AUTOMATED NATIONWIDE DEATH CLEARANCE OF PROVINCIAL CANCER REGISTRY FILES -- THE ALBERTA CANCER REGISTRY STUDY

Martha E. Smith, Statistics Canada
Howard B. Newcombe, Consultant
Ron Dewar, McGill University

Knowledge of the dead-or-alive status of registered cancer patients is needed by cancer registries for a variety of administrative and epidemiological research purposes, but is not always readily obtainable. As a result, considerable interest has been expressed in Canada regarding the possibility of large-scale death searches employing the historical Canadian Mortality Data Base (MDB) file which has been recently established at Statistics Canada. This is a machine-readable file of all deaths occurring in Canada since 1950 [1], and has been created as a by-product using records routinely collected by the various provincial vital statistics offices for compilation of annual statistics and for legal purposes. As a first test of the use of this data base for cancer registry "death clearance", i.e. the linkage of provincial cancer registrations with any relevant death registrations to determine dead-or-alive status, it was decided that the Alberta cancer registry records would be used to initiate the death searches, on a trial basis. The factors influencing the success of the operation for the one province would then be studied so as to develop strategies for an automated nationwide death clearance for the National Cancer Incidence Reporting System [1-2]. The Alberta Cancer Registry Study described here therefore constitutes a test of the practicability and value of what in future may become a general practice for cancer and other special disease registries, in Canada and elsewhere.

This paper highlights some of the main features of the undertaking, which are more fully described in three detailed operational planning documents prepared for this study [3-5]. Some of the early results from the actual production runs are given, and their implications are discussed. The specific theme of this paper is that refinement of searching, where there are no useable personal identity numbers, often requires a probabilistic approach.

Organizationally, this paper is divided into several parts. We will first give the main results and conclusions as indicating the technical feasibility of matching existing cancer registry and national death files by computer. Some detailed findings will be given regarding two specific modifications made, involving rules for preliminary eliminations of unpromising record pairs, and early cutoff of the sequence of identifier comparisons where the evidence against linkage is already conclusive. The methods and logic used to locate the matching death records, and to confirm that they are correctly paired with the appropriate cancer record are discussed. A refinement has been made to the formula used for calculating the total weights. New weights specific for disease diagnosis have been calculated and used in this project. Future plans include evaluation of the overall success of the linkage. Practical long-term implications of the study for the administration of cancer and vital statistics registries, for epidemiological research, and for data collection, will be considered.

The logic of probabilistic record linkage is much the same as would be employed by a human searcher, but is more precisely quantified. The special features of the present death searches are: (a) the probabilistic approach itself [6-7], (b) a generalized computer system for record linkage [8-10], and (c) certain strategies to make the computer operation fast, economical and accurate.

1. MAIN RESULTS AND CONCLUSIONS

The chief limiting factor in both computerized probabilistic death searches and the corresponding manual searches is the amount of the personal identifying information entered into the records and made available for the searching process. Computerized searching can be accomplished on a large-scale wherever the cancer and death files have individuals identified by full current and birth names, birth date, birthplace, sex, marital status, mother's maiden name, and so forth.

Although it is prudent to carry out manual checks where the computer is uncertain about the correctness of a match, these will only slightly improve the accuracy unless additional identifying information is made available to the human. Where that is the case, much time and effort would be saved if such additional identifiers were collected in machine-readable form in the first place, on the vital statistics and cancer records. Appropriate channels of communication need to be established between the provincial and national agencies, as well as health researchers, to ensure adequate data collection and uniform data entry.

The results of the Alberta study indicate that death linkages should be preceded by a computerized "internal" linkage of duplicate entries within the cancer file, both to simplify the subsequent death linkages, and also to expose the more important limitations of the availability of identifying particulars in some of the cancer record files. For example, where complete birth date information is not available, and where surname changes at marriage are not indicated on the cancer records for women, many potential death matches will be missed altogether.

