RELIABILITY OF COMPUTERIZED VERSUS MANUAL DEATH SEARCHES IN A STUDY OF THE HEALTH OF ELDORADO URANIUM WORKERS **

H. B. NEWCOMBE*, M. E. SMITH†, G. R. HOWE‡, J. MINGAY§, A. STRUGNELL§ and J. D. ABBATT§ ‖

* P.O. Box 135, Deep River, Ontario, KOJ 1PO, Canada; † Vital Statistics and Disease Registries, Statistics Canada; ‡ National Cancer Institute of Canada Epidemiology Unit, University of Toronto; §Eldorado Nuclear Limited, Ottawa

Abstract—An epidemiological follow-up study of 16,000 uranium mine and refinery employees has made use of computerized techniques for searching a national death file. The accuracy of this computerized matching has been compared with that of corresponding manual searches based on one-eighth of the worker file. The national death file—Canadian Mortality Data Base—at Statistics Canada includes coded causes of death for all deaths back to 1950. The machine search was carried out using a generalized record linkage system based upon a probabilistic approach. The machine was more successful than the manual searchers and was also less likely to yield false linkages with death records not related to the study population. In both approaches accuracy was strongly dependent on the amount of personal identifying information available on the records being linked.

Uranium Radium Cancer Risks Follow-up Epidemiology
Industrial cancer Death searches Computer searches Automated follow-up

INTRODUCTION

Eldorado Nuclear Limited (E.N.L.) is conducting a retrospective epidemiological study of the health of its former employees. Eldorado operations involve the mining, milling and refining of uranium and these activities have been carried on continually from the early 1930s. Initially radium was extracted for medical and other purposes, and more recently uranium metal and nuclear fuel materials have become the main products.

The objectives of this study are:

(a) to identify former employees who may have a potential compensation claim, and to inform them or their survivors of these potential compensation claim rights, and

(b) to obtain dose-response data for evaluation of the risks to workers, especially with respect to atmospheres containing radon and radon-daughters.

The main study design and details regarding the assembly of the nominal roll have been described elsewhere [1]. The purpose of the present study, which serves both the short-term and the long-term aims of the broader investigation and of other similar studies, was to investigate the reliability of searches of all relevant death registration material using the study nominal roll and the Canadian Mortality Data Base (C.M.D.B.) operated by Statistics Canada. In an attempt to assess the reliability of machine record linkage for which the C.M.D.B. was designed [2, 3], the results of rapid computer searching and file linkage have been compared with manual searching and file linkage.

It has rarely if ever been possible to judge, much less quantify, how many false positive (incorrect) and false negative (missed) linkages result from conventional manual searches for death registrations where the dead or alive status of the members of the nominal roll is unknown. The present study is designed to provide quantitative information on both manual and machine file searching. The comparison has demonstrated the extent of the influence of an abundance or scarcity of personal identifiers on the efficiency of both types

**Reprinted with permission from Computers in Biology and Medicine, Vol. 13, No. 3, Copyright ©1983 by Pergamon Press Ltd., pp. 157-169.

Table 1. Manual matches of worker records with death records, by degree of assurance

Degree of assurance	Category	Number of worker records	
A	definite link	137	219
B+	very good possible	35	
B	good possible	47	
B−	unlikely possible	23	
C	poor possible	17	
D	not enough identification	10	
other	no link	1602	

From a sample of 1871 male worker records in which the surnames begin with the letters A or B.

of file matching. It has also demonstrated the greater efficiency of machine than manual matching.

The Eldorado study, although retrospective in nature, is being carried out with the intention of merging it into a prospective health monitoring instrument. It is the hope of many that similar prospective undertakings will come to be regarded in the future as desirable and feasible. Only thus can full use be made of available records to assess the adequacy of current standards of protection against delayed harm from the working experience.

MATERIALS AND METHODS

The Eldorado nominal roll used for the present study of linkage accuracy consists of a total of 16,658 names. These relate to past workers at the Port Radium mine (4526), Beaverlodge mine (9336), the Port Hope refinery (2514) and Research and Development (282), and involve employment as far back as 1932.

The Canadian Mortality Data Base file contains over five million death registrations with coded cause of death for the years 1950 to 1977.

For the computer linkage study, only E.N.L. records with a sex code equal to male or unknown (15,937) were used to initiate searches of the male half of the C.M.D.B. Searches for deaths relating to female workers (721) were not attempted because of the small numbers and the practical problems associated with changes of name at marriage. Such searches should be possible in the future, however, using the maiden surnames which occur on the death registrations of ever-married women, in the form of fathers' surnames.

For the manual linkage part of the operation, a sample of the E.N.L. file was used to initiate the searches representing all surnames of males beginning with the letters A and B (1871). A and B were chosen because they are known to provide a good sample of common and uncommon names (Andersons and Browns), and there is no evidence that they introduce a bias. The manual search used the C.M.D.B. microfiche listings.

