MPH Advanced Epidemiology
February 2019

Using routine data for research

Dr Michael Fleming
Research Fellow
Public Health

Learning objectives

- Gain an understanding of
  - Sources of routine data that may be useful for public health research
  - Advantages and limitations of routine data
  - Data linkage
  - Ethical considerations when using routine data
  - How to access routine data for research purposes

What is ‘routine data’?

What is ‘routine data’

- Not collected specifically for a one off research project
- Ongoing / repeated collection
- Available for secondary analysis

- Data collection a by-product of ‘core business’ – administrative data
  - *Health records, prescriptions, benefits payments, exam results, prison records*
- Data collection the primary purpose – survey data
  - *Scottish Health Survey, Scottish Longitudinal Study*

- What kinds of routine data are available that may be useful for public health research?

Kinds of routine data

- Population
- Health sector data
- Non health sector data
- Population based data on health status, determinants, and outcomes
- Environmental data

Population Data

- Census
  - population, demographics, ethnicity, identity, language, religion, health, education, housing, transport, labour market, households
- Live births
- Deaths
- Mid year population estimates (census + births/deaths/migration)
- CHI database
  - register of all patients in NHS Scotland (dob+4 digits)
  - ensures that patients can be correctly identified
  - Estimates migration between council areas and below
- NHSCR
  - used to calculate moves between NHS Board areas within the UK

- National Records for Scotland http://www.nrscotland.gov.uk/

Health sector data

- Births and Deaths (NRS)
- Primary care
  - GP consultations (SPIRE)
  - Prescriptions (PIS)
  - Dental
  - Ophthalmic
- Secondary care
  - General hospital admission (SMR01)
  - Mental health admission (SMR04)
  - Obstetric admission (SMR02)
  - Neonatal admission (SMR11/SBR)
  - Outpatient attendance (SMR00)
- Unscheduled care
  - NHS24
  - A&E attendance
- Community care
  - Childhood immunisations (CHSP)
  - Child health reviews (CHSP)
- Event notification/registration
  - Cancer (SMR06)
  - Diabetes data (SC-DC)
  - Abortions
  - Stillbirths/Infant Deaths (SSBID)
  - Congenital anomalies (SMR11/SBR)
  - Smoking cessation
  - Drugs misuse data (SMR25)
  - Blood borne viruses (HIV/Hep C)
  - Healthcare Associate Infection

Deaths – National Records for Scotland http://www.nrscotland.gov.uk/
National health sector data – NHS Information Services Division http://www.isdscotland.org/
Infectious and environmental hazards https://www.hps.scot.nhs.uk/
Local health sector data – NHS Boards https://www.scot.nhs.uk/organisations/

Non health sector administrative data

- Education
  - ScotXEd pupil census
  - Leaver destination data, SQA exam results
- Welfare
  - Department of Work and Pensions
  - Benefits and credits (housing, unemployment, disability etc.)
- Social care
  - Home care, telecare, self directed support, meals
- Criminal justice
  - Scottish Prisoner Service databases

Health determinants, status, and outcomes

- Population based surveys
  - Scottish Health Survey (SHS)
    - Representative sample of the Scottish population
    - Asks questions about lifestyle factors and health outcomes
  - Scottish Longitudinal Study (SLS)
    - 5.3% random sample of the Scottish population
    - Large scale linkage study using administrative sources
    - cultural, demographic, economic, health, education, ecological, housing and social data.
  - Growing Up in Scotland (GUS)
    - longitudinal research study, launched in 2005 tracking the lives of thousands of children and their families from birth through teenage years and beyond
  - Health Behaviour in School Aged Children (HBSC)
    - collects data every four years on 11-, 13- and 15-year-old boys' and girls' health and well-being, social environments and health behaviours.

Sources – various!

UK Data Archive holds a wide range of archived survey data http://www.data-archive.ac.uk/about

Environmental data

- Not personal data
- Provides information on small areas (which can then be assigned to individuals according to their postcode of residence)
- Area deprivation measures eg Scottish Index of Multiple Deprivation
  http://www.scotland.gov.uk/Topics/Statistics/SIMD
- Environmental data eg air pollution, UV levels
  http://www.scottishairquality.co.uk/data/
- https://uk-air.defra.gov.uk/data/

- Advantages and limitations of using routine (administrative) data for research?

Advantages of using routine (administrative) data for research

- Low cost
- Quick
- Real world
- Long time trends
- National coverage
- Population based
- High completeness
- Hard to reach groups
- High volume
- Low selection bias
- Avoid recall bias in case control studies
- Low loss to follow up in cohort studies

Limitations of using routine (administrative) data for research

- Low detail
- Limited info on confounders
- Variable timeliness
- Variable accuracy/validity/quality
- Variable classification error
- Variable metadata
- Surveys bring a different set of issues e.g. non response bias, validity of measurement instruments, etc.

Questions to ask of routine (administrative) data

- Is proposed dataset fit for research purpose (relevance)
- Who/what event is captured by the data?
- Coverage (time period, population wide?)
- What are the data collection methods?
- Definitions? inbuilt validation? inbuilt coding?
- Accuracy and quality?
- Completeness/Robustness?
- Timeliness

- Is the data linkable to other sources!!

Coding classification systems

Low detail

- Healthcare groupings
- Statistical classifications
- Clinical terminologies

High detail

- ➢ Service planning and management
- ➢ Epidemiology
- ➢ Electronic clinical records

Adapted from Donaldson & Scally. Donaldson's Essential Public Health. 3rd ed. Radcliffe, 2009

Clinical terminologies

- Read code system used in primary care in Scotland (to be replaced by Systematised Nomenclature of Medicine – SNOMED)
- Standard terminology / thesaurus for (almost) all aspects of health care
- Health related circumstances/risk factors, symptoms, signs, diagnoses, treatment, procedures
- Clinicians enter clinical terms in free text into electronic records and the terminology system automatically suggests/assigns codes
- Used to construct searchable electronic patient records

Statistical classifications

- WHO International Statistical Classification of Diseases and Related Health Problems, 10th revision (ICD10) used for coding cause of death and diagnostic information on hospital discharge records
- Operations/procedures – OPCS Classification of Surgical Operations and Procedures, 4th revision (OPCS4) used for coding operation/procedure information on hospital discharge records
- Used to support statistical analysis

Healthcare groupings

- Healthcare Resource Groups (HRGs) used as a secondary classification system for healthcare events eg hospital discharge records
- Draw on a range of information available in the primary record (patient age, sex, main diagnosis, comorbidities, procedures done, complications) to classify records into a relatively small number of groups
- Any one group includes episodes that are similar in terms of both clinical characteristics and resource requirements
- Used mainly for service planning but also sometimes in research

Characteristics of classification systems

- Stable over time
- Process for additions / revisions
- All encompassing
- Mutually exclusive
- Supporting coding rules and index

- Clinical coding systems used in NHS Scotland
  http://www.isdscotland.org/Products-and-Services/Terminology-Services/Coding-and-Terminology-Systems/

Validation

- ‘To establish the soundness, accuracy, or legitimacy of’
- Often inherent in the data life cycle
  http://www.isdscotland.org/Products-and-Services/Data-Quality/
- External gold standard required
- Additional study specific validation required?

