DataSHIELD: free access to information while keeping primary data secure

Paul Burton
University of Bristol, D2K Research Program
McGill University, OICR, Maelstrom Research
The Norwegian Institute of Public Health, Dept of Epidemiology

Nationales Biobanken-Symposium, Berlin, 2013

11th December 2013

Combining data from multiple sources is fundamental to modern bioscience

- Need for large sample sizes and deep high quality phenotyping
- Environmental heterogeneity
- Checks for consistency (replication)
- Cost containment

- Additional variables (record linkage – *e.g.* health events)
- Longitudinal or familial extension of data collection
- Universal controls

7 COOPERATION BioSHaRE .eu

D2K data to knowledge

Constraints and barriers to sharing and combining raw individual-level data

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

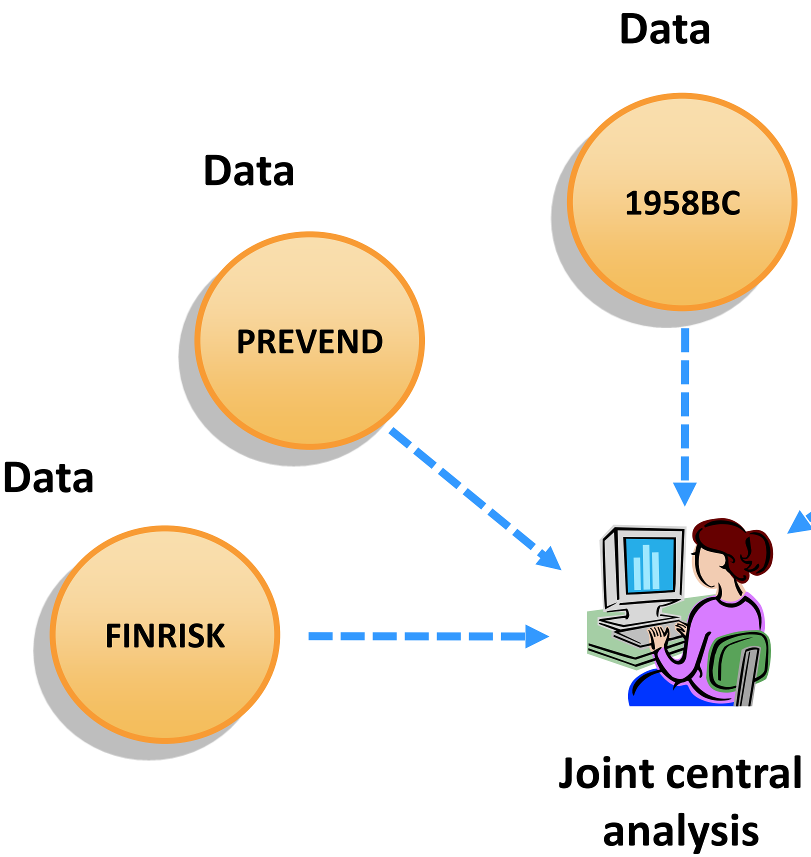
- How can we deal with these problems?

What actually needs combining, in what context, and how?

Individual-level data

- Often need to work with the 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



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

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”

data to knowledge

Study level meta-analysis

- Quick, easy <u>and it works</u>
- 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

BioSHaRE.eu

data to knowledge

Constraints on sharing individual-level data

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 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
  - Particular concerns about participant identification
- THEN:
  - Data generators may wish to fully collaborate and freely share <u>information</u> in a dataset, but not the <u>raw data</u> themselves

Physical size issues

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

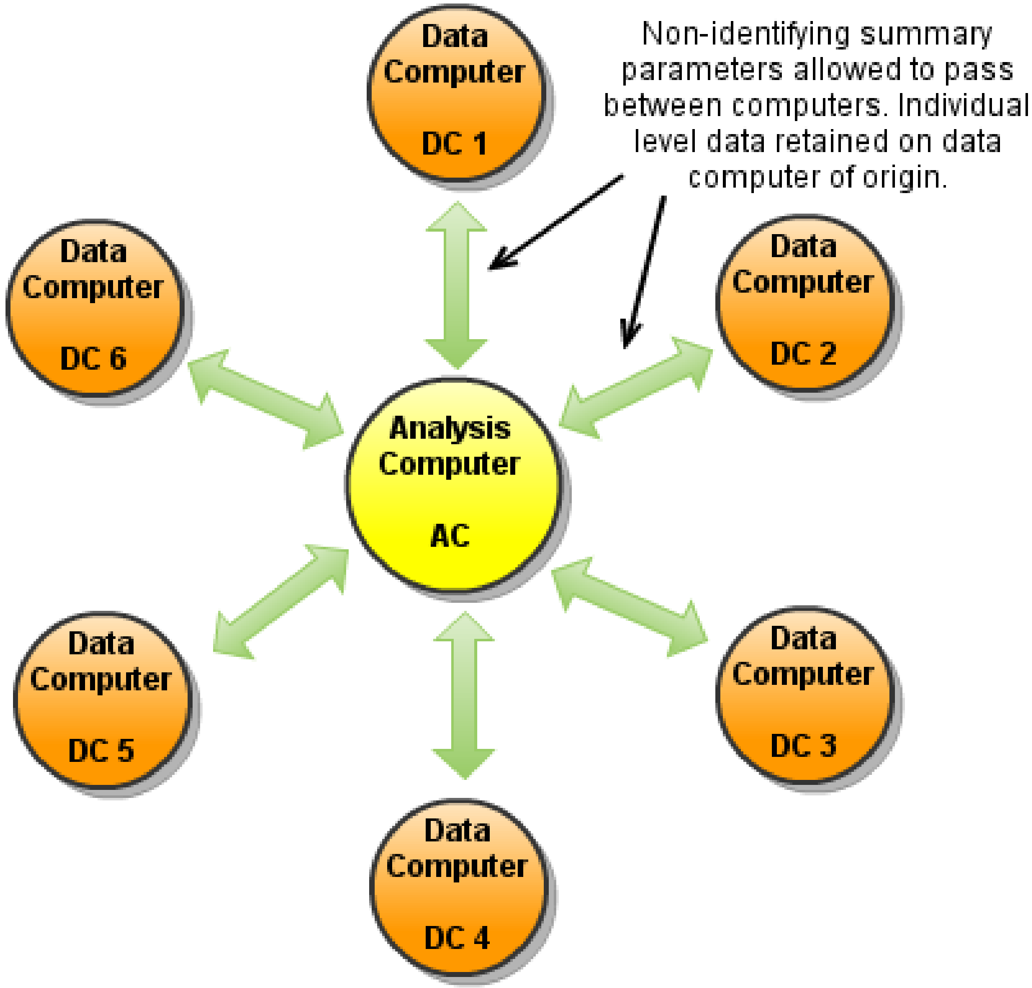
A radically different approach

- 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 simultaneous parallelised analyses in all studies simultaneously
- Tie analyses together with non-disclosive “summary statistics” so the overall analysis is equivalent to working on a single dataset

DataSHIELD:

