Risk Factor Redistribution of the National HIV/AIDS Surveillance Data: An Alternative Approach

Kathleen McDavid Harrison, PhD, MPH[a]
Tebitha Kajese, MSPH[b]
H. Irene Hall, PhD[a]
Ruiguang Song, PhD[a]

SYNOPSIS

Objective. The purpose of this study was to assess an alternative statistical approach—multiple imputation—to risk factor redistribution in the national human immunodeficiency virus (HIV)/acquired immunodeficiency syndrome (AIDS) surveillance system as a way to adjust for missing risk factor information.

Methods. We used an approximate model incorporating random variation to impute values for missing risk factors for HIV and AIDS cases diagnosed from 2000 to 2004. The process was repeated M times to generate M datasets. We combined results from the datasets to compute an overall multiple imputation estimate and standard error (SE), and then compared results from multiple imputation and from risk factor redistribution. Variables in the imputation models were age at diagnosis, race/ethnicity, type of facility where diagnosis was made, region of residence, national origin, CD-4 T-lymphocyte cell count within six months of diagnosis, and reporting year.

Results. In HIV data, male-to-male sexual contact accounted for 67.3% of cases by risk factor redistribution and 70.4% (SE=0.45) by multiple imputation. Also among males, injection drug use (IDU) accounted for 11.6% and 10.8% (SE=0.34), and high-risk heterosexual contact for 15.1% and 13.0% (SE=0.34) by risk factor redistribution and multiple imputation, respectively. Among females, IDU accounted for 18.2% and 17.9% (SE=0.61), and high-risk heterosexual contact for 80.8% and 80.9% (SE=0.63) by risk factor redistribution and multiple imputation, respectively.

Conclusions. Because multiple imputation produces less biased subgroup estimates and offers objectivity and a semiautomated approach, we suggest consideration of its use in adjusting for missing risk factor information.

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

[b]Business Computer Applications, Inc., Atlanta, GA

Address correspondence to: Kathleen McDavid Harrison, PhD, MPH, Centers for Disease Control and Prevention, MS E-47, 1600 Clifton Rd. NE, Atlanta, GA 30333; tel. 404-639-6034; fax 404-639-2980; e-mail <KMcDavid@cdc.gov>.

Since the early 1980s, recording behavioral risk factors associated with acquired immunodeficiency syndrome (AIDS) and then with human immunodeficiency virus (HIV) infection has been critical in elucidating the infectious nature of the epidemic, identifying areas in which prevention efforts are essential (i.e., screening blood donations), and focusing prevention and treatment programs on the basis of major transmission routes. However, throughout the 1990s, the proportion of cases that were reported to the Centers for Disease Control and Prevention (CDC) without an identified risk factor for HIV infection increased. In 2005, approximately 40% of HIV cases, compared with less than 20% in 1994, were reported to CDC without risk factor information.[1,2]

In the U.S., the legal authority to collect and store information on cases of HIV and AIDS resides with the governments of the 50 states, the District of Columbia (DC), and U.S.-dependent areas. Government agencies voluntarily forward HIV and AIDS surveillance data to CDC after removing personally identifying information, including the patient's name, from the record for each case.

The combination of expansion in reporting volume (a result of integrating HIV with AIDS reporting), reliance on laboratory reports as the initial case notification to health departments, and decreased access to detailed documentation for follow-up of newly reported cases resulted in larger case loads for follow-up by surveillance staff and decreasing success in acquiring the necessary information.[3–5]

The first case report forms, developed by CDC in the early 1980s, were designed to collect clinical, demographic, and risk factor information for each case. Initially, only information about intravenous drug use, blood transfusion, and sexual preference (the term used on the first form) was requested. Today, the standardized risk factors for adults that are collected for public health surveillance purposes are male-to-male sexual contact; injection drug use (IDU); high-risk heterosexual (HRH) contact (contact with a person known to have, or to be at high risk for HIV infection, with high risk based on, for example, a history of male-to-male sexual contact, IDU, or receipt of blood products); and receipt of a blood product, transfusion, or transplant.[5]

Although a person can have multiple risk factors, for the purposes of analysis and presentation in reports, the risk factor information on each surveillance record is summarized according to hierarchical categories. In descending order of priority, these hierarchical categories are:

- Male-to-male sexual contact
- IDU
- Male-to-male sexual contact and IDU
- HRH contact (contact with a person known to have, or to be at high risk for HIV infection)
- Other (much less prevalent and includes hemophilia, blood transfusion, perinatal exposure, and risk factor not reported or not identified)

This hierarchy is based on the probability of transmission per act as well as the prevalence of infection among people to whom these categories apply. For a classification of HRH contact, the case report form must bear an indication of a sex partner with, or at high risk for HIV infection.

