Methods and tools for co-analysing biobanks

Paul Burton

Professor of Genetic Epidemiology
Departments of Health Sciences and Genetics
University of Leicester

SHIP Conference
St Andrews, Sept 11th 2011

DataSHIELD and DataSHaPER

Harmonized analysis of individual-level data without sharing the data

Combined analysis: why bother?

THEORY AND METHODS

Size matters: just how big is BIG?

Quantifying realistic sample size requirements for human genome epidemiology

Paul R Burton,[1,2,3*,†] Anna L Hansell,[4,†] Isabel Fortier,[3,5] Teri A Manolio,[6] Muin J Khoury,[3,7] Julian Little[3,8] and Paul Elliott[4]

- **Sample size**
- **Depth of phenotyping**
- **Quality of measurement**

ALL critical

International Journal of Epidemiology 2009;38:263–273
doi:10.1093/ije/dyn147

How big is BIG? ("typically" and "realistically")

- The direct effect of a gene
  - 2,000 cases minimum, 10,000 cases better
- Environmental and life-style factors
  - Highly context specific
- Gene-lifestyle and gene-gene "interactions"
  - Absolute minimum 10,000, usually need at least 20,000, a comprehensive platform needs at least 50,000
  - Scientifically fundamental

The bottom line

- Effective access to well-measured data is vital
- Efficient valid analysis is crucial
- Effective combined analysis is also important
- Two fundamental challenges
  - Streamlined ethico-legally acceptable access to multiple data sets
  - Scientific harmonization
- New systems, methods and tools
  - DataSHIELD, DataSHaPER

Combined analysis *in practice*

Seven steps to combined analysis

- Define scientific question
- Identify information required
  - What is needed?
  - On how many people?

- Use catalogues to identify studies/cohorts/biobanks or biorepositories that can service the informational needs
- Develop “DataSchema” and “formal pairing rules”

- Create algorithms to generate required data in each study
- Generate the harmonized data sets
- Undertake an *appropriate* joint analysis (science and ELSI)

DataSHIELD

http://www.p3gobservatory.org/

- Study catalogue
  - 162 studies
  - More than 12.5m participants
- Questionnaire catalogue
  - 76 questionnaires (cut and key-worded)
  - From 48 studies

http://www.datashaper.org/
Dr Isabel Fortier – University of McGill

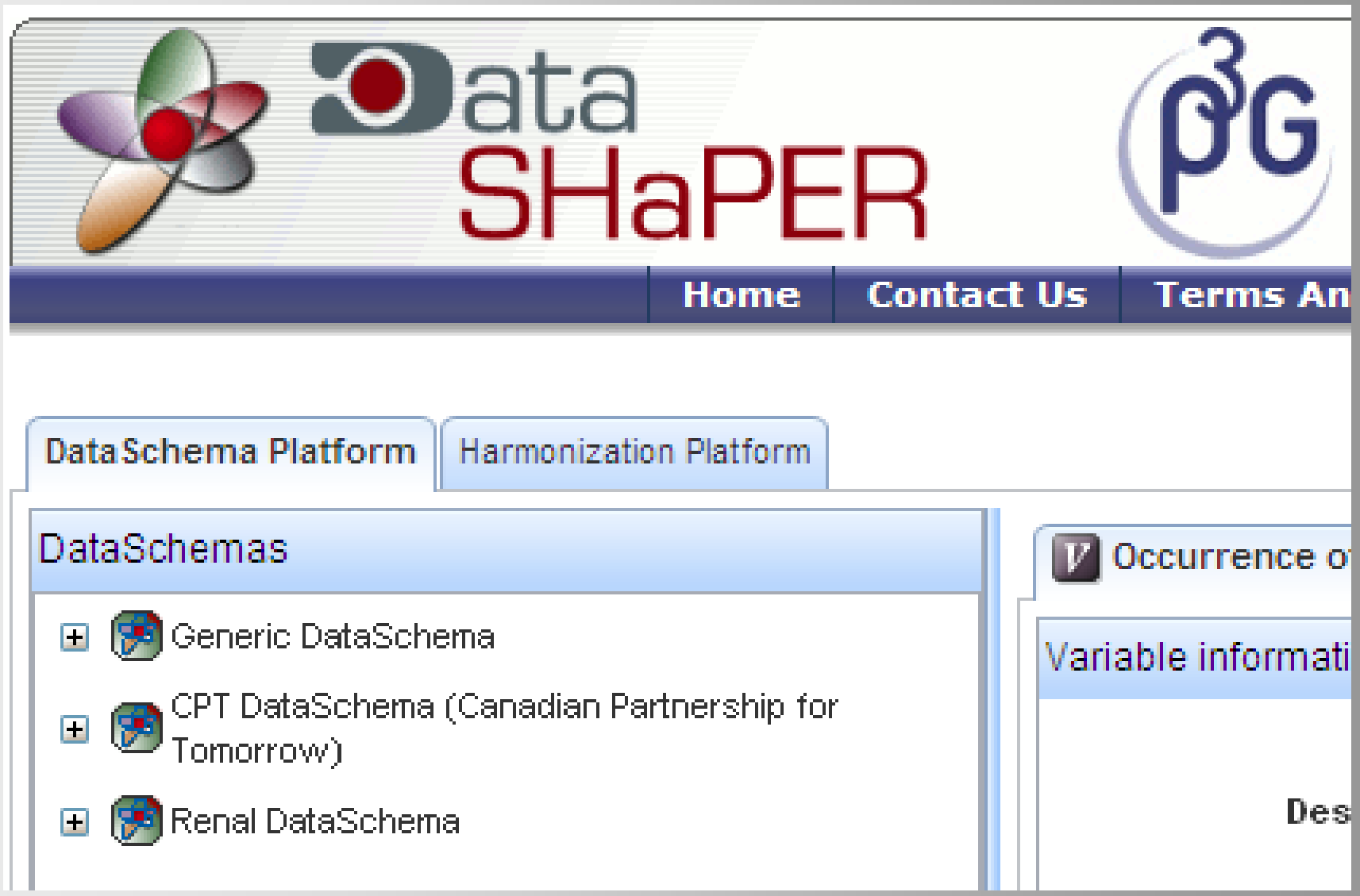
WELCOME TO DATASHAPER WEBSITE

The DataSHaPER (Data Schema and Harmonization Platform for Epidemiological Research) is both a scientific approach and a suite of practical tools. Its primary aims are to facilitate the prospective harmonization of emerging biobanks, provide a template for retrospective synthesis and support the development of questionnaires and information-collection devices, even when pooling of data with other biobanks is not foreseen.

- **What is the DataSHaPER?**

The development of these tools has been jointly funded by **P³G**, **CPT**, **PHOEBE**, and **Generation Scotland**.

A DataSchema identifies and describes a thematic set of core variables that are of particular value in a specified scientific setting. It also contains associated support material including variable definitions, links to relevant Ontologies and Classifications, and access to reference questionnaires and operating procedures.


Data SHaPER
Home
Contact Us
Terms An
DataSchema Platform
Harmonization Platform
DataSchemas
Generic DataSchema
CPT DataSchema (Canadian Partnership for Tomorrow)
Renal DataSchema
Occurrence o
Variable informati
Des

DATASCHEMA PLATFORM

A DataSchema contains a thematic set of core variables that are of significant relevance to the specific scientific area addressed by a specific study or group of studies.

The core variables in each DataSchema are grouped a four level nested hierarchy:

1. Module:

Assessment modes or type of element measured or collected. Each module subsumes one or more themes.

2. Theme:

General area of interest. Each theme subsumes one or more domains.

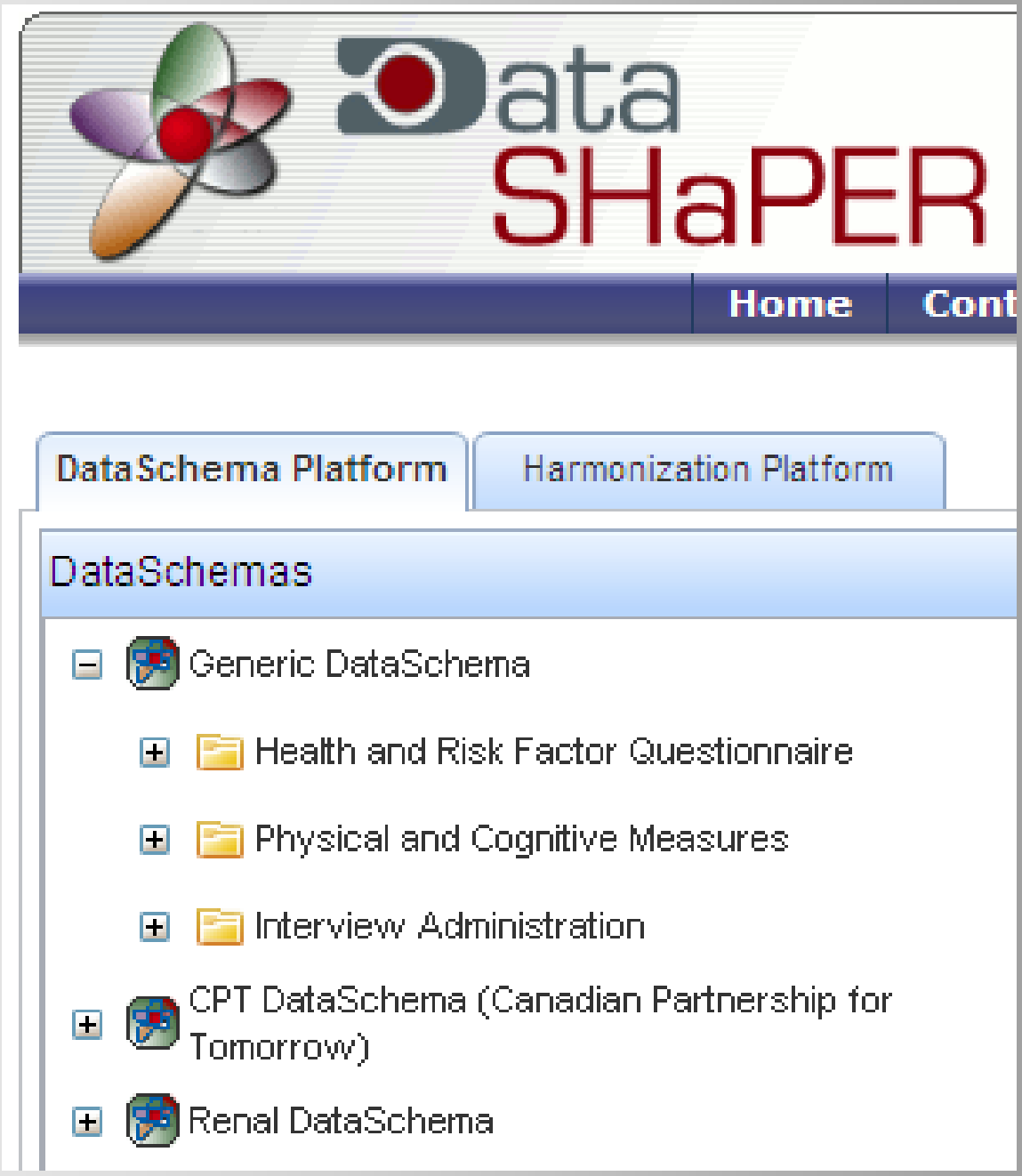
3. Domain:

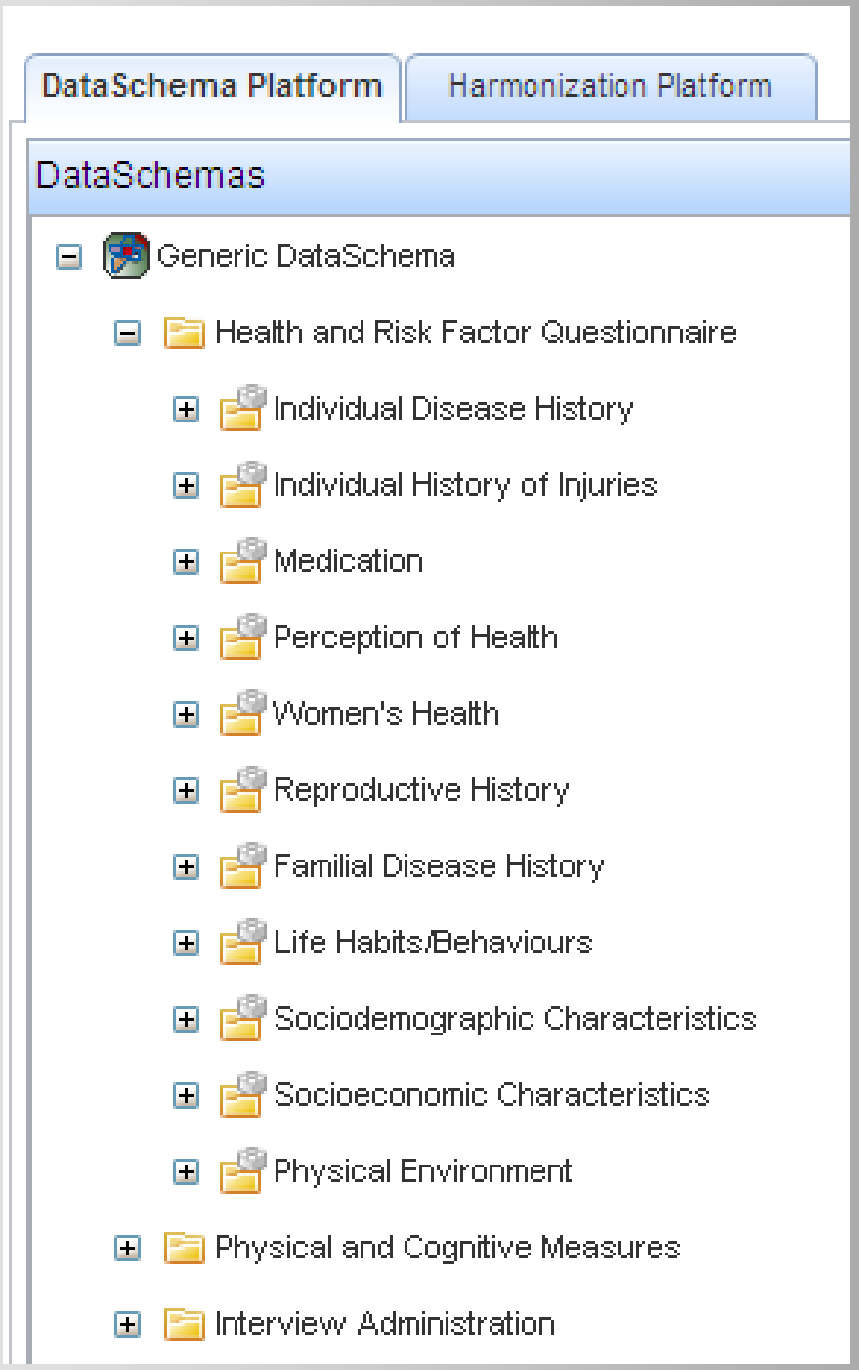
Risk factor or outcome of interest. Each domain subsumes a number of variables.

