Confidentiality in practice
A user perspective

David Lawrence

Senior Statistician

Centre for Developmental Health

Overview

- What does "confidentialising" a data set mean?
- How linked data sets are confidentialised in WA
- What is the impact of confidentialisation on research?
- An example – the ABS National Survey of Mental Health and Wellbeing

Terminology

Data integration – the process of combining information from two or more data sources based on information common to the data sets

Data linkage – the process of creating links between data from different sources based on common features present in those sources

Confidentialise – to ensure a data set does not provide sufficient information as to make it likely that an individual within the data set can be identified, directly or indirectly

Statistical linking – involves linking records that do not necessarily belong to the same individual, to create a file which accurately represents the population characteristics of both data sets

Integrating authority – single agency ultimately accountable for the Statistical Data Integration project. This agency may work with a network of agencies to achieve the data integration, for example, it might use another agency to undertake linkage or to support dissemination

De-identification and confidentialisation

A data set is said to be **de-identified** when identifiers have been removed

A data set is said to be **confidentialised** when identifiers have been removed and it has undergone risk assessment and management to determine that it is unlikely to enable identification

Other definitions of the terms

Re-identifiable or pseudonymised data

A data set where identifiers have been removed

Confidentialised data

"Data that have been degraded by being subjected to a process to 'blur' the information itself, to reduce the potential to identify individuals. This process could involve, for example, changing all dates of birth to the first day of the month."

— Kelman CW, Bass AJ, Holman CDJ (2002) Research use of linked health data — a best practice protocol. *Australian and New Zealand Journal of Public Health*. 26: 251-5.

Statistical disclosure control

Aim

- To help protect the confidentiality of individual subjects' data in files that are shared, while simultaneously preserving the analytic value of data for secondary users

—Sieber JE (2006) Data sharing and disclosure limitation techniques. *Journal of Empirical Research on Human Research Ethics*. 1: 47-50.

- Goal is to balance utility and risk, recognising that risk can never be completely eliminated

"When producing microdata files, one should always keep the user perspective in mind. It is fundamental that the released file meet the researcher's requirements. Both information content and the choice of protection methods have to focus as much as possible on the user's needs. Knowledge of the statistical analysis the users generally want to perform helps deciding the anonymization strategy."

— International Household Survey Network. *Anonymization principles*. 2006.

Considerations in statistical disclosure control

- Work has a manual element, normally overseen by a Disclosure Review Committee of some type
- Risk assessment phase:
  - Existence of other similar data sets that are publicly available
  - Does the data set address rare conditions or small and/or highly visible groups?
- Confidentialisation phase
  - Choice of methods will depend on the analytical goals for the data set
  - Quality and sources of error in the data set

— Zarate AO, Zayatz LM (2006) Essentials of the disclosure review process. *Journal of Empirical Research on Human Research Ethics* 1: 51-62.

Approaches to minimising exposure risk

- Restricting the content of a file
  - Suppressing cells, variables, or records
  - Grouping or collapsing data items
  - Introducing small amounts of random error
- Restricting access to the file
  - Limit distribution
  - Control sites where it can be accessed
  - Agree to conditions regarding access and use

Software for risk assessment

- μ-Argus – Statistics Netherlands
- DIS-SUDA – University of Manchester
- International Household Survey Network – Microdata Management Toolkit
- UNESCAP provide Stata/SPSS plug-ins for developing countries
- PARAT

Procedures for disclosure control

- Rounding
- Categorising or re-coding
- Top- or bottom-coding
- Random perturbation
- Data swapping
- Post-randomisation
- Micro-aggregation
- Cell suppression
- Variable suppression
- Record suppression

— Doyle P, Lane JW, Theeuwes JJM, Zayatz LM (2001) *Confidentiality, disclosure and data access. Theory and practical applications for statistical agencies*. North-Holland.

