Adding Stage of Infection to HIV Back-Calculation in WA State, 2005-2014

Martina Morris and Jeanette Birnbaum

January 14, 2016

1 Overview

This report contains the initial results comparing the standard testing history model results for WA State undiagnosed estimates to an extended model incorporating stage of infection data.

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.[1]

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.

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.

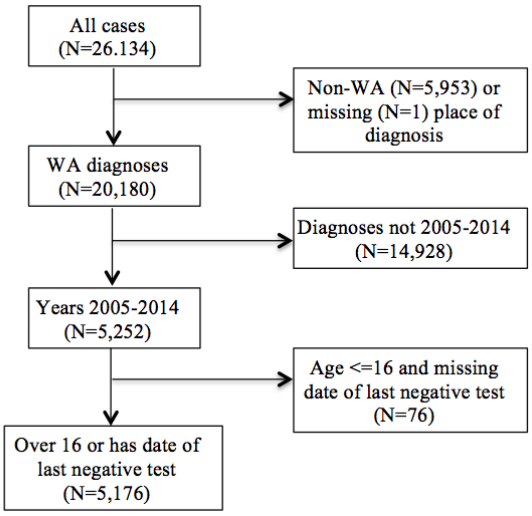

Figure 1: Construction of analytic sample

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

2.1.2 Sample characteristics

This section focuses on the new characteristics we are working with: BED result and dual diagnosis.

Table 1: Cross-tabulation of BED and dual diagnoses

	AIDS w/in 1yr of HIV	no AIDS or not w/in 1yr of HIV
BED -	382	1029
BED +	101	779
BED Miss	1238	1647

Table 2: Composition of analytic sample by BED and dual diagnosis status. 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)

stageGroup	N	Column %	% Yes	% No	% Missing
All	5176	100	46	12	42
BED+DD+	101	2	43	18	40
BED+DD-	779	15	75	6	19
BED-DD+	382	7	29	21	50
BED-DD-	1029	20	58	14	28
BEDmDD+	1238	24	33	18	49
BEDmDD-	1647	32	39	7	54

The presence of an LNT ranges pretty dramatically across groups, from 29 to 75%. Missingness also varies substantially, from 19 to 54%.

What are the implications of these associations, for example the fact that LNTs are much more common among the BED+DD- than the BEDm? Just that the BED information won't be as useful as it could be if it wasn't correlated with missing/no LNT?

2.2 Stage of infection impact on infection window

The purpose of this section is to understand how many individuals will have their individual probabilities of infection altered due to a BED result and/or dual diagnosis, and to what extent.

We have 6 stage categories, 3 BED categories x 2 dual diagnosis subgroups. The table in Figure 2 is our original plan for the Base Case for each subgroup.

BED Result	Dual Diagnosis	Status of LNT Data: LNT Date Known	No Previous Test
+	Yes	$x_i^* = \min(x_i, x_{BED})$ $p(i,t)^* = Cp(i,t)p(AIDS)$	$x_i^* = x_{BED}$ $p(i,t)^* = Cp(i,t)p(AIDS)$
	No	$x_i^* = \min(x_i, x_{BED})$	$x_i^* = x_{BED}$
-	Yes	If $x_i > x_{BED}$, $p(i,t) = p_{BED\text{-}}(i,t)$ $p(i,t)^* = Cp(i,t)p(AIDS)$	$x_i = \min(age-16, 18)$ $p(i,t) = p_{BED\text{-}}(i,t)$ $p(i,t)^* = Cp(i,t)p(AIDS)$
	No	If $x_i > x_{BED}$, $p(i,t) = p_{BED\text{-}}(i,t)$	$x_i = \min(age-16, 18)$ $p(i,t) = p_{BED\text{-}}(i,t)$
Missing	Yes	$p(i,t)^* = Cp(i,t)p(AIDS)$	$p(i,t)^* = p(AIDS)$
	No	No change: $p(i,t) = 1/x_i$	No change: $p(i,t) = 1/x_i$ and $x_i = \min(age-16, 18)$

Table 2. Impact of stage of infection data on the Base Case probability model of time from infection to diagnosis. All stages of infection have a modified infection window x_i^* and/or modified probability model of time from infection to diagnosis $p(i,t)^*$ except when BED status is missing and there is no dual diagnosis (final row). For the BED+, the modified x_i^* has a maximum value of the BED window x_{BED}=162 days. Dual diagnoses have a $p(i,t)$ modified by *p(AIDS)*, the AIDS incubation distribution (Figure 2), with a scalar *C* to constrain cumulative probability of infection over the window to 1. For the BED- with $x_i > x_{BED}$, their $p(i,t) = p_{BED\text{-}}(i,t) = I(t > x_{BED})*(1/(x_i - x_{BED}))$, i.e. all probability falls between x_{BED} and x_i. When LNT data are missing, they are considered missing at random conditional on the stage of infection.

