Epidemiol. Infect. (2003), **131**, 669–674. © 2003 Cambridge University Press
DOI: 10.1017/S0950268803008914 Printed in the United Kingdom

Estimating the sensitivity and specificity of matching name-based with non-name-based case registries

P. ETKIND[1], Y. TANG[1]*, M. WHELAN[1], S. RATELLE[1], J. MURPHY[2], S. SHARNPRAPAI[3] AND A. DEMARIA[4]

[1] *Division of STD Prevention, Bureau of Communicable Disease Control, Massachusetts Department of Public Health,*
[2] *HIV/AIDS Surveillance Program, Bureau of Communicable Disease Control, Massachusetts Department of Public Health,*
[3] *Division of Tuberculosis Prevention and Control, Bureau of Communicable Disease Control, Massachusetts Department of Public Health,*
[4] *Bureau of Communicable Disease Control, Massachusetts Department of Public Health, 305 South Street, Jamaica Plain, MA 02130, USA*

(*Accepted 17 April 2003*)

SUMMARY

Because non-name-based case registries have recently been used for reporting human immunodeficiency virus infection, this study attempted to define the sensitivity, specificity and accuracy of case registry matches using non-name-based registries. The AIDS, sexually transmitted disease (STD), and tuberculosis (TB) case registries were matched using all available information to establish the standard. The registries were then matched again using five increasingly less specific criteria to compare sensitivity, specificity and accuracy. The registries were then also transformed into non-name-based codes as if they were the HIV registry and matched again. With name-based registries, sensitivities increased as the matching criteria became less exacting, while the accuracy declined slightly. Specificities remained close to 100% due to the relatively small number of matched cases. Results from matches of non-name-based registry matches were similar to those of the name-based registry matches. Non-name reporting can be used for data matching with acceptable accuracy.

INTRODUCTION

Matching disease registries can be a powerful tool for enhanced surveillance and programme evaluation [1–4]. The extent and kind of concordance can suggest unrecognized risk factors, as well as possible points of intervention. Identifiers common to registries, in combination with other information, are used in matching. Identifiers include name (or some variation of initials and/or name), address, birth date, gender, social security number or patient identification number.

* Author for correspondence: Bureau of Communicable Disease Control, 305 South Street, Jamaica Plain, MA 02130, USA.

Although name-based case registries have been the usual practice in public health surveillance, non-name-based surveillance registries have recently been put into place for reporting of human immunodeficiency virus (HIV) infection in response to concerns regarding confidentiality and potential misuse of such information. Instead of names, cases are identified by soundex methodologies, code systems external to the names, or by a code generated by features of the case's name and perhaps other identifying information. The utility of non-name based registries for matching purposes needs to be assessed. The purpose of this study was to define the sensitivity, specificity, and accuracy of matches using non-name-based case registries.

Table 1. *Matching criteria for establishing the sensitivity, specificity and positive predictive value of non-name-based registry matching, Massachusetts, 2001*

Name-based registry match criteria
1. Full last name + first name + gender + DOB
2. First four letters of last name + first two letters of first name + gender + DOB
3. Soundex of last name + first two letters of first name + gender + DOB
4. First four letters of last name + first two letters of first name + gender + DOB (allowing for a one number discrepancy in year, month or day of DOB)
5. First letter of last name + first letter of first name + gender + DOB
Non-name registry match criteria
6. Number of letters in last name + first two letters of first name + gender + DOB
7. Number of letters of last name (+/−1 letter) + first two letters of first name + gender + DOB

METHODS

AIDS has been reportable in Massachusetts since 1983. People diagnosed with AIDS are reported by full name, date of birth, their residence at time of diagnosis, and social security number. More than 16000 cases were entered into that registry between 1984 and 2001. HIV infection became reportable in 1999. Cases are reported by an identifier which is a combination of several variables: the first two letters of the first name, the number of letters in the last name, gender, date of birth, zip code of residence, and the last four digits of the social security number (a randomly generated number). Since initiation of HIV infection reporting, AIDS case reports have included the HIV coded identifier, as well as name, address, date of birth, social security number and gender.

