A New HIV Testing History-Based Approach for Estimating the Undiagnosed Fraction of Persons with HIV Infection: Findings Suggest That Few HIV-Infected Men Who Have Sex with Men in King County, WA, U.S.A. Are Undiagnosed

Ian Fellows PhD

Martina Morris PhD

Julia Dombrowski MD, MPH

Susan Buskin PhD

Amy Bennett MPH

Matthew R. Golden MD, MPH

Abstract

Objective(s): To develop a simple back-calculation approach for estimating the undiagnosed fraction of HIV cases based on patient reported and medical record derived HIV testing history, and to use that method to estimate the proportion of HIV-infected men who have sex with men (MSM) who are undiagnosed in a single area.

Design: Back-calculation modeling of public health data on HIV testing history among persons with newly diagnosed HIV infection.

Methods: HIV testing history information for newly diagnosed cases of HIV from 2006-2012 was obtained from three separate reporting mechanisms, one of which was verified though medical records. Two new methods were developed for estimating the number of undiagnosed HIV cases using this data. The first is based on a back-calculation type methodology. The second uses the simplifying assumption of constant incidence, allowing the estimate to be expressed as a simple to implement formula.

Results: The model estimates that 6% of HIV-infected MSM in King County are undiagnosed, with an upper bound estimate of 11%. The undiagnosed percentage of cases varied by race/ethnicity, with an estimated 4.9% of White HIV-infected MSM, 8.6% of African American HIV-infected MSM, and 9.3% of Hispanic HIV-infected MSM being undiagnosed.

Conclusions: A relatively simple back-calculation method may be useful in estimating the proportion of cases of HIV infection that are undiagnosed. Estimates from King County, WA suggest that the undiagnosed fraction of HIV cases among MSM is less than one-third of the national estimate.

Introduction

HIV prevention is increasingly focused on identifying infected persons and ensuring that they initiate and continue treatment[1]. The success of these efforts can be tracked through the HIV care cascade or care continuum[2], and many public health authorities now seek to measure each step in that continuum as a routine epidemiologic monitoring activity [3].

The first step in the care continuum – the proportion of HIV infected persons who have not yet been diagnosed – is perhaps the most challenging to estimate. Directly estimating this number requires HIV testing a representative sample of persons at risk for infection and identifying the fraction of infected persons who are unaware of their status. In the U.S., Centers for Disease Control and Prevention (CDC) investigators have adopted this direct approach as part of the National HIV Behavioral Surveillance system (NHBS) and their experience highlights some of the difficulties involved. The NHBS of men who have sex with men (MSM) relies on venue-based sampling, and the representativeness of the surveyed population is unknown. The system measures awareness of HIV status based on self-report, which is vulnerable to under-reporting by respondents who may be reluctant to reveal their HIV positive status to interviewers[4, 5]. In addition, this approach is costly, labor intensive and not replicable by many local health jurisdictions.

Alternatively, estimates of the undiagnosed fraction can be indirectly obtained by using some form of back-calculation to estimate underlying HIV incidence. Variants of this method are increasingly being used for this purpose [6]. CDC has published back-calculation estimates for the undiagnosed fraction of MSM in the US of 19-26% (the range reflects differences by race/ethnicity) [7]. These estimates are based on an “extended back calculation” methodology and relies on surveillance data on the number of AIDS and HIV diagnoses in each year. The methodology requires a set of reasonably complex assumptions regarding testing frequency and the distributions of time from infection to AIDS diagnosis (see the web appendix to [8]), though the complete mathematical details of the model have not been published. As a result, local health departments currently do not have an accepted method for estimating the undiagnosed fraction of HIV infections in their area.

