RECENT DEVELOPMENTS AT STATISTICS CANADA IN THE LINKING OF COMPLEX HEALTH FILES

Martha E. Fair
Statistics Canada

Many of the common social, economic, occupational and environmental health concerns of today are complex and multi-faceted. Census of Population, vital statistics, cancer morbidity, radiation exposure and work histories are important national data sources because of their comprehensiveness and inclusiveness over time. Record linkage is the process of bringing together two or more data sources relating to the same individual, family or entity.

This paper examines some practical methods recently developed to improve the searching, comparison and grouping of records in a typical generalized linkage process. The data sources used to illustrate this are from linkages using the Canadian Birth Data Base (from 1985), the Canadian Cancer Data Base (from 1969) and the Canadian Mortality Data Base (from 1950). In particular, the procedures developed for a birth-death linkage have been used to improve: a) the quality of the data files themselves (e.g. by identifying duplicate and missing registrations), b) the accuracy of carrying out generalized linkages, c) the analytical value of the vital statistics data (e.g. for a study of maternal education and fetal and infant mortality) and d) the development of new health indicators (e.g. using birth weight, gestational age and age at death variables).

Introduction

The increasing capacity of users to incorporate statistical information in their decision-making and research challenges statistical agencies to improve the quality and breadth of their information. This increasing demand is also accompanied by a rapidly changing technological environment and the increasing skill of the workforce for bringing together and analyzing data. For example, vital statistics, census, administrative and survey data are used for a multiplicity of purposes, and data are integrated from a number of different sources. One important tool to achieve the integration of these data is by computerized record linkage. Record linkage is the bringing together of two or more records, usually in order to match up those records relating to the same individual, family, event or entity. This paper will examine some of the practical methods developed at Statistics Canada to improve the searching, comparison and grouping of records in a typical generalized linkage process. Some background is also provided as to why record linkage of health data is needed, and some of the current initiatives that are furthering this work in Canada.

Background

Statistics Canada is the centralized statistical organization in Canada. The Health Statistics Division of Statistics Canada is authorized under the Statistics Act to collect, compile, analyze, abstract and publish statistics relating to the health and well being of Canadians. The Division's primary objective is to provide statistical information and analyses about the health of the population, determinants of health, occupational and environmental health, vital statistics and cancer, and the scope and utilization of Canada's health care sector.

In the health statistics area, there are many forces that are changing the type and breadth of statistics required. Recently, Statistics Canada, the Canadian Institute for Health Information

and Health Canada have joined forces to create a National Health Information Roadmap. These organizations conducted a broadly based national consultation on health information needs in 1998. Participants stressed that national agencies must build on and contribute to the considerable investments and expertise at local, regional and provincial/territorial levels (1).

Canadians are keenly interested in how to improve their health, and how to prevent disease. The health care system is a major contributor, but there are many factors beyond the health care system influencing health. Our health care system has suffered from a lack of information for too long. It has been stated that managing an enterprise which accounts for 10% of the economy and 30% of the provincial/territorial budgets without good information is like flying a jumbo jet without an instrument panel (2).

The health information system of the future should be secure, consistent, relevant, integrable, flexible, user-friendly and accessible (2). Our health information system should also provide us with information to answer the following two crucial questions: i) How healthy is the health care system? and ii) How healthy are Canadians?

There are a number of forces changing the health care environment today, such as: increasing and changing demand; demographic shifts; the increasing political importance of health care; health care consumerism; new technology developments; and the continuous pressure regarding costs (3). If one looks at the forecast of Canadian populations, there will be a significant increase in demand for health care services predicted because of the doubling of the 65 and over population in the next 25 years and the significant increase in the 40-65 age range. Longer survival will increase the number of patients with high-cost chronic diseases.

The Roadmap initiatives (4) are divided into six themes: i) National consultation; ii) Standards; iii) Health Services Data Gaps; iv) Population Health Data Gaps; v) Canadian Population Health Initiative; and vi) Health Reports and Indicators. At present, data in the health care system tend to be organized around who is providing the service, not around who is receiving the service. To help fill various population health data gaps, we need a person-oriented perspective of what happens to individuals over time.

