Medical Applications

M. WOODBURY, Editor

Identification or Record Linkage by Name, Address, Date of Birth, etc.—*Because of the key role identification routines will play in long-term medical follow-up studies, this department would like to review manuscripts describing systems in use or under development. Theoretical and analytic papers are particularly helpful to readers undertaking new problems. Programs unsuitable for your application may be appropriate for others. Extra-medical experience is particularly invited. Generalized programs that are self-revising according to the error frequencies encountered need development.*

N. C. Webb, Jr.,
BIO ad hoc Committee On Patient Registers and Patient Identification
Max Woodbury, Editor, Medical Applications

Person-Matching by Electronic Methods

William Phillips, Jr., Anita K. Bahn, Mabel Miyasaki*

National Institute of Mental Health, Bethesda, Maryland

Record linkage in the updating of files is accomplished in many establishments through the use of a preassigned number, such as payroll number, customer number, or social security number. In vital and health records, however, a unique number is generally not preassigned to an individual for purposes of reporting services received to the health department. In order to determine whether different physician reports refer to the same individual, name and other identification must be compared. This is a laborious operation which is subject to various errors because of name misspellings, changes of name upon marriage, and other problems.

We are interested in the maintenance of a psychiatric case register in Maryland, where many of the reports from over a hundred psychiatric agencies refer to the same patient. These records must be linked in order to provide unduplicated counts of individuals under care and longitudinal records of psychiatric history. An earlier paper [1] describes our general procedures for register maintenance by use of a digital computer (Honeywell 800). Here we present in more detail our initial procedures for the person-matching process in order to elicit comments and suggestions from persons who have had experience in matching.

Problem

It is necessary to establish rules for automatic search and decision which duplicate as far as possible the best clerical judgment regarding person-matching. Although rules cannot cover all situations they should provide a logic which would be equally applicable to non-automatic operations, if there were sufficient clerks to handle the job.

At present we lack empirical data on the frequency in our patient population of specified surnames and first names, nicknames, misspellings, incorrectly reported birth years, and the like. Our rules for initial matching, therefore, reflect only our own judgment and the experience we have gained from 'hand" matching a sample list of psychiatric clinic patients [2] and from comparing identification data in multiple reports for patients readmitted to the same facility.

Our initial procedures will be largely "trial and error.' Therefore: (1) the outcomes of each matching operation (see reference codes in Tables 1 and 2) will be listed and clerically reviewed; (2) Our rules are conservative; "possible matches" are referred for clerical investigation when there is categorical uncertainty whether to accept or reject a match. A record will be kept of the proportion of these "possible matches" that turn out to be positive.

From (1) and (2) we can empirically determine the wisdom of each rule, by its yield of positive and false matches. Based on these findings, we will modify our automatic procedures so as to eliminate unprofitable sorting and searching operations, increase the resolution achieved by the computer, and reduce the amount of clerical decision.

Identifying Information

The rules for person-matching depend upon the identifying information available for comparison. The information requested on psychiatric reports is limited to surname, given name, middle name, maiden name if a married woman, birth date, sex, race, street address, and social security number. Although maiden name of mother and birthplace would be desirable, this information cannot be requested at present.

In addition to misspelling of name and the usual inconsistencies, other problems are likely to be encountered because of the nature of our population. For example, birth date and address are sometimes not reported for emergency cases. The Baltimore City population is highly mobile and change of address is frequent. Aliases may be used.

Factors of identification have a varied role. Concordance in address can verify a match, but discordance does not necessarily disprove it. On the other hand, maiden name and social security number should rarely be discordant. Maiden name, social security number, birth date or

* Mr. Phillips and Miss Miyasaki are Digital Computer Programmers and Dr. Bahn is Chief, Outpatient Studies Section, Biometrics Branch.

address, in addition to confirming agreement, can associate two records whose common identity would not otherwise be detected by the manner in which the surname is misspelled (soundex code cannot overcome all misspellings), by the use of an alias, or by a change of name due to marriage.

