HIV MORTALITY RATE

James Altman, Carson Mohr, Corey Coole,

Brandon Vermeer, and Danny Breyfogle

March 2017

INTRODUCTION

Per the Center for Disease Control (CDC): HIV stands for human immunodeficiency virus. In 2000, there were 59,807 deaths attributed to HIV. In 2014, there were 12,333 deaths attributed to HIV. We want to study if there has been a significant decrease in HIV deaths over time, as this is the sixth deadliest disease in the United States (CDC), making this an important issue. The problem that we were asked to consider is the mortality rate in people with HIV as compared to those of other diseases. Provided to research this topic was information from the Center for Disease Control (CDC) of the United States, more specifically a CMF, "The Compressed Mortality File (CMF) is a county-level national mortality and population database spanning the years 1968-2015" (CDC WONDER). Each year this file is then updated and revised so that it is consistent, and allows for the closest approximation,

> *"The mortality data on the Compressed Mortality File are based on information from all death certificates filed in the fifty states and the District of Columbia. Deaths of nonresidents (e.g. nonresident aliens, nationals living abroad, residents of Puerto Rico, Guam, the Virgin Islands, and other territories of the U.S.) and fetal deaths are excluded". (CDC WONDER).*

The data that we were most interested in though, was that from the years 1999-2015. Given this set we were then tasked with sorting and analyzing the statistics provided, from the following CMF file, with an emphasis to find the mortality rates in people of different age groups. The goal being to find data correlation between their age and how that is a factor in the death of an individual that were to contract and then die from HIV or other non-related HIV diseases. Some of the questions that we were seeking to find answers on are: Are there certain age groups that are more likely to die given they have HIV? Is there a correlation between the number of deaths

by disease in certain age groups, (meaning is there an across the board age group that is more apt to die in diseased circumstances)? If so, how much more likely is this to happen? How many deaths are there per 100,000 cases? Has there been a change in mortality rate over the time between either HIV or other non-HIV related diseases? These questions are ones that we are tasked with finding, and in doing so, below is the findings of our data.

DATA SOURCE

Produced by The National Center for Health Statistic, the data set used in the analysis came from Centers for Disease Control and Prevention (CDC). Data was imported from the CDC website which contain "mortality and population counts for all U.S. counties. Counts and rates of death can be obtained by underlying cause of death, state, county, age, race, sex, and year" (CDC WONDER). To better reflect what we want, the data was filtered on site by age group, year, and ICD Sub-Chapter. Once imported, the data was then shortened to remove unnecessary parts and filtered for the year 1999 to 2015. One thing to note that is extremely important about this data, is that it was truly only collected during the census years of 2000 and 2010. The years in between it went through what is called an intercensal process, one by which they predict and estimate numbers based on the actual census data. So, in doing this process, there is actual data that could end up being more an estimate than it is an actual prediction. Another note as well, is that the data compiled is also based upon the physician, meaning that the death certificate, was based upon the physician's belief as to how the person died. There may be some fluctuation in data, if the professional accidentally diagnosed the death wrong, or made cause of death unknown. In these cases, there had to be extreme caution with what data was then allowed to be processed. As stated by the Center for Disease Control (CDC);

> *"Cause of death on the CMF is the underlying cause-of-death, which is defined by the World Health Organization (WHO) as "the disease or injury which initiated the train of events leading directly to death, or the circumstances of the accident or violence which produced the fatal injury." Underlying cause-of-death is selected from the conditions entered by the physician on the cause of death section of the death certificate" (CDC).*

Furthermore, with this data there must be consideration into this idea that it can possibly be changed later to correct information that possibly could have been entered wrong, or if a death report came back to then be attributed to another cause. In either case, this could lead to future problems, but later it will be explained as to how they were accounted for, and then made minimalized by certain methods.

