Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

______________________________ _______________

Noël V. Hatley Date

Using County-Level Socio-Demographics to Estimate HIV Diagnoses in Maryland, North Carolina and Virginia

By

Noël V. Hatley

Degree to be awarded: Master of Public Health

Epidemiology

____

Travis Sanchez, DVM, MPH

Committee Chair

Using County-Level Socio-Demographics to Estimate HIV Diagnoses in Maryland, North Carolina and Virginia

By

Noël V. Hatley

Bachelor of Science
Boston University
2010

Thesis Committee Chair: Travis Sanchez, DVM, MPH

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 Epidemiology
2014

Abstract

Using County-Level Socio-Demographics to Estimate HIV Diagnoses in Maryland, North Carolina and Virginia
By Noël V. Hatley

Background: The rate of new HIV diagnoses has ceased to increase over the last decade, yet those rates have remained stable but not decreasing. Paramount to HIV prevention is the accurate and systematic surveillance systems capturing timely reports of new HIV diagnoses. This study seeks to determine whether differences between county-level reported new HIV diagnoses can be explained by demographic factors.

Methods: Using publicly available HIV diagnosis data from 2008-2011 and socio-demographic factors potentially associated with HIV, we created models stratified by Maryland and North Carolina combined and Virginia. We used coefficients from the models to estimate new HIV diagnoses in Virginia and Maryland/North Carolina. We mapped the reported diagnoses and visually compared with mapped expected diagnoses.

Results: The 134 counties of Virginia had 4,466 new HIV diagnoses from 2008-2011, with 3,804 occurring in 29 counties with 20 or more total cases (unsuppressed). The 122 counties of Maryland/North Carolina combined had 14,400 cases with 13,854 occurring in 70 unsuppressed counties. Our two final reduced models fit Maryland/North Carolina and Virginia respectively. After mapping the expected HIV diagnoses, we found that diagnoses in 6 counties on the Virginia border with North Carolina had the largest differences. These counties had reported between 0 and 20 diagnoses (suppressed counties), but after estimation have 20-50 HIV diagnoses each.

Conclusion: After accounting for social demographic characteristics of the counties, the Virginia socio-demographic data predicted different diagnoses counts than what was reported, indicating the significance of unmeasured factors. We hypothesize that the unmeasured factors are underreporting or HIV diagnosis issues. Moving forward, additional research should be conducted to assess the extent of reporting bias and determine the steps to mediate the problem.

Using County-Level Socio-Demographics to Estimate HIV Diagnoses in Maryland, North Carolina and Virginia

By

Noël V. Hatley

Bachelor of Science
Boston University
2010

Thesis Committee Chair: Travis Sanchez, DVM, MPH

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 Epidemiology
2014

ACKNOWLEDGEMENTS

I would like to extend my gratitude to Dr. Travis Sanchez, my thesis faculty advisor, for his exceptional support, guidance, and patience with me throughout this process. A special thanks to my parents, whom I love dearly, for supporting me and my goals, every step of the way. I could not have made it this far without them. Lastly, I would like to dedicate this thesis in the memory of my mother who passed away far too soon and just days before its completion.

TABLE OF CONTENTS

CHAPTER 1: LITERATURE REVIEW

Since its discovery over 30 years ago, HIV continues to be a public health problem with an estimated 1.1 million people living with HIV in the United States (1, 2). Even though the yearly number of new diagnoses in the past decade has remained stable (around 50,000), the number of new infections among young persons, especially younger black men has increased (1, 3, 4). Most troubling are the estimated 180,000 people (among the 1.1 million living with HIV) unaware of their infection (5). Many of those unaware of their infection remain undiagnosed until they present with AIDS-related conditions (6). Additionally, only 37% of the population aged 18-64 report ever receiving an HIV test, an estimate that varies by state from 23.4% to 66.3% (7).

HIV in the South

The Southern states are known for having the worst health in the nation on many health indicators, including infant mortality, heart disease and diabetes (8, 9). They also have high rates of HIV infection. From 2008 to 2011, the rate of new HIV diagnoses in the Southern U.S. remained between 19.5 and 22.4 per 100,000 people, the highest in the nation (1). The Southern states accounted for between 48% and 50% of all new HIV diagnoses and represented only 37% of the entire U.S. population in the same time-period (1, 10). The southern states also have the highest HIV-related mortality rates in the country, accounting for half of all HIV-related deaths in 2008-2011 (1, 8, 10).

Factors that may contribute to the high rates of new HIV diagnoses in the South include rates of other sexually transmitted diseases, poverty rates, race/ethnicity and stigma (2, 8, 10, 11). The southern states are disproportionately affected by sexually transmitted diseases. In 2009, nine of the 10 states with the highest syphilis rates were in the South. STDs have consistently been found to increase the risk of HIV transmission (2, 9, 10). Poverty is also highest in the South, where nine of the 10 states with the lowest median incomes were in the South (10). The states with the highest HIV case-fatality rates also had the lowest incomes. The high rates of disease and poverty may impact the way states respond to health issues, like HIV. With limited resources, the HIV epidemic cannot be adequately addressed, perpetuating the continued high rates of HIV diagnoses.

HIV infection rate differences by race/ethnicity are signs of more complex issues associated with race. Black/African Americans are disproportionately affected by HIV in the US and even more so in the South (8). African Americans also have a poverty rate twice that of Whites (8, 10). Black/African Americans also face poorer health care access, even after controlling for income (8, 10). Many have theorized that unstable housing, higher rates of incarceration, lack of trust in health care and government, and HIV-related stigma issues contribute to the higher rate of HIV disease among African Americans (10). Blacks in the South are not the only group disproportionately affected. Hispanics/Latinos are also disproportionately affected, with over half of the new diagnoses among Hispanic/Latinos occurring in the South (1, 5, 10).

Many of the laws and policies in the Southern states have been connected with the continued spread of HIV. For instance, many southern states have abstinence only programs in school, which are ineffective in STD prevention (11). Also common in the South are laws criminalizing HIV behaviors and prohibiting the exchange of syringes, which further marginalize people at high risk and discourage HIV testing (8, 11).

HIV Testing

HIV testing is the cornerstone of current HIV prevention strategies in the United States, especially testing at earlier stages of disease. Researchers have found that minorities, women, heterosexuals, young people and people with low education had less frequent early detection of HIV (12). For the entire U.S. population, the rate of late HIV diagnoses (AIDS diagnosis occurring within 12 months of initial HIV diagnosis) was 32% in 2010. Late HIV testing occurred among 29.1% of new HIV diagnoses in Maryland, 30.2% in Virginia, and 27.4% in North Carolina (13).

Many studies have found that late HIV testing and diagnosis is most significant among older (older than 30 years old), heterosexual males. Most HIV prevention and testing interventions do not specifically target heterosexual males, making opportunities for early HIV diagnoses less than among injection drug users (IDU), men who have sex with men (MSM), and women (14). HIV treatment works most effectively when accessed early, placing a substantial amount of importance on early HIV testing (14).

The Centers for Disease Control and Prevention (CDC) estimates that approximately 1.1 million people are living with an HIV infection (15). At the end of 2008, 20% of the estimated 1.1 million people living with HIV were undiagnosed and unaware of their infection. In order to increase HIV testing and promote early detection of HIV infection, in 2006 the CDC recommended routine screening for all patients aged 13-64 years in health-care settings (15-17).

HIV Reporting

Reporting cases of HIV is required in all States of the U.S. Each state is responsible for collecting HIV surveillance data based on CDC recommendations and reporting that data to the CDC. All states, Washington D.C. and five U.S. dependent areas were using confidential name-based reporting by April 2008 (12, 18). The accuracy and completeness of reporting varies from state to state, despite following recommended guidelines. Accurately collecting and reporting HIV surveillance data is a vital public health procedure. The allocation of federal funds for HIV prevention and care, such as those from the Ryan White Care Act, relies heavily on surveillance data (19-21). This in turn affects the availability and ease of access to testing and treatment, especially in rural areas. Underreporting is more likely to occur in rural areas with lower HIV incidence due to inefficient surveillance infrastructure, HIV testing and treatment availability (22, 23).

Assessing the completeness of reporting HIV diagnoses has been assessed using various techniques, including capture-recapture methods (16). One study completed during 2002-2004 estimated completeness of reporting of HIV infections diagnosed within a one-year period and reported up to six months after initial diagnosis was on average 76%, ranging from 72% to 95% (16). Additionally, 32%-78% of reports were from laboratories (ranges depend on reporting site), with the next most common source of reports from outpatient and inpatient facilities. Approximately 39% of HIV diagnoses were reported by two or more sources. The CDC requires a minimum performance standard of completeness of HIV reporting greater than or equal to 85% for states (17). Even though reporting completeness is quite high in various areas, there is always room for improvement.

Even when the CDC expanded the recommendations for yearly testing among 13-64 year olds, the date of HIV diagnoses has been found to vary considerably between sources including self-report, medical record and surveillance data (17). Medical record documentation is widely considered the gold standard of HIV diagnoses among the medical community, and yet one study found that the diagnosis date in surveillance systems occurs on average 9 months after the patient self-reported diagnosis date and medical record diagnosis date (24). Furthermore, researchers found that of all HIV infected patients from 2000-2008 in a large North Carolina HIV-STD clinic based in a large academic hospital setting only 81% were successfully matched to records in the North Carolina HIV surveillance. Some may have been from out of state and simply not updated in North Carolina's system, however 51% were diagnosed before 1995 when anonymous testing was still available (24).

Social Determinants of HIV

HIV disproportionately affects minority populations, including Blacks/African Americans, and Hispanics/Latinos. In 2009, While Blacks represented 12% of the population, they constituted 44% of the new HIV diagnoses (25). While Latinos represented 16% of the population, they made up 20% of the new HIV diagnoses (25, 26). Blacks/African Americans are at a significantly higher risk of morbidity and pre-mature mortality as compared to Whites. Socioeconomic status (SES) can account for much of the difference, however racial/ethnic disparities continue to persist after adjusting for SES (27).

The distribution of income is a key determinant of health. As income inequality increases, residential concentrations of affluence and poverty increase, creating residential segregation and diminishing social cohesiveness (28). Consequently, this increases inequalities in many societal factors including access to health care, crime and violence, economic growth, and health indicators (28, 29). Included in those factors are HIV rates and stigma. Among low-income men and women living with HIV, the most perceptible spheres of social stigma included blame and stereotypes of HIV, fear of contagion, disclosure of a stigmatized role and readjusting social status and integration (30, 31).

Many have found education to be a significant factor in HIV morbidity and mortality: as educational attainment increases, the rate of HIV mortality decreases (30, 32). Education has been found to be so strongly predictive of safer behavior and reduced infection rates that it has been described as the social vaccine and one of the most effective weapons against HIV (33).

As the HIV epidemic has progressed, rural people are affected more than ever before. The southern states have the highest percentage (27%) of HIV-infected individuals living in rural areas as compared with other geographic areas (23). Additionally, people with HIV in rural settings are more likely than their urban counterparts to be diagnosed at a later stage of disease, suggesting missed opportunities for HIV testing (22, 34).

As illustrated above, the issues surrounding new HIV diagnoses are expansive and varying. Variations and issues with HIV reporting (excluding reporting completeness) have been largely unexplored among the States. State HIV reporting mechanisms are another aspect of the complex HIV epidemic in the U.S. As all states are participating in confidential name-based reporting methods, state-based reporting inaccuracies can be assessed and compared.

CHAPTER 2: MANUSCRIPT

INTRODUCTION

In the United States, more than 50,000 people are newly diagnosed with HIV every year with around 49% of those new diagnoses occurring in the South (1). More than 180,000 additional people are unaware of their HIV infection. With the passage of the National HIV/AIDS Strategy in 2010, the United States is focusing on reducing those numbers by focusing on three overarching goals: reduce new HIV infections; increase access to care and improve health outcomes; and reduce HIV-related health disparities (35).

Paramount to HIV prevention is the accurate and systematic surveillance systems capturing timely reports of new HIV diagnoses. Though the Centers for Disease Control and Prevention has a uniform case surveillance definition and report form that all 50 states, the District of Columbia, and 6 U.S. dependent areas currently follow, each state is responsible for collecting that data, allowing for variations. Variations in what is reported can occur due to missed diagnoses, delayed reporting and reporting completeness. Researchers have found that the average time difference between the diagnoses date in the electronic HIV/AIDS Reporting System and the diagnosis date in medical record (often considered the gold standard for accurate diagnosis date) is approximately 9 months (24).

While mapping the 2008-2011 counts of new HIV diagnoses aggregated by county, Virginia appeared to be markedly different from the neighboring states Maryland and North Carolina (figure 1). Where Maryland and North Carolina were reporting that

56% of their counties had more than 20 cases each, only 21% of Virginia's counties had more than 20 cases each (AIDSVu.org). This variation is also not entirely explained by differences in HIV testing rates between states because 45.6% of Maryland adults, 42.2% of North Carolina adults and 41.3% of Virginia adults report ever having an HIV test (7). While the completeness of reporting has been studied (16-17), little investigation has been done on reporting biases and estimating how many undiagnosed people are in Virginia.

Using county-level socio-demographics known to be associated with HIV diagnosis and prevalence, we modeled the counts of HIV diagnoses from 2008-2011 stratified by state to determine whether the variation in reported HIV diagnoses between the states can be accounted for by these factors. We used those factors that may account for differences in HIV cases to then project the expected number of new HIV diagnoses. The differences between the observed and expected cases may indicate issues with unmeasured factors, such as surveillance case reporting and HIV testing.

METHODS

Data

We used publicly available data to create statistical models of county-level new HIV cases as a function of social determinants stratified by Maryland and North Carolina combined, and Virginia. We combined Maryland and North Carolina HIV case counts and socio-demographic data as the comparison group because they share borders with Virginia, are also coastal states with large cities, are considered part of the Southern U.S., and have similar geographically distributed demographics (Table 1). We were particularly interested in the Virginia-North Carolina border, a political boundary not determined by any natural geographic boundaries, such as mountains or rivers, and where one would expect similar distributions of HIV cases.

We included all Maryland, Virginia and North Carolina county level new HIV diagnosis counts among persons ages >13 years from 2008 through 2011. HIV counts were obtained from national new HIV diagnosis data (Centers for Disease Control and Prevention, presented through AIDSVu.org). To maintain confidentiality, the CDC suppressed newly diagnosed HIV counts of 0 to 20 at the county level prior to release. Part of this analysis estimated the number of new HIV diagnoses in all counties of Virginia, including the suppressed counties. Maryland and North Carolina had 124 counties, of which 70 (56.5%) were unsuppressed and included in the analyses. Virginia had 134 counties of which 29 (21.6%) were unsuppressed and included in the analyses.

County-level estimates of socio-demographic covariates, including population density, total population, housing density, median age, race, sex, population in prison, income inequality (Gini coefficient), population over 25 with a high school diploma, population of male same-sex households, proportion of people without health insurance, median income and the proportion living in poverty were obtained from the United States Census Bureau for 2008 through 2011. Average estimates over the four-year observation period for each covariate were calculated. County level drug use data was obtained from the Substance Abuse and Mental Health Services Administration (SAMHSA). Each county within the SAMHSA-defined sub-state region was assigned the same value (percent of population age 12 or older who used an illicit drug other than marijuana in the past month).

Normality was assessed using histograms of each covariate and Kolmogorov-Smirnov statistics. Due to non-normal distributions of the total population, population density, housing density, median income, rate of male same-sex couples living together, and the population in prison rate were transformed by taking the log of each covariate. Race/ethnicity was also transformed into the log of the rate of Black/African Americans compared to the rate of Whites, the log of the rate of Hispanics compared to the rate of Whites and the log of the rate of all other races compared to the rate of Whites.

The numbers of people living in poverty, people over 25 with a high school diploma, people living without health insurance, people with drug and/or alcohol dependence and people with past 30 day drug use (excluding marijuana), were normally distributed and not transformed. Age remained median age per county. Gini remained the average gini coefficient for each county. Sex was modified to the ratio of males to females for each county.

Description of Analyses

Exploration of the data compared Virginia covariates with Maryland/North Carolina covariates using two sample t-tests. Additionally, we completed simple linear regressions of the dependent variable, new HIV diagnoses, with each of the covariates. Pearson correlation coefficients were calculated for each state at the 0.05 significance level to assess linear associations between HIV and each covariate.

The outcome variable, number of new HIV diagnoses from 2008-2011, was over-dispersed (mean=178, variance= 135247.16) so the negative binomial was chosen as the most appropriate distribution for the model. We developed two stratified negative binomial linear regressions to assess covariate differences by state. The first model, which we call the full model, included all covariates in the models stratified by state. The second model, called the reduced model, included only those covariates significantly correlated with HIV in either Virginia or Maryland/North Carolina. As stated previously, the goal of this step of the analysis was to determine which covariates account for the variation of the reported new HIV diagnoses.

If the significant covariates differed between states then we projected estimated HIV diagnoses weighted by the covariate coefficients in the reduced model of Maryland/North Carolina. If the significant covariates did not differ between the states then we projected estimated HIV diagnoses weighted by the covariate coefficients in the reduced model of Virginia. One map was created to show the reported HIV diagnoses. Two additional maps displaying the distribution of cases weighted by the two models (VA reduced model, and MD/NC reduced model) were created. For all tests, significance was determined using a two-sided p-value at the 0.05 level.

All analyses were conducted using SAS version 9.3 for Windows (SAS Institute Inc., Cary NC). This analysis used summarized county-level data and was therefore not considered to be research involving human subjects.

RESULTS

County-Level Characteristics

The 134 counties of Virginia had 4,466 cases of new HIV diagnoses from 2008-2011 with 3,804 occurring in 29 counties with 20 or more new cases total (unsuppressed counties, Table 1). In the 29 unsuppressed counties of Virginia, the mean number of cases was 131 cases (Standard Deviation [SD] 131.5) and median number of cases was 78 (Interquartile Range [IQR] 159). The 122 counties of Maryland and North Carolina combined had 14,400 cases with 13,854 occurring in 70 unsuppressed counties during the same time period. The 70 counties of Maryland/North Carolina had on average 198 cases per county (SD 428.7) and a median of 56 cases per county (IQR 69).

The following covariates were significantly different between counties in Virginia and Maryland/North Carolina: Age, race/ethnicities categorized as Other, people living in poverty, past month drug use, drug dependence, high school graduate, people without health insurance, median income, population density and housing density (Table 1).

Correlation Analyses

Correlation analyses revealed similarities and differences in the county-level factors correlated with HIV diagnoses by state (Table 2). For Virginia and Maryland/North Carolina total population, population density, housing density, ratio of Hispanics to Whites, ratio of Other race to Whites and male-male households were significant and positively correlated with the distribution of HIV diagnoses. Rates of past month drug use were significant and negatively correlated with the distribution of HIV diagnoses in Virginia and Maryland/North Carolina. In Maryland/North Carolina alone, county-level covariates significant and positively correlated with HIV diagnoses included ratio of Blacks to Whites and covariates significant and negatively correlated with HIV diagnoses included ratio of males to females and median age. There were no covariates significantly correlated in Virginia alone.

Multivariate Analyses

Neither full model fit the data well (data not shown), but the reduced models for Virginia and Maryland/North Carolina fit the data well (Table 3a). Significant covariates in the reduced model for Virginia included sex, total population, population density, housing density, ratio of Blacks to Whites and past month drug use. Significant covariates in the reduced model for Maryland/North Carolina included sex, total population, ratio of Blacks to Whites, and ratio of Other races to Whites were statistically significant (Table 3b).

Projection of Expected HIV Counts

Since the reduced model for Maryland/North Carolina fit the data well, the observed covariates in Virginia and Maryland/North Carolina were weighted by the coefficients of that model to estimate one set of expected counts of HIV per county. The reduced model for Virginia also fit the data well, so a second set of expected counts of HIV were calculated by weighting Virginia and Maryland/North Carolina covariates by the coefficients of the reduced model for Virginia. The maps of county-level HIV counts were created to visually compare the observed HIV diagnoses and the two sets of expected HIV diagnoses (Figures 1 and 2). Both maps of expected HIV diagnoses depict changes in reported HIV diagnoses in 6 Virginia counties on the North Carolina border.

These counties had reported less than 20 diagnoses (suppressed counties), but after estimation have 20-50 HIV diagnoses each.

