Automatic Linkage of Vital Records*

Computers can be used to extract "follow-up" statistics of families from files of routine records.

H. B. Newcombe, J. M. Kennedy, S. J. Axford, A. P. James

The term *record linkage* has been used to indicate the bringing together of two or more separately recorded pieces of information concerning a particular individual or family (*1*). Defined in this broad manner, it includes almost any use of a file of records to determine what has subsequently happened to people about whom one has some prior information.

The various facts concerning an individual which in any modern society are recorded routinely would, if brought together, form an extensively documented history of his life. In theory at least, an understanding might be derived from such collective histories concerning many of the factors which operate to influence the welfare of human populations, factors about which we are at present almost entirely in ignorance. Of course, much of the recorded information is in a relatively inaccessible form; but, even when circumstances have been most favorable, as in the registrations of births, deaths, and marriages, and in the census, there has been little recognition of the special value of the records as a source of statistics when they are brought together so as to relate the successive events in the lives of particular individuals and families. The chief reason for this lies in the high cost of searching manually for large numbers of single documents among vast accumulations of files. It is obvious that the searching could be mechanized, but as yet there has been no clear demonstration that machines can carry out the record linkages rapidly enough, cheaply enough, and with sufficient accuracy to make this practicable.

The need for various follow-up studies such as might be carried out with the aid of record linkage have been discussed in detail elsewhere (*1*, *2*), and there are numerous examples of important surveys which could be greatly extended in scope if existing record files were more readily linkable (*3*). Our special interest in the techniques of record linkage relates to their possible use (i) for keeping track of large groups of individuals who have been exposed to low levels of radiation, in order to determine the causes of their eventual deaths (see *4*, chap. 8, para. 48; *5*), and (ii) for assessing the relative importance of repeated natural mutations on the one hand, and of fertility differentials on the other, in maintaining the frequency of genetic defects in human populations (see *4*, chap. 6, para. 36*c*).

Our own studies (*6*) were started as part of a plan to look for possible differentials of family fertility in relation to the presence or absence of hereditary disease (through the use of vital records and a register of handicapped children). The first step has been the development of a method for linking birth records to marriage records automatically with a Datatron 205 computer. For this purpose use has been made of the records of births which occurred in the Canadian province of British Columbia during the year 1955 (34,138 births) and of the marriages which took place in the same province over the 10-year period 1946-55 (114,471 marriages). Fortunately, these records were already in punch-card form as a part of Canada's National Index, and from them could be extracted most of the necessary information on names and other identifying particulars. An intensive study of the various sources of error in the automatic-linkage procedure has now been carried out on approximately one-fifth of these files.

Technical Problems

One of the chief difficulties arises from the unreliability of the identifying information contained in successive records which have to do with the same individual or married pair. The spellings of the surnames may be altered, the first Christian name on one record may become the second on another, and the birthplaces and ages may not be correctly stated. Much of the design effort must be directed toward ensuring that records can be linked in spite of such discrepancies, which in our files occurred with frequencies of about 10 percent of all record linkages involving live births and 25 percent of all linkages involving stillbirths.

A second problem relates to ambiguous linkage, in which it is uncertain whether or not a birth has arisen out of a particular marriage, or where there are two or more marriages any one of which might be that of the parents. These problems tend to occur when the husband's surname and the wife's maiden name are both common in the region studied, but they can also be associated with rarer family names, as in the marriage of two brothers to two sisters, and in certain racial minority groups. The difficulty increases with the size of the population under study.

At first sight these considerations might seem to preclude any extensive use of automatic record linkage as a source of statistics, since it is not at all obvious that the rules of judgment as exercised by a human being can be adapted to machine use. Also, partially mechanized record-linkage operations have proved laborious in the past (*7*).

Nevertheless, satisfactory procedures were eventually developed. These began with a series of small-scale attempts to link records visually, and thus to gain insight into the causes of any failures. The first of these studies was carried out at the Bureau of Statistics by one of us (S.J.A.) and made use of one of the standard phonetic name-coding systems to reduce the undesirable consequences of spelling discrepancies in linking records of sibling stillbirths. The gradual evolution of the method since that time has served to make it evident that further refinements can undoubted-

*Reprinted with permission from Science, Copyright 1959, by the American Association for the Advancement of Science, Vol. 130, No. 3381, October 16, 1959, pp. 954-959.