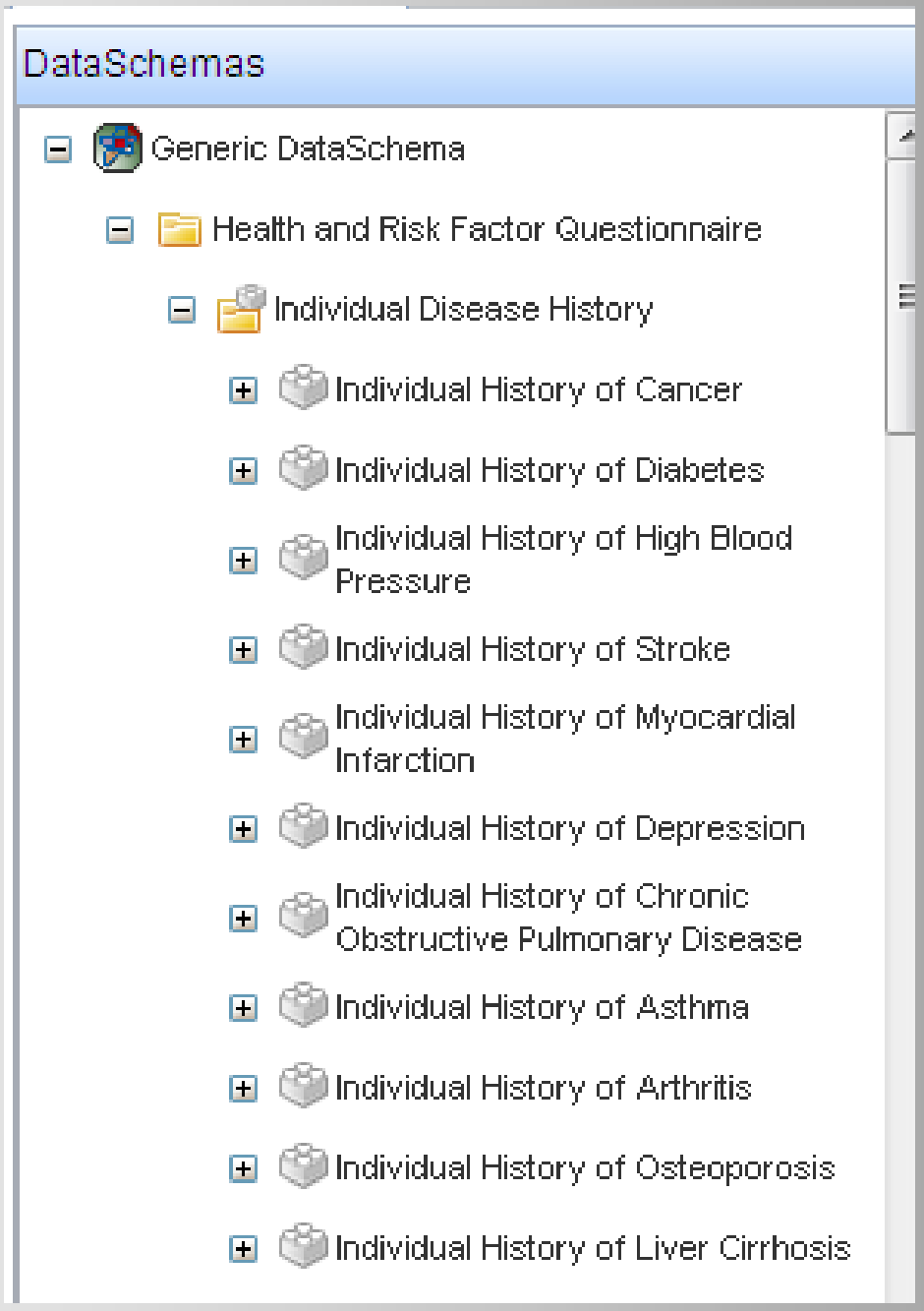
4. Variable:

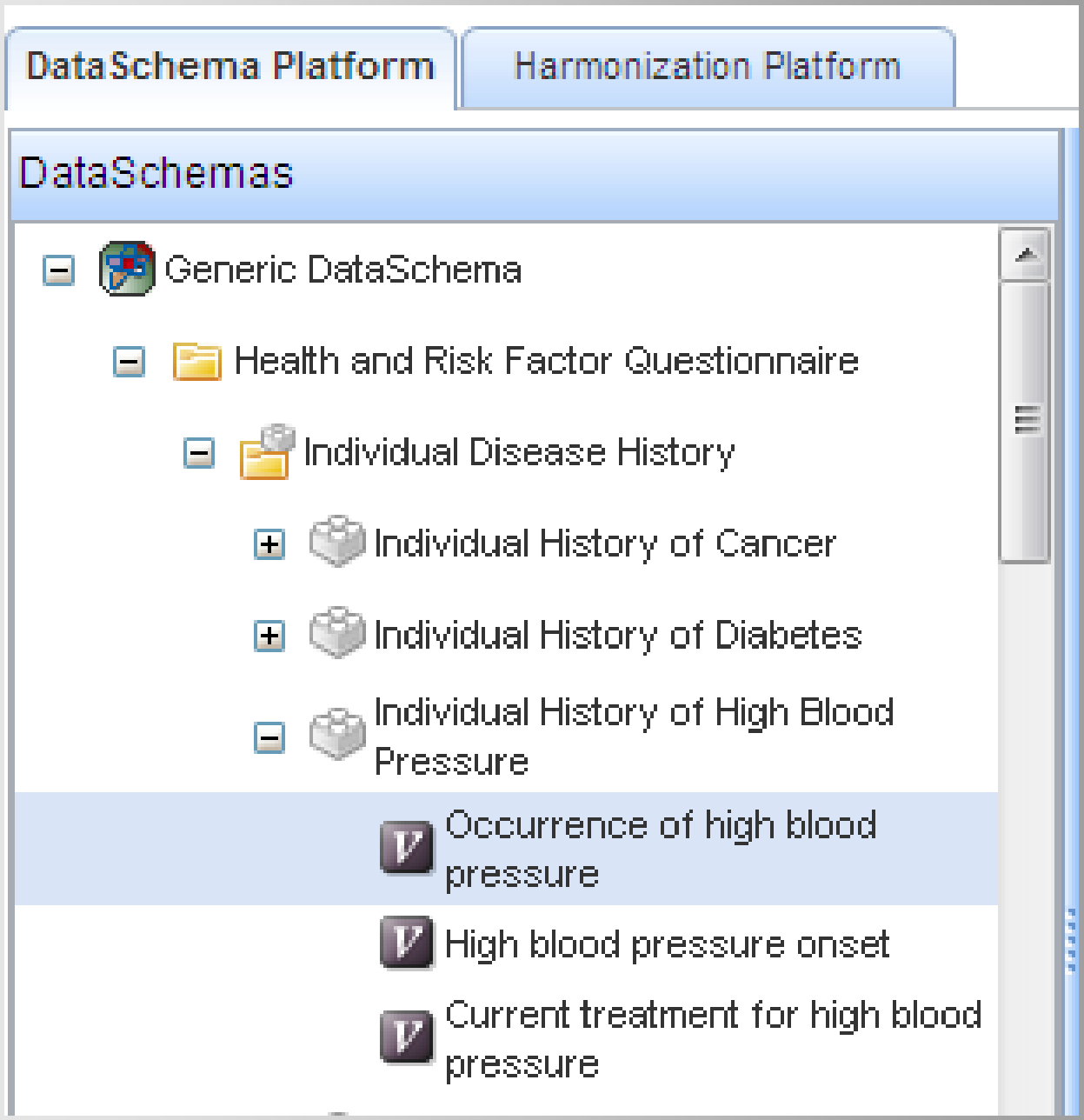
Primary unit of interest for a statistical analysis.

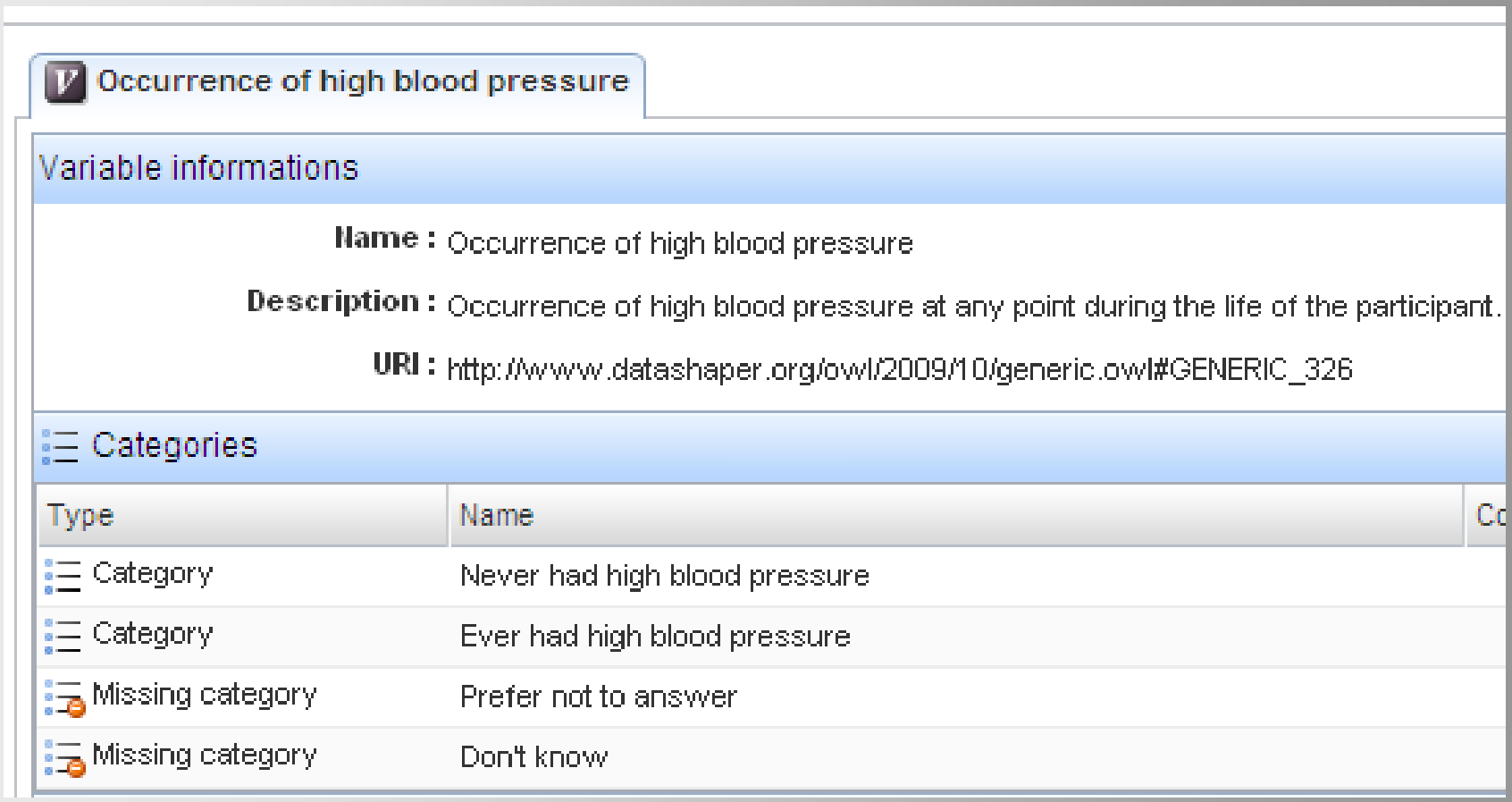
Each DataSchema contains, where relevant, descriptive information to complements the list of variables itself, such as variable definitions, links to relevant standard classifications, reference questionnaires, operating procedures and so on.


