Opportunities and Challenges of Data Linkage for Longitudinal Surveys

Ray Chambers[1], Prerna Banati[2] & Natasha Codiroli McMaster[3]

1 University of Wollongong
2 UNICEF Office of Research - Innocenti, Florence
3 UCL & IoE, London

(with thanks to Bridget Taylor, Maria Sigala & James Fenner of ESRC)

Workshop on The Future of the HILDA Survey - Opportunities and Challenges
Melbourne, September 7, 2017

NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

UNIVERSITY OF WOLLONGONG

What is Data/Record Linkage/Matching?

OECD Glossary of Statistical Terms (http://stats.oecd.org/glossary/):

"Record linkage refers to a merging that brings together information from two or more sources of data with the object of consolidating facts concerning an individual or an event that are not available in any separate record."

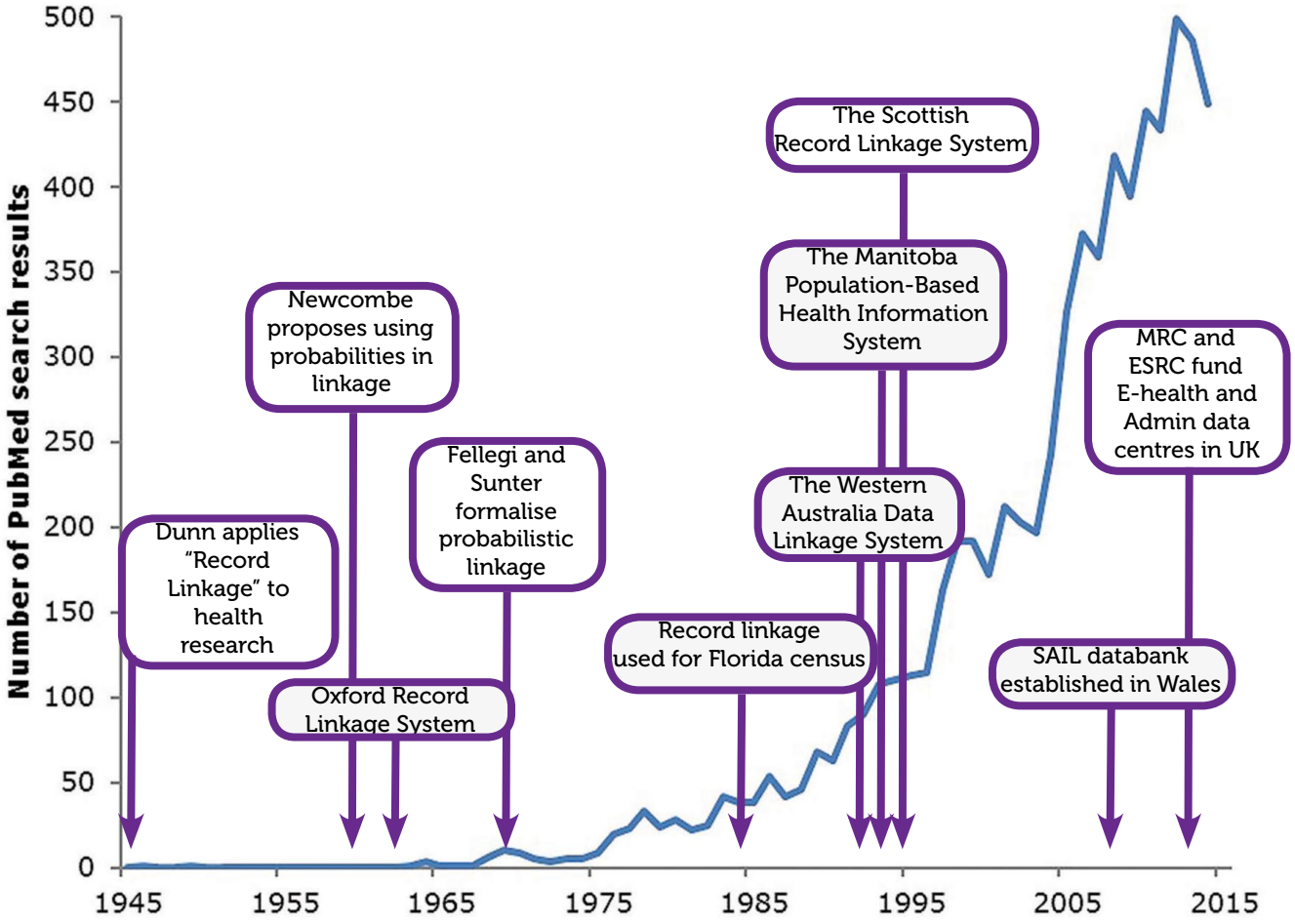
ADRN Definition (http://www.esrc.ac.uk/files/research/administrative-data-taskforce-adt/improving-access-for-research-and-policy/)

