

## Record Linkage in an Information Age Society

*Martha E. Fair, Statistics Canada*

---

### **Abstract**

*As we move into the 21st century the acquisition, generation, distribution, and application of statistical knowledge in a timely fashion will become more important. Required are innovations in terms of the products, technologies, and the way in which we generate, disseminate, and use statistical data and information. It is anticipated that work units will shrink, funding will be limited, and there will be greater analytical uses of administrative, as well as survey and census data. There may need to be a fundamental rethinking and radical redesign of business processes and workplaces. Today's market, customer values and technologies are changing rapidly. Standards, cooperation and collaboration of various agencies, and software developments are very important. Access and control of sensitive information as well as the technical aspects of confidentiality are necessary. Data integration of a number of different sources, including census, survey, registry and administrative files in a variety of economic and social areas are sometimes required. The quality of the statistical information is also of concern.*

*One useful tool that has been developed for generating and using statistical data is computerized record linkage. Anticipated new developments and applications of this methodology for the 21st century are described. Emphasis is placed on the health area, particularly in these times of health reform.*

*Over the past 15 years, generalized systems have been developed at Statistics Canada. Briefly described is a new version of a generalized record linkage system (GRLS.V3) that is being put into place to carry out internal and two-file linkages. With an earlier mainframe system, large-scale death and cancer linkages for members of survey and other cohorts have been shown to be practicable using the Canadian Mortality Data Base, the Canadian Cancer Data Base and the Canadian Birth Data Base. This approach has greatly reduced respondent burden, lowered survey costs, and greatly refined the detection and measurements of differences in geographic, socio-economic and occupational groups. Some of the past successes are described, particularly where longitudinal follow-up and creation of new sampling frames are required. For example, the Nutrition Canada Survey, the Canada Health Survey and Fitness Canada Surveys have been linked with mortality data. Some examples of the use of follow-up of census data are discussed (e.g., a study of farmers using 1971 Census of Agriculture and Census of Population).*

---

This paper was reprinted with permission from the *Proceedings of the Census Bureau's Conference and Technology Interchange*, March 17-21, 1996.

Introduction -- Statistical Data Needs for the 21st Century

---

## Purpose

The purpose of this article is to discuss some of the issues surrounding **statistical uses of record linkage**, with a view to the expanded uses of **probabilistic record linkage** in the 21st century, particularly with respect to the generation and use of administrative and survey data. Record linkage is the bringing together of two or more pieces of information relating to the same entity (e.g., individual, family, business). In probabilistic record linkage, the comparison or matching algorithm yields for each record pair, a probability or “weight” which indicates the likelihood that record pairs relate to the same entity (Fair, 1995).

In the 21st century, it is anticipated that those carrying out and requiring record linkage of data should be prepared for change. Hardware and software needs for record linkage will range from global statistical systems for giant organizations on large super computers, to requests for linkages of small area data sets on small laptops. Integration of a variety of statistical survey and administrative data sources may be required. There is a move to reduce the complexity of data, to avoid unnecessarily duplicating data, and to have a single, unified view of an organization’s information, with the data’s physical location being almost transparent to the user. There is considerable re-engineering of data acquisition processes, including the editing, manipulating and grouping of files. This should improve the quality of the input files. Data models may be centered around the same individual, family or entity over time rather than a cross-sectional snapshot of an event. It is anticipated that databases will become more comprehensive and inclusive. There will be a need to develop and revise international data standards, such as for disease, geographic, industrial coding, and data exchange. Timeliness is important with many organizations moving to electronic data capture and optical imaging. Dissemination of products will be via a spectrum of medium, with emphasis on the usefulness to the customer. On-line access may be required for inquiry, downloading and reporting. New links between agencies and countries may be required, and hence confidentiality issues will be of prime importance. Here, it is useful that statistical and administrative record linkage applications be differentiated.

Today, we will examine some **general topics** first, namely:

- evolving in response to customer needs in changing times;
- some comments regarding the “information age;”
- characteristics/indicators of success for an effective statistical system; and
- moving from data to information.

We will then look at **record linkage** in more depth and examine:

- today’s situation;
- examples of present uses of record linkage;
- preparing for the future journey -- the life cycle of events;
- making the right connection; and
- summary.

I will use examples of statistical applications of record linkage, with emphasis on those from Statistics Canada and the health research area in particular.

## Evolving in Response to Customer Needs in Times of Change

Many of the common social, economic, occupational and environmental concerns of today are complex and multi-faceted. **Change** seems to have become the operative word. **Policies, institutions, communities and businesses** are changing at the global, national, provincial/state, regional and local levels. Institutions in North America and worldwide have undergone an unprecedented wave of consolidation. There is concern to identify and strive toward **global statistical systems** that can produce national statistical services that are comparable and readily accessible (Haberman, 1995). The capabilities of **technology**, especially communication and information technology are changing daily.

The **tools and options for dissemination** are expanding. There is a corresponding rising consumer expectation, particularly with respect to timeliness and quality of statistical data products. This has implications in terms of standard data concepts, definitions, coding, methodology used for record linkage, and development of national and provincial/state data bases. Communication and collaboration of various countries, particularly with respect to software and methodological developments, have benefited through a series of seven workshops regarding record linkage held in Canada (e.g., Carpenter and Fair, 1989) and others held in the United States (Kilss and Alvey, 1985).

