Robertson and Cao Project

Part 1:Teenage HPV Vaccination Coverage and Socioeconomic Factors

Kelly Cao and Rachel Robertson

2024-02-04

# 1. Introduction

## 1.1 Data Set

The data we chose to use is the National Immunization Survey (NIS) of 2022. It consists survey data that was collected to monitor vaccination coverage for teenagers age 13-17 in the United States of America. The surveys were conducted by the National Center for Immunization and Respiratory Diseases of the Center (NCIRD) for Disease Control and Prevention (CDC). The survey itself consists of two parts: (1) the household telephone (random digit dialing) survey answered by a parent or guardian and (2) a mailed survey for the vaccination provider, called “Immunization History Questionnaire.” The original data set is a fixed width file (FWF). In the data set, there are 43,738 rows and 672 columns, with each row representing an individual and each column representing the answer to a question. The questions pertain to the teenager’s immunization history, demographics, and additional household-reported health information. We have truncated the data set to include 26-selected columns for the ease of observation. We have also filtered the data set to include one year, 2022, which is the most recent year of data that is published. The provided survey data will require cleaning, as there are signs of human errors present. ## Opening the Data Set

First, we will load the readr package to open the file

Next, we will define the column widths (beginning and end positions) and applying the column names to each position.

#Using the fwf\_position() in the readr package to identify the columns widths  
column\_position<- fwf\_positions(  
 start = c(1, 89, 287, 295, 298, 314, 330, 331, 334, 336, 343, 344, 345,  
 347, 355, 359, 362, 472, 473, 495, 496, 497, 502, 504, 1296, 1297  
 ),  
 end = c(5, 92, 288, 295, 313, 329, 330, 332, 334, 336, 343, 344, 346,  
 347, 356, 359, 362, 472, 473, 495, 496, 497, 502, 504, 1296, 1297  
 ),  
 col\_names = c("SEQNUMT", "YEAR", "AGE", "EDUC1", "INCPORAR", "INCPORAR\_I",   
 "INCPOV1", "INCQ298A", "LANGUAGE", "MOBIL\_1", "RACEETHK",   
 "RACE\_K", "RENT\_OWN", "SEX", "STATE", "FACILITY", "WELLCHILD",   
 "P\_U13HPV", "P\_U13HPV3", "P\_UTDHPV", "P\_UTDHPV\_15",   
 "P\_UTDHPV\_15INT", "P\_UTDHPV2", "P\_UTDHPV3", "INS\_STAT2\_I",   
 "INS\_BREAK\_I")  
)

Create the data frame “data” with read\_fwf function.

data\_file\_path <- here("products", "manuscript", "NISTEENPUF22.DAT")  
  
data\_frame <- read\_fwf(data\_file\_path, col\_positions = column\_position)

#Read the statistics in the data file.  
 summary(data\_frame)

SEQNUMT YEAR AGE EDUC1   
 Length:43738 Min. :2022 Min. :13.00 Min. :1.000   
 Class :character 1st Qu.:2022 1st Qu.:14.00 1st Qu.:3.000   
 Mode :character Median :2022 Median :15.00 Median :4.000   
 Mean :2022 Mean :15.03 Mean :3.242   
 3rd Qu.:2022 3rd Qu.:16.00 3rd Qu.:4.000   
 Max. :2022 Max. :17.00 Max. :4.000   
 INCPORAR INCPORAR\_I INCPOV1 INCQ298A   
 Length:43738 Min. :0.500 Min. :1.000 Min. : 3.0   
 Class :character 1st Qu.:1.660 1st Qu.:1.000 1st Qu.:12.0   
 Mode :character Median :3.000 Median :1.000 Median :14.0   
 Mean :2.356 Mean :1.713 Mean :18.9   
 3rd Qu.:3.000 3rd Qu.:2.000 3rd Qu.:14.0   
 Max. :3.000 Max. :4.000 Max. :99.0   
 LANGUAGE MOBIL\_1 RACEETHK RACE\_K   
 Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.00   
 1st Qu.:1.000 1st Qu.:2.000 1st Qu.:2.000 1st Qu.:1.00   
 Median :1.000 Median :2.000 Median :2.000 Median :1.00   
 Mean :1.099 Mean :1.762 Mean :2.129 Mean :1.41   
 3rd Qu.:1.000 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:2.00   
 Max. :3.000 Max. :2.000 Max. :4.000 Max. :3.00   
 RENT\_OWN SEX STATE FACILITY   
 Min. : 1.000 Min. :1.000 Length:43738 Length:43738   
 1st Qu.: 1.000 1st Qu.:1.000 Class :character Class :character   
 Median : 1.000 Median :1.000 Mode :character Mode :character   
 Mean : 2.262 Mean :1.482   
 3rd Qu.: 2.000 3rd Qu.:2.000   
 Max. :99.000 Max. :2.000   
 WELLCHILD P\_U13HPV P\_U13HPV3 P\_UTDHPV   
 Min. :1.000 Length:43738 Length:43738 Length:43738   
 1st Qu.:3.000 Class :character Class :character Class :character   
 Median :3.000 Mode :character Mode :character Mode :character   
 Mean :2.677   
 3rd Qu.:3.000   
 Max. :3.000   
 P\_UTDHPV\_15 P\_UTDHPV\_15INT P\_UTDHPV2 P\_UTDHPV3   
 Length:43738 Length:43738 Length:43738 Length:43738   
 Class :character Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character Mode :character   
   
