Global data for diabetes and obesity research

New ways to collaborate – federated analysis
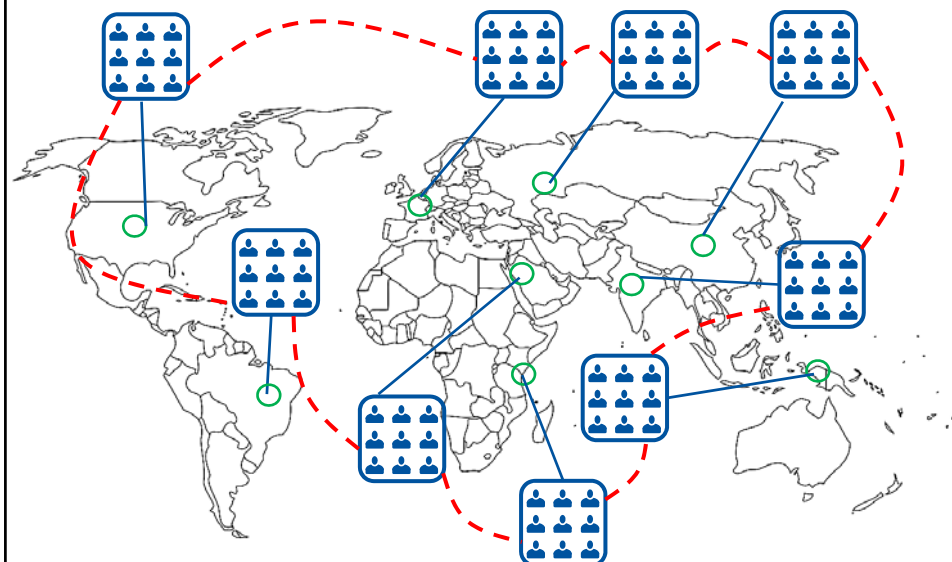
Tom Bishop
Senior Data Scientist
MRC Epidemiology Unit, University of Cambridge

Overview

- What is InterConnect and federated analysis?
- A real example, focus on harmonisation
- Review of current situation and what we are doing next

InterConnect offers a different way to run analyses over multiple studies

Meta analysis of published results – minimal/no harmonisation

	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 harmonise, run and re-run analyses
- Each group needs someone available to do work
- Harmonisation & analysis plan are open to different interpretations
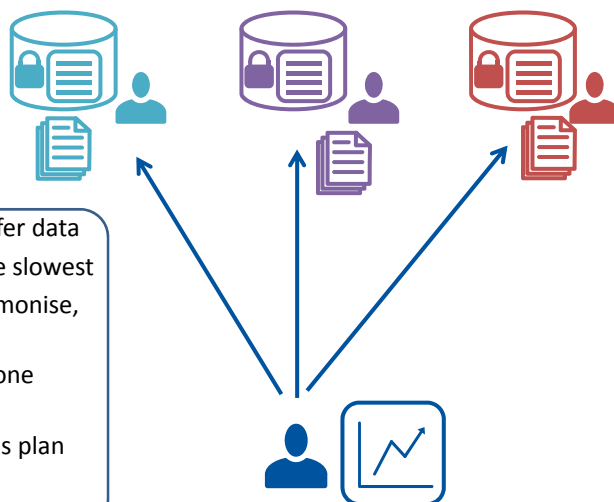
Data sharing

- 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

InterConnect addresses the challenges by taking the analysis to the data (federated analysis)
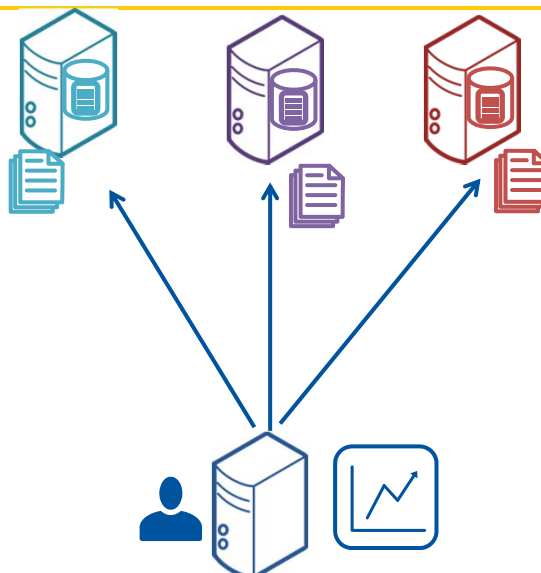
- Data stays on each study's server – **no data transfer agreement**
- Analytical commands passed to each server
- Summary results passed back – **no access to individual values**
- No waiting for others to run analysis
- No publication bias

