An Overview of Record Linkage in Canada

Martha E. Fair, Statistics Canada

Canada's health care system is best described as an interlocking set of 10 provincial and 2 territorial health insurance schemes. Each is universal and publicly funded. As part of the national health care system, each provincial and territorial scheme is linked through adherence to national standards set at the Federal level. Provincial and territorial hospital and medical care insurance plans are Federally regulated in areas such as the financial extent of coverage, accessibility, portability of benefits, and the requirement that the plans be non-profit.

The 1984 Canada Health Act defines the criteria and conditions that each provincial health insurance plan must meet to receive full Federal contributions. The Act discourages direct charges to patients for physician and hospital services. The five legislated principles of the Canada Health Act include: accessibility, comprehensiveness, public administration, portability, and universality. The Act is built on the Hospital Insurance and Diagnostic Service Act (1957) and the Medical Care Act (1966-67). The Hospital Insurance and Diagnostic Services Act had made a range of hospital and diagnostic services available at little or no direct cost to the patient. The purposes of these acts were to increase the supply of health personnel and facilities and to make the services available regardless of socio-economic circumstances and geography. The Medical Care Act required the Federal government to make payments to provinces and territories operating medical care insurance plans.

How the provinces organize, finance and administer their health insurance plans varies. Some administer their plans directly through provincial health departments; other plans are managed by separate public agencies reporting to the provincial health minister. Some provinces have two plans, one controlled by the province and another by a public agency (Minister of Industry, Science and Technology, 1993).

A recent article has presented an overview of health care systems in Canada and selected OECD countries (Australia, France, Finland, Germany, Sweden, the United States, and the United Kingdom). It discusses the organization of the health care systems, health care expenditure, the availability and utilization of health services, and the health status of the population (Nair and Karim, 1993).

A publication by the World Bank (World Bank, 1984) points out that the ultimate goal of public policy is to improve living standards, to increase individual choice, and to create conditions that enable people to realize their potential.

It has been stated that improvements in longevity and health policy in particular are very much a success story in Canada; life expectancy has progressed steadily and there is only a handful of countries with better health conditions than Canada (Beaujot, 1991). However, certain issues are subject to continued discussion, such as cost, the best curative and preventive approaches, better methods for the quality of prolonged life, and the delayed effects of occupational, lifestyle, socio-economic, and environmental factors on health. Health policy should take into account the demographics of health, morbidity, and mortality.

In Canada neither the Federal nor provincial authorities nor researchers have yet made more than a modest start at tapping the immense data resources and statistical tools at their disposal for health services management and research in areas such as health care, health status, epidemiology, demography, genetics, occupational, and environmental health. The desire to understand and improve the performance of the health care system has been evident in other countries, such as the United States (Donaldson and Lohr, 1994) and the United Kingdom (Gill et al., 1993).

One useful tool that has been developed over the years in Canada is computerized record linkage. This paper will briefly consider the definition, history, methods, and uses of record linkage, drawing on a variety of experiences in Canada. Probabilistic methods, generalized systems, provincial and national files have been developed to help bring together records relating to the same individual (e.g., to create patient histories) and to carry out two-file linkages (e.g., linkage of breast screening and mortality records). Several examples will be described that use national files at Statistics Canada, which is Canada's centralized statistical agency. Emphasis will be placed on use of this tool for the administrative and statistical data needs for the development of health care policies.

■ The Definition of Record Linkage

Record linkage is simply the bringing together of information from one or more independent source records that are believed to relate to the same individual, family, or entity. With successive linkages, the information may take on the characteristics of a collection of personal or family histories.

The term "record linkage" was first used in 1946 by H.L. Dunn, Chief of the United States National Bureau of Vital Statistics (Dunn, 1946). Dunn introduced the new term to a group of Canadian vital statisticians in this way: "Each person in the world creates a Book of Life. This Book starts with birth and ends with death. Its pages are made up of the records of the principal events in life. Record linkage is the name given to the process of assembling the pages of this Book into a volume." Dunn envisaged the assembly of records of the principle events of life into a personal "Book of Life" over a lifetime, with linkage implying records brought together by means of common identification data (for example, birth number), which ensure that records are assigned to the correct file. He envisaged that numerous national, state, and local official organizations would rely heavily on knowing certain chapters in the "Life Records Index." He applied the record linkage principle to many social and economic services and benefits so that they might be better directed at the real needs of people. About the time when this speech was given, Canada was introducing a national scheme of Family allowance, whereby the Federal Government was to pay a monthly allowance on behalf of each child. Some 3.5 million children were eligible for claims and verification of the fact of birth was required by linkage with birth records.

