Austrian Journal of Statistics
Volume 33 (2004), Number 1&2, 37-53

Generalized Record Linkage System – Statistics Canada's Record Linkage Software

Martha Fair
Statistics Canada, Ottawa

Abstract: The Generalized Record Linkage System is a probabilistic record linkage system designed for use by a wide range of statistical applications. It is part of a toolbox of generalized systems developed at Statistics Canada in response to business and health needs. This article gives an overview of the record linkage system and its application in the health area. The system is used to improve data quality and coverage, for long term medical follow-up of cohorts, for creating patient-oriented rather than event-oriented data, for building new data sources, and for a range of other statistical purposes. Described are the three main stages in carrying out the linkage, namely 1) the searching stage; 2) the decision stage; and 3) the grouping stage. Outlined are the essential criteria for the evaluation of such software.

Zusammenfassung: Das generalisierte Rekord Linkage System ist ein probabilistisches System, das für einen weiten Bereich von statistischen Anwendungen bestimmt ist. Es ist ein Teil einer Toolbox von generalisierten Systemen, die von Statistics Canada entwickelt werden, um Anforderungen aus Wirtschaft und dem Gesundheitswesen nachzukommen. Der Artikel gibt einen Überblick über das Rekord Linkage System und seine Anwendung im Bereich des Gesundheitswesens. Das System wird benutzt, um die Qualität und den Geltungsbereich der Daten zu verbessern, für langfristiges medizinisches follow-up von Kohorten, zum Erstellen von Patienten-orientierten anstelle von Fall-orientierten Daten, zum Aufbauen neuer Datenquellen und für einen weiten Bereich anderer statistischer Zwecke. Es werden die drei zentralen Stadien beschrieben, in denen des Rekord Linkage ausgeführt wird, nämlich 1) das Stadium des Suchens; 2) das Entscheidungs-Stadium; und 3) das Stadium des Gruppierens. Die wesentlichen Kriterien für die Bewertung einer solchen Software werden kurz dargestellt.

Keywords: Matching, Probabilistic Linkage; Longitudinal follow-up; Record linkage software.

1 Introduction

Record linkage is the process of bringing together two or more records relating to the same individual, family or entity (e.g. event, business and geography). The insertion of address and telephone changes into a mailing or telephone list and the removal of duplicate entries from a mailing list are basic examples of record linkage. Historically

record linkage was assigned to clerks who would search and review lists to bring together the appropriate pairs of records for comparison, seek additional information when there were questionable matches, and finally make decisions regarding the linkages based on established rules.

The term *record linkage* was first used by the chief of the U.S. National Office of Vital Statistics, Dr. Halbert L. Dunn in a talk given in Canada in 1946. Dunn advocated the use of a unique number (e.g. birth registration number) to facilitate such linkages. Dunn (1946) is worth quoting for his imagery:

„Each person in the word creates a book of life. The book starts with birth and ends with death. Its pages are made up of the principle events of life. Record linkage is the name given to the process of assembling the pages of the book into a volume. "

Today computerized record linkage problems arise where there are no permanent universal unique lifetime identifiers available on records. Even where numerical identifiers exist (e.g. health insurance number or business number), there may be errors in reporting, legal restrictions regarding their use, or the identifiers may change when people or businesses move. Identifying variables (names, birth dates, addresses) available on records may vary, and the consistency and manner in which they are reported often make them difficult to standardize and parse. Some names are very common, or else they may be rare, but cluster in certain geographic areas.

Computer systems and probabilistic linkage techniques have been developed at Statistics Canada to improve the ease, speed and accuracy of carrying out linkages in spite of these difficulties. This is the main focus of this paper.

The outline of this paper is as follows. As background, we examine the historical vision of record linkage and identify some critical factors affecting its evolution in Canada (e.g. development of large administrative files, improved computer technology, increased demand for detailed statistics, and the concerns of privacy and confidentiality). In the third section, we describe some software and methodology requirements for record linkage software. In the fourth section, we examine the current generalized record linkage system used at Statistics Canada, and then describe its features and the major phases in carrying out a linkage. In the fifth section, specific applications in health and business are highlighted, along with various features that have made each case important in terms of the experience gained in carrying out the linkages. The final section describes some future developments and concluding remarks.

2 Background

The main drivers of record linkage have been researchers and the user community's demand for detailed statistical information, along with the work of statisticians interested in serving clients needs, reducing respondent burden and costs, plus improving data quality and timeliness.

The idea of computerized record linkage emerged in Canada with the vision of using existing administrative and health records to answer research questions relating to genetics, occupational and environmental health and medical research. It was recognized early that diverse data sources (e.g. vital records and hospital records), that by themselves give limited information at one point or event in time, could be linked

together to generate longitudinal person-oriented information with outcome information generated over the life course.

