DataSHIELD: an introduction

Paul Burton

University of Bristol, D2K Research Program, ALSPAC
Farr Institutes Wales and Scotland
McGill University, OICR, Maelstrom Research
The Norwegian Institute of Public Health, Dept of Epidemiology

DataSHIELD Developers Workshop

University of Bristol: 17-19 June, 2015

Why now?

- DataSHIELD is at a critical juncture
  - Underpinning theory proven
  - Technology for platform developed and implemented in open source software
  - Real-world potential roles undoubted and increasing

Why now?

- DataSHIELD is at a critical juncture
  - Underpinning theory proven
  - Technology for platform developed and implemented in open source software
  - Real-world potential roles undoubted and increasing

BUT!!!

Why now?

- The funding and resources we have so far obtained was never intended to have taken us so far
- Need a period of intense focus on making the software usable by non-developers
- Two key options (need both):
  - Keep seeking additional funding
  - Broaden the developers group – particularly to include technical specialists from groups with a direct methodological or applied interest

HDS – where are we now?

- Priorities
  - Working to enhance current functions and ease of use
  - Creating and enhancing documentation and tutorial material
  - Extending functionality:
    - Automate data access protocols
    - Server status monitoring and alerting
    - Survival models
    - Large scale genomics (Random effects SLMA, Opal)
    - Generalized linear *mixed* models
    - Textual data
  - Formal governance for DataSHIELD project itself

Exemplar questions and problems: thanks to Tom Bishop and the InterConnect Project

Fundamental questions and problems

- One technique (missing indicator method) was not easily implemented using existing DataSHIELD functionality.
  - Need new functions and enhanced utility of pre-existing functions
- Simple procedures such as generating new variables for use in analyses based on the values of existing variables are difficult using existing DataSHIELD functions.
  - Need new functions and enhanced utility of pre-existing functions
- The balance between an acceptable level of security and provision of an environment that is easier to use may also be considered in the future
  - One of most important issues facing us: particularly for tabulation and sub-setting rules

Fundamental questions and problems

- The processing time for some operations was much longer with DataSHIELD than with standalone R analysing the same data
  - Avoid defaulting most “checks” and unnecessary “assign” procedures during analysis (*e.g.* ds.meanByClass)
- Users need to learn to use R and DataSHIELD; this may prove a barrier to its widespread uptake. Presently many researchers use commercial software such as Stata.
  - New user interfaces and possibility of extending platform to new database implementations and analytic tools
- Other studies will have to provide support to users when things go wrong or do not work
  - Need more formal – explicitly resourced - service-level agreements with users

Fundamental questions and problems

- How do you achieve the best balance between doing a lot of pre-processing during harmonisation against being able to manipulate data in DataSHIELD?
  - Enhanced flexibility in assignment
- What are the tips and tricks for working with data where you cannot see them? How do I validate and trust my output without being able to see the data?
  - Enhanced exploration and use of logical checks
  - Better missing data handling
  - DANGERfunctions

More specific issues

- How do I save my analysis in RStudio Server?
- What do I do if I need a new function to be developed for DataSHIELD?
- How can I tell if one server is the performance bottleneck in my analysis?
- Who do I contact if I need help?

- Many such questions will be covered in the workshop
- We will keep a record on additional issues that arise during discussion so we can prioritise subsequent work

What problem does DataSHIELD set out to solve?

What problem does DataSHIELD set out to solve?

The modern biomedical and social sciences are critically dependent on data sharing and pooling. But there are many (reasonable) constraints and barriers to sharing and combining raw individual-level data

Fundamental barriers

- Ethico-legal or other governance restrictions
- Maintaining control of intellectual property
- Physical size of data

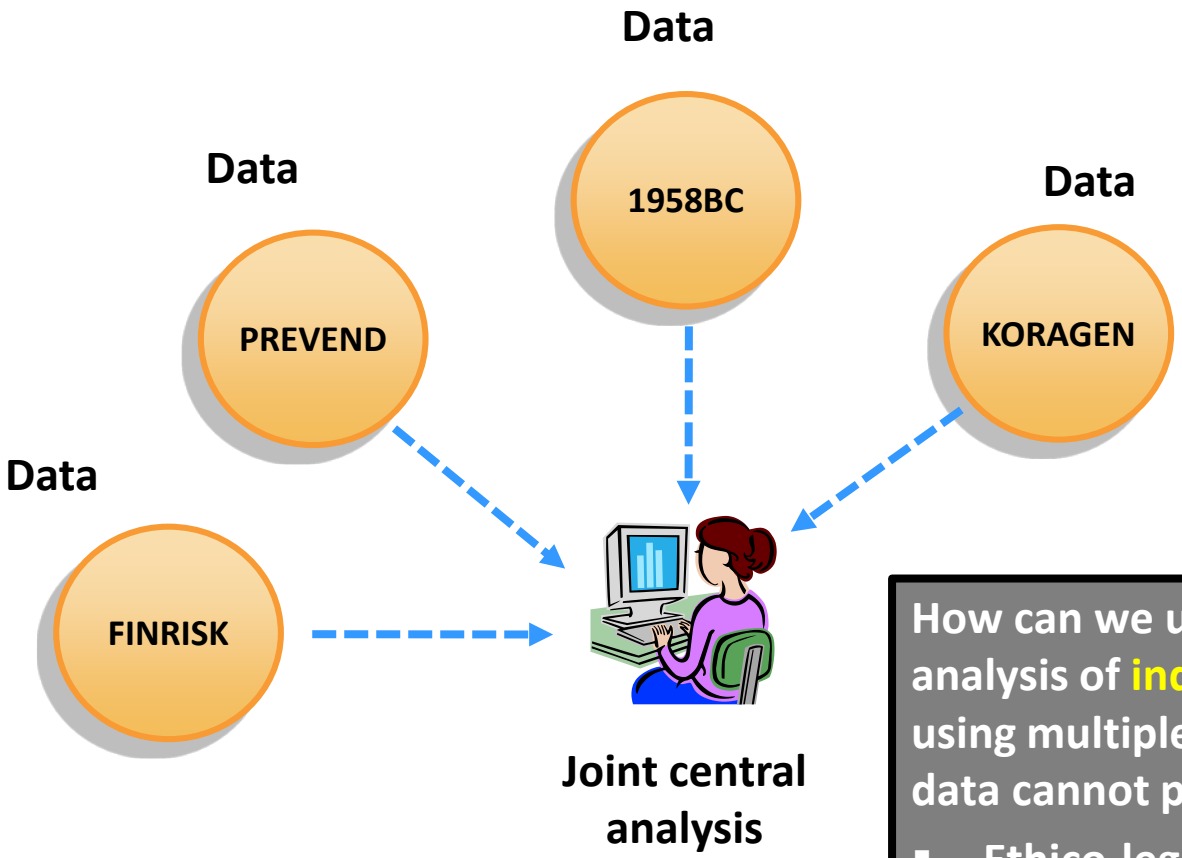
- How can we deal with these problems?

