A Guide to Public-Use Research Data Sources Containing Race and Health Outcomes

Proposed list of authors in alpha order: Marcela Alsan, Cong Gian and Kosali Simon [1]

Table of Contents

Introduction

In this document, we provide a detailed guidance on obtaining and analyzing data from five nationally representative, public-use data sources collected by U.S. federal agencies, that contain information about race and health outcomes. This file accompanies the Excel document that provides a comparison of various features of the data sources. This is a "live" document to which we will continue adding more data sources in the following dimensions 1) ones that contain health insurance and health care access data (these five all do as well, but there are other data sets that, say, contain only health insurance and race); 2) ones that are non-public use but are not prohibitively expensive or prohibitively administratively difficult to obtain, and we will describe the data protection protocols in place; 3) ones that contain socioeconomic determinants of health (this will be a large potential set, including education, housing, correctional systems, etc). Please send us any suggestions you have for making this a useful resource.

A. Description

1. Behavioral Risk Factor Surveillance System (BRFSS)

1.1. <u>Dataset Description:</u>
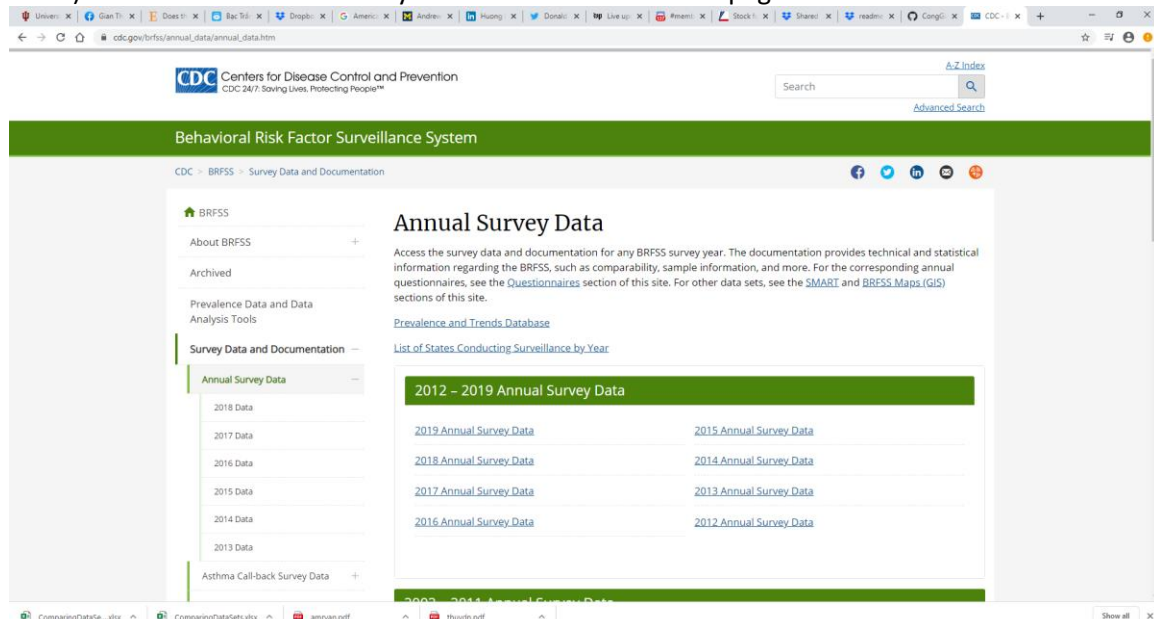(*From BRFSS website at https://www.cdc.gov/brfss/index.html*)
"The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world"

1.2. <u>Years Available:</u> 2007-2018

1.3. <u>Download instructions</u>

a) Go to https://www.cdc.gov/brfss/annual_data/annual_data.htm
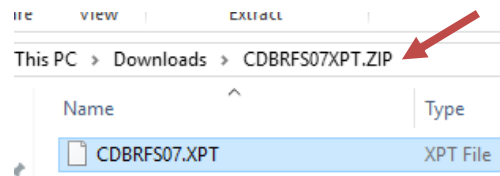b) Click "*Year* Annual Survey Data." You will then see a page like this.

