Small-Area Estimation of HIV Diagnoses in the United States: Cross-Validation of a Transmission Model using Jurisdictional Data

By

Kevin Weiss

MPH

Global Epidemiology

____

Dr. Eli Rosenberg
Committee Chair

Small-Area Estimation of HIV Diagnoses in the United States: Cross-Validation of a Transmission Model using Jurisdictional Data

By

Kevin Weiss

B.S.

University of Michigan - Ann Arbor

2014

Thesis Committee Chair: Eli Rosenberg, Ph.D

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Global Epidemiology
2016

Abstract

Small-Area Estimation of HIV Diagnoses in the United States: Cross-Validation of a Transmission Model using Jurisdictional Data

By Kevin Weiss

Background:
Human immunodeficiency virus (HIV) infection remains a major public health issue in the United States, with an estimated 1.2 million persons living with HIV at the end of 2012. Accurate estimation of HIV incidence at state and sub-national levels is a key goal for further characterization of and resource allocation to the HIV epidemic. The HIV care continuum, a framework for care and treatment engagement, is a rich data source for national and local data that can inform incidence estimation.

Methods:
National and state HIV surveillance data from AIDSVu.org and the Centers for Disease Control and Prevention were used to populate a web tool, reflecting a published biological transmission probability model, with estimates of the percent of individuals at each level of the HIV care continuum in order to produce state-specific estimates of HIV transmissions. New diagnoses were used as a proxy for incidence. Multiple linear regression was used to investigate and predict the association between 2012 transmissions estimates and new diagnoses in 2013, accounting for social determinants of health, including, income inequality, racial makeup, health insurance, education, and poverty. Models fit included three classes: prevalence, predicted transmissions, and predicted transmissions reflecting state-specific estimates of the proportion of undiagnosed HIV infection.

Results:
Prevalence-only models accounted for 89% of the variation in 2013 HIV diagnoses, while models reflecting state-level variability in the proportion of cases undiagnosed accounted for up to 35% of the unexplained variation. Inclusion of social determinants of health improved predictive ability to account for 50% of the variation unexplained by prevalence (adjusted R^2=0.95). Regression coefficients for predicted HIV transmissions were statistically significant, and could not rule out a 1:1 correspondence between transmissions and new diagnoses.

Conclusions:
A continuum-based approach to estimation can accurately predict new HIV diagnoses in 2013 with a limited amount of local data on social determinants of health. Further continuum stratification, reflecting the targeting of prevention efforts to particular groups, is supported by this methodology. Incorporation of information from state and local jurisdictions can allow for refinement of models and methods to maximize the predictive value for HIV diagnoses.

Small-Area Estimation of HIV Diagnoses in the United States: Cross-Validation of a Transmission Model using Jurisdictional Data

By

Kevin Weiss

B.S.

University of Michigan – Ann Arbor

2014

Thesis Committee Chair: Eli Rosenberg, Ph.D

A thesis submitted to the Faculty
of the Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health/Master of Science in Public Health
in Global Epidemiology
2016

Acknowledgements

I would like to thank Dr. Eli Rosenberg, my faculty thesis advisor, for all of his support, comments, and thoughtful and insightful advice throughout this process. From the day when we first discussed this idea, I have been fascinated by the possibilities that it might have, and for that, I am thankful to you for supporting my love of learning. I am grateful to CDC and the various state health departments represented here for tracking all of this information and making it publically available. I am thankful to Dr. Patrick Sullivan and the entire PRISM team for the time I have had, the wonderful work and professional experience, and the endless support and guidance you have provided over the last two years. I am appreciative of the Epidemiology and Rollins cohort for their support, entertainment, and much more over the course of this program and this process. Here’s to a successful step for all of us into public health practice, and I know I will see everyone again as we work to understand, reduce, and eliminate health disparities. Lastly, I would like to thank my family for their unconditional love, support, and guidance. You have made this academic journey and my life possible, and I cannot begin to express my gratitude for all that you have allowed for and encouraged.

Table of Contents

Background

Human immunodeficiency virus (HIV) infection burden has remained at a substantial level in the United States, with the Centers for Disease Control and Prevention (CDC) estimating that the number of persons living with HIV increased nearly 9% from 2007 to the figure of 1.2 million persons at the end of 2012 (1). National estimates of the percentage of persons living with HIV who are aware of their status range from 80-85%, with a goal of the National HIV/AIDS Strategy (NHAS) being to increase that percentage to 90% (1-3).

The number of new infections estimated to occur every year is thought to have remained stable at around 50,000 infections since the early 2000's (4). Most recently, CDC estimates of annual new infections were 47,500 in 2008 and 2010, with 45,000 predicted in 2009 (5). CDC has not published annual HIV incidence estimates since 2010, rather releasing diagnosis data adjusted for reporting delays and missing information. NHAS plans to adopt this approach to monitoring new infections in 2016 (3, 5).

One recently proposed incidence estimation method uses routinely collected HIV testing history from newly diagnosed cases to model the number of individuals who are undiagnosed with infection in a local jurisdiction (6). The second method, accounting for undiagnosed HIV infection at national and state levels, uses back-calculation incorporating disease severity at diagnosis, death records, lab results, and the estimated yearly total of HIV diagnoses (1, 4, 7). State-level estimates of the undiagnosed proportion are available for each year from 2008-2012, while national estimates of the undiagnosed proportion are available for each year between 2009 and 2012. A third method, focusing on transmissions rather than incidence rates, is the basis of the calculations in this analysis. This model applies biological transmission probabilities, calculated from national surveys, to 2009 national care continuum data to estimate rates and the number of HIV transmissions at each level of the HIV care continuum (8). This model was calibrated to the 2009 national estimate of 45,000 transmissions per year and thus may not say anything new about national incidence, but may have some novel application for prediction of subnational incidence. Different methods have led to different conclusions, with estimation remaining a key barrier to diagnosing the true scope of the HIV epidemic in the United States.