We present in this paper two new methods for estimating the undiagnosed fraction that are designed for use by local health departments. The first is a relatively simple back-calculation based approach which requires no assumptions about trends in HIV incidence, and the second is an even simpler formula that can be used when HIV incidence counts are stable. Both methods use data on HIV testing history for newly diagnosed cases, which already is, or can be, collected as part of routine public health activities. The calculations do not rely on AIDS diagnoses, so do not require assumptions regarding the distribution of time from infection to AIDS, which is a major potential point of model fragility for previous back-calculation approaches. The testing frequency distribution is estimated from patient reported testing history data, and thus avoids the need to assume a constant testing rate. Calculations are done using a package written in the programming language R, and both the R software and this package are open source and publicly available. We demonstrate the methods by estimating the undiagnosed fraction of HIV cases among MSM in King County, WA, overall and broken down by race.

Methods

The methods developed here rely on the date of a last negative test for all newly diagnosed HIV positive cases to estimate the distribution of time from infection to HIV diagnosis (TID), and then use the TID to estimate the undiagnosed fraction. We start by describing the sources of HIV testing data available in King County. Because approximately 85% of all cases of HIV in King County occur in MSM, and almost all MSM diagnosed with HIV infection in the area have previously tested HIV negative, here we restrict our analysis to MSM.

Sources of HIV testing data: King County, WA

HIV testing data were all collected as part of routine public health activities undertaken by Public Health – Seattle & King County (Public Health). Public Health has three sources of HIV testing history data: The enhanced HIV/AIDS reporting system (eHARS), the CDC treatment and testing history questionnaire (HIS), and data collected through HIV partner services. The eHARS data only include dates of HIV negative tests that have verified documentation ascertained through review of medical records. HIS and partner services testing history data are collected by Public Health Disease Intervention Specialists (DIS) using standardized instruments, and the date of last test may be based on self-report by persons with newly diagnosed HIV infection and/or medical record reviews. Partner services data were only available for the period 2010-2012 and are based on client self-report. eHARS data has the advantage of including only verified information, but in some instances persons with HIV report having recent HIV negative tests that cannot be verified, and eHARS then records older but verifiable HIV test results. As part of our evaluation of these different sources of data, we used Pearson's correlations to assess the agreement between the three data sources for time since last HIV negative test.

Estimating the time of last negative HIV test

This is directly estimated from the testing history data. We used the eHARS date when available, because it was a chart validated (if somewhat conservative) measure. In the absence of eHARS data, we used the most recent date observed in the HIS or partner services. If the day and/or month of the last HIV negative test was missing, we assumed that testing occurred in the middle of the month or on July 1, respectively.

Estimating the possible infection interval

For each individual, the period during which they were possibly infected was defined as the time between subjects' last negative HIV test and their HIV diagnosis, with a maximum assigned value of 18 years. This maximum value is consistent with prior analyses suggesting that 95% of persons will develop AIDS within 18 years of infection [9]. When testing history data are missing, we defined the period of possible infection as their age at time of HIV diagnosis minus 16, the median age of sexual debut in the U.S.; here again we assume that no one is HIV infected for more than 18 years prior to diagnosis.

Estimating the distribution of time from infection to diagnosis (TID)
The methods start by estimating the distribution of time from infection to diagnosis. This requires an assumption about the distribution of infection within the possible infection interval, so we conducted a sensitivity analysis using two different assumptions:

Base Case Estimate: Here we assume that infection occurred at a random point distributed uniformly during the period of possible infection. If some men test for HIV in response to specific risks, their TID will shift toward the end of the possible infection interval (as was observed in [10]), so this estimate is likely to be somewhat conservative.

Upper bound Estimate: Here we assume that HIV acquisition occurred at the beginning of the possible infection interval. This represents the most conservative possible estimate of the time from infection to diagnosis consistent with the data.

Both estimated distributions assume that the distribution of time from infection to diagnosis does not change appreciably from year to year, though testing frequency for individuals may vary. We tested this assumption using a Welsh ANOVA.