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
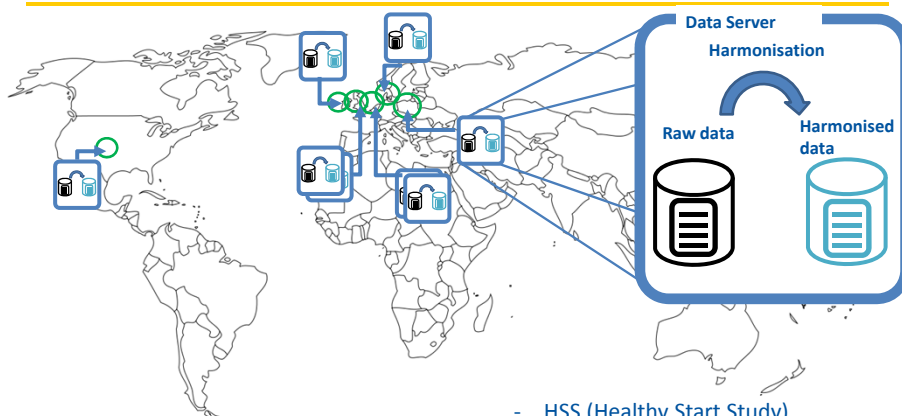
Activity ideas

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

Data Server

Harmonisation

Raw data → **Harmonised 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)

Harmonisation process

Review & refine

1 Confirm which studies will participate

2 Obtain meta-data for participating studies

vs.

3 Assess whether raw variables in each study can be transformed into target variables (harmonisation potential)

Requires studies to:
- **make meta-data available**
- **give assistance to enable data to be understood**

No access to the individual-level data

Review & refine

4 Design harmonisation algorithms for each study to convert raw data into harmonised data

7 Code catalogued for future use

Review & refine

Review & refine

6 Run summaries & then analysis on harmonised data

5 { /… } Transform harmonisation algorithms into code

Harmonisation example: Duration of moderate-vigorous leisure time physical activity

ALSPAC:

Question in questionnaire	Q1. HOW MUCH do you do the following at present? • Jogging • Aerobic • Antenatal exercises • Keep fit exercises • Yoga • Squash • Tennis/badminton • Swimming • Brisk walking • Weight training • Cycling • Other exercises
Frequency/duration	>7h/w, 2-6h/w, <1h/w, Never h/w = hours/week
Harmonised variable	**Duration of moderate/vigorous LTPA (h/w)**
Harmonisation rule	1. Convert: >7h/w to 7h/w 2-6h/w to 4h/w <1h/w to 0.5h/w Never to 0h/w 2. Sum up hours/w for all activities over 3 MET* (i.e. exclude antenatal exercises)
Missing data rule	Count single missing activity durations as 0 duration. If all activity durations missing, then mark participant as missing.

*MET = Metabolic Equivalent of Task

Example code for ALSPAC

```
var LTPA_dur = ['b050', 'b051','b052','b053','b054','b055','b056','b057','b058','b059','b060','b061'];
var i;
var LTPA_dur_val = newValue(0, 'decimal');

var val_to_add = newValue(0, 'decimal');
var LTPA_sum = newValue(0, 'decimal');
var LTPA_miss = newValue(0, 'integer');

//no imputation - missing values counted as 0
// if they have a -7 for every answer, they didn't do the questionnaire so count as missing
// if they have -1s then they did the questionnaire but didnt fill anything in - assume no activity

for (i = 0; i < LTPA_dur.length; i++) {
  LTPA_dur_val = $(LTPA_dur[i]).map({
    '1': '7',
    '2': '4',
    '3': '0.5',
    '4': '0',
    '9': '0',
    '-1': '0',
    '-9999': '-1',
    '-7': '-7'
    },
  null,
  null);
  if(LTPA_dur_val != -7 && LTPA_dur_val != -1 && !LTPA_dur_val.isNull().value()){
    val_to_add = Number(LTPA_dur_val);
    LTPA_sum = LTPA_sum + val_to_add;
  } else if(LTPA_dur_val == -7) {
    LTPA_miss = LTPA_miss + 1;
  } else {
    LTPA_sum = LTPA_sum;
  }
}
if(LTPA_miss == LTPA_dur.length) {
  LTPA_sum = -1;
}
LTPA_sum;
```

Harmonisation example: Duration of moderate-vigorous leisure time physical activity

ABCD:

