BST 210 Project Report

Amine Abdeljaoued, Rindala Fayyad, Marius Weidmann

December 2021

1 Abstract

OBJECTIVES: We are analyzing data from the Women's Interagency HIV Study (WIHS) to assess which factors influence a woman's risk of having HIV, whether HIV status and marital status affect a woman's household income, what factors predict a woman's CD4 count, and what factors predict a woman's household size. WIHS is a prospective cohort study of women living in the US that are either at risk of or already have HIV. METHODS: A logistic regression model was used to predict HIV status, a multinomial regression model to predict a woman's salary, a linear model to predict CD4 counts, and a Poisson model to predict household size. SOME RESULTS: Having sex with an HIV positive male almost triples the odds of a woman having HIV (OR = 2.74, 95% CI = [1.84, 4.09]), and women who never used sex toys have higher odds of having HIV than women who use sex toys (OR = 1.88, 95% CI = [0.67, 5.26]). The odds of being in a high income category are 1.62 (95% CI = [0.99, 2.80]) times higher for women who are HIV positive vs HIV negative. Having HIV significantly decreases the number of CD4 positive cells per cubic millimeter of blood in the first visit (Slope = 12.20, 95% CI = [11.15, 13.26]). We found no statistically significant association between HIV status and the number of people a woman lives with (p-value = 0.46).

2 Introduction

Infectious diseases have always been a major public health issue. Some infectious diseases that were once responsible for the death of many, are no longer a problem today thanks to the constant medical advancement and creation of vaccines. One can think of polio, smallpox, measles, yellow fever and so many more. However, there are some infectious diseases for which vaccines don't work and for which it is almost impossible to find a cure today. One of the viruses that has been the reason behind millions of deaths in the world is the Human Immunodeficiency Virus (HIV). In fact, so far, around 80 million people have been infected with HIV and 36 million have died globally. In the United States, HIV has killed more than 700,000 people, among which approximately 50% are women (ODPHP, 2021).

Today, 1.2 million Americans have HIV and although infections have decreased in the past decade, people are still catching HIV every day, and some do not even know it. Early Antiretroviral Therapy (ART) has been proven to be effective in helping reduce HIV transmission, especially among sex and drug partners. However, not everyone who catches HIV can afford ART, and some do not care enough to adhere to therapy. It is important to raise awareness concerning this matter, and many organizations have done a great job in educating people about HIV transmission. But even though there are ways to reduce these transmissions, there is no definite cure or vaccine for it yet. Today, people with HIV are living longer lives, which makes the prevalence of people living with HIV higher. HIV has been an important public health and economic problem for decades, and we believe that it is crucial to continue doing research on the subject, even though a lot has already been done (Fonkwo, 2008).

It is also important to look at some statistics. For instance, in 2015, 81% of HIV infections occurred in men among which the majority were gay or bisexual. Also, 45% of HIV infections were among African Americans. This explains why the majority of the past HIV-

focused research was done around men and racial minorities, and very few focused on HIV among women, especially women in the United States (ODPHP, 2021).

Acquired Immune Deficiency Syndrome (AIDS) reached its peak in the United States in 1995. The number of people with AIDS had been strictly increasing during the years leading up to 1995, which explains why the Women's Interagency HIV Study (WIHS) started in the period. The study began in 1993 because of the worry of what impact HIV can have on women. Today, the WIHS dataset is made publicly available to anyone who is interested in public health research focused on HIV. We plan to use the data provided in order to answer the following questions:

1. Can we predict HIV status among women in the United States?
2. Do HIV status and marital status affect a woman's salary in the United States?
3. Can we predict the CD4 count in the blood samples of the patients at the first visit?
4. Can we predict the number of people that a woman in our study is living with?

3 Review of Literature and Domain Expertise

We did a literature review of similar problems, data and questions. A few modelling studies have been made on the WIHS cohort. Most of them talk about very specific technical subjects that we don't plan to address. There aren't a lot of studies on the WIHS cohort approaching the subject the way we do, by asking general questions including factors and variables that aren't too technical and that can be understood by the general public. One of the studies talks about neurocognitive factors of HIV based on many factors, which is not exactly what we are trying to study. However, they are still doing HIV based predictions, so their predictors are relevant to us. In fact, we included the two main predictors they used: ethnic group and educational experience. They also provided some useful data processing such as: "Total years of education is the sum of years completed in elementary, high school,

and post-high school phases of education, with a minimum of 0 and a maximum of 20", which we also used in our research (Manly et al, 2011).

