Backcalculation of Undiagnosed HIV in WA State, 2005-2014

Martina Morris and Jeanette Birnbaum

July 10, 2015

1 Background

This report uses the approach developed by Fellows et al[1] to estimate HIV incidence and undiagnosed cases. The method combines data on the number of diagnoses per quarter with information on the distribution of the time between HIV infection and diagnosis, or TID. These two elements are used to "backcalculate" the number of incident cases per quarter that must have occurred to result in the observed number of diagnoses. The number of undiagnosed cases per quarter are those cases who are estimated to have already been infected but not yet diagnosed in a given quarter.

Because TID is not directly observed, the method uses the time between last negative HIV test and diagnosis to approximate TID. The features of this approximation will define the uncertainty in the results.

2 Data

2.1 Description of analytic sample

Data from the advanced HIV/AIDS reporting system (eHARS) and the CDC treatment and testing history questionnaire (HIS) provided records for 26,134 HIV cases in WA state.[2]

2.1.1 Exclusions

Figure 1 diagrams the construction of the analytic sample. We first restricted to cases diagnosed in WA state in the years 2005-2014. We further excluded cases diagnosed at age 16 or younger if their date of last negative test was missing, because the assumptions we use when date of last negative test is missing are not applicable to this age group (details given in Section 3).

The final sample includes 5,176 cases. In the 2014 report there were 4744 cases in the final sample across diagnosis years 2005-2013. Of the additional 447 diagnoses reported in 2014 eligible for this anaylsis, 432 met all our inclusion criteria.

2.1.2 Sample characteristics

Table 1 describes the sample by age, race and mode of transmission. Column % sums to 100% within each characteristic. Six race/ethnicity groups are represented, White, Black, Hispanic, Asian, Native (NHoPI and AI/AN) and Multiracial, and three modes of transmission, MSM (including MSM/IDU), Hetero (including NIR and Female Presumed Hetero) and Blood/Needle (IDU, Ped, Hemo and Transfusion).

For each level of these three characteristics, the table provides the breakdown of responses to the testing history question "Have you ever had a prior negative HIV test?" If a person ever had a negative test prior to diagnosis, they are in the % Yes column. If they never had a negative test prior to diagnosis, they are in the % No column. Those in the % Missing column did not answer the question. These are row %s that sum to 100% across the % Yes, % No and % Missing columns for each row. For example, 57% of MSM have had a negative test, while 9% have not. For 34% of MSM, testing history is unknown. (Note, some %s do not sum to exactly 100% due to rounding error.)

Table 2 further breaks down the sample into racial composition of cases within transmission modes.

[1] Fellows I, Morris M, Dombrowski J, Buskin S, Bennett A, and Golden M. *A new method for estimating the number of undiagnosed HIV infected based on HIV testing history, with an application to men who have sex with men in Seattle/King County, WA*. In press at PLoS One, 2015.

[2] Provided by Jason Carr, Washington State Department of Health, June 2015

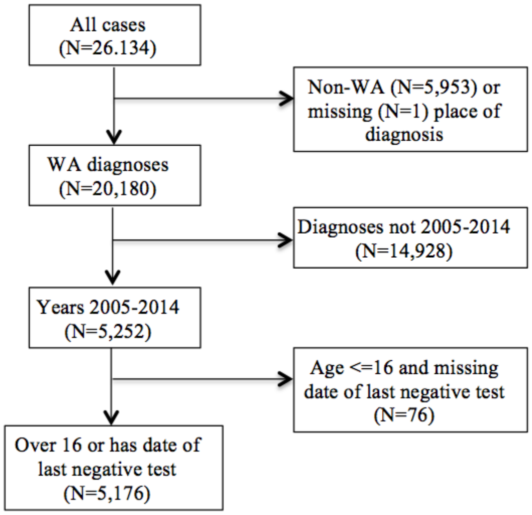
All cases
(N=26.134)
Non-WA (N=5,953) or missing (N=1) place of diagnosis
WA diagnoses
(N=20,180)
Diagnoses not 2005-2014
(N=14,928)
Years 2005-2014
(N=5,252)
Age <=16 and missing date of last negative test
(N=76)
Over 16 or has date of last negative test
(N=5,176)