Back calculation allowing for time-varying incidence
We use a back-calculation based method to estimate quarterly HIV incidence using observed testing history, with progression from infection to diagnosis defined by either the base case or upper bound TID assumption. Incidence counts are modeled as a (possibly time-varying) Poisson process with rate λ_i per quarter, using the standard convolution equation [11]:

$$E(D_t) = \sum_{i<t} \lambda_i \, f_{i,t-i}$$

where D_t is the number diagnosed with HIV during quarter *t,* and $f_{i,t-i}$ is the probability that an individual infected during quarter *i* is diagnosed $t - i$ quarters later, at quarter *t*. The unknown incidence rates λ_i are traditionally estimated using an expectation-maximization (EM) algorithm [8]. Once the quarterly incidence rates have been estimated, the number undiagnosed at quarter t (U_t) may be estimated as

$$E(U_t) = \sum_{i<t} \lambda_i \left(\frac{1}{2} f_{i,t-i} + \sum_{k>t-i} f_{i,k}\right).$$

Each term in the outer sum represents the expected number infected during quarter i who are diagnosed after quarter t plus half the expected number of cases diagnosed during quarter t.

We made two refinements to the usual EM algorithm for fitting this model. First, since our data do not extend back to the beginning of the HIV epidemic but only back to 2006, we adjusted for our limited surveillance window. Second, we added a quadratic smoothing penalty to the likelihood to stabilize the parameter estimates and enforce the *a priori* assumption that any changes in incidence would be relatively gradual. The details of the fitting algorithm are described in the supplemental appendix.

In contrast to the back calculation methods reviewed in [6], this method does not rely on AIDS diagnoses or other biomarkers, and it does not require assumptions about the rates of testing, since the relevant testing interval is directly estimated from data. This considerably simplifies the back-calculation itself, and reduces the impact of assumptions related to disease severity and progression on the estimates.

Direct calculation assuming constant incidence

If both HIV incidence (λ) and the distribution of TID are constant over time, a much simpler formula for estimating the number of undiagnosed can be used:

$$(1) \qquad E(U) = \lambda \int_0^\infty P(TID > t)dt$$

Under these conditions the expected number of new diagnoses (D) will be equal to the expected incidence, so equation (1) can be estimated by substituting the average number diagnosed in each quarter for λ, and either the base case or upper bound estimate of TID for P.

Subpopulation estimates

Some sub-populations, such as groups defined by race or ethnicity, may have different testing behaviors and risks for infection. We tested for sub-population differences in time from last negative HIV test to diagnosis using Welsh ANOVAs, and estimated the number of undiagnosed persons by applying the methods outlined above to the subset of the data containing only those in the sub-population.

Estimates of the undiagnosed fraction of cases

The above equations provide an estimate of the number of undiagnosed cases of HIV infection. The proportion of cases with undiagnosed HIV infection is the estimated number of persons with undiagnosed HIV divided by the sum of the number persons with diagnosed and undiagnosed HIV infection. We used Public Health HIV surveillance data to estimate the number of persons with diagnosed HIV infection living in the area. Medical providers in Washington State are legally required to report all new cases of HIV infection to Public Health – Seattle & King County, and laboratories are required to report all positive HIV test results, all CD4 lymphocyte counts and all HIV RNA test results. Public health surveillance staff investigate all newly identified cases, and surveillance data integrate both in- and out-migration of persons to the area.

Results

From 2006 through the end of 2012, 1522 MSM were diagnosed with HIV in King County, WA. Multiple sources of testing history data were often available, as shown in Table 1. HIS, HIV partner services and eHARS data on date of last HIV negative test were available on 1080 (71%), 476 (31%) and 382 (25%) men, respectively. Combining data from the three sources, among 1522 men, information on date of last HIV negative test was available for 1233 (81%), 101 (6.6%) men were diagnosed with HIV at their first test, and no data were available on 188 (12.4%). The HIS showed excellent agreement with the eHARS and partner services measures (correlations of 0.76 and 0.85 respectively). Among the 113 subjects with both eHARS and partner services data the agreement was less strong (correlation: 0.37), largely due to a cluster of outliers showing much more recent eHARS test date than the date reported in partner services interview. Given the hierarchical assignment, the last negative date was taken from eHARS for 382 men (31% of those with an observed last test date), from HIS for 787 men (64%), and from partner services for 64 men (5%).