We used the AIDS registry as a surrogate for the HIV registry in both name-based and non-name-based methods. Name-based matches of AIDS cases and reported cases of sexually transmitted disease (STD) and tuberculosis (TB) were performed and matched case lists were created using all of the information contained in the five matching criteria listed in Table 1. In addition to the variables listed in the table, we looked for matches between the registries in four additional ways. These were (1) inverting the first and last names, as often occurs in people from non-English speaking parts of the world; (2) matching women by first name and date of birth, in order to identify women who might have been reported once by a maiden name as well as by a married name; (3) matching only be social security number allowing a difference in one number, followed by a look at the name, date of birth, address and phone number in order to verify a match; and (4) matching by home address alone. The result was considered to be a complete set of matches and became our standard by which sensitivities, specificities and positive predictive values were calculated. The registries were then matched again by the five different criteria marked by decreasing specificity (Criteria 1–5, Table 1). Some of these criteria were nested within each other. Criteria 1 and 2 were nested within Criterion 4, since these were based on spelling of the name, gender and date of birth. Criteria 3 and 5 were more independent as these were less dependent on spellings. The cumulative totals of the matches were used to establish standards for sensitivity, specificity, and positive predictive values.

Identifiers in the AIDS name-based registry were transformed into the non-name-based codes as if they were the HIV case registry. Two independent matches of the two registries were then done, each with a different set of criteria (Criteria 6 and 7, Table 1). The sensitivity, specificity and positive predictive value of each criterion, as well as the cumulative values of each of the name-based matches according to the different sets of criteria, were calculated using the named match results that used all available information.

Only personnel authorized for access to identifiers in each data set had access to those data sets. Only one person (YT) conducted the matches. This was done in a locked office on a stand-alone computer that was not connected to the local or wide area network. Output was stripped of all identifiers. No personal identifiers were shared between programs. Aggregate results were used in analyses.

The STD Case Registry used in this investigation comprised 138047 individual patients reported between 1986–2001. The AIDS Case Registry included the 16370 individual people reported between 1983 through the end of April 2001. The Tuberculosis Case Registry used in this investigation contained 2392 individual people reported between 1993–2000.

Table 2. *Estimating sensitivity, specificity, and positive predictive value of STD/AIDS matches by different criteria, Massachusetts, 2001*

	Number of matches	Verified matches	Total verified matches	Sensitivity	Specificity	PPV
Name-based match criteria						
1	1015	115	1015	67·4 (1015/1506)	100·0 (14864/14864)	100·0 (1015/1015)
2	567	221	1236	82·1 (1236/1506)	97·7 (14518/14864)	78·1 (1236/1582)
3	137	54	1290	85·7 (1290/1506)	97·1 (14435/14864)	78·1 (1290/1719)
4	479	148	1438	95·5 (1438/1506)	94·9 (14104/14864)	65·4 (1438/2198)
5	340	19	1457	96·8 (1457/1506)	92·7 (13783/14864)	57·4 (1457/2538)
Total	2538	1457	1457			
Non-name-based criteria						
6	2051	1319	1319	87·6 (1319/1506)	95·1 (14132/14864)	64·3 (1319/2051)
7	1548	1270	1270	84·3 (1270/1506)	98·1 (14586/14864)	82·0 (1270/1548)

Table 3. *Estimating sensitivity, specificity, and positive predictive value of TB/AIDS matches by different criteria, Massachusetts, 2001*

	Number of matches	Verified matches	Total verified matches	Sensitivity	Specificity	PPV
Name-based match criteria						
1	160	160	160	66·1 (160/242)	100·0 (2150/2150)	100·0 (160/160)
2	50	50	210	86·8 (210/242)	100·0 (2150/2150)	100·0 (210/210)
3	13	12	222	91·7 (222/242)	99·9 (2142/2150)	99·6 (222/223)
4	16	13	235	97·1 (235/242)	99·8 (2146/2150)	98·3 (235/239)
5	6	3	238	98·3 (238/242)	99·7 (2142/2150)	97·1 (238/245)
Total	245	238	238			
Non-name-based criteria						
6	223	215	215	88·8 (215/242)	99·6 (2142/2150)	96·4 (215/223)
7	206	205	205	84·7 (205/242)	99·9 (2149/2150)	99·5 (205/206)