Another study revealed factors linked to HIV infection (AJPH, 2000). The study supports "the hypothesis of a continuum of risk, with early childhood abuse leading to later domestic violence, which may increase the risk of behaviors leading to HIV infection". We tried to incorporate the information on violence (domestic violence, childhood abuse etc...) and number of male sexual partners, just like they did. We believed it would be interesting to combine these factors with other sociodemographic ones, and then see how these factors would affect a woman's chances of getting HIV. These two articles give us already some confidence in our choice of specific predictors, as they have been used in previous literature. More importantly, they show us that there is a statistically significant relationship between HIV incidence and some of the reported measures in the WIHS study.

We also spoke to Dr. Caitlin Dugdale, infectious disease physician with expertise in HIV research. She advised us to include the following predictors in our models, as they are important variables and confounders when studying HIV status: ethnicity, drug injection, sex with HIV positive male, household income, number of sex partners, and history of incarceration. We also checked with her if our work was coherent and took her advice on how to improve our models.

4 Research and Analysis Methods

We used the statistical software R to conduct our analyses. Before building our models, we constructed a correlation matrix with all of our predictors to check for multicollinearity. We chose to include only the most relevant out of the highly correlated variables in our model. We also did some exploratory analysis to look at how our variables are distributed

and to study pairwise relationships between our variables of interest.

In our first attempt to answer our first question, we built a logistic regression model with HIV status as our outcome variable and all the variables that the domain experts told us to include as predictors. We then attempted to build a new model using backwards selection. We used the step function in R and set the direction to "backward". We also built a model with forward selection, using the step function in R and setting the direction to "forward". For each model we conducted a Hosmer and Lemeshow Goodness of Fit test to assess the goodness of fit of the model, plotted the ROC curve and calculated the AUC to check if the model discriminates well between cases and non-cases.

We then conducted several Chi-square tests to check whether the number of sex partners or household income are effect measure modifiers of the effect of the other covariates on HIV status. After agreeing on the best model, we checked if there were any outliers and points of high influence, and ran the model again on data that excludes these points.

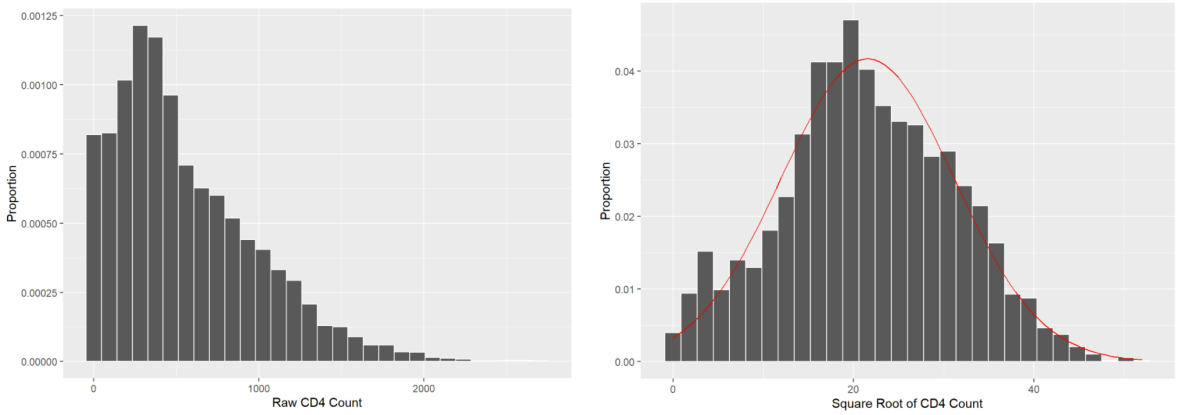
We then looked at how our model performs if we assign "1" (i.e., HIV status = 1) to the fitted values greater than 0.5, and 0 otherwise. We compared the predicted results to the actual results of our dataset by constructing a confusion matrix and looking at the accuracy, specificity and sensitivity. We repeated the process again by setting the threshold equal to 0.74 instead of 0.5.

After that, we focused our study on how HIV status and marital status affect a woman's salary in the United States. Our outcome variable of interest here is "household_income" (Check Table 1 for category descriptions). As our income variable is categorical but with ordered categories (ranges of income), we initially thought about fitting ordinal logistic models. However, the proportional odds assumption does not hold when using HIV status as

our predictor. This is due to the fact that income categories define social categories that are different across a wide range of covariates. As HIV status is our main predictor that we will be carrying throughout this analysis, we won't be fitting ordinal models and instead restrict ourselves to multinomial and simple logistic regression models. For the latter, we merged our income categories into 2: 0 for monthly household income less than or equal to 2000 and 1 for incomes higher than 2000.

We started by fitting multinomial logistic and simple logistic regressions with HIV status and marital status alone. We then studied how marital status can influence the effect of HIV status on household income both in terms of confounding and effect modification. For the former, we used the structural and statistical definitions of confounding and for the latter, we studied the significance of the interaction term and did a chi-square test with models with and without that interaction term. Our final model then included HIV status, marital status and their interaction term.