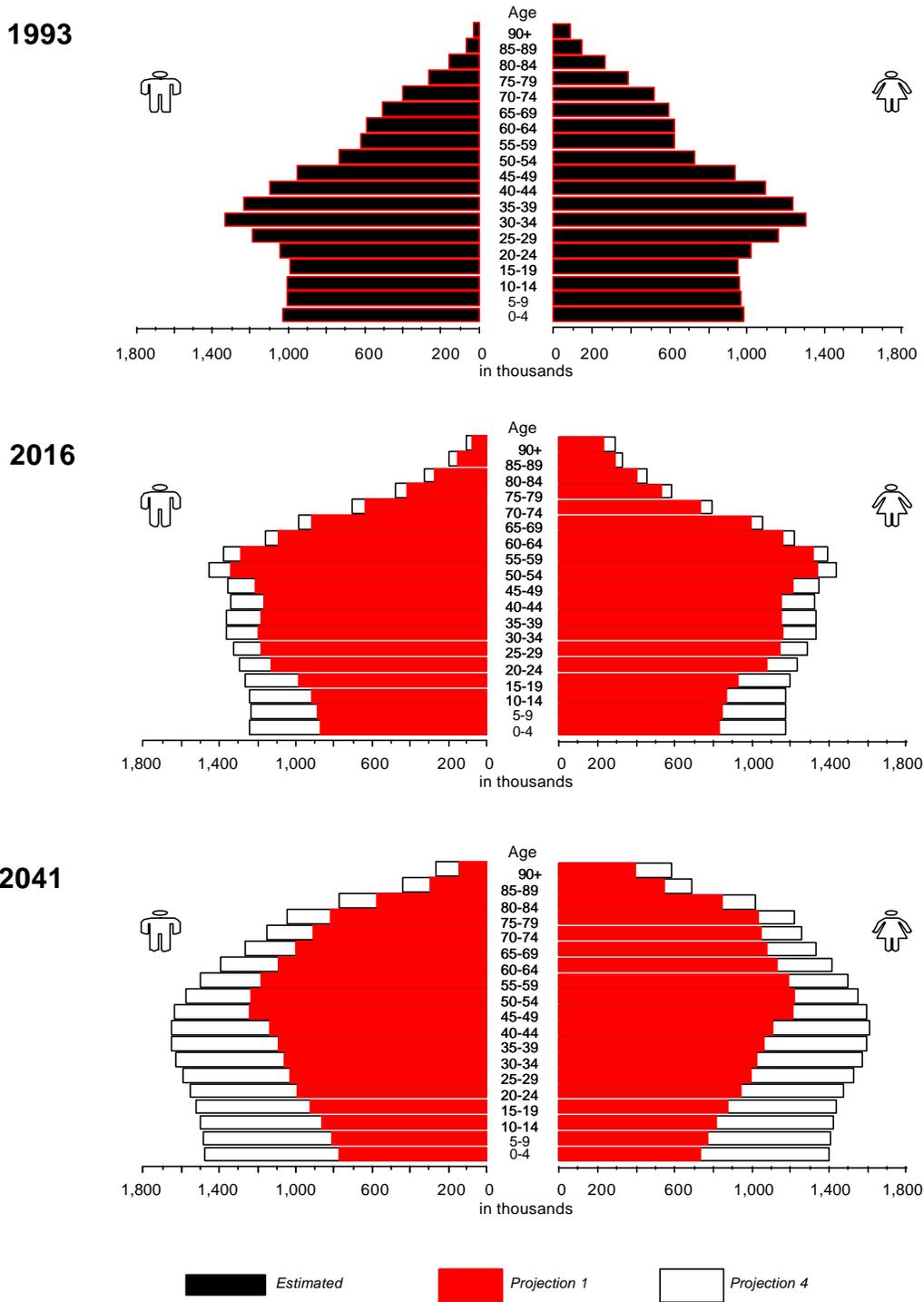
**Analysis** of data sources from different countries is helpful and comparative international statistics are required. Joint analysis of data from different countries is common (e.g., a joint analysis of 11 underground miners studies to examine radon and lung cancer risks). There is a need for **international collaborative works**, such as the United Nations Scientific Committee on the Effects of Atomic Radiation, which aims to provide to the scientific and world community, its latest evaluations of the sources of ionizing radiation and the effects of exposures (United Nations, 1993). Here the major aim is to assess the consequences to human health of a wide range of doses of ionizing radiation and to estimate the dose of people all over the world from natural and man-made radiation sources. Linkage of a variety of data sources are required.

The **social and economic structure** is changing. There are **new concepts of family, childhood and parenthood**. This has important implications for the follow-up of households and individuals for longitudinal surveys and for administrative files. *The Health of Canada's Children - A CICH Profile* discusses some of the recent trends in Canada (Canadian Institute of Child Health, 1994). Some of the examples given are as follows. **Families are changing** -- the structure of the families are different from what they used to be. In 1967, 65% of all Canadian families consisted of a male wage earner and a stay-at-home spouse. In 1990, this traditional family structure accounted for only 15% of families.

Our **society is becoming more diverse**. Families are rooted in more diverse cultural, religious, linguistic and ethnic backgrounds than in the past. In 1991, 13% of the Canadian population spoke a language other than French or English at home. Where surnames and forenames are used in probabilistic matching, we have found that special tables of weights have had to be developed by region, and sometimes over time. For example, there is quite a different distribution of name frequencies in Quebec, which is predominantly French, in British Columbia, where the number of Asian names have increased in recent years, and in Canada overall. Naming conventions are changing, with women often retaining their maiden name, as is particularly common in Quebec.