In the United States, the HIV care continuum cascade is a framework for understanding care engagement (3, 9, 10). It outlines sequential steps of care and treatment from the time of diagnosis to the achievement of viral suppression, including: diagnosis of HIV infection, linkage to care, engagement or retention in care, prescription of antiretroviral therapy (ART), and suppression of viral load (11). The reduction of disease burden, morbidity, and mortality due to HIV infection relies on providing effective medical treatment to and averting new transmissions from individuals at each stage of the continuum. Recent transmission modeling research posits that nearly one-third of ongoing HIV transmissions are contributed by those unaware of their infection (8). Valid and precise estimation of individuals present at each level of the cascade is a vital component of designing prevention programs aimed at reducing HIV transmission (12).

There is a need for local data to estimate and determine the small-area, local burden of HIV incidence, as state and local health departments may determine how resources are deployed and allocated based on cases or new diagnoses. These departments serve as a source of CDC continuum estimates through collection of HIV diagnoses within their borders and, for participating sites, through the provision of laboratory data and participation in the Medical Monitoring Project (MMP) (13, 14). The care continuum can provide one locally available data source to inform incidence estimation, and recent models have demonstrated how the national continuum can be leveraged to make inferences about incidence (8).

In addition to HIV care and treatment data, social determinants of health play an important role in public health practice, reflecting underlying environments and health services that can influence the likelihood of an individual to be susceptible to health issues or to access care (15). These social and structural factors, including race, education, inequality, and insurance, can influence and explain a significant proportion of variance in HIV and AIDS diagnoses, incidence, and outcomes at both state and county levels (16-18).

Utilizing published social determinants of health and HIV continuum data from national and state jurisdictions as input data for the biological transmission model, the viability of using publicly available data and existing HIV estimation models for small areas can be evaluated. This analysis will aim to serve as an example for how subnational and national data and methodology for HIV can be integrated to better understand short-term state and national trends in HIV diagnosis. Validation of small-area estimation methods for HIV diagnoses can serve as an example of the application of existing models to state and national data to accurately generate estimates of new HIV diagnoses.

Extra Data Source Considerations

The linkage to care indicator in the continuum measures the proportion of individuals who, in a given calendar year, have at least one documented viral load or CD4+ test within three months of their diagnosis date (19). Engagement in care uses MMP data to estimate the proportion of those living with HIV who had at least one HIV medical care visit in the previous year, while the National HIV Surveillance System (NHSS) differentiates between “in care,” the proportion of those diagnosed with at least one viral load or CD4+ test in that observation year, and “retained in care”, which measures the proportion of diagnosed persons with two or more viral load or CD4+ tests performed at least three months apart in a given observation year. MMP data is used to estimate the proportion of individuals receiving HIV medical care with a documented ART prescription in the observation year, while either MMP or NHSS data can be used to estimate the proportion of individuals living with HIV whose most recent HIV viral load was below 200 viral copies per milliliter.

Although estimates at each level of the HIV care continuum are regarded as having an inherent uncertainty, the proportion of HIV-infected persons who have not yet been diagnosed is perhaps the most important and most difficult to evaluate (6). Current surveillance sources, such as the National HIV Behavioral Surveillance (NHBS) system or the group of 25 jurisdictions collecting laboratory and testing data that serve as the basis of national HIV incidence calculations, likely do not capture the true proportion of HIV-infected individuals who have not yet been diagnosed. The aforementioned back-calculation method for estimation of incidence and the proportion of undiagnosed individuals differentiates between two approaches with different denominators: prevalence-based, which use the total number of individuals living with HIV (both diagnosed and undiagnosed), and diagnosis-based, which is limited to those aware of their infection (19).

Methods

New Diagnoses

State-level new HIV diagnoses were considered the outcome for this analysis. Data on new HIV diagnoses in 2013 were obtained from AIDSVu.org, an Emory University database providing state, national, and county-level information on socioeconomic measures, HIV prevalence, and annual HIV diagnoses. HIV prevalence and diagnosis information at state and county levels are obtained from the national CDC National HIV Surveillance System, which summarizes HIV case reports from state and local health departments (20). Data represent individuals 13 years and older who were diagnosed with an HIV infection between January 1, 2013 and December 31, 2013. CDC estimates incorporate statistical adjustment for delays in reporting, as well as missing information on risk factors, but do not adjust for incomplete reporting of information.

Social Determinants of Health Covariates

Variables reflecting social determinants of health were included as covariates in this analysis. These variables, obtained from AIDSVu.org, were initially gathered on an annual basis from American Community Survey estimates from the U.S. Census Bureau, including population estimates, poverty and median income estimates (21-25). Model covariates included: the proportion of cases attributable to black individuals and men who have sex with other men, the proportion of individuals under age 65 lacking health insurance, the proportion of individuals with a high school education, the proportion living in poverty, Gini coefficient of income inequality, and median household income (Table 1).

Transmissions and Care Continuum Estimates

Model-based transmission estimates were obtained using an RShiny web tool populated with national and state continuum data, reflecting predictions about the distribution of individuals at each level of the HIV care continuum. This web tool, built from the biological transmission probability model created in Skarbinski et al., uses input prevalence data and parameter values to estimate the rate and number of transmissions occurring at each level of the continuum (8). Parameter estimates for each level of the care continuum were obtained from CDC's 2012 national HIV continuum and reflect persons living with diagnosed or undiagnosed HIV infection (26) (Table 2).

All models included a transmissions term and used observed cases in a state and applied percentages from national and/or state data to create care continua. There were two classes of models with a transmissions term built. Models using an overall average value of 87.2% for percentage of individuals diagnosed with infection were referred to as National Average (NA) models. Models incorporating state-specific estimates of the percent of individuals diagnosed with infection were referred to as Morbidity and Mortality Weekly Report (MMWR) models, with values being obtained from a recent MMWR report (Figure 1) (7). All other parameter estimates, including the proportions of individuals engaged in care, retained in care, and with viral suppression, were identical between the two models after adjusting these values to reflect a denominator of all diagnosed individuals, rather than all diagnosed and undiagnosed individuals (Table 2).

The difficulty in monitoring the HIV care continuum and the numbers of individuals present along the steps is partially due to the existence of multiple methods of estimation. HIV care continua can differ by how levels of the cascade are defined and are dependent on the data source from which the information is drawn. The continuum is often conceptualized in two ways: 1) a prevalence-based denominator that includes an estimate of individuals who are unaware of their infection 2) a diagnosis-based denominator that only includes individuals aware of their infection. Information for both the prevalence- and diagnosis-based continua can reflect two main data sources: 1) the National HIV Surveillance System (NHSS), which uses data reported to the CDC by state and local health departments 2) the Medical Monitoring Project (MMP), which reports weighted data from individuals receiving HIV care (19).