Newcombe and his associates (1959) required quantitative data regarding the effects of radiation in human populations. The researchers had the foresight to look at the possibility of using computerized record linkage of vital statistics and health surveillance records to help answer this question. The first issues that had to be addressed were key technical issues (Newcombe 1988). There were no unique identifiers on the files, and the problem of discrepancies in variables did exist. There was also the question of processing the large volume of data with reasonable computer time. Ad hoc computer programs were developed to carry out the linkages of vital records into individual and family groupings.

This example also illustrates the importance of a multi-disciplinary approach and collaboration among agencies to address record linkage projects. Dr. Newcombe was a geneticist at Atomic Energy of Canada, Al James was a geneticist working in the laboratory, Jim Kennedy was a theoretical physicist, and Sam Axford worked at the then Dominion Bureau of Statistics.

At this time, social, tax and health administrative programs were being put into place in Canada. Today these administrative data sources form essential ingredients for record linkage of large population based files at the provincial and national level. These administrative data were also being used for statistical purposes.

The mathematical theory of record linkage work of Drs. Fellegi and Sunter (1969) at Statistics Canada was not motivated by health research issues. Rather, it was explicitly oriented to the problem of merging the information content of large administrative files in order to create a statistically useful source of new information (see Fellegi, 1997). In the presence of identifying information on records, how does one decide which record pairs of potential comparisons should be regarded as linked? They (1) rigorously described the comparison space of record pairs consisting of all possible comparisons; (2) provided a calculus for comparing the evidence contained in different record pairs about the likelihood that they refer to the same underlying unit; (3) defined a linkage rule as partitioning the comparison space into subsets that we called „linked“ i.e. record pairs about which the inference is that they indeed link to the same underlying unit, a second subset for which the inference is that the record pairs refer to different underlying units; and a complementary third area where the inference cannot be made without further evidence; (4) identified that characteristic Type I and Type II errors associated with a given linkage rule; and (5) showed that if the space of record pair comparisons is ordered according to the metric, this will result in a linkage rule that is optimal for any pre-specified Type I and Type II error levels.

In early 1980s a generalized linkage system was developed at Statistics Canada in collaboration with the National Cancer Institute of Canada (Howe and Lindsay, 1981). The statistical framework used in developing the system was the Fellegi- Sunter model. This system allows for the linkages between any two files of interest or it can be used to internally linking files (e.g. to create individual histories or to remove duplicate records). Communication and feedback between the system developers, methodologists and users improved the development of the generalized system software. At this time, new computer technology facilitated the maintenance, integration and extraction of information from large databases.

Another critical factor that has impacted the development of record linkage is the public concern for privacy and confidentiality (Fellegi, 1997). The legal authority and mandate of an organization, along with the development of transparent record linkage policies and procedures are important.

Unlike many countries, most statistical activity in Canada is carried out within a single national agency. This agency is legislated by a uniform and strong Statistics Act. Currently, in addition to conducting a census every five years, there are about 350 active surveys on virtually all aspects of Canadian life carried out. Statistics Canada is committed to protect the confidentiality of all information entrusted to it and to ensuring the information is relevant to Canadians.

Statistics Canada has two main objectives:

1. To provide statistical information and analysis about Canada's economic and social structure. These data are used to develop and to evaluate public policies and programs and to improve public and private decision-making for the benefit of Canadians.
2. To promote sound statistical standards and practices by:

- using common concepts and classifications to provide better data,
- working with the provinces and territories to achieve greater efficiency in data collection and less duplication,
- reducing the burden on respondents through greater use of data sharing agreements (sources used include annual tax records), and
- improving statistical methods and systems through joint research studies and projects.

Dr. Ivan Fellegi, the Chief Statistician for Canada, described three main indicators of success of a statistical system:

- The adaptability of the system in adjusting its product line to evolving needs;
- How effective is the system in exploiting existing data to meet client needs; and
- How credible is the system in terms of statistical quality of its outputs and its non-political objectivity?

Statistics Canada will carry out linkages of different records only for statistical purposes and only when the results of the linkage would yield a potential public good that clearly outweighs the potential invasion of the privacy rights. This activity is conducted in accordance with the Agency's Policy on Record Linkage that has been in place since 1986. A description of all linkages approved since 2000 and a copy of our Policy are available on the website http://www.statcan.ca.

Statistics Canada carries out a wide range of record linkage work given the broad mandate of the organization. There are two major categories of record linkage conducted by Statistics Canada:

- linkages to support the design, maintenance, evaluation, research and re-design of ongoing data collection and methodological studies within the Agency; and
- linkages to provide statistical information in aggregate or anonymous format in support of research studies.

Many of the research studies are conducted on a cost recovery basis.

Statistics Canada maintains a large complement of specialists in the related areas of information technology and statistical methodology as well as subject matter experts. Internally, this critical mass of experts is organized to provide a broad range of centralized services to support Statistics Canada's programs.

