Tetherless World Constellation

SciKG Part 1: Studies, Data, and Documentation

Henrique Santos, Paulo Pinheiro, Jamie P. McCusker, Sabbir M. Rashid, Deborah L. McGuinness

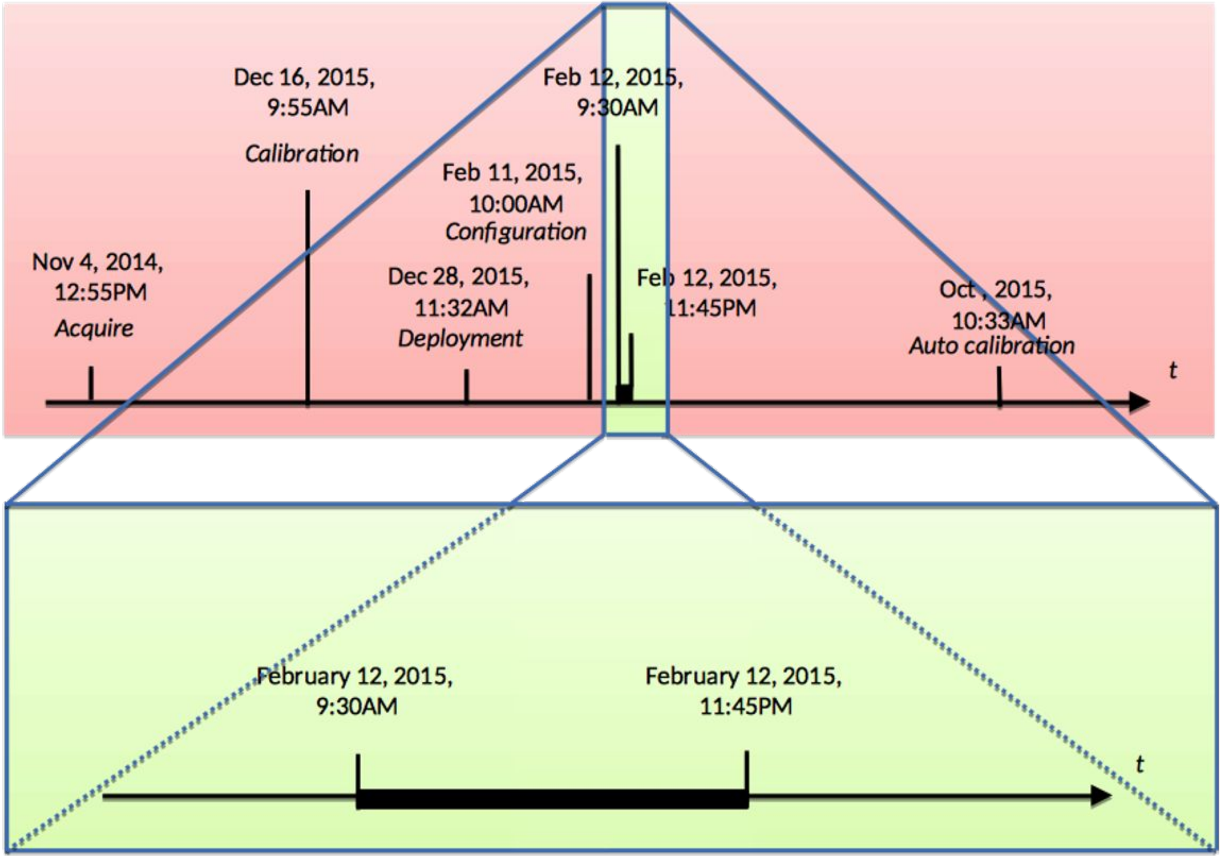
May 28th 2023

The 20th Extended Semantic Web Conference (ESWC-23)