The **role of women** has changed. In the early 1960's less than a third of Canadian women worked outside the home. By 1996, about 80% of women are expected to be in the work force. **Lifelong learning has become a necessity**. Increasingly, the workplace requires a higher level of skills and a different set of skills than in the past. More and more jobs require people who can work in teams, who have high literacy, numeracy and computing skills, who can then critically and creatively solve problems -- and most of all, continue to learn new skills.

Figure 1. -- Population by Age Group and Sex, Canada, 1993, 2016 and 2041



Source: George, M.V.; Norris, M.J.; Nault, F.; Loh, S.; and Dai, S.Y. (1994), *Population Projections for Canada, Provinces and Territories 1993-2016*, Statistics Canada Catalogue No. 91-520, page 74.

Between 1971 and 1991, the **age profile** of the population changed from a traditional pyramid shape to a wide column, with fewer younger people and dramatically more **older** people. By 2041 the column will be top-heavy (see Figure 1-- Source: George et al., 1994). Similarly in the United States, by 2025 more than

30% of the population will be over 55 year old. Persons aged 80 and over will outnumber any younger 5-year age group (UNDIESA, 1991).

The **economy is restructuring**. There is a heightened sense of economic anxiety. Driven by technological innovation, global competition and new trade arrangements, the economy is undergoing a fundamental restructuring. **Governments are restructuring**. At all levels, they are tightening up their spending to make programs more cost effective and more relevant to changing needs. This has been most apparent in the health sector in term of **health reform**. (Blomquist and Brown, 1994). At the same time, there are **major reforms of social programs**, not only to improve efficiency, but also to remake these programs.

## **The Information Age**

The Tofflers have described our times as being that of the third wave (Toffler and Toffler, 1995). The **First wave** was agricultural and it lasted thousands of years until the 18th century. Then the Industrial Revolution created a novel concept of massification -- mass production, mass markets, mass consumption, mass media, mass political parties, mass religion and weapons of mass destruction. This **Second wave** lasted about three hundred years. The **Third wave** is that of an **information-age society**. Because of the computer chip we are moving from an age in which we produce things to an age in which we produce information. But paradoxically, the more that national boundaries are usurped by our universal hook-up to the global computer network, the more we segment (Grant, 1996). With the complexities of the new system we require more and more information exchange among the various units of companies, government agencies, hospitals, associations, institutions and individuals. Factories, cities, even nations are receding and being replaced by smaller units of consumption and by minority political and religious interests. In the Tofflers' words, the world "de-massifies."

We are in a time of **redefining the workplace -- and work itself**. Work units are shrinking. The home may be the workplace of the future for many more people. **Customized and semi-customized, highly diversified statistical products** will be required -- yet the cost of producing these diversified products must be minimal. There is a requirement of flexibility and choice by many clients.

There is a growing **time crunch**. Time itself is one of the most important economic resources. The ability to shorten time -- by communicating swiftly or by bringing products in a timely fashion -- may mean the difference between profit and loss (Toffler and Toffler, 1995). In the health area, there is a need for more flexible, fast-paced, information-rich systems which can act as surveillance systems and assist in identifying present and emerging health issues.

We may have to **rethink and re-image our relationships**. Amidst societal change, people more than ever need an anchor, a refuge, a place where they belong (Bank of Montreal, 1995). Traditionally, a sense of **community** has helped fill that need. This in the past, was often built around a common geographic location, a common workplace, a common history or tradition. Individuals now form commitments to a wide variety of communities based on shared experiences and values -- family, profession, neighbourhood, age, ethnic background, talent, language. Barna (1990) notes that in the process of redefining what counts in life, many of us have decided that **commitment** is not in our best interest. Traditional concepts such as loyalty and the importance of memberships in various groups have been thrown out in favour of personal interest and self-preservation. This may have important implications for the workforce and for negotiations.

## **Characteristics and Indicators of an Effective Statistical System**

Dr. Fellegi, the Chief Statistician at Statistics Canada, gave a 1995 Morris Hansen Lecture at the Washington Statistical Society. He described an effective statistical system as being characterized by its ability to:

- illuminate issues, not just monitor them;
- evolve in response to needs;
- be aware of priority needs;
- set priorities;
- have a high level of public credibility, since few in society can verify national statistics; and
- be free from undue political interference (Fellegi, 1995).

Three main indicators of success of statistical systems noted in this paper were:

- How adaptable is the system in adjusting its product line to evolving needs?
- How effective is the system in exploiting existing data to meet client needs?
- How credible is the system in terms of the statistical quality of its outputs and its non-political objectivity? (Fellegi, 1995)

### **Moving from Data to Information**

Two recent methodology symposium topics held at Statistics Canada are relevant. The XIIth International Symposium on Methodological Issues, held at Statistics Canada on November 1-3, 1995, was entitled "From Data to Information." At this symposium topics included the role of statistics in making social policy, data integration, analytical methods, access and control of data, quality of statistical information, technical aspects of confidentiality, making data accessible to the general public, data warehousing, and electronic information dissemination. An earlier symposium dealt with re-engineering for statistical agencies (Statistics Canada, 1994). Re-engineering is a rethinking and radical redesign of the way business is carried out by an agency or corporation. The desired end results are lower production costs, quicker dissemination, and higher customer satisfaction.

