A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-up Studies*

G. R. HOWE

NCIC Epidemiology Unit, Faculty of Medicine, McMurrich Building, University of Toronto, Toronto, Ontario M5S 1A5, Canada

AND

J. LINDSAY

Vital Statistics and Disease Registries Section, Health Division, Statistics Canada, Ottawa, Ontario K1A 0T6, Canada

The development of a generalized iterative record linkage system for use in follow-up of cohorts in epidemiologic studies is described. The availability of this system makes such large-scale studies feasible and economical. The methodology for linking records is described as well as the different modules of the computer system developed to apply the methodology. Two applications of record linkage using the generalized system are discussed together with some considerations regarding strategies for conducting linkages efficiently.

The primary focus of epidemiologic studies of chronic disease is the determination of factors which may be associated with increased risk of such diseases. Two classic approaches to identifying such factors are the case-control and cohort studies (*1*).

In a cohort or follow-up study one starts with a group of individuals some or all of whom may have been exposed to the factor under study, and ascertains their subsequent morbidity or mortality experience. In order to accumulate sufficient person-years of experience to provide a sufficiently powerful statistical test of any association between exposure and disease, it may be necessary to follow large groups of individuals for many years, and this is particularly true if the excess risk in question is a small one. However, even in the latter case it is possible that if exposure to the factor is widespread, the population attributable risk can be substantial and consequently the factor can be a significant health hazard. Conventional methods for following cohorts include personal contact, telephone, and mail inquiries (*1*) and when the cohort is large such methods can be prohibitively difficult, expensive, and time consuming.

---

*Reprinted with permission from Computers and Biomedical Research 14, Copyright © 1981 by Academic Press, Inc., pp. 327-340.

An alternative method for following cohorts is the use of computerized record linkage in which records of individual members of a cohort are compared with records from files of morbidity and mortality data (*2–4*). When a unique identification number (such, for example, as the Canadian Social Insurance Number or the U.S. Social Security Number) is present on both the exposure records and the morbidity or mortality records, such linkages simply involve sorting both files using the unique identifier as key and then directly matching records from the two files. However, such unique identifiers rarely exist, especially on data which have been assembled retrospectively. In this case, it is necessary to use identifying characteristics such as surname, given name, date of birth, etc. in order to link records from the two files, and this involves two practical problems. In the first place, such identifying items are not unique to a particular individual and even combinations of identifying items may not be unique; and in addition, identifying items may be misrecorded or missing on certain records. It is therefore necessary to devise algorithms for comparing the two records in order to produce some quantitative measure which is a function of the probability that those two records do indeed refer to the same individual. Secondly, given such algorithms, it is necessary to devise a computer system in order to efficiently carry out the data processing involved.

Considerable attention has been paid to the first of these two problems and the methods most widely used are those which have been developed by Newcombe and his associates (*5*) and Fellegi and Sunter (*6*). However, the implementation of these methods in terms of computer programs has generally been done on an ad hoc basis for each specific application. This paper describes some extensions of the Newcombe methodology, in particular to cope with the problem of partial agreement of identifying items, and also a generalized computer system which has been developed in order to carry out linkages between any two files of interest. The system may also be used to internally link records from a single file, where one individual may have more than one record, but again no unique identifier exists. The application of the system to two studies in cancer epidemiology is also described.

Methodology

A. Basic Principles

Conceptually carrying out a record linkage between two files A and B involves the following steps:

Step 1. Every record on file A is compared with every record on file B. The result of each comparison is a series of outcomes, one outcome resulting from each identifying item being used for linkage such as surname, first given name, year of birth, etc. An outcome may be defined as specifically as desired; for example, the two records agree on the first five characters of the surname and the value is SMITH, or the first given name agrees on first character irrespective of value, but remaining characters disagree.

Step 2. A statistic called the total weight (W^*) is calculated for the comparison of any two particular records. The weight is an estimate of the odds that the two records under consideration do in fact refer to the same individual, i.e., that they are linked (L) as opposed to referring to different individuals, i.e., they are not linked ($\bar{L}$).

