International Journal of Epidemiology, 2022, 1–6
https://doi.org/10.1093/ije/dyac201
Software Application Profile

Software Application Profile

Software Application Profile: ShinyDataSHIELD—an R Shiny application to perform federated non-disclosive data analysis in multicohort studies

Xavier Escribà-Montagut,[1,2] Yannick Marcon,[3] Demetris Avraam,[4] Soumya Banerjee,[5] Tom RP Bishop,[5] Paul Burton [4] and Juan R González [1,2,6,*]

[1]Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain, [2]Universitat Pompeu Fabra (UPF), Barcelona, Spain, [3]Epigeny, St Ouen, France, [4]Population Health Sciences Institute, Newcastle University, Newcastle, UK, [5]MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge, UK and [6]Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública, Barcelona, Spain

*Corresponding author. Barcelona Biomedical Research Park (PRBB), Doctor Aiguader, 88, 08003 Barcelona, Spain. E-mail: juanr.gonzalez@isglobal.org

Received 9 September 2021; Editorial decision 20 September 2022; Accepted 10 October 2022

Abstract

Motivation: DataSHIELD is an open-source software infrastructure enabling the analysis of data distributed across multiple databases (federated data) without leaking individuals' information (non-disclosive). It has applications in many scientific domains, ranging from biosciences to social sciences and including high-throughput genomic studies. R is the language used to interact with (and build) DataSHIELD. This creates difficulties for researchers who do not have experience writing R code or lack the time to learn how to use the DataSHIELD functions. To help new researchers use the DataSHIELD infrastructure and to improve the user-friendliness for experienced researchers, we present ShinyDataSHIELD.
Implementation: ShinyDataSHIELD is a web application with an R backend that serves as a graphical user interface (GUI) to the DataSHIELD infrastructure.
General features: The version of the application presented here includes modules to perform: (i) exploratory analysis through descriptive summary statistics and graphical representations (scatter plots, histograms, heatmaps and boxplots); (ii) statistical modelling (generalized linear fixed and mixed-effects models, survival analysis through Cox regression); (iii) genome-wide association studies (GWAS); and (iv) omic analysis (transcriptomics, epigenomics and multi-omic integration).
Availability: ShinyDataSHIELD is publicly hosted online [https://datashield-demo.obiba.org/], the source code and user guide are deposited on Zenodo DOI 10.5281/zenodo.6500323, freely available to non-commercial users under 'Commons Clause' License Condition v1.0. Docker images are also available [https://hub.docker.com/r/brgelab/shiny-data-shield].

Key words: R Shiny, federated analysis, non-disclosive analysis, DataSHIELD, multicohort studies, genetic epidemiology

Introduction

Data privacy continues to be a central concern in contemporary research[1]. There are many ethico-legal considerations, including requirements under General Data Protection Regulation (GDPR), which must be taken into account when planning and configuring an analysis involving sensitive data,[2] in particular mitigating the risk of individual identification. Such considerations have a huge impact on the feasibility and time required to carry out multicohort studies and genomic studies, which rely on obtaining permissions to access and share sensitive data. DataSHIELD is an open-source software infrastructure aimed at facilitating an effective and appropriate response to such challenges[3,4]. To achieve this, DataSHIELD represents an infrastructure where the researchers only ever receive sufficient statistics (low-dimensional data transformations/aggregations containing all of the information needed to drive whatever analysis is required) from each of the different data servers, and the servers themselves manage their data using Obiba's Opal technology[5]. The data owners/custodians manage these servers and have sole control of the disclosure filters that are applied to the outputs, as well as the set of DataSHIELD functions that can be used on their data. This enables researchers to perform analyses on federated data without the need to possess physical copies of the data from each source. On the DataSHIELD website [https://www.datashield.org/help] the reader can find information on how to:

i. conduct basic statistical analyses;
ii. administrate the servers;
iii. deploy DataSHIELD functions and packages.

