
DataSHIELD workshop

Block 1 - Introduction to DataSHIELD

Juan R Gonzalez


Bionformatics Research Group in Epidemiology

e-mail: juanr.gonzalez@isglobal.org

ISGlobal
Barcelona
Institute for
Global Health



A partnership of:

 "la Caixa" Foundation

CLÍNIC
BARCELONA
Hospital Universitari



 UNIVERSITAT DE
BARCELONA

 Universitat
Pompeu Fabra
Barcelona

 Generalitat
de Catalunya



FUNDACIÓN
RAMÓN ARECES

Biomedical Research Park of Barcelona

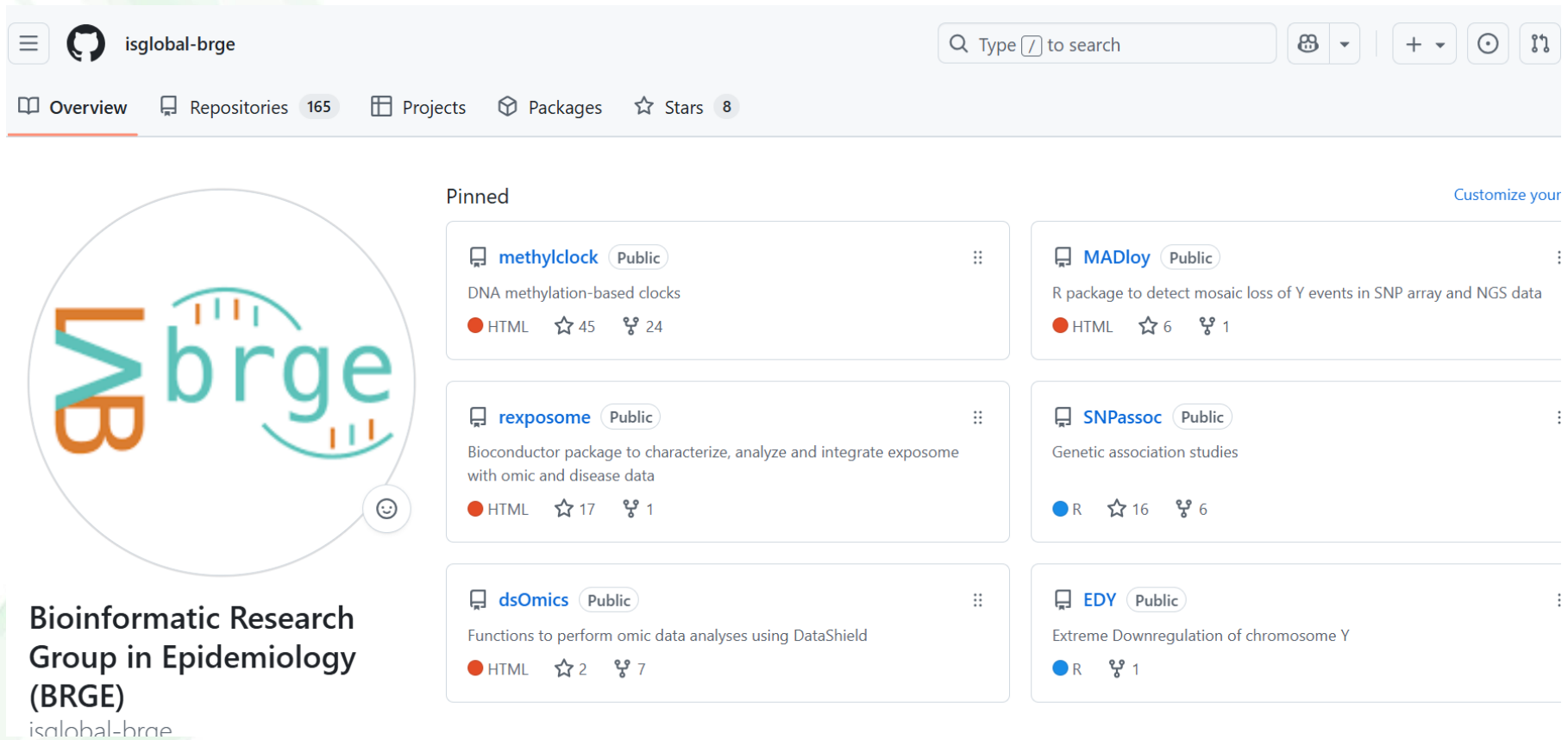


- ❑ Barcelona Institute for Global Health (ISGlobal)
- ❑ University Pompeu Fabra (UPF)
- ❑ Center for Genomic Regulation
- ❑ European Molecular Laboratory (EMBL)
- ❑ Institute of Medicine (Hospital)
- ❑ Evolutive Biology Laboratory



Bioinformatics Research Group in Epidemiology (BRGE)

<https://github.com/isglobal-brge>



isglobal-brge

Overview Repositories 165 Projects Packages Stars 8

Bioinformatic Research Group in Epidemiology (BRGE)
isglobal-brge

Pinned

- methylclock** (Public)
DNA methylation-based clocks
HTML 45 stars 24 forks
- MADloy** (Public)
R package to detect mosaic loss of Y events in SNP array and NGS data
HTML 6 stars 1 fork
- rexposome** (Public)
Bioconductor package to characterize, analyze and integrate exposome with omic and disease data
HTML 17 stars 1 fork
- SNPassoc** (Public)
Genetic association studies
R 16 stars 6 forks
- dsOmics** (Public)
Functions to perform omic data analyses using DataShield
HTML 2 stars 7 forks
- EDY** (Public)
Extreme Downregulation of chromosome Y
R 1 fork

[Customize your](#)

Health Analytics Hub @ISGlobal

RESEARCH

Health Analytics Hub



What is DataSHIELD

- ❑ A privacy-preserving, non-disclosive, federated analysis software
- ❑ What it is
 - ❑ way of analyzing sensitive data, at a individual level, without providing direct access
- ❑ What it is not
 - ❑ data harmonization platform
 - ❑ GUI based analysis tool (yet possible!)

