Record Linking: The Design of Efficient Systems for Linking Records into Individual and Family Histories

HOWARD B. NEWCOMBE

Biology Branch,
Chalk River Nuclear Laboratories,
Chalk River, Ontario.

INTRODUCTION

THE APPLICATIONS of computer technology to genetic problems discussed so far in this Supplement make use, primarily, of the ability of the machines to carry out involved mathematical procedures. In contrast, the application which I shall describe uses the computer as a kind of filing clerk. The task given it is that of building family histories of births, marriages, procreations, deaths, and ill health from the individual registrations of these events, and of doing so on a substantial scale.

Although the computer is at no point asked to carry out any mathematical operation more complicated than simple addition and subtraction, it must nevertheless perform a function that is much more unconventional for machines. It is required to simulate the judgment of a human clerk who attempts to file correctly the incoming correspondence from people who are careless about the way they spell their family names, who may sometimes use their middle names as if these were their first, and who may be writing from places that are not their usual addresses.

Provided that a computer can be instructed to carry out an operation of this kind with a degree of accuracy similar to that of a human filing clerk, the special talent which it may be expected to apply to the task is its speed. Current experience with this sort of computer application is particularly encouraging, in terms of accuracy, speed, and cost, and the capabilities of the machines will undoubtedly increase as time goes on. Thus, it is not unrealistic to think of integrating, in due course, some major fraction of the routine personal documentation dealing with reproduction and health into the form of individual and family histories.

CONCEPTS

A number of concepts will be discussed that are inherently simple, but the implications of these concepts will not necessarily be self evident.

The idea of linking records, for example, is particularly simple—the phrase *record linking* just means bringing together information from two independent sources about the same person—but with successive linkings the information may take on the characteristics of a collection of personal or family histories.

AMERICAN JOURNAL OF HUMAN GENETICS, VOL. 19, NO. 3, PART I (MAY), 1967

Even such familiar file upkeep operations as the insertion of address changes into a mailing list are elementary forms of record linking. However, the process as applied to human genetics will involve successive linkings of routinely collected records of procreative and health events to derive, eventually, multigeneration pedigrees for whole populations.

The two principal steps in any linking operation, namely, those of searching out the potentially linkable pairs of records for detailed comparison and of deciding whether or not a given pair is correctly matched, are commonplace in almost any operation by which a file is kept up-to-date. However, both of these steps, if they are to be carried out efficiently by machines, involve the use of stratagems of kinds that are employed almost unconsciously by a human filing clerk. For the *searching step,* the aim must be to reduce the number of failures to bring potentially linkable records together for comparison, such as may occur as a result of discrepancies in the file sequencing information, but this must be done without resorting to excessive amounts of additional searching. For the *matching step,* the problem is that of enabling the machine to apply in numerical form the rules of judgment by which a human clerk would decide whether or not a pair of records relates to the same person when some of the identifying information agrees and some disagrees.

Similarly, the idea of arraying pedigree information in linear fashion to facilitate storage, updating, and retrieval by machines using magnetic tapes as the storage medium is simple and by no means new. Nevertheless, the forms which such linear arrays may take bear little resemblance to the conventional pedigree charts with which geneticists are most familiar. The great flexibility of the *linear pedigrees* and the ease with which family relationships of unlimited complexity may be represented in such a fashion are, for this reason, not generally appreciated. In comparison, however, the usual two-dimensional representations are exceedingly cumbersome (Fig. 1).

Finally, it has not been uncommon in the past to derive partial histories of individuals and families from the *routine vital and health records,* on a small scale, by manual means. However, the idea that some substantial fraction of these enormous files might be so organized and that we are at the point now where this would be technically feasible and not too expensive is one that has been slow in gaining acceptance. Nevertheless, the inherent possibilities are beginning to be recognized. A colleague of mine is reported to have remarked recently that we are still using old data on hemophilia, that there are many hemophiliacs in Canada, almost all of whom will wind up in a computer sooner or later, and "what a shame if it is only opposite a dollar sign."

The concepts may not be new, but such implications are.

METHODS OF RECORD LINKING

The two essential steps in the linking of records by computer, that is, the *searching* step and the *matching* step, have precise counterparts in many manual filing operations. Although the accuracies of such operations and the times required are generally regarded as important, it is unusual to judge the efficiencies in numerical terms or to set down the conditions under which

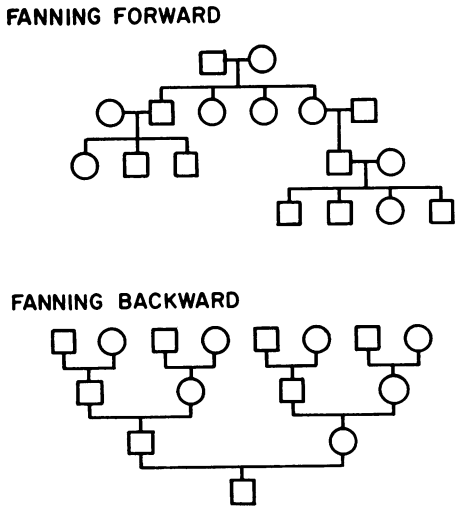


FIG. 1. Conventional pedigree charts. Note the difficulty of representing in a single chart the ancestors, descendants, cousins, and in-laws.

an optimum balance may be achieved between the level of accuracy and its cost as indicated by time required to achieve that level. Where such an undertaking is to be carried out on a very large scale by a computer, however, some thought may profitably be given to the efficiency of the operation in these terms.

1. *Optimizing the Searching Step*

In the case of the searching step, errors in the form of failures to bring potentially linkable pairs of records together for comparison could be reduced to zero simply by comparing each incoming record with all of the records already present in the master file. Where the files are large, however, such a procedure would generally be regarded as excessively costly in terms of the enormous numbers of wasted comparisons of pairs of records that are unlinkable.

For this reason, it is usual to arrange the file in some orderly sequence, using identifying information that is common to both the incoming records and those already present in the master file. Detailed comparisons then only need to be carried out within the small portions of the master file for which the sequencing information is the same as that on the incoming records (Fig. 2). For many purposes, it is common practice to use the alphabetic surnames and first given names for sequencing a file of personal records. The price that must be paid for the saving of time is an increase in the failures to bring potentially linkable pairs of records together for comparison, owing to discrepancies in the sequencing information on pairs that in fact relate to the same person. However, different kinds of information that might be used for the sequencing differ widely, both in their reliability and in the extents to which they subdivide a file.

