The Effect of Case Rate and Coinfection Rate on the Positive Predictive Value of a Registry Data-Matching Algorithm

Qiang Xia, MD, MPH[a]
Sarah L. Braunstein, PhD, MPH[a]
Laura E. Stadelmann, MPH[a]
Preeti Pathela, DrPH[b]
Lucia V. Torian, PhD[a]

ABSTRACT

Objective. Statistical modeling has suggested that the prevalence of false matches in data matching declines as the events become rarer or the number of matches increases. We examined the effect of case rate and coinfection rate in the population on the positive predictive value (PPV) of a matching algorithm for HIV/AIDS and sexually transmitted disease (STD) surveillance registry data.

Methods. We used LinkPlus™, a probabilistic data-matching program, to match HIV/AIDS cases diagnosed in New York City (NYC) from 1981 to March 31, 2012, and reported to the NYC HIV/AIDS surveillance registry against syphilis and chlamydia cases diagnosed in NYC from January 1 to June 30, 2010, and reported to the NYC STD registry. Match results were manually reviewed to determine true matches.

Results. With an agreement/disagreement comparison score cutoff value of 10.0, LinkPlus identified 3,013 matches, of which 1,582 were determined to be true by manual review. PPV varied greatly in subpopulations with different case rates and coinfection rates. PPV was the highest (91.6%) in male syphilis cases, who had a relatively low case rate but a high HIV coinfection rate, and lowest (18.0%) in female chlamydia cases, who had a high case rate but a low HIV coinfection rate. When the cutoff value was increased to 15.0, PPVs in male syphilis and female chlamydia cases increased to 98.3% and 90.5%, respectively.

Conclusions. Case rates and coinfection rates have a significant effect on the PPV of a registry data-matching algorithm: PPV decreases as the case rate increases and coinfection rate decreases. Before conducting registry data matching, program staff should assess the case rate and coinfection rate of the population included in the data matching and select an appropriate matching algorithm.

[a]New York City Department of Health and Mental Hygiene, Bureau of HIV/AIDS Prevention and Control, Long Island City, NY

[b]New York City Department of Health and Mental Hygiene, Bureau of STD Control, Long Island City, NY

Address correspondence to: Qiang Xia, MD, MPH, New York City Department of Health and Mental Hygiene, Bureau of HIV/AIDS Prevention and Control, HIV Epidemiology and Field Services Program, 42-09 28th St., Long Island City, NY 11101; tel. 347-396-7664; fax 347-396-4463; e-mail <qxia@health.nyc.gov>.

©2014 Association of Schools and Programs of Public Health

In 2009, the Centers for Disease Control and Prevention (CDC) National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention released a white paper on program collaboration and service integration (PCSI). PCSI is a mechanism for promoting collaboration across disease surveillance programs working on interrelated health issues, activities, and prevention strategies, with the ultimate goal of facilitating the delivery of comprehensive services.[1] PCSI focus areas are human immunodeficiency virus (HIV)/acquired immunodeficiency syndrome (AIDS), viral hepatitis, other sexually transmitted diseases (STDs), and tuberculosis (TB).

One of the priorities of PCSI is to conduct electronic matching of case surveillance registries. Matching case registries is an efficient and effective way to identify people with coinfection for case management and epidemiologic monitoring; matching can also improve the completeness of registry data by adding demographic, clinical, or behavioral information obtained from other registries on cases of coinfection.[2] A number of health departments have matched their HIV/AIDS registry data with STD and TB registries to address the syndemics of HIV/AIDS and STDs, and HIV/AIDS and TB.[3–7]

All data-matching projects use either a deterministic or probabilistic method, or a combination of the two, to identify matches. Some published analyses have reported the sensitivity, specificity, and positive predictive value (PPV) of their method;[5,8] however, as of this writing, none has addressed the effect of the case rate and coinfection rate in the population on the PPV of a registry data-matching algorithm. One statistical modeling study reported that the prevalence of false matches declined as the events became rarer (i.e., lower case rate) or the number of matches increased (i.e., higher coinfection rate).[9]