2. DETAILED FINDINGS FROM THE ALBERTA STUDY

The files of the Alberta Cancer Registry [11] used in this study were for the period 1953-1978 and contained a total of 178,856 records of registered patients; 98,749 of these patients had malignancies, and the remainder were benign cases. The files of the Mortality

Data Base relate to approximately 4.6 million deceased persons whose deaths were registered in Canada over the period 1950-1981.

The major human tasks involved in linkage of records have been simulated in various phases of the Generalized Iterative Record Linkage System (GIRLS) [10] as shown in Figure 1. There is: (1) the searching for the appropriate pairs of records for comparison (the COMPARE phase), (2) making a decision as to whether the same individual is involved (WEIGHTS, LINK), (3) grouping all the information relating to the individual and selecting, if necessary, the best match (GROUP, MAPPING), and (4) retrieval of appropriate information for the user.

Certain strategies were developed to make the computer operation fast, yet flexible to allow for mis-reporting of items. Theoretically one should consider all possible combinations of pairs of records from the two files. In practical terms, however, one does not do this, but rather one tries to "funnel" and "sift" the record pairs, so that only the promising pairs will undergo detailed assessment, and only the good links will remain at the end (see Figure 1).

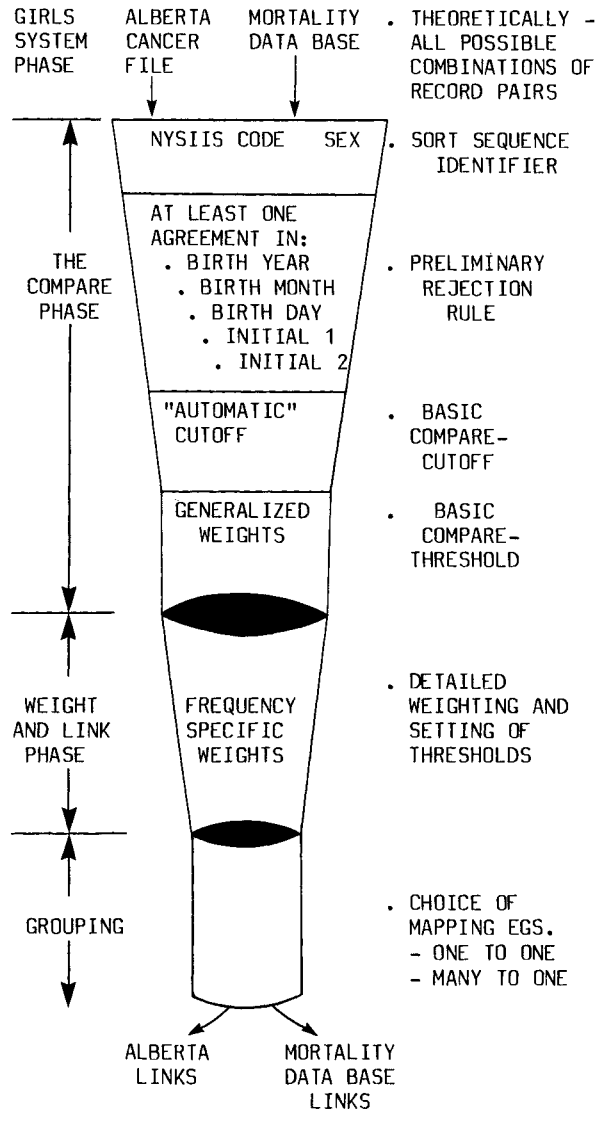
To reduce the number of record pairs passing through the successive comparison steps in the COMPARE phase, two modifications have been made to the methods often used previously. Both modifications, to date, appear to have been particularly successful. These involved rules for preliminary eliminations of unpromising record pairs, and for the automatic generation of an early cutoff during the COMPARE phase, when the chance of subsequent agreements is quite unlikely to rescue a record pair so that it will pass the threshold set at the end of the COMPARE phase.

