

National Health and Nutrition Examination Survey

2017-March 2020 Data Documentation, Codebook, and Frequencies

Demographic Variables and Sample Weights (P_DEMO)

Data File: P_DEMO.xpt

First Published: May 2021

Last Revised: NA

Component Description

The NHANES program suspended field operations in March 2020 due to the coronavirus disease 2019 (COVID-19) pandemic. As a result, data collection for the NHANES 2019-2020 cycle was not completed and the collected data are not nationally representative. Therefore, data collected from 2019 to March 2020 were combined with data from the NHANES 2017-2018 cycle to form a nationally representative sample of NHANES 2017-March 2020 pre-pandemic data. These data are available to the public. Please refer to the Analytic Notes section for more details on the use of the data.

This demographics file provides information on the following topics:

- Survey participant's household interview and examination status;
- Interview and examination sample weights;
- Masked variance units;
- Language of questionnaires used for the interviews conducted in the household and in the mobile examination center;
- Use of proxy or interpreter during the interviews;
- The six-month time period when the examination was performed;
- Pregnancy status;
- The ratio of family income to poverty guidelines; and
- Other selected demographic information, such as gender, age, race/Hispanic origin, education, marital status, country of birth, and years of U.S. residence.

Eligible Sample

All participants in the NHANES 2017-March 2020 pre-pandemic sample are included in this dataset. The target age group for individual demographic variables in this file varies by the topic. Please review the codebook carefully.

Interview Setting and Mode of Administration

The family and sample person demographics questionnaires were asked, in the home, by trained interviewers using Computer-Assisted Personal Interview (CAPI) system. The respondent selected the language of interview (English or Spanish) or requested that an interpreter be used. Hand cards, showing response choices or information that survey participants needed to answer the questions, were used for some questions. The hand cards were printed in English, Spanish, Mandarin Chinese (both traditional and simplified), Korean, and Vietnamese. The interviewer directed the respondent to the appropriate hand card during the interview. When necessary, the interviewer further assisted the respondent by reading the response choices listed on the hand cards.

Persons 16 years and older and emancipated minors were interviewed directly. A proxy provided information for survey participants who were under 16 and for participants who could not answer the questions themselves.

The NHANES [2017-2018](#) and [2019-2020](#) demographics questionnaires are available on the NHANES website.

Quality Assurance & Quality Control

The CAPI system is programmed with built-in consistency checks to reduce data entry errors. CAPI also uses online help screens to assist interviewers in defining key terms used in the questionnaire.

After collection, interview data were reviewed by the NHANES field office staff for accuracy and completeness of selected items. The interviewers were required to audio record interviews and the recorded interviews were reviewed by NCHS staff and interviewer supervisors.

Data Processing and Editing

Frequency counts were checked, "skip" patterns were verified, and the reasonableness of question responses was reviewed. Edits were made to some variables to ensure the completeness, consistency, and analytic usefulness of the data. Edits were also made, when necessary, to address data disclosure concerns.

SDDSRVYR: This variable identifies the specific data release cycle. A value of "66" denotes NHANES 2017–March 2020 pre-pandemic data.

RIDSTATR: This variable is used to identify whether a participant was both interviewed at home and examined in the mobile examination center (MEC) or was only interviewed in the home but never went through the examination.

RIDAGEYR: Age in years, at the time of the screening interview, is reported for survey participants between the ages of 1 and 79 years of age. Due to disclosure concern, all responses of participants aged 80 years and older are coded as '80.' In NHANES 2017-March 2020 pre-pandemic sample, the weighted mean age for participants 80 years and older is 85 years.

RIDAGEYR was calculated based on the participant's date of birth. In rare cases, if the actual date of birth was missing but the participant's age in years was provided, the reported age was used.

RIDAGEMN: Age in months, at the time of the screening interview, is provided for participants who were less than 25 months of age at the time of examination (RIDEXAGM < 25). If the exact date of birth was not provided by the respondent, the age in months was calculated based on the imputed age in years at the time of the screening interview.

RIDEXMON: This variable indicates the six-month time period when the examination was performed. A value of "1" indicates November 1st through April 30th; a value of "2" indicates May 1st through October 31st.

RIDRETH3: This is the race-ethnicity variable included in the demographics file since the 2011-2012 survey cycle to accommodate the oversample of Asian Americans. It was derived from responses to the survey questions on race and Hispanic origin. Respondents who self-identified as "Mexican American" were coded as such (i.e., RIDRETH3=1) regardless of their other race-ethnicity identities. Otherwise, self-identified "Hispanic" ethnicity would result in code "2, Other Hispanic" in the RIDRETH3 variable. All other non-Hispanic participants would then be categorized based on their self-reported races: non-Hispanic white (RIDRETH3=3), non-Hispanic black (RIDRETH3=4), non-Hispanic Asian (RIDRETH3=6), and other non-Hispanic races including non-Hispanic multiracial (RIDRETH3=7). Code "5" was not used in RIDRETH3.