Thus the weight is an estimate of:

$$\frac{P(L/{}_1O_2O_3O. \ . \ .)}{P(\bar{L}/{}_1O_2O_3O. \ . \ .)}, \qquad [1]$$

where $P(L/{}_1O_2O_3O. \ . \ .)$ is the probability that the two records are linked conditional that the outcome from comparing the first identifying item is ${}_1O$, etc. If one assumes that the values of the identifying items on the records are statistically independent then it follows that:

$$W^* = {}_1w + {}_2w + {}_3w \ . \ . \ . + \log_2 \frac{P(L)}{P(\bar{L})}, \qquad [2]$$

where ${}_1w$ is $\log_2$ of the estimate of the odds of obtaining outcome ${}_1O$ conditional upon the two records being linked. It is convenient as is customary in information theory to use $\log_2$ in Eq. [2] in order to make the equation additive.

In practice the final term in Eq. [2] is usually impossible to evaluate since it requires a priori knowledge of the number of links among the set of all comparisons and this is usually unknown. Thus a modified total weight may be defined as:

$$W = {}_1w + {}_2w + {}_3w \ . \ . \ . \ . \qquad [3]$$

If W can be estimated from Eq. [3] for all possible comparisons between the records on the two files and these comparisons are then ordered by the value of W, they represent potential links in decreasing order of believability, and, in particular, the difference $W1 - W2$ for any two particular comparisons is an estimate of $\log_2$ of the odds ratio. Thus, if two comparisons result in W's which differ by 1.0 the odds in favor of the first comparison being a true link are twice the odds for the second comparison being a true link. Details of weight calculations including the case of partial agreements are given below.

Step 3. Having ordered the comparisons by W, upper and lower threshold values are chosen. These are used to divide the set of all comparisons into three; namely, the "definite links"—those with a weight above the upper threshold, the "nonlinks"—those below the lower threshold, and the "possible links"—those between the thresholds. The possible links may be manually inspected and if possible resolved. If further identifying information is available which is not in machine-readable form, this may be used to supplement the data for the possible links in order to resolve them. If no such data are available, manual resolution is probably undesirable and one possible approach is to choose a single threshold value (*2*). Fellegi and Sunter (*6*) have developed a likelihood ratio test based upon the total weight statistic which leads to optimum values of the upper and lower thresholds. Alternatively, and

frequently more conveniently, their values may be empirically assigned from inspection of the set of potential links.

B. Blocking

In order to compute W it is therefore only necessary to estimate ${}_1w, {}_2w, {}_3w$, etc. for each identifying item, for each possible outcome from comparing the possible values of that item. There is, however, a further practical consideration. When dealing with files of any appreciable size the total number of possible comparisons between records becomes extremely large and resulting computer costs are inordinate. It is therefore necessary to block the files using a combination of identifying items or derivatives of identifying items to define the blocks. Comparisons are then only carried out between records in corresponding blocks on the two files. The block identifier used in the applications described in the last section of this paper, for example, was the combination of sex and the NYSIIS code of surname (*7*). The NYSIIS code is an alphabetic code designed so that surnames of similar sound have the same code and frequently encountered errors of misreporting do not result in change in the NYSIIS code. Thus this blocking system will generally bring together records belonging to a single individual even when errors of recording have occurred. The effect of blocking on the calculation of weights is taken into account in the general formulation given below.

C. Derivation of Formulas for Weights

The w's of Eq. [3] may now be computed from simple probability theory. The general formulation proposed leads to slight modifications of the original formulas of Newcombe and Fellegi and Sunter as discussed subsequently.

It is convenient for illustrative purposes to consider a specific identifying item; the most useful one in the present context is surname since this involves a consideration of the blocking factor, namely, the NYSIIS code. Although the number and types of outcome in comparing the surnames from two records is arbitrary, we have found it most convenient to consider five possible types of outcome defined as follows. The subscript used to identify the particular identifying item is omitted from these formulas. (For outcomes 1 to 4 surname is assumed to be present on both records.)

(1) $O_{1=i}$: Surname agrees on first seven characters with value i.

(2) $O_{2=j}$: Surname agrees on first four characters with value j, but disagrees within next three characters.

(3) $O_{3=k}$: Surname agrees on NYSIIS code with value k, but disagrees within the first four characters.

(4) O_4: Surname disagrees on NYSIIS code.

(5) O_5: Surname is missing on one or both records.