In this article, we describe the method currently used to redistribute risk factors when risk factor information is missing from the national HIV/AIDS reporting system (HARS),[6] present and evaluate an alternative method to address the growing proportion of HIV/AIDS cases reported without risk factors,[3,4] and suggest an approach for handling future missing risk factor information in HARS. We focused on the risk factors among adults and adolescents; fortunately, cases of HIV infection in children have become rare in the U.S., and most are attributable to perinatal exposure.[1]

METHODS

During the 1990s, CDC developed a statistical method to address the problem of the increasing proportions of cases of HIV reported without a risk factor.[6] This method, which assigns a risk factor distribution to cases without a reported risk factor, is based on reporting patterns (four to 10 years before the date the dataset was created) among cases that were originally reported without a risk factor, but that were later reclassified as having a known risk factor, which was obtained from follow-up investigations and chart reviews. Reclassified cases are divided into 16 groups representing the cross-classification of four regions (Northeast, Midwest, South, West), two sexes (female, male), and two races (white, other). Proportions of risk factor reclassification are calculated for all transmission categories for each of the 16 combinations of region, sex, and race. These proportions are combined with reporting delay weights and applied to cases for which risk factor information is missing.

Calculations of the proportions of redistributed risk factors are based on two assumptions: (*1*) the distribution of risk factors among cases initially submitted with no reported risk factor (NRR) does not change during the period used in calculating weights, and (*2*) cases reclassified as NRR are representative of all NRR cases.

Both of these assumptions are increasingly unlikely to be valid. The pattern of risk factors has changed since the beginning of the epidemic,[1,7] and reclassified cases usually represent cases for which risk factors are easiest to find (Personal communication, Eve Mokotoff, Michigan Department of Community Health, and Judith Sackoff, New York City Department of Health and Mental Hygiene, June 2005). In addition, a recent reabstraction study found that for males, the current method overestimated the number of cases attributed to male-to-male sex and IDU and underestimated the number of cases attributed to HRH contact; for females, it overestimated IDU and underestimated HRH contact.[8] Until the ascertainment and reporting of HIV risk factors improve significantly, surveillance is likely to rely on statistical approaches to adjust for missing risk factor information.

Missing data is an ongoing problem in routinely collected data or large-scale epidemiologic studies.[9] Some frequently used, but less sound ways of handling missing data are list-wise deletion, pair-wise deletion, and mean substitution.[10–14] More statistically rooted methods of handling missing data are concentrated not on merely replacing a missing value but on attempting, by using available data, to preserve the relationships inherent in the dataset.[10,12–14]

Multiple imputation, the method of choice for large datasets,[15] is one such method. It requires specification of a statistical model and is considered a sound approach.[12,13] Multiple imputation does not attempt to estimate each missing value. Instead of estimating the risk factor distribution probabilities for cases with missing risk factors by the current redistribution approach, the multiple imputation approach draws a random sample of the missing values from its distribution. This process results in valid statistical inferences that properly reflect the uncertainty due to missing values. Instead of filling in a single value for each missing value, multiple imputation[16] replaces each missing value with a set of plausible values that reserve the statistical distribution of the imputed variable and the relationship with other variables in the imputation model. The multiply imputed datasets are then analyzed by using standard procedures for complete data. Results from these analyses are then combined to get the final estimates.

Specifically, multiple imputation follows these steps:

1. Impute missing values by using an approximate model incorporating random variation; repeat M times, generating M datasets.
2. Perform standard statistical analyses on each dataset.
3. Combine results from the datasets to compute overall multiple imputation estimate and SE.