■ The History of Record Linkage

The process of systemizing the approach to computerized record linkage was first undertaken by the geneticist Howard Newcombe and his associates at Atomic Energy of Canada (Newcombe et al., 1959). He recognized the full implication of extending the principle to the arrangement of personal files and into family histories. Early studies linked British Columbia health and vital records, showing that various source records could be linked into individual, family, and multi-generation pedigrees (Newcombe, 1967).

Development of the theory of record linkage in a more rigorous mathematical fashion was carried out in 1969 at Statistics Canada (Fellegi and Sunter, 1969).

Key technical issues were identified early, namely: using personal identifiers to discriminate between the individual to whom the record refers and all other persons in the population; deciding whether discrepancies in identifiers are due to mistakes in reporting for a single individual or to the presence of additional individuals; and processing the large volume of data required for record linkage, while using a reasonable amount of computer time.

A generalized record linkage system was developed at Statistics Canada in collaboration with the Epidemiology Unit at the National Cancer Institute of Canada in the early 1980's (Howe and

Lindsay, 1981; Smith and Silins, 1981). Its primary benefit is that it enables record linkage to be carried out for both an internal (e.g., to create patient histories) and a variety of two-file linkages (e.g., linkage of a cancer file with death records) without reprogramming each time. When Statistics Canada used probabilistic linkage in the 1980s, based on the Fellegi-Sunter theory, it was to search the newly established Canadian Mortality Data Base, which is a file of all deaths in Canada extending back to 1950 (Smith and Newcombe, 1980).

Over the past fifteen years numerous health studies have been carried out at Statistics Canada using probabilistic matching techniques incorporated in the generalized linkage system. Various applications have been initiated in provinces, such as Ontario, Manitoba, and British Columbia. The software utilized has varied. More recent work on generalized software at Statistics Canada (GRLS.V2) has been for computers having UNIX operating system (Nuyens, 1993).

There has been a series of seven workshops describing the methods, software and results of such research (e.g., Howe and Spasoff, 1986; Carpenter and Fair, 1989). Statistics Canada also has a collection of preprocessing routines and generalized systems which are useful for the preparation of files for record linkage applications. The last record linkage workshop was held at Niagara-on-the Lake in April 1994 with the North American Association of Central Cancer Registries. This included discussions regarding the development and use of cancer registries, software, and collaboration and cooperation.

Communication and collaboration with other agencies in the various provinces in Canada, in United States, and the United Kingdom have aided record linkage developmental work (e.g., Tepping, 1968; Kilss and Alvey, 1985; Newcombe, 1988; Jaro, 1989; Newcombe, Fair, and Lalonde, 1992; Gill et al., 1993; Scheuren and Winkler, 1993). In particular, there have been efforts to study undercoverage in the United Stated census. A comparison of three different computer matchers has recently been undertaken. There is considerable record linkage work being carried out in the United Kingdom at the University of Oxford. The types of research being carried out in Oxford include such topics as trends in the workload in hospital specialties, descriptive epidemiology of hospitalised diseases, geographic variation in admission rates, social class, marital status, and place of birth, and inter-relationships between diseases.

The Methods Used in Probabilistic Record Linkage

The emphasis in this paper is on the matching of records where no unique identifier is available. In probabilistic linkage, the comparison or matching algorithm yields for each record pair, a probability or "weight" which indicates whether the records relate to the same entity. Where linkages can be based on some sort of personal identifying number, there is much greater certainty of achieving a correct match. However, that number may be incorrectly recorded on some records so that disagreement of it is not necessarily proof of a non-match. Such numbers are also occasionally improperly "borrowed" from other people, so that agreement is not always positive proof of a correct match. Although rare, such occurrences make it prudent to check names and other usual identifiers as a backup when doing numerical linkages. Thus, the methods of probabilistic record linkage should still be applied to improve the accuracy of even the numerical linkages.

