GENERALIZED ITERATIVE RECORD LINKAGE SYSTEM

Martha E. Smith and John Silins
Statistics Canada

I INTRODUCTION

Epidemiological follow-up of large cohorts, exposed to potentially harmful agents and circumstances, has been greatly facilitated in Canada by the development of a generalized iterative record linkage system. The availability of this system makes large-scale cohort studies feasible and economic, and has opened up a whole new era for research. The system has been developed and successfully used in several applications which have had much to do with cancer and its prevention. We are rather excited about the potential these studies have in their humanitarian objective, their scientific output, and in the development of new computer techniques to carry out such statistical studies.

The idea of 'record linkage' is by no means new. The term 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 is worth quoting verbatim for his imagery:

> "Each person in the world creates a book of life. The 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 the book into a volume." (Dunn 1946).

What is new to most people, however, is that modern computers, machine readable data sources, storage devices, and the development of advanced software for record linkage have made many more applications economic and feasible on a scale very much larger than was conceived to be possible by manual procedures. Advantages also arise from the consistency of decision-making and the use of complex matching rules (Acheson 1967 and 1968).

II DELAYED RISKS - THE GROWING DEMAND FOR DATA

The mandate under which Statistics Canada operates requires that it produce, analyze, disseminate, plan and evaluate statistics concerning, among other things, the health status of Canadians, and the social, economic, and environmental influences which affect this health status.

Because such influences frequently have long-term delayed effects on health (as distinct from the more immediate effects such as accidental deaths), the obligations are by no means fully discharged through the production of statistics relating to a point in time or a single year. If one goes through the tables in the Canada Year Book, for example, virtually all of these tables may be compared with snapshots at a point in time, or perhaps, in some cases, series of snapshots showing trends over a period of time. They describe the state and characteristics of the population.

The kinds of statistics we will be talking about are of a quite different sort, and have to do, in essence, with establishing statistical associations which may serve as pointers to possible 'troublespots', and with investigating selected suspected high-risk groups. The public keeps asking for these kinds of statistics, often without fully understanding how one must go about collecting them.
For example:

- If one works with asbestos for ten years of ones life, what are the chances of getting cancer as a result of the experience?
- How good are the present standards of protection for radiation workers?

And similar questions may be asked with respect to many individual groups exposed to hazardous substances that might cause delayed effects in the exposed individuals, including such materials as nickel, various petroleum products, and even coal. Cancer is not the only delayed effect one may be concerned with. One may be just as interested in the long-term safety of foods, drugs and the home environment as one is in the safety of the workplace.

III INDIVIDUAL LONG-TERM FOLLOW-UP

Most of you by this time will say that it is just a matter of identifying a few 'exposed' people, plus a suitable control group, and waiting to see what happens to them. One must use individual long-term follow-up to see what happens.

This is quite feasible, except that where the concern is about low levels of individual risk, as it now usually is, there can be little hope of collecting significant amounts of information unless the study populations are very large, say of the order of 10's of thousands or 100's of thousands of people. And there are the twin problems of first, identifying such large numbers of 'exposed' people, and second, as many as two or three decades later, determining which ones have died or developed cancer, or whatever, and when they did so. Current widespread demands to know the risk to health associated with living and working conditions have coincided with an increased population mobility that greatly complicates the task of pinpointing potential health trouble spots. At this stage, if one has in mind only the conventional methods of

epidemiologists (e.g., follow-up by personal contact, telephone and mail inquiries), it will be apparent that such a study is likely to be excessively laborious and expensive.

Under such circumstances, the only solution would seem to be to make better use of existing centralized files of personal records, in their machine-readable form, and of computers to match up, or link the 'starting-point' records with 'endpoint' records, specific for the individuals involved; and to do so under conditions which protect the privacy of the individual. Only thus can anyone ever expect to monitor for low levels of risk in very large populations, and to do so in a cost-effective manner.

IV THE UNIQUE OPPORTUNITY IN CANADA

Canada possesses certain major advantages for medical record linkage studies. It is very fortunate in the manner in which its provincial vital and health records are collected and organized. In addition, good working relationships have existed between the provincial and federal agencies which have collaborated to make computerized record linkage feasible. It was anticipated that a number of existing population-based files regarding health events would be extremely useful as 'endpoint' records, and therefore organization of these files has been going on in parallel with the development of the generalized record linkage system. Canada now has a computer-searchable Mortality Data Base file which includes all deaths in the country with coded cause of death, extending back to over three decades. A National Cancer Incidence Reporting system exists and dates back to 1969 (for details see Smith and Newcombe 1980).

In the absence of universal lifetime personal identity numbers on all vital and health records, it has been necessary to develop special computer methods to maximize the use of personal identifying information that is already present on the records, for the purpose of linking existing records from multiple sources into individual and family groups. Much hinges on making these computer techniques fast, accurate, and economical. Thus, substantial effort has been directed toward the development, testing, and application of appropriate methods (Newcombe et al. 1959; Kennedy et al. 1965; Newcombe 1967; Sunter 1968; Fellegi and Sunter 1969; Smith and Newcombe 1975; Smith and Newcombe 1979; Howe and Lindsay 1981).

Computer storage techniques and the speed of computers have advanced markedly in the last few years, and this makes it easier to store large volumes of data over long periods of time. We are now in a position to implement some of the studies planned in early years.

Some similar advantages exist in other countries, but not all in the same country and not always to the same degree. (OPCS 1973; Beebe 1980; Kinlen 1980; Patterson 1980).

V AN OVERVIEW OF RECORD LINKAGE

Prior to the mid-seventies in Canada, much attention was focussed on the feasibility and methodology required for computerized record linkage.

There are three major difficulties to be overcome to achieve efficient record linkage. The first difficulty arises because the personal identifying items obtained on a single record are often inadequate to discriminate between the person to whom the record refers, and all other persons in the population. Different personal identifiers and different combinations of them have different discriminating power. There has also been little uniformity in the manner in which any one person is identified by different vital and health record institutions, and in the manner and format in which such information has been put into machine-readable form over the years.