Our data set includes data on the CD4 cell count per mm^3 of women in the study. The CD4 count was measured multiple times for some of the participants at different time points. To achieve independence of the data points we restricted ourselves to the CD4 count value measured at the first visit where the CD4 count was measured. CD4 count is a continuous variable, so we used a linear model to predict this outcome variable of interest. However, the CD4 cell count data was right skewed, so we performed a square root transformation on the data (See Figure 1). After making sure the LINE assumptions hold, we built the model by starting with a model that includes all our covariates and using the step function in R to perform backward and forward selections. After picking the most appropriate model, we checked for confounders and effect modifiers, we performed a residual analysis to check for goodness of fit, and searched for significant outliers.



(a) raw CD4 count data

(b) transformed CD4 count data

Figure 1: We can see that the raw CD4 count data is heavily right skewed. After taking the square root the data looks a lot more normally distributed. This is also clearly visible when comparing the histogram to the overlaid normal distribution in red.

Finally, we created a model to predict the number of people a woman is living with. As this is a count data, and some exploratory analysis showed that the data might be Poisson distributed, we used a Poisson regression model. We again started with a full model including all covariates and used the step function in R to perform backward selection. After selecting the final model we calculated the dispersion parameter and performed an overdispersion test. We also performed a Hosmer and Lemeshow Goodness of Fit test to assess the goodness of fit of the model, and checked if some variables are confounders or effect modifiers of the effect of some covariates on the outcome variable.

5 Findings and Analysis

5.1 Predicting HIV status

To answer the first question about what variables are the "best" in predicting HIV status, we chose to stick to the model from the backward selection process, but we added drug injection as a confounder. Our model selection choice was based on multiple factors such as AIC, Hosmer and Lemeshow GOF test (p-value = 0.41), and AUC (0.73). In addition, the

Chi-square tests provided evidence that the number of sex partners and household income are not effect measure modifiers of the effect of the other covariates on HIV status (p-value > 0.05).

Our model predicting HIV status is:

$$\begin{aligned} \hat{HIV Status} = \hat{\beta}_0 + \hat{\beta}_1 sex\_partners + \hat{\beta}_2 drug + \hat{\beta}_3 ever\_marijuana + \hat{\beta}_4 ever\_sex\_toys + \\ \hat{\beta}_5 household\_income + \hat{\beta}_6 ever\_sex\_with\_hiv\_male + \hat{\beta}_7 condom\_use \end{aligned} \tag{1}$$

Where the coefficients are found in the output below and each predictor category is described in Table1.

```
Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                        2.31079    0.49192   4.698 2.63e-06 ***
sex_partners                       0.17382    0.08907   1.951 0.051005 .
as.factor(drug)2                  -0.17312    0.17461  -0.991 0.321476
as.factor(ever_marijuana)2         0.44495    0.22150   2.009 0.044561 *
as.factor(ever_sex_toys)1         -1.25720    0.42375  -2.967 0.003008 **
as.factor(ever_sex_toys)2         -0.62716    0.36629  -1.712 0.086860 .
household_income                   0.12461    0.04290   2.904 0.003680 **
as.factor(ever_sex_with_hiv_male)1 -0.26848    0.40613  -0.661 0.508557
as.factor(ever_sex_with_hiv_male)2 -1.27604    0.37595  -3.394 0.000688 ***
as.factor(condom_use)1             0.02673    0.23162   0.115 0.908136
as.factor(condom_use)2            -0.57438    0.24815  -2.315 0.020635 *
as.factor(condom_use)3            -1.53385    0.23523  -6.521 7.00e-11 ***
```

We ran our model again on data that excludes the outliers and points with high influence, and in terms of goodness of fit, the model performed significantly better (AIC = 413.22 vs 973.71).

Our final model led to the following results:

The odds of having HIV among those who have never injected drugs is 0.84 (95% CI: [0.60, 1.18]) times the odds of having HIV among those who have injected drugs before, on average, keeping all other covariates constant. The odds of having HIV among those who never used sex toys is 1.88 (95% CI: [0.67, 5.26]) times the odds of having HIV among those who have used sex toys, on average, keeping all other covariates constant. The reasoning behind is

that women who are bisexual or lesbian are more likely to use sex toys. And since having sex with males increases a woman's chances of getting HIV, then those who don't use sex toys seem to have higher chances of being HIV positive. Finally, the odds of having HIV among those who ever had sex with an HIV positive male is 2.74 (95% CI: [1.84, 4.09]) times the odds of having HIV among those who never had sex with an HIV positive male, on average, keeping all other covariates constant.

