Comparing the National Death Index and the Social Security Administration's Death Master File to Ascertain Death in HIV Surveillance

David B. Hanna, MS[a]
Melissa R. Pfeiffer, MPH[a]
Judith E. Sackoff, PhD[a]
Richard M. Selik, MD[b]
Elizabeth M. Begier, MD, MPH[a]
Lucia V. Torian, PhD[a]

SYNOPSIS

Objectives. New York City (NYC) maintains a population-based registry of people with human immunodeficiency virus (HIV) infection to monitor the epidemic and inform resource allocation. We evaluated record linkages with the National Death Index (NDI) and the Social Security Administration's Death Master File (SSDMF) to find deaths occurring from 2000 through 2004.

Methods. We linked records from 32,837 people reported with HIV and not previously known to be dead with deaths reported in the NDI and the SSDMF. We calculated the kappa statistic to assess agreement between data sources. We performed subgroup analyses to assess differences within demographic and transmission risk subpopulations. We quantified the benefit of linkages with each data source beyond prior death ascertainment from local vital statistics data.

Results. We discovered 1,926 (5.87%) deaths, which reduced the HIV prevalence estimate in NYC by 2.03%, from 1.19% to 1.16%. Of these, 458 (23.78%) were identified only from NDI, and 305 (15.84%) only from SSDMF. Agreement in ascertainment between sources was substantial (kappa = [K] 0.74, 95% confidence interval [CI] 0.72, 0.76); agreement was lower among Hispanic people (K=0.65, 95% CI 0.62, 0.69) and people born outside the U.S. (K=0.60, 95% CI 0.52, 0.68). We identified an additional 13.62% of deaths to people reported with HIV in NYC; white people and men who have sex with men were disproportionately likely to be underascertained without these linkages ($p<0.0001$).

Conclusion. Record linkages with national databases are essential for accurate prevalence estimates from disease registries, and the SSDMF is an inexpensive means to supplement linkages with the NDI to maximize death ascertainment.

[a]HIV Epidemiology and Field Services Program, Bureau of HIV/AIDS Prevention and Control, New York City Department of Health and Mental Hygiene, New York, NY

[b]Division of HIV/AIDS Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, GA

Address correspondence to: David B. Hanna, MS, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe St., PO Box 130, Baltimore, MD 21205; tel. 917-723-4377; fax 212-925-2701; e-mail <dhanna@jhsph.edu>.

©2009 Association of Schools of Public Health

Population-based disease registries aim to contain all occurrences of a disease within a defined jurisdiction to accurately describe the burden of the disease.[1] For example, name-based reporting of all cases of human immunodeficiency virus (HIV) infection (with or without acquired immunodeficiency syndrome [AIDS]) is mandated and conducted in every U.S. state and dependency to monitor the epidemic of this infection.[2,3] Accurate counts of people living with HIV infection are essential for resource allocation, as they are used to determine the provision of millions of dollars distributed annually for federally funded HIV-related care and treatment services.[4]

Having accurate counts of living (prevalent) cases in a jurisdiction requires both identifying new cases of disease and, for those already in the registry, maintaining current knowledge of their vital status and whether they still reside in the area. One challenge to the accuracy of case counts is out-migration of people originally reported in the defined area. For example, increased out-migration of people living with AIDS over time has been documented in New York City (NYC) since highly active antiretroviral therapy (HAART) was made widely available,[5] possibly as a consequence of HAART's positive effects on survival and quality of life.[6]

Conducting electronic record linkages for death ascertainment has been shown to be essential to more accurately measure HIV prevalence, and most state HIV/AIDS surveillance programs perform electronic record linkage to ascertain deaths.[7] The National Death Index (NDI) is one well-known source that is used to link records of deaths to records of people in disease registries, especially deaths occurring outside the jurisdiction that cannot be identified through matches with the local vital statistics registry.[8,9] Because searching the NDI can be costly, alternative sources have been considered to obtain death information, such as the Social Security Death Master File (SSDMF) maintained by the Social Security Administration (SSA). Previously published comparisons between the NDI and the SSA's data files based on cohort studies have generally found the NDI to have higher sensitivity in identifying deaths and more accurate dates of death.[10–12] However, to our knowledge, no comparison has been made between the two databases for people with HIV, who differ from the general population in many ways with respect to demographic, behavioral, and clinical characteristics.

We evaluated the NDI and the SSDMF to ascertain deaths among cases reported to the NYC HIV Surveillance Registry who died between January 1, 2000, and December 31, 2004. This article describes our linkage process and demonstrates the effects of linkage with each data source on HIV prevalence numbers for NYC for the year 2004. We assessed agreement between the two sources and evaluated the usefulness of the SSDMF in comparison with the NDI, as the SSDMF was a novel data source to us. Finally, we repeated these analyses within subpopulations of people with HIV, which can inform other disease registries on the utility of each database for certain groups.

METHODS

Data sources

The NYC HIV Surveillance Registry is a population-based registry of people diagnosed with HIV infection as defined by the Centers for Disease Control and Prevention (CDC),[13,14] and reported in NYC to the Department of Health and Mental Hygiene (DOHMH). Reporting of AIDS cases by name began in New York State in 1983, and reporting of non-AIDS cases of HIV infection by name was implemented in 2000.[15] The DOHMH learns of potential new HIV infections primarily through electronic laboratory test reports and physician reports. Field surveillance personnel review medical records to confirm reports and to collect data on sociodemographic characteristics.