The Types of Linkages Required

The types of linkages required to change the type and breadth of health statistics required are as follows:

1) *internal linkages*, e.g. linkage within the same file to create health histories by person;
2) *two-file linkages*, e.g. linkage of a cohort file, such as a group of workers in a particular industry, with mortality records;
3) *intermediate linkages*, e.g. linkage back to a master file containing unique numbers and name identifiers, when only a number may be available on one of the source files;
4) *linking reference files* (e.g. geographic files) to add new data to data sets;
5) linkage as *part of an operational system* environment (e.g. cancer registry); and
6) linkage for *adding analytical variables* (e.g. work and exposure histories).

Longitudinal follow-up of persons, over time and place, with a particular interest in the outcomes, such as causes of disease or death, is required.

Existing National Data Sources – Some Examples

The Provincial and Territorial Registrars across Canada collect vital statistics data on live births, fetal deaths and deaths occurring in Canada as well as some deaths of Canadian residents occurring in the United States. Surname and alternate surname fields are assigned a phonetic New York State Intelligence Information System (NYSIIS) code (5). The live birth and stillbirth data are stored in the *Canadian Birth Data Base* (CBDB) (6) and mortality data are stored in the *Canadian Mortality Data Base* (CMDB) (7). Income tax summary files are available from 1984 onward to help evaluate death searches and confirm whether an individual is *alive*.

The *Canadian Cancer Data Base* is an historic file held at Statistics Canada. It contains cancer incidence data from 1969 onward reported by all Canadian provincial and territorial cancer registries. The CCDB was created for undertaking historical cancer incidence record linkage and epidemiological studies.

The typical sizes of the data bases utilized are shown in Table 1 while Table 2 indicates the typical size of cohorts being followed up.

Table 1. Typical Sizes of Data Bases Utilized in Record Linkage Studies

Data Base	Years of Data	Volume*	Annual # of Events
Canadian Birth Data Base	1985 – 1995	4.8 million	400,000/year
Canadian Cancer Data Base	1969 – 1995	3.2 million	125,000/year
Canadian Mortality Data Base	1950 - 1995	9.7 million	200,000/year
Alive Follow-up File	1984 – present	24.6 million	21 million/year

*The volume of records includes alternate entries generated for the births, cancer and mortality file (e.g. maiden name, alternate spelling)
The alive follow-up file is a summary file and counts the individual once.

Linkage and Related Software Available

The mainframe software currently available are: Generalized Record Linkage System V1, Match360 (both of which have been developed at Statistics Canada), and SAS programs. GRLS V4 has been recently developed (8). Some of the features of the system are:

- Runs using UNIX and ORACLE
- Based on Fellegi-Sunter linkage methodology (9)
- Has graphical interface
- Allows multiple concurrent users
- Allows user-defined rules which are programmed in C
- Linked records can be grouped into "weak" and "strong" groups
- Allows refinements of weights and thresholds
- Bilingual
- On-line help is available
- Allows for sampling, and
- Has NYSIIS and Soundex rules built-in.

Table 2. Typical Size of Cohorts Along with Some Examples

A.	**Large Cohorts** (100,000 records or greater)	
	Occupational Cohort	700,000
	National Dose Registry	400,000
	Census of Agriculture	325,000
	Births	400,000 per year
B.	**Medium Cohorts** (10,000 - 99,999)	
	Breast Screening Cohort	
	British Columbia Pulp and Paper	
	Ontario Pulp and Paper	
	INCO workers	
	INCO-Manitoba workers	
	Imperial ESSO workers	
C.	**Small Cohorts** (1,000 - 10,000)	
	Stirling County survey	
	Sherritt International workers	
D.	**Manual Searches** (1000 or less)	
E.	**Typical Sizes of Exposure/Work History Data Files** (100,000 or greater)	
	Occupational Cohort - work histories	
	National Dose Registry work histories	
	INCO work histories	

Although name and address standardization may not form part of the available linkage software, this is an important first step. Generally there are two components involved: i) separating a string format field into its component parts and ii) optionally standardizing the components. Addresses may be present in a variety of formats. Historic data may be different than that of the current time (for example, earlier data will not have the postal code available). The data may have abbreviations, rural routes, box numbers, apartment numbers, inconsistencies and misspellings. A number of stand-alone routines and software have been developed to aid solving these problems at Statistics Canada (e.g. NSKGEN5 and NSKGEN7 for personal and business names; ENCODA, ASKGEN2, and PAAS and PCODE for addresses and postal codes) (10).