Practical issues that impact confidentiality processes

The research funding cycle

- Research funders such as NHMRC and ARC generally take applications in February each year, put these to peer review and panel consideration before making funding decisions in November, for projects to commence in January
- Funding applications are more favourably considered if some preliminary investigation has been undertaken
  - Assessment of number of cases and statistical power
  - Quality of the coding of variables of interest
- This may require some preliminary tables to be run from the data sets which is not feasible under current linkage practice

Practical issues that impact confidentiality processes

Conflicting goals

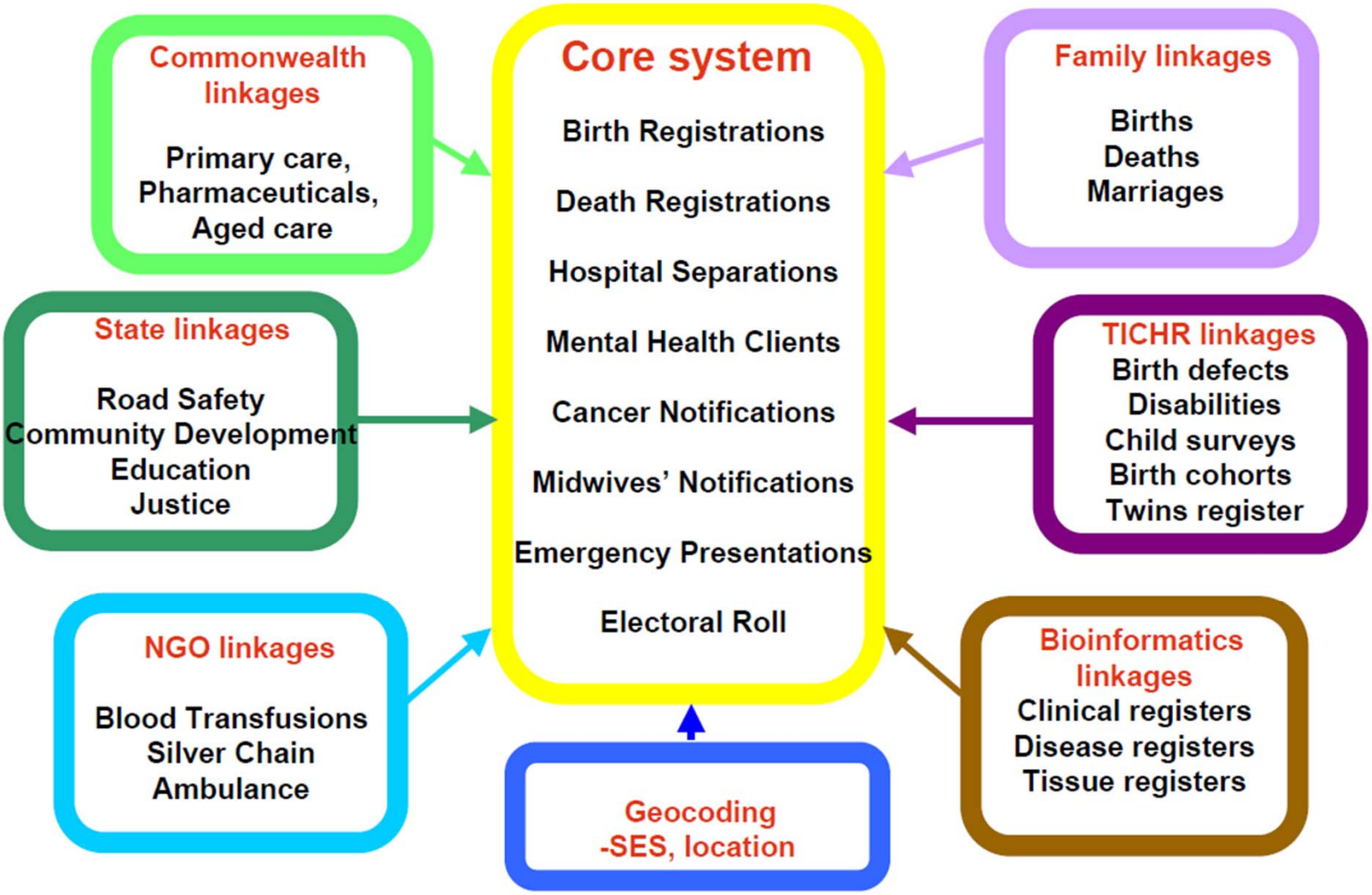
- Custodians want disclosure control practices that protect the confidentiality of their data sets and business models
- Researchers are concerned about the impact on their ability to conduct their research

- Do the various parties understand what impact, if any, confidentiality processes have on research and analysis?
- Are there enough examples of the impact of confidentialisation on analyses to understand in what conditions it may have an impact?