Individual-level data - terminology

- Data relating to *individual subjects* held in a dataset
  - = microdata
  - = IPD, *i.e.* “individual patient data”
- Contrast with study-level data
  - e.g. study level meta analysis (SLMA)

Horizontally Partitioned Data

Horizontally partitioned data



How can we undertake a full joint analysis of individual-level data using multiple data sources if the data cannot physically be pooled?

- Ethico-legal constraints
- Intellectual property issues
- Physical size of the data objects

BioSHaRE.eu

data to knowledge

Two approaches to data synthesis

- Study level meta-analysis (SLMA)
  - Obtain result for each study separately – *e.g.* odds ratio for a SNP. Calculate an appropriately weighted mean and standard error for that odds ratio across *all* studies
  - = “Conventional meta-analysis”
- Individual level meta-analysis (ILMA)
  - Pool all of the individual level data from each of the studies into one large data set and then analyse that data set as if it was one single study (with parameters for heterogeneity)
  - = “Direct pooling”

Study level meta-analysis

- Quick, easy and (generally) efficient
- But SERIOUS lack of flexibility - for example:
  - One million SNPs on a GWA chip are successfully analysed
  - But, then you want to study interaction of all apparently associated SNPs with age and sex
  - Impossible unless these analytic results provided up-front
- Contemporary bioscience is getting more complex
- Exploratory analysis needs flexibility

ILMA (direct data pooling) therefore preferable

Individual level meta-analysis: *i.e.* sharing and pooling individual data

- Analytically optimal
- But the important constraints are real:
  - Ethico-legal or other governance restrictions
  - Maintaining control of intellectual property
  - Physical size of data

Where are we now?

- Analytic flexibility greatly favours ILMA
- But many potential barriers to sharing individual level data
  - → Most current GWASs based on SLMA
  - BUT: this situation is not sustainable as things become more complex, unpredictable and exploratory
- Neither commonly used solution is really ideal

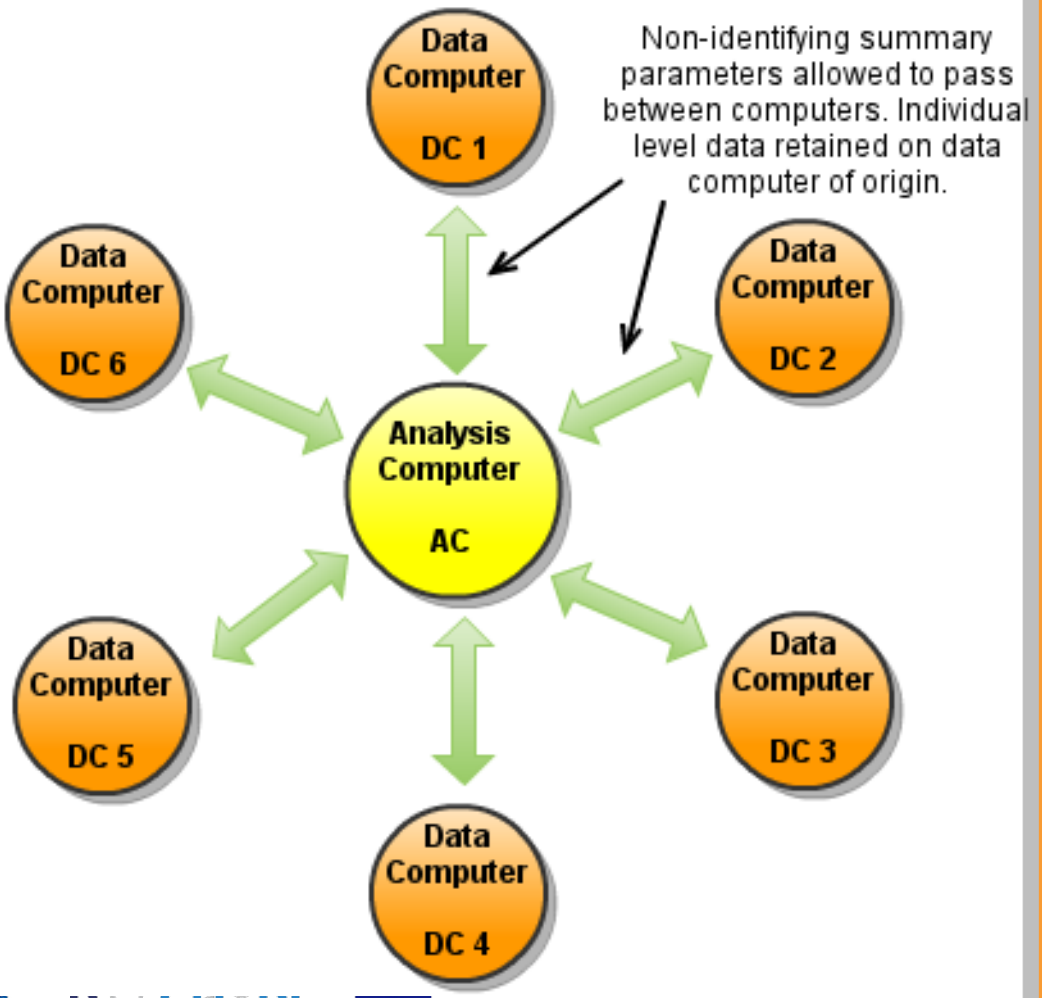
An alternative approach

- Take “analysis to data” ….. not “data to analysis”
- Leave the raw data from each study on a local server at that study
- Analysis centre co-ordinates parallelised analyses in all studies simultaneously
- Tie analyses together with non-disclosive *statistics* of an appropriate nature (ideally *sufficient* statistics)
- CRUCIALLY – get’s around key challenges both of ILMA and SLMA!!!

