Ensuring new solutions meet the real challenges: the role of DataSHIELD

Paul Burton

University of Newcastle, *Institute of Health & Society*

Data to Knowledge Research Group

Connected Health Cities Project Team

Tom Bishop

University of Cambridge, *MRC Epidemiology Unit*

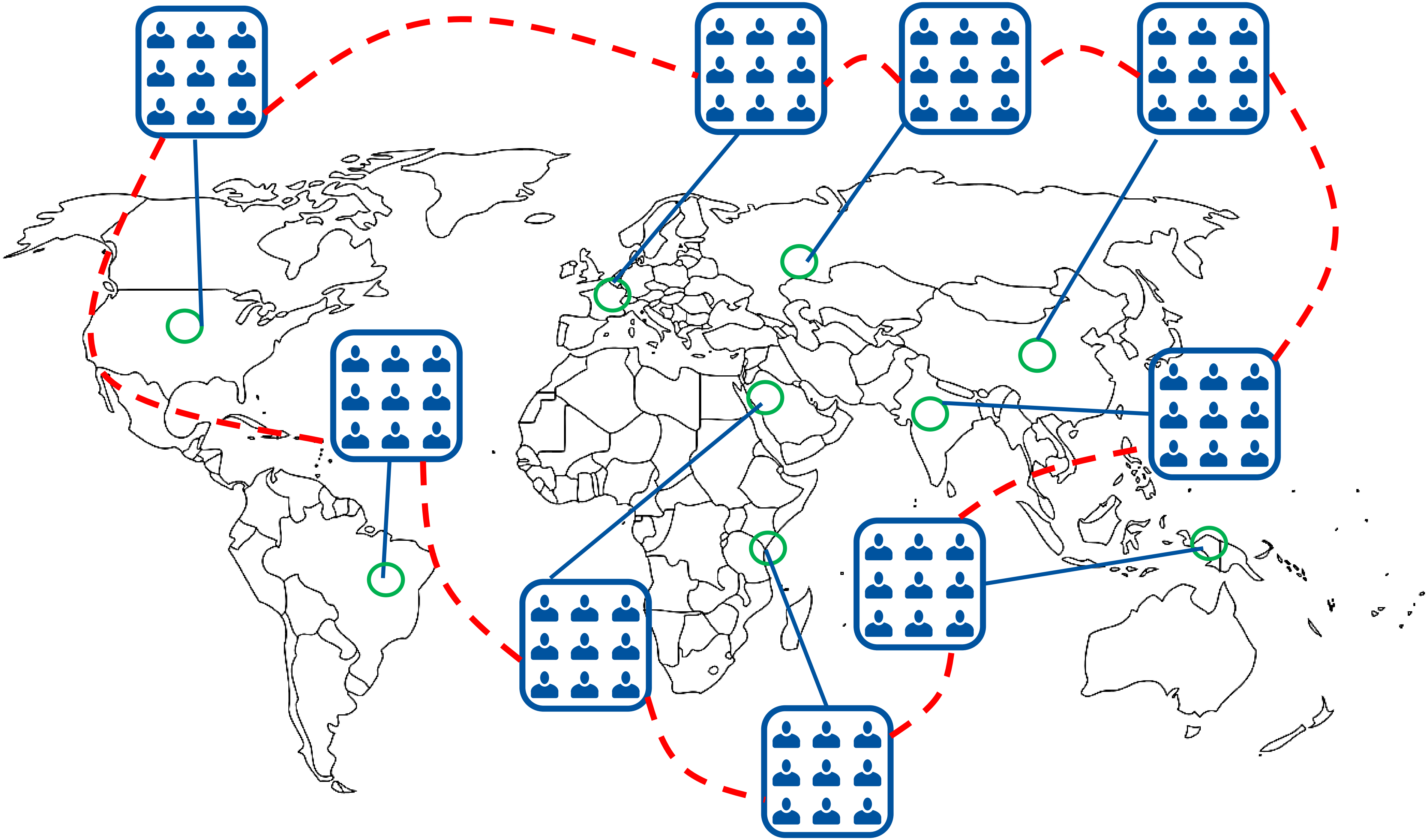
McGill University, OICR, Maelstrom Research

The role of DataSHIELD: Overview

- Scientific motivation for analysing across distributed data sets (e.g. when studying metabolic disorders)
- Existing approaches and how the DataSHIELD approach of taking analysis to the data is different
- Results from research we have done using DataSHIELD