There is a desire to understand and improve the performance of the health system. As noted in *Health Data in the Information Age -- Use, Disclosure and Privacy* (Donaldson and Lohr, 1994) this in turn motivates proposals for the creation and maintenance of comprehensive, population-based health care data bases. Regional health care databases are being established around the United States and Canada. Guidelines are needed to realize the full potential of these files, as well as to reduce respondent burden.

Two critical dimensions of databases are their **comprehensiveness and inclusiveness**. Comprehensiveness describes the completeness of the records (i.e., the amount of information one has for each patient and for an individual over time). Inclusiveness refers to which populations in a geographic area are included in a database. The more inclusive a database, the more it approaches coverage of 100 percent of the population. The Census of Population, the vital statistics and morbidity files are important data sources for a variety of national health studies because of their comprehensiveness and inclusiveness.

## Record Linkage

### Today's Situation

Just as we have just looked into the future in a more general fashion, it is also good to reflect on some of the past development of record linkage methods. Some of today's data sources were created by individuals with a view to record linkage in the future (e.g., in Canada the vital statistics birth records were linked with Family Allowance files to determine the eligibility of applicants when this program was first implemented).

The initial definition of record linkage was in terms of the book of life (Dunn, 1946). The early development work had to do with investigating the feasibility of probabilistic linkage (Newcombe et al., 1959), the theory of record linkage (Fellegi and Sunter, 1969), the development of specific computer programs, followed by the development of generalized software (Hill, 1981) and national files, commercial software (Jaro, 1995) and other software (e.g., Chad, 1993). Communication and collaboration with agencies in various provinces in Canada, in the United States, the United Kingdom and Australia have aided record linkage developmental work (e.g., Kilss and Alvey, 1985; Gill et al., 1993; Jaro, 1995; Winkler and Scheuren, 1995).

One key technological development is the shift from a paper-based system of records to an **electronic process** for creating, transmitting and disseminating products. At Statistics Canada, the 1990s brought about a major revolution in advanced technology with the wide-scale introduction of Computer Assisted Interviewing (CAI) for household, agriculture and business surveys. Computer Assisted Personal Interviewing (CAPI) has been introduced with the Labour Force Survey supplements and longitudinal household surveys covering a wide range of topics including Survey of Income and Labour Dynamics, the National Population Health Survey and the National Longitudinal Survey of Children (Gosselin, 1995). Vital statistics (Starr and Starr, 1995), census and cancer registries are additional examples where re-engineering and change may be anticipated in the future. There has been a move from microfilming of source documents to **optical imaging**.

A **generalized system** initiative at Statistics Canada was started in response to the use of repetitive processes, particularly in survey taking. This includes sampling, data collection and capture, automated coding, edit and imputation, estimation, and record linkage (Doucet, 1995). This **suite of software products** has been developed with technologies that make them highly portable across major computing platforms.

The original version of generalized record linkage software (GRLS.V1) that was developed at Statistics Canada was for a mainframe environment. Currently under development is GRLS.V3 which runs in a client-server environment with ORACLE and a C compiler (Statistics Canada, 1996). GRLS will run on a PC or workstation which supports the UNIX operating system. This software allows for an internal linkage within a file (e.g., to create health histories in a cancer registry) or a two-file linkage (e.g., linkage of a survey file to mortality). This software is particularly useful where there is no unique, reliable, lifetime identifier on the files being linked.

GRLS has three important stages:

- In the **searching stage** screens are used to specify the files, indicate the records to be compared (e.g., within pockets with similar phonetic code of the surname), specify the rules for comparison (e.g., agree, disagree, partially agreement, or user-defined functions), and specify the weights to be assigned to the outcomes.

- In the **decision stage**, the weights can be adjusted and threshold weights selected to define whether pairs are linked, possibly linked, or unlinked.
- In the final **grouping stage**, the records are brought together appropriately. You can have conflicts resolved automatically (e.g., two records linking to one death record). This is called mapping, and one can select the appropriate type (e.g., 1-1, 1-many, many-1, many-many). You may also have the option to resolve conflicts manually via on-screen updates. The final output of GRLS is an ORACLE table containing the GROUP information.

It is very important to note that GRLS V3 **does not modify** the files it is linking. This means that the same file may participate concurrently in several two-file linkages. For example, one might want to link several (and unrelated) files against the same master file.

### Record Linkage in the Toolbox of Software -- Some Examples of its Use

Statistics Canada uses a common set of software products in re-engineering its administrative and statistical programs. This set of products is collectively referred to as the toolbox. Each toolbox product has a current release, an identified support level and a designated support centre. Currently the generalized record linkage software is part of this toolbox.

Record linkage is an important tool for the creation of statistical data, particularly in relation to census taking. Some of the important uses are as follows:

- **Data Quality.**--Some European countries use population registers instead of a census (e.g., Denmark). It is also possible to use administrative data and record linkage to help impute missing or inconsistent data. Data sources can be examined to eliminate duplicate records for individuals and to identify missing records in databases (e.g., by the linkage of infant deaths and birth records or by the linkage of births and deaths with census records).
- **Bias.**--The advantage of population-based record linkage includes the avoidance of selection bias, which can occur in cohort and case-control studies. Recall bias is usually avoided because the data are collected before the outcome or in ignorance of the outcome.
- **Coverage.**--In Canada record linkage data is used to improve the census coverage (e.g., address register) as well as to estimate its coverage (e.g., reverse record check). With disease-specific registries, it is possible to use linkage to identify underreporting of cases (e.g., by linkage of cancer registries with death registrations, the linkage of hospital records with deaths for heart disease). This has important implications for diseases such as AIDS and cancer.
- **Tracing Tool.**--Record linkage and administrative records are often used to follow-up cohorts to determine the individuals' vital status. Tracing is often needed for follow-up of industrial cohorts and for longitudinal surveys to obtain the cause of death and/or cancer. Mobility patterns of persons are important for the allocation of health resources.
- **Benchmarking/Calibration.**--Combining results from several data collection sources may give improved estimates (e.g., use of income from tax, survey and census sources).
- **Sampling Frame.**--Record linkage may be involved in setting up a sampling frame for surveys (e.g., census of agriculture farm register is used for the sampling of intercensal farm surveys).