Data linkage

- Bringing together data from different records or databases that relate to the same individual
- Can involve administrative data, survey data, environmental data, and/or specially collected primary data (eg from a clinical trial)
- Deterministic (exact) matching on unique ID numbers e.g. CHI
- Probabilistic matching on a range of available identifiers eg name, DOB, gender, postcode and accepting those with a reasonable probability of relating to the same individual - probability weights and threshold setting
- Probability matching overcomes data quality issues
- Increases research potential but also increases privacy risk

Background

- When was record linkage first done ?
  - Initially developed in 1950's
  - Pioneered computer assisted 'record linkage' techniques
  - Originally a Geneticist

Howard B. Newcombe
(Statistical Society of Canada)

Background

- When was record linkage first done in Scotland?
  - Initially outlined by Heasman in 1968
  - National datasets held centrally in machine readable form
  - Routine Recording of Patient Identifiers
- How has record linkage evolved?
  - Computing capacity & data storage
  - Increasing demand for epidemiological and health service research
  - Creation of permanent, dynamic “linked” datasets
- Why use “Probability Matching” ?
  - Exact matching leads to inexact results
  - Quantifies the implications of levels of agreement and disagreement

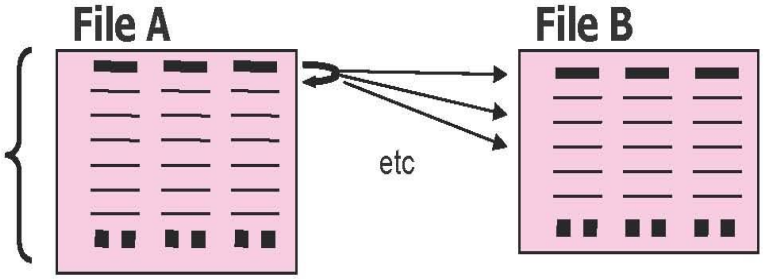
Linking records – Examples

Surname	Soundex	Forename	Date of birth	Post code	CHI number
Duck	D200	Donald	1 January 1955	DL1 2BY	0101551234
Dukk	D200	Donald	1 January 1955	DL2 3SL	0101551234
✗	✓	✓	✓	✗	✓

Surname	Soundex	Forename	Date of birth	Post code	CHI number
Swan	S500	Daisy	24 December 1959	AZ1 2BY	2412591234
Duck	D200	Daisy	24 December 1959	DL2 3SL	2412591234
✗	✗	✓	✓	✗	✓

Calculating probability weights

1. All A x B record pairs are created



2. For each variable estimate the probability that the variable agrees among matches (m_i) and the probability that the variable agrees among non-matches (u_i)

3. Assign (dis-)agreement weight for each linking variable based on m_i and u_i probabilities

$$\text{Agreement weight} = \log_2 \frac{m_i}{u_i} \qquad \text{Disagreement weight} = \log_2 \frac{1-m_i}{1-u_i}$$

4. Sum the individual weights to obtain a total linking weight for each record pair

What is a Soundex code?

‘a phonetic algorithm for indexing names by sound, as pronounced in English”

1. **Apply NYSIIS (New York State Intelligence Information System)**

   MAC → MCC
   PH → FF
   IE → EE
   SH → SCH

- **Translates commonly confused characters and removes all vowels**

What is a Soundex code?

2. The 'Soundex' code consists of the 1st letter of the name followed by 3 digits:

Soundex Table

1. **b, f, p, v**
2. **c, g, j, k, q, s, x, z**
3. **d, t**
4. **l**
5. **m, n**
6. **r**

(i) If 2 or more consecutive letters have the same code, code as 1 letter

(ii) Any remaining digits are set to zero

Examples:

