MEASURING THE UNDIAGNOSED FRACTION:

Understanding the UW and CDC back-calculation models

Martina Morris, PhD Director, UW CFAR SPRC
Jeanette K Birnbaum, PhD Research Scientist, UW CFAR

Based on work originally developed by Ian Fellows, PhD

Outline

1. Back-calculation, the basics

- The original method, developed in the 1980s

2. Current back calculation methods

- UW: “Testing history” back-calculation
- CDC: “Extended” back-calculation

3. Comparing the results for WA State

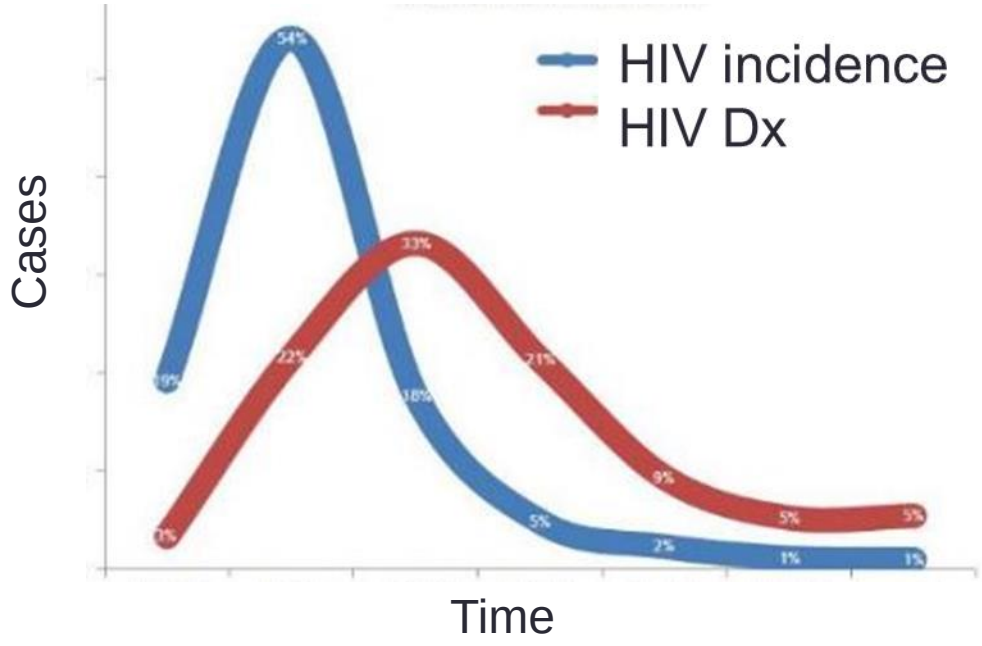
- Future work

BACK-CALCULATION

The basics

Basic idea

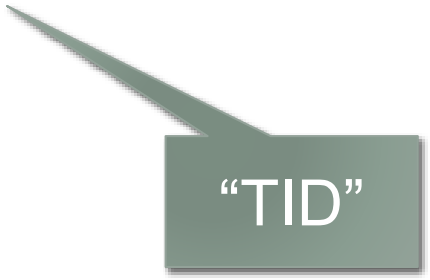
- What you see now
- Is based on infections that happened in the past



Can you use new diagnoses to back-calculate past incidence?

Time from Incidence to Diagnosis

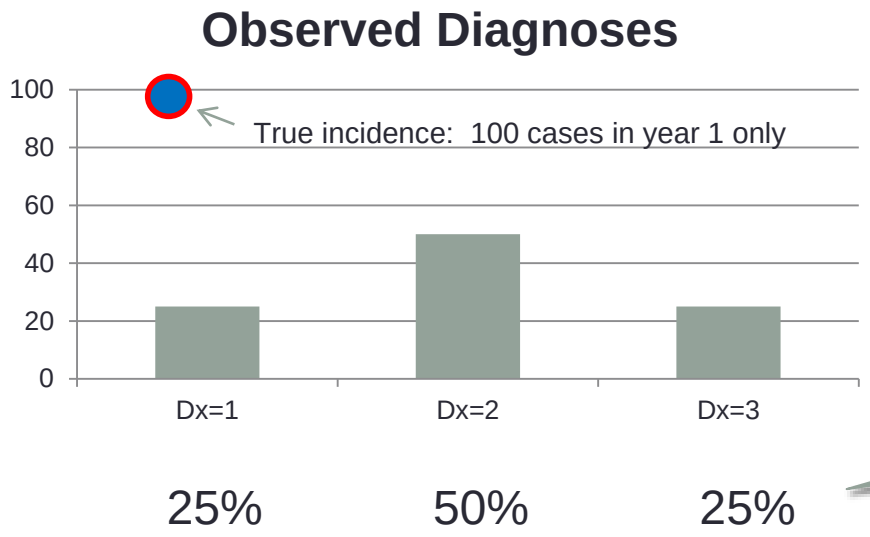
- Imagine an HIV Dx always happens within 3 years of infection
  - 25% get Dx in first year
  - 50% in second year
  - 25% in third year



distribution of "Time from Infection to Diagnosis"

Time from Incidence to Diagnosis

- With this TID (25%, 50%, 25%)
- And 100 new infections this year
- The *observed* HIV Dx curve in the future would look like this:



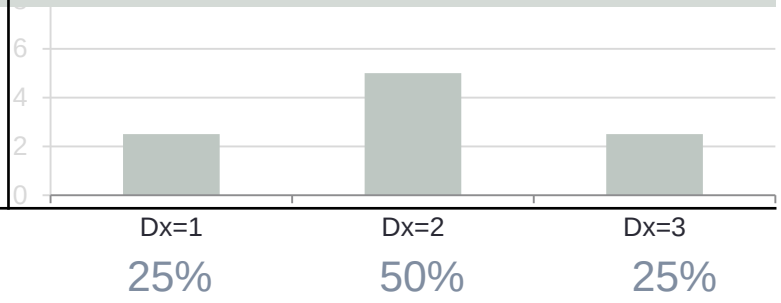
One year of incidence (n=100), distributed over 3 years of Dx

TID

Tracking in tabular form

- From infections (unobserved) to diagnoses (observed)

		Year of Diagnosis (t+z)		
Year of Incidence (t)	# New Infections	2013 (z=0)	2014 (z=1)	2015 (z=2)
2013	**100**	**25**	**50**	**25**
2014				
2015				

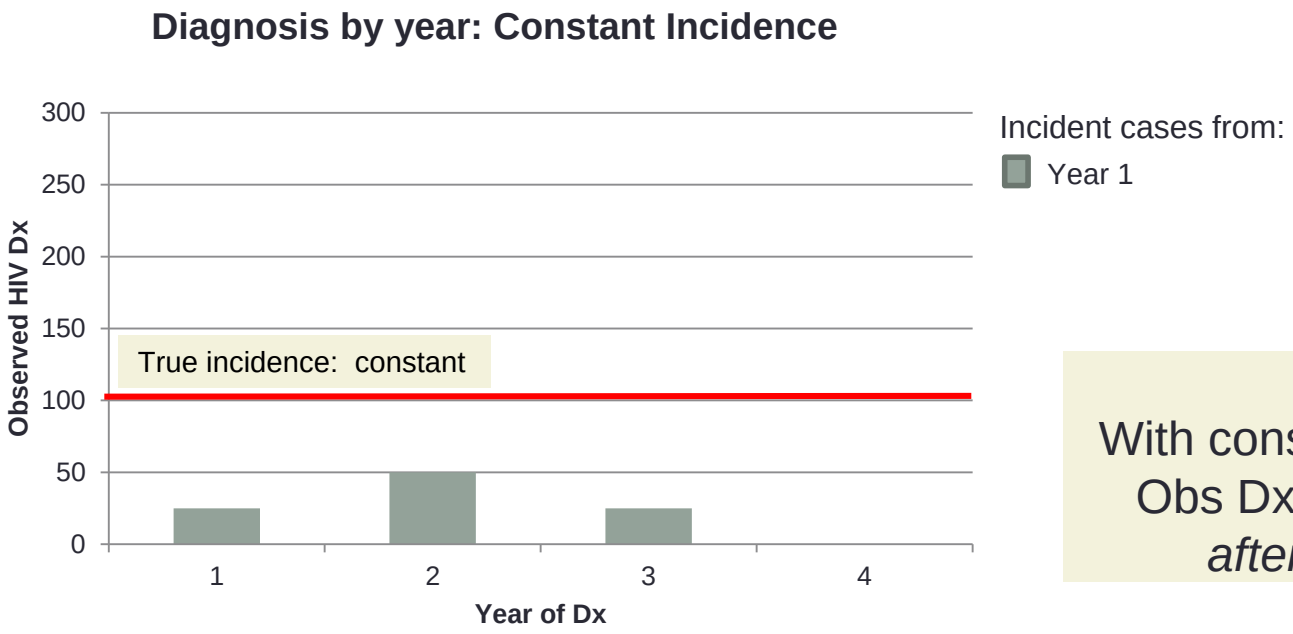


Obs Dx(t+z) = **New Infections(t)** * **TID(t+z)**

With multiple years of incidence?

- Assume constant incidence (100 cases each year)
- And the same TID (25-50-25%)

- Annual observed HIV Dx is now a mix of cases from previous (up to 3) years

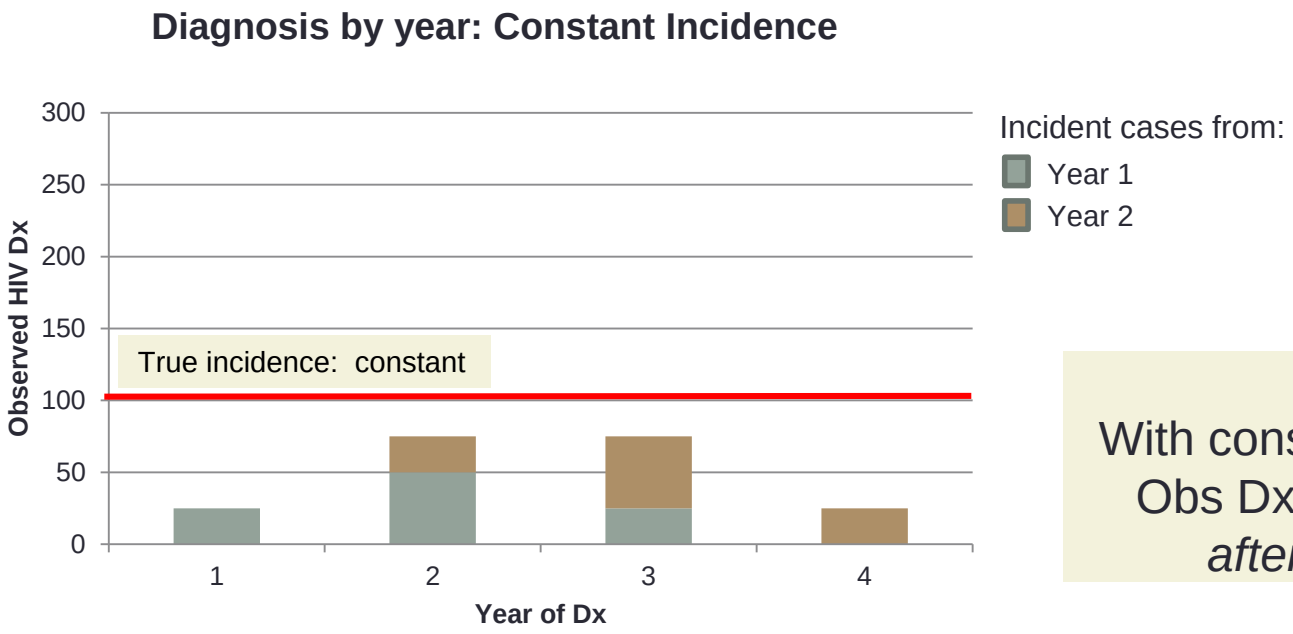


Note:
With constant incidence
Obs Dx = Incidence
after max TID

With multiple years of incidence?

