Health Informatics Journal

http://jhi.sagepub.com/

Matching AIDS and tuberculosis registry data to identify AIDS/tuberculosis comorbidity cases in California

Qiang Xia, Janice L. Westenhouse, Alan F. Schultz, Atsuko Nonoyama, William Elms, Nancy Wu, Lisette Tabshouri, Juan D. Ruiz and Jennifer M. Flood

Health Informatics Journal 2011 17: 41
DOI: 10.1177/1460458210380524

Article

Health Informatics Journal

Health Informatics Journal
17(1) 41–50

DOI: 10.1177/1460458210380524
jhi.sagepub.com
SAGE

Matching AIDS and tuberculosis registry data to identify AIDS/tuberculosis comorbidity cases in California

Qiang Xia, Janice L. Westenhouse, Alan F. Schultz, Atsuko Nonoyama and William Elms
California Department of Public Health, USA

Nancy Wu
University of California, Davis, USA

Lisette Tabshouri, Juan D. Ruiz and Jennifer M. Flood
California Department of Public Health, USA

Abstract

The purpose of this study was to evaluate the sensitivity and positive predictive value (PPV) of a registry data linkage procedure used in the California AIDS and Tuberculosis (TB) Registry Data Linkage Study to identify AIDS/TB comorbidity cases in California. The California AIDS registry data from 1981 to 2006 were linked to the California TB registry data from 1996 to 2006 using LinkPlus, a probabilistic record linkage program developed by the Centers for Disease Control and Prevention, and matched results were manually reviewed to determine true or false matches. We estimated the sensitivity of this procedure to range from 98.0 per cent (95% confidence interval, CI: 97.3%, 98.7%) to 98.8 per cent (95% CI: 98.1%, 99.2%), and the PPV to be 100 per cent (95% CI: 96.8%, 100.0%). Our study demonstrated the feasibility of using this linkage procedure to match AIDS and TB registry data with a very high degree of accuracy.

Keywords
acquired immunodeficiency syndrome, comorbidity, positive predictive value, registries, sensitivity, tuberculosis

Corresponding author:
Qiang Xia (*Chiang Hsia*) MD MPH, RTI International, PO Box 12194, Research Triangle Park, NC 27709–2194, USA
Email: qiangxia@post.harvard.edu.

The HIV/AIDS epidemic in the past 25 years has had a dramatic impact on the epidemiology of tuberculosis (TB), especially in populations where both infections are prevalent.[1–4] The World Health Organization estimates that one-third of the world's population is infected with TB, and approximately 10 million people are coinfected with TB and HIV.[5, 6] In the United States, the Centers for Disease Control and Prevention (CDC) estimated that 14 per cent of TB cases reported in 2004 among people aged 25 to 44 were coinfected with HIV.[7]

HIV infection is the most potent risk factor for the progression of TB infection to TB disease.[1, 8, 9] Before the availability of highly active antiretroviral therapy, people dually infected with HIV and TB had more than 100 times increased risk of developing active TB disease and becoming infectious compared with people who were only infected with TB.[8, 10] In addition, people living with HIV/AIDS are at greater risk for developing multi-drug-resistant TB, which is extremely difficult to treat and can be fatal.[11–15]

To monitor the impact of HIV infection on the epidemiology of TB in a region, AIDS and TB registry data linkage had been widely used by state health departments in the United States.[16–22] However, all studies used a deterministic method without reporting its sensitivity and positive predictive value (PPV). A deterministic method uses a combination of algorithms and rules to look for exact agreement on one or more matching variables between data files with the capability to catch some common errors such as typos, phonetic variations, and transpositions.[23, 24] A probabilistic method, by contrast, calculates the probability that a comparison pair is a true match based on a group of matching variables with the capability to catch more complex typographical errors and error patterns.[24–29]

In order to reduce the bias caused by misclassification, a very high degree of accuracy is needed in the AIDS and TB registry data linkage. The purpose of this study was to evaluate the sensitivity and PPV of an AIDS/TB registry data linkage procedure using a combination of probabilistic and deterministic methods.