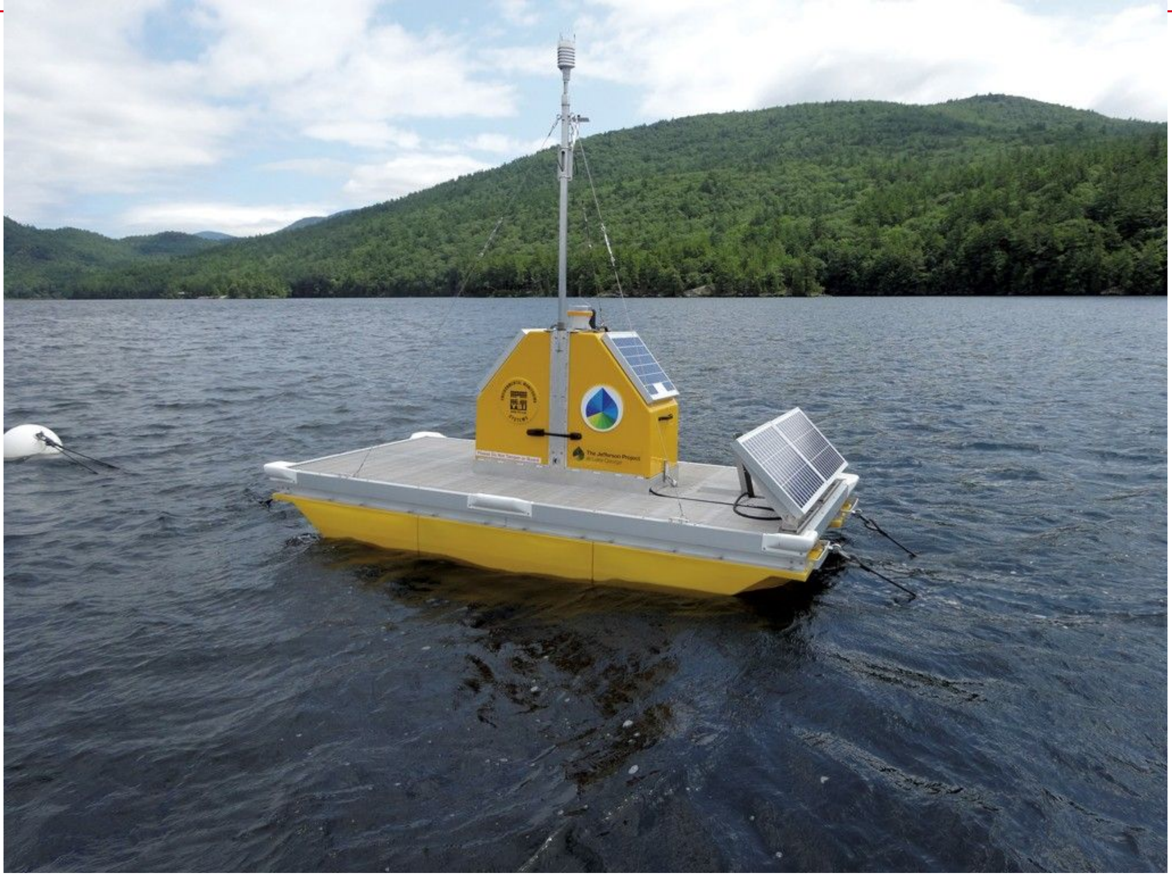
Part 1: Studies, Data, and Documentation

- Scientific studies and their data acquisition activities
- Scientific data organization
- Scientific data publishing
- Documentation
  - Data dictionaries
  - Codebooks
  - Methods
- National Health and Nutrition Examination Surveys (NHANES)
  - Semantics of NHANES data