- **Supplementary Surveys.**--Several postcensal surveys have been carried out following the Canadian census. Examples include the aboriginal peoples, and the health and activity limitation surveys. Data from the survey can be linked with that available on the census.
- **Release of Public-Use Tapes.**--Linkage can be used to examine public-use tapes for potential problems in their release (e.g., data crackers).
- **Building New Data Sources (e.g., Registries).**--Some cancer registries combine a variety of data sources using record linkage to generate their registry. Some of the data sources include hospital admissions, pathology reports, records from clinics, and death registrations.
- **Creation of Patient-Oriented, Rather than Event-Oriented Statistics.**-- (e.g., for hospital admissions, for cancer registries, (Dale, 1989)).

The uses of linkage in analytical studies have often been varied, and are generally tied in with increased use of administrative records for statistical purposes and with the reduction of respondent burden. (A roundtable luncheon of the Social Statistics Section at the 1995 American Statistical Association, chaired by G. Hole, discussed some of the above and future uses of administrative records to complement/ supplement data from household surveys.)

A more complete list of some of the uses of record linkage have been described earlier (Fair, 1995; Newcombe, 1994). Some examples are as follow:

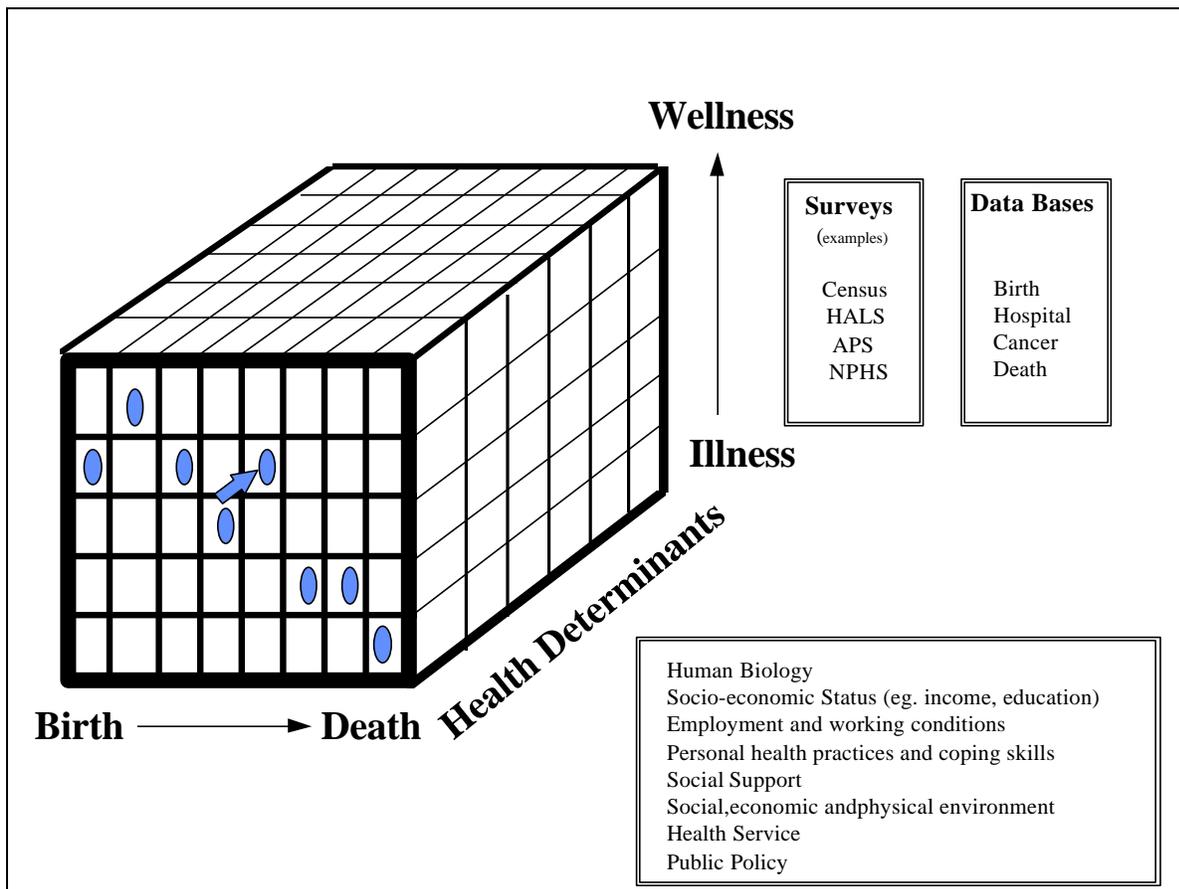
- Mortality, cancer and/or birth follow-up of
  - cohorts (e.g., miners, asbestos workers)
  - case/control studies
  - clinical trials (e.g., Canadian Breast Screening study);
- Building, maintaining and using registries (e.g., cancer and AIDS);
- Creation of patient-oriented histories;
- Follow-up of surveys (e.g., Nutrition Canada, Canada Health Survey, Fitness Canada);
- Occupational and environmental health studies;
- Examining factors which influence health care usage and costs; and
- Regional variations in the incidence of disease.