After projecting new HIV diagnoses using the reduced model for Virginia, there were 43 (32.1%) unsuppressed counties in Virginia (compared to the reported 29) with a total of 4,398 and an average 102 (110.9 SD) new diagnoses (Table 4). The sum of all new diagnoses in Virginia was 5,147, a 15% increase from the reported count. In Maryland and North Carolina, there were 86 (69.4%) unsuppressed counties with a total 11,295 and an average 131 (276.0 SD) new diagnoses. The sum of all new diagnoses in Maryland/North Carolina was 11,639, a 19% decrease from the reported count.

The second set of estimated HIV diagnoses were projected using the reduced model for Maryland/North Carolina. There were 39 unsuppressed counties in Virginia (compared to the reported 29) with a total of 4,698 new diagnoses and an average of 120 (142.2 SD) new diagnoses per county (Table 4). The sum of all new diagnoses in Virginia was 5,286, an 18% increase from the reported count. In Maryland and North Carolina there were 77 (62.1%) unsuppressed counties with a total of 14,305 new diagnoses and an average of 186 (450.2 SD) new diagnoses per county. The sum of all new diagnoses in Maryland/North Carolina was 14,636, a 2% increase from what was reported.

DISCUSSION

Preliminary analyses of county-level distributions of socio-demographic factors indicate county differences between states. Correlation and multivariate modeling further corroborate this finding. Interestingly, we did find the rate ratio of Blacks to Whites was significantly correlated in Maryland and North Carolina, but not significant in Virginia. There is no epidemiologic reason for this difference. African Americans/Blacks make up 19.1% of Virginia's population, 29% of Maryland's population and 21.3% of North Carolina's population (13). Additionally, the Male to Female ratio was not significantly correlated with Virginia diagnoses but was significant in Maryland/North Carolina. The lack of significant correlation between Black population and the Male-Female ratio with new HIV diagnoses in Virginia and the significant correlation between Male-Male Households with HIV diagnoses suggests underreporting or missed diagnoses among Black, heterosexual individuals and females in Virginia. It also conveys the possibility of confounding with gender, population size and population density.

Though county-level demographics do differ between the counties, it was not the outcome of this study. The purpose of this study was to determine whether population demographics would accurately predict the reported new HIV diagnoses in Virginia. Both the full model and reduced model for Maryland/North Carolina had different significant covariates than the Virginia models. Both models fit the Maryland/North Carolina data better than the Virginia HIV data indicating that the county-level variance in reported HIV cases is only partially explained by the county-level socio-demographic factors. However, as both reduced models fit the data well, we expected the projected number of new diagnoses for Maryland and North Carolina to be similar to what was reported and the projected number of new diagnoses for Virginia to also be similar to what was reported. We found that the Maryland/North Carolina projected diagnoses were very close to what was reported (MD/NC data predict MD/NC diagnoses), however the projected Virginia diagnoses in both models were not similar to what was reported. This leads us to believe unmeasured factors, such as underreporting and missed diagnoses are involved. Much of the increase appeared to be in the counties along the North Carolina border, confirming our preliminary suspicions form the reported cases maps.

Furthermore, the map of the reported diagnoses and the two maps of the projected diagnoses show counties where the reported cases do not match the expected cases after weighting the observed county-level covariates by the distribution of covariates in Maryland/North Carolina. In particular, the Virginia counties that border North Carolina have noticeable differences of reported case counts. They report having less than 20 diagnoses (suppressed counties) to having between 20 and 52 diagnoses. The results further indicate the presence of issues with testing/diagnosis or reporting in Virginia, when predicted using Virginia socio-demographics and when predicted after standardizing on Maryland/North Carolina socio-demographics. Interestingly, the Virginia counties with increased diagnoses from what was reported are the same in both reduced models.

Additionally, the North Carolina and Virginia border county HIV variations we see could also point to a reporting issue in North Carolina. The North Carolina counties on the border have higher reported HIV cases than the Virginia border counties. After projecting HIV cases using the two multivariate models, those border counties in Virginia had increases in case counts compared to what was reported. The border between North

Carolina and Virginia started to look like the border between Maryland and Virginia: border counties with similar and high case counts. Again, this corroborates our hypothesis that demographics alone do not explain why there are fewer reported cases in these counties and when demographics were accounted for we estimated more cases than what was reported, leading us to question what is happening on the North Carolina-Virginia border.

Assessing reporting issues of infectious diseases, particularly HIV, is not well studied. Many studies that review reporting and surveillance assess reporting completeness, but not potential reporting biases (36). Though this study is preliminary, it successfully identifies a potential issue with reported HIV diagnoses in Virginia. The results may suggest that the people in the Virginia border counties seek testing and treatment services in North Carolina, are counted in the North Carolina surveillance system and are not reported back to Virginia. In contrast, diagnoses in these counties may not be counted by the local health departments in Virginia and never reported to the state.

There could also be a problem in Virginia with making HIV diagnoses, even though the proportion of the population aged 18-64 who reported ever receiving an HIV test was very similar between Virginia, North Carolina and Maryland (41.3%, 42.2%, and 45.6% respectively) (7). Testing services may need to be reassessed and renovated. Of the three states, Virginia had the highest rate of late diagnoses in 2010 with 30.2% of new diagnoses developing AIDS within 1 year of diagnosis (versus 29.1% in MD and 27.4% in NC) (13). The high rate of late diagnoses suggests missed opportunities for earlier testing and also indicate longer periods for potential HIV transmission among those diagnosed late.

Strengths and Limitations

This analysis only controlled for variations based on demographics, precluding the ability to quantify the extent of unmeasured factors contributing to the variation of reported HIV diagnoses in Virginia. A limitation of this study involves the selection of covariates, as we used all publicly available data. Though we included as many social and population demographic factors as was possible there may be additional factors associated with HIV diagnosis and reporting that may have provided better model fit and variance explanation.

Conclusion

This study focused on identifying whether social and demographics can account for the variation between counties of the reported new HIV diagnoses and predict diagnoses. After controlling for significant social demographic characteristics of the counties, the reported county-level HIV diagnoses were not well predicted, highlighting the importance of unmeasured factors. We hypothesized that underreporting or HIV diagnosis issues may cause the variations of reported diagnoses. Moving forward, additional investigation should be conducted to assess the extent of potential reporting issues, particularly in the Virginia counties along the North Carolina border.

TABLES AND FIGURES

Table 1. Demographics of Unsuppressed Counties (new HIV cases ≥20) in Maryland, North Carolina and Virginia, 2008-2011

	Unsuppressed Counties (≥20 Cases per County)				
	Virginia Counties (n=29)		Maryland and North Carolina Counties (n=70)		T-test
County Characteristics	No.	%	No.	%	p-value [5]
Total New HIV Diagnoses, 2008-2011	3,804		13,854		
County Level Mean (SD)	131	131.5	198	428.7	0.25
County Level Median (IQR) [1]	78	159	56	69	
4 Year Population Averages For Each County					
Total Population					
Mean (SD)	186,018	(122,146)	191,816	(225,982)	0.83
Median (IQR) [1]	203,848	(137,963)	110,378	(141,763)	
Mean Age (years)					
Median Age (SD)	35.7	(4.06)	37.8	(3.56)	0.03
Male Median Age (SD)	34.2	(3.91)	36.3	(3.46)	
Female Median Age (SD)	37.2	(4.17)	39.2	(3.56)	
Sex					
Males	90,987	48.9%	93,027	48.5%	0.68
Females	95,031	51.1%	98,789	51.5%	0.45

Race/Ethnicity					
Hispanic/Latino	17,686	9.5%	16,397	8.5%	0.68
Black/African American, Non-Hispanic	40,368	21.7%	50,038	26.1%	0.84
White, Non-Hispanic	108,767	58.5%	112,962	58.9%	0.84
Other, Non-Hispanic [2]	19,196	10.3%	12,419	6.5%	0.01
Social Determinants					
People Living in Poverty	17,321	12.5%	24,978	16.4%	0.01
People Living in Prison	780	0.4%	1,123	0.6%	0.18
Past Month Drug Use	5,425	2.9%	6,444	3.4%	<0.0001
Drug Dependence	17,227	9.3%	15,143	7.9%	<0.0001
Gini, Income Inequality (SD)	0.42	(0.05)	0.44	(0.03)	0.11
HS Graduate or Higher	165,817	89.1%	164,509	85.8%	<0.01
Male-Male Households	212	0.1%	187	0.1%	0.41
People Living Without Health Insurance	24,266	13.0%	30,079	15.7%	<0.0001
Median Income, USD (IQR) [1]	59,407	(32,733)	43,027	(14,195)	0.01
Geographical Determinants					
Median Population Density [1,3]	1,313.40	(2,203)	489.9	(260.6)	<0.0001
Median House Density [1,4]	568.2	(951.4)	96.1	(107.3)	<0.0001

1. Median and IQR reported in place of Mean and Standard Deviation
2. Other race category includes Non-Hispanic Asian, Native American/Alaska Native, Native Hawaiian/Pacific Islander, and Two or more races
3. People per Sq. Mile
4. Housing Units per Sq. Mile
5. Comparing the Virginia mean to the Maryland/North Carolina Mean
6. Chi-Square test of differences

Table 2. Pearson Correlations of Covariates with the Distribution of HIV by State

Variable	Virginia		Maryland & North Carolina	
	Pearson Correlation	P-value[1]	Pearson Correlation	P-value[1]
Male to Female RR[2]	0.118	0.54	-0.246	**0.04**
Total Population, rate	0.737	**<.0001**	0.668	**<.0001**
Population Density	0.564	**<0.01**	0.717	**<.0001**
Housing Density	0.532	**<0.01**	0.728	**<.0001**
Median Age	-0.218	0.26	-0.238	**<0.05**
Black to White RR[2]	0.237	0.22	0.393	**<0.01**
Hispanic to White RR[2]	0.583	**<0.01**	0.393	**<0.01**
Other race to White RR[2]	0.623	**<0.01**	0.441	**<0.01**
Male-Male Households	0.446	**0.02**	0.243	**0.04**
Median Income	0.136	0.48	0.221	0.07
Poverty Rate	0.006	0.97	-0.149	0.22
Income Inequality (Gini)	0.189	0.33	0.092	0.45
Education Rate	0.145	0.45	0.178	0.14
Prison Rate	-0.033	0.86	-0.160	0.19
No Health Insurance Rate	0.125	0.52	-0.148	0.22
Past Month Drug Use Rate	-0.484	**0.01**	-0.332	**0.01**
Drug Dependence Rate	-0.330	0.08	0.054	0.66

1. Bold p-values significantly correlated with the distribution of HIV at the 95% confidence level
2. Rate Ratio

Table 3a. Model Fit Statistics - Deviance

	Virginia			Maryland & North Carolina		
	Chi-Square	df	P-Value	Chi-Square	df	P-Value
Model 1: Full Model	25.5474	11	<0.01	70.1128	52	0.05
Model 2: Reduced Model	27.0275	18	**0.08**	69.3778	59	**0.17**

*Bold p-values indicate Good Fit at the 0.05 significance level

Table 3b. Reduced Model Variables and Coefficient P-Values

Virginia – Reduced Model		Maryland & North Carolina – Reduced Model
Variable	P-Value	P-Value
Intercept	<.0001	<.0001
Sex	<.0001	0.05
Total Pop.	<.0001	<.0001
Pop. Density	<.0001	0.72
House Density	<.0001	0.52
Median Age	0.11	0.16
BlackRR	<.0001	<.0001
HispRR	0.69	0.71
OtherRR	0.09	0.04
Male-Male Housholds	0.64	0.23
Drug Use	0.04	0.30

Table 4. Comparing Reported New HIV Diagnoses with Projected New HIV Diagnoses in Virginia and Maryland/North Carolina, 2008-2011

	Reported Diagnoses		Projected Diagnoses by MD/NC Model		Projected Diagnoses by VA Model	
	VA (n=134)	MD&NC (n=124)	VA (n=134)	MD&NC (n=124)	VA (n=134)	MD&NC (n=124)
Suppressed Counties [1]	105 (78.3%)	54 (43.5%)	95 (70.9%)	47 (37.9%)	91 (67.9%)	38 (30.6%)
Total Diagnoses	662	546	588	331	749	344
Unsuppressed Counties [2]	29 (21.6%)	70 (56.5%)	39 (29.1%)	77 (62.1%)	43 (32.1%)	86 (69.4%)
Total Diagnoses	3804	13854	4698	14305	4398	11295
Mean (SD)	131 (131.5)	198 (428.7)	120 (142.2)	186 (450.2)	102 (110.9)	131 (276.0)
Median (IQR)	78 (159)	56 (69)	53 (167)	54 (64)	50 (132)	47 (53)
Total New Diagnoses	4466	14400	5286	14636	5147	11639

1. < 20 new diagnoses
2. ≥ 20 new diagnoses

Figure 1. Reported new diagnoses versus Maryland/North Carolina reduced model projected diagnoses

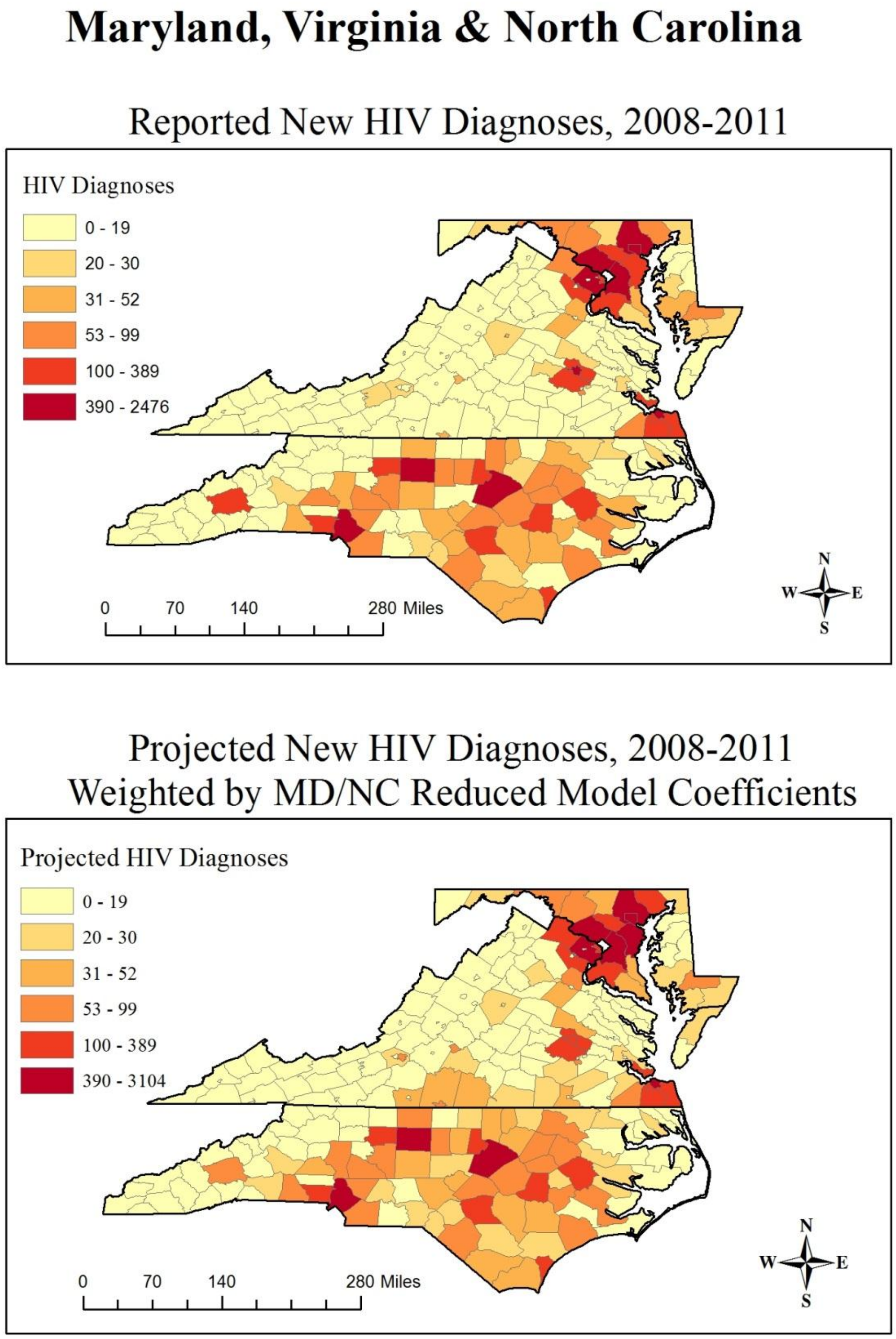
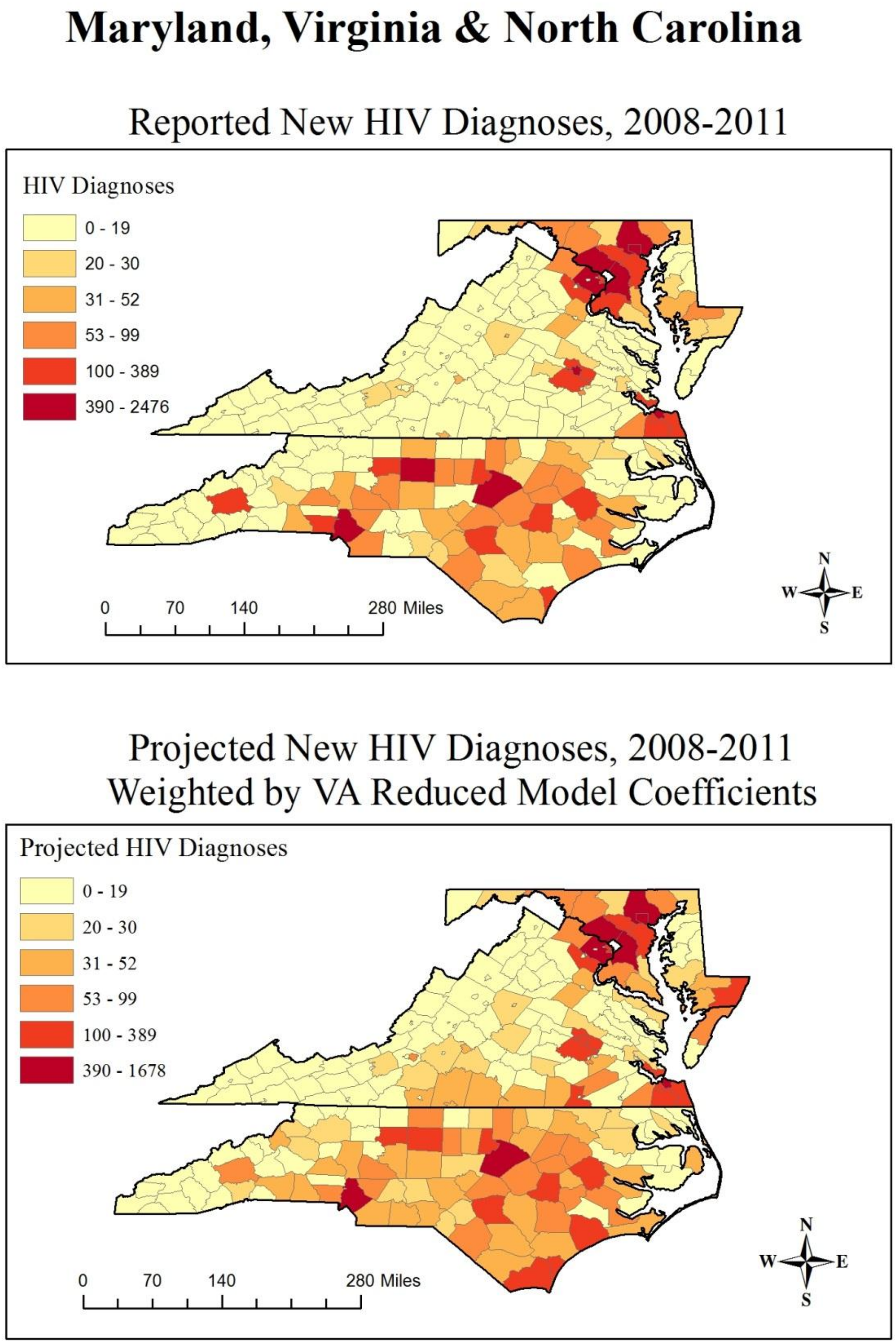


Figure 2. Virginia reduced model projected diagnoses



CHAPTER 3: PUBLIC HEALTH IMPLICATIONS AND SUGGESTIONS

The implications of this study are significant for public health practice. The reported HIV diagnoses in Virginia cannot be accurately predicted using socio-demographic factors that were able to predict reported diagnoses in Maryland and North Carolina. After accounting for either Virginia's or Maryland/North Carolina's demographics there are counties in Virginia that should have more cases that they are actually reporting. There is no biologic plausibility for people who are demographically similar on both sides of a political border to have different rates of HIV infection. The issue must be related to how cases are identified and how they are reported, leading to the conclusion that reporting and/or diagnosis biases are occurring. The effects of such biases lead to false conclusions about the HIV diagnoses and number of unidentified HIV infected individuals in Virginia, and possibly in North Carolina. Even though Virginia may be reaching the CDC required 85% reporting completeness mark, the distribution of reported cases versus expected cases within the state hints at differential completeness. Such conclusions may lead to under-funding of HIV prevention and care in Virginia relative to other states and potentially miss-distribution of available funds within Virginia.