The second difficulty arises because when people report personal identifiers they frequently make mistakes. For example, if one record relates to 'Mary Smith born 20 January 1961' and a second record specifies 'Mary Smith born 28 January 1961,' a decision must be made whether there is a discrepancy in one of the personal identifiers or whether these records indeed relate to different individuals. In such circumstances, additional items must be examined to reach a decision.

The third difficulty arises because of the large volume of records involved in record linkage. In theory, each incoming 'starting' point record must be compared, in terms of personal identifiers, with each record on the 'endpoint' file and a decision must be made whether or not a given match relates to the same person or family.

With reasonable selection of personal identifiers on each record, and in spite of these three difficulties of discrimination, discrepancy, and volume, systems were developed for matching records by computer. The essence of these systems involved optimizing three major tasks. The first is the search operation in which potentially linkable pairs of records are brought together for scrutiny. If two files of records are to be linked, then all possible combinations of record pairs on both files need to be tested. The number of comparisons using this method is large, and hence, in practice, we select an item or items to sequence the files. This results in a comparison space of small blocks or pockets in which pairs of records are compared.

The second is the optimizing the decision-making step - the comparison of two sets of identifying information to decide whether or not they relate to the same individual or family, or whether there is insufficient evidence to justify either of these decisions. The final outcome of a match is one of definitely linked, rejected, or possibly linked. The selection of an optimal matching process involves decisions on the order of comparison, and on the decision rules to be used to minimize the errors in matching.

When records are being compared, one needs to

attempt to arrange for the computer to simulate the kinds of procedures and subjective judgement that a filing clerk would employ intuitively. Agreements of various identifying items will generally argue in favour of a correct linkage, whereas disagreements will argue that the records relate to different people.

The mathematical basis for such intuitive assessments is really quite simple. In general, agreements of initials, birth dates, and such will be more common in genuinely linked pairs than in pairs brought together for comparison and rejected as non-linkable. The greater the ratio of these two frequencies, the greater will be the weight attached to the particular kind of agreement.

If we wish to obtain numerical weights that can be added to other such weights, the above ratio can simply be converted to logarithms. In practice, it has been convenient to employ the logarithm to the base two as in decision theory. These so called binit weights are:

$$W_t = \log_2 \frac{A}{B}$$

where

A = the frequency of a particular agreement (or disagreement) defined as specifically as one wishes among linked pairs, and

B = the corresponding frequency of the same agreement (or disagreement) among pairs that are rejected as non-links.

For each detailed comparison of a pair of records, the positive and negative weights for appropriate agreements and disagreements are added together, and the total weight is used to indicate the degree of assurance that the pair does, or does not, relate to the same person. The procedure assumes as a tolerable approximation that the weights for the individual agreements or disagreements are uncorrelated with each other; where the items are not independent, some refinements in the manner in which weights are calculated or are assigned may be required to ensure that one does not weigh twice what is essentially the same information (e.g., change of address and change of hospital).

There is a great deal of flexibility in the manner in which weights can be employed (e.g., partial agreement, cross comparisons). They permit the introduction of numerous refinements so as to make nearly full use of the discriminating power inherent in the identifying information. One often requires an opportunity to experiment with the data files to become familiar with their characteristics in order to devise the best comparison rules. This forms the required 'iterative' aspect of designing the appropriate rules and weights for particular files.

There are two distinct frequency distributions typically created when the number of record pairs being matched is plotted against the total binit weight - one distribution for the linked pairs and another for the rejected non-linked pairs. These distributions are likely to be skewed and to overlap. This produces an area of possible links. Threshold values may be set defining the boundaries for rejected and linked pairs.

The third major concern is the retrieval problem; i.e., how the files should be structured in order to record the linked information in the most convenient manner for subsequent examination. Here the availability of new data base packages and utilities have assisted greatly. The user may want to see the results grouped in various manners, and flexibility needs to be provided here.

Several practical applications demonstrated that these methods did indeed work in a variety of areas using Canadian vital and health records. Marriage and birth records were linked into family groups for a complete province (Newcombe 1967; Smith 1980), techniques for creating cousin and multi-generation pedigrees were developed (Newcombe 1969); individual health histories for children were created using multiple sources of records (Smith and Newcombe 1975; Smith 1977); a test was carried out to determine the feasibility of linking infant death and birth records; studies involving the potential effect of the drug isoniazid (Howe et al. 1979) were carried out; and finally individual work histories relating to over 700,000 Canadians were created and linked to the mortality and cancer incidence files (Howe and Miller 1980). It was also believed that the probabilistic approach was preferable to that of using stringent criteria for matching of records (e.g., using a series of absolute agreement of selected items, or combination of items, as had been used in a study of Ontario miners (Ham 1976)).

These escalating needs and demands for statistical information were also coincident with scarcity of government funds. Thus, special efforts had to be placed on fully exploiting available resources - data bases, facilities, and human analytical talent - in order to respond to such needs and to reduce the cost and response burden required to produce the kinds of statistics being requested. Emphasis had also to be placed on the organization of the 'endpoint' files required to do long-term follow-up studies at a national scale; this is a function which other institutions are unable to perform due to the confidentiality laws governing the use of such information.

Implementation of the computer methods that had been developed earlier, in terms of computer programs, had been mainly on an ad hoc basis for each specific application. It was then necessary to devise a computer system to efficiently carry out the data processing involved, and to do so for a wide variety of research requests. Outside organizations generally come to us with detailed 'starting' point records which relate to some specific group under study. We carry out these epidemiological studies on a cost-recovery basis. The analytical interpretation of results is also often done by an outside organization.

At Statistics Canada there appeared to be at least

three approaches one might consider in providing computer facilities. When requests are received for matching of files: (1) one could develop new computer programs each time; (2) one could develop a completely generalized program (if this is indeed feasible); or (3) one could try to do a synthesis of these two methods.

