Montreal Bioinformatics Users Group Meetup

DataSHIELD: R library that enables the remote and non-disclosive analysis of sensitive research data

Lead by Professor Paul Burton, University of Bristol.

mælstrom

R

DataSHIELD
Secure Bioscience Collaboration

DataSHIELD (Data Aggregation Through **Anonymous Summary-statistics** from **Harmonised** Individual levEL Databases)

...uses distributed computing and parallelized analyses to enable full joint analyses of individuel level data from differents sources without the need for those data to move, or even be seen, outside the study were they usually reside.

Co-analyses principles

Very large sample sizes are required for the estimation of effects which are known to be small. This is particularly so for genetic epidemiology studies where associations between individual genetic variant and phenotypes of interest are generally weak.

Sample sizes that are sufficiently large for adequalely powered analysis are often only achievable py pooling data from multiple studies.

Three differents principles:

- Individual level meta-analysis(ILMA): Data are pooled as a single database
- Study level meta-analysis (SLMA) : most common approach of meta-analysis
- Federated analysis: DataSHIELD

Individual Level Meta Analysis (ILMA) :

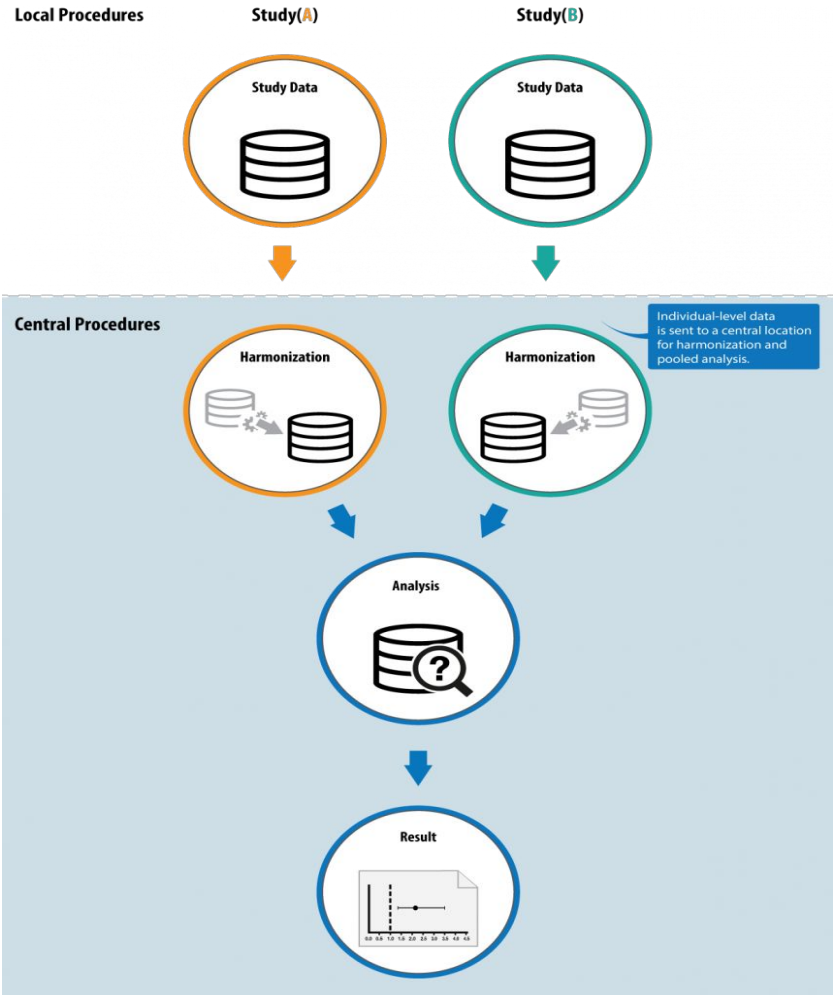
Data from every individual in each participating study are combined into one large dataset and analysed as a single study (allowing for study-to-study heterogeneity)

Pros:

➔ Great flexibility in statistical analysis

Cons:

➔ Highly disclosive to subject identity, raises ethical, legal and social issues.
➔ Not effective to move around huge genomic data.