The rate of undiagnosed individuals in Virginia may be even higher than previously estimated. In the United States, 49% of transmissions occur among the estimated 20% of persons with undiagnosed HIV (37, 38). According to the 2012 Virginia Statewide Coordinated Statement of Need and Comprehensive HIV Service Plan and using the CDC estimated back calculation methodology, approximately 5,916 people (74 per 100,000 people) living in Virginia in 2009 were unaware of their HIV infection

(39). Using the same methodology, North Carolina estimated that 7,372 (77 per 100,000 people) were undiagnosed and Maryland estimated there were 7,400 (128 per 100,000 people) undiagnosed people in 2010 (40, 41). These estimates are based on the total number of people living with HIV. As this study shows that the number of new diagnoses between 2008 and 2011 may be more than originally reported, the estimated number of undiagnosed people in Virginia may be even higher than previously thought.

Issues among state border counties may be occurring in other states. The techniques and methods used in this study can be applied elsewhere to investigate variations in HIV diagnoses and even variations of other reported diseases. The estimates of people living with HIV may be underreported on a much larger scale than just Virginia. The equitable distribution of resources, namely funding and testing services, are dependent on accurate reporting of disease. Accurate and timely reports of new HIV diagnoses are vital in the allocation of funds, program planning, estimating the burden of disease, and monitoring and evaluation efforts (16). This analysis suggests that inaccurate reports may contribute to the continued spread of HIV. If the variations are due to a diagnosis issue, then testing availability and referral services may be inadequate. Where diagnoses are not occurring, less money and support is provided, continuing the cycle of under-diagnosing. On the other hand, if reports are not being completed or sent to the state health department, then it suggests lack of funding and support for health departments in Virginia. Additionally, it could also imply that North Carolina may not be sending reports of out of state diagnoses to the correct state of residence. An additional analysis would be to look at the border of North Carolina- South Carolina and North

Carolina-Georgia, to see if the distribution of new HIV diagnoses is similar to that of the Virginia-North Carolina border.

The Virginia Department of Health should carefully assess whether more attention and priority is placed on the larger counties, with large cities to test and report HIV infections. Since the more noticeable variations in what was reported and what was expected occurred in less populated, smaller counties without large cities, the state needs to focus on increasing HIV testing/diagnosis and reporting efforts in those smaller counties. An internal audit of testing availability and surveillance priorities needs to be completed.

Given that the border counties had much higher expected counts of HIV than was reported, Virginia should open lines of communication with bordering state health departments to collaborate on investigating the reasons for these variations. While state borders are unrestricted and state populations can freely cross borders, state-based policies should also be more fluid and work with neighboring states.

REFERENCES

1. Centers for Disease Control and Prevention. Diagnoses of HIV infection, by year of diagnosis and selected characteristics, 2008-2011—United States. *HIV Surveillance Report,* 2011; 23. Published February 2013.
2. Wasserheit, J. (1992). Epidemiological synergy. Interrelationships between human immunodeficiency virus infection and other sexually transmitted diseases. *Sexually Transmitted Diseases*, 19(2), 61–77.
3. Prejean J, Song R, Hernandez A, et al. (2011). Estimated HIV incidence in the United States, 2006-2009. *PLoS One*. 6:e17502.
4. Centers for Disease Control and Prevention. HIV surveillance–United States, 1981-2008. *MMWR Morb Mortal Wkly Rep*. 2011;60:689–693.
5. Centers for Disease Control and Prevention. Monitoring selected national HIV prevention and care objectives by using HIV surveillance data—United States and 6 U.S. dependent areas—2011. *HIV Surveillance Supplemental Report*, 2013;18(No. 5). Published October 2013.
6. Burke, R. C., Sepkowitz, K. A., Bernstein, K. T., Karpati, A. M., Myers, J. E., Tsoi, B. W., & Begier, E. M. (2007). Why don't physicians test for HIV? A review of the US literature. *Aids, 21*(12), 1617-1624.
7. Kaiser Family Foundation. Percentage of Persons Ages 18-64 Who Reported Ever Receiving and HIV Test. 2012. Accessed March 15, 2014. Retrieved from http://kff.org/hivaids/state-indicator/hiv-testing-rate-ever-tested/
8. Kaiser Family Foundation. State Health Facts. Accessed March 15, 2014. Retrieved from http://kff.org/statedata/

9. White, B. L., Carter, Y. L., Records, K., & Martin, I. B. K. (2013). Routine HIV Screening in North Carolina in the Era of the Affordable Care Act: Update on Laws, Reimbursement, and Tests. *Southern Medical Journal*, 106(11), 637-641
10. Reif, S. S., Whetten, K., Wilson, E. R., McAllaster, C., Pence, B. W., Legrand, S., & Gong, W. (2014). HIV/AIDS in the Southern USA: a disproportionate epidemic. *AIDS Care, 26*(3), 351-359.
11. Human Rights Watch. (2010). Southern exposure: Human rights and HIV in the southern United States. Accessed March 15, 2014. Retrieved from http://www.hrw.org/sites/default/files/related_material/BPapersouth1122_6.pdf
12. Johnson, D. F. (2003). Frequent Failed Early HIV Detection in a High Prevalence Area: Implications for Prevention. *AIDS Patient Care and STDs*, 17(6), 277-282.
13. Emory University. AIDSVu. (AIDSVu.org). (Accessed February 2014)
14. Mukolo, A., Villegas, R., Aliyu, M., & Wallston, K. A. (2013). Predictors of Late Presentation for HIV Diagnosis: A Literature Review and Suggested Way Forward. *Aids and Behavior, 17*(1), 5-30.
15. Chen, M., Rhodes, P. H., Hall, H. I., Kilmarx, P. H., Branson, B. M., & Valleroy, L. A. (2012). Prevalence of Undiagnosed HIV Infection Among Persons Aged ≥ 13 Years – National HIV Surveillance System, United States, 2005-2008. *MMWR*, 61(02), 57-64.
16. Hall, H. I., Song, R., Gerstie III, J. E., & Lee, L. M. (2006). Assessing the Completeness of Reporting of Human Immunodeficiency Virus Diagnoses in 2002-2003: Capture-Recapture Methods. *Am J Epidemiol*, 164(4), 391-397.

17. Branson, B. M., Handsfield, H. H., Lampe, M. A., Janssen, R. S., Taylor, A. W., Lyss, S. B., & Clark, J. E. (2006). Revised Recommendations for HIV Testing of Adults, Adolescents, and Pregnant Women in Health-Care Settings. *MMWR*, 55(RR14), 1-17.

18. Glynn, M. K., Ling, Q., Phelps, R., Li, J. M., & Lee, L. M. (2008). Accurate monitoring of the HIV epidemic in the United States - Case duplication in the national HIV/AIDS surveillance system. *Jaids-Journal of Acquired Immune Deficiency Syndromes*, 47(3), 391-396.

19. Buehler, J. W., Berkelman, R. L., & Stehr-Green, J. K. (1992). The completeness of AIDS surveillance. *Journal of acquired immune deficiency syndromes*, 5(3), 257-264.

20. Hall, H. I., Mokotoff, E. D., & Advisory, Grp. (2007). Setting standards and an evaluation framework for human immunodeficiency virus/acquired immunodeficiency syndrome surveillance. *Journal of Public Health Management and Practice*, 13(5), 519-523.

21. Nash, D., Andreopoulos, E., Horowitz, D., Sohler, N., & Vlahov, D. (2007). Differences Among U.S. States in Estimating the Number of People Living with HIV/AIDS: Impact on Allocation of Federal Ryan White Funding. *Public Health Reports*, 122, 644-656.

22. Ohl, M. E., & Perencevich, E. (2011). Frequency of human immunodeficiency virus (HIV) testing in urban vs. rural areas of the United States: Results from a nationally-representative sample. *Bmc Public Health,* 11:681.

23. Penner M, Leone PA.(2007). Integration of testing for, prevention of, and access to treatment for HIV infection: state and local perspectives. *Clin Infect Dis*, 45: S281–S286.

24. McCoy, S. I. (2010). Variability of the Date of HIV Diagnosis: A Comparison of Self-Report, Medical Record, and HIV/AIDS Surveillance Data. *Annals of Epidemiology*, 20(10), 734-742.

25. An, Q., Prejean, J., Harrison, K. M., & Fang, X. (2013). Association Between Community Socioeconomic Position and HIV Diagnosis Rate Among Adults and Adolescents in the United States, 2005 to 2009. *American Journal of Public Health,* 103(1), 120-126.

26. An, Q., Prejean, J., & Hall, H. I. (2012). Racial Disparity in U.S. Diagnoses of Acquired Immune Deficiency Syndrome, 2000-2009. *American Jounral of Preventive Medicine*, 43(5), 461-466.

27. Franks, P., Muennig, P., Lubetkin, E., & Jia, H. M. (2006). The burden of disease associated with being African-American in the United States and the contribution of socio-economic status. *Social Science & Medicine, 62*(10), 2469-2478

28. Kawachi, I., & Kennedy, B. P. (1997). Socioeconomic determinants of health .2. Health and social cohesion: Why care about income inequality? *British Medical Journal*, 314(7086), 1037-1040.

29. Fife, D., & Mode, C. (1992). AIDS Incidence and Income. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 5(11), 1105-1110.

30. Simard, E. P., Fransua, M., Naishadham, D., & Jemal, A. (2012). The Influence of Sex, Race/Ethnicity, and Educational Attainment on Human Immunodeficiency

Virus Death Rates Among Adults, 1993-2007. *Archives of Internal Medicine, 172*(20), 1591-1598.

31. Simon, P. A., Hu, D. J., Diaz, T., & Kerndt, P. R. (1995). Income and AIDS Rates in Los Angeles County. *Aids*, 9(3), 281-284.
32. Diaz, T., Chu, S. Y., Buehler, J. W., Boyd, D., Checko, P. J., Conti, L., Davidson, A. J., Hermann, P., Herr, M., Levy, A., & Hersh, B. S. (1994). Socioeconomic Differences Among People With AIDS – Results From a Multistate Surveillance Project. *American Journal of Preventive Medicine*, 10(4), 217-222.
33. Hasnain, M., Levy, J. A., Mensah, E. K., & Sinacore, J. M. (2007). Association of educational attainment with HIV risk in African American active injection drug users. *Aids Care-Psychological and Socio-Medical Aspects of Aids/Hiv, 19*(1), 87-91
34. Sutton, M., Anthony, M. N., Vila, C., McLellan-Lemal, E., & Weidle, P. J. (2010). HIV Testing and HIV/AIDS Treatment Services in Rural Counties in 10 Southern States: Service Provider Perspectives. *Journal of Rural Health*, 26(3), 240-247.
35. White House Office of National AIDS Policy (2010). National HIV/AIDS strategy for the United States. Washington, DC. Accessed March 15, 2014. Retrieved from http://aids.gov/federal-resources/national-hiv-aids-strategy/nhas.pdf
36. Jajosky, R. A., & Groseclose, S. L. (2004). Evaluation of reporting timeliness of public health surveillance systems for infectious diseases. *Bmc Public Health*, 4(29).

37. Lansky, A., Prejean, J., & Hall, I. (2013). Challenges in Identifying and Estimating Undiagnosed HIV Infection. *Future Virol.*, 8(6), 523-526.

38. Hall, H. I., Holtgrave, D. R., & Maulsby, C. (2012). HIV Transmission Rates from Persons Living with HIV Who are Aware and Unaware of their Infection. *AIDS*, 26, 887-896.

39. Virginia Department of Health. (2012) 2012 Virginia Statewide Coordinated Statement of Need and Comprehensive HIV Service Plan. Accessed April 12, 2014. Retrieved from http://www.vdh.virginia.gov/epidemiology/DiseasePrevention/HCS/documents/2012/pdf/Virginia%202012%20SCSN%20SCP.pdf

40. North Carolina Department of Health and Human Services. (2012) State of North Carolina 2012 HIV Care and Prevention Statewide Coordinated Statement of Need Needs Assessment and Comprehensive Plan. Accessed April 12, 2014. Retrieved from http://epi.publichealth.nc.gov/cd/hiv/docs/NC_SCSN-NA-CP_2012.pdf

41. Maryland Department of Health and Mental Hygiene. (2012). 2012-2014 Maryland HIV Plan. Accessed April 12, 2014. Retrieved from http://phpa.dhmh.maryland.gov/OIDPCS/CHP/SiteAssets/SitePages/regional-advisory-committee/Maryland%202012-2014%20HIV%20Plan.pdf

APPENDICES

Appendix A: Data Sources

Variable	Source	Website
County-level New HIV Diagnoses, 2008-2011	AIDSVu.org via Centers for Disease Control and Prevention	www.aidsvu.org
Total Population Population Density Housing Density Age Race Sex	County Characteristics Datasets: Intercensal Estimates of the Resident Population by Five-Year Age Groups, Sex, Race, and Hispanic Origin for Counties: April 1, 2000 to July 1, 2010 Annual County Resident Population Estimates by Age, Sex, Race, and Hispanic Origin: April 1, 2010 to July 1, 2012	http://www.census.gov/popest/
Poverty Median Income	U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE)	http://www.census.gov/did/www/saipe/
Health Insurance	U.S. Census Bureau's Small Area Health Insurance Estimates (SAHIE)	http://www.census.gov/did/www/sahie/
High School	U.S. Census Bureau, American Community Survey 1-Year	http://factfinder2.census.gov/

Graduate	Estimates, 2008, 2009, 2010, 2011 Table C15003: Educational Attainment	
Income Inequality (Gini)	U.S. Census Bureau, American Community Survey 1-Year Estimates, 2008, 2009, 2010, 2011 Table B19083: Income Inequality	http://factfinder2.census.gov/
Past Month Drug Use Drug Dependence	Substance Abuse and Mental Health Services Administration (SAMHSA) National Survey on Drug Use and Health (NSDUH)	http://www.samhsa.gov/data/NSDUH.aspx
Male-Male Households	U.S. Census Bureau, American Community Survey 1-Year Estimates, 2008, 2009, 2010, 2011 Table S1101: Households and Families	http://factfinder2.census.gov/
Prison Population	U.S. Census Bureau, American Community Survey 1-Year Estimates, 2008, 2009, 2010, 2011 Table PCT20: Group Quarters Population by Group Quarters Type	http://factfinder2.census.gov/

Appendix B: Full Model Coefficients

	Virginia – Full Model					Maryland & North Carolina – Full Model				
Variable	Estimate	95% CI		Wald Chi-Square	P-Value	Estimate	95% CI		Wald Chi-Square	P-Value
Intercept	-18.551	-29.763	-7.340	10.52	<0.01	-0.172	-11.171	10.826	0.00	0.98
Sex	3.587	2.309	4.864	30.28	<.0001	0.847	-0.596	2.291	1.32	0.25
Tot Pop	1.127	0.999	1.254	299.78	<.0001	1.156	0.988	1.323	183.04	<.0001
Pop. Density	-2.781	-4.146	-1.416	15.94	<.0001	0.456	-0.184	1.096	1.95	0.16
House Density	2.846	1.580	4.113	19.40	<.0001	-0.342	-0.953	0.269	1.20	0.27
Median Age	-0.033	-0.063	-0.004	4.84	0.03	0.046	0.016	0.076	9.03	<0.01
Median Income	0.944	0.072	1.816	4.50	0.03	-1.206	-2.069	-0.343	7.49	0.01
BlackRR	0.626	0.503	0.750	98.95	<.0001	0.521	0.402	0.639	74.68	<.0001
HispRR	-0.108	-0.320	0.104	1.00	0.32	0.168	-0.048	0.384	2.32	0.13
OtherRR	-0.135	-0.325	0.054	1.96	0.16	-0.069	-0.178	0.039	1.57	0.21
Male-Male Households	-0.038	-0.094	0.018	1.75	0.19	0.040	-0.087	0.167	0.38	0.54
Prison	-0.001	-0.028	0.027	0.00	0.97	0.048	0.006	0.090	4.96	0.03
Gini	2.643	0.181	5.105	4.43	0.04	1.513	-1.531	4.558	0.95	0.33
Rate_HSgrad	-0.002	-0.006	0.001	2.07	0.15	0.002	0.000	0.004	2.61	0.11
Rate_NoIns	0.002	-0.005	0.009	0.25	0.62	-0.004	-0.009	0.001	2.29	0.13
Rate_Poverty	-0.001	-0.005	0.003	0.18	0.67	-0.001	-0.006	0.003	0.31	0.58
Rate_DrugUse	-0.020	-0.058	0.019	1.02	0.31	-0.008	-0.022	0.006	1.23	0.27
Rate_DrugDep	0.007	-0.005	0.019	1.18	0.28	0.001	-0.007	0.009	0.04	0.84
Dispersion	0.000					0.030	0.018	0.049		
Model Fit				25.547	0.01				70.113	0.05

Appendix C: Reduced Model Coefficients

	Virginia – Reduced Model					Maryland & North Carolina – Reduced Model				
Variable	Estimate	95% CI		Wald Chi-Square	P-Value	Estimate	95% CI		Wald Chi-Square	P-Value
Intercept	-8.490	-11.116	-5.865	40.17	<.0001	-11.043	-13.653	-8.432	68.72	<.0001
Sex	3.767	2.353	5.181	27.27	<.0001	1.354	0.006	2.702	3.88	0.05
Total Pop.	1.110	0.985	1.234	304.51	<.0001	1.167	0.994	1.341	174.28	<.0001
Pop. Density	-2.519	-3.638	-1.400	19.46	<.0001	-0.106	-0.682	0.471	0.13	0.72
House Density	2.576	1.550	3.602	24.22	<.0001	0.189	-0.384	0.762	0.42	0.52
Median Age	-0.020	-0.044	0.004	2.60	0.11	0.019	-0.008	0.045	1.96	0.16
BlackRR	0.526	0.449	0.604	176.97	<.0001	0.632	0.535	0.729	163.41	<.0001
HispRR	0.033	-0.128	0.193	0.16	0.69	-0.024	-0.147	0.100	0.14	0.71
OtherRR	-0.189	-0.404	0.027	2.95	0.09	-0.118	-0.228	-0.008	4.38	0.04
Male-Male Housholds	-0.012	-0.064	0.040	0.21	0.64	0.083	-0.054	0.219	1.41	0.23
Drug Use	-0.033	-0.064	-0.001	4.14	0.04	-0.007	-0.021	0.007	1.08	0.30
Dispersion	0.007	0.002	0.024			0.046	0.030	0.072		
Model Fit				27.03	0.08				69.38	0.17

Appendix D: Number of Counties in Virginia that had Different Projected Diagnoses compared to Reported Diagnoses

Table 5. Number of Counties in Virginia that had Higher/Lower Projected Diagnoses than Reported Diagnoses

	Number of Virginia Counties (n=134)	
Amount of Change from Reported Diagnoses	Projected using MD/NC Model	Projected using VA Model
>1 and <1.5 times	23 (17.2%)	32 (23.9%)
>1.5 and <2 times	14 (10.4%)	11 (8.2%)
>2 times	11 (8.2%)	17 (12.7%)
>0.67 and <1 times	20 (14.9%)	29 (21.6%)
>0.5 and <0.67 times	18 (13.4%)	15 (11.2%)
<0.5 times	45 (33.6%)	27 (20.1%)