5.2 Predicting Salary

To answer the question about whether HIV status and marital status affect a woman's salary in the United States, we built a multinomial model and a logistic model including HIV status, marital status and their interaction term as covariates. When studying the association between household income and HIV status using a multinomial model, we saw that most of the relative risk ratios comparing levels of household income to the lowest one are higher than 1. This means that the risk ratio of being in high income categories vs the lowest one is higher for HIV positive women than for HIV negative women. This multinomial model's relative risk ratios comparing levels of household income vs the lowest one are counterintuitive, as one would guess that women with HIV have lower income than women without HIV. Also, the relative risk ratios vary a lot across different categories and have very wide confidence intervals. We try to resolve this issue by using a logistic regression model with only 2 categories: low income and high income. The results are coherent with the previous model and with our EDA where we noticed that there seemed to be a little bit of a higher household income among those with HIV than those without HIV in our dataset as depicted in Figure 2. The logistic model tells us that the odds of being in a high income category are 1.62 (95% CI : [0.99, 2.80]) times higher for women who are HIV positive vs HIV negative, on average, holding marital status constant.

Concerning the association between marital status and household income, we saw that the

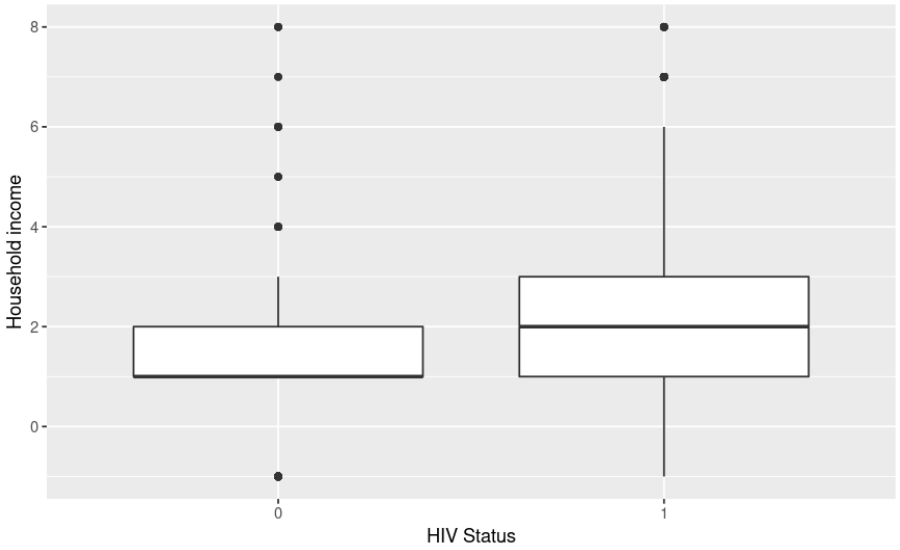


Figure 2: Distribution of household_income according to HIV status

relative risk ratio of being in high income categories vs low income ones are almost always less than 1 when comparing women who are not married, widowed, divorced or separated vs legally married women. This makes sense as the household income increase when it takes into account two or more people. The risk ratio of being in household income category 8 (> \$6250) vs category 1 (< \$500) for separated women is 0.12 (95% CI: [0.015, 0.99]) times that same risk ratio for married women, on average, holding HIV status constant. This means that separated women are less likely to be in this high income category. When looking at income category 2 (\$501-1000) which is closer to category 1, the risk ratios are not very different: the risk ratio of being in household income category 2 vs category 1 for never married women (category 6) is 0.83 (95% CI: [0.52, 1.31]) times that same risk ratio for married women, on average, holding HIV status constant.

As for the logistic regression model with "high income" and "low income" categories, most of our odds ratios' confidence intervals have their upper bound lower than one. This aligns with our previous findings where the odds of having a high income are higher for married women than women in other marital status categories. Thus as we could have guessed, the odds

of having a high household income depends significantly on the woman's marital status. In terms of confounding and effect measure modification, marital status cannot confound the effect of HIV status on household income because marital status can be a downstream consequence of HIV status or household income. For instance, if a person is HIV positive, then this might affect their chances of getting married. So, marital status does not meet the classical definition of a confounder. It is however an effect modifier of the effects of HIV status on household income (p-value = 0.0088).

5.3 Predicting CD4 Counts

Our model predicting CD4 counts is:

$$\begin{aligned}\hat{\sqrt{CD4N}} =& \hat{\beta}_0 + \hat{\beta}_1\, marital\_status + \hat{\beta}_2\, household\_income + \hat{\beta}_3\, sex\_partners + \\ & \hat{\beta}_4\, shared\_needles + \hat{\beta}_5\, ever\_jail + \hat{\beta}_6\, ever\_sex\_with\_hiv\_male + \hat{\beta}_7\, hivstat \quad (2)\end{aligned}$$

The estimates for the coefficients can be found in the output below.