"Data linkage is the joining of two or more administrative or survey datasets using individual reference numbers/identifiers or statistical methods such as probabilistic matching."

- The sample survey paradigm underpinning data acquisition for scientific research is evolving, and data linkage is now an important research tool
  - … analysts use data linked from multiple sources to improve inference
  - … most notable in the health sector, where linkage is frequently employed to enhance data on clinical performance and patient health outcomes

- Longitudinal data linkage is the ability to link longitudinal survey data to a range of other (often also longitudinal) administrative data, such as health, tax, welfare and educational records, open or free data, as well as to 'big data' such as digital footprints

Data linkage timeline and number of PubMed search results by year of publication with search term “record linkage”



Source: Harron (2016)

Why Link Longitudinal Data?

- Some of the benefits of linkage (Boyd, 2017)
  - … more efficient data collection and lower participant burden
  - … increased information for correction of participant bias e.g. due to missing data. Linkage can increase completeness
  - … collection of information that cannot be obtained from participants, expanding the potential of a singular data set
  - … increase in representativeness and coverage
  - … allows study of sub-populations who are inadequately covered by the traditional data collection process, but still have substantial contact with service providers
  - … allows cohorts to be nested within populations and so facilitates population level analysis for highly specified sub-samples

- "The whole is greater than the sum of the parts"

A Typology of Longitudinal Data Linkage

1. Linking an established longitudinal data set to one or more administrative registers

Avon Longitudinal Study of Parents and Children (ALSPAC) is linked to DfE registers (National Pupil Database, Annual School Census), NHS registers (Hospital Episode Statistics, Clinical Practice Research Datalink), and ONS registers (Cancer Registry, Death Registry)