Appendix E: SAS Code

```
*************************************************;
*     Thesis Code part 1                        *;
*     Data Management                           *;
*     Written By: Noel Hatley                   *;
*     Date January 27, 2014                     *;
*************************************************;


*****************************************************;
***************      Import HIV Data   **************;
*****************************************************;

OPTIONS nofmterr;
libname b 'T:\EpiProjs\Sullivan_data\AIDSVu\AIDSVu 2013\Data for AIDSVu
2013\NewDxData';

data countyHIV;
      set b.County_hivdx_2008_2011;
run;

data stateHIV;
      set b.state_hivdx_2008_2011;
run;

/*
proc print data=a.County_hivdx_2008_2011;
run;

proc print data=a.state_hivdx_2008_2011;
run;
*/

OPTIONS nofmterr;
libname a 'H:\Classes\Thesis\Data\SocialDeterminants';
/*
%include "H:\Classes\Thesis\Data\Raw_NewDx.sas";
*/
data a.CountyHIV (rename=(county=ctyname));
      set countyHIV;

run;

data a.StateHIV;
      set a.StateHIV;

      keep state statecase;
run;

************************************************************;
***************   Import County FIPS codes      ************;
************************************************************;
```

```
PROC IMPORT OUT= work.allfips
            DATAFILE= "H:\Classes\Thesis\Data\County_FIPS_Codes.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=5000;
RUN;

data fips;
      set allfips;
      where state='MD' or state='NC' or state='VA';

      newfips=PUT(county_fips, z3.);
      fips=trim(state_fips)||trim(newfips);
      geo_id2=fips*1;

      keep geo_id2 state county;
run;

data a.fips;
      set fips;
run;


************************************************************;
***************   Import County Demographics    ************;
************************************************************;


******AGE RACE SEX******;
* Import one file for each state of the age, race and sex composition
from 2000-2010;

*MARYLAND, 2000-2010;
PROC IMPORT OUT= WORK.AgeRaceSexMD2000
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\AgeRaceSex\MD_AgeRaceSex_200
0-2010.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=10000;
run;

*VIRGINIA, 2000-2010;
PROC IMPORT OUT= WORK.AgeRaceSexVA2000
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\AgeRaceSex\VA_AgeRaceSex_200
0-2010.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=10000;
run;

*NORTH CAROLINA, 2000-2010;
PROC IMPORT OUT= WORK.AgeRaceSexNC2000
```

```
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\AgeRaceSex\NC_AgeRaceSex_200
0-2010.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=10000;
run;

*Combine all Age, Race and Sex data for 2000-2010;
*ARS refers to Age, Race and Sex;
data ARSCombine2008;
      length stname $ 25;
      set AgeRaceSexMD2000 (in=a) AgeRaceSexVA2000 (in=b)
AgeRaceSexNC2000 (in=c);

      COfips=PUT(county, z3.);
      STfips=PUT(state, 2.);
      fips=trim(STfips)||trim(COfips);
      geo_id2=fips*1;

      if stname = "Maryland" then st = "MD";
      if stname = "Virginia" then st = "VA";
      if stname = "North Carolina" then st = "NC";

      if year = 1 or year = 2 or year =3 or year = 4 or year =5 or year
= 6 or year = 7 or year = 8 or year = 9 then delete; *2000-2007
Resident pop est.;
      if year = 10 then year1 = 2008; *2008 Resident Population
Estimate 7/1/2008;
      if year = 11 then year1 = 2009; *2009 Resident Population
Estimate 7/1/2009;
      if year = 12 then year1 = 2010.5; *2010 Census population
4/1/2010;
      if year = 13 then year1 = 2010.4; *2010 Resident Population
Estimate 7/1/2010;

      format stname $15.;
      drop year;
run;



* Import one file for each state of the age, race and sex composition
from 2010-2012;
*MARYLAND, 2010-2012;
PROC IMPORT OUT= WORK.AgeRaceSexMD2010
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\AgeRaceSex\MD_AgeRaceSex_201
0-2012.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=10000;
run;

*VIRGINIA, 2010-2012;
```

```
PROC IMPORT OUT= WORK.AgeRaceSexVA2010
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\AgeRaceSex\VA_AgeRaceSex_201
0-2012.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=10000;
run;

*NORTH CAROLINA, 2010-2012;
PROC IMPORT OUT= WORK.AgeRaceSexNC2010
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\AgeRaceSex\NC_AgeRaceSex_201
0-2012.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=10000;
run;

*Combine all Age, Race and Sex data for 2010-2012;
*ARS refers to Age, Race and Sex;
data ARSCombine2010;
      length stname $ 25;
      set AgeRaceSexMD2010 (in=a) AgeRaceSexVA2010 (in=b)
AgeRaceSexNC2010 (in=c);

      COfips=PUT(county, z3.);
      STfips=PUT(state, 2.);
      fips=trim(STfips)||trim(COfips);
      geo_id2=fips*1;

      if agegrp=0 then agegrp=99;

      if stname = "Maryland" then st = "MD";
      if stname = "Virginia" then st = "VA";
      if stname = "North Carolina" then st = "NC";

      if year = 1 then year1 = 2010.1; *2010 Census population
4/1/2010;
      if year = 2 then year1 = 2010.2; *2010 Population Estimates Base
4/1/2010;
      if year = 3 then year1 = 2010.3; *2010 Population Estimate
7/1/2010;
      if year = 4 then year1 = 2011; *Population Estimate 7/1/2011;
      if year = 5 then year1 = 2012; *Population Estimate 7/1/2012;

      format stname $15.;
      drop year;
run;

data ARSallYears;
      set ARSCombine2010 (in=a) ARSCombine2008 (in=b);
run;

* Create dataset for total population and total population by race;
```

```
data TotalPop;
      set ARSallYears;
      where agegrp=99;

      keep geo_id2 stname ctyname year1 tot_pop tot_male tot_female
nh_male nh_female nhwa_male nhwa_female nhba_male nhba_female
            nhia_male nhia_female nhaa_male nhaa_female nhna_male
nhna_female nhtom_male nhtom_female h_male h_female;
run;

* Separate total pop and race into 4 datasets, one for each year and
rename variables;
data Pop2008;
      set TotalPop;
      where year1=2008;

      tot_pop08=tot_pop*1;
      tot_m08=tot_male*1;
      tot_f08=tot_female*1;
      nh_m08=nh_male*1;
      nh_f08=nh_female*1;
      white_m08=nhwa_male*1;
      white_f08=nhwa_female*1;
      black_m08=nhba_male*1;
      black_f08=nhba_female*1;
      indian_m08=nhia_male*1;
      indian_f08=nhia_female*1;
      asian_m08=nhaa_male*1;
      asian_f08=nhaa_female*1;
      pacific_m08=nhna_male*1;
      pacific_f08=nhna_female*1;
      tworace_m08=nhtom_male*1;
      tworace_f08=nhtom_female*1;
      hispanic_m08=h_male*1;
      hispanic_f08=h_female*1;

      drop year1 tot_pop tot_male tot_female nh_male nh_female
nhwa_male nhwa_female nhba_male nhba_female
            nhia_male nhia_female nhaa_male nhaa_female nhna_male
nhna_female nhtom_male nhtom_female h_male h_female;
run;

data Pop2009;
      set TotalPop;
      where year1=2009;

      tot_pop09=tot_pop*1;
      tot_m09=tot_male*1;
      tot_f09=tot_female*1;
      nh_m09=nh_male*1;
      nh_f09=nh_female*1;
      white_m09=nhwa_male*1;
      white_f09=nhwa_female*1;
      black_m09=nhba_male*1;
      black_f09=nhba_female*1;
      indian_m09=nhia_male*1;
      indian_f09=nhia_female*1;
```

```
	asian_m09=nhaa_male*1;
	asian_f09=nhaa_female*1;
	pacific_m09=nhna_male*1;
	pacific_f09=nhna_female*1;
	tworace_m09=nhtom_male*1;
	tworace_f09=nhtom_female*1;
	hispanic_m09=h_male*1;
	hispanic_f09=h_female*1;

	drop year1 tot_pop tot_male tot_female nh_male nh_female
nhwa_male nhwa_female nhba_male nhba_female
		nhia_male nhia_female nhaa_male nhaa_female nhna_male
nhna_female nhtom_male nhtom_female h_male h_female;
run;

data Pop2010;
	set TotalPop;
	where year1=2010.3;

	tot_pop10=tot_pop*1;
	tot_m10=tot_male*1;
	tot_f10=tot_female*1;
	nh_m10=nh_male*1;
	nh_f10=nh_female*1;
	white_m10=nhwa_male*1;
	white_f10=nhwa_female*1;
	black_m10=nhba_male*1;
	black_f10=nhba_female*1;
	indian_m10=nhia_male*1;
	indian_f10=nhia_female*1;
	asian_m10=nhaa_male*1;
	asian_f10=nhaa_female*1;
	pacific_m10=nhna_male*1;
	pacific_f10=nhna_female*1;
	tworace_m10=nhtom_male*1;
	tworace_f10=nhtom_female*1;
	hispanic_m10=h_male*1;
	hispanic_f10=h_female*1;

	drop year1 tot_pop tot_male tot_female nh_male nh_female
nhwa_male nhwa_female nhba_male nhba_female
		nhia_male nhia_female nhaa_male nhaa_female nhna_male
nhna_female nhtom_male nhtom_female h_male h_female;
run;

data Pop2011;
	set TotalPop;
	where year1=2011;

	tot_pop11=tot_pop*1;
	tot_m11=tot_male*1;
	tot_f11=tot_female*1;
	nh_m11=nh_male*1;
	nh_f11=nh_female*1;
	white_m11=nhwa_male*1;
	white_f11=nhwa_female*1;
	black_m11=nhba_male*1;
```

```
	black_f11=nhba_female*1;
	indian_m11=nhia_male*1;
	indian_f11=nhia_female*1;
	asian_m11=nhaa_male*1;
	asian_f11=nhaa_female*1;
	pacific_m11=nhna_male*1;
	pacific_f11=nhna_female*1;
	tworace_m11=nhtom_male*1;
	tworace_f11=nhtom_female*1;
	hispanic_m11=h_male*1;
	hispanic_f11=h_female*1;

	drop year1 tot_pop tot_male tot_female nh_male nh_female
nhwa_male nhwa_female nhba_male nhba_female
		nhia_male nhia_female nhaa_male nhaa_female nhna_male
nhna_female nhtom_male nhtom_female h_male h_female;
run;


* Create one dataset of Race and Sex for All years;
proc sort data=pop2008;
	by geo_id2;
proc sort data=pop2009;
	by geo_id2;
proc sort data=pop2010;
	by geo_id2;
proc sort data=pop2011;
	by geo_id2;
data TotRaceSex;
	merge pop2008 pop2009 pop2010 pop2011;
	by geo_id2;

	* Total Pop;
	tot_pop_avg=(tot_pop08+tot_pop09+tot_pop10+tot_pop11)/4;
	tot_males_avg=(tot_m08+tot_m09+tot_m10+tot_m11)/4;
	tot_females_avg=(tot_f08+tot_f09+tot_f10+tot_f11)/4;

	* Non-Hispanic;
	nh_males_avg=(nh_m08+nh_m09+nh_m10+nh_m11)/4;
	nh_females_avg=(nh_f08+nh_f09+nh_f10+nh_f11)/4;
	tot_nh_avg=(nh_m08+nh_f08+nh_m09+nh_f09+nh_m10+nh_f10+nh_m11+nh_f
11)/4;

	*White Non-Hispanic;
	white_males_avg=(white_m08+white_m09+white_m10+white_m11)/4;
	white_females_avg=(white_f08+white_f09+white_f10+white_f11)/4;
	tot_white_avg=(white_m08+white_f08+white_m09+white_f09+white_m10+
white_f10+white_m11+white_f11)/4;

	*Black Non-Hispanic;
	black_males_avg=(black_m08+black_m09+black_m10+black_m11)/4;
	black_females_avg=(black_f08+black_f09+black_f10+black_f11)/4;
	tot_black_avg=(black_m08+black_f08+black_m09+black_f09+black_m10+
black_f10+black_m11+black_f11)/4;

	*Indian Non-Hispanic;
	indian_males_avg=(indian_m08+indian_m09+indian_m10+indian_m11)/4;
```

```
	indian_females_avg=(indian_f08+indian_f09+indian_f10+indian_f11)/
4;
	tot_indian_avg=(indian_m08+indian_f08+indian_m09+indian_f09+india
n_m10+indian_f10+indian_m11+indian_f11)/4;

	*Asian Non-Hispanic;
	asian_males_avg=(asian_m08+asian_m09+asian_m10+asian_m11)/4;
	asian_females_avg=(asian_f08+asian_f09+asian_f10+asian_f11)/4;
	tot_asian_avg=(asian_m08+asian_f08+asian_m09+asian_f09+asian_m10+
asian_f10+asian_m11+asian_f11)/4;

	*Pacific Islander/Native Hawaiian Non-Hispanic;
	pacific_males_avg=(pacific_m08+pacific_m09+pacific_m10+pacific_m1
1)/4;
	pacific_females_avg=(pacific_f08+pacific_f09+pacific_f10+pacific_
f11)/4;
	tot_pacific_avg=(pacific_m08+pacific_f08+pacific_m09+pacific_f09+
pacific_m10+pacific_f10+pacific_m11+pacific_f11)/4;

	*Two Races Non-Hispanic;
	tworace_males_avg=(tworace_m08+tworace_m09+tworace_m10+tworace_m1
1)/4;
	tworace_females_avg=(tworace_f08+tworace_f09+tworace_f10+tworace_
f11)/4;
	tot_tworace_avg=(tworace_m08+tworace_f08+tworace_m09+tworace_f09+
tworace_m10+tworace_f10+tworace_m11+tworace_f11)/4;

	*Hispanic;
	hispanic_males_avg=(hispanic_m08+hispanic_m09+hispanic_m10+hispan
ic_m11)/4;
	hispanic_females_avg=(hispanic_f08+hispanic_f09+hispanic_f10+hisp
anic_f11)/4;
	tot_hispanic_avg=(hispanic_m08+hispanic_f08+hispanic_m09+hispanic
_f09+hispanic_m10+hispanic_f10+hispanic_m11+hispanic_f11)/4;

	keep geo_id2 stname ctyname tot_pop_avg tot_males_avg
tot_females_avg hispanic_males_avg hispanic_females_avg
tot_hispanic_avg tworace_males_avg
			tworace_females_avg tot_tworace_avg pacific_males_avg
pacific_females_avg tot_pacific_avg asian_males_avg asian_females_avg
tot_asian_avg indian_males_avg
			indian_females_avg tot_indian_avg black_males_avg
black_females_avg tot_black_avg white_males_avg white_females_avg
tot_white_avg nh_males_avg nh_females_avg
			tot_nh_avg;
run;

data a.TotRaceSex;
	set TotRaceSex;
run;


****************************************;
***************   AGE       ************;
****************************************;

data Age;
```

```
	set arsallyears;
	where (agegrp ne 99 and agegrp ne 0 and agegrp ne 1 and agegrp ne 
2 and agegrp ne 3)
			and (year1 eq 2008 or year1 eq 2009 or year1 eq 
2010.3 or year1 eq 2011);

	keep geo_id2 stname ctyname year1 agegrp tot_pop tot_female 
tot_male;
run;

* Separate out by year then by tot_pop, tot_female and tot_male;

*********************************************;
***************   2008 AGE      *************;
*********************************************;
* 2008 Total Pop Age Distribution*;
data age08pop;
	set age;
	where year1=2008;

	keep geo_id2 agegrp tot_pop;
run;

proc sort data=age08pop;
	by geo_id2 agegrp;
proc transpose data=age08pop out=pop08age;
	by geo_id2;
	id agegrp;
run;

data TotPop08Age;
	set pop08age;

	p08age15_19=_4*1;
	p08age20_24=_5*1;
	p08age25_29=_6*1;
	p08age30_34=_7*1;
	p08age35_39=_8*1;
	p08age40_44=_9*1;
	p08age45_49=_10*1;
	p08age50_54=_11*1;
	p08age55_59=_12*1;
	p08age60_64=_13*1;
	p08age65_69=_14*1;
	p08age70_74=_15*1;
	p08age75_79=_16*1;
	p08age80_84=_17*1;
	p08age85=_18*1;

	drop _name_ _4 _5 _6 _7 _8 _9 _10 _11 _12 _13 _14 _15 _16 _17 
_18;
run;


* 2008 Total Male Pop Age Distribution;
data age08male;
	set age;
```

```
      where year1=2008;

      keep geo_id2 agegrp tot_male;
run;

proc sort data=age08male;
      by geo_id2 agegrp;
proc transpose data=age08male out=male08age;
      by geo_id2;
      id agegrp;
run;

data TotMale08Age;
      set male08age;

      m08age15_19=_4*1;
      m08age20_24=_5*1;
      m08age25_29=_6*1;
      m08age30_34=_7*1;
      m08age35_39=_8*1;
      m08age40_44=_9*1;
      m08age45_49=_10*1;
      m08age50_54=_11*1;
      m08age55_59=_12*1;
      m08age60_64=_13*1;
      m08age65_69=_14*1;
      m08age70_74=_15*1;
      m08age75_79=_16*1;
      m08age80_84=_17*1;
      m08age85=_18*1;

      drop _name_ _4 _5 _6 _7 _8 _9 _10 _11 _12 _13 _14 _15 _16 _17
_18;
run;

* 2008 Total Female Pop Age Distribution;
data age08female;
      set age;
      where year1=2008;

      keep geo_id2 agegrp tot_female;
run;

proc sort data=age08female;
      by geo_id2 agegrp;
proc transpose data=age08female out=female08age;
      by geo_id2;
      id agegrp;
run;

data TotFemale08Age;
      set female08age;

      f08age15_19=_4*1;
      f08age20_24=_5*1;
      f08age25_29=_6*1;
      f08age30_34=_7*1;
```

```
        f08age35_39=_8*1;
        f08age40_44=_9*1;
        f08age45_49=_10*1;
        f08age50_54=_11*1;
        f08age55_59=_12*1;
        f08age60_64=_13*1;
        f08age65_69=_14*1;
        f08age70_74=_15*1;
        f08age75_79=_16*1;
        f08age80_84=_17*1;
        f08age85=_18*1;

        drop _name_ _4 _5 _6 _7 _8 _9 _10 _11 _12 _13 _14 _15 _16 _17
_18;
run;

proc sort data=totpop08age;
        by geo_id2;
proc sort data=totmale08age;
        by geo_id2;
proc sort data=totfemale08age;
        by geo_id2;
data Age2008Totals;
        merge totpop08age totmale08age totfemale08age;
        by geo_id2;
run;


***********************************************;
****************  2009 AGE      ***************;
***********************************************;

* 2009 Total Pop Age Distribution*;
data age09pop;
        set age;
        where year1=2009;

        keep geo_id2 agegrp tot_pop;
run;

proc sort data=age09pop;
        by geo_id2 agegrp;
proc transpose data=age09pop out=pop09age;
        by geo_id2;
        id agegrp;
run;

data TotPop09Age;
        set pop09age;

        p09age15_19=_4*1;
        p09age20_24=_5*1;
        p09age25_29=_6*1;
        p09age30_34=_7*1;
        p09age35_39=_8*1;
        p09age40_44=_9*1;
        p09age45_49=_10*1;
```

```
        p09age50_54=_11*1;
        p09age55_59=_12*1;
        p09age60_64=_13*1;
        p09age65_69=_14*1;
        p09age70_74=_15*1;
        p09age75_79=_16*1;
        p09age80_84=_17*1;
        p09age85=_18*1;

        drop _name_ _4 _5 _6 _7 _8 _9 _10 _11 _12 _13 _14 _15 _16 _17
_18;
run;


* 2009 Total Male Pop Age Distribution;
data age09male;
        set age;
        where year1=2009;

        keep geo_id2 agegrp tot_male;
run;

proc sort data=age09male;
        by geo_id2 agegrp;
proc transpose data=age09male out=male09age;
        by geo_id2;
        id agegrp;
run;

data TotMale09Age;
        set male09age;

        m09age15_19=_4*1;
        m09age20_24=_5*1;
        m09age25_29=_6*1;
        m09age30_34=_7*1;
        m09age35_39=_8*1;
        m09age40_44=_9*1;
        m09age45_49=_10*1;
        m09age50_54=_11*1;
        m09age55_59=_12*1;
        m09age60_64=_13*1;
        m09age65_69=_14*1;
        m09age70_74=_15*1;
        m09age75_79=_16*1;
        m09age80_84=_17*1;
        m09age85=_18*1;

        drop _name_ _4 _5 _6 _7 _8 _9 _10 _11 _12 _13 _14 _15 _16 _17
_18;
run;

* 2009 Total Female Pop Age Distribution;
data age09female;
        set age;
        where year1=2009;
```

```
      keep geo_id2 agegrp tot_female;
run;

proc sort data=age09female;
      by geo_id2 agegrp;
proc transpose data=age09female out=female09age;
      by geo_id2;
      id agegrp;
run;

data TotFemale09Age;
      set female09age;

      f09age15_19=_4*1;
      f09age20_24=_5*1;
      f09age25_29=_6*1;
      f09age30_34=_7*1;
      f09age35_39=_8*1;
      f09age40_44=_9*1;
      f09age45_49=_10*1;
      f09age50_54=_11*1;
      f09age55_59=_12*1;
      f09age60_64=_13*1;
      f09age65_69=_14*1;
      f09age70_74=_15*1;
      f09age75_79=_16*1;
      f09age80_84=_17*1;
      f09age85=_18*1;

      drop _name_ _4 _5 _6 _7 _8 _9 _10 _11 _12 _13 _14 _15 _16 _17
_18;
run;

proc sort data=totpop09age;
      by geo_id2;
proc sort data=totmale09age;
      by geo_id2;
proc sort data=totfemale09age;
      by geo_id2;
data Age2009Totals;
      merge totpop09age totmale09age totfemale09age;
      by geo_id2;
run;



