This readme file was generated on 2024-08-05 by Julie Fennell

GENERAL INFORMATION

Title of Dataset: NHIS (National Health Interview Survey), as processed by IPUMS

Abstract: For our analysis, we will be using multiple waves of the National Health Interview Survey (NHIS), which is a free and publicly available nationally representative cross-sectional survey of the non-institutionalized U.S. adult population up to age 85 (https://nhis.ipums.org/). The average annual sample is around 100,000 people in 45,000 households. Key dependent variables are DEMENTIAEV ("have you ever been told by a health professional that you have dementia, including alzheimer's") from 2019-22; LAMEMORCON ("have you ever had difficulty remembering, concentrating, or both?"), which can be focused solely on respondents who had difficulty remembering, from 2010-22; LAMEMCONDIF, which asks for level of difficulty remembering for the same years; LAMEMDIFOFT, which asks for frequency of difficulty remembering for the same years. We believe using these broader measures of cognitive decline will help capture people all along the dementia spectrum. One of the common problems in previous studies of dementia has been attempting to study this condition as a black-and-white diagnosis when it is clearly a spectrum disorder.

The NHIS contains a vast array of potential independent variables pertaining to both demographics and health that can be used to look for predictive factors associated with cognitive decline, from cardiovascular issues to mental health and even to the largely under-researched variable of sexual orientation. These questions address physical and psychological health, with many questions on health behaviors as well. With previous analyses suggesting that factors associated with dementia have also shifted across time, having more than a decade of data will allow us to analyze the potential effects of time, alone and in combination with many other key predictor variables.

Author/Principal Investigator Information

Name: Julie Fennell

ORCID: 0009-0001-9500-2898

Date of data collection: 2010-2022

Geographic location of data collection: Centers for Disease Control (CDC) collected nationally representative data on the non-institutionalized population of the United States.

Information about funding sources that supported the collection of the data: The CDC collects this data with funding and mandate from the U.S. government.

ABSTRACT:

The National Health Interview Survey (NHIS) is a free and publicly available data set sponsored by the Centers for Disease Control (CDC) and conducted by the United States Census Bureau annually since 1957. Relevant questions for the purposes of studying cognitive decline have been asked every year since 2010, resulting in 12 years of data for key dependent variables. The survey is conducted throughout the year to minimize seasonal bias, with computer assisted face-to-face interviewing occurring whenever possible (although at least a third of respondents are usually contacted by phone). As a major Census Bureau project, it is a gold standard of representation for the U.S. non-institutionalized population up to age 85 (excluding current prisoners, wholly military families, and most importantly for these purposes, those in residential homes). It utilizes a complex stratification sampling system (which has been simplified since 2019), requiring the use of sample weights to achieve representative estimates of the population. Prior to 2019, information was collected from one household adult on all members of the household; after 2019, it solely focuses on the queried adult and one child in the household if present (although children are irrelevant for this analysis).

IPUMS (the Integrated Public Use Microdata Series) has "harmonized" the data across years, making it possible to easily download manageable data sets and compare questions and answers across time. Sample sizes vary greatly across survey waves. During the years for which good data on cognitive decline are available, sample size varies greatly from a low of 28,854 households/35,115 individuals in 2022 to a high of 60,347 households/112,053 individuals in 2014 for the total sample. However, from 2010-2018, only a quarter subsample was asked the relevant questions about cognitive decline, and specific sampling details for those questions are described below.

BACKGROUND:

https://nhis.ipums.org/nhis/userNotes_sampledesign.shtml

The NHIS design is complex and has changed across the major years of interest. Although it has remained large and nationally representative throughout, before 2016 it employed oversamples by race and ethnicity, which require the use of sampling weights for analysis. The nature of the oversample has changed across time, becoming increasingly simple, although weights are still required.

The greatest challenge for the survey collection was during the COVID-19 pandemic, and the survey was forced to simplify its methods considerably in order to continue throughout the 2020 survey year. Many respondents were repeated in the 2020 sample from the 2019 sample, and nearly all surveys were administered over the phone, whereas previous years were often administered in person.

METHODS:

https://www.cdc.gov/nchs/nhis/about_nhis.htm
https://www.cdc.gov/nchs/nhis/about_nhis.htm#procedures
https://nhis.ipums.org/nhis/userNotes_sampledesign.shtml
https://nhis.ipums.org/nhis-action/faq#ques0
https://nhis.ipums.org/nhis/userNotes_size.shtml

The NHIS is a major government survey, and as such, undergoes many standard procedures to collect representative data which are then stripped of key identifying information before becoming available to the public.

