University of Glasgow

CHIEF SCIENTIST OFFICE

MRC Medical Research Council

My experience with administrative data

Catherine Stewart & Ruth Dundas

MRC/CSO Social and Public Health Sciences Unit

9 February 2017

Overview

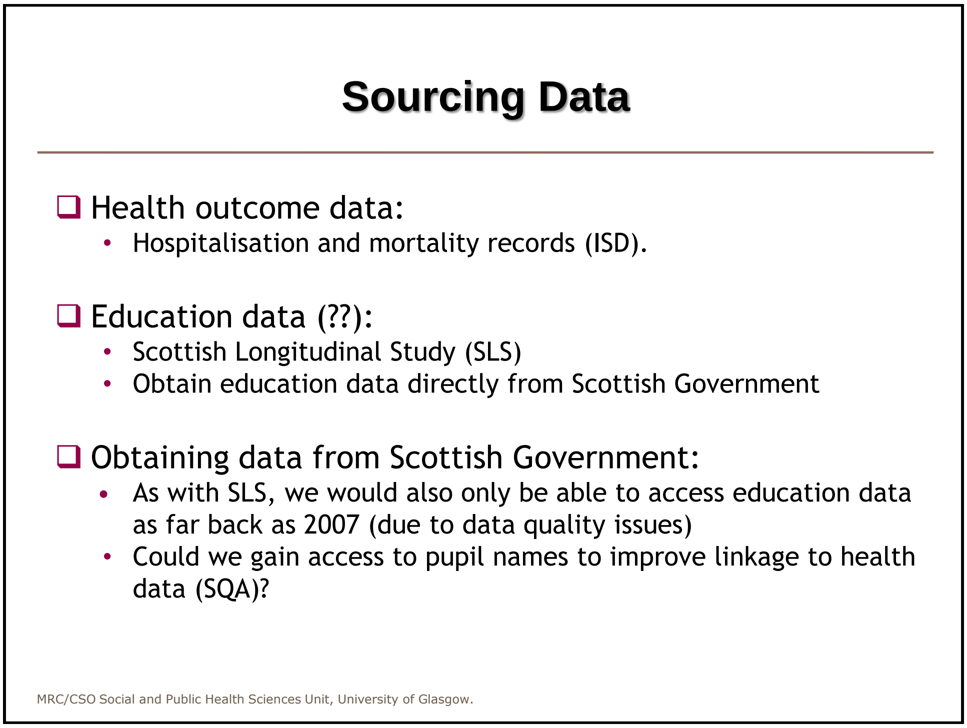
- ❑ Sourcing data
- ❑ Data application process
- ❑ Data linkage & transfer
- ❑ Data cleaning
- ❑ Benefits of using linked data
- ❑ Final reflections

Example

The importance of secondary school education in the patterning of health outcomes in Scotland

Background

- ❑ Broad aim: Investigate how various health outcomes in Scotland are patterned according to educational status.
- ❑ Particular focus on educational attainment at school-leaving.
- ❑ Several ways in which education may influence health:
  - Better education can lead to better job opportunities and income.
  - Better education can improve knowledge of how to live a healthy life and have a better understanding of how certain behaviours can affect health.



Sourcing Data

❑ Health outcome data:

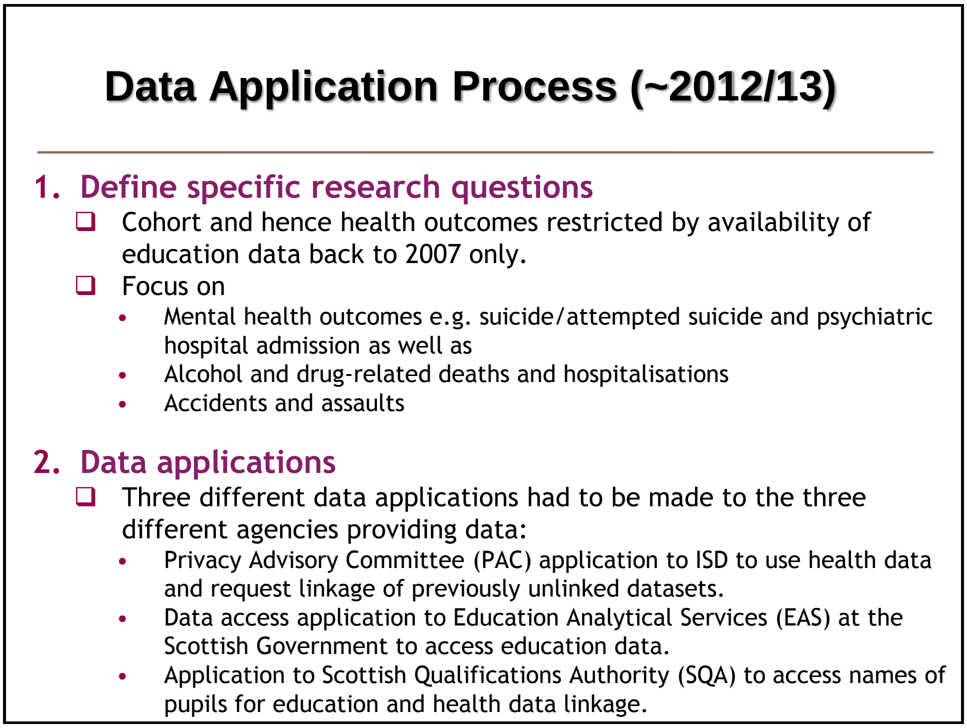
- Hospitalisation and mortality records (ISD).