DataSHIELD:

Data Aggregation Through Anonymous Summary-statistics from Harmonized Individual-Level Databases

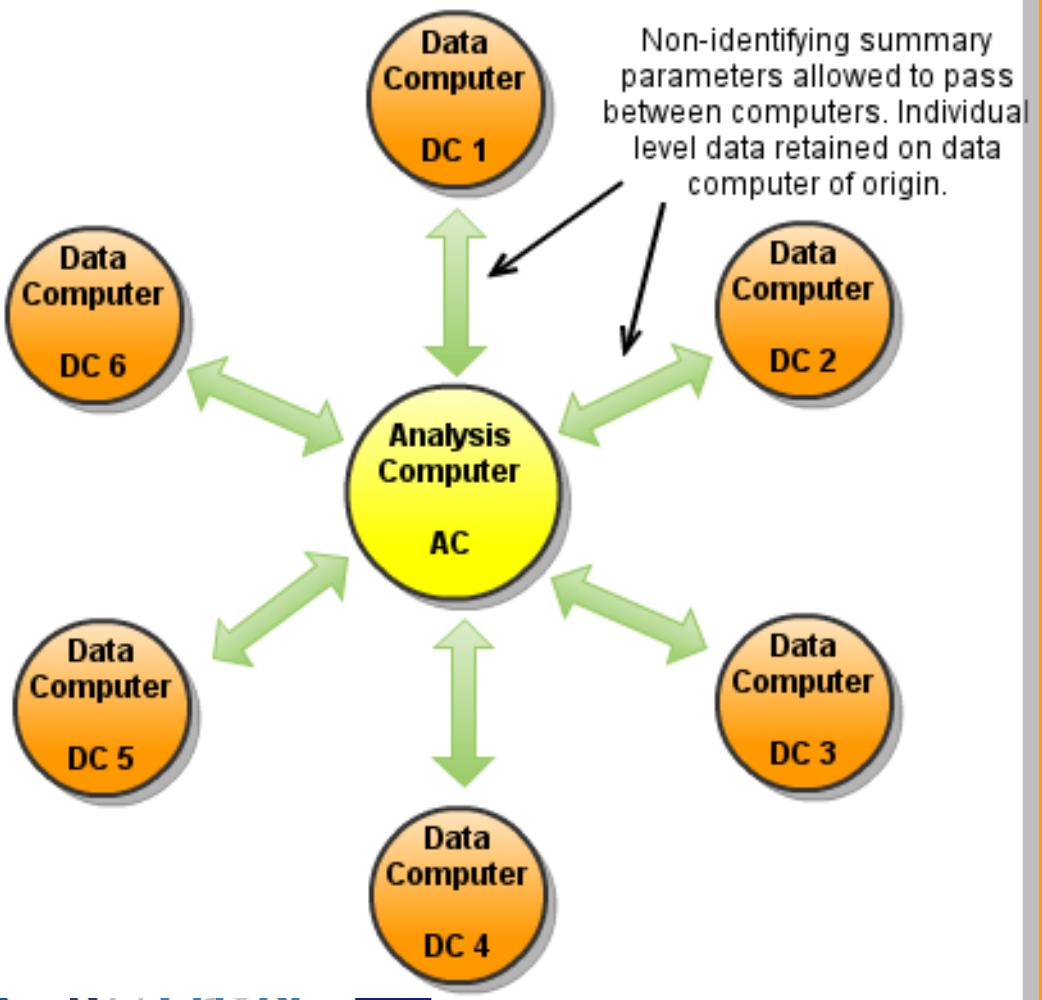
DataSHIELD: a novel solution



- Take analysis to data … not data to analysis
- One step analyses: simply combine non-disclosive output from all sources
- Iterative analyses: parallel processes linked together by entirely non-identifying summary statistics – *e.g.* for glm = score vectors and information matrices

data to knowledge

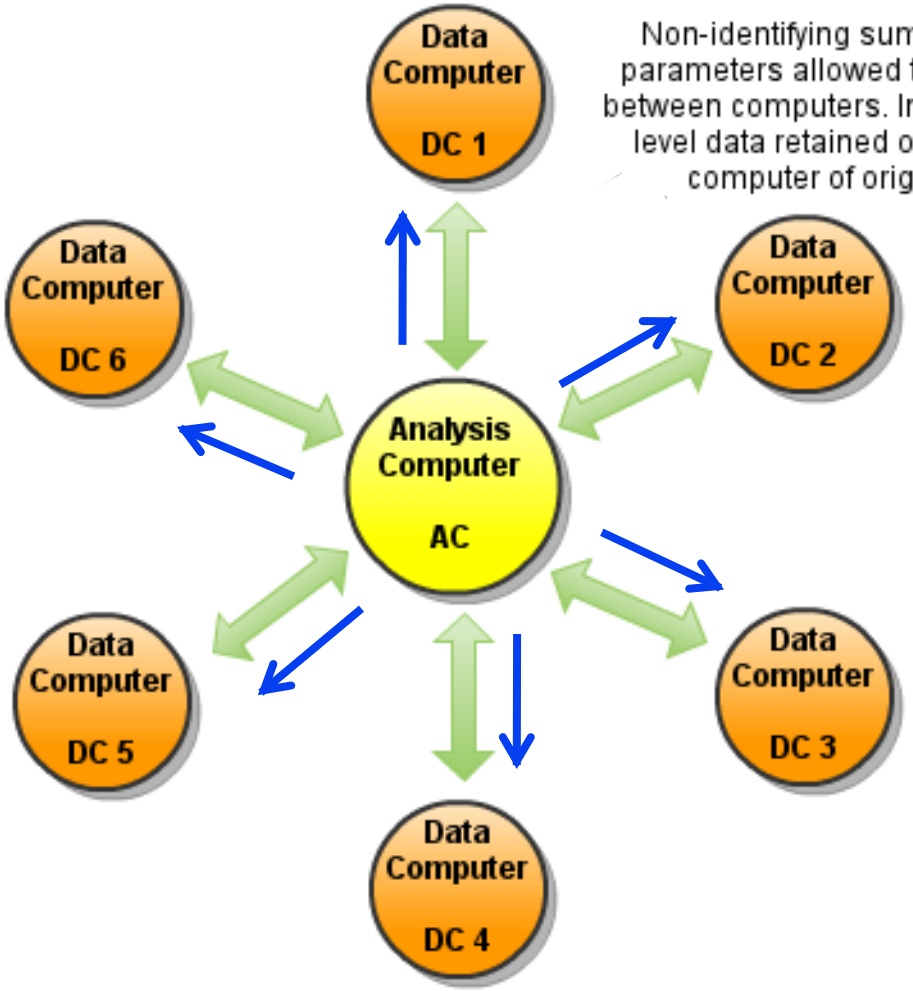
DataSHIELD: a novel solution



- Take analysis to data … not data to analysis
- One step analyses: simply combine non-disclosive output from all sources
- Iterative analyses: parallel processes linked together by entirely non-identifying summary statistics – *e.g.* for glm = score vectors and information matrices
- Can be used as equivalent to full individual level analysis or to study level meta-analysis