What is DataSHIELD

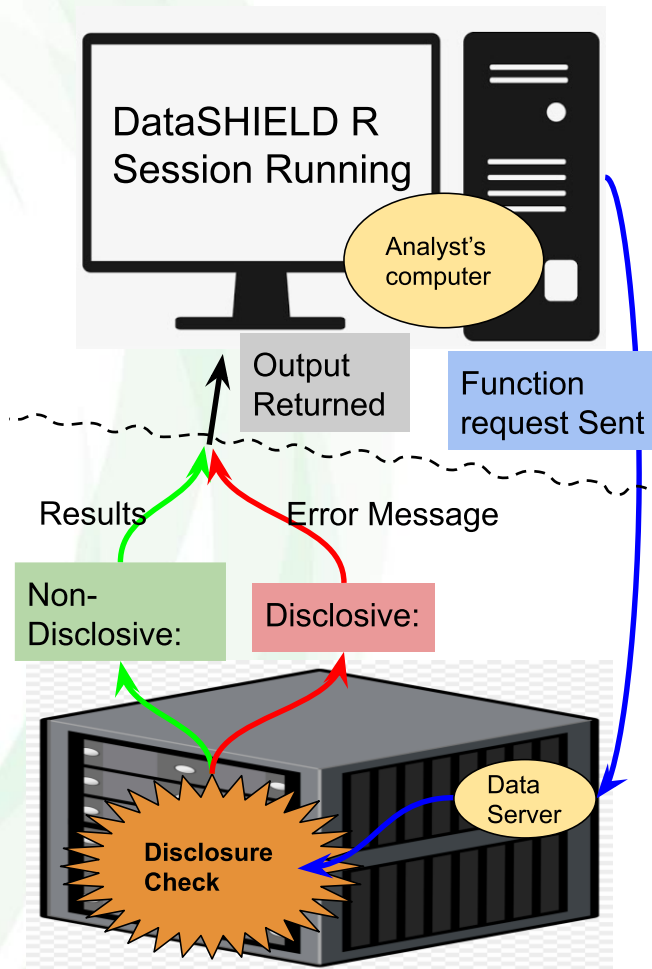
<https://datashield.org/>



About ▾ For analysts For data custodians For developers News Support

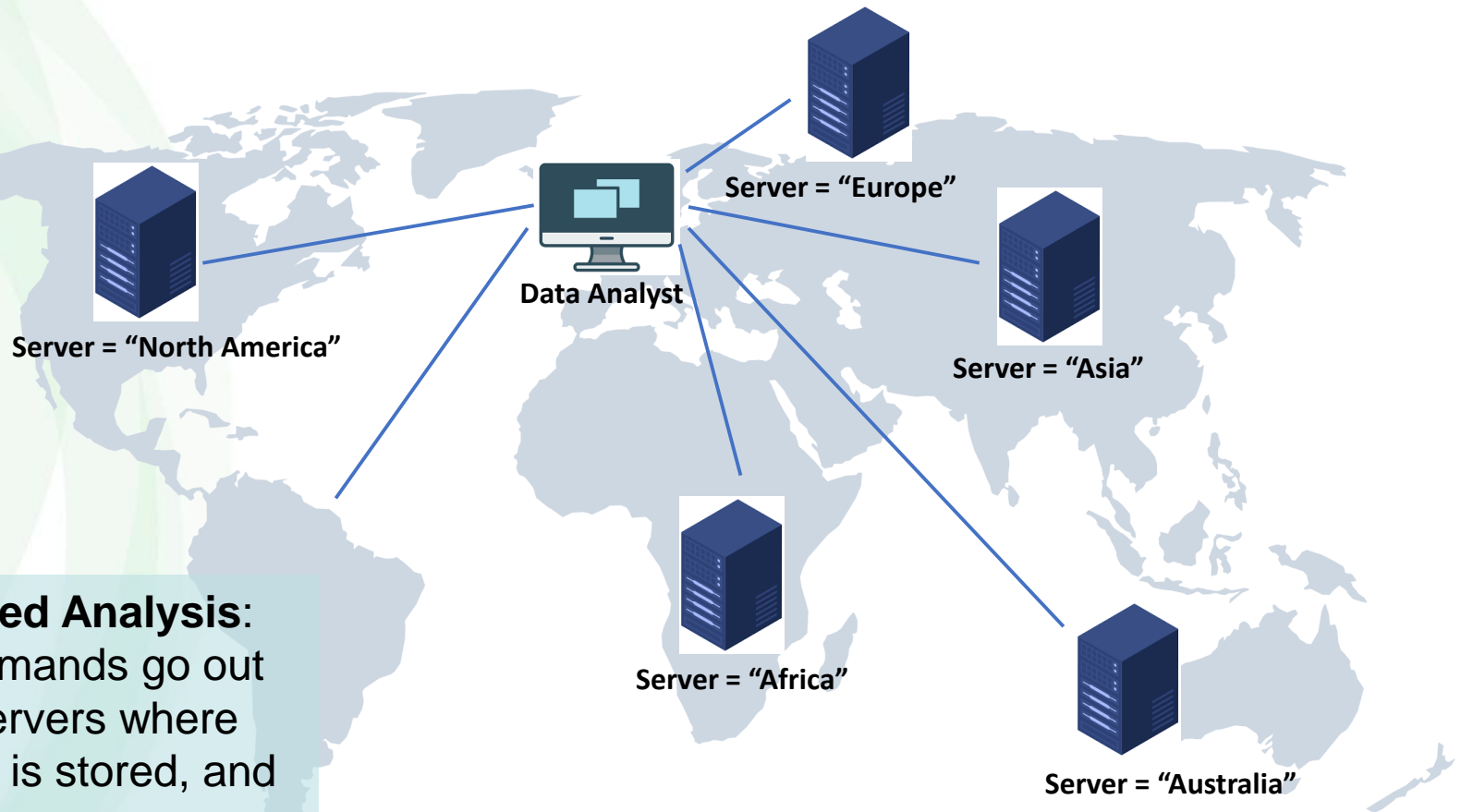


Privacy-preserving



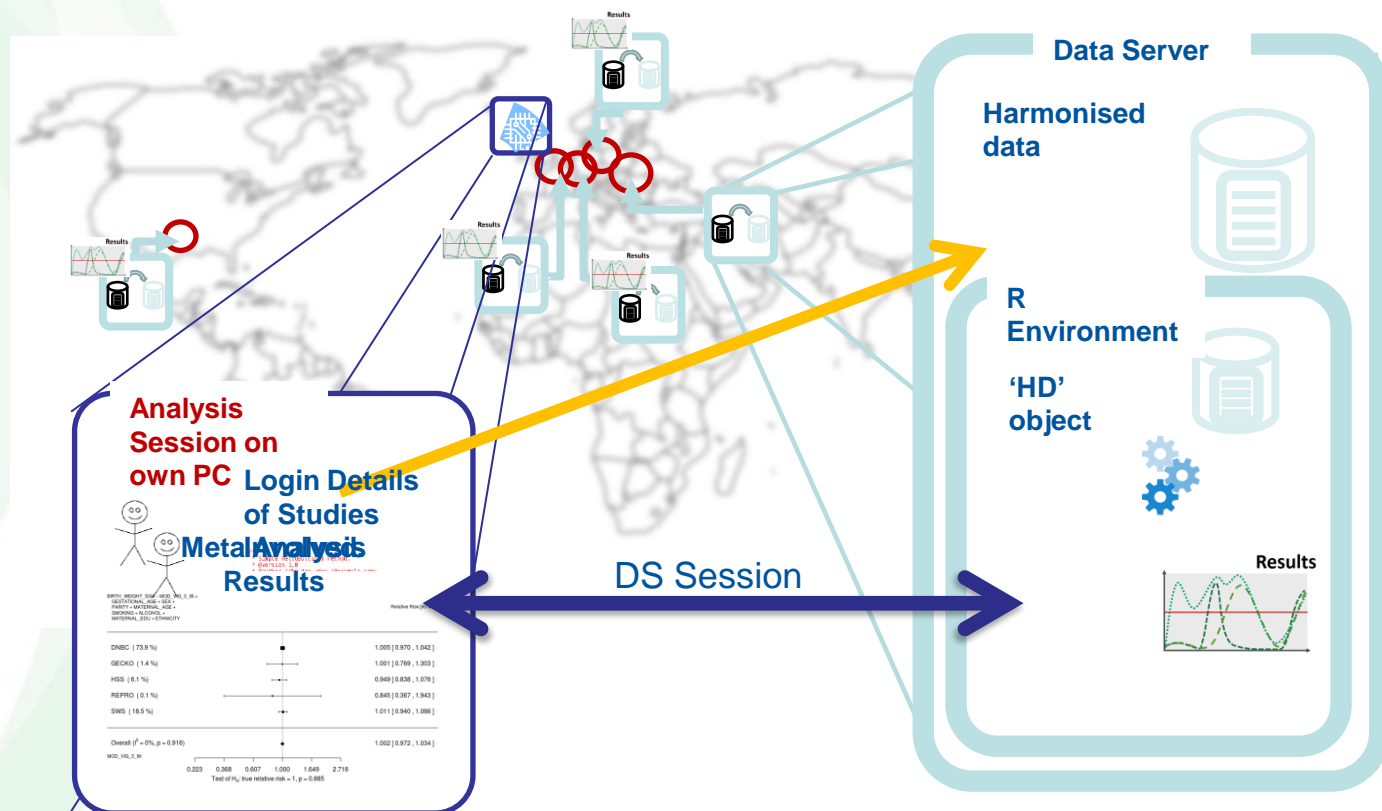
- ☐ Functions are sent to data server
- ☐ Server runs function
- ☐ Built into function is disclosure check
- ☐ If disclosive, check will discover and only return error message, not results.

Federated analysis



Federated Analysis:
the commands go out
to the servers where
the data is stored, and
run there.

Animation of a DataSHIELD Analysis



What kind of analysis can we do?

- ❑ There are 120 functions within the DataSHIELD clientside package to choose from, in version 6.3 (released November 2024)

- ❑ A full list can be found on this wiki:

<https://data2knowledge.atlassian.net/wiki/spaces/DSEV/pages/1184825438/List+of+all+DataSHIELD+functions+v6.3>

- ❑ There are external developers creating new packages (ggplot, omics, mediation, tydiverse, ...)