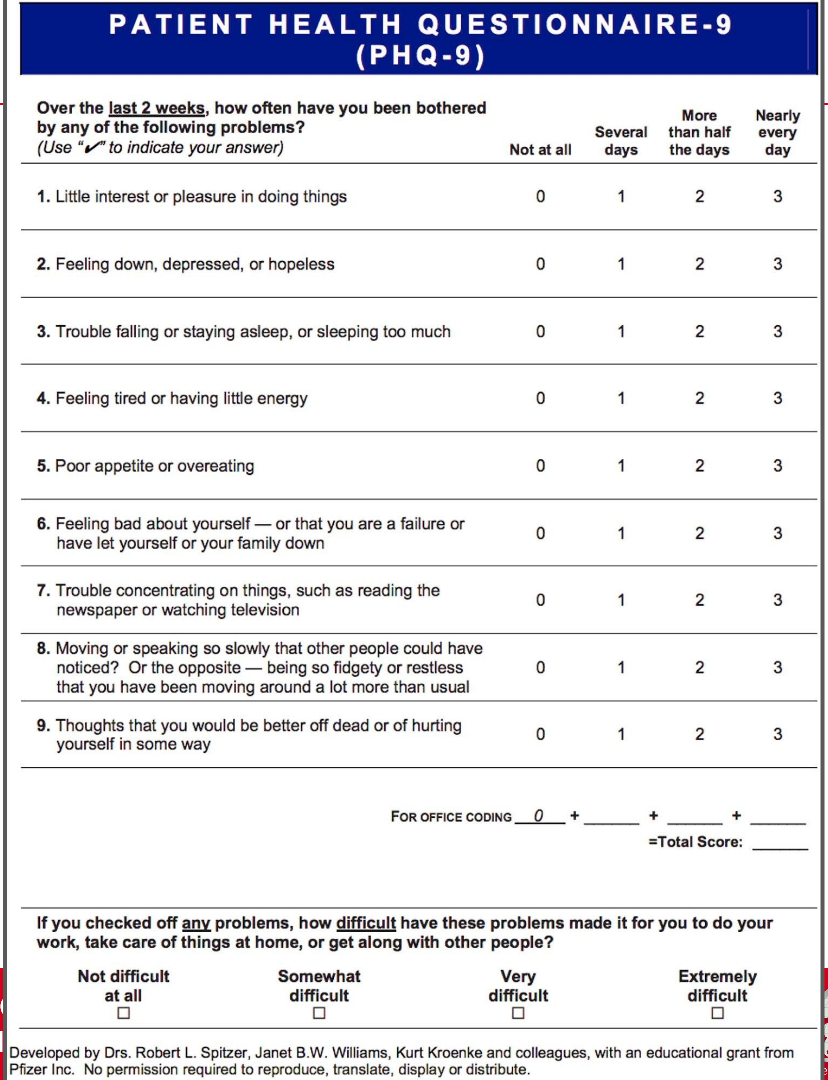
Data Acquisition context



Platforms, Instruments and Detectors



Platforms, Instruments and Detectors



PATIENT HEALTH QUESTIONNAIRE-9 (PHQ-9)

Over the last 2 weeks, how often have you been bothered by any of the following problems? *(Use "✓" to indicate your answer)*	**Not at all**	**Several days**	**More than half the days**	**Nearly every day**
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself — or that you are a failure or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed? Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
9. Thoughts that you would be better off dead or of hurting yourself in some way	0	1	2	3

FOR OFFICE CODING __0__ + ______ + ______ + ______

=Total Score: ______

If you checked off any problems, how difficult have these problems made it for you to do your work, take care of things at home, or get along with other people?

Not difficult at all	**Somewhat difficult**	**Very difficult**	**Extremely difficult**
□	□	□	□

Developed by Drs. Robert L. Spitzer, Janet B.W. Williams, Kurt Kroenke and colleagues, with an educational grant from Pfizer Inc. No permission required to reproduce, translate, display or distribute.

Platforms, Instruments and Detectors

- **Platform**: an object that keeps the instrument in a specific location to ensure that it is recording data about the selected location
- **Instrument**: an object that receives sensed signals from detectors and processes these signals into numerical values
- **Detector**: An object that it is capable of sensing environmental properties by collecting physical signals about these properties, translating these physical signals into (most often electrical) signals, and forwarding these electrical signals to instruments

Data dictionaries

"*A data dictionary is used to catalog and communicate the structure and content of data, and provides meaningful descriptions for individually named data objects.*" (USGS.gov)

