Statistics Canada Statistique Canada

Health Record Linkage at Statistics Canada

www.statcan.gc.ca

Telling Canada's story in numbers

Nicole Aitken, Philippe Finès
Statistics Canada

Thursday, November 16th 2017

Canada

Why use linked data?

- Harnessing the full potential of data
- Innovations in linking data
- Improve the care and health of Canadians
- High analytical potential: allows researchers to fill data gaps

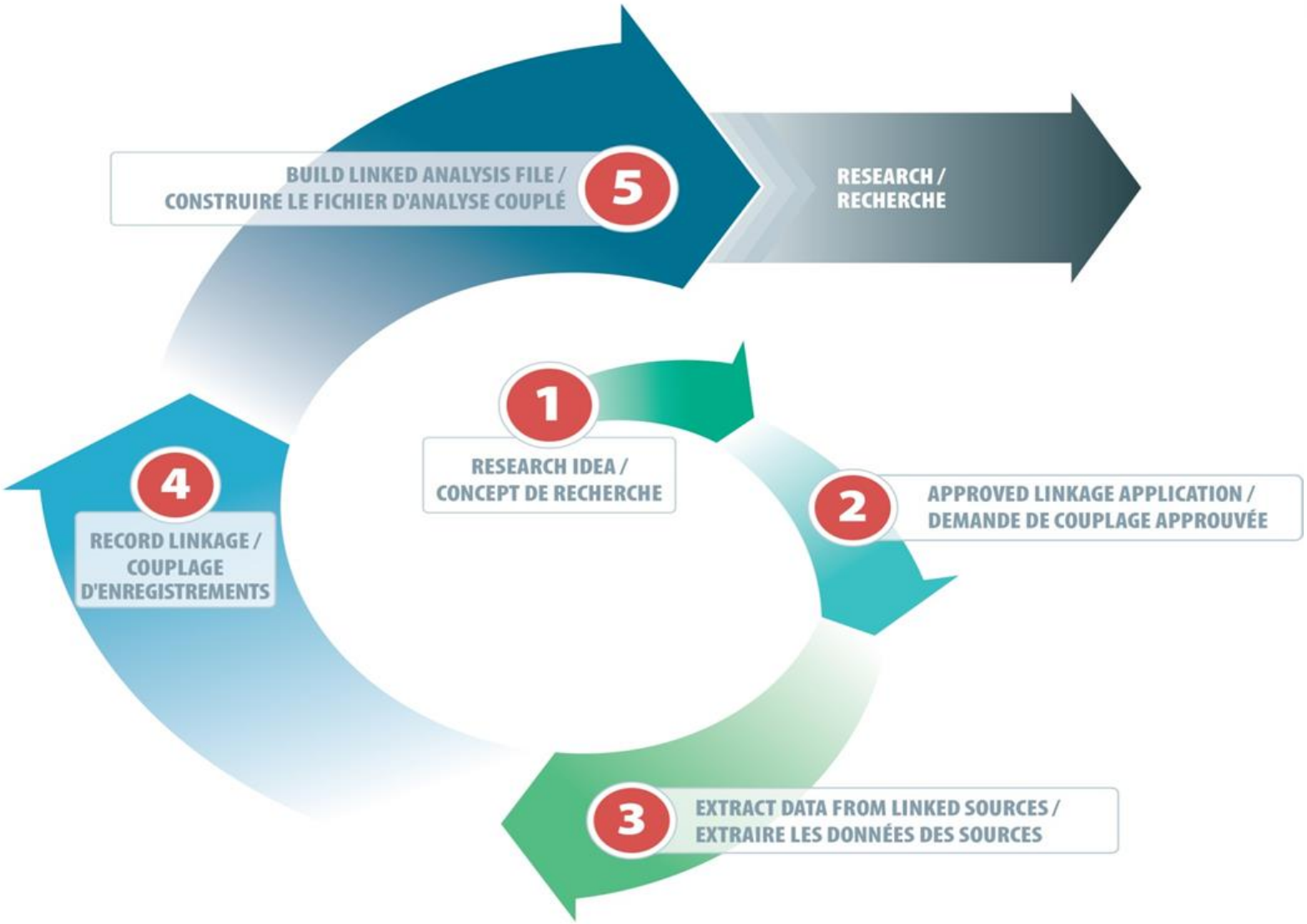
What is Record Linkage?

- A process whereby personal identifiers are used to identify the same people in different data-sources
  - Name, date of birth, health card number, postal code
  - Canadian Health Measures Survey to Canadian Cancer Registry

Record linkage at Statistics Canada

- Secure *virtual linkage environment* that stores only personal identifiers in a protected depository that is used to generate linkage keys across data sources.
- Keys are stored separately from data.
- Do NOT create large integrated data bases of survey information about individuals.
- Strong governance, adherence to policy and privacy requirements.
- Suite of services, tools and support for analysts and external researchers conducting record linkage activities within the social domain.

How does it work?



Linked Data Available to All in the RDCs

Process to access linked RDC data

- Secondary use of existing linked data-sources
  - Have a research question
  - Access the data in an Research Data Center (RDC) following standard RDC procedures
    - Submit a project proposal
    - Complete the application form

What linked health data are available in the RDC now?