Data Analysis

Linear regression models were built using a hierarchical regression approach from continuum data with state-level estimated new diagnoses in a given year as an outcome. A hierarchical approach was used to compare fully specified models, containing all possible covariates, to reduced models, as well as to compare models containing different types of transmissions terms, MMWR and NA. Two initial models were built, one containing all descriptive covariates (M0a) and a subset of those that used selection techniques that maximized the adjusted R^2 (AR^2) metric, the Akaike's Information Criterion (AIC), and Bayesian Information Criterion (BIC) (M0). Three models containing a predictive term for 2012 HIV prevalence were built, one with all covariates (M1b), one with a subset that maximized AR^2, AIC, and BIC (M1a), and one with solely prevalence as a predictor (M1).

This process of creating three models was repeated for models including a term for transmissions predicted using national average continuum data input into the web tool. One model included all covariates (M2b), a second included only a subset that maximized AR^2, AIC, and BIC (M2a), and one included solely the estimated transmissions term as a predictor (M2). Finally, three models containing a term for transmissions predicted using MMWR data input into the web tool were built. One model contained all covariates (M3b), one contained a subset that maximized AR^2, AIC, and BIC (M3a), and a third solely included the estimated transmissions term as a predictor (M3). Descriptive characteristics of each model, including AR^2, AIC, BIC, were summarized (Table 3). The regression coefficient of the transmissions term was recorded, while the gain in AR^2 of each model compared to the simple regression model containing only 2012 prevalence as a predictor was used by dividing the increase in AR^2 in each subsequent model by the proportion of variance unexplained by the prevalence term, which was 10.5%. Comparison of nested models (M1 to M1a to M1b) was done using multiple partial F-tests, comparing a model with additional covariates to its preceding reduced form to evaluate the statistical significance of including additional covariates.

Statistical and regression analysis were completed in SAS (Cary, NC, version 9.4). IRB exempt status was obtained from the Emory University Institutional Review Board. Accounting for state-specific estimates of undiagnosed cases improved the fit of a simple linear regression model of new diagnoses to transmissions overall, with smaller variation from the number of new diagnoses in 2013 observed for 38 jurisdictions with the MMWR model (Figures 2 and 3).

Results

Data from all fifty states and Washington, D.C. were used. The proportion of prevalent cases attributed to MSM behavior ranged from 45.8% to 83.7%, with a mean percentage of 70.4%. The proportion of prevalent cases attributed to black individuals ranged from 3.3% to 75.1%, with a mean percentage of 34.1% (Table 1). On average, 15.2% of individuals were living in poverty, ranging from 9.7% to 23.8% by state, and 15.6% lived without any health insurance, ranging from 4.5% to 25.2% on a state by state basis.

The RShiny web tool, upon input of national continuum data, predicted an average of 759 transmissions per state (minimum=8, maximum=5489), which increased to an average of 809 transmissions per state (minimum=8, maximum=5195) upon accounting for variability in state-specific proportions of undiagnosed cases (Tables 1 and 2). Engagement in care, retention in care, and viral suppression percentages were calculated using a denominator of all diagnosed individuals in a particular state using a national average of 87.2% diagnosed (4). These state-level proportions varied from 0% to 25.2%, with a national average of 12.8% (Figure 1, Table 2).

After model selection, 11 different models were compared. In the covariates-only models, the adjusted R^2 for the all-covariates model was 0.4467, while the best covariates-only subset model had an adjusted R^2 of 0.4666 (Table 3). The addition of three covariates to the subset model did not result in a statistically significant increase in prediction (p=0.8109). The adjusted R^2 for the prevalence-only model (M1) was 0.8947, which marginally improved to 0.8949 and 0.9320 upon replacement by terms representing web-tool-predicted transmissions from the national continuum (M2) and the national continuum with state-level variation in undiagnosed (M3), which accounted for less than 1% and about 35% of the unexplained variation, respectively (Table 3). Although the prevalence model including a subset of covariates significantly increased the adjusted R^2 value to 0.9207 (M1a, $p<0.0001$), the addition of three covariates to the subset model was not significant ($p=0.9864$). In total, nearly 25% of the unexplained variation in the prevalence-only model was subsequently explained by the addition of covariates.

The adjusted R^2 for the model with national continuum-predicted transmissions and all possible covariates was 0.9170 (M3b), while the best subset model had an R^2 performance of 0.9208 (M2b). The subset model showed statistically significant increase in prediction over the predicted transmissions-only model (M2), with the five covariates explaining nearly an additional 25% of the variation unexplained in the prevalence-only model ($p<0.0001$). The addition of the other three covariates was not significant ($p=0.9861$).

The adjusted R^2 for the model with MMWR-predicted transmissions and all possible covariates was 0.9451 (M3b), while the best subset model had an R^2 performance of 0.9480 (M3a). This subset model contained only three covariates and showed a statistically significant increase in prediction over the predicted transmissions-only model (M3) ($p<0.0001$), and the addition of the five extra covariates to this subset model did not significantly improve its predictive ability ($p=0.8051$). Of the nearly 10% of state-level variation in diagnoses unexplained in a prevalence-only model, accounting for transmissions, state-level variability in the proportion of cases who are diagnosed, and a subset of demographic covariates improved the adjusted R^2 value to nearly 95%, a nearly 50% increase in predictive ability that correlated with a drop in AIC from 610.56 to 566.70.

The coefficient of a transmissions term, representing the change in new diagnoses associated with an increase of one predicted transmission, varied from 0.8771 to 0.9609 in these models and was a significant predictor in all models in which it was present. Notably, 95% confidence intervals for the transmissions coefficient for Models 2, 3, 3a, and 3b included a value of 1, implying that that a one-unit increase in the number of predicted transmissions could still potentially result in a one-unit increase in the number of new diagnoses, implying a 1:1 correspondence. This was not the case for the NA and either a subset of or all covariates models (M2a, M2b).