In this study, we used HIV/AIDS and STD matching as an example to examine the effect of the case rate and coinfection rate on the PPV of a matching algorithm. HIV/AIDS and STD registry matching has been used to identify HIV/AIDS and STD coinfection cases. In this capacity, it acts like a special test that can detect not only HIV infections among STD patients, but also STD among HIV-infected individuals. HIV prevalence, STD incidence, and HIV/STD coinfection rates vary across regions, but, more importantly, they vary greatly among subgroups in a region or within a population (e.g., by age and sex). CDC estimated that HIV prevalence was 719.5 per 100,000 population among males vs. 230.0 per 100,000 population among females in the United States at the end of 2008.[10] In 2010, the overall rate of reported chlamydia infection among women in the U.S. (610.6 cases per 100,000 females) was approximately two and a half times the rate among men (233.7 cases per 100,000 males), and highest among females aged 15–19 years (3,378.2 cases per 100,000 females) and 20–24 years (3,407.9 per 100,000 females).[11] HIV prevalence among STD patients ranges widely but has been found to be substantial among syphilis-infected patients.[12–15]

METHODS

Datasets

The New York City Department of Health and Mental Hygiene (NYC DOHMH), Division of Disease Control, Bureau of HIV/AIDS Prevention and Control maintains the HIV/AIDS registry for all reported HIV/AIDS cases. The Bureau of STD Control maintains the STD registry for all cases of reportable STDs in NYC. Both the HIV/AIDS and STD registries contain comprehensive demographic and clinical information.

AIDS diagnoses have been reportable in New York State since 1981; the law that expanded AIDS case reporting to include diagnoses of non-AIDS HIV infection took effect on June 1, 2000. This analysis included AIDS cases diagnosed and reported from 1981 to March 31, 2012, and HIV cases reported from 2000 to March 31, 2012.

NYC requires reporting to the NYC DOHMH of all diagnoses of chlamydia, gonorrhea, syphilis, chancroid, granuloma inguinale, neonatal herpes, and lymphogranuloma venereum. To better demonstrate the effect of case rate and coinfection rate on the PPV of a matching algorithm, only syphilis and chlamydia cases were chosen because, of the three most commonly reported STDs (i.e., chlamydia, gonorrhea, and syphilis), syphilis had the lowest case rate but the highest HIV coinfection rate, and chlamydia had the highest case rate but the lowest HIV coinfection rate.[13] Syphilis and chlamydia cases diagnosed from January 1 to June 30, 2010, and reported to the NYC DOHMH were included in the matching. As an HIV-infected individual had to be alive in 2010 to have an STD diagnosis in 2010, HIV/AIDS cases that had a confirmed date of death before January 1, 2010, were removed before data matching to reduce computer processing time and false matches.

Data matching

We used LinkPlus® 2.0,[16] a probabilistic record linkage program developed by CDC for cancer registry database linkage and de-duplication, for this analysis. The Figure shows a list of the variables used for blocking, matching, and manual review. A blocking variable is a variable common to both registry data files that is used to block (or partition) the two files into mutu-

Figure. List of variables used for blocking, matching, and manual review in a New York City HIV/AIDS and STD registry data-matching analysis, 2010

Variable	*Blocking*	*Matching*	*Manual review*
First name (soundex)	√		
Last name (soundex)	√		
Year of birth	√		
First name (full)		√	√
Last name (full)		√	√
Date of birth		√	√
Sex		√	√
Social Security number		√	√
Middle name (full or initial)			√
Race/ethnicity			√

HIV = human immunodeficiency virus

AIDS = acquired immunodeficiency syndrome

STD = sexually transmitted disease

ally exclusive and exhaustive blocks to perform comparisons only on records within each block, thereby reducing manual review burden (i.e., reducing false matches) and computing time. LinkPlus provides a simple blocking mechanism by indexing the variables for blocking and comparing only pairs with identical values on at least one of those variables. In this analysis, first and last name (using soundex code, which is a phonetic algorithm for indexing names by sound) and year of birth were selected as the blocking variables. Five variables—first name, last name, date of birth, sex, and Social Security number—were chosen to be the matching variables in LinkPlus to calculate the agreement/disagreement comparison scores and were also included in the manual review. Two additional variables, middle name (full or initial) and race/ethnicity, were included in the manual review.