Data SHaPER
Home
Cont
DataSchema Platform
Harmonization Platform
DataSchemas
Generic DataSchema
Health and Risk Factor Questionnaire
Physical and Cognitive Measures
Interview Administration
CPT DataSchema (Canadian Partnership for Tomorrow)
Renal DataSchema


DataSchema Platform
Harmonization Platform
DataSchemas
Generic DataSchema
Health and Risk Factor Questionnaire
Individual Disease History
Individual History of Injuries
Medication
Perception of Health
Women's Health
Reproductive History
Familial Disease History
Life Habits/Behaviours
Sociodemographic Characteristics
Socioeconomic Characteristics
Physical Environment
Physical and Cognitive Measures
Interview Administration


DataSchemas
Generic DataSchema
Health and Risk Factor Questionnaire
Individual Disease History
Individual History of Cancer
Individual History of Diabetes
Individual History of High Blood Pressure
Individual History of Stroke
Individual History of Myocardial Infarction
Individual History of Depression
Individual History of Chronic Obstructive Pulmonary Disease
Individual History of Asthma
Individual History of Arthritis
Individual History of Osteoporosis
Individual History of Liver Cirrhosis


DataSchema Platform
Harmonization Platform
DataSchemas
Generic DataSchema
Health and Risk Factor Questionnaire
Individual Disease History
Individual History of Cancer
Individual History of Diabetes
Individual History of High Blood Pressure
Occurrence of high blood pressure
High blood pressure onset
Current treatment for high blood pressure


Occurrence of high blood pressure
Variable informations
Name : Occurrence of high blood pressure
Description : Occurrence of high blood pressure at any point during the life of the participant.
URI : http://www.datashaper.org/owl/2009/10/generic.owl#GENERIC_326
Categories
Type
Name
Category
Never had high blood pressure
Category
Ever had high blood pressure
Missing category
Prefer not to answer
Missing category
Don't know

The Harmonization Platform provides a template for the formal estimat ion of the potential to synthe size information between studies. At present, access to the Harmonization Platform is limited to collaborative context.

DataSchema Platform | Harmonization Platform

HARMONIZATION PLATFORM

The process follows a rigorous approach including:

1. The development of rules providing a formal assessment of the potential for each individual study to generate each of the variables in a given DataSchema.
2. The application of these rules to determine and tabulate the ability of each study to generate each variable, thereby identifying the information that can be shared.

	Study A	Study B	Study C	Study D
Variable 1	IMPOSSIBLE	PARTIAL	COMPLETE	COMPLETE
Variable 2	COMPLETE	COMPLETE	COMPLETE	COMPLETE
Variable 3	PARTIAL	IMPOSSIBLE	IMPOSSIBLE	IMPOSSIBLE

3. The development and application of a processing algorithm enabling that study to generate the required variable in an appropriate form.
4. Access to the Harmonization Platform is limited to collaborative context. Please contact us to see how we can work together.

Evaluating what can be harmonized

	KORA-gen	LifeGene	UK Biobank
Variable	**Match**	**Match**	**Match**
Occurrence of diabetes	Complete	Complete	Complete
Current treatment for high blood pressure	Complete	Impossible	Complete
Occurrence of diabetes in the family	Impossible	Impossible	Partial
Occurrence of high blood pressure in the family	Partial	Impossible	Partial
Current cigarette smoker	Complete	Partial	Complete
Current quantity of cigarettes smoked	Complete	Complete	Complete
Household income	Partial	Impossible	Partial
Country of birth-ISO-3166	Complete	Complete	Complete
Gender	Complete	Complete	Complete
Age at recruitment	Complete	Complete	Complete
Standing height	Complete	Complete	Complete
Weight	Complete	Complete	Complete
Waist circumference	Complete	Complete	Complete
Hip circumference	Complete	Complete	Complete
Body mass index	Complete	Complete	Complete
Diastolic blood pressure at rest	Complete	Complete	Complete
Systolic blood pressure at rest	Complete	Complete	Complete
Date of interview	Complete	Complete	Complete

Occurrence of high blood pressure

Target

Participant

Disease

High blood pressure

Hypertension

Source of information

Participant

Period

Through all life

Format

Categorical

Collection mode

Questionnaire

Class

Occurrence

Ontology references

URI
http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#Hypertension

Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies

Isabel Fortier,[1,2]* Paul R Burton,[1,3] Paula J Robson,[4] Vincent Ferretti,[5] Julian Little,[1,6] Francois L'Heureux,[1] Mylène Deschênes,[1] Bartha M Knoppers,[1,7] Dany Doiron,[1] Joost C Keers,[8] Pamela Linksted,[9] Jennifer R Harris,[10] Geneviève Lachance,[1] Catherine Boileau,[11] Nancy L Pedersen,[12] Carol M Hamilton,[13] Kristian Hveem,[14] Marilyn J Borugian,[15,16] Richard P Gallagher,[15,16] John McLaughlin,[17] Louise Parker,[18] John D Potter,[19] John Gallacher,[20] Rudolf Kaaks,[21] Bette Liu,[22] Tim Sprosen,[23] Anne Vilain,[1] Susan A Atkinson,[3] Andrea Rengifo,[3] Robin Morton,[9] Andres Metspalu,[24] H Erich Wichmann,[25,26,27] Mark Tremblay,[28,29] Rex L Chisholm,[30] Andrés Garcia-Montero,[31] Hans Hillege,[32] Jan-Eric Litton,[33] Lyle J Palmer,[34] Markus Perola,[35,36] Bruce HR Wolffenbuttel,[37] Leena Peltonen[38] and Thomas J Hudson[5,39,40]

International Journal of Epidemiology 2010;1–11
doi:10.1093/ije/dyq139

Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies

Isabel Fortier,[1,2]* Dany Doiron,[2] Julian Little,[1,3] Vincent Ferretti,[4] François L'Heureux,[2] Ronald P Stolk,[5] Bartha M Knoppers,[1,6] Thomas J Hudson,[4,7,8] and Paul R Burton[2,9,10] on behalf of the International Harmonization Initiative†

International Journal of Epidemiology 2011;1–15
doi:10.1093/ije/dyr106

Table 1 Pairing results (%) for selected variables presenting a high proportion of 'Complete/Partial Proximate' matches*

Variable name	Complete/Partial proximate (%)
Occurrence of diabetes	89
Current use of alcohol	89
Level of physical activity	85
Occurrence of high blood pressure	83
Occurrence of menopause	83
Occurrence of cancer	81
Occurrence of stroke	81
Employment status	81
Menopause onset	75
Occurrence of asthma	74
Current quantity of cigarettes smoked	74
Occurrence of myocardial infarction	72
Living with partner	68
Type of cancer	66
Standing height	66
Weight	66

*For full-pairing results, please refer to supplementary ials (Supplementary data available at *IJE* online).

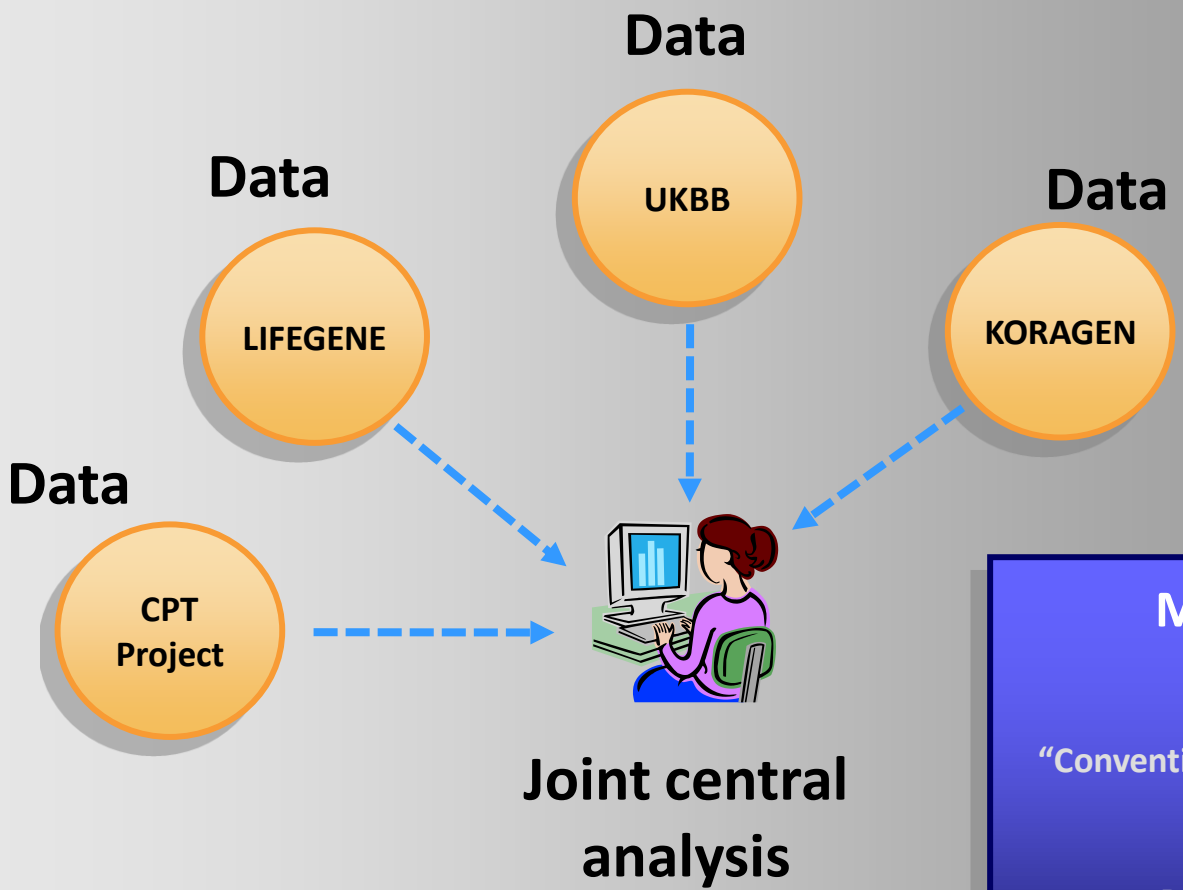
Table 3 Univariate and multivariate models; variable characteristics and pairing results

		Average across studies of the percentage of variables presenting as:			
	Number of variables	Complete match	Partial Proximate match	Partial Tentative match	Impossible match
Univariate analysis					
Variable importance					
Essential (1)	38	60	2	15	23
Important (2)	45	38	5	13	44
Useful (3)	65	21	1	14	64
Targeted individual					
Participant (0)	101	46	3	12	39
Participant's family members (1)	47	14	1	19	66
Targeted period					
Current status (0)	40	50	4	11	36
All other periods (1)	108	31	2	15	52

DataSHIELD: full joint analysis without physically sharing the data

With: Philippe Laflamme
Vincent Ferretti

Horizontally partitioned data



METHODS

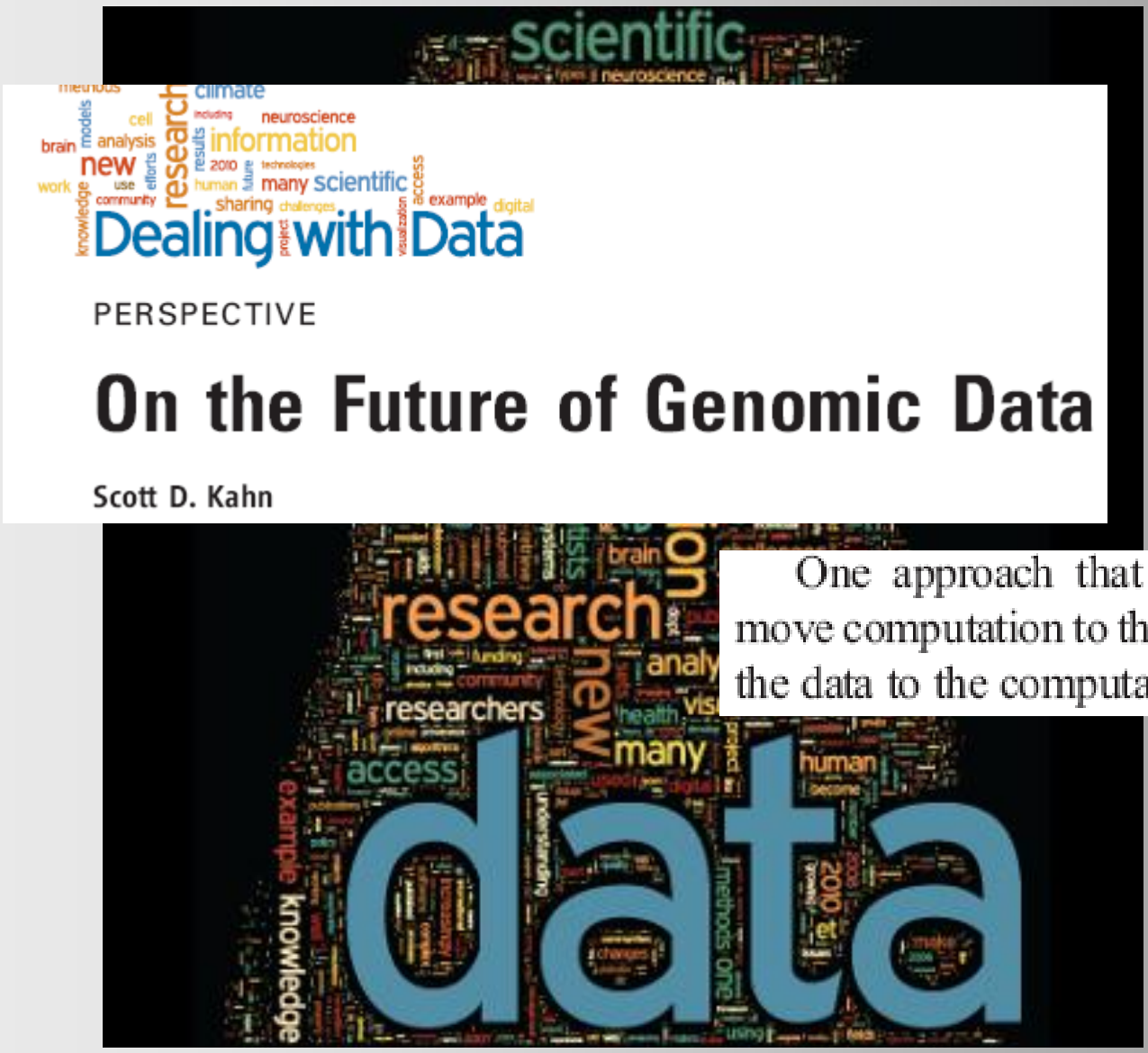
SLMA
"Conventional" Meta-analysis

ILMA
Direct pooling

DataSHIELD
Parallelized analysis

A real and growing problem

- How can we undertake a joint analysis using multiple data sources if the data cannot physically be pooled?
  - Ethico-legal constraints
  - Physical size of the data objects
  - Intellectual property issues



One approach that is being explored is to move computation to the data rather than moving the data to the computation.

DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data

Michael Wolfson,[1] Susan E Wallace,[2,3] Nicholas Masca,[4] Geoff Rowe,[1] Nuala A Sheehan,[4] Vincent Ferretti,[3,5] Philippe LaFlamme,[3,6] Martin D Tobin,[4] John Macleod,[7] Julian Little,[3,8] Isabel Fortier,[3,8,9] Bartha M Knoppers[2,3] and Paul R Burton[3,4,8,10]*

International Journal of Epidemiology 2010;39:1372–1382

BioSHaRE.eu

CHILDREN OF THE 90S

p3g

Two approaches to data synthesis

- Study level meta-analysis (SLMA)
  - Estimate a summary statistic from each study. Calculate an appropriately weighted mean and standard error 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

- If summary statistics available, quick, cheap, and convenient (*e.g.* from published literature)
  - RCTs, Health Care Assessment *etc*
- If the required summary statistics are straightforward to define and easy to calculate, also very convenient
  - Genome Wide Association Studies (GWASs)

Study level meta-analysis

- Quick, easy and it works
- 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

ELSI restrictions on sharing

- Exemplar wording

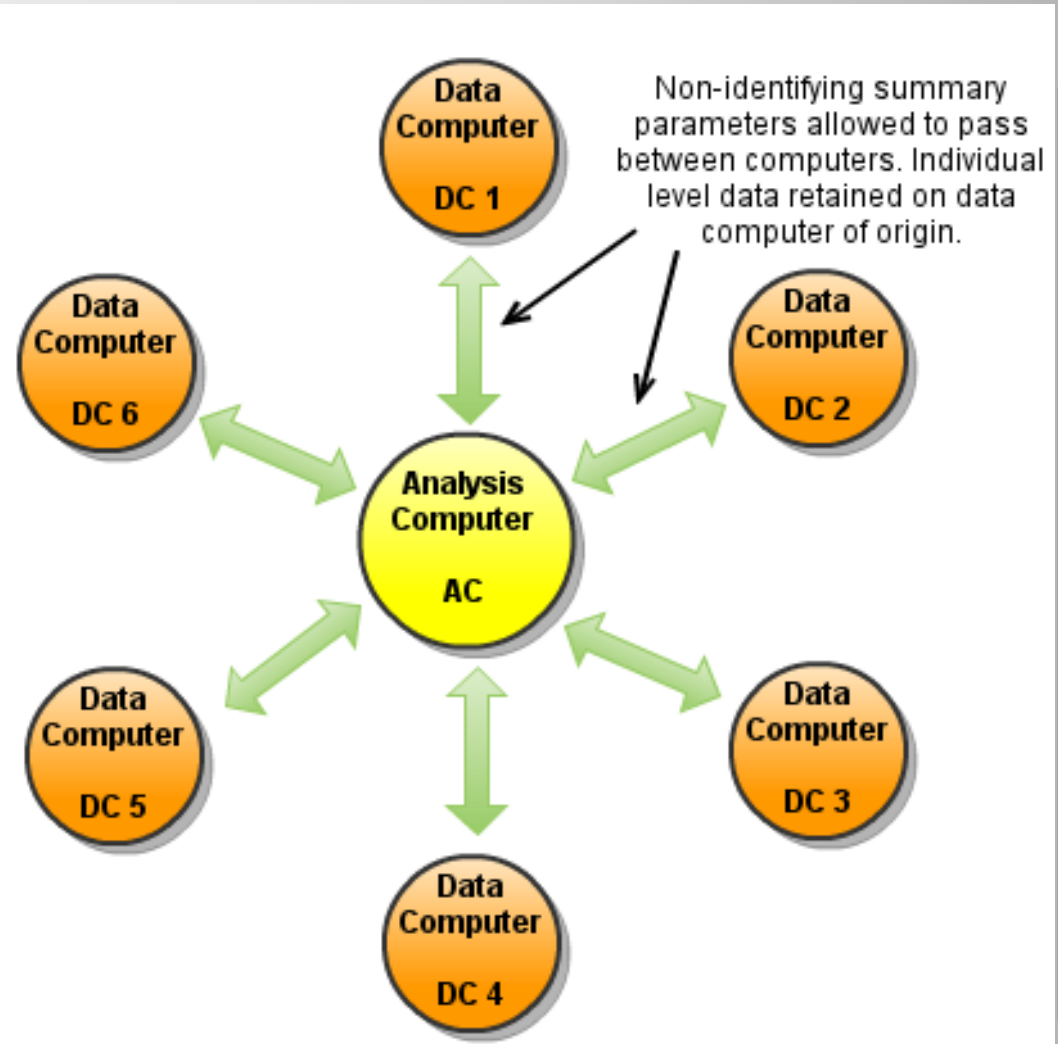
  Wallace S, Lazor S, Knoppers BM. Chapter in Kaye J and Stranger M. Principles and Practice in Biobank Governance. Ashgate, Farnham 2009

- 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 a protracted and time consuming process
  - Often needs multiple clearances

Competing public goods

- Analytic flexibility greatly favours ILMA
- ELSI can prohibit or discourage ILMA
  - → Most current GWASs based on SLMA
  - BUT: this situation is not sustainable as things become more complex, unpredictable and exploratory

DataSHIELD: a novel solution

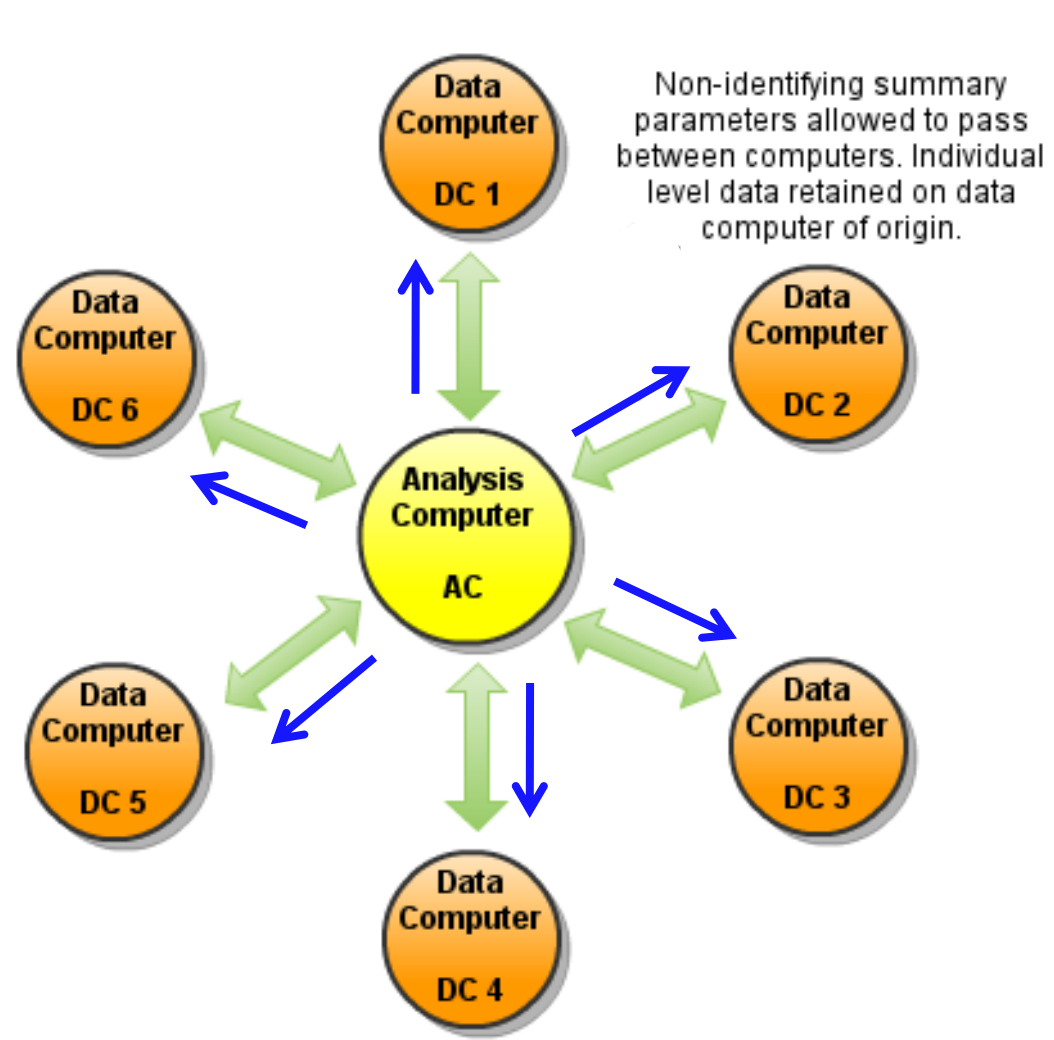


Use parallel processing to fit equivalent models on each study simultaneously

Parallel analyses linked together by transmitting entirely non-identifying summary statistics

In many settings this produces mathematically identical results to fitting a single model to all the data held in one pooled data set

DataSHIELD: a novel solution

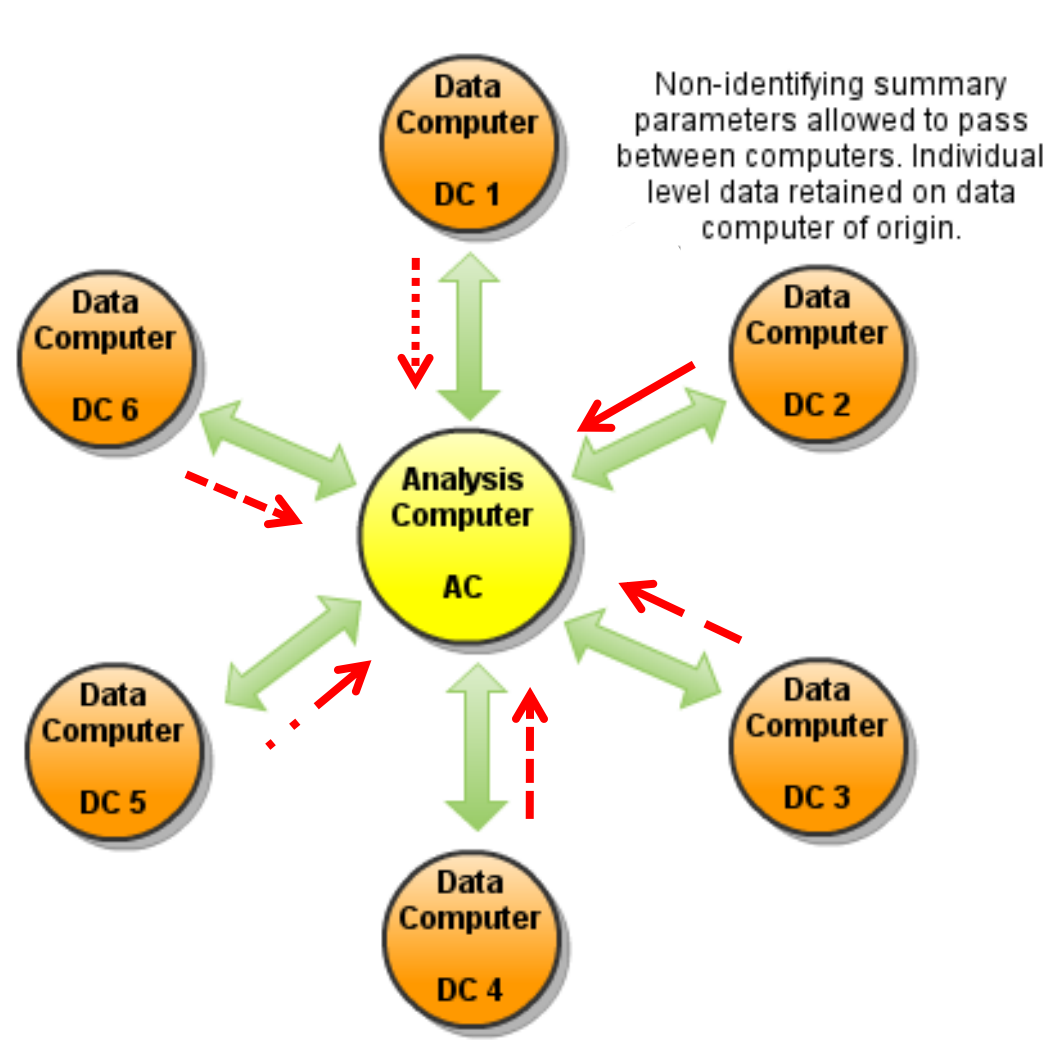


Analysis commands (1)

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)