Often the decision to "link" particular records depend on the similarities of names, birth dates, birth places and other items. Each of these identifiers carries only a limited discriminating power and is fallible, so the records may sometimes seem dissimilar when, in fact, the same person is involved. One must first **search** the files to bring pairs of records together for comparison. Conceptually, each record on file A (e.g., a cohort of women who have participated in a national breast screening study) is compared with each record on file B (e.g., the Canadian Mortality Data Base) to form a set of records C, which one attempts to classify as links or nonlinks. In practice, the files are blocked using identifiers (e.g., by the phonetic code of surname and gender code) to limit the number of pairs compared. Multiple passes of the file are possible, using different blocking items. The process of separating out the true matches is, in reality, a stepwise elimination of the false ones. It is not so much a matter of picking needles out of the haystack, as of progressively getting rid of the haystack without losing the needles. The problem of uniquely identifying an individual in a file of 28 million records is more difficult than identifying the person in a file of 100,000, because one has to discard from consideration far more unlinkable pairs. In estimating the information needed for linkage, the file size needs to be taken into account. Second, a **decision** is made as to whether the link is true, and then finally, a **group** of the appropriate records relating to the same individual or entity is formed. The group can consist of: a one-to-one match, such as one event record with a death; a one-to-many match, such as one event linking to cancer records, where an individual may have several cancer incidence records; a many-to-one match, such as a cancer incidence file linking to a death file; or a many-to-many match.

The generalized system used at Statistics Canada estimates how likely it is that a pair of records refers to the same entity. It does this by comparing corresponding fields one at a time between records, and seeing if the values agree, partially agree, disagree or are missing. The type of comparison outcomes can be similar to those that a human clerk would do in carrying out the same task.

Quite briefly, when comparing values Ax from a Record A (e.g., the record which is used to initiate the search) with value By from a record B (e.g., a death record which is the file being searched), the ODDS in favour of a correct LINK associated with the outcome Ax.By (i.e., the comparison pair of values) may be written in terms of the relative probability of occurrence of the particular outcome in LINKS as compared with NONLINKS; that is

$$\text{ODDS} = \frac{(P\,Ax\,.\,By \mid \text{LINKS})}{P\,(Ax\,.\,By \mid \text{NONLINKS})}\ .$$

As in information theory, the odds are usually expressed as logarithms to the base 2, and are often multiplied by ten and rounded to avoid decimals.

$$\text{Outcome Weight} = 10 * \log_2 (\text{ODDS})\ .$$

A rule is created to compare each of the fields (x,y) in the records. The comparisons can be straight comparison, cross comparisons (e.g., comparing first forename on record x with the second forename on record y), or specially written functions.

The total odds in favour of a match may be expressed as the sum of a number of "outcome weights."

$$\text{Total Weight} = \text{Outcome Weight}(O_1) + \text{Outcome Weight}\ (O_2)... + \text{Outcome Weight}\ (O_n)$$

where $O_1, O_2 \ ...\ O_n$ are the outcomes for the rules 1 to n (including any used for blocking) used to compare the fields on the records. The outcomes are assumed to be statistically independent.

The total weight becomes the overall "relative probability" that the potential link is in fact a definite link. As well, one may take into account the greater need for discriminating power, where the population being searched is large and/or the particular individual is unlikely to be represented in it. Details of the calculation of absolute odds are described elsewhere (Newcombe et al., 1983; Newcombe, 1993).

By comparing the total weight against two thresholds, this estimate is converted into a decision as to whether or not the link is a "true" one. If the total weight is above the upper threshold, the link is assigned a temporary status of "definite link;" if it is below the lower threshold, the temporary status is "unlinked;" if it is between the two thresholds, the temporary status is "possible."