We designated the HIV/AIDS registry data as File 1 and the STD registry data as File 2. To compute the *m* probability, which is the probability of agreement for a given matching variable, we used the default direct method. This method uses the data in File 1, the HIV/AIDS registry data, to generate the frequencies of last names and first names and then computes weights for last name and first name based on the frequencies of their values.

When a data match is performed, LinkPlus generates a score for all comparisons of a case in File 1 with a case in File 2. The score is the sum of the agreement and disagreement weights for each matching variable. If a comparison has a score equal to or higher than the user-specified cutoff value, the HIV/AIDS and STD case record will be written to the matching report as a potential match, which then usually requires manual review to assign a status of match or nonmatch. The higher the cutoff value, the fewer false positives but more false negatives the matching algorithm will produce. For this HIV/AIDS and STD registry data match, we first set a relatively low cutoff value of 10.0 and then increased it to 12.0 and 15.0 to examine its impact on PPV.

Manual review

Two reviewers independently reviewed all potential matches with a score of ≥10.0 to decide whether the match was true or false. A third reviewer was called in to resolve all discrepancies through independent review followed by discussion with the two initial reviewers.

A match was determined to be true if it matched on all five matching variables. If it was not a perfect match, the two reviewers made a personal judgment based on the score and other information, such as (*1*) an obvious typographical error in one or more of the matching variables, (*2*) transposed first and last name, (*3*) transposed first and middle name, (*4*) variants of first name, (*5*) the same middle names in the two data files, (*6*) hyphenated last name, and (*7*) transposed birth month and day.

The purpose of the analysis was to examine the PPVs for subgroups with different case rates and coinfection rates. To avoid the bias that a manual reviewer may be more likely to call a matched syphilis case a true match than a chlamydia case because of the known lower overall syphilis case rate and higher HIV/syphilis coinfection rate in the population,[11,13] STD diagnosis information was not included in the dataset for manual review.

PPV and HIV coinfection rate

We calculated PPV as the number of true matches determined by manual review divided by the number of total matches identified by LinkPlus with a specific

cutoff value. Two HIV coinfection rates among STD patients were also calculated. The "true" HIV coinfection rate was calculated as the number of true matches divided by the total number of STD cases assuming no missed matches (i.e., no false negatives). For example, the true HIV coinfection rate among syphilis patients was the number of true HIV-syphilis matches determined by manual review divided by the total number of syphilis cases. The estimated HIV coinfection rate was calculated as the number of total matches divided by the number of STD cases by accepting all matches reported by LinkPlus to be true without manual review. Incongruity between the two HIV coinfection rates serves as an indication of the extent of false matches.

RESULTS

As of March 31, 2012, 221,013 cumulative cases of HIV and AIDS had been reported to the NYC DOHMH; 104,335 cases were removed because of a confirmed date of death earlier than January 1, 2010, leaving 116,678 prevalent HIV/AIDS cases for matching. A total of 2,124 syphilis cases and 31,479 chlamydia cases diagnosed from January 1 to June 30, 2010, had been reported to the Bureau of STD Control; 33,593 of these cases were included in the match after removing one syphilis and nine chlamydia cases with missing information on the patients' sex.

LinkPlus identified 3,013 matches; 1,582 were determined to be true by manual review. The Table shows the number of true and false matches, the PPVs of the matching algorithm, and "true" and estimated HIV coinfection rates among STD patients by cutoff value and subpopulation.