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     30.75746    1.09656  28.049  < 2e-16 ***
as.factor(marital_status)2       0.14361    0.63717   0.225  0.82172
as.factor(marital_status)3      -1.92371    0.77146  -2.494  0.01278 *
as.factor(marital_status)4      -1.44645    0.71250  -2.030  0.04257 *
as.factor(marital_status)5       0.54656    0.70762   0.772  0.44003
as.factor(marital_status)6      -0.22992    0.54311  -0.423  0.67212
as.factor(marital_status)7      -3.47014    1.89204  -1.834  0.06689 .
household_income                -0.28237    0.09433  -2.993  0.00281 **
sex_partners                     0.44902    0.19836   2.264  0.02377 *
as.factor(shared_needles)1       3.21174    1.10449   2.908  0.00371 **
as.factor(shared_needles)2       1.04335    0.66487   1.569  0.11685
as.factor(ever_jail)1            1.91701    1.00261   1.912  0.05611 .
as.factor(ever_jail)2            0.96537    0.42346   2.280  0.02280 *
```

as.factor(ever_sex_with_hiv_male)1	−0.75783	0.72395	−1.047	0.29540	
as.factor(ever_sex_with_hiv_male)2	0.23150	0.69538	0.333	0.73925	
hivstat	−12.20432	0.53319	−22.889	< 2e−16	***

This led to the following results:

A one category increase in sex partners is associated with an estimated 0.45 (95% CI: [0.060, 0.84]) decrease in the square root of the number of CD4 cells per cubic millimeter, on average, holding all other variables constant. We also see that being divorced is associated with an estimated 1.45 (95% CI: [0.05, 2.85]) decrease in the square root of the number of CD4 cells per cubic millimeter compared to being married, on average, holding all other variables constant.

In addition, having sex with an HIV positive male is associated with an estimated 0.98 (95% CI: [0.16, 1.80]) decrease in the square root of the number of CD4 cells per cubic millimeter compared to never having sex with an HIV positive male, on average, holding all other variables constant. Sharing needles is associated with an estimated 2.17 (95% CI : [-0.24, 4.58]) decrease in the square root of the number of CD4 cells per cubic millimeter compared to never having shared needles, on average, holding all other variables constant.

Finally, the variable with the biggest effect on CD4 counts is HIV status. Being HIV positive is associated with a 12.20 (95% CI: [11.15, 13.26]) decrease in the square root of the number of CD4 cells per cubic millimeter compared to being HIV negative, on average, holding all other variables constant. The large effect HIV status has on the CD4 count can also be seen in Figure 3 and in Figure 4. In this plot, we can clearly distinguish the group of HIV positive participants from the group of HIV negative participants.

