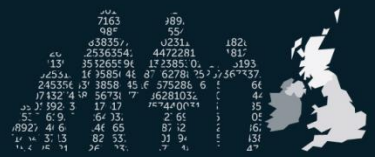


SYLLS

SYnthetic data estimation for uk LongitudinaL Studies

THE UNIVERSITY OF EDINBURGH

Running an analysis of combined data when the individual records cannot be combined:

practical issues in secure computation}

Gillian Raab, Chris Dibben, & Paul Burton

UNECE-Eurostat Work Session on Statistical Data Confidentiality, Helsinki , 2015

Administrative Data Research Network

An ESRC Data Investment

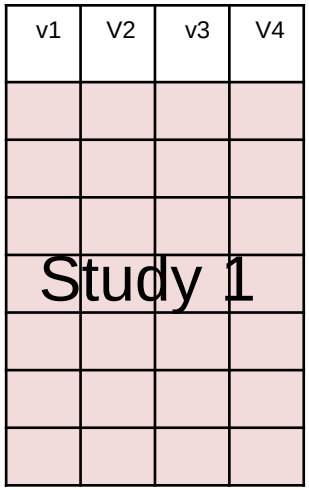
What is multi-party computation?

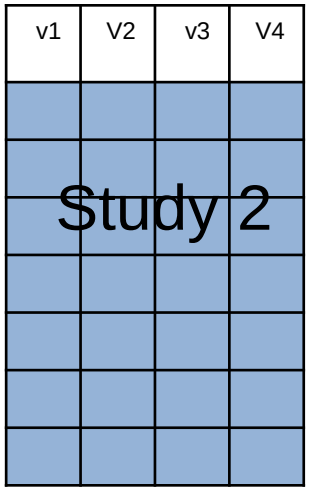
- The data to contribute to an analysis are distributed across several sites
- Confidentiality concerns prevent them being pooled for analysis
- Methods have been developed to allow analyses to be carried out without pooling the data

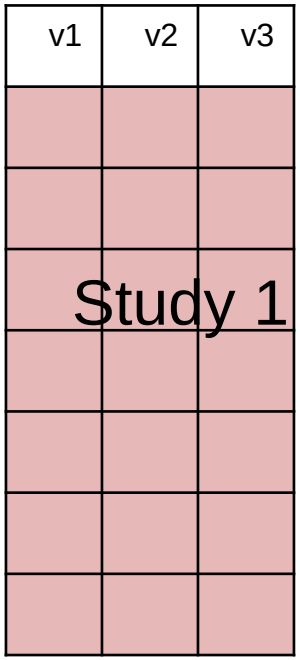
Methods developed for this

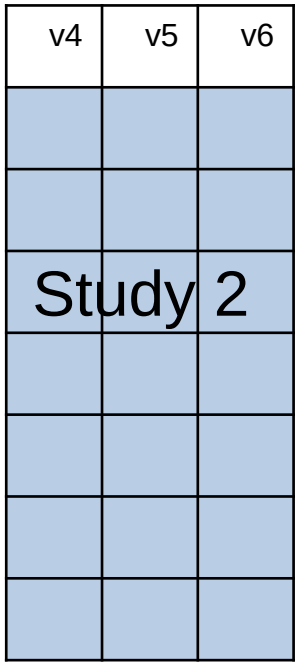
- Since around 2005
- By statisticians who call it the analysis of distributed data
- And computer scientists who call it privacy-preserving data mining (PPDM)

Horizontally or vertically partitioned data?









Horizontally partitioned data - examples

- Genomic data
- Adverse drug effects
- Comparable surveys or censuses collected by different agencies

The UK Longitudinal Studies

- Three studies each run by a different agency
  - England and Wales – (ONS LS)
  - Scottish Longitudinal Study – (SLS)
  - Norther Ireland Longitudinal study (NILS)

- Each run by a different National Agency
- Data needs to be held very securely (Census act)
- The servers that hold the data have no internet access

The UK Longitudinal Studies

- Three studies each run by a different agency
  - England and Wales – (ONS LS)
  - Scottish Longitudinal Study – (SLS)
  - Norther Ireland Longitudinal study (NILS)

- Each run by a different National Agency
- Data needs to be held very securely (Census act)
- The servers that hold the data have no internet access

Methods for multi-party computation

- Rely on exchanging and combining sufficient statistics from the different studies
- For horizontally partitioned data, the sufficient statistics for the pooled data are just the sum over those from the individual studies
- E.g. For regression inferences can be obtained from the sum of the information matrices (I)(X'X) and the score statistics (S) X'Y.
- Coefficients are $(\sum I)^{-1} \sum S$
- For iterative methods (e.g. GLMs) these quantities need to be exchanged at each iteration until the models converge.