c) For each year, under the "Data Files" section, select the appropriate "*Year* BRFSS Data (SAS Transport Format)" Zip File. The correct download will look like below (for the year 20007)

2007 BRFSS Data (SAS Transport Format) [ZIP - 103

MB]

Data updated October 23, 2008

This file was exported from SAS V8.2 in the XPT

transport format. This file contains 342 variables. This

format can be imported into SPSS or STATA. Please

d) The Zip file should download automatically. Once it is finished, extract the .XTP data file and add it to the folder that contains raw data. Rename the raw files as needed to be consistent and to clearly denote the file year. An XPT file looks something like this

e) Repeat steps c) and d) for other years
f) Refer to the code in the Appendix to convert .xpt file to .dta file
g) Save .dta files to a raw data folder

1.4. Notes for users of this dataset
"The BRFSS Data User Guide is intended to provide a brief overview of BRFSS to data users. Specific information regarding data quality, response and/or cooperation rates, or calling outcome can be found in the Summary Data Quality Report produced each year in conjunction with the annual data release. Data users needing more information about comparability across years should refer to the annual Comparability of Data document, particularly before using the data to conduct trend analyses"

1.5. References and papers
Disability, gender, and intimate partner violence: Relationships from the behavioral risk factor surveillance system (Smith 2008)
Demographic characteristics and health status of transgender adults in select US regions: Behavioral Risk Factor Surveillance System, 2014 (Meyer et al. 2017)
Time trends in racial and ethnic disparities in asthma prevalence in the United States from the Behavioral Risk Factor Surveillance System (BRFSS) Study (1999-2011) (Bhan et al. 2015)

2. Rand Health Retirement Study Longitudinal (HRS)

2.1. Dataset Description
(From HRS documentation)
"The RAND HRS Longitudinal File is a cleaned, easy-to-use, and streamlined version of the Health and Retirement Study (HRS) core interviews, with derived variables covering a very large range of measures (demographics, health, health insurance, out-of-pocket medical expenditures, income, Social Security, pensions, wealth, family structure, retirement plans, expectations, employment history, etc.). Variables are named and derived consistently across waves, and any cross-wave differences are documented"

"The file includes imputations for income, assets, and medical expenditures developed at RAND. Every individual interviewed by HRS has a record in this file. As such, the RAND HRS Longitudinal File meets essential needs of most if not all research projects. However, we note that only a large subset of the core interview information is included. The file can easily be merged with other RAND HRS data products, as well as those released separately by HRS"

"The RAND HRS Longitudinal File 2016 (V1) includes 13 waves (1992-2016; and 1993 and 1995 of AHEAD) and all five entry cohorts: the initial 1992 Health and Retirement Study (HRS) cohort; the 1993 Study of Assets and Health Dynamics (AHEAD) cohort; the Children of Depression and War Baby cohorts entering in 1998; the Early Baby Boomer cohort entering in 2004; and the Mid Baby Boomer cohort entering in 2010)".

2.2. <u>Years available</u>: 1992-2016

2.3. <u>Download Instruction</u>

a) Go to https://hrs.isr.umich.edu/data-products and scroll down to the bottom left, then click the link to Register and Access Public Data; either log in if an existing user or register if a new user

b) Once logged in, go to the data downloads page which is linked to on the log in page

c) Click "RAND HRS Longitudinal File 2016 (v.1)" under the heading "Rand Contributed Files" on the right side of the page

d) Download randhrs1992_2016v1_STATA.zip

e) Unzip the .zip file and store the resulting .dta file in the "input_data" directory

f) Run panelrandhrs1992_2016.do. This results in a dataset called panelrandhrs1992_2016.dta. Variables are labeled almost identically to in the raw data, except changes due to data being in long rather than wide form. Variables are, with the exception of the wave variable and the year variable, in the order in which the documentation lists them. A note is attached to each variable describing from which section of the documentation they come, e.g., "Section A Demographics, identifiers, and weights".