Using a cutoff value of 10.0, the matching algorithm yielded the highest PPV (91.6%) in the male syphilis subpopulation, with 900 of 982 matches being true, and estimated HIV coinfection rate (54.6%) was close to the true rate (50.0%). The PPVs in the male chlamydia and female syphilis subpopulations were 45.1% and 48.3%, respectively. In contrast, using the same cutoff value, 870 female chlamydia cases were matched to an HIV case, of which 157 were determined to be true by manual review, giving the matching algorithm a PPV of only 18.0%. The estimated HIV coinfection rate for female chlamydia cases (4.1%) was more than five times the true coinfection rate (0.7%) (Table).

When the cutoff value was increased to 12.0, the PPV in the female chlamydia, male chlamydia, female syphilis, and male syphilis subpopulations increased to 54.6%, 81.4%, 87.5%, and 97.0%, respectively. When the cutoff value was increased to 15.0, the PPV was

Table. Positive predictive value of an HIV/AIDS and STD surveillance registry data-matching algorithm and estimated HIV coinfection rate among STD patients in New York City, 2010, by cutoff value in LinkPlus® and STD subpopulation

Cutoff value	*Subpopulation*	N	*True matches (a)*	*False matches (b)*	*Total matches (a+b)*	*PPV (percent) a/(a+b)*	*"True" HIV coinfection rate (percent) a/*N	*Estimated HIV coinfection rate (percent) (a+b)/*N
All ages								
10.0	Female chlamydia	21,312	157	713	870	18.0	0.7	4.1
	Male chlamydia	10,158	511	621	1,132	45.1	5.0	11.1
	Female syphilis	323	14	15	29	48.3	4.3	9.0
	Male syphilis	1,800	900	82	982	91.6	50.0	54.6
12.0	Female chlamydia	21,312	153	127	280	54.6	0.7	1.3
	Male chlamydia	10,158	506	116	622	81.4	5.0	6.1
	Female syphilis	323	14	2	16	87.5	4.3	5.0
	Male syphilis	1,800	892	28	920	97.0	49.6	51.1
15.0	Female chlamydia	21,312	153	16	169	90.5	0.7	0.8
	Male chlamydia	10,158	501	23	524	95.6	4.9	5.2
	Female syphilis	323	14	0	14	100.0	4.3	4.3
	Male syphilis	1,800	864	15	879	98.3	48.0	48.8
15–19 years of age								
10.0	Female chlamydia	7,759	23	233	256	9.0	0.3	3.3
12.0	Female chlamydia	7,759	22	38	60	36.7	0.3	0.8
15.0	Female chlamydia	7,759	22	6	28	78.6	0.3	0.4

HIV = human immunodeficiency virus

AIDS = acquired immunodeficiency syndrome

STD = sexually transmitted disease

PPV = positive predictive value

further increased to 90.5%, 95.6%, 100.0%, and 98.3% for the female chlamydia, male chlamydia, female syphilis, and male syphilis subpopulations, respectively. However, when PPV increased, the number of missed matches (i.e., false negatives) increased as well. For example, among 900 matched true male HIV-syphilis coinfection cases identified when the cutoff value was 10.0, eight and 36 cases were missed, respectively, when the cutoff value was increased to 12.0 and 15.0 (Table).

We further examined the PPV of the matching algorithm in young female chlamydia cases aged 15–19 years who had the highest case rate and lowest HIV coinfection rate.[11,13] When the cutoff value was 10.0, the PPV in this group was only 9.0%, which meant that more than 90% were false matches; the PPV increased to 36.7% and 78.6%, respectively, when the cutoff value was increased to 12.0 and 15.0 (Table).