It was during the GROUPING and MAPPING phases, that the importance of an internal linkage to identify all records relating to the same individual was noted to be of particular importance. Some registries may endeavour to do such a linkage for a particular hospital, but may fail to detect redundant entries when the patient is seen at treatment centres in different cities. If the internal linkage is not done first, one has problems differentiating instances in which two or more clinic records for the same individual are correctly linked to one death record, versus cases where records for two or more different individuals are trying to link to the same death.

3. THE METHODS AND LOGIC

To design the procedures for the Alberta death clearance, it was necessary: (a) to single out the identifiers on which the linkages are to be based, (b) to specify the comparisons to be made where these are not obvious (as when comparing place of current residence or place of residence at diagnosis, as stated on the cancer record, with place of residence or death as stated on the death registration), and (c) to provide the means by which the various agreements, disagreements and partial agreements are "weighted" as indicating the strength with which they argue for or against a correct linkage.

FIGURE 1. Reducing the Number of Record Pairs Passing Through the Comparison Steps and Retrieving the Appropriate Links



For this latter purpose use is made of sets of weights for specified agreements, disagreements and partial agreements of the various identifiers. Positive components of these weights are derived from the frequencies of various identifier "values" in the files themselves. These represent the positive discriminating powers of various names, initials, place names and such, when they agree. Negative components of the weights are derived from frequencies of agreements, disagreements and such among the matched pairs of records out of an earlier test run. In general, rare names, place names etc. carry high positive discriminating power when they agree, and rare disagreements carry high negative discriminating power when they occur. The tables of weights are analagous to the

kinds of information which influence an experienced human searcher when performing the same task, but they represent a more precise and thorough quantification of the data used in the matching of the records.

Kinds of linking information available - - For the purpose of searching a file pertaining to individual people, those items of personal identification of greatest value tend to be those which are most distinctive, routinely available, consistently reported, and permanent. The kinds of identifiers and the natures of the comparisons used in the Alberta study are described in [3]. In each step, one tries to simulate the reasoning processes used subjectively by a human searcher, but to base the judgments on better quantitative data than the searcher would have. The computer thus attaches greater calculated positive weights to agreements of rare names and rare diagnoses, and greater calculated negative weights to rare kinds of disagreements, than to the commoner counterparts of these comparison outcomes.

For the present death searches, for example, it may be said that:

(a) the larger the file, the more likely it is that a similarity in the identifying information could have arisen by chance;
(b) the more serious the diagnosis the more likely it is that a matching death record will be found in the years closely following the year of diagnosis, and not in the much later years [4];
(c) rare surnames, forenames, initials, place names and such, argue more strongly for a match than do the more common surnames etc., because the rarer they are the less likely they are to agree just by chance;
(d) rare disagreements argue more strongly against a match than do commoner disagreements;
(e) birth date agreements argue for linkage, disagreements argue against it, and mixed agreements and disagreements, and partial agreements may argue less strongly and in either direction;
(f) marital status, because it is prone to change with time, usually carries little negative weight when it disagrees; unless the direction of change (e.g. from "married" to "single") is inherently improbable;
(g) a city or place of residence is prone to change, but can carry considerable positive weight when it agrees, especially if it is a small town;
(h) diagnosis can be used in the way indicated in (b) above, and also as an identifier in its own right [4]; with the latter kind of use the rarer the diagnosis the greater the discriminating power when it agrees.

The comparison procedures - - For computer searching of a large file of named individuals, it has become customary to array the records being searched, and those initiating the searches, in a phonetically coded surname sequence, separately for males and females. The phonetic coding reduces problems arising out of variant spellings. When the surname code and sex code agree, other identifiers are then compared and the outcomes from these comparisons are assigned positive or negative weights which are later summed. The GIRLS system recognizes two stages in this process, known respectively as the COMPARE phase and the WEIGHT phase.