<-------------(Lower ------------Upper) --------->
Unlinked Possibly Linked Linked

Possible links are then examined in more detail, perhaps on a sample basis, to fine tune the setting of the thresholds. In smaller projects, manual resolution can be carried out on these links. Further reference is usually made to the original source documents, where further identifying information may be available. In large projects, one threshold value is sometimes chosen to classify records as links and nonlinks. The analysis may be run using different threshold values. Considerable work has also been carried out in looking at the availability and type of identifiers used to identify entities (e.g., persons, families, households, farms). In some studies there is a need for a better separation of personal and family linkages when thresholds are being chosen (Newcombe, 1993). The ultimate uses of the files and the required accuracy must be taken into account.

The selection of thresholds involves two types of error. A Type I error occurs if an unlinked pair is erroneously classified as linked because it falls above the upper threshold (i.e., false positives). A Type II error occurs if a pair is erroneously classified as nonlinked (i.e., false negatives). The upper and lower thresholds are determined by the setting of acceptable errors bounds to limit the number of errors for the analysis, as is determined to be appropriate.

■ The Files Being Linked

To a large extent, the quality of record linkage is dependent on the quality of the files being linked (i.e., quality in - quality out). Substantial benefits may result from efforts to improve the accuracy, completeness and standardization of source files. There is a need for standard coding procedures (e.g., causes of disease, geographic and other coding) and data collection (Carpenter and Fair, 1990).

There is a need to have uniform data sets not only for diagnostic information, but also sufficient information is required to identify the individual and entities. Cases where one might anticipate problems include distinguishing twins, babies who die shortly after birth, persons who are similarly named in families, hyphenated names, and various groups where the naming conventions may not be the same. The ideal criteria for personal identifying information on a medical record should satisfy the requirements as shown in Table 1.

There is a requirement to have data available not only on a regional basis, but also at the national level. For example, traditionally Canada's population has been very mobile; over each five-year period since 1961, almost half the population moved from one neighbourhood, town, city, province, or territory to another. The most mobile groups of people in Canada between the 1986 and 1991 censuses were those aged 25 to 29, with seven of every ten people in this age group reporting that they lived at different addresses in the 1991 census than they had in 1986.

At Statistics Canada, efforts have been placed not only in having a national file of mortality, which goes back to 1950, but also to develop the files required in carrying out alive (since 1984), birth (since 1987), and cancer (since 1969) follow-ups.

■ The Linkage Review and Approval Process

It is important to mention that 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 Statistics Act has clauses regarding the secrecy of in-

Table 1.--Criteria for Assessing the Value of Personal Identifying Information

In ideal circumstances, personal identifying information on a medical record should satisfy the following requirements (Cheeseman and Butler, 1972; Smith, 1973).

1. The identifying information should be **permanent;** that is, it should exist at the birth of a person to whom it relates or be allocated to him/her at birth, and it should remain unchanged throughout life.

2. The identifying information should be **universal**; that is, similar information should exist for every member of the population.

3. The identifying information should be **reasonable**; that is, the person to whom it relates and others, should have no objection to its disclosure for medical purposes.

4. The identifying information should be **economical**; that is, it should not consist of more alphabetic or numeric digits and other characters than necessary.

5. The identifying information should be **simple**; that is, it should be capable of being handled easily by a clerk and computers.

6. The identifying information should be **available**.

7. The identifying information should be **known**; that is, either the person to whom it relates or an informant acting on his/her behalf should be able to provide it on demand.

8. The identifying information should be **accurate**; that is, it should not contain errors that could result in its discrepancy on two records relating to the same person.

9. The identifying information should be **unique**; that is, each member of the population should be identified differently.

No single identifier or identity set has been devised to satisfy all of these items. The efficiency of the record linkage operation depends on how well the items selected for comparison satisfy this standard.

formation. The record linkage activity must have demonstable cost or respondent burden savings over other alternatives or be the only feasible option. It must be shown to be in the public interest.

■ Major Uses of Linkage

Record linkage is being carried out at a number of provincial centres as well as at the national level in Canada. There are numerous uses. It is being used for longitudinal mortality and cancer follow-up of various cohorts, as well as for clinical trial and case control studies; for the creation of registries; for the creation of patient-oriented, rather than event-oriented, statistics (e.g., to examine the number of patients admitted to hospital rather than the number of events); for follow-up of surveys; for the preparation of sampling frames; and for examining factors which influence health care costs, curative and preventive approaches. Details of some specific examples are as follows.