Figure 2: Plan for the extended model, for each of the 6 stages of infection

The following subsections detail the changes to the Base Case for each subgroup.

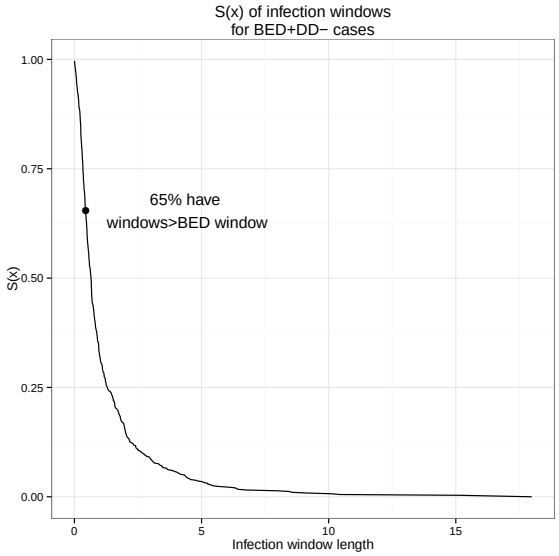

Figure 3: Infection windows of the BED+DD- who have observed windows

2.2.1 BED+, DD+

These 101 (2% of the sample) cases are highly likely to be false BED+, so let's treat them as BEDm (missing), DD+.

2.2.2 BED+, DD-

These 779 (15% of the sample) need their infection window modified to the BED window if it is longer than the BED window or they have no LNT.

We know that 6% of these 779 cases have no LNT and will get a LNT of the BED window (Table 2). Another 75% of the cases have an infection window.

Figure 3 shows that 65% of the observed windows will be shortened by using the BED window as the max.

2.2.3 BED-, DD+

All of the 382 cases (7% of sample) will have their probabilities of infection distributed according to the AIDS incubation distribution rather than a uniform distribution.

The CDC model uses a probability distribution for annual diagnosis of AIDS that they say is "derived from the AIDS incubation distribution" and is gamma(shape=2, scale=4). For our PLoS One paper, we sourced a reference that estimated a Weibull(shape=2.516, scale=1/0.086). The two curves are compared in Figure 4. We'll have to look at the references more closely to understand the differences.

Regarding the infection window, those with no LNT will get the usual assumption of $min(age - 16, 18)$. But for those assumed windows as well as the observed windows, those longer than the BED window will have their probability distributed outside the window, that is, zero probability until 162 days (Figure 5).

2.2.4 BED-, DD-

These 1,029 cases (20% of sample) will have their probabilities of infection distributed outside of the BED window if their windows are longer than 162 days, as with the BED-DD+ (Figure 6).

2.2.5 BEDm, DD+

The 1,238 cases (24% of sample) who are BEDmDD+ will either have their probability of infection distributed using the AIDS incubation distribution across their infection window (33%) or, if they have no LNT, across the full length of the incubation window (14%).

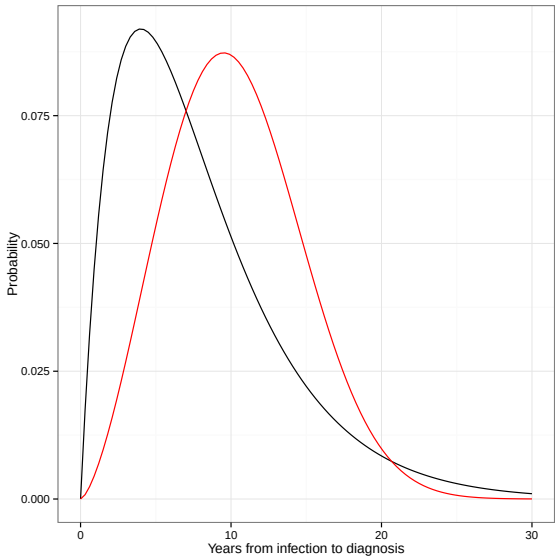

Figure 4: Probability of AIDS diagnosis by years since infection for gamma(2,4) and weibull(2.516, 1/0.086) - weibull in red