DISCUSSION

We demonstrated the effect of case rate and coinfection rate on the PPV of a given matching algorithm. Specifically, the PPV was higher in a population with a lower case rate and lower in a population with a higher case rate, and increased as the coinfection rate increased. In a match of the NYC HIV/AIDS and STD registries, PPV was highest in male syphilis cases, who had a low STD case rate (22.94 per 100,000 population) and the highest coinfection rate (50.0%), and lowest in female chlamydia cases, who had the highest STD case rate (980.95 per 100,000 population) and the lowest coinfection rate (0.7%).[17]

To reduce false positives, the specificity of a matching algorithm can be increased by applying a stricter matching criterion (e.g., a higher cutoff value in LinkPlus for comparison pairs to be accepted as matches). The score in LinkPlus is the sum of the agreement weights (positive values) and disagreement weights (negative values) for each matching variable. The higher the score a comparison pair receives, the more likely the pair is a true match; thus, the higher the cutoff value, the higher the PPV. As PCSI-related registry data matching across the U.S. becomes more routine and automated (i.e., with reduced opportunity for manual review), jurisdictions may need to apply different matching criteria in populations with different case rates and coinfection rates to achieve an acceptable PPV. Depending on program needs, resources, and objectives, it may even be reasonable to exclude from data matching certain subgroups known to have a high case rate but low coinfection rate (e.g., young female chlamydia cases).

Limitations

This study was subject to several limitations. We did not examine four important matters: (*1*) independent effects of case rate and coinfection rate on PPV, (*2*) completeness of the registry data for the match, (*3*) sensitivity and specificity of a matching algorithm, and (*4*) size of the population within which the matching was taking place. We demonstrated the combined effect of case rate and coinfection rate on the PPV of a matching algorithm but not their independent effects, because no subgroups in our dataset had the same case rate or coinfection rate. Statistical models can be used to demonstrate the independent effects.[9] Second, a reduction or imbalance in completeness has the potential to dramatically alter results. Therefore, prior to performing any registry data linkage, it is important to evaluate the completeness of each dataset.

Third, the sensitivity and specificity of a registry-matching algorithm are not totally independent of the population of interest because in data matching, some characteristics of the population (e.g., names) are included in the "test" and may affect its sensitivity and specificity. The same matching algorithm may therefore result in a higher sensitivity and lower specificity in one racial/ethnic group than in another.[13] Thus, programs should evaluate the sensitivity and specificity of a new algorithm against the composition of their local population before applying it to a local electronic data match of case surveillance registries for PCSI.

Lastly, the PPV of a matching algorithm varies inversely with the size of the population from which the people being matched are drawn.[9] For example, if we were to limit the HIV/AIDS and STD matching to cases residing in a single NYC borough, the PPV would increase with a corresponding increase in the number of false negatives/missed matches. This increase is important for PCSI in that it may not be reasonable to use the same matching algorithm in all jurisdictions because population size varies so widely. Before conducting registry data matching in PCSI, program staff should evaluate the sensitivity and specificity of the matching algorithm, and assess case rates, expected coinfection rate, and size of the populations included in the match.

CONCLUSION

Using HIV/AIDS and STD matching as an example, we demonstrated the effect of case rate and coinfection rate on the PPV of a matching algorithm: PPV decreases as the case rate increases and the coinfection rate decreases. Before conducting registry data matching in PCSI, program staff should assess the case rate

and coinfection rate of the population included in the data matching and select an appropriate matching algorithm.

This analysis was supported in part by a cooperative agreement with the Centers for Disease Control and Prevention (CDC), PS08-80202, #UC62/CCU223595, and an appointment to the Applied Epidemiology Fellowship Program administered by the Council of State and Territorial Epidemiologists (CSTE) and funded by CDC Cooperative Agreement #5U38HM000414.

The funding organizations played no role in the study design, data collection and analysis, or manuscript approval.

The findings and conclusions in this article are those of the authors and do not necessarily represent the views of the New York City Department of Health and Mental Hygiene.

The analysis used surveillance data and was a public health practice activity, not human subject research. Thus, it was exempt from institutional review board approval.