S362 - Strachan, Streaking
S530 - Smith, Smythe
F450 – Fleming, Flaming, Fliming

Distribution of scores for pair comparisons

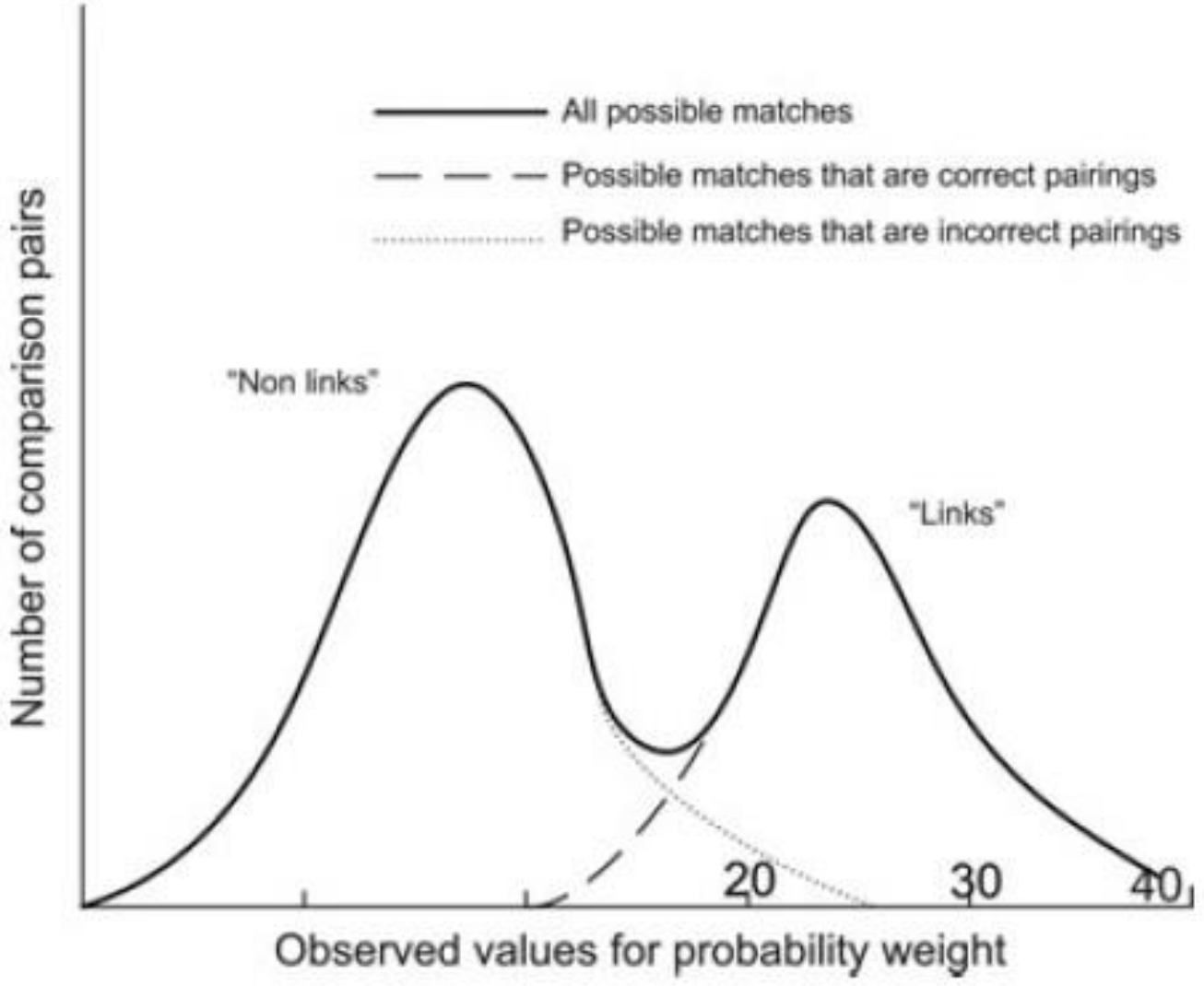


Figure 1 Distribution of probability weights (Mason & Tu 2008).

Setting the threshold

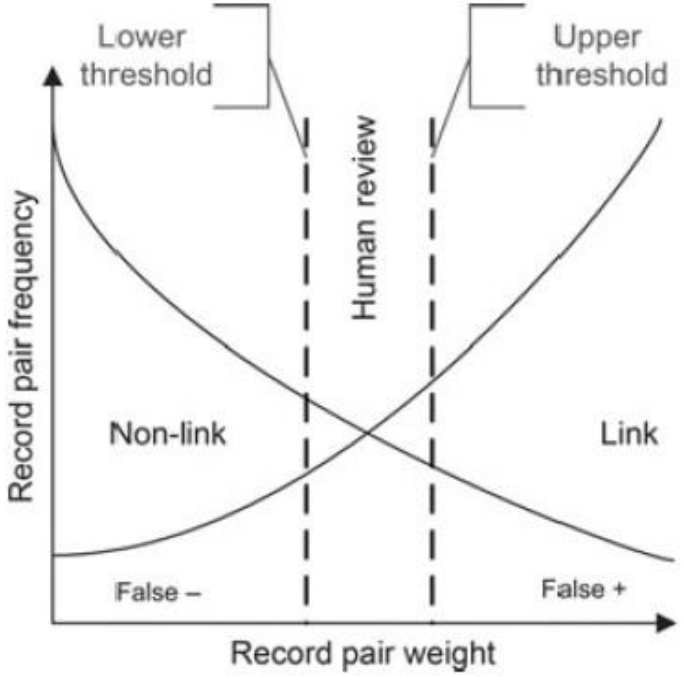


Figure 2 Two-threshold scheme for probabilistic scores using human review (Grannis *et al.* 2003).

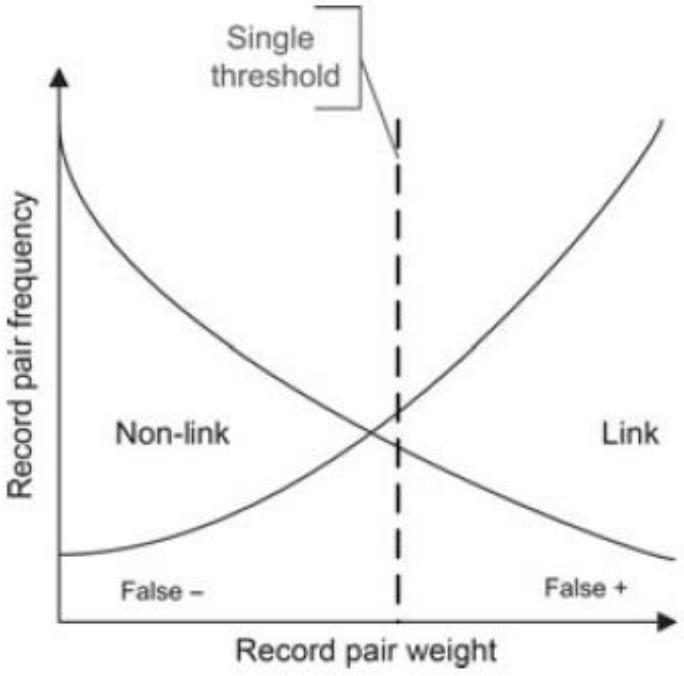
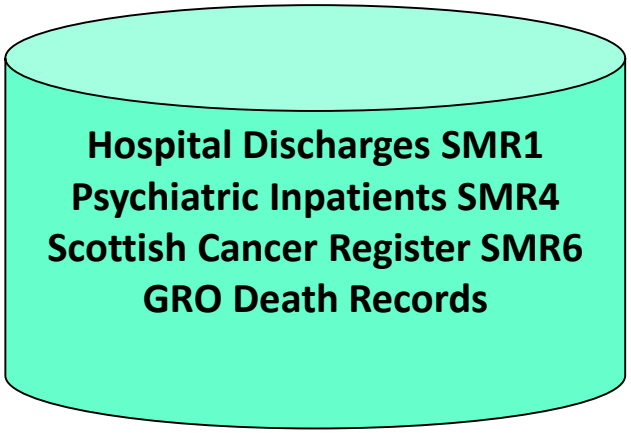