Longitudinal Studies of Mortality

Long term medical follow-up has been carried out with a number of different occupational groups, such as uranium miners (Kusiak et al., 1993), asbestos workers, nickel miners, farmers (Fair, 1993) workers at Atomic Energy of Canada (Gribbin et al., 1993), synthetic textile workers (Goldberg et al., 1993), a ten percent sample of the Canadian labour force and so forth. Linkage of a "starting point" file, which will identify an individual as belonging to an "exposed" or "at-risk" population, is linked to an endpoint, such as cancer and/or death (Statistics Canada, 1992).

Collaboration is often required among various countries to carry out analysis. For example, radon is a chemically inactive, colourless radioactive gas which occurs naturally. Underground miners who are heavily exposed to the radioactive decay products of radon gas suffer especially high rates of lung cancer. There was a recent joint analysis of 11 underground miners studies published (Lubin et al., 1994). In this joint analysis, three

Canadian cohort studies were included. In the March 10, 1994 issue of *Nature*, Julian Peto and Sarah Darby have stated that extrapolation from underground miners' data suggest that radon in people's homes may cause some 2,000 lung cancers each year in Britain and 15,000 in the United States, figures which imply that radon is among the most serious environmental (as distinct from self-inflicted) causes of cancer mortality yet identified in the world. They go on to say that there is a dearth of direct evidence on the effect of domestic radon exposure. Other such international studies have included nickel workers, and plans are being made for further studies of radiation workers, as well as individuals in the pulp and paper industry.

Follow-up of Clinical Trials

A generalized linkage capacity is important for the follow-up of clinical trials. For example, the Canadian National Breast Screening Study is an individually randomized trial designed to evaluate the efficacy of the combination of annual mammography, physical examinations of the breasts and the teaching of self-examination in reducing the rate of death from breast cancer in women (Miller et al., 1992a and 1992b). The study recruited approximately 90,000 women, aged 40-59 from across Canada to participate for a five year period from 1980. The main endpoint for assessing the results of screening is the mortality rate from breast cancer in the study groups compared with the controls.

A number of important issues arise from such studies. These include, for example, the comparative value of mammography and physical examination among different age groups and the value of screening women under 50 years of age.

Building, Maintaining and Using Disease Registries

Record linkage has been important in building, maintaining, and using disease registries. For example, in Ontario, a number of data sources are utilized for the creation of a cancer registry (Dale, 1989). The four major sources of information are: hospital separations with any mention of cancer; pathology reports with any mention of cancer; death certificates in which cancer is the underlying cause of death; and reports describing patients referred to the Regional Cancer Centres and Princess Margaret Hospital. The generalized record linkage system GRLS.V1 has been used in creating the Ontario Cancer Registry.

Canada is one of the few countries in the world with a cancer reporting system covering the complete population. This coverage is achieved through the cooperation of the various provincial/territorial registries, which have provided data to Statistics Canada since 1969. The National Cancer Incidence Reporting System has recently been processed into a form suitable for record linkage. This forms what we call the Canadian Cancer Data Base. This file is an endpoint for a number of studies (e.g., follow-up of a study of farmers, and tuberculosis patients who have been exposed to fluoroscopy). The file is also being internally linked for a study of childhood cancers.

Death clearance of cancer registries can be achieved by linking existing cancer patients with mortality. Ascertainment of new cancer cases from mortality may result. Statistical uses, such as the calculation of survival rates, require the linkage of cancer and death information. Registries can be used for the follow-up of cohort studies, in clinical trials, for case control studies, for the follow-up of screening programs, and in genetic studies.

Other types of disease registries have been created and utilized for a variety of purposes (e.g., the British Columbia Health Surveillance Registry for the investigation of reproductive problems).