Recently, use has been made of a preliminary "tough" rejection rule to weed out a substantial number of unpromising matched pairs before subjecting the remainder to the full comparison procedures. The coarse rejection rule used in the Alberta study was to reject from further comparison any records where there was no agreement at all (either because of disagreement or because of missing values) of any of the following five identifiers: year of birth, month of birth, day of birth, first initial and second initial. The results of the elimination of matched cancer records with death records for males is shown in Table 1.

In the earlier tests using the Alberta files, the procedures first used allowed all record pairs for which the NYSIIS [14] surname and sex coded agreed, to enter the main comparison steps of the COMPARE phase. Later, some preliminary reduction was achieved through the use of a "soft" rejection rule; that is, the pair was rejected whenever there was disagreement with respect to all elements of the birth date plus the two initials, both when compared directly and cross compared (but not when any of these were missing).

The test results, using the same sample from the Alberta file, showed as follows:

(a) no preliminary rejection 100% remain
(b) preliminary rejection - "soft" rule 59.75% remain (636,672/1,065,571)
(c) preliminary rejection - "tough" rule 24.11% remain (256,926/1,065,571)

that is, the "tough" rule is 2.5 times better than the "soft" rule and 4.2 times better than no preliminary eliminations at all.

Tests on both the Alberta death sample, and for another Eldorado project, have failed to show losses of likely good links. This will be more fully examined in our proposed evaluation after all the production runs have been completed.

The design of the GIRLS COMPARE phase originally provided for a CUTOFF, set by the user and based on an accumulated negative component of weight and a fixed negative cutoff value. This had several disadvantages: (a) it failed to use the discriminating power of the positive component of the weight at all; (b) the negative component became meaninglessly high in the case of partial agreements, especially when narrowly defined; (c) the eliminations were apt to take place early, before enough identifiers had been compared to base a judgment on; (d) there was sometimes a tendency for too many pairs to get through so that the burden of the COMPARE phase was not substantially reduced, i.e. the threshold at the end still did most of the eliminating; and (e) a user working on a new project sometimes had difficulty selecting an appropriate negative value for the CUTOFF, without resorting to time-consuming preparation of distributions of the values from test runs.

TABLE 1. Elimination of matched pairs of Alberta cancer records with death records -- males (Based on 77012 cancer records for males, matching against 1950-1981 Mortality Data Base records for male deaths numbering 2,815,208 wherever the the NYSIIS surname code agree.)

Stage in the Rejection Procedure	Pairs Remaining	Reduction Factor Single Step
Total possible pairs at outset (ignoring NYSIIS)	216,804,798,496	
Initial record pairs matched on NYSIIS	171,602,980	1,263.
INITIAL PROCESSING (separately for different death year groups)		
After first "tough" rejection[1]	36,180,992	4.7
After first BASIC COMPARE phase (cutoff = -999, thresholds = -20,50)	118,049	306.
PRODUCTION PROCESSING[3] (re-run with just the matched deaths but from all years)		
All record pairs matched on NYSIIS	11,539,531	-
After second "tough" rejection[2]	2,858,499	
After second BASIC COMPARE phase (cutoff = - 999, thresholds = -20,50)	118,049	-
After WEIGHT phase (threshold = -20)	83,629	1.4
After WEIGHT phase (threshold = 0)	64,194	1.3

1,2 The "tough" rejections exclude all record pairs in which there was not complete agreement in at least one item of (a) year of birth, (b) month of birth, (c) day of birth, (d) first initial, or (e) second initial.

3 The production processing is required because of the manner in which our computer production runs were carried out. The MDB is divided into five years of deaths. After the BASIC COMPARE, outputs are merged and a smaller 1950-81 BASIC COMPARE is executed. This second "tough" rejection relates to this last production run.

For these reasons, some refinements were made to the system, and an optional "automatic" cut-off system devised and implemented. The order of the outcome comparisons specified by the user is now less critical, and the need for trial and error at this stage is reduced.