One such service is the support of generalized software products that includes repetitive processes associated with survey and statistical processing. Some of the generalized products developed include: 1) sampling; 2) automated coding; 3) edit and imputation; 4) estimation; 5) record linkage; and 6) information retrieval. Each product has been designed using a tool-kit approach, which provides individual application designers with considerable flexibility, while working within a framework of standardized generalized components. It is the generalized record linkage system that is being described here from a user perspective and its use for carrying out a variety of applications.

3 Requirements for the Record Linkage System

Statistics Canada carried out an evaluation study regarding record linkage software (1998). We examined in general terms some of the requirements for record linkage systems, with the main focus being on its use for health research applications. Some factors affecting the choice of software for record linkage from a users perspective include: 1) the types of linkages required; 2) the frequency of the linkage; 3) the purposes of the linkages (e.g., statistical versus administrative) and the level of accuracy required; 4) the nature and size of the data files; 5) the data processing environment (hardware, software); and 6) the time schedule, personnel and budget available.

3.1 Requirements

Key requirements for probabilistic record linkage systems included consideration of the following items:

- the types of linkages required (see examples below),
- whether the linkages is performed in batch and/or interactive mode,
- the security provisions for confidential data files,
- the speed of operation needed,
- the volume of records that can be linked with the system,
- the initial cost of software including licensing and maintenance costs,
- whether the software is bundled with other software packages,
- the simplicity and flexibility in defining the rules used for linkages,
- the accuracy and statistical defensibility of the product,

- the availability of documentation and training, and
- the maintenance and support of the software.

Examples of the types of record linkage Statistics Canada carries out are as follows:

- *internal linkages*, e.g., linkages within the same file to create histories by person or entity ;
- *two-file linkages*, e.g., linkage of a cohort file, such as a group of workers in a particular industry, with mortality records;
- *intermediate* or „*bridge*“ *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;
- linking *reference* files (e.g., geographic files) to add new data to data sets; and
- linkage as part of an *operational system* environment (e.g., cancer registry).

3.2 Record Linkage Methodology

Examples of some of the methodological requirements that should be considered in examining probabilistic record linkage software are the following. There should be flexibility in the choice of blocking variables or criteria that are used in bringing record pairs together. Phonetic coding routines or string comparator functions may be available within the record linkage software, whereas in other cases it may be bundled and sold separately. The user should be able to select the linkage variables, and methods used by the computer system for making comparisons of items in order to make use of the discriminating power of the linkage variables. For example, it is important not only to indicate the agreement of surname, but also to examine the value of the agreement of the name taking into account its frequency in the population (e.g. agreement of Smith does not have as much discriminating power as the agreement of the surname Quigley). Ideally users should have the option of writing their own customized rules. For example, it may be necessary to write specific code and rules to compare agreement or partial agreement of place names, based on specific mobility patterns of workers for a specific project.

To save costs and ensure consistency, use of previous linkage results and previous manual resolution information may be needed when completing an update to a project. The system should make it easy for the user to accomplish the task of estimating the appropriate thresholds to accept and reject linkages. Suitable reports should also be available, such as online reports of matches, reports and histograms giving weight distributions, reports for clerical review, and reports grouping records together. Reports should also be available giving the match parameters, linkage weights and rules used. Backup and restore procedures should be available.

3.3 Data Handling and Pre-processing

Data handling and pre-processing may be done outside the linkage software itself. It is important to have software available to facilitate the pre-processing of the data files. This is often the most time consuming part of the whole process. The first step is to do an assessment of the data files themselves, and in the Health area we normally use SAS to carry out this step. The items that are common to the two records that are used in the linkage process are systematically examined to see the availability and coding of variables. This is compared with any documentation available for the files.

We phonetically code surname fields and separate names into their appropriate component parts (e.g. surname, given names, titles). Otherwise the identifying information will not be consistently compared and this will affect the linkage results. The New York State Intelligence and Information System (NYSIIS) and Russell Soundex code routines are available in the generalized record linkage software. However, we usually generate a six-character NYSIIS code from ten letters of the name at the pre-processing phase and stored this as part of the data file.

Address information on administrative record is typically in free format, that is to say there is no fixed position or even order to the components of the address. A postal address analysis system (DeGuire, 1988) has been created at Statistics Canada to facilitate the process of breaking up of addresses into its component parts (street name, street number, apartment number, and so forth). In business applications, software aimed at parsing and standardizing business names is used to generate names search keys (NSKGEN) that are used for record linkage.

A postal code conversion file (PCCF) is available within Statistics Canada. This provides a correspondence between the six character postal code and Statistics Canada's standard geographic area for which census data and other statistics are produced. Through the link between postal code and standard geographic areas, the PCCF permits the integration of data from various sources. The geographic co-ordinates attached to each postal code on the PCCF are commonly used to map the distribution of data for spatial analysis.