Methodological papers

JOURNAL ARTICLE









DataSHIELD: mitigating disclosure risk in a multi-site federated analysis platform

Demetris Avraam , Rebecca C Wilson, Noemi Aguirre Chan, Soumya Banerjee, Tom R P Bishop, Olly Butters, Tim Cadman, Luise Cedervik, Liesbeth Duijts, Xavier Escribà Montagut ... [Show more](#)

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD

Yannick Marcon ^{1*}, Tom Bishop ², Demetris Avraam ³, Xavier Escriba-Montagut ^{4,5}, Patricia Ryser-Welch ³, Stuart Wheeler ⁶, Paul Burton ³, Juan R. González ^{4,5,7,8*}


1 Epigeny, St Ouen, France, **2** MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom, **3** Population Health Sciences Institute, Newcastle University, Newcastle, United Kingdom, **4** Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain, **5** Universitat Pompeu Fabra (UPF), Barcelona, Spain, **6** Arjuna Technologies, Newcastle, United Kingdom, **7** Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain, **8** Dept. of Mathematics, Universitat Autònoma de Barcelona (UAB), Bellaterra (Barcelona), Spain

* yannick.marcon@obiba.org (YM); juanr.gonzalez@isglobal.org (JRG)



JOURNAL ARTICLE

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

Xavier Escribà-Montagut, Yannick Marcon, Demetris Avraam, Soumya Banerjee, Tom R P Bishop, Paul Burton, Juan R González 

PLOS COMPUTATIONAL BIOLOGY


RESEARCH ARTICLE

Federated privacy-protected meta- and mega-omics data analysis in multi-center studies with a fully open-source analytic platform

Xavier Escriba-Montagut ^{1,2}, Yannick Marcon ³, Augusto Anguita-Ruiz ^{1,2}, Demetris Avraam ⁴, Jose Urquiza ^{1,2,5}, Andrei S. Morgan ^{6,7}, Rebecca C. Wilson ⁴, Paul Burton ⁸, Juan R. Gonzalez ^{1,2,5*}

BMC Research Notes

[Home](#) [About](#) [Articles](#) [Submission Guidelines](#) [Collections](#) [Join The Editorial Board](#)

[Submit manuscript](#) 

Research Note | [Open access](#) | Published: 06 June 2023

dsSurvival 2.0: privacy enhancing survival curves for survival models in the federated DataSHIELD analysis system

[Soumya Banerjee](#)  & [Tom R. P. Bishop](#)

Use of DataSHIELD in science



Environment International
Volume 190, August 2024, 108853



Full length article

Green spaces and respiratory, cardiometabolic, and neurodevelopmental outcomes: An individual-participant data meta-analysis of >35.000 European children

Amanda Fernandes^{a,b,c}, Demetris Avraam^{d,e}, Tim Cadman^{a,b,c,e}, Payam Dadvand^{a,b,c}, Mònica Guxens^{a,b,c,f}, Anne-Claire Binter^{a,b,c}, Angela Pinot de Mouro^{a,g}, Mark Nieuwenhuijsen^{a,b,c}, Liesbeth Duijts^{h,i}, Jordi Julvez^{a,b,c,j}, Montserrat De Castro^{a,b,c}, Serena Fossatti^{a,b,k}, Sandra Márquez^{a,b,k}, Tanja Vrijkotte^{k,l,m}, Ahmed Elhakeem^{n,o}, Rosemary McEachan^r, Tiffany Yang^s, Marie Pedersen^t, Johan Vinther^u, Johanna Lepeule^v, Martine Vrijheid^{a,b,c}

Practice of Epidemiology

Associations of Maternal Educational Level, Proximity to Green Space During Pregnancy, and Gestational Diabetes With Body Mass Index From Infancy to Early Adulthood: A Proof-of-Concept Federated Analysis in 18 Birth Cohorts

Tim Cadman^a, Ahmed Elhakeem, Johan Lerbech Vinther, Demetris Avraam, Paula Carrasco, Lucinda Calas, Marios Cardol, Marie-Aline Charles, Eva Corpeleijn, Sarah Crozier, Montserrat de Castro, Marisa Estarlich, Amanda Fernandes, Serena Fossatti, Dariusz Gruszfeld, Kathrin Guerlich, Veit Grote, Sido Haakma, Jennifer R. Harris, Barbara Heude, Rae-Chi Huang, Jesús Ibarluzea, Hazel Inskip, Vincent Jaddoe, Berthold Koletzko, Sandrine Lioret, Verónica Luque, Yannis Manios, Giovenale Moirano, George Moschonis, Johanna Nader, Mark Nieuwenhuijsen, Anne-Marie Nybo Andersen, Rosie McEachan, Angela Pinot de Mouro, Maja Popovic, Theano Roumeliotaki, Theodosia Salika, Loreto Santa Marina, Susana Santos, Sylvain Serbert, Evangelia Tzorovili, Marina Vafeiadi, Elvira Verduci, Martine Vrijheid, T. G. M. Vrijkotte, Marieke Welten, John Wright, Tiffany C. Yang, Daniela Zugna, and Deborah Lawlor

PLOS MEDICINE

nature communications

Explore content ▾ About the journal ▾ Publish with us ▾

nature > nature communications > articles > article

Article | [Open access](#) | Published: 03 May 2023

Identification of biomarkers for glycaemic deterioration in type 2 diabetes

Roderick C. Sliker, Louise A. Donnelly, Elina Akalestou, Livia Lopez-Noriega, Rana Melhem, Ayşim Güneş, Frederic Abou Azar, Alexander Efanov, Eleni Georgiadou, Hermine Muniangi-Muhitu, Mahsa Sheikh, Giuseppe N. Giordano, Mikael Åkerlund, Emma Ahlqvist, Ashfaq Ali, Karina Banasik, Søren Brunak, Marko Barovic, Gerard A. Boulard, Frédéric Burdet, Mickaël Canouil, Julian Dragan, Petra J. M. Elders, Celine Fernandez, ... Guy A. Rutter [+ Show authors](#)

OPEN ACCESS

RESEARCH ARTICLE

Gestational age at birth and body size from infancy through adolescence: An individual participant data meta-analysis on 253,810 singletons in 16 birth cohort studies