For many years, the Minnesota Population Center at the University of Minnesota has worked to make the NHIS data clean and easily available across the survey's many years. This data harmonization project includes recoding different questions about race and education so that they can be used in analyses across time even when the exact questions and answers have changed (while also preserving the original unique variables). The grant that fuels this project ensures that the data are freely and easily publicly available through a single website that allows users to access hundreds of variables and many hundreds of thousands of cases: https://nhis.ipums.org

On the website, users should narrow the selection of variables to the relevant years for this analysis, from 2010-22, with a focus on 2019-22. After selecting the variables they want (which could potentially be all of them), they should then select the relevant sample years. Data can be exported as .dat, .dta (Stata), .sav (SPSS), .sas (SAS), and .csv.

To access our recoded file, users should download the file from our [website]. All variables have been logically recoded based on a combination of logic and prevalence. That is, to preserve case counts, our recoding system usually creates a category for "unknowns" that includes all respondents with don't know, not ascertained, and refused responses. Exceptions are for variables with basically complete response rates.

All of our data recoding was conducted using SPSS (.sav), and all other data forms (.dta, .sas, .csv) available from our website have been exported from SPSS. *We have not been able to test the exported data, and we encourage users to double check a few of our recoded variables against the raw data from IPUMS/NHIS as SPSS conversions are sometimes imperfect.*

The attached "Recommended variables" document contains some information on the original variables from IPUMS/NHIS and more detailed information on the provided recoded variables. Case counts for each response are provided so that users can double-check their tallies against our own when using converted data files.

The chief individual sampling weight variable is PERWEIGHT.

The NHIS represents population survey data, and as such, there are no relevant experimental conditions.

As a gold-standard national government survey, the NHIS undergoes many quality checks upon its initial completion, and then many further checks upon its integration with existing data sources at the Minnesota Population Center files.

DATA DESCRIPTION:

Whereas so many data sets try to predict whether a person "has dementia" based on much-debated criteria about what "counts" as dementia, this data set allows for much more nuanced calculations based on scaled experiences of difficulty remembering, in addition to a direct measure of whether the respondent was diagnosed with dementia. The NHIS allows analysts to construct a more sophisticated scaled measure of subjective cognitive decline with its available measures.

Many studies and reviews have attempted to look at a wide variety of factors in predicting who will get dementia, but very few data sets have the vast sample size and variable set of the NHIS. With such large sample sizes, the experiences of merely 1% of people become possible to mathematically predict, especially with pooled data. As some of these questions have been asked since 2010, we can use them to look at potentially changing trends across time.

The NHIS is a nationally representative data set which until 2018 intentionally oversampled some racial and ethnic minorities, necessitating the use of sampling weights to produce population estimates for data analysis. However, prior research suggests that most effects of race and ethnicity on dementia can be controlled for with several factors that the NHIS measures: marriage, diabetes, and blood pressure (unfortunately, the NHIS does not have many measures for social contact for more than a scant handful of years, which is another relevant variable here). The NHIS also measures foreign-born status, age, and education, which previous research suggests are the three most significant demographic factors in predicting cognitive decline. Moreover, since 2013, the NHIS contains data on sexual orientation, which is a category that has been largely unexplored in analyses of ADRD to date, despite speculative hints in the literature that bisexuals might be at particular risk.

More detailed information on the recommended and recoded dependent and independent variables can be found in the "Recommended Variables" document.

All original files can be accessed from https://nhis.ipums.org. Recoded data with recommended variables can be downloaded from [our website].

IPUMS periodically updates its files across the year to add new data. As of this writing in April 2024, it is my understanding that the 2022 files are still getting new variables added to them. Additional updates from 2023 should begin soon.

RELEASE NOTES:

As previously mentioned, the samples for the NHIS have changed considerably over time. Most importantly, the sample changed in 2019 from asking a single householder about all members of the household to a single householder being asked about themselves and one focal child if relevant. The sample universes for three of our key dependent variables of interest shifts over time, from a quarter subsample prior to 2019, to the entire sample from 2019 on.