The Infant Birth-Death Linkage

As an example, we will now examine in detail a recent birth-death linkage study, outline some of the practical lessons learnt in carrying out this project, and point out utilization of some of the desirable software features.

Data Handling and Preprocessing

Live births for the years 1985-1994 were split from the CBDB. All birth records for these years were included in the linkage, and necessary exclusions were made later (e.g. Newfoundland data for the years 1985-1990 due to missing identifiers). Records were selected from the CMDB for the years 1985-1995 for infants born in the years 1985-1995. To ensure that all infant deaths had indeed been included, records were also selected if the cause of death field was coded to

perinatal death (ICD9=760 to 779) or to congenital anomaly (ICD9=740 to 759). Geographical data (e.g. postal code and census subdivision) and birth weight (Quebec only) were added to the death records used in the linkage.

In the birth-death linkage, name formats were separated into their component parts. In addition to format, there is the problem of nicknames, titles, alias names, multiple names, double-barreled names, initials only, suffixes (such as Jr.), synonyms for missing (Baby, twin etc.). Computer programs were developed to handle these.

The linkage

The birth-death linkage was done using the mainframe version of the Generalized Record Linkage System (GRLS. V1). The system is based on the Fellegi-Sunter model of record linkage (9). To work out the methodology for the linking of the files, data for the 1990-1991 birth cohorts were linked to 1990-1992 deaths. Following this, a full production run was prepared where almost 4.6 million live birth records for the years 1985-1994 representing 3.8 million individuals were compared to 62,285 death records representing 32,994 individuals for the years 1985-1995.

Three passes of the data were done. The sex code and NYSIIS surname code were defined as the GRLS.V1 "*pocket*" or "*blocking variables*" on the first data pass. The pocket fields allow all records in the same pocket to be compared. The sex code and date of birth fields were defined as the pocket on the second data pass; and the date of birth field alone was used in the final pass. The pair selected is the one receiving the highest weight from all three passes (Table 3).

Table 3. 1985 – 1993 Results (1990-1991 Results)

Pass	Blocking Variables	Frequency	Per Cent	Cumu-lative	Per Cent
1	Sex code and NYSIIS	17,471	98.2 (97.0)	17,471	98.2
2	Sex code and Birthdate	286	1.6 (2.3)	17,757	99.8
3	Birthdate alone	42	0.2 (0.7)	17,799	100.0

The infant's identifying information on the live birth file was compared to the decedent's identifying information on the death file. The *linkage variables* included: surname, alternate surname, given names or initials, date of birth, place of birth, place of event, place of residence, parental given names, mother's maiden surname, sex code, birth weight, and parental birth place codes. Logarithmic weights reflecting probabilities for or against a linkage were assigned to the results of the comparisons and summed. Each item is assumed to be independent. Where this is not true, user-defined rules were written to help take this into account (e.g. birthplace of the mother and father, residence information including postal code, census division and sub-division).

A basic comparison of the birth and death files was carried out. All record pairs achieving a total weight ≥ -100 were retained. Frequency weights were applied to the resulting record pairs in order to utilize the *discriminating power of the linkage variables* more fully. New *thresholds* can be set to determine levels at which linkages are regarded as good links, possibly linked or nonlinked pairs (11).

The generalized record linkage system offers a variety of *mapping and reporting options* to the user. For example, a one-to-one mapping selects the birth-death record pair with the highest weight, whereas a one-to-many mapping would retain one birth record and give all the death records to which it linked. Similarly, a many-to-one mapping would give all the birth records that link to one death, whereas a many-to-many option would give all the birth and death records that found a link. *Reports* can be prepared where the best link and the runner up are within a given range of weights.

