Balancing access and confidentiality – perspectives from the Manchester conference

Alan Lee
Department of Statistics
University of Auckland

- The tension between researchers (who want microdata access) and statistical agencies (who want to preserve confidentiality) is well known
- The conference “Census Microdata: Findings and Futures” in Manchester 1-3 Sept 2008 explored some of this tension and discussed ways to resolve it.
- In this talk I report on some of these issues and the steps agencies are taking to improve access.

The researcher's point of view

- From Denise Lievesley's talk….

According to the International Household Survey Network established as one of the six action points of the Marrakech Action Plan for Statistics, national and international micro databases should be established to:

Databases should…..

- promote the acquisition, documentation, dissemination and preservation of microdata essential for the production of national statistics, for research and for instruction in the social sciences,
- promote the effective use of existing survey and census data,
- ensure the continued viability and usability of microdata now and in the future, and,
- promote equitable access to these data within the framework of the national statistical legislation.

The official statistics view

- Source of the next few slides:
- Denis Trewin, *Managing statistical confidentiality and microdata access, principles and guidelines of good practice*
- UN Economic Commission for Europe, Conference of European Statisticians, 2007

Access as a civil right

"Open access to official statistics provides the citizen with more than a picture of society. It offers a window on the work and performance of government itself, showing the scale of government activity in every area of public policy and allowing the impact of public policy to be assessed"

UK 1993 white paper on Open Government in the United Kingdom

Access as a social good

The privacy paradox:

The rush to ensure complete levels of privacy in the research context paradoxically results in less social benefit, rather than more.

… people will recognise that while they surely have a right to privacy, they may also come to the realisation that they have a duty to share information, if the common good is to be furthered.

Peter Madsen, NSF Workshop on Confidentiality Research, 2003.

Other benefits

- Providing researchers with access to microdata can be a way of extracting additional value from the cost of collecting official statistics
- Takes up slack if NSO budgets shrink
- Access permits policy makers to pose and analyse complex questions, fit models
- Enables replication of important research
- Reduces reporting burden if researchers need not replicate existing data

(all from CES guidelines)

The other side of the coin

- Individual data collected by statistical agencies for statistical compilation… are to be strictly confidential and used exclusively for statistical purposes

  Sixth UN principle of official statistics

Interpreted to mean

1. It is appropriate for microdata collected for official purposes to be used to support research as long as confidentiality is protected
2. Microdata should only be made available for statistical purposes
3. Provision of microdata should be consistent with national legal arrangements
4. Procedures for access should be transparent and publically available

Arguments against access

- Must maintain trust of respondents by protecting confidentiality
- Cost of providing secure access
- But Guidelines agree these are outweighed by the benefits

How might the tension be managed

- Move from risk avoidance to risk management
- Current levels of microdata access are not controversial (compared to leaving CD's on a train)
- Have transparent procedures for release

Managing disclosure risk -options

- Open slather: Rely on sanctions, pass onus onto research community, appropriate retribution for confidentiality breaches, education, instill ethical behaviour
- Keep microdata in secure facility (data fortress), control interrogation
- Perturb (confidentialise) data and release
- Release tables (data cubes)

Carrots and sticks

- Instill good behaviour: ensure researchers understand why NSO's care about confidentiality
- Have clear protocols around release
- Make researchers aware of consequences of a breach - prison!!!
- Then provide microdata subject to an approval process (Scandinavian approach)

Data fortresses

- Either remote access (RAF) or data laboratories
- Microdata stays in NSO facility
- Users submit programs
- Output is returned after vetting
- Sounds familiar to the over 50s??
- Quick turnaround vital (automatic vetting)

Perturbed data sets

- Confidentialised microdata supplied on CD in various forms under various restrictions
  - Public use. Released under no conditions, small disclosure risk
  - Licensed files – released under conditions, higher disclosure risk

Tables (data cubes)

- Freely available, no disclosure risk (microdata for a small number of variables)
- Web or paper dissemination (Table Builder)
- Staple under the risk-avoidance regime
- Usual mode of access for general public

