Research Paper ■

Record Linkage of Health Care Insurance Claims

TIMOTHY W. VICTOR, PHD, ROBERTINO M. MERA, MD, PHD

Abstract **Objective:** This paper provides a detailed description of a method developed for purposes of linking records of individual patients, represented in diverse data sets, across time and geography.

Design: The procedure for record linkage has three major components—data standardization, weight estimation, and matching. The proposed method was designed to incorporate a combination of exact and probabilistic matching techniques.

Measurements: The procedure was validated using convergent, divergent, and criterion validity measures.

Results: The output of the process achieved a sensitivity of 92 percent and a specificity that approached 100 percent.

Conclusions: The procedure is a first step in addressing the current trend toward larger and more complex databases.

■ **J Am Med Inform Assoc.** 2001;8:281–288.

Linkage of records in the interest of science has a long pedigree. The term *record linkage*[1] indicates the bringing together of two of more separately recorded pieces of information concerning a particular person or family.

In 1929, R. A. Fisher argued for the usefulness, in human genetics research, of public records supplemented by (and presumably linked with) family data.[2] Earlier, Alexander Graham Bell exploited apparently linked genealogical records and administrative records from marriages, census results, and other sources to sustain his familial studies of deafness.[3,4]

Affiliations of the authors: Healthcare Informatics (TWV) and Health Economics and Outcomes Research (RMM), SmithKline Beecham, Collegeville, Pennsylvania.

Correspondence and reprints: Timothy W. Victor, Assistant Director, Research and Biostatistics, Healthcare Informatics, SmithKline Beecham, MS UP4305, 1250 South Collegeville Road, Collegeville PA 19426-2990; e-mail: <timothy.w.victor@sbphrd.com>.

Received for publication: 9/19/00; accepted for publication: 1/9/01.

Newcombe and Kennedy[5] undertook the first rigorous treatment of record linkage. They introduced the concept of matching weights based on probabilities of chance agreement of component value states. Fellegi and Sunter[6] extended and formalized these concepts into a more rigorous mathematical treatment of the record linkage process. The Fellegi–Sunter model provides an optimal means of obtaining weights associated with the quality of the match for pairs of records. Most probabilistic matching procedures in use today are derived from the techniques described in the Fellegi-Sunter paper.

Practical Implications

Problems that hinder the implementation of record linkage methods include poor data quality, lack of computational efficiency and complex software in the case of probabilistic matching, limitations of deterministic matching, and legal restrictions.

In addition, data often originate in heterogeneous computer systems, so that the analyst has no control or influence over the data collection or the data entry process. This heterogeneity yields data of variable

quality that are prone to error. Another problem is the enormous size of the data sets to be linked. Linking must occur both within and between data sets. The unreliability of the identifying information contained in successive records of the same subject presents a significant challenge.

Specifically, limitations of either matching technique include the difficulty in handling large and heterogeneous data sets, the need to provide a priori weights for probabilistic matching, and the need to generate unique identifiers. Furthermore, exact matching has low sensitivity, and probabilistic matching is computationally expensive in large data sets.

This paper describes a method developed for linking records of individual patients and health care providers across time and geography. The method was designed to incorporate a combination of exact and probabilistic matching techniques. Finally, this method provides a static unique identifier for every patient and health care provider.

Methods

Data Sources

The data used in this research come from various commercial insurance claim transaction databases used in various health-care-related research programs. These streams include eligibility, pharmacy, laboratory, hospital, and doctor claims. Individual patients and physicians may have multiple nonunique identifiers both within and among databases.

The full demographic data set contains more than 52 million rows of data, representing more than 20 million persons over a five-year period. Each person is represented by an average of 2.3 identifiers. Figure 1 illustrates the data elements used in the matching process.

Nomenclature

Consider two files, A (input) and B (reference), containing demographic data for patients from two populations, which will be denoted by a and b, respectively. We assume that some patients are common to A and B. Consequently, the set of ordered pairs

$$A * B = \{(a,b);\ a \in A,\ b \in B\}$$

is the union of two disjoint sets, the matched set (M) and the unmatched set (U), where

$$M = \{(a,b);\ a = b,\ a \in A,\ b \in B\}$$

and

$$U = \{(a,b);\ a \neq b,\ a \in A,\ b \in B\}.$$

In the matching process, each record in file A can be compared with each record in file B. The comparison of any such pair of records can be viewed as a set of outcomes, each of which is the result of comparing a specific attribute from the record in file A with the same attribute in the record from file B. Outcomes may be defined as specifically as desired.