Methods

Datasets

The California Department of Public Health, Center for Infectious Diseases, Office of AIDS maintains the AIDS registry for all reported AIDS cases, and the TB Control Branch maintains the TB registry for all reported TB cases in California. The AIDS and TB registry data contain both demographic and clinical information. All cases in the AIDS registry were assigned an HIV/AIDS Reporting System (HARS) number, and all cases in the TB registry were assigned a Report of Verified Case of Tuberculosis (RVCT) number. Some AIDS cases with a verified TB diagnosis may have their RVCT number entered in the AIDS registry, and some TB cases with a verified AIDS diagnosis may also have their HARS number documented in the TB registry. These are AIDS/TB cases, whose statuses are known. Therefore, we can identify some AIDS/TB cases without performing any data linkage.

Both the California AIDS and TB registry data were de-duplicated before data linkage. A patient with multiple episodes of TB in different years was considered multiple cases, not a duplicate. AIDS cases diagnosed between 1981 and 2006, and TB cases diagnosed between 1996 and 2006, were included in the study. AIDS cases that had a confirmed date of death earlier than 1 January 1996, were removed before data linkage to reduce the computer time and false matches.

Before the data linkage, three categorical matching variables were recoded. Sex was recoded with two categories, male and female; race/ethnicity was recoded with six categories, Hispanic, Non-Hispanic White, Non-Hispanic Black, Asian/Native Hawaiian/Pacific Islander, American

Indian/Alaskan Native, and other; and origin was recoded with two categories, US born and foreign born. The date of birth variable was coded in an mm/dd/yyyy format.

Computerized data linkage

LinkPlus 2.10, a probabilistic record linkage program developed by CDC for cancer registry database linkage and de-duplication, was used for the California AIDS and TB Registry Data Linkage Study.[30] Table 1 shows a list of variables used for blocking, matching, and manual review. A blocking variable is a variable common to the AIDS and TB registry data files that are used to 'block' (or partition) the two files. LinkPlus provides a simple blocking ('OR blocking') mechanism by indexing the variables for blocking and comparing the pairs with the identical values on at least one of those variables. Blocking is a way to reduce the computing cost by portioning files into mutually exclusive and exhaustive blocks and performing comparisons only on records within each block. Blocking also reduces manual review time because of fewer false positive links. In this study, first name, last name, and year of birth were selected to be the blocking variables, and Soundex code was used for the first name and last name.

In LinkPlus version 2, a record in the first comparison file, file 1, can match multiple records in the second, file 2, but not vice versa (i.e., only one-to-many linkages are supported, not many-to-many linkages). As an AIDS patient can have multiple episodes of TB, we specified the AIDS registry data as file 1, and the TB registry data as file 2.

LinkPlus has an option to choose either a direct or an indirect method to compute the *m*-probability, which is the probability of agreement for a given matching variable. Because many patients in the TB registry were born outside the United States and their names were not common in the US Census or AIDS registry data, an indirect method was chosen to compute the *m*-probability in the

Table 1. List of variables used for blocking, matching and manual review in the California AIDS and TB Registry Data Linkage Study

	Blocking	Matching	Manual review
First name (Soundex)	√		
Last name (Soundex)	√		
First name		√	√
Last name		√	√
Date of birth		√	√
Sex		√	√
Race		√	√
Origin		√	√
Year of birth	√		√
HARS number			√
RVCT number			√
Middle name			√
Residence (street)			√
Residence (city)			√
Residence (zip code)			√
Date of TB diagnosis			√
Date of AIDS diagnosis			√
History of TB			√

AIDS, acquired immunodeficiency syndrome; TB, tuberculosis; HARS, HIV/AIDS reporting system; RVCT, report of verified case of tuberculosis.

study to be more reflective of the true probability. Thus, the m-probability was computed by capturing and utilizing the information dynamically from the actual data being linked.

When a data linkage is run, LinkPlus generates a score for all comparisons of a case in file 1 (the AIDS registry file) with a case in file 2 (the TB registry file). The score is the sum of the agreement and disagreement weights for each matching variable.[27] If a comparison has a score equal to or higher than the specified cutoff value, the AIDS and TB case records will be written to the linkage report as a potential link, which requires manual review to assign a match status of match or non-match. We set a cutoff value of 10.0.

Record linkage was used even before computers became widely available, but record linkage terms were not defined the same way.[24, 27, 31] In this study, we defined a comparison pair as any possible comparison of a record in the AIDS registry dataset with a record in the TB registry dataset; a potential link as a comparison pair with a score of the cutoff value or higher generated by the probabilistic record linkage software, LinkPlus; a match as a potential link that is accepted to be the same individual after manual review; and a non-match as a potential link that is not accepted to be the same individual after manual review.