Recently, DataSHIELD has seen a major upgrade focused on expanding the scope of which types of data can be analysed[6], which results in the ability to analyse high-volume, potentially non-tabular data such as genomics data structures, among many others. Many science fields can now make use of this extension, and therefore it is important to make DataSHIELD easier to use for non-technical users.

The DataSHIELD infrastructure includes a series of R packages that enables the remote and non-disclosive analysis of sensitive data [https://www.datashield.org/help/community-packages]. The software described in this article uses a subset of functionalities from the dsBase[7] package for data shaping, analysis and presentation methods, the dsSurvival[8] package for survival analysis and the dsOmics[9] package for omics analysis.

We present ShinyDataSHIELD, an R Shiny[10] application that enables interaction with the DataSHIELD analysis infrastructure via a web application, providing capabilities to perform federated non-disclosive analysis for non-technical users. We have designed the application to provide a user-friendly experience that frees the researcher from writing any analysis code.

Implementation

The following list describes all the functionalities of the software presented, which can be used in two configurations: (i) single data sources; and (ii) multiple data sources in a federated configuration. The single-source configuration invokes all of the privacy protection features of DataSHIELD but, by definition, has no need to activate the routines and algorithms permitting federated co-analysis across multiple sites. For further information we have made available an online user guide with different use cases and technical information [https://isglobal-brge.github.io/ShinyDataSHIELD_bookdown/]. All the functionalities use DataSHIELD disclosure controls.

Tabular data functionalities

i. Data column types. When dealing with tabular data, a researcher may be interested in: (a) assessing the column class; and (b) transforming the class of a column. Both functionalities are available.
ii. Descriptive statistics. There are a number of functions that can display descriptive statistics in two different ways.
    a. Summary statistics. Available for numerical and categorical variables. It displays a table with summary statistics. For categorical variables, it displays the number of counts in each category, and for numerical variables, the quartiles and the mean values.
    b. Graphical representations.: (1) scatter plot, to visualize the relationship between two numerical variables; (2) histogram, to visualize the distribution of a numerical variable; (3) heatmap, to visualize the density of counts in grids formed by two numerical variables; and (4) boxplot, to visualize the locality and spread of one or more numerical variables, with the option of performing two groupings using

categorical variables. All graphical representations preserve data privacy as they are generated through anonymization techniques or disclosure controls[11].

iii. Statistical modelling: There are three classes of statistical models that can be fit.
   a. Generalized linear models (GLM): GLM models can be fitted using pooled techniques or meta-analysis techniques (study-level meta-analysis fitting will also yield a forest plot of the results as well as the regression results table). For both approaches, the user can specify the error distribution to be either Gaussian, Poisson or binomial for linear, Poisson and logistic regressions, respectively.
   b. Generalized linear mixed effects models (GLMer): GLMer models are fitted using meta-analysis techniques (a forest plot is displayed alongside the results table). The user can specify the error distribution to be either Poisson or binomial.
   c. Survival analysis: Survival analysis can be performed via a study-level meta-analysis of Cox regression models. The models can be fitted for different types of censoring: left, right, counting and intervals. The regression results are displayed in a table and a forest plot and privacy-preserving survival curves are also displayed.

Resources functionalities

i. Genomics. Genome-wide association studies (GWAS) can be performed using two types of resources: VCF (Variant Call Format) files and PLINK containers. The VCF files are analysed using BioConductor libraries (GWASTools[12] mainly), and PLINK containers are analysed using the PLINK software[13]. The GWAS results can be visualized as a table or as a Manhattan plot[14].
ii. Omics. Perform association analysis can be performed using Limma[15]. The accepted resources for this analysis are ExpressionSet and RangedSummarizedExperiment containers[16].

Miscellaneous functionalities

The miscellaneous functionalities are not part of DataSHIELD, they are part of the software that improves the user experience.