- Assume constant incidence (100 cases each year)
- And the same TID (25-50-25%)

- Annual observed HIV Dx is now a mix of cases from previous (up to 3) years

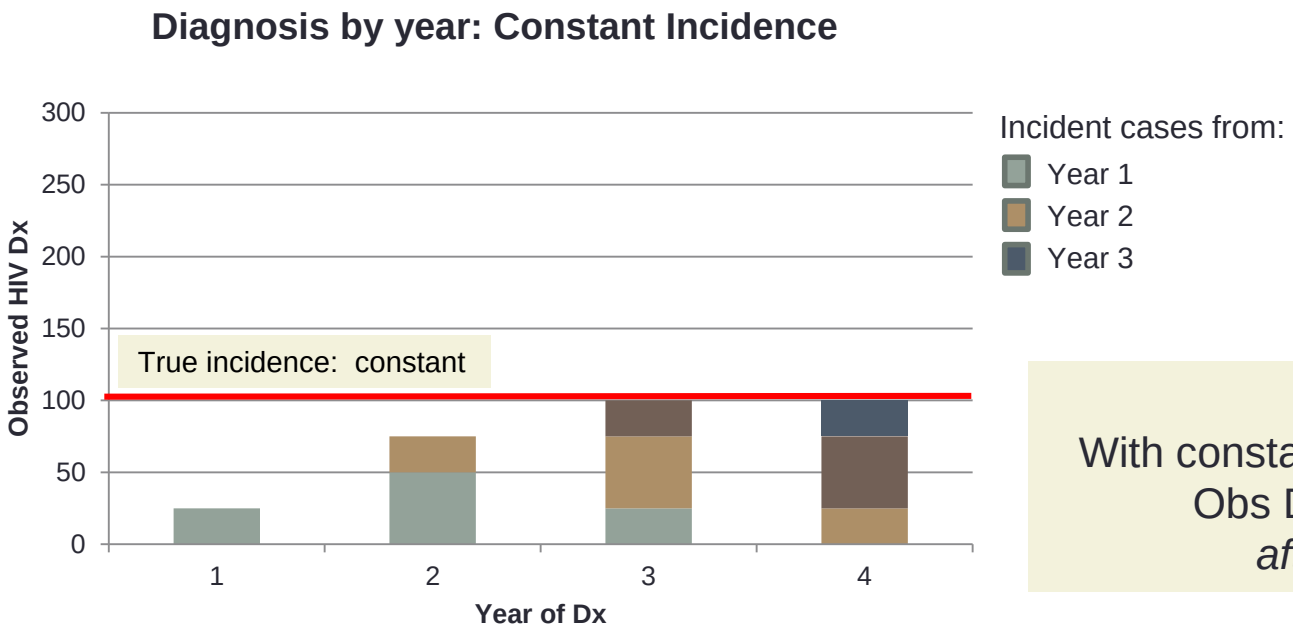


Note:
With constant incidence
Obs Dx = Incidence
after max TID

With multiple years of incidence?

- Assume constant incidence (100 cases each year)
- And the same TID (25-50-25%)

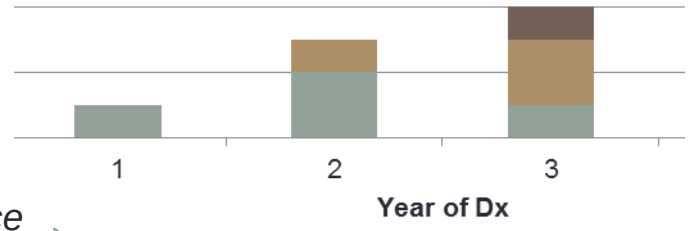
- Annual observed HIV Dx is now a mix of cases from previous (up to 3) years



Note:
With constant incidence and TID
Obs Dx = Incidence
after max TID

Multi-year incidence, tabular form

		Year of Diagnosis (t+z)			*Future years*	
Year of Incidence (t)	New Infections	2013	2014	2015	2016	2017
2013	100	25	50	25		
2014	100		25	50	25	
2015	100			25	50	25
Total Dx	300	25	75	100		



TID t+0 = 25%
t+1 = 50%
t+2 = 25%

We're assuming constant incidence

$$\text{Obs Dx}(t+z) = \sum_{z=0}^{Z} \text{New Infections} * \text{TID}(t+z)$$

Undiagnosed cases calculation

		Year of Diagnosis		
Year of Incidence	New Infections	2013	2014	2015
2013	100	25	50	25
2014	100		25	50
2015	100			25
Totals	300	25	75	100

300 incident by 2015

200 diagnosed by 2015

Undiagnosed (2015) = Cumulative incidence – Cumulative diagnosed

= 300 – 200

= **100**

This would be straightforward, … *if*

- If you could observe everything
  - Incidence
  - TID
  - Dx cases

		Year of Diagnosis		
Year of Incidence	New Infections	2013	2014	2015
2013	100	25	50	25
2014	100		25	50
2015	100			25
Totals	300	25	75	100

- But we only observe Dx cases…

- *If we can estimate the TID from some other data*, then we can "back calculate" the new infections, and the undiagnosed cases

$$\text{Obs Dx}(t+Z) = \sum_{z=0}^{Z} \text{New Infections} * \text{TID}(t+z)$$

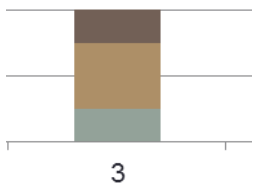
Start with observed Dx

	Year of Diagnosis (t+z)				
Year of Incidence (t)	2013	2014	2015	2016	2017
2013					
2014					
2015					
Total Dx	25	75	100		

$$\text{Obs Dx}(t+z) = \sum_{z=0}^{Z} \text{New Infections} * \text{TID}(t+z)$$

Use an estimated TID

	Year of Diagnosis (t+Z)				
Year of Incidence (t)	2013	2014	2015	2016	2017
2013	25	50	*0.25*NI*		
2014		25	*0.50*NI*	25	
2015			*0.25*NI*	50	25
Total Dx	25	75	100		



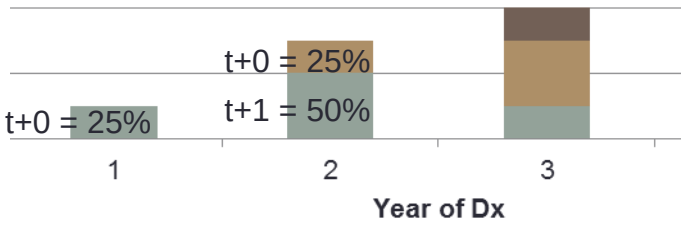
t+0 = 25%
t+1 = 50%
t+2 = 25%

TID

$$\text{Obs Dx}(t+z) = \sum_{z=0}^{Z} \text{New Infections} * \text{TID}(t+z)$$

Back-fill in all the cells

Year of Incidence (t)	Year of Diagnosis (t+z) 2013	2014	2015	2016	2017
2013	*0.25*NI*	*0.50*NI*	*0.25*NI*		
2014		*0.25*NI*	*0.50*NI*	*0.25*NI*	
2015			*0.25*NI*	*0.50*NI*	*0.25*NI*
Total Dx	25	75	100		

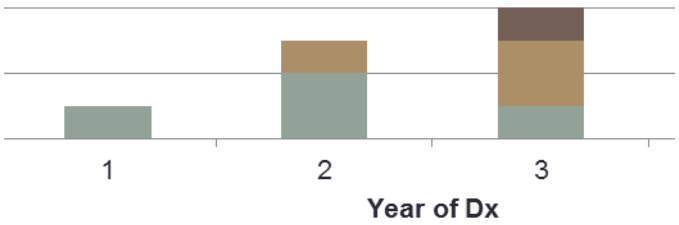


NI=New Infections

$$\text{Obs Dx}(\,t+z\,) = \sum_{k=0}^{Z} \text{New Infections} * \text{TID}(t+z)$$

And solve for the NI (New Infections)

		Year of Diagnosis (t+z)				
Year of Incidence (t)	New Infections	2013	2014	2015	2016	2017
2013	100	25	50	25		
2014	100		25	50	25	
2015	100			25	50	25
Total Dx	300	25	75	100		



$$\text{Obs Dx}(t+z) = \sum_{z=0}^{Z} \text{New Infections} * \text{TID}(t+z)$$

Note (1)

		Year of Diagnosis (t+z)				
Year of Incidence (t)	New Infections	2013	2014	2015	2016	2017
2013	100	25	50	25		
2014	100		25	50	25	
2015	100			25	50	25
Total Dx	300	25	75	100		

We assume constant incidence here, but the approach also works if incidence is changing

Note (2)

		Year of Diagnosis (t+z)				
Year of Incidence (t)	New Infections	2013	2014	2015	2016	2017
2013	100	25	50	25		
2014	100		25	50	25	
2015	100			25	50	25
Total Dx	300	25	75	100		

We assume a constant TID here too, but the approach also works if the TID is changing

Note (3)

		Year of Diagnosis (t+z)				
Year of Incidence (t)	New Infections	2013	2014	2015	2016	2017
2013	100	25	50	25		
2014	100		25	50	25	
2015	100			25	50	25
Total Dx	300	25	75	100		

We can also use this method to project diagnoses forward

Summary

Observe

$$Dx(t+z) =$$

Back calculate *Estimate from other data*

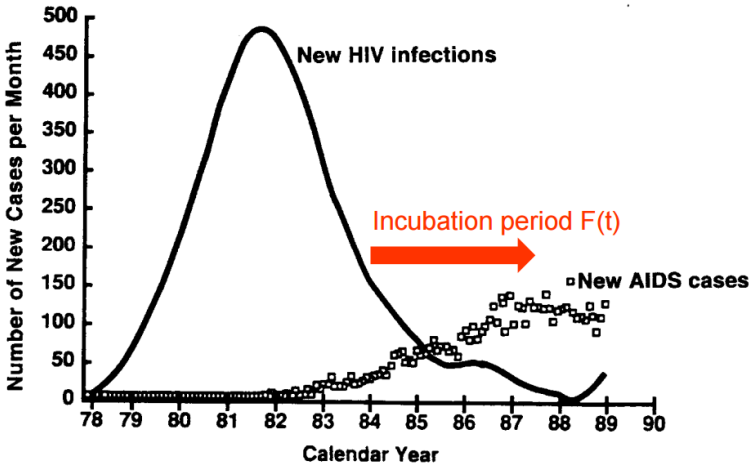
$$\sum_{z=0}^{Z} \text{New Infections} * \text{TID}(t+z)$$