Strategies for carrying out manual resolution were developed to ensure that siblings were linked correctly. For example, children in the same family having the same surname may achieve a high weight because all the parental identifiers are similar, and the surnames are rare. This is particularly true for twins and multiple births. A one-to-many mapping was done to account for multiple birth events linking to the same death registration and other data quality issues (e.g. duplicate entries for births). Many birth registrations linking to the same death registration were reviewed if the difference between the total weight values of the competing links was ≤ 175 points.

A *one-to-one mapping* was then done at a weight of -90 for the complete file. Computer printouts listing linkage identifiers and other items such as street address and spouse's given names were generated to review selected links manually. One listing was prepared for the review of singleton births if the total weight was ≤ 300. A sample of the singleton links greater than 300 was reviewed and it was decided that these links were correct. Live birth and death registration forms were consulted in the manual review of the printouts. Other steps were taken to ensure that the multiple live birth events were correctly linked. The final threshold was set at +50 and manual updates were made.

Results - Record Linkage and Data Quality

Internal linkage of the birth and death files

During the course of the manual resolution, it was noted that two or more birth or death records could actually be referring to the same individual, with each record having a unique registration number. It was also noted that some multiple births had been entered as a singleton under the birth type code. Accordingly, an internal linkage to bring together all records by person was done on the live birth and stillbirth files. The program identified potential duplicate entries and miscoded birth types. Computer printouts were prepared and manually reviewed. Original live birth and stillbirth registrations were consulted where required.

A "duplicate entry" flag was added to the file with a coded value. An example of a duplicate entry is an incomplete birth registration, followed up later by a second, more complete registration. However, the same event is assigned two unique registration numbers and remains on the file as two distinct records. For multiple births coded as singleton events, a new "birth type code" was created.

An internal linkage was carried out on the infant death file where only a few cases of duplicate entries were found.

Missing Links
Death registrations of infants born in Canada aged 0 to 364 days that did not link to a live birth were reviewed. Decedents born in Newfoundland for 1985-1990 were excluded. The CBDB was queried and the provincial and territorial registrars were contacted. A subset of the Ontario Physician's Notice of Birth (PNOB) forms for 1990 and 1991were reviewed, finding an under-reporting of live births in Ontario. The PNOB forms without a corresponding birth registration were found for some of the unlinked death registrations.

Post-processing – creation of analysis files
Live births were followed for one year for infant mortality. Stillbirth data were also available. A composite birth-death record was generated for the linked pairs to cross-classify items from the death registration with items on the live birth registration forms. A birth-death analysis file was created consisting of the following: composite birth-death linkages, survivors (unlinked birth records), unlinked death records of infants age 0 to 364 days born in Canada, and fetal deaths in the years 1985-1994. The analysis file did not include records for Newfoundland for the birth years 1985-1990. The unlinked death file excluded decedents born outside of Canada as well as those born in Newfoundland prior to 1991.

Validation Study

A validation study was initiated to test the validity of the probabilistic linkage methods used at Statistics Canada for the birth-death linkage. This was initiated after the 1990-1991 study was completed. It compared the results of the vital statistics data linkages of infant deaths in Canada with information collected from hospitals in the provinces of Nova Scotia and Alberta. Particulars of the mother and child were collected at the time of birth. It also compared the availability of fetal deaths on the national and provincial files. Finally, it examined the values of two important analytical variables – gestational age and birth weight – that were used for descriptive and analytical purposes. Highlighted in this study (12) were data definition problems, particularly with respect to fetal deaths. Overall the comparable linkage rate was 99% for infant deaths in Nova Scotia, and over 99% for Alberta neonatal deaths. Agreement on fetal deaths between Nova Scotia and the Statistics Canada file was 92% whereas for the Alberta file it was nearly 99%. The study findings point to the importance of complete and correct identifiers, as well as the uniform application of event definition (e.g. for livebirth and stillbirth).