This method maintains the original variability of the missing data by creating imputed values, which are based on variables correlated with the missing data and the reasons the data are missing. Uncertainty is accounted for by generating iterations of the missing data and observing the variability between the imputed datasets.[14]

Assumptions of the multiple imputation method include the following: the data must be missing at random (the probability of being missing depends on observed variables), the model used to generate the imputed values must be "correct" in some sense (i.e., must include all anticipated predictor variables), and the model used in the analysis must be consistent with the model used in the imputation.[15,16]

The use of multiple imputation is desirable in adjusting for missing HIV risk factor information because it produces unbiased parameter estimates, which reflect the uncertainty associated with estimating missing data. In addition, multiple imputation methods are available in easy-to-use software.[17–19] We used SAS® procedure MI[19] with a discriminant function analysis, based on multivariate normal theory. We compared the results from multiple imputation and the results from the risk factor redistribution method currently used by CDC.

In our analysis, we included AIDS data from all 50 states and DC and HIV data from 32 states (as of 2004). All data, after collection by state and local health departments, were reported to CDC without personally identifying information.

AIDS data

We used information in HARS from the 50 states and DC about people whose diagnosis of AIDS had been made from 2000 to 2004 and who had been reported through June 2005 to assess the variables that were missing in ≤20% of cases, those that were thought to be correlated with the lack of reported risk factors, and those that will be used in future analyses of surveillance data. We tested the correlation of covariates with reported risk factor and with the absence of risk factor information by using Cramer's V statistic[20] and *p*-values from Chi-square tests. The variables considered control variables in analyses and the variables with a Cramer's V statistic of approximately ≥0.1 for males and females were retained for further analyses. All of the variables that were correlated with the absence of risk factor information were included in our analysis. Data were imputed 10 times both for males and females, HIV, and AIDS, based on relative efficiency of about 95% or better.

Multiple imputation models were calculated for each combination of males and females, and transmission categories; only the missing values for risk factors were imputed. No interaction terms were included in the models. A sensitivity analysis of case frequency by time (in months) to reclassify a case resulted in our decision to use data from the past five years (sufficient to capture approximately 85% of the cases that were eventually reclassified).

HIV data

In the HIV analysis, we included data on diagnoses made from 2000 to 2004 (reported to CDC through June 2005) from 32 states with name-based HIV reporting. All inclusion criteria and analyses of AIDS data were repeated with HIV data.

RESULTS

The variables retained and used in multiple imputation models to impute values for missing risk factors for HIV and AIDS analyses included age at diagnosis, race/ethnicity, type of facility where diagnosis was made, region of residence, national origin, T-lymphocyte cell count (CD4) within six months of diagnosis (AIDS cases only), and reporting year (Tables 1a and 1b and Tables 2a and 2b).

In the AIDS data, male-to-male sexual contact accounted for 57.8% by risk factor redistribution, compared with 60.6% by multiple imputation (SE=0.30) (Table 3a). Also among males, IDU accounted for 19.5% by risk factor redistribution and 18.2% by multiple imputation (SE=0.26), and HRH contact accounted for 15.5% by risk factor redistribution and 14.2% by multiple imputation (SE=0.22).

In the AIDS data on females, the estimates of cases attributable to IDU were very close: 28.8% by risk factor redistribution and 29.6% by multiple imputation (Table 3b). HRH contact accounted for 68.8% by risk factor redistribution and 68.3% by multiple imputation.

In the HIV data, male-to-male sexual contact accounted for 67.3% by risk factor redistribution and 70.4% by multiple imputation (SE=0.45) (Table 4a). Also among males, IDU accounted for 11.6% by risk factor redistribution and 10.8% by multiple imputation (SE=0.34), and HRH contact accounted for 15.1% by risk factor redistribution and 13.0% by multiple imputation (SE=0.34).

Among females, the distribution of HIV was similar to the distribution of AIDS. IDU accounted for 18.2% by risk factor redistribution and 17.9% by multiple imputation (SE=0.61) (Table 4b). In the other major transmission category for females—HRH contact—estimates were 80.8% by risk factor redistribution and 80.9% (SE=0.63) by multiple imputation.

DISCUSSION

For data on HIV infection and AIDS in males, the multiple imputation estimates for male-to-male sexual contact were slightly higher than the proportions from the risk factor redistribution method and were slightly smaller for HRH contact and IDU. For females, however, the multiple imputation estimates for IDU were very similar to those from risk factor redistribution, and they were slightly higher for IDU. Overall, the differences are not of public health significance. We could not test statistical significance between results from the two methods because there is no simple way to estimate the uncertainty associated with the result derived from the risk factor redistribution method. Whether or not they are statistically significant is not important, as some differences between the two methods are expected. No difference in the overall results does not mean that there is no difference in the results for subpopulations. One advantage of the multiple imputation method is that it provides appropriate (unbiased or less biased) estimates not only for the overall risk factor distribution, but also for the risk factor distribution within each subpopulation group that can be characterized by variables included in the imputation model.