Manual review

Two reviewers independently reviewed all potential links with a score of 10.0 or higher to decide whether the link was a match or a non-match. A third reviewer was called in to resolve all the discrepancies through discussion with the two reviewers.

A link was considered a match if it met any of the following criteria: (1) a perfect match on all six matching variables; (2) not a perfect match but identical RVCT numbers; (3) not a perfect match but identical HARS numbers; and (4) not a perfect match but identical residence addresses (street number, city, and zip code). If a link did not meet any of the above criteria, two reviewers made a personal judgment based on the score and other information, such as: (1) an obvious typo in 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, (7) transposed birth month and day, (8) documentation of TB diagnosis in the AIDS registry, and (9) closeness of the dates of AIDS and TB diagnoses.

Sensitivity

It is difficult to evaluate sensitivity for this study because we do not know of all AIDS/TB cases in the registries, only a proportion. We know that AIDS cases in the AIDS registry with a valid 1996–2006 RVCT number, and TB cases in the TB registry with a valid HARS number, are AIDS/TB cases, but we also know that many AIDS/TB cases did not have such information documented in the registries. In order to evaluate the sensitivity, we had to use the matching results from the AIDS/TB cases whose statuses are known to estimate how many true AIDS/TB cases can be captured by this linkage method.

First, we identified all cases with a known AIDS/TB status, and compiled a list of such cases after de-duplication. Then, we compared the list with the matching results to see how many such cases were captured and separated them into three categories: perfect match, good match, and partial match. We also identified how many such cases were missed and put them into the fourth category: non-match. A perfect match is a link that matches on all matching variables; a good match is a link that matches the majority of matching variables, and has obvious typos, transpositions, and/or missing values in other variables that did not match; a partial match is a link that matches a few matching variables, but has identical HARS or RVCT numbers. Different from a partial match, which requires identical HARS or RVCT numbers in two files to be accepted as a match, a perfect match and a good match can be accepted as a match solely based on matching variable information. A non-match means an AIDS/TB case did not find a match in the other registry because of the limitations of the data linkage software, or reporting errors.

Using the information from cases with a known AIDS/TB status, we present the matching results of all AIDS/TB cases in Table 2, assuming that we had a known number *m* and an unknown number *n* of AIDS/TB cases in the two registries, and the quality of data entry for known AIDS/TB cases was the same as for unknown AIDS/TB cases in terms of data entry errors and missing values. Of those perfect and good matches, all AIDS/TB cases ($m1 + n1$) were captured regardless of their HARS or RVCT number information. Of those partial matches, only AIDS/TB cases ($m2$) with a known status (RVCT numbers recorded in the AIDS registry, and/or HARS numbers documented in the TB registry) were captured because of the requirement of identical HARS or RVCT numbers to be accepted as a match, and AIDS/TB cases ($n2$) without HARS and RVCT information were missed. Of those non-matches, all of them ($m3 + n3$) were missed. Thus, in total, we are able to capture $m1 + m2 + n1$ out of $m + n$ AIDS/TB cases.

Equation 1 was then used to compute the sensitivity. If m is much greater than n ($m >> n$), the sensitivity closes to $(m1 + m2)/m$; if m is much smaller than n ($m << n$), the sensitivity closes to $m1/m$. As the size of n was unknown, we reported the sensitivity in a range format between $m1/m$ and $(m1 + m2)/m$.

$$sensitivity = \frac{m1 + m2 + n1}{m + n} = \frac{(m1 + m2) + \frac{m1}{m} n}{m + n} \qquad \text{(equation 1)}$$

PPV

As the number of matches was so large and patients were diagnosed over an 11-year span, it was not feasible to review all patient records to confirm their HIV and TB infection status. Thus, only matches with a TB diagnosis in 2006 were sent to local health jurisdictions (LHJs) for confirmation to assess PPV by checking patients' medical records, registry records, or both. LHJs reported their confirmation results as a true match, a false match, or indeterminate.