Regional Variation in the Incidence of Disease

Generalized linkage facilities can aid in looking at regional variation in the incidence of disease. For example, cardiovascular diseases (CVD) are the leading causes of death, potential years of life lost, and utilization of health care services in Canada. Since the late 1960's, Canada and many other western countries have experienced a noticeable decline in mortality from CVD, and particularly from ischemic heart disease (IHD). Efforts are under way worldwide to study the determinants of change in cardiovascular mortality in prospective studies. The idea of linking routinely collected hospital separation records of IHD to the Canadian Mortality Data Base was initiated and field tested in the provinces of Nova Scotia and Saskatchewan. A feasibility study relating to acute myocardial infarction (AMI) was initiated and it used information from provincial morbidity files and the Canadian Mortality Data Base. A sample of AMI cases from the hospital admission databases of Nova Scotia and Saskatchewan were field validated and individuals histories created to differentiate recurrent from incident cases. Incidence rates for fatal and non-fatal AMI's were higher in Nova Scotia than in Saskatchewan over three time periods (1977, 1981, 1985). A number of provincial surveys and related studies are now being carried out across the country.

Comprehensive Multifile Databases

Clinical data collection and surveys are often episodic, depending upon such factors as funding and investigator interest. Alternatively, administrative databases are generally continuous and are characteristically available to researchers a few months after the fiscal year end.

In Manitoba, the population health database records all patient contacts with physicians, hospitals and nursing homes. It was derived from information contained in the population registry and from health insurance claims routinely filled out by physicians and health care facilities with the Manitoba Health Services Commission (Roos et al., 1992). Information from the various files have been described earlier. Seven files provide the bulk of the information contained in the Manitoba health database, including the registration file, hospital file, medical claims file, personal care home file, the Manitoba Immunization Monitoring system, cancer registry and mortality file. These data files have been used to address many research issues, including those of surgery (Roos and Roos, 1987).

Roos, Shapiro, and Roos (1984) have observed that dying is a much more important factor than aging, per se in the high usage of hospitals. For the elderly, as for the non-elderly, a dramatic increase in use occurs in the relatively short period before death. Persons aged forty-five and over in Manitoba used an average of forty-two hospital bed days in the last year of life, while even persons aged eighty-five and over, who survived the next four years had less than seven bed days per year. Projections of hospital needs might profitably take into account expected deaths as well as population aging (Beaujot, 1991).

Discussion and Future Directions

There are a number of areas where record linkage could be used for additional administrative applications. This includes the creation, maintenance, and use of files for persons enrolled in the health care system. For example, many provinces have found the need to issue health cards. This may require creation of an up-to-date file of eligible registrants. There must be checks to verify that duplicate entries are not present on the file. Some provinces in Canada have had problems with ensuring that the registrant file is up-to-date and without duplication.

With advanced medicine and medical insurance, one might expect that social difference in mortality might disappear in Canada. However, this is not the case. Mortality differences by income or social class, as well as education, are important. There is a need to study mortality, births, and cancer by income and education level in greater depth.

Cancer and other registries are likely to become greater users of record linkage techniques. There will be a need to create and compare data across wider geographic areas. The recent meeting of the North American Central Cancer Registries, with its recent name change as well as its chosen theme -- "co-operation and collaboration" -- is evidence of this. Other types of registries include registries containing exposure data, such as the Canadian National Dose Registry of persons exposed to radiation, registries of specific diseases, and those used for studies of reproductive outcomes.

Quality control will play a greater part in every aspect of record linkage applications. This includes such things as the quality of the data sources, timeliness (e.g., production speed), customer satisfaction, relevance of the output, measurement of errors (e.g, detecting and correcting errors), and the security of sensitive information. Availability of resources (personnel, time, money, computer hardware, and software) are important factors.

Advances in information technology hold great potential not only for increasing efficiency, but for also redefining organizational and geographic boundaries and for improving information management. For some applications, one may have to implement a re-engineering plan.

It is anticipated that more applications will be conducted using software that is developed for a variety of personal computers that may be in a network environment. There is a trend to move from the mainframe computing environments, as was discussed at the recent Record Linkage Workshop held at Niagara-on-the Lake.

There is also a trend to acquire and maintain information from a variety of sources and to put databases to multiple uses. Agencies will require careful attention in their stewardship of data for policy decisions and research. A comprehensive list of recommendations for Federal statistical agencies in the United States is given in the recent book *Private Lives and Public Policies -- Confidentiality and Accessibility of Government Statistics* (Duncan et al., 1993). Administrative and statistical research uses of the files will need to be differentiated.