There is a built-in plot editor that allows some simple customization for the generated plots. This editor has been built using the *ggplot2*[17] and the *ggthemr*[18] R packages. The options that the plot editor offers are: (i) change text size; (ii) change x-axis text angle; (iii) add title, subtitle and caption; (iv) custom labels for x- and y-axes; (v) custom legend label (if there is a legend to be customized); and (vi) colour themes.

ShinyDataSHIELD has been implemented using a modular approach where all the modules have the same design, so there is no confusion when using different functionalities. Each one of the different modules performs a single task: a module for statistical modelling, a module for descriptive analysis etc. This guarantees that the application will be easy to upgrade and the source code will be easier to read for future maintainers.

The language for interacting with the modules follows a similar pattern, which in our application is the following.

a. When entering a module, the researcher must select the tables or resources to be used. Internal checks are always performed on the selected items to ensure the functions of the module will not crash.
b. The buttons are disabled until the researcher performs an operation that requires them, e.g. the visualization buttons for the survival models are not available until a survival model is fitted. This shared language ensures a consistent experience.

Operating in the background, R Shiny provides executive control of the functions in the different DataSHIELD packages, thus providing a seamless experience for the researcher, who simply receives the aggregated results as tables and figures interacting with a web application.

Use

In this section we will explain how to use the software. As previously mentioned, there is a common structure across all the functionalities described, which provides for a user-friendly experience. With this provision, the typical experience of a ShinyDataSHIELD user should be both easy and intuitive. A key component of our software is that there are multiple checks that display human-readable error messages, helping the users understand the reason something is not working. If the reader wishes to reproduce the displayed screenshots using our software, he/she can refer to the online user guide.

The first step is to define which Opal servers will be used, the credentials to be applied to them and which tables or resources have to be loaded into the remote R sessions hosted at the Opal servers. It is in this step that we define whether we want to use a single data source or multiple sources. To define multiple data sources, we just have to add more servers. This allows us to perform pooled analysis—combining inferences across all the specified servers. Once we have performed the connections, we only

have access to the particular set of datasets we have selected. If we wish to use different ones, we have to disconnect and reconnect again specifying the new set of datasets we now require. This step is illustrated in Figure 1.

The next step is to select which of the array of data available on each server should be used. This can sound counterintuitive, given we just selected the data to load in the previous step. However it adds flexibility, e.g. we can load multiple datasets and study them separately without having to disconnect. Moreover, it is at this point that the data are checked for integrity for doing pooled analysis. This ensures that the tables contain equivalent variables. This step is illustrated in Figure 2.

With this last step completed, we are inside the module that we have selected. Now we can finally use our data to undertake the desired statistical data analyses. All the modules have multiple functionalities, e.g. the descriptive analysis module has different visualization options (see Section 3.3. in the user guide bookdown), all the functionalities available are displayed along the navigation tabs of each module (see Figure 2). Linear, generalized linear and survival models can also be fitted using our Shiny app as described in Section 3.4. in the bookdown.

There are other modules that have functionalities that can only be used when a certain action has been performed. This is very easy and intuitive for the users, e.g. the genomics module can perform a GWAS and it can also plot the results (e.g. Manhattan plot). But the plot can only be created once a GWAS analysis has been performed (see Section 3.5. in the bookdown).

Having followed all the steps described above, the user is now in a position to perform any of the analyses described in the Implementation section.

Discussion

ShinyDataSHIELD is a novel R Shiny application that enables federated non-disclosive analyses for non-technical users. The goal of our software is to make the DataSHIELD infrastructure more accessible to researchers without R skills, as well as providing a platform for researchers experienced in DataSHIELD to perform quick hypothesis prototypes and quick analyses without the burden of writing a new analysis pipeline within R. For reference, even a basic pipeline aimed at fitting a linear model and necessarily incorporating appropriate check code, may typically require anywhere from 25 to 50 lines of code. However, despite benefiting from the simplicity provided by ShinyDataSHIELD, the researcher is not freed from the need to ensure that he/she is correctly interpreting the statistical results obtained from the application and that all models are built upon correct assumptions.