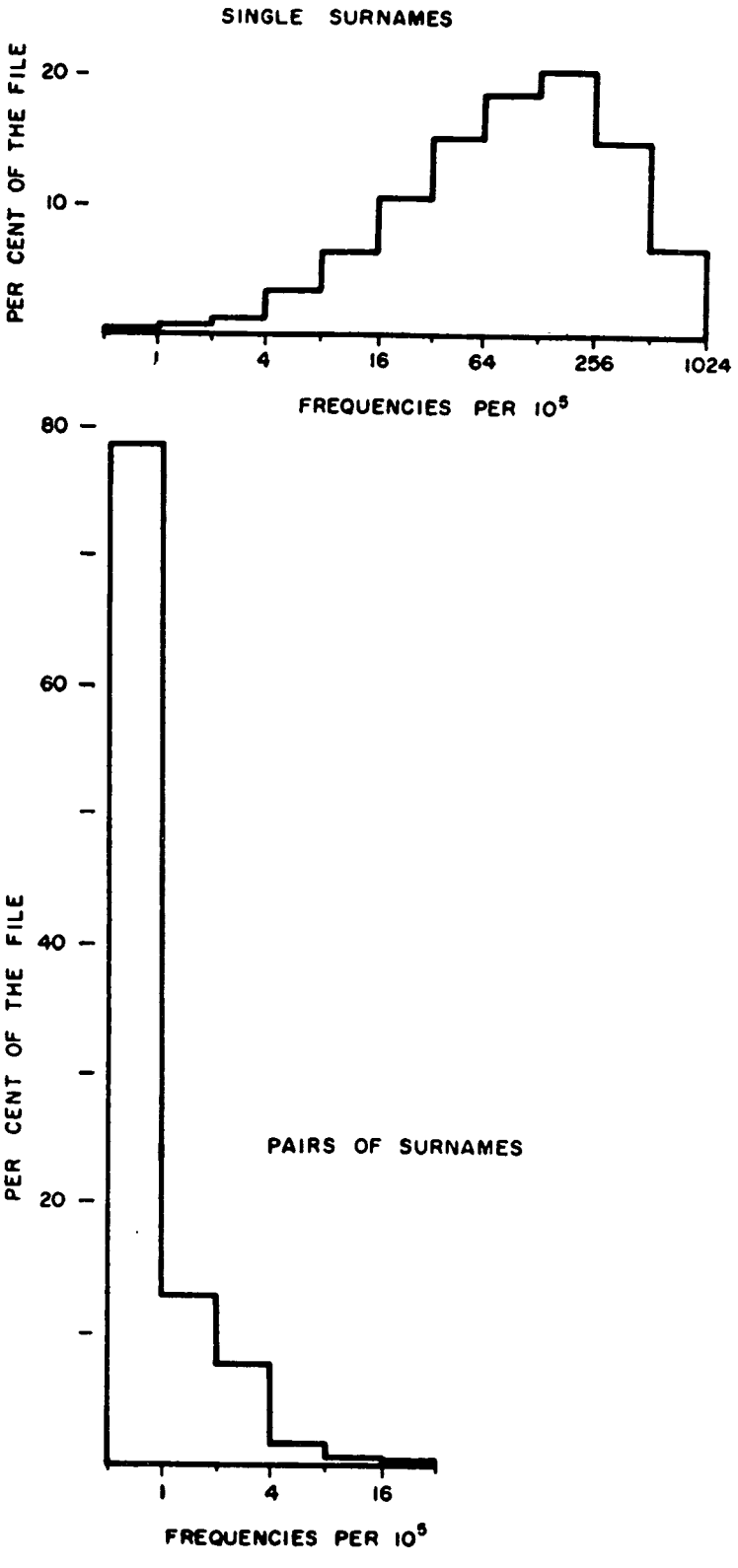

Fig. 1. (Top) Frequency distribution of brides' maiden names, in Soundex coded form, from records of 114,471 marriages in British Columbia for 1946–55. (Bottom) Frequency distribution of family-name pairs for married couples, in Soundex coded form, from the same records. Two East Indian names, of which one is customarily passed from mother to daughter and the other from father to son, were omitted. These occurred together in the same combination in approximately 100 marriages.

ly be developed and that no limit to the possible reliability of the linkages is yet in sight.