************************************************;
****************    2010 AGE    ***************;
************************************************;

* 2010 Total Pop Age Distribution*;
data age10pop;
      set age;
      where year1=2010.3;

      keep geo_id2 agegrp tot_pop;
```

```
run;

proc sort data=age10pop;
      by geo_id2 agegrp;
proc transpose data=age10pop out=pop10age;
      by geo_id2;
      id agegrp;
run;

data TotPop10Age;
      set pop10age;

      p10age15_19=_4*1;
      p10age20_24=_5*1;
      p10age25_29=_6*1;
      p10age30_34=_7*1;
      p10age35_39=_8*1;
      p10age40_44=_9*1;
      p10age45_49=_10*1;
      p10age50_54=_11*1;
      p10age55_59=_12*1;
      p10age60_64=_13*1;
      p10age65_69=_14*1;
      p10age70_74=_15*1;
      p10age75_79=_16*1;
      p10age80_84=_17*1;
      p10age85=_18*1;

      drop _name_ _4 _5 _6 _7 _8 _9 _10 _11 _12 _13 _14 _15 _16 _17
_18;
run;


* 2010 Total Male Pop Age Distribution;
data age10male;
      set age;
      where year1=2010.3;

      keep geo_id2 agegrp tot_male;
run;

proc sort data=age10male;
      by geo_id2 agegrp;
proc transpose data=age10male out=male10age;
      by geo_id2;
      id agegrp;
run;

data TotMale10Age;
      set male10age;

      m10age15_19=_4*1;
      m10age20_24=_5*1;
      m10age25_29=_6*1;
      m10age30_34=_7*1;
      m10age35_39=_8*1;
      m10age40_44=_9*1;
```

```
        m10age45_49=_10*1;
        m10age50_54=_11*1;
        m10age55_59=_12*1;
        m10age60_64=_13*1;
        m10age65_69=_14*1;
        m10age70_74=_15*1;
        m10age75_79=_16*1;
        m10age80_84=_17*1;
        m10age85=_18*1;

        drop _name_ _4 _5 _6 _7 _8 _9 _10 _11 _12 _13 _14 _15 _16 _17
_18;
run;

* 2010 Total Female Pop Age Distribution;
data age10female;
        set age;
        where year1=2010.3;

        keep geo_id2 agegrp tot_female;
run;

proc sort data=age10female;
        by geo_id2 agegrp;
proc transpose data=age10female out=female10age;
        by geo_id2;
        id agegrp;
run;

data TotFemale10Age;
        set female10age;

        f10age15_19=_4*1;
        f10age20_24=_5*1;
        f10age25_29=_6*1;
        f10age30_34=_7*1;
        f10age35_39=_8*1;
        f10age40_44=_9*1;
        f10age45_49=_10*1;
        f10age50_54=_11*1;
        f10age55_59=_12*1;
        f10age60_64=_13*1;
        f10age65_69=_14*1;
        f10age70_74=_15*1;
        f10age75_79=_16*1;
        f10age80_84=_17*1;
        f10age85=_18*1;

        drop _name_ _4 _5 _6 _7 _8 _9 _10 _11 _12 _13 _14 _15 _16 _17
_18;
run;

proc sort data=totpop10age;
        by geo_id2;
proc sort data=totmale10age;
        by geo_id2;
proc sort data=totfemale10age;
```

```
      by geo_id2;
data Age2010Totals;
      merge totpop10age totmale10age totfemale10age;
      by geo_id2;
run;



***********************************************;
****************    2011 AGE    ***************;
***********************************************;

* 2011 Total Pop Age Distribution*;
data age11pop;
      set age;
      where year1=2011;

      keep geo_id2 agegrp tot_pop;
run;

proc sort data=age11pop;
      by geo_id2 agegrp;
proc transpose data=age11pop out=pop11age;
      by geo_id2;
      id agegrp;
run;

data TotPop11Age;
      set pop11age;

      p11age15_19=_4*1;
      p11age20_24=_5*1;
      p11age25_29=_6*1;
      p11age30_34=_7*1;
      p11age35_39=_8*1;
      p11age40_44=_9*1;
      p11age45_49=_10*1;
      p11age50_54=_11*1;
      p11age55_59=_12*1;
      p11age60_64=_13*1;
      p11age65_69=_14*1;
      p11age70_74=_15*1;
      p11age75_79=_16*1;
      p11age80_84=_17*1;
      p11age85=_18*1;

      drop _name_ _4 _5 _6 _7 _8 _9 _10 _11 _12 _13 _14 _15 _16 _17
_18;
run;


* 2011 Total Male Pop Age Distribution;
data age11male;
      set age;
      where year1=2011;

      keep geo_id2 agegrp tot_male;
```

```
run;

proc sort data=age11male;
      by geo_id2 agegrp;
proc transpose data=age11male out=male11age;
      by geo_id2;
      id agegrp;
run;

data TotMale11Age;
      set male11age;

      m11age15_19=_4*1;
      m11age20_24=_5*1;
      m11age25_29=_6*1;
      m11age30_34=_7*1;
      m11age35_39=_8*1;
      m11age40_44=_9*1;
      m11age45_49=_10*1;
      m11age50_54=_11*1;
      m11age55_59=_12*1;
      m11age60_64=_13*1;
      m11age65_69=_14*1;
      m11age70_74=_15*1;
      m11age75_79=_16*1;
      m11age80_84=_17*1;
      m11age85=_18*1;

      drop _name_ _4 _5 _6 _7 _8 _9 _10 _11 _12 _13 _14 _15 _16 _17
_18;
run;

* 2011 Total Female Pop Age Distribution;
data age11female;
      set age;
      where year1=2011;

      keep geo_id2 agegrp tot_female;
run;

proc sort data=age11female;
      by geo_id2 agegrp;
proc transpose data=age11female out=female11age;
      by geo_id2;
      id agegrp;
run;

data TotFemale11Age;
      set female11age;

      f11age15_19=_4*1;
      f11age20_24=_5*1;
      f11age25_29=_6*1;
      f11age30_34=_7*1;
      f11age35_39=_8*1;
      f11age40_44=_9*1;
      f11age45_49=_10*1;
```

```
	f11age50_54=_11*1;
	f11age55_59=_12*1;
	f11age60_64=_13*1;
	f11age65_69=_14*1;
	f11age70_74=_15*1;
	f11age75_79=_16*1;
	f11age80_84=_17*1;
	f11age85=_18*1;

	drop _name_ _4 _5 _6 _7 _8 _9 _10 _11 _12 _13 _14 _15 _16 _17
_18;
run;

proc sort data=totpop11age;
	by geo_id2;
proc sort data=totmale11age;
	by geo_id2;
proc sort data=totfemale11age;
	by geo_id2;
data Age2011Totals;
	merge totpop11age totmale11age totfemale11age;
	by geo_id2;
run;


**********************************************************************;

* Merge all years together and calculate average age distributions;
proc sort data=Age2008Totals;
	by geo_id2;
proc sort data=Age2009Totals;
	by geo_id2;
proc sort data=Age2010Totals;
	by geo_id2;
proc sort data=Age2011Totals;
	by geo_id2;
data FinalAge;
	merge Age2008Totals Age2009Totals Age2010Totals Age2011Totals;

	*Total POPULATION Avg Age Distribution;
	Tot_15_19avg=(p08age15_19+p09age15_19+p10age15_19+p11age15_19)/4;
	Tot_20_24avg=(p08age20_24+p09age20_24+p10age20_24+p11age20_24)/4;
	Tot_25_29avg=(p08age25_29+p09age25_29+p10age25_29+p11age25_29)/4;
	Tot_30_34avg=(p08age30_34+p09age30_34+p10age30_34+p11age30_34)/4;
	Tot_35_39avg=(p08age35_39+p09age35_39+p10age35_39+p11age35_39)/4;
	Tot_40_44avg=(p08age40_44+p09age40_44+p10age40_44+p11age40_44)/4;
	Tot_45_49avg=(p08age45_49+p09age45_49+p10age45_49+p11age45_49)/4;
	Tot_50_54avg=(p08age50_54+p09age50_54+p10age50_54+p11age50_54)/4;
	Tot_55_59avg=(p08age55_59+p09age55_59+p10age55_59+p11age55_59)/4;
	Tot_60_64avg=(p08age60_64+p09age60_64+p10age60_64+p11age60_64)/4;
	Tot_65_69avg=(p08age65_69+p09age65_69+p10age65_69+p11age65_69)/4;
	Tot_70_74avg=(p08age70_74+p09age70_74+p10age70_74+p11age70_74)/4;
	Tot_75_79avg=(p08age75_79+p09age75_79+p10age75_79+p11age75_79)/4;
	Tot_80_84avg=(p08age80_84+p09age80_84+p10age80_84+p11age80_84)/4;
	Tot_85avg=(p08age85+p09age85+p10age85+p11age85)/4;

	*Total MALE pop Average Age Distribution;
```

```
	M_15_19avg=(m08age15_19+m09age15_19+m10age15_19+m11age15_19)/4;
	M_20_24avg=(m08age20_24+m09age20_24+m10age20_24+m11age20_24)/4;
	M_25_29avg=(m08age25_29+m09age25_29+m10age25_29+m11age25_29)/4;
	M_30_34avg=(m08age30_34+m09age30_34+m10age30_34+m11age30_34)/4;
	M_35_39avg=(m08age35_39+m09age35_39+m10age35_39+m11age35_39)/4;
	M_40_44avg=(m08age40_44+m09age40_44+m10age40_44+m11age40_44)/4;
	M_45_49avg=(m08age45_49+m09age45_49+m10age45_49+m11age45_49)/4;
	M_50_54avg=(m08age50_54+m09age50_54+m10age50_54+m11age50_54)/4;
	M_55_59avg=(m08age55_59+m09age55_59+m10age55_59+m11age55_59)/4;
	M_60_64avg=(m08age60_64+m09age60_64+m10age60_64+m11age60_64)/4;
	M_65_69avg=(m08age65_69+m09age65_69+m10age65_69+m11age65_69)/4;
	M_70_74avg=(m08age70_74+m09age70_74+m10age70_74+m11age70_74)/4;
	M_75_79avg=(m08age75_79+m09age75_79+m10age75_79+m11age75_79)/4;
	M_80_84avg=(m08age80_84+m09age80_84+m10age80_84+m11age80_84)/4;
	M_85avg=(m08age85+m09age85+m10age85+m11age85)/4;


	*Total FEMALE pop Average Age Distribution;
	F_15_19avg=(f08age15_19+f09age15_19+f10age15_19+f11age15_19)/4;
	F_20_24avg=(f08age20_24+f09age20_24+f10age20_24+f11age20_24)/4;
	F_25_29avg=(f08age25_29+f09age25_29+f10age25_29+f11age25_29)/4;
	F_30_34avg=(f08age30_34+f09age30_34+f10age30_34+f11age30_34)/4;
	F_35_39avg=(f08age35_39+f09age35_39+f10age35_39+f11age35_39)/4;
	F_40_44avg=(f08age40_44+f09age40_44+f10age40_44+f11age40_44)/4;
	F_45_49avg=(f08age45_49+f09age45_49+f10age45_49+f11age45_49)/4;
	F_50_54avg=(f08age50_54+f09age50_54+f10age50_54+f11age50_54)/4;
	F_55_59avg=(f08age55_59+f09age55_59+f10age55_59+f11age55_59)/4;
	F_60_64avg=(f08age60_64+f09age60_64+f10age60_64+f11age60_64)/4;
	F_65_69avg=(f08age65_69+f09age65_69+f10age65_69+f11age65_69)/4;
	F_70_74avg=(f08age70_74+f09age70_74+f10age70_74+f11age70_74)/4;
	F_75_79avg=(f08age75_79+f09age75_79+f10age75_79+f11age75_79)/4;
	F_80_84avg=(f08age80_84+f09age80_84+f10age80_84+f11age80_84)/4;
	F_85avg=(f08age85+f09age85+f10age85+f11age85)/4;

	keep geo_id2
	Tot_15_19avg M_15_19avg F_15_19avg
	Tot_20_24avg M_20_24avg F_20_24avg
	Tot_25_29avg M_25_29avg F_25_29avg
	Tot_30_34avg M_30_34avg F_30_34avg
	Tot_35_39avg M_35_39avg F_35_39avg
	Tot_40_44avg M_40_44avg F_40_44avg
	Tot_45_49avg M_45_49avg F_45_49avg
	Tot_50_54avg M_50_54avg F_50_54avg
	Tot_55_59avg M_55_59avg F_55_59avg
	Tot_60_64avg M_60_64avg F_60_64avg
	Tot_65_69avg M_65_69avg F_65_69avg
	Tot_70_74avg M_70_74avg F_70_74avg
	Tot_75_79avg M_75_79avg F_75_79avg
	Tot_80_84avg M_80_84avg F_80_84avg
	Tot_85avg M_85avg F_85avg;
run;

data a.FinalAge;
	set work.FinalAge;
run;
```

```
************************************************;
***************   MEDIAN AGE      ***************;
************************************************;
PROC IMPORT OUT= WORK.MedAge2008
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\AgeRaceSex\MedAge2008.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=3000;
run;

PROC IMPORT OUT= WORK.MedAge2009
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\AgeRaceSex\MedAge2009.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=3000;
run;

PROC IMPORT OUT= WORK.MedAge2010
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\AgeRaceSex\MedAge2010.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=3000;
run;

PROC IMPORT OUT= WORK.MedAge2011
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\AgeRaceSex\MedAge2011.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=3000;
run;

proc sort data=medage2008;
      by geo_id2;
proc sort data=medage2009;
      by geo_id2;
proc sort data=medage2010;
      by geo_id2;
proc sort data=medage2011;
      by geo_id2;

data MedAge;
      merge medage2008 medage2009 medage2010 medage2011;
      by geo_id2;

      MedAge=(MedianAge08+MedianAge09+MedianAge10+MedianAge11)/4;
      if MedianAge08=. then
MedAge=(MedianAge09+MedianAge10+MedianAge11)/3;
```

```
	M_MedAge=(Male_MedianAge08+Male_MedianAge09+Male_MedianAge10+Male
_MedianAge11)/4;
	if Male_MedianAge08=. then
M_MedAge=(Male_MedianAge09+Male_MedianAge10+Male_MedianAge11)/3;

	F_MedAge=(Female_MedianAge08+Female_MedianAge09+Female_MedianAge1
0+Female_MedianAge11)/4;
	if Female_MedianAge08=. then
F_MedAge=(Female_MedianAge09+Female_MedianAge10+Female_MedianAge11)/3;

	keep geo_id geo_id2 geo_display_label MedAge M_MedAge F_MedAge;
run;

data a.MedianAge;
	set MedAge;
run;



****************************************************;
****************    PRISON POP       ***************;
****************************************************;

PROC IMPORT OUT= WORK.Corrections
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\Correctional\Correctional_Po
p_2010.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=5000;
run;

data prison;
	set corrections;

	keep geo_id2 geo_display_label Corr_Pop;
run;

data a.prison;
	set work.prison;
run;



****************************************************;
****************    DRUG USE         ***************;
****************************************************;

* Drug Use during last month, average for 2008, 2009, 2010;
PROC IMPORT OUT= WORK.DrugUse
            DATAFILE= "H:\Classes\Thesis\Data\SocialDeterminants\Drug
Use\DrugUsePastMonth.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=5000;
```

```
run;

* Drug and Alcohol dependence, average for 2008, 2009, 2010;
PROC IMPORT OUT= WORK.DrugDep
            DATAFILE= "H:\Classes\Thesis\Data\SocialDeterminants\Drug
Use\DrugAlc_UseDependence.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=5000;
run;

* Drug Use (no MJ) during last month, average for 2008, 2009, 2010;
PROC IMPORT OUT= WORK.DrugUsenoMJ
            DATAFILE= "H:\Classes\Thesis\Data\SocialDeterminants\Drug
Use\DrugUsePastMonth_noMJ.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=5000;
run;


* Counties in each region;
PROC IMPORT OUT= WORK.DrugRegions
            DATAFILE= "H:\Classes\Thesis\Data\SocialDeterminants\Drug
Use\DrugUseRegions.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=5000;
run;

* Sort and  merge Drug Data with regions files by region;
proc sort data=druguse;
      by region;
proc sort data=drugdep;
      by region;
proc sort data=drugusenomj;
      by region;
proc sort data=drugregions;
      by region;
data DrugsA;
      merge druguse drugdep drugusenomj drugregions;
      by region;
run;
data DrugsCity;
      set DrugsA;
      where county contains ' City';

      county=lowcase(county);
run;
data DrugsB;
      set DrugsA;

      co='County';
      county=trim(county)||' '||trim(co);
```

```
        county=lowcase(county);
run;

data Drugs;
        set DrugsCity (in=a) DrugsB (in=b);

        keep county state region drugusemonth drugusemonthnomj
drugalc_usedep;
run;

data fips;
        set fips;
        county=lowcase(county);
run;

proc sort data=Drugs;
        by county;
proc sort data=fips;
        by county;
data drug;
        merge Drugs fips;
        by county;
run;

data a.drugs;
        set work.drug;
        where state eq 'MD' or state eq 'VA' or state eq 'NC';
run;



****************************************************;
****************    EDUCATION        ***************;
****************************************************;

PROC IMPORT OUT= WORK.educ2008 (RENAME=(Est_HS_Over25=HS08))
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\Education\Education2008.csv"
            DBMS=CSV REPLACE;
        GETNAMES=YES;
        DATAROW=2;
        guessingrows=5000;
run;

PROC IMPORT OUT= WORK.educ2009 (RENAME=(Est_HS_Over25=HS09))
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\Education\Education2009.csv"
            DBMS=CSV REPLACE;
        GETNAMES=YES;
        DATAROW=2;
        guessingrows=5000;
run;

PROC IMPORT OUT= WORK.educ2010 (RENAME=(Est_HS_Over25=HS10))
```

```
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\Education\Education2010.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=5000;
run;

PROC IMPORT OUT= WORK.educ2011 (RENAME=(Est_HS_Over25=HS11))
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\Education\Education2011.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=5000;
run;

proc sort data=educ2008;
      by geo_id2;
proc sort data=educ2009;
      by geo_id2;
proc sort data=educ2010;
      by geo_id2;
proc sort data=educ2011;
      by geo_id2;
data Education;
      merge educ2008 educ2009 educ2010 educ2011;
      by geo_id2;

      HSgrad=(HS08+HS09+HS10+HS11)/4;
      if HS08 = . then HSgrad=(HS09+HS10+HS11)/3;

      keep geo_id geo_id2 geo_display_label HSgrad;
run;

data a.Educ;
      set work.Education;
run;



****************************************************************;
****************   GINI INCOME INEQUALITY       ***************;
****************************************************************;
PROC IMPORT OUT= WORK.gini2008
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\Gini\Gini2008.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=5000;
run;

PROC IMPORT OUT= WORK.gini2009
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\Gini\Gini2009.csv"
            DBMS=CSV REPLACE;
```

```
	GETNAMES=YES;
	DATAROW=2;
	guessingrows=5000;
run;

PROC IMPORT OUT= WORK.gini2010
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\Gini\Gini2010.csv"
            DBMS=CSV REPLACE;
	GETNAMES=YES;
	DATAROW=2;
	guessingrows=5000;
run;

PROC IMPORT OUT= WORK.gini2011
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\Gini\Gini2011.csv"
            DBMS=CSV REPLACE;
	GETNAMES=YES;
	DATAROW=2;
	guessingrows=5000;
run;

* Sort Gini data by county fips code and merge the four datasets to
create one Gini dataset;
* Calculate average Gini Coefficient over four years;
proc sort data=gini2008;
	by geo_id2;
proc sort data=gini2009;
	by geo_id2;
proc sort data=gini2010;
	by geo_id2;
proc sort data=gini2011;
	by geo_id2;
data Gini;
	merge gini2008 gini2009 gini2010 gini2011;
	by geo_id2;

	Gini=(gini08+gini09+gini10+gini11)/4;

	* Counties missing data for 2008-2009;
	if gini08=. and gini09=. then Gini=(gini10+gini11)/2;

	keep geo_id geo_id2 geo_display_label gini gini08 gini09 gini10
gini11;
run;

data a.Gini;
	set work.gini;
run;


*********************************************************;
******       HEALTH EXPENDITURES and INSURANCE    *******;
*********************************************************;

*per capita health expenditures;
```

```
PROC IMPORT OUT= WORK.HealthCosts
            DATAFILE= "H:\Classes\Thesis\Data\SocialDeterminants\Health
Expenditures\CMS_StateSpending.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=5000;
run;


PROC IMPORT OUT= WORK.Insured2008 (rename=(_stcou=geo_id2))
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\Insurance\sahie2008.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=200000;
run;

PROC IMPORT OUT= WORK.Insured2009 (rename=(_stcou=geo_id2))
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\Insurance\sahie2009.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=200000;
run;

PROC IMPORT OUT= WORK.Insured2010 (rename=(_stcou=geo_id2))
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\Insurance\sahie2010.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=200000;
run;

PROC IMPORT OUT= WORK.Insured2011 (rename=(_stcou=geo_id2))
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\Insurance\sahie2011.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=200000;