The manner in which this "automatic" cutoff works is as follows. The CUTOFF is now based on the net weights (i.e. the negative and positive components together). As identifying items are being compared, there is often a certain point at which, due to too many disagreements, the pair needs to be disregarded. After each identifier comparison outcome, reference is made to a table (set up at the very beginning) giving the maximum positive weight that could possibly be assigned for the remaining identifiers assuming all of them agree. If the sum of a current negative accumulated weight, part way through the comparison, plus the maximum positive weights for the remaining identifiers assuming they all agree, is less than the lower threshold, the record is dropped from further comparison (see Table 2). (The user can still specify a particular constant value for a cutoff if he choses to do so, but this will now be compared to the accumulated net total weight.)

Mathematical basis for the weights. - - The probabilistic approach to record linkage has been described earlier [6-9]. The extent or "factor" (F) by which a particular outcome from a given identifier comparison influences the overall odds is proportional to:

F = Frequency of that outcome in linked pairs / Frequency of the same outcome in randomly matched pairs.

To derive addable weights from such ratios one can convert to logarithms, and the base 2 is often used as in information theory. For a given pair of records there will be a series of weights (w_1, w_2, ----w_n) one for each of the successive identifier comparisons and their outcomes, and these may usually be summed to get a total weight for all of the identifier comparison outcomes together.

Since the overall odds in favour of a correct death linkage are influenced also by the likelihood of the patient being represented in the national file covering a specified period, and by the possibility that the death file for that period is large enough to produce purely fortuitously the observed combination of outcomes, these two factors must be taken into account as well. The complete formula is thus:

$$W^* = W + \log_2 \frac{N_A(L)}{N_A(\bar{L})} + \log_2 \frac{1}{N_B - 1}$$

where

W^* = $\log_2$ of the overall odds in favour of a correct linkage;

W = the sum of the individual weights for the various identifier comparisons;

$N_A(L)/N_A(\bar{L})$ = the ratio of linked/unlinked records in the cancer file after the search of the MDB over a specified period has been completed, or, the estimate numbers from file A (i.e. the Alberta cancer file) who will have died, and who will not have died, over the specified period covered by the death file, e.g. estimated from

TABLE 2. The computer "automatic" cutoff generated after each outcome

OUTCOME	ONAME	NEGWT[1]	REMWT[2]	CUTWT[3]
1	INIT11	-48	679	-700
2	INIT22	-83	642	-663
3	INIT12	-83	602	-623
4	INIT21	-83	563	-584
5	YRDIFF	-155	499	-520
6	MNDIFF	-217	465	-486
7	DYDIFF	-276	420	-441
8	XDATES	-276	385	-406
9	SURNAME	-347	385	-406
10	YRLKA	-431	385	-406
11	GIVEN11	-484	361	-382
12	GIVEN22	-537	337	-358
13	GIVEN12	-565	320	-341
14	GIVEN21	-593	303	-324
15	MARST	-593	291	-312
16	NCLASS	-593	106	-127
17	FILSIZE	-753	106	-127
18	AGE	-753	106	-127
19	KNDEAD	-813	106	-127
20	RESPROV	-833	86	-107
21	RESCDIV	-833	47	-68
22	DIAGTWN	-833	17	-38
23	ICDA	-844	0	-21
24	DIAGAGE	-844	0	-21
25	DTHRATE	-864	0	-21
26	DEATHYR	-864	0	-21
27	DEATHMN	-864	0	-21
28	DEATHDY	-864	0	-21
29	VSSRCE	-864	0	-21
30	LOWGHT	-864	0	-21

[1] NEGWT is defined as the maximum negative weight that can be accumulated up to and including a given outcome step.

[2] REMWT is the maximum positive weight that can conceivably be accumulated after a given outcome step.