In date of birth comparisons, we place greater reliance on month and day than on year. Experience indicates that mothers reporting for children are more likely to remember month and day than birth year which they frequently must compute from age. Adult respondents reporting for themselves may, because of vanity, state a more recent birth year than is true.

The first or given name poses a problem, since a nickname may be used in place of the given name. We are therefore exploring methods whereby a common nickname may be called into memory for comparison.

Tentative Procedures

Since it is not practical to compare every record against every other record, it is necessary to find a primary factor or key for grouping records for purposes of comparison. Any factor which is present in all records and divides the file into workable segments can be used. It is our plan to use several different factors for primary sorts — name, birth date, social security number, and address.

Detailed records of all accretions (new admissions to psychiatric facilities and readmissions with case numbers not previously reported) will be automatically assigned a soundex code according to the Russell Soundex System with the modification that the first letter as well as the next three consonant sounds are coded.[1] The records will then be processed through a name (Soundex) check program which will compare these accretions with the master name file of previously reported cases. If a match is not accepted (i.e., register number not located), the case will then be compared with the master file using a birth date check program. Subsequent checks by social security number and address are under consideration. Tables 1 and 2 enumerate for the first two programs the specific matching procedures, possible outcomes, and automatic decision for each outcome. The Tolerance Rules of Concordance for each program are also specified. Figure 1 (p. 406) outlines the general computer processing for these programs.

In the Soundex program (Table 1), the accretions are compared with each name in the master file with the same four-digit soundex code. The first check in this program compares surname, address, first name and birth year. Secondary factors to aid in positive identification are the social security number and maiden name if given. The final group of factors or third check if agreement is still in doubt consists of sex, race and complete birth date.

[1] In deciding upon Soundex, we were influenced by the fact that the Social Security Administration maintains a file of over 195 million account number holders with this system.

TABLE 1. SOUNDEX CHECK FOR PERSON-MATCHING

		FIRST CHECK					SECOND CHECK			THIRD CHECK			
Reference code	Soundex code	Surname	First name	Addr.	Birth year range	Decision	Soc. Sec.#	Maid. name	Decision	Sex & race	Birth mo. & day	Birth year	Decision
A0000	0	0	0	0	0	Accept							
A0100	0	0	0	0	1	Possible	0	0	Accept				
A0110							0	1	Accept				
A0120							1	0	Accept				
A0130							1	1	Possible	0	0	1	Accept
A0131										0	1	1	Reject
A0132										1	0	1	Possible
A0133										1	1	1	Reject
A0200	0	0	0	1	0	Possible	0	0	Accept				
A0210							0	1	Accept				
A0220							1	0	Accept				
A0230							1	1	Possible	0	0	0	Accept
A0231										0	0	1	Accept
A0232										0	1	0	Possible
A0233										0	1	1	Possible
A0234										1	0	0	Accept
A0235										1	0	1	Possible
A0236										1	1	0	Reject
A0237										1	1	1	Reject
A0300	0	0	0	1	1	Reject							
A0400	0	0	1	0	0	Possible	0	0	Accept				
A0410							0	1	Accept				
A0420							1	0	Accept				
A0430							1	1	Possible	0	0	0	Accept
A0431										0	0	1	Accept
A0432										0	1	0	Possible
A0433										0	1	1	Reject
A0434										1	0	0	Possible
A0435										1	0	1	Possible
A0436										1	1	0	Reject
A0437										1	1	1	Reject
A0500	0	0	1	0	1	Possible	0	0	Accept				
A0510							0	1	Accept				
A0520							1	0	Accept				
A0530							1	1	Possible	0	0	1	Possible
A0531										0	1	1	Reject
A0532										1	0	1	Possible
A0533										1	1	1	Reject
A0600	0	0	1	1	0	Possible	0	0	Accept				
A0610							0	1	Accept				
A0620							1	0	Accept				
A0630							1	1	Possible	0	0	0	Accept
A0631										0	0	1	Possible
A0632										0	1	0	Possible
A0633										0	1	1	Reject
A0634										1	0	0	Possible
A0635										1	0	1	Possible
A0636										1	1	0	Reject
A0637										1	1	1	Reject
A0700	0	0	1	1	1	Reject							
A0800	0	1	0	0	0	Accept							
A0900	0	1	0	0	1	Possible	0	0	Accept				
A0910							0	1	Accept				
A0920							1	0	Accept				
A0930							1	1	Possible	0	0	1	Accept
A0931										0	1	1	Reject
A0932										1	0	1	Possible
A0933										1	1	1	Reject
A1000	0	1	0	1	0	Reject							
A1100	0	1	0	1	1	Reject							
A1200	0	1	1	0	0	Possible	0	0	Accept				
A1210							0	1	Accept				
A1220							1	0	Accept				
A1230							1	1	Possible	0	0	0	Possible
A1231										0	0	1	Possible
A1232										0	1	0	Possible
A1233										0	1	1	Reject
A1234										1	0	0	Possible
A1235										1	0	1	Possible
A1236										1	1	0	Reject
A1237										1	1	1	Reject
A1300	0	1	1	0	1	Reject							
A1400	0	1	1	1	0	Reject							
A1500	0	1	1	1	1	Reject							