data to knowledge

DataSHIELD: a novel solution

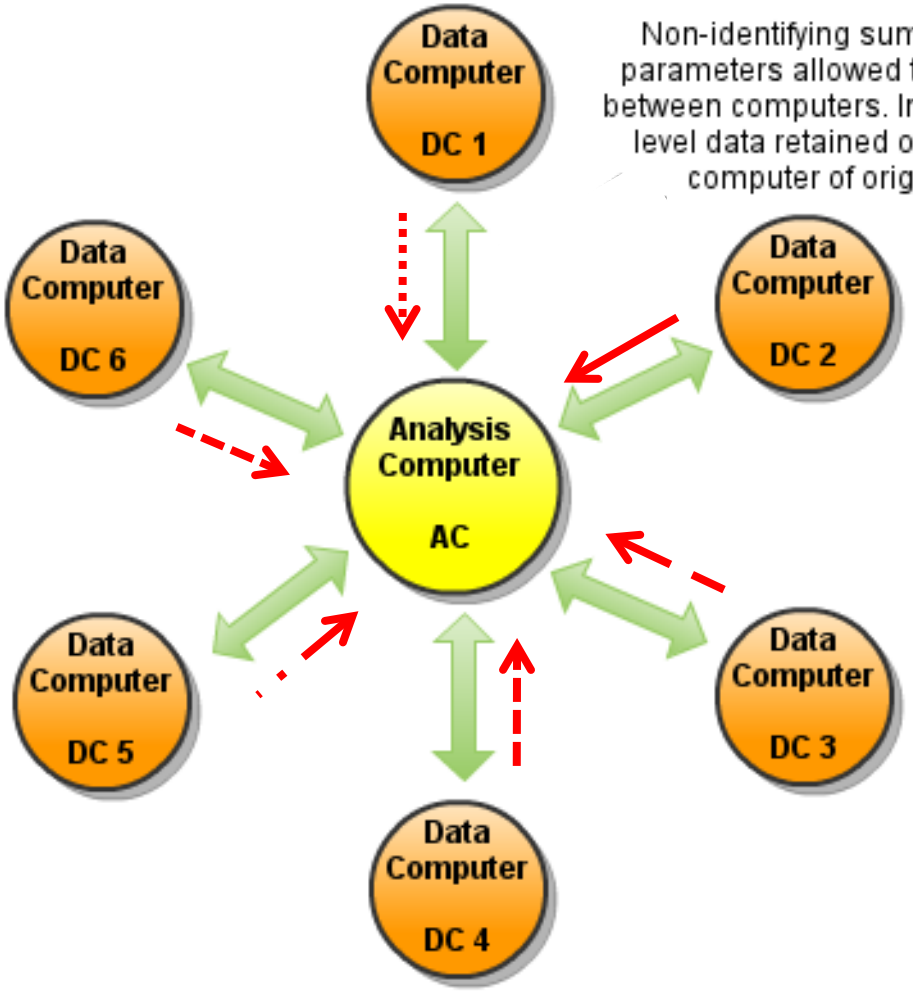


Analysis commands (1)

b.vector<-c(0,0,0,0)

glm(cc~1+BMI+BMI.456+SNP, family=binomial, start=b.vector, maxit=1)

DataSHIELD: a novel solution



Non-identifying summary parameters allowed to pass between computers. Individual level data retained on data computer of origin.

Summary Statistics (1)

Score vector Study 5

[36, 487.2951, 487.2951, 149]

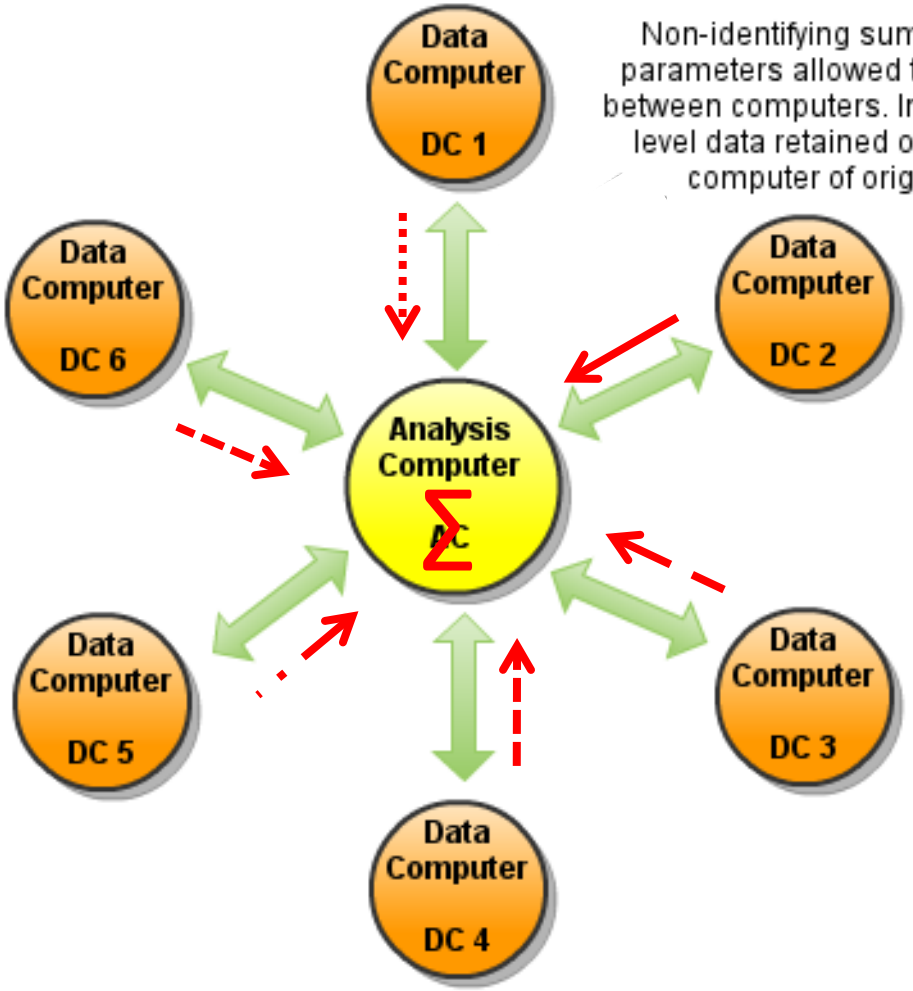
Information Matrix Study 5

500	70.56657	70.56657	297
70.56657	7646.29164	7646.29164	65.39412
70.56657	7646.29164	7646.29164	65.39412
297	65.39412	65.39412	382

BIOSHARE.eu

D2K data to knowledge

DataSHIELD: a novel solution



Non-identifying summary parameters allowed to pass between computers. Individual level data retained on data computer of origin.