2.4. <u>References and papers</u>

3. National Health and Nutrition Examination Survey (NHANES)

3.1. <u>Datasets Description</u>

(From NHANES website https://wwwn.cdc.gov/nchs/nhanes/default.aspx)

"The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations. NHANES is a major program of the National Center for Health Statistics (NCHS). NCHS is part of the Centers for Disease Control and Prevention (CDC) and is responsible for producing vital and health statistics for the nation. In 1999, the survey became a continuous program that has a changing focus on a variety of health and nutrition measurements to meet emerging needs. The survey examines a nationally representative sample of about 5,000 people each year. These people are located in counties across the country, 15 of which are visited each year. The NHANES interview includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel."

3.2. Years available: 1999-2018

3.3. Download instructions
   Health Condition Measures
   a) For health condition measures, the data are from the main questionnaire. Go to the link below for 1999-2000 data
https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire&CycleBeginYear=1999
   b) Download the .xpt file (as indicated by the arrow) for each of the section of data that you are interested in. you can also download the .doc with variables descriptions for each of section
   c) Repeat a) and b) for data in other years

Details are as follows:

Acculturation	ACQ Doc	ACQ Data [XPT - 629.2 KB]
Alcohol Use	ALQ Doc	ALQ Data [XPT - 314.5 KB]
Analgesic Pain Relievers	RXQ_ANA Doc	RXQ_ANA Data [XPT - 279.2 KB]
Audiometry	AUQ Doc	AUQ Data [XPT - 966.6 KB]
Balance	BAQ Doc	BAQ Data [XPT - 601.2 KB]
Blood Pressure & Cholesterol	BPQ Doc	BPQ Data [XPT - 1.7 MB]
Cardiovascular Health	CDQ Doc	CDQ Data [XPT - 250.9 KB]
Cognitive Functioning	CFQ Doc	CFQ Data [XPT - 116.5 KB]
Current Health Status	HSQ Doc	HSQ Data [XPT - 553.8 KB]
Dermatology	DEQ Doc	DEQ Data [XPT - 3.5 MB]
Diabetes	DIQ Doc	DIQ Data [XPT - 1.2 MB]
Diet Behavior & Nutrition	DBQ Doc	DBQ Data [XPT - 3.5 MB]
Drug Use	DUQ Doc	DUQ Data [XPT - 112.3 KB]
Early Childhood	ECQ Doc	ECQ Data [XPT - 462.3 KB]
Food Security	FSQ Doc	FSQ Data [XPT - 1.2 MB]
Health Insurance	HIQ Doc	HIQ Data [XPT - 780.6 KB]
Hospital Utilization & Access to Care	HUQ Doc	HUQ Data [XPT - 780.6 KB]
Housing Characteristics	HOQ Doc	HOQ Data [XPT - 1.3 MB]
Immunization	IMQ Doc	IMQ Data [XPT - 312.7 KB]
Kidney Conditions	KIQ Doc	KIQ Data [XPT - 804.2 KB]
Medical Conditions	MCQ Doc	MCQ Data [XPT - 12.1 MB]
Mental Health - Depression	CIQMDEP Doc	CIQMDEP Data [XPT - 1.1 MB]
Mental Health - Generalized Anxiety Disorder	CIQGAD Doc	CIQGAD Data [XPT - 765.2 KB]
Mental Health - Panic Disorder	CIQPANIC Doc	CIQPANIC Data [XPT - 1.2 MB]
Miscellaneous Pain	MPQ Doc	MPQ Data [XPT - 2.2 MB]
Occupation	OCQ Doc	OCQ Data [XPT - 2 MB]
Oral Health	OHQ Doc	OHQ Data [XPT - 1012.2 KB]