❑ Education data (??):

- Scottish Longitudinal Study (SLS)
- Obtain education data directly from Scottish Government

❑ Obtaining data from Scottish Government:

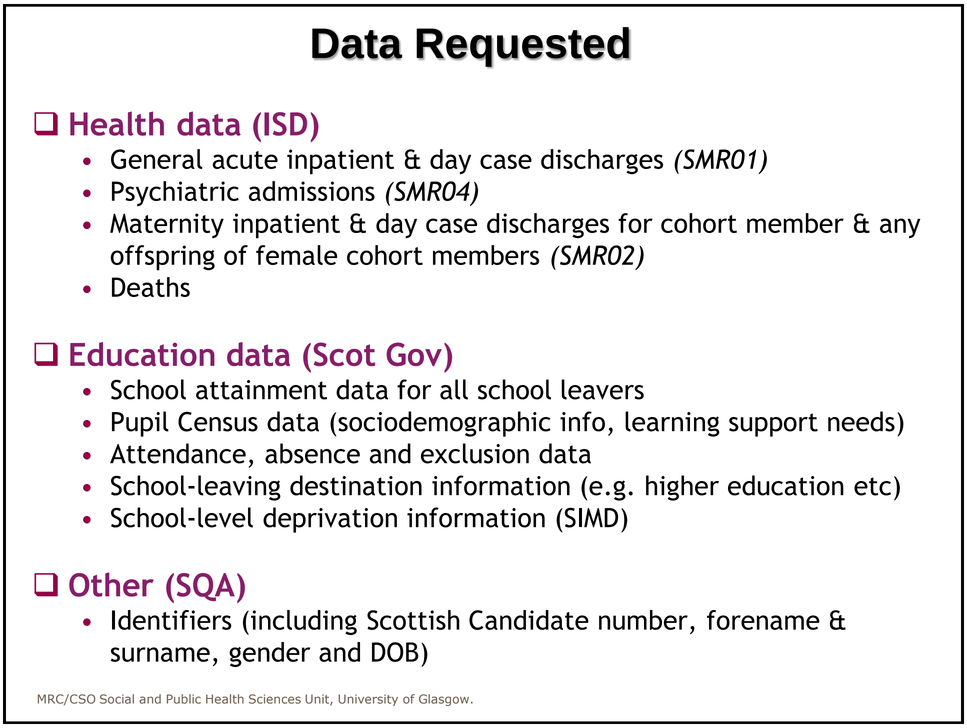
- As with SLS, we would also only be able to access education data as far back as 2007 (due to data quality issues)
- Could we gain access to pupil names to improve linkage to health data (SQA)?



Data Application Process (~2012/13)

1. **Define specific research questions**
   - ❑ Cohort and hence health outcomes restricted by availability of education data back to 2007 only.
   - ❑ Focus on
     - Mental health outcomes e.g. suicide/attempted suicide and psychiatric hospital admission as well as
     - Alcohol and drug-related deaths and hospitalisations
     - Accidents and assaults

2. **Data applications**
   - ❑ Three different data applications had to be made to the three different agencies providing data:
     - Privacy Advisory Committee (PAC) application to ISD to use health data and request linkage of previously unlinked datasets.
     - Data access application to Education Analytical Services (EAS) at the Scottish Government to access education data.
     - Application to Scottish Qualifications Authority (SQA) to access names of pupils for education and health data linkage.