4 Generalized Record Linkage System

Large-scale record linkage using probabilistic matching techniques is done at Statistics Canada using the Generalized Record Linkage System (GRLS). The current version of the system GRLS (version 4) runs in a client-server environment with ORACLE and a C compiler. The software will also run on a PC or workstation that supports the UNIX operating system. The GRLS is particularly suited to applications where there are no unique identifiers available to carry out the linkage. GRLS improves both the quality and the ease of your linkage.

Based on statistical decision theory, GRLS breaks the linkage operation into three major phases as shown in Figures 1 and 2. 1) A searching phase is implemented where a table of all the potential pairs is generated using the initial criteria set by the user. 2) A decision phase is carried out where linkage rules are applied to the potential pairs and weights or odds ratios are assigned and the potential pairs are divided into sets of

definite (Def), possible (Pos) and excluded (Excl) pairs by the setting of upper (TU) and lower (TL) thresholds. Each linkage rule produces an outcome of missing, agreement, partial agreement, or disagreement. It is possible to refine the linkage weights and reset the threshold values using the possible and definite pairs that are then classified as definite, possible or rejected (Rej). 3) In the grouping phase, the pairs of possible and definite links which could relate to the same entity are grouped together in a manner specified by the user, and the final groups generated. Manual resolution may also be carried out during the process and the results updated.

The GRLS record linkage system provides a convenient framework for testing linkage parameters. It allows concurrent users for each linkage project; and permits background or interactive linkages. Detailed queries may be made to identify problems with either the pairs or groups. It provides options for sampling for testing purposes; and allows for the export and import of linkage tables. Different types of bilingual documents (English and French) are available to aid the user (e.g. record linkage concepts manual, user guide, strategy guide and tutorial). Bilingual courses discussing record linkage methodology and software are offered at Statistics Canada's training centre. The development of the software has focused on Statistics Canada users. On-site statistical services are bundled with the licensing of the software outside of Statistics Canada. Tutorials have been presented at various international workshops, such as that held in Washington in 1997 (Fair and Whitridge, 1997)

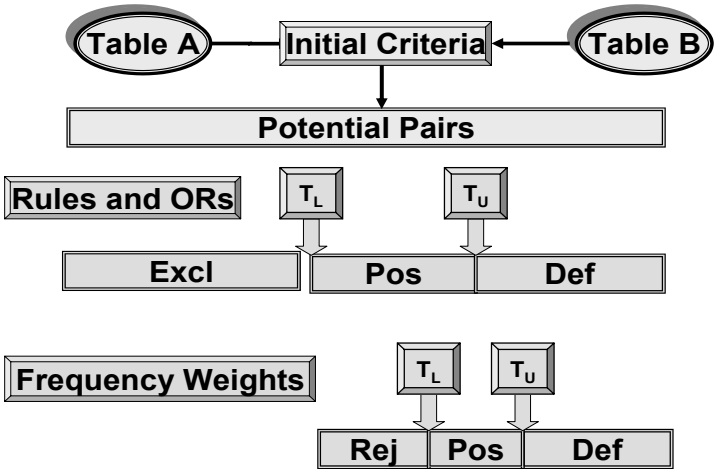

Figure 1: Implementation of linkage operation

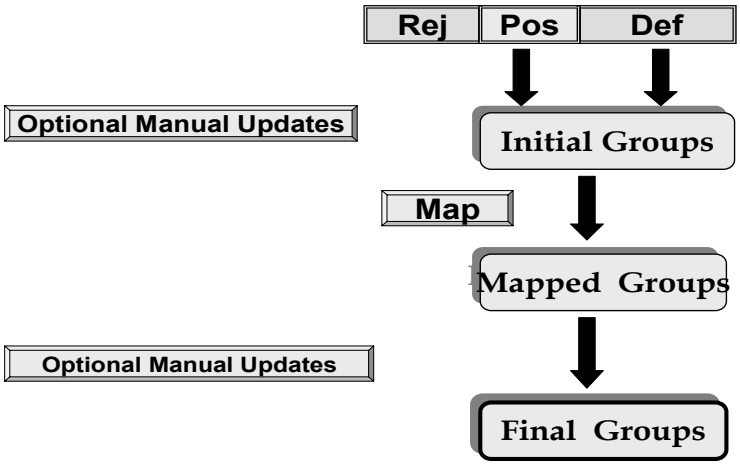

Figure 2: The grouping of pairs of records

5 Applications

Record linkage is an important technique in the development, production, analysis and evaluation of statistical data. It is an important tool for the creation of statistical data, particularly in relation to census taking, health research and in survey taking for social and economic statistics.