Osteoporosis	OSQ Doc	OSQ Data [XPT - 2 MB]
Pesticide Use	PUQ Doc	PUQ Data [XPT - 1.1 MB]
Physical Activity	PAQ Doc	PAQ Data [XPT - 1.5 MB]
Physical Activity - Individual Activities	PAQIAF Doc	PAQIAF Data [XPT - 508.5 KB]
Physical Functioning	PFQ Doc	PFQ Data [XPT - 6.7 MB]
Prescription Medications	RXQ_RX Doc	RXQ_RX Data [XPT - 3 MB]
Prescription Medications - Drug Information	RXQ_DRUG Doc	RXQ_DRUG Data [XPT - 2.6 KB]
Reproductive Health	RHQ Doc	RHQ Data [XPT - 2.5 MB]
Respiratory Health	RDQ Doc	RDQ Data [XPT - 1.4 MB]
Sexual Behavior	SXQ Doc	SXQ Data [XPT - 402.3 KB]
Smoking - Adult Recent Tobacco Use & Youth Cigarette/Tobacco Use	SMQMEC Doc	SMQMEC Data [XPT - 2.2 MB]
Smoking - Cigarette/Tobacco Use - Adult	SMQ Doc	SMQ Data [XPT - 1.7 MB]
Smoking - Household Smokers	SMQFAM Doc	SMQFAM Data [XPT - 702.7 KB]
Social Support	SSQ Doc	SSQ Data [XPT - 290 KB]
Tuberculosis	TBQ Doc	TBQ Data [XPT - 520.9 KB]
Vision	VIQ Doc	VIQ Data [XPT - 229.2 KB]
Weight History	WHQ Doc	WHQ Data [XPT - 2 MB]

Demographic information

a) Demographic information is from a different questionnaire. In order to retrieve information of age, gender, race, ect.,go to the link below
https://wwwn.cdc.gov/nchs/nhanes/Search/DataPage.aspx?Component=Demographics&CycleBeginYear=1999
b) Download the XPT file (as indicated by the arrow)
c) Repeat a) and (b) for data in other years

Data File Name	Doc File	Data File	Date Published
Demographic Variables & Sample Weights	DEMO Doc	DEMO Data [XPT - 11 MB]	Updated September 2009

Other health related information

NHANES also provide data from other modules including Dietary, Examination, Laboratory, and Other Limited Access data. Each module has a corresponding data sets for all the years NHANES was conducted

3.4. References and Papers

4. National Health Interview Survey (NHIS)

4.1. Dataset Description

"The National Health Interview Survey (NHIS) has monitored the health of the nation since 1957. NHIS data on a broad range of health topics are collected through personal household interviews. For over 50 years, the U.S. Census Bureau has been the data collection agent for the

National Health Interview Survey. Survey results have been instrumental in providing data to track health status, health care access, and progress toward achieving national health objectives"
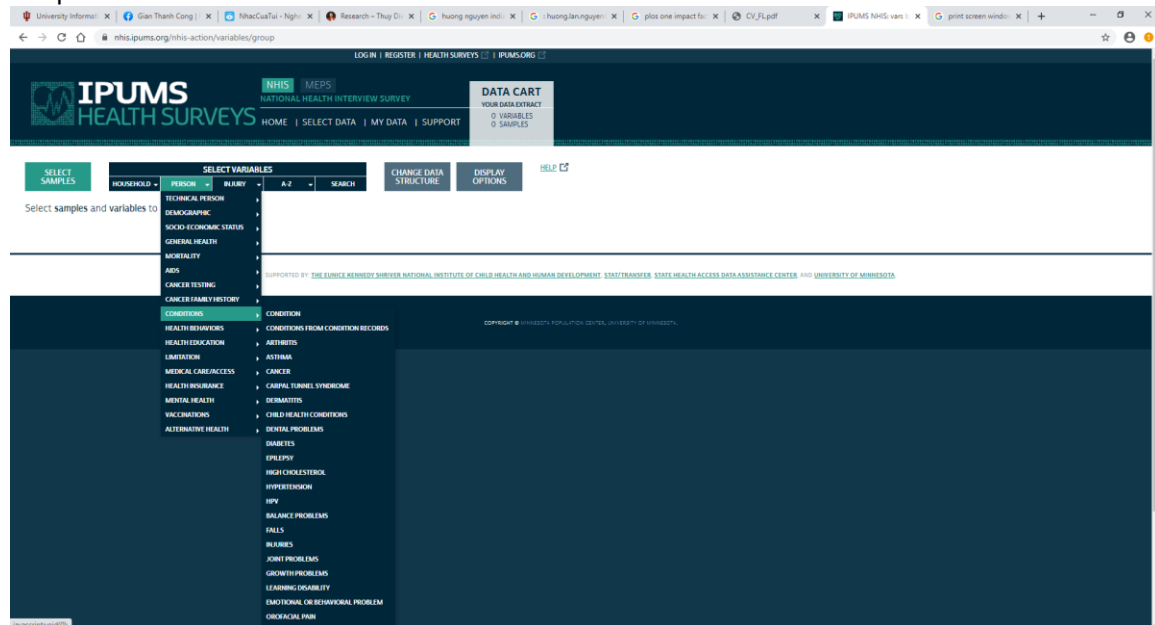
4.2. Years available: 1994-2018