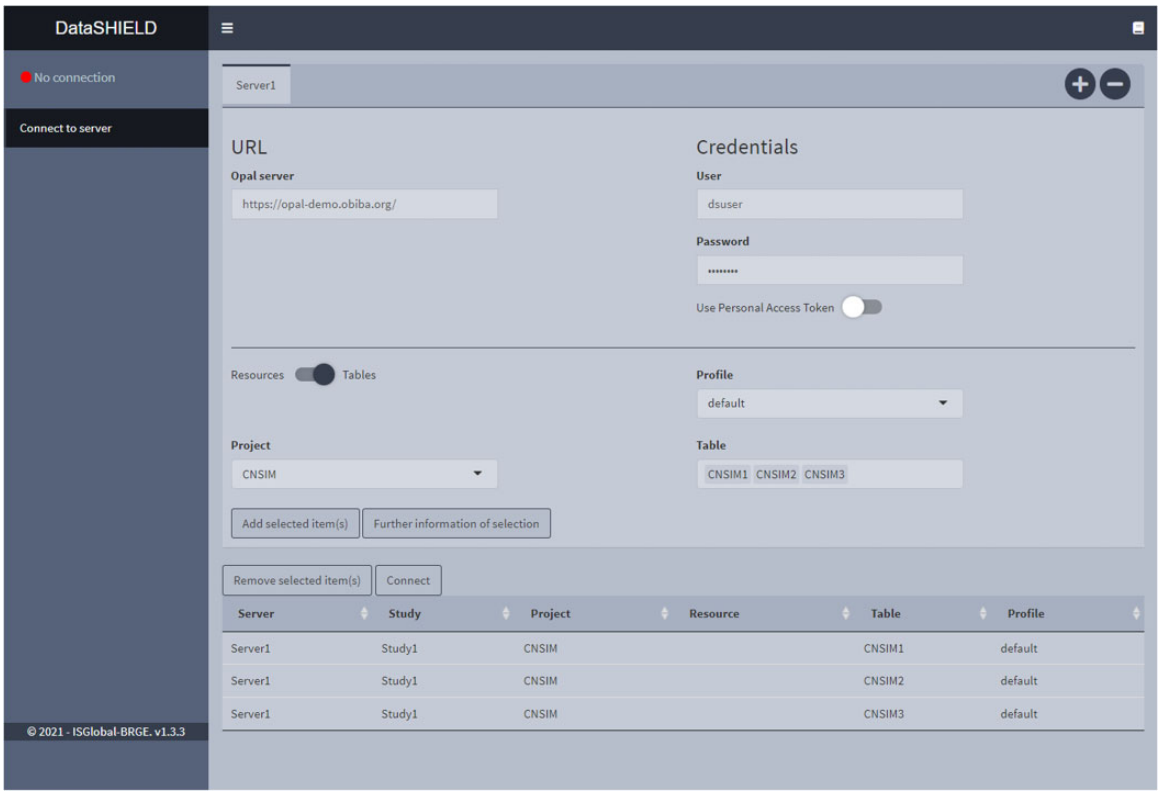

Figure 1 Connections interface. The illustrated configuration is a single server data source configuration with three different tables selected