RIDRETH1: This is the race-ethnicity variable that can be linked to the previous NHANES race-ethnicity variable in 1999-2010. Non-Hispanic Asian participants are grouped with other non-Hispanic races in code "5" (other non-Hispanic race including non-Hispanic multiracial) in RIDRETH1. Codes "6" and "7" were not used in RIDRETH1. Coding procedure for other categories in RIDRETH1 was compatible to RIDRETH3.

DMDBORN4: Due to the concerns of disclosure risk, starting in 2011, country of birth was recoded into two categories: 1) Born in 50 U.S. states or Washington, DC; and 2) Born in other countries, including U.S. territories.

DMDMARTZ: The marital status question was asked of persons 14 years of age and older. Due to disclosure risks, marital status is only released for persons 20 years of age and older, and recoded from the original 6 categories to 3 categories (1 = Married/Living with partner; 2 = Widowed/Divorced/Separated; 3 = Never married).

RIDEXPRG: Pregnancy status at the time of the health examination was ascertained for females 8–59 years of age. Due to disclosure risks pregnancy status is only released for women 20–44 years of age. The information used to code RIDEXPRG values included self-reported pregnancy status and urine pregnancy test results. Persons who reported they were pregnant at the time of exam were assumed to be pregnant (RIDEXPRG=1). Those who reported they were not pregnant or did not know their pregnancy status were further classified based on the results of the urine pregnancy test. If the respondent reported "no" or "don't know" and the urine test result was positive, the respondent was coded as pregnant (RIDEXPRG=1). If the respondent reported "no" and the urine test was negative, the respondent was coded not pregnant (RIDEXPRG=2). If the respondent reported did not know her pregnancy status and the urine test was negative, the respondent was coded "could not be determined" (RIDEXPRG=3). Persons who were interviewed, but not examined also have an RIDEXPRG value = 3 (could not be determined).

DMDYRUSZ: This variable is the number of years the participant has lived in the United States. Participants who were born outside the U.S. were asked the month and year when they came to the U.S. to live (DMQ.160). The responses to the question were used to calculate the length of these participants lived in the US and then recoded into 4 categories: 1 = Less than 5 year, 2 = 5 year or more, but less than 15 years, 3 = 15 year or more, but less than 30 years, and 4 = 30 year or more. For participants who did not report the month of their arrival, a month value of 7 (July) was imputed in the calculation of the residence length.

DMDEDUC2: This variable is the highest grade or level of education completed by adults 20 years and older.

The response categories are: less than 9th grade education, 9-11th grade education (includes 12th grade and no diploma), High school graduate/GED, some college or associates (AA) degree, and college graduate or higher.

SIALANG: This variable indicates the language (English or Spanish) used during the sample person questionnaire interview conducted at the participant's home.

SIAPROXY: This variable denotes whether a proxy respondent was used during the sample person questionnaire interview.

SIAINTRP: This variable denotes whether an interpreter was used during the sample person questionnaire interview. The language spoken by the respondent is only available through the [NCHS RDC](#).

FIALANG: This variable indicates the language used during the family questionnaire interview conducted at the participant's home.

FIAPROXY: This variable denotes whether a proxy respondent was used to complete the family questionnaire interview.

FIAINTRP: This variable denotes whether an interpreter was used to complete the family questionnaire interview. The language spoken by the respondent is only available through the [NCHS RDC](#).

MIALANG: This variable indicates the language (English or Spanish) used for the CAPI portion of the MEC interview.

MIAPROXY: This variable denotes whether a proxy respondent was used during the CAPI portion of the MEC interview.

MIAINTRP: This variable denotes whether an interpreter was used during the CAPI portion of the MEC interview. The language spoken by the respondent is only available through the [NCHS RDC](#).

AIALANGZ: This variable indicates the language used for the audio-computer-assisted self-interview (ACASI) portion of the MEC interview. The ACASI portion was translated into Chinese (traditional/Mandarin, simplified/Mandarin, and traditional/Cantonese), Korean, and Vietnamese to accommodate the Asian oversampling as of 2011. The three categories reported are: 1) English, 2) Spanish, and 3) Asian languages.

INDFMPIR: This variable is the ratio of family income to poverty. The Department of Health and Human Services (HHS) poverty guidelines were used as the poverty measure to calculate this ratio. These guidelines are issued each year, in the Federal Register, for determining financial eligibility for certain federal programs, such as Head Start, Supplemental Nutrition Assistance Program (SNAP), Special Supplemental Nutrition Program for Women, Infants, and Children (WIC), and the National School Lunch Program. The poverty guidelines vary by family size and geographic location (with different guidelines for the 48 contiguous states and the District of Columbia; Alaska; and Hawaii).

INDFMPIR was calculated by dividing total annual family (or individual) income by the poverty guidelines specific to the survey year.

During the household interview, the respondent was asked to report total income for the entire family in the last calendar year in dollars. A family is defined as a group of two people or more related by birth, marriage, or adoption and residing together. Annual individual income was asked for households with one person or households comprised of unrelated individuals.