Although alphabetic surnames are commonly employed, they are not particu-

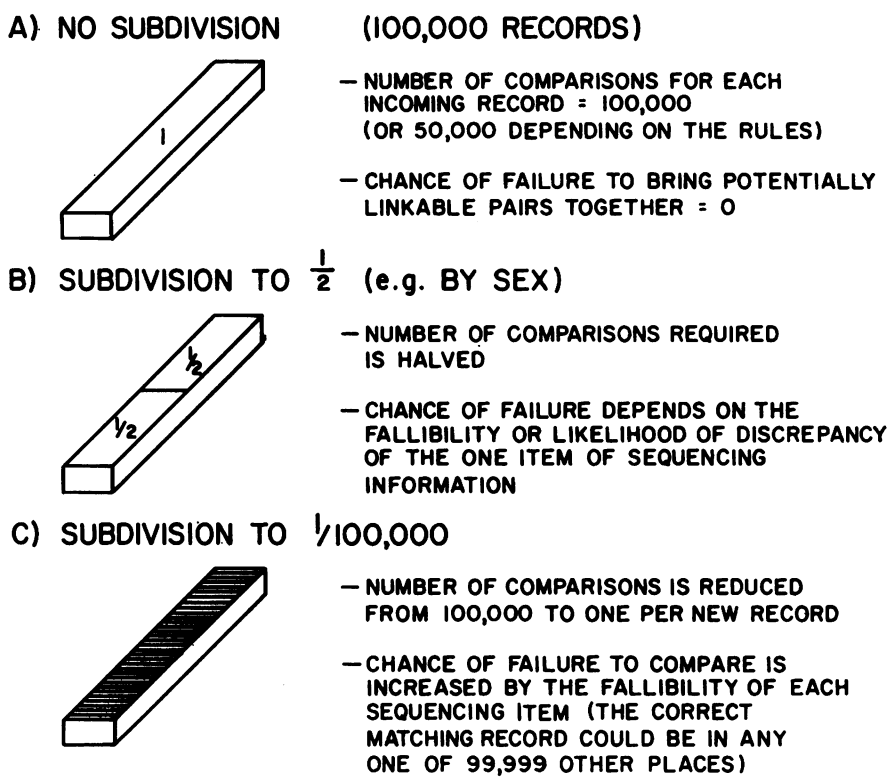


FIG. 2. Optimizing a single sequence search. Subdivision must be based on items of identifying information with the highest efficiency ratios and must be adjusted to an acceptable low level of losses or of wasted comparisons.

larly efficient for sequencing, because of the high frequency with which they are misspelled or altered. Considerable improvement can be achieved by setting aside temporarily the more fallible or labile parts of the information which the surnames contain, while retaining as much as possible of the inherent discriminating power. There are a number of systems for doing this, the most common of which is known as the Russell Soundex code. This is essentially a phonetic coding, based on the assignment of code digits which are the same for any of a phonetically similar group of consonants. (Details of a number of such surname coding systems are given in the Appendix.)

In practice, we have found that the Soundex code remains unchanged with about two-thirds of the spelling variations observed in linked pairs of vital records, and that it sets aside only a small part of the total discriminating power of the full alphabetic surname. The system is designed primarily for Caucasian surnames, but works well for files containing names of many different origins (such as those appearing on the records of the U. S. Immigration and Naturalization Service). This particular code is less satisfactory, however, where the files contain names of predominantly Oriental origin, because much of the discriminating power of these resides in the vowel sounds which the code ignores.

Any kind of identifying information that is available on all of the records may, of course, be used for sequencing the files, and it should not be assumed that surnames necessarily possess special merit for this purpose. The qualities required are reliability and discriminating power, both of which may be measured numerically. Usually, where the discriminating power of any one kind of information alone is insufficient to divide the file finely enough, two or more kinds of information may be used together to achieve a required degree of subdivision. However, each additional kind of information carries its own likelihood of discrepancy and thus contributes to the over-all tendency for the sequencing information to be reported differently on successive records relating to the same person, with a resulting increase in the frequency with which potentially linkable records will fail to be brought together for comparison. It is important, therefore, to choose the most appropriate kinds of information from among those that are available.

Fortunately, there are numerical tests which will indicate the relative merits of the different items of identifying information for the purpose of sequencing the files. Three values will be discussed, the *coefficient of specificity,* the *discriminating power,* which is simply another way of describing the specificity, and a so-called *merit ratio,* which may be used to indicate the amount of discriminating power per unit likelihood of discrepancy. This latter value can be used in selecting the most appropriate information to be employed in sequencing a file.

The fineness with which a file will be divided by a particular kind of identifying information may be represented by a single number, the *coefficient of specificity,*

$$C_s = \Sigma P_x^2 \tag{1}$$

where P_x is the fraction of the file falling in the xth block (see Fig. 3). C_s may be thought of as the fraction of the file falling within a block of strictly representative size. Since most identifying information divides a file unevenly into a mixture of small and large blocks, it is convenient to be able to indicate the effective degree of division of the file in this simple manner.

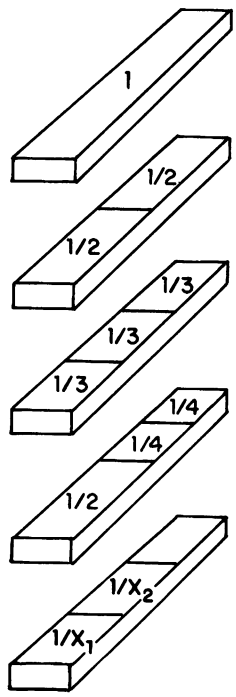
Unlike the coefficient of specificity, which gets smaller as a file becomes more finely divided, the *discriminating power* increases with the extent of the subdivision. Furthermore, it is usually regarded as an "addable" quantity. Thus, the discriminating power may be taken as the logarithm of the inverse of the coefficient of specificity, and in practice we have found it convenient to use logarithms to the base two (see Table 1):

$$D_p = \log_2 (1/C_s) \tag{2}$$

Finally, the merit of any particular kind of identifying information for sequencing the files may be taken as the ratio of the discriminating power to the likelihood of discrepancy or inconsistency of such information in linkable pairs of records:

$$M_t = D_p/I \tag{3}$$

In calculating this so-called *merit ratio,* we normally use the percentage likelihood of inconsistency as the numerical value of I.



$C_s = (1/1)^2 = 1/1 = 1$

$C_s = (1/2)^2 + (1/2)^2 = 2/4 = 1/2$

$C_s = (1/3)^2 + (1/3)^2 + (1/3)^2 = 3/9 = 1/3$

$C_s = (1/2)^2 + (1/4)^2 + (1/4)^2 = 6/16 = 1/2.7$

$C_s = (1/x_1)^2 + (1/x_2)^2 + \ldots\ldots = \Sigma P_x^2$
(where P_x is the proportion in the Xth block)

FIG. 3. Examples of coefficients of specificity.