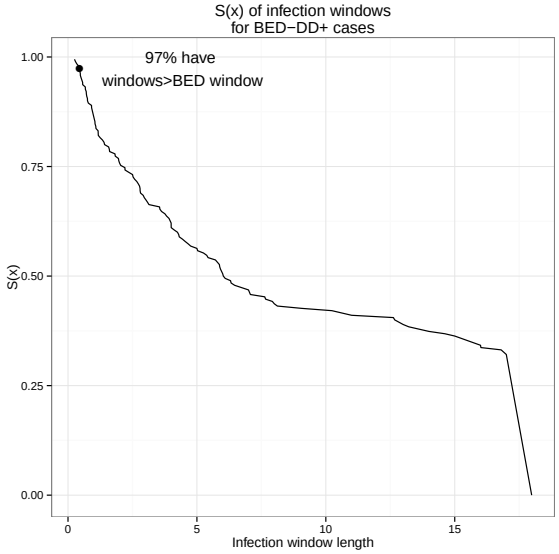

Figure 5: Infection windows of the BED-DD+, either observed or assumed as min(age-16,18) for those with no LNT

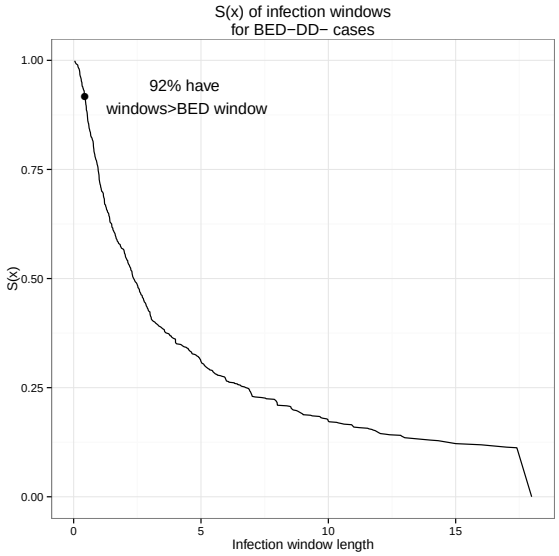

Figure 6: Infection windows of the BED-DD-, either observed or assumed as min(age-16,18) for those with no LNT

2.2.6 BEDm, DD-

These 1,647 cases (32% of sample) have no change to their infection window or probability of infection compared to the original Base Case.

2.2.7 Impact on Upper Bound

There is no impact of dual diagnosis on the upper bound case, since all probability is assumed to be immediately after the LNT. For the BED+DD-, the LNT will be modified to min(LNT, BED window=162 days).

3 Methodological Notes

1. The results in the next section are an abbreviated form of the changes described in the previous section. Only the BED-related modifications were made to the infection windows. The full plan for the Base Case involves altering the uniform probability distribution to the AIDS incubation curve, truncated to the infection window. This is not yet implemented.

2. We decided to treat missing testing histories as missing conditional on stage of infection. My current method for executing this assumption is to run the model on each subgroup. This approach computes the TID separately for each group and applies that TID to estimate incidence only on the diagnosis counts from that subgroup.

3. This is conceptually reasonable but poses sample size problems. The BED+DD+ and BED-DD+ groups were both too small to estimate quarterly incidence. Hence the decision to roll BED+DD+ in with BEDmDD+ and combine BED-DD+ with BED-DD-. I doubt this decision affected the total results much, given the relatively small sizes of those groups.

4. We will need to investigate the sample size limits of the method–they are currently unknown. When quarterly diagnoses are very small, we may want to consider increasing the time step to half or full years.

5. I also want to highlight that the way the method currently works is that the TID is generated from all the data used in the estimation. So here we use WA State data for 2005-2014.

6. A small percentage of BED- cases have LNTs that are shorter than 162 days, indicating false negatives or reporting error. (Figures 5 and 6). This leads to the next point:

7. I think we should work towards a way of incorporating BED results that better reflects the fact that it is intended for population-level interpretation, since the false positive and false negative rates are high but approximately equal. Something where we use the proportions of BED+/- cases rather than look at the individual BED results. However...

8. BED result is correlated with LNT presence/absence and length, so we'll have to think about that too.

4 Aggregating from six to four stage subgroups

We have to aggregate the two smallest BED-DD groups in order to have sufficient sample size. This is a short-term solution. Ultimately we should explore the sample size limits of the method, the impact of using a longer time-step than quarter-year, and ways that the TID can fairly reflect missing data without needing to stratify the estimation.

