  
  
Project BRFSS Weekly Update

Joshua J. Cook, M.S. DS, M.S. CRM, ACRP-PM, CCRC,

Swann A. Adams Ph.D., M.S., FACE

2024-07-20

Table of contents

[Setup 1](#_Toc172342607)

[Purpose 2](#_Toc172342608)

[Temporal Amalgamation 2](#_Toc172342609)

[What’s Missing? 3](#_Toc172342610)

[Codebook / Data Dictionary 9](#_Toc172342611)

[Variable Breakdown for Potential Harmonizing 12](#_Toc172342612)

[Survey Design, Survey Stratification, Survey Weighting - GOT IT 12](#_Toc172342613)

[Sex - GOT IT (see notes) 12](#_Toc172342614)

[Sexual Orientation - GOT IT (see notes) 12](#_Toc172342615)

[Gender Identity - GOT IT 12](#_Toc172342616)

[Urban v. Rural - GOT IT (see notes) 12](#_Toc172342617)

[HPV Vaccination - GOT IT (see notes) 12](#_Toc172342618)

[HPV Testing (MAY BE A PART OF PAP TEST) - MISSING 1 YEAR (see notes) 13](#_Toc172342619)

[Cervical Cancer Screening (AKA PAP TEST) - MISSING 1 YEAR (see notes) 13](#_Toc172342620)

[Cervical Cancer Diagnosis - MISSING 1 YEAR (see notes) 13](#_Toc172342621)

[Summary 13](#_Toc172342622)

## Setup

library(here) # Managing nested directories   
library(haven) # Importing SAS XPT files  
library(readxl) # Importing Excel files  
library(tidyverse) # Data wrangling, viz  
library(reshape2) # For restructuring data  
library(gt) # For table generation

## Purpose

Temporal analysis (at least 10 years) of HPV screening/vaccination and cervical cancer diagnosis by sexual orientation and gender identity (SOGI).

* 2014-2023 (2023 not yet available as of 20JUL2024)
* Straight v. non-straight
  + urban v. rural
  + racial/ethnic
* Cis v. trans
  + urban v. rural
  + racial/ethnic

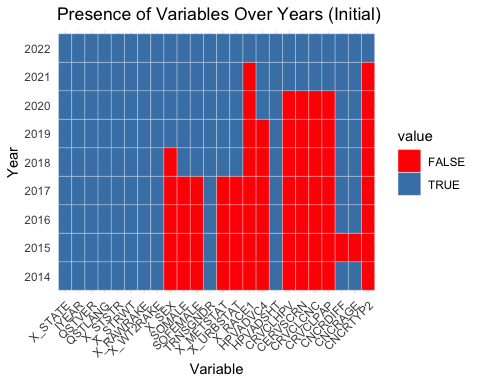
## Temporal Amalgamation

|  |
| --- |
| Warning |
| *\*Switched to SAS Transport Files for Faster / More Efficient Reading.* |

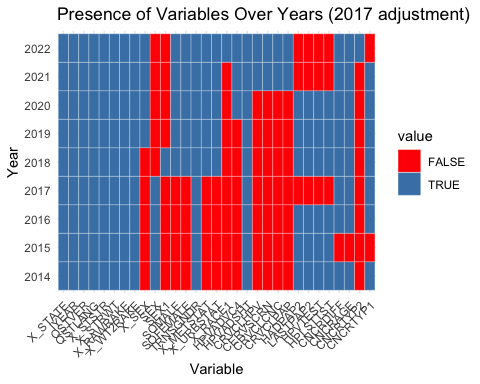
# Define a function to read, process, and save each SAS file individually to a compressed R RDS file  
process\_and\_save\_xpt\_files <- function(start\_year, end\_year, subdirectory, file\_extension, temp\_directory) {  
 for (year in start\_year:end\_year) {  
 # Construct the file path  
 file\_path <- here(subdirectory, paste0(year, "\_data", file\_extension))  
   
 # Read the XPT file  
 data <- read\_xpt(file\_path)  
   
 # Ensure consistent data types  
 data <- mutate(data, SEQNO = as.character(SEQNO))  
   
 # Save the processed data to a temporary file  
 saveRDS(data, file = here(temp\_directory, paste0(year, "\_data.rds")))  
 }  
}  
  