Diabetes & Obesity – global health challenges

Studies are costly to run & tend to focus on a single population

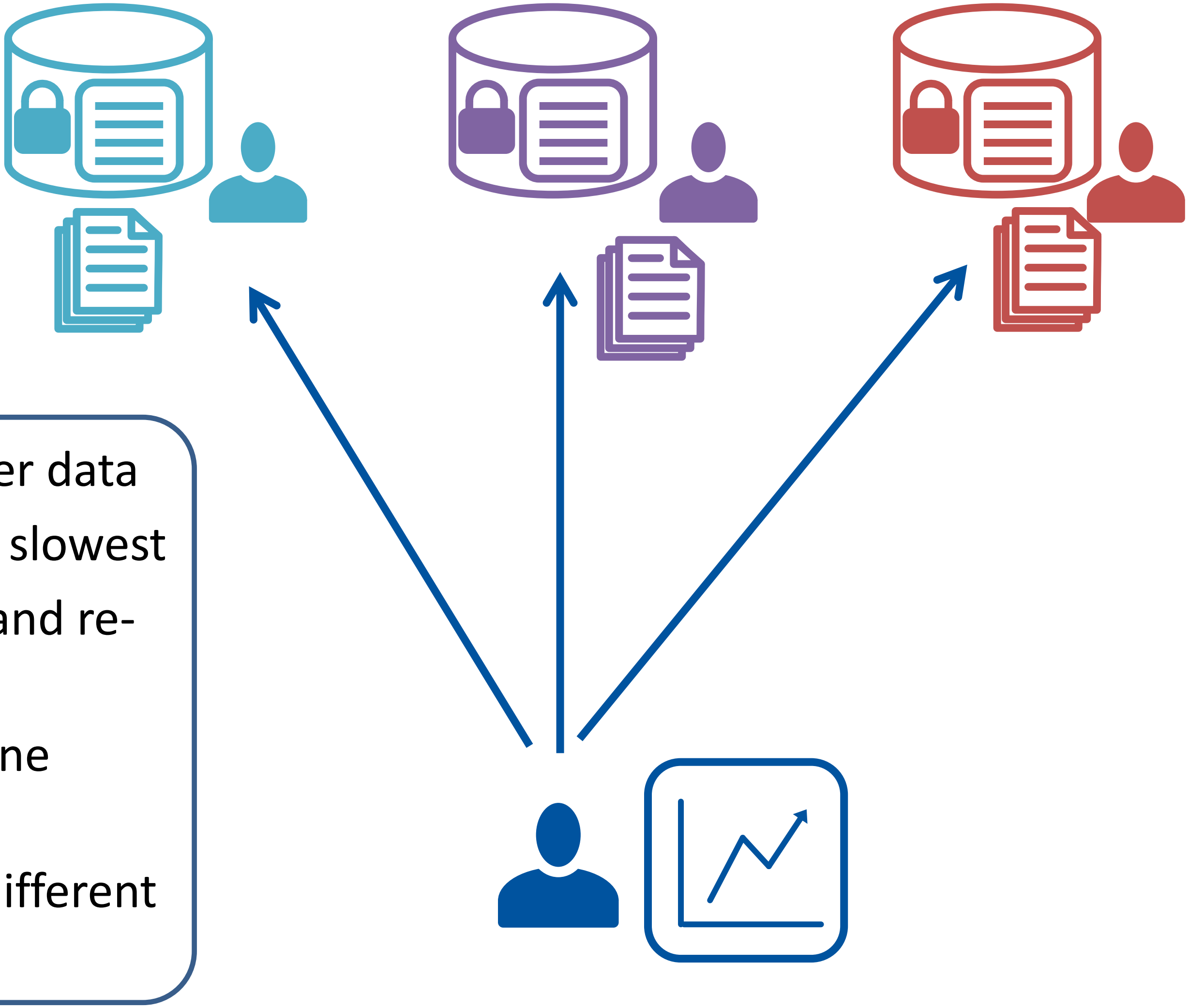


Use study results from existing publications (literature-based meta analysis)