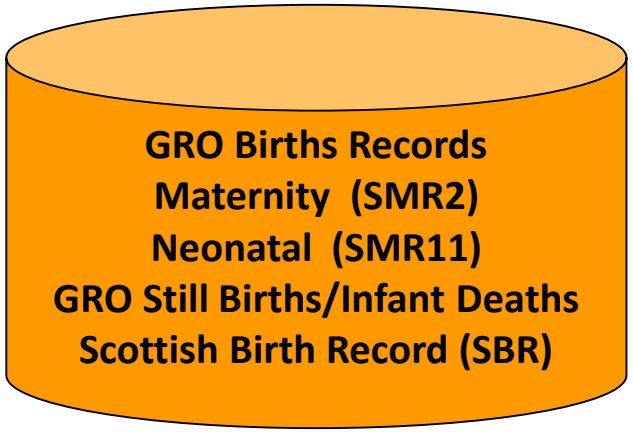
Figure 3 Single probabilistic score threshold (Grannis *et al.* 2003).

Record Linkage in Scotland

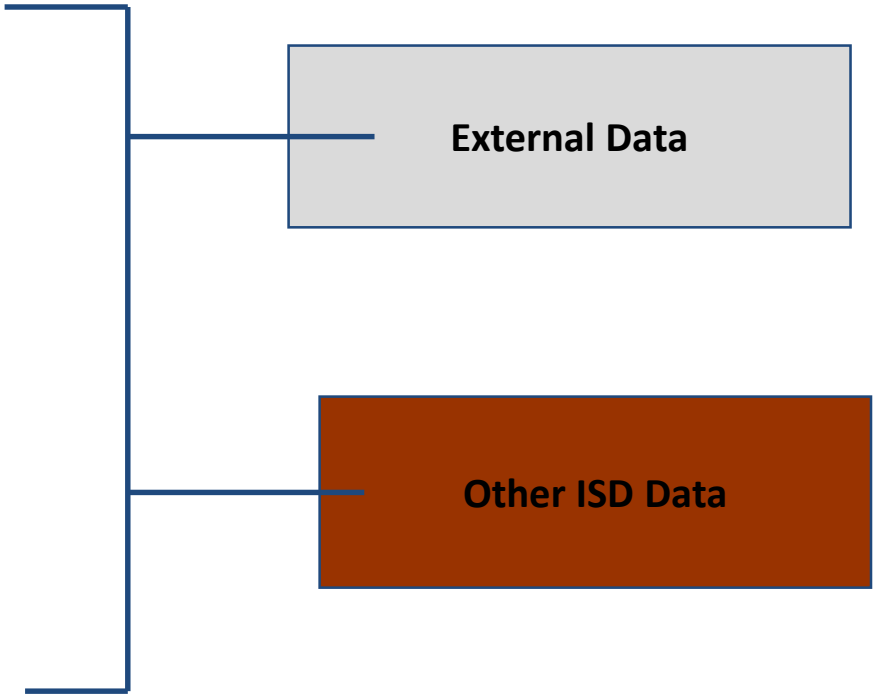
- **Linked "Acute" Database**



- **Linked "Maternity" Database**



Linkable data



CHI-seeding methodology

- Community Health Index (CHI) – Population register of patients registered with a GP Practice
- Full coverage in Scotland since 1988
- Unique Patient Identifiers for 10 million patients
- Conformed ISD Datasets seeded with CHI
- Automated CHI-seeding on sufficient external datasets = simpler & more efficient process

From conception through childhood..

Linking health

- Maternal antecedents
- Mothers ante-natal records
- Abortion Notifications
- Still birth/infant death
- Maternity record – CHI Number
- Neonatal record
- Register birth - NHS number
- Register with GP
- Child-health surveillance
- Immunisation
- GP Appointments
- Dental Appointments

Linking externally

- Education records
- Survey data
- Environmental data

…through adulthood till death

Linking health

- Outpatients
- NHS 24
- Prescribing
- GP appointments
- Sexual health
- Maternity events
- A&E attendance
- General hospital admission
- Well woman/man screening
- Cancer registration
- Cancer treatment
- Audit of clinical treatment
- Stroke - hospitalisation & rehab
- Community care
- Death