Data Analysis

Fetal and infant mortality rates have declined rapidly in industrialized nations in the past decades. In Canada, these rates are among the lowest in the world. Nevertheless, there are still marked disparities in infant mortality by socio-economic status. A study was completed to examine the differences in fetal and infant mortality by maternal education in the province of Quebec, where the rates are among the lowest in Canada. The data used were from the linked birth and infant death records (including stillbirths) for the 1990-1991 cohorts in Quebec.

The main results (13) indicate that marked differences in fetal and infant mortality by maternal education persist despite many years of universal access to publicly funded, high-quality health care. If all mothers were able to attain the low rate of fetal and infant mortality already achieved

by Quebec mothers with the highest educational attainment, one-fifth of all fetal and infant deaths and nearly one-third of post-neonatal deaths could be avoided.

Maternal education affects fetal and infant mortality largely through the key intervening factors of low birth weight, or pre-term and small-for-gestational-age births. Consequently reductions in fetal and infant mortality will require addressing the causes of these intervening factors. The highest potential for improvement exists among mothers with less than 12 years of education, who accounted for a disproportionately large share of excess deaths, especially in the post-neonatal period and for non-congenital-related conditions.

Indicators

Aggregate information and indicators can be used to assess the health and well being of children during the prenatal period and infancy (14). Some of the key direct indicators from vital statistics are measures of infant mortality and measures of low birth weight. With a linked birth-death file, one can also examine birth weight and gestational age-specific mortality by the age of death.

Combining Data from Different Sources

Several record linkage workshops have been held earlier in Canada and the focus of the recent Statistics Canada's XVIth Annual International Symposium held May 4-7, 1999 (15) was on the techniques and methods for combining data from different sources and on the analysis of the resulting data sets. For example, discussed was a collaborative study between the Manitoba Centre for Health Policy and Evaluation and Statistics Canada that was conducted to link provincial administrative health care utilization with census data for a sample of Manitobans. Earlier, mortality and health service utilization have been described in relation to the socio-economic status measures, mortality and use of health care services at seven different stages in the life course (16).

Statistics Canada conducts the National Population Health Survey (NPHS). The NPHS collects both *cross sectional and longitudinal survey data* on the physical and mental health of Canadians and their use of the health care services (17). The longitudinal sample of 17,276 randomly selected individuals will be re-interviewed every 2 years for up to 20 years. Three NPHS cycles have been completed: cycle 1 in 1994/95; cycle 2 in 1996/1997; and cycle 3 in 1998/1999. Record linkage is important in the tracing of individuals, in the validation of items collected and in the additional of variable for analysis. Efforts to locate NPHS respondents who have moved between cycles without notifying Statistics Canada have generally been successful. Only 1.7% of respondents from Cycle 1 could not be found for cycle 2. A mortality record linkage search of the NPHS was carried out to match longitudinal respondents who were reported as deceased with the Canadian Mortality Data Base to confirm their death and to collect the cause of death.

To increase the analytic usefulness of the survey data, research is under way to link external provincial records with NPHS responses about the use of health care services. Such linkages are done only with the respondent's permission. About 94% of the respondents agreed to have their data followed up in Cycle 2.

The Future

- Generalized record linkage software has been used in a variety of research applications. A number of additional large studies are planned.
- Address and name standardization computer programs and software are useful tools to facilitate linkage and reduce error rates and their development should be encouraged.
- It is now recommended that an internal file linkage be done on the live birth and stillbirth records as part of the routine processing of files being linked.
- The problem of missing birth registrations in Ontario is under further investigation and a special workshop was held in Ontario to address data quality issues in vital statistics registrations.
- Health Canada is leading the development of the Canadian Perinatal Surveillance System. A Fetal-Infant Mortality Study Group, chaired by Dr. Alexander Allen, is planning a number of research articles using the linked birth-death file. In addition, deaths for women of childbearing age have been linked to birth records to examine maternal mortality.
- A study is examining the feasibility of linking births to a file of registered Indians.
- Considerable effort is being placed on how patient-oriented cancer data can be created with emphasis on studies of cancer survival and research into this disease and its risk factors.
- A person oriented information project is being developed to increase the capacity to combine health care data and to use this capacity to provide information on the health of Canadians and the effectiveness, efficiency and responsiveness of our health care system.