The Generalized Iterative Record Linkage System (GIRLS) is basically the latter. It uses new data base technology, operates in either batch or on-line mode, is modular in development, has been designed to be simple to use, and utilizes weights to produce a quantitative measure of the total probability that two records being compared do or do not relate to the same entity. The particular rules used in the linkage are tailored to the files coming to us from a wide variety of research areas. The system is modular in development and allows the user to adjust parameters and weights easily until optimum linkage results are achieved. The system is also designed such that unwanted repetition of successfully completed steps is unnecessary, thus facilitating an iterative approach to the testing of appropriate rules and weights. A flexible command language has been developed to control the individual steps in linkage process (Hill, 1981). All commands are logged for management control, to aid the user, and to provide input to the system designers. For example, if it was noted that syntax errors were being made by many users, one could examine the user's guide to verify the instructions were clear. The system has been designed and tested using real data from a fluorscopy study, with Dr. G. Howe, from the Epidemiology Unit of the National Cancer Institute of Canada, collaborating with Statistics Canada in its development (Howe and Lindsay 1981). Since 1979, production runs from several large projects have been successfully completed using this system.

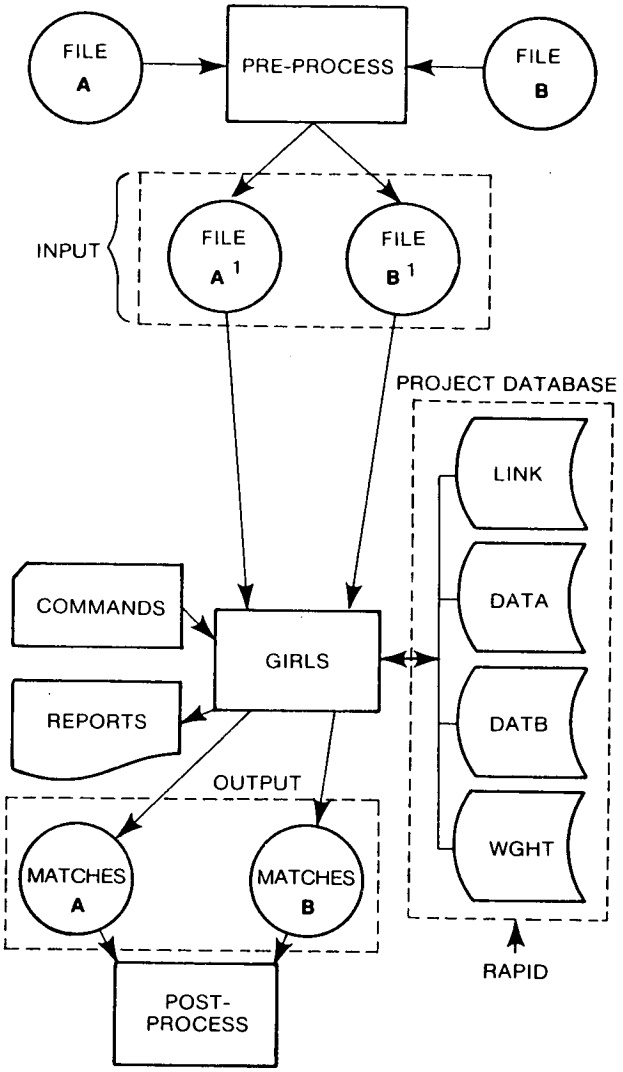
VI AN OVERVIEW OF THE SYSTEM

An overview of the flow of information through the GIRLS system is as shown in Figure 1. Note that the system can handle an 'internal' linkage where one input file is used (e.g., to create individual health histories), as well as a two-file linkage (e.g., matching a specific occupational group against the Mortality Data Base file). The system relies heavily on the use of an existing data base management system available at Statistics Canada - the Relational Access Processor for Integrated Databases, known by the acronym RAPID. The Statistical Analysis System (SAS) is also available and has been found extremely useful.

The user of the system enters the required commands according to the syntax. The user can receive a variety of reports back from the system. The final output from the GIRLS system is usually sequential magnetic tape file(s) containing information about the relevant matches. The user can then retrieve the desired information from these files in the so called 'post-processing' phase in the format required for the specific project.

Figure 1

GIRLS FLOWCHART



VII THE GIRLS MODULES

Ten modules have been developed for creating linked files of records and obtaining data from them:

(1) pre-process;
(2) ANALYSIS;
(3) weight generation;
(4) COMPARE;
(5) WEIGHT;
(6) LINK;
(7) GROUP;
(8) REPORT;
(9) INFORMATION; and
(10) post-processing.

The pre-processing, weight generation, and post-processing phases are currently regarded as specialized activities which very much depend on the user, so they are regarded as outside the current GIRLS system. We have however developed programs to handle these phases for our specific applications in the health field.

I will now go through the details of these modules, using our current Ontario miners study to illustrate how the system is being used.

The purpose of Ontario miners study is the quantitative evaluation of the delayed mortality risks involved in mining. The study group consists of about 17,000 uranium miners, 33,400 non-uranium miners, 750 salt miners, plus various other industrial groups. This file of about 57,000 records is being matched against a file of 2.4 million death records for the period 1950-77.

(1) Pre-process

Pre-processing of the files involves editing the input files; standardizing the manner in which names are handled; recoding; correcting data items, if required; encoding surname fields; and sorting the file into the sequence required for the linkage. We have found that the bulk of the effort in large-scale linkage has been taken up in this stage - mainly because the data formats, coding techniques, have varied considerably over the years. When a new file, such as the Ontario miners study comes to us, we normally use SAS to prepare frequency tabulations of the values of variables stored in the fields used for linkage. We use this output to look for inconsistencies in coding (e.g., blanks and zeros being confused), use of special characters in surnames, clustering of records pertaining to specific ethnic groups, etc.

Searching for potentially linkable pairs of records may be done by the selection of appropriate items for pocket identifiers. Items which are available, reliable, and have a high discriminating power should be used. To get around spelling variations in surnames, a phonetic New York Identification and Intelligence System (NYSIIS) surname code is used (Lynch and Arends 1977). This phonetic code was assigned to surnames on the Ontario miners file, and had already been assigned to the Mortality Data Base. For the death file over the period 1950-77, we had 211,028 different surnames that produce 38,816 unique NYSIIS surname codes.