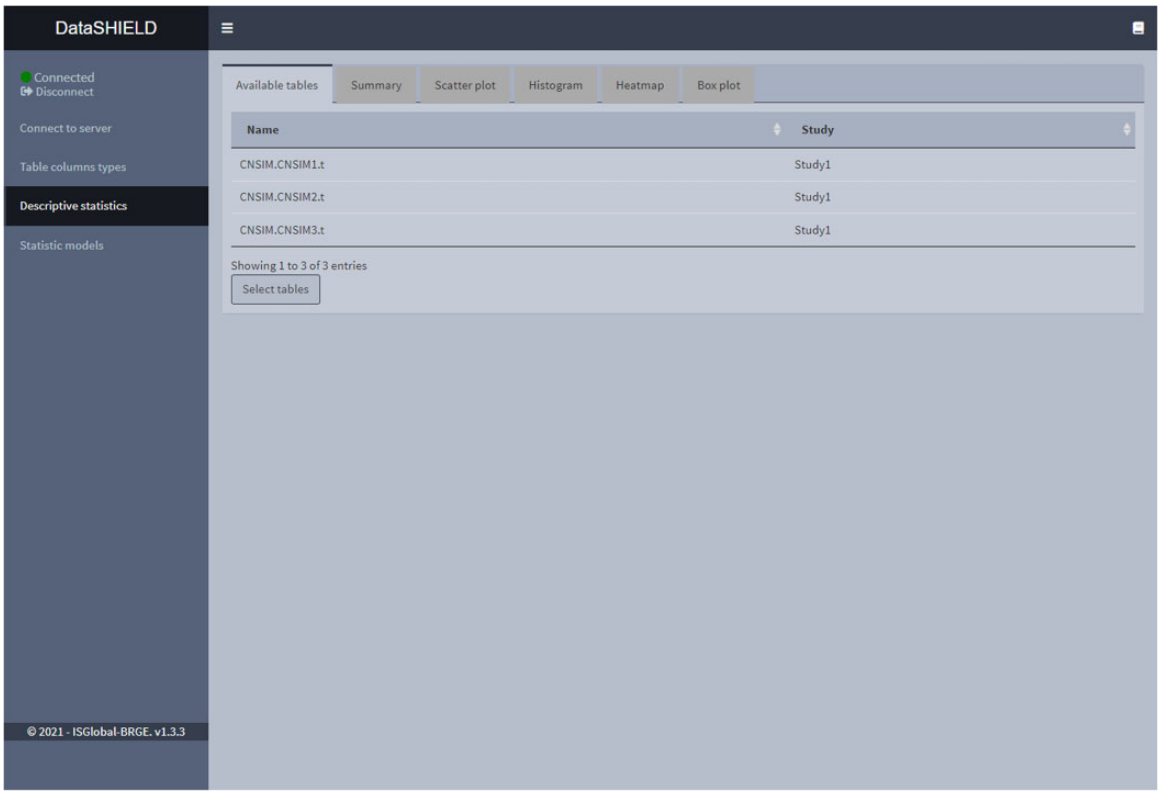

Figure 2 Selecting the tables to use. We can see, following Figure 1, that we have three available tables

Our software is designed around core DataSHIELD functionalities and will be expanded as new functionalities and packages are available. With the current version of the Shiny app, the plot editor can only be used for customizing boxplots, but we aim to make this functionality available for all other types of plots in a future release. Also, new functionalities will be added if researchers request them. Moreover, other research groups could take the source code and modify it to suit their particular needs. Since our software wraps DataSHIELD functions, periodical revisions from the maintainers will be required when new versions of DataSHIELD are released, to ensure that no wrappers are broken.

In conclusion, helping researchers to adopt non-disclosive methods for potentially federated analyses is inevitably a challenging task. When using DataSHIELD via its traditional integrated development environment, this involves learning DataSHIELD syntax. By providing a user-friendly Shiny R-based tool that can simplify the procedures and shorten the learning curve, we believe that this article and the technical work program underpinning it can contribute towards an accelerated adoption of such methods as well as demonstrating the capabilities of DataSHIELD. We hope that this will encourage researchers interested in the technology to explore its capabilities and test them out for themselves. Given ongoing developments in ShinyDataSHIELD as well as rapid evolution of the DataSHIELD infrastructure itself, ShinyDataSHIELD will be actively maintained and upgraded in the years to come, so that all the novel functionalities introduced by the DataSHIELD community can be used and be utilized in on our application.

Ethics approval

Not applicable.