Figure 1: Construction of analytic sample

Table 1: Composition of analytic sample by age, race and mode of transmission. Column % sums to 100 within each characteristic. Availability of testing history data within each subgroup level is shown as row percents of % Yes, % No, and % Missing)

Characteristic	Subgroup	N	Column %	% Yes	% No	% Missing
All	All	5176	100	46	12	42
Age Group	¡=20	187	4	52	17	32
	21-25	718	14	55	11	33
	26-30	725	14	55	10	35
	31-35	811	16	51	9	40
	36-40	767	15	44	10	46
	41-45	665	13	42	12	46
	46-50	570	11	37	12	51
	51-55	334	6	37	16	46
	56-60	218	4	39	21	40
	61-65	114	2	27	18	54
	66-70	51	1	33	16	51
	71-85	16	0	50	19	31
Race/Ethnicity	White	2994	58	52	9	39
	Black	879	17	37	15	47
	Hisp	792	15	41	13	45
	Asian	256	5	33	21	46
	Native	105	2	31	23	46
	Multi	150	3	51	14	35
Mode of Transmission	MSM	3403	66	57	9	34
	Hetero	1479	29	24	18	58
	Blood/Needle	294	6	32	16	53

Minor assumptions made during data cleaning are given in Section A.1.

2.2 Time trends in diagnoses and testing history

Figure 2 shows a downward trend in quarterly diagnosis counts over time, and Figure 3 shows the overall trend in testing history responses over time. The percent of missing responses appears to have increased in recent years.

Table 2: Composition of racial groups within modes of transmission. Column % sums to 100 within each mode. Availability of testing history data by mode-race subgroup levels is shown as row percents of % Yes, % No, and % Missing

Mode of Transmission	Race/Ethnicity	N	Column %	% Yes	% No	% Missing
MSM	White	2300	44	58	7	35
MSM	Black	307	6	58	11	31
MSM	Hisp	500	10	57	11	32
MSM	Asian	134	3	51	14	35
MSM	Native	51	1	49	20	31
MSM	Multi	111	2	59	14	28
Hetero	White	500	10	30	16	54
Hetero	Black	535	10	26	17	57
Hetero	Hisp	254	5	12	19	69
Hetero	Asian	116	2	14	29	57
Hetero	Native	44	1	11	30	59
Hetero	Multi	30	1	27	17	57
Blood/Needle	White	194	4	34	15	52
Blood/Needle	Black	37	1	32	19	49
Blood/Needle	Hisp	38	1	24	16	61
Blood/Needle	Asian	6	0	0	33	67
Blood/Needle	Native	10	0	30	10	60
Blood/Needle	Multi	9	0	44	11	44

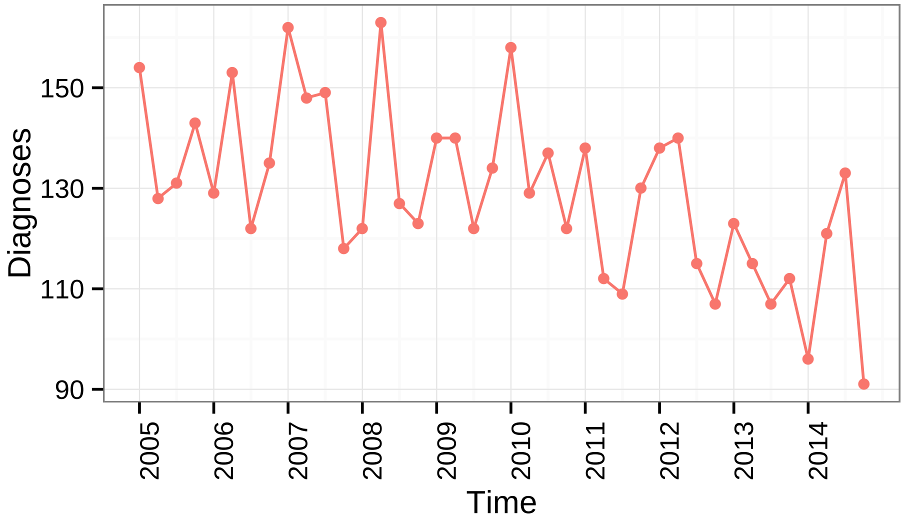

Figure 2: Quarterly diagnosis counts over time

3 Scenarios

We consider two alternative scenarios to approximate the TID from the testing history data. The essential differences are described below, with more details in Section A.2.