TABLE 1. RELATIONSHIP OF COEFFICIENT OF SPECIFICITY AND DISCRIMINATING POWER

Coefficient of specificity $C_s = \Sigma P_x^2$	Discriminating power $\log_2(1/C_s)$	Equivalent number of blocks if file equally divided
1	0	$2^0 = 1$
1/2	1	$2^1 = 2$
1/4	2	$2^2 = 4$
1/8	3	$2^3 = 8$
1/16	4	$2^4 = 16$
1/1024	10	$2^{10} = 1024$
$1/10^6$	20	$2^{20} = 10^6$

The most efficient sequencing of a file will be based on the items of identifying information that have the highest merit ratios, using enough different items to achieve a combined discriminating power that will subdivide the file to the required degree of fineness. In this manner, the minimum total likelihood of discrepancy or inconsistency will have been introduced into the sequencing items for any required degree of subdivision.

By means of such numerical values, the usefulness of surname information in its Soundex coded form can be shown to be considerably greater than

TABLE 2. RELATIVE MERITS OF ALPHABETIC VERSUS SOUNDEX CODED SURNAMES FOR SEQUENCING FILES

Surname information	Discriminating power D_p	Equivalent number of blocks of equal size $1/C_s$	Percentage likelihood of discrepancy* I	Merit ratio $Mt = D_p/I$
Alphabetic	+9	512	2.2	4.1
Soundex	+8	256	0.8	10.0
Residual	+1	2	1.4	0.7

*Average for husbands' and wives' birth surnames.

that of the full alphabetic surnames for the purpose of sequencing the files, the merit ratio being about two or three times as large (Table 2). The residual information that is omitted from the Soundex codes is of very low quality indeed, having a merit ratio that is less than one-tenth that of the Soundex codes.

The approach permits the searching step of a linkage operation to be optimized, in terms of the numbers of (1) wasted comparisons to which an incoming record must be subjected in order to be brought together with a potentially linkable counterpart from the master file, and (2) failures to bring such records together. A tolerable level may be set for either the wasted comparisons or the failures, and the other value may then be minimized. Adjustment is achieved by adding or deleting an item from the sequencing information, thus increasing or decreasing the fineness of subdivision and the errors simultaneously until the required balance is struck. At no time should the sequencing information include an item with a lower merit ratio where one with a higher ratio is available. The cost of the searching step is thus balanced against its precision with a view to getting the best possible bargain.

In practice, we have found that by sequencing a master file of 114,000 marriage records in order of the pairs of surname codes for the grooms and brides, the number of wasted comparisons was kept at a very low level, i.e., 0.6 per incoming birth record where the births had arisen from marriages represented in the master file and 1.6 for all other incoming birth records. The number of failures to bring potentially linkable records together for comparison due to spelling discrepancies that altered one or other of the Soundex codes amounted to 1.6% of the potentially possible linkages.

The discussion so far has assumed that all of the linkings will be carried out using files arranged in a single sequence. However, the cost of sorting by computer is rapidly diminishing. Where more than one sequence is permitted, an even better bargain may be struck in terms of the precision that can be achieved for any given number of wasted comparisons. Linkings may then be carried out using very fine subdivisions of the file sequences, based on information of quite limited reliability, with the assurance that potentially linkable pairs of records which are not brought together on the first search will be compared in one of the alternative sequences based on other identifying information.

One quite large manual test of such a procedure has been carried out in

TABLE 3. IDENTIFYING INFORMATION ON VITAL RECORDS

Event and individual	Birth name	Birth-place*	Birth date (or age)
Marriage			
Groom	+	+	(+)
Bride	+	+	(+)
Father of groom	+	+	
Mother of groom	+	+	
Father of bride	+	+	
Mother of bride	+	+	
Birth			
Child	+	+	+
Father	+	+	(+)
Mother	+	+	(+)
Death			
Deceased	+	+	+
Spouse	+		
Father	+	+	
Mother	+	+	

*i.e., city or place, and province or country.

which initials and provinces of birth were substituted in the secondary sequences for one or other of the two surname codes. This test showed that a reduction in errors by more than tenfold could be achieved at the price of a two- to three-fold increase in wasted comparisons.

Where the avoidance of "lost" linkages is of special importance, the use of multiple alternative sequences represents an ultimate in refinement.

2. *Optimizing the Matching Step*

When pairs of records are brought together for comparison, decisions must be made as to whether these are to be regarded as linked, not linked, or possibly linked, depending upon the various agreements and disagreements of items of identifying information. It is also desirable that such decisions be based on numerical estimates of the degrees of assurance that the records do or do not relate to the same persons. The computer is asked, in effect, to simulate the processes of human judgment and to make the best use it can of the items of identifying information that are individually unreliable but collectively of considerable discriminating power.

The extent of the personal information that is usually entered in the vital registration makes the potential accuracy of the linkings of these records high indeed. Newborn children, grooms and brides, and deceased persons are commonly identified by their full birth names, their birth dates or ages, and their birthplaces. Together with this personal identification, there is a substantial amount of family information. The full names of the parents, including the maiden surname of the mother, are usually given, as well as their birthplaces. In addition, the ages of married couples are entered in the records of their marriages and the records of the births of their children (Table 3).

Thus, there is an abundance of overlapping information that may be used to link (1) deaths to births, (2) births to the parental marriages and to the births of older siblings, and (3) marriage records of brides and grooms to their birth records, to the marriage records of their parents, and to the birth and marriage records of their siblings (Table 4). Even where some of the items fail to agree, the combined discriminating power of such information is almost always large.

A human filing clerk attempting to carry out such a grouping operation would intuitively attach greater positive weight to some of the agreements than to others and greater negative weight to some of the disagreements than to others. In each instance, the question that is asked, almost unconsciously, is, "Would such an agreement be likely to have occurred by chance if the pair of records *did not* relate to the same person?" or "Would such a disagreement be likely to have occurred by chance if the pair of records *did* in fact relate to the same person?" The answer in each case will depend upon prior knowledge gained from experience. An initial known to be rare, such as "Z," will be regarded as less likely to agree by chance on a pair of records than would a commonly occurring initial such as "J." Similarly, a highly reliable and stable item of identification, such as sex, when it fails to agree, will argue more strongly that the people referred to are *not* the same than would, for example, disagreement of province of birth, which is known from our own experience to be discordant in about one per cent of genuinely linked pairs of records.