4.3. Download instructions
Health conditions
a) Go to https://nhis.ipums.org/nhis-action/data_requests/download.
b) Select "all samples" to get a long series of years
c) Select the tab "person" for person-level variables (not "household")
d) Select all health conditions that seemed to be asked all years from 1994 onwards. Please see the picture below

Demographic information
d) Select demographic tab that contains age, race and gender variables
e) Submit a request and wait until IPUMS notifies of the availability of download
f) Download .dat and the Stata code from the download page
g) Open the Stata code to read the .dta file.
h) Save the .dta file into a raw data folder

4.4. References and Papers
Lynn A. Blewett, Julia A. Rivera Drew, Miriam L. King and Kari C.W. Williams. IPUMS Health Surveys: National Health Interview Survey, Version 6.4 [dataset]. Minneapolis, MN: IPUMS, 2019.
https://doi.org/10.18128/D070.V6.4

5. Medical Expenditure Panel Survey (MEPS)
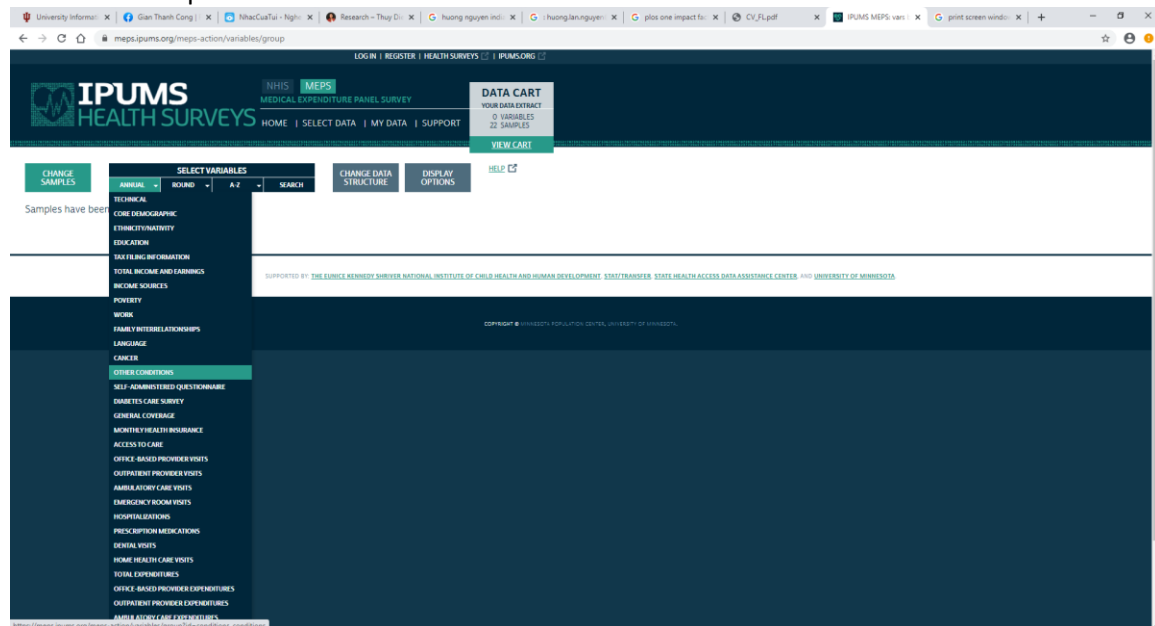5.1. Dataset Description

From MEPS website: "The Medical Expenditure Panel Survey, which began in 1996, is a set of large-scale surveys of families and individuals, their medical providers (doctors, hospitals, pharmacies, etc.), and employers across the United States. MEPS collects data on the specific health services that Americans use, how frequently they use them, the cost of these services, and how they are paid for, as well as data on the cost, scope, and breadth of health insurance held by and available to U.S. workers"

5.2. Years Available: 1996-2017

5.3. Download Instructions

a)   Go to https://healthsurveys.ipums.org/
b)   Select "all samples" to get a long series of years
c)   Select the tab "person" for person-level variables (not "household")
d)   Select all health conditions. Selected variables from annual, not round option. That should make this like a repeated cross section of sorts. Of course person id is available to link peoples answers from one year to the next but if doing longitudinal work, it would be much better to do round level
   Please see the picture below

Demographic information
d)   Select demographic tab that contains age, race and gender variables