A longitudinal National Population Health Survey is currently in progress in Canada. In the original survey approximately 95% of the respondents agreed to have their survey data linked to their provincial health records. This linkage will strongly enhance the data set's potential usefulness because it will add respondents' interaction with the health care system.

## Preparing for the Journey Ahead -- The Life Cycle of Events

In the health area, this usually involves the linkage of various data sources over time. Figure 2 is an example of how we can view the **life cycle** of events from birth to death, **health determinants**, and **outcomes in ranges of “illness” to “wellness.”** Piecing together the various important components may involve gathering data from a number of different sources such as surveys to estimate the degree of “illness” or “wellness” of the population (e.g., Census, Health and Activity Limitation Survey (HALS), Aboriginal Peoples Survey (APS), National Population Health Survey (NPHS)), national databases of existing administrative records (e.g., Canadian Birth Data Base, the Canadian Cancer Data Base and the Canadian Mortality Data Base), and from a number of different perspectives. For example, within health determinants one may be interested in human biology, socio-economic status (e.g., income, education), employment and working conditions, personal health practices and coping skills, social support, social, economic and physical environment, health services, and public policy. As the population/individual progresses through the different stages of the life cycle, the degree of “wellness” can vary as indicated in the diagram. (See also Hertzman, Frank and Evans, 1994).

Figure 2.--The Life Cycle of Events



HALS -- Health and Activity Limitation Survey  
 APS -- Aboriginal Peoples Survey  
 NPHS -- National Population Health Survey

Some examples, involving the use of census data, are as follows:

- **Maternal Health and Infant Birth - Death Linkages.**--A study of regional differences in perinatal and infant mortality in the province of Ontario has been carried out. Infant and perinatal mortality in the 53 counties of Ontario were studied in two time periods -- 1970-72 and 1978-79. A considerable regional variation in the range of rates was found. Socio-economic factors were found to have an important influence on the maternal and infant determinants of mortality and in this way contributed to the variations in mortality over the province. Recently, there has been interest in establishing a Canadian Perinatal Health Surveillance system.
  
- **Occupational Studies.**--There are strong pressures from society to determine and reveal the health risks to which it is exposed, especially where the harm is cumulative or latent for an extended period of time. These pressures come from three main sources. **Organized labour** has a special interest in conditions in the workplace which might lead to delayed effects, such as cancers among its members. Both the **general public and environmental groups** have frequently expressed concern over the possible consequences of exposure of the population at large to chemical and other agents. These agents are being produced in increasing numbers and quantities, and distributed both as commercial products and as contaminant wastes in ways that may result in ingestion or inhalation. The third source of pressure originates with **professional groups** whose work involves them in the detection and measurement of health risks and in setting safety regulations. Cancer incidence and mortality data are a main source of information to assist in the determination of health risks.

In light of urgent demands to protect workers' health, there is a need for a broad-based occupation-cancer database containing information on both cancer incidence and a wide range of occupations. A current feasibility study is examining the possibility of linking cancer, mortality and occupational, household and socio-economic data derived from the 1986 census data. The sample, consisting of seven geographic regions (4 urban and 3 rural), were selected based on census geography.

As an occupational group, farmers have low overall mortality. However, a number of epidemiological studies suggest increased risk of certain cancers among farmers, including cancer of the stomach, lip, prostate, brain and skin, leukemia, Hodgkin's disease, multiple myeloma, and non-Hodgkin's disease.

A mortality and cancer cohort study of about 326,000 Canadian male farm operators enumerated in the 1971 Census of Agriculture has been carried out in collaboration with Health Canada (Fair, 1993). Seven major files were linked to create the data required for the analysis file in this study, namely:

- the 1971 Census of Agriculture;
- the 1971 Census of Population;
- the 1971 Central Farm Register;
- the 1981 Central farm Register;
- the Canadian Mortality Data Base;
- the 1966-71-76-81-86 Census of Agriculture Longitudinal file; and
- the Canadian Cancer Data Base.

Analyses of these data have examined prostate (Morrison et al., 1993) and brain cancer (Morrison et al., 1992) in particular.

- **Socio-Economic Gradients.**--There has been an increasing awareness of the importance of supporting basic research designed to identify determinants of health in order to inform policy makers about how best to improve the population's health and how best to accomplish this goal efficiently and cost-effectively. As a result, the Manitoba Centre for Health Policy and Evaluation has collaborated with Statistics Canada to determine the feasibility of linking provincial administrative health care utilization with census data for a sample of Manitobans (Mustard et al., 1995).

Mortality and health services utilization have been described in relation to the socio-economic status measure, mortality and the use of health care services at seven different stages in the life course (ages 0-4, 5-14, 15-29, 30-49, 50-64, 65-74, 75+). The objective of the study was to identify those classes of morbidity which dominate utilization of health care services at each stage of life course and simultaneously, the classes of morbidity which show the greatest disparities in relation to socio-economic status. The research resource of this project was created at a fraction of the cost of a population survey. Some of the public policy responses indicated by these data were:

- to consider directing an even greater share of health care services to lower socio-economic groups;
- to more aggressively target preventive medical and health services, especially in early adulthood; and
- to formulate explicit public policies addressing health inequalities. (Mustard et al., 1995, p. 67).