- Census 2006 to Discharge Abstract Database (2006 to 2008)
- Canadian Community Health Survey (CCHS) Annual (2000 to 2011) and Focus (1.2, 2.2 and 4.2) to:
  - Canadian Vital Statistics Deaths (CVSD; 2000 to 2015)
  - Discharge Abstract Database (1999/2000 to 2012/13)
- 1991 and 2001 Canadian Census Health and Environment Cohorts (CanCHEC)
  - Weights will be available by the end of the calendar year
- Perinatal Outcomes (2006 Canadian Birth-Census Cohort)

What linked health data are coming to the RDC?

- DAD (2000/01-2014/15); NACRS (2000/01-2014/15) and; OMHRS (2005/06-2014/15) to CVSD (2000-2012)
- CVSD (2008-2014) to DAD (2004/05-2014/15); NACRS (2004/05-2014/15)
- Canadian Cancer Registry (1992 to 2014) to deaths (1992 to 2014)
- 1996 CanCHEC followed for mortality to 2013 (with weights)
- 2001 CanCHEC followed for cancer to 2013 (with weights)

Note:

DAD= Discharge Abstract Database

NACRS=National Ambulatory Care Reporting System

OMHRS=Ontario Mental Health Reporting System

CVSD= Canadian Vital Statistics Death Database

CanCHEC = Canadian Census Health and Environment Cohort

What linked health data are coming to the RDC?

- Canadian Cancer Registry (CCR; 1992 to 2014) to
  - DAD and NACRS
  - Tax (income data)
  - 2016 Census
  - Longitudinal Immigration Database (IMDB)
  - CVSD

- Canadian Community Health Survey (CCHS) Annual (2003-2014) and Focus (1.2, 5.2) to Longitudinal Immigration Database (1980-2013)

Occupational Cohorts:

- National Dose Registry to CVSD and CCR
- Newfoundland Fluorspar Miners' cohort to CVSD and CCR

For more information

- HSD Record Linkage Mailbox
  statcan.hsdrecordlinkage-dsscouplageenregistrements.statcan@canada.ca

- Evan Green
  Evan.Green@Canada.ca

… Transition to part II …

Summary

- Details on the databases
- 3 linked databases
  - The Census – DAD linked database
  - The CCHS – CMDB – DAD linked database
  - The Census – Tax – Mortality – Cancer linked database

Details on the databases

Canadian Community Health Survey (CCHS)

- Large, biennial, cross-sectional survey (~130,000); after 2007, annual survey (~65,000);
- Covers the household population aged 12+ representing ~98%
- Excludes members of the regular Forces, institutionalized, Indian Reserves, and some remote areas
- Regular collection since 2000/01
- Core Content: health status, Risk behaviours, chronic conditions, socio-economic indicators

- Focus content since 2002
  - Topics include mental health (Cycle 1.2), food intake (Cycle 2.2), aging (Cycle 4.2)
  - Sample size (~30,000)

Census

- Long form (20% representative sample of the Canadian household population)
  - Income – personal, household, source
  - Immigration – time of immigration, world region of birth, generational status
  - Ethnicity
  - Household composition - marital status, relationship of occupants, living arrangements
  - Housing – type, tenure, need of repair
  - Collective dwellings - rooming houses, hotels and shelters
  - Language - mother tongue, home language, knowledge of official language
  - Disability status
  - Rural-urban residence
  - Indigenous status
  - ….and on and on….

Discharge Abstract Database (DAD)

- Obtained from the Canadian Institute of Health Information (CIHI)
- DAD 2005/06 through 2008/09 used for pre-processing
- DAD 2006/07 through 2008/09 used for record linkage
- Census of discharges from acute care hospitals (~3 million records per yr) (excludes Quebec)
- Contains demographic, non-medical administrative and clinical information (diagnostics and interventions)
- Use of resources via the Resource Intensity Weights which used in combination with costs of hospital stays (per day) can be used to derive costs.
- Able to count events but also create patient histories by linking hospitalizations at the person-level using personal health numbers

Mortality and place of residence

- Canadian Vital Statistics Death Database (CVSD)
  - 2000 to 2009
  - Census of deaths in Canada
  - Underlying cause of death, date of death, age at death

- Tax file
  - 1990 to 2009
  - Tax filers
  - Annual place of residence (postal code on tax return)

Some words on Validation

- Two parts of validation:
  - Internal validation – quality of the linkage (error rates)
    - *Do the linked pairs represent good links?*
    - *Are there any missed links among the non-linked pairs?*
  - External validation – quality of the linked data (representativeness of analytical file)
    - *Do the outcomes in the linked data file represent the experiences of the population of interest?*