Methods

Of primary interest was the development of a procedure which would be fully automatic and free from piecemeal operations which might later limit the usefulness of the approach. This aim was achieved, chiefly because the use of a computer made it possible to compare each birth record in turn with all of the marriage records in appropriate sections of the marriage file. Since groups of marriages were sometimes scanned a number of times, it is apparent that this operation could not have been carried out with conventional card-handling equipment. Thus, without the computer, a visual search through printed lists would have been required to achieve some of the linkages.

To reduce the number of marriage records with which the computer must compare a birth record, it was decided to make use of both the husband's surname and the wife's maiden name, these being present on both the marriage and the birth cards. The surnames were first reduced to phonetic codes, consisting in each case of the first letter of the name followed by three numeric digits and known as the Russell Soundex Code (*8*), the computer being used for the coding operation. The codes served two purposes: They were designed to remain unchanged with many of the common spelling variations and in the present application were thus expected to bring together linkable records which would have been widely separated if arranged in a strictly alphabetic sequence. The coding also simplified the subsequent use of the Datatron computer, which is essentially a mathematical instrument and works more readily with numbers than it does with letters.

The extent to which two surnames are more efficient than one for identifying a family group has probably not been generally recognized. Thus, of the various brides' maiden names encountered in the marriage file, more than half recurred (in their coded forms) with frequencies in the range from 64 up to 1024 per 10^5. In contrast to this, nearly 80 percent of the pairs of family names (in their coded forms) were unique; that is, they occurred only once in our file in that particular combination, and extremely few had frequencies exceeding 4 per 10^5 (see Fig. 1). This

meant that we could mechanically compare each birth for the entire year with all of the marriages, using the same pair of surname codes, and that only rarely would the number of code matchings exceed one or two per birth.

To enable the computer to decide whether or not a birth and a marriage relate to the same married pair, use must be made of other identifying particulars. We relied chiefly on six items: the full alphabetic family names of the husband and wife (limited to nine letters each), their provinces or countries of birth (each coded as a two-digit number), and their first initials. In addition, the ages of the married pair were available on our cards for all of the birth records and for about half of the marriage records (that is, for marriages in the period 1951-56); the second initials were present in the case of the birth file; and the name of the city or place of the event (restricted to six letters) was available throughout both files.

As mentioned earlier, no one piece of information was entirely reliable. Usually it was obvious on inspection that the two events did, or did not, relate to the same married pair, but occasionally the decision was difficult. For this reason the computer had to calculate a probability that the couples were the same, or were different. The operation was performed automatically when the files were first matched.

The principle on which such a probability was based is fairly simple. If, for example, the province or country of birth of both the husband and wife agree on the two records, these facts may influence somewhat our belief that these records relate to the same married pair. Of course, the weight which one attaches to the information will be small if both have been born in the home province of British Columbia, but it will be large if they happen to have been born in, let us say, Switzerland and New Zealand, respectively. To give this a mathematical form it is necessary to know the frequencies for the various birthplaces of brides and grooms, and these can be determined quite readily either from published statistics or from the files themselves.

Similar reasoning can be applied to any item of identifying information, and to both agreements and disagreements. In order that the probabilities may be added together they must be converted to logarithms, and it is conventional practice in information theory to use logarithms to the base 2 of the probabilities expressed in the form of the "odds," for or against. The units are known as "binits." Thus, if the odds were 16 to 1 in favor of a genuine linkage, this would be represented as plus 4 binits, and odds of 16 to 1 against would be minus 4 binits. It is convenient to remember that a value of 10 binits is equivalent to odds of approximately 1000 to 1.

For present purposes, the probability or odds associated with a given agreement or disagreement may be obtained in binit units from the expression:

$$\log_2 p_L - \log_2 p_F \qquad (1)$$

where p_L and p_F are the frequencies with which the agreement or disagreement occurs, respectively, in the linked pairs of records and in pairs which have been brought together by accident. The expression will have a positive value in the case of agreement and a negative value in the case of disagreement.

As applied to agreements of initials and birthplaces, the expression can usually be simplified without any great loss of accuracy, since the particular letter or place should agree in the linked records almost as often as it appears in the individual records, and the chance of a fortuitous agreement will in most cases be approximately the square of this frequency. By substitution, expression 1 thus becomes:

$$\log_2 p_R - \log_2 (p_R)^2 = -\log_2 p_R \qquad (2)$$

where p_R is the frequency of the particular initial or birthplace in the individual records.