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

DataSHIELD: a novel solution



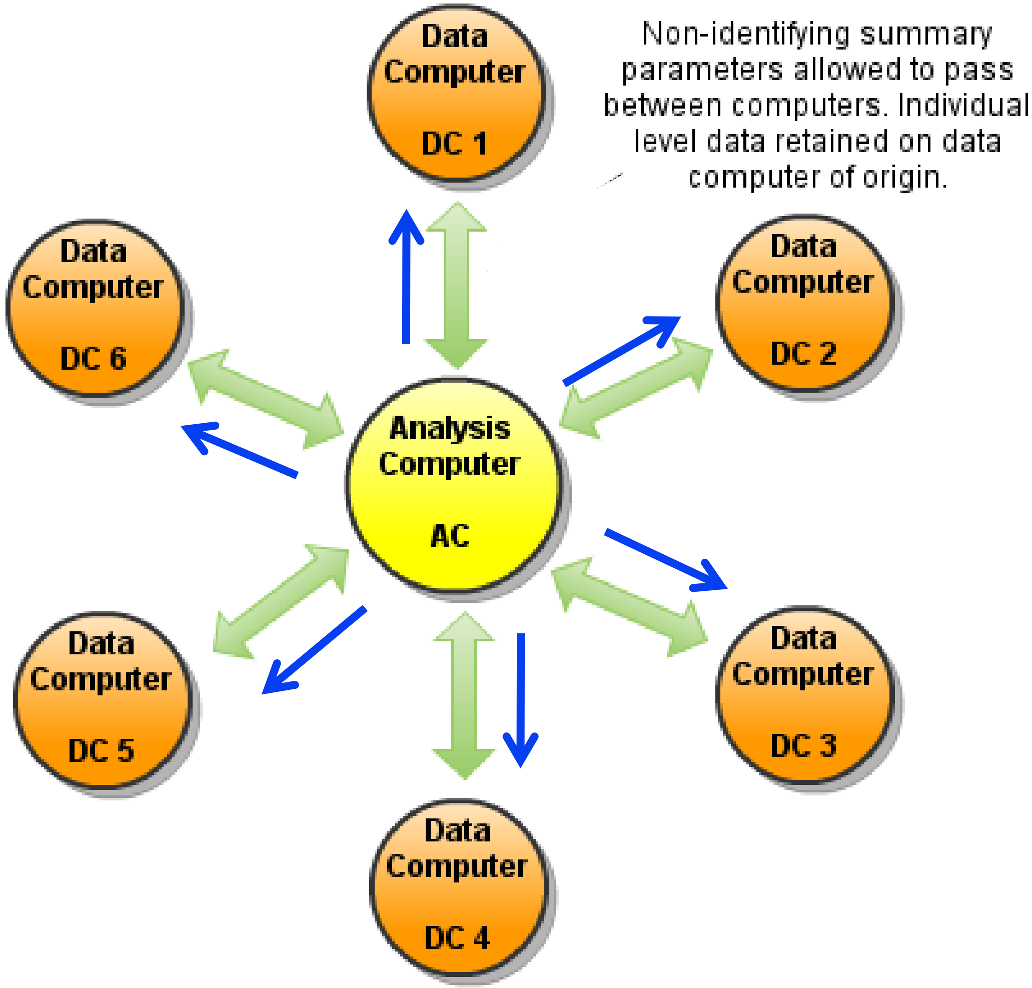
Take analysis to data … not data to analysis

One step analyses: simple

Iterative analyses: parallel processes linked together by entirely <u>non-identifying summary statistics</u>

Typically produces mathematically <u>identical results</u> to fitting a single model to all the data held in one pooled data set

DataSHIELD: a novel solution

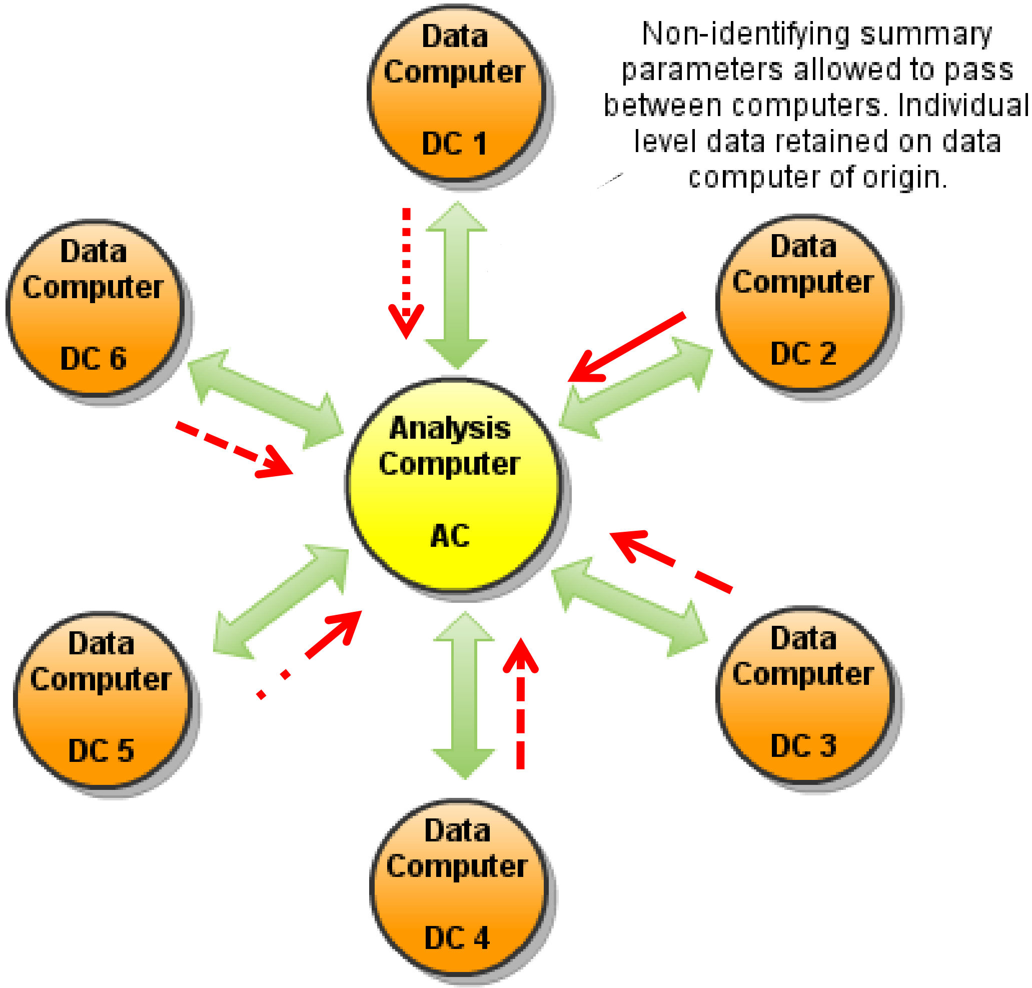


Analysis commands (1)

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

glm(cc~1+sex+snp+bmi, family=binomial, start=b.vector, maxit=1)

DataSHIELD: a novel solution



Summary Statistics (1)