📚	Review published papers
🔍	Extract relevant results
📈	Perform overall analysis

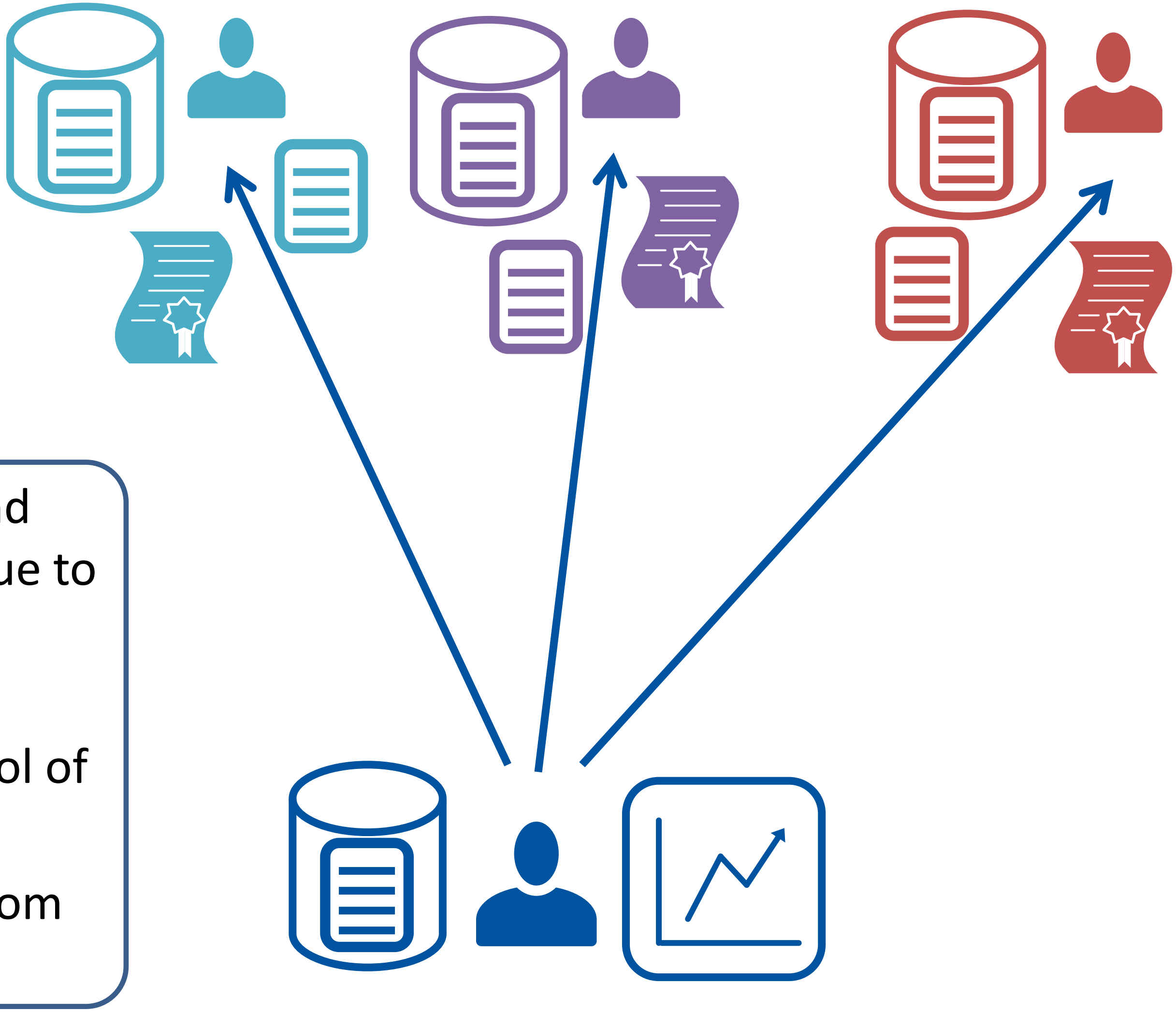
- Widely used - simple
- Can only analyse published results – potential bias
- Uncertainty in how the results were derived – inconsistencies between papers
- Results available are fixed

Results sharing



- Useful if unable to transfer data
- Moves at the pace of the slowest
- Takes a long time to run and re-run analyses
- Each group needs someone available to run analyses
- Analysis plan is open to different interpretations