5 Results

5.1 Time from infection to diagnosis (TID)

Figure 7 shows, for each original stage subgroup, the estimated distribution of TID in the analytic sample for the Base Case under the original method and the Base Case under the extended method. Figure 8 shows the same information for the 4 final stage subgroups used in the analysis. Figure 9

5.2 Unstratified, Without-Stage Results

The estimated incidence and undiagnosed counts for each scenario are shown as quarterly counts in Figure 10 and summarized over all quarters in Table 3. These results are not stratified by any group, although we do have a version of the total results that reflects stratification by MSM and non-MSM.

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

Table 4: Estimated true prevalence and the undiagnosed fraction in WA state, limited to just 2014

Year	Diagnoses/Case	Estimate	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
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

5.3 Stratified, Without- and With-Stage Results

When we run the model allowing stage (so far, just BED) to impact the TID, we also stratify by stage subgroups in order for the missing testing histories to be missing conditional on stage subgroup.

Quarterly incidence and undiagnosed counts are plotted in Figure 11. The summary results over 2005-2015 and for 2014 alone are given in Table 5. The mean without- and with-stage undiagnosed estimates for those two time periods are compared in Table 6. Table 7 shows the 2014 undiagnosed fraction results as well.

Regarding the impact on uncertainty, from Table 6 we can additionally calculate that the undiagnosed range in 2014 was 1,263-2,467 and adding stage decreased that to 1,205-2,354, which amounts to a difference of 55 cases. From Table 7 that amounts to a decrease from 7.2% to 6.9% for the range of the mean undiagnosed fraction.

Table 5: Observed diagnoses and estimated quarterly incidence and undiagnosed counts over 2005-2014 and just 2014 in WA state, using stage-subgroup strata

Year	Case	Estimate	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2005-2014	Base Case	Undiagnosed Cases	1254	1371	1460	1417	1473	1489
2005-2014	Base Case using Stage	Undiagnosed Cases	1197	1315	1449	1388	1462	1487
2005-2014	Upper Bound	Undiagnosed Cases	2449	2625	2859	2768	2915	2939
2005-2014	Upper Bound using Stage	Undiagnosed Cases	2336	2481	2737	2658	2838	2849
2014	Base Case	Undiagnosed Cases	1254	1261	1267	1263	1268	1268
2014	Base Case using Stage	Undiagnosed Cases	1197	1202	1208	1205	1209	1211
2014	Upper Bound	Undiagnosed Cases	2449	2461	2474	2467	2476	2478
2014	Upper Bound using Stage	Undiagnosed Cases	2336	2348	2360	2354	2363	2367

Table 6: Impact of using BED result to modify the TID on mean undiagnosed estimates

Year	Case	With Stage	Without Stage	Difference	Percent Change
2005-2014	Base Case	1388.0	1417.0	-29.0	-2.0
2005-2014	Upper Bound	2658.0	2768.0	-110.0	-4.0
2014	Base Case	1205.0	1263.0	-58.0	-5.0
2014	Upper Bound	2354.0	2467.0	-113.0	-5.0

Table 7: Estimated true prevalence and the undiagnosed fraction for 2014 in WA state, using stage-subgroup strata

Year	Diagnoses/Case	Estimate	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2014.0	PLWHA	PLWHA				12691.0		
2014.0	Base Case	Undiagnosed Cases	1254.0	1261.0	1267.0	1263.0	1268.0	1268.0
2014.0	Base Case using Stage	Undiagnosed Cases	1197.0	1202.0	1208.0	1205.0	1209.0	1211.0
2014.0	Upper Bound	Undiagnosed Cases	2449.0	2461.0	2474.0	2467.0	2476.0	2478.0
2014.0	Upper Bound using Stage	Undiagnosed Cases	2336.0	2348.0	2360.0	2354.0	2363.0	2367.0
2014.0	Base Case	True Prevalence	13945.0	13952.0	13958.0	13954.0	13959.0	13959.0
2014.0	Base Case using Stage	True Prevalence	13888.0	13893.0	13899.0	13896.0	13900.0	13902.0
2014.0	Upper Bound	True Prevalence	15140.0	15152.0	15165.0	15158.0	15167.0	15169.0
2014.0	Upper Bound using Stage	True Prevalence	15027.0	15039.0	15051.0	15045.0	15054.0	15058.0
2014.0	Base Case	Undiagnosed Fraction (%)	9.0	9.0	9.1	9.1	9.1	9.1
2014.0	Base Case using Stage	Undiagnosed Fraction (%)	8.6	8.7	8.7	8.7	8.7	8.7
2014.0	Upper Bound	Undiagnosed Fraction (%)	16.2	16.2	16.3	16.3	16.3	16.3
2014.0	Upper Bound using Stage	Undiagnosed Fraction (%)	15.5	15.6	15.7	15.6	15.7	15.7