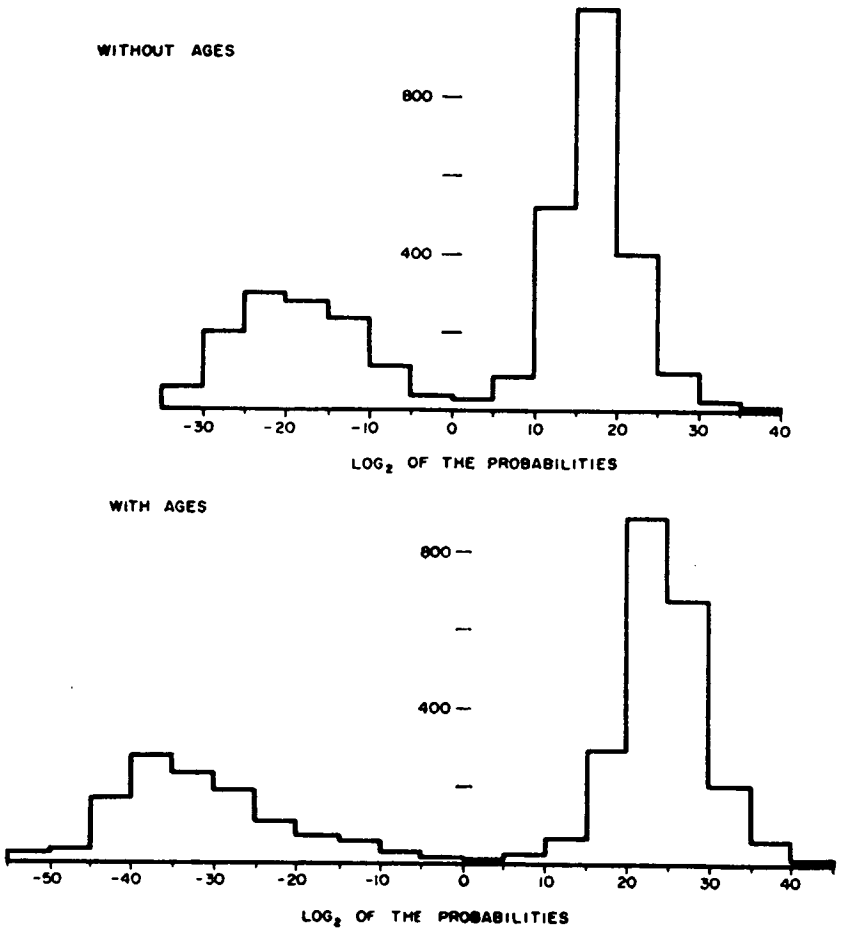

Fig. 2. (Top) Frequency distribution of the probabilities (in binits) obtained on comparing birth and marriage records having identical Soundex code pairs (calculated without using ages), based on records contained in the first fifth of the birth and marriage files (husband's surname beginning with *A*, *B*, or *C*). For this comparison only legitimate live births and marriages recorded in 1951-55 (a period for which ages are available) were considered. There were 2174 cases of genuine linkage and 1232 cases of accidental Soundex agreement. (Bottom) Same as above, except that the ages were used in calculating the probabilities.

The approach also lends itself to comparisons of the ages as stated on the two records, the lapse of time between the two events, and whether a discrepancy, if present, is slight or large, being taken into account. Even such an unlikely item as the place of the event can be used; if the marriage and the birth occurred in different places the fact carries little weight, but if they occurred in the same place (provided it was not the largest city in the province) the fact is important.

The items from which the probabilities were calculated in our study were the two alphabetic surnames, the two birthplaces, the two first initials, the two ages (where these were given on the cards), and the place of the event. For possible future use the computer also compared the birth order with the apparent duration of the marriage at the time of the birth, and wherever a first initial failed to agree, the computer looked for agreement between the first initial on the marriage record and the corresponding second initial on the birth record.

This sort of treatment can be adapted to linking almost any types of records where the information in common is sufficient for the purpose. Although tables of probabilities (in binits) containing over 300 items were used in the present study, they did not exhaust the capacity of the computer's memory unit. The limiting factor is the discriminating power inherent in the information supplied, and it is apparent that additional items of information can be of use even where they are of limited reliability.