These results for the major transmission categories (four for males and two for females) compare favorably with detailed reviews of the medical records of females in three states during the late 1990s[8] and interviews conducted with females during the mid-1990s.[21] The Enhanced HIV Risk Factor Assessment Project, a review of medical records in three states, concluded that compared with medical record reviews, for females, risk factor redistribution overestimated IDU and underestimated HRH contact; for males, it overestimated male-to-male sexual contact and IDU and underestimated HRH contact.[8]

Given the increase in the proportion of HIV cases that have occurred in females, estimates from multiple imputation appear not only plausible but more realistic than risk factor redistribution, for which data from the past four to 10 years are used. No interviews or medical record reviews have been conducted recently enough to serve as a comparison with our results. The multiple imputation methodology itself, however, is being used for national datasets generated by the National Center for Health Statistics: National Health Interview Survey, State and Local Area Integrated Telephone Survey, National Health and Nutrition Examination

Table 1a. Cramer's V, *p*-values, and percent of missing statistics for AIDS variables of interest from adult male AIDS cases, 2000–2004, in 50 areas and DC

	Correlations of variables associated with an identified risk factor (n=120,692)			*Correlations of variables with missing risk factor (n=149,749)*		
Variable	*Cramer's V*[a]	*P-value*	*Percent missing*	*Cramer's V*[b]	*P-value*	*Percent missing*
Age at diagnosis	0.317	<0.0001	0	0.116	<0.0001	0
Race/ethnicity	0.143	<0.0001	0.44	0.146	<0.0001	0.43
Diagnosis facility type	0.100	<0.0001	11.50	0.185	<0.0001	13.95
Vital status	0.077	<0.0001	0.41	0.054	<0.0001	0.46
Region	0.122	<0.0001	0	0.154	<0.0001	0
U.S. or other country of origin	0.090	<0.0001	12.14	0.150	<0.0001	14.23
Diagnosis facility setting	0.088	<0.0001	11.32	0.147	<0.0001	13.60
Death at diagnosis	0.036	<0.0001	0	0.054	<0.0001	0
First CD4 count within six months of diagnosis	0.098	<0.0001	13.33	0.123	<0.0001	13.88
Metropolitan population size	0.050	<0.0001	0.78	0.055	<0.0001	0.71
Immunologic AIDS at report	0.030	<0.0001	0	−0.001	0.7572	0
Year of report	0.033	<0.0001	0	0.057	<0.0001	0

[a]Cramer's V is a measure of association derived from the Pearson Chi-square test. (Kendall M, Stuart A. The advanced theory of statistics: vol. 2—inference and relationship. New York: Macmillan; 1979. p. 588.)

[b]Variables correlated with MISSING. MISSING variable has two levels (0 = risk factor is not missing; 1 = risk factor is missing).

AIDS = acquired immunodeficiency syndrome

DC = District of Columbia

CD4 = T-lymphocyte cell count

Table 1b. Cramer's V, *p*-values, and percent of missing statistics for potential AIDS variables of interest from adult female AIDS cases, 2000–2004, in 50 areas and DC

	Correlations of variables associated with an identified risk factor (n=37,081)			*Correlations of variables with missing risk factor (n=54,649)*		
Variable	*Cramer's V*[a]	*P-value*	*Percent missing*	*Cramer's V*[b]	*P-value*	*Percent missing*
Age at diagnosis	0.378	<0.0001	0	0.088	<0.0001	0
Race/ethnicity	0.057	<0.0001	0.48	0.068	<0.0001	0.45
Diagnosis facility type	0.063	<0.0001	13.70	0.201	<0.0001	18.44
Vital status	0.079	<0.0001	0.23	0.005	0.4974	0.24
Region	0.077	<0.0001	0	0.126	<0.0001	0
U.S. or other country of origin	0.115	<0.0001	10.62	0.151	<0.0001	13.77
Diagnosis facility setting	0.035	<0.0001	12.78	0.186	<0.0001	17.26
Death at diagnosis	0.033	<0.0001	0	0.039	<0.0001	0
First CD4 count within six months of diagnosis	0.142	<0.0001	13.85	0.139	<0.0001	14.20
Metropolitan population size	0.056	<0.0001	0.44	0.074	<0.0001	0.40
Immunologic AIDS at report	0.048	<0.0001	0	0.024	<0.0001	0
Year of report	0.031	<0.0001	0	0.083	<0.0001	0