Data Requested

□ **Health data (ISD)**

- General acute inpatient & day case discharges *(SMR01)*
- Psychiatric admissions *(SMR04)*
- Maternity inpatient & day case discharges for cohort member & any offspring of female cohort members *(SMR02)*
- Deaths

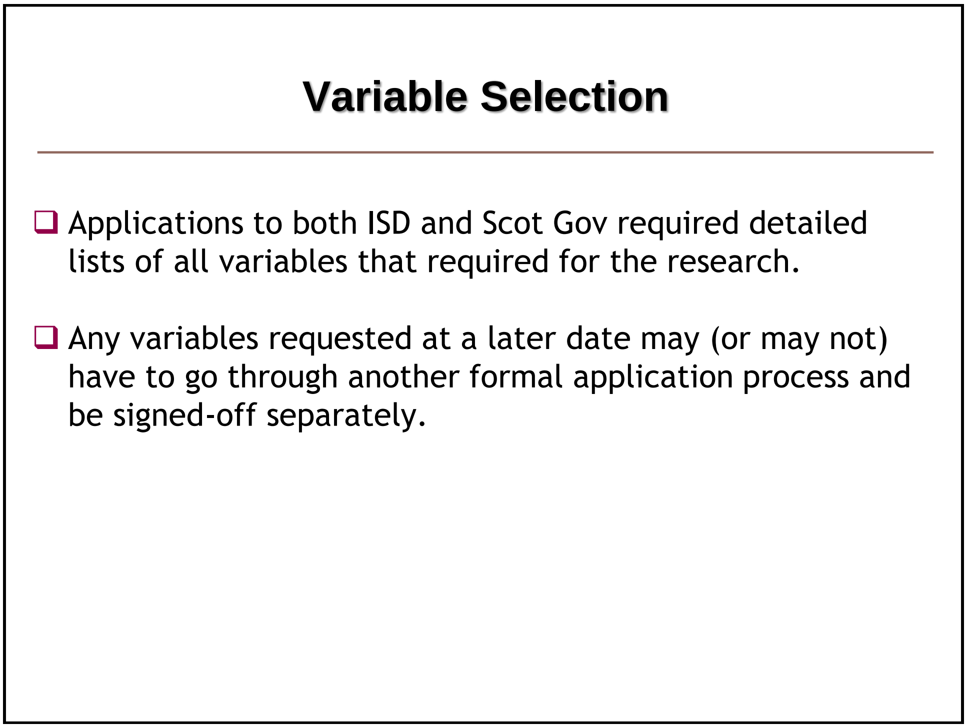
□ **Education data (Scot Gov)**

- School attainment data for all school leavers
- Pupil Census data (sociodemographic info, learning support needs)
- Attendance, absence and exclusion data
- School-leaving destination information (e.g. higher education etc)
- School-level deprivation information (SIMD)

□ **Other (SQA)**

- Identifiers (including Scottish Candidate number, forename & surname, gender and DOB)

MRC/CSO Social and Public Health Sciences Unit, University of Glasgow.



Variable Selection

□ Applications to both ISD and Scot Gov required detailed lists of all variables that required for the research.

□ Any variables requested at a later date may (or may not) have to go through another formal application process and be signed-off separately.