run;

* Get rid of observations for other states and stratified observations;
data insured2008 (rename=(_name=name));
      set insured2008;
      where  (_agecat=0 and _racecat=0 and _sexcat=0 and _iprcat=0)
            and (geo_id2 ge 24001 and geo_id2 le 24510)
            or (_agecat=0 and _racecat=0 and _sexcat=0 and _iprcat=0)
and (geo_id2 ge 37001 and geo_id2 le 37195)
            or (_agecat=0 and _racecat=0 and _sexcat=0 and
_iprcat=0)and (geo_id2 ge 51003 and geo_id2 le 51810);
run;
```

```
data insured2009 (rename=(_name=name));
      set insured2009;
      where  (_agecat=0 and _racecat=0 and _sexcat=0 and _iprcat=0)
            and (geo_id2 ge 24001 and geo_id2 le 24510)
            or (_agecat=0 and _racecat=0 and _sexcat=0 and _iprcat=0)
and (geo_id2 ge 37001 and geo_id2 le 37195)
            or (_agecat=0 and _racecat=0 and _sexcat=0 and
_iprcat=0)and (geo_id2 ge 51003 and geo_id2 le 51810);
run;

data insured2010 (rename=(_name=name));
      set insured2010;
      where  (_agecat=0 and _racecat=0 and _sexcat=0 and _iprcat=0)
            and (geo_id2 ge 24001 and geo_id2 le 24510)
            or (_agecat=0 and _racecat=0 and _sexcat=0 and _iprcat=0)
and (geo_id2 ge 37001 and geo_id2 le 37195)
            or (_agecat=0 and _racecat=0 and _sexcat=0 and
_iprcat=0)and (geo_id2 ge 51003 and geo_id2 le 51810);
run;

data insured2011 (rename=(_name=name));
      set insured2011;
      where  (_agecat=0 and _racecat=0 and _sexcat=0 and _iprcat=0)
            and (geo_id2 ge 24001 and geo_id2 le 24510)
            or (_agecat=0 and _racecat=0 and _sexcat=0 and _iprcat=0)
and (geo_id2 ge 37001 and geo_id2 le 37195)
            or (_agecat=0 and _racecat=0 and _sexcat=0 and
_iprcat=0)and (geo_id2 ge 51003 and geo_id2 le 51810);
run;

* Sort Insurance data by county fips code and merge the four dataset;
* to create one Insurance dataset;
* Calculate average number and percentage of insured and uninsured over
* four years;
proc sort data=insured2008;
      by geo_id2;
proc sort data=insured2009;
      by geo_id2;
proc sort data=insured2010;
      by geo_id2;
proc sort data=insured2011;
      by geo_id2;
data insured;
      merge insured2008 insured2009 insured2010 insured2011;
      by geo_id2;

      Num_Ins=(Num_Insured08+Num_Insured09+Num_Insured10+Num_Insured11)
/4;
      Num_Unins=(Num_Uninsured08+Num_Uninsured09+Num_Uninsured10+Num_Un
insured11)/4;
      Pct_Unins=(PCT_Uninsured08+PCT_Uninsured09+PCT_Uninsured10+PCT_Un
insured11)/4;

      if geo_id2 ge 24001 and geo_id2 le 24510 then State='MD';
      if geo_id2 ge 37001 and geo_id2 le 37195 then State='NC';
      if geo_id2 ge 51003 and geo_id2 le 51810 then State='VA';
```

```
	keep geo_id2 state name Num_Ins Num_Unins Pct_Unins;
run;

data a.insuredcosts;
	merge insured healthcosts;
	by state;
run;



*******************************************************;
****************   MSM POPULATION        *************;
*******************************************************;
PROC IMPORT OUT= WORK.msm2008
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\MSM\MSM2008.csv"
            DBMS=CSV REPLACE;
	GETNAMES=YES;
	DATAROW=2;
	guessingrows=5000;
run;

PROC IMPORT OUT= WORK.msm2009
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\MSM\MSM2009.csv"
            DBMS=CSV REPLACE;
	GETNAMES=YES;
	DATAROW=2;
	guessingrows=5000;
run;

PROC IMPORT OUT= WORK.msm2010
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\MSM\MSM2010.csv"
            DBMS=CSV REPLACE;
	GETNAMES=YES;
	DATAROW=2;
	guessingrows=5000;
run;

PROC IMPORT OUT= WORK.msm2011
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\MSM\MSM2011.csv"
            DBMS=CSV REPLACE;
	GETNAMES=YES;
	DATAROW=2;
	guessingrows=5000;
run;

* Sort MSM data by county fips code and merge the four datasets to
create one MSM dataset;
* Calculate average percentage of MSM population over four years;
proc sort data=msm2008;
	by geo_id2;
proc sort data=msm2009;
	by geo_id2;
proc sort data=msm2010;
```

```
        by geo_id2;
proc sort data=msm2011;
        by geo_id2;
data MSM;
        merge msm2008 msm2009 msm2010 msm2011;
        by geo_id2;

        MSM=(msm08+msm09+msm10+msm11)/4;

        * Counties missing data for some of the years;
        if msm08=. then MSM=(msm09+msm10+msm11)/3;


        keep geo_id geo_id2 geo_display_label MSM msm08 msm09 msm10
msm11;
run;

data a.MSM;
        set work.MSM;
run;





***************************************************************;
****************    POVERTY and MEDIAN INCOME    ***************;
***************************************************************;
PROC IMPORT OUT= WORK.income2008
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\PovertyMedIncome\saipe2008.c
sv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=5000;
run;

PROC IMPORT OUT= WORK.income2009
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\PovertyMedIncome\saipe2009.c
sv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=5000;
run;

PROC IMPORT OUT= WORK.income2010
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\PovertyMedIncome\saipe2010.c
sv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=5000;
run;
```

```
PROC IMPORT OUT= WORK.income2011
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\PovertyMedIncome\saipe2011.c
sv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=5000;
run;

* Get rid of observations from other states;
data income2008;
      set income2008;
      where (postal='MD' or postal='NC' or postal='VA');

      newfips=PUT(county_fips, z3.);
      fips=trim(state_fips)||newfips;

      keep fips postal name poverty08 pctpoverty08 medincome08;
run;

data income2009;
      set income2009;
      where postal='MD' or postal='NC' or postal='VA';

      newfips=PUT(county_fips, z3.);
      fips=trim(state_fips)||newfips;

      keep fips postal name poverty09 pctpoverty09 medincome09;
run;

data income2010;
      set income2010;
      where postal='MD' or postal='NC' or postal='VA';

      newfips=PUT(county_fips, z3.);
      fips=trim(state_fips)||newfips;

      keep fips postal name poverty10 pctpoverty10 medincome10;
run;

data income2011;
      set income2011;
      where postal='MD' or postal='NC' or postal='VA';

      newfips=PUT(county_fips, z3.);
      fips=trim(state_fips)||newfips;

      keep fips postal name poverty11 pctpoverty11 medincome11;
run;


* Sort Income and Poverty data by county fips code and merge the four
datasets to create one dataset;
* Calculate average number and percentage of people living in poverty
and median income over four years;
```

```
proc sort data=income2008;
	by fips;
proc sort data=income2009;
	by fips;
proc sort data=income2010;
	by fips;
proc sort data=income2011;
	by fips;
data Income;
	merge income2008 income2009 income2010 income2011;
	by fips;

	poverty=(poverty08+poverty09+poverty10+poverty11)/4;
	pctpoverty=(pctpoverty08+pctpoverty09+pctpoverty10+pctpoverty11)/
4;
	medincome=(medincome08+medincome09+medincome10+medincome11)/4;

	geo_id2=fips*1;

	keep geo_id2 postal name poverty pctpoverty medincome;
run;

data a.Income;
	set work.income;
run;



**************************************************;
***************   URBANICITY      **************;
**************************************************;

/*
2006 Urbanicity Classifications:
	1=Large Central Metro
	2=Large Fringe Metro
	3=Medium Metro
	4=Small Metro
	5=Micropolitan
	6=Noncore
*/

PROC IMPORT OUT= WORK.urbanicity (rename=(ST=STATE))
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\Urbanicity\CountyUrbanicity2
006.csv"
            DBMS=CSV REPLACE;
	GETNAMES=YES;
	DATAROW=2;
	guessingrows=5000;
run;

proc sort data=urbanicity;
	by geo_id2;
proc sort data=fips;
	by geo_id2;
data urban;
```

```
      merge urbanicity (in=a) fips (in=b);
      by geo_id2;
      where STATE='MD' or STATE='NC' or STATE='VA';

      if a and b;
run;

data a.Urban;
      set work.Urban;
run;


***********************************************************;
****************    POPULATION DENSITY       ***************;
***********************************************************;
PROC IMPORT OUT= WORK.popdensity
            DATAFILE=
"H:\Classes\Thesis\Data\SocialDeterminants\Density\PopDensity2010.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=5000;
run;

proc sort data=popdensity;
      by geo_id2;
proc sort data=fips;
      by geo_id2;
data density;
      merge popdensity (in=a) fips (in=b);
      by geo_id2;

      PopDensity=PopSqMile*1;
      HouseDensity=HouseSqMile*1;

      if a and b;
      keep geo_id2 geo_display_label county state PopDensity
HouseDensity;
run;

data a.density;
      set work.density;
run;

*====================================================;
****************************************************;
*====================================================;

* Merge all social determinants datasets into one dataset;
proc sort data=a.fips;
      by geo_id2;
proc sort data=a.density;
      by geo_id2;
proc sort data=a.TotRaceSex;
      by geo_id2;
proc sort data=a.FinalAge;
      by geo_id2;
```

```
proc sort data=a.MedianAge;
      by geo_id2;
proc sort data=a.Prison;
      by geo_id2;
proc sort data=a.Drugs;
      by geo_id2;
proc sort data=a.Educ;
      by geo_id2;
proc sort data=a.Gini;
      by geo_id2;
proc sort data=a.InsuredCosts;
      by geo_id2;
proc sort data=a.MSM;
      by geo_id2;
proc sort data=a.Income;
      by geo_id2;
proc sort data=a.Urban;
      by geo_id2;

data a.social;
      merge a.fips a.density a.TotRaceSex a.FinalAge a.MedianAge
a.Prison a.Drugs a.Educ a.Gini a.InsuredCosts a.Msm a.Income a.Urban;
      by geo_id2;

      where geo_id2 ne 24000 and geo_id2 ne 37000 and geo_id2 ne 51000;
      CostsAvg=(Costs08+Costs09)/2;
run;


* Create Final Dataset for Analysis. Make a permanent dataset;
********************************************************************;

proc sort data=a.social;
      by state ctyname;
proc sort data=a.countyhiv;
      by state ctyname;
data FinalHIV;
      merge a.social (in=a) a.countyhiv (in=b);
      by state ctyname;

      where state='NC' or state='MD' or state='VA';


      if state='VA' then exp=1;
      if state='MD' or state='NC' then exp=0;

      drop state;
run;

data a.ThesisData;
      set FinalHIV;

      state=exp*1;
      drop exp;
run;
```

```
***********************************************;
*     Thesis Code part 2                       *;
*     Data Analysis                            *;
*     Written By: Noel Hatley                  *;
*     Date Febraury 10, 2014                   *;
***********************************************;
ods html close;
ods html;

OPTIONS nofmterr MPRINT SYMBOLGEN mlogic;
libname a 'H:\Classes\Thesis\Data\SocialDeterminants';
%include "H:\Classes\Thesis\Data\collin_2011.sas";

/*
      VA = 1
      MD & NC = 0
*/

* FORMATS;
proc format;
      value sexf   1="Majority of Pop Male"
                         0="Majority of Pop Female";
      value statef 1='VA'
                         0='MD & NC';
      value urbanf 1="Large Central Metro"
                         2="Large Fringe Metro"
                         3="Medium Metro"
                         4="Small Metro"
                         5="Micropolitan"
                         6="Noncore";
run;

data thesisdata;
      set a.thesisdata;
      costsavg=(costs08+costs09)/2;
run;

* Create two datasets, one containing population counts the other with
population proportions;

*Dataset #1: POPULATION COUNTS;
*****************************************************;
data temp_counts;
      set thesisdata;

      *Counts;
      Tot_Pop=tot_pop_avg*1;
      Pop_Density=PopDensity*1;

      Num_Males=tot_males_avg;
      Num_Females=tot_females_avg;

      Num_Poverty=poverty;
      Num_Prison=Corr_pop*1;
```

```
	Num_HSgrad=(HSgrad/100)*tot_pop_avg;
	Num_MSM=MSM*1;
	Num_NoIns=(Pct_Unins/100)*tot_pop_avg;

	Num_Asian=tot_asian_avg*1;
	Num_Black=tot_black_avg*1;
	Num_Hisp=tot_hispanic_avg*1;
	Num_Indian=tot_Indian_avg*1;
	Num_NH=tot_nh_avg*1;
	Num_Pacific=tot_Pacific_avg*1;
	Num_TwoRace=tot_tworace_avg*1;
	Num_White=tot_White_avg*1;
	Num_Other=(tot_asian_avg+tot_indian_avg+tot_pacific_avg+tot_twora
ce_avg)*1;

	Num_DrugUse=DrugUseMonthNoMJ*tot_pop_avg;
	Num_DrugDep=DrugAlc_UseDep*tot_pop_avg;

	*Categorical variables;
	Urban=Urban_2006*1;

	* Urban Dummy Variables;
	if Urban = 6 then urban6=1;
	else urban6=0;
	if Urban = 2 then urban2=1;
	else urban2=0;
	if Urban = 3 then urban3=1;
	else urban3=0;
	if Urban = 4 then urban4=1;
	else urban4=0;
	if Urban = 5 then urban5=1;
	else urban5=0;
	if Urban = 1 then urban6=urban2=urban3=urban4=urban5=0; *ref
group;

	keep geo_id2 state CTYcase Tot_pop Pop_Density Num_Males
Num_Females MedAge M_MedAge F_MedAge Num_Asian Num_Black
		Num_Hisp Num_Indian Num_Pacific Num_TwoRace Num_White
Num_NH Num_Other Num_Poverty Num_Prison CostsAvg Num_DrugUse
		Num_DrugDep Gini Num_HSgrad Num_MSM Num_NoIns MedIncome
HouseDensity Urban Urban2 Urban3 Urban4 Urban5 Urban6;
run;

* Dataset #2: POPULATION RATES;
*****************************************************;
data temp_rates;
	set thesisdata;

	*Counts;
	Tot_Pop=tot_pop_avg*1;
	Pop_Density=PopDensity*1;

	*Proportions;
	Rate_Males=(tot_males_avg/tot_pop_avg)*1000;
	Rate_Females=(tot_females_avg/tot_pop_avg)*1000;

	Rate_Poverty=(PctPoverty/100)*1000;
```

```
	Rate_Prison=(Corr_pop/tot_pop_avg)*1000;
	Rate_HSgrad=(HSgrad/100)*1000;
	Rate_MSM=(MSM/tot_pop_avg)*1000;
	Rate_NoIns=(Pct_Unins/100)*1000;

	Rate_Asian=(tot_asian_avg/tot_pop_avg)*1000;
	Rate_Black=(tot_black_avg/tot_pop_avg)*1000;
	Rate_Hisp=(tot_hispanic_avg/tot_pop_avg)*1000;
	Rate_Indian=(tot_Indian_avg/tot_pop_avg)*1000;
	Rate_NH=(tot_nh_avg/tot_pop_avg)*1000;
	Rate_Pacific=(tot_Pacific_avg/tot_pop_avg)*1000;
	Rate_TwoRace=(tot_tworace_avg/tot_pop_avg)*1000;
	Rate_White=(tot_White_avg/tot_pop_avg)*1000;
	Rate_Other=((tot_asian_avg+tot_indian_avg+tot_pacific_avg+tot_two
race_avg)/tot_pop_avg)*1000;

	Rate_DrugUse=DrugUseMonthNoMJ*1000;
	Rate_DrugDep=DrugAlc_UseDep*1000;

	*Categorical variables;
	Urban=Urban_2006*1;

	* Urban Dummy Variables;
	if Urban = 6 then urban6=1;
	else urban6=0;
	if Urban = 2 then urban2=1;
	else urban2=0;
	if Urban = 3 then urban3=1;
	else urban3=0;
	if Urban = 4 then urban4=1;
	else urban4=0;
	if Urban = 5 then urban5=1;
	else urban5=0;
	if Urban = 1 then urban6=urban2=urban3=urban4=urban5=0; *ref
group;

	keep geo_id2 State CTYcase Tot_pop Pop_Density Rate_Males
Rate_Females MedAge M_MedAge F_MedAge Rate_Asian Rate_Black
		Rate_Hisp Rate_Indian Rate_Pacific Rate_TwoRace Rate_White
Rate_NH Rate_Other Rate_Poverty Rate_Prison CostsAvg Rate_DrugUse
		Rate_DrugDep Gini Rate_HSgrad Rate_MSM Rate_NoIns
MedIncome HouseDensity Urban Urban2 Urban3 Urban4 Urban5 Urban6;
run;

* Create Data Sets including only available data (unsuppressed HIV
data);

* Dataset #1a;
data thesis_counts;
	set temp_counts;

	log_pop=log(tot_pop);

	newvar=input(CTYcase,comma6.);
	drop CTYcase;
	HIV=round(newvar,1);
run;
```

```
* Dataset #2a;
data thesis_rates;
      set temp_rates;

      log_pop=log(tot_pop);

      newvar=input(CTYcase,comma6.);
      drop CTYcase;
      HIV=round(newvar,1);
run;

* TRANSFORM NON-NORMAL variables;
/*
proc print data=thesis_rates;
      where rate_msm = 0;
run;

proc means data=trans_rates n sum mean;
      var rate_noins;
      where state=1;
run;
*/

* Variables to transform: Black, Hisp, Other, Total Pop, Pop Density,
House Density, Urbanicity, Med INcome, Costs, Males;
* MSM, and Prison;
Data trans_rates;
      set thesis_rates;

      Log_BlackRR=log(Rate_Black/Rate_White);
      Log_HispRR=log(Rate_Hisp/Rate_White);
      Log_OtherRR=log(Rate_Other/Rate_White);

      log_Tot_Pop=log(tot_pop);
      log_popDensity=log(Pop_Density);
      log_houseDensity=log(housedensity);
      log_MedIncome=log(medIncome);
      log_msm=log(rate_msm);
      log_prison=log(rate_prison);
      log_costs=log(costsavg);

      Tot_Costs=(costsavg*tot_pop)/1000000;
      log_costs=log(tot_costs);

      sexmf=rate_males/rate_females;


      * since 4 counties have zero people in prison, convert them to
zero on log scale;

      if geo_id2=51670 then log_msm=-10;
      if geo_id2=37029 or geo_id2=37043 or geo_id2=37073 or
geo_id2=37099 or geo_id2=37113 or geo_id2=37117 or geo_id2=37121 or
geo_id2=37125     or geo_id2=37143 or geo_id2=37173 or geo_id2=37187
```

```
      or geo_id2=51005 or geo_id2=51007 or geo_id2=51011 or
geo_id2=51017 or geo_id2=51530 or geo_id2=51035 or geo_id2=51036 or
geo_id2=51540 or geo_id2=51570 or geo_id2=51045 or geo_id2=51049
      or geo_id2=51595 or geo_id2=51057 or geo_id2=51600 or
geo_id2=51610 or geo_id2=51063 or geo_id2=51630 or geo_id2=51640 or
geo_id2=51071 or geo_id2=51077 or geo_id2=51079 or geo_id2=51091
      or geo_id2=51670 or geo_id2=51093 or geo_id2=51099 or
geo_id2=51101 or geo_id2=51097 or geo_id2=51678 or geo_id2=51109 or
geo_id2=51685 or geo_id2=51683 or geo_id2=51115 or geo_id2=51125
      or geo_id2=51133 or geo_id2=51720 or geo_id2=51735 or
geo_id2=51750 or geo_id2=51177 or geo_id2=51790 or geo_id2=51181 or
geo_id2=51820 or geo_id2=51193 or geo_id2=51830 or geo_id2=51840
      or geo_id2=51197 or geo_id2=51199 then log_prison=-10;

      if geo_id2=37007 or geo_id2=37011 or geo_id2=37015 or
geo_id2=37041 or geo_id2=37073 or geo_id2=37095 or geo_id2=37177 or
geo_id2=51007     or geo_id2=51017 or geo_id2=51515 or geo_id2=51021
      or geo_id2=51530 or geo_id2=51037 or geo_id2=51570 or
geo_id2=51049 or geo_id2=51057 or geo_id2=51610 or geo_id2=51063 or
geo_id2=51620 or geo_id2=51073 or geo_id2=51091 or geo_id2=51670
      or geo_id2=51097 or geo_id2=51115 or geo_id2=51133 or
geo_id2=51720 or geo_id2=51135 or geo_id2=51149 or geo_id2=51167 or
geo_id2=51193     or geo_id2=51195 then log_msm=-10;

      if  geo_id2=37197 or geo_id2=37199 then log_costs=log(6321.5);
      if  geo_id2=51001 or geo_id2=51820 or geo_id2=51830 or
geo_id2=51840 then log_costs=log(6167.5);

      if geo_id2=51001 or geo_id2=51820 or geo_id2=51830 or
geo_id2=51840 then rate_noIns=159.392;
      if geo_id2=37197 or geo_id2=37199 then rate_noIns=181.676;

