Comparison of Probabilistic and Deterministic Record Linkage in the Development of a Statewide Trauma Registry

David E. Clark and David R. Hahn
Department of Surgery, Maine Medical Center, Portland, Maine

We have been working to develop a statewide injury surveillance system using not only hospital-based trauma registries but also other sources of data (including ambulance run reports, hospital discharge abstracts, and death certificates). For this purpose, a commercially available probabilistic matching program was compared to the deterministic program described previously. Using the same data preprocessing and linkage strategy, we programmed the probabilistic software to perform the matching step and compared the results with those obtained from the previously tested program. The outcomes using our data were similar, but we expect the probabilistic program to be more adaptable for general use, especially if large amounts of data must be linked.

INTRODUCTION

In a previous report, we described the use of multiple linked data sources in the development of a trauma registry for the state of Maine [1]. Using algorithms written in a general-purpose programming language, data from hospital trauma registries, a voluntary "trauma tracking study", vital statistics, hospital discharge abstracts, and ambulance run reports were combined in a single list without apparent duplication or failure to match obvious cases. Questionable matches were handled by progressively increasing the allowable discrepancies for certain variables.

Despite the attainment of a functioning program, this result has not been completely satisfactory for several reasons: The algorithms contain several subroutines which are dependent on knowledge of Maine geography and sociology or the idiosyncracies of different data sources; program maintenance for the future depends not only on general programming ability but also a familiarity with this specific program (which is complex and may have undetected "bugs"); there are limitations on file sizes which affect even our small state; and there is no easy way to specify or modify the assumptions inherent in the matching process.

The purpose of the present study was to determine whether, if properly customized, a commercially available general-purpose probabilistic matching program would be as accurate and easy to use as the existing deterministic program. If so, it would be more adaptable for ongoing consolidation of population-based data for injury surveillance and other purposes in Maine and elsewhere.

MATERIALS AND METHODS

Programming and data storage for this study were performed using a IBM-compatible 486-DX33 microcomputer with 234 MB of disk space and 8MB of RAM using MS-DOS 6.22. General purpose programs were written in Turbo Pascal Version 6.0 or Turbo C/C++ Version 3.0 (Borland International, Scotts Valley CA).

Table 1: Sources of data, in order of linkage:

1: Registries at largest hospitals
2: Trauma tracking study
3: Interhospital ambulance run reports
4: Hospital discharge abstracts (for first hospitals)
5: Hospital discharge abstracts (for second hospitals)
6: Death certificates
7: Prehospital ambulance run reports

0195-4210/95/$5.00

Automatch (Matchware Technologies, Silver Spring MD) was used for probabilistic matching. All source code is available from the authors.

The data used for this study were the same as those described previously [1], and the sequence of matching was unchanged (Table 1). Files contained from 259 to 11645 records. As before, separate programs were written for each source which extracted the common data elements (Table 2) and converted them to a standard ASCII array of characters. This string was then appended to the original file in each case. In the case of ambulance run report data, separate files were created for interhospital transports and for prehospital transports.

From this point forward, the data linkage was done in two ways: In the first method, the existing deterministic program was utilized exactly as previously described, comparing each source in turn to a master list, and appending any unmatched records. In the second method, Automatch was programmed to do the same thing [2], using a DOS batch file to allow repeated comparisons and updating of the master list.

Both matching methods involve the decomposition of each record into its most basic data elements (sex, age, date of admission, etc.). The deterministic method matches each data element character by character, and although some discrepancies may be allowed, two records must either be classified as the same or not the same. The probabilistic method derives a weight for each element based upon the probability that agreement or disagreement on this element increases or decreases the probability that the two records refer to the same person; the likelihood that two records match is related to the sum of these weights.

Table 2: Data abstracted from each source and used for matching records from one source to another:

Sex
Age
First hospital
- Date of admission
- Length of stay
- Outcome

Second hospital, if any
- Length of stay
- Outcome

Date of disposition from last hospital, or date of death if no hospital

The standard probabilistic approach is to derive a log-likelihood ratio: If m is the probability that a particular element agrees for records which are true matches and u is the probability that that element agrees for non-matches (by chance alone), then if that element does agree between the two records its weight will be proportional to log (m/u). Conversely, if the element disagrees, its weight will be proportional to log ((1-m)/(1-u)). The weights for each element are added to determine a composite weight for comparison of these two records [2,3].

For example (using base 2 logarithms), if we were to assign m = 1 and u = 0.5 for agreement on sex and m = 0.88 and u = 0.04 for age, then comparing a hypothetical record for a forty-year-old man on one list to a fifty-year-old man on another list would result in a composite weight of log (1/.5) + log (0.12/0.96) = 1 - 3 = -2. However, if we also assign m = 0.999 and u = 0.001 for date of injury, agreement on this element adds log (0.999/0.001) = 10 (approximately) for a composite weight of 8.

The probabilistic method requires that a subjective human decision be made to accept as true matches all those comparisons receiving above a certain composite weight and to reject all those below a certain weight. For comparisons with intermediate weight, human review of the actual records can be undertaken or an explicit decision can be made to reject or accept depending upon the goals and requirements of the particular study or application. Since our goal was to completely automate the repetitive matching process we did not plan for human review except for program development and periodic maintenance or modification.

We recognized that failing to link cases from different lists would cause overestimation of the total number of cases; however, falsely matching records from different sources could lead to major misunderstandings about the epidemiology and functioning of the trauma system. We therefore sought to maximize the specificity of the matching, that is, to minimize the possibility of making false matches, while accepting some level of failures to match.

Programming Automatch required pre-processing programs written in C to extract from each record the key used for matching and convert it to the proper form. Within Automatch, "data dictionaries" were prepared for each source,

consisting of record length, path and file, and the name, starting column, length, and missing value code for each variable in the key. Match specifications require the names of the data dictionaries, the block definitions, the acceptable matches, the cutoff weights, and other information concerning estimation and frequency analysis. After each pass through one set of blocking variables, the output can be inspected and cutoff weights can be set. The program can be customized to generate a list of possible matches for clerical review, but for our purposes we did not use this option. Finally, a report is generated, which in our case was exported to a post-processing program in C which continued the batch process by adding unmatched records from the new source to the master file in preparation for linking the next data source.

As with the previous program, special attention was required to detect hospital discharge abstracts of patients who had been transferred from one hospital to another and thus appeared on two abstracts.

Since the output of the deterministic program using 1991 data had already undergone close human scrutiny, this output was used as a standard for evaluation of the probabilistic method. In addition to a general, subjective comparison of the methods, data sources 2-7 were run against the hospital trauma registries from the two largest hospitals (the most reliable and complete source), and the differences between the two methods were inspected and analyzed by human review.

RESULTS

Automatch proved to be relatively easy to understand and program [2], and the previous strategy could be duplicated almost exactly using the probabilistic software along with auxiliary DOS and C programs.

Running the entire batch program produced apparently similar matches to the previous method. In the detailed comparison, results were as given in Table 3. On human review, the reasons for discrepancies included the following:

In 38 cases, the earlier program allowed an outcome of "transfer" with no receiving hospital to be matched to an outcome of "other", and in 34 cases, the earlier program allowed both age and date to differ by one; Automatch had not been programmed to allow these discrepancies. Almost all of these disagreements involved hospital discharge abstracts, which did not have enough identifiers to allow definite matching, even on human inspection, and it was impossible to say which program was "correct". In keeping with the philosophy of avoiding false matches, we did not undertake the more complicated reprogramming for Automatch to allow such discrepancies.

Table 3: Comparison of deterministic and probabilistic record linkage

MATCHES MADE FROM GIVEN SOURCE TO SOURCE 1

Source	Deterministic	Probabilistic	Common
2	70	70	70
3	141	139	136
4	468	445	444
5	338	316	316
4 and 5	106	83	80
6	50	49	49
7	419	402	397

In 14 cases, the deterministic program matched to a different record (many of which are similar) than the probabilistic program, thus creating a false match; in nine cases, the reverse occurred. In two cases, the deterministic program failed to match because of an identifiable programming error.

In seven cases, a match was made by the deterministic program, but missed the cutoff weight set by Automatch.

The deterministic program made five false matches for unexplainable reasons, and the probabilistic program made one such apparent error.

Although our original intent had been to perform ROC analysis of the cutoff weights for the probabilistic method using the deterministic method as a standard, this was unnecessary because a very high sensitivity and specificity were both obtained over a wide range of cutoff weights.

DISCUSSION

Obtaining accurate population-based data is highly desirable for epidemiologic research or medical system evaluation. However, a frequent problem is that individual cases in the population of interest may be contained in one or more of several lists, such as vital statistics, hospital records, or voluntary reports. None of these lists is sufficient by itself, but together they may be assumed to be nearly complete. The major complicating problem is that the lists cannot be simply combined because they are in different formats and there is no name or universal identifying information to distinguish each separate case.

For many applications involving multiple databases, enough information is present to allow an accurate human judgement about whether a record from one source refers to the same case as a record from other sources. However, this is an extremely time-consuming, error-prone, and unreproducible method except for small data sets. In general, computer methods are necessary to perform this task.

One area of interest where these problems are evident is the study of serious injuries [4]. Data collection is particularly difficult in rural areas, where there may be dozens of hospitals and pre-hospital services involved in a geographic region. While intentional injuries are less common, the death rate from vehicular trauma is significantly higher than in urban areas [5]. An effective system of injury surveillance for rural America is, therefore, a pressing need.

The increasing availability of microcomputers has led to the development of trauma registries at many major hospitals, which have been useful for research and quality assurance within these institutions. One approach to regional injury surveillance could be to encourage or compel all hospitals to maintain such a registry and then to combine the data contained in these registries at some central location. However, this approach leads to major problems involving the completeness, compatibility, and confidentiality of the data. Maintenance of a high-quality hospital registry is very expensive and may not be cost-effective for a system involving a large number of hospitals over a wide geographic area.

One compromise which has been tried in Maine (our "Tracking Study") and elsewhere is a voluntary system in which a data form is forwarded to a central location for entry into a computer database. This extends the registry concept at lower cost, but is even less likely to provide information of sufficient completeness to allow any valid assessment of the system of trauma care for an entire region. Furthermore, even if all hospitals were compelled to collect complete and high-quality data, most traumatic deaths in rural areas occur before medical attention is possible, and these, as well as other non-hospitalized cases, would not be included.

In Maine and most rural areas, medical examiners rarely perform autopsies for traumatic deaths. However, death certificates are recorded as a matter of legal routine and are therefore an inexpensive source of data which is relatively uniform across the country. Unfortunately, traumatic causes are generally under-reported on death certificates, particularly in hospitalized patients where one of the comorbid factors may be listed as the cause of death [6,7].

Combination of emergency room records and death certificates has provided useful population-based data in small geographic areas. However, the accuracy of emergency room records is poor when compared to hospital charts available at discharge [8], and many problems of clinical interest occur after the emergency room phase of care.

Hospital discharge abstracts have been mandated in many states as a result of financial concerns and there have been several attempts to use these data for epidemiologic study of trauma [9,10,11]. However, the usefulness of discharge abstract data is limited by a lack of uniformity and scant clinical detail.

Although individual databases containing injured patients collected for various purposes are

flawed and incomplete just as hospital trauma registries must be, combining multiple sources with the hospital registries as a foundation may provide a reasonably complete picture of injury epidemiology and the system of trauma care in a region. Others have recognized the value of a more comprehensive approach to trauma system evaluation in rural areas [12], but its widespread application depends upon the development of effective and easily applied computer methods.

Probabilistic record linkage is not a new concept [3], but has only recently been applied to ongoing regional injury surveillance [13]. We became aware of its possibilities through the involvement of Maine Emergency Medical Services with the Crash Outcome Data Evaluation System project of the National Highway Traffic Safety Administration, in which several databases were successfully linked using Automatch to study the effect of seatbelts in vehicular crashes [2].

As a result of the current study, we intend to use probabilistic methodology in pursuit of our goal of comprehensive data collection for trauma system evaluation and the study of injury epidemiology. We plan to make this system adaptable for additional data sources which might become available. We anticipate that it will be possible to import the data into a standard trauma registry software package in order to use its analytic and reporting capabilities.

The techniques found to be successful in this application could be transferrable to other regions attempting to achieve an inexpensive system of injury surveillance, particularly in rural areas. The deterministic program written specifically for Maine would have been much more difficult to generalize. We recommend the use of probabilistic record linkage methodology for similar purposes in epidemiologic and health services research.

Acknowledgements

This work was supported by a grant from the Maine Medical Center Research Trust Fund. The authors are indebted to numerous professional colleagues involved in data collection at Maine hospitals and state agencies.

References

[1]. D. E. Clark. Development of a statewide trauma registry using multiple linked sources of data. Proc Annu Symp Comput Appl Med Care 1993; 17: 654-658

[2]. M. A. Jaro. Probabilistic linkage of large public health data files. Statistics Med 1995; 14:491-498

[3]. H. B. Newcombe. Handbook of Record Linkage. Oxford: Oxford University Press, 1988.

[4]. Committee on Trauma Research, Commission on Life Sciences, National Research Council. Injury in America: A continuing public health problem. Washington DC: National Academy Press, 1985

[5]. S. P. Baker, R. A. Whitfield, B. O'Neill. Geographic variations in mortality from motor vehicle crashes. New Engl J Med 1987; 316: 1384-1387

[6] T. Kircher, J. Nelson, H. Burdo. The autopsy as a measure of accuracy of the death certificate. New Engl J Med 1985; 313:1263-1269

[7]. L. A. Moyer, C. A. Boyle, D. A. Pollock. Validity of death certificates for injury-related causes of death. Am J Epidemiol 1989; 130:1024

[8]. E. J. MacKenzie, S. Shapiro, J. N. Eastham Jr. Rating AIS severity using emergency department sheets vs. inpatient charts. J Trauma 1985; 25:984-988

[9]. E. J. MacKenzie, D. M. Steinwachs, A. I. Ramzy. Evaluating performance of statewide regional systems of trauma care. J Trauma 1990; 30:681-688

[10]. G. S. Smith, J. A. Langlois, J. V. Buechner. Methodologic issues in using hospital discharge data to determine the incidence of hospitalized injuries. Amer J Epidemiol 1991; 134:1146-1158

[11]. J. C. Young, D. P. Macioce, W. W. Young. Identifying injuries and trauma severity in large databases. J Trauma 1990; 30:1220-1230

[12]. D. C. Grossman, L. G. Hart, F. P. Rivara, R. V. Maier, R. Rosenblatt. From roadside to bedside: The regionalization of trauma care in a remote rural county. J Trauma 1995; 38:14-21

[13]. A. M. Ferrante, D. L. Rosman, M. W. Knuiman. The construction of a road injury database. Accid Anal Prevent 1993; 25:659-665