*MET = Metabolic Equivalent of Task

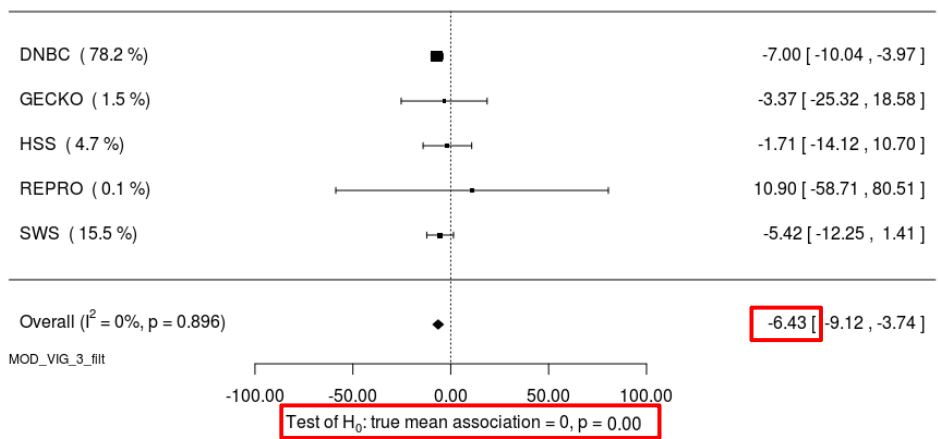| Question in questionnaire	In your spare time did you: **1.** Did you take **walks** for fun in the past week? **2.** Did you **ride a bicycle** in the past week? **3.** Did you **play sport**s in the past week? (for example: tennis, handball, gymnastics, fitness, skating, and swimming) **For each question: At what PACE do you usually do this?** • relaxed pace • average pace • brisk pace **For each question: FOR HOW LONG do you usually do this?**
Frequency/duration	Continuous value (mins/week)
Harmonised variable	**Duration of moderate/vigorous exercise (h/w)**
Harmonisation rule	1. Using Q1-3, sum up mins/w for all activities over 3 MET* (i.e. exclude relaxed walking) 2. Convert to hours (divide by 60)
Missing data rule	Count single missing activity durations as 0 duration. If pace is missing, assume relaxed. If all activity durations missing, then mark participant as missing.

How is an analysis run using InterConnect?
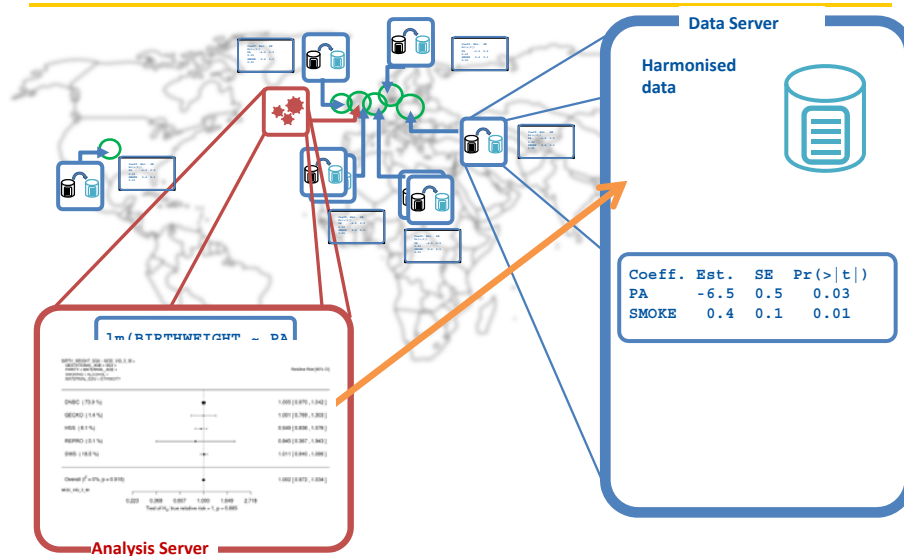
Association between 3rd trimester physical activity and birthweight

BIRTH_WEIGHT ~ MOD_VIG_3_filt + GESTATIONAL_AGE + SEX + PARITY + MATERNAL_AGE + SMOKING + ALCOHOL + MATERNAL_EDU + ETHNICITY

g per (hour PA per week) [95%CI]

DNBC (78.2 %)　　-7.00 [-10.04 , -3.97]
GECKO (1.5 %)　　-3.37 [-25.32 , 18.58]
HSS (4.7 %)　　-1.71 [-14.12 , 10.70]
REPRO (0.1 %)　　10.90 [-58.71 , 80.51]
SWS (15.5 %)　　-5.42 [-12.25 , 1.41]

Overall (I^2 = 0%, p = 0.896)　　-6.43 [-9.12 , -3.74]