Where an individual may have more than one surname (e.g., women with a maiden name), alternate entries are generated on the file. A duplicate flag is set to indicate this. An unique sequence number must be added to the file at this stage for each entry for the GIRLS system (i.e., duplicate entries will all contain the same sequence number).

To get around spelling variation in forenames, a modified Soundex (Smith 1973) and/or NYSIIS code may be assigned. (We have modified the Soundex routine so as to ignore the rule about retention of the first letter of forenames in order that common variations like 'CATHIE' and 'KATHY' will yield the same code.)

To overcome problems of discrepant, incomplete or variant sequencing items, linkage may be carried out in a series of alternative sequences, each new iteration of the file bringing together as many as possible of the remaining non-linked, but potentially linkable records. The files are duplicated, one or more times, corresponding to the number of different pocket identifiers used. The pocket identifiers in each case are pre-fixed with a one-character PASS - NUMBER which is incremented with each duplication. For example, when comparing files of individuals for the FALCONBRIDGE miners study (Shannon, et al. 1980), one pocket identifier was NYSIIS code of surname, while another was given name concatenated with birth year. The second pass yielded approximately an extra 0.5% of the genuine matches.

We have examined the study requests and found it advisable to split the Mortality Data Base file by sex code. Very often in specialized studies, only the appropriate half of the file needs to be examined for potential matches. Our current pre-processing program creates two output files separated by sex code.

An 'availability word' is created for each record which indicates the presence or absence of certain key items used in the linkage. This word is stored as a bit string, with zero indicating the item is absent; one, indicating it is present. A typical word might include bits referring to the presence or absence of surname, first initial, remainder of the first name, second initial, remainder of the second name, birth year, month, day, birthplace, etc. Tabulations on this word give one a very good idea regarding the overall availability of items on the file plus combinations of items (e.g., presence of complete forename and birth date on the records). This variable may be useful in the final subject-matter analysis phase. Records with blank birth date information may, for example, give a false impression that these individuals live longer, simply because of the fact the records lack adequate identifying information to link to the death file.

(2) ANALYSIS

The ANALYSIS phase takes the user's input record declaration(s), field comparison rule specifications, and user-defined function code to generate PL/1 source code. This is compiled and link-edited with the GIRLS comparison framework to produce a COMPARE program source code plus tables for subsequent modules of GIRLS tailored to the user's application.

In the Ontario miners study, the data items considered to be useful in linkage were: surname, first and second given name, birth date, birth place, year last known alive, place of work, and mother's maiden surname.

A specific user function was written so that the 'place of work' could be compared to the 'place of death' in a refined manner. If the exact spelling of the place disagreed on the two records being compared, a list of mining towns was examined. It

was found by empirical testing that miners often move, but when they do so, they tend to go to another mining centre. PL/1 code was written to incorporate these specific rules into the COMPARE module.

(3) Weight Generation

Weights may be generated using a number of different methods depending on the assumptions made. The derivation of the formulae used for weights has been described in detail earlier (Newcombe 1967; Fellegi and Sunter 1969; Howe and Lindsay 1981).

In order to start a new project, one can calculate weights for agreement of items based on the frequencies of the item as observed on the records in the files themselves. A decision may have to be made whether the frequencies on the linked set of records are most similar to the 'starting point' records or to the 'endpoint' file. In practice, it will often be found that for many items the frequency distributions may be very similar from one file to another (e.g. surname, forenames). In the Ontario miners study, for example, we calculated the frequency weights for all surnames and forenames using the 1950-77 Mortality Data Base file. For other items, such as birth year, the distribution will vary considerably from file to file.

Genuine change of the value of an item between the creation of two records can occur; also, errors of recording, keypunching, and so forth will happen in the transmission of information. The set of assumptions proposed by Howe (Howe and Lindsay 1981) led to the idea of using weight formulae which can be written as a product of frequencies, and as a function of transmission and error rates. These components are separable, and their logarithms are additive.

The important implications in terms of systems implementation are that the "frequency" and "transmission" components of the weight can be calculated, stored, accessed, and updated in completely separate ways.

In the GIRLS system, estimates of the transmission component can be obtained. To do this, threshold values are set and empirically tested using a sample file. Transmission coefficient estimates are obtained from a report produced by the INFORMATION module. These estimates are based on the non-rejected links. New values may be calculated in an iterative fashion, and the process repeated until relatively stable values are obtained. The user may use the output from previous runs to get better estimates of disagreement, partial agreements, and full agreement for the particular file(s) being used.

(4) The COMPARE

The function of the COMPARE module is to create a GIRLS database of all potential links, and to eliminate obvious non-links. Facilities are provided to modify any weights specified in the ANALYSIS phase, and to select a sample of pockets to be compared.

Identifying items are compared in the order in which they are specified at execution time. Efficiency can be gained by comparing fields on the basis of which pairs of records can be rejected immediately. For example, on certain files a 'last known alive date' may be used to create a rejection rule. If an individual was known to be alive in 1975, and this record is being compared to the mortality file, any deaths prior to 1975 may be rejected.

It is important, in terms of efficiency, to order the comparison rules such that simple numeric comparisons and those with high discriminating power come earlier. More complex character comparisons should fall near the end. Any user-defined code involving character manipulation should be optimized.

To eliminate obvious non-links, a CUTOFF parameter is used. This weight (usually specified at a conservative value) is used to stop further comparisons for a pair of records. When the running total of disagreement weights fall below the CUTOFF value, the pair of records will be rejected. In addition, if the final total weight is lower than the lower threshold value specified, the pair will be rejected.

With the Ontario miners study 57,000 male miners were compared with 2.4 million male deaths using NYSIIS code as the pocket identifier within which pairs of records were compared. 140 million pairs of records were generated and compared, but because of the CUTOFF value selected and the lower THRESHOLD value set, only 69,000 potential links were written to disk. The weights used at this phase are overall estimates - fine tuning of these occurs in the next module.