[a]Cramer's V is a measure of association derived from the Pearson Chi-square test. (Kendall M, Stuart A. The advanced theory of statistics: vol. 2—inference and relationship. New York: Macmillan; 1979. p. 588.)

[b]Variables correlated with MISSING. MISSING variable has two levels (0 = risk factor is not missing; 1 = risk factor is missing).

AIDS = acquired immunodeficiency syndrome

DC = District of Columbia

CD4 = T-lymphocyte cell count

Table 2a. Cramer's V, *p*-values, and percent of missing statistics for HIV variables of interest from adult male HIV (not AIDS) cases, 2000–2004, in 32 areas with confidential name-based HIV infection reporting

	Correlations of variables associated with an identified risk factor (n=47,261)			*Correlations of variables with missing risk factor (n=62,505)*		
Variable	*Cramer's V*[a]	*P-value*	*Percent missing*	*Cramer's V*[b]	*P-value*	*Percent missing*
Age at diagnosis	0.120	<0.0001	0	0.113	<0.0001	0
Race/ethnicity	0.144	<0.0001	0.47	0.201	<0.0001	0.72
Diagnosis facility type	0.102	<0.0001	3.98	0.105	<0.0001	4.09
Vital status	0.065	<0.0001	0.18	0.059	<0.0001	0.22
Region	0.097	<0.0001	0	0.138	<0.0001	0
U.S. or other country of origin	0.098	<0.0001	16.77	0.145	<0.0001	19.82
Diagnosis facility setting	0.118	<0.0001	3.06	0.042	<0.0001	3.20
Death at diagnosis	0.043	<0.0001	0	0.062	<0.0001	0
First CD4 count within six months of diagnosis	0.072	1.00	97.36	0.074	0.5656	97.56
Metropolitan population size	0.136	<0.0001	2.48	0.093	<0.0001	2.06
Year of report	0.035	<0.0001	0	0.047	<0.0001	0

[a]Cramer's V is a measure of association derived from the Pearson Chi-square test. (Kendall M, Stuart A. The advanced theory of statistics: vol. 2—inference and relationship. New York: Macmillan; 1979. p. 588.)

[b]Variables correlated with MISSING. MISSING variable has two levels (0 = risk factor is not missing; 1 = risk factor is missing).

HIV = human immunodeficiency virus

AIDS = acquired immunodeficiency syndrome

CD4 = T-lymphocyte cell count

Table 2b. Cramer's V, *p*-values, and percent of missing statistics for HIV variables of interest from adult female HIV (not AIDS) cases, 2000–2004, in 32 areas with confidential name-based HIV infection reporting

	Correlations of variables associated with an identified risk factor (n=16,779)			*Correlations of variables with missing risk factor (n=27,063)*		
Variable	*Cramer's V*[a]	*P-value*	*Percent missing*	*Cramer's V*[b]	*P-value*	*Percent missing*
Age at diagnosis	0.111	<0.0001	0	0.070	<0.0001	0
Race/ethnicity	0.095	<0.0001	0.72	0.103	<0.0001	0.81
Diagnosis facility type	0.080	<0.0001	3.73	0.064	<0.0001	3.64
Vital status	0.066	<0.0001	0.13	0.015	0.0591	0.16
Region	0.100	<0.0001	0	0.106	<0.0001	0
U.S. or other country of origin	0.107	<0.0001	14.10	0.114	<0.0001	17.46
Diagnosis facility setting	0.081	<0.0001	2.52	0.038	<0.0001	2.56
Death at diagnosis	0.030	<0.0001	0	0.030	<0.0001	0
First CD4 count within six months of diagnosis	0.092	1.00	97.05	0.092	0.829	97.37
Metropolitan population size	0.091	<0.0001	1.29	0.067	<0.0001	0.86
Year of report	0.033	<0.0001	0	0.113	<0.0001	0

[a]Cramer's V is a measure of association derived from the Pearson Chi-square test. (Kendall M, Stuart A. The advanced theory of statistics: vol. 2—inference and relationship. New York: Macmillan; 1979. p. 588.)