This implies that every record from file A is compared with every record from file B. In practice, with large files this would require an extremely large number of comparisons, the vast majority of which would not be matches. In fact, the number of comparisons would be $A * B$. To make the size of the problem more manageable, files are generally "blocked" using one or more of the available attributes. These blocked record pairs are assumed to be possible matches and subject to the detailed attribute comparison. When using a blocking procedure, the number of records with unmatched blocking attributes is necessarily higher, and these records are automatically rejected as possible matches. However, the achieved gain, in the form of reduced processing, is significant.

Definitions

An *exact* or *deterministic* match is defined where specified attributes in data vector A match the same specified attributes in data vector B. A data vector is defined as a collection of patient attributes, such as demographic data.

A *probabilistic* match is defined where some attributes of the data vector A match some attributes of data vector B, under a level of agreement that surpasses a certain a priori defined threshold.

Procedure for Record Linkage

The procedure includes three major components— data standardization, weight estimation, and matching.

Data Standardization

The first step of this process involves the standardization of data in an input file. Data standardization is suggested for increased matching precision and reliability. The input file can contain any number of variables, of which one or more may be unique to a particular entity. Examples of useful variables are member identifier, driver's license number, Social Security number, insurance company code number, name, gender, date of birth, street address, city, state, postal code, citizenship. In addition, some identifiers can be

