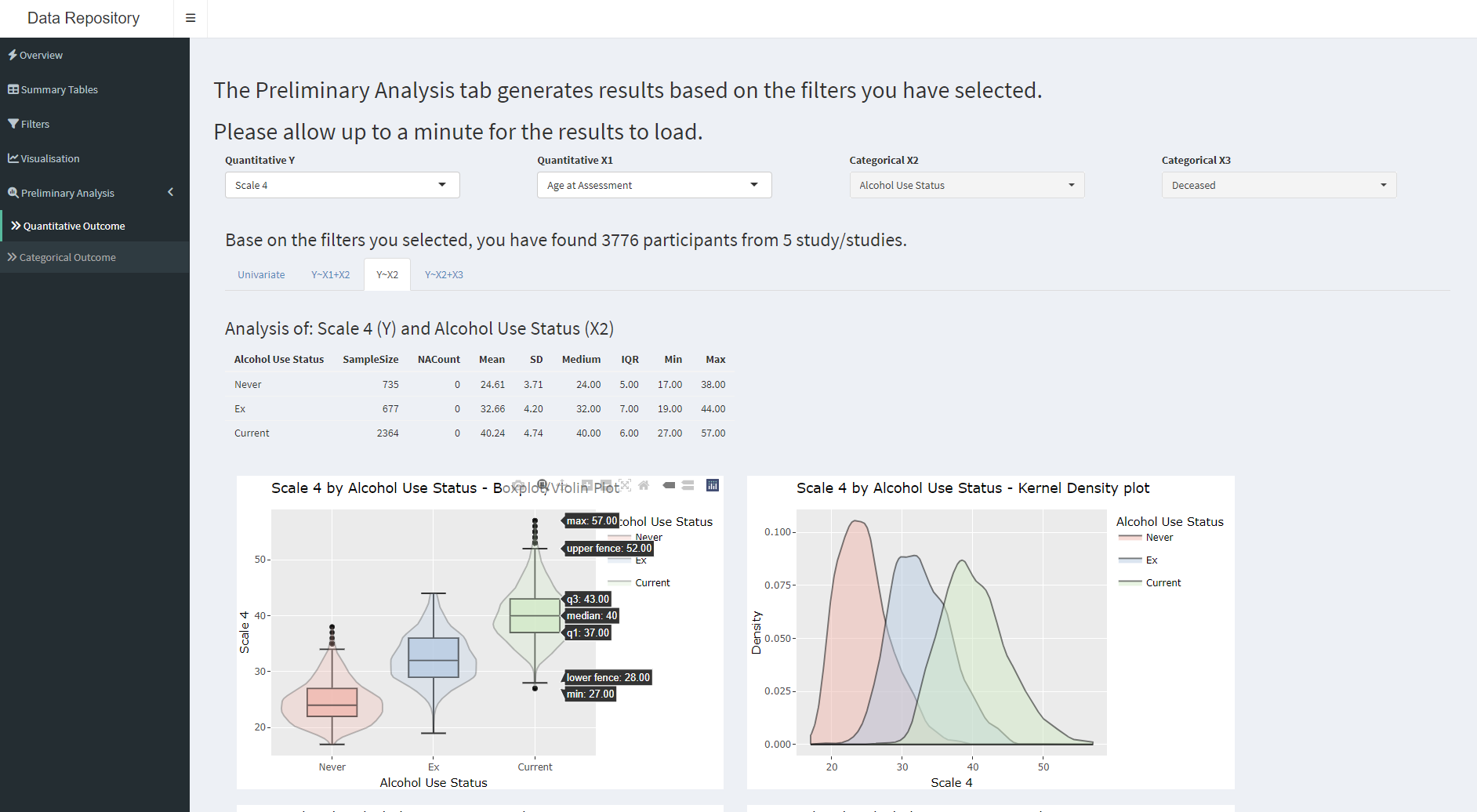
DataRepExp was written using R [10] and JavaScript using the following libraries:

* Shiny: shiny [11], shinydashboard [12], shinyWidgets [13], shinyjs [14].
* Data manipulation: dplyr [15], tidyr [16], tidyverse [17], forcats [18], useful [19], magrittr [20], purrr [21].
* Data Report and Visualisation: ggplot2 [22], plotly [23], scales [24], DT [25], htmltools [26], fontawesome [27].

## Overview

The application layout features a side menu, through which the users can navigate through tabs, and the main view which displays the content of the selected tab (Figure 1).

* First tab – Overview: includes statement and navigation instruction.
* Second tab – Summary Tables: three metadata tables for high-level comparison
* Third tab – Filters and Filter Reports: users can adjust and apply filters to identify participants and studies that match selected criteria. Then download Filter Report with the list of studies that matched the filters selected.
* Fourth tab – Visualisation: plots organised by different domains, generated using filtered dataset.
* Fifth tab – Preliminary Analysis: run preliminary analysis with user selected variables



Application features include:

* Simulation: For demonstration purposes, we generate simulated data. Scripts and reference documents used to generate the data can be found in the GitHub repository.
* Modularisation: DataRepExp was built in Shiny modules. Modularity makes the app easy to test, maintain, and deploy. The features can be easily further expanded with loose coupling module design.
* Interactive: DataRepExp provides an interactive interface that allows users to engage with the data and output. Elevated user experience with integrative charts and figures, which include functions such as sort, filter, zoom, select, adjust axis, hover for information, reset, etc.