# Create a temporary directory to save intermediate files  
temp\_directory <- "temp\_data"  
dir.create(temp\_directory, showWarnings = FALSE)  
  
# Process and save each XPT file individually  
process\_and\_save\_xpt\_files(  
 start\_year = 2014,  
 end\_year = 2022,  
 subdirectory = "Data",  
 file\_extension = ".XPT",  
 temp\_directory = temp\_directory  
)

## What’s Missing?

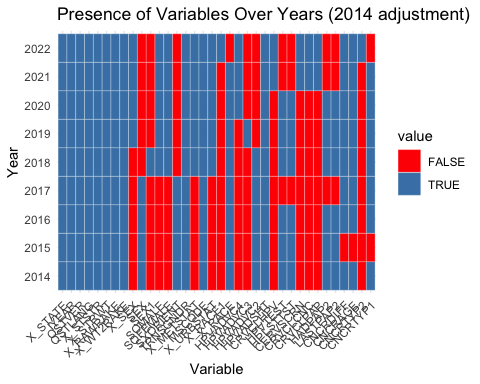
# Define a function to check the presence of variables and combine the results  
check\_and\_combine\_variable\_presence <- function(start\_year, end\_year, temp\_directory, variables\_of\_interest) {  
 # Initialize an empty list to store results  
 presence\_list <- list()  
   
 for (year in start\_year:end\_year) {  
 # Construct the file path for the RDS file  
 file\_path <- here(temp\_directory, paste0(year, "\_data.rds"))  
   
 # Read the RDS file  
 data <- readRDS(file\_path)  
   
 # Check for the presence of each variable of interest  
 presence <- sapply(variables\_of\_interest, function(var) var %in% names(data))  
   
 # Create a data frame with the results for the current year  
 presence\_df <- data.frame(Year = year, t(presence))  
   
 # Add the data frame to the list  
 presence\_list[[length(presence\_list) + 1]] <- presence\_df  
   
 # Remove the original data to free up memory  
 rm(data)  
 gc() # Call garbage collection to free up memory  
 }  
   
 # Combine the list of data frames into a single data frame  
 combined\_presence <- bind\_rows(presence\_list)  
   
 return(combined\_presence)  
}  
  
# Variables of interest  
variables\_of\_interest1 <- c(  
 "\_STATE",  
 "IYEAR",  
 "QSTVER",  
 "QSTLANG",  
 "\_STSTR",  
 "\_STRWT",  
 "\_RAWRAKE",  
 "\_WT2RAKE",  
 "\_SEX",  
 "SOMALE",   
 "SOFEMALE",   
 "TRNSGNDR",  
 "\_METSTAT",  
 "\_URBSTAT",  
 "\_RACE1",  
 "HPVADVC4",  
 "HPVADSHT",  
 "CRVCLHPV",  
 "CERVSCRN",  
 "CRVCLCNC",  
 "CRVCLPAP",  
 "CNCRDIFF",  
 "CNCRAGE",  
 "CNCRTYP2"  
)  
  
variables\_of\_interest2 <- c(  
 "\_STATE",  
 "IYEAR",  
 "QSTVER",  
 "QSTLANG",  
 "\_STSTR",  
 "\_STRWT",  
 "\_RAWRAKE",  
 "\_WT2RAKE",  
 "\_SEX",  
 "SEX",  
 "SEX1",  
 "SOMALE",   
 "SOFEMALE",   
 "TRNSGNDR",  
 "\_METSTAT",  
 "\_URBSTAT",  
 "\_RACE1",  
 "HPVADVC4",  
 "HPVADSHT",  
 "CRVCLHPV",  
 "CERVSCRN",  
 "CRVCLCNC",  
 "CRVCLPAP",  
 "HADPAP2",  
 "LASTPAP2",  
 "HPVTEST",  
 "HPLSTTST",  
 "CNCRDIFF",  
 "CNCRAGE",  
 "CNCRTYP2",  
 "CNCRTYP1"  
)  
  