e)   Submit a request and wait until IPUMS notifies of the availability of download
f)   Download .dat and the Stata code from the download page
g)   Open the Stata code to read the .dta file.
h)   Save the .dta file into a raw data folder

5.4. References and Papers
Lynn A. Blewett, Julia A. Rivera Drew, Risa Griffin and Kari C.W. Williams. IPUMS Health Surveys: Medical Expenditure Panel Survey, Version 1.1 [dataset]. Minneapolis, MN: IPUMS, 2019.

B. Comparison and harmonization of health variables across data sets

	Behavioral Risk Factor Surveillance System	National Health Interview Survey	National Health and Nutrition Examination Survey	Medical Expenditure Panel Survey	Health Retirement Survey
Health Measures					
Asthma	**Variable/Question** *(asthma2): Have you ever been told by a doctor, nurse, or other health professional that you had asthma?*	**Variable/Question** *(Asthmaev): Ever told to have asthma before?*	**Variable/Question** *(MCQ010): Ever been told you have asthma?*	**Variable/Question** *(Asthmaev): Ever told to have asthma before?*	**Variable/Question** (*rheart*) *if new interview "Ever told to have diabetes before?"* *If reinterviewed and no condition "Since we last talked to you, that is since [last interview date], has a doctor told you that you have"* *if reinterviewed and had condition " "Our records from your last interview show that you have had"*
	Categories: 1- Yes; 2- No; 7-Don't Know/Not Sure; 9- Refused to answer	*Categories 0-niu; 1- no or not mentioned; 2-- Yes; 7-Unknown-Refused; 8-Unknown-Not certain; 9-Unknown- Don't know*	*Categories: 1-Yes; 2-No; 7-Refused; 9-Don't know*	*Categories: 0- niu; 1- no or not mentioned; 2- Yes; 7-Unknown-Refused; 8-Unknown-Not certain; 9-Unknown- Don't know*	*Categories 1-no; 2-- Yes;3- Disp prev record and has cond; 4- Disp prev record and no cond; 5- Disp prev record (DK if cond)*
Cancer	**Variable/Question**: (*cncrhave* for 2009-2010 and *chcocncr* for 2011-	**Variable/Question** (*cancerev*): *Ever told you had cancer before ?*	**Variable/Question** (*mcq220*) *Ever told you*	**Variable/Question** (*cancerev*): *Ever told you had cancer before ?*	**Variable/Question** (*rheart*) *if new interview "Ever told to have diabetes before?"*