Data Extraction, Linkage & Transfer Process

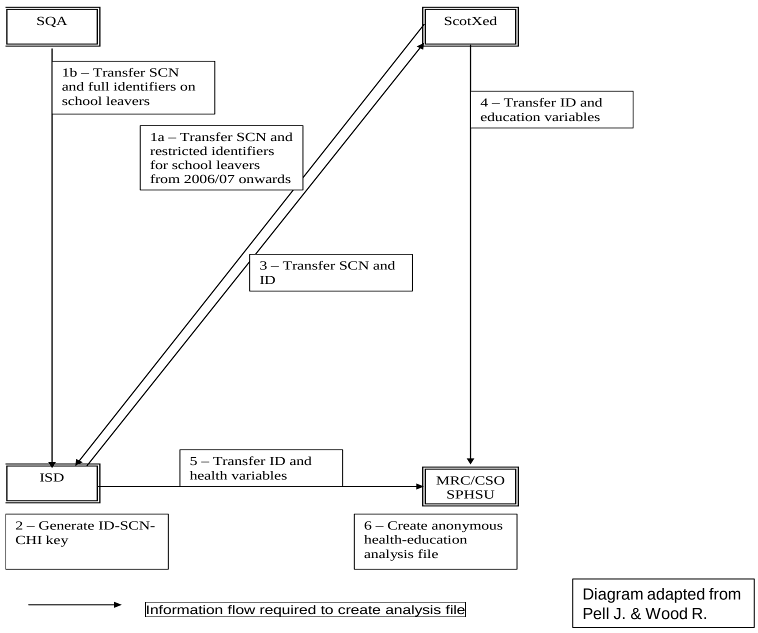
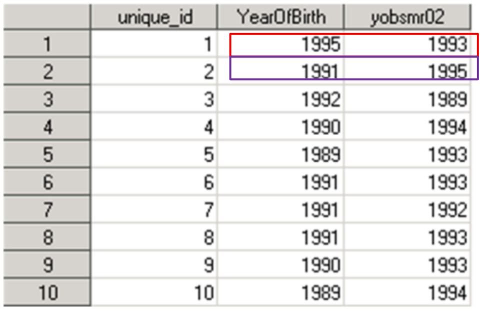

SQA
ScotXed
1b – Transfer SCN and full identifiers on school leavers
1a – Transfer SCN and restricted identifiers for school leavers from 2006/07 onwards
4 – Transfer ID and education variables
3 – Transfer SCN and ID
ISD
5 – Transfer ID and health variables
MRC/CSO SPHSU
2 – Generate ID-SCN-CHI key
6 – Create anonymous health-education analysis file
Information flow required to create analysis file

Diagram adapted from
Pell J. & Wood R.

Problems with the data (Received June 2013)

Major problems

- ❑ Health and education data did not appear to be referring to the same person when cross-checking on variables like gender and year of birth.



	unique_id	YearOfBirth	yobsmr02
1	1	1995	1993
2	2	1991	1995
3	3	1992	1989
4	4	1990	1994
5	5	1989	1993
6	6	1991	1993
7	7	1991	1992
8	8	1991	1993
9	9	1990	1993
10	10	1989	1994

- ❑ ISD had sent an old version of the anonymised ID to ScotXed for them to attach to the education data.

- ❑ Education data was very messy - inconsistencies within education data – having to check for consistency within individuals for all variables (very time-consuming!!).

Unique ID	Gender
1	M
1	M
1	M
1	M
2	M
2	M
2	M
3	F
3	F
3	F

☺

Unique ID	Gender
1	M
1	F
1	F
1	M
2	M
2	M
2	M
3	F
3	F
3	M

☹

- ❑ Data extraction problems – delete all education data (January 2014)!!
- ❑ New (cleaner!!) dataset received end February 2014.

Minor Problems (some examples)

- ❑ Death records for individuals who had further records (health and/or education) after date of death.
  - Most of these death records had been linked to individuals who were multiple birth babies and the death record was actually for their twin: delete death record.

- ❑ Mismatch between education and health records based on gender/YOB cross-checks: full exclusion

- ❑ Attainment data where the date of award was after supposed date of school-leaving.
  - Keep the attainment record if the date of award within 1 year of school-leaving.
  - Assumed this would capture courses that had been taken at school, but had been awarded at a later date due to late submission, but would exclude any courses taken at college.

Benefits

- ❑ Large datasets
  - Rare outcomes
- ❑ Range of confounders
- ❑ Natural experiments
  - Causal relationships

Opportunities for Publications

- ❑ **Inequalities in Perinatal outcomes**
  - Fairley L, Leyland AH. Social class inequalities in perinatal outcomes: Scotland 1980-2000. *Journal of Epidemiology & Community Health* 2006;60:31-36
  - Fairley L, Dundas R, Leyland AH. The influence of both individual and area based socioeconomic status on temporal trends in Caesarean sections in Scotland 1980-2000. *BMC Public Health* 2011;11:330
- ❑ **Educational effects on health of young adults**
  - Stewart CH, Leyland AH. The role of educational attainment in explaining the relationship between perinatal conditions and suicidal behaviour in young adults in Scotland: a prospective cohort study
  - Cohort profile paper
  - 4 conference presentations
- ❑ **Evaluation of the Health in Pregnancy Grant policy**
  - NIHR Report in press
  - 5 conference presentations

Research in Progress

- ❑ **Evaluation of the Healthy Start Voucher Scheme**
  - Linking survey data to routine data
  - NIHR Report; academic journals
  - 4 conference presentations
- ❑ **The health of Looked After Children in Scotland**
  - ❑ Linking administrative routinely collected data across sectors
    - ❑ Education and health
  - ❑ UBDC project
    - ❑ Facilitating application process
    - ❑ Liaising with data controllers
    - ❑ Providing expertise in data access agreements

Final Reflections: What I’ve Learned

- ❑ Linking previously unlinked data is a long process, but it can provide access to large, rich datasets.
- ❑ Document all the data cleaning decisions that have to be made and any cases that have to be excluded.
- ❑ Get in touch with data custodians sooner rather than later if data seem more ‘messy’ than expected.