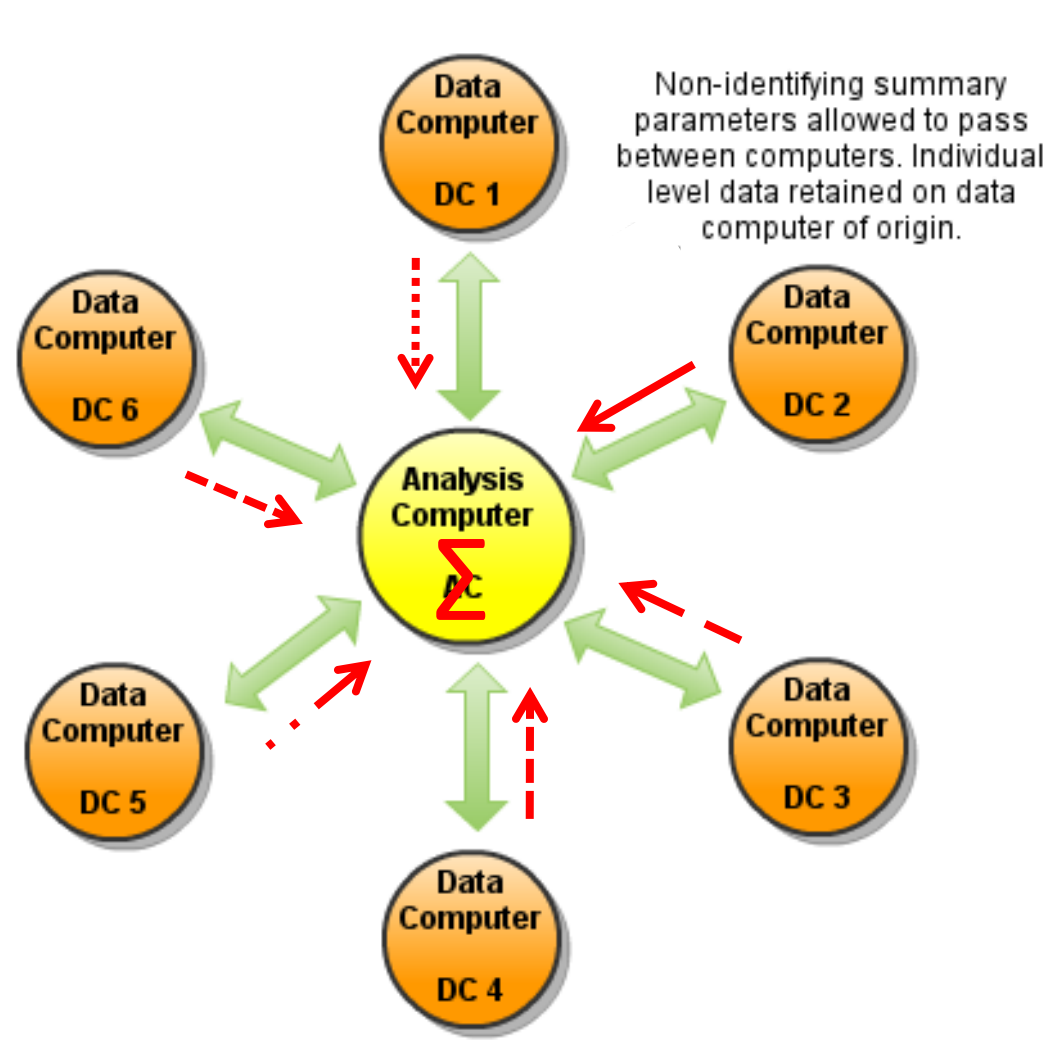
Score vector Study 5

[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

DataSHIELD: a novel solution



Summary Statistics (1)