Johan L. Vinther^{a,1,4}, Tim Cadman^{a,2}, Demetris Avraam^{a,3}, Claus T. Ekstrøm^{a,4}, Thorkild I. A. Sørensen^{a,1,5}, Ahmed Elhakeem^{a,2}, Ana C. Santos^{a,6,7}, Angela Pinot de Mouro^{a,1}, Barbara Heude^{a,8}, Carmen Iñiguez^{a,10,11}, Costanza Pizzi^{a,12}, Elinor Simons^{a,13,14}, Ellis Voerman^{a,15,16}, Eva Corpeleijn^{a,17}, Faryal Zariouh^{a,18}, Gillian Santorelli^{a,19}, Hazel M. Inskip^{a,20,21}, Henrique Barros^{a,6,7}, Jennie Carson^{a,22,23}, Jennifer R. Harris^{a,24}, Johanna L. Nader^{a,25}, Justina Ronkainen^{a,26}, Katrine Strandberg-Larsen^{a,1}, Loreto Santa-Marina^{a,10,27,28}, Lucinda Calas^{a,8}, Luise Cederkvist^{a,1}, Maja Popovic^{a,12}, Marie-Aline Charles^{a,18}, Marieke Welten^{a,15,16}, Martine Vrijheid^{a,10,29,30}, Meghan Azad^{a,13,31,32}, Padmaja Subbarao^{a,33,34,35}, Paul Burton^{a,36}, Puishkumar J. Mandhane^{a,36}, Rae-Chi Huang^{a,22,37}, Rebecca C. Wilson^{a,38}, Sido Haakma^{a,39}, Silvia Fernández-Barrés^{a,29,30,40}, Stuart Turvey^{a,41}, Susana Santos^{a,15,16}, Suzanne C. Tough^{a,42}, Sylvain Serbert^{a,26}, Theo J. Moraes^{a,33}, Theodosia Salika^{a,21}, Vincent W. V. Jaddoe^{a,15,16}, Deborah A. Lawlor^{a,2,43}, Anne-Marie Nybo Andersen^{a,1}



The Lancet Regional Health - Europe
Volume 45, October 2024, 101036



Articles

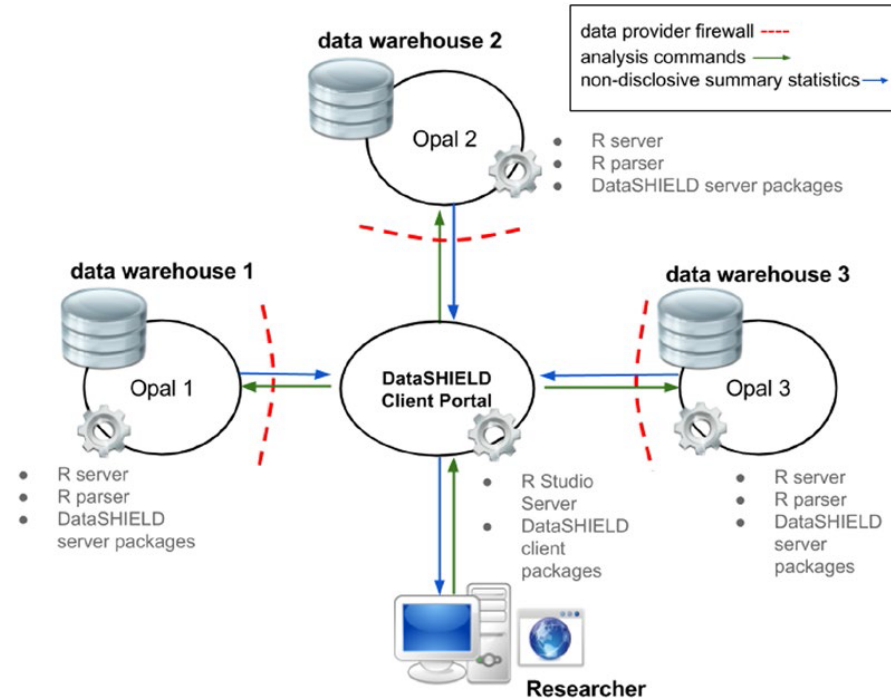
Early childcare arrangements and children's internalizing and externalizing symptoms: an individual participant data meta-analysis of six prospective birth cohorts in Europe

Katharine M. Barry^{a,b}, Demetris Avraam^c, Tim Cadman^c, Ahmed Elhakeem^d, Hanan El Marroun^{e,f}, Pauline W. Jansen^{a,f}, Anne-Marie Nybo-Andersen^c, Katrine Strandberg-Larsen^c, Lúcia González Safont^{g,h}, Raquel Soler-Blasco^{i,j,k}, Florencia Barreto-Zarza^{b,k,t}, Jordi Julvez^{l,u,j}, Martine Vrijheid^{l,u,j}, Barbara Heude^{m,n}, Marie-Aline Charles^{a,p}, Alexandre Ramchandarr Gomajee^{a,b,q}, Maria Melchior^{b,r}