In order for statistics to be relevant to decision-makers, Dr. Fellegi, Statistics Canada's chief statistician, has proposed that researchers need to identify social problems of recognized importance, determine through analysis the factors related to such problems, find out which of these can be influenced through decisions, and effectively communicate the results (Beaujot, 1991).

References

Beaujot, R. (1991). *Population Change in Canada: The Challenges of Policy Adaptation*, McClelland and Stewart Inc., The Canadian Publishers, 481 University Avenue, Toronto, M5G 2E9.

Carpenter, M. and Fair, M. E. (eds.) (1989). *Canadian Epidemiology Research Conference -- Proceedings of the Record Linkage Session and Workshop.**

Carpenter, M. and Fair M. E. (1990). A Standard Data Collection Package for Medical Follow-up Studies, Ottawa, Ontario, K1A 0T6: Statistics Canada, Health Statistics Division; *Health Reports*, Cat. 82-003, 1990; 2, 157-173.

Cheeseman, E. A. and Butler, A. R. (1972). *Computer-based Routine Systems of Linked Medical Records with Particular Reference to Northern Ireland*, Northern Ireland Med. Record Linkage Research Unit, Report No. RLU 1, Queen's University at Belfast.

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*, (eds.) M. Carpenter and M.E. Fair.*

Donaldson, M. S. and Lohr, K. N. (eds.) (1994). *Health Data in the Information Age -- Use, Disclosure, and Privacy*, National Academy Press, Washington, D.C.

Duncan, G. T.; Jabine, T. B.; and de Wolf, V. A. (eds.) (1993). *Private Lives and Public Policies -- Confidentially and Accessibility of Government Statistics*, National Academy Press, Washington, D.C.

Dunn, H. L. (1946). Record Linkage, *American Journal of Public Health*, 36, 1412-1416.

Fair, M. (1993). Recent Advances in Matching and Record Linkage from a Study of Canadian Farm Operators and Their Farming Practices, in *Proceedings of the International Conference on Establishment Surveys*, American Statistical Association.

Fellegi, I. P. and Sunter, A. B. (1969). A Theory of Record Linkage, *Journal of the American Statistical Association*, 40, 1183-1210.

Gribbon, M. A.; Weeks, J. L.; and Howe, G. R. (1993). Cancer Mortality (1956-1985) Among Male Employees of Atomic Energy of Canada Limited with Respect to Occupational Exposure to External Low-Linear-Energy-Transfer Ionising Radiation, *Radiation Research*, 133, 375-380.

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.

Goldberg, M. S.; Carpenter, M.; Theriault, G.; and Fair, M. (1993). The Accuracy of Ascertaining Vital Status in a Historical Cohort of Synthetic Textile Workers Using Computerized Record Linkage to the Canadian Mortality Data Base, *Canadian Journal of Public Health*, 84, 201-204.

Howe, G. R. and Lindsay J. (1981). A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-Up Studies, *Computers and Biomedical Research*, 14, 327-340.

Howe, G. R. and Spasoff, R.A. (eds.) (1986). *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, Ottawa, Ontario, May 21-23,1986, University of Toronto Press, Toronto.*

Jaro, M. (1989). Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida, *Journal of the American Statistical Association*, 84, 414-420.

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, Department of Treasury, Internal Revenue Service, Washington, D.C.

Kusiak, R. A.; Ritchie, A. C.; Muller, J.; and Springer, J. (1993). Mortality from Lung Cancer in Ontario Uranium Miners, *British Journal of Industrial Medicine*, 50, 920-928.

Lubin, J. H.; Boice, J. D.; Edling, C.; Hornung, R. W.; Howe, G.; Kunz, E.; Kusiak, R. A.; Morrison, H. I.; Radford, E. P.;Samet, J. M.; Tirmarche, M.; Woodward, A.; Xiang Y. S.; and Pierce, D. A. (1994). *Radon and Lung Cancer Risk: A Joint Analysis of 11 Underground Miners Studies*, U.S. Department of Health and Human Services, Public Health Services, National Institutes of Health, NIH Publication No. 94-3644.

