Supplementary Online Content


eMethods.
**eMethods**

**Stratified Extrapolation Approach**

Since males have more human immunodeficiency virus (HIV) transmission categories than females, multiple imputation was carried out separately for males and females. Variables associated ($P<.05$ by $\chi^2$ test) with having had a BED HIV 1 capture enzyme immunoassay (BED) test or previous testing status or associated with missing values in these variables were included in the imputation models. The variables included in imputing BED values were race/ethnicity, age at diagnosis, transmission category, facility type where HIV was diagnosed, and having ever tested negative for HIV. The variables included in imputing previous testing status were race/ethnicity, age at diagnosis, transmission category, facility type where HIV was diagnosed, AIDS diagnosis within 6 months after diagnosis, whether the BED result is imputed, and BED result. After imputation, all cases have a BED result and information on previous testing.

To control for heterogeneity in testing frequency, newly diagnosed cases were stratified by sex, race/ethnicity, age group, and transmission category. Age groups were formed based on the age at HIV infection. Since the exact age at HIV infection was unknown, it was estimated based on the HIV status at diagnosis. For individuals with a previous negative test result, the age at HIV infection was assumed to be the age at the midpoint of the interval from the last negative test result to the first positive test result. For individuals with no test prior to HIV diagnosis, we assigned the age at HIV infection either as 8 years younger than the age at HIV diagnosis if AIDS was diagnosed at the time of HIV diagnosis, as 4 years younger than the age at HIV diagnosis if AIDS was not diagnosed and the BED result indicated long-term infection, or as the age at diagnosis if the BED result indicated recent HIV infection. Due to the small numbers of cases, those occurring in the race groups Asian/Pacific Islander and American Indian/Alaska Native were not stratified by other variables, and those occurring in the transmission category men who have sex with men/injection drug use were not stratified by age.

Within each stratum, cases were further divided into 2 subgroups based on previous testing status: repeat testers and first-time testers. Within each subgroup, incidence was
estimated by the number of BED-recent specimens divided by the probability of being classified as BED-recent. Because all persons without AIDS within 6 months after their HIV diagnosis have a BED result in the imputed data, the probability of these persons being classified as BED-recent is the product $p_1 \times p_w$, where $p_1$ is the probability of being tested within 1 year after infection, and $p_w$ is the probability of having a BED test result indicating recent infection if the test is performed within 1 year after infection. The latter probability is approximately equal to the mean window period for the BED testing algorithm (156/365 years; R. H. Byers, PhD, unpublished data, July 2005). The window period is the time from seroconversion to the point at which the individual’s serum, if tested using the BED test, would reach an optical density level predetermined to distinguish recent from long-standing infections. For repeat testers, $p_1$ is estimated based on the time from the last negative test result to the first positive result for each individual (reported or imputed) in the group. For first-time testers, $p_1$ is determined by the testing hazard, which is based on the proportion of individuals with AIDS diagnosed at the time of HIV diagnosis in this group. These estimates are approximately 0.60 (range, 0.41-0.71) for repeat testers and 0.24 (range, 0.13-0.51) for first-time testers.

The standard errors for the incidence estimates (derived using the delta method) incorporate uncertainties associated with imputation, the observed number of BED test results indicating recent infection, estimates of $p_1$ and the mean BED window period, adjustments for reporting delay, risk redistribution, extrapolation to the nation, and the covariance among groups for which estimates were made resulting from the inclusion of $p_w$ in each estimate.

Crude incidence rates per 100 000 population were calculated by sex, race/ethnicity, and age (population denominators were not available by transmission category). Population denominators for rates were based on official postcensus estimates for 2006 from the US Census Bureau and on bridged-race estimates for 2006 obtained from the National Center for Health Statistics. The bridged estimates were based on counts from the 2000 Census and produced under a collaborative agreement with the US Census Bureau. These estimates result from regrouping the 31 race categories used in the 2000 Census (1997 standard of the
Office of Management and Budget) for the classification of data on race/ethnicity to the 4 race categories of the 1977 standard and, therefore, to correspond to the HIV data.