The weight corresponding to O_5 is obviously zero unless the linked and unlinked set of records have different frequencies for the reporting or nonreporting of identifying items. If an estimate can be made of any differential reporting for the two sets, w_5 may be computed correctly from its definition. No further consideration need be given to missing data, as all probabilities and frequencies are assumed to be conditional upon a value for the identifying item in question being present.

In order to compute w_1 to w_4 it is necessary to specify the frequency with which surname is misreported. These frequencies, referred to as transmission rates, are defined as follows:

t_1: The probability that the surname on a particular record has the same first seven characters as the "true" value.

t_2: The probability that the surname has at least the first same four characters as its "true" value.

t_3: The probability that the surname has the same NYSIIS code as its "true" value.

By this definition there is a single set of transmission coefficients, t_1 to t_3, for each identifying item. It should be noted that the transmission coefficients correspond to the various possible outcomes listed above in the sense that if both records in a particular comparison are transmitted from the "true" value to the recorded value so that the first seven characters remain the same the outcome will be O_1 and the probability of such a transmission is t_1 for each record. It should also be noted that various components can contribute to the transmission coefficients, such as a genuine change in the "true" value of surname between the creation of the two records, errors of recording, etc. If such components can be identified and numerical values estimated, these values can be used to compute the transmission coefficients. The approach we have used is to compute the transmission coefficients in an iterative fashion from the records themselves as described subsequently.

In order to calculate the weights corresponding to each possible outcome the basic definition is used. For example, the probability of exact agreement on the first seven specific characters of a certain surname when the two records originate from the same individual is given by

$$t_1^2 f_i,$$

where f_i is the relative frequency of occurrence of the particular seven-character value among the individuals who give rise to the linked set. In order to estimate such frequencies it is usually necessary to use the frequencies as observed on the records in the files themselves. This involves a decision as to whether the frequencies on the linked set are most similar to the frequencies on file A or file B, and this obviously depends on the particular data sets under consideration and involves essentially an empirical decision. Given the particular file to be used for estimating the frequencies there are two possible models. In the first, it is assumed that errors in recording are such that the original "true" value is transmitted to some value that does not already exist

within the linked set. This leads to the observed frequency value within the file being set equal to $t_1^2 f_i$, which is the formulation proposed by Fellegi and Sunter. Alternatively it may be assumed that when a recording error is made it results in some value which already exists within the linked set. If this process happens randomly the observed frequency within the file will be equal to f_i. We have used the second model since we feel it to be more realistic and since it leads to a formulation in which transmission and frequency components of the weights are separable and the weight for any particular outcome can be factorized into these two components.

The probability for any outcome with the unlinked set of comparisons is most simply determined from consideration of frequencies as they occur on the files. Thus the probability of agreement by chance on the first seven characters of surname in the unlinked set is given by:

$$ {}_A f_i \; {}_B f_i, $$

where ${}_A f_i$ and ${}_B f_i$ refer to the relative frequencies on files A and B, respectively. (The contribution to all possible comparisons from the linked set is negligibly small and is therefore ignored in this formulation.) Using this approach the weights for 1–4 above can be shown to be:

$$ w_{1=i} = \log_2 t_1^2 + \log_2 \frac{1}{{}_B f_i}, \qquad [4] $$

$$ w_{2=j} = \log_2 (t_2^2 - t_1^2) + \log_2 \left[\frac{{}_A g_j}{{}_A g_j \; {}_B g_j - \sum_{i \epsilon j} {}_A f_i \; {}_B f_i} \right], \qquad [5] $$

$$ w_{3=k} = \log_2 (t_3^2 - t_2^2) + \log_2 \left[\frac{{}_A h_k}{{}_A h_k \; {}_B h_k - \sum_{j \epsilon k} {}_A g_j \; {}_B g_j} \right], \qquad [6] $$

$$ w_4 = \log_2 (O), \qquad [7] $$

where ${}_B f_i$ is as before; ${}_A g_j$ is the relative frequency of first four characters of surname equal to j, and ${}_A h_k$ is the relative frequency of NYSIIS code equal to k (for file A). Equation [7] is applicable only to the item used as a pocket identifier.

These formulas apply when the frequency distributions in the linked set are taken as being the same as those on file A.