RESULTS

There were a total of 1506 verified matches between the AIDS and STD registries when all available information was used. After then conducting the five matches according to the different criteria listed in Table 1, only 1457 case matches were identified. Sensitivities were calculated according to the cumulative number of matches derived from the criteria at each step compared to the total number of true matches (Table 2). Specificities were calculated using the number of unmatched cases (16370 − 1506, or 14864) in the AIDS case registry as the denominator. The comparison utilizing full name, date of birth, and gender (Criterion 1) had a sensitivity of 67·4%, a specificity of 100% and a positive predictive value of 100%. The sensitivities of matching increased from 67·4% to 96·8% as the criteria became less exacting, while the positive predictive values correspondingly declined from 100% to 57·4%. The specificities of each set of criteria remained high due to the relatively small proportion of matching cases, but also declined from 100% to 92·7%. The results using the non-name match criteria were similar to those of Criterion 2 of the name-based matching scheme (Table 1). However, the positive predictive value of the Criterion 7 non-name match exceeded those of all name-based matches except Criterion 1.

In the match between the tuberculosis and AIDS case registries, the AIDS case registry used was the same as noted above. There were 242 verified matches when using all available information for the subsequent name-based TB/AIDS matches (Table 3).

The sensitivities were, as before, calculated by comparing the number of matched cases from those sets of criteria (ultimately totaling 239) to the number of matches established when using all available information (242). The number of unmatched cases

(2392 − 242, or 2150) in the TB case registry was used as the denominator for the specificity calculations. In the comparison utilizing full names (Criterion 1), the sensitivities of matching increased from 66·1% to 98·3% as the matching criteria became less exacting, while the positive predictive values correspondingly declined from 100% to 97·1%. The specificity of each remained close to 100% due to the relatively small number of matching cases but decreased slightly as well. The matches using the non-name criteria (Criteria 6 and 7) had sensitivities between 84–88% and both had very high positive predictive values. The sensitivities of the non-name criteria again compared favourably to that resulting from named-Criterion 2. The positive predictive values were similar throughout.

DISCUSSION

Matching case registries is a tool for evaluating programmatic activities and can help to direct policy decisions and priorities. The use of a non-name-based surveillance registry in a match for such purposes requires developing an estimate of sensitivity, specificity and positive predictive value. Problems encountered in the course of matching case registries are often in trying to evaluate minor differences in similar information that is requested and recorded within the respective registries. Transforming a case registry into non-name-based codes according to a set of defined rules may not allow for expressions of these minor differences that often exist between case registries. Thus, the resulting match may be considered to be somewhat artificial. However, conducting the match in this fashion allows for comparing these two different forms of case registries and to calculate the sensitivities, specificities and positive and negative predictive values. Use of a surrogate data set from a similar population converted into a non-name-based data set and matched with name-based registries expected to have overlap in population allows such estimation and to verify the likelihood that the matches are true or false.

Using the name-based AIDS case registry as a surrogate for the non-name-based HIV database was valid. The individuals reported to the registries were from the same population. Both matches, between AIDS and STD as well as between AIDS and TB, gave similar estimates of sensitivity, specificity and positive value of a match with this surrogate case registry.

Likely differences for non-matches of case reports for the same individuals are changes in the last name (as resulting from name change upon marriage) and reversal of first and last name in one or the other data sets (especially among foreign-born individuals). We tried to deal with these possibilities in conducting our matches according to the various sets of criteria.