Discussion

Stable and valid incidence estimation is essential to designing prevention strategies that enhance serostatus awareness and link individuals newly infected with HIV to care. Approaches to provide estimation of sub-national HIV new HIV diagnoses, in lieu of incidence, can inform this goal. Renewed attention to, and research on, HIV-infected individuals and incident transmission gain growing importance under the lens of treatment as prevention (TasP), HIV prevention methods that use ART to suppress viral loads and thus reduce the risk of transmission (27-30).

The transmissions prediction model used in this analysis is unique in incorporating both biological and statistical likelihood of transmission from individuals at each level of the care continuum. Using standardized data sources for input data and for the model, this analysis accounts for the varying risk and numbers of transmissions at each level of the cascade, and the validation of this model using small-area estimation can provide an alternative methodology for incidence approximation. This analysis additionally demonstrates the feasibility of using publicly available surveillance and demographic data to accurately generate new diagnosis estimates in a 1-year horizon.

In these models, assuming that the proportion of diagnosed individuals in care is static, the number of predicted transmissions is predominantly driven by the proportion of individuals who are unaware of or undiagnosed with an HIV infection, as well as the proportion of those diagnosed who have achieved viral suppression and thus significantly reduced their risk of transmission. The percent of individuals retained in care who have suppressed viral loads is a key driver of transmissions prediction in this model when allowed to vary as well.

In this hierarchical modeling approach, important comparisons can be made between models. Modeling new diagnoses with solely demographic covariates managed to account for approximately 45% of the adjusted variation in new diagnoses. The MMWR-transmissions-only model had better fit than the National-Continuum-transmissions-only model, accounting for approximately 93% of adjusted variation compared to nearly 90%. Including prevalence was comparable to the alternative use of predicted transmission using the national average continuum, both accounting for nearly 90% of the variation in new diagnoses in simple linear regression models.

No as-of-yet discernable trend in the discrepancy between the number of predicted transmission and new diagnoses for particular jurisdictions, including by population size, geographic region, or any of the covariates, has been elucidated. Further study of other covariates, such as a quantification of delays in reporting, may shed light on why simple comparisons of transmission estimates to new diagnoses in states may over- or under- predict new diagnoses. In evaluating the utility of prevalence- or transmission-based incidence or diagnosis prediction, there may be more definitive value in determining a national or state-level diagnosis correction factor accounting for the discrepancy between incident transmissions and diagnoses, particularly that could account for state-level variation that is subsumed under a single coefficient for a transmissions term.

It is noteworthy that the best predictive model, the subset of the larger MMWR and covariates model, only included three additional covariates: the proportion of cases attributed to MSM, the proportion of cases who are black, and the proportion of individuals in a state who are uninsured. This more simplistic formulation of the model outperformed all other models in adjusted R^2, AIC, and BIC measures. Explaining nearly 95% of the variation in new diagnoses in 2013 with a limited number of input variables is an important step forward. Utilizing prevalence and social determinants of health data, alongside an estimate of transmissions, can allow for states to predict HIV burden for a subsequent year and aid in resource allocation and decision-making. It is important to note that Model 3a performs well in predicting the year following, and is better than a prevalence-only model, but further work needs to be done before extrapolation to further years can be accurately done.

The value of incorporating further covariates to prevalence or transmission simple linear regression models is demonstrated by the increase in adjusted R^2. Partial F-tests determined whether the addition of covariates provided statistically significant increases in predictive value, and, in all three sets of models incorporating prevalence and transmissions terms, a subset of covariates was preferred over the simple and fullycovariate-specified models. Adding covariates to the NA subset model (M2a) explained just as much variation as adding a subset of covariates (M1a) to the prevalence-only model, each accounting for 25% of the unexplained variation increase in adjusted R^2. Thus, the added predictive effect of adding covariates to a simple model was greater in the MMWR model (M3a) incorporating state-level variability in the undiagnosed proportion than in prevalence and NA models.

In evaluating multiple models for prediction of new diagnoses, it was important to account for prevalence in some way, whether through prevalent cases or a transmissions term. The coefficient of the transmissions term remained below one in each model that it was present in, highlighting that, in a given year, more transmissions are estimated to occur than diagnoses. However, 42,376 new diagnoses were made in 2013, and the national continuum model predicted 38,710 transmissions, while the MMWR model predicted 41,260 transmissions across the fifty states and Washington, D.C. The sum of transmissions from the state continua is less than the number of transmissions predicted from the national continua, likely due to the calibration of the initial biological transmission model, and accompanying web tool, to reflect an estimated 45,000 transmissions per year when the model was created to represent the 2009 national HIV care continuum. An adjustment or a re-calibration of the model to a 2012 continuum would likely yield the expected case in which transmissions are predicted to outnumber diagnoses and address a limitation of the methods presented here. These findings illustrate the continued disparity between reported new diagnoses and actual infections, as well as a growing understanding of the burden of transmissions attributable to individuals at all levels of the HIV care continuum.

The absence of appropriate nested state-specific continua hindered the fulfillment of one of the original main aims of this analysis. State HIV surveillance reports are published annually and some states publish a continuum of care, but it was observed that these continua do not align with previously published continua, as each subsequent level was not necessarily nested within a prior level of the continuum. As an example, the number of individuals experiencing viral suppression was larger than the number of individuals engaged or retained in care for some jurisdictions. The web tool and its source model, in its current format, are not able to evaluate non-nested continuum data, preventing the usage of state-specific continua as parameters for transmission prediction. Continued work in characterizing state-specific continua can improve the predictive ability of the tool, accounting for important differences in viral suppression and other levels that may influence care and transmission.

State-level or jurisdictional estimates along steps of the HIV care continuum can differ greatly from NHSS estimates, indicating a large variation in figures across subnational areas, but there is uncertainty about whether this variation can be attributed to data quality or true differences (31, 32). The comparison of data across jurisdictions is limited, as these jurisdictions are discouraged from reporting prevalence and encouraged to use diagnoses as their continuum denominator (33). Additionally, diagnoses reported by local jurisdictions are reliant on resource allocation, lab supply availability, and other factors such as in-migration and out-migration that can result in a misclassification of the proportion of individuals diagnosed (33). Diagnoses data used in this analysis are taken from CDC, having been adjusted for reporting delays and missing data, and may differ from state-reported diagnoses which have not been adjusted.