Summary Statistics (1)

Score vector Study 5

[36, 487.2951, 487.2951, 149]

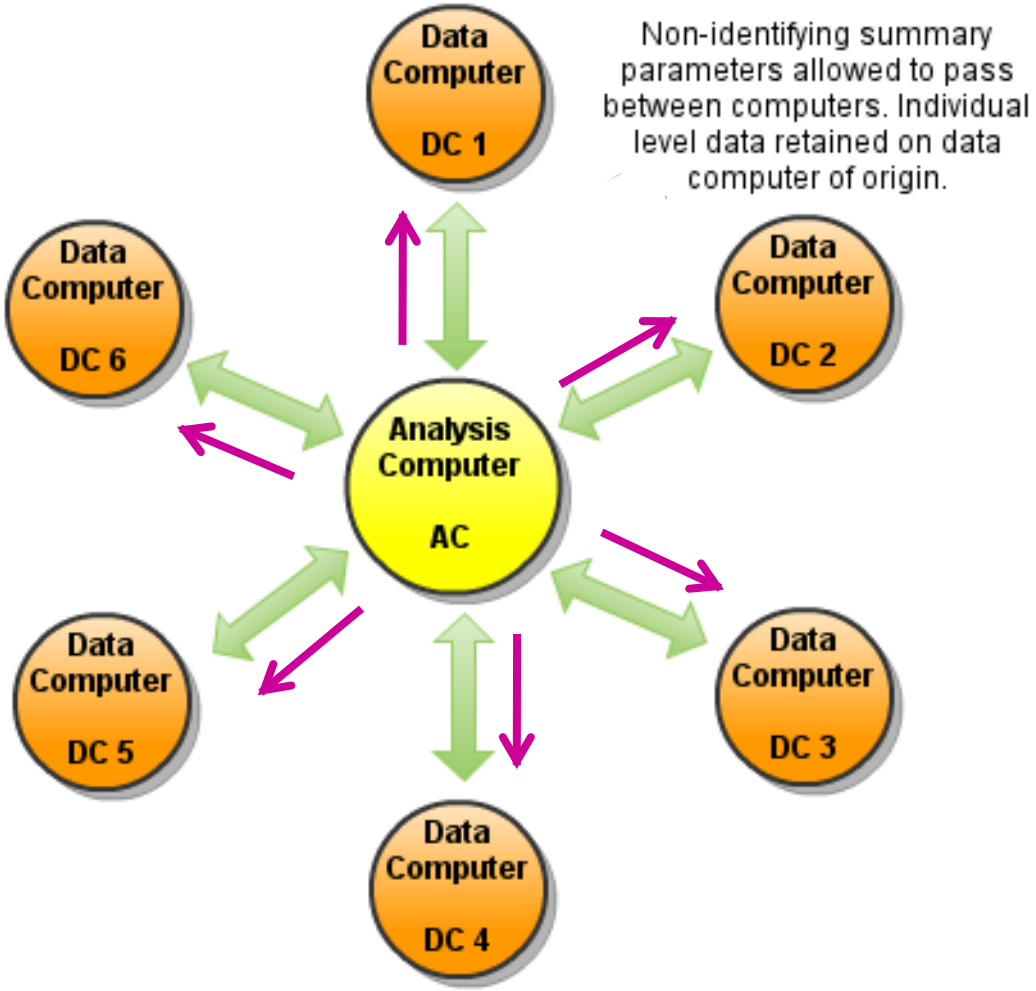
Information Matrix Study 5

500	70.56657	70.56657	297
70.56657	7646.29164	7646.29164	65.39412
70.56657	7646.29164	7646.29164	65.39412
297	65.39412	65.39412	382

BIOSHARE.eu

D2K data to knowledge

DataSHIELD: a novel solution



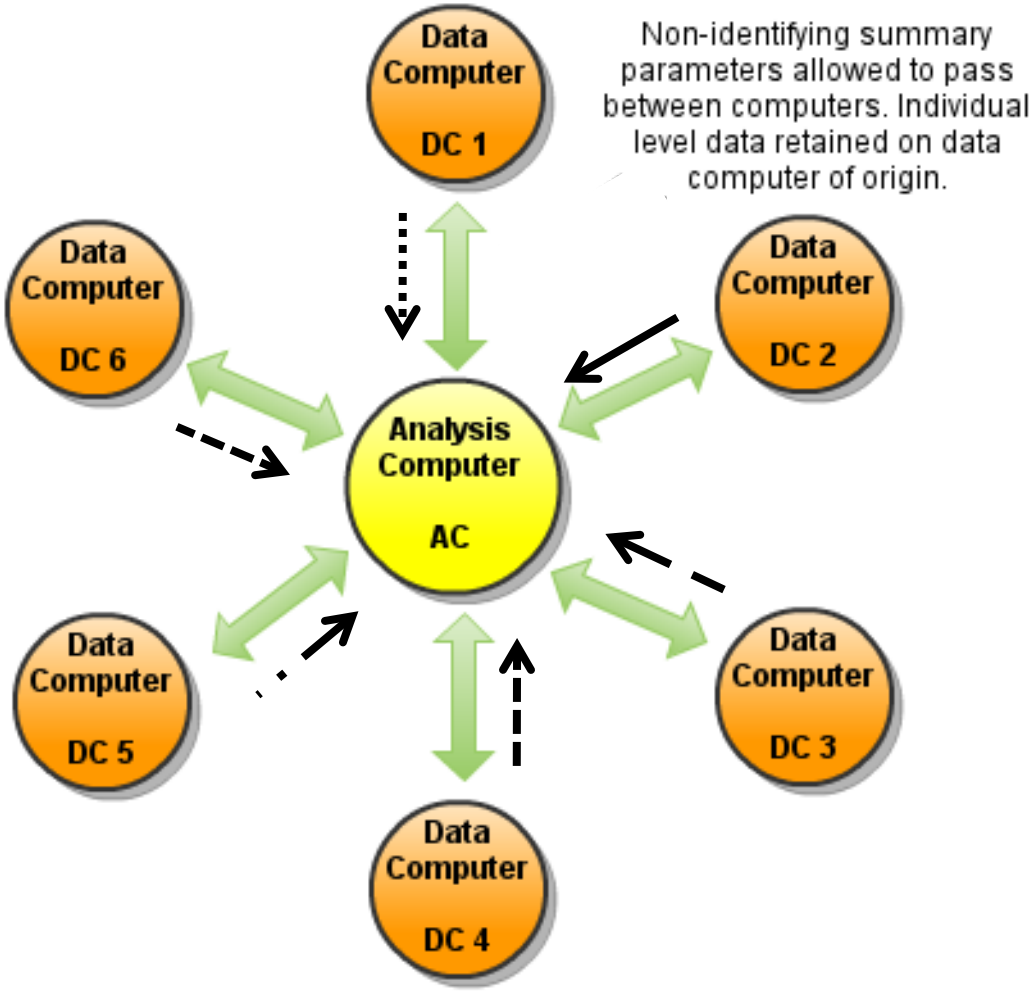
Analysis commands (2)