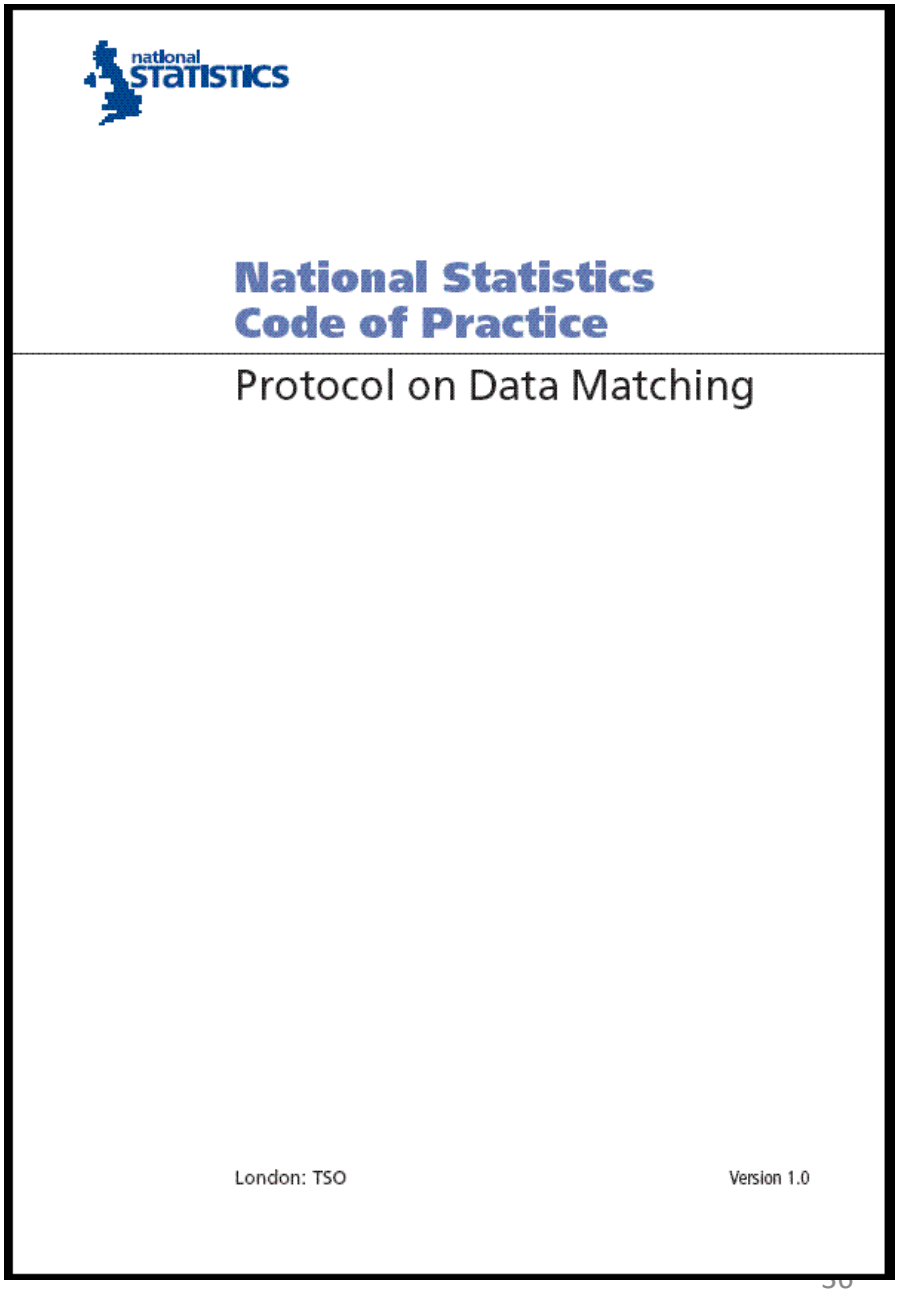
Linking externally

- Survey data
- Environmental data
- Education
- Welfare
- Criminal justice
- Disease registries

Rationale for data linkage

- Existence of disparate databases
- Added value of combining data to produce richer data sources
- Obligation not to waste public resource re-collecting data

"In order to reduce the burden on data providers and to fully exploit the value of existing statistical sources, data matching should be used in preference to creating new statistical sources, wherever possible, and where the results are likely to be of comparable quality."



Record linkage versus patient contact

- Record linkage can be as effective as reporting based on direct contact with patients….

- West of Scotland Coronary Prevention study Group. Computerised record linkage: Compared with traditional patient follow-up methods in clinical trials and illustrated in a prospective epidemiological study. Journal of Clinical Epidemiology, 1995, Volume 48 , Issue 12 , 1441 – 1452

- Nitsch D, Morton S, DeStavola BL, Clark H, Leon DA. How good is probabilistic record linkage to reconstruct reproductive histories? Results from the Aberdeen Children of the 1950s study. BMC Med Res Methodol. 2006 Mar 22;6:15

The age of big data!

- Real drive to link data across sectors
  - Education, welfare, justice, social care
- Wider determinants of health
- Emerging health datasets
  - Genetic data

Challenges

- Technical
  - Lack of sufficient identifiers
  - Ever more complex linkages
- Ethical
  - Consent
  - Increased risk of disclosure
  - Issues around ownership

JCN Journal of Clinical Nursing

Journal of
Clinical Nursing

GENERATION OF ROBUST DATA SETS

Record linkage in Scotland and its applications to health research

Michael Fleming, Brad Kirby and Kay I Penny

Aims and objectives. This paper will focus on the key concepts behind record linkage and describe how probability matching of Scottish health records can be used for national health research.

Background. Record linkage can bring together two or more records relating to the same individual. This allows information from multiple sources to be joined together to produce richer data sets for research purposes and has wide applicability in public health and epidemiological research. The probability matching techniques underpinning record linkage bring together records on a patient basis using key identifying information on each record. Scotland has a strong track record for performing linkage for research purposes owing to routinely collected and well-maintained national administrative health data sets, the emergence of the Scottish record linkage system and organisations like the Information Services Division of NHS National Services Scotland who centrally hold permanently linked patient-based databases.

Design. A record linkage retrospective population cohort study is described within this paper.

Methods. The paper will describe current linkage methodology before discussing typical applications in the setting of Information Services Division and focusing on a particular linkage study investigating rates and risk factors for gastroschisis.

Results. Conclusions from the gastroschisis study are typical of the types of important findings drawn from analysing linked health data.

Conclusions. Scotland's good track record for linking records for health research is evidenced by the high volume of research projects, publications and findings resulting from probability matching of national health data.

Relevance to clinical practice. Record linkage allows information relating to the same person held across different data sources to be brought together. Probabilistic record linkage can overcome data quality issues, producing accurate matches. This allows linked, analysable, patient-based databases, capable of answering complex research questions, to be produced from several data sources with wide applications in the field of health research.