ORIGINAL BACK-CALCULATION

From the way way back

Context

- First used in 1986 by Brookmeyer and Gail for the HIV epidemic
  - At that time, only AIDS Dx were available
  - So the goal was to back calculate HIV incidence from AIDS Dx



Not much data available

http://www.stat.wisc.edu/~yandell/stat/50-year/Brookmeyer_Ron.pdf

Original back-calculation

Observe

AIDS Dx(t+z) = ← Uses AIDS Dx, and AIDS TID

↓

Back calculate *Estimate from data**

$$\sum_{z=0}^{Z} \text{HIV Incidence}(t+Z-z) * \text{AIDS TID}(t+z)$$

Incidence is not assumed to be constant

* Data sources (all from the 1980s):
Multicenter Hemophilia Cohort Study (N=373)
International Registry of Seroconverters (MSM, N=1020)
Amsterdam cohort studies (IDU, N=173; MSM, N=348)
Multicenter AIDS Cohort Study (MSM, N=1861)

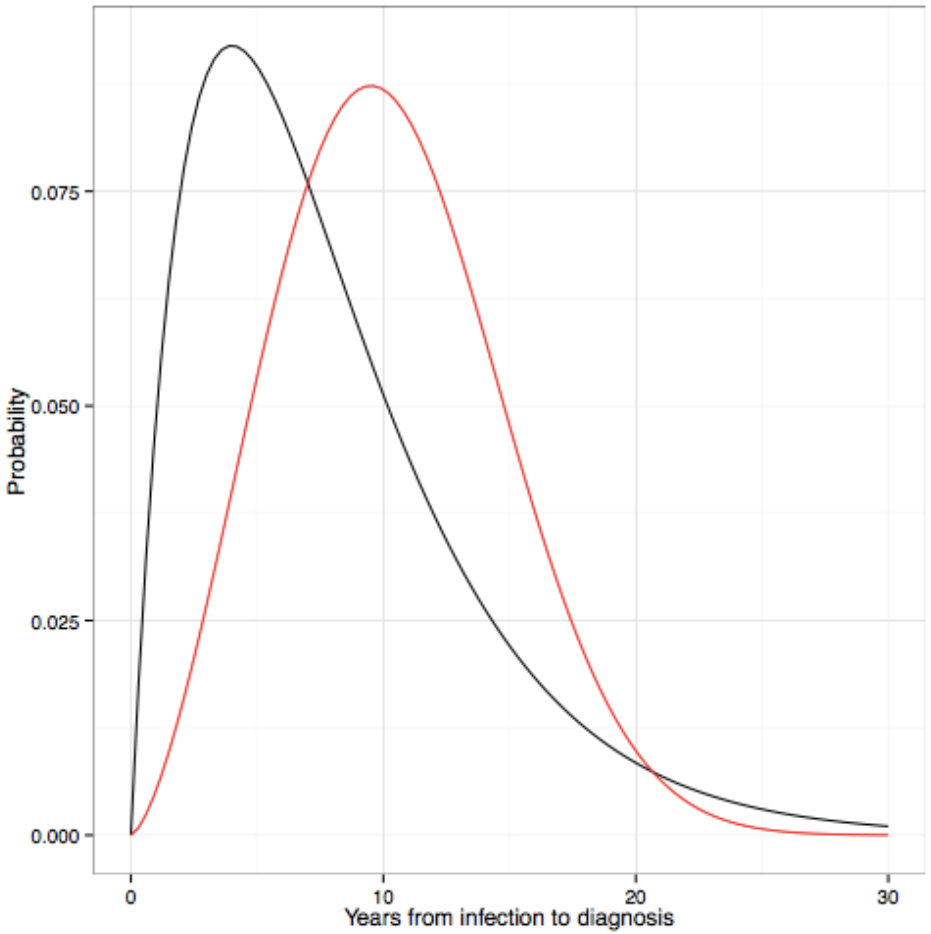
Estimating the AIDS TID ("Incubation Period")

Multiple approaches

- Different parametric forms (Weibull, Gamma)
- Different # stages of infection (2-5)

Problems

- Time of infection usually not known
- Loss to follow-up
- Representativeness of cohorts
- Treatment delays AIDS onset

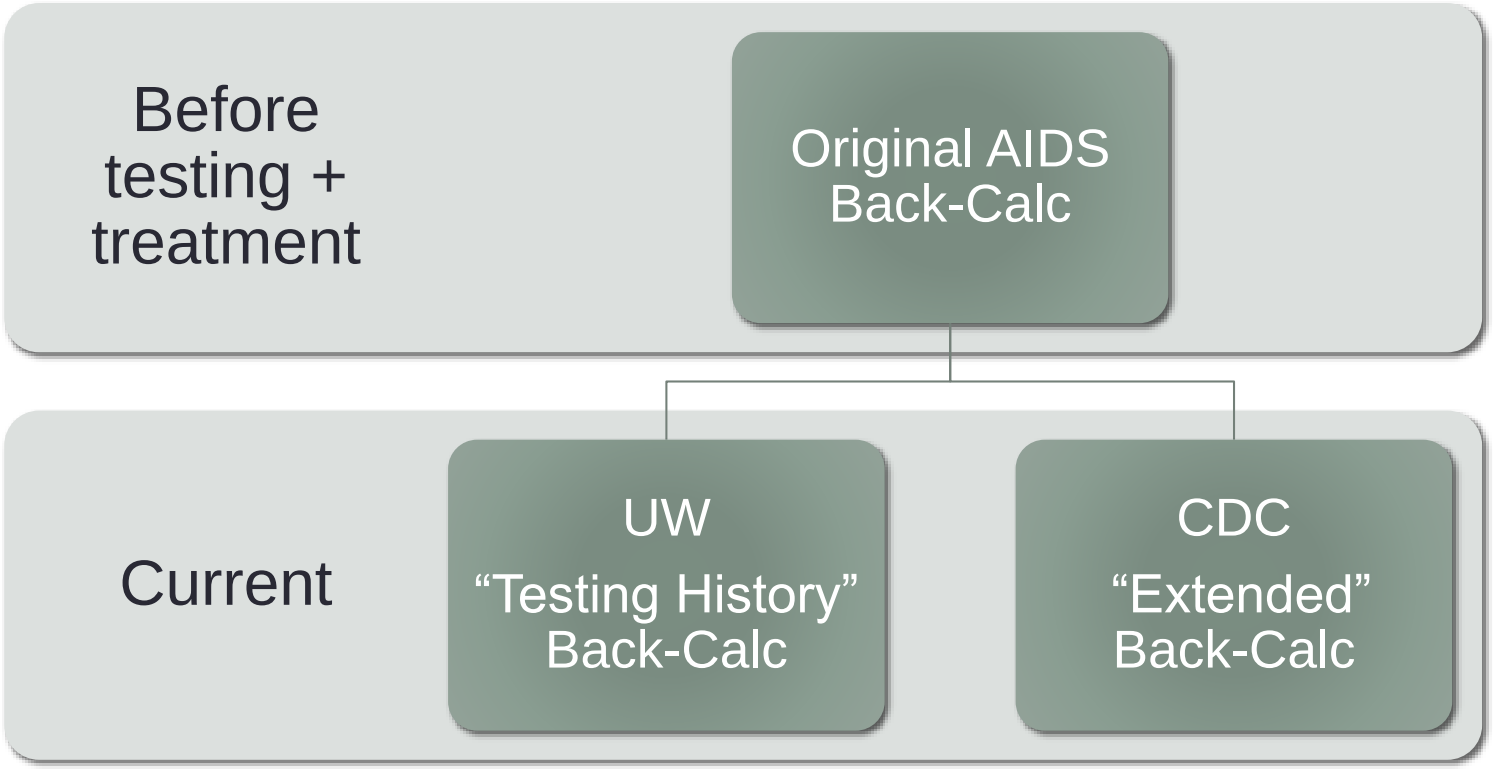


Key improvement to the method

Incorporating data on HIV Dx

- Necessary because treatment dramatically reduced AIDS Dx
- Possible because HIV Dx became reportable

Fast forward to now



UW vs CDC: Overview

- Similarities:
  - Both incorporate data on observed HIV Dx
    - But use this in very different ways
  - Both use complex algorithms for estimation
    - UW : EM Algorithm (EM = Expectation-Maximization)
    - CDC : Bayesian MCMC (MCMC = Markov Chain Monte Carlo)
- Differences:
  - Use different information in the back-calc process
  - Select and weight cases from the Dx populations differently

UW TESTING HISTORY METHOD

Originally developed by Ian Fellows, PhD

Fellows, I., M. Morris, J. Dombrowski, S. Buskin, A. Bennett and M. R. Golden (2015). "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." PLOS One **10(7): e0129551.**

UW Testing History back-calculation

HIV Dx(t+z) =

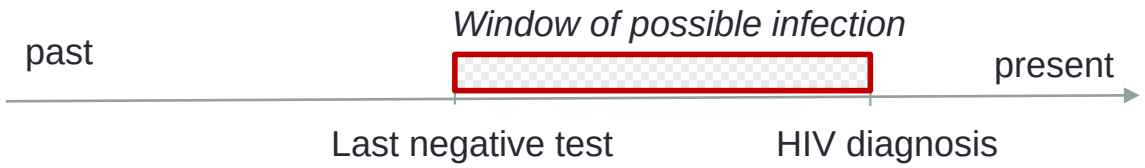
A simple approach: Use HIV Dx to back calculate HIV incidence