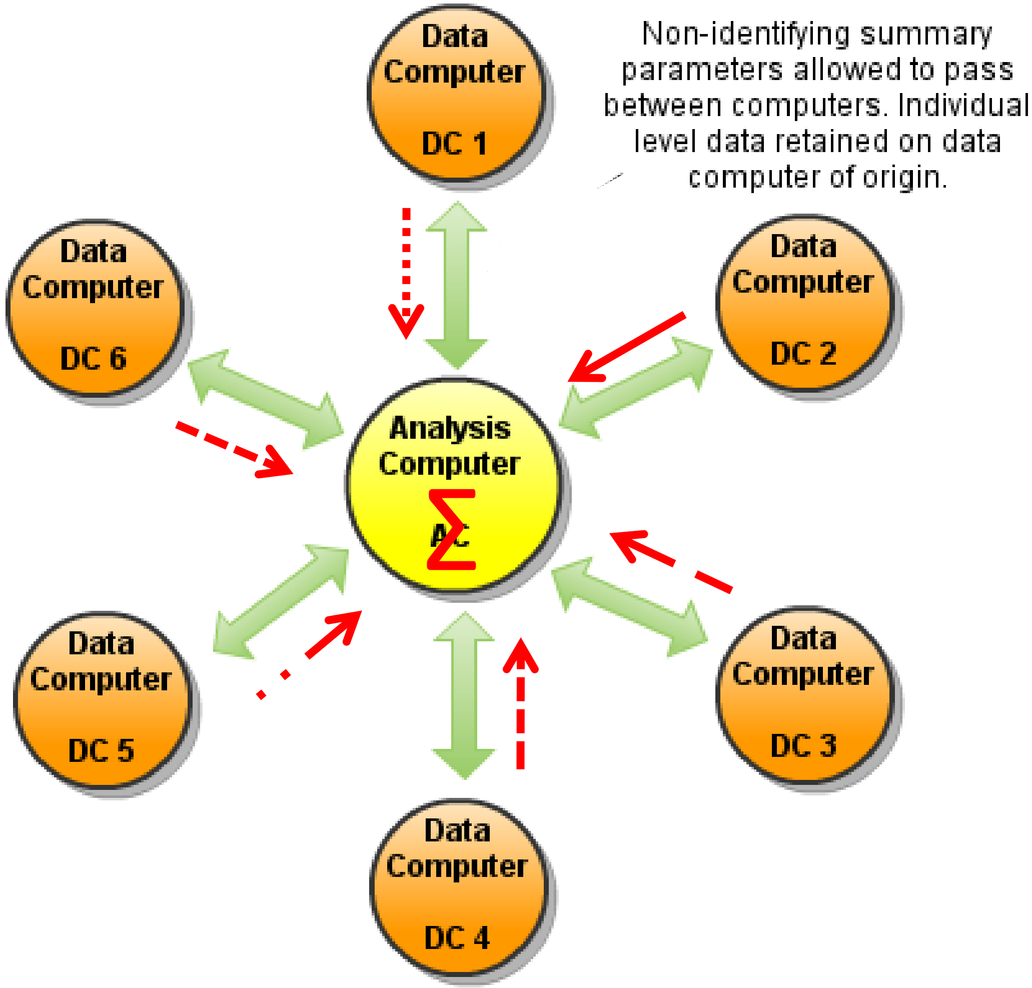
Score vector $_{\text{Study 5}}$

[36, 487.2951, 487.2951, 149]

Information Matrix $_{\text{Study 5}}$

$$\begin{vmatrix} 500 & 70.56657 & 70.56657 & 297 \\ 70.56657 & 7646.29164 & 7646.29164 & 65.39412 \\ 70.56657 & 7646.29164 & 7646.29164 & 65.39412 \\ 297 & 65.39412 & 65.39412 & 382 \end{vmatrix}$$

DataSHIELD: a novel solution



Summary Statistics (1)

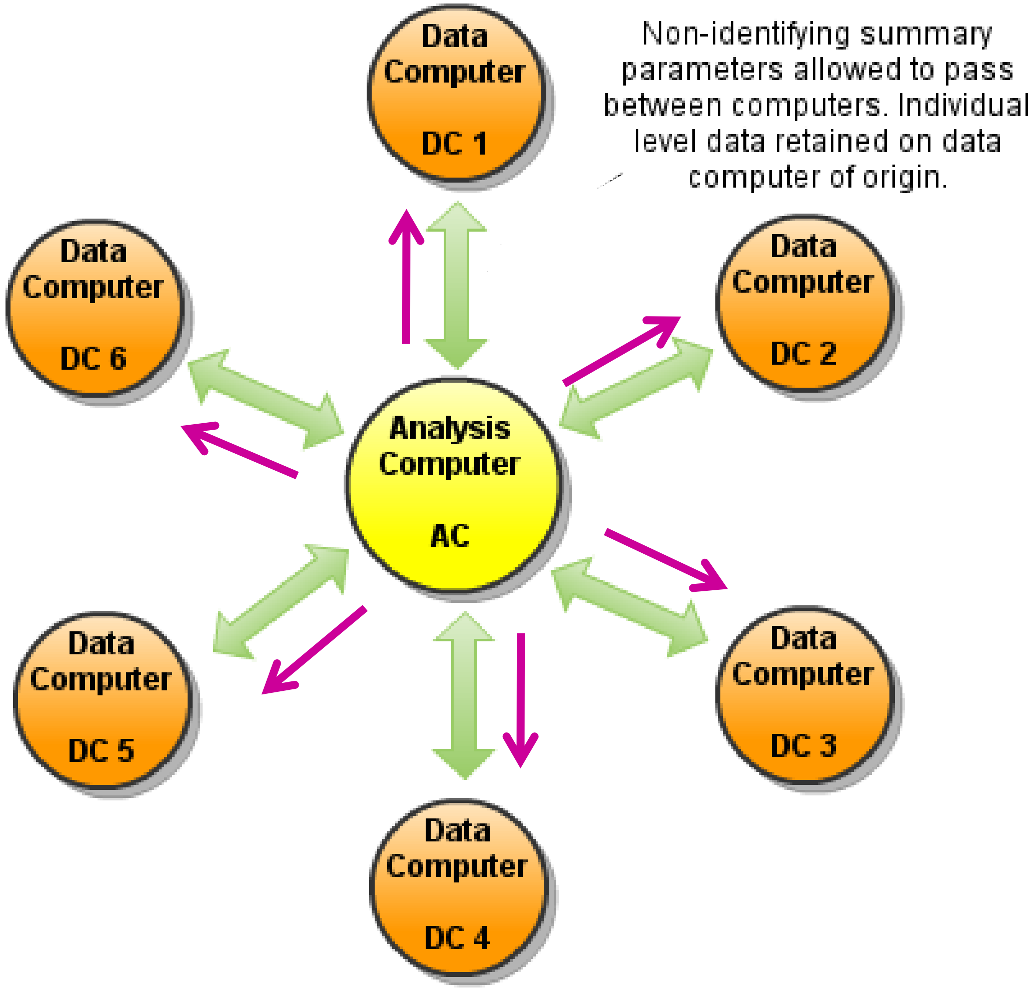
Score vector $_{\text{Study 5}}$

[36, 487.2951, 487.2951, 149]

Information Matrix $_{\text{Study 5}}$

$$\begin{vmatrix} 500 & 70.56657 & 70.56657 & 297 \\ 70.56657 & 7646.29164 & 7646.29164 & 65.39412 \\ 70.56657 & 7646.29164 & 7646.29164 & 65.39412 \\ 297 & 65.39412 & 65.39412 & 382 \end{vmatrix}$$