REFERENCES

1. Centers for Disease Control and Prevention (US). Program collaboration and service integration: enhancing the prevention and control of HIV/AIDS, viral hepatitis, sexually transmitted diseases, and tuberculosis in the United States. Atlanta: CDC; 2009.
2. Newman LM, Samuel MC, Stenger MR, Gerber TM, Macomber K, Stover JA, et al. Practical considerations for matching STD and HIV surveillance data with data from other sources. Public Health Rep 2009;124 Suppl 2:7-17.
3. Gollub EL, Trino R, Salmon M, Moore L, Dean JL, Davidson BL. Co-occurrence of AIDS and tuberculosis: results of a database "match" and investigation. J Acquir Immune Defic Syndr Hum Retrovirol 1997;16:44-9.
4. Moore M, McCray E, Onorato IM. Cross-matching TB and AIDS registries: TB patients with HIV co-infection, United States, 1993–1994. Public Health Rep 1999;114:269-77.
5. Xia Q, Westenhouse JL, Schultz AF, Nonoyama A, Elms W. Matching AIDS and tuberculosis registry data to identify AIDS/tuberculosis comorbidity cases in California. Health Inform J 2011;17:41-50.
6. Stenger MR, Courogen MT, Carr JB. Trends in *Neisseria gonorrhoeae* incidence among HIV-negative and HIV-positive men in Washington State, 1996–2007. Public Health Rep 2009;124 Suppl 2:18-23.
7. Manning SE, Pfeiffer MR, Nash D, Blank S, Sackoff J, Schillinger J. Incident sexually transmitted infections among persons living with diagnosed HIV/AIDS in New York City, 2001–2002: a population-based assessment. Sex Transm Dis 2007;34:1008-15.
8. Etkind P, Tang Y, Whelan M, Ratelle S, Murphy J, Sharnprapai S, et al. Estimating the sensitivity and specificity of matching name-based with non-name-based case registries. Epidemiol Infect 2003;131:669-74.
9. Karmel R, Gibson D. Event-based record linkage in health and aged care services data: a methodological innovation. BMC Health Serv Res 2007;7:154.
10. HIV surveillance—United States, 1981–2008. MMWR Morb Mortal Wkly Rep 2011;60(21):689-93.
11. Centers for Disease Control and Prevention (US). Sexually transmitted disease surveillance, 2010. Atlanta: Department of Health and Human Services (US); 2011.
12. Blocker ME, Levine WC, St Louis ME. HIV prevalence in patients with syphilis, United States. Sex Transm Dis 2000;27:53-9.
13. Olson N, Samuel MC, Brodsky J, Gilson D, Damesyn M, Bernstein K, et al. STD and HIV/AIDS case registry matching to estimate California STD-HIV/AIDS co-infection. Richmond (CA): California Department of Public Health, Center for Infectious Diseases, Division of Communicable Disease Control, Sexually Transmitted Diseases Control Branch; 2011.
14. Torian LV, Makki HA, Menzies IB, Murrill CS, Weisfuse IB. HIV infection in men who have sex with men, New York City Department of Health sexually transmitted disease clinics, 1990–1999: a decade of serosurveillance finds that racial disparities and associations between HIV and gonorrhea persist. Sex Transm Dis 2002;29:73-8.
15. Torian LV, Makki HA, Menzies IB, Murrill CS, Benson DA, Schween FW, et al. High HIV seroprevalence associated with gonorrhea: New York City Department of Health, sexually transmitted disease clinics, 1990–1997. AIDS 2000;14:189-95.
16. Centers for Disease Control and Prevention (US). Registry Plus® LinkPlus® Version 2.0. Atlanta: Registry Plus; 2007.
17. New York City Department of Health and Mental Hygiene. Bureau of Sexually Transmitted Disease Control quarterly report, January 1, 2011–June 30, 2011. New York: NYC DOHMH; 2011. Also available from: URL: http://www.nyc.gov/html/doh/downloads/pdf/std/std-quarterlyreport2011-2.pdf [cited 2013 Feb 27].