The COMPARE is generally the most costly module to run in the system. The Ontario miners' runs cost about $2,500 to do all the comparisons and write the selected records to disk.

(5) WEIGHT

The purposes of the WEIGHT module are to perform maintenance (add, delete, and change) of the weight sets, and to assign more representative statistical weights to specific items.

Several refinements have been made in the manner in which forenames are weighted in the Ontario miners study. Tables of weights were generated for all male names on the 1950-77 death file at the seven-character, four character, and initial level.

A user-defined function has been written for the Ontario miners study to compare two given names from two records. The actual comparison of the given name is done in three phases, each earlier one generally being more reliable than the next. At each step, two direct straight comparisons are made, and if this fails, then two cross-comparisions are carried out.

Phase 1: A comparison of seven, and then four characters;

Phase 2: A comparison is made using a three-character NYSIIS code at the seven character level;

Phase 3: A multi-level comparison looking for agreement on:
- a three digit Soundex code;
- the first three characters only;
- the first two characters only;
- the first character only;
- one initial only versus the first character of a complete forename.

(6) LINK

The LINK module is used to set the match status of the record pair to rejected, possible or definite.

For the purposes of the linkage program, threshold values for the total binit weight are defined. The setting of these threshold values will affect the frequencies of two kinds of errors which may be present (i.e., of false linkages and of failures to link). While it is desirable that both errors be kept to a minimum, these are competing aims, and a conscious compromise is usually required. For some statistical purposes, false linkage may perhaps be regarded as more serious than failures to link. In such circumstances, an increase in the threshold value (or values) will serve to reduce the ratio.

The user must empirically test the weights and thresholds used. Link reports are also available to examine each component used in calculation of the total weight.

In the Ontario miners study, we as yet do not have precise measurements of these error values. It is, however, planned that such an evaluation will take place in the near future. Independent sources of information, plus additional information available on the source documents, are being used to aid in getting these estimates.

(7) GROUP

The purpose of the GROUP module is to bring together all records which have linked with each other. For the internal linkage, there is no limitation upon the number of records that can constitute a group corresponding to an individual. For a two-file linkage, there may be reasonable limitations, especially if one is linking specific individuals to the mortality file. The GROUP MAPPING will automatically resolve "conflicts" based on highest total weight if a one-to-one mapping is selected.

The various types of mapping, along with an example, are shown in Table 1.

For the Ontario miners study, a one-to-many mapping was selected. With the threshold value set, this resulted in 6,750 miners records linking to 6,850 deaths. GROUP reports were prepared and the 'conflict' cases examined.

Where a miner's record has linked to two or more deaths, the source documents which had additional identifying information available were used to help resolve such cases. The typical kinds of situations which may cause such multiple matches are various family members with rare surnames clustering in a given area or occupational group. It should be pointed out that the manual searcher will also have difficulty resolving such cases.

Studies regarding the accuracy of manual versus computer linkage have indicated that the success rate for the computer operation was higher (98.3 versus 96.7 per cent) and the proportion of false linkages very much lower (0.1 versus 2.3 per cent). This test related to the linkage of vital and health records back to the birth records for children (Smith 1979).

TABLE 1

Mapping Options Available in the GIRLS System

Kind of Mapping	Example
One - One	Individual worker records linking to death records.
One - Many	Individual worker records linking to cancer incidence records.
Many - One	Worker records, with duplicate entries when the employee worked at different sites, linking to death records.
Many - Many	Worker records, with duplicate entries, linking to the cancer incidence record.

(8) REPORT

Facilities are provided for the production of a wide variety of reports. These may be produced for links, groups, weight sets, and data records. Histograms may be obtained to aid in the selection of appropriate thresholds.

(9) INFORMATION

The user can list information of a general nature regarding the system using this module. One very useful report, for example, is information on how often agreements, partial agreements and disagreements occurred for a given comparison.

For example, in another study, it was found among all the definite links, there was:

- full agreement on surname spelling 97%
- full agreement on first forename 83%
- full agreement on birth year 81%
- full agreement on birth month 95%
- full agreement on birth day 88%

In early years, birth year was calculated from age, and hence birth month and day are blank (about 25% of the links). The full agreement percentages stated here refer to cases where both variables were present and agree.

(10) Post-processing

This task is usually carried out using SAS. The user creates a file in a format suitable for subject-matter analysis.

The efficiency and cost of any follow-up depends on the availability and quality of identifying information that is common to the two files being brought together. In certain mortality studies, cases known to be alive are included in the death search to assist estimation of possible linkage errors.

In general, the larger the number of records in a cohort file, the lower the cost per search. Unit costs in the vicinity of a dollar or two per search may be encountered. This includes the overall steps of preprocessing the files, encoding surnames, sorting and preparing a file in a format suitable for analysis, in addition to the linkage steps. There are fixed costs for the personnel time required to do the computer programming and tailoring of the linkage technique to make maximum use of the items available.

VIII THE PRODUCTS

Most of the studies undertaken to date relate to groups of people numbering in the tens of thousands to hundred of thousands. The studies are concerned with the long-term consequences of various occupations, medical treatments and diagnostic procedures, reproductive problems, lifestyle and other environmental factors, plus organizing cancer incidence and survival data. Table 2 indicates the wide range of topics of interest and also the size of the studies being carried out. An expanded list of references for these studies is given in (Smith 1979; Smith 1980; Smith et al. 1980, Smith and Newcombe 1981).

(1) Occupational Groups

The industrial groups include uranium miners, hard rock and salt miners, nickel workers, uranium refinery workers, radiation workers, plus some who are exposed to asbestos, fibreglass, vinyl chloride, and formaldehyde vapour.