As described in an earlier workshop on record linkage techniques in Washington (Fair, 1997) the procedures used in bringing together different data sources can be viewed as a cube with three dimensions (see Figure 3). The first side represents the life cycle of the entity from birth to death. This could be for an individual, family or an entity such as a business or farm. The second dimension represents the degree of wellness to ill health. The third side represents the determinants. For example some of the health determinants are human biology, socio-economic status, employment and working conditions, social support, social economic and physical environment, health services, and public policy. There may be changes to our health status as we move through the different stages of our life cycle.

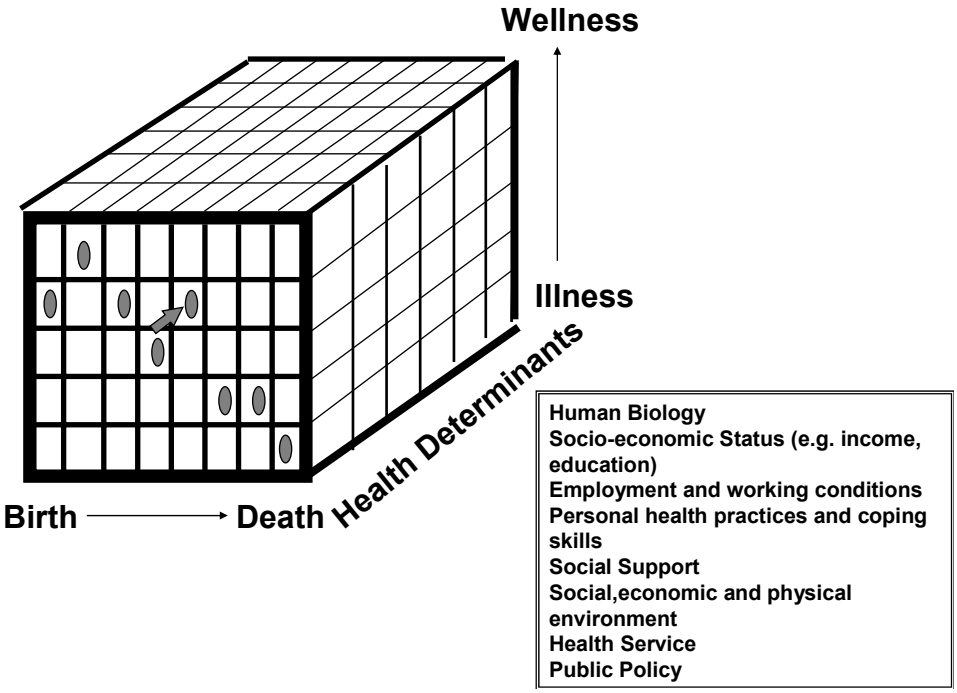

Figure 3: Record linkage over the life cycle

Some of the data sources that are used to indicate the health status of the individual in the population across time include vital statistics, survey and hospital morbidity information. Examples of surveys include the Canada fitness survey, the Canada Health Survey, and the National Population Health Survey.

National data sources have be developed for use in record linkage. 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. The live birth and stillbirth data are stored in the Canadian Birth Data Base (CBDB) dating back to 1985, cancer information in the Canadian Cancer Data Base dates back to 1969, and mortality data in the Canadian Mortality Data Base (CMDB) dates back to 1950. Income tax summary files from 1984 onward are used to help evaluate death searches and to confirm whether an individual is alive.

Births and deaths for businesses are considerably more difficult to establish than those for human populations (Colledge, 1995). They depend on the type of unit being considered. Birth and death criteria for a business' legal and administrative units are determined by legal and administrative rules and can be quite different from the criteria for the birth and death of the organizational units. These criteria may not be appropriate for defining the births and deaths of statistical units.

5.1 Health

The generalized record linkage system using probabilistic record linkage techniques is used for most health applications where identifiers are limited. Figure 4 gives an overview of a typical medical follow-up study. A cohort of individuals is usually supplied by an outside organization to Statistics Canada - this may include records for individuals dating back to the 1950s and the interest is in knowing whether there is excess cause-specific risk of mortality or cancer in the group. A series of linkages are required - the cohort is first linked to determine if the individuals are alive, followed by linkages to the Canadian Mortality Data Base to determine whether the individuals are deceased and if so the cause of death. Finally the cohort is linked to the Canadian Cancer Data Base to determine if the individuals have cancer. This information is then brought together with work and exposure history data. If data on other factors are available, such as smoking histories, these are added for each person.

In these linkages, the process of separating out the true links is, in reality, a stepwise elimination of the false ones. For example, if on file A there are 10,000 records and on file B there are 3 million records, then the total number of potential pairs is 30 billion. However, it may be known that there are only 1000 true links expected in this linkage. Thus the pairs to be excluded are 30 billion minus the 1000. Steps are used to reduce the number of pairs compared. This is achieved by specifying initial criteria to determine which pairs of records to consider for comparison – these are often referred to as blocking variables. Reject rules may also be used to reduce the number of comparisons made (e.g. there must be at least two items agreeing on the record before a comparison will be considered).