One other concern was the seemingly low number of matches between the STD Case Registry (containing 138047 cases) and the AIDS Case Registry (containing 16370 cases). There were 1506, matches, which was 1·1% of the STD cases and 9·2% of the AIDS cases. There have been five matches between these two registries since the early 1990's, and these two percentages have remained consistent throughout the decade. Outside of Boston, the HIV/AIDS epidemic in Massachusetts is primarily related to injection drug use according to the CDC surveillance hierarchy [5]. While it is acknowledged that this hierarchy is somewhat artificial and provides greater weight to the drug use rather than possibly concurrent sexual risk behaviours, it does suggest that it is reasonable to see relatively few matches. One other possibility to explain the apparent low number of matches is under-reporting of STD among people reported with HIV or AIDS. The extent of STD screening supported by the Department of Public Health and the extent of laboratory-based reporting in Massachusetts provides confidence that such under-reporting is minimized. There may be a greater degree of concordance between the STD and HIV case registries, which would be a reflection of the current outbreak of sexually transmitted diseases among men who have sex with men, many of whom are co-infected with HIV [6–9].

Concerns regarding stigma and discrimination, coupled with distrust of possible misuse of a long-term HIV infection case registry, have resulted in debate over HIV infection reporting since the initial availability of HIV antibody tests [10–15]. Evaluations of HIV reporting in states with name-based reporting have indicated that HIV reporting did not appear to be a deterrent to testing among at-risk individuals [16]. A debate has continued with a focus on name-based reporting versus non-name-based reporting [17, 18]. Results of a 3-year evaluation of two states using HIV case surveillance conducted by using non-name-based unique identifiers were not supportive of this approach to surveillance [19], concluding that name-based reporting demonstrated superior performance in terms of completeness of reporting and documentation of risk factors. Although subsequent

re-evaluations of the unique identifier approach were more supportive of its use [20, 21], the Council of State and Territorial Epidemiologists (CSTE) endorsed name-based HIV reporting [22].

The Massachusetts HIV infection surveillance system does not require that providers create a coded unique identifier that is external to the name. Instead, elements of the individual's identifying information are used. These are

- first two letters of the first name;
- number of letters in the last name;
- gender;
- date of birth;
- last four digits of the Social Security number; and,
- zip code of residence.

Completeness of HIV infection reporting is similar to that of AIDS in Massachusetts as well as nationally [23, 24]. The documentation of risk exposures in the Massachusetts HIV Infection Case Registry (through October 2002) is slightly below the minimum standard of 85% for reported HIV cases (4926 of 6201, or 79·4%) and is less than that for reported AIDS cases (15641 of 17626 reported cases, or 88·7%) [25]. In addition, HIV-infection reporting is more timely than name-based AIDS case reporting. For cases diagnosed in 1999, 75% of all HIV cases were reported within 6 months. By 2001, this had increased to 81% reporting within 6 months. In comparison, for cases diagnosed in 1999, 68% of all AIDS cases were reported within 6 months. In 2001, that had increased to 74% reporting within 6 months. The findings reported here extend the demonstrated utility of non-name HIV infection reporting to show that it can be used in case registry matching. Indeed, the sensitivity of the non-name match exceeded that of a match using the full name and date of birth as the matching variables. Non-name reporting can be used for data matching with acceptable accuracy. Additional work will be necessary to evaluate the utility of non-name case registries in meeting other standards of public health surveillance and HIV/STD prevention [26, 27].

ACKNOWLEDGEMENT

This work was made possible by a grant from the federal Centers for Disease Control and Prevention (CDC) as part of the Outcome Assessment through Systems of Integrated Surveillance ('OASIS') Project (Program Announcement 02211).