The estimated mean and median period of possible infection were 3.12 years and 1.25 years, respectively, with no statistically significant difference across years (F= 1.64 (6, 523), p > 0.1). Figure 1 shows the estimated TID distributions for the base case (Median=0.5 years) and upper bound (Median=1.3 years).

The back-calculation estimates of HIV incidence yielded almost identical quarterly counts for both the base case and upper bound TID models: 49.7-57.5 and 49.6-56.8 respectively. Figure 2 (upper panel) shows the estimated incidence count curves for each model, along with the observed quarterly diagnosis counts. Both models find a relatively stable incidence over the 2006 to 2012 period, suggesting that the simpler constant incidence model can be used to estimate the undiagnosed fraction for these data.

Figure 2 (lower panel) shows the estimated number of undiagnosed cases of HIV among MSM based on each model. The number remains stable throughout the period of observation, ranging from 333-368 in the base case, and about double that, 662-713 for the upper bound. This base case estimates correspond to about 6% of HIV-infected MSM being undiagnosed; the upper bound to about 11%. Table 1 compares the back calculation based estimates to those generated by applying Equation 1 assuming constant incidence.

The testing history data show significant variation by race/ethnicity in the estimated time from infection to diagnosis. The mean possible infection interval estimates were 2.8, 4.3 and 3.4 years, for Whites, African Americans and Hispanics respectively (F=4.5 (2,199) , p < 0.05). The medians are 1.2, 1.8 and 1.5 respectively. As shown in Figure 3, this contributes to relatively large differences in estimates of the undiagnosed fraction between White MSM, and MSM of color.

Using the base case model and assuming constant incidence, we estimate that 4.9% of White HIV-infected MSM, 8.6% of African American HIV-infected MSM, and 9.3% of Hispanic HIV-infected MSM in King County are undiagnosed (the full table of estimates under each model is provided in the supplemental web appendix). It seems counterintuitive that the undiagnosed fraction would be slightly higher among Hispanics than African Americans when their mean TID is lower. The reason is that shape of the TID for Hispanic MSM is distinctively different: their early diagnoses are later than both Whites and African Americans, while their later diagnoses are more frequent than whites, but less extreme than African Americans. For example, 11% of African Americans had periods greater than 15 years, versus 6% among Hispanics, but 25% of African Americans had tested negative within 6 months of diagnosis, versus 20% for Hispanics.

Discussion

We present a new approach to estimating the proportion of persons with undiagnosed HIV infection based on HIV testing history data. Applying the methods to data from King County, WA, our best estimate suggests that about 6% of HIV-infected MSM are undiagnosed. If we use the most conservative assumptions that are consistent with the observed data, the upper bound overall estimate is 11%. Under both models, the estimated undiagnosed fraction is almost twice as high among African Americans and Hispanics than among Whites.

Our estimates for King County, WA are much lower than the most recent national estimates: using an extended back calculation method, CDC investigators estimated that in 2008 19% of HIV-infected MSM were undiagnosed [7]. The extent to which our findings reflect true differences between King County and the U.S. as a whole, versus differences due to methodology is not clear. Ideally, we would compare our results to estimates generated using CDC's extended back calculation method [8, 12] [7]. However, at present, the details of that method are not available in sufficient detail to undertake such a comparison. However, two other findings suggest that at least some part of the difference may be real. First, based on laboratory reported surveillance data, Public Health currently estimates that 68% of HIV diagnosed persons in the area are virologically suppressed compared to national estimate of 31%[3, 12]. Second, among MSM participants in the 2008 cycle of NHBS, 15% of HIV-infected MSM in Seattle vs. 44% of MSM nationally were reported to be unaware of their HIV status[13]. While the validity of NHBS findings on this subject have been questioned[4], the very large difference observed within NHBS suggests that there is a real difference of some magnitude between King County, WA and the U.S.