So for this we need to estimate the HIV TID

Back calculate

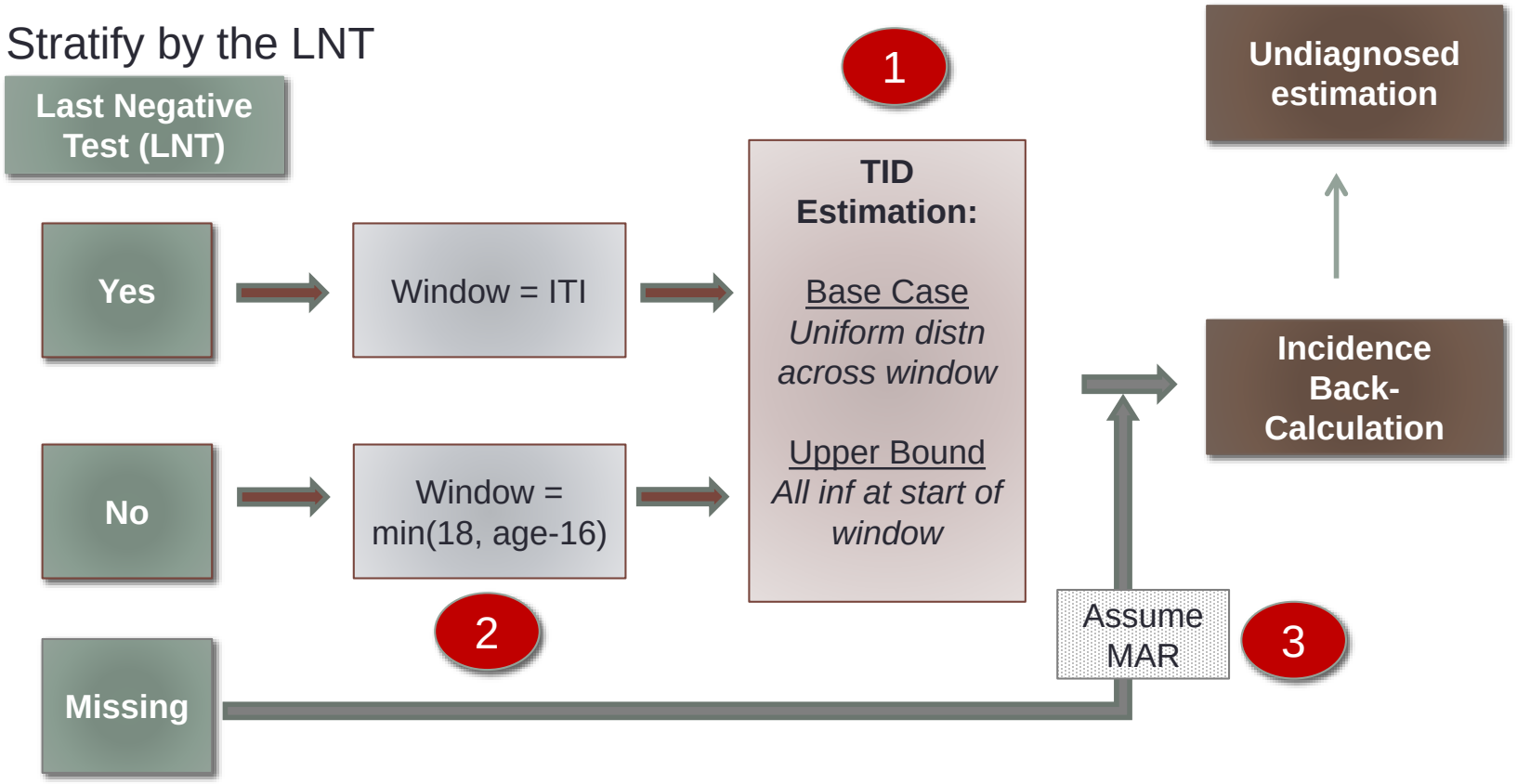
$$\sum_{k=0}^{Z} \text{HIV Incidence}(t+Z-z) * \textbf{HIV TID}(t+z)$$

Use testing histories to estimate the TID



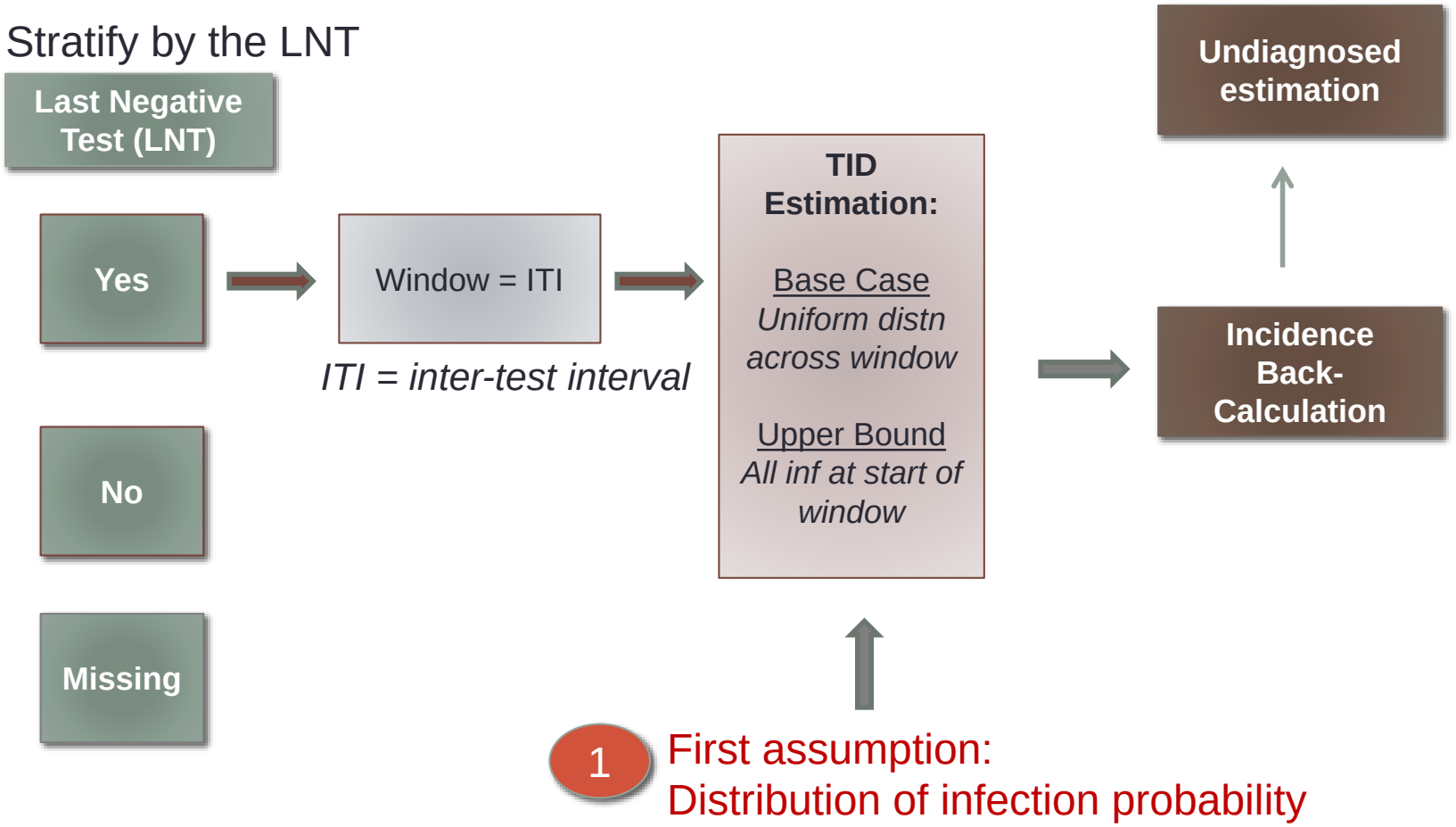
- Testing histories give us an infection window, but…
  - When did infection occur in the window?
  - What if people never had a LNT, or have missing data?

Full Model overview (we'll take it in pieces)



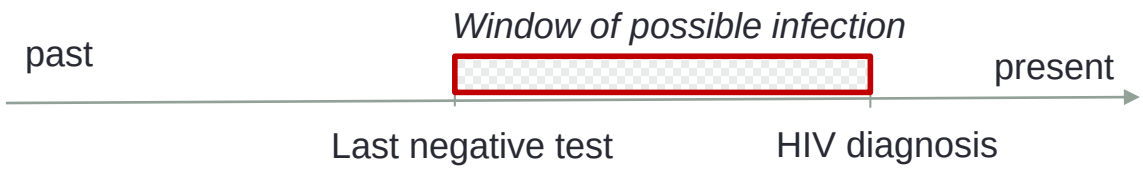
3 important assumptions

1. Model for repeat testers

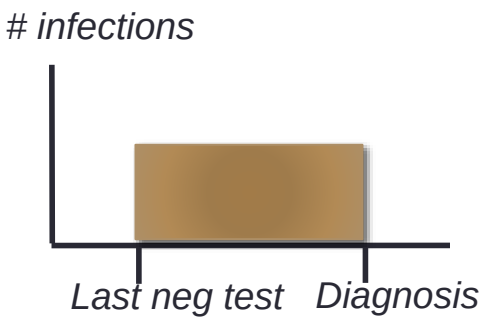


Assumption 1:

Infection probability distribution

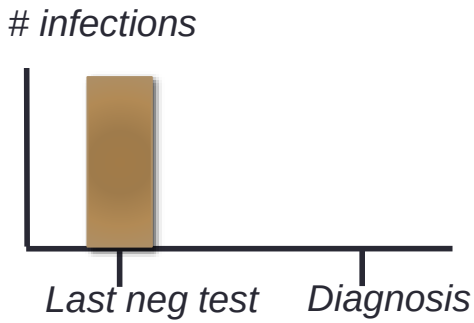


Base Case



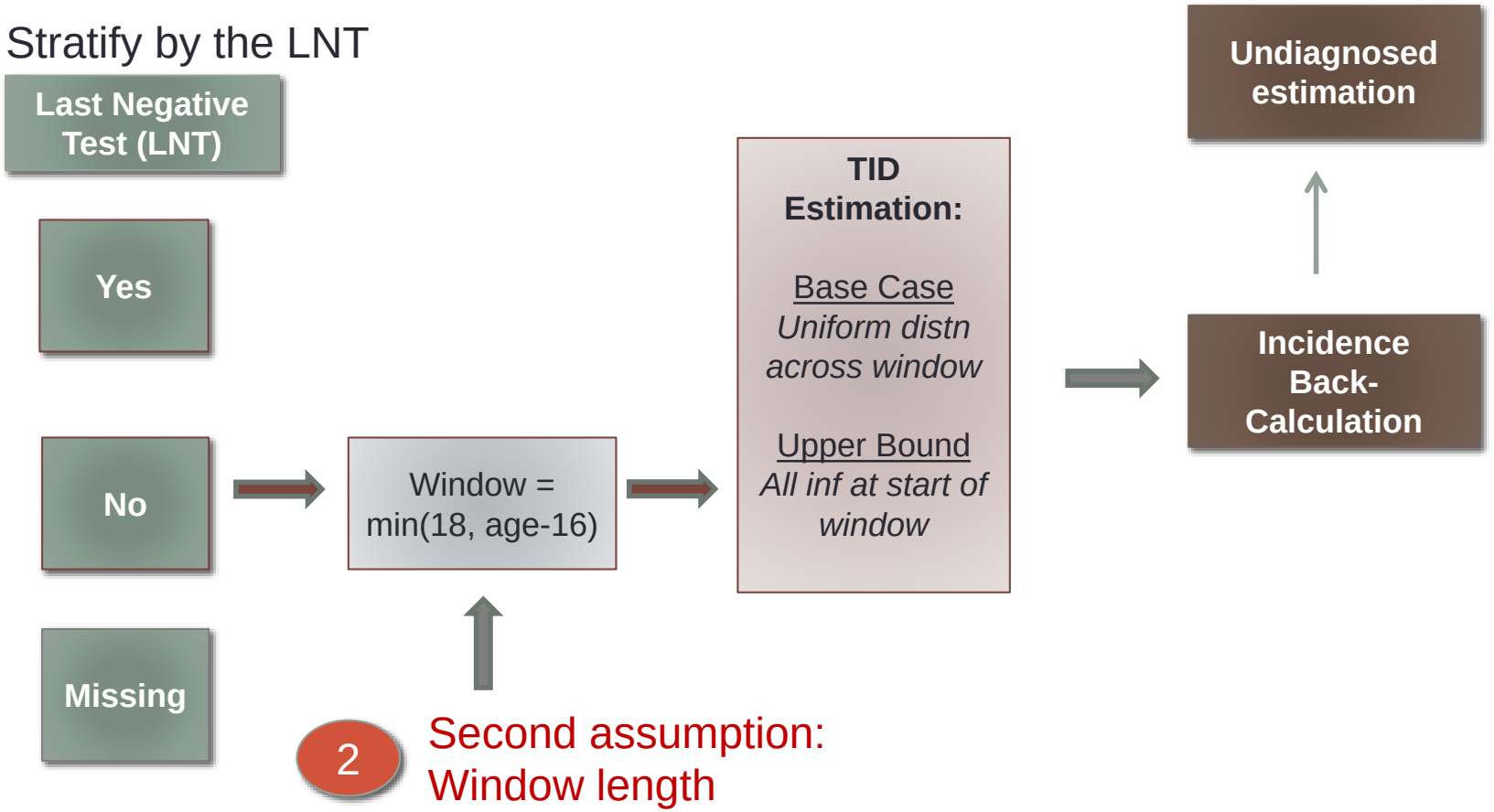
Uniform:
Distributes the probability of infection uniformly across the possible interval