In all the above expressions it will be seen that the transmission and frequency components of the weight are separable and their $\log_2$s are additive. It should be noted that the value for w_4 means that no two records from different blocks can link. In order to estimate the various values of t, we have used an iterative procedure as follows. The linkage is carried out using estimates for t, usually based on previous experience. Given an estimate of the upper threshold value, a sample of links may be drawn from the linked set and estimates made of the transmission coefficients from the number of times that

full or partial agreements on surname occur within the linked set. These new values may then be used as the basis for another linkage and the process repeated iteratively until reasonably stable values for the transmission coefficients are obtained. Alternatively, as previously mentioned, the transmission coefficients may be estimated empirically.

System Design

The particular series of programs, which were written in order to apply the above methodological principles to specific data sets, relies heavily upon use of a data base system (Relational Access Processor for Integrated Data Bases (RAPID)) which is available within the facility where the programs were developed (Statistics Canada). The programs as such, therefore, are of no direct use in any other environment, but the principles of the system involved are readily generalizable to any other computer environment, and may be programmed within the particular limitations of the hardware/software available.

The system has been deliberately designed to be modular in nature. In particular, the most time-consuming element, namely, the comparison of all records within each block, was developed as a single module. Only one pass of the complete data is necessary, which will eliminate any comparisons which result in any obvious nonlinks and will produce a file of potential links with their corresponding outcomes. These potential links may then be subjected to a number of different weighting runs in order to refine the linkage results at a much lower cost than would be incurred by rerunning comparisons between the entire data files. This modular approach also facilitates the iterative process of calculating transmission weights. The modules involved in the system are shown in block diagram form in Fig. 1 and their specific functions are now described.

A. Preprocessing

This step involves editing and correcting of the original data files, including such functions as creating a unique sequence number for each record and the NYSIIS code of surname, left justifying fields such as given name, removing blanks within names, recoding variables, etc. Following the editing step the files are sorted by whichever identifying item is to be used as the pocket identifier, e.g., NYSIIS code.

B. Calculation of Frequency Component of Weights

Frequency counts are carried out on the preprocessed files for all levels of agreement and partial agreement for all identifying items. From these frequency distributions are computed the frequency components of the weights as given in Eqs. [4] to [7]. In practice it will often be found that for many items the frequency distribution is similar from one file to another and consequently a

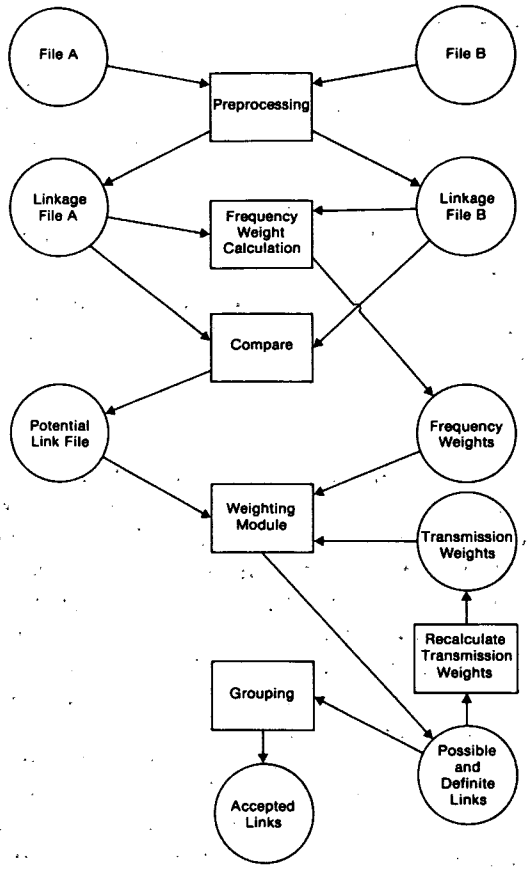

FIG. 1. Generalized iterative record linkage system.

single set of frequency weights will suffice. For other items, such as birth year, the distribution will vary considerably from file to file and may need recomputing each time.

C. Comparison Module

The function of the compare module as stated is to create a file of potential links and their corresponding outcomes and to eliminate all obvious nonlinks. In this module all records within a given pocket are compared with each other, each comparison giving rise to a series of outcomes such as, e.g., "seven character agreement on surname, and the value is Smith." Identifying items on the two records are compared in an order which is specified at execution time. This ordering is decided by two factors, the discriminating power of the identifying item and the CPU time necessary to make the comparison. An option is provided to carry a crude "running total of disagreement weights."