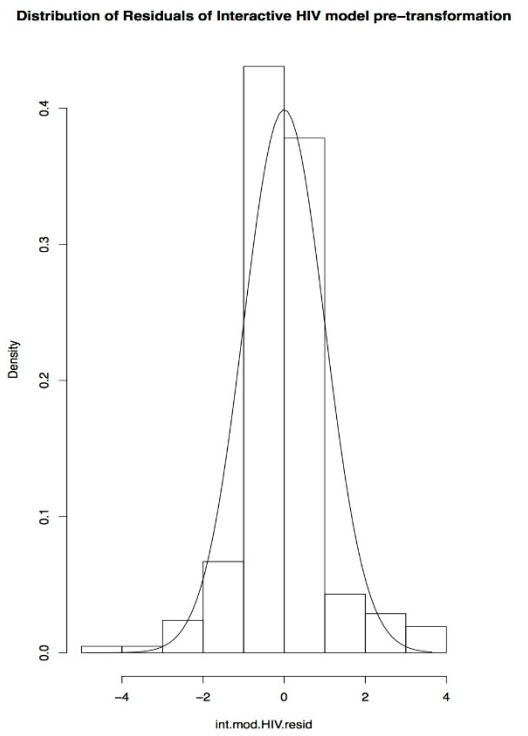
DATA

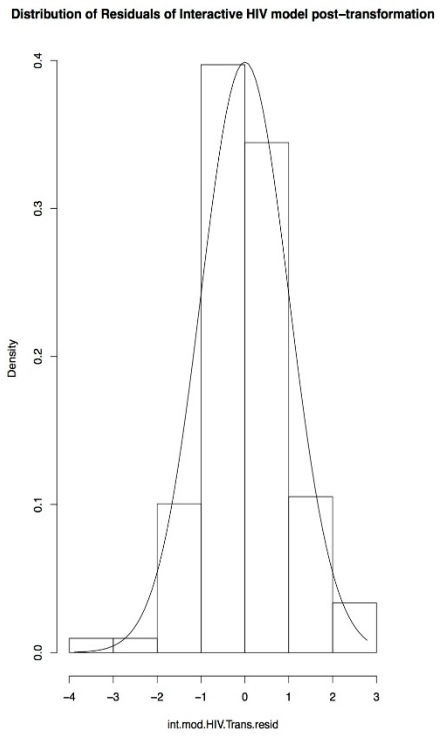
The variables: notes, crude rate, year code and age group code were removed from the data set. The actual part of the data set used contained the following variables: age groups, years, ICD, number of death, and population. Observations with values "Not Applicable" and "Not Stated" were removed to generate a consistent model. The data set contained mortality information about a plethora of diseases, not just HIV. Therefore, after the data had been properly sorted, it was necessary to introduce a "dummy variable" into the data for two cases: HIV related cases, and non-HIV related cases. To do this, a Boolean variable was created assigning a value of 1 to cases that were HIV related, and 0 to all others. The aggregate function was then used to add up all the deaths and population sizes. This aggregated data was then used to generate a variable rate by dividing the total number of deaths by the population and

multiplying by 100,000 to scale the rate to match that of the CDC and to ensure consistency. This rate will be used as the response variable for generating models.

METHODS

Given the complexity of our dataset's factors and consequent factor-levels, we chose the interactive form of regression modeling as it accounts for the interaction between the HIV mortality rate and the relevant age groups. The additive models we tested did not account for a reasonable amount of correlation between the two factors' and therefore addressed a lesser percentage of the population. Furthermore, upon utilizing a Shapiro test on the interaction model's residuals, the model's P-value was calculated at 3.934e-14. Thusly, we could not conclude the model's residual errors to be anywhere on the reasonable spectrum of normality. The method of a square-root transformation was applied to the interactive model by the creation of a new dependent variable, which was the square-root of the interactive model's output rate. The transformation model's residuals were Shapiro tested and the residuals of the transformed model were calculated at a P-value of .02008. Although the second Shapiro test did not yield a value greater than .05, the calculated value is much closer to achieving normality than the non-transformed model.