6 Conclusion

It's my feeling that we should explore a more population-based approach to incorporating the BED results and one that doesn't involve stratifying the model runs by subgroup. We should also consider introducing BED and DD in a stepwise fashion to understand their relative contributions, once the AIDS incubation distribution is incorporated.

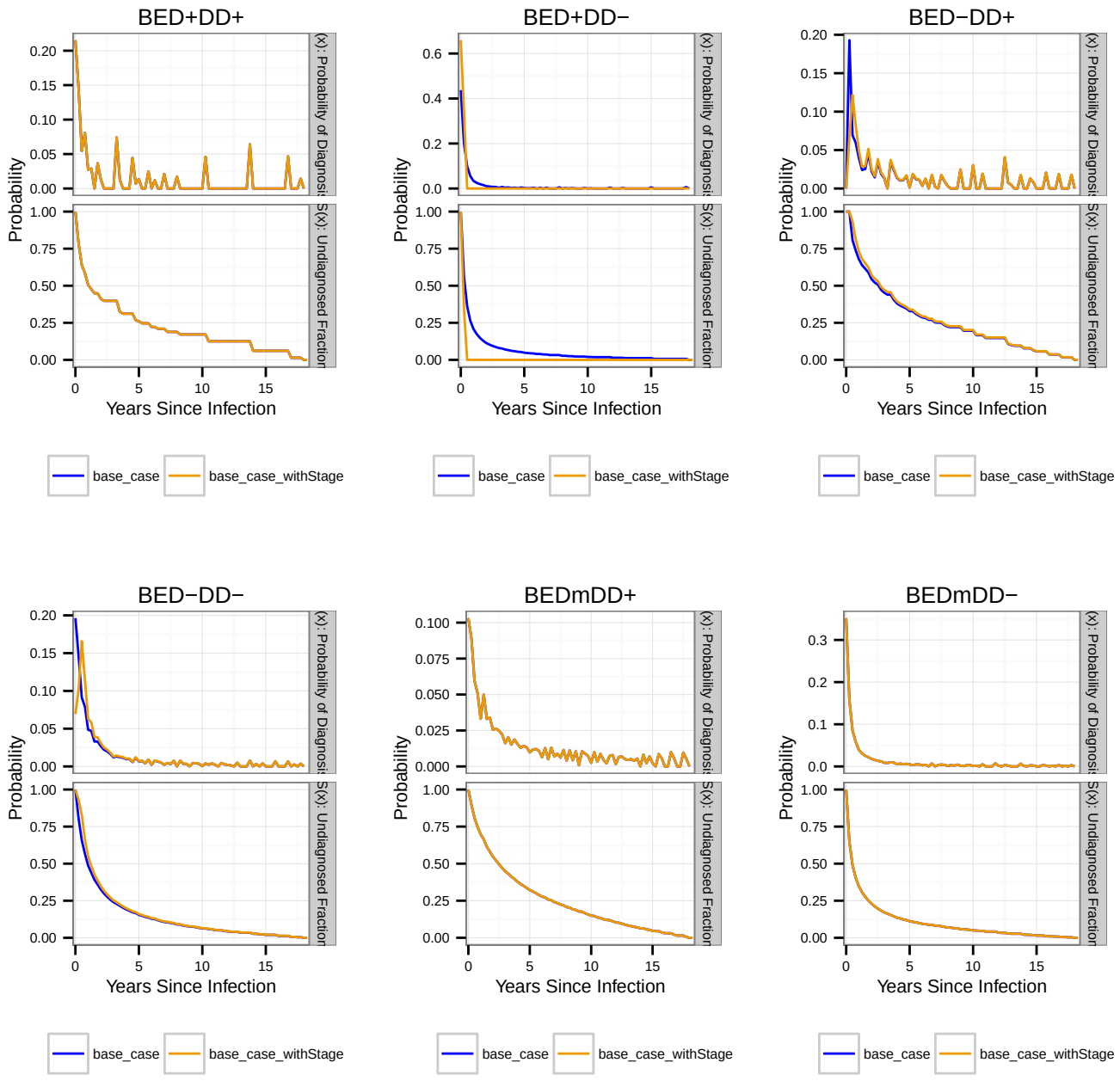

Figure 7: Time from infection to diagnosis (TID) for base case without and with stage, 6 groups