b.vector<-
c(-0.322, 0.0223, 0.0391, 0.535)

glm(cc~1+BMI+BMI.456+SNP,
family=binomial,
start=b.vector, maxit=1)

data to knowledge

DataSHIELD: a novel solution



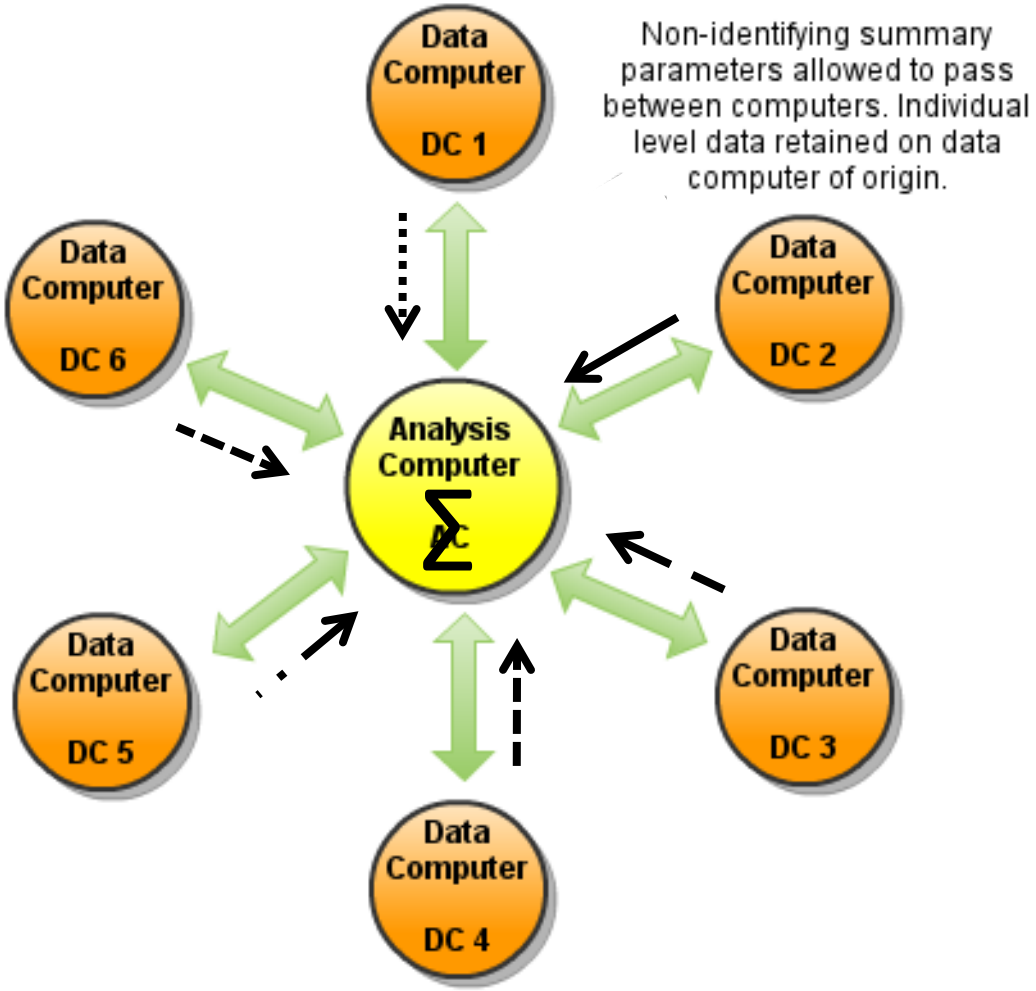
Summary Statistics (2)

Score vectors

Information Matrices

data to knowledge

DataSHIELD: a novel solution



Summary Statistics (2)