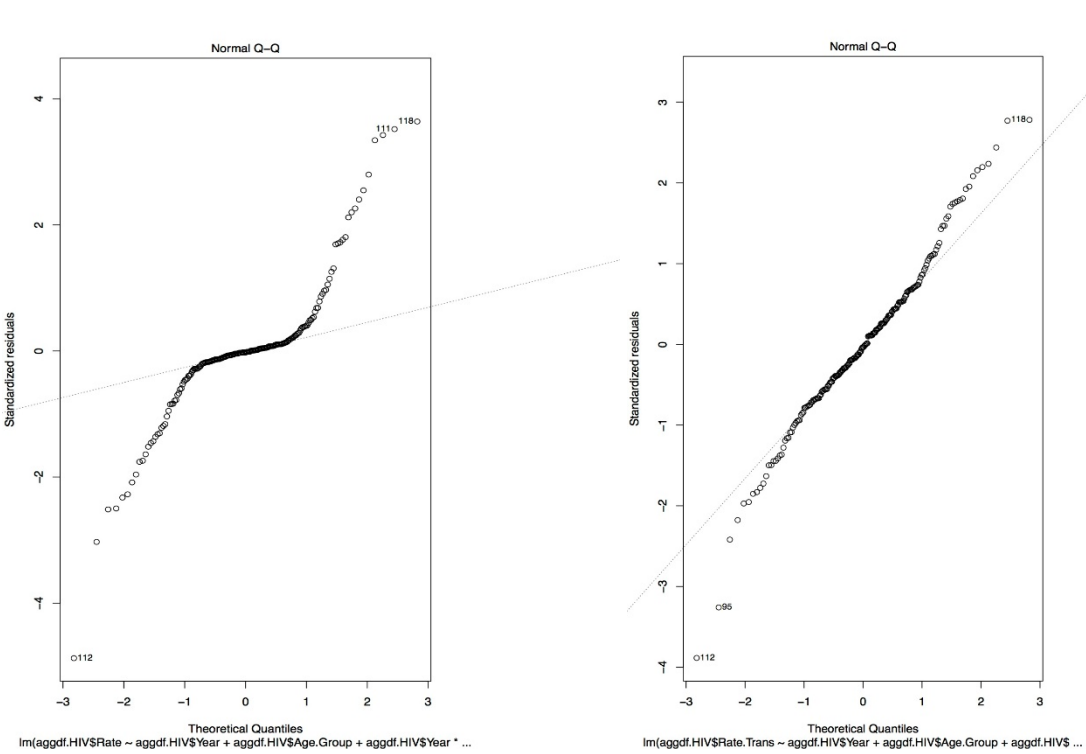




Graph 1: Residuals of Interactive Pre-Transformed (Left)

Graph 2: Residuals of Interactive Post-Transformation (Right)

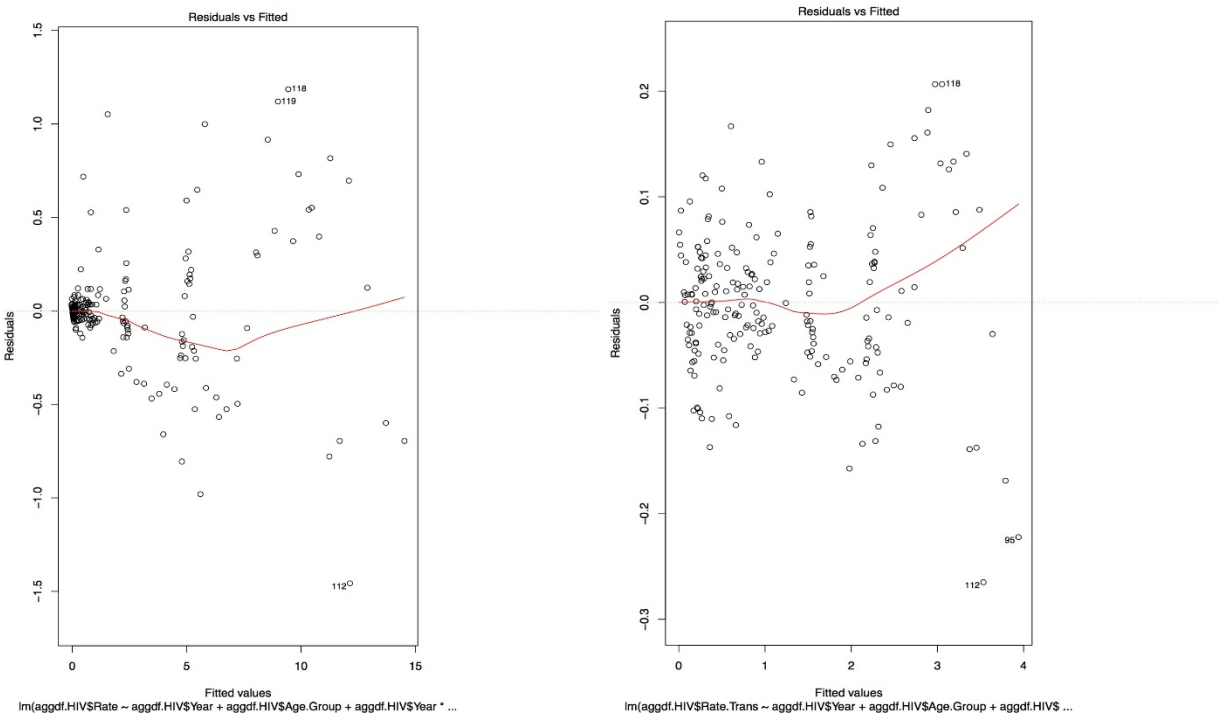
The histogram on the left depicts the distribution of the residual of the pre-transformed interactive model, while they look to follow a partial bell curve, indicative to the characteristic of a normal distribution, the histogram's curve is weighted towards the zero value. The histogram depicts the post-transformed model's residual distribution and appeared much more robust in the bell curve shape.