Table 2. Linkage results with a known number of *m* and an unknown number of *n* of HIV/TB coinfection cases in the AIDS and TB registries

Matching status	Cases with a known coinfection status	Cases with an unknown coinfection status	Coinfection cases identified after linkage
Perfect/good matches	$m1$	$n1 = (m1/m)n$	$m1 + n1$
Partial matches	$m2$	$n2 = (m2/m)n$	$m2$
Non-matches	$m3$	$n3 = (m3/m)n$	0
Total	m	n	$m1 + m2 + n1$

AIDS, acquired immunodeficiency syndrome; TB, tuberculosis.

Results

By 31 December 2006, 151,978 AIDS cases diagnosed between 1981 and 2006 and living in California were reported to the California Department of Public Health, Office of AIDS, and 68,619 cases were removed before the linkage because of a confirmed date of death earlier than 1 January 1996, leaving 83,359 AIDS cases for the study. Between 1996 and 2006, 37,105 TB cases diagnosed and living in California were reported to the California Department of Public Health, TB Control Branch, all of which were included in the study.

The record linkage results are shown in Table 3. LinkPlus identified 9408 potential links with a score of 10.0 or higher. After manual review, 2236 of them were accepted as matches (i.e. AIDS/TB cases). As expected, the higher the score a potential link had, the more likely it was accepted as a match. Links with a score between 25.0 and 33.5 were all accepted as matches, but only 1.2 per cent of links with a score between 10.0 and 14.9 were accepted as matches. One link with a score as low as 10.0 was accepted as a match based on identical RVCT numbers and residence addresses.

In the AIDS and TB registries, we identified a total of 1692 cases with a known AIDS/TB status: AIDS cases in the AIDS registry with a valid 1996–2006 RVCT number, and TB cases in the TB registry with a valid HARS number. Twenty-one of these AIDS/TB cases were missed after the matching process (software run and manual review), and 12 more were missed after repeating the matching process while withholding their HARS and RVCT number information. Based on these two numbers, the values of *m1*, *m2*, *m3*, and *m* were computed as 1659, 12, 21, and 1692, respectively. Using equation 1, we obtained the sensitivity of our method in a range format from 98.0 per cent (95% confidence interval, CI: 97.3%, 98.7%) to 98.8 per cent (95% CI: 98.1%, 99.2%).

LinkPlus version 2 allows one-to-many matching, but not many-to-many matching. A few matches can be missed because of this restriction, when two or more records in one registry are similar to each other but not duplicates. To reduce the number of missed matches, one of the options is to switch file 1 and file 2 and repeat the matching process. By doing this, nine more matches were captured, and the sensitivity was increased to 98.6–99.3 per cent.

A total of 129 matches with a TB diagnosis in 2006 were identified, and sent to LHJs for confirmation by checking patients' medical records, registry records, or both. One hundred and fourteen matches were returned with 113 true matches, one indeterminate, and no false matches. The match with an indeterminate status was because the AIDS and TB diagnoses were made in and reported by two different counties in California and there were some difficulties in confirming AIDS/TB cases across counties, despite the fact that the pair had identical information in matching

Table 3. Number of links identified and number of matches accepted in the California AIDS and TB Registry Data Linkage Study

Score	Links	Matches	Matches/links (%)
30.0–33.5	108	108	100.0
25.0–29.9	586	586	100.0
20.0–24.9	1175	1141	89.5
15.0–19.9	661	321	57.2
10.0–14.9	6878	80	1.2
Total	9408	2236	23.8

AIDS, acquired immunodeficiency syndrome; TB, tuberculosis.

variables (first name, last name, race/ethnicity, gender, origin, and date of birth) in the two registries. Treating this match and other non-returned matches as missing values, we estimated the PPV of our method to be 100 per cent (95% CI: 96.8%, 100.0%).

Discussion

Knowing the TB status of HIV-infected individuals and the HIV status of TB patients is essential not only to patients' HIV and TB care, but also to public health interventions seeking to reduce the comorbidity of HIV/TB. Because of the under-reporting of TB status in the HIV/AIDS registry and of HIV status in the TB registry, neither registry provides accurate surveillance data of HIV/TB coinfections: one-third of TB patients reported to the United States National TB Surveillance System in 2005 contained no information about HIV status,[32] and one-fifth of AIDS/TB comorbidity cases reported to the Florida AIDS registry in 1981–1993 had no documentation of TB.[19] Thus, linking the AIDS and TB registry data becomes a very important tool to identify AIDS/TB cases, and has been widely used by the state health departments in the United States since the mid 1980s.[16–22]

Previously, a deterministic method was used in all AIDS and TB registry data linkage studies using a combination of algorithms and rules to look for exact agreement on one or more matching variables between data files with the capability to catch some common errors such as typos, phonetic variations, and transpositions.[16–24] However, none of these studies reported the sensitivity and PPV of the methods applied.