Country-by-country

New Zealand has three of these modes of access

- Table builder
- CURFS
- Data Lab

  E.g 2001 census CURF has about 76,000 records, about a 2% sample, comes on a CD

Data Labs - at SNZ offices, turnaround not instant

Table Builder – has prepopulated tables, up to 4-dimensional

Finland

- A researcher’s paradise
- Statistics Act governs access
- Licenses granted to approved researchers
- Some confidentialisation done according to circumstances, then microdata released
- Violation of rules carries prison sentence

USA

- **Data labs (9 Research Data Centers)**
  - Access granted only if project benefits the Bureau of the Census
  - Proposal Review process slow
- **Public-Use Microdata Samples (PUMS)**
  - These files contain records for a sample of housing units with information on the characteristics of each unit and each person in it. While preserving confidentiality (by removing identifiers), these microdata files permit users with special data needs to prepare virtually any tabulation.

USA (cont)

- Public use files (cont)
  - IPUMS – Integrated Public Use Microdata Series

    An amazing collection of international census data assembled by the Minnesota Population Center at the University of Minnesota

Sample	Sample Fraction (%)	Households	Persons	Weighted	Notes
Argentina 1970	2	129,728	466,892	-	
Argentina 1980	10	672,062	2,667,714	yes	
Argentina 1991	10	1,148,351	4,143,727	yes	Missing data for several key variables requires alternative weight variable
Argentina 2001	10	1,040,852	3,626,103	-	
Austria 1971	10	264,655	749,894	-	
Austria 1981	10	283,693	756,556	-	
Austria 1991	10	310,099	780,512	-	
Austria 2001	10	341,035	803,471	-	
Belarus 1999	10	385,508	990,706	-	
Brazil 1960	5	613,273	3,001,439	-	Excludes 11 states in the north

Australia (courtesy of Jenny Telford)

2006 Census Microdata – Modes of Access

- CD-Rom
  - 1% confidentialised sample file
- Remote Access Data Laboratory
  - 5% confidentialised sample file
- ABS Onsite Data Laboratory
  - customised sample file
- Table Builder
  - tabular access to unit record file.

1% Sample via CD-Rom

- Available since the 1981 Census
- No technical restrictions on use
- Least amount of detail
- Subject to undertakings and review
- Most data items available (classifications collapsed e.g. Age)
- Minimum population size of 250,000 persons per geographic unit.

5% Sample via Remote Access Data Laboratory (RADL)

- 5% – biggest Australian sample ever
- More data items than 1% (e.g. Indigenous Status)
- More detailed (less collapsing e.g. Age)
- Minimum population size of 125,000 persons per geographic unit
- Different sample source from 1%
- Available via RADL only

RADL

- Internet based query system
- Keeps unit records within the ABS
- Allows for analysis in SAS, SPSS and STATA
- Layers of protection allowing for more detail to be available

ABS Onsite Data Laboratory

- Onsite supervised access to detailed microdata
- Users subject to conditions of use
- Cost recovered service
- All output is audited prior to release
- Only considered in cases where other modes are insufficient

2006 Census TableBuilder

- New application due for release 2009
- Full access to counts based on unit record file
- Allows for tabulations only
- No direct access to record level data
- Uses new perturbation algorithm to dynamically confidentialise data

Canada (courtesy of Gustave Goldman and Sri Kanagarajah, Statistics Canada)

- Research Data Centers (RDC) – data labs located in universities
- Public use data files