Acknowledgements

Thanks to M. Cyr, P. Lalonde and J. Carswell who worked on the birth-death linkages. Health Canada, through the Canadian Perinatal Surveillance System, supported the linkage of the birth and death records on the Canadian Birth Data Base and the Canadian Mortality Data Base. The co-operation of provincial/territorial cancer and vital statistics registrars, who supply cancer, birth and mortality data to Statistics Canada, is gratefully acknowledged.

References

1. Canadian Institute for Health Information, Health Canada, Statistics Canada. *Health Information Needs in Canada.* Available from: CIHI, 377 Dalhousie Street, Suite 200, Ottawa, Ontario K1N 9N8, 1998. Also on the CIHI website (http://www.cihi.ca).

2. Canadian Institute for Health Information, Health Canada, Statistics Canada. *Health Information Roadmap Responding to Needs.* Available from: CIHI, 377 Dalhousie Street, Suite 200, Ottawa, Ontario K1N 9N8, 1999. Also on the CIHI website (http://www.cihi.ca).

3. Guerriere, M. *Why do we need better Health Information?* A paper presented at the CIHI spring conference: The Journey towards Better Health Information, Montreal, May 12, 1999.

4. Canadian Institute for Health Information, Health Canada, Statistics Canada. *Health Information Roadmap Beginning the Journey.* Available from: CIHI, 377 Dalhousie Street, Suite 200, Ottawa, Ontario K1N 9N8, 1999; also on the CIHI website (http://www.cihi.ca).

5. Lynch BT, Arends WL. *Selection of surname coding procedures for the SRS record linkage system.* Sample Survey Research Branch. Research Division, Statistical Reporting System. U.S. Department of Agriculture. Washington, D.C. February 1977.

6. Fair M, Cyr M. The Canadian Birth Data Base. A new research tool to study reproductive outcomes. *Health Reports* (Statistics Canada, Catalogue 82-003) 1993; (3): 281-290.

7. Smith ME, Newcombe HB. Use of the Canadian Mortality Data Base for epidemiological follow-up. *Canadian Journal of Public Health,* 1982; 73:39-45.

8. Statistics Canada. GRLS V4 *Generalized Record Linkage System Tutorial,* Systems Development Division, (1999).

9. Fellegi IP, Sunter AB. A theory of record linkage. *Journal of the American Statistical Association,* 1969; 40:1183-1210.

10. Statistics Canada. General Systems Sub-division of Systems Development Division. *Record Linkage Software.* Statistics Canada internal report, September 1989.

11. Newcombe HB. *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business.* Oxford, U.K.: Oxford University Press, 1988.

12. Fair M, Cyr M, Allen A, Guyon G, Wen SW. Fetal Infant Mortality Study Group. *Validation Study for a Record Linkage of Births and Infant Deaths in Canada.* (in preparation).

13. Chen J, Fair M, Wilkins R, Cyr M, and the Fetal and Infant Mortality Study Group of the Canadian Perinatal Surveillance System. Maternal education and fetal and infant mortality in Quebec. *Health Reports* (Statistics Canada Catalogue 82-003), 1998; 10 (2), 53-64.

14. Lantz P, Partin M. Population indicators of prenatal and infant health. In: Hauser RM, Brown BV, Prosser WR (eds.) *Indicators of Children's Well-Being.* Russell Sage Foundation, New York, 1997; 47-74.

15. Statistics Canada. Symposium 99: *Combining Data from Different Sources.* May 4-7, 1999. Ottawa. (Proceedings in preparation.)

16. Mustard C, Derksen S, Berthelot J-M, Wolfson M, Roos LL, Carriere KS. *Socio-economic Gradients in Mortality and the Use of Health Care Services in the Life Course.* Manitoba Centre for Health Policy and Evaluation, Department of Community Health Sciences, Faculty of Medicine, University of Manitoba, 1995.

17. Swain L, Catlin G Beaudet MP. The National Population Health Survey – its longitudinal nature. *Health Reports (*Statistics Canada Catalogue 82-003) 10(4): 69 – 82.