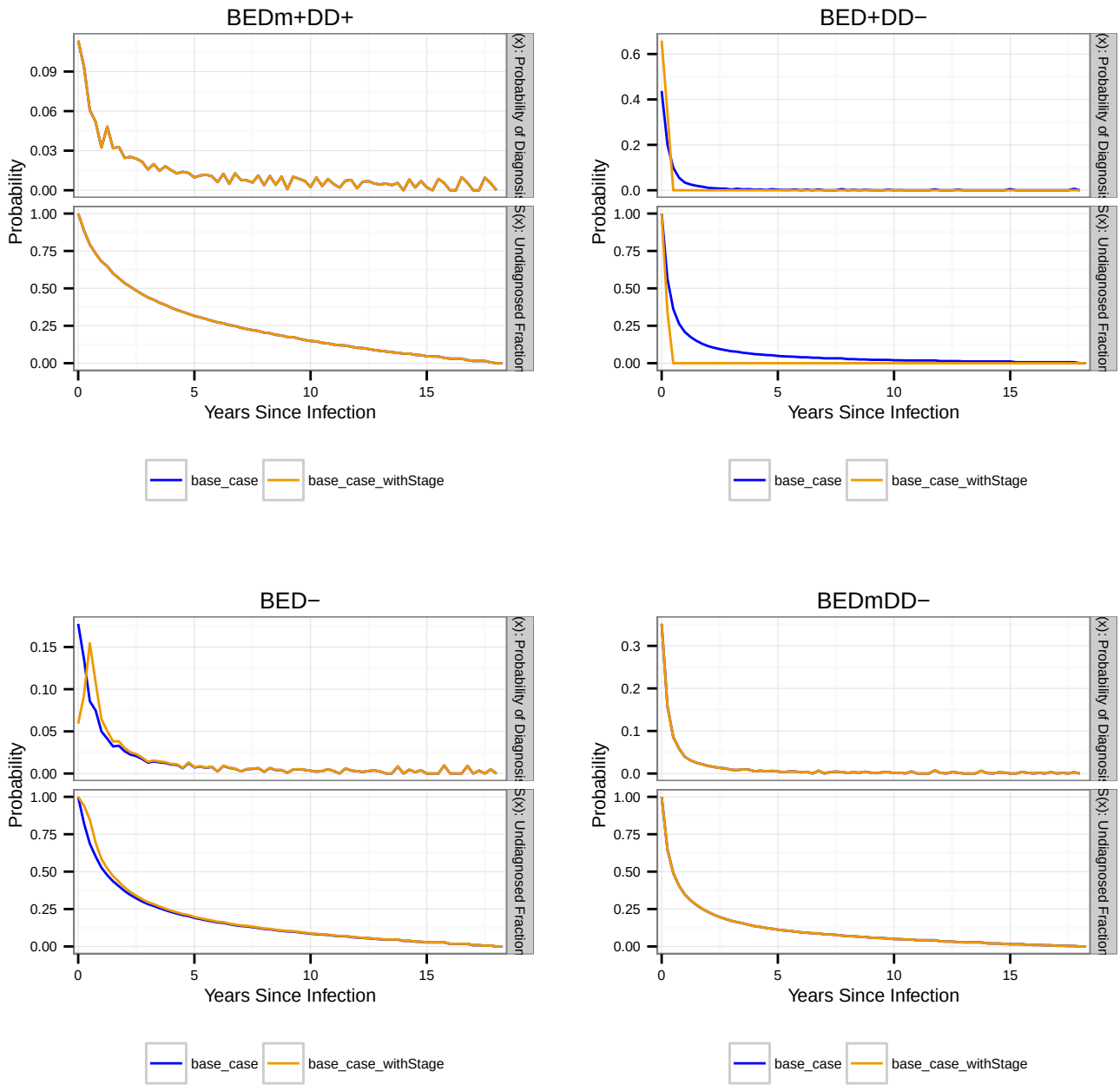

Figure 8: Time from infection to diagnosis (TID) for base case without and with stage, 4 groups