Miller, A. B.; Baines, C. J.; To, T.; and Wall, C. (1992a). Canadian National Breast Screening Study: 1. Breast Cancer Detection and Death Rates Among Women Aged 40 to 49 Years, *Canada Medical Assoc., J*, 147, 1459-1476.

Miller, A. B.; Baines, C. J.; To, T.; and Wall, C. (1992b). Canadian National Breast Screening Study: 2. Breast Cancer Detection and Death Rates Among Women Aged 50 to 59 Years, *Canada Medical Assoc., J*, 147, 1477-1488.

Ministry of Industry, Science and Technology (1993). *Canada Year Book 1994*, Catalogue No. 11-402E/1994. Available from: Publication Sales and Services, Statistics Canada, Ottawa, K1A 0T6.

Nair, C. and Karim, R. (1993). An Overview of Health Care Systems: Canada and Selected OECD Countries, *Health Reports*, 5, 259-279.

Newcombe, H. B. (1967). Record Linking: The Design of Efficient Systems for Linking Records into Individual and Family Histories, *American Journal of Human Genetics*, 19, 335-359.

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

Newcombe, H. B. (1993). Distinguishing Individual Linkages of Personal Records from Family Linkages. *Methods of Information in Medicine*, 32, 358-364.

Newcombe, H. B.; Fair M. E.; and Lalonde, P. (1992). The Use of Names for Linking Personal Records, *Journal of the American Statistical Association*, 87, 1193-1204.

Newcombe, H. B.; Kennedy, J. M.; Axford, S. J.; and James, A. P. (1959). Automatic Linkage of Vital Records, *Science*, 130, 954-959.

Newcombe, H. B.; Smith, M. E.; Howe, G. R.; Mingay, J.; Strugnell, A.; and Abbatt, J. D. (1983). Reliability of Computer Versus Manual Death Searches in a Study of the Health of Eldorado Uranium Workers, *Computers in Biology and Medicine*, 13, 157-169.

Nuyens, C. (1993). Generalized Record Linkage at Statistics Canada, in *Proceedings of the International Conference of Establishment Surveys*, American Statistical Association, pp. 927-930.

Peto, J. and Darby, S. (1994). Radon Risk Assessed, *Nature*, 368, 97-98.

Roos, L. L. and Roos, N. P. (1987). Using Large Data Bases for Research on Surgery, *Statistical Uses of Administrative Data An International Symposium*, eds. J.W. Coombs and M.P. Singh, Statistics Canada, Ottawa, 75-94.

Roos, N. P.; Shapiro, E.; and Roos, L. L. (1984). Aging and the Demand for Health Services: Which Aged and Whose Demand?, *The Gerontologist*, 24: 1, 31-36.

Roos, L. L.; Wajda, A.; Nicol, J. P.; and Roberts, J. (1992). Record Linkage: An Overview, *Medical Effectiveness Research Data Methods*, eds. M.L. Grady and H.A. Schwartz, AHCPR Pub. No. 92-0056, U.S. Department of Health and Human Services, Rockville, MD, 119-129.

Scheuren, F. and Winkler, W.E. (1993). Regression Analysis of Data Files that Are Computer Matched, *Suvey Methodology*, 19, 39-58.

Smith, M. E. (1973). *Record Linkage of Hospital Admission-Separation Records*, Atomic Energy of Canada, AECL-4507, Chalk River, Ontario.*

Smith, M. E. and Newcombe, H. B. (1980). Automated Follow-up Facilities in Canada for Monitoring Delayed Health Effects, *American Journal of Public Health*, 73, 39-46.

Smith, M. E and Silins J. (1981). Generalized Iterative Record Linkage System, *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 128-137.

Statistics Canada (1992). *Studies and References Relating to the Uses of the Canadian Mortality Data Base.**

Tepping, B.J. (1968). A Model for Optimum Linkage of Records, *Journal of the American Statistical Association*, 63, 1321-32.

World Bank (1984). *World Development Report*, The World Bank, Washington, D.C.

---

* Copies available from Statistics Canada, Occupational and Environmental Health Research Section, R.H. Coats Building, 18th Floor, Tunney's Pasture, Ottawa, Ontario K1A 0T6. ■