[b]Variables correlated with MISSING. MISSING variable has two levels (0 = risk factor is not missing; 1 = risk factor is missing).

HIV = human immunodeficiency virus

AIDS = acquired immunodeficiency syndrome

CD4 = T-lymphocyte cell count

Table 3a. Estimated number and percentage of AIDS cases in males, by risk factor redistribution and multiple imputation methods, 2004, in 50 states and DC[a]

Transmission category	*Observed*[b]	*Risk factor redistribution*	*Multiple imputation*	*SE*[c]
Male-to-male sexual contact	15,436 (48.3)	18,408 (57.8)	19,364 (60.6)	0.30
Injection drug use	4,266 (13.4)	6,243 (19.5)	5,815 (18.2)	0.26
Male-to-male sexual contact and injection drug use	1,653 (5.2)	2,069 (6.5)	2,038 (6.4)	0.15
Hemophilia/coagulation disorder	70 (0.2)	99 (0.3)	86 (0.3)	0.03
High-risk heterosexual contact[d]	3,202 (10.0)	4,953 (15.5)	4,537 (14.2)	0.22
Receipt of blood transfusion, blood components, or tissue	85 (0.3)	136 (0.4)	124 (0.4)	0.04
Other risk factor or not identified[e]	7,252 (22.7)	55 (0.2)	0 (0)	0.02
Total[f]	31,964 (100.0)	31,964 (100.0)	31,964 (100.0)	

[a]Multiple imputation method imputed missing risk factor values only. Final multiple imputation dataset includes cases reported in a five-year period, 2000–2004.

[b]Cases in people born after 1991 were deleted.

[c]SE is for the percentage estimate of risk factor (in parentheses) based on multiple imputation.

[d]Heterosexual contact with a person known to have, or to be at high risk for HIV infection

[e]Includes cases attributable to perinatal transmission in people aged 13 years or older.

[f]Because of rounding, column percentages may not total 100.

AIDS = acquired immunodeficiency syndrome

DC = District of Columbia

SE = standard error

HIV = human immunodeficiency virus

Table 3b. Estimated number and percentage of AIDS cases in females, by risk factor redistribution and multiple imputation methods, 2004, in 50 states and DC[a]

Transmission category	*Observed*[b]	*Risk factor redistribution*	*Multiple imputation*	*SE*[c]
Injection drug use	2,287 (19.8)	3,324 (28.8)	3,416 (29.6)	0.52
Hemophilia/coagulation disorder	21 (0.2)	51 (0.4)	75 (0.7)	0.16
High-risk heterosexual contact[d]	5,089 (44.1)	7,949 (68.8)	7,891 (68.3)	0.53
Receipt of blood transfusion, blood components, or tissue	105 (0.9)	161 (1.4)	170 (1.5)	0.15
Other risk factor or not identified[e]	4,050 (35.1)	67 (0.6)	0 (0)	0.00
Total[f]	11,552 (100.0)	11,552 (100.0)	11,552 (100.0)	

[a]Multiple imputation method imputed missing risk factor values only. Final multiple imputation dataset includes cases reported in a five-year period, 2000–2004.

[b]Cases in people born after 1991 were deleted.

[c]SE is for the percentage estimate of risk factor (in parentheses) based on multiple imputation.

[d]Heterosexual contact with a person known to have, or to be at high risk for HIV infection

[e]Includes cases attributable to perinatal transmission in people aged 13 years or older.

[f]Because of rounding, column percentages may not total 100.