Upper Bound



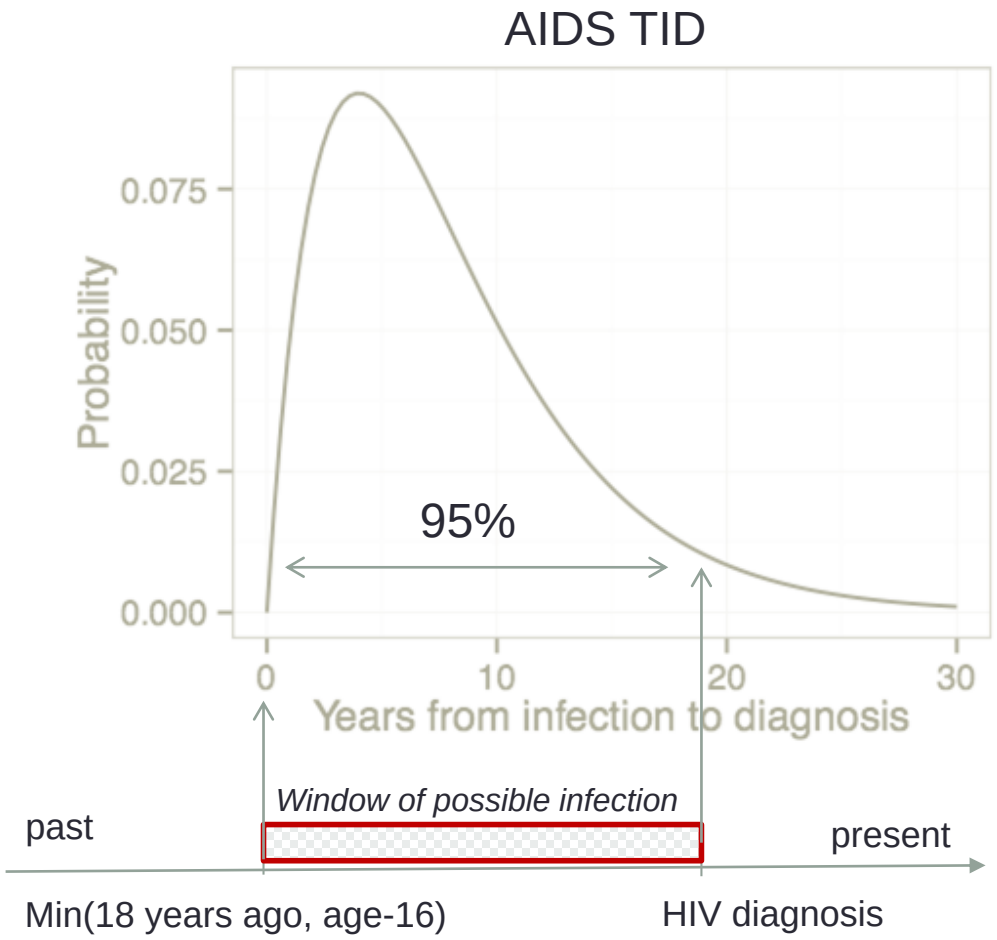
At last neg test:
Probability=1 that infection occurred on the day after the last negative test

2. Model for Dx with no previous test



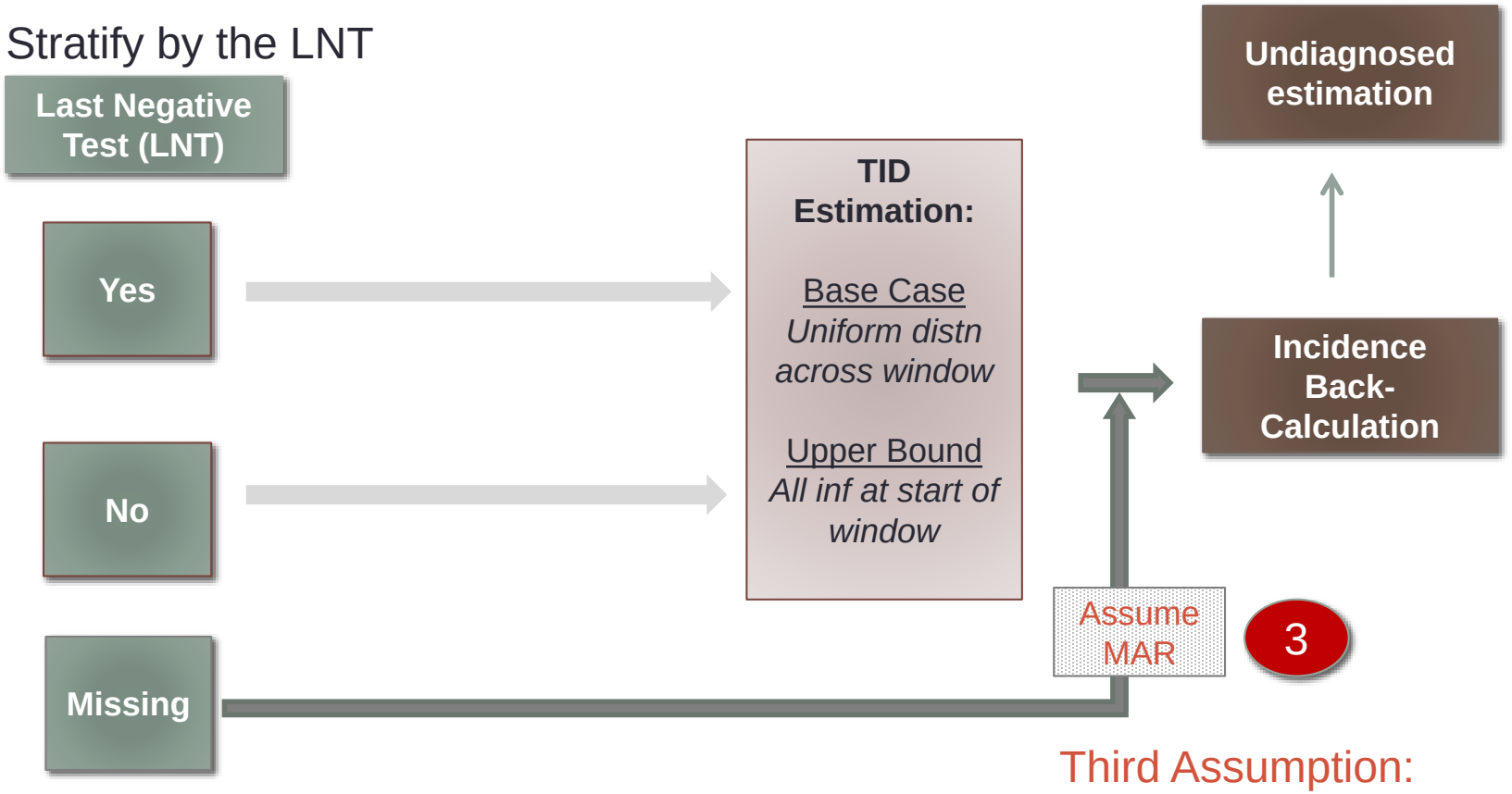
Assumption 2:

Window length if no previous test



- 95% of HIV+ progress to AIDS in 18 years (Lui 1996)
- Age 16 is the median age of sexual debut in the US
- So we take the minimum of these as the window length
- And then apply base case or upper bound assumption for the distribution of infection probability

3. Model for cases missing test info



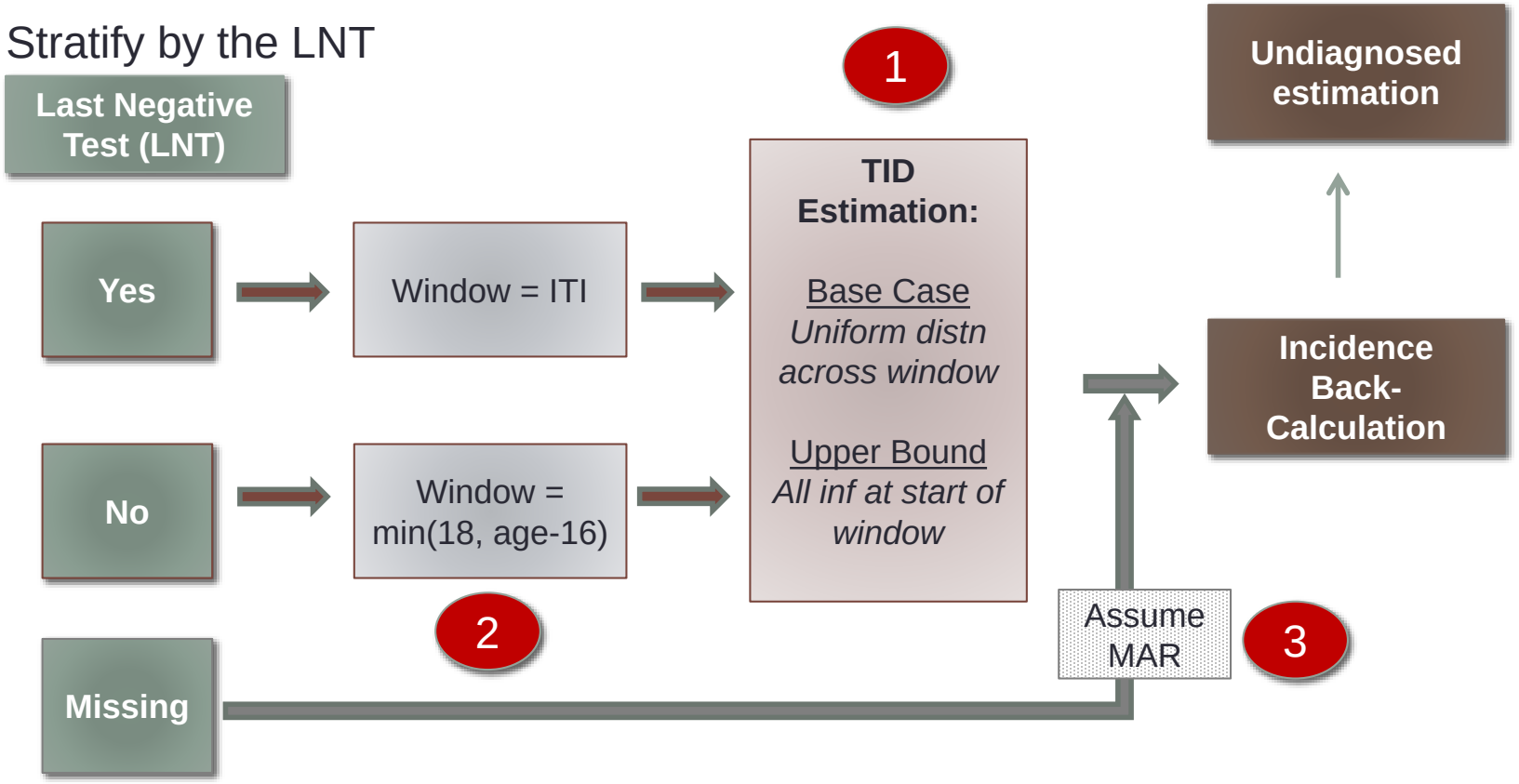
Third Assumption:
Treatment of missing cases