2. Linking central registers to administrative data to create a population level longitudinal data set

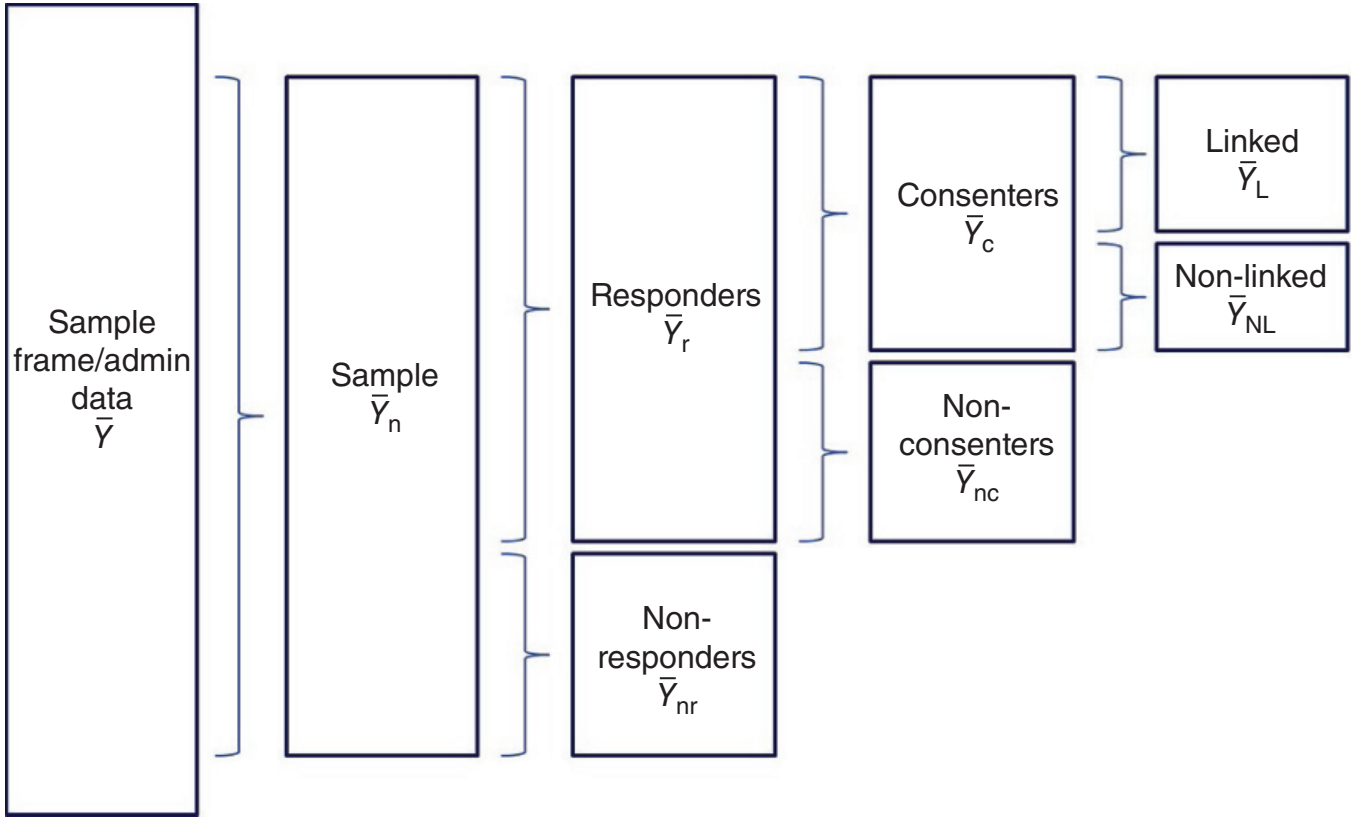
Brazilian 100 million cohort constructed by linking the Cadastra Unico register (all individuals receiving Bolsa Familia cash payments) to the registers making up the Brazilian Unified Health System

3. Contextual linkage used to enhance a longitudinal data set

Netherlands Cohort Study on Diet and Cancer links to geospatial coordinates recording particulate air pollution

Sources of Error in the Record Linkage Process

Note: Further errors in the linked cases (incorrect links)

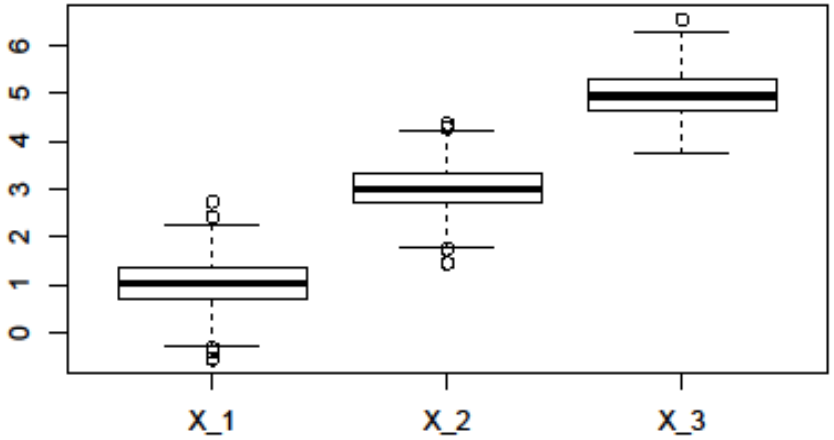


Source: Errors in Linking Survey and Administrative Data (Sakshaug, J.W. and Antoni, M. in Total Survey Error In Practice, eds. Biemer et al., Wiley)