Ionizing radiation is a well-established risk factor for human cancer. Epidemiological studies of atomic bomb survivors, uranium workers and patients treated with ionizing radiation have revealed excess risks, particularly for leukemia and for lung cancer and for other solid tumours. Centralized registries of occupational radiation exposure, established in a number of countries for regulatory purposes, contain comprehensive records of radiation doses that are useful for epidemiological investigations.

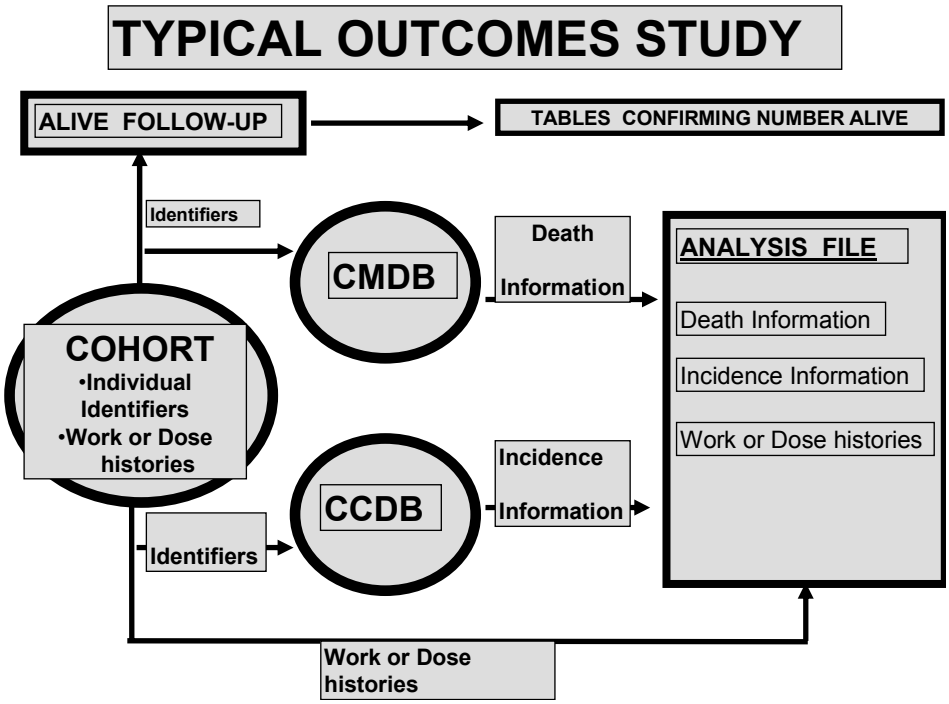

Figure 4: Typical follow-up study

The Canadian National Dose Registry (Ashmore et al. 1998, and Sont et al. 2001) is a centralized registry of records of occupational exposure to ionizing radiation dating back to 1950. Health Canada maintains it. It currently includes records on more than 500,000 individuals from over 24,000 organizations. This large cohort enables one to investigate sub-groups such as dentists, nurses, and medical workers. Other related studies are being conducted of Eldorado Nuclear workers and Newfoundland fluorospar miners, where the interest is in exposure to radon, and for the follow-up of Gulf War veterans.

The National Dose Registry mortality and cancer study illustrates the complexities of linkage over long periods of time (about 50 years). Individual histories had to be generated from the National Dose Registry. The formats, identifiers and coding practices changed over time. These records also had to be matched with millions of exposure records. The Canadian population is very mobile, and therefore many of the individuals moved between the time of initial registration on the file, plus individuals could have changes jobs and companies of employment. Cancer or death may have occurred years later.

Perinatal health surveillance is a necessary component of managing the health system to improve the health status of pregnant women, mothers and infants. Part of our ongoing work with Health Canada is to routinely link birth, infant death and foetal birth records (Health Canada, 2000). There have been over 18 papers prepared using the linked files, currently 3 have been submitted for publications, and 8 are in preparation. This record linkage study was challenging because of the variation in data quality of different provinces across the country on the national file. It also illustrates the

multiplicity of uses of the linked files that become possible when the data have been integrated.

Indicators, such as infant mortality and low birth weight can be developed using unlinked birth and mortality files. With a birth-infant death linked file one can then look at birth weight and gestational age specific mortality by the age at death. There is interest in being able to break down these indicators for sub-populations, such as the aboriginal population.