The validity of estimates along each subsequent level of the care continuum is threatened by the potential of misclassification or measurement error of the baseline number of estimated prevalent and incident infections (33). Estimates and measures present in the HIV care continuum for a particular jurisdiction can vary due to the source of data used, either MMP or NHSS. This key source of bias may also influence reporting on both national and sub-national levels. A structural reason for this might be the reliance on data from the MMP, which uses narrow definitions of retention and other continuum categories, such as retention signifying at least one visit to a provider in the first four months of the calendar year. As the number of individuals infected with HIV reflects an estimated prevalence among diagnosed and undiagnosed individuals, it is possible that the cascade category estimates may undervalue the numbers of individuals truly present at each level (8, 34).

Quantitative input data, such as the proportion of individuals who engage in cross-border sexual relations or who migrate in and out of a particular jurisdiction in a particular year influence the proportion of individuals who would be diagnosed with HIV in a jurisdiction of interest. This analysis assumes that migration is balanced, with the inflow matching the outflow for a particular state. Although already incorporated into a national-level model, jurisdictional breakdown and quantification of testing, sexual behavior, and migration are vital for this analysis of smaller areas. These data are not readily accessible or available, and may require review of extra-disciplinary research and novel estimation efforts.

A potential route forward may be the adoption of yearly state-to-state migration flows estimated through the American Community Survey as a proxy for testing, but the correlation between moving residences and testing practices is not likely to be perfect (35). When relying on small area estimation or attempting to generalize the approach to as-of-yet unstudied or international contexts, parameterization of migration and cross-border population trends is vital to valid estimation. Competent case-based and disease surveillance must be supported by attention to demographic trends that might affect the likelihood to get tested, remain in care, and suppress one's viral load in a specific location.

A lack of data about the number of individuals retained in care who received ART for the state jurisdictions included in this study leaves one of the continuum levels unaddressed in this analysis. Above all, the dynamic nature of the cascade for individuals is not addressed by this analysis (34). There are delays in reporting in a given year, as not all incident cases may be diagnosed in the year in which HIV is acquired. As a snapshot of state jurisdictions at the end of a particular year, the model and analysis do not fully address situations in which an individual retained in care and with a suppressed viral load might fall out of care and cease to have a suppressed viral load.

Arresting transmission and preventing HIV acquisition are vital goals, worthy of sustained and increased focus, research, and resources. The proportion of individuals who are undiagnosed when the continuum is stratified may differ vastly, as awareness of infection may differ by age, race or ethnicity, or a number of other factors. Thus, this web-tool-based approach can align with continued research into stratified continua. Finer stratification of HIV care continua is desired and can be supported by this web-tool-based approach, provided that necessary parameters can be calculated. Greater stratification can support prevention efforts that focus on particular groups, such as increasing testing in younger individuals.

The National HIV/AIDS Strategy has set objectives to reduce the number of new infections and the proportion of those unaware of their infection, as well as to boost linkage to, and retention in care among those diagnosed. Thus, ultimately, a continuum-based approach to estimation and response is important, as boosting the engagement of individuals at each level of the care continuum in prevention in treatment is the key to addressing the burden of HIV and achieving the goal of zero new infections. Given that two-thirds of new infections are estimated to result from those with known HIV infection, the need for data to support comprehensive HIV prevention and treatment packages to halt the spread of HIV in state and local jurisdictions is great. Meaningful prediction was accomplished with a limited amount of local data, and the incorporation of further information from state and local levels can allow for refinement of models and methods to maximize the predictive value for HIV diagnoses.

References

1. Monitoring selected national HIV prevention and care objectives by using HIV surveillance data - United States and 6 dependent areas - 2013. HIV Supplemental Surveillance Report. Vol. 20 (No. 2). Centers for Disease Control and Prevention; 2015.

2. Monitoring selected national HIV prevention and care objectives by using HIV surveillance data - United States and 6 dependent areas - 2012. HIV Supplemental Surveillance Report. Vol. 19 (No. 3). Centers for Disease Control and Prevention; 2014.

3. National HIV/AIDS Strategy for the United States: Updated to 2020. Office of National AIDS Policy (ONAP). Washington, D.C.: The White House; 2015.

4. Hall HI, Song R, Rhodes P, Prejean J, An Q, Lee LM, et al. Estimation of HIV incidence in the United States. JAMA 2008;300(5):520-9.

5. National HIV Prevention Progress Report. Centers for Disease Control and Prevention; 2015. www.cdc.gov/hiv/pdf/policies/progressreports/cdc-hiv-nationalprogressreport.pdf.

6. Fellows IE, Morris M, Birnbaum JK, Dombrowski JC, Buskin S, Bennett A, et al. 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 2015;10(7):e0129551.

7. Hall HI, An Q, Tang T, Song R, Chen M, Green T, et al. Prevalence of Diagnosed and Undiagnosed HIV Infection--United States, 2008-2012. MMWR Morb Mortal Wkly Rep 2015;64(24):657-62.

8. Skarbinski J, Rosenberg E, Paz-Bailey G, Hall HI, Rose CE, Viall AH, et al. Human immunodeficiency virus transmission at each step of the care continuum in the United States. JAMA Intern Med 2015;175(4):588-96.

9. Exec. Order No. 13649. 3 C.F.R.; 2013. p. 43057-43059.

10. Gardner EM, McLees MP, Steiner JF, Del Rio C, Burman WJ. The spectrum of engagement in HIV care and its relevance to test-and-treat strategies for prevention of HIV infection. Clin Infect Dis 2011;52(6):793-800.

11. What is the HIV Care Continuum? AIDS.Gov. www.aids.gov/federal-resources/policies/care-continuum/

12. Monitoring HIV Care in the United States: Indicators and Data Systems. Washington, D.C.: Institute of Medicine; 2012.

13. Blair JM, McNaghten AD, Frazier EL, Skarbinski J, Huang P, Heffelfinger JD. Clinical and behavioral characteristics of adults receiving medical care for HIV infection --- Medical Monitoring Project, United States, 2007. MMWR Surveill Summ 2011;60(11):1-20.