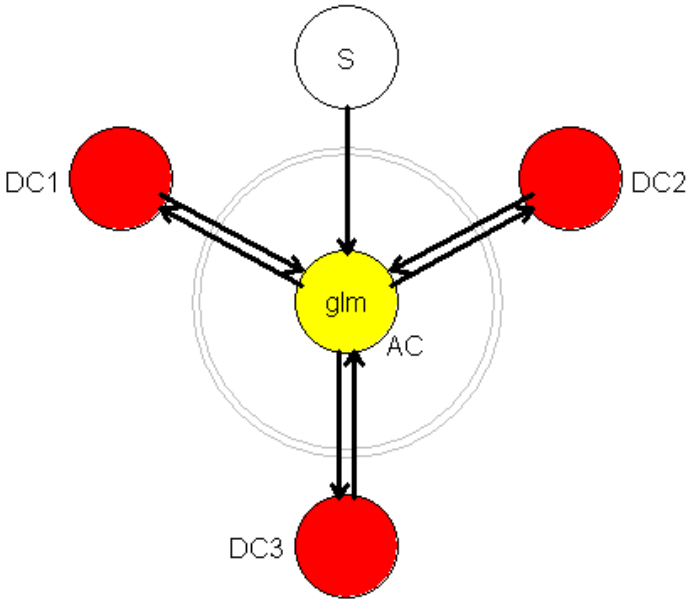
Iterative estimation for GLMs

- An analysis computer (AC) can control the estimation
- Data is held on several data computers (DCs)
- The following steps are needed
  - Starting values for the coefficients are sent to all DCs
  - Each DC returns their current I and S to the AC
  - The AC calculates a new value of the coefficients and checks if model converged
- These steps are repeated until convergence

The DataSHIELD project

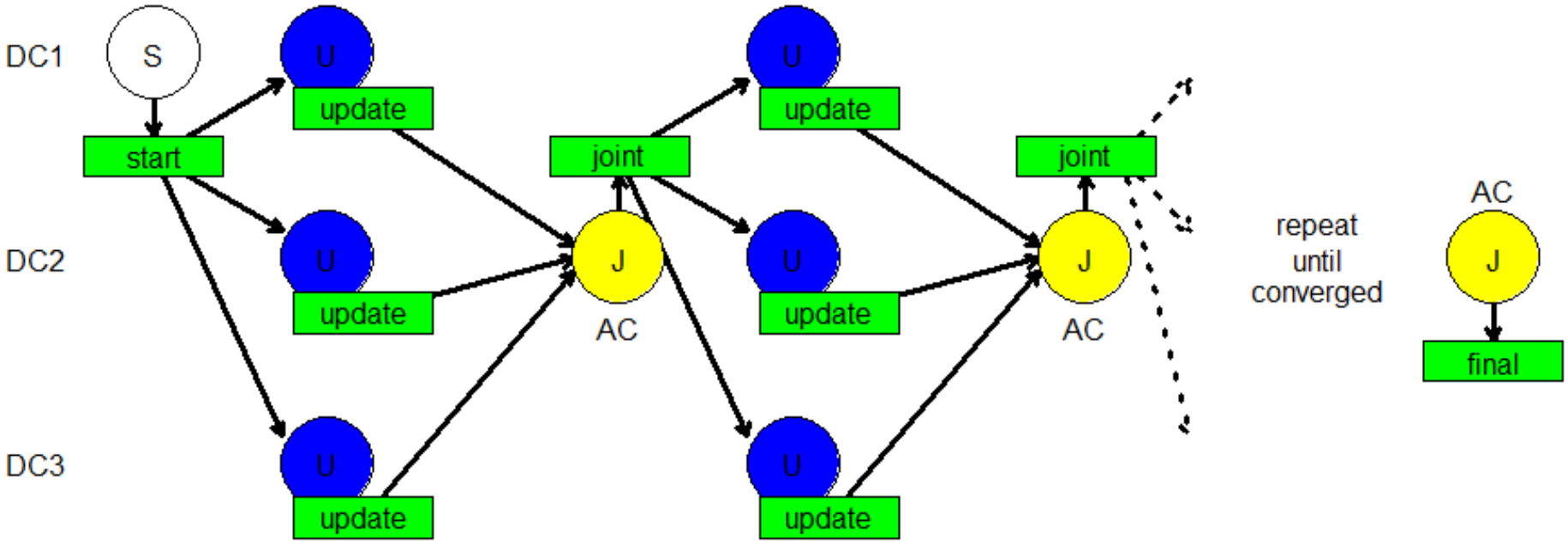
- www.datasheild.org
- Provides open source code to carry out these computations in the R package
- They also provide a means of setting up an interface (opal server) between the AC and the DCs that restricts what objects can be transferred
- In this protocol the AC controls all the analyses
- After **data harmonisation** between the sites the AC establishes communication with each one and can then run a limited number of commands to extract data summaries from the DCs