DataSHIELD standard platform

Preliminary: data located at Opal server

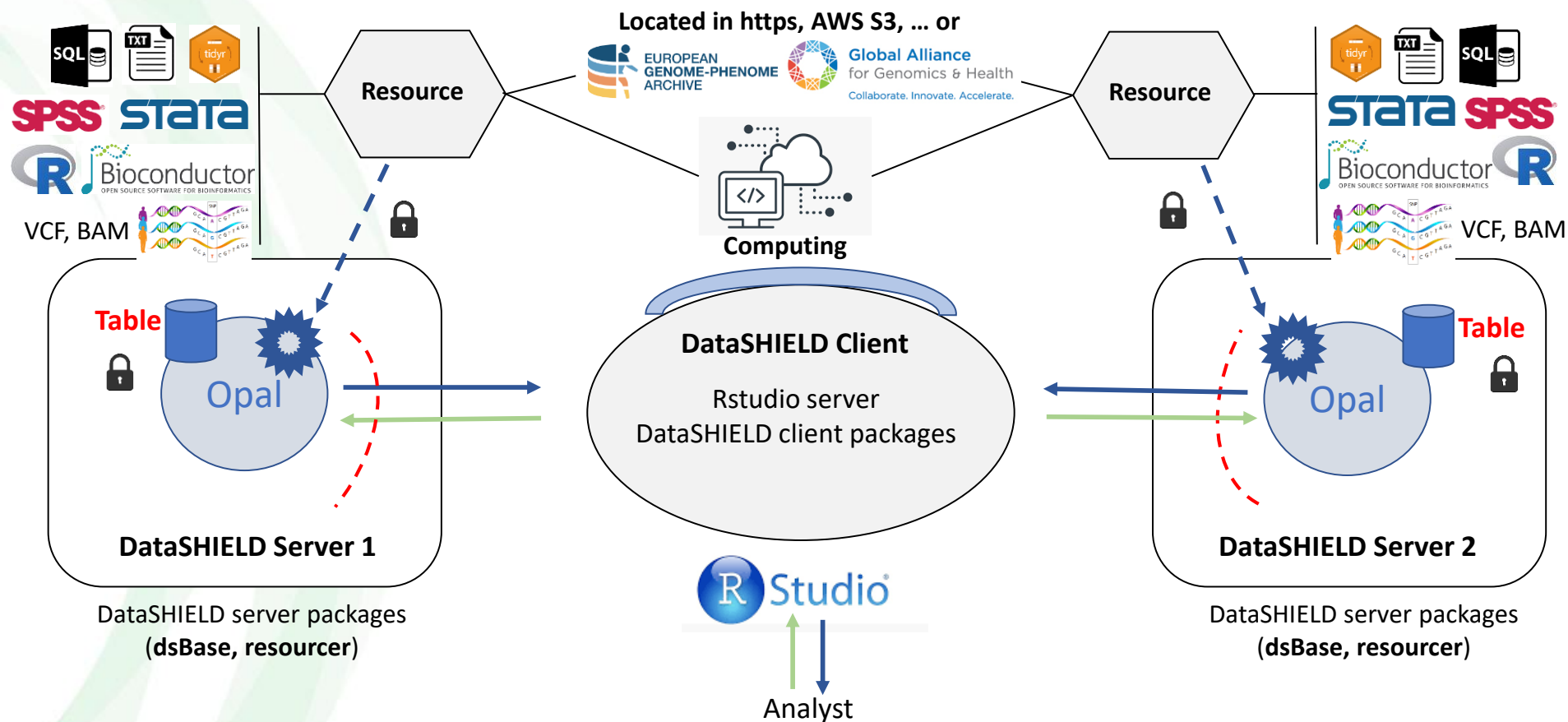
- ❑ Authenticate and authorize user
- ❑ Assign Opal table into the R server (data transfer)
- ❑ Execute DataSHIELD-verified R commands in R server



DataSHIELD standard platform

Open Opal demo server

The resources: idea



The resources: rational

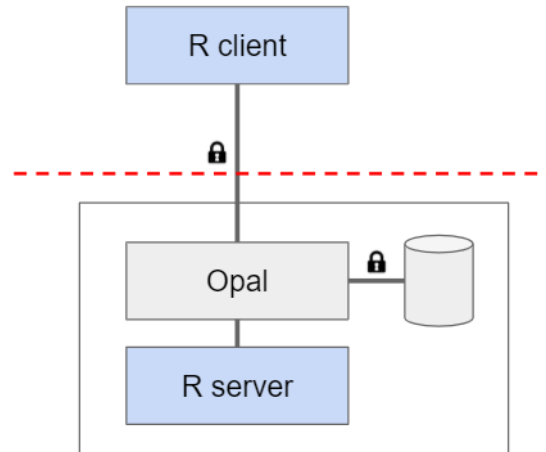
- ☐ use data at their original location (do not move/copy data when possible)
- ☐ use data in their original format (no pre-processing, no loss of information)
- ☐ use external computation facilities (no R limitations)
- ☐ => use DataSHIELD with large/big datasets (omics etc.)

The resources ...

Opal with tables

DataSHIELD client

DataSHIELD server 1..n

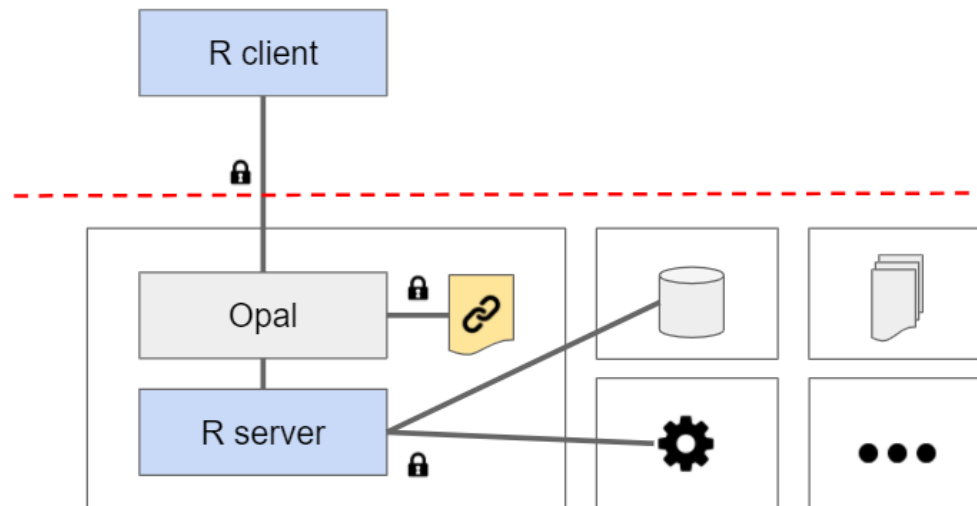


Opal with resources

DataSHIELD client

DataSHIELD server 1..n

- Databases
- File repositories
- Big data analytics systems
- etc.