In our study, a combination of probabilistic method and deterministic method (manual review) was used to match the California AIDS and TB registry data by calculating the probability of a comparison pair as a potential link based on a group of matching variables with the capability to catch more complex typographical errors and error patterns,[24–29] and manually reviewing all potential links to determine true or false matches. Using this method to match the California AIDS and TB registry data, we obtained a very high degree of accuracy with a sensitivity of 98.6–99.3 per cent and a PPV of 100 per cent. However, in LinkPlus, m-probabilities are generated based on the data being linked. It is important to note that the linkage scores and calculations of sensitivity and PPV in the current study are specific to this particular linkage and datasets, and may not be extrapolated to all linkages performed with LinkPlus.

In addition to the sensitivity and PPV, there are a few important issues that must be considered in an AIDS and TB registry data linkage study: (1) completeness of the AIDS and TB registry data; (2) validity of the AIDS and TB registry data; and (3) delay in AIDS diagnosis.

Incompleteness of the AIDS and TB registry data would seriously alter final results. For example, if both AIDS and TB registry data were only 50 per cent complete, even a highly sensitive method would have only captured 25 per cent (50% × 50% = 25%) of AIDS/TB cases. Therefore, before performing any registry data linkage, it is very important to evaluate the completeness of the datasets. In our study, both California AIDS and TB registry data had been previously examined and were found to be very complete.[33–35]

The quality of registry data linkage depends on the quality of the data collection (i.e. the validity of the AIDS and TB surveillance data). Missed matches are mostly caused by poor data collection and data entry (e.g. missing values and typos). Studies have demonstrated the high validity of the California AIDS and TB surveillance data.[34, 36]

Delay in AIDS diagnosis may also affect the linkage results. For example, if a patient was diagnosed with TB and reported to the TB registry in 2000 and his HIV-positive status was later found out and reported to the AIDS registry in 2002, we would not be able to capture this AIDS/TB case by matching the 1981–2000 AIDS registry data with the 2000 TB registry data. The registry data linkage then would miss some true AIDS/TB cases, and we would underestimate the AIDS/TB prevalence, especially in an area where HIV testing is not routinely taking place in TB programs. To reduce the number of AIDS/TB cases missed due to delay in AIDS diagnosis, in our study we performed one record linkage using 11-year TB registry data (1996–2006) to match the whole AIDS registry dataset (1981–2006) rather than performing 11 matches with one-year TB registry data for each.

By matching multi-year rather than one-year TB registry data with AIDS registry data, we should be aware of a possible scenario of true match but a false AIDS/TB case due to temporal sequence of the infections. When we identify a match with TB diagnosis earlier than AIDS diagnosis, we need to decide whether this is a true AIDS/TB case because of delay in AIDS diagnosis, or a false coinfection case because the patient acquired HIV shortly after his or her TB was cured. As we are unable to find out when the patient acquired HIV infection, we have to make a decision based on available information (dates of TB diagnosis and AIDS diagnosis) and the knowledge that the median time of progression from HIV infection to AIDS is 9–10 years in the absence of antiretroviral treatment.[37] In our study, all matches (n = 37) with a TB diagnosis earlier than an AIDS diagnosis were accepted as AIDS/TB cases.

LinkPlus is a probabilistic record linkage program developed by CDC for cancer registry database linkage and de-duplication.[30] Our study demonstrated that it could also be used for the AIDS and TB registry data linkage with extremely high accuracy. The software is easy to use, and provides a very important and useful feature: manual review. Without manual review, we could have included many false positives and false negatives in our final results depending on the cutoff value we set (Table 3). For the purpose of evaluation, we manually reviewed 9408 potential links with a linkage of 11-year TB data. Given the size of AIDS and TB registry data in California, it is feasible to perform manual review in the future when we match AIDS registry data with only one-year TB registry data.