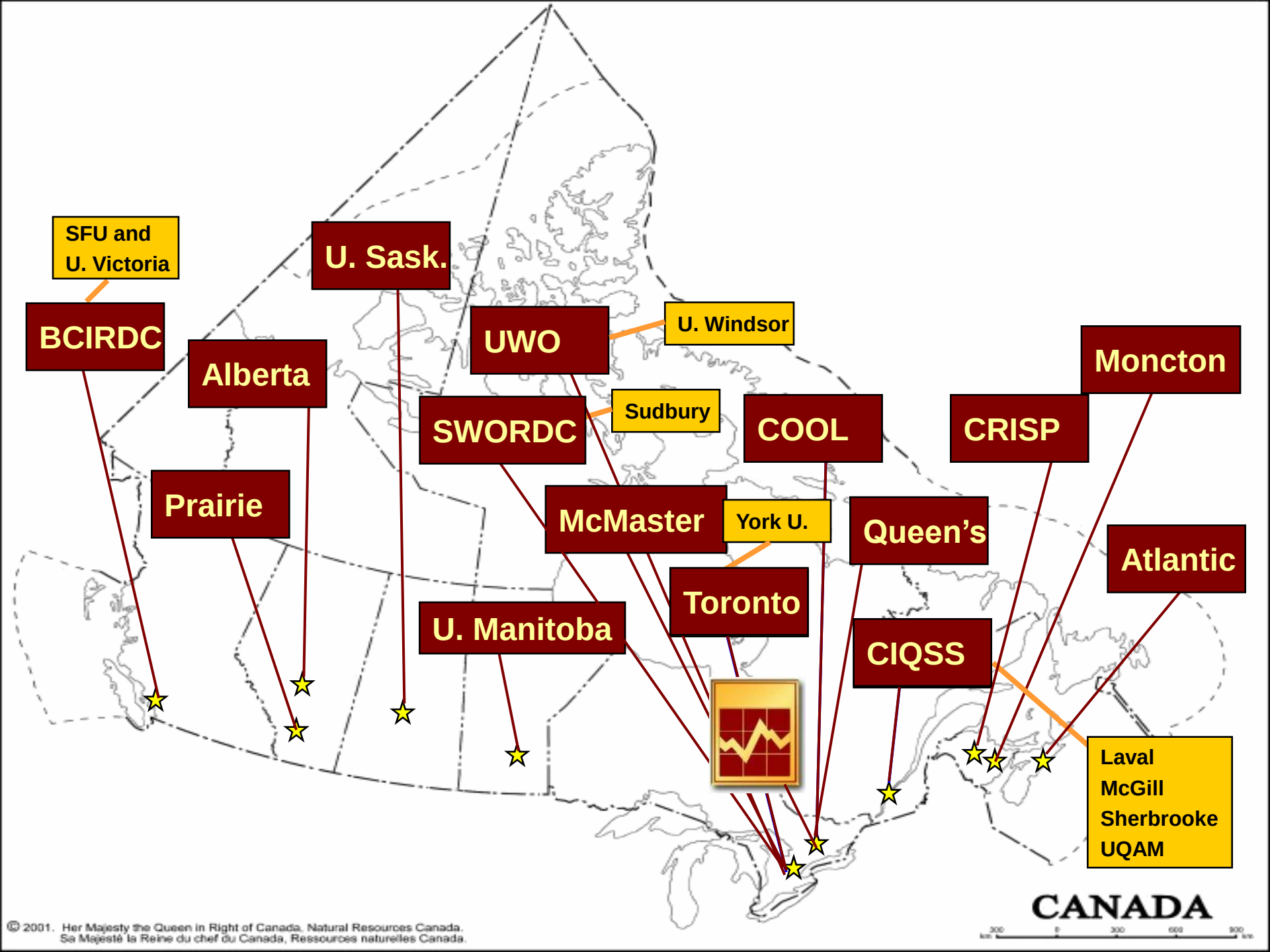
What is the RDC Network?

It is a partnership that includes:

- More than 40 Canadian universities
- Major Granting Councils
- Provincial governments
- Statistics Canada

What is a Research Data Centre?