NB: 0 indicates agreement;
1 indicates discrepancy between the records

TOLERANCE RULES FOR CONCORDANCE:

Surname—In a one-to-one correspondence of the first 8 letters, only one disagreement allowed

First name—In a one-to-one correspondence of the first 8 letters, only one disagreement allowed

Address—Agreement on street number and first 8 letters of street name

Birth year range—

If current age is:	Range must be within:
0–17	2 years
18–29	5 years
30–49	10 years
50 and over	15 years

Complete agreement required for *social security number, maiden name, sex, race, birth month and day,* and *birth year*

TABLE 2. DATE OF BIRTH CHECK FOR PERSON-MATCHING
To be applied to all "Rejects" and "Possibles" from Soundex Check

Reference code	Birth mo. & day; Sex	Social Sec. #	Birth year	Maiden name	First name	Address	Birth year range	Decision
B01	0	0						Accept
B02	0	1	0	0				Accept
B03	0	1	0	1	0			Accept
B04	0	1	0	1	1	0		Possible
B05	0	1	0	1	1	1		Reject
B06	0	1	1	0	0			Accept
B07	0	1	1	0	1			Possible
B08	0	1	1	1	0	0	0	Accept
B09	0	1	1	1	0	0	1	Possible
B10	0	1	1	1	0	1		Reject
B11	0	1	1	1	1	0	0	Possible
B12	0	1	1	1	1	0	1	Reject
B13	0	1	1	1	1	1		Reject

NB: 0 indicates agreement;
1 indicates discrepancy between the records

TOLERANCE RULES FOR CONCORDANCE:

Social security number—Complete agreement
Maiden name—Complete agreement in either maiden names or in cross-check with surname
Birth year, first name, address, birth year range—Same as in Table 1

Whenever a match is accepted, further checking of the accretion is discontinued. The previously assigned register number is added to the accretion record which is then placed in a "located" tape file. Several "possibles" may be encountered before an "acceptance" decision is made. In this case, the "possibles" are nullified. An accretion is not considered "unlocated" or rejected as a match until it has been compared to each name of the master file with same soundex code. All accretions which do not encounter a "positive" match are entered on an "unlocated" tape file.

The "unlocated" file and the master file are then rearranged by month and day of birth, and by sex. Each accretion is checked to every master record having the same month and day of birth and sex (Table 2). Social security number, birth year, maiden name, surname and address are used for verification. Positive matches are processed as in the Soundex program.

