Big Data and Health Research
Health sector data

- Routine administrative data
  - Hospital admissions
  - Birth & Death certificates
  - Outpatient attendances
  - GP consultations
- Clinical data
  - Lab tests, immunisation, drug prescriptions etc
- Disease / operation registers
- Health surveys
Research

Epidemiology
- How much?
- Who gets it?
- What is it associated with?

Intervention Studies
Can you change it?
Epidemiology ....
Time trends in hospitalisations for suspected heart disease

**Under 75**

- Chest pain (M)
- Chest pain (F)
- AMI (M)
- US angina (M)
- AMI (F)
- US angina (F)

**Over 75**

- AMI (M)
- AMI (F)
- Chest pain (M)
- US angina (M)
- Chest pain (F)
- US angina (F)
Time trends in 30 day survival following heart attack

![Graph showing time trends in 30 day survival following heart attack for women and men. The graph illustrates a decrease in odds ratio over years, with women generally having a higher survival rate than men.]
Heart attacks by day of week
Children with special educational needs by month of conception

per 1,000 livebirths
Grandmaternal heart disease / stroke by infant weight
Interventional studies ...
“back to sleep” campaign for cot death
Smoke-free legislation and preterm deliveries
Neonatal deaths by date and time of birth

weekday 9-5
weekday 5-9
weekend
all out of hours
### Randomised Controlled Trials Follow-Up

<table>
<thead>
<tr>
<th>Record linkage</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>WOSCOPS</td>
<td>Yes</td>
<td>1043</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>137</td>
</tr>
</tbody>
</table>

**Graph A: CHD-Related Death or Nonfatal MI**

- **Placebo**
- **Pravastatin**

**Graph B: CHD-Related Death or Hospitalization**

- **Placebo**
- **Pravastatin**

**No. at Risk**

- **Placebo**
- **Pravastatin**
“Health-related” data from other sectors

- Education
- Housing
- Environment
  - pollution
  - weather
- Social services
- Word, pensions & benefits
- Criminal justice
The evolution of cross-sectoral record linkage
# Maternity-Education Linkage

<table>
<thead>
<tr>
<th>Health</th>
<th>Maternity record</th>
<th>Maternal factors (eg age, SIMD, smoking)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All deliveries</td>
<td>Obstetric factors (eg mode/gestation of delivery, urgency)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Infant factors (eg sex, birthweight, APGAR)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Education</th>
<th>Annual pupil census</th>
<th>Special educational need (learning/physical disabilities)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2006 onwards</td>
<td>Need type</td>
</tr>
<tr>
<td></td>
<td>Primary, secondary and special schools</td>
<td>Truancy/absence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Level of English</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Looked after status</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Free school meals</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annual school leavers destination survey</th>
<th>Employed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unemployed</td>
</tr>
<tr>
<td></td>
<td>Higher education</td>
</tr>
</tbody>
</table>

| Examination grades | All SCQF levels |
• 2006 requested pupil census data
• 3 years, and 20 meetings later ….
• approval to access centralised data denied
• approvals granted by 19/32 Local Authorities
Education produces two different extracts for ISD and investigators

<table>
<thead>
<tr>
<th>Pupil ID</th>
<th>Identifiers</th>
<th>Education data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>name, sex, dob, postcode</td>
<td>SEN, type school etc</td>
</tr>
</tbody>
</table>

Pupil ID
Identifiers
name, sex, dob, postcode

Education data
SEN, type school etc

ISD extract contains no education data only identifiers

ISD links extract to health data and then removes
The identifiers to produce an extract for the investigators

Investigators’ extract contains no identifier only education data

Investigators merge two extracts to produce one file containing health and education data, Investigators have no access to identifiers

<table>
<thead>
<tr>
<th>Pupil ID</th>
<th>Health data</th>
<th>Education data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pregnancy complications, vaccination status, antenatal markers etc</td>
<td>SEN, type school etc</td>
</tr>
</tbody>
</table>

Pupil ID
Health data
Pregnancy complications, vaccination status, antenatal markers etc
But finally ....

- 19 Local Authorities
- 75% Scottish population
- N=407,503 pupils
- SEN in 17,784 (4.9%) children
  - 1,565 (8.4%) of preterm
  - 16,219 (4.7%) of term
Gestation overall 10%
Preterm 3.6%
Early term 5.5%
2012 …..

- Cross-sectoral “collaboration”
- Data sharing agreement
- Central data

- No permission to release names
- Probability matching on DOB, sex, postcode
• Pupils 2006-2011
• Linkage achieved for
  – 94% 1 census year
  – 97% >1 census year
• N=803,275 pupils
• 674,705 singletons
  – 2,130 (0.3%) breech vaginal
  – 13,054 (1.9%) breech caesarean section
  – 461,571 (68.4%) cephalic vaginal
• % of breech delivered vaginally fell:
  – 23% (2006 pupils) to 7% (2011 pupils)
Delivery of breech infants
Vaginal versus Caesarean section

[Graphs showing data related to low Apgar score and special educational need versus low exam grades.]
Safe havens