The degree of assurance that a correct match has been achieved is assessed quantitatively by the computer. The decision is based upon prior information about the discriminating powers of various possible agreements and disagreements of the personal identifying information. The manual searchers assessed the degree of assurance subjectively and ranked the matches (links) they achieved on a scale that was qualitative (Table 1).

The principles are the same in both cases. Greater weight is attached to agreements of rare names, rare birthplaces, etc., than to agreements of their commoner counterparts. Similarly disagreements that occur only rarely, in a pair of records, argue more strongly against a correct match than will disagreements that are common. These fairly obvious inferences are taken into account by both the computer and the searcher. The chief difference is that the computer works from look-up tables that tell it by how much a given agreement, or disagreement, will shift the odds in favour of, or against, a correct match. The man relies on judgement with regard to the same matter, based on similar information and reasoning.

Table 2. Coincident identifiers in potentially matching worker records and death records (estimated)

Identifiers for searching and linkage	Percentage available in		
	Worker records alone	Death records alone	Both simultaneously (est.)
Surname plus at least one given name	100	100	100
plus a middle initial or name	50	47	23
Birth date in full	79	95	75
province or country	55	98	54
Parental initials, one or more	23	87	20
birth province/country, one or both	8	87	7

The system used for searching the death records was developed by Statistics Canada and the Epidemiology Unit of the National Cancer Institute of Canada for use in medical studies at Statistics Canada [4] and is described as a Generalized Iterative Record Linkage System (GIRLS). It is an extension of the probabilistic approach to record linkage developed at Chalk River [5–8]. Record linkage has been described in detail in numerous other publications (see references [9–13] and for a complete bibliography [14]). The mathematical derivation of 'weighting factors', from the frequencies of the various identifier comparison outcomes (agreements, disagreements, etc.), in linked vs unlinked pairs of records, has been described in detail elsewhere [4–7]. The weighting factors serve to represent in numeric form the discriminating powers of different identifier comparisons and their outcomes.

The assurances calculated by the computer are conveniently expressed on a logarithmic scale using the base 2 as in information theory. On such a scale, zero represents odds of 1:1 that the linkage is a correct one, each added unit doubling the odds and each subtracted unit halving them. For example, +1 and +2 represent odds of 2:1 and 4:1 respectively, in favour of a correct match; whereas −1 and −2 represent odds of 1:2 and 1:4 and so argue against a correct match. With an abundance of personal identifying information common to a pair of records, the evidence for or against a correct match tends to become more decisive, and stronger positive or negative 'weights', as they are called, are likely to be associated with the comparisons. Thus, for genuinely linkable pairs of records, total weights of +10 to +20 may be common, representing favourable odds of 1000:1 to 1,000,000:1. For unlinkable pairs, the weights and the odds will tend to be similar in magnitude but opposite in direction.

The degrees of assurance of a correct match, in both approaches, may be expected to vary widely. In large part this is due to differences in the amount of personal identifying information common to a potentially linkable pair (Table 2). For example, without the full birth date, the name information alone will usually not carry enough discriminating power to enable the correct death record to be selected from among a million or so others. And in part it is due to differences in the rarity or commonness of the names, birthplaces and such. Assurance is similarly affected whether the search is carried out by computer or by man.

A major purpose in performing the analysis of the data yielded by the combined efforts of the computer and the human searchers is to determine to what degree the accuracy of the death searches depends upon the amount of personal identifying information which can be applied to the problem of distinguishing good matches from bad.

RESULTS AND DISCUSSION

Assurances associated with the computer and manual searches

As a result of the computer search, approximately 2000 of 15,937 Eldorado worker records were linked to matching death registrations with varying degrees of assurance (Table 3). As a result of the manual search, somewhat over 200 of the 1871 records from

Table 3. Computer matches of worker records with death records, by degree of assurance

Weight range	Category	Range of odds (inferred from weights)	Number of worker records	
+4 and over	positive link	(11:1 and over)	1490	2023
+1 to +3	probable link	(1.4:1 to 11:1)	362	
zero	possible	(1:1.4 to 1.4:1)	171	
−1 to −3	probable non-link	(1:11 to 1:1.4)	794	
−4 to −8	positive non-link	(1:256 to 1:11)	2339	
other	no link	—	10,781	

From a total of 15,937 records where sex is male or unknown.

the sample (relating to surnames beginning with A or B) were similarly linked (Table 1). In each case, the precise number of 'acceptable' links depends upon where one sets the 'threshold' for acceptability. If one places it where the implied odds in favour of a correct match are 50:50 or better, either as calculated by the computer or as judged subjectively by the manual searchers, the precise number of 'acceptable' links would be 2023 and 219 respectively.

Because the setting of the threshold for acceptance is necessarily arbitrary in both cases, one must consider how best to estimate the numbers of accepted links that are in fact wrong, and the numbers of rejected matches that were correctly paired.