MOD_VIG_3_filt

-100.00　　-50.00　　0.00　　50.00　　100.00

Test of H_0: true mean association = 0, p = 0.00

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

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
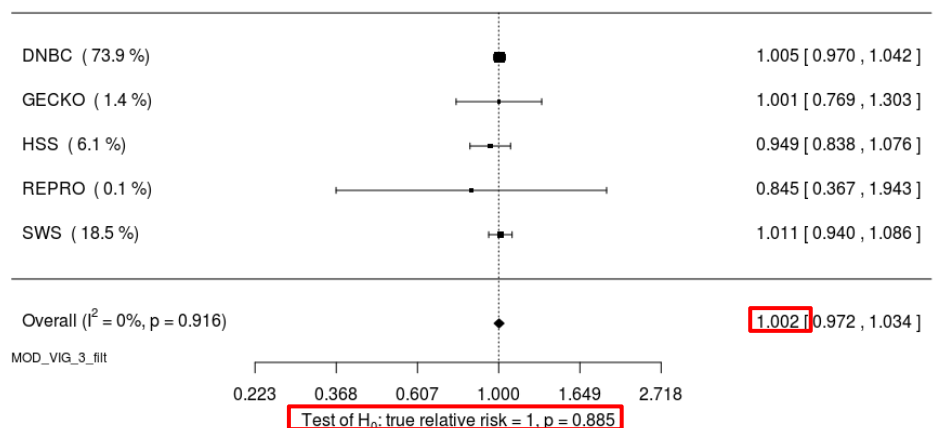
InterConnect - current status

- 18 servers set up
- 28 studies participating
- 1 publication, 2 papers in preparation, 2 analyses in progress
- Not just physical activity – diet, diabetes etc.

The future – EUCAN-Connect (Horizon 2020)
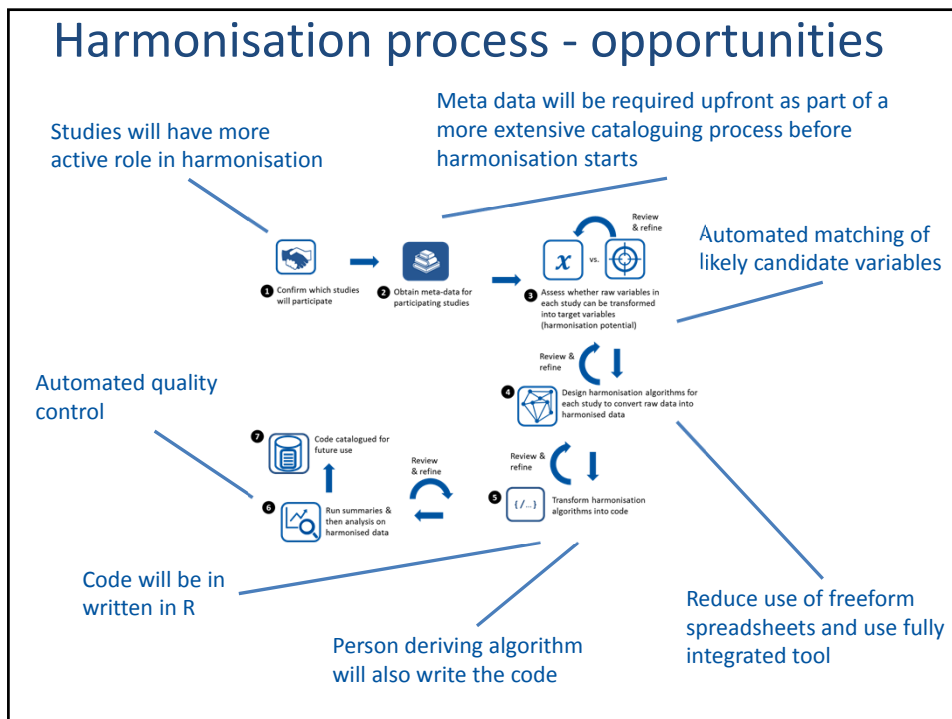
- €6M over 5 years
- Consortium of 13 institutions
- **EUCAN** – collaboration between EU and CANada
- **Connect** - Enable reuse of existing data sets to enable personalised medicine approaches

Harmonisation process - opportunities

Studies will have more active role in harmonisation

Meta data will be required upfront as part of a more extensive cataloguing process before harmonisation starts

Automated matching of likely candidate variables

Automated quality control

Code will be in written in R

Person deriving algorithm will also write the code

Reduce use of freeform spreadsheets and use fully integrated tool

① Confirm which studies will participate
② Obtain meta-data for participating studies
③ Assess whether raw variables in each study can be transformed into target variables (harmonisation potential)
④ Design harmonisation algorithms for each study to convert raw data into harmonised data
⑤ Transform harmonisation algorithms into code
⑥ Run summaries & then analysis on harmonised data
⑦ Code catalogued for future use

Review & refine

InterConnect - summary

- InterConnect dynamically takes analysis to the data (federated analysis) to reuse existing data
- We harmonised data across studies collaboratively without having access to the individual level values to obtain results for publication
- EUCAN-Connect gives opportunities to continue and extend this work

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