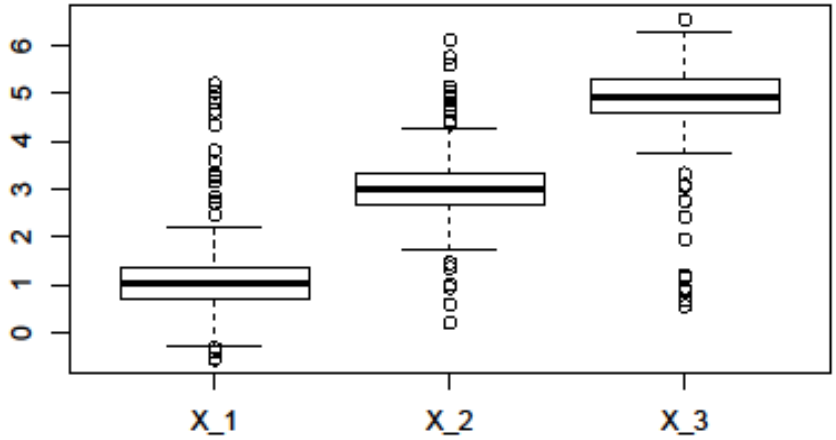
Simulation of Biases Due to Incorrect Linkage: 3 domains distributed differentially across 30 blocks (exchangeable linkage errors within blocks, overall 94% probability of correct linkage)

Block	Pr(Correct Linkage)	Domain Proportions			N
		1	2	3	
1-20	1.0	0.5	0.3	0.2	1000
21-26	0.9	0.3	0.5	0.2	300
27-30	0.7	0.2	0.3	0.5	200
Domain Size (Expected attribute value)		630 (1)	510 (3)	360 (5)	1500 (2.64)

Original data



Probability-linked data



Sources of Error in Linked Data

- Small amounts of linkage error can result in substantially biased results
  - … non-consents & missed links reduce sample size and result in a loss of power which can mean a potential selection bias.
  - … differential exclusion bias (Harron et al. 2016) - ethnic populations, women, and socially disadvantaged groups are less likely to be part of a linked cohort
  - … incorrect links (false matches) introduce variability and weaken association between variables, biasing towards the null
- Strategies for evaluating bias due to linkage error
  - … comparing linked data sets with a subset of 'gold standard' data
  - … undertaking sensitivity analyses using different linking criteria
  - … imputing uncertain links using missing data methods

- Ethics and Privacy
  - … linking data provides more information that can be used for identification, and potential breaches of confidentiality
  - … protecting the confidentiality of data can reduce statistical usefulness
  - … consent to linkage is not always sufficient to ensure public acceptability (Boyd, 2007)

- Governance
  - … linkage is inherently risky, but its benefits are large
  - … safe havens (secure analysis labs), accreditation, risk-based proportionate governance and improved researcher training can help with the risk - benefit tradeoff

Methodology of Data Linkage

- Deterministic Linkage (unique identifier)
  - … exact agreement on a common identifier (or a set of identifiers)
  - … minimises incorrect links
  - … problem with missed links