In Ontario, the recent report of the Royal Commission on the Health and Safety of Workers in Mines (Ham 1976) indicated an excess of lung cancer among uranium miners. The study currently in progress is more comprehensive and will attempt to calculate the relationship between the time course of exposure to short-lived radon daughters and the time course of excess lung cancer risk. Several control populations will be used: Ontario males, Northern Ontario males and Ontario miners who have not worked in a uranium mine. The study is ongoing and the data are to be reviewed regularly.

TABLE 2

The Size and Type of Long-Term Follow-up Studies

Study	Number of Individuals
A. Occupational Groups	
1. Ontario Uranium miners	16,000
2. Canadian labour force - 10% sample	700,000
3. INCO nickel workers	62,000
4. Falconbridge nickel workers	12,000
5. Eldorado uranium workers	16,000
6. Ontario miners nominal roll - uranium miners - non-uranium miners - salt miners - asbestos workers - insulation workers (fibreglass) - nickel sinter workers - nickel carbonyl workers - polyvinyl chloride workers - vinyl chloride workers	57,000
7. Newfoundland fluorspar miners	2,000
8. Railway workers	18,000
9. Ontario morticians	1,500
10. Asbestos workers	2,000
B. Medical Diagnosis and Treatment	
1. Isoniazid and cancer in tuberculosis patients	64,000
2. Fluoroscopy and cancer in tuberculosis patients	100,000
C. Reproductive Problems	
1. Infant death - birth linkage (1971 births)	6,000
2. Infant death - birth linkage (1978 births)	4,200
3. Breast cancer and age at first birth	300,000
D. Lifestyle and Other Environmental Factors	
1. Nutrition Canada Survey participants	20,000
E. Cancer Incidence, Prevalence, Evaluation of Care, and Survival	
1. Ontario Cancer Registry Reporting System	125,000
2. Alberta Cancer Registry Death Clearance	175,000

There is existing scientific evidence, although by no means conclusive, suggesting that the various sulphide forms of nickel are carcinogenic. The initiative for an INCO study of nickel workers came from a joint union-management committee on occupational health set up by INCO Metals Limited and the United Steel workers of America, and was funded as a result of the 1975 negotiations over renewal of their three-year agreement. McMaster University was chosen to conduct the work (Roberts et al. 1980). Falconbridge Nickel Limited made a similar request to the university (Shannon et al. 1980). The employment records for the cohort populations of about 62,000 and 12,000 individuals, respectively, have been matched against the Mortality Data Base file. Since INCO has produced, in earlier years, up to 75% of the western world's nickel (and still accounts for 40%), its potential contribution to the study of the effect of nickel on humans should be great.

The Eldorado Nuclear uranium workers, who are being followed in a similar fashion, include those employed in mining, milling, and the extraction of radium for medical purposes, and the refining of uranium as a nuclear fuel (Abbatt et al. 1980). As the operator of Canada's first uranium mine in Port Radium, and the country's only refinery over a 50-year period at Port Hope, they are in a good position to make an important contribution to the scientific knowledge of the health effects of the uranium industry on its workforce.

Plans are being made to also follow-up other radiation workers (Newcombe 1976; Weeks 1979; Myers et al. 1981; Ashmore 1979; Newcombe 1980). This study, like many others, serves to stress the need for industry to retain adequate name rosters of past employees in compact, easily retrievable form. Extensive work has been involved in compiling the nominal roll file. Many institutions now routinely destroy old personnel files a few years after employees have terminated. In an ongoing study of the employees of a company, a cumulative name roster should be maintained and updated routinely.

Ontario morticians are of special interest as a study group because of their exposure to formaldehyde vapour (Gunby 1980). Use is being made of records for 1,500 licences issued to embalmers and funeral directors in Ontario from the period 1928-57 (Levine 1979). We have done a manual search using the early microfiche file, as well as a computer search for deaths after 1950.

(2) Medical Diagnosis and Treatment

Fluoroscopy and cancer in tuberculosis patients is being studied to determine the relationship between radiation dose and the magnitude of the subsequent risk of cancer. Of particular interest are cancer of the breast, lung and thyroid, and leukaemia (Myrden and Hiltz 1969; Cook et al. 1974; Newcombe 1975). This study is being carried out by the Epidemiology Unit of the National Cancer Institute of Canada (Howe et al. 1980). The generalized iterative record linkage system was developed and first tested using data from this project.

(3) Reproductive Problems

A cohort study is being carried out to determine whether the risk of breast cancer differs according to the age at which a woman bears her first full-term child. About 300,000 British Columbia first-births have been selected as the study group, and their birth records will be matched against female deaths represented in the Mortality Data Base file (Stavraky et al. 1979).

(4) Lifestyle and Other Environmental Factors

There is a growing interest in the extents to which other circumstances including diet, food additives, lifestyle, pollution, and other variables may contribute to killing conditions such as cancer and cardiovascular disease. About 20,000 Canadians participated in a Nutrition Canada survey conducted in 1971-72, and the subsequent mortality experience of these persons is being investigated (Verdier 1979).

(5) Cancer Incidence, Prevalence, Evaluation of Care and Survival

The Ontario Cancer Treatment and Research Foundation uses available health records in which cancer is mentioned to generate incidence data. Each year the Foundation receives 65,000 hospital separation forms, 15,000 death abstracts and 25,000 pathology reports, together with 15,000 new patient registrations from the Princess Margaret Hospital and the Foundation's regional treatment centres. Currently, these reports are being linked at Statistics Canada for the period 1972-76 to produce new Ontario Cancer Incidence data (Clarke and Spengler 1980). The linkage uses both the two-file capability of the linkage system (e.g., linkage of hospital separations to pathology reports) plus the internal-linkage option whereby a variety of record sources relating to the same individual are brought together so that a composite record can then be created.

IX THE LAW

Statistics Canada does carry responsibility for the confidentiality of the vital and health records which are entrusted to it. There are several pieces of legislation which define what we may or may not do. Most important are the Statistics Act, the recent federal Human Rights Act, Orders-in-Council pertaining to vital statistics, and the various statistics acts or their equivalents. (For further details regarding confidentiality of records see: Medical Research Council 1968; Rowebottom 1979; Krever 1980; Hansen 1981).