Graph 3: Normal QQ Plot Pre-Transformed (Left)

Graph 4: Normal QQ Plot Post-Transformation (Right)

The qq-normal graphs depict the residual of the pre-transformed (left) and the post-transformed model (right). Clearly the square-root transformation corrected much of the skewing in the tail ends of the residual vs theoretical quantities data points, in turn making for a model of better fit to our original data set.



Graph 5: Residuals vs. Fitted Pre-Transformation (Left)

Graph 6: Residuals vs. Fitted Post-Transformation (Right)

Assessing the heteroskedasticity or homoscedasticity of the pre-transformation (left) and post-transformation (right), we saw that the pre-transformed model had residuals weighted about a mean of zero for the lower fitted values of the interaction model. However, the residual vs fitted values graph of the post-transformed model indicates a much more uniform distribution of residual vs fitted value data points which indicates a more uniform variance on output values for our transformed model. The transformed model's graph also indicates a much more uniform mean as the residual vs fitted value line of fit is weighted about the zero-mean value in a much more consistent fashion the pre-transformation model. By these interpretations of our model's residual and test of validity, as well as recognizing the non-linear relationship held between the

factors of our data set, the interactive model with a square-root transformation was the best model to apply to our specific data set.

RESULTS

Of the model's twenty-five coefficients, sixteen were modeled with a high degree of significance with a P-value of each specific term less than an alpha level of .01, which implies a high degree of utility throughout the transformed model.

Our transformed interactive model is as follows;

EXPECTED RATE = SQRT((4.684e+01) + (-2.318e-02)(YEAR) + (-1.531e+01)(AGE:1-4) + (-1.7e+00)(AGE:10-14) + (-9.997e+00)(AGE:15-19) + (-1.156e+01)(AGE:20-24) + (1.424e+02)(AGE:25-34) + (2.587e+02)(AGE:35-44) + (1.165e+02)(AGE:45-54) + (-2.437e+00)(AGE:5-9) + (-2.629e+01)(AGE:55-64) + (-5.634e+01)(AGE:65-74) + (-8.641e+01)(AGE:75-84) + (-6.885e+01)(AGE:85+) + (7.544e-03)(YEAR * AGE:1-4) + (7.875e-04)(YEAR * AGE:10-14) + (4.988e-03)(YEAR * AGE:15-19) + (6.000e-03)(YEAR * AGE:20-24) + (-7.025e-02)(YEAR * AGE:25-34) + (-1.277e-01)(YEAR * AGE:35-44) + (-5.678e-02)(YEAR * AGE:45-54) + (1.125e-03)(YEAR * AGE:5-9) + (1.406e-02)(YEAR * AGE:55-64) + (2.868e-02)(YEAR * AGE:65-74) + (4.337e-02)(YEAR * AGE:75-84) + (3.444e-02)(Year * AGE:85+)