1) 2006 Canadian Census and Discharge Abstract Database Linkage

Context

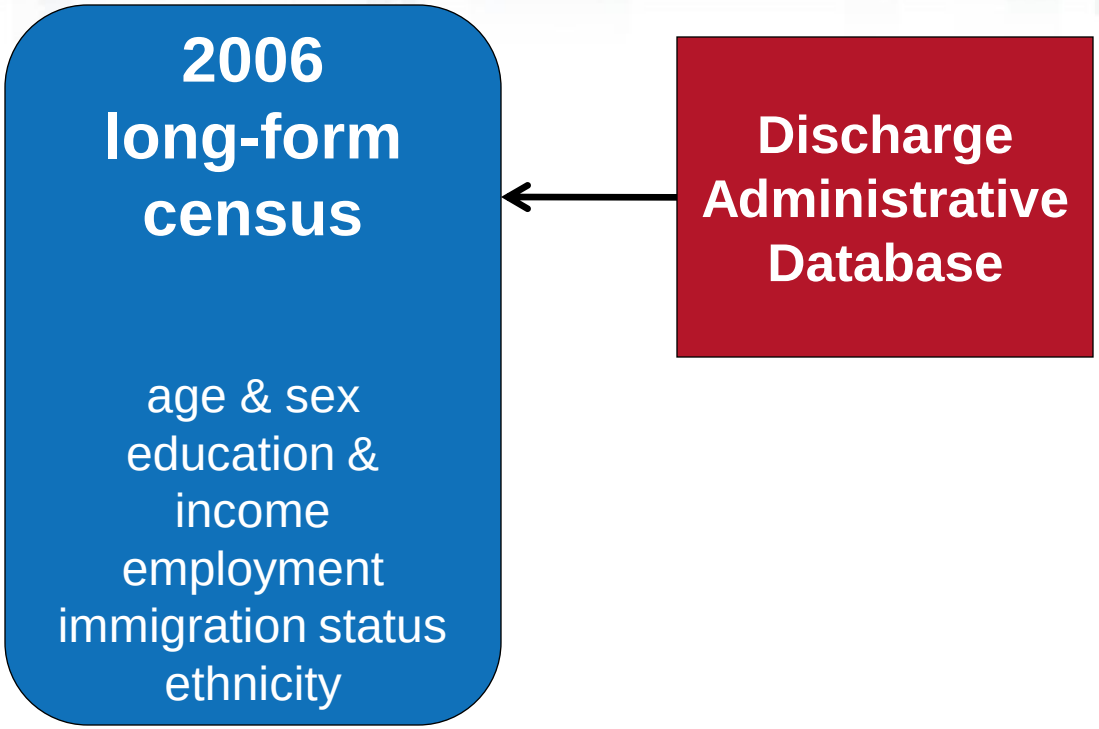
- To better understand the health outcomes and healthcare use of specific sub-populations –
  - Immigrants, Indigenous groups
  - Identify and quantify differences
  - Understand differences in the context of other social determinants of health

Research areas

- **Immigrant research**
  - Comparative analysis of hospitalizations by immigrant status, source country and time since immigration;
  - Use of hospital services among immigrant seniors;
  - Multi-generational analysis of cardiovascular related hospitalizations – is the health advantage lost among second generation?

- **Aboriginal research**
  - Comparative analysis of hospitalization rates among Indigenous groups, on and off reserve
  - Impact of housing condition on respiratory related hospitalizations among First Nations on reserve