X CONCLUSION

The approach being taken in Canada for the development of the files and facilities for use in epidemiological studies is much like the development of the electron microscope for the biologist. Here a tool was developed to serve a multiplicity of users. The facilities should not only meet present needs, but favour inquiry into

all sorts of future questions without foreknowledge of what these questions will be. The overall strategy for problem solving has been multi-disciplinary. A combination of laboratory, clinical, genetic and epidemiological approaches are required to get a clearer understanding of our health problems and their solution.

The recent developments at Statistics Canada will render many more epidemiological studies possible to a degree which has not been seriously contemplated in the past. The technology and systems development for bringing records together have been well tested. The expansion and enhancement of our current capabilities very much depend on public demand, financial support, and close collaboration and co-operation of a number of federal, provincial and other agencies.

ACKNOWLEDGEMENTS AND NOTES

The authors would like to thank the members of the GIRLS project team, and especially Ted Hill and Pierre Lalonde, for their help in the preparation of this paper.

The views expressed in this paper are those of the authors and do not necessarily represent the views of Statistics Canada.

REFERENCES

Abbatt, J.D. and E.N.L. Project Team (1980) The Eldorado Epidemiology Project: Health Follow-up of Eldorado Uranium Workers. Eldorado Nuclear Ltd., Ottawa.

Acheson, E.D. (1967) Medical Record Linkage. Oxford University Press, London - New York - Toronto.

Acheson, E.D. (Ed.) (1968) Record Linkage in Medicine. E. & S. Livingstone Ltd., Edinburgh & London.

Ashmore, P. (1979) The National Dose Registry, file content, and potential uses. (A paper presented at the NCIC Workshop on 'Computerized Record Linkage in Cancer Epidemiology', Ottawa, 8-10 August 1979).

Beebe, G.W. (1980) Record linkage systems - Canada vs the United States, AJPH 70, 1246 1248.

Clarke, E.A. and Spengler, R.F. (1980) Review of cancer incidence, mortality, treatment and survival in Canada. In: Ontario Cancer Treatment and Research Foundation annual report, Cancer in Ontario 1980, Toronto, Ontario, pp. 40-60.

Cook, D.C., Dent, O. and Hewitt, D. (1974) Breast cancer following multiple chest fluoroscopy: The Ontario experience. Can. Med. Assoc. J. 111, 406-410.

Dunn, H.L. (1946) Record linkage. AJPH 36, 1412-1416.

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

Gunby, P. (1980) Fact or fiction about formaldehyde? JAMA 243, 1697-1703.

Ham, J.M. (1976) Report of the Royal Commission on the Health and Safety of Workers in Mines. Ministry of the Attorney General, Province of Ontario, Toronto.

Hansen, I. (1981) Report of the Privacy Commissioner on the Use of the Social Insurance Number. Canadian Human Rights Commission, Ottawa, Ontario.
Copies available from: Canadian Govt. Publ. Centre, Supply and Services Canada, Hull, Quebec K1A 0S9 Cat. No. HR21-7-1981-B.

Hill, T. (1981) Generalized iterative record linkage system user guide. An unpublished report available from: Systems Development Division, 12-K, R.H. Coats Bldg. Tunney's Pasture, Statistics Canada, Ottawa Ontario K1A 0T6.

Howe, G.R. and Lindsay, J.P. (1981) A generalized iterative record linkage computer system for use in medical follow-up studies. Comp. Biomed. Res. (In press).

Howe, G.R., Lindsay, J., Coppock, E. and Miller, A.B. (1979) Isoniazid exposure in relation to cancer incidence and mortality in a cohort of tuberculosis patients. Int. J. Epidemiol. 8, 305-312.

Howe, G.R. and Miller, A.B. (1980) Cancer incidence and mortality in relation to occupation in 700,000 members of the Canadian Labour force. (A paper presented at the Fourth International Symposium on Prevention and Detection of Cancer, London, England, July 1980).

Howe, G.R., Miller, A.B. and Sherman, G.J. (1980) Breast cancer mortality following fluorscopic irradiation in a cohort of tuberculosis patients. (A paper presented at the Fourth International symposium on Prevention and Detection of Cancer, London, England, July 1980).

Kennedy, J.M., Newcombe, H.B., Okazaki, E.A. and Smith, M.E. (1965) Computer Methods for Family Linkage of Vital and Health Records. Atomic Energy of Canada Ltd., Publ. No. AECL-2222, Chalk River, Ontario.

Kinlen, L.J. (1980) The death registration system and the national health service center register in Britain: their value to epidemiological research. In: J. Cairns, J.L. Lyon, and M. Skolnick (Eds.) Banbury Report No. 4 - Cancer Incidence in Defined Populations. Cold Spring Harbor Laboratories, pp. 437-442.

Krever, H. (1980) Report of the Commission of Inquiry into the Confidentiality of Health Information. Printed by: J.C. Thatcher, Queen's Printer for Ontario. Copies available from: Publication Services Section, Ontario

Government, 880 Bay Street, 5th Floor, Toronto, Ontario M7A 1N8.

Levine, R.J. (1979) Retrospective cohort mortality study of Ontario embalmers and funeral directors. (A description of a research project. Mimeographed document. Chemical Industrial Institute of Toxicology, Research Triangle Park, North Carolina 27709).

Lynch, B.T. and Arends, W.L. (1977) Selection of a Surname Coding Procedure for the SRS Record Linkage System. Sample Survey Research Branch, Research Division, Statistical Reporting Services, Dept. of Agriculture, Washington, D.C.

Medical Research Council of Canada (1968) Health Research Uses of Record Linkage in Canada. MRCC Report No. 3, Ottawa, Ontario.

Myers, D.K., Newcombe, H.B. and Marko, A.M. (1981) Long-term follow-up of radiation workers in Canada. In: Proceedings of REAC-TS International Conference on the Medical Basis of Accident Preparedness. Oak Ridge, Oct. 1979. (In press).