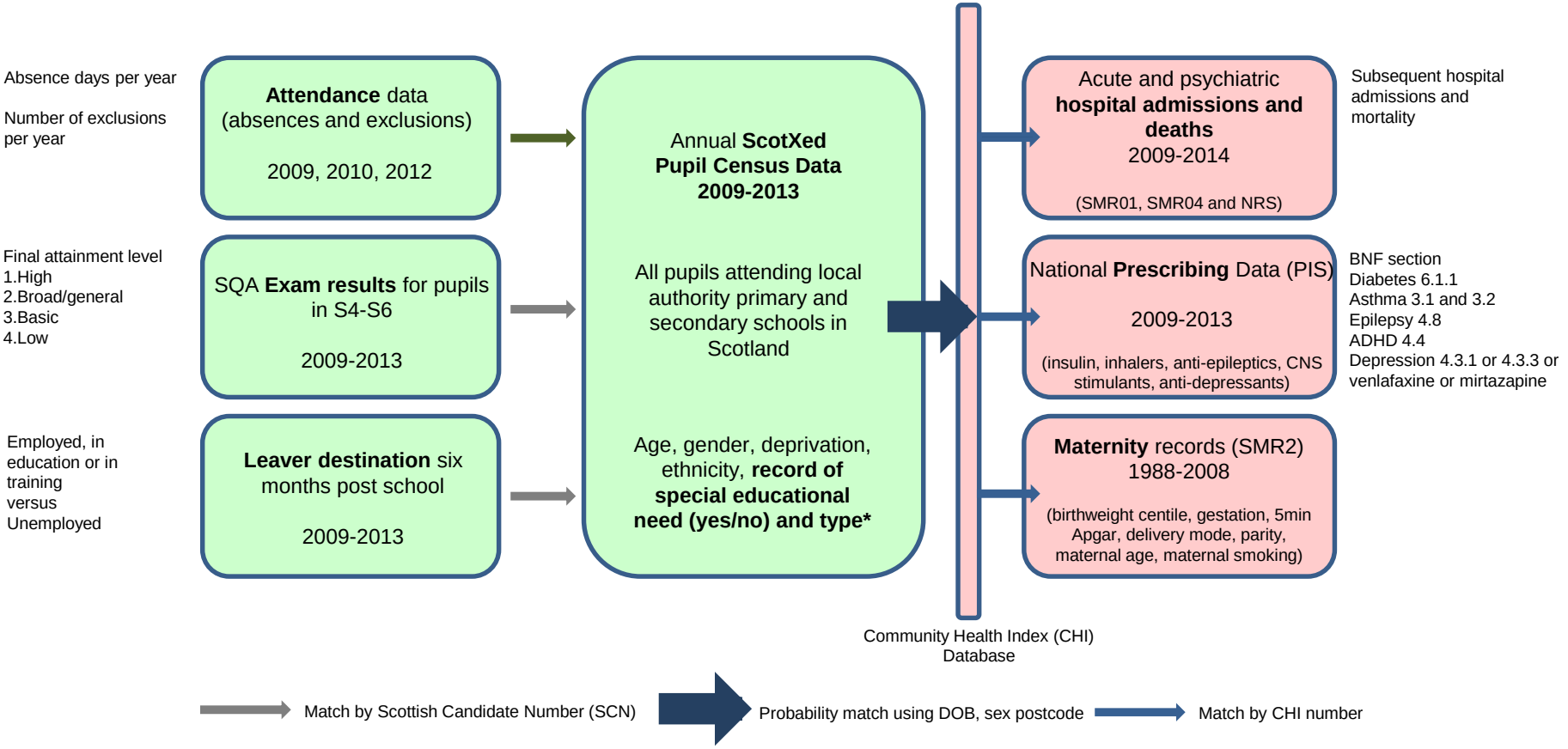
Key words: gastroschisis, nurses, nursing, probability matching, record linkage, statistics

Examples of public health research in Scotland based on linked routine data…..

- Maternal risk factors for congenital anomalies in offspring

- Educational and health outcomes of children treated for chronic conditions
  - Do schoolchildren with treatable chronic conditions have poorer educational and health outcomes compared to peer?
  - Conditions investigated were diabetes, asthma, epilepsy, ADHD and depression

Example: Linking education to heath data



* learning disability, learning difficulty, sensory impairment, physical motor impairment, communication problems, autism spectrum disorder, social/emotional/behavioural difficulty, physical health problem, mental health problem

- Ethical issues inherent in using routine data for research purposes?

Ethical issues

- ‘Consent or anonymise’ doesn’t really work for research on routine data
- Using personal data without explicit consent
- Risk of political misuse
- Risk of intrusion into private life and associated distress

Ethical issues v2

- Consent
  - for use of patient data
- Privacy
  - Should be maintained e.g. cannot identify people from outputs
- Autonomy
  - Patients have control and can decide independently if /how their data is used
- Property
  - Who owns the data? Does your data belong to you?
- Confidentiality
  - should be maintained – only the people who are supposed to be seeing it are seeing it
- Trust
  - From the public/owners – what is happening to your data?
- Public benefit
  - What are we doing with the data? Does benefit outweigh the risk?
- Social solidarity
  - Researchers/Governance professionals/data owners, data custodians/patients/public working together for the common good
- Reciprocity
  - exchanging things with others for mutual benefit especially privileges granted by one organisation to another
  - Organisations sharing information. Also patients and researchers working together

Legal framework governing use of personal data

- Common law duty of confidentiality
  - The general position is that, if information is given in circumstances where it is expected that a duty of confidence applies, that information cannot normally be disclosed without the information provider's consent.

- General Data Protection Regulation (GDPR)
  - is a regulation in EU law on data protection and privacy for all individuals
  - Supersedes Data Protection Act 1998 but regulates the collection, storage, and use of personal data significantly more strictly
  - Gives control to individuals over their personal data
  - Controllers of personal data must put in place appropriate technical and organisational measures to implement the data protection principles.

- Human Rights Act 1998
  - Act of parliament inncorporaing into UK law the rights contained in the European Convention on Human Rights
  - Unlawful for any public body to act in a way which is incompatible with the Convention

- (Section 60, Health & Social Care Act 2001 and Section 251, NHS Act 2006 – England only)

Debate within research community

- Recognition of value of routine data within research community
- Increased frustration about access difficulties – particularly around using identifiable data without explicit consent
- Complex legislative framework and professional guidance
- Unclear and bureaucratic access processes and inconsistent decision making
- More emphasis on privacy and autonomy than public benefit

Articulating the perceived problem

Lowrance W. Learning from experience: Privacy and the secondary use of data in health research. Research report. Nuffield Trust, 2002
http://www.nuffieldtrust.org.uk/sites/files/nuffield/publication/learning-from-experience-nov02.pdf

Personal data for public good: using health information in medical research. Academy of Medical Sciences, 2006.
http://www.acmedsci.ac.uk/viewFile/publicationDownloads/Personal.pdf

Suggesting a solution

A focus on proportionality and balancing public benefit and privacy concerns

Rumbold B et al. Access to person-level data in health care: Understanding information governance. Research summary. Nuffield Trust, 2011.

http://www.nuffieldtrust.org.uk/publications/access-person-level-data-health-care-understanding-information-governance

Public benefit vs. privacy and autonomy

Maximising public benefit

- Important issue
- High quality research
- Robust design/methods
- Lack of commercial interest
- Quality outputs
- Engage with policy makers
- Public engagement

Minimising privacy concerns

- Safe data – consent where feasible, anonymise as soon as possible, use the minimum data possible, control of disclosure risk
- Safe environment – secure data transfer, separate identifiers and rest of data, separate indexing, linking and analysis functions, safe haven facility
- Safe people – training, experience, accreditation
- Public engagement and public/patient involvement (PPI)

What does that mean in practice for researchers?