Estimating the false positive and false negative computer matches

There are two ways in which the accuracy of the computer linkages may be judged without reference to parallel manual searches. The first approach is based on the simple fact that where a worker's record links 'acceptably' to two different death records, only one of these links can be correct; the frequency of such instances tells us something about the potential for producing false positive outcomes. The second approach takes at face value the calculated odds, in favour of or against a correct match, and derives both an estimated number of false matches that lie above the threshold for acceptance, as well as another estimated number of potential correct matches that fall below the threshold for rejection.

Table 4. 'Runners up' as indicators of the potential for false positive linkages (computer searching)

Weight range	Range of odds (inferred from weights)	Number of worker records ('best' match for each)		Number of matches not the 'best' ('runners up')		'Runners up' (% of 'best')	
+10 and over	(724:1 and up)	1057	2023	10	325	1	16%
+4 to +9	(11:1 to 724:1)	433		64		15	
+1 to +3	(1.4:1 to 11:1)	362		150		41	
zero	(1:1.4 to 1.4:1)	171		101		59	
−1 to −3	(1:11 to 1:1.4)	794		680		86	
−4 to −8	(1:256 to 1:11)	2339		5053		216	

Note: (1) Weighting factors are rounded for simplicity, the precise dividing lines in the above table being +9.5, +3.5, +0.5, −0.5, and −3.5.

(2) In the '+10 and over' group, a substantial fraction carry weights in the region of +20 and even +30, representing odds of a million-to-one and a billion-to-one in favour of a correct linkage.

(3) Where such high weights occur among the 'runners up', which cannot be true links, they nevertheless correctly refer to similarities of identifying information which are exceedingly unlikely to have occurred by chance alone. Sometimes, such a pair of records will relate to two members of a family, one of whom was named after the other. Also, twins, who share the same birth date, are apt to turn up in such pairs of records, and so do members of small ethnic groups who share the same rare birth places and rare surnames. Manual searchers and the computer, both correctly tend to pay special attention to such non-random pairings of records, which signify correlations other than those due to the identity of the individual.

Table 5. Calculated 'weights' as indicators of probable false positives and false negatives (computer searching)

Weight range	Range of odds (inferred from weights)	Number of worker records ('best' matches)		Probable correct matches (est.)		Probable false matches (est.)	
+10 and over	(724:1 and up)	1057		1057		–	
+4 to +9	(11:1 to 724:1)	433	2023	424	1845	9	177
+1 to +3	(1.4:1 to 11:1)	362		279		83	
zero	(1:1.4 to 1.4:1)	171		85		85	
−1 to −3	(1:11 to 1:1.4)	794	3133	153	204	641	2929
−4 to −8	(1:256 to 1:11)	2339		51		2288	

Note: Whichever weight one chooses as representing a threshold for acceptance, those 'false matches' which fall above the threshold will become 'false positives', and those 'correct matches' which fall below the threshold will become 'false negatives'.

For the first approach, one may compare the numbers of 'best' matches with the numbers of 'runners up', broken down by the calculated 'weight' or odds in favour of a correct match (Table 4). The number of runners up increases with progressively lower weights. With the threshold for acceptance set just below zero, the 'runners up' (representing death records to which workers' records might have linked 'acceptably' if they hadn't found a better match) number sixteen per hundred 'best' matches. These are *potential* rather than actual false positives, but they indicate what might happen to the record of a worker who hadn't yet died and for whom there was therefore no correct matching death registration. This problem arises chiefly where the personal identifying information is limited.

For the second approach, the calculated weights (and their associated odds) were used to derive the probable numbers of links and non-links. For example, a weight of zero represents odds of 1:1 in favour of a correct linkage. Therefore half of the matches which have been assigned this weight, probably do relate to the same person and the other half do not. Taking the weighting factors at face value, the likely proportions of correct and false matches associated with each value of the total weights were calculated (Table 5). From this sort of calculation it was inferred that, for a threshold set just below zero weight, and with 2203 'accepted' links, 178 of these or just under 9% are likely to be false positives. In addition there are a probable 205 potential correct links that were not accepted, represent-

Table 6. Numbers of matches achieved by manual vs computer searching, by degree of assurance (based on worker records having surnames beginning with A or B)

Computer weight range	Degree of manual assurance						No man. match	Total
	A	B+	B	B−	C	D		
+10 and up	121	16	7	1	2	–	14	161
+4 to +9	13	8	9	1	1	–	21	53
+1 to +3	2	4	8	3	2	–	23	42
zero	–	1	3	1	–	–	11	16
−1 to −3	1	4	3	3	2	–	79	92
−4 to −8	–	1	9	10	5	9	266	300
no comp. match	–	1	6	5	7	–	1188	1207
Total	137	35	45	24	19	9	1602	1871

Note: (1) Where the thresholds for acceptance are set at zero and above for the computer, and at B and above for the manual searches, the following would be the result:
accepted by both = 192
accepted by computer only = 80
accepted by manual only = 25
rejected by both = 1574.

(2) The table includes cases in which the death record selected by the computer differs from that selected by the manual searcher (see next table).

Table 7. Computer – manual disagreements with respect to the death record selected (Parentheses indicate which were judged correct on subsequent review.)