Assumption 3:

If Dx case is missing test information

- We have two options:
  - Include these when estimating the TID distribution
    - Assuming the maximum possible infection window
    - **IMPACT**: A conservative (longer) estimate of the time spent undiagnosed.
  - Exclude these when estimating the TID distribution
    - We still use them in the back-calculation, we just give them the TID estimated from the other cases
      - Assumes these cases are “missing at random” (MAR)

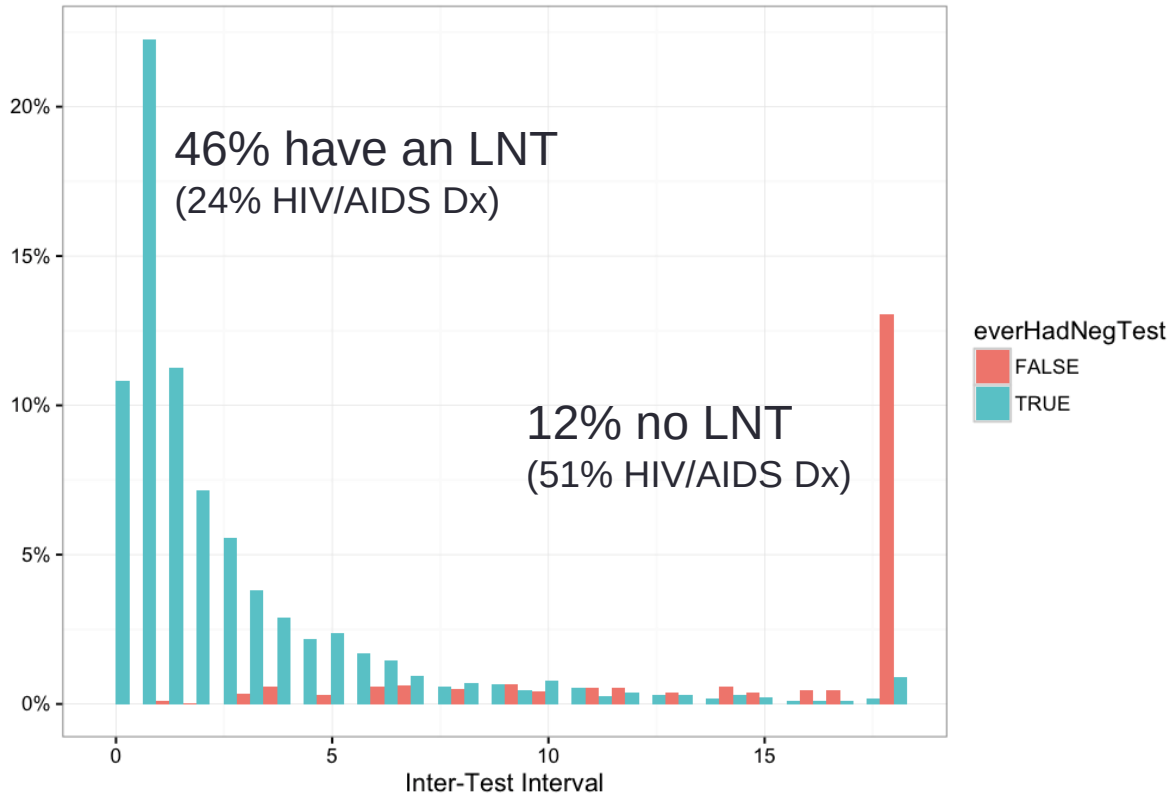
Full Model overview



3 important assumptions

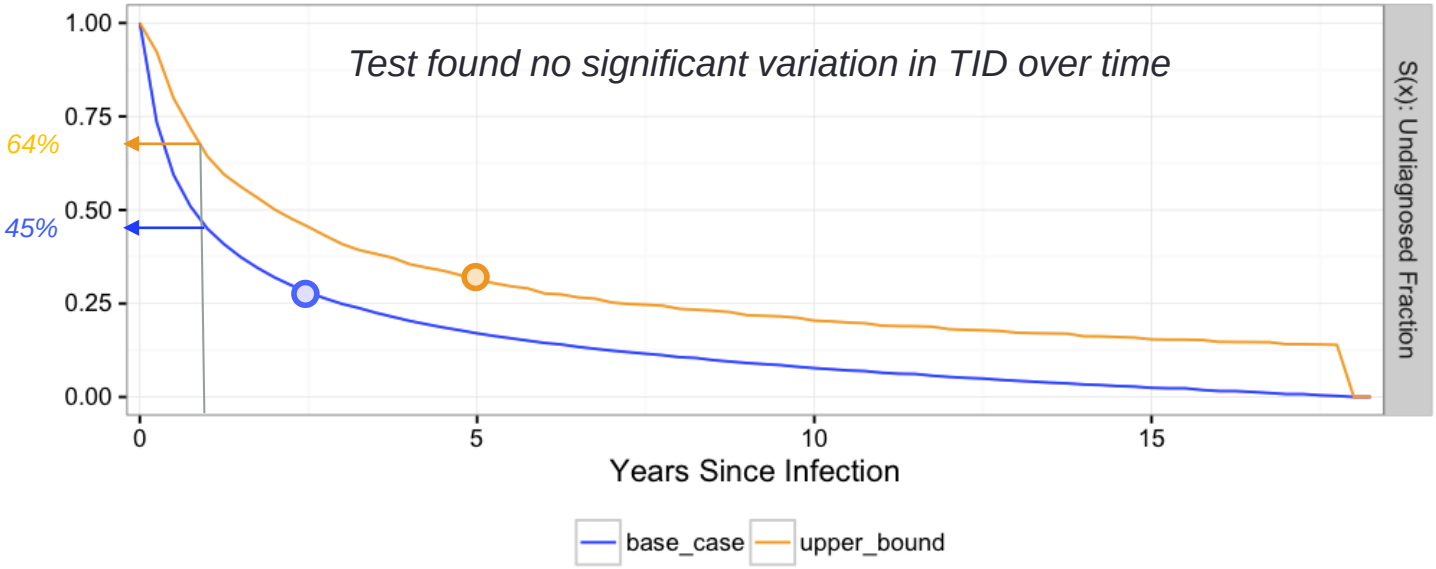
Results WA State: ITI dist'n (2006-2015)

- The distribution of inter-test intervals
  - For all non-missing cases



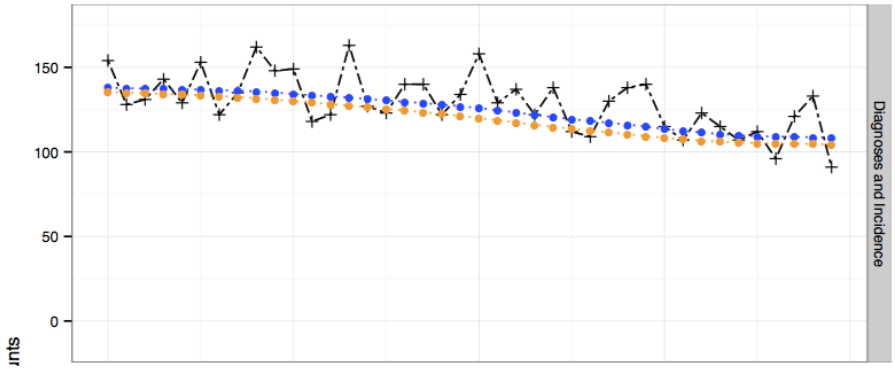
42% of cases are missing, so not included in TID estimation (39% HIV/AIDS Dx)

Results WA State: HIV TID (2006-2014)



Estimate	Mean TID	% UnDx at 1yr
Base Case	2.5 years	45%
Upper Bound	5 years	64%

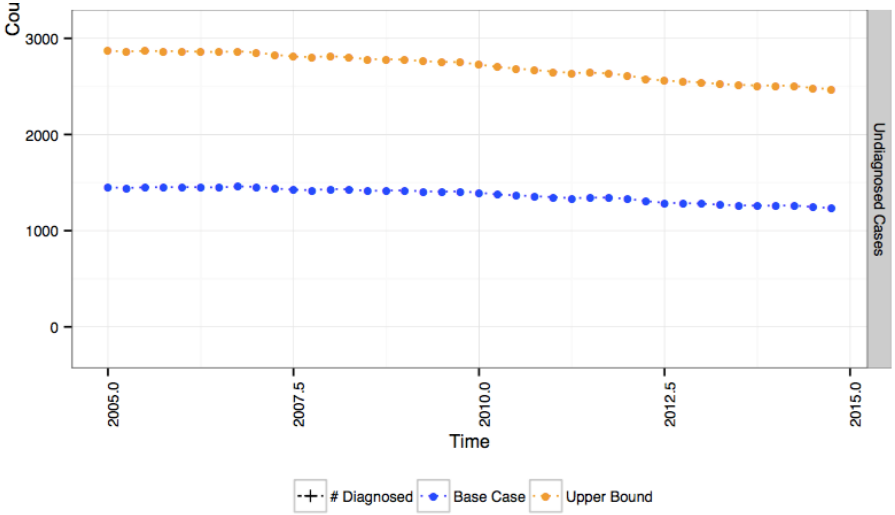
Results WA State: Incidence & UnDx



Observed Dx (black)

Estimated Incidence (colors)

Obs=Est and Base=UB suggests relative stability in the dynamics



Estimated UnDx cases

Upper bound ~ 2x Base Case

2014:
~100 new cases/qtr
1200-2500 UnDx cases

CDC EXTENDED BACK CALC

An, Q., J. Kang, R. Song and H. I. Hall (2015). "A Bayesian hierarchical model with novel prior specifications for estimating HIV testing rates." Statistics in Medicine **(in press).**