[3] CUTWT is the accumulated negative weight at which there is no hope of exceeding or even reaching the lower threshold value set (here -21). In this example, there can be no cutoff until after the YRLKA outcome, since this is the first stage at which NEGWT can possibly exceed CUTWT.

appropriate survival tables on the basis of age and sex for a given single year of death;

N_B = the size of the national death file covering the same specified period.

The weights -- The principles on which a system of weighting factors is based are relatively simple, but the application of these principles necessarily involves arbitrary choices, and some approximations, so that the computer operation will not become too involved or time-consuming.

In the Alberta study, the derivation and refinements of the weights used for the surnames, birth dates, places of residence, and diagnoses deserve special mention. The kinds of detailed partial agreement and cross comparisons being used are discussed in [3].

To test the initial set of rules and weights, a sample of Alberta records with surnames starting with the initial "A" were linked to deaths for the period 1950-79. This test file was used to examine and refine the threshold values selected for acceptance or rejection of a linkage. The final values chosen were -20 and +50. This means that if the total accumulated weight fell below -20 the records pairs would be "rejected" and not included in the output; pairs which fall above +50 would be flagged as "definite" links, and those in the range -20 to +50 would be considered as "possibles". The system is very flexible and the values may be altered [10].

4. FUTURE PLANS - EVALUATION OF THE RESULTS

Following the automated death search, it is informative to compare the results of the automated search with those of the corresponding manual searches done routinely in the province. To carry out this comparison, one would like to tabulate the results first for Alberta deaths in the period 1953-78, and later for deaths outside Alberta. In each case the following three questions would be asked.

(1) How many matches were found by both the computer and manual searcher?

(2) How many matches were found only by the computer?

(3) How many were found only by the manual searcher?

Unfortunately, however, the Alberta file does not have the province or country of death recorded. For deaths not found by the computer, but found by the Alberta registry, an effort will be made to confirm the validity of these manual matches. But some of the deaths could have occurred outside of Canada and not be available on the MDB file.

In addition, it is possible to extract from the MDB file information on those cancer deaths of Alberta residents who were not represented on the Alberta Cancer Registry file.

5. LONG-TERM IMPLICATIONS -- FUTURE LINKAGES INVOLVING CANCER RECORDS

As cancer records become more readily linkable, and as familiarity is gained with what is involved in making a linkage operation successful, proposals for further epidemiological uses of cancer registry files, and improvement of the cancer files for these and other purposes, will become more common.

One particularly attractive idea is that centralized data on the diagnosis and treatment of cancer, already being collected routinely by various agencies for hospital and medical care insurance, might be used to enrich the cancer registries. In certain cases, existing records may even be used to create a cancer registry file, as is currently being carried out in Ontario [12]. The major problem here is that of having timely data; the obstacles to date are mainly of a technical, legal, and organizational nature.

For those concerned with associations between cancer risks and the prior circumstances of people's lives, linkages with records containing particulars of these prior circumstances have a special importance. Many ad hoc sources, such as company employment records and the records of patients in particular hospitals have been used on a modest

or medium scale as starting points for the follow-up of epidemiological cohorts. For example, it is proposed that the participants of the Nutrition Canada Survey of the early 1970's, which gave data dealing with nutrition for a sample of Canadians, be followed to determine the later risks of developing cancer. Similarly, a number of industrial cohorts, such as persons working in the nuclear industry, may be similarly followed. One study currently being planned is follow-up of employees of Eldorado Nuclear Limited using cancer registry files in addition to the Mortality Data Base [13].

A number of such epidemiological studies may be done with the Alberta cancer registry file once the death information has been added. One can, moreover, document the survival of patients after diagnosis with respect to various sites of cancer.

6. CONCLUSION

To sum up, refinements of searching, where there are no useable personal identity numbers, often requires a probabilistic approach. Certain features, such as the implementation of preliminary rejection rules, the automatic generation of cutoff values by the computer, the development of sets of weights for disease diagnosis, and the use of the total weight to reflect the absolute rather than the relative odds are some new procedures that have been developed, implemented and tested in the course of the Alberta Cancer Registry linkage study. Plans have been made to evaluate all of these more fully after the production jobs have been completed.