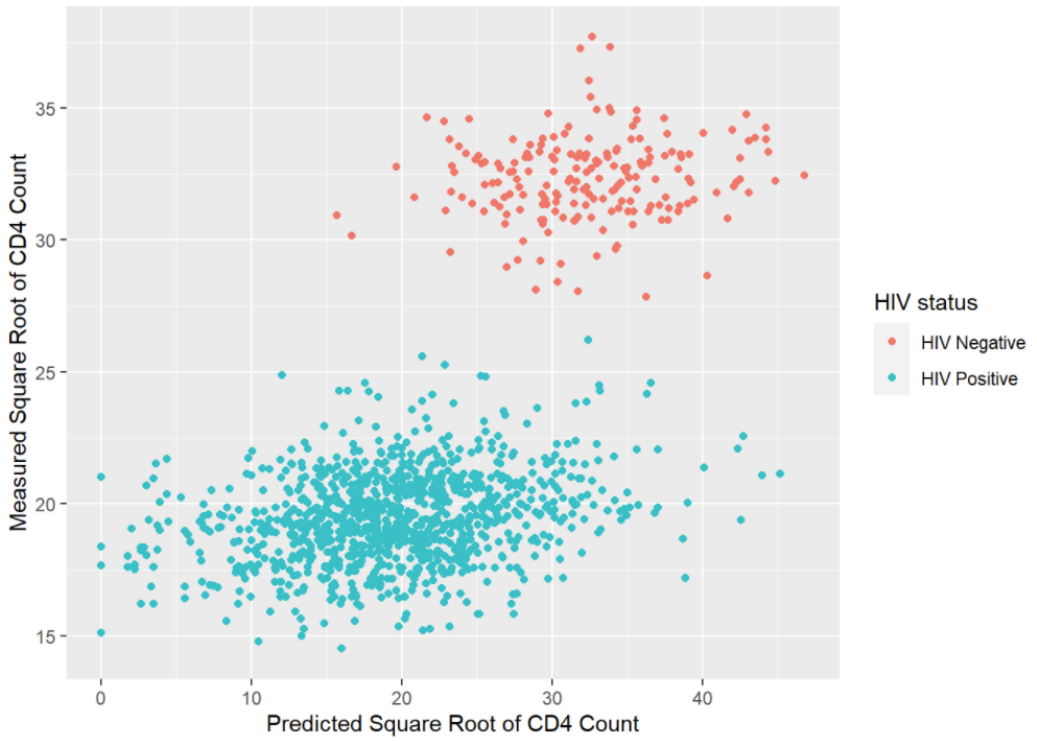


Figure 3: Predicted vs measured square root of the CD4 counts. HIV positive participants in red and HIV negative participants in blue. We can clearly distinguish the two groups and see the large effect HIV status has on CD4 count

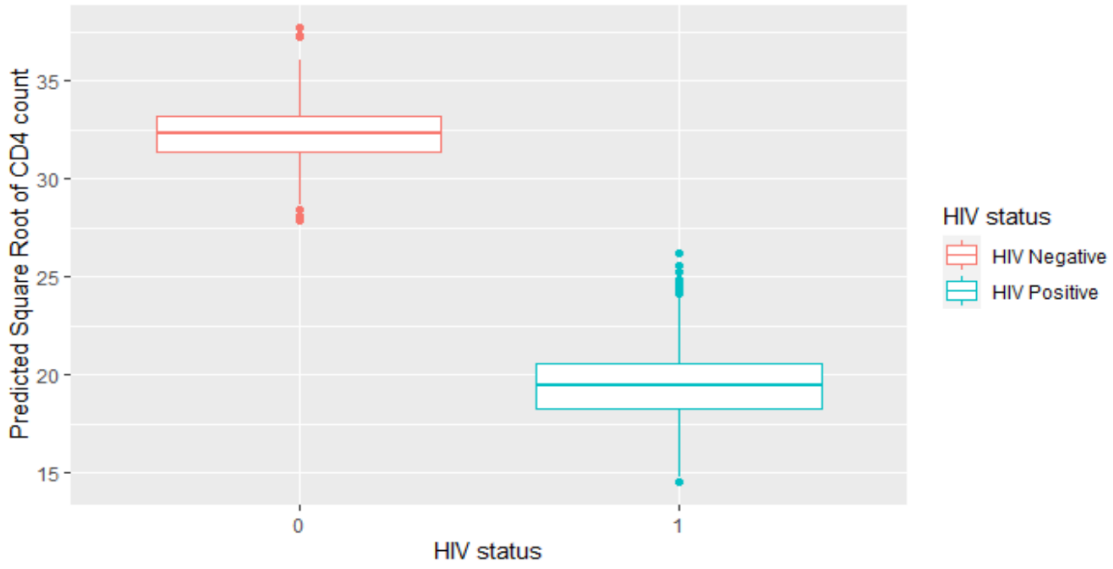


Figure 4: Predicted Square Root of CD4 count vs HIV status. Our model predicts high CD4 counts for HIV negative people and low CD4 counts for HIV positive people

5.4 Predicting Household Size

Our model predicting the number of people a women is living with has the following model statement:

$$\begin{aligned} \hat{log(living\_with)} =& \hat{\beta}_0 + \hat{\beta}_1\, ethnicity + \hat{\beta}_2\, educ\_level + \hat{\beta}_3\, marital\_status \\ &+ \hat{\beta}_4\, drug + \hat{\beta}_5\, shared\_needles + \hat{\beta}_6\, sexual\_orientation \\ &+ \hat{\beta}_7\, age + \hat{\beta}_8\, household\_income + \hat{\beta}_9\, sex\_partners \end{aligned} \tag{3}$$

The estimates for the coefficients can be found in the output below.

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     2.526865   0.605858   4.171 3.04e-05 ***
as.factor(ethnicity)1           0.298977   0.431438   0.693 0.488323
as.factor(ethnicity)2          -0.027514   0.435258  -0.063 0.949596
as.factor(ethnicity)3           0.119854   0.520135   0.230 0.817759
as.factor(ethnicity)4           0.010407   0.509702   0.020 0.983710
as.factor(ethnicity)5           0.433986   0.432109   1.004 0.315213
as.factor(educ_level)2         -0.364989   0.295763  -1.234 0.217181
as.factor(educ_level)3         -0.405040   0.269072  -1.505 0.132242
as.factor(educ_level)4         -0.370510   0.268891  -1.378 0.168228
as.factor(educ_level)5         -0.457149   0.270006  -1.693 0.090435 .
as.factor(educ_level)6         -0.718895   0.298468  -2.409 0.016013 *
as.factor(educ_level)7         -1.351766   0.385314  -3.508 0.000451 ***
as.factor(marital_status)2     -0.079093   0.077724  -1.018 0.308863
as.factor(marital_status)3     -0.216008   0.101888  -2.120 0.034001 *
as.factor(marital_status)4     -0.104069   0.091748  -1.134 0.256673
as.factor(marital_status)5     -0.022642   0.086424  -0.262 0.793329
as.factor(marital_status)6     -0.288733   0.069695  -4.143 3.43e-05 ***
as.factor(marital_status)7     -0.109549   0.257948  -0.425 0.671058
as.factor(drug)2               -0.090200   0.056615  -1.593 0.111107
as.factor(shared_needles)1      0.014325   0.163749   0.087 0.930290
```