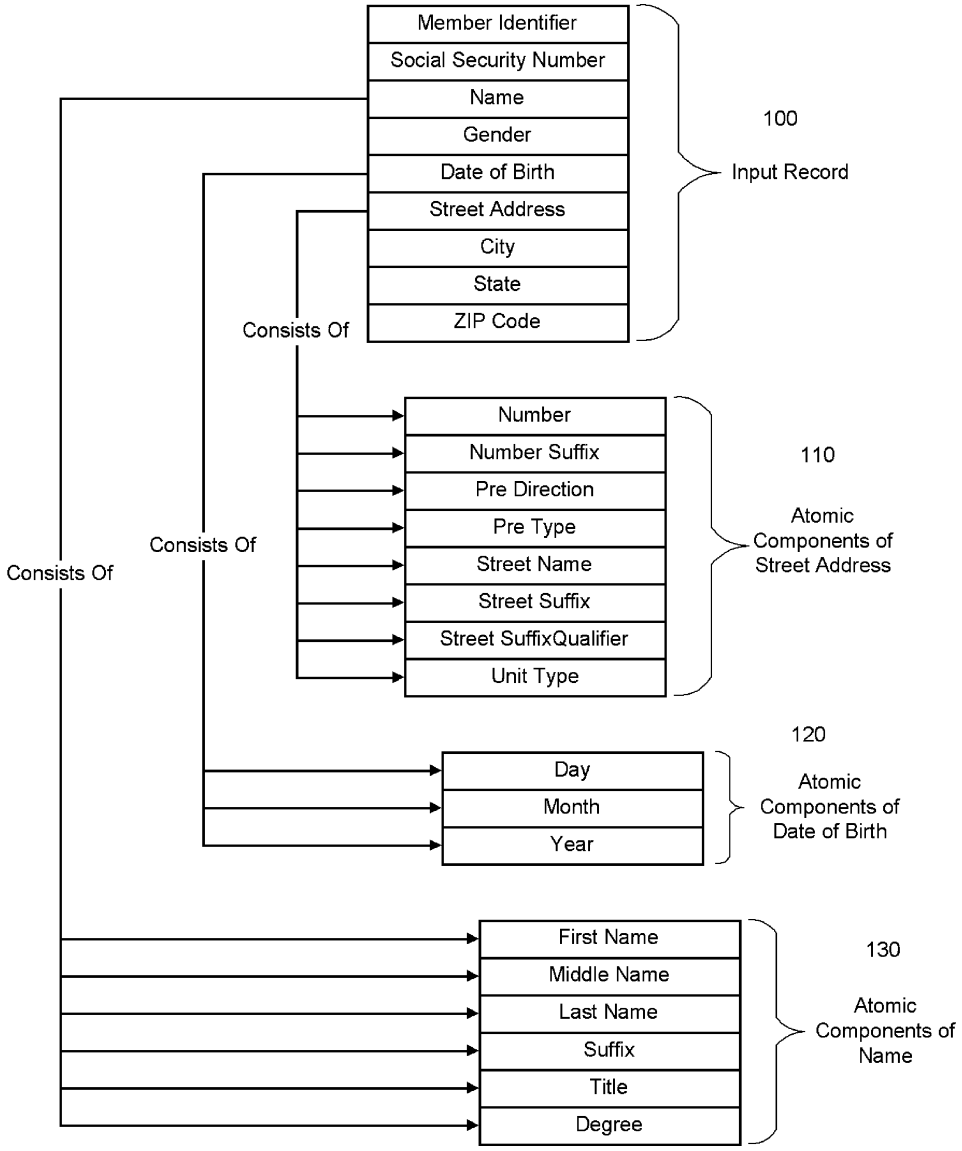

Figure 1 Input record and atomic parts.

further distilled down into their basic, or atomic, components. Figure 1 illustrates the use of selected input record components and atomic components of some records that are amenable to such further distillation. Input Record 100 illustrates data that can be used as the basis for assigning a unique identifier and how the data can be broken out into atomic and subatomic components, exemplified by Street Address 110, Date of Birth 120, and Name 130.

During the standardization process, all character data are preferably transformed to a single case. For example, they may be transformed to uppercase. In addition, given names, street, and city are standardized. So, for instance, first names are standardized, e.g., {BOB, ROB, ROBBY} = ROBERT. Common names for cities and streets may be transformed, {KENSINGTON, FISHTOWN, PHILADELPHIA} = PHILADELPHIA, to the postal code—e.g., to the U.S. Postal Service standard in the United States. In the latter instance this can be done using industry standard Coding Accuracy Support System (CASS) certified software.

Weight Estimation

A match weight is the measure that conveys the discriminating power of a variable. For example, the match weight assigned to a Social Security number variable is greater than that of a sex variable because the amount of information provided to the decision process is superior.

Agreement and disagreement weights need to be estimated and are required for the probabilistic match. Weights are calculated on the basis of proba-