The transformed model calculated an adjusted R-squared value of .9941, as well as a correlation coefficient of .997. Looking further into the transformed model's summary output, the intercept's explanation while necessary for the model, represents a coefficient not found in real world context. The intercept of our model represents the expected rate of cases attributed to HIV at a

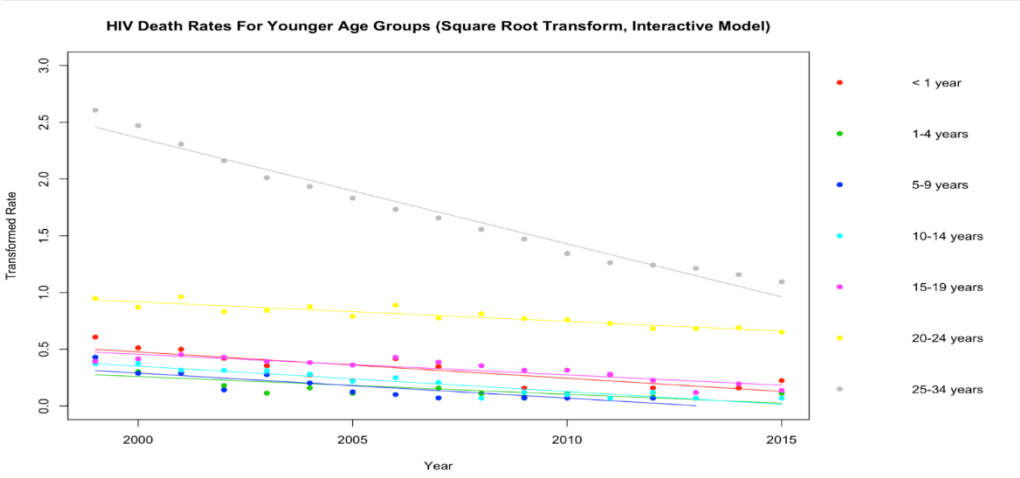
category zeros years of age and a period year zero, the rate of such a circumstance expected by our transformed model is 4.684e+01 one hundred thousand HIV related deaths per population coefficient. The significant figures with negative, or descending coefficients are also notable in our model's summary output. All age categories, 0 to 24 years old have a negative slope, as well as all age categories of 65 years old onward. The slope becomes positive when we look at the categories of 25-34 years old, 35-44 years old, and 45-54 years old. The factors with positive rates of change are calculated at a p-value less than .05, designating them as significant and are thereby intrinsic for the model's benefit. However not all the values related to a descending rate of change are deemed significant, a large majority are, citing the age categories; less than 4 years old, and 55 to 85+ years old. While these coefficients tell us a lot about what the transformed model is doing, we must look at the terms of interaction as well to get a complete perspective into the model's utility. 58% the interaction terms of Year and Age were found to be statistically significant, these terms, weighted towards the categories of age greater than 25 years. Overall, considering the high number of significant figure, a p-value much smaller in value than the alpha level, and a very high adjusted R-squared coefficient, the summary suggests a model of very high utility.

```
Call:
lm(formula = aggdf.HIV$Rate.Trans ~ aggdf.HIV$Year + aggdf.HIV$Age.Group +
    aggdf.HIV$Year * aggdf.HIV$Age.Group)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26486 -0.04150 -0.00247  0.03819  0.20656

Coefficients:
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     4.684e+01  8.197e+00   5.713 4.42e-08 ***
aggdf.HIV$Year                                 -2.318e-02  4.086e-03  -5.673 5.40e-08 ***