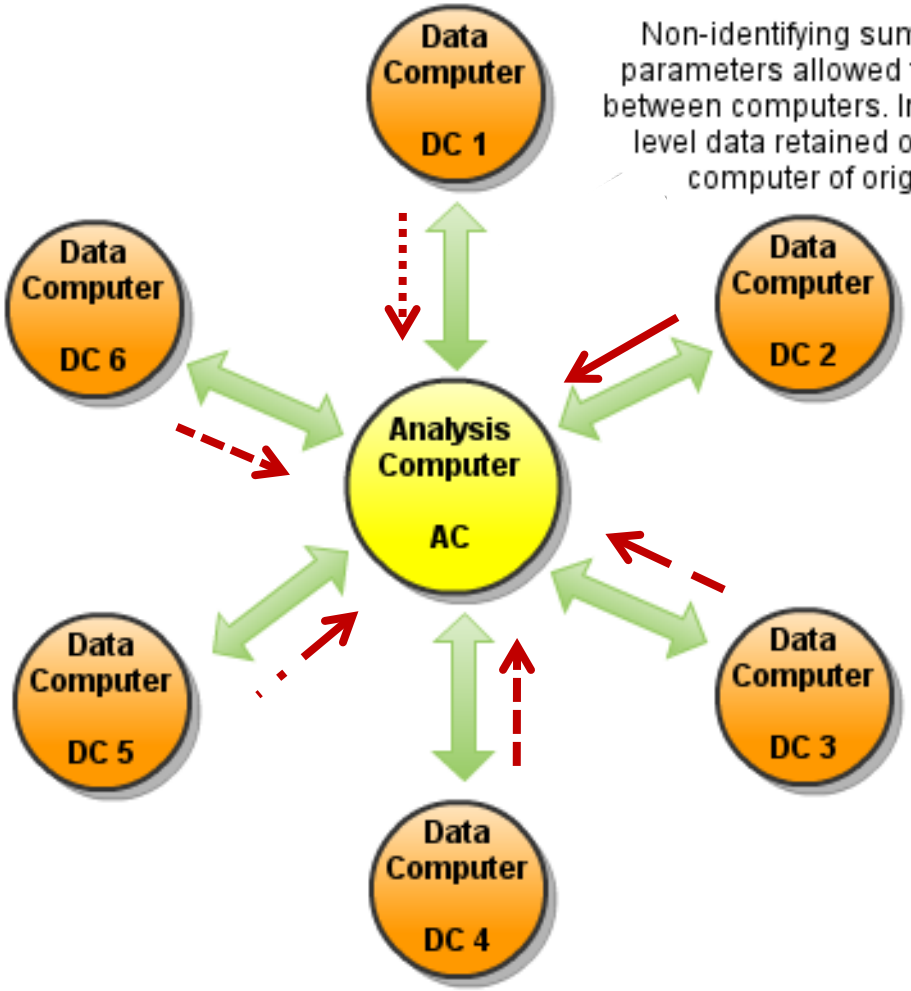
Score vectors

Information Matrices

data to knowledge

and so on .....

DataSHIELD: a novel solution



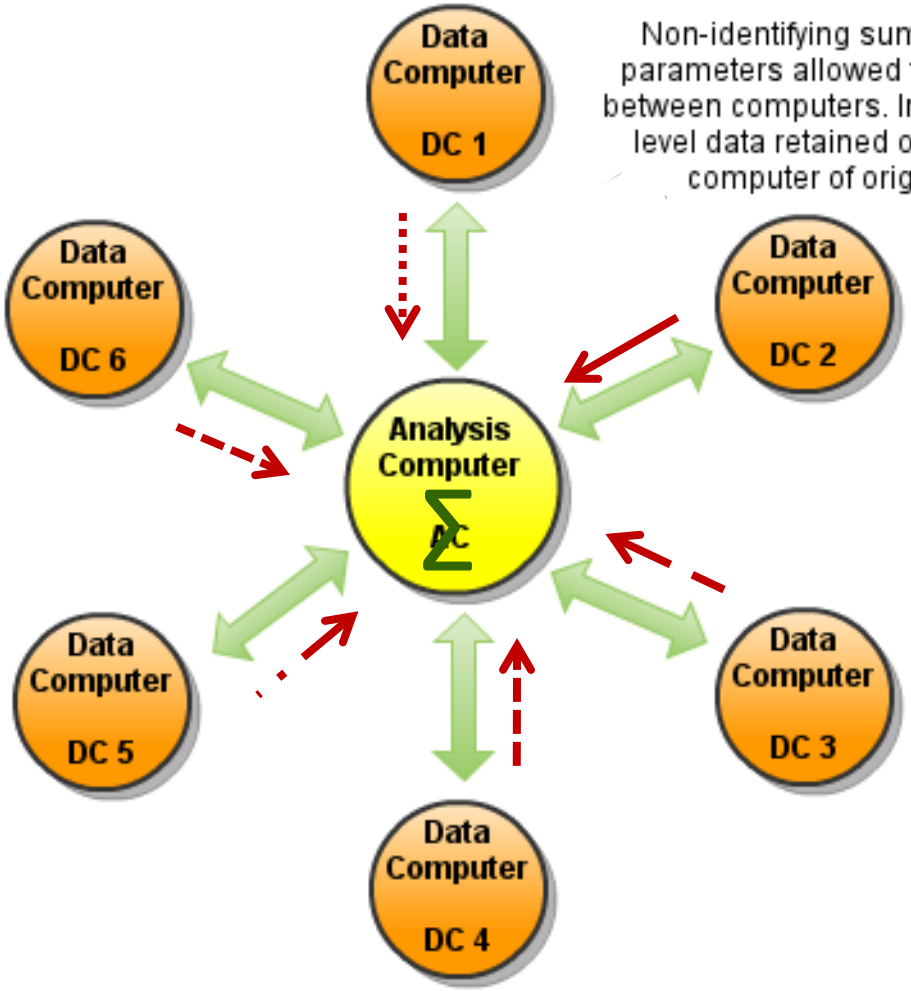
Updated parameters (4)

Final parameter estimates

BIOSHARE.eu

D2K data to knowledge

DataSHIELD: a novel solution



Updated parameters (4)

Final parameter estimates

Coefficient	Estimate	Std Error
Intercept	-0.3296	0.02838
BMI	0.02300	0.00621
BMI.456	0.04126	0.01140
SNP	0.5517	0.03295

data to knowledge

ILMA: Conventional analysis

Coefficients:

	Estimate	Std. Error
(Intercept)	-0.32956	0.02838
BMI	0.02300	0.00621
BMI.456	0.04126	0.01140
SNP	0.55173	0.03295