variables\_of\_interest3 <- c(  
 "\_STATE",  
 "IYEAR",  
 "QSTVER",  
 "QSTLANG",  
 "\_STSTR",  
 "\_STRWT",  
 "\_RAWRAKE",  
 "\_WT2RAKE",  
 "\_SEX",  
 "SEX",  
 "SEX1",  
 "SOMALE",   
 "SOFEMALE",  
 "SXORIENT",  
 "TRNSGNDR",  
 "\_METSTAT",  
 "MSCODE",  
 "\_URBSTAT",  
 "\_RACE1",  
 "\_RACE",  
 "HPVADVC4",  
 "HPVADVC3",  
 "HPVADVC2",  
 "HPVADSHT",  
 "CRVCLHPV",  
 "HPVTEST",  
 "HPLSTTST",  
 "CERVSCRN",  
 "CRVCLCNC",  
 "CRVCLPAP",  
 "HADPAP2",  
 "LASTPAP2",  
 "CNCRDIFF",  
 "CNCRAGE",  
 "CNCRTYP2",  
 "CNCRTYP1"  
)  
  
# Use the function to check variable presence and combine results from 2014 to 2022 (initial)  
combined\_presence1 <- check\_and\_combine\_variable\_presence(  
 start\_year = 2014,  
 end\_year = 2022,  
 temp\_directory = "temp\_data",  
 variables\_of\_interest = variables\_of\_interest1  
)  
  
# Use the function to check variable presence and combine results from 2014 to 2022 (2017 identified variables)  
combined\_presence2 <- check\_and\_combine\_variable\_presence(  
 start\_year = 2014,  
 end\_year = 2022,  
 temp\_directory = "temp\_data",  
 variables\_of\_interest = variables\_of\_interest2  
)  
  
# Use the function to check variable presence and combine results from 2014 to 2022 (2015 identified variablbes)  
combined\_presence3 <- check\_and\_combine\_variable\_presence(  
 start\_year = 2014,  
 end\_year = 2022,  
 temp\_directory = "temp\_data",  
 variables\_of\_interest = variables\_of\_interest3  
)  
  
  
# Optionally, clean up temporary files  
#unlink(temp\_directory, recursive = TRUE)  
  
# Reshape data (long) to create heatmaps of missing / present variables  
column\_presence\_df\_long1 <- melt(combined\_presence1, id.vars = "Year")  
column\_presence\_df\_long2 <- melt(combined\_presence2, id.vars = "Year")  
column\_presence\_df\_long3 <- melt(combined\_presence3, id.vars = "Year")  
  
# Generate heatmaps given each selection of variables to see where variables were swapped over the years  
ggplot(column\_presence\_df\_long1, aes(x = variable, y = factor(Year), fill = value)) +  
 geom\_tile(color = "white") +  
 scale\_fill\_manual(values = c("TRUE" = "steelblue", "FALSE" = "red")) +  
 theme\_minimal() +  
 labs(title = "Presence of Variables Over Years (Initial)", x = "Variable", y = "Year") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



ggplot(column\_presence\_df\_long2, aes(x = variable, y = factor(Year), fill = value)) +  
 geom\_tile(color = "white") +  
 scale\_fill\_manual(values = c("TRUE" = "steelblue", "FALSE" = "red")) +  
 theme\_minimal() +  
 labs(title = "Presence of Variables Over Years (2017 adjustment)", x = "Variable", y = "Year") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



ggplot(column\_presence\_df\_long3, aes(x = variable, y = factor(Year), fill = value)) +  
 geom\_tile(color = "white") +  
 scale\_fill\_manual(values = c("TRUE" = "steelblue", "FALSE" = "red")) +  
 theme\_minimal() +  
 labs(title = "Presence of Variables Over Years (2014 adjustment)", x = "Variable", y = "Year") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