2018): *Ever told you had cancer before ?*		*had cancer or malignancy?*		*If <u>reinterviewed and no condition</u> "Since we last talked to you, that is since [last interview date], has a doctor told you that you have"* *if <u>reinterviewed and had condition</u> " "Our records from your last interview show that you have had"*
Categories*: 1- Yes;2- No; 7-Don't know/Not Sure—Go to next module; 9-Refused—Go to next module; Blank- Not asked or Missing*	***Categories****: 0- niu;1-Yes;2- No; 7- Unknown - Refused; 8- Unknown- Not ascertained; 9- Unknown - Don't know*	***Categories****: 1 Yes;2 No;7 Refused;9 Don't know; . Missing*	***Categories****: 0- niu; 1- Yes; 2- No; 7- Unknown - Refused; 8- Unknown - Not ascertained; 9- Unknown- Don't know*	***Categories:*** *1-no; 2-- Yes; 3- Disp prev record and has cond; 4- Disp prev record and no cond; 5- Disp prev record (DK if cond)*
Skin cancer exclusion*: reliance on type of cancer (cncrtype for 2009-2010) and ever being told to have skin cancer (chcscncr for 2011-2018)*	***Skin cancer exclusion****: reliance on having skin cancer questions (cnskdk cnsknm cnsoft): ever being told to have skin cancer, or non-melanoma, or soft tissue cancer)*	***Skin cancer exclusion****: Reliance on cancer type questions(mcq230a, mcq230b, mcq230c, mcq230d). Skin cancer categories-non-melanoma, soft tissue excluded*	***Skin cancer exclusion****: reliance on having skin cancer questions (cnskdk cnsknmag cnstom): ever being told to have skin cancer, or non-melanoma, or soft tissue cancer))*	***Skin cancer exclusion*** *is not feasible*
Obesity	**Variable/Question** *(diabete2)*: *Have you ever been told by a doctor that you have diabetes?*	**Variable/Question** *(diabeticev) : Ever told to have diabetes before*	**Variable/Question***(DIQ 010)*: *Doctor told you have diabetes?*	**Variable/Question** *(diabeticev) : Ever told to have diabetes before*	**Variable/Question** *(rdiab)* *if <u>new interview</u> "Ever told to have diabetes before?"* *If <u>reinterviewed and no condition</u> "Since we last talked to you, that is since [last interview date], has a doctor told you that you have"*

				if reinterviewed and had condition " "Our records from your last interview show that you have had"
Categories: 1- Yes; 2- Yes, but female told only during pregnancy; 3- No; 4- No, pre-diabetes or borderline diabetes; 7- Don't know; 9- Not Sure	**Categories:** 0-niu; 1- no or not mentioned; 2-- Yes; 3- Borderline; 7- Unknown- Refused; 8- Unknown- Not certained; 9- Unknown- Don't know	**Categories:**1-Yes; 2-No; 3-Borderline; 7- Refused; 9-Don't know; . Missing	**Categories** 0-niu; 1- no or not mentioned; 2-- Yes; 3- Borderline; 7- Unknown- Refused; 8- Unknown- Not certained; 9- Unknown- Don't know	**Categories:** 1-no; 2-- Yes; 3- Disp prev record and has cond; 4- Disp prev record and no cond; 5- Disp prev record (DK if cond)

Heart Disease	**Variable/Question** (CVDCRHD4): Has a doctor, nurse, or other health professional ever told you had angina or coronary heart disease?	**Variable/Question** (cheartdiev) : Ever told to have corona heart disease before?	**Variable/Question** (MCQ160c) - Ever told you had coronary heart disease	**Variable/Question** (cheartdiev): Ever told to have corona heart disease before	**Variable/Question** (rheart) *if new interview* "Ever told to have diabetes before?" If *reinterviewed and no condition* "Since we last talked to you, that is since [last interview date], has a doctor told you that you have" *if reinterviewed and had condition* " "Our records from your last interview show that you have had"
	Categories: 1-Yes; 2- No; 7- Don't know/Not Sure; 9- Refused	**Categories** 0-niu; 1- no or not mentioned; 2-- Yes; 3- Borderline; 7- Unknown- Refused; 8- Unknown- Not certain; 9- Unknown- Don't know	**Categories:**1 Yes; 2 No; 7 Refused; 9 Don't know; . Missing	**Categories:** 0-niu; 1- no or not mentioned; 2-- Yes; 3- Borderline; 7- Unknown- Refused; 8- Unknown- Not certain; 9- Unknown- Don't know	**Categories:** 1-no; 2-- Yes; 3- Disp prev record and has cond; 4- Disp prev record and no cond; 5- Disp prev record (DK if cond)