If the respondent was not willing or able to provide an exact dollar figure, the interviewer asked an additional question to determine whether the income was < \$20,000 or ≥ \$20,000. Based on the respondent's answer to this question, he/she was asked to select a category of income from a list on a hand card. For respondents who selected a category of income, their family incomes were set as the midpoints of the selected ranges. INDFMPIR was not computed if the respondent only reported income as < \$20,000 or ≥ \$20,000 but no additional details provided. INDFMPIR values at or above 5.00 were coded as 5.00 or more because of disclosure concerns. The values were not computed if the income data was missing.

Analytic Notes

The COVID-19 pandemic required suspension of NHANES 2019-2020 field operations in March 2020 after data were collected in 18 of the 30 survey locations in the 2019-2020 sample. Data collection was cancelled for the remaining 12 locations. Because the collected data from 18 locations were not nationally representative, these data were combined with data from the previous cycle (2017-2018) to create a 2017-March 2020 pre-pandemic data file. A special weighting process was applied to the 2017-March 2020 pre-pandemic data file. The resulting sample weights in the present file should be used to calculate estimates from the combined cycles. These sample weights are not appropriate for independent analyses of the 2019-2020 data and will not yield nationally representative results for either the 2017-2018 data alone or the 2019-March 2020 data alone. Please refer to the NHANES website for additional information for the NHANES 2017-March 2020 pre-pandemic data, and for the previous 2017-2018 public use data file with specific weights for that 2-year cycle.

Age at screening: Age at screening was used to determine eligibility for an examination component and should be used for most analyses.

DMDMARTZ: Marital status is only released for persons 20 years of age and older because of potential disclosure risks. Prior to 2007, marital status was released for participants aged 14 and older. In NHANES 2017-March 2020 pre-pandemic sample, the percentage of persons aged 14-19 who are married or living with a partner is less than 2.5%.

RIDEXPRG: Because of possible disclosure risks, pregnancy status is only released for women aged 20-44 years. The percentage of pregnant women/girls aged 8-19 or 45-59 years is less than 1% in the 2017-March 2020 pre-pandemic sample.

Masked Variance Units (MVUs): Twenty-four masked variance strata and 49 masked primary sampling units (PSUs) are included in the NHANES 2017-March 2020 pre-pandemic demographics file. Each stratum has 2-3 PSUs. These MVUs are a collection of secondary sampling units that are aggregated into groups for the purpose of variance estimation. The variance estimates that are produced, using the MVUs, closely approximate the variances that would have been estimated using the “true” sample design variance units that are based on the actual survey sample strata and primary sampling units. MVUs are used to protect the confidentiality of information provided by survey participants and to reduce disclosure risks. The use of MVUs is described in the [NHANES Analytic Guidelines](#). Analysts should review the Guidelines carefully prior to analyzing the survey data.

Sample Weights: The special sample weights (WTINTPRP, WTMECPRP) should be used for all NHANES 2017-March 2020 pre-pandemic data analyses. Detailed instructions for combining datasets from previous NHANES cycles are provided in the [NHANES Analytic Guidelines](#).

Please also refer to the [NHANES Analytic Guidelines](#) and the on-line [NHANES Tutorial](#) for further details on the use of sample weights and other analytic issues.

Disclosure risks and issues pertaining to confidentiality protection prevent NCHS from releasing all of the NHANES demographic variables publicly. Additional information and instructions for requesting use of these data are available through the [NCHS RDC](#).

References

- U.S. Department of Health & Human Services. Poverty Guidelines, Research, and Measurement. Washington, DC: U.S. Department of Health & Human Services. <https://aspe.hhs.gov/prior-hhs-poverty-guidelines-and-federal-register-references>

Codebook and Frequencies

SEQN - Respondent sequence number

Variable Name:	SEQN
SAS Label:	Respondent sequence number
English Text:	Respondent sequence number.
Target:	Both males and females 0 YEARS - 150 YEARS

RIAGENDR - Gender

Variable Name: RIAGENDR
SAS Label: Gender
English Text: Gender of the participant.
Target: Both males and females 0 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Male	7721	7721	
2	Female	7839	15560	
.	Missing	0	15560	

RIDAGEYR - Age in years at screening

Variable Name: RIDAGEYR

SAS Label: Age in years at screening

English Text: Age in years of the participant at the time of screening. Individuals 80 and over are topcoded at 80 years of age.

Target: Both males and females 0 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
0 to 79	Range of Values	14878	14878	
80	80 years of age and over	682	15560	
.	Missing	0	15560	

RIDRETH1 - Race/Hispanic origin

Variable Name: RIDRETH1
SAS Label: Race/Hispanic origin
English Text: Recode of reported race and Hispanic origin information
Target: Both males and females 0 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Mexican American	1990	1990	
2	Other Hispanic	1544	3534	
3	Non-Hispanic White	5271	8805	
4	Non-Hispanic Black	4098	12903	
5	Other Race - Including Multi-Racial	2657	15560	
.	Missing	0	15560	