AIDS = acquired immunodeficiency syndrome

DC = District of Columbia

SE = standard error

HIV = human immunodeficiency virus

Table 4a. Estimated number and percentage of HIV cases in males, by risk factor redistribution and multiple imputation methods, 2004, in 32 states[a]

Transmission category	*Observed[b]*	*Risk factor redistribution*	*Multiple imputation*	*SE[c]*
Male-to-male sexual contact	7,792 (53.1)	9,874 (67.3)	10,340 (70.4)	0.45
Injection drug use	1,059 (7.2)	1,699 (11.6)	1,584 (10.8)	0.34
Male-to-male sexual contact and injection drug use	596 (4.1)	817 (5.6)	777 (5.3)	0.21
Hemophilia/coagulation disorder	13 (0.1)	13 (0.1)	21 (0.1)	0.04
High-risk heterosexual contact[d]	1,227 (8.4)	2,217 (15.1)	1,906 (13.0)	0.34
Receipt of blood transfusion, blood components, or tissue	24 (0.2)	25 (0.2)	53 (0.4)	0.09
Other risk factor or not identified[e]	3,970 (27.0)	35 (0.2)	0 (0)	0.00
Total[f]	14,681 (100.0)	14,681 (100.0)	14,681 (100.0)	

[a]Multiple imputation method imputed missing risk factor values only. Final multiple imputation dataset includes cases reported in a five-year period, 2000–2004.

[b]Cases in people born after 1991 were deleted.

[c]SE is for the percentage estimate of risk factor (in parentheses) based on multiple imputation.

[d]Heterosexual contact with a person known to have, or to be at high risk for HIV infection

[e]Includes cases attributable to perinatal transmission in people aged 13 years or older.

[f]Because of rounding, column percentages may not total 100.

HIV = human immunodeficiency virus

SE = standard error

Table 4b. Estimated number and percentage of HIV cases in females, by risk factor redistribution and multiple imputation methods, 2004, in 32 states[a]

Transmission category	*Observed[b]*	*Risk factor redistribution*	*Multiple imputation*	*SE[c]*
Injection drug use	593 (10.2)	1,063 (18.2)	1,043 (17.9)	0.61
Hemophilia/coagulation disorder	0 (0)	0 (0)	19 (0.3)	0.09
High-risk heterosexual contact[d]	2,503 (42.8)	4,719 (80.8)	4,727 (80.9)	0.63
Receipt of blood transfusion, blood components, or tissue	24 (0.4)	26 (0.4)	53 (0.9)	0.16
Other risk factor or not identified[e]	2,722 (46.6)	35 (0.6)	0 (0)	0.00
Total[f]	5,842 (100.0)	5,842 (100.0)	5,842 (100.0)	

[a]Multiple imputation method imputed missing risk factor values only. Final multiple imputation dataset includes cases reported in a five-year period, 2000–2004.

[b]Cases in people born after 1991 were deleted.

[c]SE is for the percentage estimate of risk factor (in parentheses) based on multiple imputation.

[d]Heterosexual contact with a person known to have, or to be at high risk for HIV infection

[e]Includes cases attributable to perinatal transmission in people aged 13 years or older.

[f]Because of rounding, column percentages may not total 100.

HIV = human immunodeficiency virus

SE = standard error

Survey, and by the Federal Reserve for the Survey of Consumer Finances.

In the future, data from CDC's Medical Monitoring Project (MMP) can be used to evaluate the performance of multiple imputation or other alternatives (methods or classification schemes) to risk factor redistribution. The MMP is a national, population-based surveillance project collecting information on clinical outcomes and behaviors of HIV-infected individuals receiving care in the U.S. In addition, CDC and its state surveillance partners are exploring the addition of a female presumed heterosexual contact category.

Results from multiple imputation and risk factor redistribution methods may differ because more variables are included in multiple imputation analysis (seven for multiple imputation; three for risk factor redistribution). In addition, unlike the risk factor redistribution method, the multiple imputation method takes into account the relationships between those variables and the variable being imputed (risk factor) so that overall variability of the missing data is maintained and parameter estimates are unbiased.

Unlike risk factor redistribution, multiple imputation is not based on assumptions about the data that are no longer valid. In addition, multiple imputation could be automated as part of CDC's annual processing of national HIV/AIDS data, resulting in the use of a documented method that is accepted at the national and state levels. Another advantage of multiple imputation is that the method could be reassessed every three or so years (a less automated assessment involving reassessing the variables and determining the number of imputations) instead of the annual labor-intensive work needed to determine the proportions for risk factor redistribution.

Among the practical considerations in adopting the multiple imputation method are the training needs of CDC staff and state surveillance coordinators, the development of SAS programs for national and state use, and the development of procedures for disseminating information about changes.