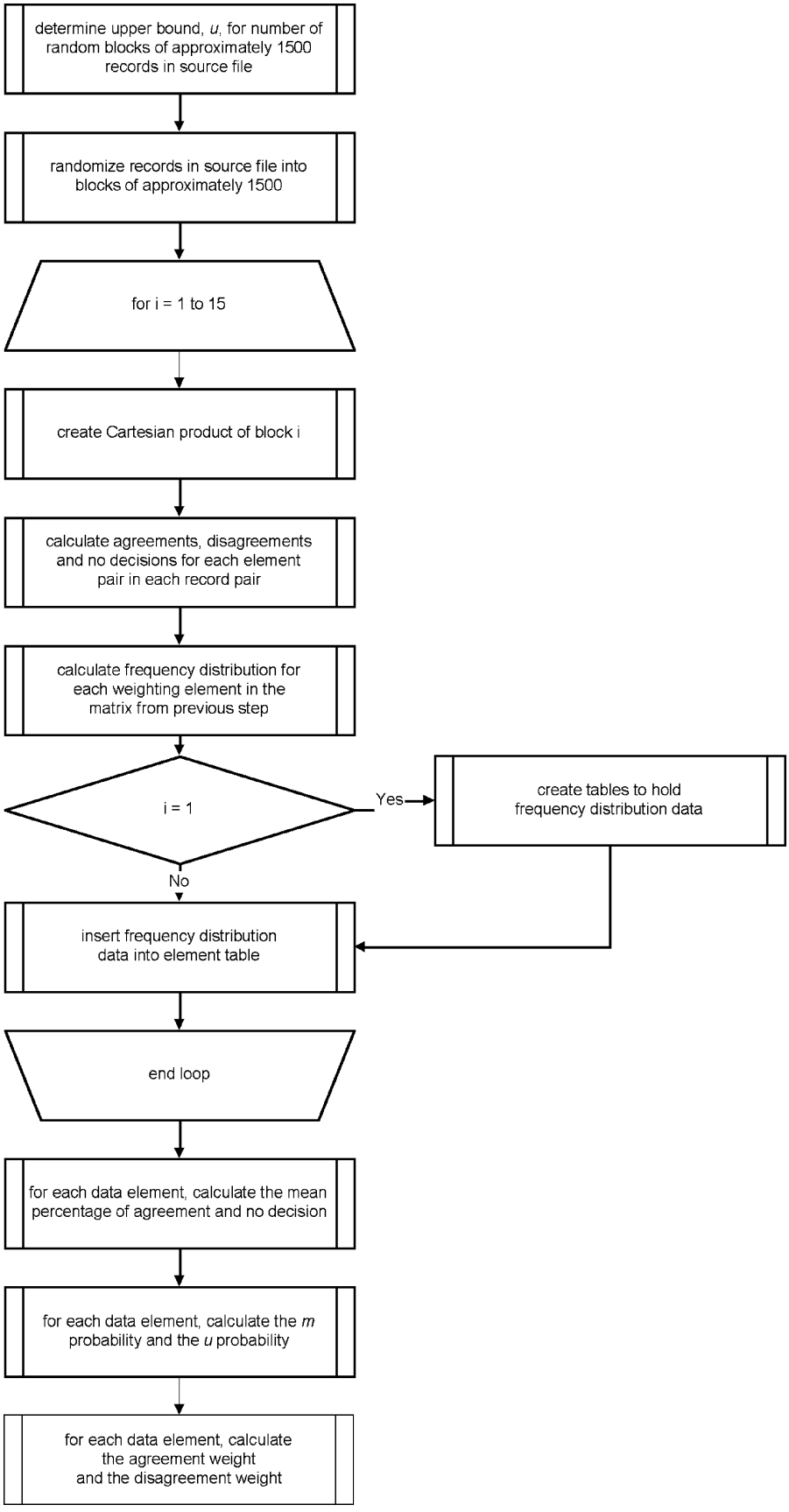

Figure 2 Weight calculation process.

bilities of chance agreement and the reliability of data employing a resampling technique. Figure 2 shows the flow of the process.

Chance agreement is defined as the likelihood of a match between a pair of elements. *Data reliability* is defined as the consistency of the attributes of the elements. For example, in the gender element we may have a consistent rate of 2 percent of attributes that are not *M* or *F*, e.g., null. The reliability index for this example is 100 minus 2, or 98, percent.

The first step in the weight estimation process is to determine the number of strata (υ) required such that the data set can be divided into approximately equal blocks of 1,500* rows:

$$\upsilon = \text{int}\left(\frac{\text{number of records in data set}}{1{,}500}\right) \quad (1)$$

The source file is then scanned, and the records are assigned a random number between 1 and υ. A data set is created from those records with a random number of 1. Note that $\upsilon * 1{,}500$ is approximately equal to the total number of records in the data set. A Cartesian product (the result of joining two relational tables, producing all possible ordered combinations of rows from the first table with all rows from the second table) is created from these sampled data. The resulting matrix is then scanned. Each element pair within each record pair is assessed and assigned a value, e_n, in the following manner:

$$e_n = \begin{cases} 1 \text{ if } A_{e_n} = B_{e_n} \text{ (agreement)} \\ 0 \text{ if } A_{e_n} = \textit{null} \text{ and/or } B_{e_n} = \textit{null} \text{ (no decision)} \\ -1 \text{ if } A_{e_n} \neq B_{e_n} \text{ (disagreement)} \end{cases} \quad (2)$$

where A_{e_n} is the nth element from record A.

Once the matrix has been fully assessed, percentages for each e_n are tabulated and stored. This process is repeated for, say, 15 iterations.

Mean percentages of *Agreements* and *No Decisions* are calculated for each data element. The reliability ρ for each data element is then calculated.

If we let $\mathring{a} = \bar{x}_{\text{Percent No Decision}}$

$$\rho = \begin{cases} \text{if } \mathring{a} \geq 0.99 \text{ then } 1 - \mathring{a} \\ \text{else } 0.99 - \mathring{a} \end{cases} \quad (3)$$

The μ probability, or the probability that element n for any given record pair will match by chance, is calculated:

$$\mathring{a} = \bar{x}_{\text{Percent No Decision}} \quad (4)$$

From the ρ and μ probabilities, the disagreement and agreement weights are calculated employing equations (5) and (6), respectively:

$$\textit{Disagreement} = \log_2\left(\frac{1-\rho}{1-\mu}\right) \quad (5)$$

$$\textit{Agreement} = \log_2\left(\frac{\rho}{\mu}\right) \quad (6)$$

Notice that equations 5 and 6 are ratios of the attribute reliability and the probability of a random match.