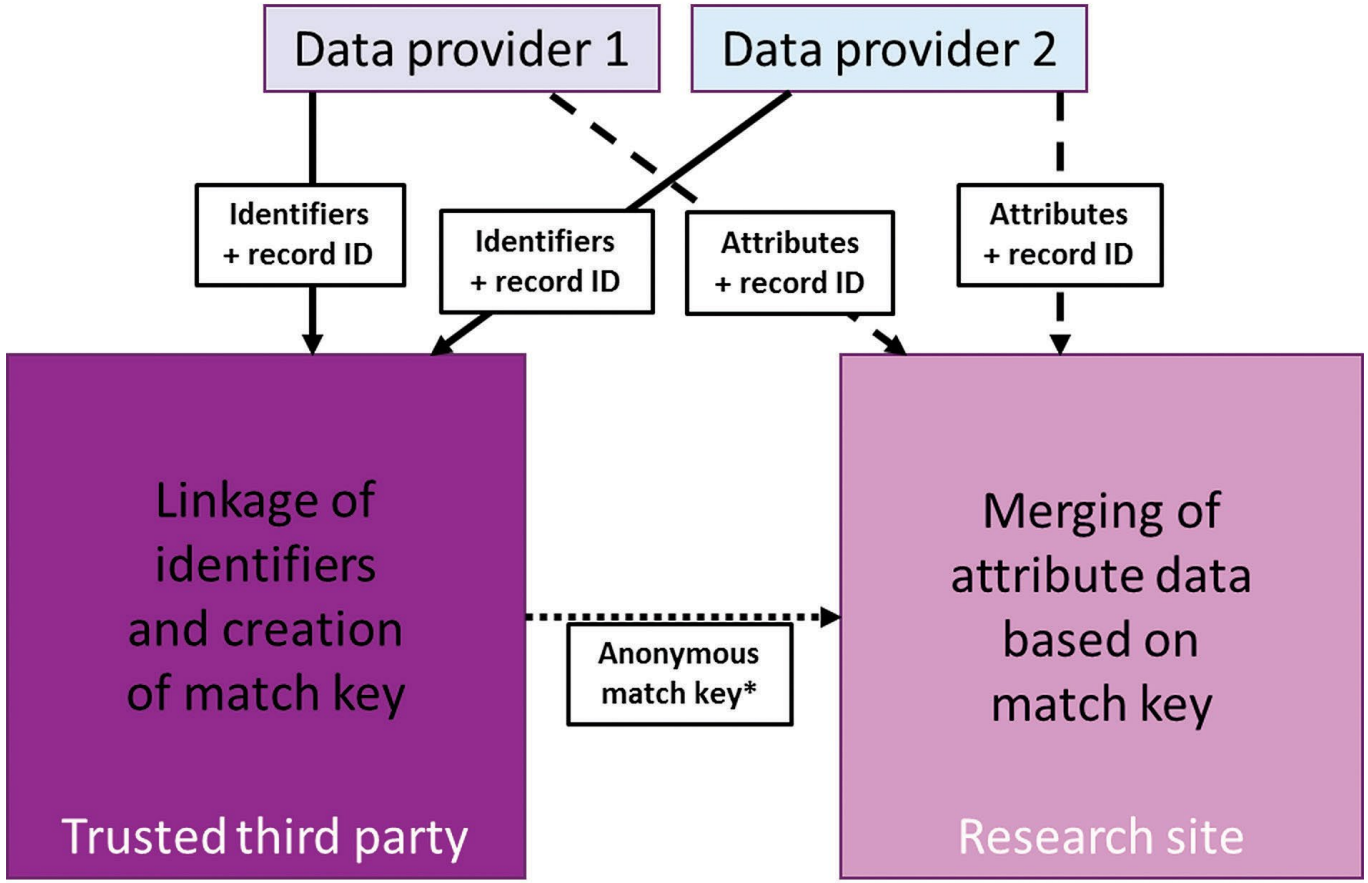
- Probabilistic Linkage (no unique identifier)
  - … well established Fellegi-Sunter linkage framework
  - … records matched on the basis of scores engineered to maximise the probability of a correct match
  - … probabilities defined in terms of "distance" between values of identifier variables common to both records
  - … link "declared" only when score is above a (subjective) threshold
  - … software widely available

- Separation Principle: Identifying information should not be included in the linked data (attribute data)

- Implementation: Trusted Third Party Linkage (many flavours)
  - … linker (TTP) given identifiers + pseudo-IDs from both contributing sources
  - … anonymous "match key" (i.e. a mapping) created, allowing pseudo-IDs from first source to be linked to pseudo-IDs from second source
  - … analyst uses match key and pseudo-IDs (but no identifiers) to link attribute data from both sources
  - … good for protecting confidentiality
  - … bad for understanding the quality of the linkage and assessing impact of linkage errors

Linkage implementation based on separation of identifiers and attributes

*Anonymous match key provides link between record IDs



Source: Harron (2016)

GUILD: GUidance for Information about Linking Data sets (Gilbert et al. 2017, Journal of Public Health)

- For the linker ...