## Making the Right Connections and Summary

We are in a time of rapid **changes** in terms of **markets, customer expectations** and technologies for record linkage **software** development, **hardware**, and **applications**. There often needs to be an optimal balance between cost, quality and timeliness. Many of the existing **data systems** are on the threshold of change. There is a shift from single data base applications to electronic data transfer and warehousing, data sharing within broad subject matter areas, and to enterprise wide systems and **data integration**. There are various hardware and software environments being used. A variety of approaches can be used to assess user's needs. These include professional advisory committees, client-oriented program evaluations, interactions with professional and other associations, market feedback, and analytic programs.

One needs to have the capacity to acquire, generate, distribute and apply knowledge strategically and operationally (Toffler and Toffler, 1995). To a large extent the quality of record linkage in the future is dependent on the **quality of the files** being linked -- quality in/quality out. There is a need to harmonize concepts and outputs. For example, it is anticipated that the Tenth International Revision of the Classification of Disease will be implemented. A restructured industry classification system known as the North American Industry Classification System is being developed. **Uniform lifetime** business and individual **numbers** are highly desirable for many of the new information systems. Further work is required in designing appropriate items for the data sets -- for example, more detail may be available at the local level than on a national basis.

There is a need to **integrate** a number of different sources of data. As governments and agencies regionalize services, there are additional requests for **small area data**. It is important to have the capacity to use multiple definitions of geographic population areas of interest (e.g., enumeration area, postal code areas, school districts, health units) depending on the nature of the investigation.

There is a need to develop **confidentiality procedures** and screening rules for the generation and release of public use data files. All studies involving record linkage at Statistics Canada must satisfy a prescribed review and approval process. For example, the purpose of the record linkage activity must be statistical or research in nature and must be consistent with the mandate of Statistics Canada as described in the Statistics Act. The record linkage activity must have demonstrable cost or respondent burden savings over

other alternatives, or be the only feasible option. It must also be shown to be in the public interest. A comprehensive list of recommendations for Federal statistics agencies in the United States is given in Duncan et al., 1993.

In conclusion, **analysis** of existing and future linked data sets is indispensable in illuminating the main social and economic issues we face not only today, but also into the future. We need to anticipate and look forward to issues of the 21st century where record linkage may serve as an important research tool.

## References

- Bank of Montreal (1995). *Bank of Montreal 178th Annual Report 1995*, Public Affairs Department of the Bank, Bank of Montreal Tower, 55 Bloor Street West, 4th Floor, Toronto, Ontario M4W 3N5.
- Barna, G. (1990). *The Frog in the Kettle: What Christians Need to Know About Life in the Year 2000*, Regal Books, Ventura, California 93006.
- Blomquist, A. and Brown D. M. (Eds.) (1994). *Limits to Care: Reforming Canada's Health Care System in an Age of Restraint*. Available from: Renouf Publishing Company Limited, 1294 Algoma Road, Ottawa, Ontario K1B 3W8.
- Canadian Institute of Child Health (1994). *The Health of Canada's Children - A CICH Profile 2nd Profile*. Available from: Canadian Institute of Child Health, 885 Meadowlands Drive, Suite 512 Ottawa, Ontario K2C 3N2.
- Carpenter, M., and Fair, M.E. (Eds.) (1989). *Canadian Epidemiology Research Conference -- Proceedings of the Record Linkage Session and Workshop*. Available from: Statistics Canada, Occupational and Environmental Health Research Section, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6.
- Chad, R. (1993). *A Comparison of Three Different Computer Matches*. Special Census/Administrative Record Match Working Group in Conjunction with the Year 2000 Researcher Development Staff, U.S. Bureau of the Census, Washington, DC, September 1993, (Matchers -- Winkler, Slaven, Jaro).
- Dale, D. (1989). Linkage As Part of a Production System. The Ontario Cancer Registry, in *Canadian Epidemiology Research Conference -- Proceedings of the Record Linkage Sessions and Workshop*, M. Carpenter and M.E. Fair, Eds. Available from: Statistics Canada, Occupational and Environmental Health Research Section, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6.
- Donaldson, M.S., and Lohr, K.N. (Eds.) (1994). *Health Data in the Information Age -- Use, Disclosure, and Privacy*, Washington, D. C.: National Academy Press.
- Doucet, E. (1995). Survey Re-Engineering: Is Our Information Technology Framework Up to It? *Proceedings of Statistics Canada Symposium 94 -- Re-Engineering for Statistical Agencies*, November 1994, Available from: Financial Operations Division, Statistics Canada, R.H. Coats Building, 6th Floor, Tunney's Pasture, Ottawa, Ontario K1A 0T6, P. 159-168.
- Duncan, G.T.; Jabine, T.B.; and De Wolf, V.A. (Eds.) (1993). *Private Lives and Public Policies -- Confidentiality and Accessibility of Government Statistics*, Washington, D.C.: National Academy Press.
- Dunn, H.L. (1946). Record Linkage, *American Journal of Public Health*, 36, 1412-1416.