as.factor(shared_needles)2	−0.215127	0.103619	−2.076	0.037880	*
as.factor(sexual_orientation)1	−0.670723	0.276099	−2.429	0.015129	*
as.factor(sexual_orientation)2	−0.635798	0.291451	−2.181	0.029147	*
as.factor(sexual_orientation)3	−0.861768	0.300200	−2.871	0.004096	**
as.factor(sexual_orientation)4	−0.735220	0.357103	−2.059	0.039509	*
age	−0.015597	0.003718	−4.195	2.73e−05	***
household_income	0.025777	0.014477	1.781	0.074993	.
sex_partners	−0.079258	0.026841	−2.953	0.003148	**
hivstat	−0.043365	0.058498	−0.741	0.458508	

This model has a dispersion parameter of 1.42. We performed an overdispersion test and found that we do have statistically significant evidence of overdispersion (p = 7.52 e-6). But since the overdispersion is not too extreme, we decided to keep our Poisson model for now. We performed a Hosmer and Lemeshow Goodness of Fit test (p-value = 0.9), meaning that the model is indeed a good fit for our data. We also found that ethnicity is a confounder of the effects of marital status on the number of people a woman lives with. However, we found no significant effect measure modifiers.

In our model, a 10 year increase in age is associated with an average decrease in the number of people a woman is living with by a factor of 0.86 (95% CI: [0.82, 0.89]), on average, holding all other covariates constant. A 1 unit increase in number of sex partners is associated with an average decrease in the number of people a women is living with by a factor of 0.92 (95% CI: [0.88, 0.97]), holding all other covariates constant.

When looking at levels of education obtained, attendance or completion of graduate school vs no schooling is associated with an average decrease in the number of people a woman is living with by a factor of 0.27 (95% CI: [0.14, 0.53]), holding all other covariates constant. Similarly, completing 4 years of college vs no schooling is associated with an average decrease in the number of people a woman is living with by a factor of 0.51 (95% CI: [0.31, 0.85]),

holding all other covariates constant. This shows that women with higher levels of education are more likely to live alone or with fewer people than women with lower levels of education. This is related to the fact that women with higher levels of education are more likely to have jobs and higher salaries, and can thus afford to live alone or with fewer people.

Being widowed vs married is associated with an average decrease in the number of people a woman is living with by a factor of 0.75 (95% CI: [0.67, 0.84]), holding all other covariates constant. This is intuitive, as married women are more likely to have bigger household sizes than women who are not married. However, looking at the association between HIV status and number of people a woman is living with, we find no significant association at the $\alpha = 0.05$ level. We find that being HIV positive vs negative is associated with an average decrease in the number of people a women is living with by a factor of 0.96 (95% CI: [0.85, 1.07]), holding all other covariates constant.

6 Discussion

Our work has led to several interesting findings. First, when predicting HIV status, we found several statistically significant predictors, but the most important one was the variable indicating whether a woman has had sex with an HIV positive male. In fact, if a woman has sex with an HIV positive male, this almost triples her odds of having HIV. We then studied if HIV status and marital status have an effect on household income. In our data, we found that there is a positive relationship between having HIV and household income. However, this result is specific to the women in our dataset and might not be generalizable, due to the way our dataset was constructed and selected. We only studied a small subset of the women in the WIHS, which might have caused this counterintuitive result. We also found that being married results in a higher household income that being divorced, widowed, separated or single. In addition, HIV status had a very statistically significant effect

on the number of CD4 positive cells per cubic millimeter of blood in the first visit. We found that being HIV positive significantly reduces the CD4 count in the blood cells, compared to being HIV negative. Naturally, one of the results we also found was that having sex with an HIV positive male is also associated with a decrease in the CD4 count, due to the fact that having sex with an HIV positive male is strongly associated with having HIV. Finally, because our Poisson model showed evidence of slight overdispersion, we believe that for the future it would be appropriate to try to fit a negative binomial model instead. However, the model was still a good fit for our data and led to multiple interesting associations. Among our findings, we found that women with higher levels of education are more likely to have smaller household sizes than women with no education, and we also found that in our study, HIV status is not associated with the number of people a woman lives with.

The questions we answered are just a small subset of what can be answered and found using the WIHS dataset. There are so many interesting variables that are worth studying, and so many interesting relationships that are worth investigating.