   
   
 INS\_STAT2\_I INS\_BREAK\_I   
 Length:43738 Length:43738   
 Class :character Class :character   
 Mode :character Mode :character

str(data\_frame)

spc\_tbl\_ [43,738 × 26] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
 $ SEQNUMT : chr [1:43738] "30397" "02792" "27914" "37271" ...  
 $ YEAR : num [1:43738] 2022 2022 2022 2022 2022 ...  
 $ AGE : num [1:43738] 17 13 15 14 13 13 16 13 15 16 ...  
 $ EDUC1 : num [1:43738] 3 4 1 4 3 2 3 4 3 4 ...  
 $ INCPORAR : chr [1:43738] "3.00000000000000" "3.00000000000000" "2.29242125532988" "3.00000000000000" ...  
 $ INCPORAR\_I : num [1:43738] 3 3 2.292 3 0.857 ...  
 $ INCPOV1 : num [1:43738] 1 1 2 1 3 1 1 2 1 1 ...  
 $ INCQ298A : num [1:43738] 14 14 11 14 5 14 14 13 14 14 ...  
 $ LANGUAGE : num [1:43738] 1 1 1 1 1 1 1 1 1 1 ...  
 $ MOBIL\_1 : num [1:43738] 2 1 2 2 1 2 1 2 2 1 ...  
 $ RACEETHK : num [1:43738] 3 4 3 4 2 2 2 3 2 2 ...  
 $ RACE\_K : num [1:43738] 2 3 2 3 1 1 1 2 1 1 ...  
 $ RENT\_OWN : num [1:43738] 1 1 1 1 3 1 1 1 1 1 ...  
 $ SEX : num [1:43738] 2 1 2 2 2 1 1 2 2 1 ...  
 $ STATE : chr [1:43738] "31" "2." "21" "31" ...  
 $ FACILITY : chr [1:43738] "6" "." "6" "5" ...  
 $ WELLCHILD : num [1:43738] 1 3 3 3 3 3 3 1 3 3 ...  
 $ P\_U13HPV : chr [1:43738] "1" "." "0" "0" ...  
 $ P\_U13HPV3 : chr [1:43738] "0" "." "0" "0" ...  
 $ P\_UTDHPV : chr [1:43738] "1" "." "0" "1" ...  
 $ P\_UTDHPV\_15 : chr [1:43738] "1" "." "0" "1" ...  
 $ P\_UTDHPV\_15INT: chr [1:43738] "1" "." "0" "1" ...  
 $ P\_UTDHPV2 : chr [1:43738] "1" "." "0" "1" ...  
 $ P\_UTDHPV3 : chr [1:43738] "0" "." "0" "0" ...  
 $ INS\_STAT2\_I : chr [1:43738] "2" "." "2" "1" ...  
 $ INS\_BREAK\_I : chr [1:43738] "2" "." "2" "2" ...  
 - attr(\*, "spec")=  
 .. cols(  
 .. SEQNUMT = col\_character(),  
 .. YEAR = col\_double(),  
 .. AGE = col\_double(),  
 .. EDUC1 = col\_double(),  
 .. INCPORAR = col\_character(),  
 .. INCPORAR\_I = col\_double(),  
 .. INCPOV1 = col\_double(),  
 .. INCQ298A = col\_double(),  
 .. LANGUAGE = col\_double(),  
 .. MOBIL\_1 = col\_double(),  
 .. RACEETHK = col\_double(),  
 .. RACE\_K = col\_double(),  
 .. RENT\_OWN = col\_double(),  
 .. SEX = col\_double(),  
 .. STATE = col\_character(),  
 .. FACILITY = col\_character(),  
 .. WELLCHILD = col\_double(),  
 .. P\_U13HPV = col\_character(),  
 .. P\_U13HPV3 = col\_character(),  
 .. P\_UTDHPV = col\_character(),  
 .. P\_UTDHPV\_15 = col\_character(),  
 .. P\_UTDHPV\_15INT = col\_character(),  
 .. P\_UTDHPV2 = col\_character(),  
 .. P\_UTDHPV3 = col\_character(),  
 .. INS\_STAT2\_I = col\_character(),  
 .. INS\_BREAK\_I = col\_character()  
 .. )  
 - attr(\*, "problems")=<externalptr>