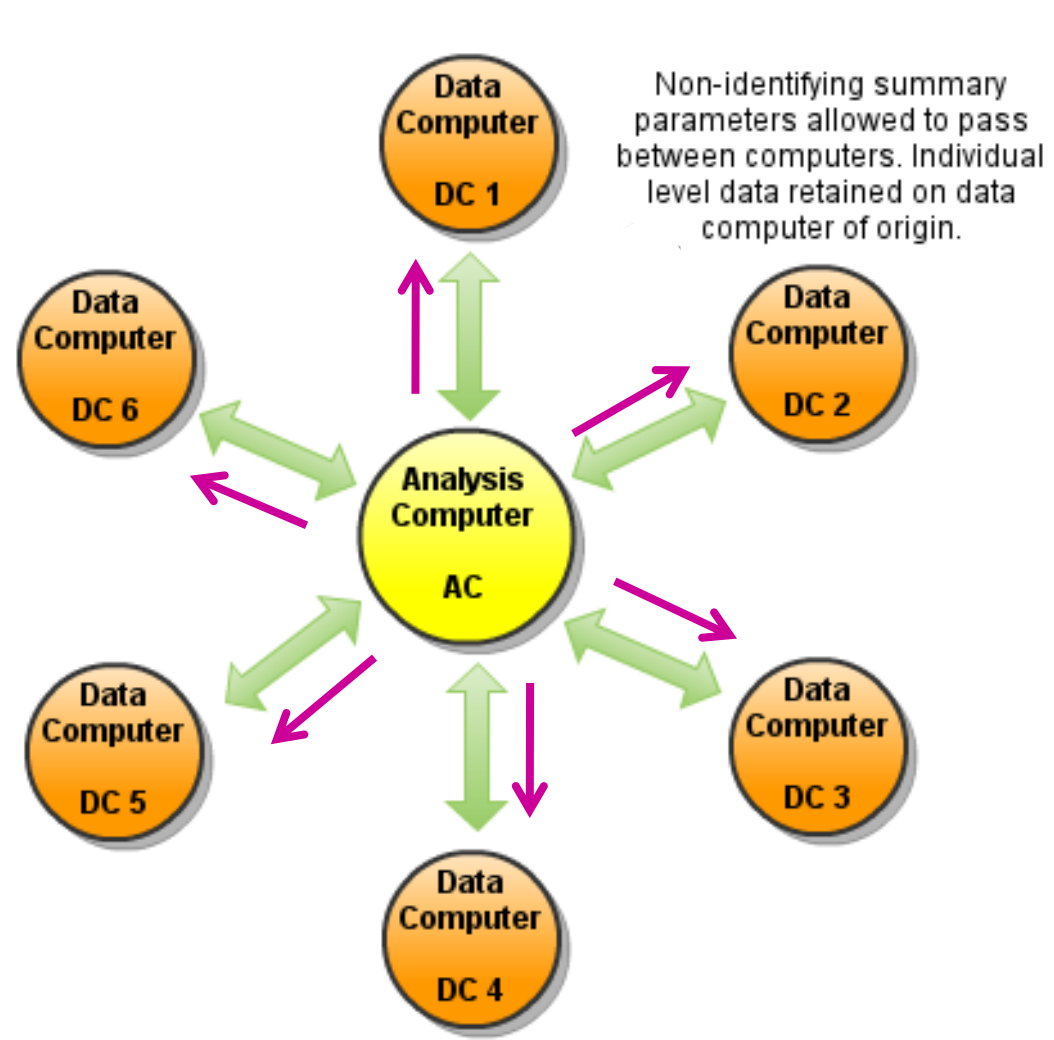
Score vector Study 5

[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

DataSHIELD: a novel solution

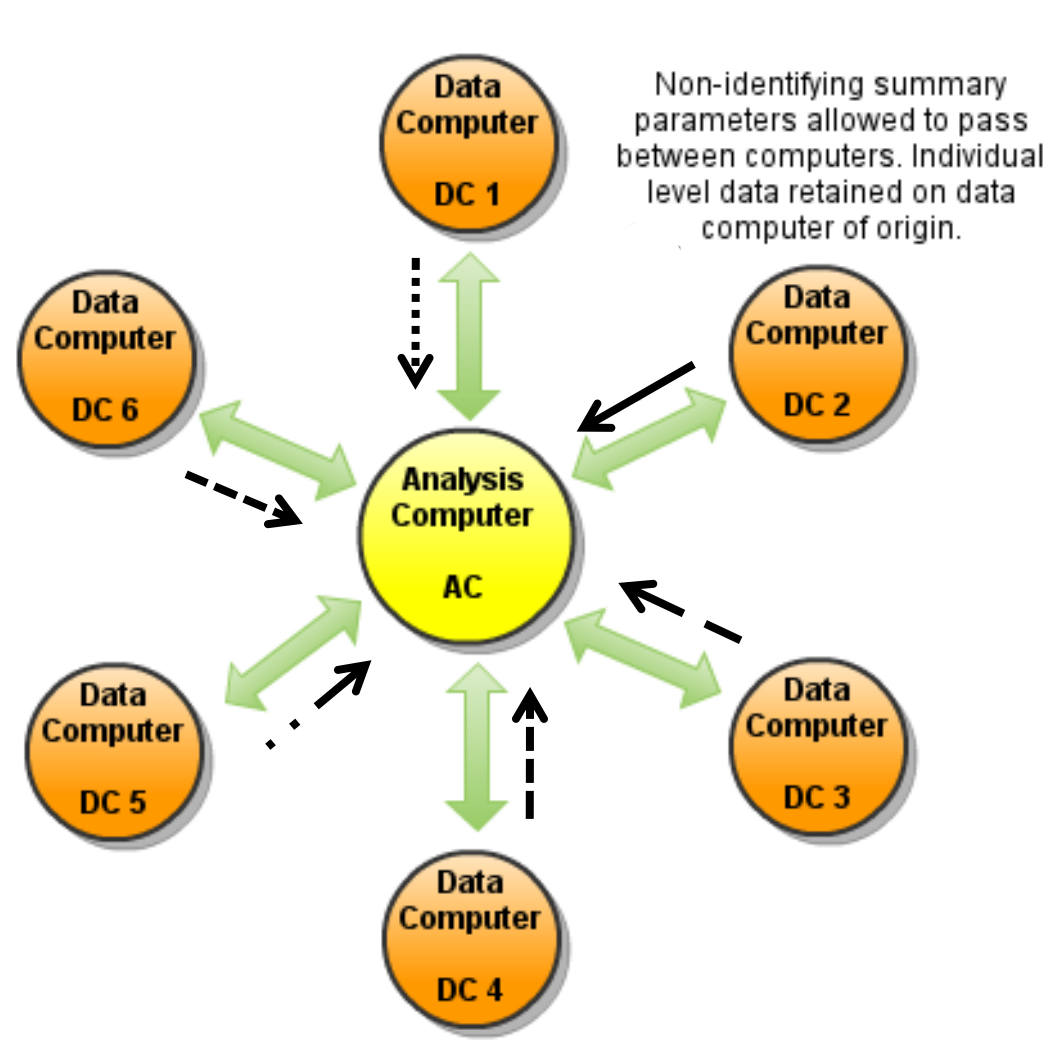


Analysis commands (2)

```
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

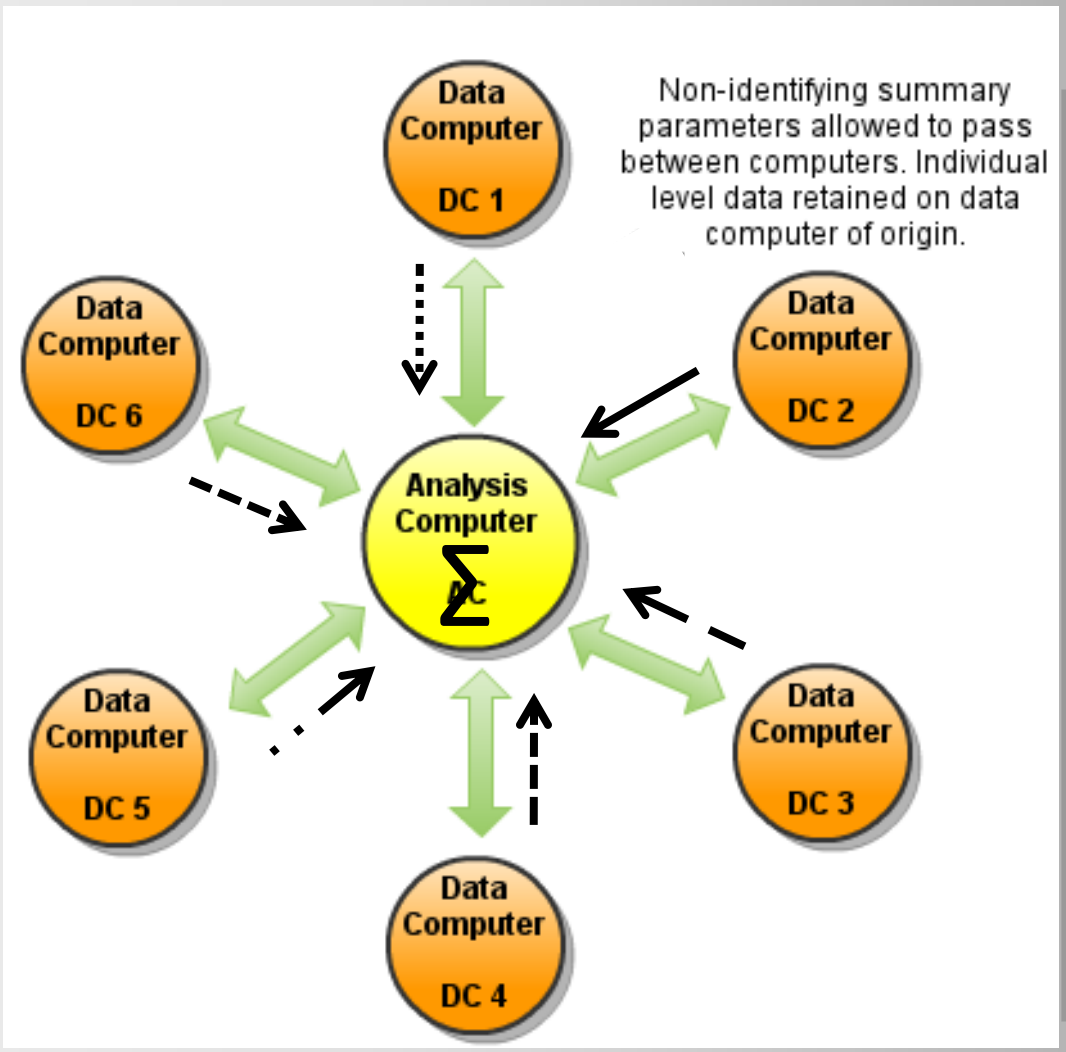


Summary Statistics (2)

Score vectors

Information Matrices

DataSHIELD: a novel solution



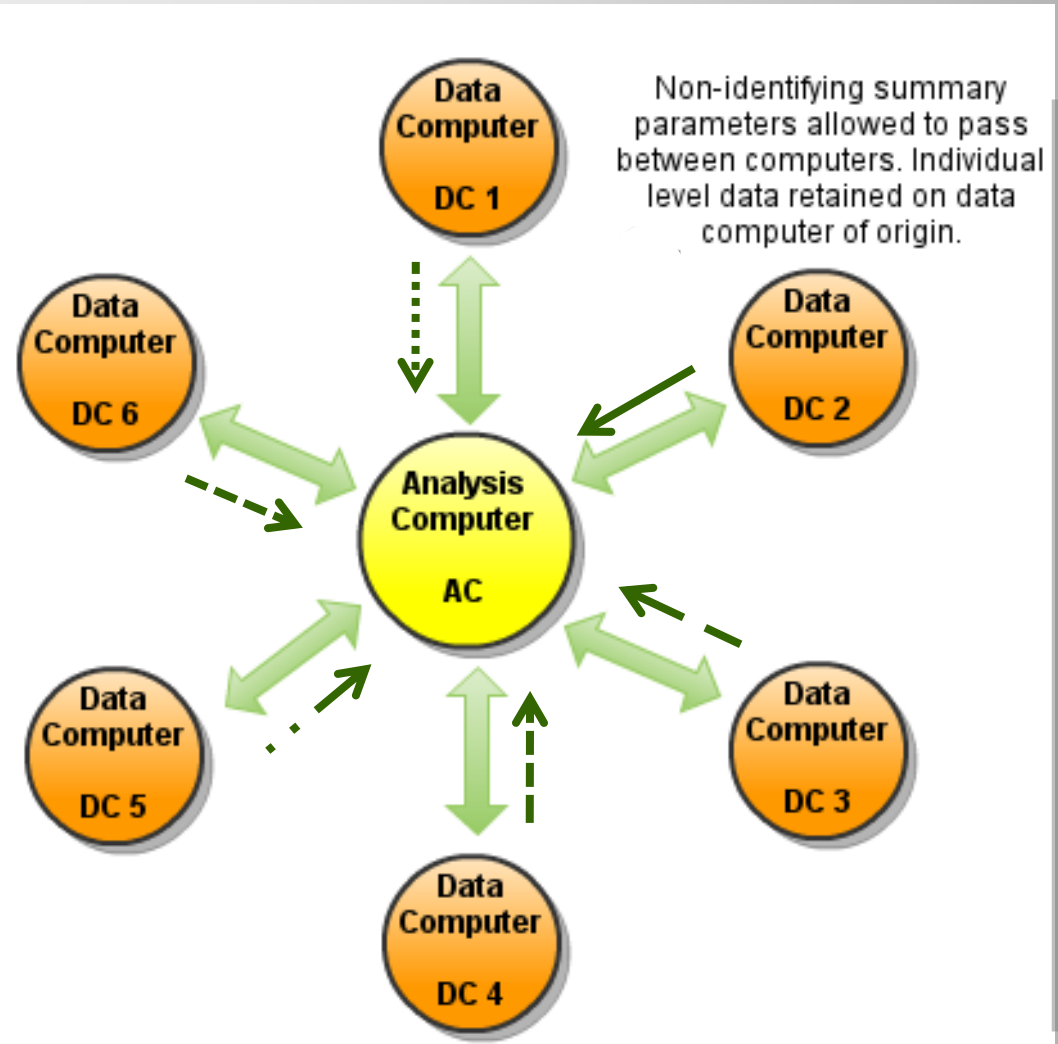
Summary Statistics (2)

Score vectors

Information Matrices

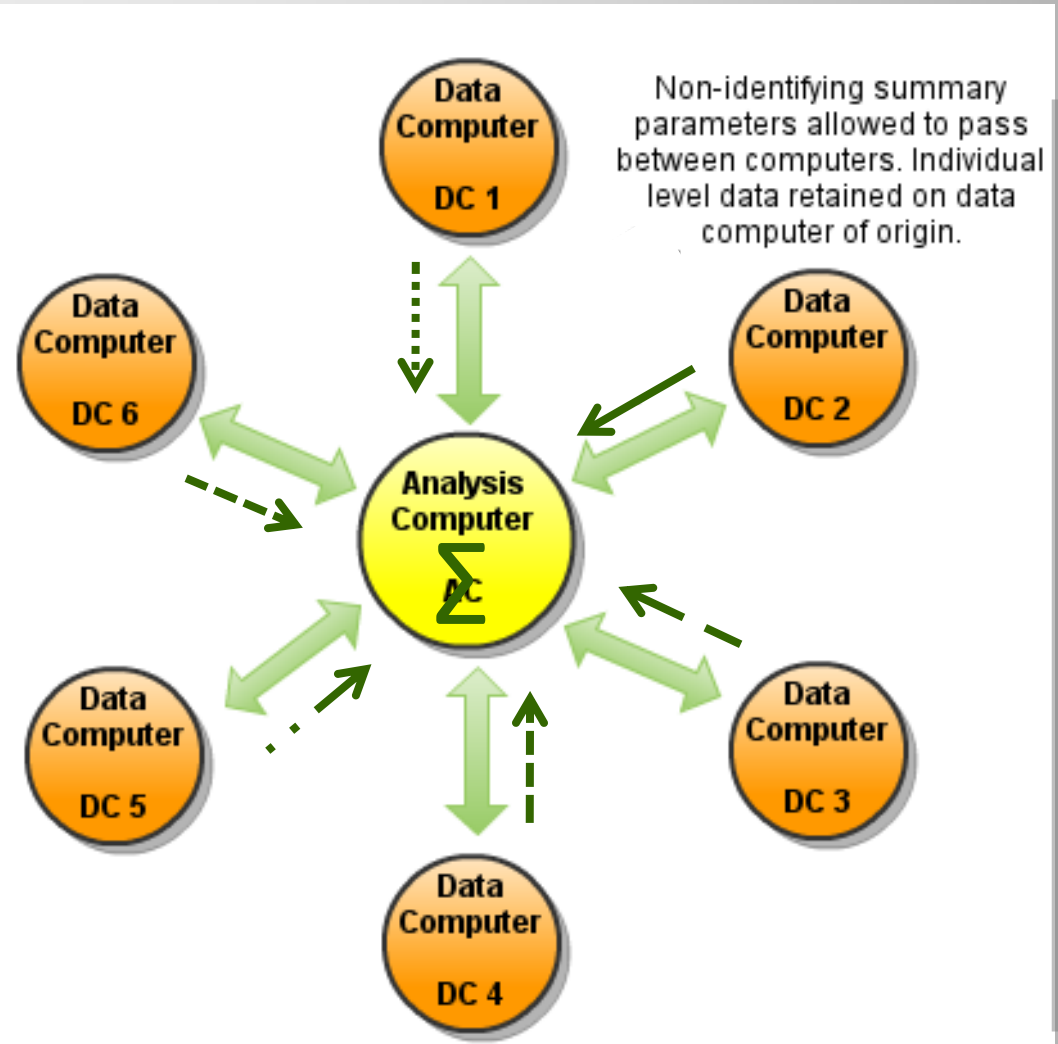
and so on .....

DataSHIELD: a novel solution



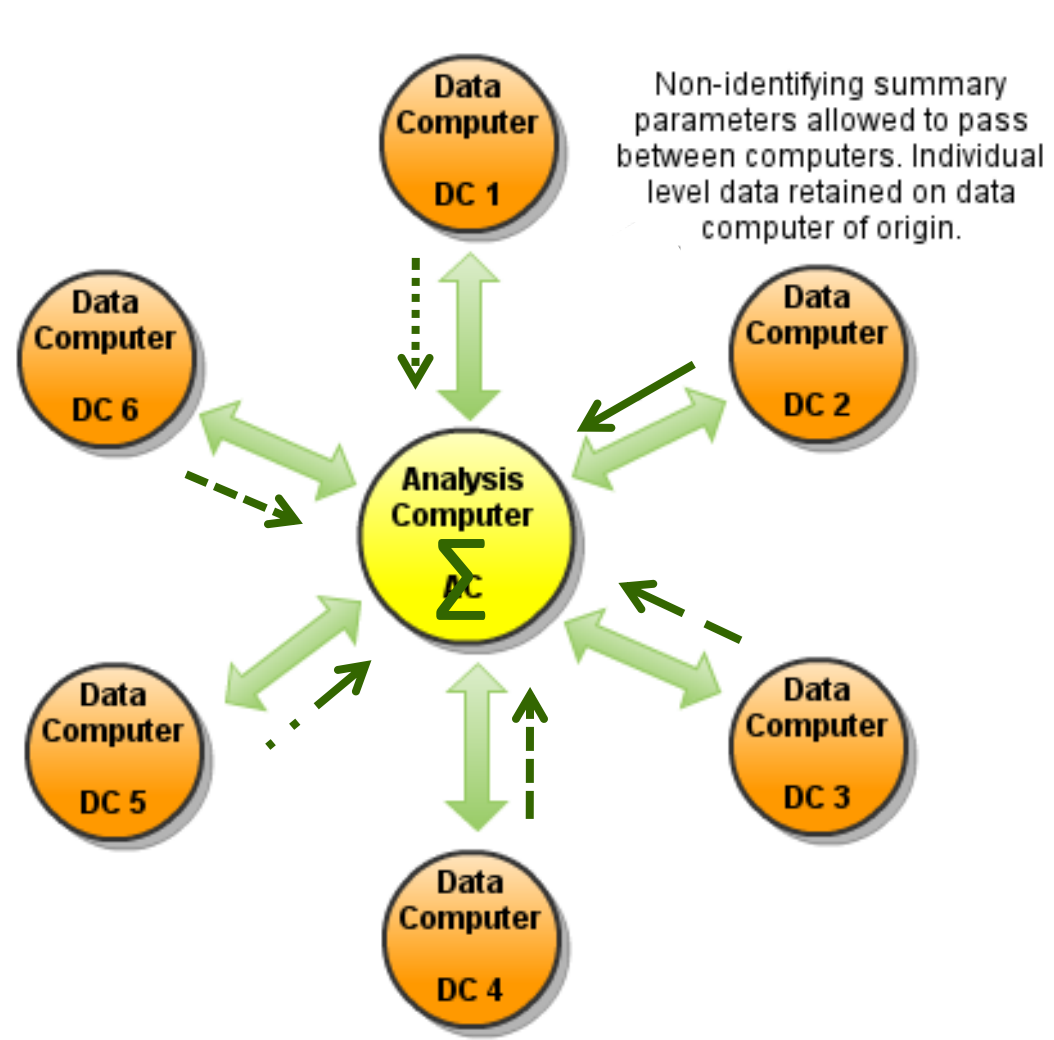
Summary Statistics (4)

DataSHIELD: a novel solution



Updated parameters (4)

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:

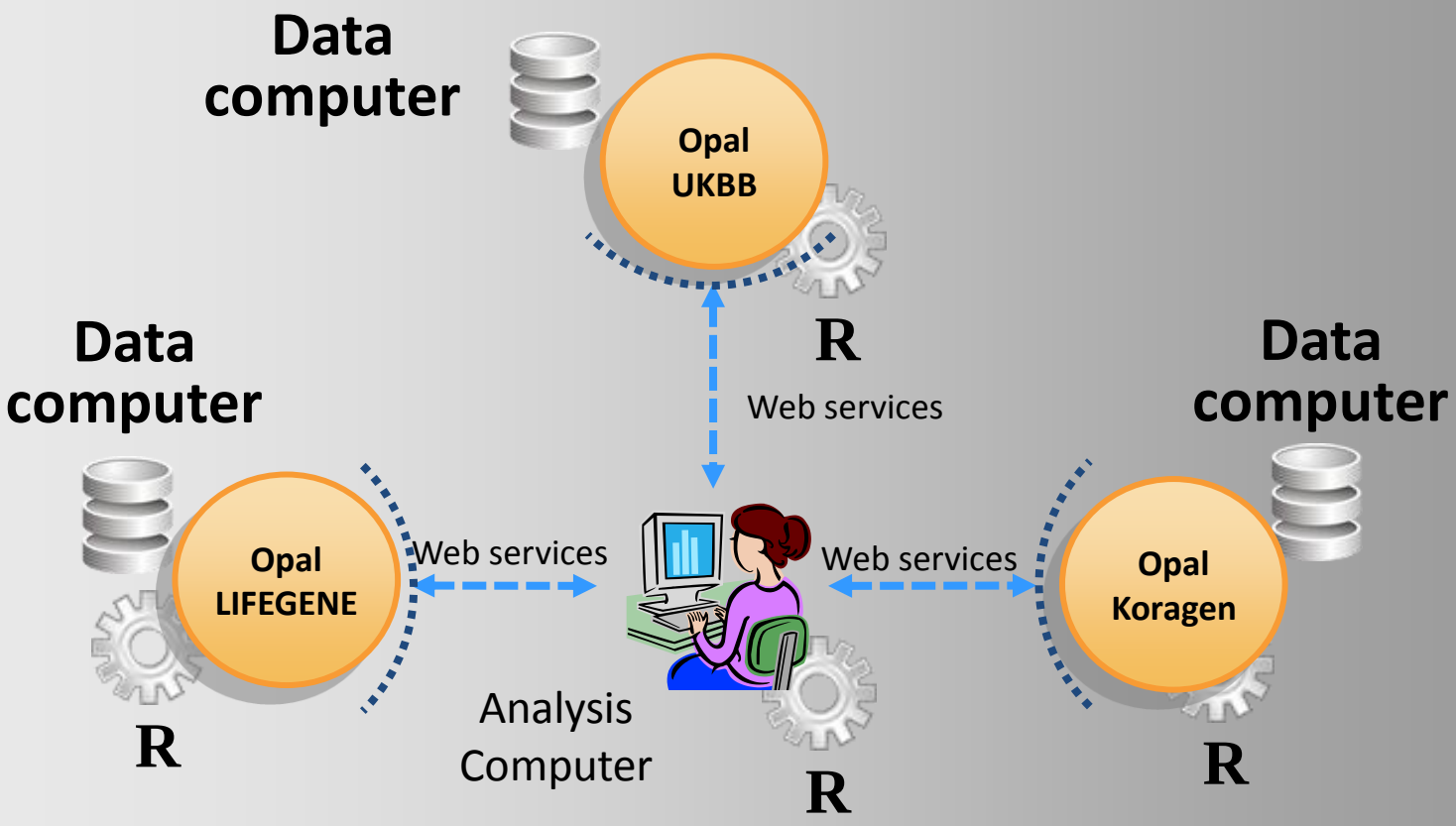
	Estimate	Std. Error
(Intercept)	**-0.32956**	**0.02838**
BMI	**0.02300**	**0.00621**
BMI.456	**0.04126**	**0.01140**
SNP	**0.55173**	**0.03295**

Does it work?

DataSHIELD analysis

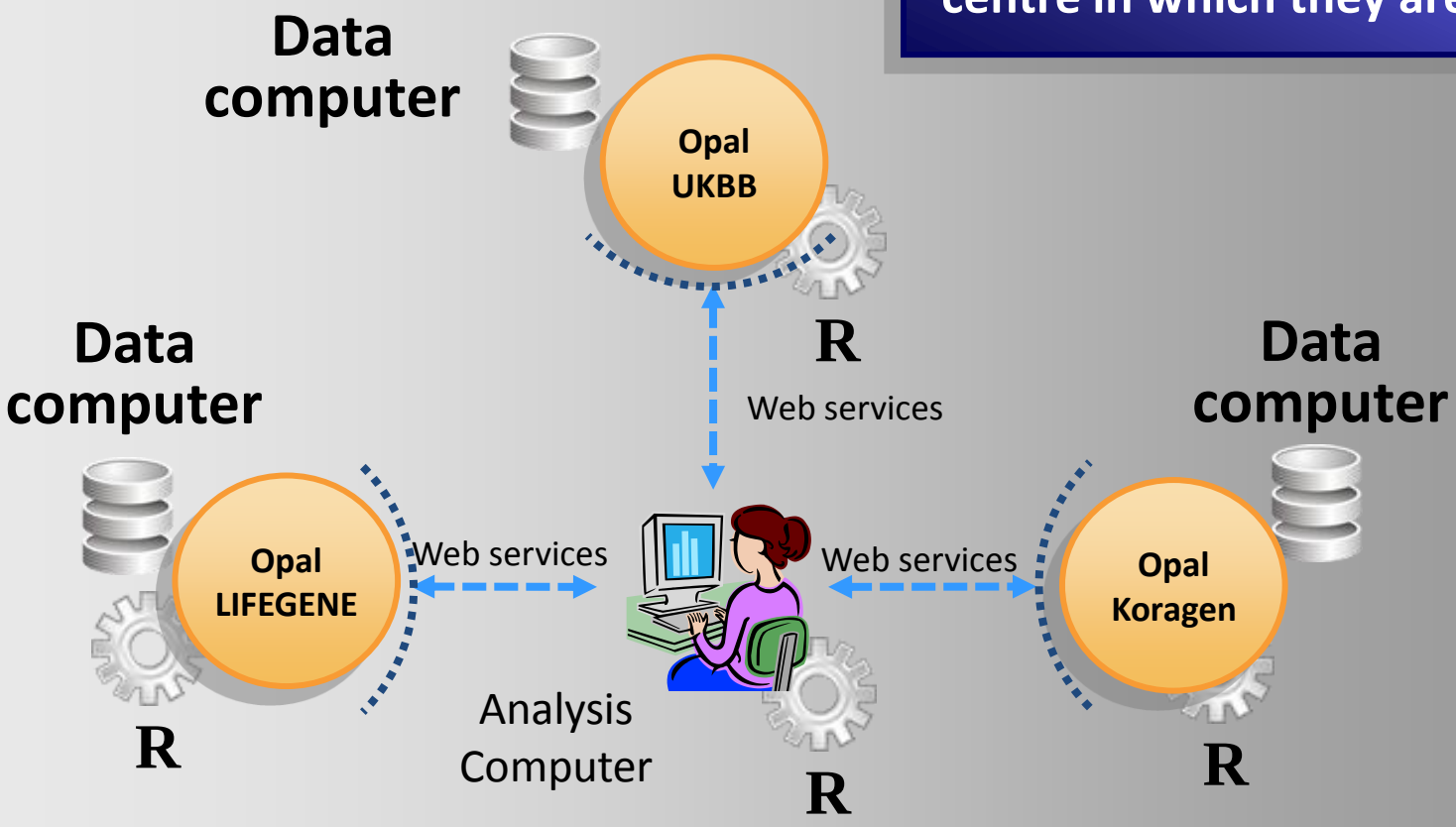
Parameter	Coefficient	Standard Error
$b_{intercept}$	**-0.3296**	**0.02838**
b_{BMI}	**0.02300**	**0.00621**
$b_{BMI.456}$	**0.04126**	**0.01140**
b_{SNP}	**0.5517**	**0.03295**

Current Implementation



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.



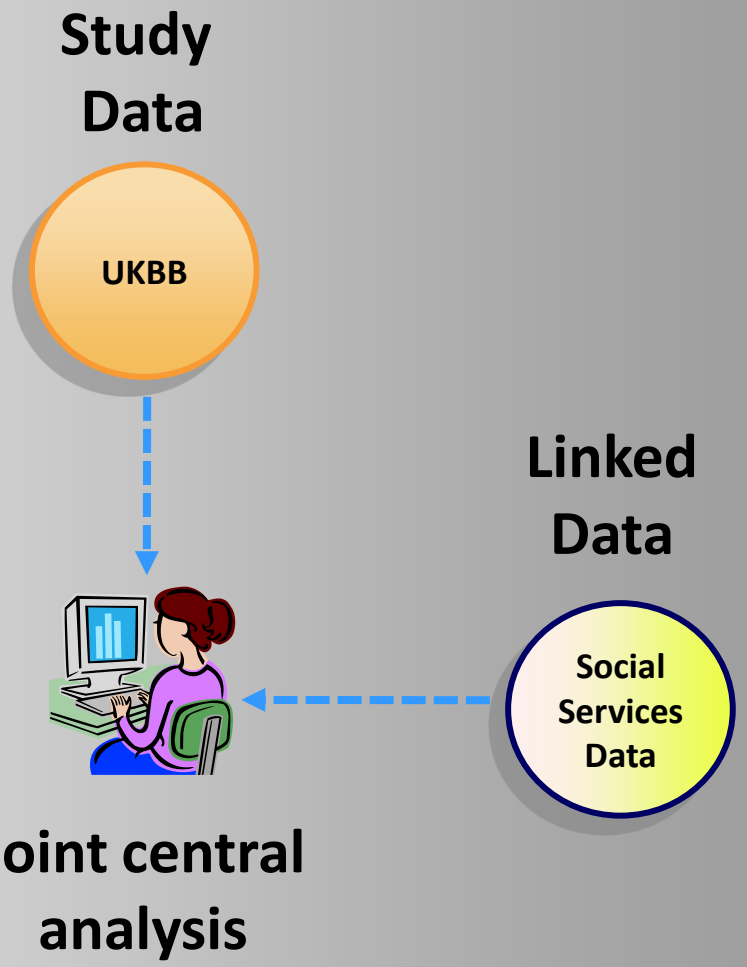
Vertically partitioned data

METHODS

Direct pooling

Secure server

DataSHIELD
Parallelized analysis



Safeguards

- Ethical approval for *all* contributing studies
- Confidentiality agreements for *all*
- Record *all* information passed to and from each DC *at the DC*
- No new models until summary statistics fully understood
  - At present GLMs OK, random effects no!!
- Harmonization essential

Safeguards

- For the future – a software wrapper
  - Monitor, scrutinise and interpret all incoming and outgoing information flows and block and identify/record any request, or series of requests, that might – either by accident or by design – generate potentially disclosive information
- For the moment
  - Suggest more than one statistician (from more than one centre) closely follows the analysis

Current state of play

- Main funding, BioSHARE-eu and ALSPAC
- DataSHIELD implemented in R
  - Refining at present
- R – based implementation inserted into OPAL
  - Part of OBiBa suite
- First workshop held in Lyon, April 2011
- Analytic validity confirmed on real IPRI data
- Three pilot studies moving forward

Thank you for listening

Combined analysis *in principle*

- Workable access
  - Integrative IT platforms
  - Working within ELSI guidelines and edicts – particularly across jurisdictions
- Scientific compatibility
  - Prospective v Retrospective
  - Stringent v Flexible

Study Catalogue

Catalogue current content

Number of Studies	162
With Summary information (to be validated)	70
With Complete information (validated by study investigators)	92

All 0-9 A B C D E F G H I J K L M N

Results 1 to 25 of 162 studies found

[First/Prev] **1**, **2**, **3**, **4**, **5**, **6**, **7**[**Next/Last**]

Questionnaire Catalogue

Catalogue current content	
Number of studies	48
Number of questionnaires	76

http://www.p3gobservatory.org/questionnaireblock/viewBlock.htm?questionnaireId=6&blockId=77 - Windows Internet Explorer

http://www.p3gobservatory.org/questionnaireblock/viewBlock.htm?questionnaireId=6&blockId=77

UK Biobank: a large-scale prospective epidemiological resource [UKB] / UK Biobank Touchscreen Questionnaire / Block 77

L3	Has a doctor ever told you that you have diabetes?	Select one from - YE Yes - NO No - UN Do not know - DA Prefer not to answer
L3A	Did you only have diabetes during pregnancy?	Select one from - YE Yes - NO No - NA Not applicable - UN Do not know - DA Prefer not to answer
L3B	What was your age when the diabetes was first diagnosed?	Enter number OR UN Do not know OR DA Prefer not to answer
L3C	Did you start insulin within one year of your diagnosis of diabetes?	Select one from

Combined analysis *in principle*