2006 Census Cohort: DAD follow-up



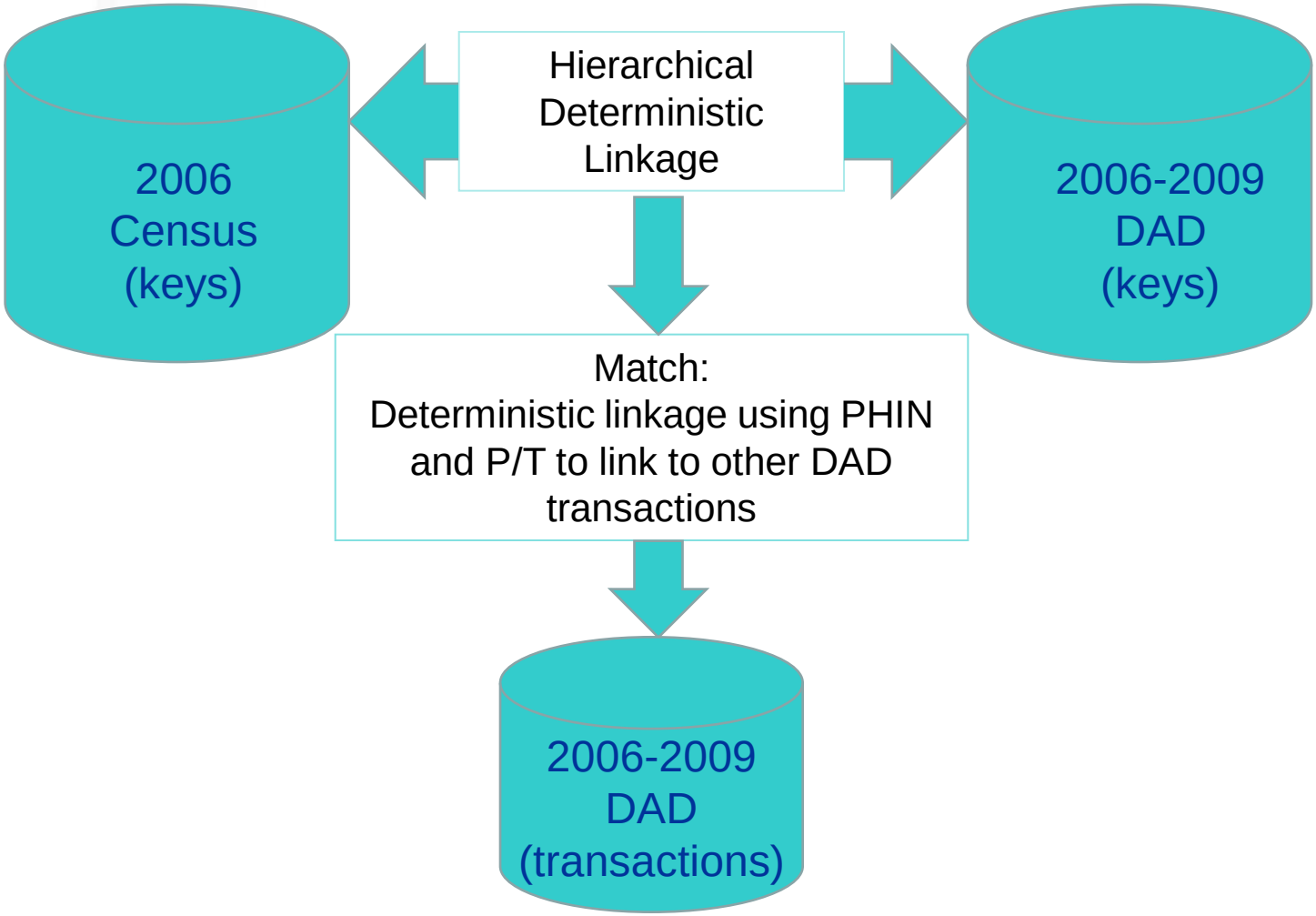
Step 1: Data Preparation

- *Eligibility of records for linkage*:
  - Complete (non-missing) date of birth in both Census and DAD;
  - Statistical linkage key must be unique in Census - no duplicates (e.g. multiple births removed)
  - Statistical linkage key associated with only one Health Insurance Number (HIN) in DAD

- *Hierarchical Deterministic Linkage*
  - **Unique statistical linkage key** – *date of birth, sex, postal code*
    - Used postal code information from HSTF as alternative to capture change in address overtime

  - Series of exact matches -conservative approach but appropriate given lack of **unique** identifying information

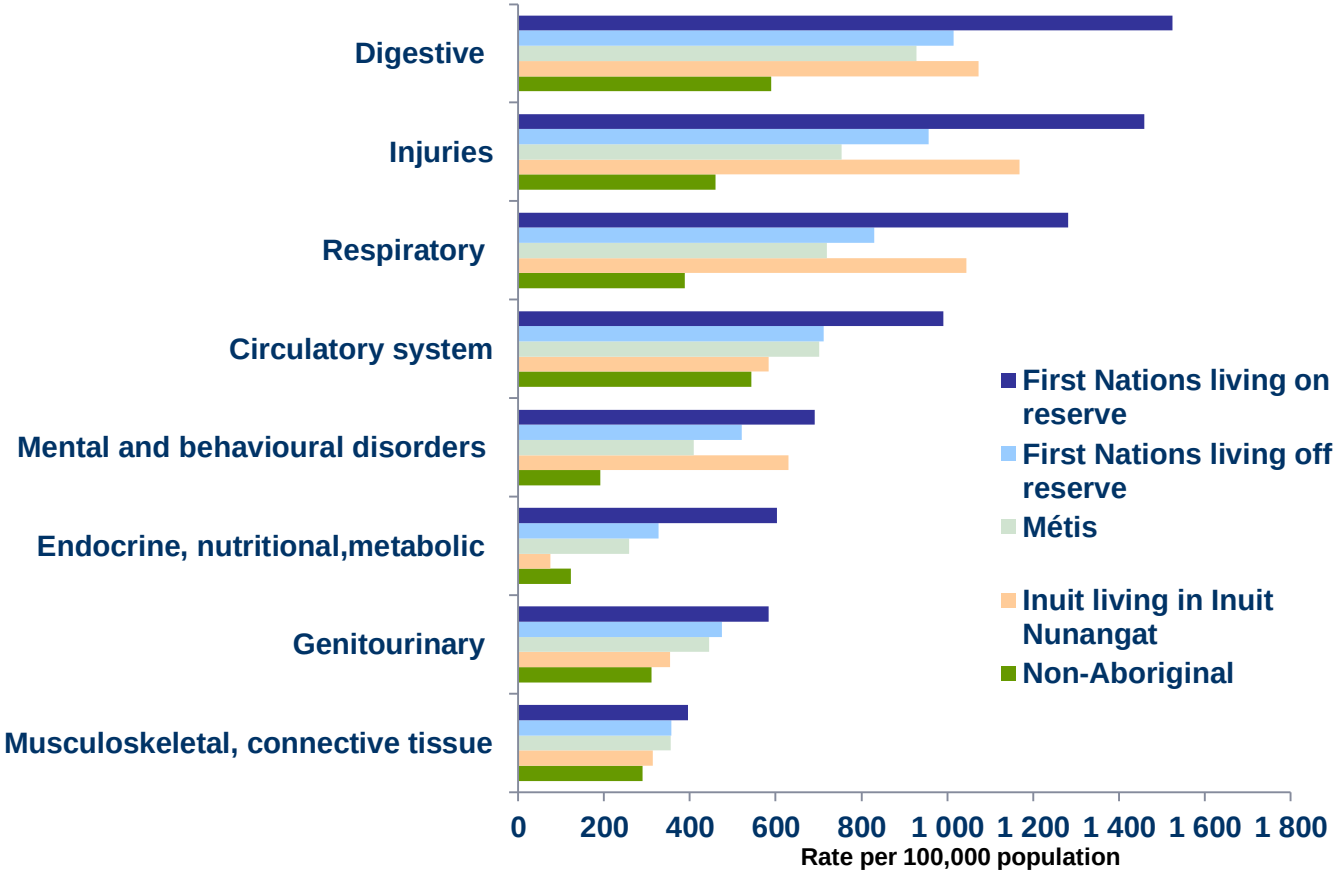
Steps 2: Record Linkage



Research Results

(Carriere G, Bougie E et al. Health Reports, August 2016)

Age-standardized acute-care hospitalization rates (ASHR) per 100,000 non-institutionalized population, by Aboriginal identity and by diagnostic chapter, Canada (excluding Quebec), combined fiscal 2006/2007 through 2008/2009



Source: Census of Population 2006, Census-linked Discharge Abstract Database 2006/2007, 2007/2008, 2008/2009 pooled.