LinkPlus does not allow many-to-many linkages, it allows one-to-many. A few matches can be missed because of this restriction. In our study, we increased our sensitivity from 98.0–98.8 per cent to 98.6–99.3 per cent by repeating the matching process (software run and manual review) after switching the order of file 1 and file 2, but the process was burdensome. The problem can be solved if the software were modified to include an option of one-to-many or many-to-many linkage, and to group all related links together for manual review.[38] LinkPlus version 3.0 will address this issue, and will allow users to specify a one-to-one or many-to-many linkage upon linkage configuration.

We acknowledge limitations to the generalizability of our findings. The study used existing software and applied it specifically to match AIDS and TB registry data, and the study findings may not be extrapolated to other linkages performed with the software, LinkPlus. Therefore, we suggest that future studies always evaluate their procedures when using LinkPlus to match other datasets.

In conclusion, our study demonstrated the feasibility of using this registry data linkage procedure to match the AIDS and TB registry data with a very high degree of accuracy, and we also provided a new tool that can be used in other studies to evaluate and report sensitivity in a range format when a common unique identifier (e.g. social security number) was entered in partial records of the files to be linked.

Acknowledgements

The authors thank HIV/AIDS surveillance coordinators from the Counties of Imperial, Kern, Los Angeles, Monterey, Orange, Sacramento, San Francisco, San Joaquin, San Diego, and Santa Clara who verified the AIDS/TB cases for this study. Special thanks are due to Michael Donahue and Teresa Lauchaire of the HIV/AIDS Case Registry Section, CDPH/OA for their assistance in coordinating the project. The authors would also like to thank Richard Selik, David Gu, and Kathleen Thoburn from the Centers for Disease Control and Prevention (CDC) for their comments and suggestions.

References

1. Cantwell MF, Snider DE Jr, Cauthen GM, Onorato IM. Epidemiology of tuberculosis in the United States, 1985 through 1992. *J Am Med Assoc* 1994; 272: 535–539.
2. Zumla A, Malon P, Henderson J, Grange JM. Impact of HIV infection on tuberculosis. *Postgrad Med J* 2000; 76: 259–268.
3. Lienhardt C and Rodrigues LC. Estimation of the impact of the human immunodeficiency virus infection on tuberculosis: tuberculosis risks re-visited? *Int J Tuberc Lung Dis* 1997; 1: 196–204.
4. Shafer RW and Edlin BR. Tuberculosis in patients infected with human immunodeficiency virus: perspective on the past decade. *Clin Infect Dis* 1996; 22: 683–704.
5. Small PM. Tuberculosis research: balancing the portfolio. *J Am Med Assoc* 1996; 276: 1512–1513.
6. Dye C, Scheele S, Dolin P, Pathania V, Raviglione MC. Consensus statement. Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project. *J Am Med Assoc* 1999; 282: 677–686.
7. *Reported tuberculosis in the United States, 2005*. Atlanta, GA: US Department of Health and Human Services, CDC, 2006, p.23.
8. Menzies D and Pourier L. Diagnosis of tuberculosis infection and disease. In: Long R (ed.) *The Canadian tuberculosis standards*. Ottawa: Canadian Lung Association and Health Canada, 2000, p.45–65.
9. Havlir DV and Barnes PF. Tuberculosis in patients with human immunodeficiency virus infection. *N Engl J Med* 1999; 340: 367–373.
10. Pitchenik AE, Fertel D and Bloch AB. Mycobacterial disease: epidemiology, diagnosis, treatment, and prevention. *Clin Chest Med* 1988; 9: 425–441.
11. Frieden TR, Sterling T, Pablos-Mendez A, Kilburn JO, Cauthen GM, Dooley SW. The emergence of drug-resistant tuberculosis in New York City. *N Engl J Med* 1993; 328: 521–526.
12. Gordin FM, Nelson ET, Matts JP, et al. The impact of human immunodeficiency virus infection on drug-resistant tuberculosis. *Am J Respir Crit Care Med* 1996; 154: 1478–1483.
13. Nolan CM, Williams DL, Cave MD, et al. Evolution of rifampin resistance in human immunodeficiency virus-associated tuberculosis. *Am J Respir Crit Care Med* 1995; 152: 1067–1071.
14. Bradford WZ, Martin JN, Reingold AL, Schecter GF, Hopewell PC, Small PM. The changing epidemiology of acquired drug-resistant tuberculosis in San Francisco, USA. *Lancet* 1996; 348: 928–931.
15. Beck-Sague C, Dooley SW, Hutton MD, et al. Hospital outbreak of multidrug-resistant Mycobacterium tuberculosis infections. Factors in transmission to staff and HIV-infected patients. *J Am Med Assoc* 1992; 268: 1280–1286.
16. Moore M, McCray E and Onorato IM. Cross-matching TB and AIDS registries: TB patients with HIV coinfection, United States, 1993–1994. *Public Health Rep* 1999; 114: 269–277.
17. Burwen DR, Bloch AB, Griffin LD, Ciesielski CA, Stern HA, Onorato IM. National trends in the concurrence of tuberculosis and acquired immunodeficiency syndrome. *Arch Intern Med* 1995; 155: 1281–1286.
18. Co-incidence of HIV/AIDS and tuberculosis: Chicago, 1982–1993. *MMWR Morb Mortal Wkly Rep* 1995; 44: 227–231.
19. Surveillance of tuberculosis and AIDS co-morbidity: Florida, 1981–1993. *MMWR Morb Mortal Wkly Rep* 1996; 45: 38–41.
20. Rieder HL, Cauthen GM, Bloch AB, et al. Tuberculosis and acquired immunodeficiency syndrome: Florida. *Arch Intern Med* 1989; 149: 1268–1273.
21. Libbus MK, Phillips L and Knudson JK. TB-HIV registry matching in Missouri, 1987–1999. *Public Health Nurs* 2002; 19: 470–474.