Matching

The final stage of this procedure is the action of uniquely identifying entities in the input data set. Figure 3 provides an overview of this process.

Each record from the input file A is evaluated against a reference file B to determine whether the person represented by the data has been previously identified using a combination of deterministic and probabilistic matching techniques. If it is judged that the person is already represented in the reference set, the input record is assigned the unique identifier (UID) from the reference record that it has matched against. If it is judged that the entity represented by data is not yet in the reference set, a new, unassigned UID is randomly generated and assigned.

After the UID assignment occurs, the input record is evaluated, in its entirety, to determine if the record is a unique representation of the entity not already contained in the reference table. If it is a new record, then it is inserted into the reference table for future use.

Exact Matching. The deterministic matching technique employs simple Boolean logic. Two records are judged to match if certain criteria are met, such as the following:

First name matches exactly

Last name matches exactly

Date of birth matches exactly

Social Security number matches exactly

If two records satisfy the criteria for an exact match, no probabilistic processing occurs. However, if no exact match occurs, the input record is presented for a probabilistic match. The deterministic intersections in our data sets have averaged approximately 60 percent.

* An arbitrary number which is large enough such that the sample is representative of the dataset, but not so large as to become computationally prohibitive.

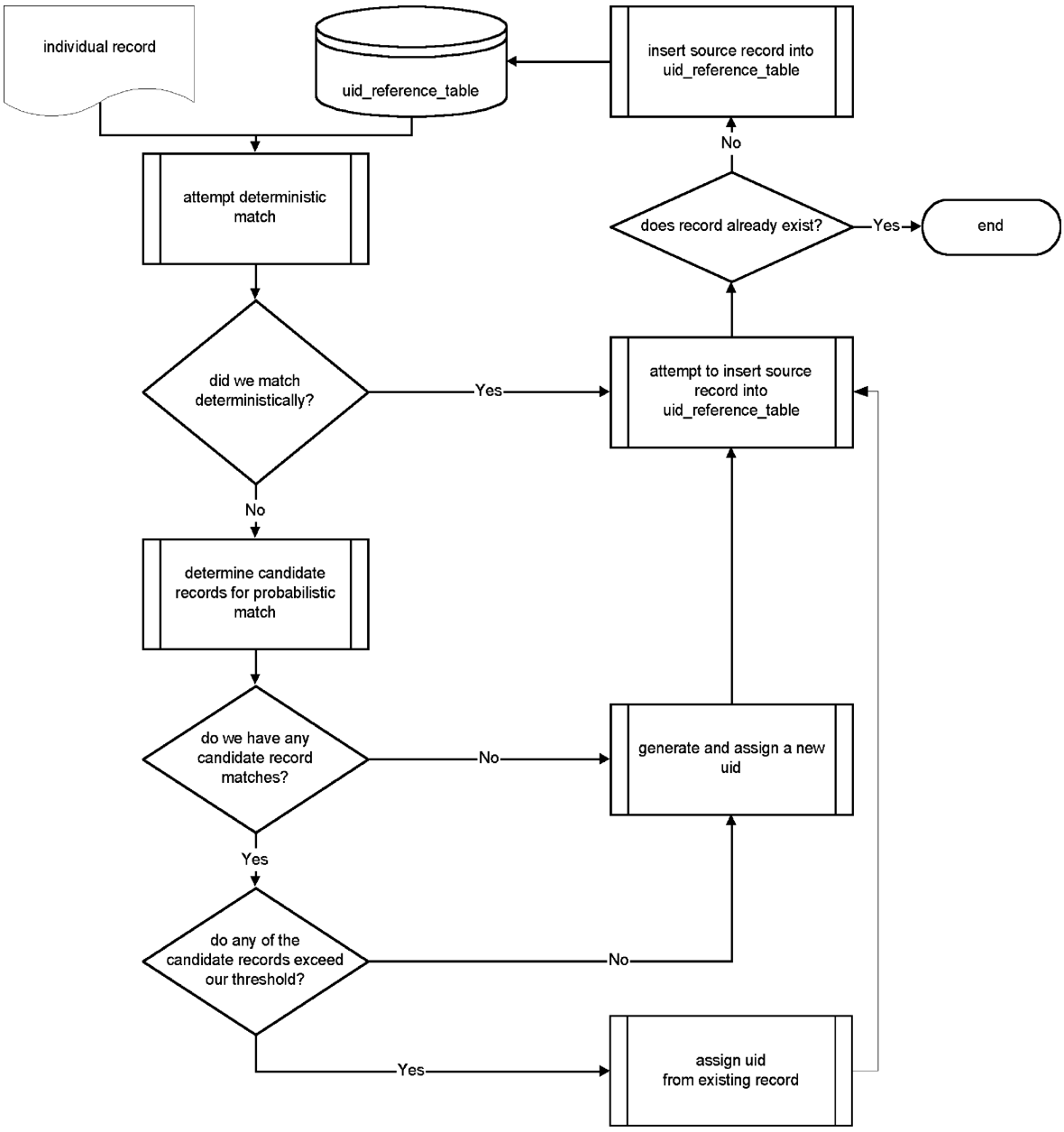

Figure 3 Record linkage process.

Probabilistic Matching. The first step in the probabilistic matching procedure is to build a set of candidate records from the reference table maintained from previous match runs, based on characteristics of specific elements of the input record. This process is referred to as blocking; the set of candidate records is referred to as the blocking table. Generally, all data sets do not use the same blocking variables; the selection of these variables depends on the characteristics of the data set. Moreover, it is suggested that blocking variables consist of those elements that are somewhat unique to an element, e.g., Social Security number or a combination of year of birth and last name.