1. **Base Case** The probability of acquiring infection is uniformly distributed across the infection period. This assumes testing is not driven by risk exposure, so is likely to be conservative (i.e., overestimate the time spent undiagnosed).

2. **Upper Bound** All infections occur immediately after the last negative test. This is an extremely conservative assumption that represents the maximum possible amount of time people could have been infected but undiagnosed.

In both scenarios, cases who reported "No" to ever having a negative test are also assumed to have a last

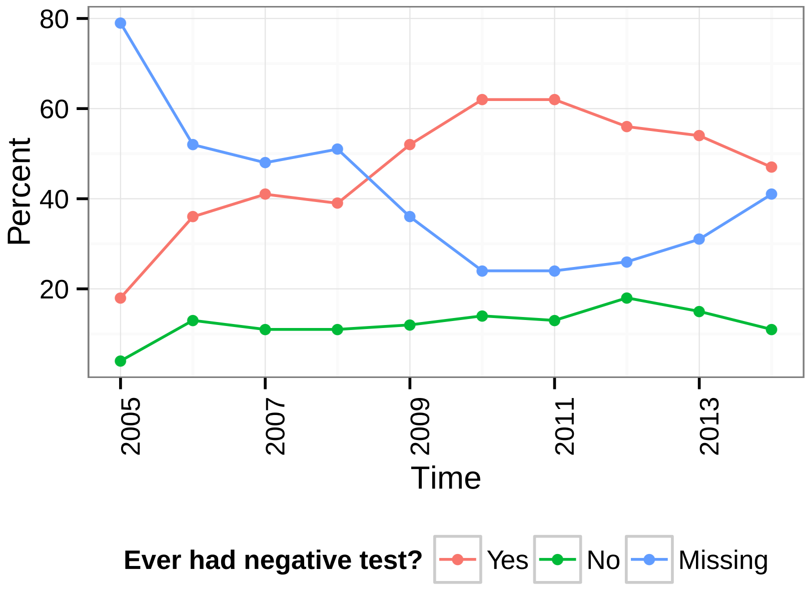

Figure 3: Testing history responses over time (y-axis is in %)

negative test either 18 years prior to diagnosis or at age 16, whichever is more recent (see Section A.2 for more details).

4 Results

4.1 Time from infection to diagnosis (TID)

Figure 4 shows the estimated distribution of TID in the analytic sample for the two scenarios. When the upper bound assumption is made, the proportion of undiagnosed cases at shorter times since infection increases. The artifical spike in the probability of diagnosis/drop in the undiagnosed fraction at 18 years is a result of the assumption that all cases are diagnosed within 18 years.

4.2 Incidence and undiagnosed cases

We use observed quarterly diagnoses with each the three TID scenarios shown in Figure 4 to perform the backcalculation for each scenario. The estimated incidence and undiagnosed counts for each scenario are shown as quarterly counts in Figure 5 and summarized over all quarters in Table 3.

Table 3: Observed diagnoses and estimated quarterly incidence and undiagnosed counts over 2005-2014 in WA state

Diagnoses/Case	Estimate	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
# Diagnosed	Diagnoses	91	120	129	129	140	163
Base Case	Incidence	108	115	126	124	134	138
Base Case	Undiagnosed Cases	1236	1303	1401	1371	1435	1461
Upper Bound	Incidence	105	109	121	120	130	135
Upper Bound	Undiagnosed Cases	2473	2575	2739	2704	2818	2870