Limitations

The most noteworthy limitation of the multiple imputation method is the need for resources to implement the change to a new system. Most of the resources needed would be at the national level. A second limitation is that we were not able to fully assess the missing-at-random assumption. However, as an alternative, we included in the analysis all variables that were collected in the national system and that we knew were of good quality and correlated with missing risk factor information.

CONCLUSION

Even though the overall results of multiple imputation and risk factor redistribution are similar, results for some subgroups may differ statistically. However, multiple imputation produces less biased subgroup estimates because it maintains the statistical relationship between variables, particularly the relationship between the risk factor and the variables determining the subgroups. This advantage, coupled with the objectivity and relatively automated approach that multiple imputation offers, lead us to recommend that the national HIV surveillance program consider adopting the multiple imputation method to adjust for missing (not reported) risk factor information.

The findings and conclusions in this article are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

REFERENCES

1. Centers for Disease Control and Prevention (US). HIV/AIDS surveillance report 2005. Vol. 17. Rev ed. Atlanta: Department of Health and Human Services, CDC (US); 2007. p. 13. Also available from: URL: http://www.cdc.gov/hiv/topics/surveillance/resources/reports/2005report/default.htm [cited 2008 Apr 29].
2. Centers for Disease Control and Prevention (US). HIV/AIDS surveillance report 1995;7(no. 2):1-38.
3. Nakashima AK, Fleming PL. HIV/AIDS surveillance report: in the United States, 1981–2001. J Acquir Immune Defic Syndr 2003;32 Suppl 1:S68-85.
4. Fleming PL, Wortley PM, Karon JM, De Cock KM, Janssen RS. Tracking the HIV epidemic: current issues, future challenges. Am J Public Health 2000;90:1037-41.
5. McDavid K, McKenna MT. HIV/AIDS risk factor ascertainment: a critical challenge. AIDS Patient Care STDS 2006;20:285-92.
6. Green TA. Using surveillance data to monitor trends in the AIDS epidemic. Stat Med 1998;17:143-54.
7. 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.
8. McDavid K, Gerstle JE 3rd, Hammett TA, Ellison DM, Stephens TG, Kirk J. Results of the Expanded HIV Risk Assessment Project (EHRAP). AIDS Care 2006;18:77-81.
9. Arnold AM, Kronmal RA. Multiple imputation of baseline data in the Cardiovascular Health Study. Am J Epidemiol 2003;157:74-84.
10. Little RJA, Schenker N. Missing data. In: Arminger G, Clogg CC, Sobel ME, editors. Handbook of statistical modeling for the social and behavioral sciences. New York: Plenum Press; 1995. p. 39-76.
11. Graham JW, Hofer SM. Multiple imputation in multivariate research. In: Little TD, Schnabel KU, Baumert J, editors. Modeling longitudinal and multiple-group data: practical issues, applied approaches, and specific examples. Hillsdale (NJ): Lawrence Erlbaum Associates Inc.; 2000. p. 201-18.
12. Graham JW, Hofer SM, Donaldson SI, MacKinnon DP, Schafer JL. Analysis with missing data in prevention research. In: Bryant KJ, Windle MT, West SG, editors. The science of prevention: methodological advances from alcohol and substance abuse research. Washington: American Psychological Association; 1997. p. 325-66.
13. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002;7:147-77.
14. Little RJA, Rubin DB. Statistical analysis with missing data. New York: John Wiley & Sons; 1987.
15. Rubin DB. Multiple imputation for nonresponse in surveys. New York: John Wiley & Sons; 1987.

16. Rubin DB. Multiple imputation after 18+ years. Journal of the American Statistical Association 1996;91:473-89.
17. King G, Honaker J, Joseph A, Scheve K. Analyzing incomplete political science data: an alternative algorithm for multiple imputation. Am Poli Sci Rev 2001;95:49-69.
18. Schafer JL. Analysis of incomplete multivariate data. London: Chapman & Hall; 1997.
19. SAS Institute, Inc. SAS/STAT: Version 9.1 for Windows. Cary (NC): SAS Institute, Inc.; 2003.
20. Kendall MG, Stuart A. The advanced theory of statistics: vol. 2—inference and relationship. New York: Macmillan; 1979. p. 588.
21. Lansky A, Fleming PL, Byers RH Jr., Karon JM, Wortley PM. A method for classification of HIV exposure category for women without HIV risk information. MMWR Recomm Rep 2001;50(RR-6):31-40.