Resources examples

- ☐ CSV file (compressed or not)
- ☐ SQL database table
- ☐ R object stored in a R data file
- ☐ HL7 FHIR dataset
- ☐ GA4GH server
- ☐ AWS server ([OMOP database](#))
- ☐ HPC server accessible through SSH
- ☐ Python script
- ☐ Big data analytics system (Apache Spark, Dremio, ...)
- ☐ Apps in docker images
- ☐ ...

The resources ...

Property	Description	Examples
url	Location of the resource	<code>https://example.org/some/file.rda</code> <code>file://path/to/file.csv</code> <code>ssh://example.org/work/dir?exec=plink</code> <code>mysql://dbhost:3306/mydb/mytable</code>
format	Data format (optional)	<code>SPSS</code> <code>ExpressionSet</code>
credentials	Data access (optional)	<code>token=Q3sDdsWq2dsx7</code> <code>username=user1 password=xxxxxx</code>

Not visible by DataSHIELD users

DataSHIELD disclosure controls

- ❑ Ensures no individual-level data leaves the data holder's server.
- ❑ Analyses are performed remotely; only non-disclosive summaries are returned.
- ❑ Protects participant privacy while enabling collaborative research.
- ❑ Balances data utility with confidentiality.

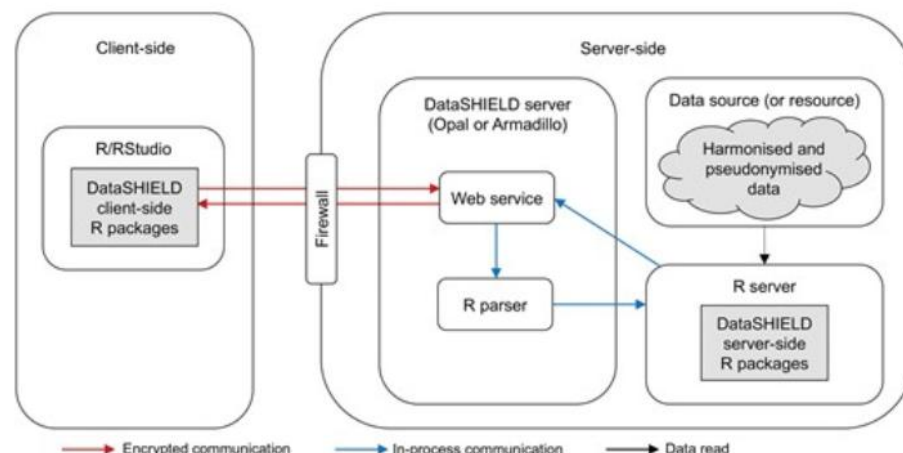
JOURNAL ARTICLE

DataSHIELD: mitigating disclosure risk in a multi-site federated analysis platform

Demetris Avraam , Rebecca C Wilson, Noemi Aguirre Chan, Soumya Banerjee, Tom R P Bishop, Olly Butters, Tim Cadman, Luise Cederkvist, Liesbeth Duijts, Xavier Escribà Montagut ... [Show more](#)

Bioinformatics Advances, Volume 5, Issue 1, 2025, vbaf046,

<https://doi.org/10.1093/bioadv/vbaf046>



DataSHIELD: mitigation disclosure risks

- ☐ System protection elements
- ☐ Analysis protection elements
- ☐ Governance protection elements

- ☐ Five Safe Framework

DataSHIELD: system protection elements

☐ Network security:

- ☐ Traffic encrypted.
- ☐ Connection requires SSL/TLS certificate.
- ☐ Firewall protection

☐ User authentication and certification:

- ☐ R server is only callable via middleware hosting DataSHIELD.

☐ R parser:

- ☐ Analyses are only able to use R server-side functions not native R

☐ Data management:

- ☐ Data comprise a snapshot not a live data (resources)

DataSHIELD: analysis protection elements

- ❑ **Only use assign and aggregate functions:**
 - ❑ Only create objects in the server-side.
 - ❑ Generate summarized data to the client-side.
- ❑ **Implementation restriction:**
 - ❑ R functions are not implemented (*print*, *max*, ...)
- ❑ **Disclose controls:**
 - ❑ Function to provide non-disclosive results (each package can have their own functions)
- ❑ **Data level obfuscation:**
 - ❑ Anonymization, data synthesis techniques, ...
- ❑ **DataSHIELD log files:**
 - ❑ log files visible for data custodians

DataSHIELD: examples of disclose controls

Name	Description
<code>nfilter.tab</code>	Prevents the return of a contingency table if any of its cells represents less than <i>nfilter.tab</i> observations. The value of <i>nfilter.tab</i> can be set to any non-negative integer. The default value is set to 3.
<code>nfilter.subset</code>	Prevents the creation of a dataset's subset if the subset has less than <i>nfilter.subset</i> rows. The value of <i>nfilter.subset</i> can be set to any positive integer. The default value is set to 3.
<code>nfilter.glm</code>	Prevents the fitting of a regression model that has more than <i>nfilter.glm</i> \times N unknown parameters in a dataset with sample size N . The value of <i>nfilter.glm</i> can be set to any numeric value in the interval (0,1). The default value is set to 0.33.
<code>nfilter.string</code> , <code>nfilter.stringShort</code>	Blocks the evaluation of a string argument that passes from the client-side to the server-side, if it has a length greater than <i>nfilter.string</i> or <i>nfilter.stringShort</i> characters. The values of <i>nfilter.string</i> and <i>nfilter.stringShort</i> can be set to any positive integers. The default values are set to 80 and 20, respectively.
<code>nfilter.levels.density</code>	Prevents the return of the unique levels of a categorical variable if their length is more than <i>nfilter.levels.density</i> \times N where N is the length of the vector of the categorical variable. The value of <i>nfilter.levels.density</i> can be set to any numeric value in the interval (0,1). The default value is set to 0.33.