Our findings have several important limitations. First, our estimate is based on HIV testing history data with varying levels of reliability. Just over 30% of the cases draw the last negative test date from the eHARS system, and eHARS records the last test date that can be validated against medical records. This places an upper bound on the last negative HIV test date, but it may not be the most recent test, so this may bias the TID toward conservative estimates. For the remainder of the cases, TID is based on self-reported date of last test. The validity of these data are not known; we found good agreement between some measures, but not all; this source of potential error could result in either an over estimate or an underestimate of the undiagnosed fraction. Second, our base case model assumes that individuals are infected uniformly between last negative test and first positive. If some people test in response to a recent risk exposure [10], our base case assumption is also conservative, and the undiagnosed fraction would be lower than what we report here. Finally, over 80% of the cases in our data had some information on the date of last negative test; the method we present here is likely to be much more uncertain, and less useful, in populations for which testing prior to diagnosis is uncommon, or data on HIV testing history are unavailable.

We developed this method in response to a request from our Center for AIDS Research Northwest Regional Public Health Consortium, a group convened to advance the integration of public health practice and university research in our region. It was designed to be relatively simple to implement and understand, and to integrate local data to create locally relevant estimates. Compared to the range of back calculation methods currently used for estimating the undiagnosed fraction [6], our approach depends on fewer, and more straightforward assumptions, and requires simpler data and statistical methods. It does not use data on AIDS diagnoses, so it does not incur the well-known uncertainties associated with assumptions about the distribution of time from infection to AIDS. The testing history data it does use can be routinely collected as part of HIV surveillance and partner services, and allows for these local data, which may vary substantially from national data, to be integrated into the model. And the statistical methods can be implemented using a package written in R, an open source software environment, and the model's uncertainty is expressed as a consequence of model assumptions (base case vs. upper bound), rather than confidence intervals (which would be conditional on the statistical model). .

In conclusion, we developed a new method for calculating the undiagnosed faction of persons living with HIV/AIDS. Applied to King County, WA, these findings suggest that relatively few MSM in our area are undiagnosed. The approach we describe requires additional study and application, but could be useful to public health agencies interested in monitoring their local HIV care continuum.

Figure 1: Distributions estimating and bounding the distribution of time between infection and diagnosis (TID)

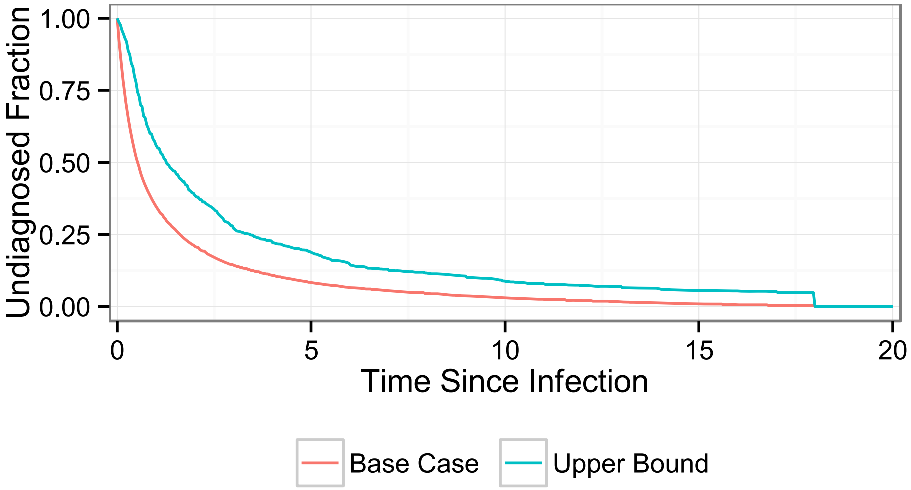

Figure 2: Poisson process estimates of the quarterly HIV+ counts among MSM in King County by quarter using a quadratic smoothing parameter of 0.1. The upper panel displays the incidence estimates along with the observed number of diagnosed cases. The lower panel displays the model-based estimates of the total number of undiagnosed HIV+ cases by quarter.