**Extended Back-Calculation Approach**

A $K \times 2$ table ($K$ = number of years) of the estimated number of new diagnoses by calendar year, and disease severity at diagnosis (whether AIDS was diagnosed within the same calendar year as HIV), served as the input data for the back-calculation model. A discrete-time probability model (calendar year) for the observed diagnosis data was based on 3 sets of parameters whose properties are described below. Because surveillance data were incomplete for a variety of reasons, a number of adjustments were necessary. For underreporting of HIV cases, we estimated the number of diagnosed but unreported HIV/not AIDS cases for areas with either AIDS-only surveillance or with incomplete combined HIV/AIDS surveillance. We defined strata based on sex, race/ethnicity, transmission risk group, and year of diagnosis. Within these strata, for the 30 states with mature HIV/AIDS surveillance systems, we computed the proportion of diagnosed cases that were still HIV/not AIDS as of the end of 2006 and adjusted the number of HIV/not AIDS diagnoses in the other states or areas to match these proportions. Adjustments also were made for reporting delay, detection and elimination of duplicate reports, and misclassification of the first diagnosis date; these adjustments were based on information from prior studies.²¹,²⁶

The model parameters for the extended back-calculation approach include (1) the number of infections per year, (2) the AIDS diagnosis (discrete) hazards, and (3) the HIV testing (discrete) hazards. The number of infections per year was estimated subject to constraints with categorical structure. Periods were defined such that the number of infections was forced to be the same for each year within a period. The AIDS diagnosis (discrete) hazards were completely specified, not estimated. These values depend only on time since infection, not on calendar time. The AIDS diagnosis hazard values used here are similar to those described by Aalen²⁷ from a Markov model that included staged decreases of CD4 cell counts, progression to AIDS by occurrence of opportunistic infections, and/or diagnosis by HIV testing. The hazards used in the model described by Aalen were modified to account for the US AIDS case definition, which is based either on the occurrence of
opportunistic infections or on immunologic criteria related to CD4 cell counts. One prominent feature of this set of hazards is the flattening of the curve at times distant from infection. The HIV testing (discrete) hazards were estimated subject to categorical constraints. The testing hazards were assumed to depend only on calendar time and not on time since infection. A categorical structure was imposed; ie, periods were defined such that years within the same period were forced to have the same testing hazard. Note that in this instance (calendar time dependence), to ensure identifiable and stable estimates, the periods defined for the HIV testing parameters cannot be identical or too similar to those specified for the number of infections.

The discrete hazards represent conditional probabilities for the 2 types of diagnosis (disease severity). Due to the discrete time framework, we specified that within the same period, an AIDS diagnosis took precedence over diagnosis by HIV testing. Thus, within each period the undiagnosed individual was at risk first to receive an AIDS diagnosis and only if no AIDS diagnosis occurred was the individual then at risk for being diagnosed by HIV testing.

The expected values of the observed data in any year (ie, the 2 types of diagnoses by time) can be written as a linear function of the incidence in years prior to and including the current year with weights that are a function of the AIDS diagnosis and HIV testing hazard values in the same set of years. We assumed that the diagnosis counts have Poisson distributions with expectations that are linear, as described above.

We used an expectation-maximization algorithm to estimate the unknown parameters in the back-calculation model. After specifying some initial starting values for the unknown parameters, the algorithm alternates between an expectation step, which calculates an “expanded” version of the observed data set that is both consistent with the specified model structure and with current “working” parameter values, and a maximization step that reestimates the parameter values using the observed and the expanded data. In this case, the expanded data set consists of the number of diagnoses by time of infection, type of diagnosis, and time of detection.
Variance estimates for estimated HIV incidence or testing hazard values took into account the variability in the (estimated) diagnosis data that served as input to the back-calculation model as well as the variability arising from the back-calculation model (including the effects of estimating other parameters). Operationally, the overall variability was estimated by a multiple imputation approach that incorporated multiple estimates of relevant values (eg, estimated diagnoses by time and disease severity at HIV diagnosis).