Computer weight range	Degree of manual assurance						Total
	A	B+	B	B−	C	D	
+10 and up	–	1(M)	1(C)	1(C)	1(C)	–	4
+4 to +9	–	1(?)	1(C)	1(?)	1(C)	–	4
+1 to +3	–	–	1(C), 1(X)	1(?)	2(?)	–	5
zero	–	–	–	–	–	–	–
− 1 to − 3	–	–	–	–	1(?)	–	1
− 4 to − 8	–	–	1(?), 1(X)	3(?), 2(X)	2(?), 1(X)	2(?)	12
Total	–	2	6	8	8	2	26

Note: These numbers are included in the previous table.
M = manual choice correct
C = computer choice correct
X = both manual + computer choices incorrect
? = uncertain

ing a false negative rate of about 10%. If the threshold were raised to get rid of the false positives the false negatives would increase, and lowering the threshold would have the opposite effect. With the threshold in the vicinity of zero the number of false positives and false negatives are expected to be about equal. The only way to simultaneously reduce the frequencies of false positives and false negatives is to obtain a greater amount of personal identifying information for each record.

The human searcher is faced with the same problem, except that in this case it is not quantified. For both the man and the computer there may be additional false negatives that arise because some of the worker records are grossly deficient in identifying information; e.g. an absent birth date may result in insufficient discriminating power to distinguish between multiple possibilities for linkage.

Comparisons of computer vs manual linkages

Further insights into the respective levels of accuracy may be gained from comparisons of the performance of the computer vs that of a human searcher. Specifically, where the two approaches fail to agree, (a) they may yield different deaths, (b) the human may appear to succeed and the computer not at all, and (c) the reverse may be the case.

It might be supposed that the ultimate test of the accuracy of the computer searching would be for a man to carry out the same searches as the machine to see where the computer had gone wrong. This assumes, without evidence, that the man is more accurate than the computer. Instead, however, the problem is actually quite symmetrical, because lack of specificity in the identifying information adversely affects the accuracy of both the computer and the human searcher, and it remains to be shown which is the more accurate in the present setting.

Direct comparisons serve to indicate where the two approaches have yielded the same

Table 8. Proportions of worker records linked with death records by the computer, when birth year is absent vs present

Birth year* (present/absent)	Linkages (weights zero and over)	Worker records	% linked
Absent	18	3323	0.5
Present	2004	12614	15.9
Total	2022	15937	12.7

* Note: Virtually all of the worker records that lack year of birth, also lack the rest of the birth date.

Table 12. Calculation of 'weighting factors' for place of death vs place of work

Place of death	Number in linked pairs	Expected for average Canadians	Ratio (inferred odds in favour of linkage)	Weighting factor ($\log_2$ of the ratio)
Port Radium and Beaverlodge workers (145 pairs)				
Que.–Atlantic	8	53	1:6.6	−2.7
Ont.	30	52	1:1.7	−0.8
Man.–Sask.	19	12	1.5:1	+0.6
Alta.–B.C.	51	27	1.9:1	+0.9
Y.T.–N.W.T.	8	0.4	20:1	+4.4
Edmonton	27	3.5	8:1	+3.0
Port Hope workers (59 pairs)				
Que.–Atlantic	–	22	1:43	−5.4
Ont.	44	21	2.1:1	+1.1
Man.–Sask.	3	5	1:1.7	−0.8
Alta.–B.C.	12	11	1.1:1	+0.1
Y.T.–N.W.T.	–	–	–	–
Port Hope	20	0.05	400:1	+8.7

Note: (1) Where no death occurred, the ratio is based on an assumed 0.5 deaths; the resulting 'weighting factor' will then tend to be conservative.

(2) The expected numbers 'for average Canadians' are based simply on the populations of the regions.

unlinkable pairs argue against linkage.) The conversion of this ratio into a logarithm to the base 2 is just a convenience to make the weights addable. The first of the two frequencies is obtained by direct observation of the linked pairs of records, and the second is normally calculated from the frequency of the particular value of an identifier in the files themselves.

Examples are given of the use of such data as derived from the present study after its completion. These have to do with (a) simple disagreement weights (Table 10), (b) weights for a spectrum of outcome values ranging from complete agreement through various degrees of partial agreement–disagreement to complete disagreement (Table 11), and (c) weights for the occurrence in matched pairs of records, of identifier combinations which are correlated but cannot be regarded as either agreeing or disagreeing (Table 12). The latter two tables represent relatively fine groupings of the full range of possible outcome values. Such breakdowns are designed to avoid unnecessary pooling of outcomes with high and with low discriminating power, which would degrade the usefulness of the identifiers (rather as the usefulness of panned gold dust is degraded by re-mixing it with the sand).