14. McNaghten AD, Wolfe MI, Onorato I, Nakashima AK, Valdiserri RO, Mokotoff E, et al. Improving the representativeness of behavioral and clinical surveillance for persons with HIV in the United States: the rationale for developing a population-based approach. PLoS One 2007;2(6):e550.

15. Dean HD, Fenton KA. Integrating a social determinants of health approach into public health practice: a five-year perspective of actions implemented by CDC's national center for HIV/AIDS, viral hepatitis, STD, and TB prevention. Public Health Rep 2013;128 Suppl 3:5-11.

16. Forsyth AD, Valdiserri RO. A State-Level Analysis of Social and Structural Factors and HIV Outcomes Among Men Who Have Sex With Men in the United States. AIDS Educ Prev 2015;27(6):493-504.

17. Gant Z LM, Hall HI, Hu X, Guo X, Song R. A County-Level Examination of the Relationship Between HIV and Social Determinants of Health: 40 States, 2006-2008. The Open AIDS Journal 2012;6:1-7.

18. Zeglin RJ, Stein JP. Social determinants of health predict state incidence of HIV and AIDS: a short report. AIDS Care 2015;27(2):255-9.

19. Understanding the HIV Care Continuum. Centers for Disease Control and Prevention; 2014. www.cdc.gov/hiv/pdf/DHAP_Continuum.pdf

20. AIDSVu. In: Rollins School of Public Health and Gilead Sciences, Inc.; 2016. www.aidsvu.org

21. U.S. Census Bureau. Small Area Income and Poverty Estimates, Table 1: 2012 Poverty and Median Income Estimates - States. 2013.

22. U.S. Census Bureau. Population Estimates, Vintage 2012 [entire data set]. In; 2012.

23. U.S. Census Bureau. American Community Survey 1-Year Estimates, 2012, Table C15003: Educational attainment for the population 25 years and over – States. In; 2013.

24. U.S. Census Bureau. Small Area Health Insurance Estimates, 2012: Health Insurance Coverage Status by Age, Race, Hispanic Origin, Sex and Income for Counties and States. 2012.

25. U.S. Census Bureau. American Community Survey 1-Year Estimates, 2012, Table B19083: Gini Index of Income Inequality. 2012.

26. HIV Care Continuum for the United States and Puerto Rico. Centers for Disease Control and Prevention. http://www.cdc.gov/hiv/pdf/Continuum_Surveillance.pdf

27. Cohen MS, Chen YQ, McCauley M, Gamble T, Hosseinipour MC, Kumarasamy N, et al. Prevention of HIV-1 infection with early antiretroviral therapy. N Engl J Med 2011;365(6):493-505.

28. Granich RM, Gilks CF, Dye C, De Cock KM, Williams BG. Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model. Lancet 2009;373(9657):48-57.

29. Montaner JS. Treatment as prevention--a double hat-trick. Lancet 2011;378(9787):208-9.

30. Montaner JS, Hogg R, Wood E, Kerr T, Tyndall M, Levy AR, et al. The case for expanding access to highly active antiretroviral therapy to curb the growth of the HIV epidemic. Lancet 2006;368(9534):531-6.

31. Gray KM, Cohen SM, Hu X, Li J, Mermin J, Hall HI. Jurisdiction level differences in HIV diagnosis, retention in care, and viral suppression in the United States. J Acquir Immune Defic Syndr 2014;65(2):129-32.

32. Dombrowski JC, Buskin SE, Bennett A, Thiede H, Golden MR. Use of multiple data sources and individual case investigation to refine surveillance-based estimates of the HIV care continuum. J Acquir Immune Defic Syndr 2014;67(3):323-30.

33. Lesko CR, Sampson LA, Miller WC, Clymore J, Leone PA, Swygard H, et al. Measuring the HIV Care Continuum Using Public Health Surveillance Data in the United States. J Acquir Immune Defic Syndr 2015;70(5):489-94.

34. Giordano TP. The HIV treatment cascade--a new tool in HIV prevention. JAMA Intern Med 2015;175(4):596-7.

35. American Commmunity Survey. State-to-State Migration Flows. U.S. Census Bureau; 2012.

Table 1: Summary Statistics for Covariates

Variable	Mean	SD	Minimum	Maximum
Transmissions (MMWR)	809.0	1209	8	5195
Transmissions (National Average)	759.0	1196	8	5489
Proportion of HIV Cases attributed to MSM behavior	0.7044	0.0873	0.4580	0.8370
Proportion of HIV cases attributed to Black Individuals	0.3413	0.2217	0.0332	0.7511
Gini Coefficient	0.4592	0.0224	0.4166	0.5343
Proportion with a High School Education	0.8801	0.0312	0.8140	0.9280
Proportion living in Poverty	0.1524	0.0326	0.0970	0.2380
Proportion Uninsured	0.1555	0.0476	0.0450	0.2520
Median Household Income	51937	8564	37179	71169

SD = Standard deviation

Table 2: HIV Care Continuum Parameters for Transmissions Predictions

Parameter	National Average Continuum Model	MMWR Model
Percent Diagnosed	87.2	*State-specific (Supp. Table 2)*
Percent of Diagnosed Engaged in Care	44.8	44.8
Percent of Diagnosed Retained in Care	41.5	41.5
Percent of Diagnosed with Suppressed Viral Load	34.6	34.6