aggdf.HIV$Age.Group1-4 years                   -1.531e+01  1.227e+01  -1.248   0.2137
aggdf.HIV$Age.Group10-14 years                 -1.700e+00  1.158e+01  -0.147   0.8835
aggdf.HIV$Age.Group15-19 years                 -9.997e+00  1.121e+01  -0.892   0.3737
aggdf.HIV$Age.Group20-24 years                 -1.156e+01  1.121e+01  -1.031   0.3039
aggdf.HIV$Age.Group25-34 years                  1.424e+02  1.121e+01  12.701  < 2e-16 ***
aggdf.HIV$Age.Group35-44 years                  2.587e+02  1.121e+01  23.074  < 2e-16 ***
aggdf.HIV$Age.Group45-54 years                  1.165e+02  1.121e+01  10.395  < 2e-16 ***
aggdf.HIV$Age.Group5-9 years                   -2.437e+00  1.279e+01  -0.191   0.8491
aggdf.HIV$Age.Group55-64 years                 -2.629e+01  1.121e+01  -2.345   0.0201 *
aggdf.HIV$Age.Group65-74 years                 -5.634e+01  1.121e+01  -5.026 1.19e-06 ***
aggdf.HIV$Age.Group75-84 years                 -8.641e+01  1.121e+01  -7.708 7.89e-13 ***
aggdf.HIV$Age.Group85+ years                   -6.885e+01  1.121e+01  -6.141 4.97e-09 ***
aggdf.HIV$Year:aggdf.HIV$Age.Group1-4 years     7.544e-03  6.114e-03   1.234   0.2188
aggdf.HIV$Year:aggdf.HIV$Age.Group10-14 years   7.875e-04  5.774e-03   0.136   0.8917
aggdf.HIV$Year:aggdf.HIV$Age.Group15-19 years   4.988e-03  5.587e-03   0.893   0.3731
aggdf.HIV$Year:aggdf.HIV$Age.Group20-24 years   6.000e-03  5.587e-03   1.074   0.2843
aggdf.HIV$Year:aggdf.HIV$Age.Group25-34 years  -7.025e-02  5.587e-03 -12.574  < 2e-16 ***
aggdf.HIV$Year:aggdf.HIV$Age.Group35-44 years  -1.277e-01  5.587e-03 -22.854  < 2e-16 ***
aggdf.HIV$Year:aggdf.HIV$Age.Group45-54 years  -5.678e-02  5.587e-03 -10.163  < 2e-16 ***
aggdf.HIV$Year:aggdf.HIV$Age.Group5-9 years     1.125e-03  6.376e-03   0.176   0.8601
aggdf.HIV$Year:aggdf.HIV$Age.Group55-64 years   1.406e-02  5.587e-03   2.516   0.0127 *
aggdf.HIV$Year:aggdf.HIV$Age.Group65-74 years   2.868e-02  5.587e-03   5.133 7.25e-07 ***
aggdf.HIV$Year:aggdf.HIV$Age.Group75-84 years   4.337e-02  5.587e-03   7.762 5.71e-13 ***
aggdf.HIV$Year:aggdf.HIV$Age.Group85+ years     3.444e-02  5.587e-03   6.165 4.39e-09 ***
---
Signif. codes:  0 ë***í 0.001 ë**í 0.01 ë*í 0.05 ë.í 0.1 ë í 1

Residual standard error: 0.07697 on 183 degrees of freedom
Multiple R-squared:  0.9948,    Adjusted R-squared:  0.9941
F-statistic:  1397 on 25 and 183 DF,  p-value: < 2.2e-16
```