On completion of populating the blocking table, each element for each candidate record is compared against its corresponding element from the input record. Equation 7 shows the scoring mechanism.

$$w_n = \begin{cases} \text{Agreement weight if } A_{e_n} = B_{e_n} \\ 0 \quad \text{if } A_{e_n} = null \text{ and/or } B_{e_n} = null \\ \text{Disgreement weight if } A_{e_n} \neq B_{e_n} \end{cases} \tag{7}$$

where A_{e_n} is the nth element from record A.

Summing the agreement and disagreement weights over all variables yields a composite weight; the formula can be seen in equation 8.

$$W = \sum_{i=1}^{N} w_i \quad (8)$$

This composite weight indicates the likelihood that any two records represent the same person. Composite weights will have high negative values when there is little or no agreement between the attributes of any two given vectors. Conversely, high positive values are observed when there is considerable agreement between any two vectors.

The candidate record with the highest composite weight, *W*, is then evaluated against a predefined threshold. If the weight meets or exceeds the threshold, the candidate record is judged to match the input record. If the weight does not exceed the threshold, it is assumed that the input record represents an entity not yet included in the reference file B.

The analyst must take care to choose a threshold that simultaneously maximizes both sensitivity and specificity. The process of choosing a threshold is as follows:

- Randomly select *n* records from file A, where *n* is large, say, 2000.
- Create a Cartesian product from the selected records.
- Calculate the composite weight for all n^2 record pairs using the method described above.
- Sort the matrix by composite weight in descending order.
- Select a weight that best distinguishes between sets U and M. This is done by the analyst.

Figure 4 illustrates a typical distribution of frequency weights.

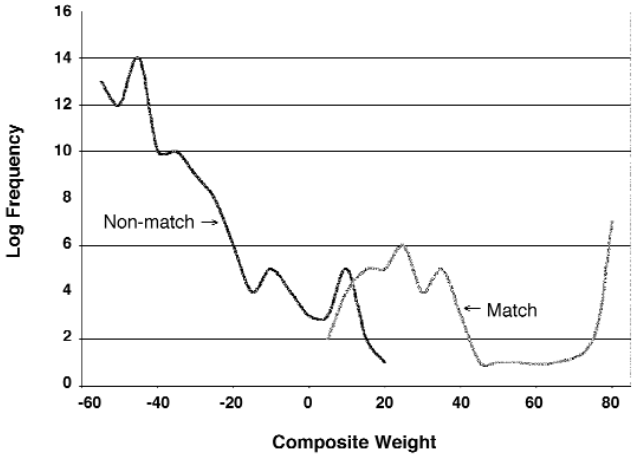

Figure 4 Distribution of frequency weights.

Validation

Validation was required to ascertain the quality of the linkage procedure. The procedure was validated using divergent, convergent, and criterion validity.

Divergent validity is an index of dissimilarity; things that are not related should be not correlated. *Convergent validity* is an index of similarity; things that are similar should be correlated in a predictable direction. *Criterion validity* uses a standard to indicate the accuracy of an indicator.