2) Canadian Community Health Survey (CCHS) linked to Canadian Vital Statistics Death Database (CVSD) and Discharge Abstract Database (DAD)

Background

- Enhance the capacity of health data to address complex questions with "value added" information - fill data gaps
  - Survey data – lots of socio-economic, risk factor information but no outcomes;
  - Administrative data – outcome information (hospitalization, mortality) but limited individual information
- Linked data allow for "population health" lens to the study of health care services and outcomes
  - Used to study a wider range of determinants of health care use and outcomes of care
- Population based studies on a representative sample of Canadians
  - Large sample sizes - study specific populations and "rare" events
- Opportunity for comparisons across provinces and territories

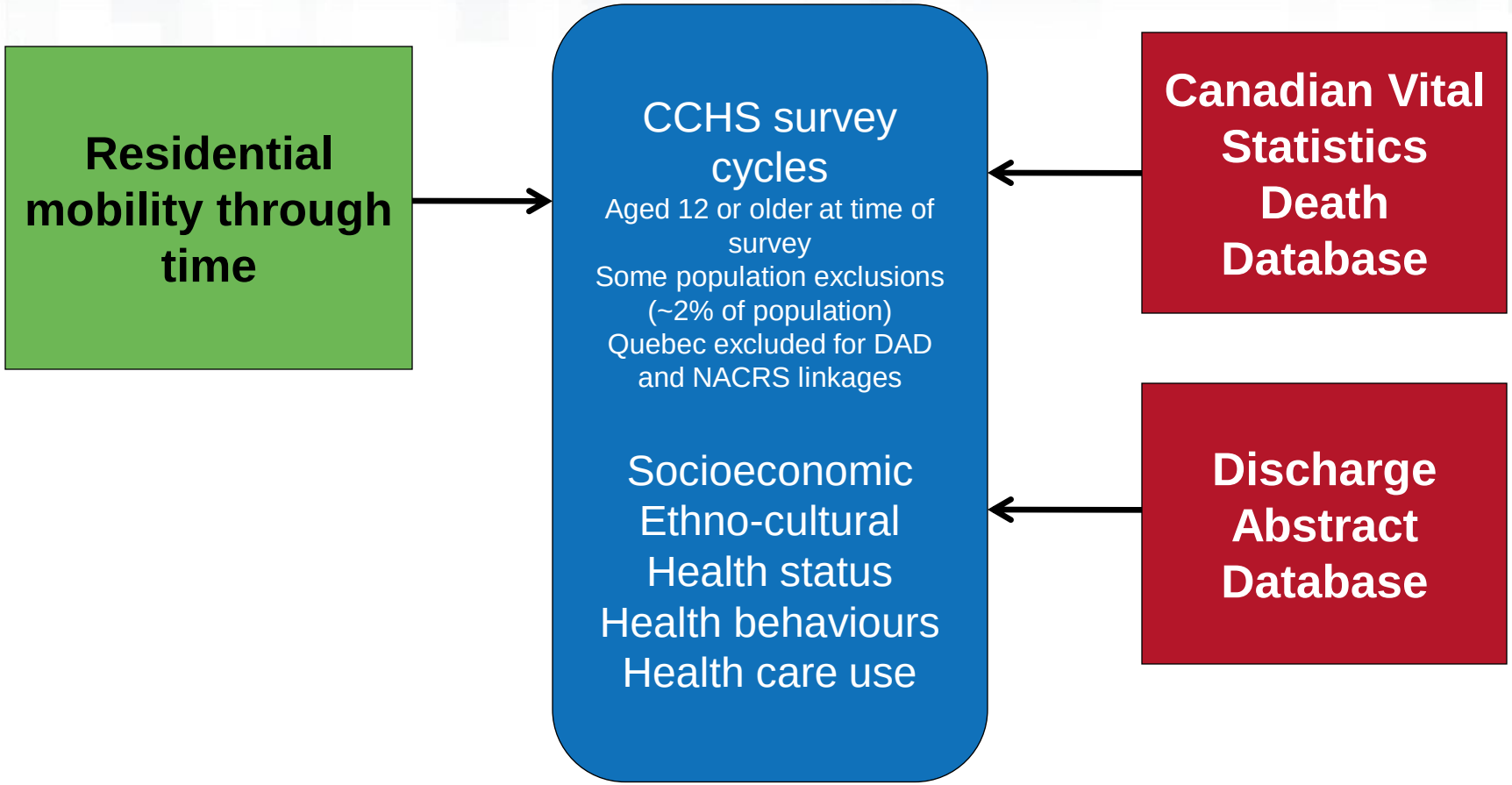
Research examples

1. To understand the interaction between socio-economic and behavioural risk factors and their effect on the use and cost of hospital services

2. To understand the extent to which differences in the prevalence of risk factors in Canada explains the variation in the use of hospital services

3. To examine the interaction between risk factors, ambient air pollution exposures, mortality, and the use of hospital services

Canadian Community Health Survey Cohorts



Main strengths & limitations

- Strengths
  - Population based
  - Rich source of information on the cohort characteristics and outcomes
  - Large sample size
  - Able to examine several variables simultaneously
  - Multilevel analysis