## Codebook / Data Dictionary

codes <- read\_xlsx ("codebook.xlsx", sheet = "Current\_Codebook")  
  
codes %>%  
 gt() %>%  
 opt\_interactive(  
 use\_search=TRUE,  
 use\_filters=TRUE  
 )

| variable\_name | variable\_label | variable\_type | variable\_type\_actual | variable\_question | variable\_section |
| --- | --- | --- | --- | --- | --- |
| \_STATE | FIPS Code | Num | Factor | FIPS Code | Record Identification |
| IYEAR | Interview Year | Char | Factor | Interview Year | Record Identification |
| QSTVER | Questionnaire Version Identifier | Num | Factor | Questionnaire Version Identifier | Questionnaire Version |
| QSTLANG | QSTLANG | Num | Factor | Questionnaire Language | Questionnaire Language |
| \_STSTR | Sample Design Stratification Variable | Num | Factor | Sample Design Stratification Variable | Weighting Variables |
| \_STRWT | Stratum weight | Num | Num | Stratum weight | Weighting Variables |
| \_RAWRAKE | Raw weighting factor used in raking | Num | Num | Raw weighting factor used in raking | Weighting Variables |
| \_WT2RAKE | Design weight use in raking | Num | Num | Design weight use in raking | Weighting Variables |
| \_SEX | Calculated sex variable | Num | Factor | Calculated sex variable | Calculated Variables |
| SEX | Calculated sex variable | Num | Factor | Calculated sex variable | Calculated Variables |
| SEX1 | Calculated sex variable | Num | Factor | Calculated sex variable | Calculated Variables |
| SOMALE | Sexual orientation | Num | Factor | Which of the following best represents how you think of yourself? | Sexual Orientation and Gender Identity (SOGI) |
| SOFEMALE | Sexual orientation | Num | Factor | Which of the following best represents how you think of yourself? | Sexual Orientation and Gender Identity (SOGI) |
| SXORIENT | Sexual orientation or gender identity | Num | Factor | Do you consider yourself to be: (We ask this question in order to better understand the health and health care needs of people with different sexual orientations.) | Sexual Orientation and Gender Identity |
| TRNSGNDR | Which of the following best represents how you think of yourself? | Num | Factor | Do you consider yourself to be transgender?  (If yes, ask “Do you consider yourself to be male-to-female, female-to-male, or gender non-conforming?) | Sexual Orientation and Gender Identity (SOGI) |
| \_METSTAT | Metropolitan Status | Num | Factor | Metropolitan Status | Urban Rural |
| MSCODE | Metropolitan Status Code | Num | Factor | Metropolitan Status Code | Weighting Variables |
| \_URBSTAT | Urban/Rural Status | Num | Factor | Urban/Rural Status | Urban Rural |
| \_RACE1 | Race/ethnicity categories | Num | Factor | Race/ethnicity categories | Calculated Race Variables |
| \_RACE | Race/ethnicity categories | Num | Factor | Race/ethnicity categories | Calculated Race Variables |
| HPVADVC4 | Have you ever had an H.P.V. vaccination? | Num | Factor | Have you ever had an H.P.V. vaccination? | HPV Vaccination |
| HPVADVC3 | Have you ever had an H.P.V. vaccination? | Num | Factor | Have you ever had an H.P.V. vaccination? | HPV Vaccination |
| HPVADVC2 | Have you ever had the HPV vaccination? | Num | Factor | A vaccine to prevent the human papilloma virus or H.P.V. infection is available and is called the cervical cancer or genital warts vaccine, H.P.V. shot, [Fill: if female “GARDASIL or CERVARIX”, if male “GARDASIL”]. Question: Have you ever had the H.P.V. vaccination? (Human Papilloma Virus (Human Pap•uh•loh•muh Virus), Gardasil (Gar•duh• seel), Cervarix (Serv a rix)) | Adult Human Papilloma Virus (HPV) - Vaccination |
| HPVADSHT | How many HPV shots did you receive? | Num | Factor | How many HPV shots did you receive? | HPV Vaccination |
| CRVCLHPV | Have an H.P.V. test and recent cervical cancer screening | Num | Factor | At your most recent cervical cancer screening, did you have an H.P.V. test? | Breast and Cervical Cancer Screening |
| HPVTEST | Have you ever had an HPV test? | Num | Factor | An H.P.V. test is sometimes given with the Pap test for cervical cancer screening. Have you ever had an H.P.V. test? | Breast and Cervical Cancer Screening |
| HPLSTTST | How long has it been since you had your last H.P.V. test? | Num | Factor | How long has it been since you had your last H.P.V. test? | Breast and Cervical Cancer Screening |
| CERVSCRN | Have you ever had a cervical cancer screening test? | Num | Factor | Have you ever had a cervical cancer screening test? | Breast and Cervical Cancer Screening |
| CRVCLCNC | Time since last cervical cancer screening test | Num | Factor | How long has it been since you had your last cervical cancer screening test? | Breast and Cervical Cancer Screening |
| CRVCLPAP | Have a PAP test and recent cervical cancer screening | Num | Factor | At your most recent cervical cancer screening, did you have a Pap test? | Breast and Cervical Cancer Screening |
| HADPAP2 | Ever Had a Pap Test | Num | Factor | Have you ever had a Pap test? | Breast and Cervical Cancer Screening |
| LASTPAP2 | How Long Since Last Pap Test | Num | Factor | How long has it been since you had your last Pap test? | Breast and Cervical Cancer Screening |
| CNCRDIFF | How Many Types of Cancer? | Num | Factor | How Many Types of Cancer? | Cancer Survivorship: Type of Cancer |
| CNCRAGE | Age Told Had Cancer | Num | Factor | At what age were you told that you had cancer?  (If Response = 2 (Two) or 3 (Three or more), ask: “At what age was your first diagnosis of cancer?”) | Cancer Survivorship: Type of Cancer |
| CNCRTYP2 | Type of Cancer | Num | Factor | What kind of cancer is it? | Cancer Survivorship: Type of Cancer |
| CNCRTYP1 | Type of Cancer | Num | Factor | What type of cancer was it? (If Response = 2 (Two) or 3 (Three or more), ask: “With your most recent diagnoses of cancer, what type of cancer was it?”) | Cancer Survivorship: Type of Cancer |