  … linking methodology should be shared with analysts
  … linker should describe and justify the identifiers used in linkage
  … linker using score-based methods should report on the threshold for designating links as matches, and grouping of records that could potentially link (blocking)
  … linker should share record-level information (e.g. linkage scores) that enables the analyst to take linkage uncertainty into account
  … linker should publish aggregate-level linkage accuracy
  … linker should provide generic information reflecting quality of linkage
  … linkers should publish their methods for disclosure control of linked data

- For the analyst ...

  … should report evaluation of linkage accuracy and how this information is used in the analysis

  … should report on record-level indicators of linkage uncertainty (e.g. linkage scores) if possible

  … otherwise comparisons of the linked data with the unlinked source populations or through external comparisons with expected rates should be provided

Where Is Longitudinal Data Linkage Going?

- Focus on current UK situation
  - … ESRC is main "owner" of longitudinal social survey data resources in the UK
  - … in 2017, ESRC launched a review of its longitudinal studies, with the goal of identifying priority areas for future focus
  - … Recommendations from this review will inform ESRC future strategy, funding, management and commissioning decisions

- Preliminary consultation exercise (Townsend, 2016) identified data linkage as the most frequently mentioned methodological and technological issue for both UK and international respondents
  - … only methodological and technological issue mentioned by respondents from across all five primary research areas
  - … first on the ‘top ten’ list for respondents from all areas except the natural environment

ESRC Research Programmes With a Data Linkage Focus

ADRN (Administrative Data Research Network) Project

https://adrn.ac.uk/

- From the website:

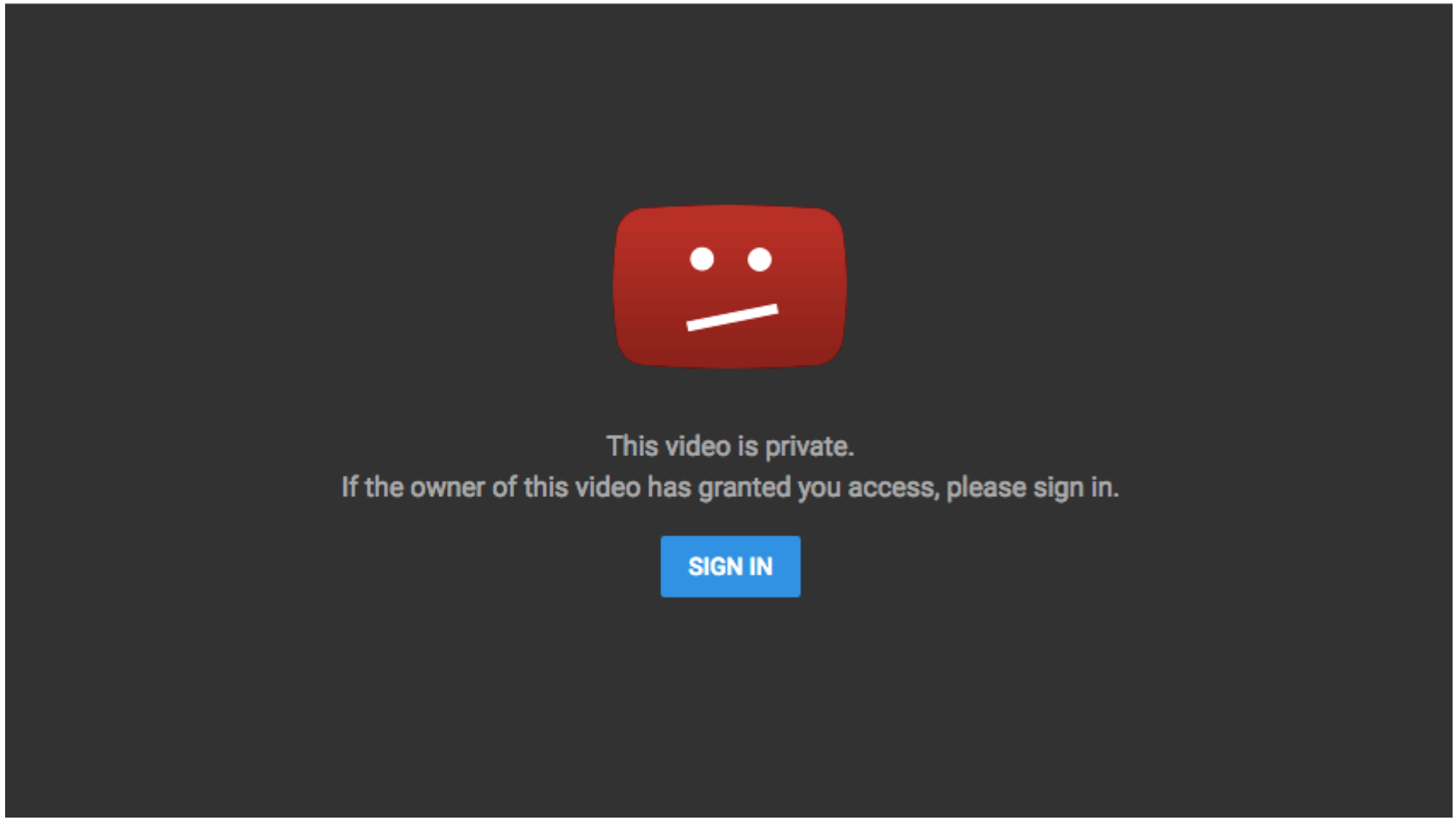
"The Administrative Data Research Network gives trained social and economic researchers access to linked, de-identified administrative data in a secure environment"