Each item is assigned an appropriate preliminary disagreement weight, and where a disagreement occurs, the running total is decremented by the disagreement weight for the item concerned. When the running total achieves a value below a preselected cutoff value, the comparison between the two records in question is abandoned and the module then proceeds to the next comparison. This procedure ensures that records which are in obvious disagreement are not considered as potential links. For any comparison which does not yield a value for the running weight below the critical, a "link record" is created consisting of the two record numbers and an outcome code and, where appropriate, a value for each identifying item in question. At the completion of this phase the link record file thus contains all potential links and further processing is concerned with this particular file.

D. Weighting Module

The function of this module is to add both frequency and transmission components of the weights to the link record file. Components may be added in separate passes as they are completely independent of each other as in the formulation of the previous section. The particular method used to add the weights will of course depend on the hardware configuration available. In general, the procedure will involve table lookups using the outcome code and value where appropriate as an index. Since the link records are ordered in the same sequence as the pocket identifier, the weights for the pocket identifier (e.g., NYSIIS of surname) may be added conveniently from a sequential file. For items with relatively limited numbers of values such as birth year the tables may be conveniently stored in core; for alphabetic data other than the pocket identifier, such as given name, random access disk files probably provide the most convenient means. As there are relatively few transmission coefficients these generally can be stored in core, and a weighting pass to change just the transmission coefficients can be carried out rapidly. Subsequent to applying the weights to the link record file, a sample of this can be printed out for manual inspection and this can be used to assign tested threshold values. Given these threshold values new estimates of transmission weights can be made using the set of links which are above the upper threshold. These new values can be applied to the links and the process repeated until some measure of consistency is achieved.

E. Grouping Module

The function of this module is to bring together all records which have linked with each other. The specific algorithm to be used is of course dependent upon the nature of the records concerned, and whether the linkage is two file or internal. For an internal linkage generally there is no limitation upon the number of records that can constitute a "group" corresponding to a single individual. Often in the case of two-file linkage only a one-to-one relationship is possible as for example in linking records for specific individuals to a file of

death records. However, in the latter case, since some links will occur by chance, it is necessary to identify records which appear in more than one link.

For grouping records from an internal linkage we utilized the following method which involves starting with a single record, identifying all links to that record, then identifying all links to those links, and so on. We defined definite groups of records as those in which each member is linked to at least one other member of the group with a weight which is above the upper threshold (a definite link). Possible groups are then defined as being composed of a series of definite groups in which there is at least one possible link between members of the definite groups concerned. Any possible groups which are formed can then be printed out for visual inspection and a decision made as to whether the definite groups which constitute them should be amalgamated into a single group or whether the original definite groups should be maintained as individuals. The reservations concerning the utility of manual resolution when no further identifying data are available, expressed in the methodology section, should be taken into account when deciding whether to adopt such a grouping procedure.

In order to group links from a two-file linkage where only a one-to-one link is permissible, the links are sorted by weight, then proceeding from the link with the largest value downward, each link is checked to see whether either record concerned has appeared in a previous link. If either has, the link may be printed out as a conflict and the situation resolved by visual inspection. Alternatively, the link with the highest weight may be accepted.

Since processing up to this point has involved record numbers rather than the actual records themselves at this stage a number is assigned to each group or pair of records that has been linked. These group numbers may then be assigned sequentially using the record number of one of the original records, and sorting the records on this group number brings together those records which have been linked so they may thus then be processed further as desired. It should be noted that although the identifying items on any particular record which has entered into a possible link are essentially contained on the link record file, and are there available for inspection if needed, it is also desirable to provide a mechanism for accessing the original complete data records. In the system we have developed this is done by maintaining a parallel file containing those data records which have formed at least one link so that they may be accessed via the data base used.

Applications

The system described has been primarily developed for use in monitoring the morbidity and mortality experience of various groups of individuals with various exposures, by linking such exposure records to national morbidity and mortality files. Two such specific applications are now described in more detail.