- eDRIS (Electronic data and research innovation service) http://www.isdscotland.org/Products-and-Services/eDRIS/
  - Research coordination
  - Support through governance/approvals processes
  - Indexing
  - Linkage
  - Data provision through safe haven

Adoption in Scotland

- Policy
  - Data Linkage Framework 2012
    - Scottish Government are enabling responsible, efficient and effective data linkage by: improving the ethical and legal governance arrangements around data linkage and supporting increases in the technical capacity to securely and efficiently link statistical and administrative data
    - http://www.scotland.gov.uk/Topics/Statistics/datalinkageframework
  - Health and Biomedical Informatics Research Strategy for Scotland 2015
    - How is Scotland respond ing to the opportunities and challenges around the secure use of routinely collected patient data for research
    - http://www.gov.scot/Publications/2015/04/6687

Academic funding for research infrastructure development

- Health Data Research UK (HDR-UK)
  - will develop and apply cutting edge data science approaches in order to address the most pressing health research challenges facing the public.
  - Joint investment led by MRC supports world-leading research to develop cutting-edge analytical tools and methodologies to address the most pressing health research challenges https://www.hdruk.ac.uk
  - Supersedes MRC funded Farr Institute http://www.farrinstitute.org/

- Administrative Data Research Network
  - ESRC funded
  - promotes wider social research using routine data http://www.adrn.ac.uk/

- Governance/approvals required for research involving routine data?

Governance/approvals required for research involving routine health data

- NHS Ethical approval
  http://www.nhsresearchscotland.org.uk/226_Research+Ethics.html

- Public Benefit and Privacy Panel approval
- http://www.informationgovernance.scot.nhs.uk/

  Replaced
  - Privacy Advisory Committee
  - Caldicott Guardians' Forum
  - CHI (community health index) advisory group

- Data sharing and data processing agreements for non health data

- Safe haven training

Approvals to link education to health data

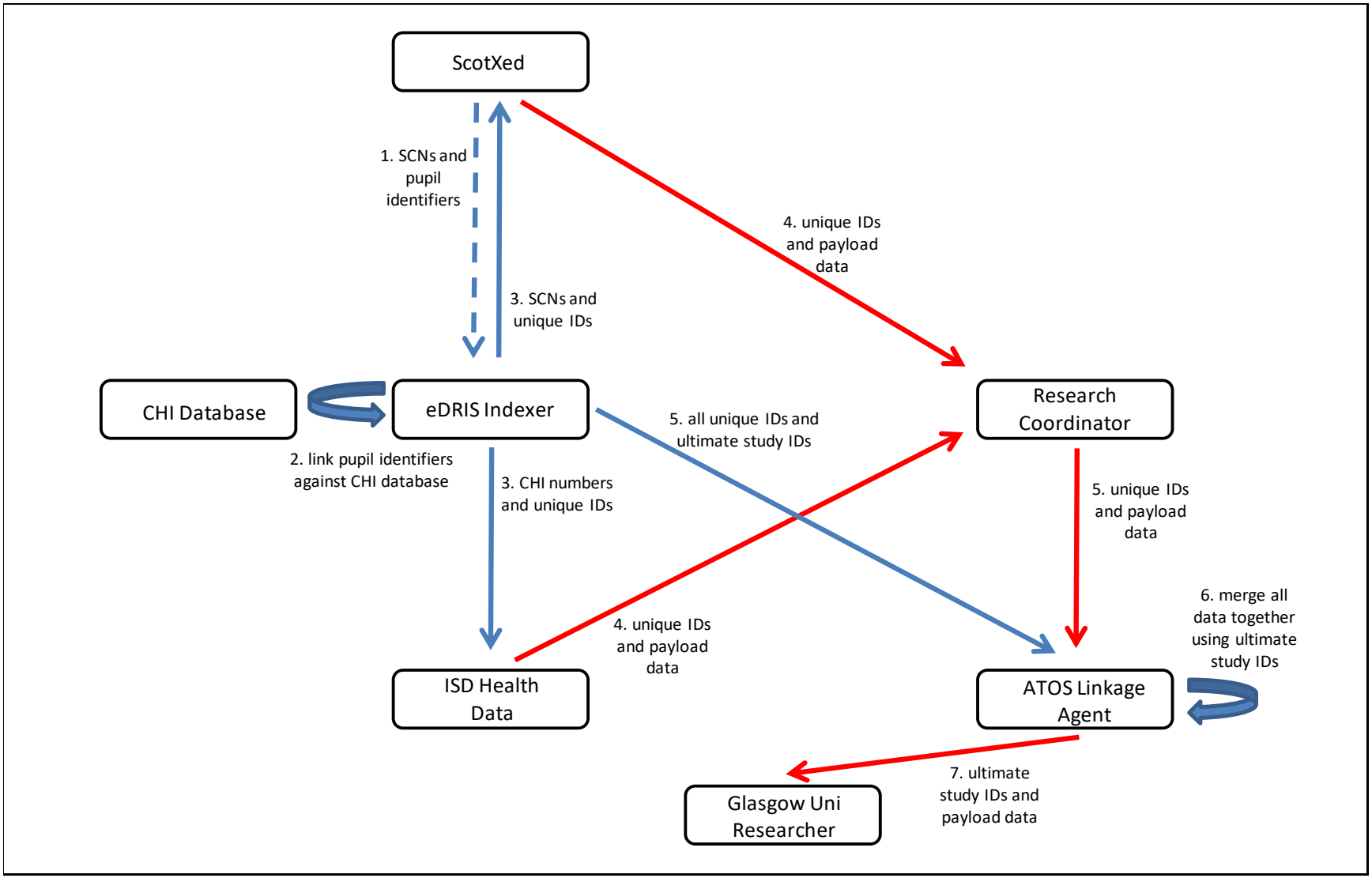
- NHS NSS Privacy Advisory Committee (PAC)
  - 9months to get approval back in 2014
- NHS West of Scotland Research Ethics Service