Author contributions

Creation of the Shiny app: X.E. and Y.M. Study concept and design: D.A., T.B. and J.R.G. S.B provided statistical tools and methods for the software. Manuscript writing: X.E., D.A. and J.R.G. Project supervision: J.R.G. Evaluation of the software and editing of the manuscript: all authors. English proofreading: T.B. and P.B. All authors read and approved the final manuscript.

Funding

This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No 874583 (ATHLETE) and No 824989 (EUCANConnect); also through Centro de Excelencia Severo Ochoa 2019–2023 Program (CEX2018-000806-S); also has received funding from the project PID2021-122855OB-I00 funded by MCIN / AEI / 10.13039/501100011033 / FEDER, UE; and through the

support of the Government of Catalonia's Secretariat for Universities and Research of the Ministry of Economy and Knowledge (2017 SGR 801).

Conflict of interest

None declared.

References

1. Petrescu M, Krishen AS. Analyzing the analytics: data privacy concerns. *J Market Anal* 2018;**6**:41–43.
2. Abouelmehdi K, Beni-Hssane A, Khaloufi H, Saadi M. Big data security and privacy in healthcare: a review. *Procedia Comput Sci* 2017;**113**:73–80.
3. Gaye A, Marcon Y, Isaeva J *et al.* DataSHIELD: taking the analysis to the data, not the data to the analysis. *Int J Epidemiol* 2014;**43**:1929–44.
4. Wilson RC, Butters OW, Avraam D *et al.* DataSHIELD: new directions and dimensions. *Data Sci J* 2017;**16**:1–20.
5. Doiron D, Marcon Y, Fortier I, Burton P, Ferretti V. Software Application Profile: Opal and Mica: open-source software solutions for epidemiological data management, harmonization and dissemination. *Int J Epidemiol* 2017;**46**:1372–78.
6. Marcon Y, Bishop T, Avraam D *et al.* Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD. *PLoS Comput Biol* 2021;**17**:e1008880.doi: 10.1371/journal.pcbi.1008880
7. DataSHIELD Core Development Team. *dsBase: v6.1.1*. 2020. https://github.com/datashield/dsBase (3 October 2022, date last accessed).
8. Banerjee S, Sofack GN, Papakonstantinou T *et al.* dsSurvival: Privacy preserving survival models for federated individual patient meta-analysis in DataSHIELD. *BMC Res Notes* 2022;**15**:197.
9. González JR, Escriba-Montagut X, Marcon Y. *dsOmics: v1.0.7*. 2020. https://github.com/isglobal-brge/dsOmics (3 October 2022, date last accessed).
10. Chang W, Cheng J, Allaire JJ et al. *Shiny: Web Application Framework for R. R package version 1.5.0*. Published 2020. https://cran.r-project.org/package=shiny (3 October 2022, date last accessed).
11. Avraam D, Wilson R, Butters O *et al.* Privacy preserving data visualizations. *EPJ Data Sci* 2021;**10**:2.
12. Gogarten SM, Bhangale T, Conomos MP *et al.* GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics* 2012;**28**:3329–31.
13. Purcell S, Neale B, Todd-Brown K *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;**81**:559–75.
14. Turner SD. Annotated Manhattan plots and QQ plots for GWAS using R, Revisited. *Nat Preced* 2011. doi:10.1038/npre.2011.6070.1.
15. Smyth GK. *Limma, Linear Models for Microarray Data: v3.48.3*. 2021. https://bioconductor.org/packages/release/bioc/html/limma.html (3 October 2022, date last accessed).
16. *Summarized Experiment for Coordinating Experimental Assays, Samples, and Regions of Interest*. https://bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html (3 October 2022, date last accessed).
17. Wickham H. ggplot2. *Wires Comp Stat* 2011;**3**:180–85.
18. Mikata-Project. *ggthemr: v1.1.0*. 2020. https://github.com/Mikata-Project/ggthemr (3 October 2022, date last accessed).