ETHICS:

Detailed information on the ethics and government authorization for the NHIS can be found at:

https://www.cdc.gov/nchs/nhis/participants/yourprivacy.htm

https://www.cdc.gov/nchs/nhis/participant.htm

CONFLICTS OF INTEREST:

No conflicts to declare.

REFERENCES:

Links to all essential information about the original data can be found at:

https://www.cdc.gov/nchs/nhis/index.htm

https://nhis.ipums.org/nhis/

ACCESS:

Access to the original data is free and available to the public following completion of a short data application:

https://uma.pop.umn.edu/nhis/user/new?return_url=https%3A%2F%2Fnhis.ipums.org%2Fnhis-action%2Fmenu

Access to our recoded data will be available from our [website].

Licenses/restrictions placed on the data:

In order to access the data, users must promise not to redistribute the data without permission, use the data for statistical analysis and reporting only, not seek to identify individuals, and that they will cite the data appropriately in publications. This restriction applies to both the original data and to our recoded data.

We were granted official permission on April 16, 2024 from Kari Williams to redistribute the data specifically for this project, subject to the aforementioned conditions.

Links to publications that cite or use the data:

https://thescholarship.ecu.edu/bitstream/handle/10342/9937/LuoSelfReportedCognitiveImpairmentAcrossRacialEthnicGroupsintheUnitedStates.pdf?sequence=1

https://academic.oup.com/innovateage/article/5/4/igab039/6375379

https://doi.org/10.1177/08982643211046466

https://doi.org/10.1044/2022_AJA-22-00087

Links to other publicly accessible locations of the data: N/A

Links/relationships to ancillary data sets: With considerable effort, IPUMS data can be connected to NHIS data that have not yet been harmonized. However, there is a sufficiently large data set available that such effort is largely unnecessary.

Was data derived from another source? No

If yes, list source(s):

Recommended citation for this dataset:

Lynn A. Blewett, Julia A. Rivera Drew, Miriam L. King, Kari C.W. Williams, Annie Chen, Stephanie Richards, and Michael Westberry. IPUMS Health Surveys: National Health Interview Survey,

Version 7.3 [dataset]. Minneapolis, MN: IPUMS, 2023. https://doi.org/10.18128/D070.V7.3
https://www.nhis.ipums.org

VERSIONS:

A detailed revision history is available at: https://nhis.ipums.org/nhis-action/revisions

FILES:

Readme.doc

Recommended Variables.doc

PREPARE draft.sav [which will also be available as .csv, .dta, and .sas]

Number of variables:

The number of variables in the original NHIS changes year to year as topics are rotated across time. A summary of that rotation can be seen in the chart on: https://nhis.ipums.org/nhis/userNotes_2019_NHIS_Redesign.shtml

In our recoded sample extract, there are 92 variables (although that is only because SPSS automatically generated a filter variable around sex, so depending on the data conversion process, there might only be 81). 39 of the variables in our set are recodes (including the filter variable), and the remainder are original IPUMS NHIS variables.

Number of cases/rows/samples:

Within our recoded data set, there are 1,019,882 rows/cases.

For the NHIS as a whole, the sample size varies considerably by year. However, it is important to note that the relevant questions pertaining to memory were only asked of adults, and children should be excluded from all raw analyses (and automatically are in our recoded data set). Case counts for each recoded variable are in the Recommended Variables document.

Unweighted, the raw pooled sample for 2010-2022 that has valid responses for the central question LAMEMCONDIF is distributed thus along major independent variables:
Education: Not completed high school--28,791 (11.4%), High school graduate--56,930 (22.3%), Some college--82,501 (32.3%), College--53,063 (20.8%), Graduate or professional degree--32,699 (12.8%)
Nativity: Born in US—209,510 (82.1%), born outside US--41,896 (16.4%), unknown--3,536 (1.4%)
Hispanic ethnicity: Not Hispanic--218,778 (85.7%), Hispanic--36,375 (14.3%)
Race: White--195,185 (76.5%), Black--31,340 (12.3%), Asian--14,592 (5.7%), Indigenous only--2,476 (1.0%), Other and multiple races--5,616 (2.2%), Unknown--5,838 (2.3%)
Sex: Male--115,761 (45.4%), Female--139,382 (54.6%)
Age: 18-30--47,216 (18.5%), 31-40--41,267 (16.2%), 41-50--38,521 (15.1%), 51-60--43,823 (17.2%), 61-70--43,429 (17.0%), 71-80--27,036 (10.6%), 81-85--13,565 (5.3%)

With such a large sample size, it is possible to use pooled NHIS data to analyze extremely small and understudied subgroups, such as immigrants and Indigenous people. With over 13,000 people aged 81-85, it is even possible to simply focus on trends in this older age group, or on the

27,000 respondents aged 71-80, who presumably are the group to target for earlier detection of AD/ADRD.