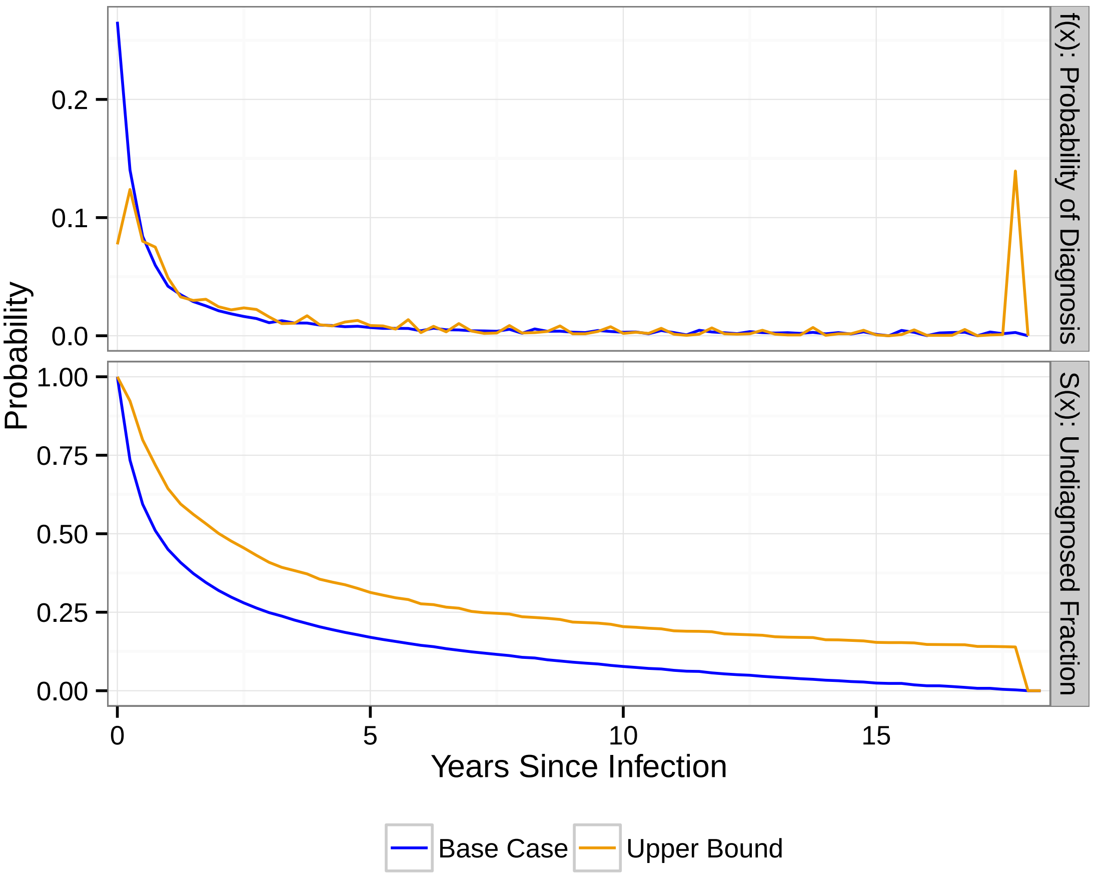

Figure 4: Time from infection to diagnosis (TID) under the three scenarios

A Assumptions

A.1 Assumptions for missing or inconsistent data

The following assumptions were made during data cleaning:

Note: the analysis assumes that that there are a negligible number of cases whose HIV/AIDS diagnosis is never captured by eHARS.

A.2 Assumptions for TID

As described in Section 3, we construct two scenarios for TID that use different assumptions for the time of infection within the window between last negative test and diagnosis.

Time of infection within the window between negative test and diagnosis There are two ways we can assign the precise time of infection within the possible infection window. The first is to assume that infections are uniformly distributed within the window, i.e. there is equal probability of infection at each time point within the window. The second is a worst case assumption, that infection occurred immediately after the negative test.

Assumptions for all scenarios We additionally make four assumptions in both scenarios.