Local deaths of people in the HIV Surveillance Registry are ascertained by quarterly linkages with electronic death certificates reported to the NYC DOHMH Vital Statistics Registry. These are supplemented with reports from medical record review by field staff. Deaths occurring outside NYC must be ascertained by record linkages with other sources.

The NDI is a database that is maintained by CDC's National Center for Health Statistics (NCHS).[8,16] Updated annually, it contains the name, social security number (SSN) (if available), demographic data, as well as date, state, and cause of death for all deaths that occurred after 1978 in the U.S. (including dependent territories), as derived from information on death certificates and provided by local vital records offices. Access to the NDI is made available to health researchers and disease registries solely for statistical purposes, with a fee charged per case per year searched to cover operating expenses, plus the annual cost of purchasing death records from all state vital records offices. Cases whose deaths are searched for in the NDI are temporarily linked by NCHS staff to the NDI, and NDI records of potential matches are returned to researchers and registries, who then decide which deaths they deem true matches. A routine search returns information on the state in which the death occurred, the date of death, the death certificate number, and the extent to which linkage variables (e.g., name, date of birth, and SSN) matched, but not the values of those variables.

Information on cause of death is available for an additional fee through an "NDI Plus" search. The DOHMH purchased an NDI Plus linkage to identify all deaths of people not previously known to be dead in the HIV Surveillance Registry through 2004.

The SSDMF contains the name, SSN, and dates of birth and death for all deaths reported to the SSA, often in connection with a claim for death benefits or termination of benefits the decedent received when alive.[17,18] It has been populated since 1998 and is updated monthly. The state in which death occurred and cause of death are not available from the SSDMF, although the zip code of last residence is available. It may be easier for the user to select the true match from potential matches in the SSDMF, compared with the NDI, because the user is permitted to see the exact values of the linkage variables in the SSDMF, unlike the NDI. A subscription to the database may be purchased for a flat fee; the version used for this analysis was purchased by CDC and provided to the DOHMH at no cost. Before this analysis, the SSDMF had never been used to ascertain deaths of people reported with HIV in NYC.

Population and eligibility criteria

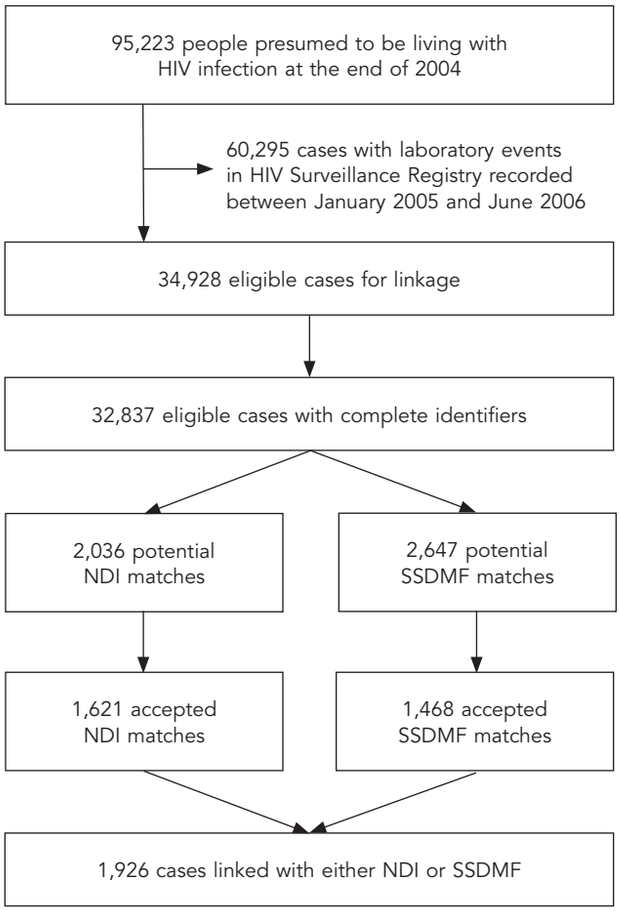
The population for this analysis comprised all known people diagnosed with HIV infection through the end of 2004, reported to the NYC DOHMH by December 31, 2005, and not yet known to have died, based on data in the Surveillance Registry. We excluded people with at least one reportable HIV-related laboratory event between January 1, 2005, and June 30, 2006, in the HIV Surveillance Registry (i.e., a positive HIV Western blot test, a CD4+ lymphocyte test, or an HIV-1 viral load measure) on the premise that such people were in care after 2004 and, therefore, could not have died during the time period of interest (Figure 1). Furthermore, we excluded any case in the HIV Surveillance Registry missing a first name, last name, or date of birth, because we required these identifiers for linkage to minimize the possibility of false matches.

Linkage process

Eligible HIV cases in the HIV Surveillance Registry were independently linked with deaths occurring between January 1, 2000, and December 31, 2004, reported in the NDI and the SSDMF. We selected this time period to include the earliest year in which the HIV Surveillance Registry had not been previously linked with the NDI (2000), and the latest year in which both databases contained data available at the time of the linkage (2004). We conducted the linkages in the second half of 2006, with differing procedures for each data source based on their unique requirements and contents.