The data will then be cleaned. As of now, the data set requires a library to translate the code.

The data will then be converted the data frame to a readable and downloadable .csv file for further analysis

output\_file<-"NIS\_Teen\_Data\_2022.csv"  
write.csv(data\_frame, file = output\_file, row.names = FALSE)

## 1.2 Research Question

We want to use the 2022 National Immunization Survey for teens to identify which socioeconomic and geographic factors are associated with HPV vaccination completion for teenagers in the U.S. Specifically, we are asking the question, What is the likelihood of a teenager being up-to-date for their HPV vaccinations in the U.S., based on socioeconomic factors and geographic distribution. The outcome of this study will be measured using the columns that pertain to HPV vaccination completion. These variables are provider-collected, and they state whether or not the teen is (1) up-to-date with 1+ HPV shot, (2) up-to-date with 2+ HPV shots, or (3) up-to-date with 3+ HPV shots (excluding all vaccinations post-survey). Measuring HPV vaccine completion with “up-to-date” variables may be more robust than vaccination rate based on those who have completed the regime (2 or more shots), because not all teens included in the survey are at the age in which they would’ve completed the full regime. This means that younger teens who are up to date with 1+ shot will not be excluded from the analysis. We aim to measure determinants of socioeconomic status that are engrained into the survey questions. These factors include: family income, poverty status, income-to-poverty ratio, insurance status, insurance breaks, maternal education, living arrangement, and geographic mobility status. Additional demographic factors that are associated with healthcare access and may be examined include: race, ethnicity, language, facility in which the vaccine was administered, and whether the teen had completed a wellness exam between the ages of 11-12. Geographic distribution of the teens will be assessed by true state of residence. We would like to examine the data for correlations between HPV vaccination completion status and the aforementioned socioeconomic determinants, demographics, and geographic location. This serves as valuable information to determine differences in healthcare access and vaccination coverage for teenagers living in the U.S.

## 1.3 Proposed Analysis

For the analysis portion, we have decided to do a correlation analysis between HPV vaccination completion and each of the socioeconomic factors. When performing this, we may choose to stratify by age or gender, if we find these to be confounders. The correlation analysis cannot be performed using a multivariate model because there may be collinearity present between many of the determinants for socioeconomic status. This will lead to spurious results. Instead, the correlation between vaccination status and each factor is examined separately.   
Next, we may stratify this by state. Comparing these correlation values between states and regions may lead us to find patterns of healthcare accessibility in the U.S. It is generally known that those of lower income and without insurance have limited healthcare access in the U.S.(1), but it will be interesting to see if this pattern is reflected at the state level. If so, policy should be examined as it may be related to healthcare access disparities. Recent studies have examined HPV vaccination completion according to social determinants, but they have not examined geographic distribution and include data prior to 2019 (2). More recent years display differences in global HPV vaccination trends due to the COVID19 pandemic (3). Our study will fill in geographical gaps and provide updated vaccination trends based on socioeconomic factors.

## 1.4 References

1. Sciences, N. A. of, Engineering, & Medicine, and. (2018, March 1). Factors that affect health-care utilization. Health-Care Utilization as a Proxy in Disability Determination. https://www.ncbi.nlm.nih.gov/books/NBK500097/
2. Mansfield, L. N., Chung, R. J., Silva, S. G., Merwin, E. I., & Gonzalez-Guarda, R. M. (2022). Social determinants of human papillomavirus vaccine series completion among U.S. adolescents: A mixed-methods study. SSM - population health, 18, 101082. https://doi.org/10.1016/j.ssmph.2022.101082
3. Casey, R. M., Akaba, H., Hyde, T. B, et al. (2024). Covid-19 pandemic and equity of global human papillomavirus vaccination: descriptive study of World Health Organization-Unicef vaccination coverage estimates. BMJ Medicine, 3, 000726. doi: 10.1136/bmjmed-2023-000726