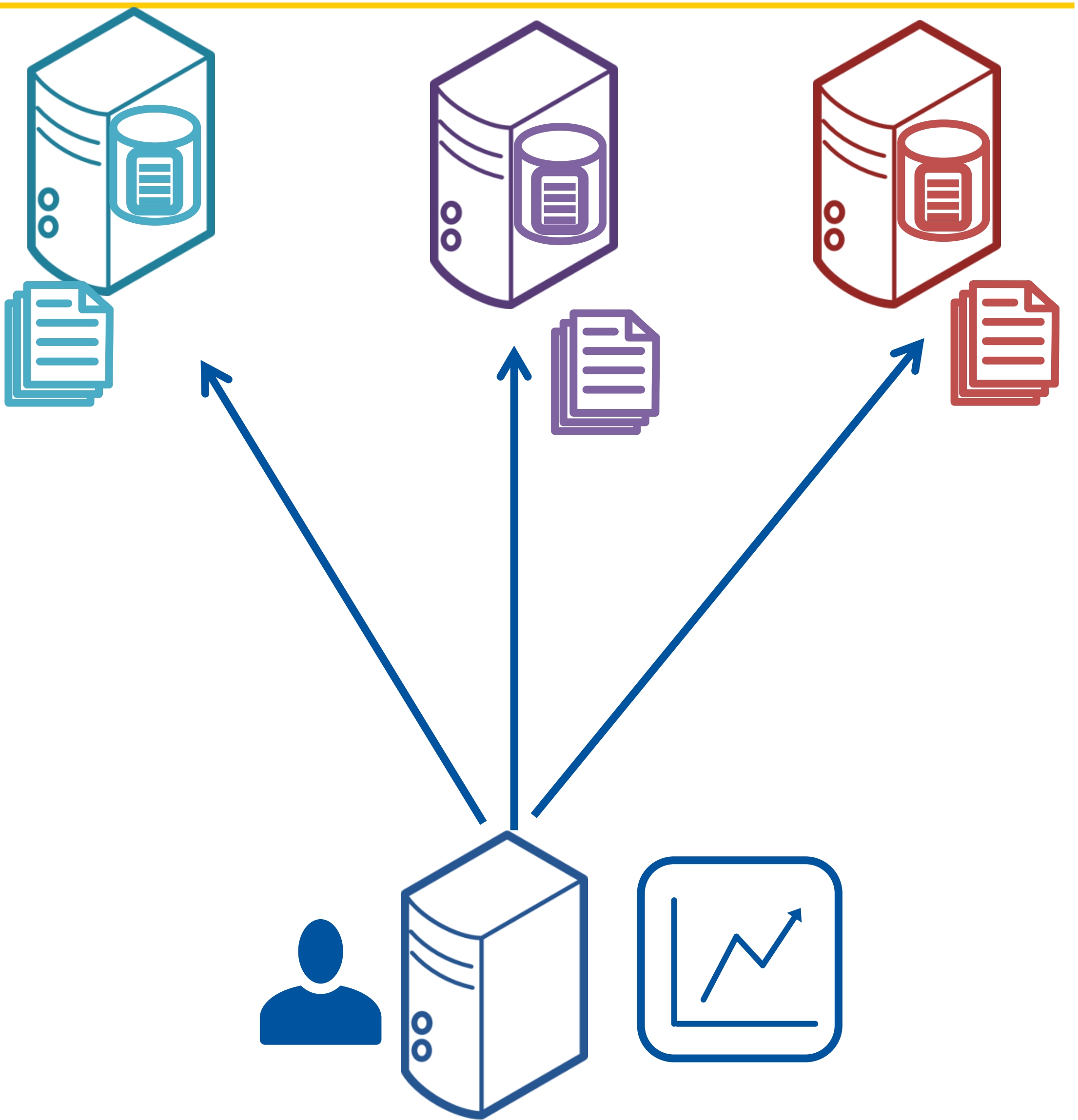
Data sharing



- Potential ethico-legal and governance problems due to moving data around
- Reluctance of data custodians to lose control of data
- Flexible and desirable from an analytical standpoint

DataSHIELD addresses the challenges by taking the analysis to the data

- Data stays on each study's server – **no data transfer agreement**
- Analytical commands passed to each server
- Summary results passed back – **no access to individual values**
- No waiting for others to run analysis
- No publication bias



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

Horizontal partitioning:

- Different sources hold all variables but on different individuals
- Secure meta-analysis (IPD and Study-Level)
- Secure single-site analysis

Vertical partitioning:

- Different sources hold different variables on the same individuals
- Secure processing and analysis of linked data without bringing the data together

DataSHIELD real world example: InterConnect

Global data for diabetes and obesity research

- Funded by the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement no 602068.
- www.interconnect-diabetes.eu

Other DataSHIELD users

- BioSHaRE-EU Healthy Obese project
- BioSHaRE-EU Environmental Core project
- SPIRIT (child health development in Canada)
- ENDAPASI (German Institute of Human Nutrition)
- Farr Institutes
- UK Data Archive
- F1000 research journal

Effect of maternal Physical Activity during pregnancy on Neonatal Anthropometric Outcomes

Your guide to staying active in pregnancy

- ✓ Physical activity in pregnancy is safe and healthy
- ✓ Being active benefits you and your baby
- ✓ Stay active: 30 minutes a day, 4 times a week

Exercising increases the blood flow to the placenta. This is great for your baby's growth and development.