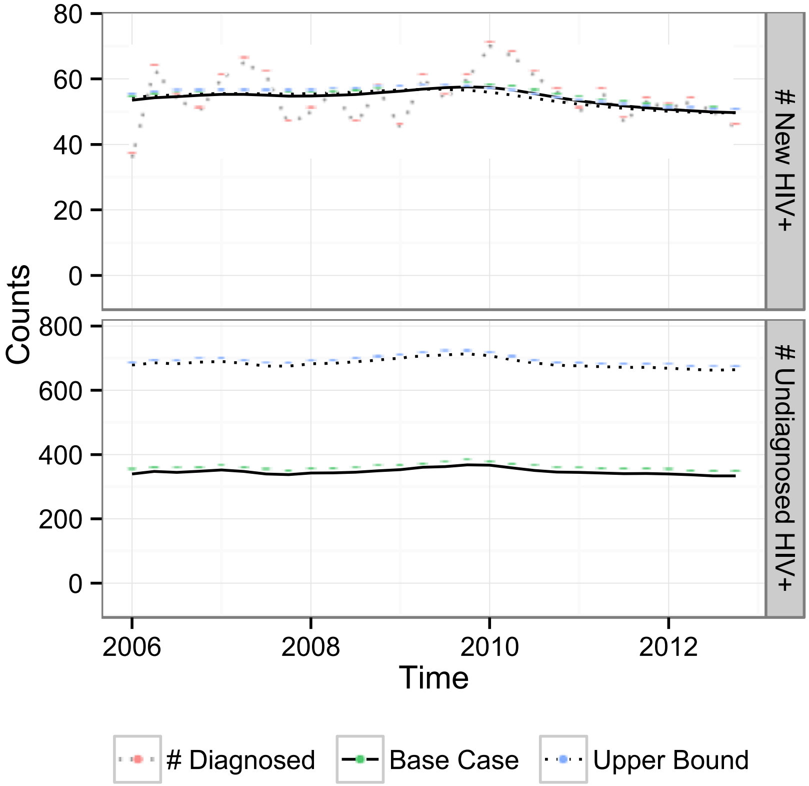

Figure 3. Racial/Ethnic disparities in the undiagnosed fraction.

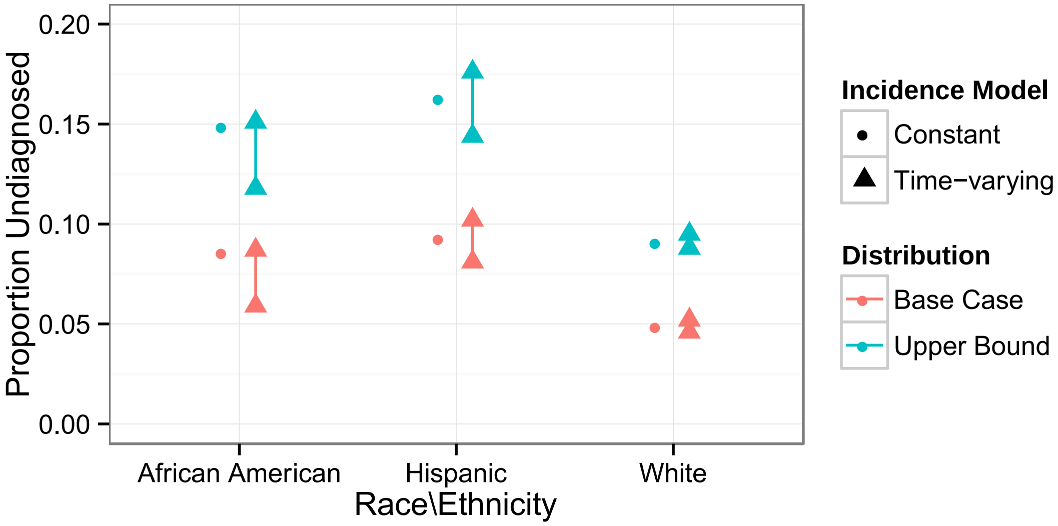

Table 1. HIV testing data sources, overlap, and used in this analysis

		Case also has:†			Used for analysis
		HIS	PS	eHARS	
Case has:	Testing history questionnaire (HIS)	1080	422	346	787
	Partner Services data (PS)††		476	113	64
	Enhanced HIV/AIDS reporting system (eHARS)			382	382
	Total with previous negative test				1233
	Diagnosed at first test				188
	Test history not available				101
	Total diagnosed				1522

† Diagonal elements represent the total number of newly diagnosed cases who had this source of testing information available.