The divergent validity measure consisted of files taken from two sources orthogonal with regard to geography and source; however, they were related with regard to time. The files were submitted to the procedure with the expected outcome of little or no overlapping records. The underlying assumption is that people cannot be in two places at the same time and that the files should, therefore, be unrelated. No records representing the same persons were found in both files.

The convergent validity measures consisted of files taken from sources related with regard to source and geography. However, the time attributes of the two files differed. The files were submitted to the procedure with the expected outcome of an intersection of the two files. The underlying assumption is that data coming from the same source but different time frames will have duplicates of sampling units. There was an intersection of approximately 92 percent between the two files.

Criterion validity was assessed using a triangulation approach. Triangulation is the process of collecting three or more types of data to help confirm, revise, or reject results. Three sources of reference were identified for sampling unit count in a given data set—a manual matching process, the results of the matching process, and a method of probabilistic population estimation (PPE) developed by Banks and Pandiani.[7]

A random sample of 1,000 unique identifiers was selected from the reference table for validation. Two data analysts, who were independent of each other and were not the authors, performed a manual match of the 1,000 records. Here, human judgment was considered the gold standard to which comparisons were made. The output of the process described in this paper achieved a sensitivity of 92 percent. A discussion of the specificity of the algorithm is fruitless,

as the true-negative rate necessarily approaches 100 percent if the input file is large. The results where found to be within a 95% confidence interval of the PPE estimate.

The method has been applied and tested on data sets ranging from 500 MB to 3 GB, with the number of records ranging from 1.7 million to 8.5 million. The code was executed on a Hewlett-Packard K series machine running version 10.1 of the HP-UNIX operating system. Version 6.12 of the SAS was the programming language used to calculate the agreement and disagreement weights. The weight estimation process averaged approximately 4 hours to complete. The matching process took approximately four continuous days to process the largest file in a Sybase environment.

Discussion

The method addresses problems with current matching techniques, e.g., handling large heterogeneous data sets, simplification of a priori weight calculation for probabilistic matching, lack of sensitivity in deterministic matching, and the computational cost of probabilistic matching in large data sets. Furthermore, the procedure is extendable in that an analyst can insert business rules for decision making. An example of this is the setting of the threshold. An organization may decide that assigning one or more identifiers to a person is more desirable than assigning the same identifier to two distinct persons. Finally, the procedure can be implemented and understood by those without a strong mathematical background.

It is expected that processing time will dramatically decrease as the reference table becomes more populated. This efficiency increase is attributed to an expected greater number of deterministic matches. The method stores all unique representations of a person's demographic data in such a way that when a new input file is processed, the number of deterministic matches should be higher.

The problem with current probabilistic algorithms is that they all require the estimation of match weights. The different procedures proposed for weight estimation are difficult to implement and understand. The method discussed in this paper employs principles of probability and mathematics that are easily implemented.

The unique identifiers assigned to patients are randomly generated. This randomness helps ensure privacy and anonymity. Sequentially assigned numbers do little to hide familial relationships.

The validity of this method was assessed against convergent, divergent, and criterion indexes; for example, the sensitivity rate achieved was 92 percent. The sensitivity of any probabilistic algorithm is driven largely by the quality of the input data.

There is a continuing trend toward the automation of large databases and the effort toward data (XML) and communication (HL-7) standardization. What is needed to meet this demand is an uncomplicated, computationally efficient process that will provide accurate record matching. More efficient algorithms and advances in computer technology may enhance the procedure we have outlined.

References ■

1. Dunn HL. Record linkage. Am J Public Health. 1946;36:1412–6.
2. Box JF. R. A. Fisher: The Life of a Scientist. New York: Wiley, 1978.
3. Bruce RV. Alexander Graham Bell and the Conquest of Solitude. Boston. Mass.: Little, Brown, 1973.
4. Bell AG. The deaf: In: U.S. Department of Commerce and Labor, Bureau of the Census. Special Reports: The Blind and the Deaf, 1900. Washington, DC: US Government Printing Office, 1906.
5. Newcombe HB, Kennedy JM. Record linkage. Commun ACM. 1962;5:563–6.
6. Fellegi IP, Sunter AB. A theory for record linkage. J Am Stat Assoc. 1969;40:1183–220.
7. Banks SM, Pandiani JA. The use of state and general hospitals for inpatient psychiatric care. Am J Public Health. 1998;88:448–51.