- Limitations
  - Information collected at one point in time (changes in risk factors are not captured)
  - Some population exclusions (reserves, children)

3) 1991 Canadian Census Health and Environment Cohort – aka CanCHEC

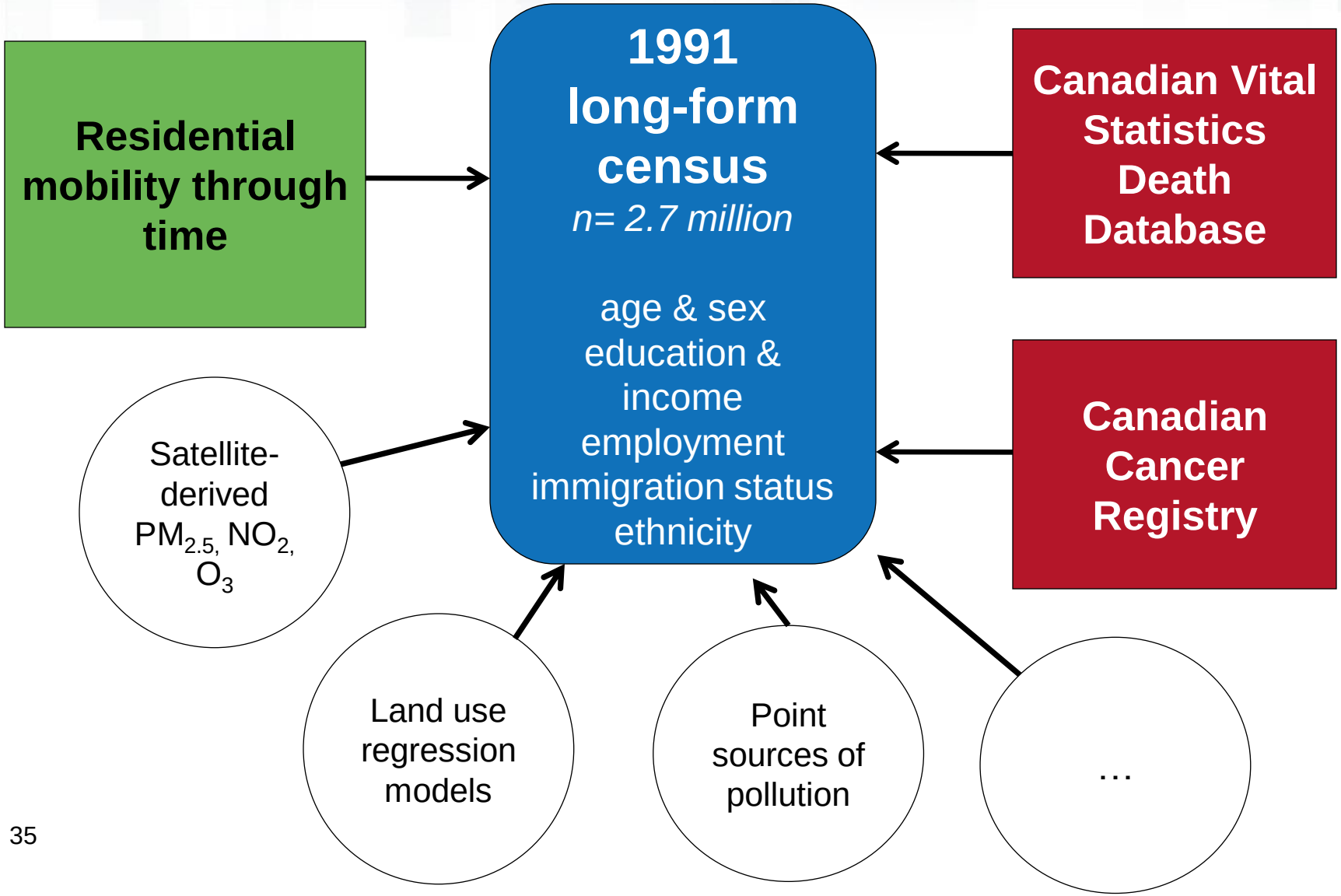
Context

- Greater focus on understanding potential inequalities in health outcomes
- Vital statistics, registries and health administrative data lack individual identifiers (ethnicity, Indigenous identity) or characteristic
- Identification of differences in mortality across socio-economic characteristics for a number of populations
  - Immigrants, ethnic origins, First Nations, Métis, and Inuit
- Produce baseline indicators of mortality for monitoring health disparities
  - Life expectancy & mortality by detailed population groups (occupation, education, income groups)

Research areas

- Sub-population analysis
  - First Nations, Métis, immigrants (year of immigration), place of birth, ethnic origin etc
- Analysis by socioeconomic status
  - Income (source, household, individual), education (years, qualifications), occupation, industry, type of housing, marital status
- Multi-dimensional analysis
- Exposure analysis
  - Assign exposure via postal code representative points

1991 Census Cohort: mortality & cancer follow-up



1991 census cohort

- Eligibility
  - Enumerated on 1991 census long form (1 in 5 households *)
  - Aged 25 or older as of June 4, 1991
  - Not a usual resident of an institution
  - N=3,576,487

  - Note that 3.4% of the Canadian population of all ages were not enumerated by the census

- Linkage approval for 15% of persons aged 25+

* Note that all residents of Indian Reserves and remote northern communities receive long form questionnaire