The setting of the 'zero point' on the weight scale has proved more complicated than originally expected. This is the point at which the total weight for a matched pair of records indicates 50:50 odds in favour of, or against, a correct linkage. The total weight as initially envisaged did not take into account either the increased likelihood of chance similarities where the file being searched is particularly large, or the degree to which age and sex may influence the likelihood that an individual will be represented in that file where it is a death file. The hope was that the zero point could be adequately pinpointed by manual examination of borderline linkages. However, the present extensive work of this sort leaves one less confident about use of the manual approach alone, for this purpose. Substantial biases are now suspected, from a human tendency to reject out-of-hand those troublesome pairs which lack sufficient identifiers on which to base a judgement but might non-the-less be correctly matched. For a total of the calculated weights to represent 'absolute odds', as distinct from just 'relative odds', components are required which will take into account (a) the size of the death file over a given period, (b) the likelihood of an individual dying in that period, and (c) the likelihood of his being alive at the start of the period so as to be 'available' to die within the period. This approach is now being developed as a result of the need indicated by the present manual studies. And ways of estimating, and perhaps correcting for, any biases in the total weights arising out of this approach are being considered.

outcomes, and where they have differed. But judgements concerning which is the correct outcome when the approaches disagree are necessarily subjective, except where an actual oversight/error of some kind can be detected, or where additional identifying information can be obtained and used. The comparisons between the outcomes of the computer vs the manual searches that will be considered relate to the sample of 1871 Eldorado worker records in which the surnames began with A or B.

The degree of assurance of a correct linkage with a death record, or of a non-linkage, was variable both for the computer and for the manual searches. To a large extent, where the computer was 'very sure' that a correct decision had been made, so was the manual searcher, but the correlation is a fairly loose one when all degrees of assurance are considered (Table 6).

The conclusions one may draw from this comparison are best described in terms of a possible arbitrary threshold for 'acceptance' as a linkage, or 'rejection' as a non-linkage. Suppose, for example, that this threshold is set so that computer weights of zero and above, and manual assurances of B and above, are taken to indicate acceptable linkages. Then for 94% of worker records the outcomes from the two types of search both indicate either an appropriate linkage (192 cases or 10.3% of the records) or a non-linkage (1574 cases or 84.1% of the records).

For about 6% of the worker records the computer and the manual searcher were in disagreement as to whether an appropriate matching death record had been found (Table 6). If the results of the human searching are believed the computer approach resulted in 80 false positives and 25 false negatives (i.e. 4.3% and 1.3%, respectively, of the 1871 worker records, or, when based on the 219 manual linkages, 37% and 11% of the potentially linkable records). If the results of the computer searching are believed, the manual approach is similarly inaccurate and results in 25 false positives and 80 false negatives (out of 1871 worker records, or, when based on the 272 computer linkages, 9% and 29% of the potentially linkable pairs). This comparison serves chiefly to suggest that both approaches may involve considerable inaccuracy where the personal identification lacks discriminating power. And, of course, such comparisons cannot indicate how many relevant death records were missed by both kinds of searching.

There is evidence, however, that the computer searching results in fewer false negatives than does the manual searching. Thus, in Table 6 there are only seven cases of 'acceptable' manual matches of which the computer was apparently unaware, as against 69 cases of 'acceptable' computer matches of which the manual searchers were seemingly unaware.

Evidence that the computer is likewise less prone to the production of false positive linkages, may be obtained from those instances in which both approaches appeared to be successful but each identified a different death record as the appropriate one. For all 26 examples of disagreement of this kind, the source documents (E.N.L. work records and death certificates) were re-examined for additional information with which to resolve alternative choice 'matches' (Table 7). The resulting 'final' judgements are not infallible, but they do show that the computer is more reliable than the manual searchers where the two find different death records. The computer 'accepted' thirteen matches for the 26 ENL records, later judged to consist of six 'right', two 'wrong', and five 'doubtful'. The manual searchers 'accepted' just eight matches, later judged to consist of one 'right', five 'wrong', and two 'doubtful'.

From the above evidence, the computer searches appear to result in substantially fewer false positive and false negative outcomes than do the manual searches. Appropriate empirical tests and procedural adjustments will further improve the quality of machine linkage. Some of the proposed procedural changes will be described in what follows.

DISCRIMINATING POWER AS A LIMITING FACTOR

Since record linkage in the absence of unique identifier numbers depends upon multiple identifiers, it follows that discrimination decreases rapidly as personal identifying inform-

Table 9. Effects of differences in the availability of identifying particulars on the estimated proportions of false positives and false negatives (matched pairs with computer weights of zero and above being 'accepted' as 'linked')

Available identifiers	Number of matched pairs	Calculated false positives		Calculated false negatives	
		No.	% of accepted	No.	% of accepted
Year of birth, but not month and day					
Accepted	291	47.8	16.4	–	–
Rejected	805	–	–	54.2	18.6
Full birth date					
Accepted	1684	122.9	7.3	–	–
Rejected	2092		–	136.6	8.1
Birth date and place, plus two given names					
Accepted	166	4.8	2.9	–	–
Rejected	89	–	–	5.2	3.1

Note: (1) Columns headed 'No.' contain estimated numbers. They will therefore not be integers. For the method of estimation, see Section on 'Estimating the false positive and false negative computer matches'.