Linkage of TB Patient File to Mortality File

Between 1930 and 1952 extensive use was made of collapse therapy in the treatment of tuberculosis. This involved considerable X-ray exposure from fluoroscopy machines which were extensively used for examination of the chest cavity. A major study of cancer mortality in relation to this radiation exposure is being conducted (*3*), by collecting data on individual patients from all existing hospital and sanitorium records in Canada.

The TB patient file was first internally linked using the generalized iterative linkage system described here to bring together treatment data from different institutions to form one complete treatment history per patient. The TB patient file containing 118,000 records was then linked to the national mortality file covering the years 1950 to 1977 containing 5,000,000 records. (1950 is the first year for which sufficiently well-identified mortality records are available in a format suitable for computerized record linkage.)

The identifying items used were the following: NYSIIS code and surname; first and second given names; day, month, and year of birth; place of birth; sex; NYSIIS of mother's maiden name; mother's first initial; mother's birthplace; father's first initial; and father's birthplace. Year of last contact on the TB records was compared with year of death on the mortality records in order to eliminate unnecessary comparisons. Use was made of the facility to incorporate partial agreements as follows: Surnames were considered to be in full agreement if they agreed on seven characters; the first level of partial agreement was on the first four characters and the second level of partial agreement, on NYSIIS only. Full agreement for given names was on the first four characters, and partial agreement, on initial only. Birth year was treated as being in full agreement if it was within plus or minus 1 year. The first level of partial agreement was within 5 years, and the second level, within 10.

The records were blocked by NYSIIS code of surname and sex. Alternate surname spellings and maiden names were also available. These were included as comparison items by creating duplicate records for alternate surnames at the preprocessing stage. Following the linkage, duplicate records were combined. The total file of TB patients was linked to 1 year of mortality records at a time. This provided the advantage of allowing the runs to be checked closely rather than risking costly errors over the entire linkage.

Initially, the number of potential links formed between the TB and mortality files was 787,800 for males and 554,800 for females, using a very conservative cutoff weight to ensure that no potential links were missed. The preliminary weights used were average values or approximations of the final weights. After the final weights were calculated and threshold values set, there were 82,828 possible and definite links generated by the male files and 67,490 by the female files. This was considered to be an application where only a one-to-one link was acceptable, i.e., one TB record could validly link with one death record. Following the application of the one-to-one rule, there remained 20,293 male links and 12,697 female links which were considered to be definite for the purpose of the subsequent statistical analysis.

The cost of this record linkage was just over $5000 (Canadian). This cost includes the comparison of the records, assignment of preliminary weights used to determine whether each link was a potential link, insertion of the final weights, setting of the thresholds and resulting classification of each link as definite, possible or rejected, the listing of a sample of links from each run, and resolution of duplicate links within each run. In addition, duplicate links involving records over different years of death were resolved. Over two-thirds of the cost was accounted for by the comparison of the records. As previously mentioned, this demonstrates the advantage of a modular system, where all other steps may be carried out iteratively at relatively minimal cost. The next most expensive step was the weighting which accounted for approximately 14%. The steps listed above took 179 min of CPU time for the males and 175 min for the females. It should be noted that testing was carried out first on a very small sample of the file consisting of a few blocks of records from the two files. At this point, the mortality records were selected from a single year of death. When preliminary testing was completed, an entire year of death records was linked with the TB records and further refinements made. For example, it was found that test runs where no cutoff weight was used were about 15% more expensive than those where a cutoff weight was used that was sufficiently low for no potential links to be missed. The cost of this linkage using the generalized system was substantially lower than the cost of linkages carried out previously using ad hoc programs.

Linkage of Occupational Cohort to Cancer Incidence

Between 1965 and 1971, data were collected by Statistics Canada for a 10% sample of the Canadian labor force (approximately 700,000 individuals). The data included identifying information together with the industry and occupation in which the individual was engaged in each particular year. In order to follow the mortality and cancer morbidity experience of this cohort with respect to their industrial and occupational exposure, these records were linked to the national mortality data base and the cancer incidence files. For the linkage to the cancer incidence files, Ontario occupational records were excluded, since identifiable cancer incidence records were not available for that province, leaving 476,174 occupational records.