Myrden, J.A. and Hiltz, J.E. (1969) Breast cancer following multiple fluoroscopies during artificial pneumothorax treatment of pulmonary tuberculosis. Can. Med. Assoc. J. 100, 1032-1034.

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

Newcombe, H.B. (1967) Record linking: the design of efficient systems for linking records into individual and family histories. Am. J. Hum. Genet. 19, 335-359.

Newcombe, H.B. (1969) The use of medical record linkage for population and genetic studies. Methods of Information in Medicine 8, 7-11.

Newcombe, H.B. (1975) Cancer following multiple fluoroscopies. Publ. No. AECL-5243. Chalk River Nuclear Laboratories, Chalk River, Ontario.

Newcombe, H.B. (1976) Plan for a Continuing Follow-up of Persons Exposed to Radiation in the Nuclear Power Industry. Publ. No. AECL-5538, Chalk River Nuclear Laboratories, Chalk River, Ontario.

Newcombe, H.B. (1980) Design and future use of national dose registers for regulatory control and epidemiology. Health Physics 39, 783-796.

Office of Population Censuses and Surveys (1973) Cohort Studies: New Developments. Studies on Medical and Population Subjects No. 25. Her Majesty's Stationary Office, London, England.

Patterson, J.E. (1980) The establishment of a national death index in the United States. In: J. Cairns, J.L. Lyon, and M. Skolnick (Eds.) Banbury Report No. 4 - Cancer Incidence in Defined Populations. Cold Spring Harbor Laboratories, pp. 443-451.

Roberts, R.S., Julian, J.A., Shannon, H.S., Muir, D.C.F. (1980) Mortality studies in Ontario nickel workers: The INCO/JOHC project. In: S.S. Brown and S.W. Sunderman Jr. (Eds.) Nickel Toxicology, Academic Press, London, pp. 27-30.

Rowebottom, L.E. (1979) Statistics Canada: The legal basis of health record linkage. (Paper presented at NCIC workshop 'Computerized Record Linkage in Cancer Epidemiology', Ottawa, 8-10 August 1979).

Shannon, H.S., Cecutti, A.G., Julian, J.A., Muir, D.C.F. and Roberts, R.S. (1980) Mortality studies in Ontario nickel workers: The Falconbridge Project. In: S.S. Brown and S.W. Sunderman, Jr. (Eds.) Nickel Toxicology, Academic Press, London, pp. 23-26.

Smith, M.E. (1973) Record Linkage of Hospital Admission - Separation Records. AECL Report No. AECL - 4507, Chalk River, Ontario.

Smith, M.E. and Newcombe, H.B. (1975) Methods for computer linkage of hospital admission - separation records into cumulative health histories. Methods of Information in Medicine 14, 118-125.

Smith, M.E. (1977) The Use of Computer for Studying the Origins and Social Cost of Ill Health. In: D.B. Shires and H. Wolf (Eds.); MEDINFO 77. Proceedings of the Second World Conference on Medical Information, Amsterdam, North Holland, pp. 37-41.

Smith, M.E. (1979) Automated medical follow-up and delayed industrial risks. In: N.E. Gentner and P. Unrau (Eds.) Proceedings of the First International Conference on Health Effects of Energy Production. Report No. AECL 6958. Chalk River Nuclear Laboratories, Chalk River, Ont. pp. 125-132.

Smith, M.E. (1980) The present state of automated follow-up in Canada. Part 1: Methodology and files. J. of Clinical Computing 9, 1-18.

Smith, M.E. (1981a) Value of record linkage studies in identifying populations at genetic risk and relating risk to exposures. In: K.C. Bora (Ed.) Proceedings of an International Symposium on Chemical Mutagenesis, Human Population Monitoring, and Genetic Risk Assessment. Amsterdam: ASP Biological and Medical Press B.C. (In press).

Smith, M.E. (1981b) Long-term medical follow-up in Canada. A paper presented at a conference on the Quantification of Occupational Cancer, Banbury Center, 29 March - 2 April 1981. Proceedings of the conference are to be published in Banbury Report No. 9 - Quantification of Occupation Cancer, Banbury Center, Cold Spring Harbor Laboratory, P.O. Box 534, Cold Spring Harbor, New York 11724.

Smith, M.E. and Newcombe, H.B. (1975) Methods for computer linkage of hospital admission - separation records into cumulative health histories. Methods of Information in Medicine 14, 118-125.

Smith, M.E. and Newcombe, H.B. (1979) Accuracies of computer versus manual linkages of routine health records. Methods of Information in Medicine 8, 89-97.

Smith, M.E. and Newcombe, H.B. (1980) Automated follow-up facilities in Canada for monitoring delayed health effects. AJPH 70, 1261-1268.

Smith, M.E. and Newcombe, H.B. (1981) Use of the Canadian Mortality Data Base for epidermiological follow-up. Can. J. of Public Health (In press).

Smith, M.E., Silins, J. and Lindsay, J.P. (1980) The present state of automated follow-up in Canada, Part II. The Products. J. of Clinical Computing 9, 19-30.

Stavraky, K., Smith, M.E., Uh, S.H. and Miller, J.R. (1979) A cohort study of the age of first birth upon the risk of death from breast cancer. (Paper presented at NCIC Workshop 'Computerized Record Linkage in Cancer Epidemiology', Ottawa, 8-10 August 1979).

Sunter, A.B. (1968) A statistical approach to record linkage. In: E.D. Acheson (Ed.) Record Linkage in Medicine. E. & S. Livingstone Limited, Edinburgh and London, pp. 89-107.

Verdier, P. (1979) Nutrition in relation to mortality. (Paper presented at the NCIC workshop 'Computerized Record Linkage in Cancer Epidemiology', Ottawa, 8-10 August 1979).

Weeks, J.L. (1979) A Registry for the Study of the Health of Radiation Workers Employed by Atomic Energy of Canada Limited. Publ. No. AECL-6914, Whiteshell Nuclear Research Establishment, Pinawa, Manitoba.