The DataSHIELD protocol



- AC (yellow) controls the whole analysis
- The interface restricts what can be extracted from the DCs

The E-DataSHIELD protocol



At each iteration an analyst at a DC must

- receive a vector of coefficients from the AC by email
- transfer it to the secure server via a secure data stick (for some LSs this transfer can only be done by agency staff, not be the individual researcher)
- read it into their R workspace
- run a routine to update the information matrix and score vector write them to a file
- export the file from the secure setting via the secure data stick
- email it to the AC.

At each iteration an analyst at a DC must

- receive a vector of coefficients from the AC by email
- transfer it to the secure server via a secure data stick (for some LSs this transfer can only be done by agency staff, not be the individual researcher)
- read it into their R workspace
- run a routine to update the information matrix and score vector write them to a file
- export the file from the secure setting via the secure data stick
- email it to the AC.

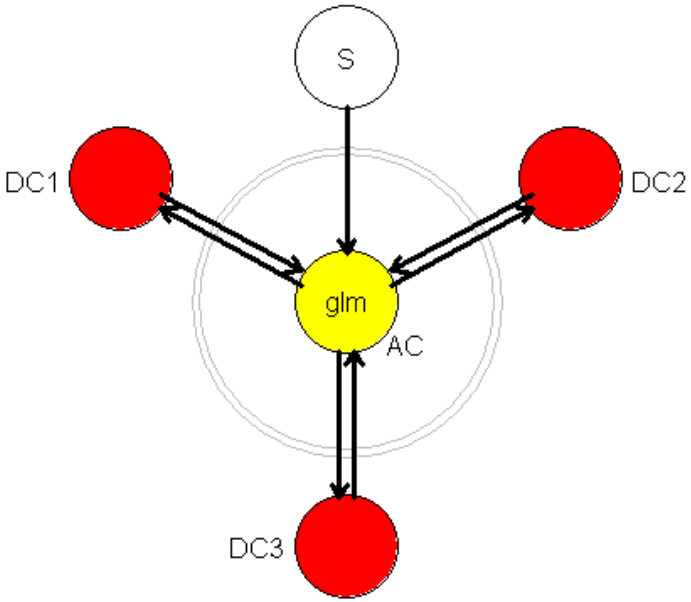
Features of the E-DataSHIELD method

- Automatic calculation of starting values
- Fits a whole set of models together and stops when they have all converged
- Generates the names of the files to transfer automatically at each iteration, and then checks that everything is as it should be.
- Other routines available for pre-analysis harmonisation and summaries of data and presentation of results
- Two projects using the LSs have used this methodology successfully