†† Partner services data collection began in 2010, and were available for 69% of the 659 MSM newly diagnosed with HIV after this time.

Table 2: Estimates of the average number of undiagnosed HIV cases among MSM in King County and percentages based on population size estimates.

Distribution Assumption	Incidence Model	Number of MSM with HIV infection†	Number of MSM with undiagnosed HIV infection	Percentage
Base case	Varying	5850-5884	333.5-367.8	5.7%-6.3%
	Constant	5863	346.7	5.9%
Upper bound	Varying	6178-6229	662.2-713.3	10.7%-11.4%
	Constant	6203.2	687.2	11.1%

† Population size estimated as the sum of HIV-infected MSM thought to reside in King County, WA based on HIV surveillance data (n=5516) plus the estimated number of undiagnosed cases.

References

1. Policy, E.H.O.o.N.A., *National HIV/AIDS Strategy for the United States*. 2010.
2. Mugavero, M.J., et al., *The State of Engagement in HIV Care in the United States: From Cascade to Continuum to Control.* Clin Infect Dis, 2013.
3. HIV/AIDS Epidemiology Unit, P.H.S.K.C. and W.S.D.o.H. and the Infectious Disease and Reproductive Health Assessment Unit, *HIV/AIDS Epidemiology Report, Second Half 2012: Volume 81*, P.H.-S.K.C.a.W.S.D.o. Health, Editor. 2013: Seattle, WA.
4. Sanchez, T.H., et al. *Lack of Awareness of HIV Infection: Problems and Solutions with Self-reported HIV Serostatus or Men Who Have Sex with Men*. in *XIX International AIDS Conference*. 2012. Washington DC.
5. Marzinke, M.A., et al., *Nondisclosure of HIV Status in a Clinical Trial Setting: Antiretroviral Drug Screening Can Help Distinguish Between Newly Diagnosed and Previously Diagnosed HIV Infection.* Clinical Infectious Diseases, 2013.
6. Lodwick, R., et al., *HIV in hiding: methods and data requirements for the estimation of the number of people living with undiagnosed HIV Working Group on Estimation of HIV Prevalence in Europe.* Aids, 2011. **25**(8): p. 1017-1023.
7. Chen, M., et al., *Prevalence of undiagnosed HIV infection among persons aged >/=13 years--National HIV Surveillance System, United States, 2005-2008.* MMWR Morb Mortal Wkly Rep, 2012. **61 Suppl**: p. 57-64.
8. Hall, H.I., et al., *Estimation of HIV incidence in the United States.* Jama-Journal of the American Medical Association, 2008. **300**(5): p. 520-529.
9. Lui, K.J., et al., *A model-based approach for estimating the mean incubation period of transfusion-associated acquired-immunodeficiency-syndrome.* Proceedings of the National Academy of Sciences of the United States of America, 1986. **83**(10): p. 3051-3055.
10. Skar, H., J. Albert, and T. Leitner, *Towards estimation of HIV-1 date of infection: a time-continuous IgG-model shows that seroconversion does not occur at the midpoint between negative and positive tests.* PLoS One, 2013. **8**(4): p. e60906.
11. Brookmeyer, R. and M.H. Gail, *A method for obtaining short-term projections and lower bounds on the size of the aids epidemic.* Journal of the American Statistical Association, 1988. **83**(402): p. 301-308.
12. Hall, H.I., et al., *Differences in human immunodeficiency virus care and treatment among subpopulations in the United States.* JAMA Intern Med, 2013. **173**(14): p. 1337-44.
13. Centers for Disease, C. and Prevention, *Prevalence and awareness of HIV infection among men who have sex with men --- 21 cities, United States, 2008.* MMWR Morb Mortal Wkly Rep, 2010. **59**(37): p. 1201-7.