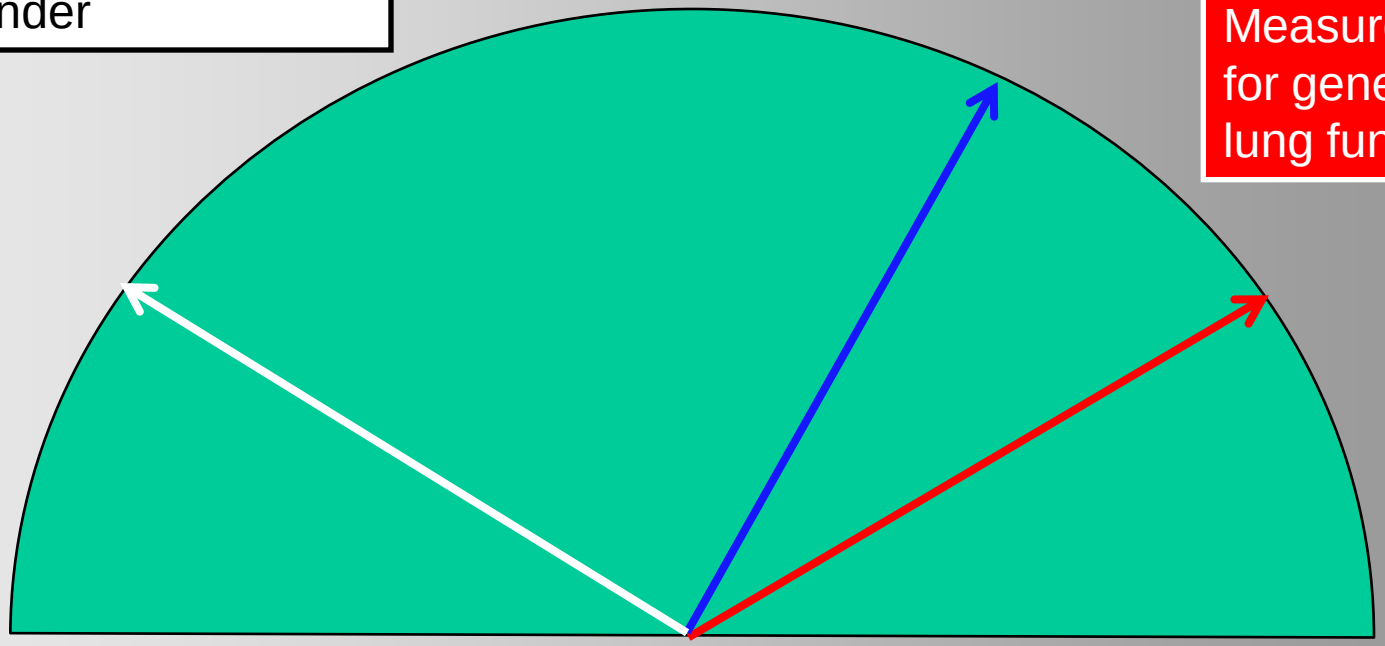
- Workable access
  - Integrative IT platforms
  - Working within ELSI guidelines and edicts – particularly across jurisdictions
- Scientific compatibility
  - Prospective v Retrospective
  - Stringent v Flexible

Context specific!!

Identify subjects with a doctor diagnosis of diabetes for an aetiological study

Assess smoker versus non-smoker as a confounder

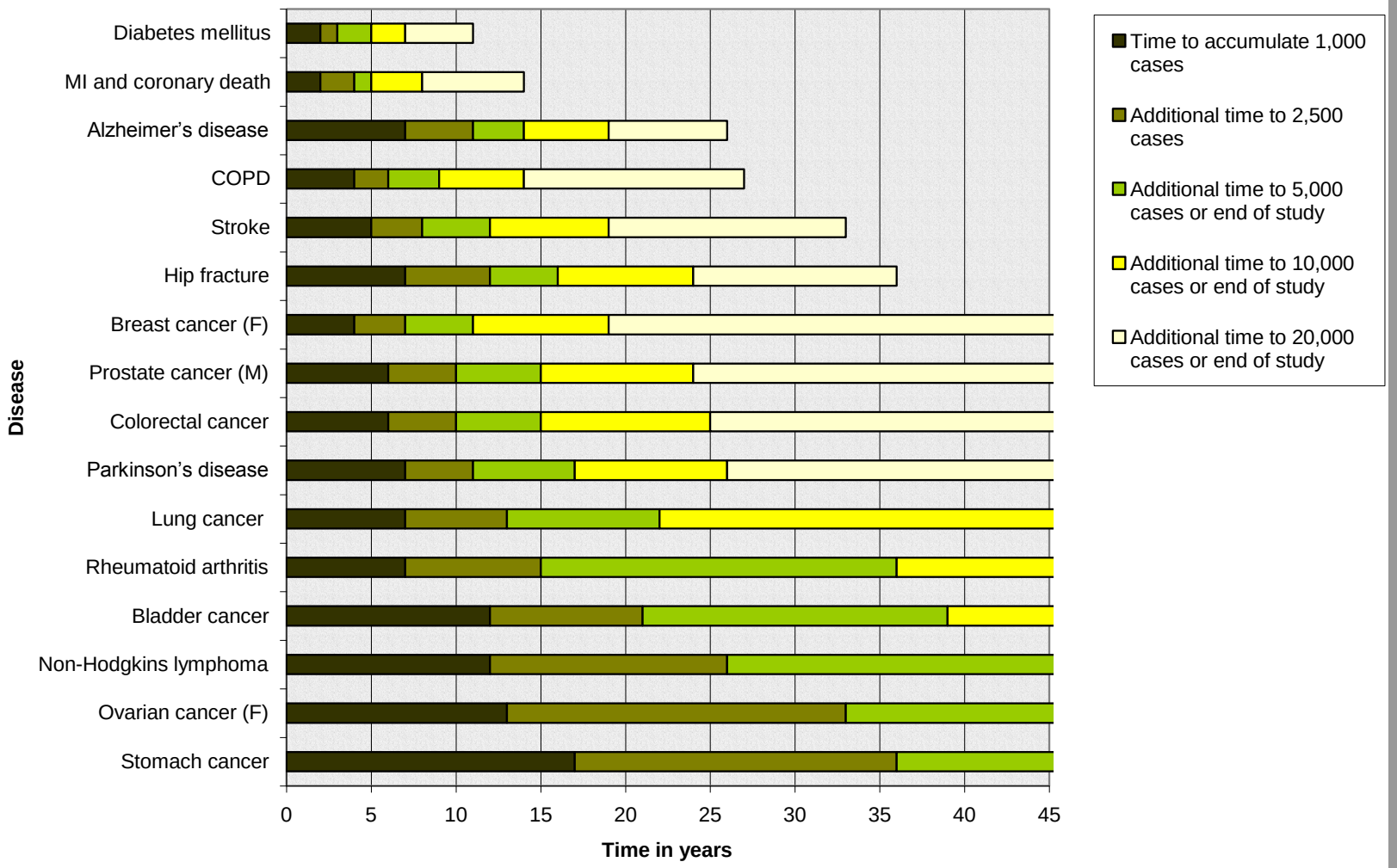
Measure FEV_1 for genetics of lung function



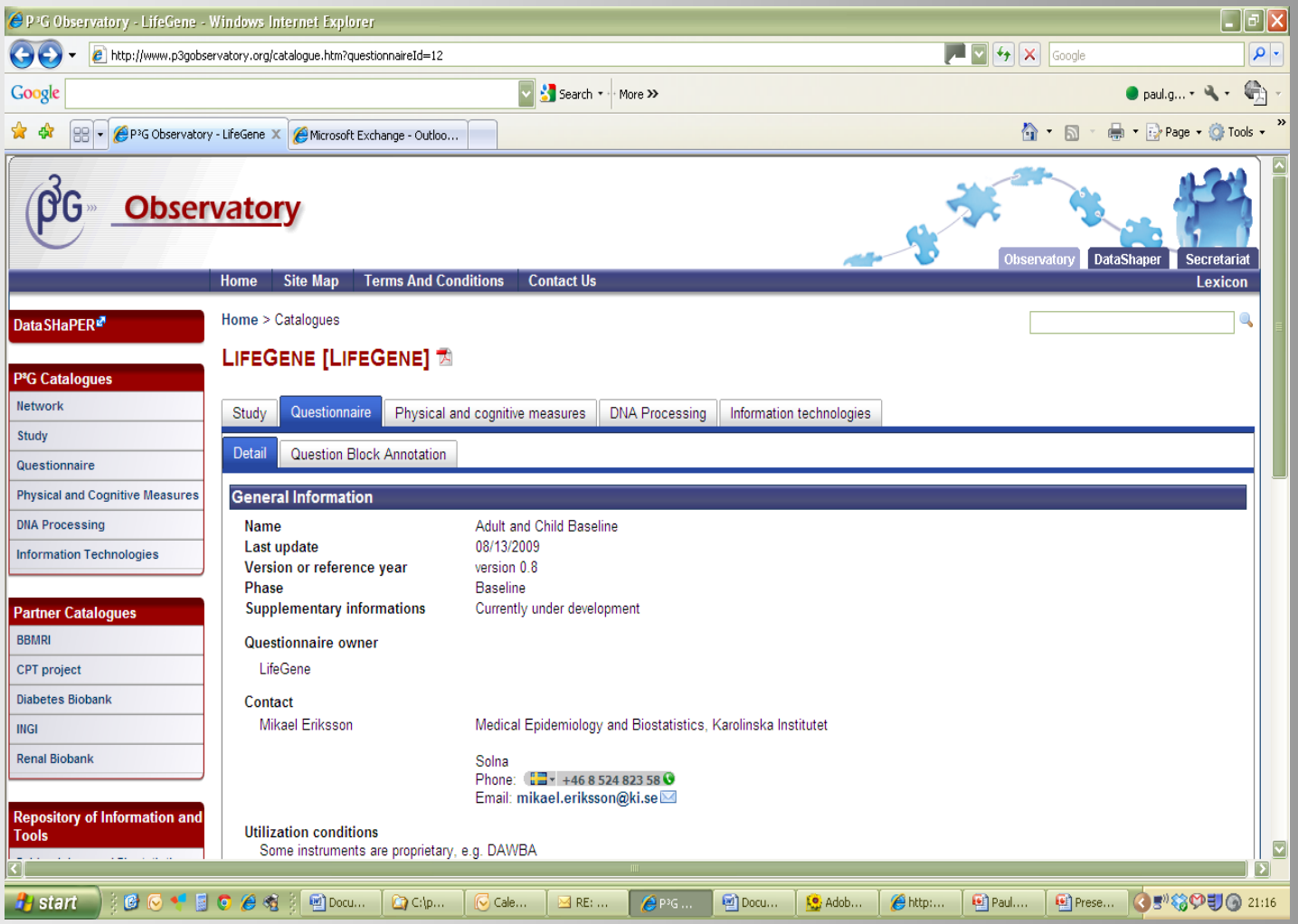
Ad hoc methods

Perfect standardisation

How long is 'LONG'?



The P³G Observatory



DataSHaPER website

Study Catalogue

File Edit View History Bookmarks Tools Help

http://www.p3gobservatory.org/catalogue.htm?studyId=21

P³G Observatory - LifeGene

Legend ✓ Yes X No ? Unknown

General design

Study design	Cohort
Type of participants	Families
Target or final number of participants	500000
Target or final number of families	No information available
Target or final number of DNA	No information available

Participant selection / Characteristics of the population

Selection criteria

✓ Age	Maximun: 55 years
✓ Country of residence	Sweden

Recruitment procedures

Initially, enrollment of subjects will be performed by use of national databases such as Swedish Twin Registry, followed by enrollment of subjects through contacts via mass e-mails, advertisements, for enrichment of the cohort with under-studied ethnic populations, etc.

Data Sources

	Cross-sectional	Repeated/continuous
✓ Questionnaires to participants/respondents	X	✓
✓ Direct physical measures	X	✓
✓ Biological samples	✓	X
✓ Medical paper		
✓ Electronic databases		
X Genealogical records		

Collection procedures

Participants will be requested to fill out web-based questionnaires and answer questions by use of cell-phones etc. Data on exposures and phenotypes will be collected repeatedly over time. Linkage to comprehensive registers with health care information will be performed yearly. Blood sampling for DNA extraction and analyses of key biomarkers will be performed according to standardized procedures. Data collection will be for research purpose only. All data will be obtained in coded format. Subject identification data registered in conjunction to biological samples will be stored in such a way that unauthorized persons will have no access and be able to link data and/or samples to the donors. Information transmitted by electronic means will be encrypted. Subject data transferred to a third country (e.g. outside EU) must be de-identified. If data cannot be de-identified before transfer, data security issues must be agreed upon in writing in advance.

Follow up procedures

No information available

Baseline principal variables of interest

Health information

- ✓ Diseases history - ICD-10 All conditions
- ✓ Familial disease history
- ✓ Early life
- ✓ Women's health
- ✓ Quality of life

Physical / Biochemical measures

- ✓ Body structure measures
- ✓ Body function measures
- ✓ Biochemical measures

Sociodemographic Characteristics

- ✓ Birth location
- ✓ Parents birth location

Language restricting the scope of data sharing (Wallace et al, 2009)

- Use of data restricted to researchers participating in the original study
  - "All research data are confidential ... they will only be used in medical research and [will] remain in the sole use of the participating researchers."
- Use of data restricted to researchers in one country
  - "Blood and DNA samples may...be distributed to laboratories...around [country] for further research."
  - "Research using the anonymous samples will be done by [researchers] ... throughout [country]."

The need to obtain both scientific and ethical approval

- "The [Project] gives approved researchers access to data and samples.... All researchers will only have access to coded data or samples, in order to protect your privacy. They also have to obtain prior scientific and ethical approval, as described above, and their research must fit the purpose of the resource/biobank."