(2) For the purpose of this table an identifier is said to be 'available' as a basis for linkage when it is present on both a worker record and the death record to which it is matched, regardless of whether it agrees or disagrees.

(3) Where not specifically mentioned, an identifier may be either available or unavailable.

ation diminishes in abundance. In other words, the number of false negatives increases disproportionately as identifying information decreases.

Some indication of the quantitative importance of different amounts of identifying information may be gained from a few comparisons. For example, where information on birth year was present on the ENL records, some 16% were successful in finding a matching death record. But when it was absent, the success rate was only 0.5% (Table 8).

A better comparison involves three different levels of discriminating power in records that have the birth year (Table 9). 'Full identifying information' results in an estimated 3% of false positives and 3% of false negatives. Records reduced to birth date without place, etc., double both error rates to 7 and 8% each. Records with year of birth only again double the error rates to 16 and 19%. The comparisons are not precise, because different data sets are involved. But, in the absence of more elaborate and expensive tests, it would be unwise to disregard the practical guidance from such internally consistent evidence, of the need for multiple identifiers.

A redundancy of identifiers may be needed for a rather different reason. Strictly speak-

Table 10. Frequency of discrepancies in personal identifying information, and the 'weighting factors' derived from these frquencies (based on 269 matched pairs of worker and death records, with weights of zero and up)

Kind of identifier	Discrepant	Total linked pairs	Frequency in linked pairs	Weight for discrepancy ($\log_2$ freq.)
Surname spelling	12	269	1/22	−4.5
First initial	27	269	1/10	−3.3
First given name	74	268	1/3.6	−1.8
Second initial	19	119	1/6	−2.6
Second given name	18	65	1/3.6	−1.8
Birth province or country	7	114	1/16	−4.0
Parental initials	18	73	1/4	−2.0
Parental birth province/ country	11	25	1/2.3	−1.2

Note: For simplicity, the frequency of the discrepancy in unlinked pairs is taken to be virtually unity. Thus, $\log_2$ of the frequency in linked pairs approximates closely, $\log_2$ of the ratio of the frequencies in linked/ unlinked pairs.

Table 11. Calculation of 'weighting factors' for birthdate discrepancies

Degree of discrepancy	Number in linked pairs	Expected in unlinked pairs	Ratio (inferred odds in favour of linkage)	Weighting factor ($\log_2$ of the ratio)
Year of birth (268 pairs)				
0	170	2	85:1	+6.4
1	45	4	11:1	+3.5
2–3	38	8	5:1	+2.3
4–9	8	24	1:3	−1.6
10+	7	230	1:33	−5.0
Month of birth (243 pairs)				
0	219	20	11:1	+3.5
1	10	37	1:3.7	−1.9
2–3	8	64	1:8	−3.0
4–9	5	112	1:20	−4.3
10–11	1	10		
Day of birth (241 pairs)				
0	189	8	24:1	+4.6
1	11	16	1:1.5	−0.6
2–3	10	29	1:2.9	−1.6
4–9	17	76	1:4.5	−2.2
10+	14	112	1:8	−3.0

Note: The numbers expected in unlinked pairs are calculated as follows:

For exact agreements the expectation is taken to be n/n^2 times the number of matched pairs, where n is the number of different values of the identifier.

For discrepancies of degree d, the expectation is taken to be $2(n-d)/n^2$ times the number of matched pairs.

These equations represent approximations based on the assumption that the different values are equal in frequency. Where they are not equal, a more detailed calculation is required and this has been carried out in the case of year of birth.

ing, total weights reflect only the likelihood or unlikelihood that the observed similarity of identifying information on pairs of records has arisen other than by chance. But the ruling out of chance does not necessarily establish that the same person is involved:

Family members may be named after each other, and twins may be confused because of a common birthplace, birth date, and perhaps because of similar given names.

There are fashions in given names with small communities, and surnames repeat in localized ethnic groups and communities.

In short, similar or identical identifiers occasionally refer to attributes associated with particular groups of people, but not uniquely with any individual person.

The above kinds of problems can be minimized by abundant information, and to some extent by manual resolution using additional identifiers.

IMPROVING THE WEIGHTING PROCEDURES

The present manual/machine matching study has revealed needs for improvements in the weighting procedures used by the machine, and has provided some of the data required for the purpose. Such improvements would have to do in particular with (a) putting to use more of the potential discriminating power that could otherwise remain latent in the available identifiers, and (b) finding a better way of setting the 'zero-point' on the weighting scale.