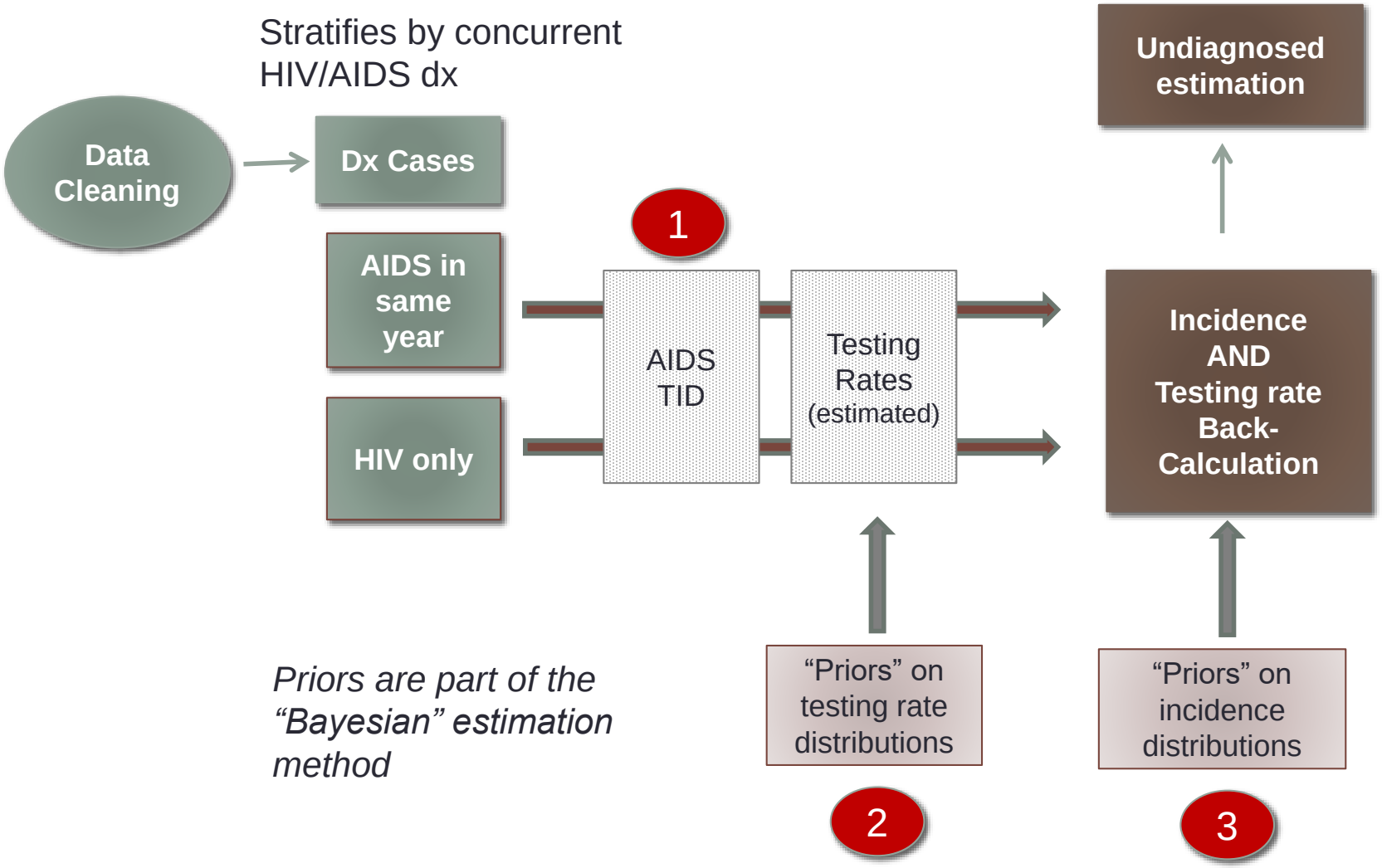
CDC Extended Back-Calc Overview

- Uses AIDS Dx, like original method
- Adds data on HIV Dx

- Stratifies observed cases by
  - HIV Dx only vs.
  - Concurrent HIV/AIDS diagnosis (AIDS Dx within 1 year of HIV Dx)

- And it uses a Bayesian estimation approach
  - So it will need “prior” distributions to start the estimation algorithm

CDC Model overview



Key difference: competing risks of Dx

- In the UW TH model
  - HIV+ person faces only one "risk": the risk of HIV Dx by testing

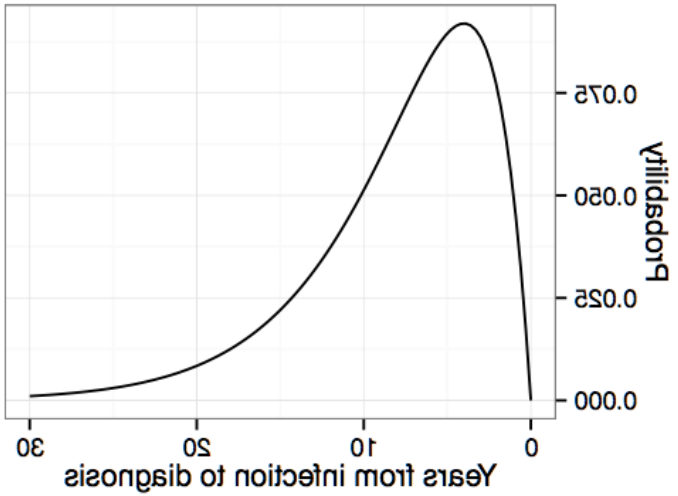
- In the CDC model
  - HIV+ person faces two risks: AIDS Dx or HIV Dx
  - So the model has to specify how these processes unfold over time

HIV+ →

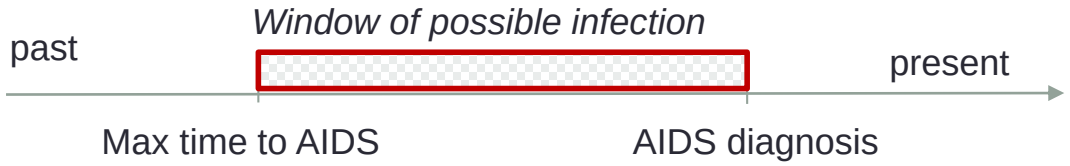
Event at time t	Depends on:
AIDS Dx	AIDS TID (Dx), not tested
HIV Dx	AIDS TID (noDx), not tested earlier, tested now
UnDx	AIDS TID (noDx), not tested

Assumption 1: Historical AIDS TID

This is how they distribute the probability of HIV infection back in time for someone diagnosed with AIDS



Gamma(2,4)



Assumption 2: HIV testing rate priors

HIV testing rates are not observed – they are estimated

Huh?

But these rates are used *to estimate incidence*.

So how can they be estimated too?

Bayesian estimation in a nutshell

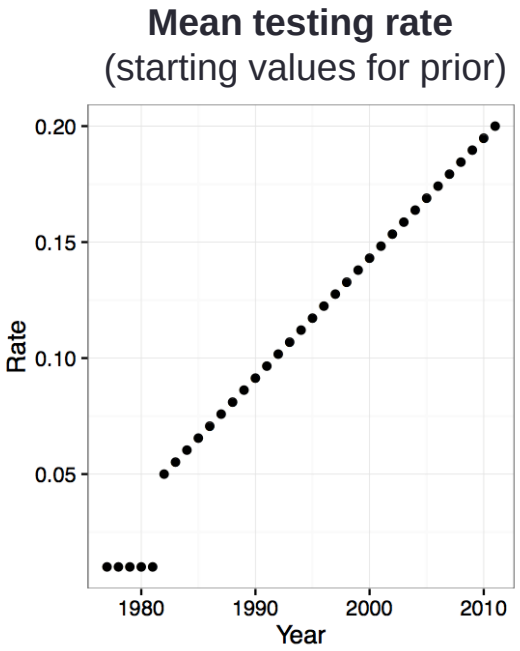
- Say you have a parameter you want to estimate, *p*
  - *But you don't have a simple formula for it in terms of the data you observe (here HIV Dx and AIDS Dx)*
- Draw a starting value from a distribution
  - Reflects what we think/know about the value of *p*
- Plug it in and calculate the predicted Dx
- Compare the predicted Dx to the observed
- Update the estimate of *p* in the prior
- Repeat until predicted=observed

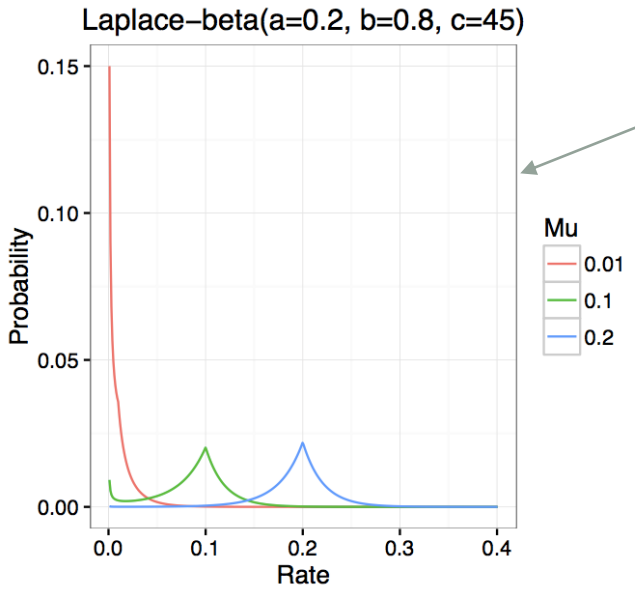
This is the prior distribution

Assumption 2: HIV testing rate priors

HIV testing rates are not observed – they are estimated

- Prior distributions for the annual rates, 1977-present





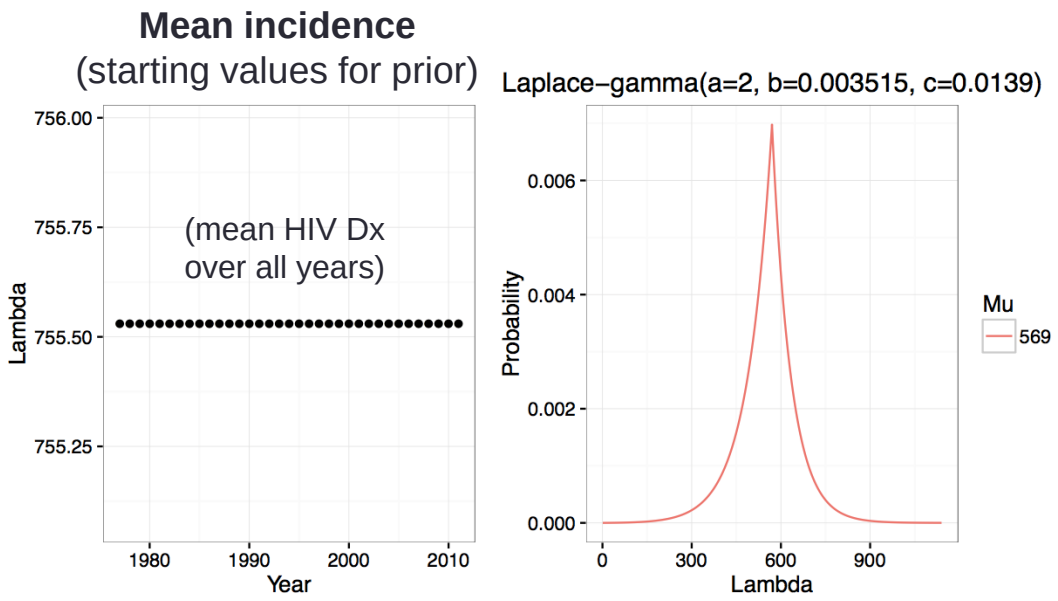
Example priors for different means

The testing rate is *for HIV+ persons* (not the pop'n)