4.5 times

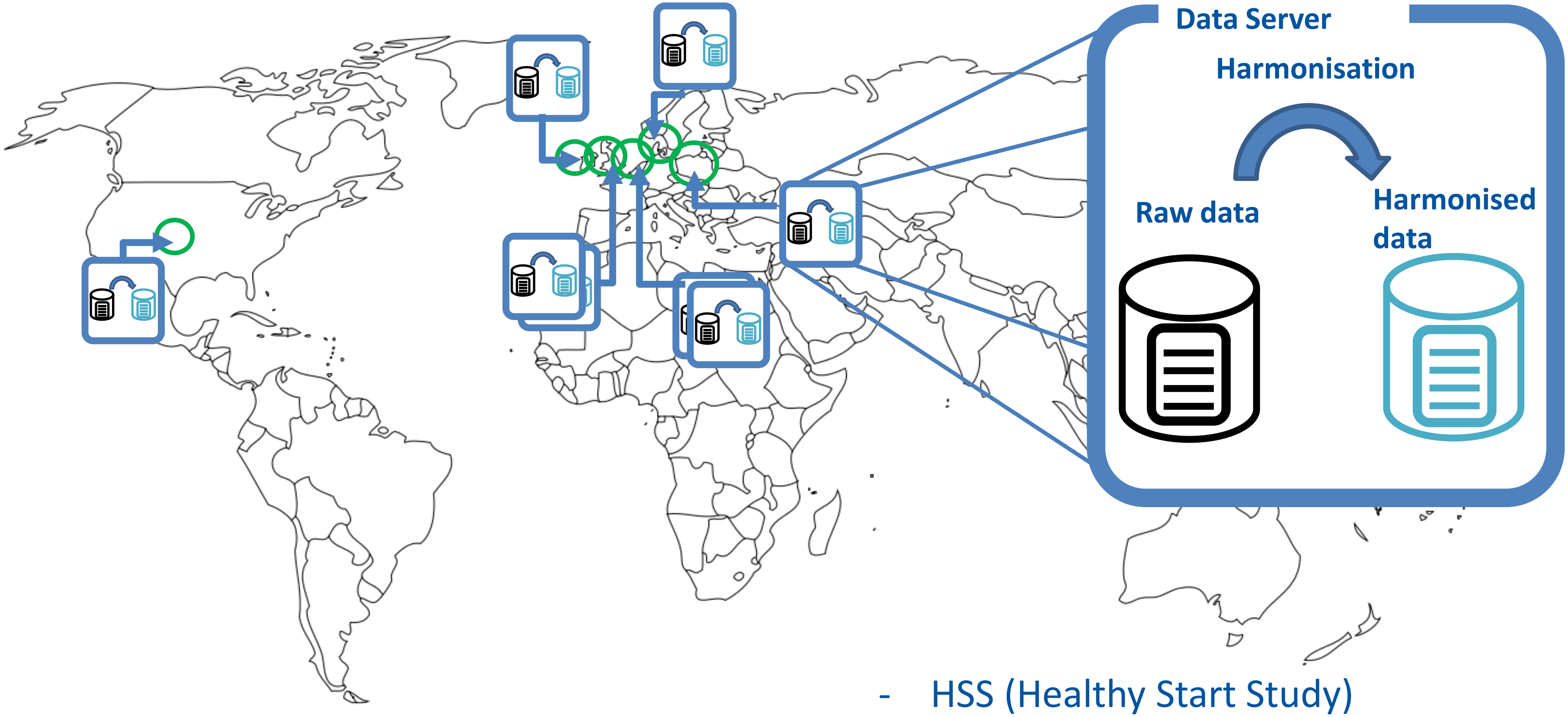
...more likely to have a caesarean section if not active during pregnancy