- Those who repond “No” to the question “Ever had a negative test?” have a date of last negative test imputed using the minimum of 18 years and age-16 approach described above. Since these cases

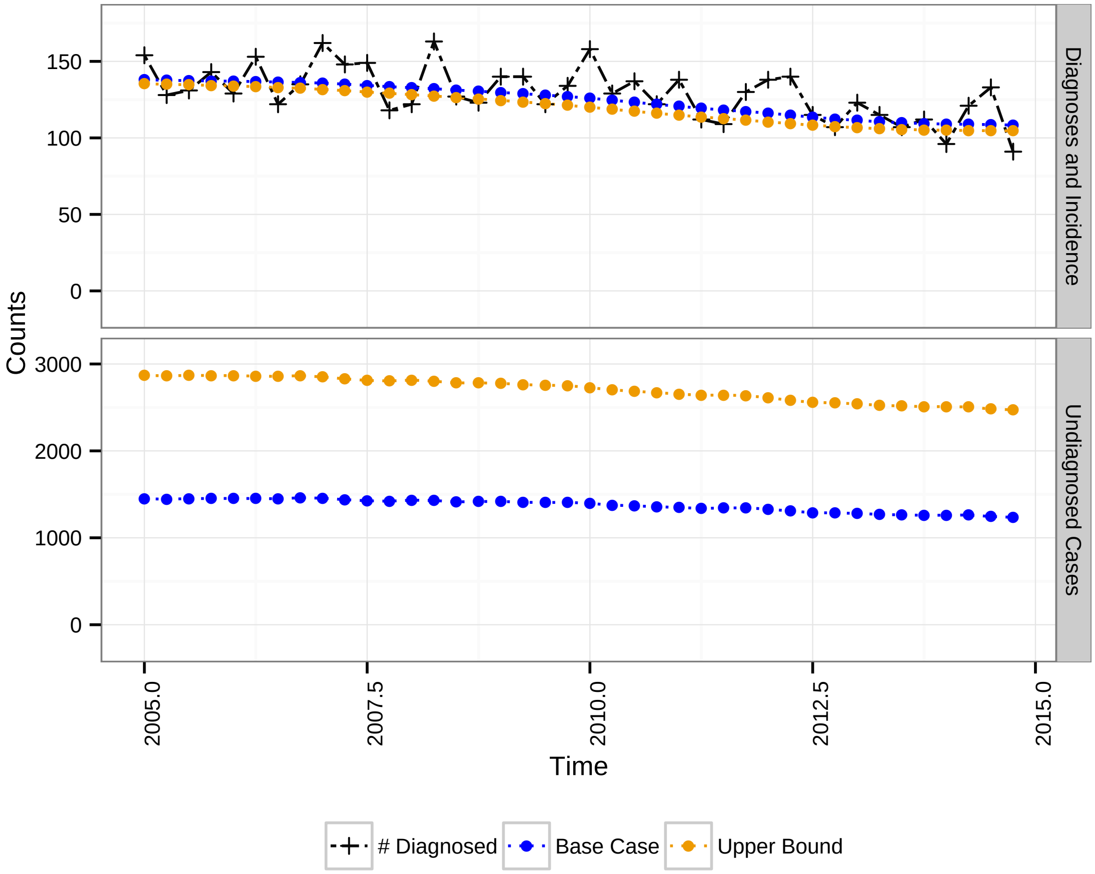

Figure 5: Observed diagnoses and estimated quarterly and undiagnosed counts over 2005-2014 in WA state

confirmed never having a negative test, we use a worst case testing history to bound their infection window.

- Dates of last negative test occurring more than 18 years prior to diagnosis are re-set to 18 years prior to diagnosis, to reflect a more likely maximum window in which infection could occur.

- We assume that the TID distribution does not change over time. In order to have enough cases to stably estimate the TID, we pool testing history data over all years. The time trends in the results are thus driven by the time trends in diagnosis counts.

- We assume that cases whose date of last negative test is not known are exclude them when TID is computed, which assumes that their data are missing at random, e.g. they are well-represented by those cases whose data is not missing. This is reasonable only if the cases who do have a date of last negative test are representative of those who do not. As we further develop our method, we will explore ways to account for non-random missingness in the testing history responses.