The extent to which ages, for example, enable the computer to separate the genuine linkages from the fortuitous Soundex agreements can be seen from the data of Fig. 2. In this case, the number of record comparisons falling in the region from minus 10 to plus 10 binits, where the degree of certainty is less than 1000 to 1, is reduced by a factor of 3 when use is made of the additional information.

Reliability of the Linkages

Studies of the accuracy of the present computer-handling procedures indicate that about 98.3 percent of the potential linkages are detected in the existing record files, and that contamination with spurious linkages is 0.7 percent [see (9)]. This degree of accuracy is considered adequate for the statistical studies which have been planned, since the loss of such a small amount of data cannot in itself constitute a source of bias. Further, both the losses and the contaminations can be detected in the majority of cases by means of a subsequent check on the continuity of birth orders within families.

Variations in the spelling of the family names occur in about 4 to 5 percent of all linkages, but the losses from this source are reduced by the use of the phonetic codings to approximately a third of that value (see Table 1). The detection of such losses was accomplished by the simple expedient of resorting the files in a sequence which ignored the suspect code but trusted other identifying items, the files then being listed and examined visually. This operation could have been performed by the computer, and since the six main identifying items all agree in about 90 percent of the linked pairs of records (see Table 2), two additional arrangements of the files, each of which ignored one of the two Soundex codes, would be sufficient to reduce losses of this kind from the present 1.6 percent to about 0.16 percent. For the projected statistical studies such a procedure would hardly be worth while, the computer time being the limiting factor. It might become of value for other purposes, however, as computer speeds increase, especially as it is customary for central registry offices to keep two separate listings of marriages for searching purposes, arranged under grooms' surnames and brides' maiden names, respectively.

Failure of the calculated probabilities to make a correct distinction contributed a few additional losses and a few spurious linkages. These were detected by comparing the full Christian names as given on the original registration forms wherever the calculated probability fell within the range from minus 10 to plus 10 binits. Where age was used in calculating the probabilities there were only one loss and four spurious linkages from this source in a sample of over 2000 linkages (see Table 3). Although this degree of accuracy is adequate for almost any purpose, to make a further reduction in the number of spurious linkages would not be difficult.

Table 1. Surname spelling discrepancies*.

Name	Number of linkages in sample	Total spelling discrepancies		Discrepancies affecting the phonetic codes	
		No.	Percentage	No.	Percentage
Husband's surname	3622	41	1.1	15	0.4
Wife's maiden name	3501	115	3.3	42	1.2
Combined			4.4		1.6

* Based on visual linkages of births with marriages. To detect spelling discrepancies in a random assortment of the family names of one partner, use was made of the parts of the files in which the family name of the spouse began with *A*, *B*, or *C*. Thus, the two samples of records each represented approximately 19 percent of the total files.

Table 2. Discrepancies in birthplaces and first initials*.

Category	Number of linkages in sample	Discrepancies	
		No.	Percentage
Birthplace of husband	2174	22	1.0
Birthplace of wife	2174	21	1.0
First initial of husband	2174	60	2.8
First initial of wife	2174	83	3.8
Total			8.6
Total, including surnames			11.4
Linkages having discrepancies in one or more of the six items			10.3

* Discrepancies in computer linkages of records contained in the first fifth of the birth and marriage files (husbands' surnames beginning with *A*, *B*, or *C*); only linkages of legitimate live births with marriages in the period 1951-56 (for which ages were available) were used. For the "total, including surnames," use was made of the data from Table 1.

Table 3. Losses and spurious linkages due to lack of sufficient identifying information, which occurred in the linkage reported in Table 2 (9).

Item	No. of linkages in sample	Losses		Spurious linkages	
		No.	Percentage	No.	Percentage
Age data used	2174	1	0.05	4	0.23
Age data not used	2174	5	0.2	26	1.2

The contamination with spurious linkages will tend, however, to vary in direct proportion to the size of the marriage file with which the births are compared. Thus, in any future studies of larger populations it might be desirable to make use of additional identifying information. Christian names (perhaps restricted to four letters each), the city of birth of the husband and of the wife, respectively (likewise restricted to a few letters), and the province and year of marriage (not shown at present on the birth registration form) would all be suitable data for this purpose. The last of these three groups of items, however, would be of special value in effectively reducing the size of the marriage file with which any one birth would have to be compared, and in this manner reducing the false linkages. Occasional inaccuracies in the additional information would not greatly alter its usefulness in view of the nature of the handling procedures.