Activity ideas

Always chat with your instructor or midwife to make sure activities work for you

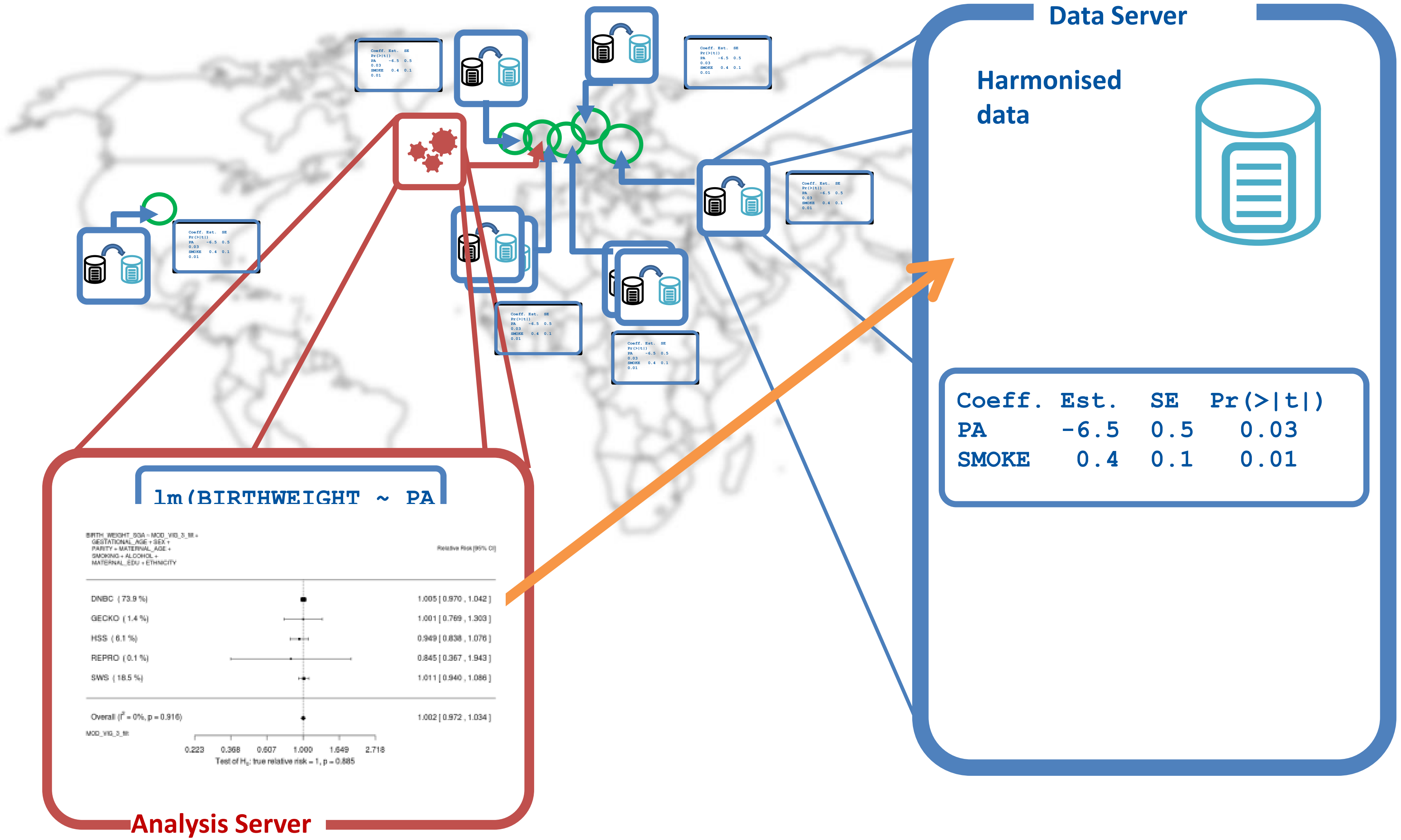
Source www.tommys.org

8 participating studies set up a server & prepared data



- ABCD (Amsterdam Born Children Development
- ALSPAC (Avon Longitudinal Study of Parents and Children)
- DNBC (Danish National Birth Cohort)
- GECKO Drenthe Study
- HSS (Healthy Start Study)
- REPRO (Polish Mother and Child Cohort Study)
- ROLO (RCT Of LOw glycaemic index diet)
- SWS (Southampton Women's Study)

How is an analysis run using DataSHIELD?



