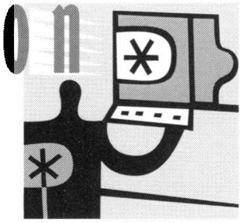
Information  Technology

PUBLIC HEALTH APPLICATIONS

An Algorithm to Match Registries with Minimal Disclosure of Individual Identities

GIRISH S. SUBRAMANYAN
DEBORAH S. YOKOE, MD MPH
SHARON SHARNPRAPAI, MS
YUREN TANG, MD MPH
RICHARD PLATT, MD MS

Identifying individuals common to datasets belonging to different organizations is necessary for many epidemiological studies. Researchers commonly use automated methods to compare large datasets. However, both automated and non-automated methods of comparison face limitations when organizations are required either to avoid disclosing the identity of any of the individuals in their datasets or to limit the amount of information they disclose. We faced this circumstance during a study of methods of surveillance for tuberculosis (TB) that required matching the membership roster of Harvard Pilgrim Health Care, the largest health maintenance organization (HMO) in New England, to the Massachusetts Department of Public Health TB registry.[1]

To meet these needs, we developed a computerized matching algorithm using partial identifiers to locate a small number of potential matches. This algorithm reduced the array of potential matches by several orders of magnitude, allowing the HMO to provide complete demographic identifiers in a highly controlled manner for an extremely small subset of its members who were potential matches. The computer algorithm we developed was implemented with commercially available database software. We report here our matching algorithm with a brief discussion of its potential uses and limitations.

Developing the Algorithm

To establish the number of HMO members with active TB who were known to the public health system but not known to the HMO-based investigators, we needed to match the entire HMO membership to the TB registry. To this end, we decided to use the first two letters of an individual's last name, the first two letters of the first name, the month of birth, the year of birth, and sex to identify potential matches. For the first author, the entry would have been coded as follows: SUGI0173M. We chose this combination because the number of individuals with each combination was small but these variables would not conclusively identify an individual.

After excluding 68 HMO members with TB already known to HMO investigators, we converted the HMO membership registry into a dBase file with entries coded as described above. Each two-character variable and the sex variable comprised separate datafields; we removed all redundant combinations. We converted the public health TB registry into the same format, including one additional datafield for an asterisk (*) used in the tagging process. The 1,421,105-member HMO membership dataset contained 1,206,406 unique combinations of identifiers, and the 1715-member TB registry had 1715 unique combinations.

The two files were sorted similarly and then compared using DOS-based FoxPro[2] Version 2.0 (see shaded box for FoxPro code). When an exact match was identified by the program, the coded entry in the TB registry dataset was moved to the HMO dataset, in effect tagging the matching coded HMO entry with an asterisk (*). Eighty-seven combinations were common to both datasets, representing 87 unique individuals in the TB registry and 124 unique HMO members. This computerized matching process resulted in a more than 10,000-fold reduction in the number of HMO members who could possibly be in the TB registry. Computer processing time for this match was approximately 20 minutes.

HMO investigators compiled a list of the first and last names, date of birth, and sex of the 124 HMO members. One HMO investigator read aloud only this limited information for the 124 HMO members to a Department of Public Health investigator, who compared the information to listings in the TB registry. The Department of Public Health investigator provided only a count of the number of matches to the HMO investigator, and in no case was the identity of any individual in the TB registry disclosed to the HMO. Nor was any clinical information regarding the HMO patients exchanged, and there was no exchange of any written information containing the names or dates of birth of any individuals.

Performance of the Algorithm

Manual matching of the 124 HMO members using the more complete demographic identifiers resulted in 29 matches to the public health TB registry, for a positive predictive value of 23%.

The sensitivity of the matching algorithm was estimated by its ability to identify individuals previously reported to have TB. The algorithm correctly identified 67 of the 68 HMO members previously known to the HMO investigators to be in the TB registry, for an estimated sensitivity of 99%. In the one case that it missed, the month of birth was discrepant in the two datasets.

FoxPro Matching Algorithm Code

PROGRAM CODE	PROGRAM DOCUMENTATION
close all	
set stat on	
sele a	
use tb	open the TB registry dataset
index on last2+first2+yr+mon+sex to tbindex	index on a string combined by the first 2 letters of the last name, first 2 letters of first name, year of birth, month of birth and sex
use hmo	open the HMO membership dataset
set relation to last2+first2+yr+mon+sex into a	perform exact character string matching in the 2 datasets
do while not eof()	set a loop to go to the end of the HMO dataset
replace match with a->match	if matched, move the entry from the TB registry dataset (which includes an "*") to the HMO dataset
skip	go to the next record
enddo	

There was no exchange of any written information containing the names or dates of birth of any individuals.

In the larger realm of medical record linkage strategies, matching algorithms fall into two categories: deterministic or probabilistic. Deterministic matching is an all-or-none matching scheme in which a computer-generated decision is made as to whether a pair of records pertains to the same person.[3] In probabilistic matching, on the other hand, a computed calculation that two records pertain to the same individual is compared to a threshold (that can be varied according to circumstances) to either accept or reject them as a true match.[3] While probabilistic matching is a powerful tool in information-poor settings, it requires considerable experience, specialized or flexible software applications, weights for each information variable, extensive programming, and the use of revealing or highly identifying patient variables.[4]

Moreover, the greater amount of work involved with probabilistic matching entails higher costs.

This study suggests that a simple deterministic matching scheme using partial identifiers is appropriate to identify individuals common to two lists when patient name, date of birth, and sex are available in both datasets and the number of individuals common to both datasets is small. The method we describe here is simple and can be implemented in many personal computer database programs.

> The method we describe here is simple and can be implemented in many personal computer data base programs.

Other methods of name matching also afford patient anonymity to varying degrees. One such popular method involves the phonetic reduction of names in a well-established and easy-to-use system called Soundex.[5,6] In a recent study, Balogun et al. used Soundex-coded last names in conjunction with patient sex and year of birth to match a TB registry with an AIDS registry to evaluate the underreporting of TB in patients with AIDS in London.[7] Using the Soundex method of coding names in conjunction with other identifiers does not guarantee same-person matches; matching by this method would have to be verified by manual matching.[8,9] Several studies have documented relatively poor performance of phonetic name reduction schemes in name matching.[10,11]

Potential limitations to our matching algorithm include errors in coding of any of the variables in either dataset. Such errors would decrease the sensitivity of the algorithm. These errors will be less common, however, than errors that would arise from deterministic matching of full names and full dates of birth. Matching individuals, especially women, who use different names at different times in their lives, is similarly problematic and would be a greater problem when the datasets to be compared are generated at widely separated times. Another important caveat is the unknown variation in performance that would result from different distributions of names in a population. For example, in populations with limited variations in names, the positive predictive value of this matching algorithm would be lower.

This computer matching algorithm can be used in many settings as a first step in identifying individuals known to different organizations in order to limit the disclosure of full identifying information for the vast majority.

This study was approved by the institutional review board of Harvard Pilgrim Health Care.

Support for the study was provided by the Centers for Disease Control and Prevention (Contract 200-95-0957-010), and the Harvard Pilgrim Health Care Foundation.

References

1. Yokoe DS, Subramanyan GS, Nardell E, Sharnprapai S, McCray E, Platt R. Tuberculosis screening in HMOs using automated data [abstract]. In: 38th ICAAC abstracts: proceedings of the 38th Interscience Conference on Antimicrobial Agents and Chemotherapy; 1998 Sep 24–27; San Diego, California. Washington: American Society for Microbiology; 1998. p. 553.
2. FoxPro. Version 2.0. New York: Fox Holdings; 1989.
3. Gill L, Goldare M, Simmons H, Bettley G, Griffith M. Computerised linking of medical records: methodological guidelines. J Epidemiol Community Health 1993;47:316-19.
4. Roos LL, Wajda A. Record linkage strategies. Methods Inf Med 1991;30:117-23.
5. Russel RC, inventor. Index. US patent 1,435,663. 1922 Nov 14.
6. Fenna D. Phonetic reduction of names. Comput Programs Biomed 1984;19:31-6.
7. Balogun MA, Wall PG, Noone A. Undernotification of tuberculosis in patients with AIDS. Int J STD AIDS 1996;7:58-60.
8. Davis KB, Fischer L, Gillespie MJ, Pettinger M. A test of the National Death Index using the Coronary Artery Surgery Study (CASS). Control Clin Trials 1985;6:179-91.
9. Goehring R. Identification of patients in medical data bases—Soundex codes versus match code. Med Inf 1985;10:27-34.
10. Zobel J, Dart P. Finding approximate matches in large lexicons. Software—Practice & Experience 1995;25:331-45.
11. Sideli RV, Friedman C. Validating patient names in an integrated clinical information system. In: Assessing the value of medical information: proceedings of the Annual Symposium on Computer Applications in Medical Care; 1991 Nov 17–20; Washington DC. New York: McGraw Hill; 1991. p. 588-592. ■

Mr. Subramanyan, Dr. Yokoe, and Dr. Platt are with the Channing Laboratory and Department of Medicine, Brigham and Women's Hospital, Boston. Mr. Subramanyan is a Research Fellow. Dr. Yokoe is the Associate Hospital Epidemiologist and an Instructor in Medicine, Harvard Medical School. Dr. Platt is the Hospital Epidemiologist; he is also the Associate Chair of the Department of Ambulatory Care and Prevention, Harvard Medical School, and Director of Research, Harvard Pilgrim Health Care. Ms. Sharnprapai and Dr. Tang are with the Massachusetts Department of Public Health; Ms. Sharnprapai is Research Coordinator, Division of TB Prevention and Control, and Dr. Tang is an Epidemiologist.

Address correspondence to: Dr. Yokoe, 181 Longwood Ave., Boston MA 02115; tel. 617-525-2689; fax 617-731-1541; e-mail <deborah.yokoe@channing.harvard.edu>.