DataSHIELD: governance protection elements

- ☐ **Formal agreements:**

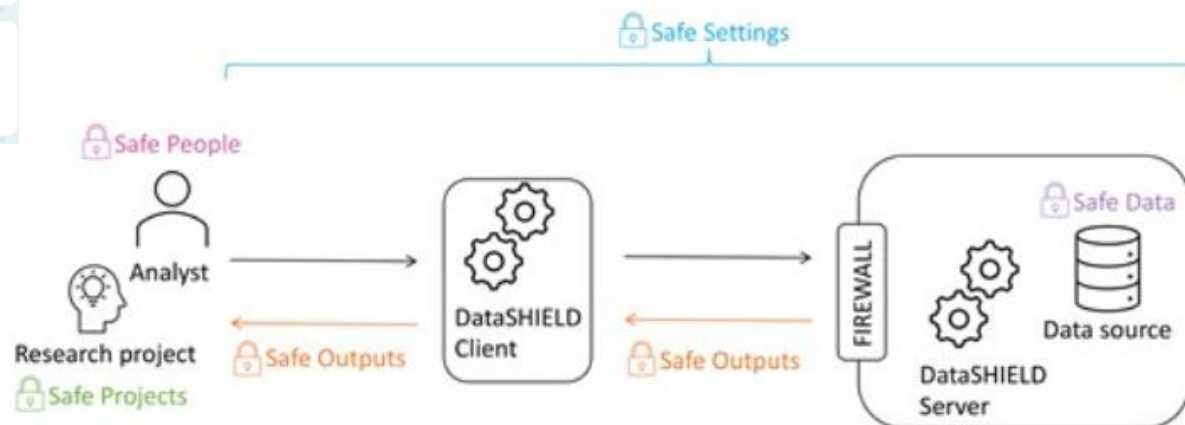
- ☐ Data access, data analysis and approval of research results prior to publication.

- ☐ **User can have access to:**

- ☐ Different tables or resources
- ☐ Different variables from a table of resource
- ☐ Different data analysis packages

- ☐ It is controlled through the ‘profiles’

DataSHIELD and the Five Safes Framework



DataSHIELD and the Five Safes Framework

Five Safes	Study mitigation	DataSHIELD mitigation
Safe People	<ul style="list-style-type: none">• Formal data access request process• Due diligence on prospective users —‘are they bona fide researchers? Have they conducted mandatory training/accreditation to work with data safely?’• Legal contracts, signed terms and conditions of data access and use• Sanctions policy• DataSHIELD users are authorized for data access by the study	<ul style="list-style-type: none">• The authorization to access DataSHIELD can be delegated, under the principle of subsidiarity to individual studies

DataSHIELD and the Five Safes Framework

Five Safes	Study mitigation	DataSHIELD mitigation
Safe projects	<ul style="list-style-type: none">• Assess the requirement for access to the data—the context of the research project or the data being used• Ensuring the data access/use does not contradict any necessary legal requirements, e.g. study consent	

DataSHIELD and the Five Safes Framework

Five Safes	Study mitigation	DataSHIELD mitigation
Safe settings	<ul style="list-style-type: none">• The operation and maintenance of robust computing infrastructure and hardware• Following IT security best practice• Log of registered users, maintaining/blocking user access• Preventative measures for unauthorized access• Deploy DataSHIELD to securely transfer information (analysis commands and outputs) via https• Each study/consortium provides unique authentication credentials for users to log onto the DataSHIELD client and to connect to each study they are authorized for	<ul style="list-style-type: none">• DataSHIELD is a client-server architecture, the user does not connect directly to the study data• Analysis of individual level data occurs server-side (at study)• Analysis environment server-side can only be called via authenticated users through Opal or Armadillo• The server-side R Parser prevents invalid characters or non-approved functions from being run.• Users can not directly view the individual-level data• User commands logged server-side, only accessible by the study. Can be manually scrutinized e.g. if data misused

DataSHIELD and the Five Safes Framework

Five Safes	Study mitigation	DataSHIELD mitigation
Safe data	<ul style="list-style-type: none">• Data Protection Impact Assessment (risk assessment)• Assessing the disclosure risk of the data• Pseudonymized data used to lower the disclosure risk	

DataSHIELD and the Five Safes Framework

Five Safes	Study mitigation	DataSHIELD mitigation
Safe outputs	<ul style="list-style-type: none">• Manual checking of analysis outputs for disclosure• Legal contracts or terms and conditions making it mandatory for the user to check their own outputs for disclosure before publishing• DataSHIELD disclosure setting thresholds are set and maintained by the study, no one else can alter these	<ul style="list-style-type: none">• DataSHIELD server-side functions prevent directly disclosive outputs being returned to the analyst• DataSHIELD server-side functions prevent viewing of individual-level data• DataSHIELD has disclosure settings based on established statistical disclosure control methods to conduct automated checks for direct disclosure in outputs