Confidentiality procedures in WA

- Linkage requests are assessed by two groups – an ethics committee who ensure compliance with national principles, and a committee of data custodians who advise on disclosure risk
- Restrictions are placed on the size of populations that can be studied. Sub-populations or random samples should be studied when possible
- There are restrictions on the number of variables that can be obtained from each file – each variable's relevance to the research questions must be justified
- Variables may be suppressed, recoded or collapsed to reduce identifiability
- Restrictions are placed on file management
  - Each linked file is project specific and cannot be combined across projects
  - Access restricted to specific personnel
  - Access restricted to approved environments
  - Researchers sign confidentiality agreement with DLU

WA Data Linkage System



Examples of linked files that are routinely confidentialised

- US National Health and Nutrition Examination Survey (NHANES) linked to:
  - Mortality
  - Medicare/Medicaid
  - Social security
  - Air quality
- US SEER-Medicare cancer data
- US Veterans affairs data
- Canada National Longitudinal Study of Children and Youth
- UK Neighbourhood Statistics Initiative
- UK Linkage of road traffic, hospital admission and police data
- Italy: Linkage of European Statistics on Income and Living Conditions with Person Tax annual Register
- ESSnet Integration of Surveys and Administrative Data project

National Survey of Mental Health and Well-being

Survey design

- Part of the World Mental Health Initiative
- Funded by Department of Health and Ageing
- Aims: to measure prevalence of mental disorders, burden of mental disorders, and use of services
- Uses the Composite International Diagnostic Interview (CIDI) — a detailed structured interview to assess presence, severity, onset and duration of mental illnesses
- Produces both ICD-10 and DSM-IV diagnoses
- Administered by trained interviewers in the home using computer-assisted interviewing
- Sample size: 8,841
- Variables on master file : 3,405

ABS procedures for confidentialising files

- Removal of identifying data items, such as name or address
- Reducing the level of detail shown for some items (eg geography, industry or occupation)
- Ranging, collapsing, or top- or bottom-coding some data items (eg age first smoked)
- changing characteristics of some records, eg introducing small random perturbations into continuous variables (eg income)

SMHWB — What was confidentialised?

Module	Items recoded	Items suppressed
Household demographics	7	5
Personal demographics	7	10
Chronic conditions	0	3
Kessler 10	0	0
Functioning	0	0
MMSE	0	5
Screener	0	0
Suicidality	6	1
Generalised anxiety	2	0
Obsessive-compulsive	2	0
Post-traumatic stress	5	0
Depression	4	1

SMHWB — What was confidentialised?

Module	Items recoded	Items suppressed
Mania	2	0
Bipolar disorder	0	0
Substance use	13	0
Social phobia	2	1
Panic	2	0
Agoraphobia	2	0
Psychosis	0	0
Medications	0	0
Social networks	0	0
Caregiving	0	0
Service use	43	8
Total	99	34

Impact on the analysis

Nil

US National Comorbidity Survey–Replication

Survey design

- Part of the World Mental Health Initiative
- Aims: to measure prevalence of mental disorders, burden of mental disorders, and use of services
- Uses the Composite International Diagnostic Interview (CIDI) — a detailed structured interview to assess presence, severity, onset and duration of mental illnesses
- Produces both DSM-IV and ICD-10 diagnoses
- Administered by trained interviewers in the home using computer-assisted interviewing
- Sample size: 9,282
- Variables on master file : 4,223

NCS-R — disclosure control process

- Survey run by Department of Health Care Policy at Harvard Medical School
- Field work conducted by Survey Research Centre, University of Michigan
- Unit record file warehoused by the Substance Abuse and Mental Health Data Archive at the Inter-University Consortium for Political and Social Research
- Data reviewed by the SAMHDA Disclosure Review Committee

NCS-R — what was confidentialised?

- CURF released more than 5 years after survey field work
- Variables removed from file: 509
- Variables recoded: 281
- Total variables on master file: 4,223

— O'Rourke *et al* (2006) Solving problems of disclosure risk while retaining key analytic uses of publicly released microdata. *Journal of Empirical Research on Human Research Ethics*. 1: 63-84.

Identifiability and usability trade-off

"Agencies and users should work together to promote legislative, regulatory, and dissemination policies and practices that facilitate timely and cost-effective access to data for statistical research and policy analysis but do not permit full and open access by all of the public for any use. If confidentiality issues are not fully addressed in constructive and proactive ways, users face the very real risk of losing access to high quality data."

— Doyle P, Lane JW, Theeuwes JJM, Zayatz LM (2001) *Confidentiality, disclosure and data access. Theory and practical applications for statistical agencies*. North-Holland.


TELETHON INSTITUTE FOR
Child Health
Research