Variable/Question (rhibp)

Hyper-tension	Variable/Question: (BPHIGH4) Have you EVER been told by a doctor, nurse or other health professional that you have high blood pressure?	Variable/Question (hypertenev) : Ever told to hyper tension before	Variable/Question: BPQ020 : Ever told you had high blood pressure	Variable/Question (hypertenev) - Ever told to hyper tension before	if new interview "Ever told to have diabetes before?" If reinterviewed and no condition "Since we last talked to you, that is since [last interview date], has a doctor told you that you have" if reinterviewed and had condition " "Our records from your last interview show that you have had"
	Categories: 1-Yes; 2- Yes, but female told only during pregnancy; 3- No; 4- Told borderline high or pre-hypertensive; 7- Don´t know/Not Sure; 9- Refused	Categories: 0-niu; 1- no or not mentioned; 2-- Yes; 3- Borderline;7- Unknown- Refused; 8- Unknown- Not certain; 9- Unknown- Don't know	Categories: 1 Yes; 2 No; 7 Refused; 9 Don't know; . Missing	Categories: 0-niu; 1- no or not mentioned; 2-- Yes; 3- Borderline; 7- Unknown- Refused; 8- Unknown- Not certain; 9- Unknown- Don't know	Categories: 1-no; 2-- Yes; 3- Disp prev record and has cond; 4- Disp prev record and no cond; 5- Disp prev record (DK if cond)

Demographic Variable					
Age	1--85 & above	0-99	1--85 & above	0-99	50 and above
Race/ Ethnicity	race/ethnicity variable	From questions of race back ground and Hispanic origin	race/ethnicity variable	From questions of race back ground and Hispanic origin	From questions of race back ground and Hispanic origin
Sex	Male/Female	Male/Female	Male/Female	Male/Female	Male/Female
Other Notes					
Years Health	2009-2018	1997-2018	1999-2018 (2-year period)	2007-2017	1992-2018

Conditions Measure Available					
Geographic variable availability	State ID available in public use version	State or county ID can be obtained by applying to https://www.cdc.gov/rdc/B2AccessMod/ACs220.htm (to use at a RDC site)	State or county ID can be obtained by applying to https://www.cdc.gov/rdc/b1datatype/Dt1222.htm	State or county ID can be obtained by applying to https://www.cdc.gov/rdc/B2AccessMod/ACs220.htm (to use at a RDC site)	State or county ID can be obtained by applying to https://meps.ahrq.gov/mepsweb/data_stats/onsite_datacenter.jsp?# (to use at a RDC site)
Sample Size	Approximately 450,000 individuals/year	Approximately 100,000 individuals / year	Approximately 10,000 individuals / year	Approximately 35,000 individuals / year	Approximately 42,000 individuals / year

C. Stata Code

Most of the files downloaded are in SAS format. This section introduces the Stata codes to convert them in to Stata file

In Stata, use *fdause* command to convert a SAS file to a DTA file.

The code below gives an example of how to use the command to convert .xpt files from NHANES data in to .dta files

Details as follows

```
* Author: Cong GIan (6/15/20)
** Converting SAS file to Stata file
** Change local macro to include other periods after 1999-2000
local yr "1999_2000"
    foreach yr in `yr' {
       dis "`yr'"
    *Change the current directory to the folder in which the .xpt files are located
            cd "I:\NHANES\Raw Data\\`yr'"
            local filelist: dir . files "*.xpt"
            dis `filelist'
}
local yr "1999_2000"
foreach yr in `yr' {
    foreach file in `filelist' {
            clear
            fdause `file'
            local filename = substr("`file'", 1, strpos("`file'", ".") -1)
            save "I:\NHANES\Gen Data\\`yr'\\`filename'.dta", replace
            }
    }
```

D. GitHub Storage and Sharing of Codes and Documents

Link https://github.com/CongGian/racehealthdatasets

(Snap shot is below)