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

If we wish to obtain numerical weights that can be added to other such weights, the above ratio may simply be converted to a logarithm. In practice, the logarithm to the base two has proved particularly convenient. These so-called *binit weights* are simply

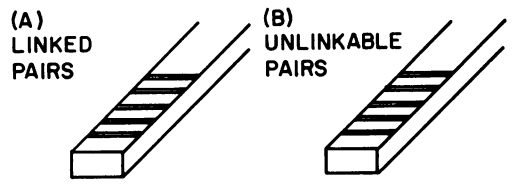
$$W_t = \log_2 (A/B) \tag{4}$$

where A and B are the frequencies of the particular agreement, defined as specifically as one wishes, among linked pairs of records and among pairs that are rejected as unlinkable. The binit weights for agreements will have positive values because A in such circumstances is always greater than B (Fig. 4), and these weights may be regarded as strictly analogous to the discriminating powers discussed earlier except that they relate to particular values of the various items of identifying information.

There is no need to alter this formula when deriving the weights for disagreements. A and B may be regarded simply as the frequencies of the particular disagreement, defined in any way, among linked and unlinked pairs of records. Usually the weights will then be negative in sign, because disagree-

TABLE 4. EXAMPLES OF KINDS OF LINKAGE

Event		Parental information (husband × wife)						Individual information	
Kind	Year	Surnames	Initials		Birthplace codes		Ages	Name	Birth date (or age)
Death to birth									
Birth	1950	Doe × Cox	JA	MB	09	09	30 25	Fred	15.6.50
Death	1955	Doe × Cox	JA	MB	09	09	— —	Fred	15.6.50
Birth to parental marriage									
Parental marriage	1945	Doe × Cox	JA	MB	09	09	25 20	—	—
Birth	1950	Doe × Cox	JA	MB	09	09	30 25	Fred	15.6.50
Marriage of a groom, to own birth and own parents' marriage									
Parental marriage	1945	Doe × Cox	JA	MB	09	09	25 20	—	—
Birth	1946	Doe × Cox	JA	MB	09	09	26 21	Andy	18.5.46
Own marriage	1966	Doe × Cox	JA	MB	09	09	— —	Andy	(age 20)



("BINIT WEIGHTS" = $\log_2 A/B$)

Examples

Kinds of agreements or disagreements	Frequency in linked pairs *A*	Frequency in unlinkable pairs *B*	Ratio *A/B*	Binit weight $\log_2 A/B$
Agreements				
Male sex	1/2	1/4	2	+1
Initial "J"	1/16	1/256	16	+4
Initial "Z"	1/1000	1/1,000,000	1000	+10
Disagreements				
City of residence	1/3	2/3	1/2	−1
Initial (any)	1/40	32/40	1/32	−5
Sex	1/8000	1/2	1/4000	−12

FIG. 4. Calculating "binit weights."

ments are, in most instances, less common among the linked than among the unlinked pairs; i.e., *A* will be less than *B*, and the logarithm of *A/B* will be negative.

Exceptions will occur in which an apparent disagreement is in reality a partial agreement. For example, a discrepancy of one year of age, after allowance is made for the interval of time between the two registered events, will frequently be a reflection of an underlying genuine agreement. Fortunately, however, it is not necessary to prejudge the issue. If the apparent discrepancy is predominantly a reflection of a partial agreement, the calculated weight will automatically turn out to be positive.

In practice, the formula is used to derive from the actual files a set of look-up tables of weights for agreements and disagreements of various items of information, broken down by the natures of these agreements and disagreements to whatever extent is necessary to make nearly full use of the discriminating powers. Such tables are stored in the memory of the computer. For each detailed comparison of a pair of records, the positive and negative weights appropriate for the different agreements and disagreements are added together, and the total weight is used to indicate the degree of assurance that the pair do, or do not, relate to the same person. The procedure assumes as a tolerable approximation that the weight for the individual agreements or disagreements are uncorrelated with each other; corrections are possible where this is not strictly true, but in our own experience these have been too small to be worth applying.

The derivation and use of the binit weighting factors have been described in greater detail elsewhere (Newcombe *et al.*, 1959; Newcombe and Kennedy, 1962). For present purposes, it is sufficient to indicate that there is great flexibility in the manner in which the weights can be employed and that they permit the introduction of numerous refinements so as to make nearly full use of the discriminating power inherent in the identifying information. For anyone planning an actual application, I would recommend that a number of small linking studies be carried out by hand to provide an opportunity to experiment with the system and become familiar with its characteristics.

The total binit weight represents the extent to which assurance of a genuine linkage is increased, or decreased, as a result of the comparisons made. Such weights are, in fact, logarithms to the base two of the factors by which the odds in favor of a linkage are increased over and above what they would have been in the absence of the comparisons.

In our own operation, the linkages are carried out within the very small "double surname pockets" of the master file, which contain on the average between one and two records apiece. Furthermore, an incoming record is quite likely to find a linkable counterpart there. Thus, even in the absence of the detailed comparisons, the probability of a match with a record drawn at random from the correct pocket of the master file will not be so very much less than 50% (i.e., odds of 1:1). In this situation, the total binit weight will closely approximate the $\log_2$ of the odds in favor of a linkage. Weights of +10 and of +20, for example, may in this situation be regarded as indicating favorable odds of approximately 1,000 to 1 and 1,000,000 to 1, respectively.

Using the double-surname sequenced files in this manner, no weights are attached to agreements of the items of sequencing information, i.e., to agreements of the surname codes. The reason is that the discriminating powers of these have already been taken into account automatically, since it is this information which determines the sizes of the pockets in the master file.

If binit weights were attached to agreements and disagreements of the sequencing information, incoming records would then have to be thought of as linking within a population of records consisting of the whole of the master file. Suppose, for example, that this contained 10^6 records and was known to include one which matched each of the incoming records. Under these conditions, the chance of an incoming record linking with a randomly chosen record from the master file would be $1/10^6$ ($= 2^{-20}$). However, if the detailed comparisons yielded a weight of +24, this would raise the odds from 2^{-20} up to 2^4, i.e., to 16:1 in favor of a genuine linkage.

Thus, to derive from the total binit weights the odds in favor of a linkage, allowance must be made for the size of the population of records within which the linkage is carried out by subtracting $\log_2$ of this population size. Similarly, allowance must also be made for the limited probability that there is, in fact, a matching record within that particular population. The $\log_2$ of this probability will be negative in sign and when added to the total binit weight will further reduce its value.