## Variable Breakdown for Potential Harmonizing

|  |
| --- |
| Warning |
| *\*Need 2023 data release to have 10 years for temporal paper.* |

### Survey Design, Survey Stratification, Survey Weighting - GOT IT

### Sex - GOT IT (see notes)

* 2022-2019 = X\_SEX
* 2018 = SEX\_1
* 2017-2014 = SEX

### Sexual Orientation - GOT IT (see notes)

* 2022-2018 = SOMALE (for sex=male), SOFEMALE (for sex=female)
* 2017-2014 = SXORIENT (need to subset by Sex to match SOMALE and SOFEMALE)

### Gender Identity - GOT IT

### Urban v. Rural - GOT IT (see notes)

* Use MSCODE for all years
* More specifics available for 2022-2018 via X\_URBSTAT and X\_METSTAT (don’t use?)

### HPV Vaccination - GOT IT (see notes)

* 2022-2020 = HPVADVC4, HPVADSHT
* 2019 = HPVADVC3, HPVADSHT
* 2018-2014 = HPVADVC2, HPVADSHT

### HPV Testing (MAY BE A PART OF PAP TEST) - MISSING 1 YEAR (see notes)

* 2022-2021 = CRVCLHPV
* 2020-2018, 2016-2014 = HPVTEST, HPLSTTST
* 2017 = MISSING

### Cervical Cancer Screening (AKA PAP TEST) - MISSING 1 YEAR (see notes)

* 2022-2021 = CERVSCRN, CRVCLCNC, CRVCLPAP
* 2020-2018, 2016-2014 = HADPAP2, LASTPAP2
* 2017 = MISSING

### Cervical Cancer Diagnosis - MISSING 1 YEAR (see notes)

* 2022 = CNCRDIFF, CNCRAGE, CNCRTYP2
* 2021-2016, 2014 = CNCRDIFF, CNCRAGE, CNCRTYP1
* 2015 = MISSING

## Summary

To complete a temporal analysis as planned, we need to wait for 2023 data. In the meantime, we can begin **cleaning and analyzing 2022-2014 data as noted above.**

We have complete and clean data for the following:

* survey design,
* survey stratification,
* survey weighting,
* gender identity.

We can engineer complete and clean data for:

* sex,
* sexual orientation,
* urban v. rural status,
* HPV vaccination.

We will have incomplete but clean data (missing 1 year each) for:

* HPV testing,
* cervical cancer screening,
* cervical cancer diagnoses.

These findings impact our timelines for the abstract and manuscript submissions, but do not impact our technical papers (which are due in the coming weeks).