The data used for calculating the weighting factors consists of the frequencies of various identifier comparison outcomes (agreements, disagreements, etc.) in pairs of records judged to be correctly linked, together with the corresponding frequencies for unlinkable pairs. Quite simply, the ratio between these two frequencies indicates the degree of assurance associated with a particular comparison outcome. (Outcomes that are more fashionable in linked pairs argue for linkage, and those that are more fashionable in

Table 13. Discrepancies of given names, by kind of discrepancies (based on 92 discrepancies of the first and second names combined, among 333 given names compared in record pairs with weights of zero and above)

Kind of discrepancy	Examples	
All discrepancies (92 cases)		
Position only, same spelling	(John – William John)	24
Different initial and name	(John – Fred)	16
Different spelling, same initial	(Louie – Louis)	52
Spelling discrepancies (52 cases)		
Vowel change only	(Ralph – Rolph)	15
Shortened only	(Fred – Frederick)	11
Nicknames, not just shortened	(John – Jack)	5
Phonetic similarities	(Ouide – Ovide)	4
Anglicizations	(Kenneth – Kazimie)	3
Double consonants	(Riser – Risser)	2
Other	(Bjom – Bjorvi)	12

Note: Of 46 disagreements of first or second initials, 11 were associated with simple reversals of the sequence on one of a matched pair of records as compared with the other (inversions), and 22 were due to one of the initials being transposed from first to second place (frame shifts).

Various other possible improvements in the weighting system, which will not be described here, are under development as a result of the present manual comparisons. Some of these have to do with (a) the handling of given name similarities where precise agreement is lacking (see examples in Table 13), (b) comparisons involving inverted sequences (e.g. of initials, and of birth month and day), and (c) practical means for making better use of the discriminating powers of very rare surnames, without recourse to excessively long look-up tables of weights.

IMPLICATIONS FOR ALL RETROSPECTIVE AND PROSPECTIVE STUDIES

Safety standards

(1) It is in everyone's interests to know where problems of safety are greatest and where they are least.
(2) Neither workers, management nor society in general benefit where undue emphasis is directed to non-problems, while real problems are neglected because they remain undetected.
(3) The limited public funds available earmarked for administration and enforcement of safety standards ought to be used so that attention to low-risk situations never results in the neglect of higher risks.

Fears about possible loss of privacy have tended recently to further reduce the specificity of personal identification on personnel records, notably on application forms for employment. At the same time, the public has increasingly demanded investigations of the delayed risks in various work situations, and has emphasized the right of the worker to know the risks.

To detect and measure delayed personal harm of almost any sort, and resulting from almost any kind of 'exposure', individual people require to be identified in a reasonably unambiguous fashion. This is true whether one follows exposed individuals forward to look for harm, or sick individuals backward in time to look for exposures. With both approaches, the most serious stumbling block is often a lack of sufficient specificity and redundancy in the personal identifiers (names, birth dates and such) by which people are known and represented on their various records, including their work records.

SUMMARY

Computerized searching of a national death file has been tested and compared for accuracy with the corresponding manual searches. The test formed a part of an

epidemiological follow-up study of some 16,000 former Eldorado employees, in which employment records are being used to initiate the searches for related death registrations contained in the Canadian Mortality Data Base at Statistics Canada. This facility includes the coded cause for all deaths back to 1950. The computer searching was guided by a generalized record linkage program, based on a probabilistic approach; the program was developed by Statistics Canada and the Epidemiology Unit of the National Cancer Institute of Canada. The corresponding manual searches used microfiche printouts from the Mortality Data Base tapes.

The results from the test showed the machine to be more accurate than the manual searchers. Not only was it more successful in extracting the relevant deaths, but it was also much less likely to yield false linkages with death records not relating to members of the study population. For both approaches, however, accuracy was strongly dependent on the amount of personal identifying information available on the records being linked.

Acknowledgments—The authors wish to thank Mr. J. Silins, Mrs. C. Poliquin and Mrs. M. Warner for their invaluable assistance. The advice and criticism of many other colleagues, particularly Mr. S. E. Frost and Mr. R. G. Dakers, is gratefully acknowledged.

REFERENCES

1. J. D. Abbatt, The Eldorado Epidemiology Project; Health Follow-up of Eldorado Uranium Workers. Eldorado Nuclear Limited, Ottawa, Ont. (1980). (Available on request by writing to Eldorado Nuclear Limited, 255 Albert Street, Suite 400, Ottawa, Ont. K1P 6A9.)
2. M. E. Smith and H. B. Newcombe, Automated follow-up facilities in Canada for monitoring delayed health effects, *Am. J. pub. Hlth* **70**, 1261–1268 (1980).
3. G. W. Beebe, Record linkage systems–Canada vs the United States, *Am. J. pub. Hlth* **70**, 1246–1247 (1980).
4. G. R. Howe and J. Lindsay, A generalized iterative record linkage computer system for use in medical follow-up studies, *Comput. biomed. Res.* **14**, 327–340 (1981).
5. H. B. Newcombe, J. M. Kennedy, S. J. Axford and A. P. James, Automatic linkage of vital records, *Science* **130**, 954–959 (1959).
6. H. B. Newcombe and J. M. Kennedy, Record linkage: making maximum use of the discriminating power of identifying information, *Communs Ass. Comput. Mach.* **5**, 363–566 (1962).
7. H. B. Newcombe, Record linking: the design of efficient systems for linking records into individual and family histories, *Am. J. hum. Genet.* **19**, 335–359 (1967).
8. M. E. Smith and H. B. Newcombe, Methods for computer linkage of hospital admission–separation records into cumulative health histories, *Meth. Inf. Med.* **14**, 118–125 (1975).
9. E. D. Acheson, Record Linkage in Medicine, E. and S. Livingstone, Edinburgh. (1968).
10. J. A. Baldwin, Linked medical information systems, *Proc. R. Soc.* **184**, 403–420 (1973).
11. P. Beauchamp, H. Charbonneau and B. Desjardins, La reconstitution automatique des familles; un fait acquis, dans la mesure des phénomènes démographiques, Homage à Louis Henry, Popul 1977, numéro spécial (mars 1977).
12. M. E. Smith, Record linkage of hospital admission–separation records, Chalk River Nuclear Laboratories, Chalk River, Ont. Publication No. AECL-4507 (Sept. 1973).
13. M. E. Smith and H. B. Newcombe, Accuracies of computer versus manual linkages of routine health records, *Meth. Inf. Med.* **18**, 89–97 (1979).
14. G. Wagner and H. B. Newcombe, Record linkage: Its methodology and application in data processing (a bibliography), *Meth. Inf. Med.* **9**, 121–138 (1970).