As globalization has increased, population has become more mobile, but comprehensive research on the health of diverse migrant populations is lacking. In Canada international migrants represent a large population (about 18% of Canadians). One study involved a cohort of immigrants who achieved landed status between 1980 and 1990. A sample of individuals including refugees was linked to the cancer and death files. This study addresses the question of how cancer incidence among subgroups of immigrants compare with that in the general Canadian population. One unique feature of this study is that it takes into account the duration of the individuals living in Canada. One challenging feature of this record linkage study was the variation of names. For example, Asian surnames are often short, and phonetic coding routines, particularly where vowels are ignored, are not always very helpful. Given names and surnames fields may be inconsistently recorded or alternate anglicized names used.

Breast cancer is one of the most common forms of cancer in women. If early detection of this disease through mammography is shown to improve the quality of life or extend life, then this practice should be encouraged. Conversely, if mammography is shown to be ineffective or deleterious to certain age groups, then it should be discontinued for these age groups.

The National Breast Screening study project began in 1980 and was designed to determine the efficacy of screening programs for women in different age groups (e.g. 40-49, 50 - 59) on entry into the study (Miller at al. 1992). All participants of the study have provided consent forms giving permission for linkage with vital statistics. This original study has been updated to include more recent years of mortality and cancer information.

Cohorts of surveys such the Canadian Health Survey, the Canada Fitness survey, and the Nutrition Canada survey have been followed up with mortality. The National Population Health Survey is also routinely followed up with mortality. The challenges with these follow-ups are to ensure accurate follow-up over an extended period of time.

5.2 Business

Business statistics offer particular challenges with record linkage (Winkler, 1995). In many respects, the applications have many similarities with linkages of individuals in the health area, and yet offer some additional challenges. For example, with businesses the population definition is often complicated because the organizations may be incorporated, family or individually owned or partnerships. The entities can operate in several provinces and the structures may change over time. Unincorporated businesses are very volatile and the same person can go in and out of business in a short time. The business name may be rare, there may be some confusion regarding whether the address and phone numbers refer to the home or business and birth dates may not be

consistently reported or may be unknown for partners. In some there may be a complex structure of multi-holding corporations with different locations, different addresses and phone numbers.

Some addresses are difficult to parse into components and there is a need to standardize items (e.g. inc, incorp, corp, lim, ltd). In Canada different languages may be used in the address (e.g. street, rue), and sometimes different naming customs are used across the country. There may be structural problems, with the units available to be matched not always the same units on both files (e.g. businesses versus owners). Unincorporated businesses may be owned by many partners, and individuals may also be involved in more than one business. The same name may be common in rural communities (clustering) and the family structures to some businesses where the parents and children are involved are difficult to match when the businesses are passed down to family members.

At Statistics Canada, a central business register provides the framework for the production of coherent statistics for National Accounts and the conduct of analytical studies linking data from the different business surveys. A central business register is less costly to operate than the maintenance of numerous local business frames from within each statistical program. A business register number was introduced by Canada Customs and Revenue Agency (CCRA) in 1994 as a means of integrating CCRA tax accounts under one number. This number has facilitated integrating the data - for further details regarding the business register see the Statistics Canada Internet site http://www.statcan.ca.

6 Future Directions

Record linkage is a powerful tool for generating more value out of existing data bases. A significant investment in record linkage can have a high rate of return in terms of knowledge, and, in turn, make possible a great deal of new research. Some of the future research applications that are envisaged within the health area involve the expansion of research within maternal and infant health to include morbidity files as well as vital statistics data. It is anticipated that further follow-ups will be required for clinical trials and for occupational cohorts particularly in relation to community concerns over exposure to hazardous substances. Additional large projects are planned for the linkage of census and mortality data.

The Canadian record linkage experience indicates that there is an opportunity for linkage using census, administrative and survey data for statistical purposes. This can be used to satisfy clients demand for more detailed information in ways that could not be generated in any other way. Many of the technological and software problems have now been solved, but improved timeliness of data is still an important issue. Some files are dynamic in nature and this can causes problem when the data are being used by a number of organizations.

The largest concern may relate to issues of privacy and confidentiality. Transparent policies and procedures need to put into place for organizations carrying out record linkages. There is a need for timely review of studies.

Record linkage is an area where a lot can be gained by sharing experiences between the business and social fields. Various national statistical agencies apply corresponding record linkage techniques and actively work on enhancements for record linkage. International workshops such as this one provide an opportunity to exchange views on new ideas and developments with respect to methodology and software development. Statistics Canada can benefit from the exchange of information held at this workshop where many European statistical agencies are represented.

7 Acknowledgements

The author gratefully acknowledges the assistance of Pierre Lalonde, Ted Hill, Evelyn Perkins, Mike Wenzowski, Brad Thomas, Patricia Whitridge and Stephanie Holland in the preparation of this report.

References

J.P. Ashmore, D. Krewski, J. M. Zielinski, H. Jiang, R. Semenciw, and P.R. Band. First analysis of mortality and occupational radiation exposure based on the National Dose Registry of Canada. *American Journal of Epidemiology*. 148 (6): 564-574.