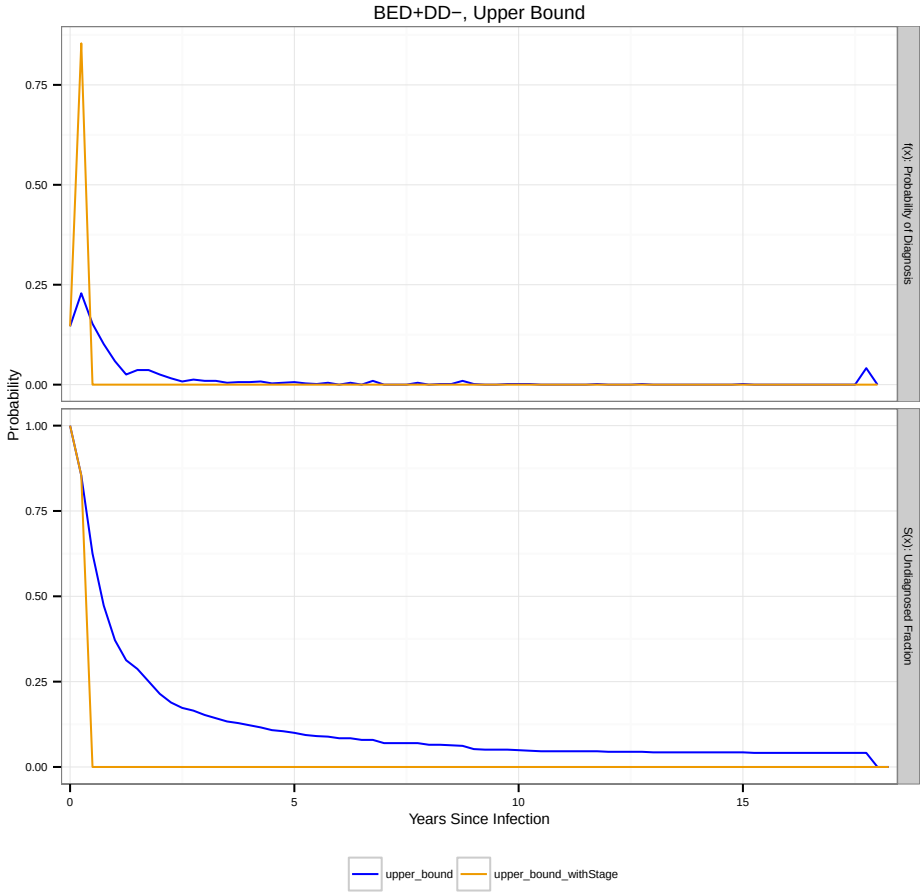

Figure 9: Time from infection to diagnosis (TID) for upper bound without and with stage, BED+DD-