NDI linkage. Initial linkages of eligible cases to deaths in the NDI were generated by NCHS staff according to established procedures based on the following linkage variables: first name, last name, date of birth, SSN, gender, race, last known state of residence, and place of birth.[16] While the minimum elements required by NCHS staff are either (*1*) first and last name, and month and year of birth; (*2*) first and last name, and SSN; or (*3*) SSN, date of birth, and gender, we limited records for linkage to those fulfilling the first criterion only. NCHS staff returned potential matches to DOHMH for evaluation. Pairs matching exactly on all linkage variables were accepted immediately, along with those that matched exactly by first name, last name, date of birth, and SSN. For the remaining potential

Figure 1. Flow chart of selection criteria and results of NDI and SSDMF linkages with eligible people in the New York City HIV Surveillance Registry



NDI = National Death Index
SSDMF = Social Security Administration's Death Master File
HIV = human immunodeficiency virus

match pairs, NCHS staff created probabilistic scores to prioritize those requiring additional review to be accepted as true matches.[19]

To minimize clerical review, we assumed that remaining pairs with a score below 28 were false matches. We transferred records with a score of 28 or higher to a secure Microsoft® Access[20] database developed specifically for this activity, and we based decisions regarding true matches on independent evaluations by two reviewers of the fields mentioned, as well as the underlying cause of death and the last date on which HIV-related care was recorded in the HIV Surveillance Registry. A third independent reviewer resolved any discordant decisions.

SSDMF linkage. To link eligible cases with the SSDMF, we used the following linkage variables: first name, last name, date of birth, and SSN, requiring at minimum the first three elements. We included variables in addition to SSN because SSN was not available for 62.6% of eligible HIV Surveillance Registry cases. Cases matching exactly either by first name, last name, and date of birth—or by SSN only—were accepted after review of fields for any obvious discrepancies. In addition, we generated keys based on combinations of first name, last name, and date of birth, or components of each, using SAS® version 9.2.[21] Potential matches based on these keys were reviewed independently in Access by the same two reviewers as in the NDI match, along with other known variables from the HIV Surveillance Registry. As before, a third reviewer resolved any discordant decisions. Evaluation of SSDMF linkage results occurred several months after evaluation of NDI results was already complete. It was unlikely that results of the NDI review influenced decisions on potential SSDMF matches due to the timing of the SSDMF review and the vast number of potential matches involved.

Analysis of linkage results

We determined the number of deaths from 2000 through 2004 identified by either linkage source, and analyzed their distribution by state of death (available from the NDI) or state of residence (from the SSDMF) and by underlying cause of death (from International Classification of Diseases, 10th Revision [ICD-10] codes in the NDI). We calculated the kappa statistic (K) to assess agreement between the NDI and the SSDMF in death ascertainment for cases found to match to the HIV Surveillance Registry. K is a measure of the proportion of agreement between two sources beyond that expected by chance,[22] with a K of 0.41–0.60 representing moderate agreement, 0.61–0.80 representing substantial agreement, and 0.81–1.00 representing almost perfect agreement.[23] We performed subgroup analyses to assess differences within subpopulations defined by demographic and other characteristics. We performed statistical analyses using SAS software.

To quantify the benefit of linking the HIV Surveillance Registry to national death databases beyond previous death ascertainment from other sources, we determined the percentage of all deaths in the Registry occurring from 2000 through 2004 ascertained by each data source: the NYC Vital Statistics Registry, routine medical record review, the NDI, and the SSDMF. Finally, to assess the impact of using national death databases to improve the accuracy of reported HIV prevalence in NYC based on the HIV Surveillance Registry, we counted the total number of living HIV cases before and after the linkages, and calculated the percentage change in population prevalence. We based denominators for prevalence on 2000 U.S. Census data for NYC.

RESULTS

Based on data reported to the NYC HIV Surveillance Registry by December 31, 2005, 95,223 people were reported with HIV infection in NYC and were not yet known to have died by the end of 2004. Of these, 60,295 had at least one HIV-related laboratory event recorded in the HIV Surveillance Registry between January 2005 and June 2006. This provided evidence that they were still alive after 2004; therefore, we excluded them from the linkage (Figure 1). We excluded an additional 2,091 people because they did not have complete identifiers recorded in the HIV Surveillance Registry, leaving 32,837 eligible people available for linkages with the NDI and SSDMF.

NDI staff initially identified 44,142 potential matches of NDI records to 15,104 eligible cases (median = two matches per case; interquartile range = 1–4). Eliminating those with a probabilistic score of <28 reduced the number of eligible cases with potential matches to 2,036. Of these, 671 were exact matches and accepted as true matches without further review. Two staff members independently reviewed the remaining potential matches for 1,365 cases clerically; 950 of these were accepted as true. Because of discordant match decisions, 9.3% of the potential matches required resolution by a third independent reviewer. In total, we accepted 1,621 matches (4.9%) from the NDI (Table 1). Similarly, we initially identified potential matches for 2,647 cases from the SSDMF. Of these, 1,320 were exact matches and accepted as true without further review, while 1,327 were inexact matches and reviewed by independent reviewers. Discordant match decisions necessitated resolution by a third reviewer for

Table 1. Record linkage results for New York City HIV Surveillance Registry cases, 2000–2004 (*n*=1,926)

	Number of cases linked with NDI	*Number of cases that did not link with NDI*	*Total number of cases*
Number of cases linked with SSDMF	1,163	305	1,468
Number of cases that did not link with SSDMF	458	30,911	31,369
Total number of cases	1,621	31,216	32,837

HIV = human immunodeficiency virus

NDI = National Death Index

SSDMF = Social Security Administration's Death Master File

5.3% of the inexact potential matches. After review, 148 inexact matches were accepted as true, resulting in a total of 1,468 matches (4.5%) accepted based on the SSDMF linkage.