Assumption 3: HIV incidence prior

HIV incidence is also estimated in this model

- Specifies another "prior distribution" to start

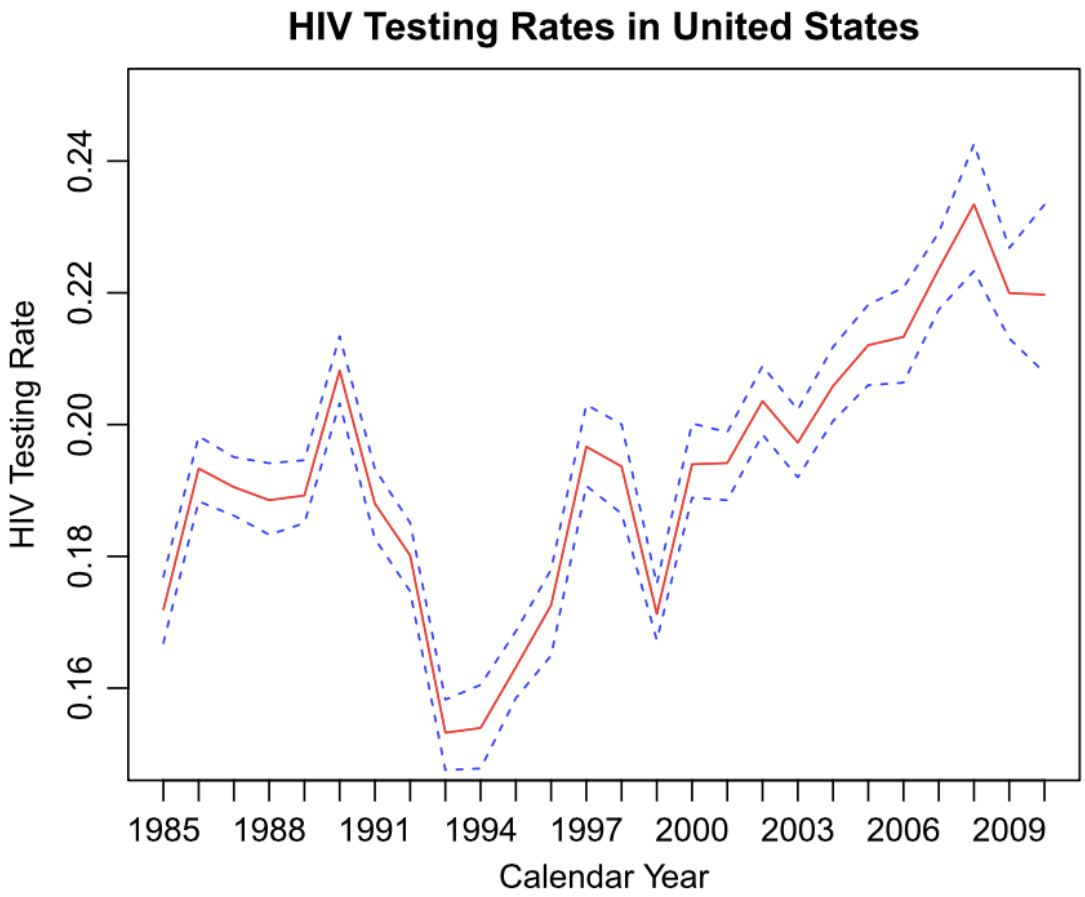


Single prior,
but posterior
estimates allowed to
vary by year

Overview of the CDC model

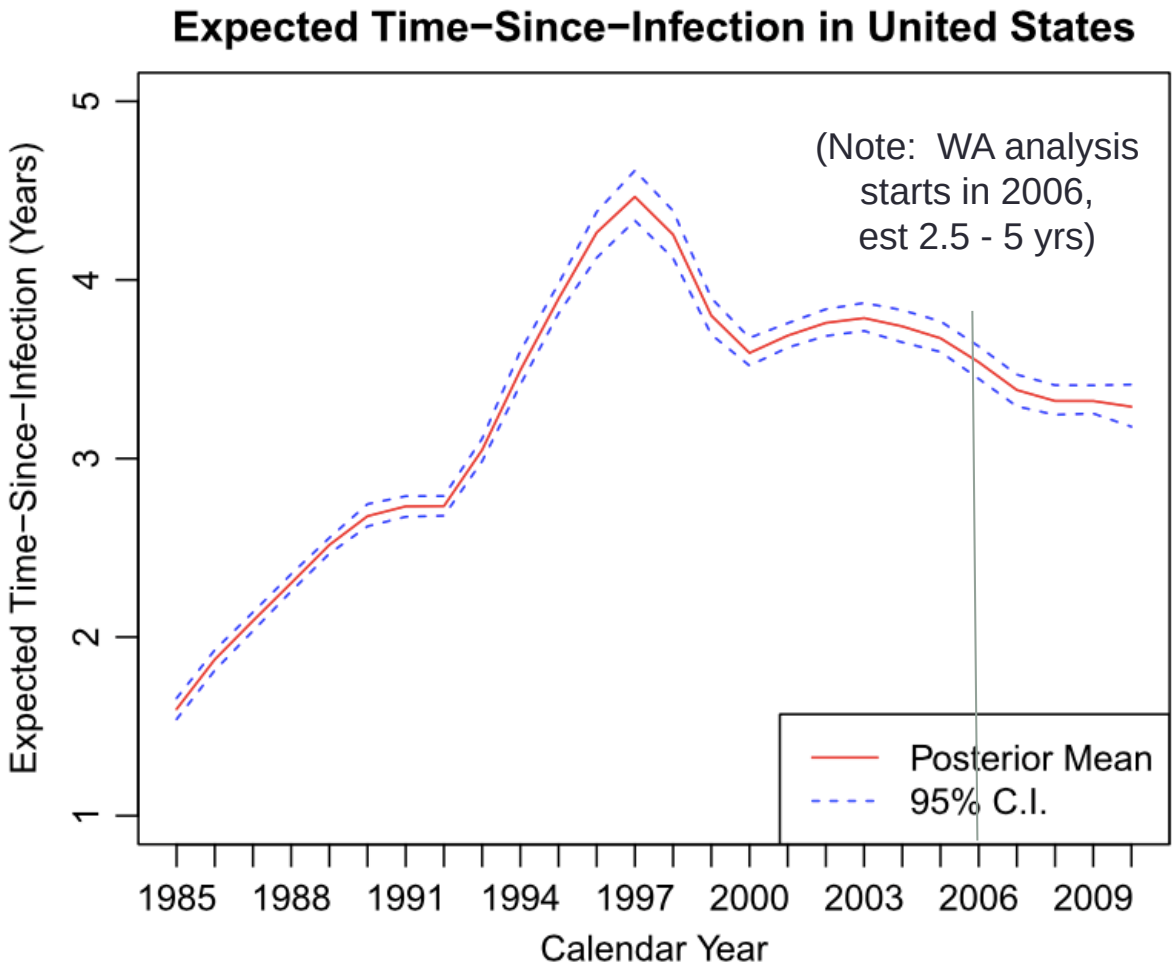
- Two unknown parameter sets to estimate:
  - Annual testing rates $\{p_i^H\}$
  - Annual HIV incidence $\{\lambda_i\}$
- Two observed data series
  - Annual HIV Dx
  - Annual AIDS Dx
- One fixed assumption (the AIDS TID)
- Two Bayesian priors
  - for testing rates and HIV incidence

CDC Results: National testing rates 1985-2010



Mean rate estimate stabilizes at around 22% per year

CDC Results: National mean time to HIV Dx



2010: 3.3 years

CDC Results: National UnDx estimate for 2012

	CDC National Model
Undiagnosed Fraction	**12.8** (12.0-13.6)
Total	1,218,400 (1,207,100—1,228,200)
Undiagnosed	156,300 (144,100—165,900)

Note: This requires estimating the number of persons living with HIV, and we're not describing how they do that here

Other aspects of both models

- Both can be used to estimate the UnDx *Fraction*
  - But need an estimate of PLWH for the denominator
- Both models assume there is some stability in the year to year changes (smoothing)
  - For UW method: Incidence counts in adjacent years are smoothed
  - For CDC method: Both incidence and testing rates are smoothed
- Both models can stratify estimation by other factors
  - Sex
  - Risk exposure
  - Geography *(though this raises issues about modeling migration)*
  - *But small sample sizes will lead to unstable estimates*

Summary of model differences

- Uses of HIV Dx
  - UW estimates the HIV TID to back-calculate HIV incidence
    - Using measured inter-test intervals for HIV Dx when available
    - Relies on the max AIDS TID window for cases Dx on their first test
    - Variation in the TID by year can be evaluated and incorporated
  - CDC estimates testing rates as part of the HIV incidence back-calculation
    - Calibrated to best fit observed HIV and AIDS Dx trends
    - Relies also on the AIDS TID
    - The annual average rate is allowed to vary over time
- Uses of AIDS Dx
  - UW does not use this (but could be adapted)
  - CDC uses this to estimate both testing rates and HIV incidence

COMPARING RESULTS FOR WA

What we know now, and plans for future investigation

WA state estimates for 2012

		CDC*	UW	% DIFFERENCE
All	**UnDx Fraction**	**11.0 (7.7—15.0)**	**10.6 (18.8)**	**-3.8%**
	Undiagnosed	1,700 (1,200--2,400)	1,410 (2,750)	**-20.6%**
	Total	15,500 (14,900--16,100)	13,310 (14,650)	**-16.5%**
MSM	**UnDx Fraction**	**11.7 (7.5—16.5)**	**6.8 (12.6)**	**-72.1%**
	Undiagnosed	1,200 (730--1,800)	647 (1274)	**-85.5%**
	Total	10,300 (9,900--10,800)	9,519 (10,147)	**-8.2%**

**Source: WA DOH*

Potential Sources of Differences

- PLWH denominator for estimating the undiagnosed fraction
  - Doesn't seem to be the driving difference since the totals are similar

So, that leaves one of:

- Case selection/weighting
  - CDC does adjustments and weighting for reporting delays and missing data
  - But I'm guessing this is not the primary driver
- Model structure and assumptions
  - If testing histories provide more precision for MSM, our estimates may be better
  - Not sure what could lead to their estimates being better

Future Work

- Compare results using identical datasets
  - Accommodate weights in testing history method to use CDC data
  - Can't run CDC method on our data since it needs to go back to 1977, which requires doing all their data cleaning relevant to the older cases
- Test both models on a mock dataset in which incidence is known
  - "Simulation study"
  - Generate the mock data from an independent model of HIV natural history

THANK YOU