About the Author—HOWARD B. NEWCOMBE, B.Sc. (Acadia University 1935), Ph.D., D.Sc., F.R.S.C. Born 1914. Dr. Newcombe was a Research Scholar at the John Innes Horticultural Institute in 1939 and after wartime service as a Lieutenant, R.N.V.R., 1941–46, from 1947–79 was Head of the Biology Branch and later Population Research Branch, Atomic Energy of Canada Limited, Chalk River, Ontario. He was Visiting Professor of Genetics to the University of Indiana in 1963, Member of the International Commission on Radiological Protection and is the author of numerous scientific papers (mutations in microorganisms; effects of ionizing radiations; methods of study of human population genetics).

Dr. Newcombe is a Past President of the American Society of Human Genetics and the Genetics Society of Canada. At the present time he is Consultant to Eldorado Nuclear Limited and Statistics Canada.

About the Author—MARTHA SMITH received her B.Sc. from the University of Manitoba and her M.Sc. in Computing and Information Science from Queen's University in 1973. She was employed for several years in the Biology and Health Physics Division at Atomic Energy of Canada Limited, working with Dr. H. B. Newcombe on the British Columbia Record Linkage Study. This work involved developing new computer record linkage techniques for studying the effects of radiation on human populations. In 1978 she joined Statistics Canada and is currently Head of the Occupational and Environmental Health Research Unit. She is involved in planning and setting up some of the national files and facilities required to do long-term medical follow-up studies.

About the Author—GEOFFREY R. HOWE, B.Sc. (University College, London 1965), Ph.D. 1969. Born 1942. Dr. Howe was initially a Research Chemist with I.C.I. in England. He has subsequently been Research Fellow at Brock University and is now Senior Biostatistician to the N.C.I.C. Epidemiology Unit, University of Toronto. In addition, Dr. Howe is Professor in the Department of Preventive Medicine and Biostatistics at the University of Toronto and a Faculty Member of the School of Graduate Studies, University of Toronto. He is the author of numerous scientific papers, mostly on epidemiology and computerized record linkage.

He is a Fellow of the Chemical Society of London, Consultant to Eldorado Nuclear Limited and Atomic Energy of Canada Limited, Member of the American Statistical Association, Biometric Society and the Scoiety for Epidemiologic Research.

About the Author—JANE MINGAY received her Bachelor of Journalism degree in 1977 from Carleton University, having received practical experience in journalism. She subsequently worked on contract on a number of data collection and editing projects including medically oriented studies. After one of these projects organizing an historical research project for Eldorado Nuclear Limited, she became an Occupational Health Researcher on the E.N.L. Epidemiology Project. She is now on the staff of Maclean Hunter.

About the Author—ARLENE STRUGNELL graduated from business college in Montreal, Quebec in 1965 and worked as Secretary in the Department of Meteorology, McGill University until 1971. After working for a number of years in Toronto and Belleville, Ontario, she subsequently moved to Ottawa and is presently Research Assistant with the Epidemiology Project at Eldorado Nuclear Limited.

About the Author—JOHN D. ABBATT, B.Sc., M.B., Ch.B. (University of Edinburgh 1945), D.M.R., C.C.B.O.M. Born 1923. After wartime service with R.A.F.V.R. and hospital appointments in Edinburgh was Member of the U.K. M.R.C. External Scientific Staff at Hammersmith Hospital and Consultant Radiotherapist. After subsequent service as a Canadian Federal Civil Servant, he retired as Director General of Laboratory Centre for Disease Control, D.N.H.&W. and is now Medical Adviser to Eldorado Nuclear Limited, Ottawa.

Author of numerous scientific papers on the early applications of nuclear medicine and the therapeutic effects of radiation in man and animals, followed by epidemiological studies on human radiation effects.