Plans for further checking of the "unlocated" records include a program using the social security account number as a primary factor and a program using the address as a primary factor in order to associate records where the surname is entirely different or the date of birth is incomplete or missing. The social security number check will be a simple collation process. Records will be read one at a time from the accretion file and from the master file. Whenever a match is encountered, the match will be verified by sex, race, year of birth and possibly first name. The address check is also a collation process but is complicated by the possibility of more than one member of a family being admitted during the same six month period. This involves holding more than one record from each file in core memory and the "forecasting" of a change of address in each file.

Listings of all matches from the above programs will be clerically reviewed and in some cases checked to the original report or to the patient case folders at the reporting facilities. The purpose is twofold: to correct discrepant information in our files, and to verify the appropriateness of the match before the accretion record is added to the registrant's earlier psychiatric history.

"Unlocated" accretions must be further processed before being assigned a new register number, since there may be several reports for the new psychiatric patient. Therefore, all non-matches must be compared with each other using programs similar to those previously described except that there will be an internal cross-check of records rather than a match of accretion to the master file. Experiences referring to the same individual will be combined, and all resulting "unlocated" records will then be automatically assigned new register numbers.

Discussion

Our file will begin with approximately 27,000 reports. The annual accretion of admissions is about 19,000. The proportion of these that link to the established master file is expected to fall between 5 and 15 percent for the first year, and to increase as the records of psychiatrically known individuals accumulate.

In our checking procedures, arbitrary weights have been assigned to each factor. After a considerable number of matches have been processed, we may be able to use these results to establish mathematical weights, as was done by Newcombe [3].

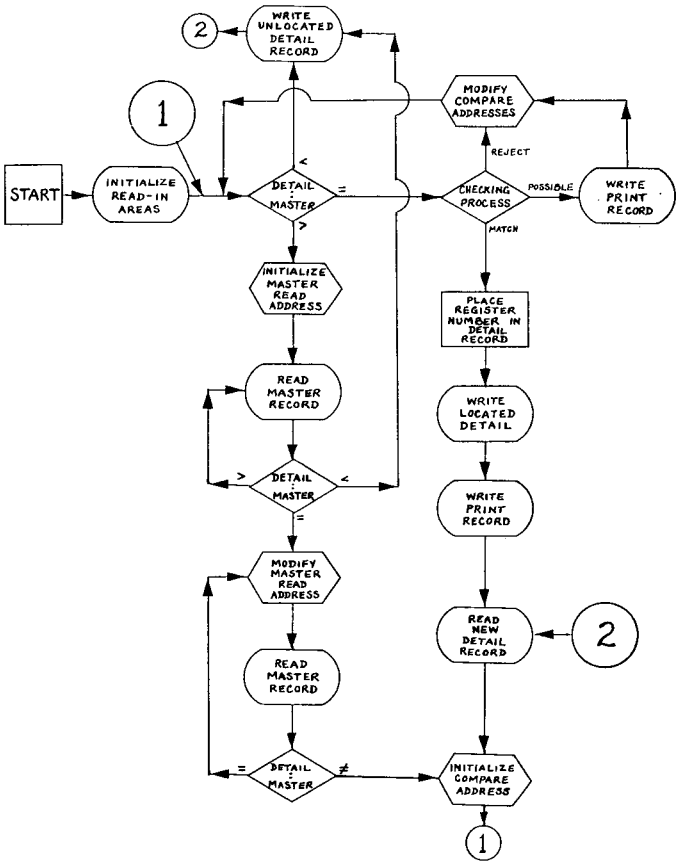

FIG. 1. Flow of processing for Soundex check and birth date check programs

Although no two projects are exactly alike, our current plans may be of use to others venturing into this field. The results of our experiences should be ready within the next year and will be available to interested parties.