In practice, thresholds must be set which specify the ranges of binit weights

TABLE 5. TYPICAL MAGNETIC TAPE FORMAT FOR A VITAL RECORD

Information	Word*
Soundex pair	1
List word	2
Event (date, etc.)	3–6
Husband (name, etc.)	7–9
Wife	10–12
Offspring	13–14
Record linkage cross reference	15–17
Sibship cross reference	18–19
Statistics	20–24
Other cross reference	25

*One word equals ten octal digits or five alphanumeric characters.

which are to be regarded as representing linkage, no linkage, and possible linkage. Initially, these thresholds may be set to what seem intuitively to be reasonable values, but empirical tests are needed to ensure that false linkages, failures to link, and tentative linkages are balanced in a reasonable fashion.

In an actual operation, the total weights for linked pairs should be recorded permanently as evidence of the degree of assurance on which the linking was based. Similarly, for pairs of records that are judged to be neither positively linkable nor positively nonlinkable but which represent the most likely linkage available, it is prudent to retain permanently information about each such doubtful link and the weight associated with it. As more information accumulates about the family groupings, such as the sequences of birth orders in the families and the intervals between the births, this further knowledge may assist with the resolution of some of these doubtful linkings, provided that the information about them is retained on the files.

3. *Factors Affecting the Speed of the Record Linking Operation*

A number of practical considerations will influence the speed of a record linking operation.

The individual magnetic tape records should not be unnecessarily large, as this will increase the times required for input and output and for sorting the records. It will also limit the number of records that can be manipulated within the available core memory at any one time. The record format chosen for our own linking operation, using the vital registrations, consists of 25 words of 30 or 32 bits each (depending upon the magnetic tape units used). Each word may contain ten octal digits or five alphanumeric characters. This size of record was found to be sufficient for the storage of the individual and family identifying information, the statistics, and the cross-referencing information pertaining to a vital registration (Table 5).

Speeds are also affected by the amount of unused space on the magnetic tapes between records or between "blocks" of records. On the tapes used with the Control Data G20 computer, on which most of the recent work was done, records are stored in addressable blocks of 800 words each, i.e., con-

TABLE 6. EXAMPLE OF LIST PROCESSING

New record	Position	Record	Links	
			Forward	Back
G	(1)	G*	0	0
B	(1)	G	0	2
	(2)	B*	1	0
D	(1)	G	0	3
	(2)	B*	3	0
	(3)	D	1	2
F	(1)	G	0	4
	(2)	B*	3	0
	(3)	D	4	2
	(4)	F	1	3
A	(1)	G	0	4
	(2)	B	3	5
	(3)	D	4	2
	(4)	F	1	3
	(5)	A*	2	0

*Indicates "flag" for head of list.

taining 32 records per block. If records are read singly onto tape rather than in blocks, a substantial fraction of the tape is used up in the inter-record gaps.

A special time-saving feature in our own linking operation has been the use of a so-called "list processing" method. Records entering a husband-wife double surname pocket in the master file are arranged, physically, simply in order of their entry or acquisition, regardless of the appropriate logical sequence in the family groups. The logical position of each record is indicated by the inclusion on it of the "entry number" (i.e., acquisition number) of the record that logically preceeds it and that of the record that logically succeeds it. These numbers are known respectively as the backward and forward links.

When a new record enters the double surname pocket, known as a "superfamily," it is placed physically at the end; backward and forward links are then entered in the incoming record, and the existing links on the records that immediately precede and succeed it in the logical sequences are updated (Table 6). The saving of time occurs because with this procedure there is no need to alter the physical positions of the records already in a pocket to make room for a new record each time one is to be interfiled. The list processing method used has been described in detail by Kennedy *et al.* (1964).

Another factor that affects the speed of a linking operation has been mentioned earlier, namely, the size of the units into which the file is broken by the sequencing information. In our own experience, the use of two phonetically coded surnames relating to the husband-wife pair has divided a master file of 114,000 marriage records into units containing on the average about 1.6 records each. For approximately 80% of the file the pairs of surname codes are unique, i.e., they occur only once in that combination throughout the whole file.

Under the various conditions described above as pertaining to our own

operation, incoming birth records have been merged and linked with a master file of parental marriages and earlier births at a rate of 2,300 per minute. Thus for the British Columbia population of 1.6 million people, with which this study is concerned, a year's crop of 35,000 birth records can be merged and linked with the master family file of ten years of marriages in somewhat less than 30 minutes of machine time, once the magnetic tape records have been prepared in the proper format and appropriately sequenced. At a machine rental of two dollars per minute this is equivalent to a cost of 0.1 cents per record, i.e., it is minute in comparison with the cost of producing the punchcards in the first place, as is done routinely for administrative and statistical purposes.

The ways in which these various time-saving devices have been employed are described in greater detail by Kennedy *et al.* (1965).

STORAGE AND RETRIEVAL

In the sections that follow, we will consider the manner in which records relating to sibship groups may be stored together, certain extensions of the procedures to permit the inclusion of pedigree information covering an indefinite number of generations, and methods of retrieving information from the sibship grouping and multigeneration pedigrees. The records pertaining to the sibships, of course, fall within the main file sequence based on the surname pairs in their phonetically coded forms (Table 7).

1. *Storage of Sibship Groupings of Records*

There is a natural sequence in which the vital and health records pertaining to a sibship group may be linked and stored. Starting with the parental marriage registration, which may be regarded as a "head-of-family" record, birth records are linked to the marriage record in chronological order, and records of the various events of ill health, including death, are linked to the birth records of the children to whom they relate, those for a particular child falling likewise in chronological order after his or her birth record (Table 8).

The experience which we have had with this kind of file organization relates to records of marriages, livebirths, stillbirths, and deaths, together with those from a special register of handicapping conditions of children and adults. In addition, detailed plans have been worked out for the possible future inclusion of substantial numbers of records from a universal scheme of hospital insurance. Off-line linkings with the birth registration records are needed in the case of the handicap and hospital records in order to pick up the mother's maiden name which is lacking on the original form. Only after this has been done can the handicap and hospital records be merged and linked with the master family file, which is arranged in order of the two parental surname codes.