1991 census cohort

- Cohort creation
  - Eligible census respondents linked to tax filer data (non-financial) in order to get names
  - Matching variables: sex, date of birth, postal code, spousal date of birth
  - Results: 80% linkage rate, 99% correct links
  - Cohort is slightly biased to those of higher socioeconomic status
- Deterministic linkage to annual place of residence and Longitudinal Worker File
- Probabilistic linkage to mortality and cancer

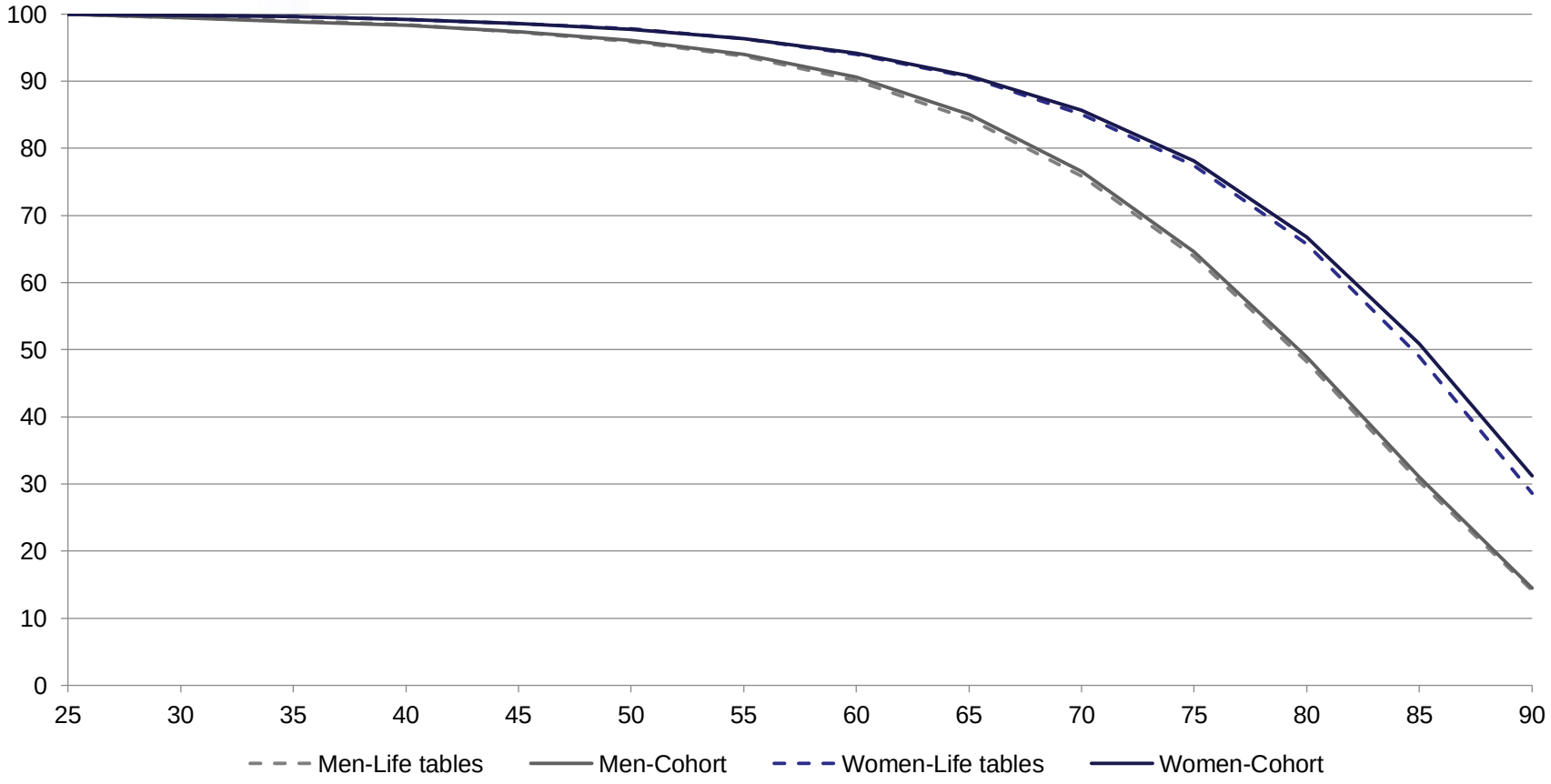
How "good" was the cohort?

Characteristic	Cohort	In-scope*
Total (count)	2,734,835	3,576,485
Sex (%)		
Male	49.7	48.6
Female	50.3	51.4
Age (%)		
25 to 44	54.5	52.6
45 to 64	30.0	30.5
65 +	15.4	16.9
Educational attainment (%)		
Less than secondary graduation	34.9	37.8
Secondary graduation or higher	65.1	62.2
Income adequacy quintile (%)		
Quintile 1-poorest	17.2	20.0
Quintile 5-richest	21.5	20.0

* In-scope refers to all individuals who were enumerated by the long-form, were aged 25+, and were not a resident of an institution