- Fair, M.E. (1995). An Overview of Record Linkage in Canada, *1995 Proceedings of the Social Statistics Section of the American Statistical Association*, American Statistical Association, 1429 Duke Street, Alexandria, Virginia 22314-3402, 25-33.
- Fair, M.E. (1993). Recent Advances in Matching and Record Linkage from a Study of Canadian Farm Operators and Their Farming Practices, *1993 ICES Proceedings of the International Conference of Establishment Surveys*, American Statistical Association, 1429 Duke Street, Alexandria, Virginia 22314-3402, 600-605.
- Fellegi, I.P. (1995). Characteristics of an Effective Statistical System, Morris Hansen Lecture, presented at the Washington Statistical Society, October 25, 1995.
- Fellegi, I.P. and Sunter, A.B. (1969). A Theory of Record Linkage, *Journal of the American Statistical Association*, 40, 1183-1210.
- George, M.V.; Norris, M.J.; Nault, F.; Loh, S.; and Dai, S.Y. (1994). *Population, Projections for Canada, Provinces and Territories 1993-2016*, Statistics Canada, Demography Division, Catalogue No. 91-520. Available from: Marketing Division, Sales and Services, Statistics Canada, Ottawa, K1A 0T6.
- Gill, L.; Goldacre, M.; Simmons, H.; Bettley, G.; and Griffith, M. (1993). Computerized Linking of Medical Records: Methodological Guidelines, *Journal of Epidemiology and Comm. Health*, 47:4, 316-319.
- Gosselin, J. F. (1995). The Operational Framework at Statistics Canada, *Proceedings of Statistics Canada Symposium '94 -- Re-Engineering for Statistical Agencies*, November 1994. Available from: Financial Operations Division, R.H. Coats Building, 6th Floor, Tunney's Pasture, Ottawa, Ontario K1A 0T6, 170-174.
- Grant, Linda (of *The Guardian*) (1996). Riding the Wave, *The Ottawa Citizen*, January 20, 1996, B4.
- Haberman, H. (1995). Towards a Global Statistical System, *Proceedings of Statistics Canada Symposium 94 -- Re-Engineering for Statistical Agencies*, November 1994. Available from: Financial Operations Division, R.H. Coats Building, 6th Floor, Tunney's Pasture, Ottawa, Ontario K1A 0T6, 53-60.
- Hertzman, C.; Frank, J.; and Evans, R.G. (1994). Heterogeneities in Health Status and the Determinants of Population Health, *Why Are Some People Healthy and Others Not? The Determinants of Health of Populations*, R.G. Evans; L. Barer; and T.M. Marmor, Eds. New York: Aldine De Gruyter, 74f.
- Hill, T. (1981). Generalized Iterative Record Linkage System, Ottawa, Canada: Statistics Canada.
- Jaro, M.A. (1995). Probabilistic Linkage of Large Public Health Data Files, *Statistics in Medicine*, 14, 491-498.
- Kilss, B., and Alvey, W. (Eds.) (1985). *Record Linkage Techniques -- 1985. Proceedings of the Workshop on Exact Matching Methodologies*, Arlington, Virginia, May 9-10, 1985, Washington, DC: Department of Treasury, Internal Revenue Service.
- Morrison, H.; Savitz, D.; Semenciw, R.; Hulka, B.; Mao, Y.; Morison, D.; and Wigle, D. (1993). Farming and Prostate Cancer Mortality, *American Journal of Epidemiology*, 137, 270-280.
- Morrison, H.I.; Semenciw, R.M.; Morison, D.; Magwood, S.; and Mao, Y. (1992). Brain Cancer and Farming in Western Canada, *Neuroepidemiology*, 11, 267-276.

- Mustard, C.; Derksen, S.; Berthelot, J.M.; Wolfson, M.; Roos, L.L.; and Carriere, K.S. (1995). *Socio-economic Gradients in Mortality and the Use of Health Care Services at Different Stages in the Life Course*, Manitoba Centre for Health Policy and Evaluation, Department of Community Health Sciences, Faculty of Medicine, University of Manitoba.
- Newcombe, H.B. (1994). Cohorts and Privacy, *Cancer Causes and Control*, 5, 287-292.
- Newcombe, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Social Studies, Administration and Business*, Oxford, U.K.: Oxford University Press.
- Newcombe, H. B.; Kennedy, J. M.; Axford, S. J.; and James, A. P. (1959). Automatic Linkage of Vital Records, *Science*, 130, 3381, 954-959.
- Starr, P. and Starr, S. (1995). Reinventing Vital Statistics, The Impact of Changes in Information Technology, Welfare Policy and Health Care, *Public Health Reports*, 110, 535-544.
- Statistics Canada (1996). *Generalized Record Linkage System Concepts*, Draft version dated 1996 February 14, Research and General Systems Development Division, Ottawa, K1A 0T6.
- Statistics Canada (1994). *Symposium '94 Re-engineering for Statistical Agencies*, Catalogue No. 11-522E, Occasional -- November 1994. Available from: Marketing Division, Sales and Service, Statistics Canada, Ottawa, K1A 0T6.
- Toffler A. and Toffler H. (1995). *Creating A New Generation -- The Politics of the Third Wave*, Atlanta: Turner Publishing, Inc.
- United Nations (1993). *Sources and Effects of Ionizing Radiation -- United Nations Scientific Committee on the Effects of Atomic Radiation*, New York: United Nations Publication, United Nations.
- United Nations Department of International Economic and Social Affairs (UNDIESA) (1991). *The Sex and Age Distribution of Population*, ST/ESA/ SER. A/122, New York.
- Winkler, W. and Scheuren, F. (1995). Linking Data to Create Information, *Proceedings of Statistics Canada Symposium '95 -- From Data to Information -- Methods and Systems*, November 1995, Statistics Canada, Ottawa K1A 0T6 (in press).