Incompleteness of a sibship grouping of records poses no special problem. In the absence of the parental marriage record, for example, the birth record of the oldest child represented in the file may serve as the head-of-family record, and records of the births of younger siblings will be linked to it. A

TABLE 7. EXAMPLE OF DOUBLE SOUNDEX FILE SEQUENCE*

Adams × Adair	A 352	A 360
Adams × Baron	A 352	B 650
Adams × Caird	A 352	C 630
Adams × Danys	A 352	D 520
↓		
Baker × Allen	B 260	A 450
Baker × Barks	B 260	B 620
Baker × Caron	B 260	C 650
Baker × Duffy	B 260	D 200
↓		
Baird × Aubry	B 630	A 160
Baird × Baker	B 630	B 260
(and so on)		

*i.e., by husband's surname code followed by the wife's maiden surname code.

TABLE 8. EXAMPLE OF A SIBSHIP GROUP OF RECORDS

	Parental couple	Child
Parental marriage	Doe × Cox	—
Birth 1	Doe × Cox	Alan
Birth 2	Doe × Cox	Carl
Ill health	Doe × Cox	Carl
Death	Doe × Cox	Carl
Birth 3	Doe × Cox	Edna

death record may serve likewise as a head-of-family record where it relates to the oldest child represented in the family group and the birth record for this child is missing. Thus, all of the available records of vital and health events may be merged and linked into sibship arrays, regardless of the degree of completeness or incompleteness of these groupings, and the master file may be updated periodically by the introduction into it of successive crops of current records.

The times required to merge and link the death and handicap records to the master file are somewhat greater than those for the corresponding operation as applied to birth records. There are two reasons for this. First, an ill health or death record must scan all of the birth records present in the appropriate double surname pocket of the master file, and these will tend to be more numerous than the head-of-family records which the incoming births must scan. Second, where an incoming ill health or death record fails to find a matching birth record, it must scan the double surname pocket a second time in an attempt to find a head-of-family record with which to link.

In our own operation, handicap and death records were merged and linked with the master file at a rate of approximately 1,100 per minute, i.e., at about one-half of the speed for the merging and linking of birth records.

2. *Storage of Multigeneration Pedigrees*

The modifications of the above procedures needed to permit the linking and

storage of the vital and health records in the form of multigeneration pedigrees are surprisingly simple. For most registration areas, the marriage records contain sufficient information to serve as bridges between the generations and between the in-law sibships.

Information from a marriage record may be treated in two ways. We have discussed already how it can be arranged into the form of a head-of-family record representing the marriage of a parental couple. Similarly, information from the registration form may also be fitted into the format of a record such as is used to describe an event in the life of an individual. The part of this latter kind of record entry that is assigned to family information would then contain the names and other identifying particulars of the parents of the newly married person, and the part of the record assigned to personal identification would contain his or her own name, age, and birthplace. This kind of entry of the marriage information is almost precisely analogous to a death record, since both relate to events in the lives of members of a sibship group. In the master file, the three entries pertaining to a particular event of marriage (i.e., the groom's entry, the bride's entry, and the head-of-family entry) will each become part of a different sibship group of records.

The only special requirement for the three marriage entry records is that each of them, before being placed in these various locations on the master tape, be cross-referenced to the other two. This is done by inserting in the cross-reference field of each record entry the double surname codes for the other two. These codes, together with the marriage registration number which is common to all three entries, provide both a means of access within the master file from one of the double surname pockets to the other two and a positive identification of the alternative entries when the pockets in which they occur have been located. The cross-referencing is illustrated in Tables 9 and 10.

The simplicity of the procedure resides in the use of essentially the same format for the marriage entries of grooms or brides as for their death records. In our own operation, the same programs that are used to build the sibship groupings of records will also be employed to insert into these groupings the grooms' and brides' marriage entries, just as they would the records of any other kinds of events in the lives of the same individuals.

The idea of thus putting family groups of records into a single linear array and of using cross references to indicate the relationships between the groupings that are filed as units is basic to any system by which computers may be employed to store and retrieve large quantities of pedigree information of unlimited complexity. The special features of the system described are merely matters of convenience. The choice of the sibship group as the unit of storage and of the surname pair as the sequencing information may have fairly wide application, but the details of the use of identifying particulars have been dictated largely by the nature of the vital records.

It would, of course, be feasible to store the same pedigree information more compactly if the family relationships were worked out in advance so that every individual could be assigned an identifying number containing as few

TABLE 9. EXAMPLE OF A MARRIAGE REGISTRATION AND OF THE MARRIAGE ENTRY RECORDS DERIVED FROM IT

Marriage registration

Groom	Dunn, Alex
Bride	Rowe, Anna
Groom's father	Dunn, Carl
Groom's mother	Bell, Edna
Bride's father	Rowe, Paul
Bride's mother	Hill, Jean

Marriage entry records

	Parental couple	Offspring
1. Head of family entry	Dunn × Rowe (Alex) (Anna)	—
2. Groom's entry	Dunn × Bell (Carl) (Edna)	Alex
3. Bride's entry	Rowe × Hill (Paul) (Jean)	Anna

TABLE 10. EXAMPLE OF CROSS-REFERENCING A SIBSHIP TO THE RELATED SIBSHIPS

Record	Parental couple	Offspring	Cross references
Parental marriage	Dunn × Bell		Dunn × Nash—father's sibship Bell × Mann—mother's sibship
Birth 1	Dunn × Bell	Alex	
Groom's entry	Dunn × Bell	Alex	Dunn × Rowe—new family Rowe × Hill—bride's sibship
Birth 2	Dunn × Bell	Stan	
Groom's entry	Dunn × Bell	Stan	Dunn × Knox—new family Knox × Fynn—bride's sibship

digits as possible, but the disadvantages of this approach where large populations are involved should perhaps be mentioned. A main objective of the present handling procedures has been to avoid entirely all manual manipulations so that full use can be made of the speeds of electronic computers. If this feature is to be preserved, the present kind of linking operation would have to be carried out anyway. A more important problem would be what to do with the borderline linkings when condensing the pedigree information into its more compact form, since both the extents of the uncertainties and the means for their later resolution would tend to be lost in the process. It might also be difficult to keep open the possibility, as the present system does, of merging at some future time the pedigrees drawn from a limited region, such as a province or a state, with those for a wider region such as the country as a whole.

3. *Retrieval of Pedigree Information*

The need for writing detailed programs does not end with the establishment

of a master family file containing the required pedigree information. For almost any kind of genetic study, the extraction of the required tabular information from a printed listing of the master file would be almost unthinkably laborious and expensive.

In general, it is necessary first to prepare programs that will summarize in a single record whatever information is required about a particular family. A further program is then written to extract information in tabular form from the resulting file of these summary records. Two examples of such procedures will be described, relating to sibship groups and to multigeneration pedigrees, respectively.