- Secure environment in a setting that is removed from Statistics Canada premises
- Houses Statistics Canada microdata files
- Staffed by a Statistics Canada employee at all times
- Operates under the provisions of the Statistics Act
- Access limited to researchers with approved projects and "**sworn-in**" under Statistics Act as "deemed employees"
- All researchers have direct access to the data


SFU and
U. Victoria
BCIRDC
U. Sask.
Alberta
Prairie
U. Manitoba
UWO
U. Windsor
SWORDC
Sudbury
McMaster
York U.
Toronto
COOL
Queen's
CIQSS
CRISP
Moncton
Atlantic
Laval
McGill
Sherbrooke
UQAM
CANADA
© 2001. Her Majesty the Queen in Right of Canada, Natural Resources Canada.
Sa Majesté la Reine du chef du Canada, Ressources naturelles Canada.

Access to the Research Data Centres
(Academic researchers)

- Project proposal
- Proposal evaluation – SSHRC
- Security clearance – enhanced reliability check
- Orientation session and “oath of office”
- Researcher agrees to provide publicly available report that falls within Statistics Canada’s mandate

A sample of the data that are in the RDCs

Aboriginal Peoples Survey (APS)

Canadian Community Health Survey (CCHS)
Cycle 3.1
Cycle 2.2 - Nutrition
Cycle 2.1
Cycle 1.2 - Mental Health and Well-being
Cycle 1.1

Census of Population
2001 Census
1996 Census
1991 Census

Ethnic Diversity Survey (EDS)

General Social Survey (GSS)
- Access to and Use of Information Communication Technology
- Education, Work and Retirement
- Family
- Health
- Social Engagement
- Social Support and Aging
- Time Use
- Victimization

Longitudinal Survey of Immigrants to Canada (LSIC)

National Graduates Survey (NGS)

National Longitudinal Survey of Children and Youth (NLSCY)

National Population Health Survey (NPHS)
- Household Component - Cross-sectional
- Household Component - Longitudinal
- North Component
- Health Institutions Component

Participation and Activity Limitation Survey (PALS)

Survey of Labour and Income Dynamics (SLID)

Workplace and Employee Survey (WES)

Youth in Transition Survey (YITS)

Program for International Student Assessment (PISA)

Differences between public files and detailed master files in the RDCs

Public files	RDC master files
Level of geography = province or CMA	Census Subdivision, Census Tract or below
Aggregates certain countries of birth or ethnic origins	All the ethno-cultural details are available
Joint analysis at individual and family level is limited	Master files can be used with full individual level information and characteristics of families
Only cross-sectional data	Panel data tracking the same respondents over time (not for Census data)
Census public file is only a sample	The Census master file with over 6 million records is available

Canadian Public Use Files (PUMFS)

- Single individual file **(2.7% of population)**
- More geography detail like provinces, CMAs
- Unit of analysis is person
- Variables as close as possible to the questionnaire to allow users the freedom to create their own derived variables; only complexity level 4 variables need to be derived (e.g. LICO – Low Income Cut Off)

Canadian Public Use Files (cont)

Hierarchical (1% of population)

- All persons within same household and family are linked
- Allows analysts to choose their unit of analysis (Individual, household and family) (not possible with the Individual File)
- International comparison possible
- Provides analyst with maximum flexibility in doing regressions and modelling with the 3 universes
- Allows users to create their own derived variables; only complexity level 4 variables need to be derived (e.g. LICO – Low Income Cut Off)

UK

- SARS – Samples of anonymised records
  - Hosted by the centre for Census and Survey Research, University of Manchester