How the ADRN process works

- Application
- Ethical review, privacy assessment, feasibility, public benefit & scientific merit
- Approvals Panel of independent experts
- Accreditation and training
- Negotiation for permission to access & link
- Linkage via Trusted Third Party
- Analysis in a secure environment
- Results of analyses checked before released
- Plain English summaries of research published

Unfortunately, there is only so far that one can drill down ...

More about how data collections are de-identified and linked



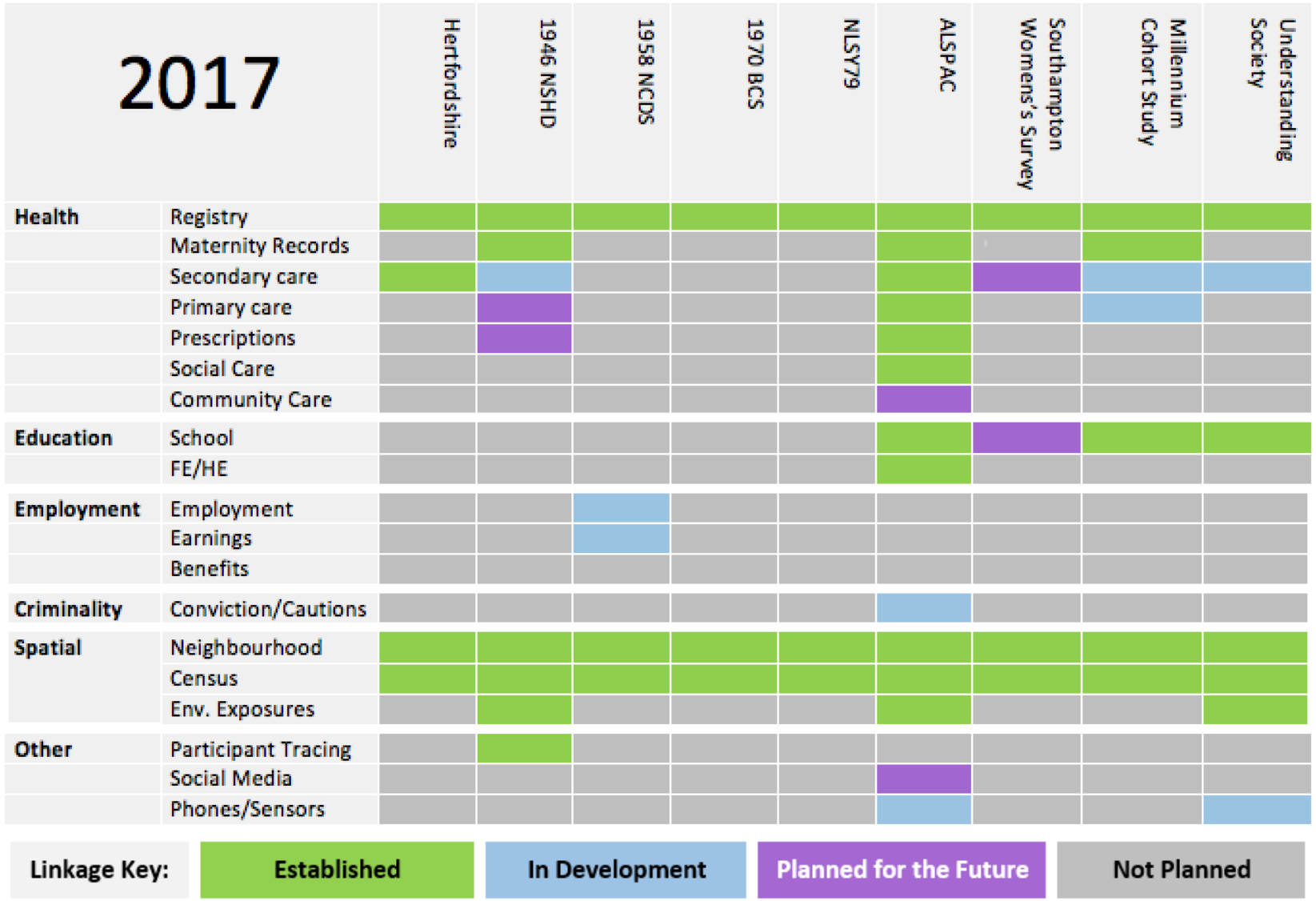
CLOSER (Cohort and Longitudinal Studies Enhancement Resources) Project

- CLOSER's mission is to maximise the use, value and impact of the UK's longitudinal studies
  - … stimulate interdisciplinary research across the major longitudinal studies
  - … provide shared resources for research
  - … assist with training and development for researchers in the use of longitudinal data at all career stages
  - … share information and expertise in longitudinal methodology'

- Facilitating researcher use of linked longitudinal resources is a crucial aspect of what CLOSER is about

http://www.closer.ac.uk/about/areas-work/data-linkage/

CLOSER linkage activity



Source: Boyd (2017)

CLOSER Experience So Far (Boyd, 2017)

- Barriers
  - … risk aversion amongst data owners, due to fear of professional/ public reputational damage
  - … lack of data owner resources for linkage activity
  - … very different data owner infrastructures that can support linkage
- Challenges - Straightforward
  - … gaining and maintaining linkage consents
  - … developing linkage strategies for participants lost to attrition
  - … demonstrating increased scientific return from linkage
- Challenges - Difficult
  - … governance requirements for linkage, including use of confidentialised data
  - … risk that research potential from linkage gets overlooked in big data/admin data world
  - … justifying the opportunity costs for investing in linkage strategies

Take Home Message(s)

- Data linkage for longitudinal surveys is a hugely active and rapidly growing area, but linking to administrative data is still a big challenge
- Creation of linked data sets is currently running far ahead of our capacity to analyse these data is a statistically rigorous fashion
- There are pressures from funding agencies for longitudinal surveys to adopt a more controlled (less "feral" and hopefully more effective) framework for linking to population registers
- It is unlikely that linking administrative data sources on their own will be able to provide the breath and depth of data currently collected in longitudinal surveys

- At the same time, it is equally unlikely that funders will persist with the self contained birth cohort / panel study models used to date

- A future where the focus of longitudinal surveys is in providing a core set of longitudinally measured characteristics that can then be combined in analysis with data from multiple external sources (including population registers) seems most likely