Table 3: Candidate Models

Model	Adjusted R^2	R^2	AIC	BIC	Root MSE	Coefficient of Transmissions	95% CI	Adjusted R^2 % gained over Model 1	Partial F-test P-value (Compared to preceding model)
M0: MSM Proportion, Gini Coefficient, Education, Poverty, Uninsured	0.4666	0.5210	683.9766	688.0566	883.3155				---
M0a: All covariates	0.4467	0.5257	687.1797	692.4526	899.6018				0.8109
M1: Prevalence Only	0.8947	0.8968	610.5626	612.7226	390.2139			---	---
M1a: Prevalence, Black Proportion, MSM Proportion, Education, Poverty, Uninsured	0.9207	0.9304	589.5045	594.4706	340.5140			24.69%	<0.0001 (5 df)
M1b: Prevalence + All Covariates	0.9169	0.9305	593.4710	599.3259	348.6038			21.08%	0.9864 (2 df)
M2: National Continuum	0.8949	0.8970	610.4995	612.6595	389.9727	0.9517	(0.8591, 1.0443)	0.19%	---
M2a: National Continuum, Black Proportion, MSM Proportion, Education, Poverty, Uninsured	0.9208	0.9305	589.4535	594.4194	340.3406	0.8771	(0.7793, 0.9749)	24.79%	<0.0001 (5 df)
M2b: National Continuum + All Covariates	0.9170	0.9306	593.4195	599.2743	348.4240	0.8744	(0.7602, 0.9886)	21.18%	0.9861 (2 df)
M3: MMWR Undiagnosed	0.9320	0.9334	588.2485	590.4084	313.5406	0.9609	(0.8872, 1.0346)	35.42%	---
M3a: MMWR Undiagnosed, Black Propoprtion, MSM Proportion, Uninsured	0.9480	0.9522	566.6985	570.3719	275.7971	0.9298	(0.8586, 1.0010)	50.62%	<0.0001 (4 df)
M3b: MMWR Undiagnosed +All Covariates	0.9451	0.9540	572.7689	578.6237	283.4156	0.9160	(0.8214, 1.0106)	47.86%	0.8051 (4 df)

Covariates: Proportion of cases attributable to black individuals (Black Proportion), Proportion of cases attributable to MSM behavior (MSM Proportion), Proportion of individuals lacking health insurance (Uninsured), Proportion of individuals with a high school education (Education), Proportion living in poverty (Poverty), Gini coefficient of income inequality (Gini), Median household income (Income)

AIC = Akaike's Information Criterion BIC = Bayesian Information Criterion MSE = Mean Squared Error df = degrees of freedom for multiple partial F-test