The 287,786 male and 188,388 female occupational records were linked to 171,628 male and 215,651 female cancer incidence records covering the years 1969 to 1976. (Cancer incidence data were first collected nationally in 1969.) The identifying items available on both files were NYSIIS of surname; surname and alternate surname; first and second given names; day, month, and year of birth; and sex. As in the previous example, the records were blocked by NYSIIS of surname and sex. In this case only two separate runs were made since the files were split by sex, but not according to the year of diagnosis of cancer. The same levels of full and partial agreement were used as for the TB–mortality linkage.

The number of potential links generated was 96,100 from the male files and 82,482 from the female files. After the insertion of final weights and the setting of threshold values, and resolution of links of multiple occupation records to single cancer records, the number of possible and definite male links was 5315 and there were 2885 female links. In this case, multiple cancer incidence links to occupation records were considered acceptable since the cancer incidence file contains one record for each primary site of cancer. The number of occupation records involved in these links or the number of individuals linking to cancer records was 4953 men and 2747 women. The cost of this linkage was approximately $600 and the CPU time used was about 30 min for the males and 23 min for the females, including the same steps for which cost was calculated for the TB–mortality linkage. The proportion of time spent on the comparison of records and weighting was comparable to the TB–mortality linkage.

Strategy for Using Linkage System

There are three main factors which affected the cost of these linkage runs using the system described. The order in which comparisons are carried out is extremely important, as has been mentioned. Obviously it would be very costly to compare alphabetic fields first, knowing that at some point later in the comparison the records could be rejected as potential links. Efficiency can be maximized by first comparing numeric fields on the basis of which pairs of records can be immediately rejected. It may be decided, for example, that the quality of the two files concerned is sufficiently high that disagreement on birth year of more than 10 years means that the link would not possibly be believed. The second factor affecting cost is the extent to which records have missing identifying items of information. If one or both files contain many records with very little information present, these records will generate large numbers of potential links because there is little or no basis on which to reject these links, i.e., there will not be a sufficient number of disagreements to bring the disagreement weight below the cutoff weight. As a result, comparison of records takes longer since more records go through the comparison of all items and weighting will also be more expensive due to the volume of potential links. The third consideration is the setting of the cutoff weight. The apparent efficiency of a linkage may be increased by using a less strongly negative cutoff weight. However, depending on the purpose of the application, this may have subsequent adverse effects. If only the definite links are of interest, no problems may arise, but if the purpose of conducting the linkage is statistical analysis, it is then important to be able to identify the records or individuals whose status is unknown. This is the case with respect to the applications described here.

Conclusion

The system which was developed provides a very powerful tool for medical research in general, and the concepts can be implemented fairly readily on any

medium-sized computer. Since the processing is sequential in general it can also be adapted to any small installation which has the facility for processing large volumes of sequential data.

ACKNOWLEDGMENTS

The authors wish to acknowledge the contributions of systems analysts Ted Hill and Steve Hobbs, and methodologists Simon Cheung, Mike Eagen, and Dave Binder.

REFERENCES

1. MACMAHON, B., AND PUCH, T. F. "Epidemiology Principles and Methods." Little, Brown, Boston, 1970.
2. HOWE, G. R., LINDSAY, J., COPPOCK, E., AND MILLER, A. B. Isoniazid exposure in relation to cancer incidence and mortality in a cohort of tuberculosis patients. *Int. J. Epidemiol.* **8**, 4, 305 (1979).
3. HOWE, G. R. Breast cancer mortality in relation to fluoroscopic X-ray exposure. Presented at the 4th International Symposium of the Detection and Prevention of Cancer, London, July 1980.
4. HOWE, G. R., LINDSAY, J., AND MILLER, A. B. A national system for monitoring occupationally related cancer morbidity and mortality. *Prev. Med.*, in press.
5. SMITH, M. E., AND NEWCOMBE, H. B. Methods for computer linkage of hospital admission-separation records into cumulative health histories. Methods of Information in Medicine **14**, 118 (1975).
6. FELLEGI, I. P., AND SUNTER, A. B. A theory for record linkage. *J. Amer. Stat. Assoc.* **64**, 1183 (1969).
7. LYNCH, B. T., AND ARENDS, W. L. "Selection of a Surname Coding Procedure for the SRS Record Linkage System." U.S. Department of Agriculture, Washington, D.C., 1977.