Study Level Meta Analysis (SLMA):

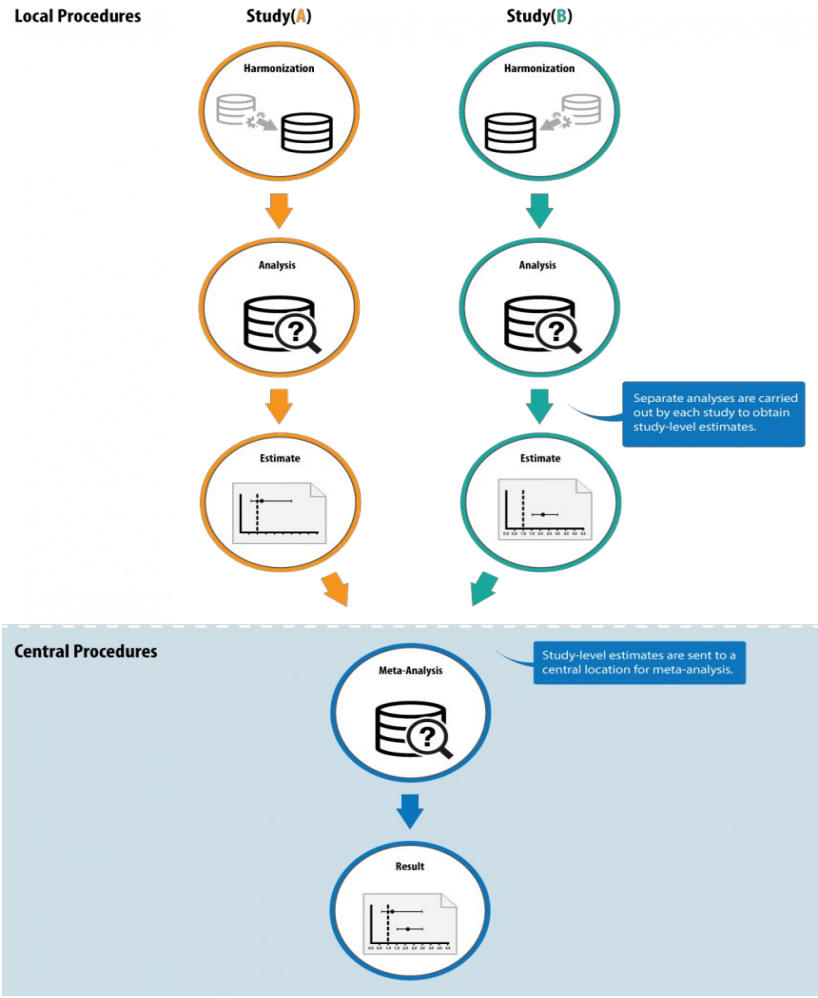
Each study performs analysis on its own data, and shares the resulting association statistics. A meta-analysis is then conducted on the these study-level statistics to obtain associational estimates accross all studies combined.

Pros:

- ➔ Consistent with ethico-legal restrictions
- ➔ Possible solution for datasharing for simple analysis, for which analysis can be pre-planned.

Cons:

- ➔ Depends on the summary statistics already available from studies. New research questions cannot be addressed until additional information is obtained and extracted from original studies



Federated analysis:

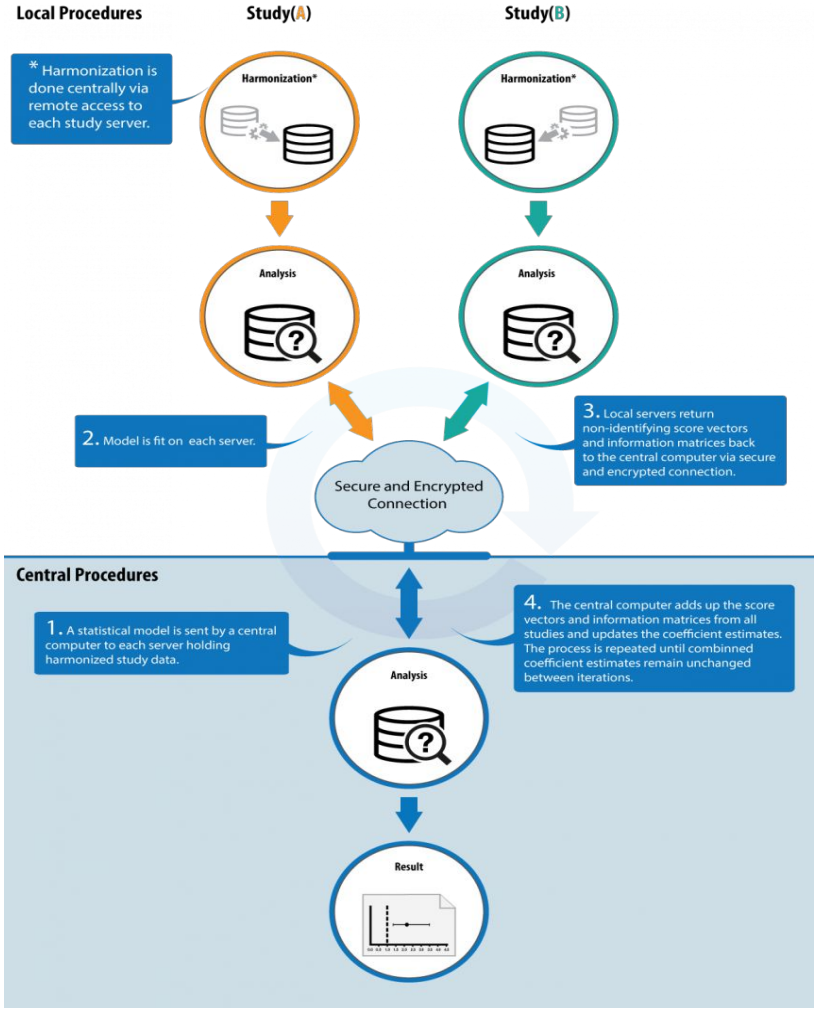
The ‘analysis computer’ (AC) send analysis instructions simultaneously to each ‘data computer’ (DC). Non disclosive low-dimensional summary statistics sent back by each DC are then combined by the AC to generate the estimate.

Pros:

- ➔ Data never move from their study location. Results are easily updatable as data is updated in each study.
- ➔ New research questions can be easily addressed.

Cons:

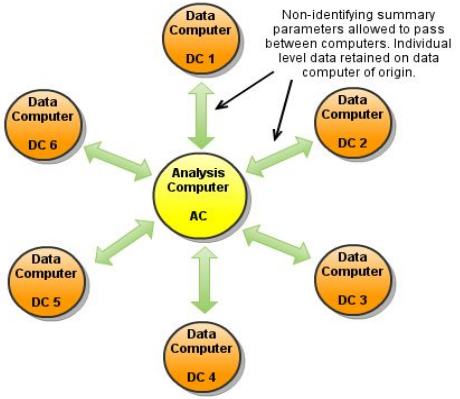
- ➔ Requires a supporting IT infrastructure.



IT infrastrusture supporting DataSHIELD

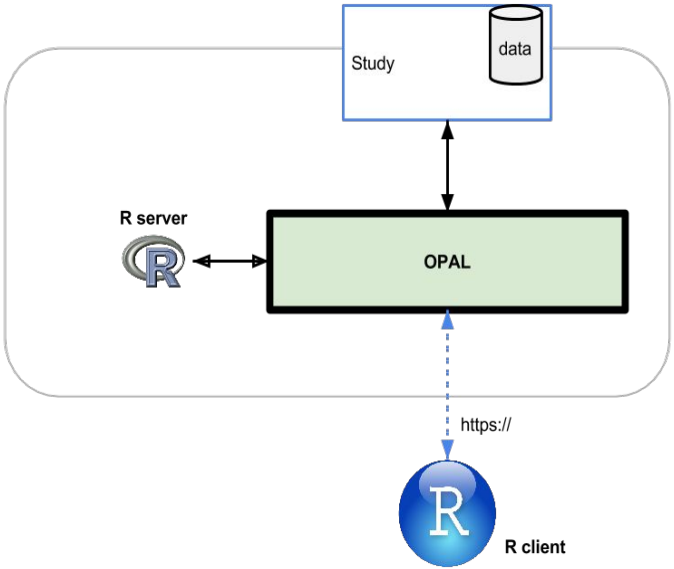
The implementation of DataSHIELD is made of 3 compenents:

- opal server(s)
- R server(s)
- an R client



The opal server component has several sub components necessary to implement DataSHIELD:

- a data and metadata module
- an R module
- a DataSHIELD module



Everything is open source

http://www.obiba.org/ *lead by Vincent ferretti, OICR*

ILMA VIA DataSHIELD

Mean of a variable:

Computing mean via DataSHIELD is straightforward.