22. 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: 44–49.
23. Muse AG, Mikl J and Smith PF. Evaluating the quality of anonymous record linkage using deterministic procedures with the New York State AIDS registry and a hospital discharge file. *Stat Med* 1995; 14: 499–509.
24. Clark DE. Practical introduction to record linkage for injury research. *Inj Prev* 2004; 10: 186–191.
25. Newgard CD. Validation of probabilistic linkage to match de-identified ambulance records to a state trauma registry. *Acad Emerg Med* 2006; 13: 69–75.
26. Dean JM, Vernon DD, Cook L, Nechodom P, Reading J, Suruda A. Probabilistic linkage of computerized ambulance and inpatient hospital discharge records: a potential tool for evaluation of emergency medical services. *Ann Emerg Med* 2001; 37: 616–626.
27. Blakely T and Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. *Int J Epidemiol* 2002; 31: 1246–1252.
28. Jaro MA. Probabilistic linkage of large public health data files. *Stat Med* 1995; 14: 491–498.
29. Reynolds P, Saunders LD, Layefsky ME, Lemp GF. The spectrum of acquired immunodeficiency syndrome (AIDS)-associated malignancies in San Francisco, 1980–1987. *Am J Epidemiol* 1993; 137: 19–30.
30. *Link Plus Version 2.10 probabilistic record linkage software*. Atlanta, GA: Centers for Disease Control and Prevention, 2006.
31. Dunn HL. Record Linkage. *Am J Public Health Nations Health* 1946; 36: 1412–1416.
32. Marks SM, Magee E and Robison V. Reported HIV status of tuberculosis patients: United States, 1993–2005. *MMWR Morb Mortal Wkly Rep* 2007; 56: 1103–1106.
33. Schwarcz SK, Hsu LC, Parisi MK, Katz MH. The impact of the 1993 AIDS case definition on the completeness and timeliness of AIDS surveillance. *Aids* 1999; 13: 1109–1114.
34. Klevens RM, Fleming PL, Li J, et al. The completeness, validity, and timeliness of AIDS surveillance data. *Ann Epidemiol* 2001; 11: 443–449.
35. Curtis AB, McCray E, McKenna M, Onorato IM. Completeness and timeliness of tuberculosis case reporting: a multistate study. *Am J Prev Med* 2001; 20: 108–112.
36. Sprinson JE, Lawton ES, Porco TC, Flood JM, Westenhouse JL. Assessing the validity of tuberculosis surveillance data in California. *BMC Public Health* 2006; 6: 217.
37. Morgan D, Mahe C, Mayanja B, Okongo JM, Lubega R, Whitworth JA. HIV-1 infection in rural Africa: is there a difference in median time to AIDS and survival compared with that in industrialized countries? *Aids* 2002; 16: 597–603.
38. Campbell KM, Deck D and Krupski A. Record linkage software in the public domain: a comparison of Link Plus, The Link King, and a 'basic' deterministic algorithm. *Health Inf J* 2008; 14: 5–15.