Data Server

Harmonised data

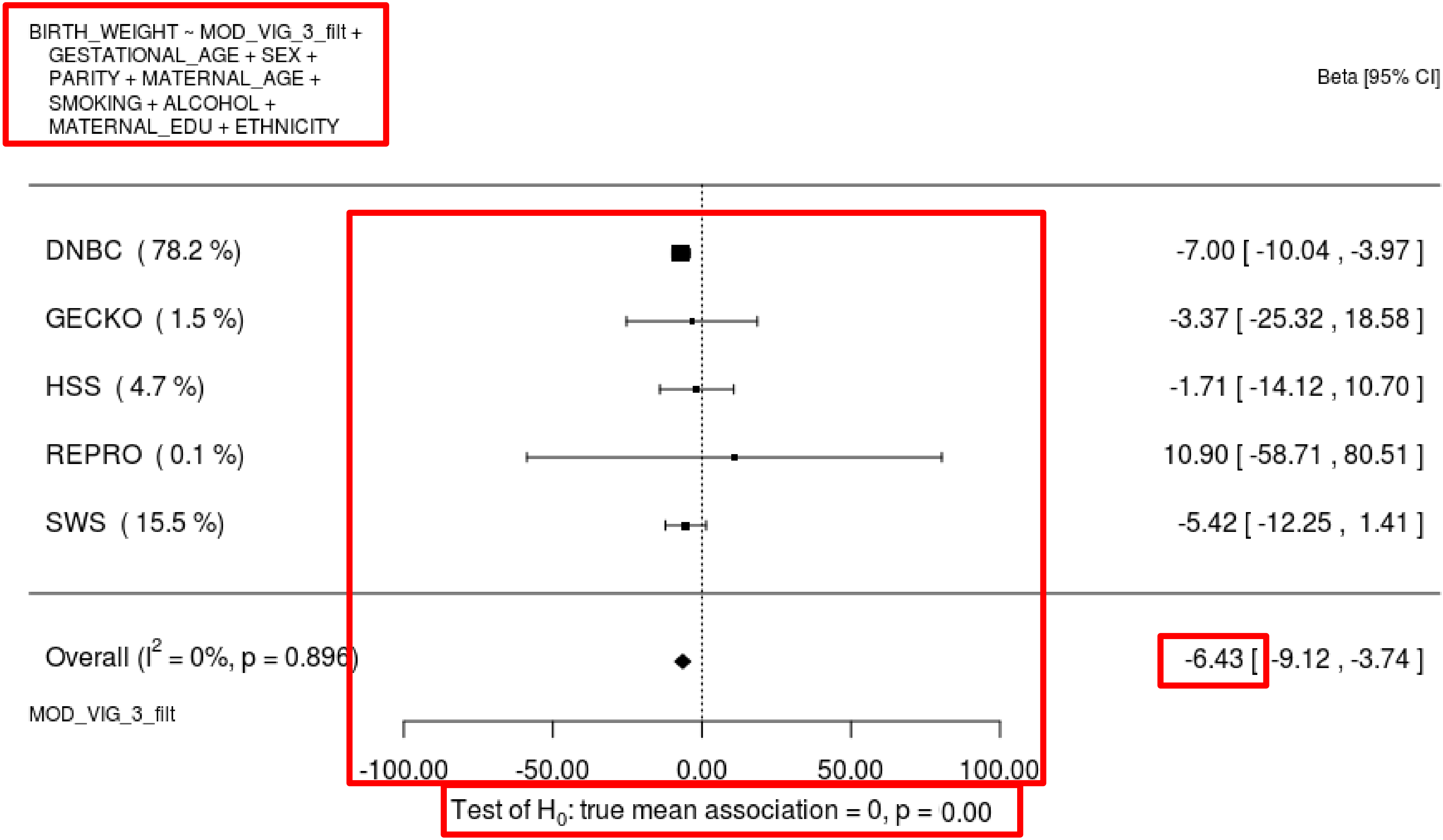
```
Coeff.  Est.   SE   Pr(>|t|)
PA      -6.5   0.5    0.03
SMOKE    0.4   0.1    0.01
```

```
lm(BIRTHWEIGHT ~ PA
```