M.J. Colledge. Frames and business registers: an overview. In B.G. Cox, D.A Binder, B. N. Chinnappa, A. Christianson, M.J. Colledge, P.S. Scott (eds.) *Business Survey Methods*.1995 Wiley -Interscience Publication Toronto, pages 21-47.

H.L. Dunn. Record linkage: *Am J Public Health*. 36:1412-1416, 1946.

Y. DeGuire. Postal address analysis. *Survey Methodology*.14: 317-325, 1988.

M.E. Fair. Record linkage in an information age society. In: W. Alvey and B. Jamerson, editors, *Record Linkage Techniques – 1997. Proceedings of an International Workshop and Exposition*, pages 427-441. Office of Management and Budget, Washington, 1997. Available at Internet website http://www.fcsm.gov/ working-papers/RLT-1997.html chapter 11.

M.E. Fair and P. Whitridge. Tutorial on record linkage. In: W. Alvey and B. Jamerson, editors, *Record Linkage Techniques - 1997. Proceedings of an International Workshop and Exposition*, pages 455-479. Office of Management and Budget, Washington, 1997. Available at Internet website http://www.fcsm.gov/working-papers/RLT-1997.html, chapter 12.

G.R. Howe and J. Lindsay. A Generalized Iterative Record Linkage System for use in medical follow-up studies. *Computers and Biomedical Research*. 14, 327-340.

I.P. Fellegi and A.B. Sunter. A theory of record linkage. *JASA*. 64:1183-1210, 1969.

I.P. Fellegi. Record linkage and public policy – a dynamic evolution. In: W. Alvey and B. Jamerson, editors, *Record Linkage Techniques – 1997. Proceedings of an International Workshop and Exposition*. Pages 3 – 12. Office of Management and Budget, Washington, 1997. Available at Internet website http://www.fcsm.gov/ working-papers/RLT-1997.html chapter 1.

Health Canada. *Canadian Perinatal Health Report, 2003*. Ottawa: Minister of Public Works and Government Services Canada, 2003. Available at Internet website: http://www.hc-sc.gc.ca/pphb-dgspsp/publicat/cphr-rspc03/

A.B. Miller, C. J. Baines, T. To and C. Wall. Canadian National Breast Screening Study: 1. Breast cancer detection and death rates among women aged 40 to 49 years. *Canadian Medical Association Journal*. 147: 1459-1476.

A.B. Miller, C. J. Baines, T. To and C. Wall. Canadian National Breast Screening Study: 2. Breast cancer detection and death rates among women aged 50 to 59 years. *Canadian Medical Association Journal*. 147: 1477 – 1488.

H.B. Newcombe, J.M. Kennedy, S.J. Axford and A.P. James. Automatic linkage of vital records. *Science*. 130: 954-959, 1959.

H.B. Newcombe. *Handbook of Record Linkage: Methods for Health and Social Studies, Administration and Business*. Oxford, Oxford University Press, Oxford, U.K.,1988.

M.E. Smith and J. Silins. Generalized Iterative Record Linkage System. In: *Proceedings of the American Statistical Association*. pages 128-137. ASA, Social Statistics Section, 1981.

W.N. Sont, J.M. Zielinski, J.P. Ashmore, H. Jiang, D. Krewski, M.E. Fair, P.R. Band and E.G. Létourneau. First analysis of cancer incidence and occupational radiation exposure based on the National Dose Registry of Canada. *American Journal of Epidemiology* 153 (4):309-318, 2001.

Statistics Canada. *Record linkage Software Evaluation Study Phase 1 Report*. Occupational and Environmental Health Research Section, Health Statistics Division, Ottawa 1998.

Statistics Canada. *Business register*. A detailed description of the Business register, its data sources and methodology, data accuracy and documentation are available on the Statistics Canada website at http://www.statcan.ca.

W. E. Winkler. Matching and record linkage. In B.G. Cox, D.A Binder, B. N. Chinnappa, A. Christianson, M.J. Colledge, P.S. Scott (eds.) *Business Survey Methods*.1995 Wiley -Interscience Publication Toronto, pages 355-384.

Author's address:

Martha Fair
Chief,
Occupational and Environmental
Health Research Section
Health Statistics Division
Main Building, Room 2200, Section G
Statistics Canada
Tunney's Pasture
Ottawa, Ontario Canada K1A 0T6

Tel. +1 613 951 / 1734
Fax +1 613 951 / 0792
Elec. Mail: martha.fair@statcan.ca
http://www.statcan.ca

Contact Information (after January 14, 2004)
3 Carter Court,
Tottenham,
Ontario
L0G 1W0
Canada

Tel. +1 905 936 3167
Elec. Mail: fairmj@magma.ca