In total, we ascertained 1,926 people in the HIV Surveillance Registry as dead from one or both linkages, or 5.9% of eligible cases. Among eligible cases, the percentage found to be dead (Table 2) was notably greater among people diagnosed at older ages, current or former injection drug users, and people born in a U.S. dependency (mainly Puerto Rico) (all $p<0.0001$). Data on the place where death occurred were available for the 1,621 deaths found in the NDI. They revealed that only 13.7% of the deaths occurred in NYC. Another 26.3% occurred in other parts of New York State, and the remaining 60.0% occurred in 46 other states or territories or the District of Columbia, with the South and Northeast regions most heavily represented (Figure 2). The three jurisdictions in which death occurred most commonly outside of New York State were New Jersey (12.0%), Florida (9.8%), and Puerto Rico (6.1%). We found similar results based on the state of residence as found in the SSDMF, but with a greater percentage of deaths in which the place of residence was NYC (17.4%).

NDI data on cause of death showed that HIV disease was the underlying cause of 64.0% of deaths overall and that this percentage had decreased over time, from 66.9% of deaths in 2000 to 61.5% of deaths in 2004. Leading non-HIV-related causes of death included cardiovascular disease and substance abuse (both 7.8%), and non-AIDS-defining cancers (5.3%).

When comparing the results of linkages with the NDI vs. the SSDMF, overall percentage agreement was 97.7%, and K was 0.74 (95% confidence interval [CI] 0.72, 0.76), representing substantial agreement between the sources (Table 2). K was lower in some subpopulations, notably among people with a history of injection drug use (K=0.68, 95% CI 0.65, 0.71) and Hispanic people (K=0.65, 95% CI 0.62, 0.69), and only fell below substantial agreement among people born outside the U.S. (K=0.60, 95% CI 0.52, 0.68).

Of the total 1,926 deaths found by linkage to the NDI or the SSDMF, 458 (23.8%) were found in the NDI but not in the SSDMF, while 305 (15.8%) were found in the SSDMF but not in the NDI. Decedents found only in the SSDMF were more likely to be aged 60 years or older (13.4%) than were those found only in the NDI (6.6%) (Table 2) ($p<0.001$). Had the NDI linkage not been conducted, 71.7% of those identified by the NDI would have been captured by the SSDMF linkage (1,163/1,621 deaths).

Before the linkages to the NDI and SSDMF, 12,219 deaths had been recorded among people reported with HIV infection in NYC from 2000 through 2004; the majority (12,049) were identified through linkages with the NYC Vital Statistics Registry, with the remainder (170) identified through medical record review. Additional deaths identified through linkages with the two national death databases resulted in a total of 14,145 deaths overall, with 13.6% of them attributable to record linkages to national databases (Table 3). Independently, the NDI contributed an additional 11.5%, and the SSDMF contributed an additional 10.4%. Had record linkages with the national databases not been conducted, deaths among certain subpopulations would have been disproportionately underascertained, including among white people (by 21.9%), men who have sex with men (by 21.6%), and people aged 30–39 years (by 16.2%) (all $p<0.0001$).

Retroactively applying results of the linkages with the national databases to the number of living HIV infection cases at the end of 2004 before linkage reduced the reported HIV population prevalence from 1.19% to 1.16%, a decrease of 2.03% (Table 4). The magnitude of the decrease in prevalence was greater within certain subpopulations, including current or former injection drug users (3.48%), people aged 60 years or older (2.92%), and people born in a U.S. dependency (2.76%).

Table 2. Characteristics of people in the NYC HIV Surveillance Registry linked with the NDI or the SSDMF, 2000–2004 (*n*=1,926)