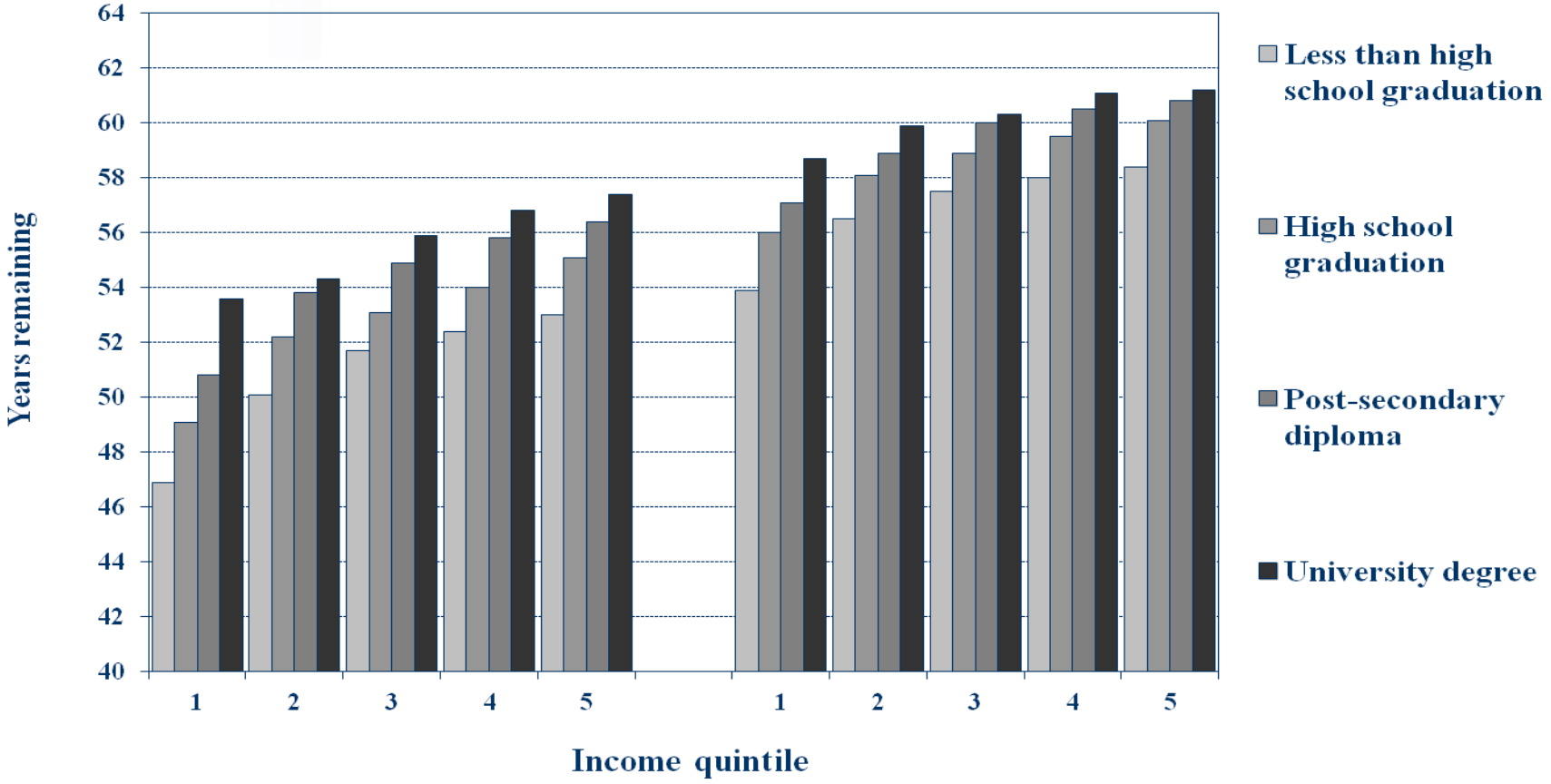
Results – survival

Percentage surviving to various ages in Canada for 1995-1997 and 2002 (average) compared to cohort for 1991-2006



Research results: Income and Education

Remaining life expectancy (at 25) by educational attainment within each income adequacy quintile and for each sex, 1991-2006 follow-up



Source: 1991 Canadian census cohort: mortality and cancer follow-up study (1991-2006)

Main strengths & limitations

- Strengths
  - Population based
  - Large sample size (rare outcomes, small population groups)
  - Able to examine several variables simultaneously
  - Long latency period required for cancer outcomes
  - Multilevel analysis
  - Captures residential mobility over a 27 year period (environmental exposure via the use of postal code representative points)

- Limitations
  - Census characteristics only measured at baseline (1991)
  - No information on health behaviours
  - Some population exclusions
    - Non tax filers, under the age of 25, institutional residents at cohort inception, those not enumerated by 1991 long form census

Thank you!

- Philippe Finès, philippe.fines@canada.ca

Record linkage at StatCan	
http://www.statcan.gc.ca/eng/record/gen	General information
http://www.statcan.gc.ca/health-sante/link-coup-eng.htm	For Health
http://www.statcan.gc.ca/eng/record/policy4-1	Statistics Canada's official directives on our record linkage activities.
http://www.statcan.gc.ca/eng/record/summ	This is a list and description of previously approved record linkage activities
Social Data Linkage Environment (SDLE)	
http://www.statcan.gc.ca/eng/sdle/index	(click on "DRD linkage status" for a list of data sources that are already linked in which you may be interested)
Research Data Centers	
http://www.statcan.gc.ca/eng/rdc/index	The Research Data Centres (RDC) Program
http://www.statcan.gc.ca/eng/rdc/network	List of RDCs
http://www.statcan.gc.ca/eng/rdc/data	List of datasets currently available in the RDCs
http://www.statcan.gc.ca/eng/rdc/process	Application process and guidelines
statcan.hsdrecordlinkage-dsscouplageenregistrements.statcan@canada.ca	HSD Record Linkage Mailbox