It is doubtful whether the present accuracy of the procedure can be matched by that of conventional survey and interview techniques, and its potential accuracy is certainly much greater than that of conventional techniques.

Speed of Record Linkage

By far the largest part of the effort in this undertaking has gone into the preparation of the card files. This has included, in the case of the marriage cards, a mechanical reproduction of the information contained in the existing National Index marriage cards for brides and for grooms, respectively, on a single card of our own format. Likewise, a part of the contents of our birth cards was obtained by reproduction from existing National Index birth cards, but in this case the maiden name of the mother and a number of other items were then added from cards which had been especially key-punched for the purpose. The family names on all cards in both files were Soundex coded by means of the computer, and the files were sorted into a Soundex sequence by pairs of codes, and listed. For the purpose of the initial record-linkage study the part of the marriage file for married pairs in which the groom's surname began with *A*, *B*, or *C* (approximately one-fifth of the total file) was transferred to magnetic tape.

This done, the computer made the necessary birth-to-marriage comparisons when presented with the birth cards, matchings with respect to the pairs of name codes being achieved at a rate of approximately one comparison every 3 seconds. About half of these code agreements represented genuine linkages (*10*). (Subsequently the whole of the birth and marriage files were put on magnetic tape and linked automatically by the computer.)

The initial steps would be largely eliminated were the format of the cards which are prepared routinely designed with a view to their possible use for record-linkage purposes. Also, an improvement in the rate at which the computer makes the comparisons can be gained in later operations by limiting the longer computations to the relatively small number of comparisons where simpler tests are inadequate. Some other short cuts might well be effected in the program if it were used sufficiently to justify the time involved. Such improvements can be thought of as reducing the cost of record linkage, in which computer rentals may be a major item, and of increasing the ease with which statistics can be derived from the linkage process.

The use of a computer especially designed to handle alphabetic information would further reduce the time required for the linkages by virtue of this special design alone, and there are larger computers in which the basic logical steps are more rapid by an order of magnitude. Thus, the present rate of something like one linkage every 6 seconds might be increased perhaps 20- or 30-fold—that is, to 200 or 300 linkages per minute, with existing equipment.

It is difficult to guess to what extent these speeds will be exceeded in the next 10 years or so. However, circuits have been described in the literature in which the basic logical steps take much less time than those in any equipment at present on the market (*11*). Research with the more novel kinds of electrical switching devices, some of which are not only fast but extremely compact, may extend the present limit by at least another order of magnitude (*12*).

Well before such equipment becomes available, however, it should be possible to develop the data-processing methods by which record linkages are achieved to the point at which the extraction of a wide variety of family and follow-up statistics becomes practicable from any records which are in an accessible form.

References and Notes

1. H. L. Dunn, *Am. J. Public Health* **36** (Dec. 1946); J. T. Marshall, *Population Studies* **1**, 204 (1947).
2. H. L. Dunn and M. Gilbert, *Public Health Repts. (U.S.)* **71**, 1002 (1956); H. B. Newcombe, in *Effect of Radiation on Human Heredity* (World Health Organization, Geneva, 1957); ———, A. P. James, S. J. Axford, "Family Linkage of Vital and Health Records," *Atomic Energy Can. Rept. No. 470* (Chalk River, 1957); H. B. Newcombe, S. J. Axford, A. P. James, "A Plan for the Study of Fertility of Relatives of Children Suffering from Hereditary and Other Defects," *Atomic Energy Can. Rept. No. 551* (Chalk River, 1957); H. B. Newcombe, A. P. James, S. J. Axford, "Genetic hazards and vital statistics," *Proc. Intern. Congr. Genet. 10th Congr., Montreal* (1958), vol. 2, p. 205.
3. S. C. Reed and J. D. Palm, *Science* **113**, 294 (1951); S. C. Reed, E. W. Reed, J. D. Palm, *Eugenics Quart.* **1**, 44 (1954); T. E. Reed, *Japan. J. Human Genet.* **2**, suppl., 48 (1957); ——— and E. L. Kelly, *Ann. Human Genet.* **22**, part 2, 165 (1958); A. B. Hill, R. Doll, T. M. Galloway, J. P. W. Hughes, *Brit. J. Prevent. & Social Med.* **12**, 1 (1958).
4. *Report of the United Nations Scientific Committee on the Effects of Atomic Radiation, Suppl. No. 17 (A/3838)* (United Nations, New York, 1958).
5. H. B. Newcombe, *Science* **126**, 549 (1957).