let i = 1, …,s. The number of studies

N_i = number of observations in study *i*

M_i = mean of the variable in study *i*

$$Total\ \ mean\ =\ \frac{\sum_{i=1}^{s} M_i \times N_i}{\sum_{i=1}^{s} N_i}$$

ILMA VIA DataSHIELD

Total variance

$$Total\ mean = \frac{\sum_{i=1}^{s} M_i \times N_i}{\sum_{i=1}^{s} N_i}$$

Let i = 1,...,s. number of studies

N_i = *number of observations in study i*

V_i = *variance in study i*

$Mean_i$ = *mean in study i*

$$Variance_{total} = \frac{\sum_{i=1}^{s}(N_i - 1)\times V_i + \sum_{i=1}^{s}(Mean_i - Mean_{total})^2}{\sum_{i=1}^{s}(N_i) - 1}$$

ILMA VIA DataSHIELD

Fitting a Generalised Linear Model(GLM) using DataSHIELD

Let the following model to be fit $y_i = \beta X_i + \varepsilon$

Assume that the relationship between Y and X can be summarized by GLM, $\eta_i := g(\mu_i) = \beta^T x_i$

where η_i is a linear predictor, g is the link function specified, μ_i is the mean of y_i with $\mu=(\mu_1,\ldots,\mu_N)$

and $\beta^T = (\beta_1, \ldots, \beta_q)$ is the parameter we wish to estimate.

IRLS algorithm $$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + \boldsymbol{I}(\boldsymbol{\beta}_t)^{-1}\boldsymbol{s}(\boldsymbol{\beta}_t)$$

fisher information $$\boldsymbol{I}(\boldsymbol{\beta}_t) = \boldsymbol{X}^T\boldsymbol{W}_t\boldsymbol{X};$$

and score vector $$\boldsymbol{s}(\boldsymbol{\beta}_t) = \boldsymbol{X}^T\boldsymbol{W}_t(\boldsymbol{Y} - \boldsymbol{\mu}(t))g'(\boldsymbol{\mu}(t)),$$

$\boldsymbol{w}_t$ = diagonal matrix with diagonal entry $w_{ii}(t)^{-1} = V_i(t)g'(\mu_i(t))^2$

ILMA VIA DataSHIELD

Fitting a Generalised Linear Model(GLM) using DataSHIELD

Let N_j the number of observations in studies $j = 1,.....,s.$ the information matrix and the score vector can be extracted from each individual study

$$\boldsymbol{I}(\boldsymbol{\beta}_t) = \sum_{j=1}^{S}\sum_{i=1}^{N_j} w_{ii}(t)\,\boldsymbol{x}_i\boldsymbol{x}_i^T = \sum_{j=1}^{S}\boldsymbol{I}_j(\boldsymbol{\beta}_t),$$ and $$\boldsymbol{s}(\boldsymbol{\beta}_t) = \sum_{j=1}^{S}\sum_{i=1}^{N_j}(y_i - \mu_i(t))g'(\mu_i(t))w_{ii}(t)\boldsymbol{x}_i = \sum_{j=1}^{S}\boldsymbol{s}_j(\boldsymbol{\beta}_t).$$

The IRLS algorithm can still be used to estimate $\boldsymbol{\beta}$ from studies information matrix and score. The algorithm is iteratively updated until a convergence criteria is reached $$\frac{|D_r - D_{r-1}|}{D_r + 0.1} < 10^{-8}$$

where $D_{r,}$ is the deviance of the Maximum likehood estimate

DEMO

More informations

Maelstrom research

- https://www.maelstrom-research.org/

OBiBa resources

- http://www.obiba.org/
- github.com/obiba

DataSHIELD

- https://github.com/datashield
- http://www.datashield.ac.uk/

Publications

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

Wolfson M, Wallace SE, Masca N, Rowe G, Sheehan NA, Ferretti V, Laflamme P, Tobin MD, Macleod J, Little J, Fortier I, Knoppers BM, Burton PR. (2010). DataSHIELD: resolving a conflict in contemporary bioscience–performing a pooled analysis of individual-level data without sharing the data. *Int J Epidemiol*, 39(5):1372-1382.

Jones, EM, Sheehan, N, Masca, N, Wallace, S, Murtagh, MJ, Burton, PR.(2012). DataSHIELD – shared individual-level analysis without sharing data: a biostatistical perspective. *Norwegian Journal of Epidemiology.* 21 (2): 231-239.

Wallace SE, Gaye A, Shoush O, Burton PR. (2014). Protecting Personal Data in Epidemiological Research: DataSHIELD and UK Law. *Public Health Genomics*, 17:149-157.

Acknowledgments