DataSHIELD: a novel solution

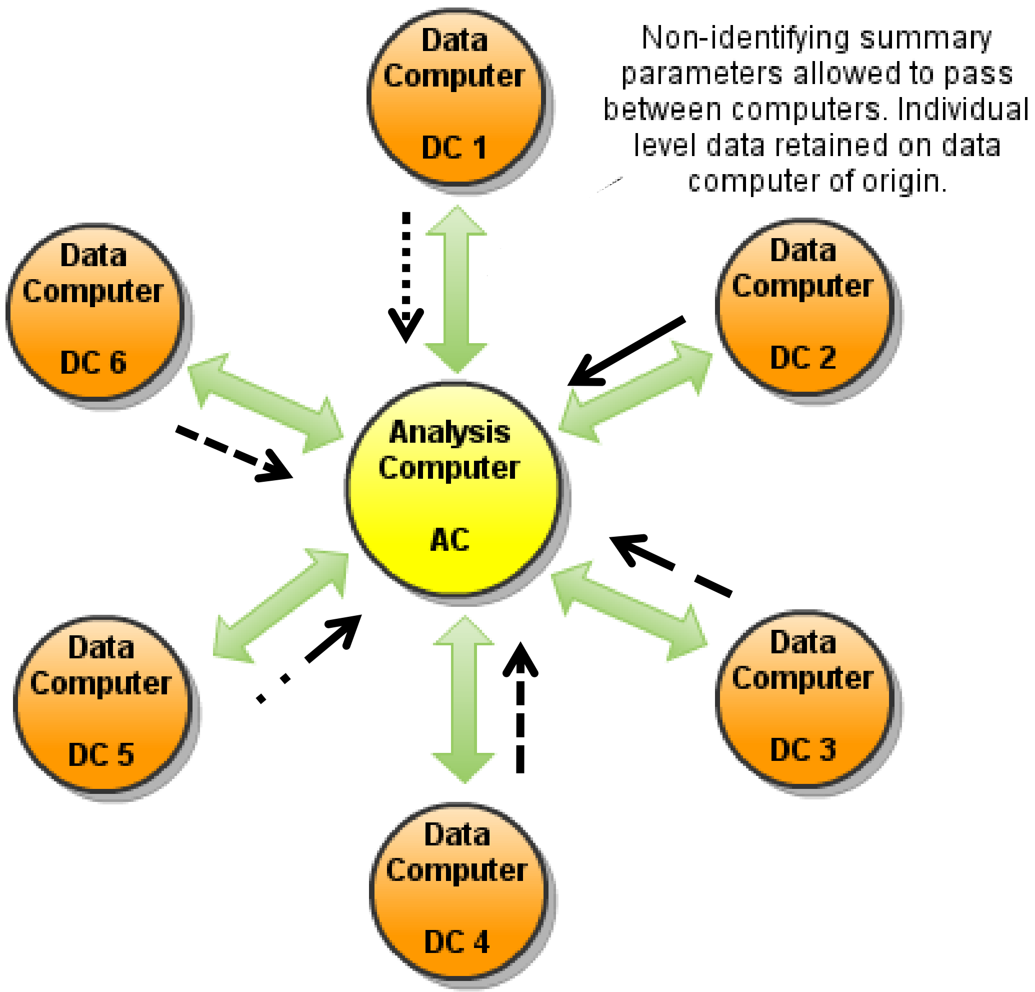


Analysis commands (2)