7 Limitations

The first limitation of our work lies in the dataset that we used. The WIHS interviewed and collected data from hundreds of thousands of women in the United States. However, the information collected was split into multiple datasets, which did not always contain data from all of the women interviewed. As such, many datasets contained missing information. The way we dealt with this issue was by taking the intersection of all the datasets we were interested in, and only including women whose data were present in all the different datasets to avoid missing values. Hence, our results are not generalizable to all the women in the study, because we took a very small subset of the original dataset. Another limitation is that the WIHS started 27 years ago, in 1994, which means that the results that we got

might not be applicable anymore today, as HIV treatment has changed a lot since then, and people with HIV live longer and differently today compared to 27 years ago. A third limitation is that our dataset did not have enough information about treatment, such as ART, so unfortunately we did not use treatment as a covariate in our models, although the domain experts told us that it is a very important confounder that should be included.

8 Future Scope

The WIHS started in 1994 and ended in 2014. So much has changed ever since in terms of HIV treatment and lifestyle. People with HIV live differently today compared to 10 and 20 years ago. It would be interesting to conduct the same study today but with more recent data, and check whether we would get the same results. We were also initially interested in modeling the time from the first positive HIV test to the development of cancer, AIDS or death, but our data did not allow for such analysis. We would like to build a Cox model in the future on data that includes information about time and check whether such time can be appropriately modeled. Finally, our study was done only around women in the United States, but it would be interesting to see if our models and predictions can also be applied to men in the United States, or to people of any gender in other countries, such as countries in the Middle East for example, where HIV research has been very limited.

9 Tables and figures

Table1 : Variable Description

Variable name	Variable possible values	Variable description
Hivstat	0: Negative, 1: Prevalent	What is the HIV status of the patient?
age	Continuous, -1 : NA	How old is the patient at the start of the study?
ethnicity	1: african american , 2: white , 3: asian/pacific islander , 4: native american/alaskan native , 5: other , -1: NA	What is the patient's ethnicity?
educ_level	1: no schooling , 2: grades 1-6 , 3: grades 7-11 , 5: some college , 6:completed 4 years of college , 7: attended/completed graduate school , -1: NA	What is the patient's highest education level?
drug	1: yes , 2: no , -1: NA	Has the patient ever injected drugs?
sex_partners	1: 0 or no sex partners , 2: 1 to 4 partners , 3: 5 to 10 partners , 4: 11 to 100 partners , 5: more than 100 partners , -1: NA	How many sex partners does the patient have?
marital_status	1: legally/common-law married , 2: not married but living with partner , 3: widowed , 4: divorced/annulled , 5: separated , 6: never married , 7: other , -1: NA	What is the patient's marital status?

living_with	Continuous , -1: NA	How many people is the patient living with during the study?
household _income	1: 500 or less , 2: 501-1000 , 3: 1001-1500 , 4: 1501-2000 , 5: 2001-2500 , 6: 2501-3000 , 7: 3001-6250 , 8: ¿6250 , -1: NA	What is the patient's household income?
ever_marijuana	1: yes , 2: no , -1: NA	Did the patient ever smoke marijuana?
shared_needles	1: yes , 2: no , -1: NA	Did the patient ever share a needle?
ever_jail	1: yes , 2: no , -1: NA	Was the patient ever put in jail?
ever_sex_with _hiv_male	1: yes , 2: no , -1: NA	Did the patient ever have sex with an HIV positive male?
Condom_use	1: always , 2: sometimes , 3: never , -1: NA	Did the patient ever use condoms?
ever_sex_toys	1: yes , 2: no , -1: NA	Did the patient ever use sex toys?
sexual _orientation	1: heterosexual, straight , 2: lesbian/gay , 3: other , -1 : NA	What is the patient's sexual orientation?
CD4N	Continuous, -1: NA	What is the number of CD4 positive cell per mm^3 ?

10 References

"Domestic violence and childhood sexual abuse in HIV-infected women and women at risk for HIV", American Journal of Public Health 90, no. 4 (April 1, 2000): pp. 560-565.

Fonkwo, Peter Ndeboc. "Pricing Infectious Disease. The Economic and Health Implications of Infectious Diseases." EMBO Reports, Nature Publishing Group, July 2008, `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3327542/`.

Manly, Jennifer J., et al. "Relationship of Ethnicity, Age, Education, and Reading Level to Speed and Executive Function among HIV and HIV– Women: The Womens Interagency HIV Study (WIHS) Neurocognitive Substudy." Journal of Clinical and Experimental Neuropsychology, vol. 33, no. 8, 2011, pp. 853–863., https://doi.org/10.1080/13803395.2010.547662.

ODPHP. "HIV." HIV — Healthy People 2020, 27 Oct. 2021, https://www.healthypeople.gov/2020/topics-objectives/topic/hiv.

11 Appendix

https://github.com/amine-abdeljaoued/HIV_Regression_Analysis/blob/main/test.Rmd