  1991 SARS 2% sample, 1.1m records

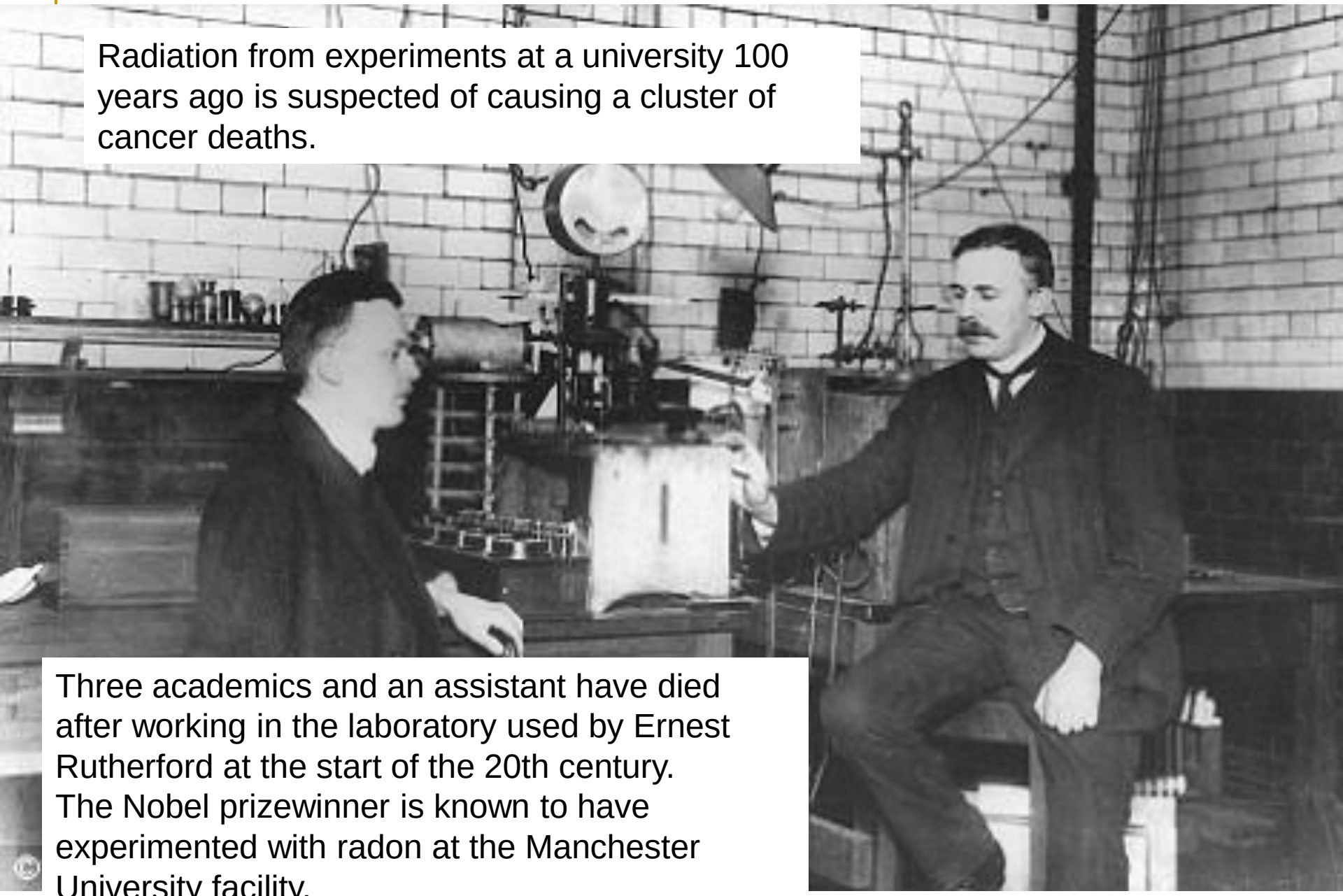
  2001 SARS 3% sample, 1.75m records

  2001 SAM (small area microdata) 5% sample, 3m records, all on CD
- CAMS: Controlled access microdata, more detailed, geography at local authority level
  - Accessed through ONS offices

UK (cont)

- ONS Longitudinal Study – links records for 1971, 1981, 1991, 2001 censuses, vital statistics registrations 1% sample for England and Wales
- Scottish Longitudinal Study – 5% sample for Scotland
- Analyses done in-house (ONS) or by remote access / in house (SLS)

Radiation from experiments at a university 100 years ago is suspected of causing a cluster of cancer deaths.



Three academics and an assistant have died after working in the laboratory used by Ernest Rutherford at the start of the 20th century. The Nobel prizewinner is known to have experimented with radon at the Manchester University facility.