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

glm(cc~1+sex+snp+bmi,
family=binomial,
start=b.vector, maxit=1)
```

DataSHIELD: a novel solution

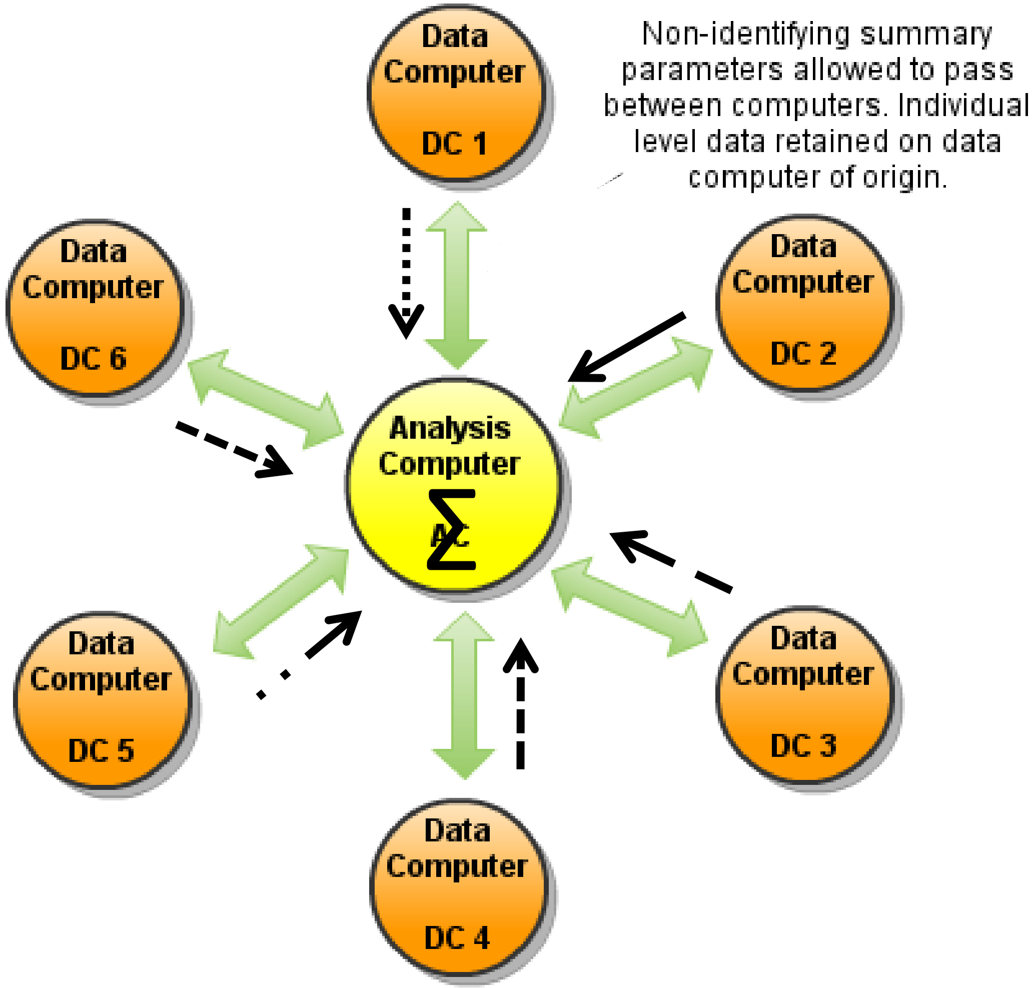


Summary Statistics (2)

Score vectors

Information Matrices

DataSHIELD: a novel solution



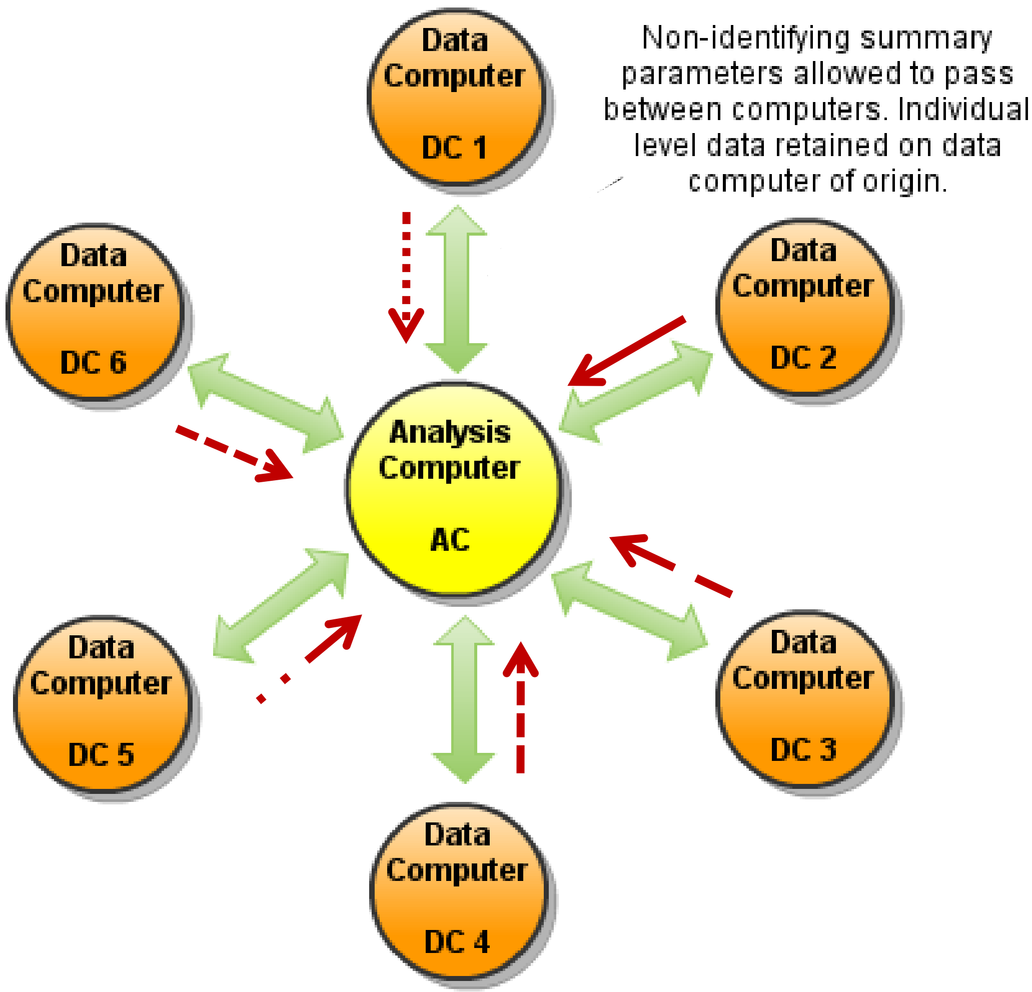
Summary Statistics (2)

Score vectors

Information Matrices

and so on .....

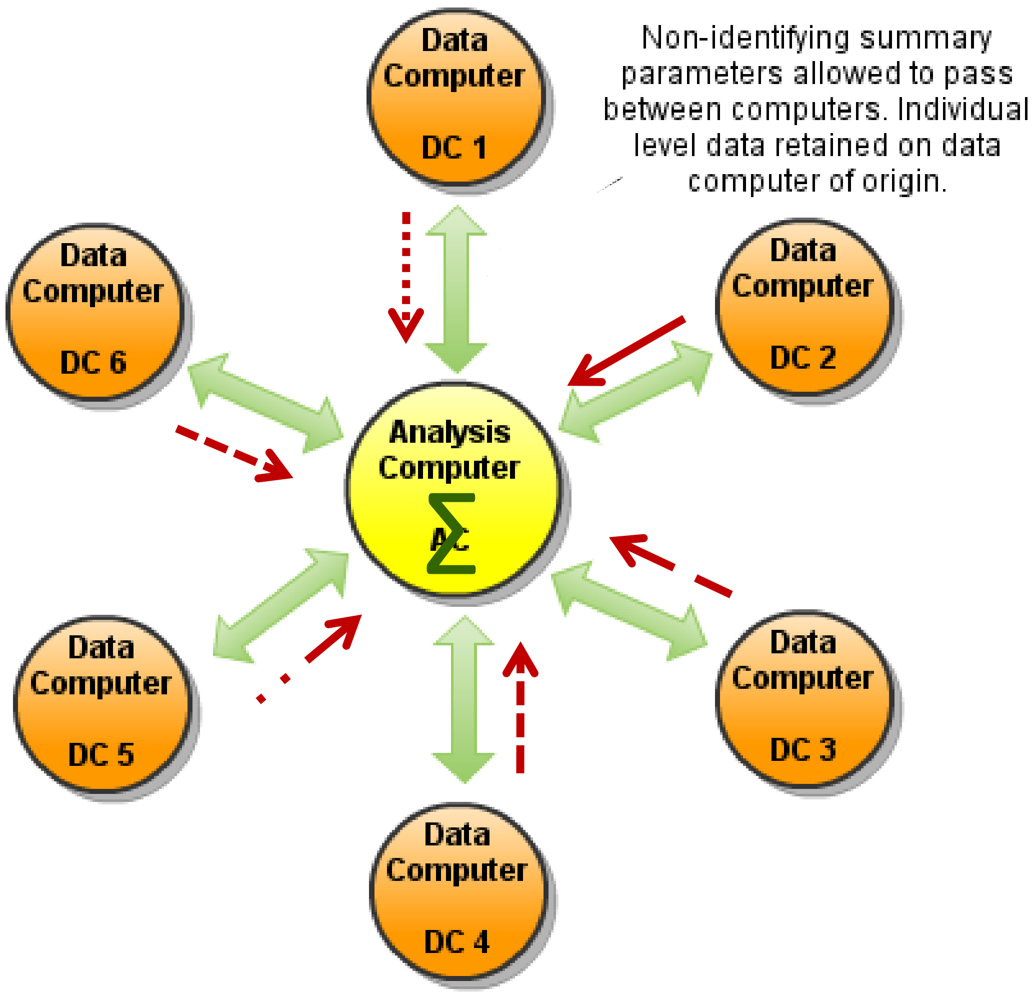
DataSHIELD: a novel solution



Updated parameters (4)

Final parameter estimates

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

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

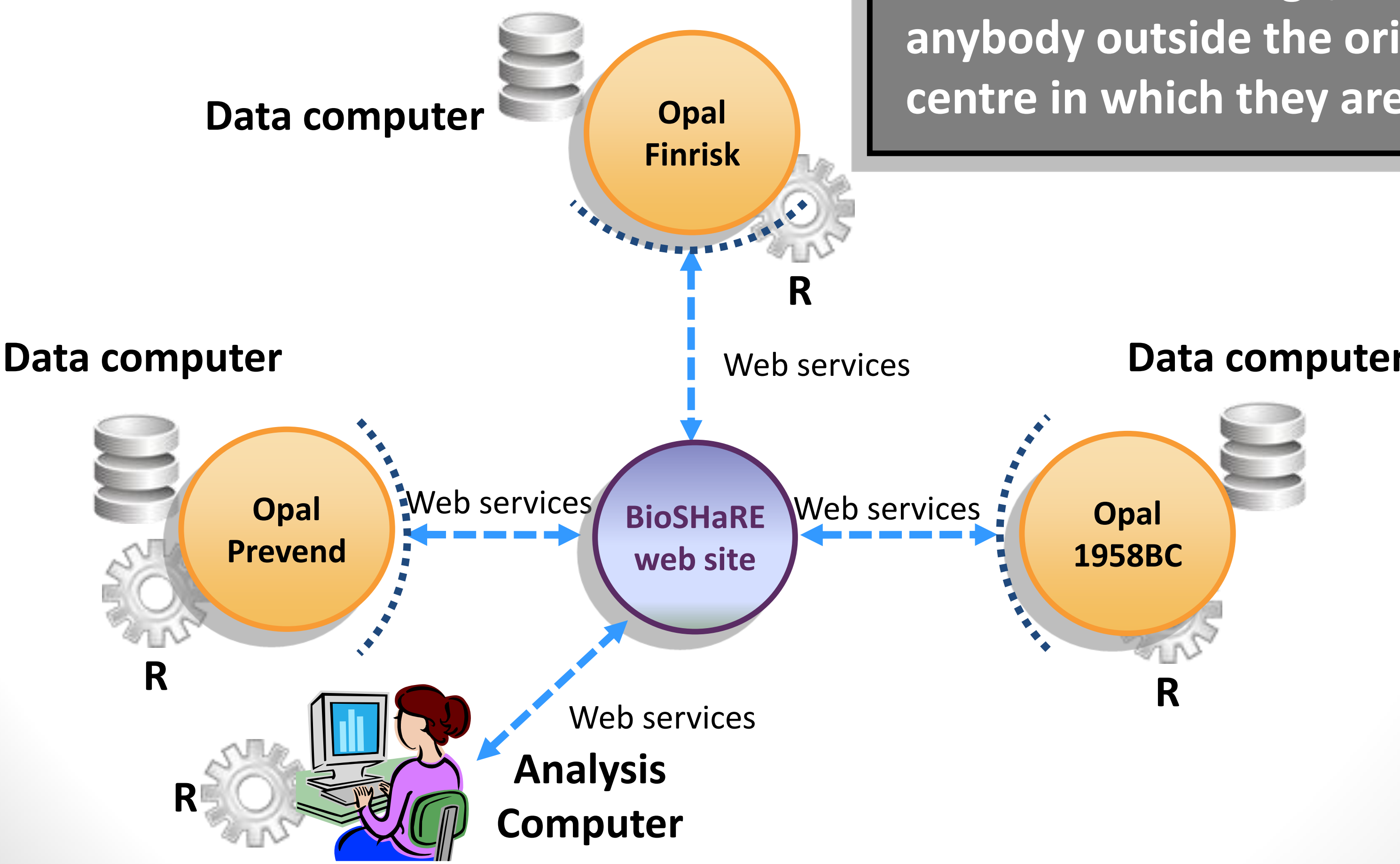
Does it work?

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

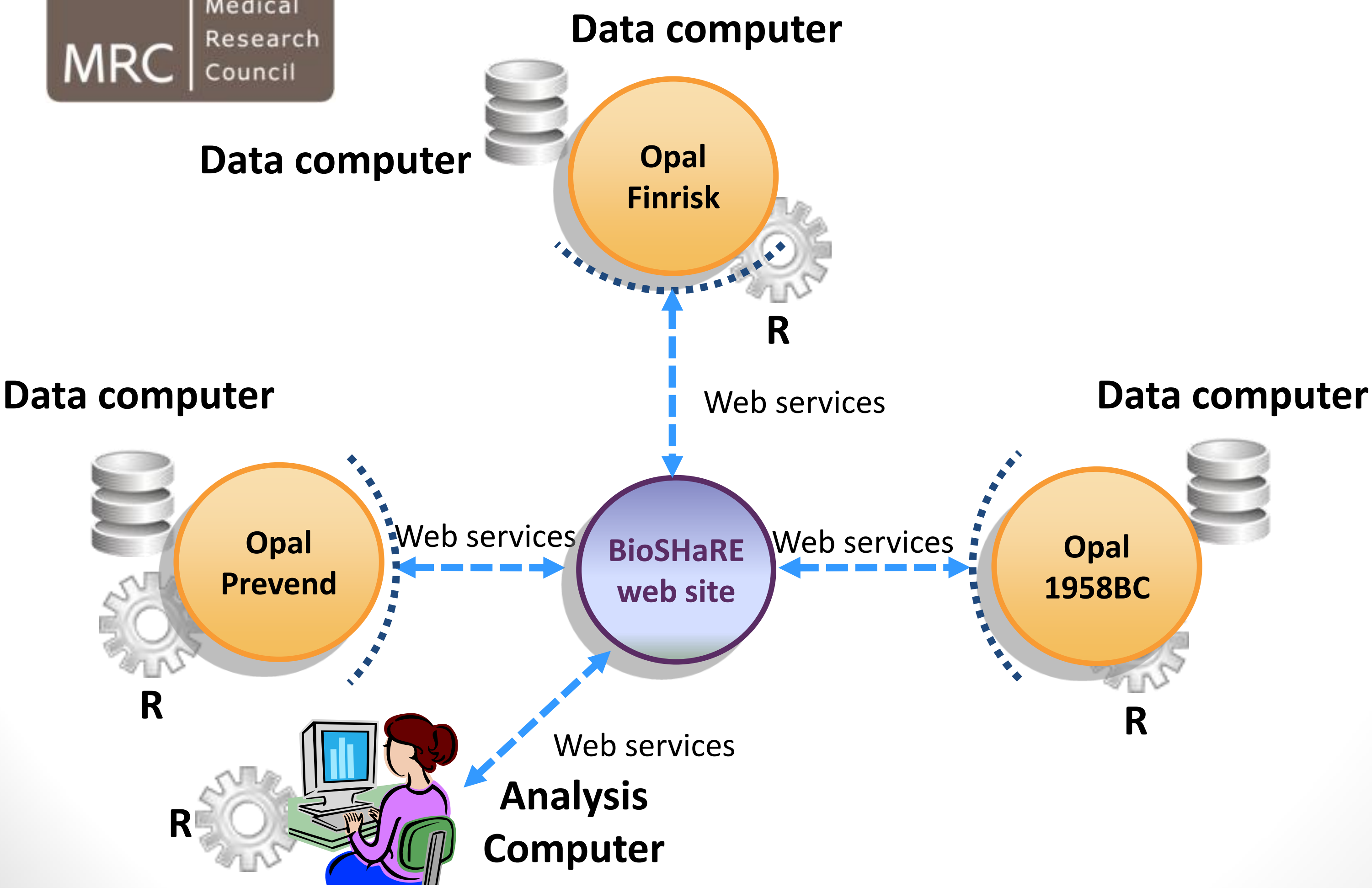
Healthy Obese Project
BioSHARE-eu

Current Implementation

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



Vertical DataSHIELD

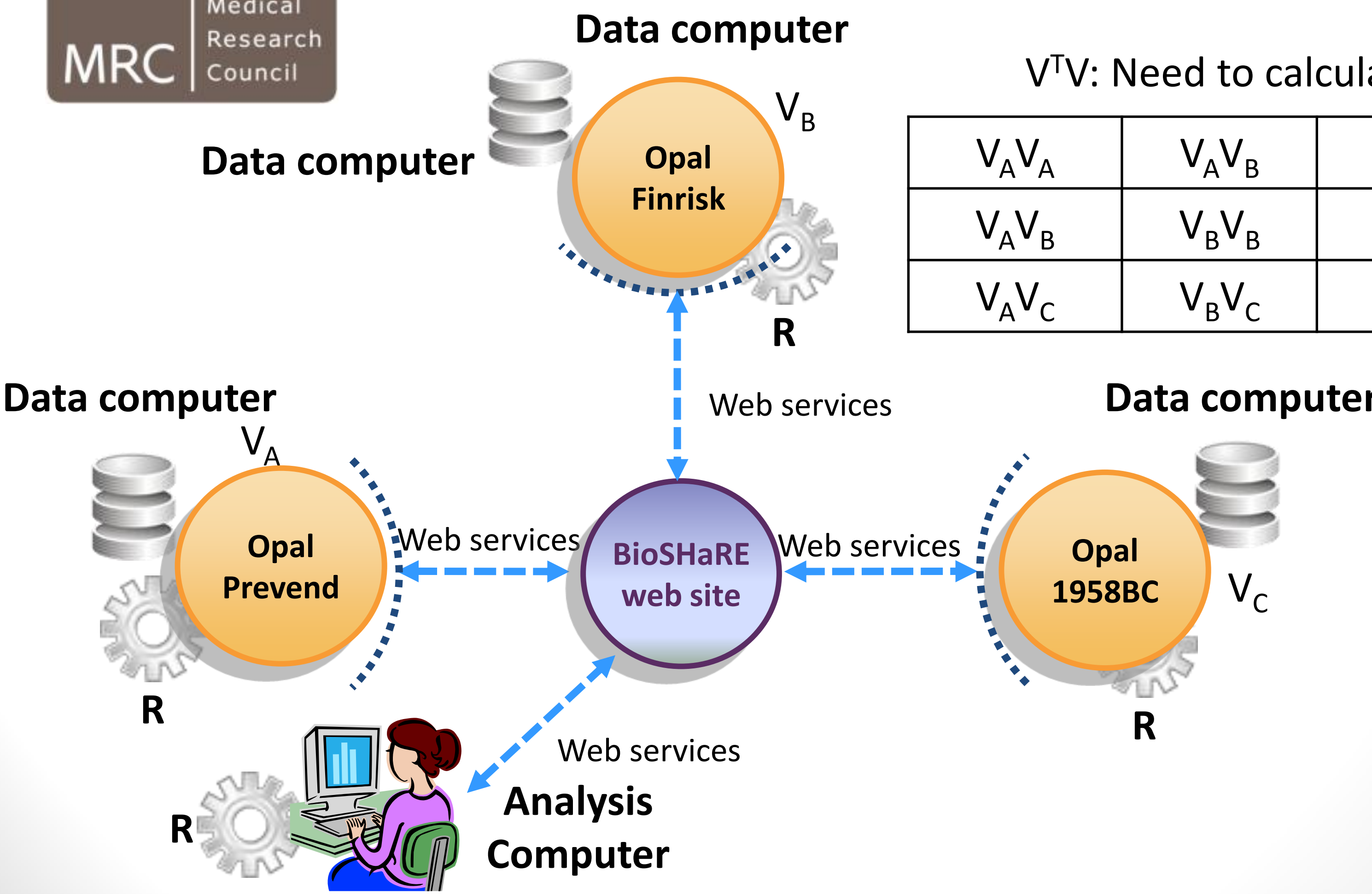


Vertical DataSHIELD

Regression coefficients = V^TY / V^TV

V^TV: Need to calculate

V_AV_A	V_AV_B	V_AV_C
V_AV_B	V_BV_B	V_BV_C
V_AV_C	V_BV_C	V_CV_C

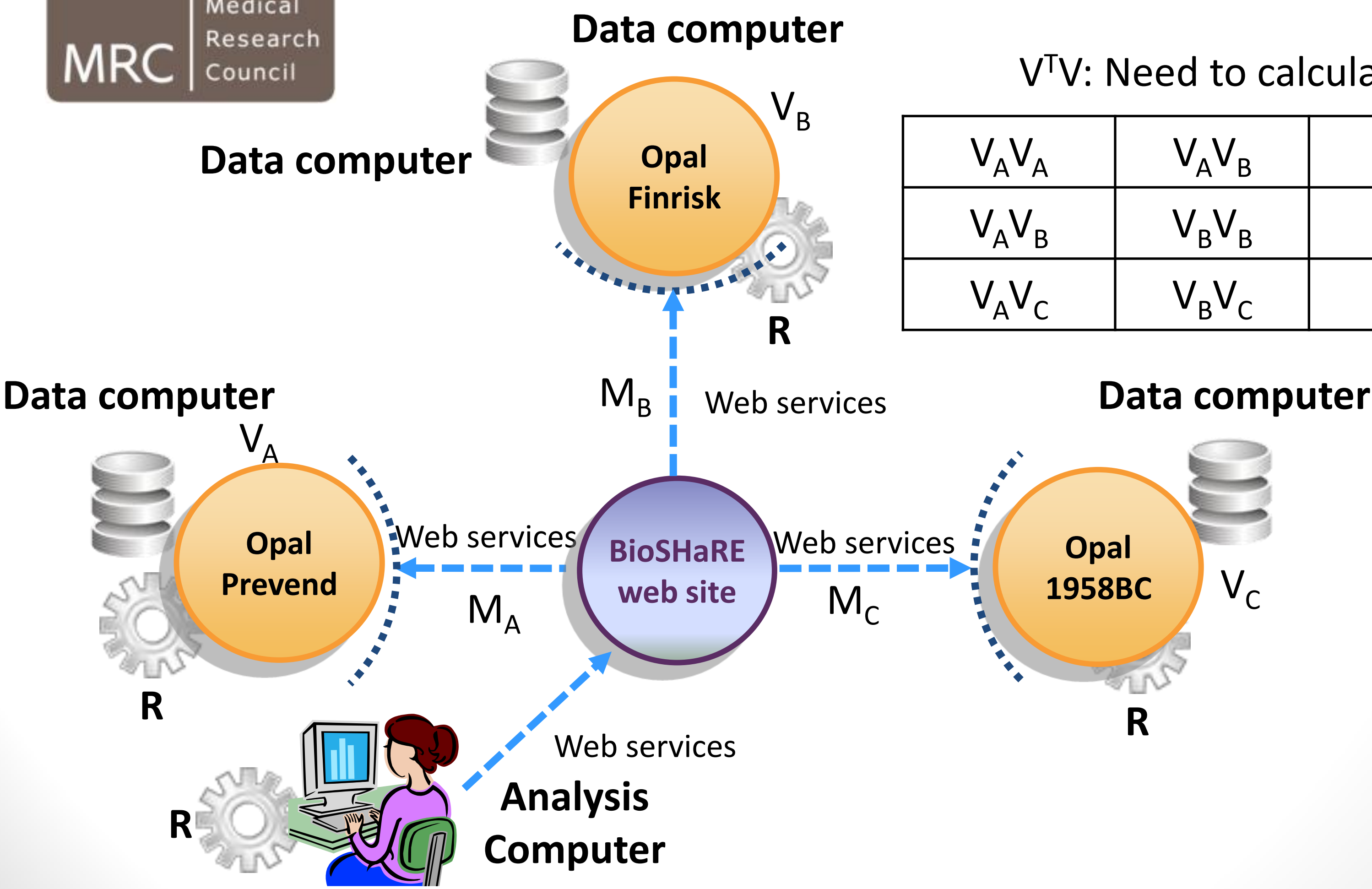


Vertical DataSHIELD

Regression coefficients = V^TY / V^TV

V^TV: Need to calculate