REFERENCES

1. Johnson RJ, Montano BL, Wallace EM. Using death certificates to estimate the completeness of AIDS case reporting in Ontario in 1985–1987. CMAJ 1989; **141**: 537–540.
2. Rosenman KD, Trimbath L, Stanbury M. Surveillance of occupational lung disease: comparison of hospital discharge data to physician reporting. Am J Public Health 1990; **80**: 1257–1258.
3. Chun HJ, O'Brien RJ, Chonde TM, Graf P, Rieder HL. An epidemiological study of tuberculosis and HIV infection in Tanzania, 1991–1993. AIDS 1996; **10**: 299–309.
4. Moore M, McCray E, Onorato IM. Cross-matching TB and AIDS registries: TB patients with HIV co-infection, United States, 1993–1994. Pub Health Rep 1999; **114**: 269–277.
5. Massachusetts Department of Public Health. HIV/AIDS Surveillance Program.
6. Massachusetts Department of Public Health. Division of STD Prevention. Annual Statistics, 2001.
7. Centers for Disease Control and Prevention. Resurgent bacterial sexually transmitted disease among men who have sex with men – King County, Washington, 1997–1999. MMWR 1999; **48**: 773–777.
8. Centers for Disease Control and Prevention. Gonorrhea among men who have sex with men – selected sexually transmitted disease clinics, 1993–1996. MMWR 1997; **46**: 889–892.
9. Centers for Disease Control and Prevention. Outbreaks of syphilis among men who have sex with men – southern California, 2000. MMWR 2000; **50**: 117–120.
10. Institute of Medicine. Confronting AIDS: directions for public health, health care, and research. Washington, DC: National Academy Press, 1986.
11. Gostin LO, Curran WJ, Clark ME. The case against compulsory case finding in controlling AIDS – testing, screening and reporting. Am J Law Med 1987; **12**: 7–53.
12. Fordyce EJ, Sambula S, Stoneburner R. Mandatory reporting of human immunodeficiency virus by testing would deter blacks and Hispanics from being tested. JAMA 1989; **262**: 349.
13. Kegeles SM, Coates TJ, Lo B, Catania JA. Mandatory reporting of HIV testing would deter men from being tested. JAMA 1989; **261**: 1275–1276.
14. Bayer R. Public health policy and the AIDS epidemic: an end to AIDS exceptionalism? N Engl J Med 1991; **324**: 1500–1504.
15. Ward JW, Fleming PL, Buehler JW. What will be the role of HIV reporting? Am J Pub Health 1994; **84**: 1888–1889.
16. Centers for Disease Control and Prevention. HIV testing among populations at risk for HIV infection – nine states, November 1995 – December 1996. MMWR 1998; **47**: 1086–1091.
17. Colefax GN, Bindman AB. Health benefits and risks of reporting HIV-infected individuals by name. Am J Public Health 1998; **88**: 876–879.

18. Burr C. The AIDS exception: privacy *vs.* public health. The Atlantic Monthly, June 1997: 57–67.
19. Centers for Disease Control and Prevention. Evaluation of HIV case surveillance through the use of non-name unique identifiers – Maryland and Texas, 1994–1996. MMWR 1998; **46**: 1254–1258, 1271.
20. Solomon L, Flynn C, Eldred L, Caldeira E, Wasserman MP, Benjamin G. Evaluation of a statewide non-name-based HIV surveillance system. J Acquir Immune Syndr 1999; **22**: 272–279.
21. Schwarcz S, Hsu L, Chu PL, et al. Evaluation of a non-name-based HIV reporting system in San Francisco. J AIDS 2002; **29**: 504–510.
22. Council of State and Territorial Epidemiologists. CSTE: position statement 1D-4. National HIV surveillance: addition to the National Public Health Surveillance System. Atlanta, GA: USPHS, 1997.
23. Rosenblum L, Buchler JW, Morgan ME, et al. The completeness of AIDS case reporting, 1988: a multi-site collaborative surveillance project. Am J Public Health 1992; **32**: 1495–1499.
24. Massachusetts Department of Public Health. HIV/AIDS Surveillance Program. Data analysis of AIDS case registry 1983–2001 and HIV case registry 1999–2001.
25. Centers for Disease Control and Prevention. HIV/AIDS Surveillance Report. Atlanta, GA: US Department of Health and Human Services, Public Health Service, 1997, vol. 8, no. 1.
26. Klaucke DN. Evaluating public health surveillance. In: Teutsch SM, Churchill RE, eds. Principles and practices of public health surveillance. New York, NY, Oxford University Press, 1994: 158–174.
27. Centers for Disease Control and Prevention. HIV partner counseling and referral services guidance. Atlanta, GA: USPHS, 1998.