			Database to which registry cases were matched											
			NDI or SSDMF (total number of deaths)		*NDI*		*SSDMF*		*NDI but not SSDMF*		*SSDMF but not NDI*			
Decedent characteristic	*Number of eligible Registry cases*	*Percent found to be dead*	N	*Percent*[a]	N	*Percent*[a]	N	*Percent*[a]	N	*Percent*[a]	N	*Percent*[a]	*K between NDI and SSDMF*	*95% CI for K*
Total	32,837	5.9	1,926	100.0	1,621	100.0	1,468	100.0	458	100.0	305	100.0	0.74	0.72, 0.76
Gender														
Male	23,665	6.2	1,456	75.6	1,215	75.0	1,132	77.1	324	70.7	241	79.0	0.75	0.73, 0.77
Female	9,172	5.1	470	24.4	406	25.0	336	22.9	134	29.3	64	21.0	0.72	0.69, 0.76
Age at death (in years)														
<30	7,071	0.9	64	3.3	52	3.2	43	2.9	21	4.6	12	3.9	0.74	0.69, 0.80
30–39	13,195	3.4	454	23.6	397	24.5	337	23.0	117	25.5	57	18.7	0.74	0.71, 0.76
40–49	8,916	8.6	770	40.0	657	40.5	572	39.0	198	43.2	113	37.0	0.75	0.72, 0.78
50–59	2,812	16.0	444	23.1	362	22.3	352	24.0	92	20.1	82	26.9	0.73	0.68, 0.78
≥60	843	23.0	194	10.1	153	9.4	164	11.2	30	6.6	41	13.4	0.72	0.63, 0.80
Race/ethnicity														
Black	14,335	5.5	791	41.1	665	41.0	602	41.0	189	41.3	126	41.3	0.74	0.71, 0.77
Hispanic	10,556	5.8	609	31.6	496	30.6	419	28.5	190	41.5	113	37.0	0.65	0.62, 0.69
White	7,141	6.9	494	25.6	433	26.7	420	28.6	74	16.2	61	20.0	0.83	0.80, 0.86
Asian/Pacific Islander	373	3.5	13	0.7	9	0.6	12	0.8	1	0.2	4	1.3	0.76	0.55, 0.96
Other/unknown	432	4.4	19	1.0	18	1.1	15	1.0	4	0.9	1	0.3	0.84	0.71, 0.98
Country of birth														
United States	18,456	5.5	1,016	52.8	854	52.7	782	53.3	234	51.1	162	53.1	0.75	0.72, 0.77
U.S. dependency	1,637	8.1	133	6.9	118	7.3	89	6.1	44	9.6	15	4.9	0.70	0.62, 0.77
Foreign country	4,567	3.0	135	7.0	99	6.1	95	6.5	40	8.7	36	11.8	0.60	0.52, 0.68
Unknown	8,177	7.9	642	33.3	550	33.9	502	34.2	140	30.6	92	30.2	0.76	0.74, 0.79
HIV transmission category														
Men who have sex with men	8,456	5.3	447	23.2	388	23.9	387	26.4	60	13.1	59	19.3	0.84	0.81, 0.87
Injection drug use history	8,568	9.1	779	40.4	645	39.8	559	38.1	220	48.0	134	43.9	0.68	0.65, 0.71
Heterosexual contact	5,474	5.0	276	14.3	227	14.0	216	14.7	60	13.1	49	16.1	0.74	0.70, 0.79
Other risk factor	1,054	3.5	37	1.9	33	2.0	29	2.0	8	1.7	4	1.3	0.80	0.69, 0.91
No identified risk factor	9,285	4.2	387	20.1	328	20.2	277	18.9	110	24.0	59	19.3	0.71	0.67, 0.75
U.S. jurisdiction of death (for NDI) or residence at death (for SSDMF)														
New York State	NA	NA	NA	NA	649	40.0	557	37.9	280	61.1	188	61.6	NA	NA
NYC only	NA	NA	NA	NA	222	13.7	256	17.4	141	30.8	175	57.4	NA	NA
Outside NYC	NA	NA	NA	NA	427	26.3	301	20.5	139	30.3	13	4.3	NA	NA
Other jurisdictions or missing	NA	NA	NA	NA	972	60.0	911	62.1	178	38.9	117	38.4	NA	NA

[a]Percentages may not total 100% due to rounding.

NYC = New York City

HIV = human immunodeficiency virus

NDI = National Death Index

SSDMF = Social Security Administration's Death Master File

K = kappa statistic

CI = confidence interval

NA = not available

DISCUSSION

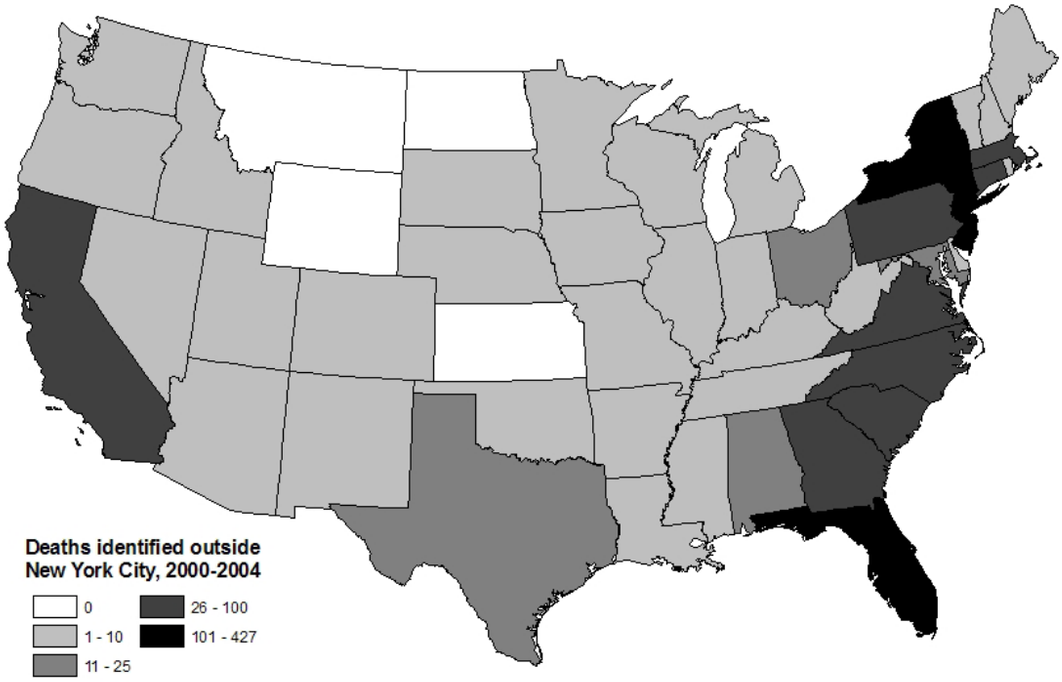
After linkages with five years of mortality data from two national databases, we found 1,926 deaths of people reported in NYC with HIV infection not previously known to be dead, most of which occurred outside NYC. These linkages substantially improved the completeness of death ascertainment and, as a result, the accuracy of living case counts used to enumerate the HIV epidemic in the city and determine federal funding for HIV-related care and treatment services. While agreement between the NDI and the SSDMF was relatively high (K=0.74), sizable percentages of newly ascertained deaths were found in only the NDI (23.8%) or only the SSDMF (15.8%), indicating that both are useful to maximize death ascertainment.