Secure multi-part computation

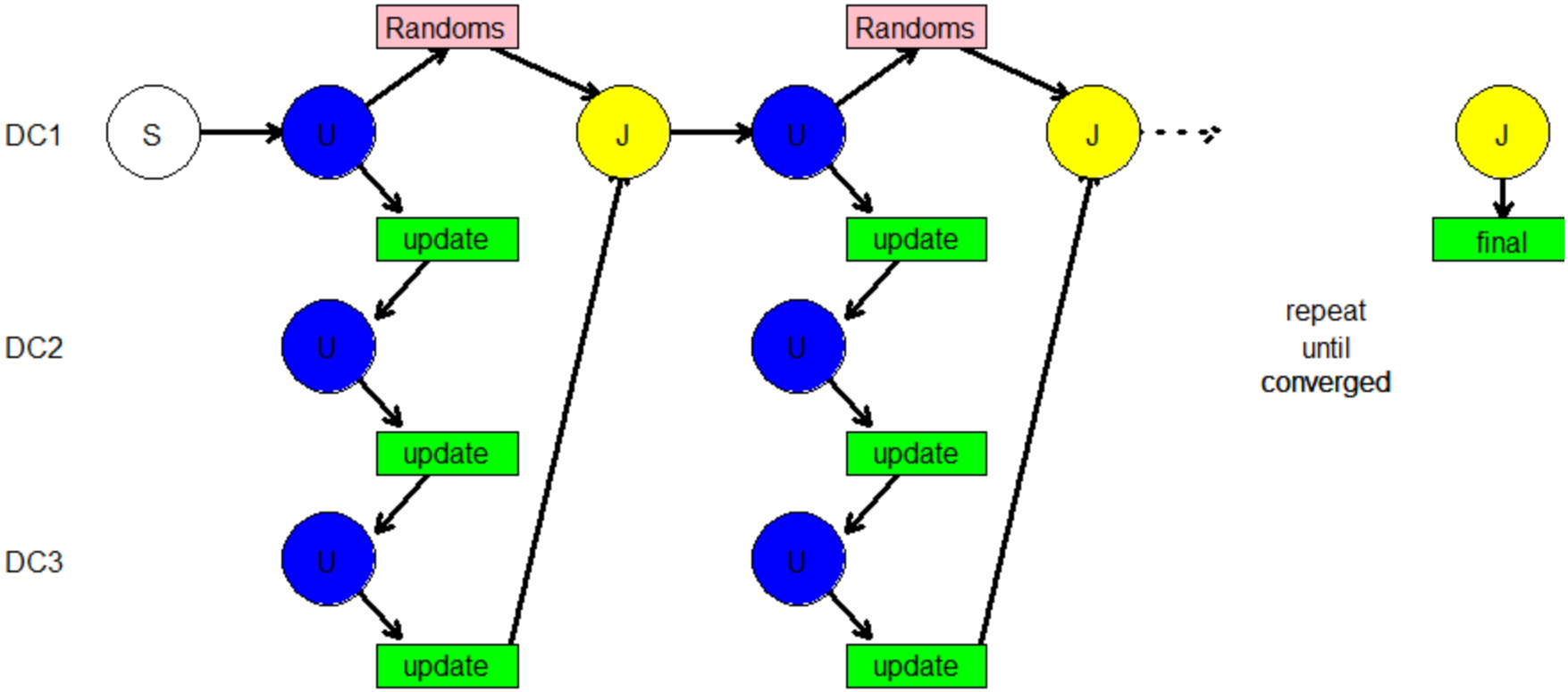
- The individual parts of the summaries are encrypted using homomorphic encryption
- Simplest example of this (used by statisticians) is for the first DS to add a random number to their result and pass it round the others, then when it comes back to the first one the random number is subtracted
- Computer scientists use more sophisticated ones
- But it wouldn’t work with DataShield

The DataSHIELD protocol



- DCs don’t communicate with each other
- But no reason why it should not work with E-DataShield

E-DataSHIELD with secure summaries



E-DataSHIELD

eds

Will shortly to be submitted to CRAN

http://cran.r-project.org/package=eds

Sample eds code

At each DC

>eds.glm.update("spine",itno=1,data=SEast,quietly=T)

File written Project_spine_Itno_1_Study_2.R

Now email file to analysis computer

At the AC

> eds.joint.update(project="spine",itno=1,studies=1:3,quietly=T)

which gives the message

0 out of 3 models converged

File written: Joint_fit_project_spine_Itno_2.R .

Sample eds code with secure summaries

At the first DC

>eds.glm.update("spine",itno=1,data=London,quietly=F,secure=T)

Result read from file Start_project_spine_Itno_1.R

Fitting 3 models for project spine at iteration 1

File with random objects is written as Project_spine_Itno_1_Randoms.R

Result for project spine iteration 1 study 1

written to file Project_spine_Itno_1_Study_1.R

Now email file to next data computer, study 2

At the next DC

Same commands but the random file is only written by the first DC

And at the next iteration the first DC reads it before the Joint update and writes a new one for the next iteration

.

Further points in the written paper

- Examples of how the results of the analyses can be computed
- How covariate by study interactions are fitted
- An example using synthetic data generated from one of the LSs
- References to other papers

Conclusions

- We have a package that works for the LSs
- We plan to make it publicly available via CRAN
- And hope it will be useful to others
- More details and future directions in the published paper

Acknowledgments

We are grateful to the DataSHIELD project for having introduced us to these ideas and shared their code with us. Also to the staff of the LSs for helping us to develop **eds**.

The DataSHIELD project is supported by funding from: the European Union's Seventh Framework Programme - BioSHaRE-EU (Biobank Standardisation and Harmonisation for Research Excellence in the European Union); a strategic award from MRC and Wellcome Trust for the ALSPAC project; and the Welsh and Scottish Farr Institutes, MRC funded E-Health Informatics Research Centres (EHIRCs).