V_AV_A	V_AV_B	V_AV_C
V_AV_B	V_BV_B	V_BV_C
V_AV_C	V_BV_C	V_CV_C

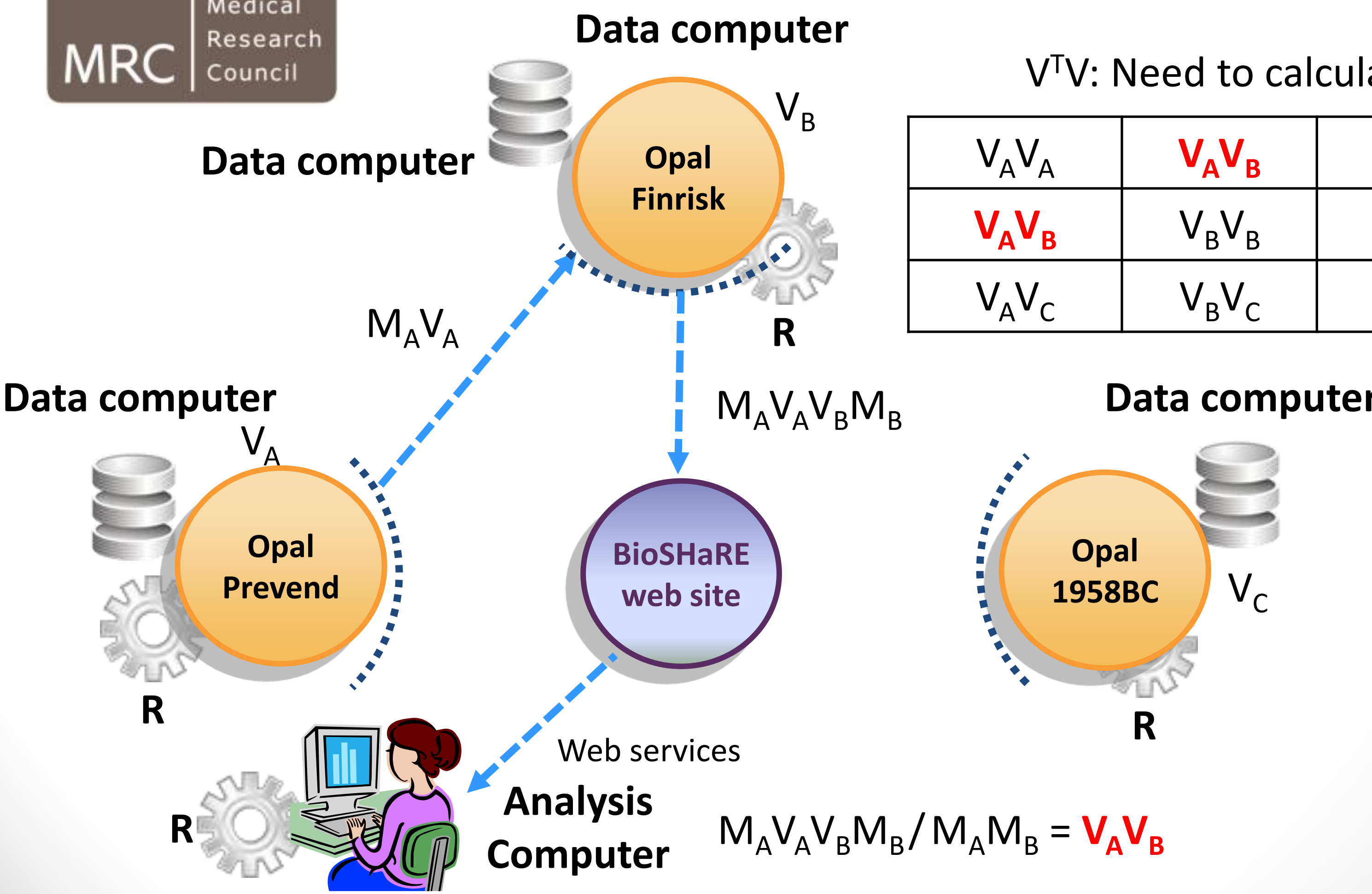


Vertical DataSHIELD

Regression coefficients = V^TY / V^TV

V^TV: Need to calculate

V_AV_A	V_AV_B	V_AV_C
V_AV_B	V_BV_B	V_BV_C
V_AV_C	V_BV_C	V_CV_C



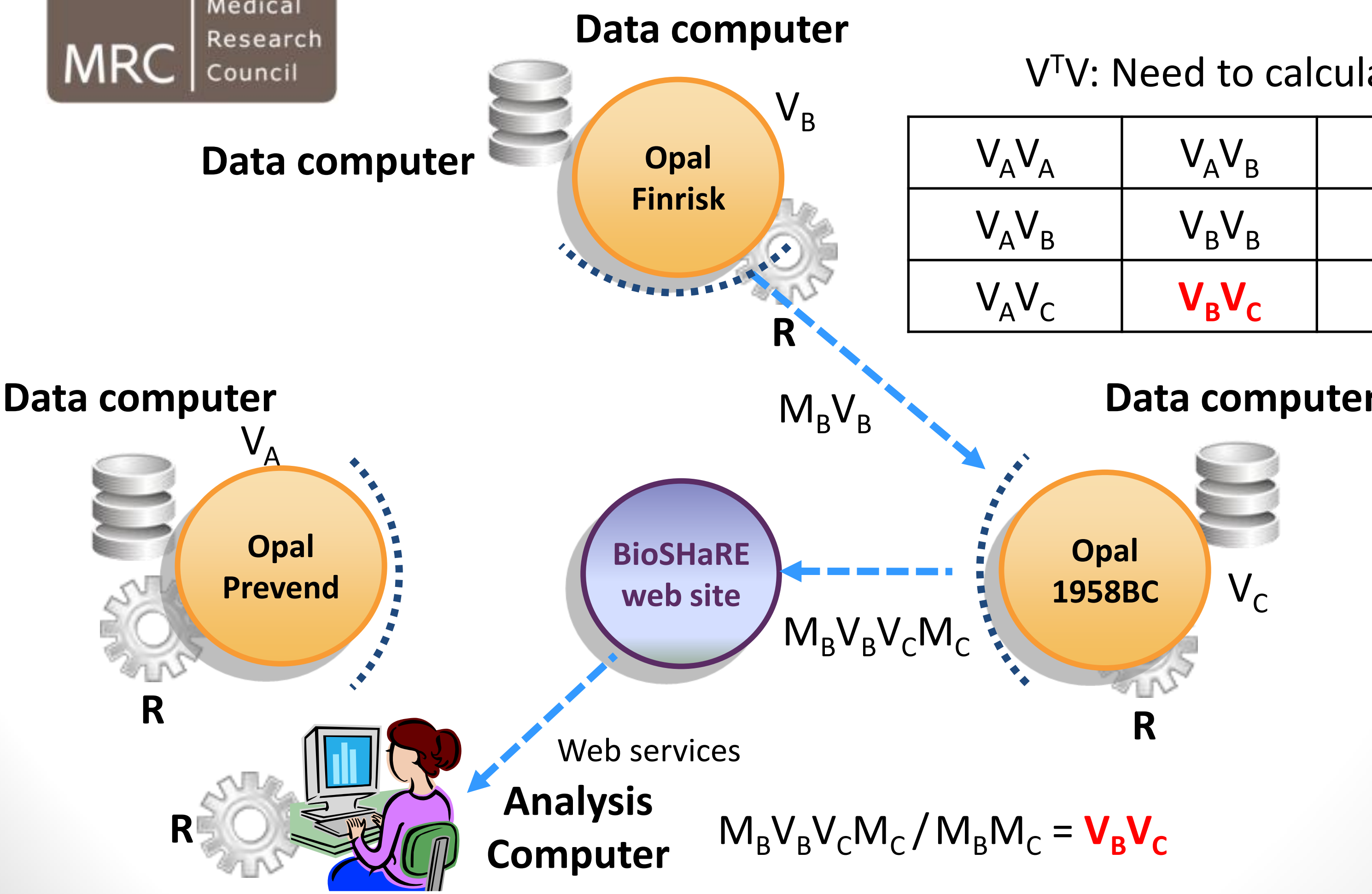
$M_AV_AV_BM_B / M_AM_B = V_AV_B$

Vertical DataSHIELD

Regression coefficients = V^TY / V^TV

V^TV: Need to calculate

V_AV_A	V_AV_B	V_AV_C
V_AV_B	V_BV_B	$\color{red}{V_BV_C}$
V_AV_C	$\color{red}{V_BV_C}$	V_CV_C



$M_BV_BV_CM_C / M_BM_C = \color{red}{V_BV_C}$

www.DataSHIELD.org

DataSHIELD Secure Bioscience Collaboration

Recent Steps

- Healthy Obese Project Analysis Workshop
  - Groningen 16-17 October 2013
- First legal paper in press
  - Wallace et al, 2013
- Active plan for Vertical DataSHIELD Development
  - Record linkage and secure matrix construction
- First thoughts on ‘Omics (particularly Genomics) capability in DataSHIELD

BioSHaRE.eu

D2K data to knowledge

Conclusions

- Many of the issues at the interface between the science/technology and the ELSI are only just starting to be explored
  - Tension between increasing ability to exploit information effectively, and need to secure the original data