Each database had advantages and disadvantages beyond case yield that should be taken into consideration when deciding whether to use one or both databases. One factor is whether the disease registry needs to obtain the cause of death. NDI Plus data include underlying and multiple causes of death, which may provide valuable information on mortality trends for a particular condition, such as recent decreases in the proportion of deaths caused by HIV disease among people with AIDS.[24] The SSDMF does not provide this information. Another important consideration is cost. As of June 2008, the cost to search the NDI was $0.15 per case per year searched, plus a $350 service charge.[16] The fee per case per year is higher ($0.21) if causes of death are also requested. This amount may make linkages between large registries and the NDI infeasible if sufficient resources are not available.

In contrast, the SSDMF was made available at no cost from CDC to state and selected local HIV surveillance programs. Even if CDC had not provided it for free, the cost of a quarterly subscription to the SSDMF was $6,900 as of June 2008,[25] which may be less expensive than the NDI for large-scale record linkages. For some programs, the cost benefits of the SSDMF may outweigh the benefits of more complete death ascertainment based on the NDI.

Most of the newly found deaths occurred outside NYC, largely reflecting change of residence to outside the city after diagnosis. Without linkage to a national death database, such as the NDI or SSDMF, ascertaining deaths would have depended mainly on the NYC Vital Statistics Registry, which is limited to deaths that

Figure 2. Jurisdiction where death occurred[a] for New York City HIV Surveillance Registry cases linked with the National Death Index, 2000–2004



[a]Not shown: Alaska (0), Hawaii (1–10), Puerto Rico (26–100)

HIV = human immunodeficiency virus

Table 3. Deaths recorded in the New York City HIV Surveillance Registry, by source of first discovery, 2000–2004

Decedent characteristic	*Number of all deaths*	*New York City Vital Statistics Registry (percent)*	*Routine medical record review (percent)*	*NDI or SSDMF (percent)*	*NDI only (percent)*	*SSDMF only (percent)*
Total	14,145	85.2	1.2	13.6	11.5	10.4
Gender						
Male	9,772	83.9	1.2	14.9	12.4	11.6
Female	4,373	88.0	1.2	10.7	9.3	7.7
Age at death (in years)						
<30	453	82.8	3.1	14.1	11.5	9.5
30–39	2,807	82.8	1.0	16.2	14.1	12.0
40–49	5,732	85.4	1.2	13.4	11.5	10.0
50–59	3,670	86.9	1.0	12.1	9.9	9.6
≥60	1,483	85.6	1.3	13.1	10.3	11.1
Race/ethnicity						
Black	7,225	87.8	1.3	10.9	9.2	8.3
Hispanic	4,507	85.4	1.1	13.5	11.0	9.3
White	2,257	77.0	1.2	21.9	19.2	18.6
Asian/Pacific Islander	82	81.7	2.4	15.9	11.0	14.6
Other/unknown	74	73.0	1.4	25.7	24.3	20.3
Country of birth						
United States	9,117	88.0	0.8	11.1	9.4	8.6
U.S. dependency	1,683	91.3	0.8	7.9	7.0	5.3
Foreign country	1,575	90.0	3.8	8.6	6.3	6.0
Unknown	1,770	60.3	1.2	36.3	31.1	28.4
HIV transmission category						
Men who have sex with men	2,072	77.7	0.7	21.6	18.7	18.7
Injection drug use history	6,377	86.8	1.0	12.2	10.1	8.8
Heterosexual contact	2,366	87.6	0.8	11.7	9.6	9.1
Other risk factor	215	80.5	2.3	17.2	15.3	13.5
No identified risk factor	3,115	85.3	2.3	12.4	10.5	8.9

HIV = human immunodeficiency virus

NDI = National Death Index

SSDMF = Social Security Administration's Death Master File

occur in NYC. Despite this, 13.7% of deaths newly ascertained from the NDI occurred in NYC, and 17.4% of deaths from the SSDMF were to city residents. This was unexpected because routine linkages between the Surveillance Registry and the NYC Vital Statistics Registry had already been conducted before this analysis. Further examination revealed that at least 26 of the 175 NYC deaths identified exclusively by the SSDMF had been flagged as possible matches in these prior linkages, but reviewers had not accepted them due to conservative match thresholds or discrepancies in identifiers. That these matches were originally rejected highlights the subjective component of the linkage process. Special software that employs probabilistic techniques to optimize match rates may be useful in this regard.[26,27] Because there will always be some uncertainty in evaluating matches, consistent standards should be used when assessing potential matches to assure reliability and replicability over time.[1]

Of deaths identified through the NDI, 427 (26.3%) occurred in New York State but outside NYC. These deaths could not have been identified through linkage with the NYC Vital Statistics Registry, but might have been identified had we decided to initiate a linkage with the state vital statistics registry. Assuming that such a record linkage could be performed at little to no cost, conducting such a linkage might have decreased both the cost and yield associated with the NDI and SSDMF linkages, as deaths occurring in New York State would have already been identified. However, this is not an option for the majority of disease registries that are already at the state level.