ACKNOWLEDGEMENTS

The authors would like to thank Leslie Gaudette, Sandra Swain, Christine Poliquin, Pierre Lalonde, Ted Hill, Gerry Hill and John Silins for their assistance in this project.

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

NOTES AND REFERENCES

[1] Smith M.E. and Newcombe, H.B., "Automated Follow-Up Facilities in Canada for Monitoring Delayed Health Effects", AJPH Vol. 70, No. 12, pp 1261-1268, 1980.

[2] Malhotra A., "A National Cancer Incidence Reporting System". A paper presented at the 13th International Cancer Congress, Seattle, September 8-15, 1982. Available from Statistics Canada - see footnote.

[3] Smith M.E., Newcombe H.B. and Dewar R., Proposed Procedures for the Alberta Cancer Registry Death Clearance, 1983. Copies available from Statistics Canada - see footnote.

[4] Smith M.E., Newcombe H.B. and Dewar R., The Use of Diagnosis in Cancer Registry Death Clearance, 1983. Copies available from Statistics Canada - see footnote.

[5] Smith M.E., Newcombe H.B. and Dewar R., Future Linkages Involving Cancer Records and Death Clearance Plus Detailed Recommendations for Data Collection in the Future, 1983. Copies available from Statistics Canada - see footnote.

[6] Newcombe H.B., Kennedy J.M., Axford S.J. and James A.P., "Automatic Linkage of Vital and Health Records". Science Vol. 130, pp 954-959, 1959.

[7] Felligi I.P. and Sunter A.B., "A Theory of Record Linkage", Journal of the American Statistical Association Vol. 64, pp 1183-1210, 1969.

[8] Howe G.R. and Lindsay J., "A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-Up Studies", Computers and Biomedical Research, Vol. 14, pp 327-340, 1981.

[9] Smith M.E. and Silins J., "Generalized Iterative Record Linkage System", American Statistical Association - 1981 Proceedings of the Section on Social Statistics, pp 128-137, 1981.

[10] Hill T., Generalized Iterative Record Linkage Sytem: GIRLS. System Development Division, Statistics Canada, Ottawa, September, 1981. (The complete document contains section on Glossary, Concepts, Strategy Guide, User Guide, and Weights, the last section by D.A. Binder.)

[11] Honeyman K.A., Taylor R.D., Gaudette L.A. and Grace M.A., "A Systems Design for Surveillance of a Chronic Disease: A Computerized Cancer Registry", Med. Inform. Vol. 2, pp 269-277, 1977.

[12] Clarke E.A. and Spengler R.D., "Cancer Incidence, Mortality and Treatment in Ontario". In: Cancer In Ontario 1981 p. 71-94. Available from: The Ontario Cancer Treatment and Research Foundation. 7 Overlea Boulevard, Toronto, Ontario. M4H 1A8, 1981.

[13] Newcombe H.B., Smith M.E. and Abbatt J.D., Linkage Procedures for the Eldorado Mortality Searches -- ENL-LINK-2. Available from: Eldorado Nuclear Limited, Suite 400, 255 Albert Street, Ottawa, Ontario. K1P 6A9, 1982.

[14] NYSIIS stands for New York State Identification and Intelligence System, a phonetic system for coding surnames. For more information, see: Lynch, B.T. and Arends, W.L., "Selection of a Surname Coding Procedure for the SRS Record Linkage System," Statistical Reporting Services, U.S. Department of Agriculture, Washington, DC., 1977.

Footnote: Copies of these papers may be obtained by writing to:

Miss Martha E. Smith, Head,
Occupational and Environmental
Health Research Unit, Vital Statistics and
Disease Registries Section,
Health Division,
Statistics Canada,
18-R, R.H. Coats Building,
Tunney's Pasture,
Ottawa, Ontario. K1A 0T6