Final Reflections: What could have been done better?

- ❑ Data custodians could have been better at suggesting further information that I would probably need e.g. continuous inpatient stay variable – chance conversation with colleague.
  - Having data agencies and ‘experts’ that know the data and what is available may help to overcome this.

Thank you

catherine.stewart@glasgow.ac.uk

ruth.dundas@glasgow.ac.uk

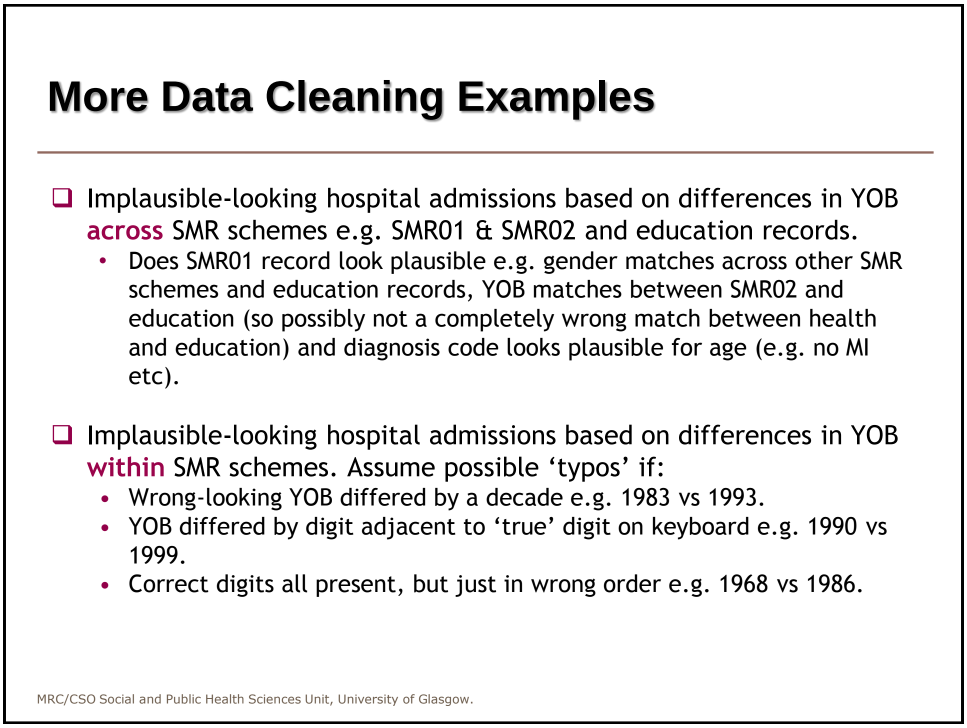
Variables that had to be requested at a later date

Health

- Continuous inpatient stay (no further PAC approval)
- Birth weights taken from SBR (no further PAC approval)
- GP de-registration date (further PAC approval required)

Education

- School deprivation measure (SIMD) (no further approval)
- Attainment data at SCQF levels 1 and 2 (further approval required and supplied with restrictions)



More Data Cleaning Examples

- ❑ Implausible-looking hospital admissions based on differences in YOB **across** SMR schemes e.g. SMR01 & SMR02 and education records.
  - Does SMR01 record look plausible e.g. gender matches across other SMR schemes and education records, YOB matches between SMR02 and education (so possibly not a completely wrong match between health and education) and diagnosis code looks plausible for age (e.g. no MI etc).

- ❑ Implausible-looking hospital admissions based on differences in YOB **within** SMR schemes. Assume possible ‘typos’ if:
  - Wrong-looking YOB differed by a decade e.g. 1983 vs 1993.
  - YOB differed by digit adjacent to ‘true’ digit on keyboard e.g. 1990 vs 1999.
  - Correct digits all present, but just in wrong order e.g. 1968 vs 1986.