DataSHIELD analysis

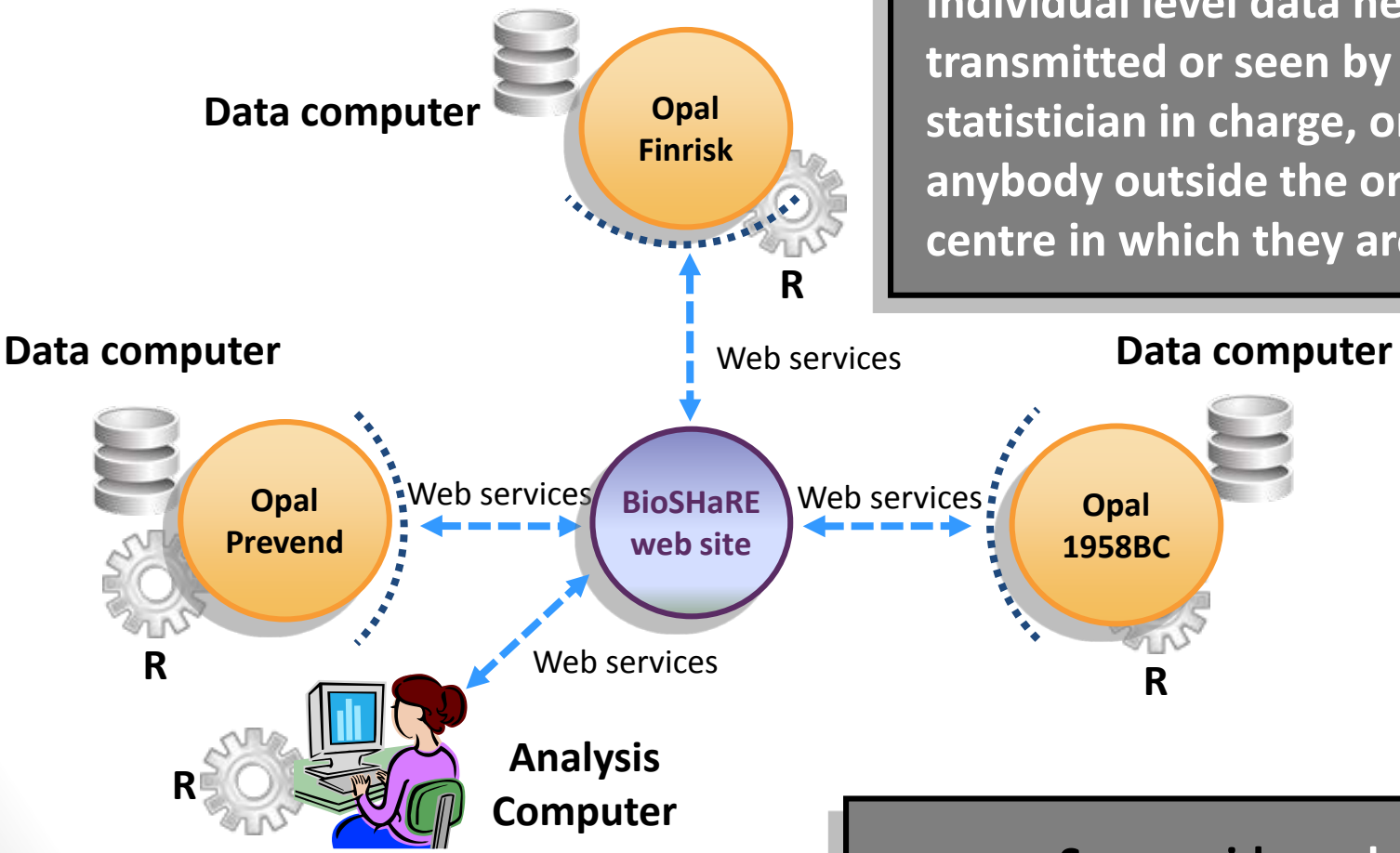
Does it work?

Parameter	Coefficient	Standard Error
$b_{intercept}$	-0.3296	0.02838
b_{BMI}	0.02300	0.00621
$b_{BMI.456}$	0.04126	0.01140
b_{SNP}	0.5517	0.03295

Current Implementation

Horizontal DataSHIELD

Individual level data never transmitted or seen by the statistician in charge, or by anybody outside the original centre in which they are stored.



Server-side and Client-side functions

Horizontal DataSHIELD – current status

- Overall
  - Proof of principle and practical implementation successful
- Multi-site horizontal DataSHIELD
  - Working to enhance current functions and ease of use
  - Creating and enhancing documentation and tutorial material
  - Extending functionality:
    - Automate data access protocols
    - Server status monitoring and alerting
    - Survival models
    - Large scale genomics (Random effects SLMA, Opal)
    - Generalized linear *mixed* models
    - Textual data
  - Formal governance for DataSHIELD project itself

Horizontal DataSHIELD – future flavours

- Single-site horizontal DataSHIELD
  - Potential currently being explored
  - Cost-effective, open source, secure data enclave
  - Controlled access to particularly sensitive intellectual property
    - H3AFRICA
  - 'Public' access to sensitive data
    - F1000 Journal
    - Easily updatable summary statistics for cohort studies freely available over web
- Vertical DataSHIELD

PostScript

- Harmonization CRITICAL
- Must understand and acceptability of DataSHIELD itself
- Must closely evaluate development of DataSHIELD
  - Technical
  - Ethicolegal
  - Social Context

MOST RECENT PUBLICATION describing practical implementation of horizontal DataSHIELD

DataSHIELD: taking the analysis to the data, not the data to the analysis.
Amadou Gaye, Yannick Marcon, Julia Isaeva, Philippe LaFlamme, ……, Madeleine J Murtagh, Vincent Ferretti and Paul R Burton *International Journal of Epidemiology, 2014, 1–16,* doi: 10.1093/ije/dyu188

ACKNOWLEDGEMENTS FROM PAPER

DataSHIELD development supported through funds from: the European Union's Seventh Framework Programme BioSHaRE-EU, grant agreement HEALTH-F4-2010-261433; the Welsh and Scottish Farr Institutes funded by MRC; joint funding from MRC and Wellcome Trust comprising a strategic award underpinning the ALSPAC project and an infrastructural grant entitled *The 1958 Birth Cohort Biomedical Resource – facilitating access to data and samples and enhancing future utility*; and the BBMRI-LPC project (EU FP7, I3 grant).

WEBSITES

www.datashield.ac.uk/ for full lists of publications and funding

www.datashield.org/ for technical detail (on horizontal DataSHIELD)

data to knowledge

THANK YOU FOR LISTENING

ELSI restrictions

- Exemplar wording
  - Wallace S, Lazor S, Knoppers BM. Chapter in Kaye J and Stranger M. Principles and Practice in Biobank Governance. Ashgate, Farnham 2009
- Use of data restricted to researchers participating in the original study
- Use of data restricted to researchers in one country
- The need to obtain ethico-legal and scientific permission to access the data
  - Often needs multiple clearances
  - Often a protracted and time consuming process

Intellectual property issues

- No issue if study originally funded on the basis that data would be freely shared and participants consented …… BUT what if:
  - Mature studies
  - Particular effort or specialist techniques used to collect data and biosamples
  - Overt non-reciprocation of access
  - Data collection in resource-poor region
- THEN:
  - Data generators may wish to fully collaborate and freely share information in a dataset, but not the raw data themselves

Physical size issues

- Genome sequence data
- Images
- Large blocks of potentially linked data – *e.g.* national hospitalization data or primary care data