      if rate_males ge 500 then sex=1;
      if rate_males lt 500 then sex=0;
run;



* Create dataset with only the transformed variables and normal
variables;
* Final Dataset;
data analysis;
      set trans_rates;


      where HIV gt 0;
      keep State HIV geo_id2 log_Pop log_PopDensity log_HouseDensity
Sex sexmf MedAge log_BlackRR log_HispRR log_OtherRR log_MSM
log_MedIncome
             rate_poverty Gini rate_HSgrad log_Prison rate_NoIns
log_Costs rate_DrugUse rate_DrugDep Urban;
run;


/*
proc contents data=analysis;
```

```
run;
*/


*******************************************************;
*=====================================================;
**********      Descriptive Statistics     ************;
*=====================================================;
*******************************************************;


* Examine Distributions of each Variables;

*HIV;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
      var HIV;
      id geo_id2;
      histogram HIV / normal;
      title 'Distribution of New Cases of HIV';
      probplot HIV / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability Plot of New Cases of HIV';
run;

*Total Population;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
      var tot_pop;
      id geo_id2;
      histogram tot_pop / normal;
      title 'Distribution of Total Population';
      probplot tot_pop / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability of Total Population Dist';
run;

*Pop Density;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
      var pop_density;
      id geo_id2;
      histogram pop_density / normal;
      title 'Distribution of Population Density';
      probplot pop_density / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability of Popualtion Density Dist';
run;

*House Density;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
      var Housedensity;
      id geo_id2;
      histogram Housedensity / normal;
      title 'Distribution of House Density';
      probplot Housedensity / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability of House Density Dist';
run;
```

```
*Median Age;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
      var MedAge;
      id geo_id2;
      histogram MedAge / normal;
      title 'Distribution of Median Age';
      probplot MedAge / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability of Median Age Dist';
run;

*Median Income;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
      var medincome;
      id geo_id2;
      histogram medincome / normal;
      title 'Distribution of Median Income';
      probplot medincome / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability of Median Income Dist';
run;

*Healthcare Expenditure per capita;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
      var CostsAvg;
      id geo_id2;
      histogram CostsAvg / normal;
      title 'Distribution of Healthcare Expenditure';
      probplot CostsAvg / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability of Healthcare Expenditure Dist';
run;

* Total Healthcare Expenditures in millions;
proc univariate data=trans_rates;
      format state statef.;
      var Tot_Costs;
      id geo_id2;
      histogram Tot_Costs / normal;
      title 'Distribution of Tot_Costs in Millions of Dollars Spent';
      probplot Tot_Costs / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability Plot';
      by state;
run;
QUIT;

*Gini;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
      var Gini;
      id geo_id2;
      histogram Gini / normal;
      title 'Distribution of Gini';
      probplot Gini / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability Plot of Gini Dist';
run;
```

```
*Males;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
      var Rate_Males;
      id geo_id2;
      histogram Rate_Males / normal;
      title 'Distribution of Males';
      probplot Rate_Males / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability Plot of Males Dist';
run;

*Females;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
      var Rate_Females;
      id geo_id2;
      histogram Rate_Females / normal;
      title 'Distribution of Females';
      probplot Rate_Females / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability Plot of Females Dist';
run;

*Hispanic/Latino;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
      var Rate_Hisp;
      id geo_id2;
      histogram Rate_Hisp / normal;
      title 'Distribution of Hispanic/Latino';
      probplot Rate_Hisp / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability Plot of Hispanic/Latino Dist';
run;

*Black;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
      var Rate_Black;
      id geo_id2;
      histogram Rate_Black / normal;
      title 'Distribution of Black';
      probplot Rate_Black / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability Plot of Black';
run;

*White;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
      var Rate_White;
      id geo_id2;
      histogram Rate_White / normal;
      title 'Distribution of White';
      probplot Rate_White / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability Plot of White';
run;

*Other Race;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
```

```
proc univariate data=thesis_rates;
	var Rate_Other;
	id geo_id2;
	histogram Rate_Other / normal;
	title 'Distribution of Other Race';
	probplot Rate_Other / normal (mu=est sigma=est color=blue w=1);
	title 'Normal Probability Plot of Other Race';
run;

*Poverty;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
	var Rate_Poverty;
	id geo_id2;
	histogram Rate_Poverty / normal;
	title 'Distribution of Poverty';
	probplot Rate_Poverty / normal (mu=est sigma=est color=blue w=1);
	title 'Normal Probability Plot of Poverty';
run;

*Education;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
	var Rate_HSgrad;
	id geo_id2;
	histogram Rate_HSgrad / normal;
	title 'Distribution of Education';
	probplot Rate_HSgrad / normal (mu=est sigma=est color=blue w=1);
	title 'Normal Probability Plot of Education';
run;

*MSM;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
	var Rate_MSM;
	id geo_id2;
	histogram Rate_MSM / normal;
	title 'Distribution of MSM';
	probplot Rate_MSM / normal (mu=est sigma=est color=blue w=1);
	title 'Normal Probability Plot of MSM';
run;

*No Health Insurance;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
	var Rate_NoIns;
	id geo_id2;
	histogram Rate_NoIns / normal;
	title 'Distribution of No Health Insurance';
	probplot Rate_NoIns / normal (mu=est sigma=est color=blue w=1);
	title 'Normal Probability Plot of No Health Insurance';
run;

*No Prison Pop;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
	var Rate_Prison;
```

```
        id geo_id2;
        histogram Rate_Prison / normal;
        title 'Distribution of Prison Pop';
        probplot Rate_Prison / normal (mu=est sigma=est color=blue w=1);
        title 'Normal Probability Plot of Prison Pop';
run;

*Drug Use;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
        var Rate_DrugUse;
        id geo_id2;
        histogram Rate_DrugUse / normal;
        title 'Distribution of Drug Use';
        probplot Rate_DrugUse / normal (mu=est sigma=est color=blue w=1);
        title 'Normal Probability Plot of Drug Use';
run;

*Drug Dependance;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
        var Rate_DrugDep;
        id geo_id2;
        histogram Rate_DrugDep / normal;
        title 'Distribution of Drug Dependance';
        probplot Rate_DrugDep / normal (mu=est sigma=est color=blue w=1);
        title 'Normal Probability Plot of Drug Dependance';
run;


*BY STATE;
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=thesis_rates;
        format state statef.;
        var HIV Tot_pop Pop_Density Rate_Males Rate_Females MedAge
M_MedAge F_MedAge Rate_Asian Rate_Black
            Rate_Hisp Rate_Indian Rate_Pacific Rate_TwoRace Rate_White
Rate_NH Rate_Other Rate_Poverty Rate_Prison CostsAvg Rate_DrugUse
            Rate_DrugDep Gini Rate_HSgrad Rate_MSM Rate_NoIns MedIncome
HouseDensity;
        id geo_id2;
        histogram HIV Tot_pop Pop_Density Rate_Males Rate_Females MedAge
M_MedAge F_MedAge Rate_Asian Rate_Black
            Rate_Hisp Rate_Indian Rate_Pacific Rate_TwoRace Rate_White
Rate_NH Rate_Other Rate_Poverty Rate_Prison CostsAvg Rate_DrugUse
            Rate_DrugDep Gini Rate_HSgrad Rate_MSM Rate_NoIns MedIncome
HouseDensity / normal;
        title 'Distribution';
        probplot HIV Tot_pop Pop_Density Rate_Males Rate_Females MedAge
M_MedAge F_MedAge Rate_Asian Rate_Black
            Rate_Hisp Rate_Indian Rate_Pacific Rate_TwoRace Rate_White
Rate_NH Rate_Other Rate_Poverty Rate_Prison CostsAvg Rate_DrugUse
            Rate_DrugDep Gini Rate_HSgrad Rate_MSM Rate_NoIns MedIncome
HouseDensity / normal (mu=est sigma=est color=blue w=1);
        title 'Normal Probability Plot';
        By state;
run;
```

```
*State;
* Use exact methods becuase some of the cells have 5 or less counties;
proc freq data=trans_rates;
      tables state*sex;
      exact pchi;
      format state statef. sex sexf.;
      title "Association between STATE and GENDER";
run;
*Mantel Haenszel;
proc freq data=trans_rates;
      tables sex_cat*state / chisq measures cl;
      format sex_cat sexf. state statef.;
      title "Prdinal Association between STATE and GENDER makeup in
counties";
run;




*Urbanicity;
* Use exact methods becuase some of the cells have 5 or less counties;
proc freq data=trans_rates;
      tables state*urban;
      exact pchi;
      format urban urbanf. state statef.;
      title "Association between STATE and URBANICITY";
run;
*Mantel Haenszel;
proc freq data=trans_rates;
      tables sex_cat*state / chisq measures cl;
      format sex_cat sexf. state statef.;
      title "Prdinal Association between STATE and GENDER makeup in
counties";
run;


* Histograms and descriptives of Transformed Variables;
proc univariate data=trans_rates;
      format state statef.;
      var log_costs;
      id geo_id2;
      histogram log_costs / normal;
      title 'Distribution of log_costs';
      probplot log_costs / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability Plot';
run;
QUIT;

proc univariate data=trans_rates;
      format state statef.;
      var log_prison;
      id geo_id2;
      histogram log_prison / normal;
      title 'Distribution of log_prison';
      probplot log_prison / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability Plot';
```

```
run;
QUIT;

proc print data=trans_rates;
      where geo_id2=37125 or geo_id2=51540 or geo_id2=51670 or
geo_id2=51177;
run;

*Sex;
* Use exact methods becuase some of the cells have 5 or less counties;
proc freq data=trans_rates;
      tables sex_cat*state;
      exact pchi;
      format sex_cat sexf. state statef.;
      title "Association between STATE and GENDER makeup in counties";
run;
*Mantel Haenszel;
proc freq data=trans_rates;
      tables sex_cat*state / chisq measures cl;
      format sex_cat sexf. state statef.;
      title "Prdinal Association between STATE and GENDER makeup in
counties";
run;

proc univariate data=trans_rates;
      format state statef.;
      var log_msm;
      id geo_id2;
      histogram log_msm / normal;
      title 'Distribution of log_msm';
      probplot log_msm / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability Plot';
run;
QUIT;

proc univariate data=trans_rates;
      format state statef.;
      var log_MedIncome;
      id geo_id2;
      histogram log_MedIncome / normal;
      title 'Distribution of log_houseDensity';
      probplot log_MedIncome / normal (mu=est sigma=est color=blue
w=1);
      title 'Normal Probability Plot';
run;

proc univariate data=trans_rates;
      format state statef.;
      var log_houseDensity;
      id geo_id2;
      histogram log_houseDensity / normal;
      title 'Distribution of log_houseDensity';
      probplot log_houseDensity / normal (mu=est sigma=est color=blue
w=1);
      title 'Normal Probability Plot';
run;
```

```
proc univariate data=trans_rates;
      format state statef.;
      var log_popDensity;
      id geo_id2;
      histogram log_popDensity / normal;
      title 'Distribution of log_popDensity';
      probplot log_popDensity / normal (mu=est sigma=est color=blue
w=1);
      title 'Normal Probability Plot';
run;

proc univariate data=trans_rates;
      format state statef.;
      var log_Tot_Pop;
      id geo_id2;
      histogram log_Tot_Pop / normal;
      title 'Distribution of Log total pop ratio';
      probplot log_Tot_Pop / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability Plot';
run;


proc univariate data=trans_rates;
      format state statef.;
      var Log_BlackRR;
      id geo_id2;
      histogram Log_BlackRR / normal;
      title 'Distribution of Black to White ratio';
      probplot Log_BlackRR / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability Plot';
run;

proc univariate data=trans_rates;
      format state statef.;
      var Log_HispRR;
      id geo_id2;
      histogram Log_HispRR / normal;
      title 'Distribution of Hispanic to White ratio';
      probplot Log_HispRR / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability Plot';
run;

proc univariate data=trans_rates;
      format state statef.;
      var Log_OtherRR;
      id geo_id2;
      histogram Log_OtherRR / normal;
      title 'Distribution of Other Race to White ratio';
      probplot Log_OtherRR / normal (mu=est sigma=est color=blue w=1);
      title 'Normal Probability Plot';
run;

******************************************************;
*====================================================;
**********       Exploratory Analysis       ***********;
*====================================================;
******************************************************;
```

```
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=analysis;
      format state statef.;
      var HIV log_Pop log_PopDensity log_HouseDensity MedAge
log_BlackRR log_HispRR log_OtherRR log_MSM log_MedIncome
             rate_poverty Gini rate_HSgrad log_Prison rate_NoIns
log_Costs rate_DrugUse rate_DrugDep;
      id geo_id2;
      By state;
run;


*CORRELATIONS for Continuous Variables;
*-----------------------------------------------------;

options ps=50 ls=64;
goptions reset=all gunit=pct border fontres=presentation ftext=swissb;
axis1 length=70 w=3 color=blue label=(h=3) value=(h=3);
axis2 length=70 w=3 color=blue label=(h=3) value=(h=3);

* Scatter Plot of HIV over Variable;
proc gplot data=analysis;
      plot HIV*(log_Pop log_PopDensity log_HouseDensity MedAge
log_BlackRR log_HispRR log_OtherRR log_MSM log_MedIncome
             rate_poverty Gini rate_HSgrad log_Prison rate_NoIns
log_Costs rate_DrugUse rate_DrugDep) / vaxis=axisa haxis=axis2;
      symbol1 v=dot h=2 w=4 color=red;
      title h=3 color=green 'Plot of New HIV Cases by Other Variables';
run;
QUIT;

* Linear correlations between HIV and other Variable;
*By State;
proc corr data=analysis;
      var HIV;
      with log_Pop log_PopDensity log_HouseDensity MedAge log_BlackRR
log_HispRR log_OtherRR log_MSM log_MedIncome
             rate_poverty Gini rate_HSgrad log_Prison rate_NoIns
log_Costs rate_DrugUse rate_DrugDep;
      by state;
      format state statef.;
run;

*All Together;
proc corr data=analysis;
      var HIV;
      with log_Pop log_PopDensity log_HouseDensity MedAge log_BlackRR
log_HispRR log_OtherRR log_MSM log_MedIncome
             rate_poverty Gini rate_HSgrad log_Prison rate_NoIns
log_Costs rate_DrugUse rate_DrugDep;
run;


* Covariance and Correlation Matrix;
ods select Cov PearsonCorr;
proc corr data=analysis noprob outp=OutCorr nomiss cov;
```

```
      var HIV state sex log_Pop log_PopDensity log_HouseDensity MedAge
log_BlackRR log_HispRR log_OtherRR log_MSM log_MedIncome
             rate_poverty Gini rate_HSgrad log_Prison rate_NoIns
log_Costs rate_DrugUse rate_DrugDep urban;
run;


proc corr data=analysis;
      var state;
      with log_Pop log_PopDensity log_HouseDensity MedAge log_BlackRR
log_HispRR log_OtherRR log_MSM log_MedIncome
             rate_poverty Gini rate_HSgrad log_Prison rate_NoIns
log_Costs rate_DrugUse rate_DrugDep;
      by state;
      format state statef. sex sexf. urban urbanf.;
run;

* ANOVAs for Categorical Variables;
*-----------------------------------------------------;

* One Way ANOVAs;
* STATE;
options ls=75 ps=45;
proc glm data=analysis;
      class state;
      model HIV = state;
      means state / hovtest;
      output out=check r=resid p=pred;
      title 'Testing for Quality of Means of HIV';
      format state statef.;
run;
QUIT;

goptions reset=all;
proc gplot data=check;
      plot resid*pred / haxis=axis1 vaxis=axis2 vref=0;
      symbol v=star h=3pct;
      axis1 w=2 major=(w=2) minor=none offset=(10pct);
      axis2 w=2 major=(w=2) minor=none;
      title 'Plot of Residuals vs. Predicted Values for New HIV
Diagnoses';
run;
quit;

proc univariate data=check normal;
      var resid;
      histogram / normal;
      probplot / normal (mu=est sigma=est color=blue w=1);
      title;
run;

* non-normal distribution, so use WILCOXON to do KRUSKAL-WALLIS test;
proc sort data=analysis;
      by state;
proc npar1way data=analysis wilcoxon median;
      class state;
      var HIV;
```

```
      format state statef.;
run;


* SEX;
options ls=75 ps=45;
proc glm data=analysis;
      class sex;
      model HIV = sex;
      means sex / hovtest;
      output out=check r=resid p=pred;
      title 'Testing for Quality of Means of HIV';
      format sex sexf.;
run;
QUIT;

goptions reset=all;
proc gplot data=check;
      plot resid*pred / haxis=axis1 vaxis=axis2 vref=0;
      symbol v=star h=3pct;
      axis1 w=2 major=(w=2) minor=none offset=(10pct);
      axis2 w=2 major=(w=2) minor=none;
      title 'Plot of Residuals vs. Predicted Values for New HIV
Diagnoses';
run;
quit;

proc univariate data=check normal;
      var resid;
      histogram / normal;
      probplot / normal (mu=est sigma=est color=blue w=1);
      title;
run;