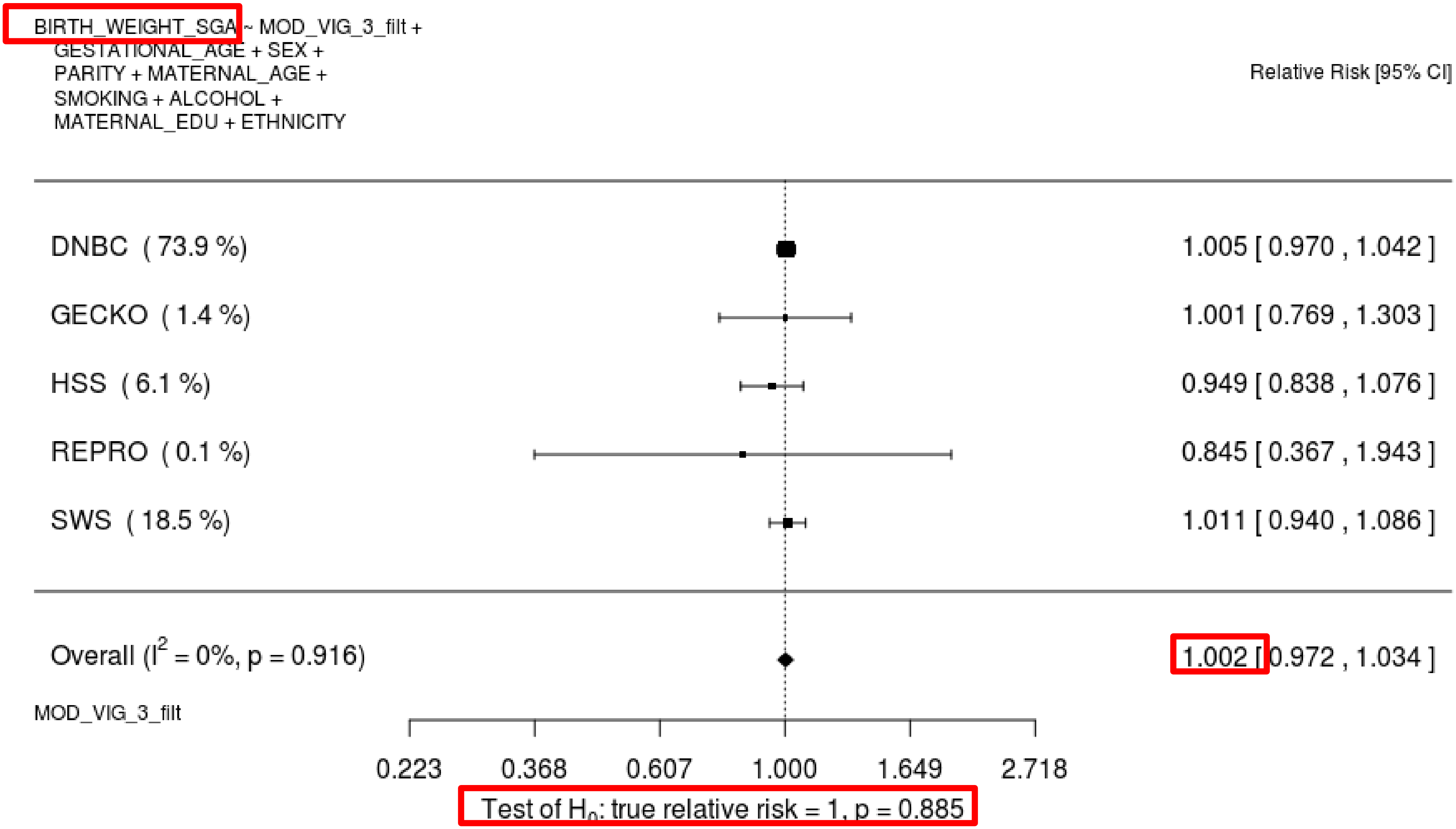
BIRTH_WEIGHT_SGA ~ MOD_VIG_3_filt +
GESTATIONAL_AGE + SEX +
PARITY + MATERNAL_AGE +
SMOKING + ALCOHOL +
MATERNAL_EDU + ETHNICITY
Relative Risk [95% CI]
DNBC (73.9 %) 1.005 [0.970 , 1.042]
GECKO (1.4 %) 1.001 [0.769 , 1.303]
HSS (6.1 %) 0.949 [0.838 , 1.076]
REPRO (0.1 %) 0.845 [0.367 , 1.943]
SWS (18.5 %) 1.011 [0.940 , 1.086]
Overall (I^2 = 0%, p = 0.916) 1.002 [0.972 , 1.034]
MOD_VIG_3_filt
0.223
0.368
0.607
1.000
1.649
2.718
Test of H_0: true relative risk = 1, p = 0.885

Analysis Server

Association between 3rd trimester physical activity and birthweight



Confirming that physical activity does not result in babies that are small for gestational age



Reflections / remaining challenges

- Harmonisation is bulk of work
- DataSHIELD is a good solution where appropriate
- Should be applied on data access and governance system that is already robust and resilient
- Inferential disclosure
- Cost recovery from projects using DataSHIELD

DataSHIELD - summary

- Cross dataset analysis is desirable, but comes with challenges if trying to share results or data
- DataSHIELD is a technology that dynamically takes analysis to the data, without ever having access to individual data values
- We have used DataSHIELD to obtain results that are relevant to public health, and are currently using it to do research in other areas

InterConnect

Global data for diabetes and obesity research

Acknowledgement

- This project is funded by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 602068.
- Software implemented in conjunction with Maelstrom Research (McGill) and Institute of Health and Society (Newcastle)

Connect with us

- www.interconnect-diabetes.eu
- InterConnect@mrc-epid.cam.ac.uk