## Deployment

The Data Repository Explorer, DataRepExp, the demo app is hosted through easy-to-use shinyapps.io while the DPAU version is hosted on AWS environment using Shiny Server for high availability, scalability, security, and compliance.

## Acknowledgements

This application was previously inspired by the visualisation tool developed by Dementias Platform UK (DPUK) using PowerBI, then developed in R-Shiny for the Dementias Platform Australia (DPAU).

## Funding

This work is supported by a grant from the National Institute on Aging/ National Institute of Health (NIA/NIH). [1RF1AG057531-01]

## Availability and Community Guidelines

The application and source code are available at the GitHub repository [<https://github.com/RoryChenXY/DataRepExp_public>]. Users and contributors are welcome to contribute, request features, and report bugs through the GitHub repository.

## References

1 Centre for Healthy Brain Ageing (CHeBA) UNSW for. Dementias Platform Australia. <https://www.dementiasplatform.com.au/>

2 Wilkinson MD, Dumontier M, Aalbersberg IjJ, *et al.* The FAIR guiding principles for scientific data management and stewardship. *Scientific data*. 2016;3:1–9.

3 Contaxis N, Clark J, Dellureficio A, *et al.* Ten simple rules for improving research data discovery. PLoS computational biology. 2022;18:e1009768.

4 Gregory SJAM Kathleen AND Khalsa. [Eleven quick tips for finding research data](https://doi.org/10.1371/journal.pcbi.1006038). *PLOS Computational Biology*. 2018;14:1–7.

5 Bauermeister S, Phatak M, Sparks K, *et al.* Evaluating the harmonisation potential of diverse cohort datasets. *European Journal of Epidemiology*. 2023;38:605–15.

6 Bauermeister S, Bauermeister JR, Bridgman R, *et al.* Ready data: The c-surv data model. *European Journal of Epidemiology*. 2023;38:179–87.

7 Bauermeister S, Orton C, Thompson S, *et al.* The Dementias Platform UK (DPUK) data portal. *European journal of epidemiology*. 2020;35:601–11.

8 Alzheimer’s Disease Data Initiative. <https://www.alzheimersdata.org/>

9 Centre for Healthy Brain Ageing UNSW for. Dementias platform Australia Data Portal. <https://portal.dementiasplatform.com.au/>

10 R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing 2023. <https://www.R-project.org/>

11 Chang W, Cheng J, Allaire J, *et al.* *Shiny: Web application framework for r*. 2023. <https://shiny.posit.co/>

12 Chang W, Borges Ribeiro B. *Shinydashboard: Create dashboards with shiny*. 2021. <http://rstudio.github.io/shinydashboard/>

13 Perrier V, Meyer F, Granjon D. *shinyWidgets: Custom inputs widgets for shiny*. 2023. <https://github.com/dreamRs/shinyWidgets>

14 Attali D. *Shinyjs: Easily improve the user experience of your shiny apps in seconds*. 2021. <https://deanattali.com/shinyjs/>

15 Wickham H, François R, Henry L, *et al.* *Dplyr: A grammar of data manipulation*. 2023. <https://dplyr.tidyverse.org>

16 Wickham H, Vaughan D, Girlich M. *Tidyr: Tidy messy data*. 2023. <https://tidyr.tidyverse.org>

17 Wickham H. *Tidyverse: Easily install and load the tidyverse*. 2023. <https://tidyverse.tidyverse.org>

18 Wickham H. *Forcats: Tools for working with categorical variables (factors)*. 2023. <https://forcats.tidyverse.org/>

19 Lander JP. *Useful: A collection of handy, useful functions*. 2023. <https://github.com/jaredlander/useful>

20 Bache SM, Wickham H. *Magrittr: A forward-pipe operator for r*. 2022. <https://magrittr.tidyverse.org>

21 Wickham H, Henry L. *Purrr: Functional programming tools*. 2023. <https://purrr.tidyverse.org/>

22 Wickham H, Chang W, Henry L, *et al.* *ggplot2: Create elegant data visualisations using the grammar of graphics*. 2023. <https://ggplot2.tidyverse.org>

23 Sievert C, Parmer C, Hocking T, *et al.* *Plotly: Create interactive web graphics via plotly.js*. 2023. <https://plotly-r.com>

24 Wickham H, Seidel D. *Scales: Scale functions for visualization*. 2022. <https://scales.r-lib.org>

25 Xie Y, Cheng J, Tan X. *DT: A wrapper of the JavaScript library DataTables*. 2023. <https://github.com/rstudio/DT>

26 Cheng J, Sievert C, Schloerke B, *et al.* *Htmltools: Tools for HTML*. 2023. <https://github.com/rstudio/htmltools>

27 Iannone R. *Fontawesome: Easily work with font awesome icons*. 2023. <https://github.com/rstudio/fontawesome>