REFERENCES

1. PHILLIPS, W., JR., GORWITZ, K., AND BAHN, A. K. The electronic maintenance of a chronic disease register. Presented at the annual meeting of the American Statistical Association, Dec. 28, 1961.
2. BAHN, A. K. Methodological study of population of outpatient psychiatric clinics, Maryland 1958–1959. PHS Pub. 821 (Pub. Health Monog. No. 65). Washington, D. C., U. S. Government Printing Office (1961), 105 pp.
3. NEWCOMBE, H. B., KENNEDY, J. M., AXFORD, S. J., JAMES, A. P. Automatic linkage of vital records. *Science 130* (Oct. 16, 1959).

A Computer Method for Radiation Treatment Planning

William Siler and John S. Laughlin

Department of Physics, Memorial Hospital for Cancer and Allied Diseases, New York, N. Y.

Automatic computation methods were first developed and applied to the problem of radiation therapy treatment planning by the Physics staff at Memorial Hospital and Sloan-Kettering Institute in 1954 and reported in 1955 [1]. The field of radiation from a single port was stored as a matrix in a library of punched cards, and a sorter and accounting machine were used to combine various fields for rotation, cycling and multi-port therapy. This system was in continuous routine use from then until 1961, when the equipment was replaced by a Bendix G15-D digital computer. Subsequent work by Sterling [2] followed essentially the same method of describing the radiation field as used by the Physics staff at Memorial Hospital [1], except that more powerful equipment has been used. An analytic expression for the dose distribution produced by rotation had been previously applied successfully in 1951 to treatment planning with high-energy X-rays [3].

In 1959, it appeared that the then existing system would shortly be made obsolete by certain changes in the equipment available for radiation therapy. After restriction of the number of field sizes and shapes to be punched for automatic calculation to a practical minimum for the accuracy required in therapy, a library of some 3,000,000 punched cards would still be required. This being impractical, a new method of automatic computation was looked for which would hopefully accomplish three long-desired ends: reduce data input required to a minimum; permit automatic plotting of isodose curves of the radiation field resulting from all practical combinations of ports; and permit taking into account the effects of body inhomogeneities (e.g. bones, lung and muscle).

The first attempt was simply to reduce the amount of data input required by the existing system. A special-purpose switching device was designed to be connected to and controlled by the accounting machine, permitting automatic rotation of the sampling grid for the dose matrix, reducing the data input by a factor of about 30. This would have reduced the punched-card library to manageable levels, but would not accomplish the aims of automatic plotting and calculation of effect of body inhomogeneities. Accordingly, a completely new description of the radiation field was devised which would permit these ends to be achieved.

If the field of radiation within a patient is given as $\bar{F} = f(p, q)$ where p and q are any convenient coordinates, then $\bar{F}$ is a function of the source-skin distance, S, the angle of the beam relative to the coordinate system α, the beam width w, and the beam length h. In the original Memorial-Sloan-Kettering system, separate matrices were punched for each variation in each of the above parameters. As mentioned above, rotation of the sampling grid may be accomplished with a suitably-equipped accounting machine and is a trivial problem for a computer, so that the parameter α is easily eliminated. The parameter S may be eliminated in two steps. First, a coordinate system is chosen for sampling the field consisting of the depth below the surface x, and the angular displacement from the beam centerline δ. Secondly, the effect of inverse-square-law attenuation is removed by defining a new field $\bar{M}$ such that

$$\bar{M} = \bar{F}\left[\frac{S + x}{S}\right]^2.$$

The field $\bar{M}$ is then digitized by sampling at equal increments in x and σ. For the primary beam, the matrix $\bar{M}_p$ is of rank one, being the outer product of two vectors, one of which represents attenuation and the other the transverse shape of the beam. (Beam quality is assumed invariant across the field.) Worthley and Wheatley [4] have reported that a similar relationship exists for scattered radiation in the range of half-value layers from 1 to 2 mm Cu and for source-skin distances from 40 to 100 cm; i.e., the scattered radiation field is given by

$$\bar{M}_s = u(x) \cdot v(\sigma)$$

and is also of rank one. The matrix of total radiation $\bar{M}$,