* non-normal distribution, so use WILCOXON to do KRUSKAL-WALLIS test;
proc sort data=analysis;
      by sex;
proc npar1way data=analysis wilcoxon median;
      class sex;
      var HIV;
      format sex sexf.;
run;

* URBANICITY;
options ls=75 ps=45;
proc glm data=analysis;
      class urban;
      model HIV = urban;
      means urban / hovtest welch;  * Welch's ANOVA bc Not normal;
      output out=check r=resid p=pred;
      title 'Testing for Quality of Means of HIV';
      format urban urbanf.;
run;
QUIT;

goptions reset=all;
proc gplot data=check;
```

```
      plot resid*pred / haxis=axis1 vaxis=axis2 vref=0;
      symbol v=star h=3pct;
      axis1 w=2 major=(w=2) minor=none offset=(10pct);
      axis2 w=2 major=(w=2) minor=none;
      title 'Plot of Residuals vs. Predicted Values for New HIV
Diagnoses';
run;
quit;

proc univariate data=check normal;
      var resid;
      histogram / normal;
      probplot / normal (mu=est sigma=est color=blue w=1);
      title;
run;



* Two Way ANOVAs;
*-----------------------------------------------------;

*STATE and SEX;
proc means data=analysis mean var std;
      class state sex;
      var HIV;
      title 'Selected Descriptive Statistics';
run;

proc gplot data=analysis;
      symbol c=blue w=2 interpol=std1mtj line=1;
      symbol2 c=green w=2 interpol=std1mtj line=2;
      symbol3 c=red w=2 interpol=std1mtj line=3;
      plot hiv*sex=state;
      title 'Illustratins the Interaction Between HIV and Sex';
run;
quit;

proc glm data=analysis;
      class state sex;
      model HIV=state sex state*sex;            *not sig;
      title 'Analyze the effects of State and Sex';
      title2 'Including Interaction';
      format state statef. sex sexf.;
run;
QUIT;


* STATE and URBAN;
proc glm data=analysis;
      class state urban;
      model HIV=state urban state*urban;
      title 'Analyze the effects of State and Gini';
      title2 'Including Interaction';
      format state statef. urban urbanf.;
run;
QUIT;
```

```
proc glm data=analysis;
      class state urban;
      model HIV=state urban state*urban;
      lsmeans state*urban / adjust=tukey pdiff=all;
      title 'Multiple Comparisons Tests for State and Urbanicity';
run;
QUIT;


* Assessing State Prev Rates;
proc genmod data=analysis descending;
      class state;
      model HIV = / link=log dist=negbin;
      by state;
      estimate 'Null Model' sex 1 -1 /exp;
run;


proc genmod data=analysis descending;
      class state;
      model HIV = sex / link=log dist=negbin;
      by state;
      estimate 'Prev Rate' sex 1 -1 /exp;
      format state statef. sex sexf. urban urbanf.;
run;

proc genmod data=analysis descending;

      model HIV = sex / link=log dist=negbin;

      estimate 'Prev Rate' sex 1 -1 /exp;
      format state statef. sex sexf. urban urbanf.;
run;

proc genmod data=analysis descending;
      class state;
      model HIV = log_pop / link=log dist=negbin;
      by state;
      estimate 'Prev Rate' log_pop 1 /exp;
      format state statef. sex sexf. urban urbanf.;
run;

proc genmod data=analysis descending;
      model HIV = log_pop / link=log dist=negbin;
      estimate 'Prev Rate' log_pop 1  /exp;
      format state statef. sex sexf. urban urbanf.;
run;

proc genmod data=analysis descending;
      class state;
      model HIV = log_popDensity / link=log dist=negbin;
      by state;
      estimate 'Prev Rate' log_popDensity 1 /exp;
      format state statef. sex sexf. urban urbanf.;
run;

proc genmod data=analysis descending;
```

```
        model HIV = log_popDensity / link=log dist=negbin;
        estimate 'Prev Rate' log_popDensity 1  /exp;
        format state statef. sex sexf. urban urbanf.;
run;

proc genmod data=analysis descending;
        class state;
        model HIV = log_HouseDensity / link=log dist=negbin;
        by state;
        estimate 'Prev Rate' log_HouseDensity 1 /exp;
        format state statef. sex sexf. urban urbanf.;
run;


proc genmod data=analysis descending;
        class state;
        model HIV = MedAge / link=log dist=negbin;
        by state;
        estimate 'Prev Rate' MedAge 1 /exp;
        format state statef. sex sexf. urban urbanf.;
run;

proc genmod data=analysis descending;
        class state;
        model HIV = log_BlackRR / link=log dist=negbin;
        by state;
        estimate 'Prev Rate' log_BlackRR 1 /exp;
        format state statef. sex sexf. urban urbanf.;
run;

proc genmod data=analysis descending;
        class state;
        model HIV = log_HispRR / link=log dist=negbin;
        by state;
run;

proc genmod data=analysis descending;
        class state;
        model HIV = log_OtherRR / link=log dist=negbin;
        by state;
run;

proc genmod data=analysis descending;
        class state;
        model HIV = log_MSM / link=log dist=negbin;
        by state;
run;

proc genmod data=analysis descending;
        class state;
        model HIV = log_MedIncome / link=log dist=negbin;
        by state;
run;

proc genmod data=analysis descending;
        class state;
        model HIV = rate_poverty / link=log dist=negbin;
```

```
      by state;
run;

proc genmod data=analysis descending;
      class state;
      model HIV = gini / link=log dist=negbin;
      by state;
run;

proc genmod data=analysis descending;
      class state;
      model HIV = rate_HSgrad / link=log dist=negbin;
      by state;
run;

proc genmod data=analysis descending;
      class state;
      model HIV = log_prison / link=log dist=negbin;
      by state;
run;

proc genmod data=analysis descending;
      class state;
      model HIV = rate_NoIns / link=log dist=negbin;
run;

proc genmod data=analysis descending;
      class state;
      model HIV = log_costs / link=log dist=negbin;
run;

proc genmod data=analysis descending;
      class state;
      model HIV = rate_DrugUse / link=log dist=negbin;
      by state;
run;

proc genmod data=analysis descending;
      class state;
      model HIV = rate_DrugDep / link=log dist=negbin;
      by state;
run;

proc genmod data=analysis descending;
      class state;
      model HIV = urban / link=log dist=negbin;
      by state;
run;



* Variables: HIV = sex log_Pop log_popdensity log_HouseDensity MedAge
log_BlackRR log_HispRR log_OtherRR log_MSM
              log_prison Gini rate_HSgrad rate_NoIns rate_DrugUse
rate_DrugDep urban log_MedIncome Rate_Poverty;
```

```
* do counties differ by state?;
proc corr data=analysis spearman;
      var state sexmf log_Pop log_popdensity log_HouseDensity MedAge
log_BlackRR log_HispRR log_OtherRR log_MSM
              log_prison Gini rate_HSgrad rate_NoIns rate_DrugUse
rate_DrugDep urban log_MedIncome Rate_Poverty;
run;

proc logistic data=analysis descending;
      model state=log_pop / expb;
run;


proc corr data=analysis;
      var HIV sexmf log_Pop log_popdensity log_HouseDensity MedAge
log_BlackRR log_HispRR log_OtherRR log_MSM
              log_prison Gini rate_HSgrad rate_NoIns rate_DrugUse
rate_DrugDep urban log_MedIncome Rate_Poverty log_costs;
      by state;
      format state statef. sex sexf. urban urbanf.;
run;


* Simple linear rate models;
* Calculate Prevalence Rate Ratios;
proc genmod data=analysis descending;
      class state (ref='MD & NC') / param=ref;
      model HIV = state / link=log dist=negbin;
      estimate 'Null Model PRR' state 1 -1 /exp;
      format state statef.;
run;

proc genmod data=analysis descending;
      class state (ref='MD & NC')/ param=ref;
      model HIV = state sex / link=log dist=negbin;
      estimate 'PRR' state 1 -1 sex 1 -1/exp;
      format state statef.;
run;

proc genmod data=analysis descending;
      class state (ref='MD & NC')/ param=ref;
      model HIV = state log_Pop / link=log dist=negbin;
      estimate 'PRR' state 1 -1/exp;
      format state statef.;
run;

proc genmod data=analysis descending;
      class state (ref='MD & NC')/ param=ref;
      model HIV = state log_PopDensity / link=log dist=negbin;
      estimate 'PRR' state 1 -1 /exp;
      format state statef.;
run;

proc genmod data=analysis descending;
      class state (ref='MD & NC')/ param=ref;
      model HIV = state log_HouseDensity / link=log dist=negbin;
      estimate 'PRR' state 1 -1 /exp;
```

```
      format state statef.;
run;

proc genmod data=analysis descending;
      class state (ref='MD & NC')/ param=ref;
      model HIV = state MedAge / link=log dist=negbin;
      estimate 'PRR' state 1 -1 /exp;
      format state statef.;
run;

proc genmod data=analysis descending;
      class state (ref='MD & NC')/ param=ref;
      model HIV = state log_BlackRR / link=log dist=negbin;
      estimate 'PRR' state 1 -1 /exp;
      format state statef.;
run;

proc genmod data=analysis descending;
      class state (ref='MD & NC')/ param=ref;
      model HIV = state log_HispRR / link=log dist=negbin;
      estimate 'PRR' state 1 -1 /exp;
      format state statef.;
run;

proc genmod data=analysis descending;
      class state (ref='MD & NC')/ param=ref;
      model HIV = state log_OtherRR / link=log dist=negbin;
      estimate 'PRR' state 1 -1 /exp;
      format state statef.;
run;

proc genmod data=analysis descending;
      class state (ref='MD & NC')/ param=ref;
      model HIV = state log_MSM / link=log dist=negbin;
      estimate 'PRR' state 1 -1 /exp;
      format state statef.;
run;

proc genmod data=analysis descending;
      class state (ref='MD & NC')/ param=ref;
      model HIV = state log_MedIncome / link=log dist=negbin;
      estimate 'PRR' state 1 -1 /exp;
      format state statef.;
run;

proc genmod data=analysis descending;
      class state (ref='MD & NC')/ param=ref;
      model HIV = state rate_poverty / link=log dist=negbin;
      estimate 'PRR' state 1 -1 /exp;
      format state statef.;
run;

proc genmod data=analysis descending;
      class state (ref='MD & NC')/ param=ref;
      model HIV = state gini / link=log dist=negbin;
      estimate 'PRR' state 1 -1 /exp;
      format state statef.;
```

```
run;

proc genmod data=analysis descending;
      class state (ref='MD & NC')/ param=ref;
      model HIV = state rate_HSgrad / link=log dist=negbin;
      estimate 'PRR' state 1 -1 /exp;
      format state statef.;
run;

proc genmod data=analysis descending;
      class state (ref='MD & NC')/ param=ref;
      model HIV = state log_prison / link=log dist=negbin;
      estimate 'PRR' state 1 -1 /exp;
      format state statef.;
run;

proc genmod data=analysis descending;
      class state (ref='MD & NC')/ param=ref;
      model HIV = state rate_NoIns / link=log dist=negbin;
      estimate 'PRR' state 1 -1 /exp;
      format state statef.;
run;

proc genmod data=analysis descending;
      class state (ref='MD & NC')/ param=ref;
      model HIV = state log_costs / link=log dist=negbin;
      estimate 'PRR' state 1 -1 /exp;
      format state statef.;
run;

proc genmod data=analysis descending;
      class state (ref='MD & NC')/ param=ref;
      model HIV = state rate_DrugUse / link=log dist=negbin;
      estimate 'PRR' state 1 -1 /exp;
      format state statef.;
run;

proc genmod data=analysis descending;
      class state (ref='MD & NC')/ param=ref;
      model HIV = state rate_DrugDep / link=log dist=negbin;
      estimate 'PRR' state 1 -1 /exp;
      format state statef.;
run;

proc genmod data=analysis descending;
      class state (ref='MD & NC') urban/ param=ref;
      model HIV = state urban / link=log dist=negbin;
      estimate 'PRR' state 1 -1 /exp;
      format state statef.;
run;

*=====================================================;
****************        Modeling      ****************;
*=====================================================;

proc genmod data=analysis;
      class state;
```

```
      model HIV = sexmf log_Pop log_popdensity log_HouseDensity MedAge
log_MedIncome log_BlackRR log_HispRR log_OtherRR log_MSM
              log_prison Gini rate_HSgrad rate_NoIns rate_poverty
rate_DrugUse rate_DrugDep / dist=negbin link=log;
      by state;
      format state statef. sex sexf. urban urbanf.;
run;

* goodness of fit p-values;

* Maryland and North Carolina;
data pvalue;
  df = 52; chisq = 70.1128;
  pvalue = 1 - probchi(chisq, df);
run;
proc print data = pvalue noobs;
title "Model fit for MD & NC";
run;

*Virgina;
data pvalue;
  df = 11; chisq = 25.5474;
  pvalue = 1 - probchi(chisq, df);
run;
proc print data = pvalue noobs;
title "Model fit for Virginia";
run;


*==============================*;
* Parsimonious model building *;
*==============================*;

proc genmod data=analysis;
      class state;
      model HIV = sexmf log_Pop log_popdensity log_HouseDensity MedAge
log_BlackRR log_HispRR
      log_OtherRR log_msm rate_DrugUse / dist=negbin link=log;
      by state;
      format state statef. sex sexf. urban urbanf.;
run;

*=====================================================;
****************      Projection     ***************;
*=====================================================;

/*
* Using Full model;
data projection_Full;
      set trans_rates;

      HIV2=exp(-2.4333 + (sex*0.3117) + (log_Pop*1.1599) +
(log_popdensity*0.3923) + (log_HouseDensity*-0.2917) + (MedAge*0.0442)
+ (log_MedIncome*-0.9215) +
            (log_BlackRR*0.5224) + (log_HispRR*0.1370) + (log_OtherRR*-
0.0728) + (log_MSM*0.0489) + (log_prison*0.0431) + (Gini*1.4188) +
(rate_HSgrad*0.0017) +
```

```
          (rate_NoIns*-0.0025) + (rate_Poverty*-0.0007) +
(rate_DrugUse*-0.0089) + (rate_DrugDep*0.0011) + (urban*-0.0107) );

     keep geo_id2 state HIV HIV2;
run;


proc means data=projection_Full sum n;
     var HIV2 HIV;
     by state;
     format state statef. sex sexf. urban urbanf.;
run;

*/

* Using Reduced model - MD_NC model;
data projection_red_MDNC;
     set trans_rates;

     HIV3=exp(-11.0425 + (sexmf*1.3539) + (log_Pop*1.1674) +
(log_popdensity*-0.1055) + (log_HouseDensity*0.1889) + (MedAge*0.0188)
+ (log_BlackRR*0.6318) +
          (log_HispRR*(-0.0235)) + (log_OtherRR*(-0.1179)) +
(log_MSM*0.0825) + (rate_DrugUse*(-0.0073)) );

     keep geo_id2 state HIV HIV3;
run;

* Using Reduced model - VA model;
data projection_Red_VA;
     set trans_rates;

     HIV4=exp(-8.4904 + (sexmf*3.7669) + (log_Pop*1.1098) +
(log_popdensity*(-2.5189)) + (log_HouseDensity*2.5763) + (MedAge*(-
0.0198)) + (log_BlackRR*0.5261) +
          (log_HispRR*0.0325) + (log_OtherRR*(-0.1887)) + (log_MSM*(-
0.0122)) + (rate_DrugUse*(-0.0325)) );

     keep geo_id2 state HIV HIV4;
run;



proc sort data=projection_red_MDNC;
     by geo_id2;
proc sort data=projection_red_VA;
     by geo_id2;
data FinalProjection;
     merge projection_red_MDNC (in=a) projection_red_VA (in=b);
     by geo_id2;

     HIV03=round(HIV3,1);
     HIV04=round(HIV4,1);
     if a and b;

     keep geo_id2 state HIV HIV03 HIV04;
run;
```

```
%macro ForMapping(dataset);
      %let x = %str(:) ;
      %let MapFile=%sysfunc(cat(H, &x,
\Classes\Thesis\Maps\ProjectedHIV0411Final.csv));
      proc export data=&dataset
            outfile= "&MapFile"
      dbms=csv replace;
      putnames=yes;
      run;
%mend;

%ForMapping(FinalProjection);

proc means data=FinalProjection sum n mean std q1 median q3;
      var HIV;
      by state;
      where HIV gt 19;
      format state statef. sex sexf. urban urbanf.;
run;

* Weighted by MD/NC model;
proc means data=FinalProjection sum n mean std q1 median q3;
      var HIV03;
      by state;
      where HIV03 gt 19;
      format state statef. sex sexf. urban urbanf.;
run;
proc means data=FinalProjection sum n mean std q1 median q3;
      var HIV03;
      by state;
      where HIV03 le 19;
      format state statef. sex sexf. urban urbanf.;
run;

*Weighted by VA model;
proc means data=FinalProjection sum n mean std q1 median q3;
      var HIV04;
      by state;
      where HIV04 gt 19;
      format state statef. sex sexf. urban urbanf.;
run;
proc means data=FinalProjection sum n mean std q1 median q3;
      var HIV04;
      by state;
      where HIV04 le 19;
      format state statef. sex sexf. urban urbanf.;
run;


* Import total HIV counts for each state;
OPTIONS nofmterr;
data stateHIV;
      set a.stateHIV;

      if state='24' then st='MD';
```

```
      if state='37' then st='NC';
      if state='51' then st='VA';

      where state='24' or state='37' or state='51';

      newvar=input(statecase,comma6.);
      drop statecase;
      stateHIV=round(newvar,1);

      keep ST stateHIV;
run;

proc means data=statehiv sum n;
      where st='VA';
      var stateHIV;
run;

proc means data=statehiv sum n;
      where st='MD' or st='NC';
      var stateHIV;
run;

PROC IMPORT OUT= work.state_oe
            DATAFILE=
"H:\Classes\Thesis\Data\ObsExp_StateHIVCounts.csv"
            DBMS=CSV REPLACE;
      GETNAMES=YES;
      DATAROW=2;
      guessingrows=5000;
RUN;

proc print data=state_oe;
run;

proc print data=projection_MDNC;
run;

proc print data=projection_VA;
run;

data log_projection_VA;
      set projection_VA;

      if HIV=. then HIV=0;

      logHIV=log(HIV);
      logHIV2=log(HIV2);
run;

goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=log_projection_VA;
      var logHIV;
      id geo_id2;
      histogram logHIV / normal;
      title 'Distribution of Observed HIV Cases in Virginia';
      probplot logHIV / normal (mu=est sigma=est color=blue w=1);
run;
```

```
goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=log_projection_VA;
      var logHIV2;
      id geo_id2;
      histogram logHIV2 / normal;
      title 'Distribution of Expected HIV Cases in Virginia';
      probplot logHIV2 / normal (mu=est sigma=est color=blue w=1);
run;

goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=log_projection_VA;
      var HIV;
      id geo_id2;
      histogram HIV / normal;
      title 'Distribution of Observed HIV Cases in Virginia';
      probplot HIV / normal (mu=est sigma=est color=blue w=1);
run;

goptions reset=all fontres=presentation ftext=swissb htext=1.5;
proc univariate data=log_projection_VA;
      var HIV2;
      id geo_id2;
      histogram HIV2 / normal;
      title 'Distribution of Expected HIV Cases in Virginia';
      probplot HIV2 / normal (mu=est sigma=est color=blue w=1);
run;
```