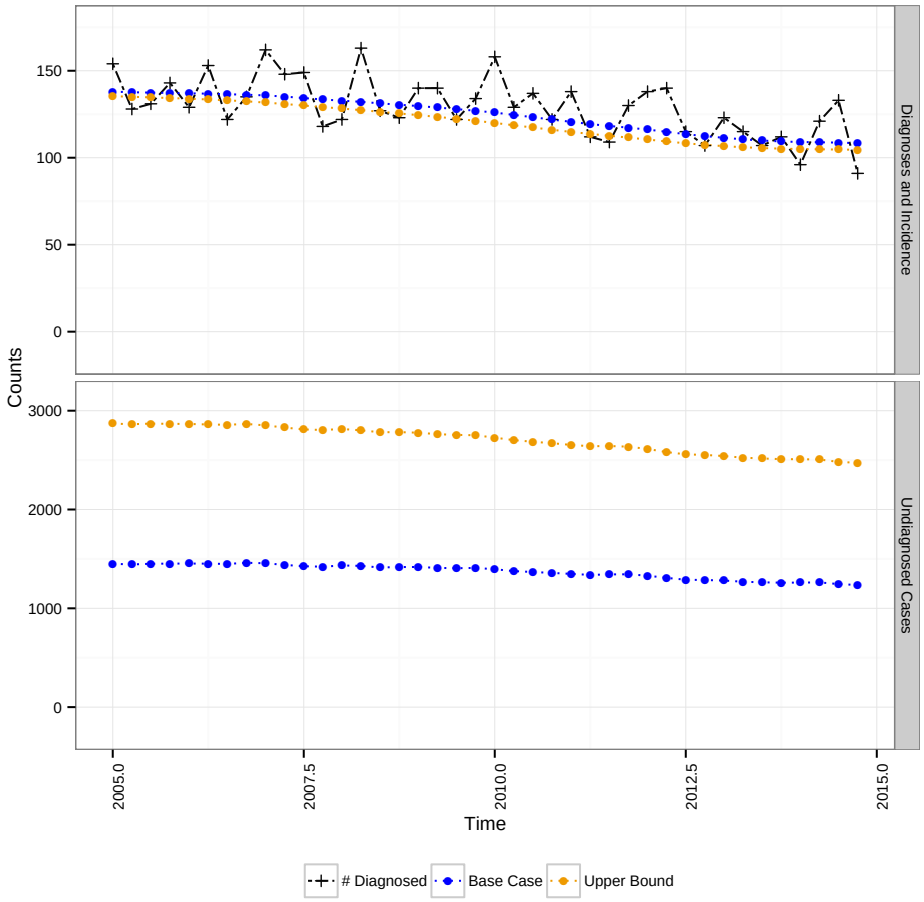

Figure 10: Observed diagnoses and estimated quarterly and undiagnosed counts over 2005-2014 in WA state

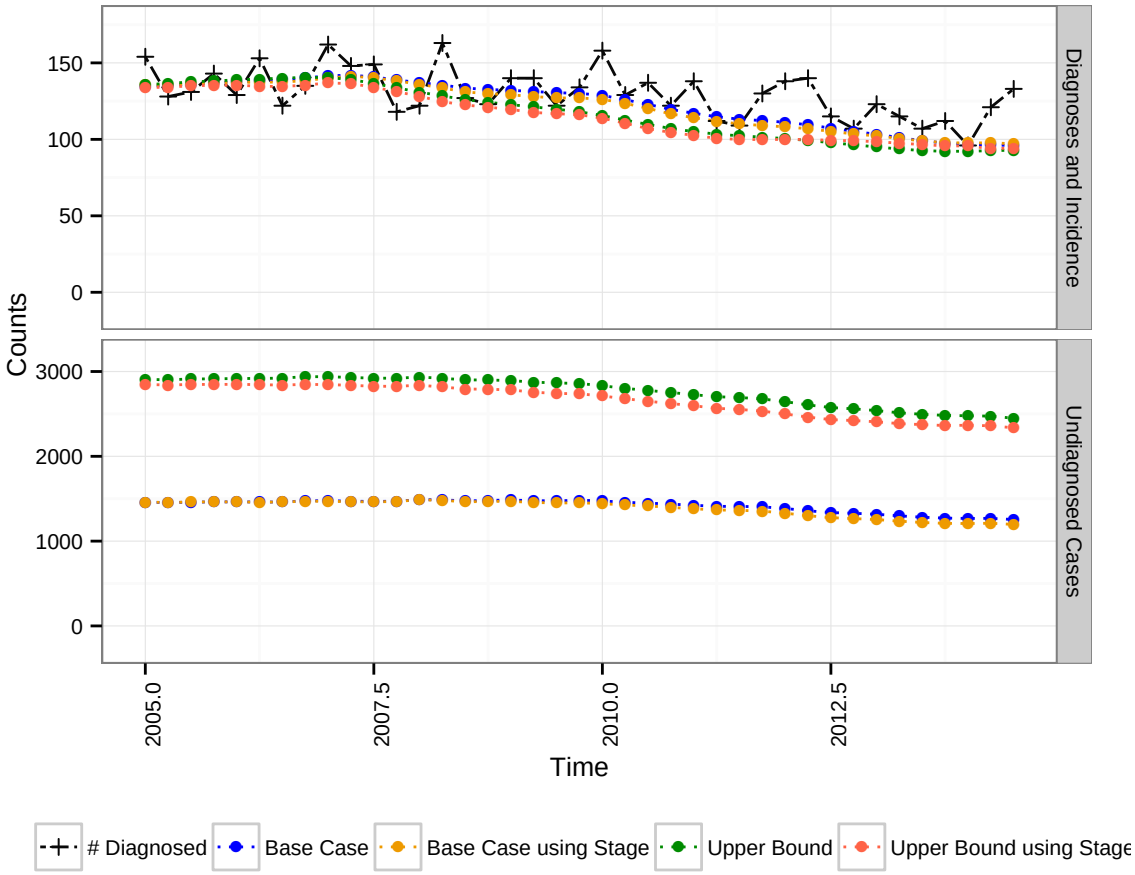

Figure 11: Observed diagnoses and estimated quarterly and undiagnosed counts over 2005-2014 in WA state, using stage-subgroup strata