6. We are indebted to John H. Doughty for his encouragement and constructive criticism in the course of this work, to Robert J. Montgomery for making available facilities for the preparation of the marriage file, and to George Selby for his help in this initial operation. We would also like to thank Elizabeth Kinsey for collaborating in the preparation of the record files and in the analysis of the results, and Arden Okasaki for her work in programming the computer. Permission to use the vital records in this study was obtained through the Dominion Bureau of Statistics, from the Health Branch, Department of Health and Welfare, Province of British Columbia. The permission was conditional upon strict observance of the oath of secrecy respecting the nonstatistical information contained in the records.

7. S. Shapiro and J. Schachter, *Estadistica* **10**, 688 (1952).

8. The rules of Soundex coding are as follows. (i) The first letter of a surname is uncoded and serves as the prefix letter. (ii) W and H are ignored completely. (iii) A, E, I, O, U, and Y are not coded but serve as separators (see v below). (iv) Other letters are coded as follows, until three digits have been used up (the remaining letters are ignored): B, F, P, V, coded 1; D, T, coded 3; L, coded 4; M, N, coded 5; R, coded 6; all other consonants (C, G, J, K, Q, S, X, Z), coded 2. (v) Exceptions are letters which follow letters having the same code, or prefix letters which would, if coded, have the same code. These are ignored in all cases unless a separator (see iii above) precedes them.
9. Since ages were available on only about half of the marriage cards, the average losses from this cause were 0.12 percent of all linkages, and the average spurious linkages were 0.7 percent. When these are added to the losses resulting from the Soundex discrepancies, as shown in Table 1, the total loss is 1.72 percent.
10. It is known that approximately 19 per cent of the surnames in the marriage file begin with A, B, or C, as determined from studies of the frequencies of brides' Soundex codes. Thus, of the 114,471 marriage records and 34,138 birth records, approximately 21,750 and 6500 records, respectively, were used in the initial linkage study. In all, 6375 comparisons (3484 with positive binit values and 2891 with negative) between birth records and marriage records having identical pairs of Soundex codes were made by the computer. Of these, 418 (20 positive and 398 negative) related to illegitimate births, 2549 (1285 positive and 1264 negative) related to legitimate births and to 1946-50 marriages, and 3408 (2179 positive and 1229 negative as determined by means of ages) related to legitimate births and to 1951-55 marriages. Since age records were available in the case of the 1951-55 marriages,

this latter group of 3408 comparisons was used for a detailed study of the reliability of the machine linkage process. (Revised tables of binit values were also derived from these comparisons.) Two of the 3408 comparison cards were removed because in each case one of the ages was missing. Of the remaining 3406 cards, 2174 represented genuine linkage (2173 positive cards plus one negative card) and 1232 represented accidental Soundex agreements (4 positive plus 1228 negative cards), as judged by comparisons of the full Christian names in all cases where the binit values fell within the range from minus 10 to plus 10. It will be noted that of the 6500 births of 1955 which were studied, 3484 (54 percent) were from marriages contracted in British Columbia during the 10-year period 1946-55. For a description of the manner in which visual record linkages (as distinct from computer linkages) were used to assess the losses due to spelling discrepancies, see footnote to Table 1.

11. R. M. Walker, D. E. Rosenheim, P. A. Lewis, A. G. Anderson, *IBM J. Research and Develop.* 1, 257 (1957).
12. R. F. Rutz, *ibid*, 1, 212 (1957); D. A. Buck, *Proc. I.R.E. (Inst. Radio Engrs.)* 44, 482 (1956); J. W. Crowe, *IBM J. Research and Develop.* 1, 295 (1957).

Editors' Note: In 1959 Dr. Newcombe and Dr. James were affiliated with the biology branch of Atomic Energy of Canada, Ltd., Chalk River, Ontario. Dr. Kennedy was affiliated with the theoretical physics branch of Atomic Energy of Canada. Dr. Axford was affiliated with the health and welfare division of the Dominion Bureau of Statistics, Ottawa.