Table 4: Estimated true prevalence and the undiagnosed fraction over 2005-2014 in WA state

Year	Diagnoses/Case	Estimate	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2010.0	PLWHA	PLWHA				11739.0		
2010.0	Base Case	Undiagnosed Cases	1358.0	1365.0	1372.0	1374.0	1381.0	1395.0
2010.0	Base Case	True Prevalence	13097.0	13104.0	13111.0	13113.0	13120.0	13134.0
2010.0	Base Case	Undiagnosed Fraction (%)	10.4	10.4	10.5	10.5	10.5	10.6
2010.0	Upper Bound	Undiagnosed Cases	2671.0	2682.0	2694.0	2696.0	2708.0	2727.0
2010.0	Upper Bound	True Prevalence	14410.0	14421.0	14433.0	14435.0	14447.0	14466.0
2010.0	Upper Bound	Undiagnosed Fraction (%)	18.5	18.6	18.7	18.7	18.7	18.9
2011.0	PLWHA	PLWHA				11745.0		
2011.0	Base Case	Undiagnosed Cases	1338.0	1342.0	1344.0	1344.0	1345.0	1349.0
2011.0	Base Case	True Prevalence	13083.0	13087.0	13089.0	13089.0	13090.0	13094.0
2011.0	Base Case	Undiagnosed Fraction (%)	10.2	10.3	10.3	10.3	10.3	10.3
2011.0	Upper Bound	Undiagnosed Cases	2636.0	2639.0	2641.0	2643.0	2645.0	2654.0
2011.0	Upper Bound	True Prevalence	14381.0	14384.0	14386.0	14388.0	14390.0	14399.0
2011.0	Upper Bound	Undiagnosed Fraction (%)	18.3	18.3	18.4	18.4	18.4	18.4
2012.0	PLWHA	PLWHA				11900.0		
2012.0	Base Case	Undiagnosed Cases	1286.0	1289.0	1299.0	1303.0	1313.0	1330.0
2012.0	Base Case	True Prevalence	13186.0	13189.0	13199.0	13203.0	13213.0	13230.0
2012.0	Base Case	Undiagnosed Fraction (%)	9.8	9.8	9.8	9.9	9.9	10.1
2012.0	Upper Bound	Undiagnosed Cases	2552.0	2557.0	2570.0	2576.0	2589.0	2613.0
2012.0	Upper Bound	True Prevalence	14452.0	14457.0	14470.0	14476.0	14489.0	14513.0
2012.0	Upper Bound	Undiagnosed Fraction (%)	17.7	17.7	17.8	17.8	17.9	18.0
2013.0	PLWHA	PLWHA				12280.0		
2013.0	Base Case	Undiagnosed Cases	1260.0	1262.0	1266.0	1268.0	1272.0	1281.0
2013.0	Base Case	True Prevalence	13540.0	13542.0	13546.0	13548.0	13552.0	13561.0
2013.0	Base Case	Undiagnosed Fraction (%)	9.3	9.3	9.3	9.4	9.4	9.4
2013.0	Upper Bound	Undiagnosed Cases	2510.0	2515.0	2521.0	2523.0	2529.0	2541.0
2013.0	Upper Bound	True Prevalence	14790.0	14795.0	14801.0	14803.0	14809.0	14821.0
2013.0	Upper Bound	Undiagnosed Fraction (%)	17.0	17.0	17.0	17.0	17.1	17.1
2014.0	PLWHA	PLWHA				12691.0		
2014.0	Base Case	Undiagnosed Cases	1236.0	1243.0	1253.0	1251.0	1261.0	1262.0
2014.0	Base Case	True Prevalence	13927.0	13934.0	13944.0	13942.0	13952.0	13953.0
2014.0	Base Case	Undiagnosed Fraction (%)	8.9	8.9	9.0	9.0	9.0	9.0
2014.0	Upper Bound	Undiagnosed Cases	2473.0	2480.0	2494.0	2492.0	2507.0	2509.0
2014.0	Upper Bound	True Prevalence	15164.0	15171.0	15185.0	15183.0	15198.0	15200.0
2014.0	Upper Bound	Undiagnosed Fraction (%)	16.3	16.3	16.4	16.4	16.5	16.5

Table 5: Assumptions for missing or inconsistent data

Issue	Assumption	Cases Affected
Year of diagnosis is recorded but quarter is not	Quarter is randomly assigned	9
Case responded “No” or missing to “Ever had negative test?” but has a date of last negative test	Change response to “Yes”	20
Case responded “Yes” to “Ever had negative test?” but has no date of last negative test	Change response to “No”	76
Case responded “Yes” to “Ever had negative test?” but the time between last negative test and diagnosis is recorded as 0	Change response and time to missing	29