- DataSHIELD provides a potential solution to a number of key issues
  - Horizontal for secure meta-analysis
  - Vertical for secure linked analysis
  - Could provide a cheap portable safe haven

Conclusions

- DataSHIELD works in theory (H and V)
- Horizontal works in practice – implemented via R in OPAL
  - BioSHaRE-eu, P^{3}G
  - *e.g.* Healthy obese project
- Vertical about to be implemented also via R in OPAL
  - MRC eHIRCs (CIPHER, Scotland), ALSPAC
- Harmonization CRITICAL
- Must check acceptability of DataSHIELD itself
- WATCH THIS SPACE …..

THANK YOU FOR LISTENING

Public Health
Genomics

Public Health Genomics 2012;15:243–253
DOI: 10.1159/000336673

Published online: June 20, 2012

Securing the Data Economy: Translating Privacy and Enacting Security in the Development of DataSHIELD

M.J. Murtagh[a] I. Demir[a] K.N. Jenkings[a] S.E. Wallace[a] B. Murtagh[a]
M. Boniol[b] M. Bota[b] P. Laflamme[c] P. Boffetta[b, e] V. Ferretti[d] P.R. Burton[a]

[a]Data to Knowledge for Practice, University of Leicester, Leicester, UK; [b]International Prevention Research Institute (iPRI), Lyon, France; [c]Public Population Project in Genomics (P³G), Montreal, Que., and [d]Ontario Institute for Cancer Research, Toronto, Ont., Canada; [e]Mount Sinai School of Medicine, New York, N.Y., USA

DataSHIELD Ethnography

- DataSHIELD as a transdiciplinary study
  - Social implications and practices
  - more on Wednesday in the discussion of the D2K approach

- The ethnographic study
  - Participant observation of meetings, workshops,
  - IPRI, Lyon, 2011
  - Murtagh et al. (2012) ‘Securing the data economy’ – combined proof of concept/social studies of science paper

Ethnography results

Central drivers of DataSHIELD development included:

- **The science**: Scientific development
- **Science in society**: Perceived concerns about privacy and confidentiality
- **The practice of science**: Career progression, funding and intellectual property

Ethnography conclusions

Central drivers of DataSHIELD development included:

- **The science**
  - DataSHIELD works!
- **Science in society**:
  - The DataSHIELD concept elides privacy concerns
    - There are no individual or identifiable data
  - Privacy concerns were transformed into a focus on technical solutions to security issues – malicious use and hacking
- **The practice of science**:
  - ‘Convincing others’
  - Scientific validity - necessary but not sufficient
  - The role of the relational in science
  - This will be our next challenge!