Investigators who analyze disease registry data should be aware of factors specific to certain subpopulations

that can result in differential death ascertainment, and bias prevalence estimates based on the data sources used. For example, we found that had we limited searches to the local vital statistics registry and not performed record linkages with national databases, we would have more often underascertained deaths in certain subpopulations, notably white people and men who have sex with men (by 21.9% and 21.6%, respectively). These findings, which may reflect higher rates of out-migration in these groups, are not inconsequential to prevalence estimates. Other groups, such as older people and injection drug users, have higher mortality rates in general[24,28] and, therefore, will have higher death ascertainment rates regardless of the data source used.

Limitations

An important limitation to our analysis was the lack of a gold standard against which to confirm the accuracy of linkage. False nonmatches may be especially likely for nonwhite and foreign-born decedents, who have been described as having lower reporting of SSNs and a higher likelihood of incorrect name spellings[19]—an assertion that was corroborated in our analysis by the lower agreement between data sources among nonwhite vs. white people, and among foreign-born vs. U.S.-born people. The possibility of false nonmatches was reduced somewhat by the use of phonetic matches of first and last names by NCHS staff for NDI record linkage.[16]

Common names also present challenges when

Table 4. HIV prevalence in New York City before and after National Death Index database linkages, 2004

	Before linkages		*After linkages*		
	PLWHA (n)	*Population prevalence (percent)*	*PLWHA (n)*	*Population prevalence (percent)*	*Percent change after linkages*
Total	95,037[a]	1.19	93,111	1.16	−2.0
Gender					
Male	66,112	1.74	64,656	1.70	−2.2
Female	28,925	0.69	28,455	0.68	−1.6
Age at end of 2004					
<30	8,037	0.23	7,973	0.23	−0.8
30–39	21,728	1.61	21,274	1.58	−2.1
40–49	37,638	3.32	36,868	3.25	−2.0
50–59	20,992	2.47	20,548	2.42	−2.1
≥60	6,642	0.53	6,448	0.51	−2.9
Race/ethnicity					
Black	42,275	2.15	41,484	2.11	−1.9
Hispanic	30,290	1.40	29,681	1.37	−2.0
White	20,350	0.73	19,856	0.71	−2.4
Asian/Pacific Islander	1,072	0.14	1,059	0.14	−1.2
Other/unknown	1,050	0.35	1,031	0.34	−1.8
Country of birth					
United States	52,974	1.03	51,958	1.01	−1.9
U.S. dependency	4,813	1.29	4,680	1.25	−2.8
Foreign country	12,257	0.43	12,122	0.42	−1.1
Unknown	24,993	NA	24,351	NA	−2.6
HIV transmission category					
Men who have sex with men	26,637	NA	26,190	NA	−1.7
Injection drug use history	22,390	NA	21,611	NA	−3.5
Heterosexual contact	17,635	NA	17,359	NA	−1.6
Other risk factor	2,990	NA	2,953	NA	−1.2
No identified risk factor	25,385	NA	24,998	NA	−1.5

[a]Excludes people with missing gender in HIV Surveillance Registry (n=186)

HIV = human immunodeficiency virus

PLWHA = people living with HIV/acquired immunodeficiency syndrome

NA = not available

evaluating potential matches. For example, the lower ascertainment we found among Hispanic people (5.8%) may have been due to a greater abundance of common names in this population, which increased the difficulty in confirming incomplete matches (e.g., with SSN missing) by our staff. This premise is supported by the lower agreement between the two data sources among Hispanic people (K=0.65). Furthermore, the experience of our field surveillance staff suggests that the use of pseudonyms or other identities when seeking HIV-related care is not uncommon, especially among those who have a history of incarceration. The use of pseudonyms in medical environments may occur for various reasons, including fraud, concealment, and changing group membership,[29] and has consequences for surveillance if linkages with other databases are attempted using these false identifiers.

Finally, while our staff are careful to ensure that data are recorded accurately, in practice the sources of registry data vary in quality, given the lack of standardization of medical records systems in the U.S. and inconsistent surveillance practices that may result from incomplete access to medical records or limited use of active surveillance in some facilities. In summary, evaluation of electronic record linkages, especially those with databases such as the NDI that return incomplete identifier information, may be project-specific; therefore, broader or narrower criteria than those used in this analysis may be more appropriate in other contexts.

CONCLUSION

Our analysis is noteworthy for several reasons. It is the first reported comparison between the NDI and the SSDMF for death ascertainment among people with HIV, who differ from the cohort populations previously used to compare the two databases demographically, behaviorally, and clinically. Because of the size and diversity of the HIV-infected population in NYC,[30] we were able to perform a robust assessment of the utility of both databases for people with HIV and report results for population subgroups, which can help other registries to identify particularly challenging groups when conducting similar linkages (e.g., Hispanic or foreign-born people). Finally, this analysis cannot be replicated nationally, because personal identifiers from local jurisdictions are not shared with CDC as part of HIV surveillance.

In conclusion, record linkages with the NDI and the SSDMF are challenging to implement and have their own unique limitations, but result in important improvements in the accuracy of prevalence estimates and the resource allocation that relies on such estimates. Linkages are especially important in jurisdictions such as NYC, with residents who are highly mobile, and in subpopulations, such as Hispanic and foreign-born people, in which death ascertainment may be more challenging. HIV surveillance registries and other disease registries that need accurate living case counts should conduct periodic death linkages using both databases if resources permit. The NDI, despite its costs, is an essential data source if the cause of death is needed to characterize mortality. However, our analysis also supports the use of the SSDMF as an inexpensive means to maximize accuracy of death ascertainment and further improve disease prevalence estimates.