Variable List:

The NHIS contains 4 key dependent variables measuring cognitive decline in slightly different ways:

DEMENTIAEV ("have you ever been told by a health professional that you have dementia, including alzheimer's") from 2019-22 and 2007
https://nhis.ipums.org/nhis-action/variables/DEMENTIAEV#survey_text_section
Sampling universe was about 30,000 cases annually from 2019-22, with around 350 people answering "yes" each year.

LAMEMCONDIF ("Do you have difficulty remembering or concentrating?"--no, some, a lot, or cannot do at all) from 2010-22. From 2010-17, this question was asked to a quarter subsample, and from 2018-2022 was asked of the full sample.
https://nhis.ipums.org/nhis-action/variables/LAMEMCONDIF#survey_text_section
The universes for the variable vary greatly by year, but the average annual number of cases for "a lot" is more than 700 in 2018-22 with the full sample, and more than 300 in 2010-2017 with the quarter-sample. Less than 20 people report annually they cannot remember or concentrate at all.

LAMEMORCON ("have you ever had difficulty remembering, concentrating, or both?"), which can be focused solely on respondents who had difficulty remembering, from 2010-22. This variable sorts out responses from the previous variable to clarify whether respondents specifically have difficulty remembering, *or* concentrating, *or* both.
https://nhis.ipums.org/nhis-action/variables/LAMEMORCON#description_section
The universes for this variable vary greatly by year, but the average number of cases reporting "difficulty remembering only" is well over 1,000 annually, and more than 1,600 for "difficulty remembering and concentrating."

LAMEMDIFOFT ("how often would you say you have difficulty remembering?"--sometimes, often, or all the time) which focuses on respondents who reported difficulty remembering, from 2010-22.
https://nhis.ipums.org/nhis-action/variables/LAMEMDIFOFT#survey_text_section
The sampling universes for this variable vary greatly by year, and the response numbers consequently vary. About 1,300 people annually report "frequently" plus "all the time" in 2018-22, and more than 600 report annually for the same categories in 2010-17.

In terms of reliability and validity, pooling the samples for the years 2019-22 (when the question on dementia diagnosis was asked), among those who say they have never been diagnosed with dementia, 84% report no difficulty remembering, 12% say they have difficulty sometimes, 3% frequently, and 1% all the time. Among those who say they have been diagnosed with dementia, 16% report no difficulty remembering, 27% say they have difficulty sometimes, 26% frequently, and 31% all the time. (Note that sample sizes here are sufficiently large that the 1% of people who say they have never been diagnosed with dementia but have difficulty remembering "all the time" still constitutes 1399 cases). Of course, people may have difficulty remembering for a variety of causes (and people with dementia may, by definition, be unreliable reporters of their own experience of memory). Regardless, concordance between these variables appears to be about 84% in both directions, and it appears that using both measures separately and together should maximize the predictive power of models.

In terms of potential independent variables for technology to scour within this vast data set, there are classic socioeconomic indicators such as race, Hispanic ethnicity, whether the respondent is foreign-born; age, sex, sexual orientation, current marital and cohabiting status; whether the respondent is a veteran; educational attainment; whether the respondent ever worked or the last time they worked for pay; whether the family owns or rents their home; poverty line ratio, as well as measures of food insecurity; overall health status and BMI; medical history of angina, asthma, any cancer, coronary heart disease, high cholesterol, diabetes, hypertension, stroke, and major breathing obstructions; alcohol consumption, smoking, and physical activity; functional limitations, such as difficulty washing, dressing, walking, seeing, or hearing; chronic pain; access to medical care and insurance; receipt of some vaccinations; and simple measures for anxiety and depression.

Missing data codes: For most of the dependent variables of interest, the important codes in the original data are 0 for not in universe (NIU), 7 for unknown-refused, 8 for unknown-not ascertained, and 9 for unknown-don't know.

In our recoded variables, missing data are mostly coded as .

Specialized formats or other abbreviations used: N/A

.