Figure 1: Percent of Cases Undiagnosed by State

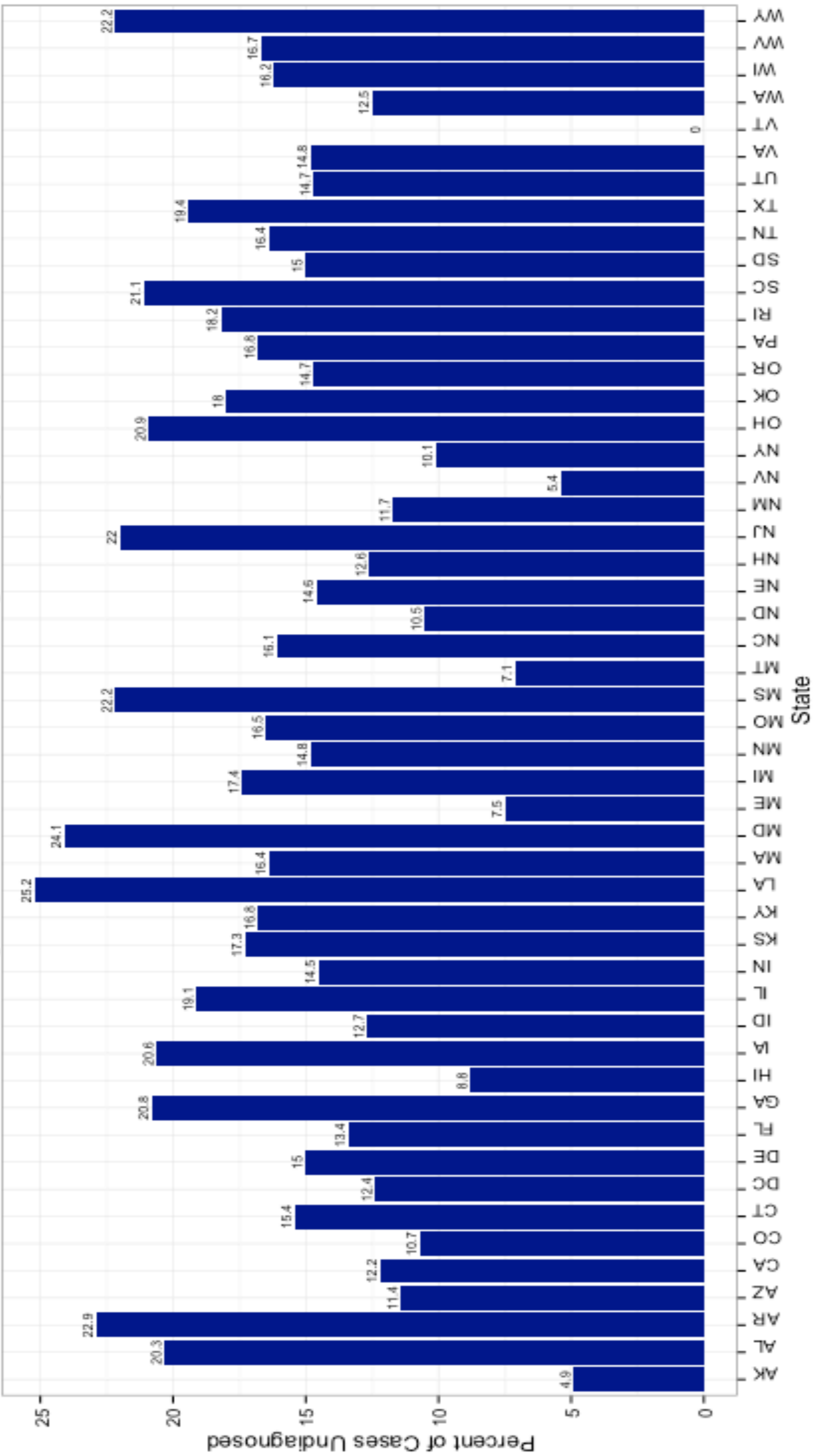


Figure 2: Transmission Prediction

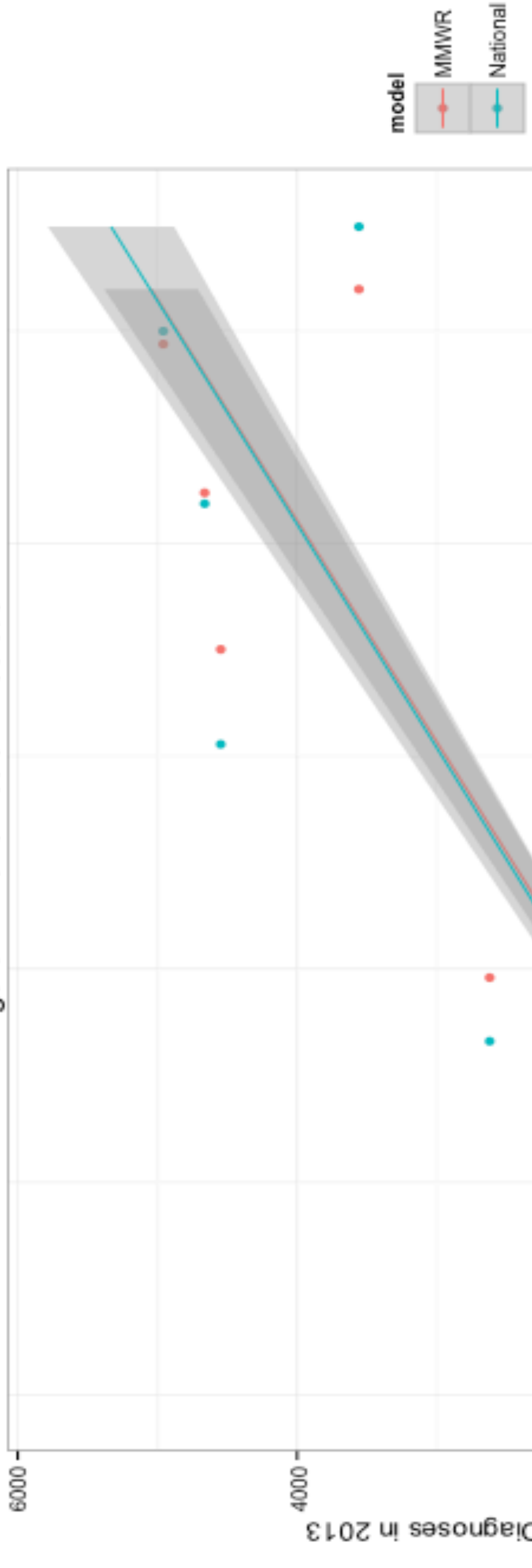
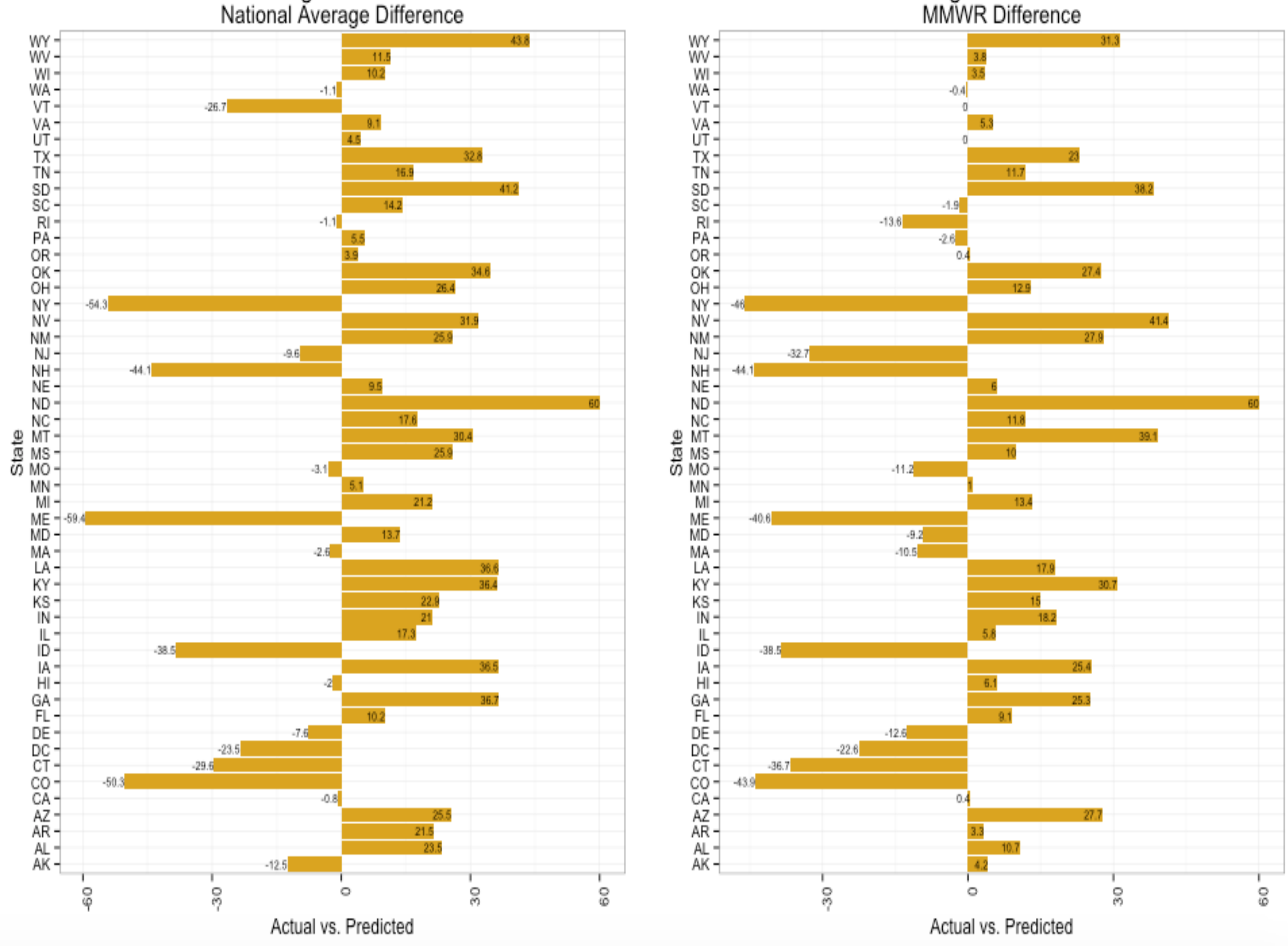


Figure 3: Difference between Predicted Transmissions and Actual Diagnoses


National Average Difference

MMWR Difference