This work was supported by a cooperative agreement between the New York City Department of Health and Mental Hygiene and the Centers for Disease Control and Prevention (CDC) (U62/CCU223595).

The authors thank members of the HIV Epidemiology and Field Services Program, especially Marcia Bryan, Wanda Davis, Selam Seyoum, and Dipal Shah for review of matches; and Sonny Ly and Julie Evans for data management.

An earlier version of this analysis was presented at the 40th Annual Society for Epidemiologic Research Meeting, Boston, Massachusetts, June 19–22, 2007.

The findings and conclusions in this article are those of the authors and do not necessarily represent the views of CDC.

REFERENCES

1. Buehler JW. Surveillance. In: Rothman KJ, Greenland S, editors. Modern epidemiology. 2nd ed. Philadelphia: Lippincott Williams & Wilkins; 1998. p. 435-57.
2. Glynn MK, Lee LM, McKenna MT. The status of national HIV case surveillance, United States 2006. Public Health Rep 2007;122 Suppl 1:S63-71.
3. Henry J. Kaiser Family Foundation. HIV name reporting, May 2007 [cited 2008 Jun 11]. Available from: URL: http://www.statehealthfacts.org/comparetable.jsp?cat=11&ind=559
4. Ryan White Treatment Modernization Act of 2006. Public Law 109-415 [109 U.S.C.] [cited 2008 Jun 11]. Available from: URL: http://hab.hrsa.gov/law/reauth06.htm
5. Hanna DB, Weglein DM, Begier EM. Increase in out-migration in persons reported with AIDS in New York City since the introduction of HAART. Presented at the XVII International AIDS Conference; 2008 Aug 3–8; Mexico City.
6. Palella FJ Jr, Delaney KM, Moorman AC, Loveless MO, Fuhrer J, Satten GA, et al. Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators. N Engl J Med 1998;338:853-60.
7. Electronic record linkage to identify deaths among persons with AIDS—District of Columbia, 2000–2005. MMWR Morb Mortal Wkly Rep 2008;57(23):631-4.
8. Patterson BH, Bilgrad R. Use of the National Death Index in cancer studies. J Natl Cancer Inst 1986;77:877-81.
9. Centers for Disease Control and Prevention (US), Council of State and Territorial Epidemiologists. Technical guidance for HIV/AIDS surveillance programs, volume I: policies and procedures. Atlanta: CDC; 2006.
10. Kraut A, Chan E, Landrigan PJ. The costs of searching for deaths: National Death Index vs. Social Security Administration. Am J Public Health 1992;82:760-1.
11. Sesso HD, Paffenbarger RS, Lee IM. Comparison of National Death Index and World Wide Web death searches. Am J Epidemiol 2000;152:107-11.

12. Lash TL, Silliman RA. A comparison of the National Death Index and Social Security Administration databases to ascertain vital status. Epidemiology 2001;12:259-61.
13. 1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults. MMWR Recomm Rep 1992;41(RR-17):1-19.
14. Guidelines for national human immunodeficiency virus case surveillance, including monitoring for human immunodeficiency virus infection and acquired immunodeficiency syndrome. MMWR Recomm Rep 1999;48(RR-13):1-28.
15. Implementation of named HIV reporting—New York City, 2001. MMWR Morb Mortal Wkly Rep 2004;52(51–52):1248-52.
16. Centers for Disease Control and Prevention (US), National Center for Health Statistics. National Death Index [cited 2008 Jun 16]. Available from: URL: http://www.cdc.gov/nchs/ndi.htm
17. Hill ME, Rosenwaike I. The Social Security Administration's Death Master File: the completeness of death reporting at older ages. Soc Secur Bull 2001–2002;64:45-51.
18. Cowper DC, Kubal JD, Maynard C, Hynes DM. A primer and comparative review of major US mortality databases. Ann Epidemiol 2002;12:462-8.
19. Horm J. Assignment of probabilistic scores to National Death Index record matches. Supplement to the National Death Index user's manual. NDI Plus: coded causes of death. Hyattsville (MD): National Center for Health Statistics; 1999. p. A7-13.
20. Microsoft Corp. Microsoft® Access 2000. Redmond (WA): Microsoft Corp.; 1999.
21. SAS Institute, Inc. SAS®: Version 9.2. Cary (NC): SAS Institute, Inc.; 2008.
22. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther 2005;85:257-68.
23. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
24. Sackoff JE, Hanna DB, Pfeiffer MR, Torian LV. Causes of death among persons with AIDS in the era of highly active antiretroviral therapy: New York City. Ann Intern Med 2006;145:397-406.
25. Department of Commerce (US). Social Security Administration's Death Master File [cited 2008 Jun 30]. Available from: URL: http://www.ntis.gov/products/ssa-quarterly.aspx
26. Jaro MA. Probabilistic linkage of large public health data files. Stat Med 1995;14:491-8.
27. Howe GR. Use of computerized record linkage in cohort studies. Epidemiol Rev 1998;20:112-21.
28. Pfeiffer MR, Hanna DB, Begier EM, Sepkowitz KA, Zimmerman R, Sackoff JE. Excess mortality among injection drug users with AIDS, New York City 1999–2004. Subst Use Misuse. In press.
29. Rendleman N. False names. West J Med 1998;169:318-21.
30. HIV Epidemiology and Field Services Program. Semiannual report vol. 2 no. 2. New York: New York City Department of Health and Mental Hygiene; 2007.