- "The [Project] expects to receive requests and, if approved, provide access to data and samples to overseas researchers and international collaborators. These researchers must follow the same procedures as all other researchers. All access is subject to the strictest scientific and ethical scrutiny..."

ELSI restrictions on third-party sharing

- Substantive reasons:
  - Relevant values held throughout many societies
  - Disclosure of identity
  - Sensitive information
  - Intellectual property
- Exemplar wording
  - Wallace S, Lazor S, Knoppers BM. Chapter in Kaye J and Stranger M. Principles and Practice in Biobank Governance. Ashgate, Farnham 2009

The Healthy Obese project (HOP): A Core Project in BioSHaRE-eu

- Create HOP DataSCHEMA for variables to be shared across three major studies in BioSHaRE-eu

LifeGene

Kooperative Gesundheitsforschung in der Region Augsburg KORA

biobank uk
Improving the health of future generations

Y13D	Has/did your father ever suffer from? (You can select more than one answer)	Select from - HE Heart disease - ST Stroke - BP High blood pressure - DB Diabetes - EM Chronic bronchitis/emphysema - AZ Alzheimers disease/dementia - NN None of the above - UN Do not know - DA Prefer not to answer	If DA go to Y16 Otherwise Go to Y13E
Y13D AD	Has/did your adopted father ever suffer from? (You can select more than one answer)	Select from - HE Heart disease - ST Stroke - BP High blood pressure - DB Diabetes - EM Chronic bronchitis/emphysema - AZ Alzheimers disease/dementia - NN None of the above - UN Do not know - DA Prefer not to answer	If DA go to Y16AD Otherwise Go to Y13EAD

Y16D	Has/Did your mother ever suffer from? (You can select more than one answer)	Select from - HE Heart disease - ST Stroke - BP High blood pressure - DB Diabetes - EM Chronic bronchitis/emphysema - AZ Alzheimers disease/dementia - NN None of the above - UN Do not know - DA Prefer not to answer	If DA go to Y17 Otherwise Go to Y16E
Y16D AD	Has/Did your adopted mother ever suffer from? (You can select more than one answer)	Select from - HE Heart disease - ST Stroke - BP High blood pressure - DB Diabetes - EM Chronic bronchitis/emphysema - AZ Alzheimers disease/dementia - NN None of the above - UN Do not know - DA Prefer not to answer	If DA go to Y17AD Otherwise Go to Y16EAD

Y19	Have any of your brothers **or** sisters suffered from any of the following illnesses? (You can select more than one answer)	Select from - HE Heart disease - ST Stroke - BP High blood pressure - DB Diabetes - EM Chronic bronchitis/emphysema - AZ Alzheimers disease/dementia - NN None of the above - UN Do not know - DA Prefer not to answer	If DA go to next section Otherwise Go to Y20
Y19AD	Have any of your adopted brothers or sisters suffered from any of the following illnesses? (You can select more than one answer)	Select from - HE Heart disease - ST Stroke - BP High blood pressure - DB Diabetes - EM Chronic bronchitis/emphysema - AZ Alzheimers disease/dementia - NN None of the above - UN Do not know - DA Prefer not to answer	If DA go to next section Otherwise Go to Y20AD

Strategic investment

- Effective *joint* data analysis involving large-scale biobanks will be an essential element of tomorrow's bioscience
- Generic tools and methods exist ***now*** that make it much easier to construct and then use harmonized data sets
- Strategic national and international funding is required to ensure that such tools continue to evolve, and are implemented in effective multipurpose user IT environments that are cheap and easy to use
- Some encouragement from funders may be needed to ensure they are actually used.

Strategic investment

- The generic integrative work I have described is undertaken and promoted by large international consortia such as P^3G, PHOEBE, BBMRI and BioSHaRE-eu
- Investment in such platform projects should therefore be international - all nations should contribute to a truly collaborative endeavour
- This investment should be equitably shared to ensure long-term sustainability