Where the family units under study are restricted to the sibships, summaries of the events of birth, ill health, and death in the lives of the various members of a sibship will usually be derived in two steps. First, individual histories will be condensed so that there is just a single summary record for each child replacing the separate records for the various events. The resulting magnetic tape file of individual or personal summaries can be used repeatedly to prepare the much more compact family summary records, which may be of a variety of kinds depending upon the natures of the studies for which they are to be used (Table 11).

To facilitate subsequent tabulations, the family summary records will have a different fixed field for each of the siblings. There must also be provision for large families, which will sometimes overrun a family summary record of modest size. This is best taken care of by arranging for trailing records to act as extensions where needed.

In one study which we have done using this procedure, the coded causes of stillbirths, handicaps, and deaths were entered into the fields of the family summary record assigned to the particular siblings who were affected, and for the unaffected siblings just the fact of birth, the birth order, and the sex of the child were entered.

In this particular study, use was made of the family summaries to derive information about the magnitudes of the risks to the later-born siblings of children who had been stillborn, handicapped, or had died, as the result of diseases of various kinds. The tabulations contained, typically, the number of index cases of a disease, the numbers of earlier and later siblings of the index cases, and the number of later-born siblings suffering from the same condition (Table 12). For details of the computer programs by which the different steps in the extraction were carried out, the reader is referred to Smith *et al.* (1965).

A more elaborate procedure is required where multigeneration pedigrees are to be summarized, because as an initial step the sibship groupings of records relating to a particular family must be brought together from different parts of the master file. Before starting this step, certain sibships whose relatives one wishes to ascertain will have been extracted from the master file. These may be called "index sibships," and they will in most instances have been chosen because they include individuals who are affected by some disease of special interest.

TABLE 11. EXAMPLES OF INDIVIDUAL AND FAMILY SUMMARY RECORDS

Event code	Birth order	Family	Child	Disease code
Event records for a sibship (one per event)				
J (birth)	1	Fox × Dow	Alan	—
J (birth)	2	Fox × Dow	John	—
J (birth)	3	Fox × Dow	Vera	—
Q (handicap)		Fox × Dow	Vera	123
J (birth)	4	Fox × Dow	Leon	—
R (death)		Fox × Dow	Leon	456
Individual summary records (one per child)				
(J)	1	Fox × Dow	Alan	—
(J)	2	Fox × Dow	John	—
(Q)	3	Fox × Dow	Vera	123
(R)	4	Fox × Dow	Leon	456
Family summary record (one per sibship)				
(Fox × Dow)	1 (J)---, 2 (J)---, 3 (Q) 123, 4 (R) 456.			

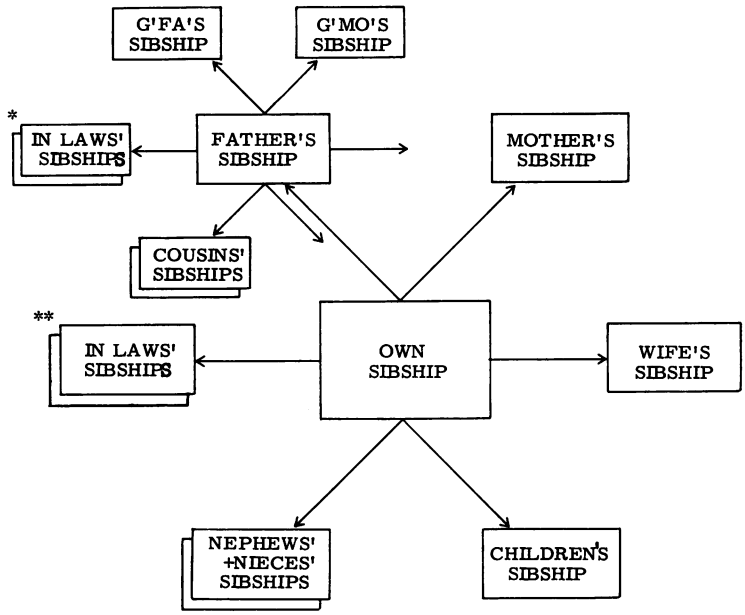
TABLE 12. EXAMPLE OF A TABULATION FROM FAMILY SUMMARY RECORDS

	Disease code 325 (*mental deficiency*)				
	Normal (J)	Stillborn (K)	Handicapped (Q)	Dead (R)	Handicapped and dead (S)
Index cases	0	0	506	9	58
Earlier sibs	208	2	6	16	0
Later sibs, same cause	0	0	11	0	1
Other later sibs	286	2	11	14	0

The records of the index sibships may contain cross-referencing information (in the form of double-surname codings and marriage registration numbers) indicating links with as many as six different kinds of related sibships, i.e.,

1. From the parental marriage (head-of-family) records to
   (a) the fathers' sibships and
   (b) the mothers' sibships.
2. From the marriage records of the "affected" individuals who got married (i.e., from the grooms' and brides' entries) to
   (c) their offspring's sibships and
   (d) their spouses' sibships.
3. From the marriage records of the brothers and sisters who got married to
   (e) the sibships of the nephews and nieces of the affected individuals and
   (f) the sibships of the spouses of the brothers and sisters who got married.

These six different kinds of cross references may be used in a single scan to draw from the master family file all of the groups of records pertaining to sibships that are removed by *one* degree of relationships from those in which the affected individuals occurred, including the in-law groups (Fig. 5).



*i.e., those of the paternal uncles and aunts by marriage.
**i.e., those of brothers' wives and sisters' husbands.

FIG. 5. Scanning the master file for related sibships.

Similarly, in a second scan of the master tape, use may be made of the further cross-referencing information contained in the sibship groups of these six different kinds to extract the sibships that are removed by *two* degrees of relationship from those in which the affected individuals occurred. Again, the in-law sibships may be extracted in the same way as those of the blood relatives. And so, with each successive scan, an expanding circle of more distant relatives may be identified and retrieved from the master file.

Each such scan will be exceedingly rapid even where large numbers of sibships groups are extracted. Thus, it is feasible to carry out the retrieval of multigeneration pedigrees on a truly massive scale.

From this point on, the making of summaries would follow much the same pattern as described earlier, except that the family summary record might be more complex than the sibship summary record.

The chief limiting factor in work of this kind is not the speed of the computer but the time required to develop the appropriate programs.

THE LIKELIHOOD OF FUTURE "TOTAL UTILIZATION" OF PEDIGREE INFORMATION

Geneticists will at first tend to think of the possible uses of record linking as applied simply to the familiar kinds of *ad hoc* studies of limited size and duration. The question arises whether it is realistic to go beyond this and to consider using for scientific purposes all of the pedigree information gathered

routinely for whole populations through the vital registration systems, of doing so on a continuing basis, and of adding an increasing amount of medical documentation as time goes on.