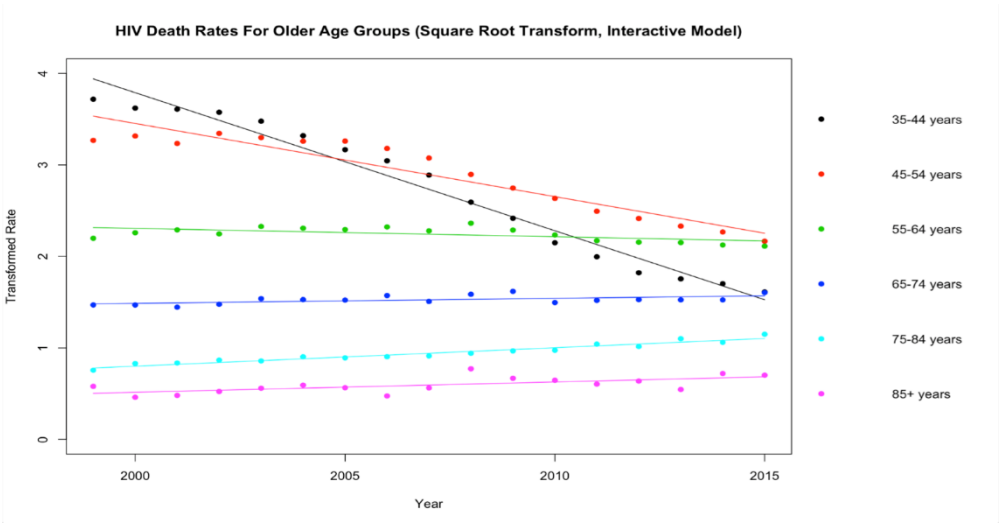
Figure 1: Summary of our Model

In graphical representation of out transformed interaction model, the first plot of data below (Expected Rate Plot 1) showcases the age categories of 0 to 34 years of age. A very clear descending trend in expected HIV related deaths can be noted from the earliest periods of data collection to most recent in the 20 to 24 and 25 to 34 year categories. Age categories of age 19 and less are also subject to a declining expected rate in our model's calculation. Looking further, into the second plot of data below (Expected Rates Plot 2), we noticed the most distinct change in expected rate related to the 35-44 years of age category. The plot also denotes a large decreasing trend with respect to the relevant data in the categories of the 45 to 55 and 55 to 64 and age groups. However, interestingly we see a definite upward trend in expected rate amongst

the categories of 65 to 74, 75-84, and 85+. Affirming the model's fit to the relevant data point in each category of age and the general absence of data point outliers, the model's plot seems to be speak strongly towards a confident inference concerning the trends of data with respect to each expected mortality rate of HIV of each age interval.



Graph 7: HIV Mortality Rate for Age Groups 1-34



Graph 8: HIV Mortality Rate for Age Groups 35-85+

DISCUSSION

Overall our model had the ability to explain 99.41% of HIV related cases, making it a very accurate model when it came to predicting the mortality rate of HIV virus. One of the most interesting findings in our study, was that there was both an overall decrease in the death rate of HIV and Non-HIV related diseases and a significant age in which more people were likely to die than at any other age group. From the period 1999-2015, there was a dramatic decrease in mortality rate, specifically in the age groups (25-34, 35-44, and 45-54 years of age), these age groups were also the ones in which the mortality rate was most significant. One major reason as to why this is, is because humans are most sexually active during those times, making it more likely that they are to contract HIV. (CDC). Almost all cases of HIV are from the sharing of bodily fluids, making it extremely contractible if precautions are not taken during sexual intercourse. Another important aspect to understand is that HIV when combined with other diseases/stressors make it dangerous, thus affecting many of the older age groups.

One other part about our model is that it is also able to be made better. If there could be a heavy trend towards eliminating HIV in older age groups, that makes for smaller amounts in younger age groups (decreasing variation among the age groups), thus making it easier to predict trends among the remaining older age groups. The CDC made a note of how many of the younger aged groups (1-14), had contradicted this disease from parents, making for higher mortality rates overall. Holding true that the more people that contract the disease, the more people that can die from it.

Lastly, the World Health Organization, United Nations, and Center for Disease Control, have called for a major fight against the HIV epidemic, hoping to have eliminated the disease by

the year 2033. We calculated that although it will not be close to that time frame, there could very well possibly be an elimination of the disease soon. Assuming that, there are programs that educate people on the dangers of the virus, and how to prevent the spreading of it. If it kills the carriers (in this case meaning the virus dies after the last of the infected die), there is a way to prevent people from contracting it in the first place. There have been large pushes in this direction already, as we can see by the trends of the mortality rate in all the age groups, but there is still a way to go before the zero point is reached.

REFERENCES

(n.d.).CDC WONDER. Compressed Mortality File 1968-2014. Retrieved March 3, 2017 from http://wonder.cdc.gov/wonder/help/cmf.html#

(n.d.). UNAIDS. Retrieved March 14, 2017, from http://unaids.org