DataSHIELD: A method for privacy protected analysis of individual-level data

Demetris Avraam[1], Rebecca Wilson[1], Andrew Turner[1], Madeleine Murtagh[1], Oliver Butters[1], Tom Bishop[2], Vincent Ferretti[3], Yannick Marcon[3], Paul Burton (PI)[1]

[1]Data to Knowledge (D2K) Research Group, School of Social and Community Medicine, University of Bristol, UK
[2]MRC Epidemiology Unit, University of Cambridge, UK
[3]Maelstrom Research Group, Research Institute of the McGill University Health Centre, McGill University, Montreal, Canada

email: datashield-research@bristol.ac.uk website: www.datashield.ac.uk

The DataSHIELD infrastructure

DataSHIELD is an open source software that allows a non-disclosive analysis of sensitive individual-level data and a simultaneous co-analysis of such data from several studies without physically sharing the data. DataSHIELD provides an effective solution for data analysis in several disciplines such as in biomedicine and social sciences where the accessibility and sharing of data is often prohibited by ethico-legal or governance constraints and other barriers such as the control of intellectual property or very large datasets.

The infrastructure of DataSHIELD enables to coordinate a parallelised analysis for microdata from different sources (Figure 1). Each study stores their data locally in Opal servers behind their firewall. Alongside Opal, the R software with the DataSHIELD server-side R functions and an R parser is installed. The analyst uses the R software in an analysis computer (client-side) where the DataSHIELD client-side R functions are installed. The analysis computer sends commands to the data computers and those commands are authorised by the parser which controls which of them are allowed to run on the servers. Each server-side conducts then, a separate analysis on their individual-level data and returns non-disclosive results to the client-side where the results from all the considered studies are finally combined appropriately to provide a "virtually pooled" analysis.

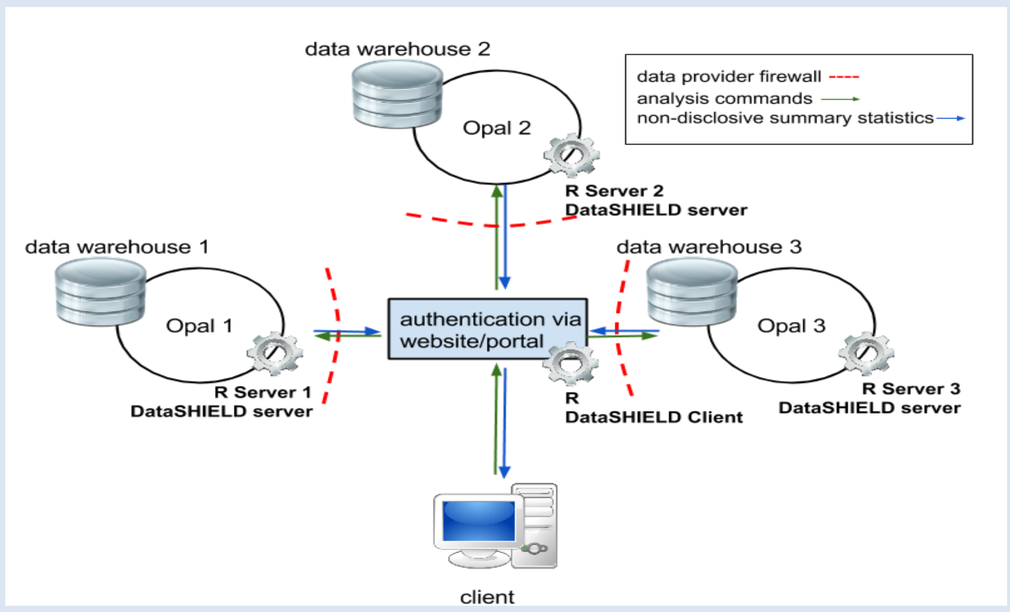

Figure 1: *The multi-site DataSHIELD infrastructure*

Under DataSHIELD, individual-level data remain always invisible and unobtainable externally and only non-disclosive summary statistics are returned to the analyst. The summary statistics are assessed by a number of statistical disclosure controls prior to their release. DataSHIELD is therefore an efficient approach for secure individual participant data analysis (single-site or multi-site) or study-level meta-analysis where different studies hold the same variables on different individuals (Horizontal DataSHIELD) and secure analysis of linked data where different studies hold different variables on the same individuals (Vertical DataSHIELD).