Clearly, the cost would appear large if it were paid wholly from budgets for scientific research. But this would not necessarily be the case, because the information that is unlocked by linking and integrating the files into individual and family histories has many statistical and administrative uses, as well as other scientific uses beyond those of the geneticist.

Those geneticists who attempt to apply the methods of record linking will be in a particularly good position to see a variety of possible uses for the linked files and to develop procedures that will serve more than one purpose. Their own long-term interest may be furthered most where they exploit the fact that there are other potential users.

Of course, with time the various files of routine records will, to an increasing extent, be linked and integrated anyway for administrative purposes, whether or not scientists take an interest in the matter. But the only way to ensure that scientific by-products will come out of this trend is for the scientists themselves to participate actively while the administrative procedures are being established.

APPENDIX
Surname Coding

Surnames may be converted into coded forms for either of two reasons: to set aside temporarily some unreliable component of the information that may vary on successive records relating to the same person, or for the sake of compactness. A number of systems have been designed to achieve one or other of these purposes, or both simultaneously. Some of the more useful of these codes will be described.

THE RUSSELL SOUNDEX CODE

This code is particularly efficient at setting aside unreliable components of the alphabetic surname information without losing more than a very small part of the total discriminating power. It is the method of choice for almost all populations, except where the names are predominantly of Oriental origin.

Rules:

1. The first letter of the surname is used in its uncoded form and serves as the prefix letter.
2. W and H are ignored entirely.
3. A, E, I, O, U, Y are not coded but serve as separators (see item 5 below).
4. Other letters are coded as follows until three digits are used up (the remaining letters are ignored):

B, P, F, 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

5. Exceptions are letters which follow prefix letters which would, if coded, have the same code. These are ignored in all cases unless a separator (see item 3 above) precedes them.

Examples:

Anderson	= A 536
Bergmans, Brigham	= B 625
Birk, Berque, Birck	= B 620
Fisher, Fischer	= F 260
Lavoie	= L 100
Llwellyn	= L 450

NAME COMPRESSION

As indicated by its name, this form of coding is designed mainly to condense surnames, given names, and place names. However, the code does remain unchanged with some of the common spelling variations, although it is less efficient in this respect than the Soundex code.

Rules:

1. Delete the second of any pair of identical consonants.
2. Delete A, E, I, O, U, Y, except when the first letter of the name.

Examples:

BENNETT	= BNT
FISHER	= FSHR

ILL-SPELLED NAME ROUTINE

Where the insertion, deletion, or substitution of a single letter of a surname alters the coded form, recognition that a pair of names are the same necessarily depends upon residual similarities in the sequences of the letters in the two, despite any interruptions in these sequences. The "ill-spelled name routine" is not, strictly speaking, a system of coding but rather a system of comparison which employs the coded forms of the names as derived by "name compression." The system was designed for use with airline bookings (Davidson, 1962).

Rules:

1. Use "name compression" procedure, up to a total of four letters.
2. Search for and count the numbers of letters or blanks, up to a total of four in all, that agree without altering the sequence.
3. Where the agreements equal 3 or 4 in a pair of names, compare other identifying information.

Examples:

		Score
BOWMANN	= B M N –	
BAUMAN	= B M N –	4
McGONE	= M C G N	
McKONE	= M C K N	3
ANGREIFF	= A N G R	
SINGER	= S N G R	3
MCGINESS	= M C G N	
MAGINNES	= M G N S	3
LU	= L – – –	
ROO	= R – – –	3

ALPHANUMERIC CONVERSION

This is a highly specific numeric coding for all surnames. It is not designed to set aside the less stable parts of the information but rather to retain virtually all of the original specificity of the alphabetic form. The numeric form of the surname is compact, is more readily sorted on an electromechanical card sorter than the alphabetic form, and is nonrevealing to anyone who lacks the relevant look-up table. Furthermore, when sorted in numerical sequence the names fall in alphabetic order or a close approximation to it.

The coding is done by computer using a look-up table containing over 8,000 different entries. (See International Business Machines, 1960.)

Examples:

ABBIT	= 0008
ADLER	= 0105
BORNE	= 1058
BRYAN	= 1070
CLARK	= 1646
COX	= 1721
	↓
ZZINA	= 9776

HOGBEN SURNAME CODE

This is a simple two-digit code for surnames based on a division of the names in a large telephone directory into 100 approximately equal parts. Although compact, it loses much of the discriminating power inherent in the full name and is therefore chiefly of historical interest. (Originally this was just a part of a much longer numeric code derived from the surname, first given name, sex, and birth date. See Hogben *et al.*, 1948.)

Examples:

00 = A A – A K
01 = A L
02 = A M – A R
03 = A S – A Z
04 = B A A – B A J
05 = B A K – B A Q
06 = B A R
(and so on)

REFERENCES

DAVIDSON, L. 1962. Retrieval of misspelled names in an airlines passenger record system. *Commun. Assoc. Computing Machinery* 5: 169–171.

HOGBEN, L., JOHNSTONE, M. M., AND CROSS, K. W. 1948. Identification of medical documents. *Brit. Med. J.* 1: 625–635.

International Business Machines. 1960. General Information Manual. *A Unique Computable Name Code for Alphabetic Account Numbering.*

KENNEDY, J. M., NEWCOMBE, H. B., OKAZAKI, E. A., AND SMITH, M. E. 1964. *List Processing Methods for Organizing Files of Linked Records.* Chalk River, Ontario: Atomic Energy of Canada, Ltd. Document AECL-2078.

KENNEDY, J. M., NEWCOMBE, H. B., OKAZAKI, E. A., AND SMITH, M. E. 1965. *Computer Methods for Family Linkage of Vital and Health Records.* Chalk River, Ontario: Atomic Energy of Canada, Ltd. Document AECL-2222.

NEWCOMBE, H. B., AND KENNEDY, J. M. 1962. Record linkage: Making maximum use of the discriminating power of identifying information. *Commun. Assoc. Computing Machinery* 5: 563–566.

NEWCOMBE, H. B., KENNEDY, J. M., AXFORD, S. J., AND JAMES, A. P. 1959. Automatic linkage of vital and health records. *Science* 130: 954–959.

SMITH, M. E., SCHWARTZ, R. R., AND NEWCOMBE, H. B. 1965. *Computer Methods for Extracting Sibship Data from Family Groupings of Records.* Chalk River, Ontario: Atomic Energy of Canada, Ltd. Document AECL-2530.