- Data sharing agreement ~ ScotXed/GU/eDRIS (ISD)
- Data processing agreement

- Data linked and prepared
  - Approximately 7months back in 2014/15

- Data analysed in national safe haven

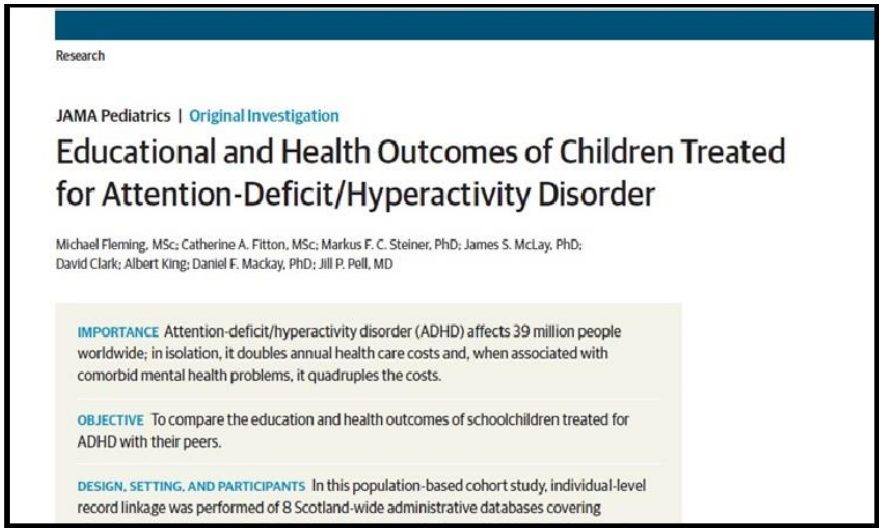
Example: Linkage of education to health data



Linking education to health data: roles and access to data

Data	Education data provider	Health data providers	eDRIS indexer	eDRIS linkage agent	Researcher
SCN / unique ID key / pupil identifiers	Yes	No	Yes	No	No
CHI / unique ID key	No	Yes	Yes	No	No
Unique IDs / overall study ID key	No	No	No	Yes	No
Health payload data	No	Yes	No	Yes	No
Education payload data	Yes	No	No	Yes	No
Linked health and education dataset with overall study IDs	No	No	No	No	Yes

Publications



Research

JAMA Pediatrics | Original Investigation

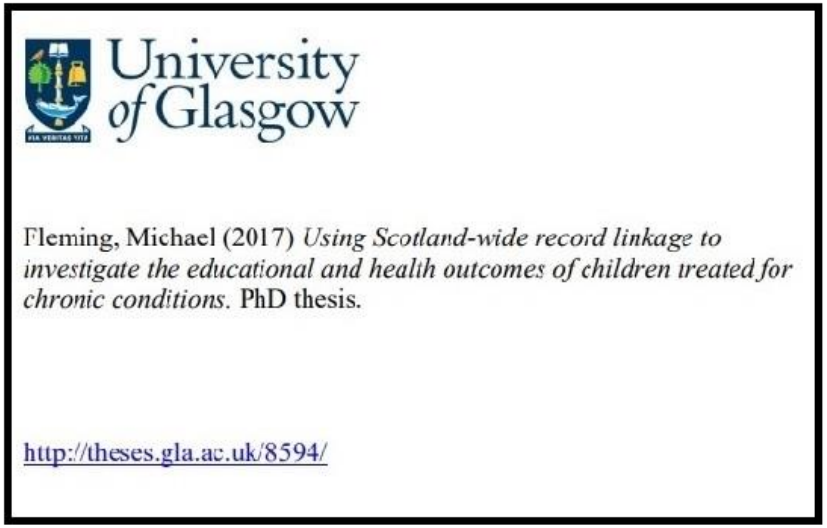
Educational and Health Outcomes of Children Treated for Attention-Deficit/Hyperactivity Disorder

Michael Fleming, MSc; Catherine A. Fitton, MSc; Markus F. C. Steiner, PhD; James S. McLay, PhD; David Clark; Albert King; Daniel F. Mackay, PhD; Jill P. Pell, MD

IMPORTANCE Attention-deficit/hyperactivity disorder (ADHD) affects 39 million people worldwide; in isolation, it doubles annual health care costs and, when associated with comorbid mental health problems, it quadruples the costs.

OBJECTIVE To compare the education and health outcomes of schoolchildren treated for ADHD with their peers.

DESIGN, SETTING, AND PARTICIPANTS In this population-based cohort study, individual-level record linkage was performed of 8 Scotland-wide administrative databases covering



University of Glasgow

Fleming, Michael (2017) *Using Scotland-wide record linkage to investigate the educational and health outcomes of children treated for chronic conditions.* PhD thesis.

http://theses.gla.ac.uk/8594/

Several papers in review for other conditions

Reviewer comments.....Big data problems?

Lots of data = lots of significance!

Suggested reading

- Basic introductions to available routine data in the UK and its strengths and limitations
  - Donaldson & Scally. Donaldson's essential public health. 3rd ed. Radcliffe Publishing Ltd, 2009. Chapter 1: Assessing the health of the population

  OR

  - Guest et al (eds). Oxford handbook of public health practice. Oxford University Press, 2013. Chapter 2.1: Understanding data, information, and knowledge

Further resources

- More detailed resources on sources of routine data in various setting and the advantages and limitations of using routine data for research
  - Detels et al (eds). Oxford textbook of public health. 5th ed. Oxford University Press, 2009.
    - Chapter 5.1: Information systems in support of public health in high income countries
    - Chapter 5.2: Information systems and community diagnosis in low and middle income countries
  - Rothman et al. Modern epidemiology. 3rd ed. Wolters Kluwer, 2009. Chapter 23: Using secondary data
  - Paper providing an overview of major health related routine data sources in the US
    - Smith et al 2011. Conducting high value secondary dataset analysis: an introductory guide and resources. J Gen Intern Med; 26(8): 920-9. doi 10.1007/s11606-010-1621-5.
  - Examples of survey based data aiming to compensate for a relative lack of routine health data in low income settings
    - Million Death Study http://www.cghr.org/index.php/projects/million-death-study-project/
    - UNICEF Multiple Indicator Cluster Survey http://www.unicef.org/statistics/index_24302.html
    - USAID Demographic and Health Surveys http://www.dhsprogram.com/

Further resources

- Guidance from the UK Information Commissioner's Office on the Data Protection Act 1998 https://ico.org.uk/for-organisations/guide-to-data-protection/
- Guidance from the General Medical Council on respecting confidentiality http://www.gmc-uk.org/guidance/ethical_guidance/confidentiality.asp
- Wellcome Trust 'spotlight' page on using personal data for research purposes http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Personal-information/index.htm