The DataSHIELD methodology

DataSHIELD includes functionality for descriptive statistics (i.e. 1D and 2D contingency table, mean, variance, covariance, correlation, quantiles, etc.), regression analysis (i.e. generalised linear modelling, survival analysis using piecewise exponential regression, etc.) and graphical illustrations (i.e. histogram, heat map plot, contour plot, etc.). The analysis through DataSHIELD is distinguished between one step analysis and iterative analysis. In one step analysis the summary statistics that are returned from the server-side to the client-side provide enough information for the calculation of combined summary statistics for the data of all corresponding sources. In iterative analysis, summary statistics are returned from the server to the client, then calculations take place at the client and some information send back to the servers and the procedure iterates until the convergence of the solution.

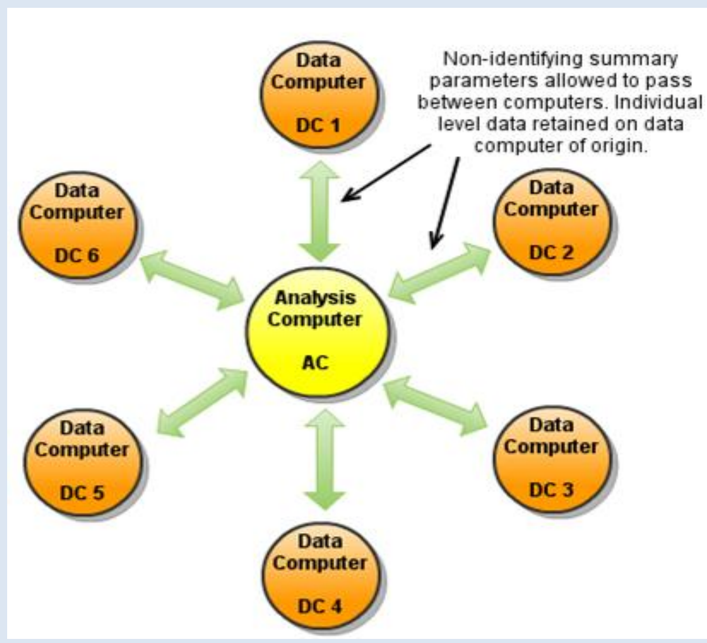

Figure 2: *Communication between the client-side and the servers*

An example of a function that performs a one step analysis in DataSHIELD, is the ds.mean function that calculates the mean of a variable from a single or from multiple sources. For a multi-site analysis, the sum of the values of the variable and the length of the variable is returned from the server-side of each study to the client-side where the weighted mean is calculated. An example of a function that performs an iterative analysis in DataSHIELD, is the ds.glm function that fits a generalised linear model to data from a single or from multiple sources. For a multi-site analysis, an initial vector of the model parameters (either specified by the user or set to be zero for all parameters by default) is send firstly from the client-side to the server-side of each study. Then, a score vector and an information matrix is calculated and send from each server to the client-side and a new vector of parameters is estimated at the client-side and send back to each server. The procedure iterates until the convergence of the solution. This iterative process is mathematically equivalent to an individual-level regression analysis if the data from different sources are physically pooled together in one central warehouse.

References:

Gaye A, et al, (2014). DataSHIELD: taking the analysis to the data, not the data to the analysis. International Journal of Epidemiology.

Jones EM, et al, (2012). DataSHIELD – shared individual-level analysis without sharing data: a biostatistical perspective. Norwegian Journal of Epidemiology 21(2):231-239.

Wolfson M, et al, (2010). DataSHIELD: resolving a conflict in contemporary bioscience–performing a pooled analysis of individual-level data without sharing the data. Int. Journal of Epidemiology 39(5):1372-1382.

Budin-Ljøsne I, et al, (2015). DataSHIELD: An Ethically Robust Solution to Multiple-Site Individual-Level Data Analysis. Public Health Genomics, 18:87-96.

Wallace SE, et al, (2014). Protecting Personal Data in Epidemiological Research: DataSHIELD and UK Law. Public Health Genomics, 17:149-157.

Murtagh MJ, et al, (2012). Securing the data economy: Translating privacy and enacting security in the development of DataSHIELD. Public Health Genomics. 15:243-253.

The project is supported by: