Development of a Computer Linkage System for a Blood Recipient Notification Program in Nova Scotia

Susie ElSaadany, MMath-Stat, PGD-Math[1]
Maura Ricketts, MD, FRCP[2]
Antonio Giulivi, MD, FRCP[1]

ABSTRACT

Objectives: To assess the potential uses of computer-assisted record linkage in the surveillance of infectious diseases, using the Nova Scotia blood recipient notification program as the example.

Methods: We developed a computer-assisted, multiple-pass, probabilistic record linkage to link records for blood recipients identified by the Nova Scotia notification program (Nova Scotia Phase I Blood Bank File information) with corresponding Nova Scotia Health Card Registration File records to obtain current mailing addresses to contact potentially living recipients. We used variables available from both files (e.g., name, date of birth, gender, and health care registration number) to link records, after eliminating duplicates/deceased cases.

Results: Among 23,925 eligible records in the Nova Scotia Phase I Blood Bank File (1984-1990), there were 1,818 (7.8%) duplications and 8,675 deceased cases, leaving 13,432 cases for linkage. 8,713 (65%) cases were successfully linked to the 1998 Health Card Registration Data File for current mailing addresses.

Interpretation: Multiple-pass linkage seems acceptable for maximizing detection of correctly matched records for lookback projects. To overcome quality/lack of information obstacles, future lookback linkages should explore the use of supplementary data files (tax files, voter lists, license files, other provincial databases) to obtain most current addresses.

La traduction du résumé se trouve à la fin de l'article.

1. Health Care Acquired Infections Division, Health Canada, Ottawa, ON
2. World Health Organization, Geneva, Switzerland
Correspondence and reprint requests: Susie El Saadany, Chief, Statistics and Risk Assessment Section, Health Care Acquired Infections Division, Health Canada, Ottawa, ON K1A 0K9, Tel: 613-952-6925, Fax: (613) 952-6668, E-mail: susie_elsaadany@hc-sc.gc.ca
Acknowledgements: The authors thank the Nova Scotia Department of Health; Natalie Hanna, Research Assistant; Statistics and Risk Assessment Section, Health Care Acquired Infections Division, Health Canada; and two anonymous reviewers for their helpful comments on an earlier version of the manuscript.

Computer-assisted record linkage has gained popularity because it makes it possible to study a relationship between two variables sited in different data-set locations, at minimum cost. During the past several decades, computer systems for automatic record linkage have been developed and have been successfully applied in chronic disease surveillance and epidemiologic research.[1-8] However, the development and application of computer record linkage systems for infectious disease surveillance and research have been sparse.

In response to the recommendations of the Commission of Blood Inquiry's February 1995 Interim Report,[9] the Nova Scotia Department of Health initiated a general look-back campaign to contact recipients of blood products regarding their potential exposure to hepatitis C transmission through the blood supply prior to June 1990. The Nova Scotia program was designed to communicate with recipients by sending letters to all Nova Scotians who received a blood transfusion or blood product in Nova Scotia hospitals between January 1984 and June 1990. The Nova Scotia program was completed in three phases, and this study covered the first phase only. The purpose of the address field in the blood bank file was for quick geographical reference and, therefore, was unable to provide a current mailing address for recipients. In addition, as some blood bank records catalogued transfusions occurring over 10 years ago, some patient addresses were outdated. On the other hand, the provincial health insurance registry maintained updated information on all Nova Scotians, including current mailing address. The two files contain some common variables, such as name, date of birth, gender, and health care registration number. It is thus possible to obtain the current mailing address for the potentially living recipients through a record linkage.

In this paper, we describe the development of a computer record linkage system to link records located in the Nova Scotia notification program (Nova Scotia Phase I Blood Bank File information) with corresponding Nova Scotia Health Card Registration File records to obtain current mailing addresses to contact potentially living recipients. The purpose of this paper is to illustrate the potential uses of

computer-assisted record linkage in the surveillance of infectious diseases, using the Nova Scotia blood recipient notification program as the example, and discusses areas of improvement for future uses.

METHODS AND MATERIALS

Data sources

Nova Scotia Phase I Blood Bank File (1984 to 1990)

A file of 23,925 records of blood and blood product recipients from the Nova Scotia Phase I Blood Bank Files, for the period of 1984 to 1990, with patient identifiers, was provided by the Nova Scotia Department of Health (NSDOH) to the Health Care Acquired Infections Division (previously the Blood Borne Pathogens Division) of the Centre for Infectious Disease Prevention and Control (previously the Laboratory Centre for Disease Control) in January of 1998. The available variables in this file included patient number, surname, given names, date of birth, gender, residence information, new and old provincial health card numbers, and earliest known date of blood transfusion.

Nova Scotia Health Card Registration Data File (1998)

The Health Card Registration Data File (HCRDF) contains information for eligible individuals who are covered by the Nova Scotia provincial health insurance plan, and is maintained by the NSDOH. The HCRDF used for the current study consists of 1,001,471 records containing new and old health card number, surname, given name, middle name, date of birth, gender, mailing address, and health card number of the head of household, and have also been made available to the BBPD by the NSDOH.

Nova Scotia Death Registration Data File (1984 to 1996)

The Death Registration Data File (DRDF) contains death records for individuals who die in Nova Scotia and are maintained by the provincial Vital Statistics. The DRDF was obtained to first remove the records of deceased persons from the blood bank file. The mortality file consists of 97,567 records describing partial or complete death certificate information for all causes of death in Nova Scotia from 1984 to 1996. The available variables in this file include surname, given names, date of birth, gender, death registration number, residence at death, social insurance number (for some years), age at death, an underlying cause of death code, and up to six additional contributing cause(s)-of-death codes. The surname, middle initial, and given names variables required further standardization to separate the components of the name into individual fields, as these fields were provided in a comma-delimited text format by the Vital Statistics office.

File preparation

Standardization of Files

We used the AutoStan program[2,3,10] to standardize the names to create fixed-length fields for each component of the name (i.e., surname, first given name and second given name). The New York State Identification and Intelligence System (NYSIIS) phonetic code[11] was generated to encode the surname values to compare records during the linkage process that have similar sounding surnames due to misspelling or key entry errors. In using the NYSIIS system, it should be recognized that some coding systems are better for some surnames than others. For example, a coding system relating MacDonald to McDonald may not relate Slavotsky to Zlavotski and so not be as useful in a community with a large number of families originated in Eastern Europe. The standardized files were then assessed for completeness of information on the same core data elements (sex, DOB, surname, middle, first names) in all files using SAS software.

Detecting and Eliminating Duplications in the Nova Scotia Phase I Blood Bank File

The Nova Scotia Phase I Blood Bank File contains duplicate records, because it is possible for one patient to have had several blood transfusions. An internal record linkage for the blood bank file was conducted to detect and eliminate duplicate registrations, using the AutoMatch record linkage system.[2,3,10] Linkage parameters were designed and coded in a format acceptable to AutoMatch. A five-pass linkage was undertaken to maximize the search for duplicate cases. The selected blocking variables (i.e., data fields which limit the number of comparisons by comparing only records agreeing exactly on a given value of a blocking variable) were as follows: i) surname initial, first given name initial, birth year and sex; ii) birth date and sex; iii) (NYSIIS) phonetic code, which was generated to encode the surname values to compare records during the linkage process that have similar sounding surnames due to misspelling or key entry errors; iv) new health card number and sex; and v) old health card number and sex. A visual inspection of all potential duplicate registrations was conducted to validate the results. Where matches are not perfect they are considered possible matches. Rules are divided to distinguish "probable" matches from "possible" matches and "improbable" matches. The rules are termed "The Grey Area Processing Decision Rules" and are located in Appendix 1.

Detecting and Eliminating Deceased Cases in the Nova Scotia Phase I Blood Bank File

The Blood Bank File contains records for deceased cases. The Nova Scotia Notification Program protocol did not require a notification letter for the families of deceased persons. A data linkage between the Blood Bank File and the Death Registration Data File was thus performed in order to identify deceased cases from the blood bank file. A three-pass linkage was undertaken using a "One to One" procedure in AutoMatch's record linkage system, which allows only one record to match to only one other record in the reference file, to maximize the search for matched cases in the mortality file. The selected blocking variables were as follows: i) surname initial, given name initial, year of birth, and sex; ii) date of birth and sex; iii) NYSIIS code and sex.

Record linkage

After eliminating duplications and deceased cases, the records in the Nova Scotia Phase I Blood Bank File were linked to corresponding records in HCRDF. A four-pass linkage was undertaken using a "Many-to-One matching" procedure in AutoMatch's record linkage system to maximize the search for matches. This pro-

TABLE I

Completeness (%) of Core Linkage Data Elements Common to All Three Files

Variables	Blood Bank File	Health Card Registration File	Death Registration File
Surname	100	99.9	100
Given Name 1	100	99.9	100
Given Name 2	78	62.4	42.3
Gender	100	100	100
Male	51.7	51.1	54.1
Female	47.8	48.9	45.8
Unknown	0.6	–	0.1
Birth Year	100	100	99.3
Birth Month	100	100	99.2
Birth Day	100	100	99.2
Death Year	n/a	n/a	100
Death Month	n/a	n/a	100
Death Day	n/a	n/a	100
Address	97	100	
Address2	n/a	0.13	3.6

TABLE II

Summary of Record Linkage Process for Nova Scotia Blood Recipient Notification Program

Steps	Number of cases
Records in original Blood Bank File	23,925
After eliminating duplications in Blood Bank File	22,107
After eliminating deceased cases	13,432
Linked with Health Card Registration File	8713

cedure allowed multiple records in the reference file HCRDF to match with the same record in the Nova Scotia Phase I Blood Bank File. The selected blocking variables were as follows: i) surname initial, first given name initial, year of birth, and sex; ii) date of birth and sex; iii) New and old provincial health card number; iv) NYSIIS code and sex. The computer system compares common fields in the two databases, assigns weights to the resulting linked records and calculates a total weight. A weight < -90 would be automatically refused and a weight > +300 would be automatically accepted as a match. For linked records with a total weight ranging from -90 to +300, manual resolution is carried out to confirm the linkage. The matched cases were flagged and kept in the same Nova Scotia Hospital Blood Bank File. The parameters and rules applied in decision-making for uncertain linkage can be found in Appendix 1.

RESULTS

Core data elements (e.g., surname, gender, dates of birth and death) critical to recording the linkage for all three files were quite complete, whereas second given name and second address fields were incomplete (Table I).

Among the 23,925 records of blood and blood product recipients in the Nova Scotia Phase I Blood Bank File from 1984 to 1990, there were 1,818 (7.8%) duplications and 8,675 deceased cases, leaving 13,432 living cases for record linkage.

For the 13,432 eligible alive cases in the Nova Scotia Phase I Blood Bank File, 8,713 (65%) cases were successfully linked to the 1998 Nova Scotia Health Card Registration Data File, and thus obtained the current address for mailing. The linkage process is summarized in Table II.

DISCUSSION

The Nova Scotia Phase I Blood Bank File (1984 to 1990), Nova Scotia Health Card Registration Data File (1998), and Nova Scotia Death Registration Data File (1984 to 1996) all have a high level of data completeness for the major linkage variables. The multiple-pass linkage developed by the current study seems an acceptable approach for maximizing the detection of correctly matched records and can be applied in future lookback projects. Unfortunately, we were unable to find addresses for approximately 4,719 for the 13,432 (35%) eligible blood recipients using our methodology. Using similar methodology to link birth and infant death records, Statistics Canada has achieved a successful record linkage rate of >95%.[8] It is very challenging to link current records with long historical records. Potential reasons for unsuccessful linkage are: changes in the province's health insurance number (from an old system to a new system) over time, lack of availability of death information for cases who died out of province, lack of address information of cases who moved out of province, change of surname (e.g., at marriage), change of provincial health card number because of marriage or re-marriage or because of change in dependency status (i.e., leaving home), lack of supplementary records, and data entry errors. Contrary to general knowledge, linkage among government databases, even within the same department, is often non-existent. The fact that we cannot follow people when they marry or change the name appearing on their health care card is a serious problem. Strategies to improve internal data co-ordination so that it is possible to track persons when their health is potentially endangered should be developed.

To overcome the obstacles and to improve the quality of the probabilistic linkage for historical records, we suggest that future lookback linkage projects should explore the uses of available supplementary data files. Such supplementary files could include tax files, voter lists, license files or other provincial databases that may have the most current addresses. For example, a higher linkage rate for blood recipients has been achieved in British Columbia using supplementary databases.[12] It may also be helpful for the Canadian Blood Service agency to collect more personal information such as social insurance number and provincial health care number from the donors and recipients. Clarification is needed, however, regarding the possibility of using such expanded personal files and of the collection of additional personal information, given the current privacy laws. Requirements for this type of information to be included on all medical records would allow tracking for serious circumstances, such as those described in this paper.

REFERENCES

1. Newcombe HB, Kennedy JM, Axford SJ. Automatic linkage of vital records. *Science* 1959;130:954-59.
2. Howe GR, Lindsay J. A generalized Iterative Record Linkage Computer System for use in medical follow-up studies. *Computers and Biomedical Research* 1981;14:327-40.
3. Smith ME, Silins J. Generalized Iterative Record Linkage System. Proceedings of the American Statistical Association, 1981;128-37.
4. Newcombe HB. *Handbook of Record Linkage: Methods for Health and Statistical Studies,*

Administration, and Business. Oxford, England: Oxford University Press, 1988.
5. Kusiak RA, Springer J, Ritchie AC, Muller J. Carcinoma of the lung in Ontario gold miners: Possible etiological factors. *Br J Industr Med* 1991;48:808-17.
6. Goldberg MS, Carpenter M, Theriault G, Fair ME. The accuracy of ascertaining vital status in a historical cohort study of synthetic textile workers using computerized record linkage to the Canadian Mortality Data Base. *Can J Public Health*1993;84:201-4.
7. Howe GR. Use of computerized record linkage in cohort studies. *Epidemiol Rev* 1998;20:112-21.
8. Fair M, Cyr M, Allen AC, Wen SW, Guyon G, MacDonald RC and the Fetal-Infant Mortality Study Group of the Canadian Perinatal Surveillance System. An assessment of the validity of a computer system for probabilistic record linkage of birth and infant death records in Canada. *Chron Dis Can* 2000;21:8-13.
9. Interim Report. Commission of Inquiry on Blood System in Canada. Ottawa: Commission of Inquiry on Blood System in Canada, 1995.
10. MatchWare Technologies, Inc. Burtonsville, Maryland. E-mail: info@matchware.com.
11. Lynch BT, Arends WL. Selection of surname coding procedure for the SRS record linkage system. Washington, DC: U.S. Department of Agriculture, February 1977.
12. British Columbia Blood Recipients Record Linkage System. Report on the Blood Recipient Notification Project. British Columbia Ministry of Health and Ministry Responsible for Seniors. May 1998;5-7.

Received: June 5, 2001
Accepted: May 21, 2002

Appendix 1

Grey Area Processing Decision Rules and Match Criteria

Records from each database were considered to be a match, upon clerical review, if one of the following criteria were present:
- Surname, given name, middle name, and date of birth match
- Surname spelling different, but Given name, middle name, and date of birth match
- Given name or middle name spelling different or transposed but Surname and date of birth match
- Surname, given name, middle name, and year of birth match but month and day of birth different or transposed
- Surname, given name, middle name, and month and day of birth match but year of birth off by 1 to 5 year(s)
- Surname, given name, middle name, year of birth match but month or day of birth or both are off by 1 to 3 month(s) or day(s)
- Surname, given name, middle name, year of birth match, month or day of birth or both are different but Nova Scotia health card number is the same
- Nova Scotia health card number, surname, given name, and middle name match, and date of birth is different by less than 6 years, or date of birth includes year 1850 (data management code)
- Given name, middle name, and date of birth match, Surname different, but Nova Scotia health card number is the same

RÉSUMÉ

Objectifs : Évaluer les possibilités d'utilisation de la liaison de dossiers assistée par ordinateur pour la surveillance des maladies infectieuses, en utilisant comme exemple le programme de notification des personnes ayant reçu du sang mis en oeuvre par la Nouvelle-Écosse.

Méthodes : Le système informatique élaboré pour le maillage des dossiers probabilistes de personnes ayant reçu du sang cernées par le programme de notification de la Nouvelle-Écosse (N.-É.) a utilisé les renseignements du fichier de la banque de sang de la phase I de la N.-É. et des dossiers correspondants du fichier d'enregistrement des cartes santé de la N.-É. afin d'obtenir les adresses postales actuelles de ces personnes qui vivent peut-être encore. On s'est servi des variables disponibles dans les deux fichiers pour mailler les dossiers, après avoir éliminé les cas en double/ de décès.

Résultats : Parmi les 23 925 dossiers admissibles dans le fichier de la banque de sang de la phase I (1984-1990), on a compté 1 818 (7,8 %) de doubles et 8 675 cas de décès, ce qui laisse 13 432 cas de maillage. 8 713 (65 %) cas ont été maillés avec succès au fichier des données d'inscription de la carte santé de 1998, obtenant ainsi les adresses postales actuelles.

Interprétation : Le maillage multipasse semble acceptable pour maximiser la détection des dossiers assortis correctement pour les projets futurs de recherche de dons antérieurs. Afin d'éliminer les obstacles relatifs à la qualité ou au manque d'information, les maillages futurs relatifs à la recherche de dons antérieurs pourraient explorer l'utilisation de fichiers de données supplémentaires (fichiers fiscaux, listes électorales, fichiers de permis, autres bases de données provinciales) afin d'obtenir l'adresse actuelle la plus récente.