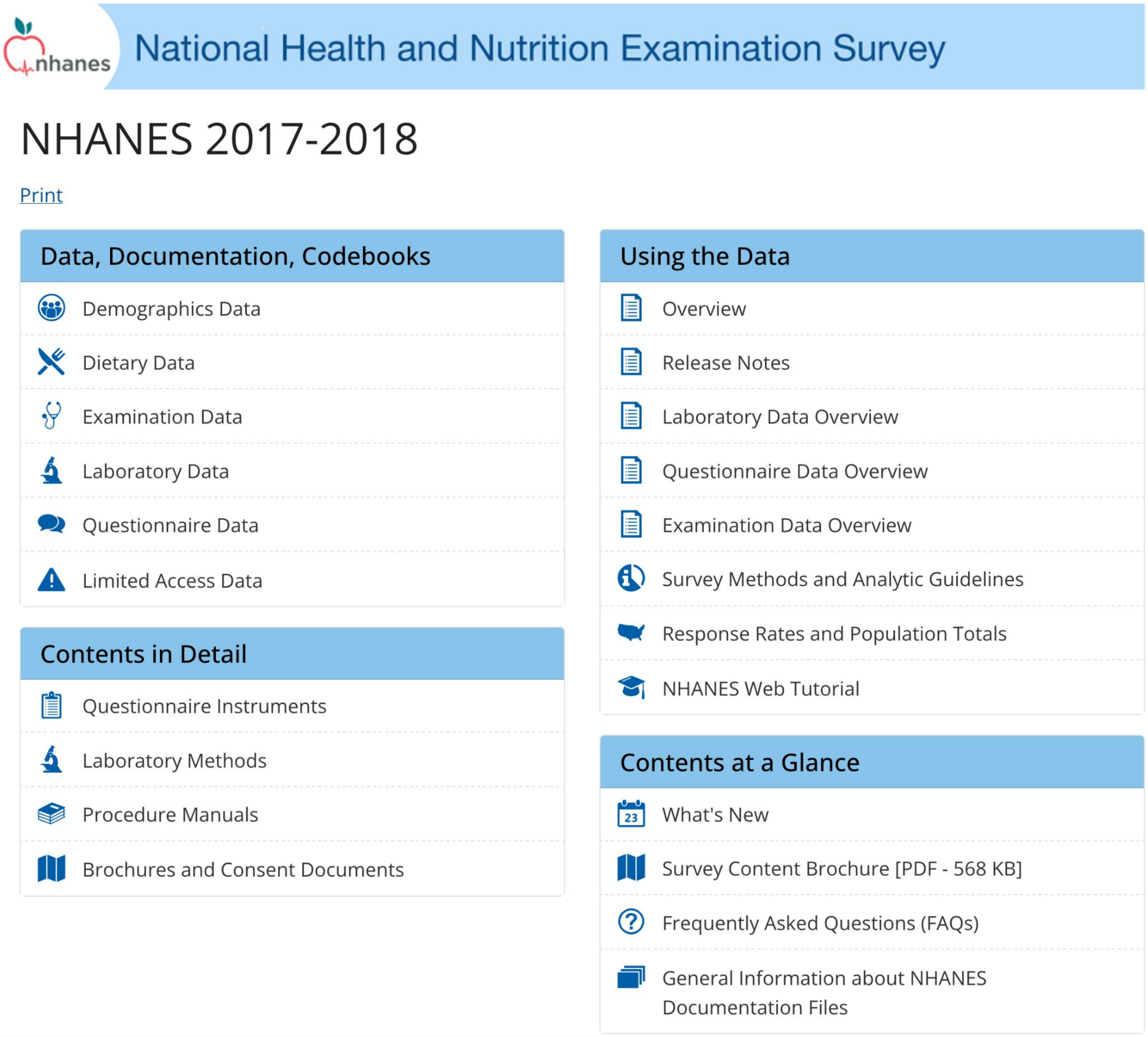
- Provide data structure details for users, developers, and other stakeholders
- Equip users with a common vocabulary and definitions for shared data, data standards, data flow and exchange, and help developers gage impacts of schema changes
- Provide the contextual understanding needed when deciding how to map one data system to another, or whether to subset, merge, stack, or transform data for a specific use

National Health and Nutrition Examination Surveys (NHANES)

- To study the relationship between diet, nutrition, and health and their roles in designated population subgroups with select diseases and risk factors
- Occurs every 2 years
  - Interview, Examination, and Laboratory results
  - Tens of datasets per survey cycle
- Challenges in understanding the data
  - Implicit objects in the data (e.g. participant, household, household reference person)
  - Questionnaires vs. Examination (e.g. “Do you have high blood pressure?” vs. Blood pressure readings)
  - Not trivial to quickly identify which types of data are there (e.g. “What are the diabetes-related data?”)

NHANES Documentation

- Documentation exists for each cycle
- Organized in
  - Data, Data Dictionnaires, Codebooks
  - Instruments, Methods, and Manuals
  - How to use the data
  - Overview and FAQs



National Health and Nutrition Examination Survey

NHANES 2017-2018

Print

Data, Documentation, Codebooks

- Demographics Data
- Dietary Data
- Examination Data
- Laboratory Data
- Questionnaire Data
- Limited Access Data

Contents in Detail

- Questionnaire Instruments
- Laboratory Methods
- Procedure Manuals
- Brochures and Consent Documents

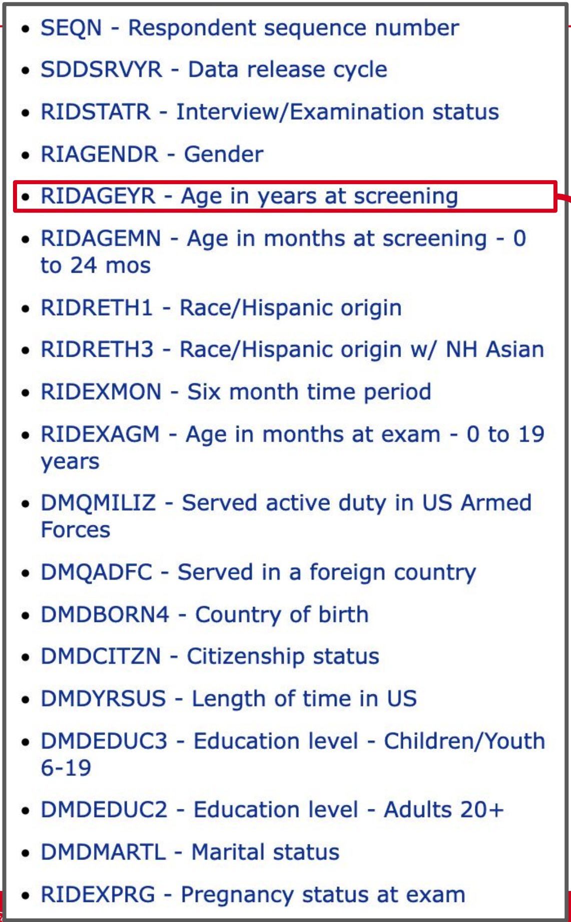
Using the Data

- Overview
- Release Notes
- Laboratory Data Overview
- Questionnaire Data Overview
- Examination Data Overview
- Survey Methods and Analytic Guidelines
- Response Rates and Population Totals
- NHANES Web Tutorial

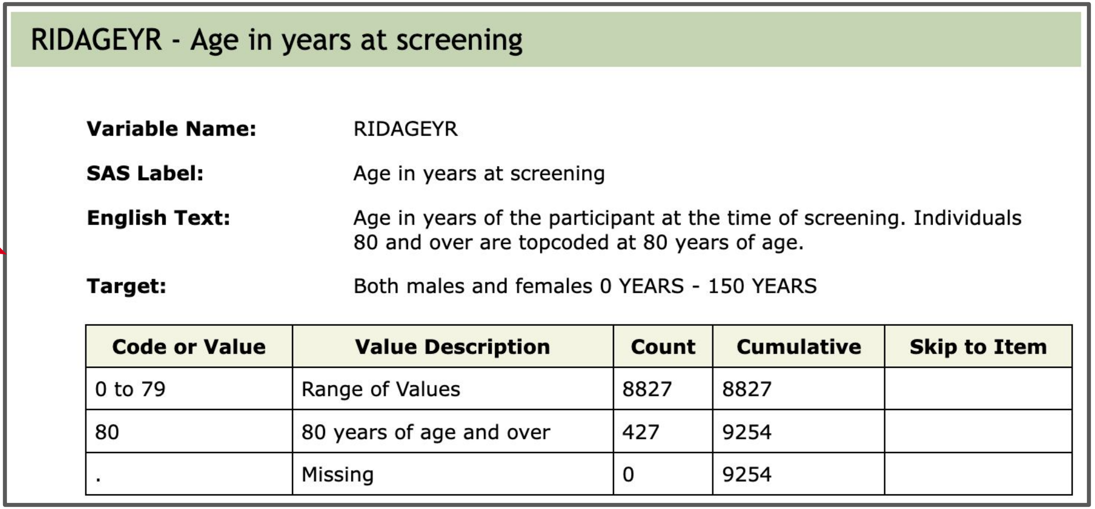
Contents at a Glance

- What's New
- Survey Content Brochure [PDF - 568 KB]
- Frequently Asked Questions (FAQs)
- General Information about NHANES Documentation Files

NHANES Data Dictionaries (2017-2018 cycle)



- SEQN - Respondent sequence number
- SDDSRVYR - Data release cycle
- RIDSTATR - Interview/Examination status
- RIAGENDR - Gender
- RIDAGEYR - Age in years at screening
- RIDAGEMN - Age in months at screening - 0 to 24 mos
- RIDRETH1 - Race/Hispanic origin
- RIDRETH3 - Race/Hispanic origin w/ NH Asian
- RIDEXMON - Six month time period
- RIDEXAGM - Age in months at exam - 0 to 19 years
- DMQMILIZ - Served active duty in US Armed Forces
- DMQADFC - Served in a foreign country
- DMDBORN4 - Country of birth
- DMDCITZN - Citizenship status
- DMDYRSUS - Length of time in US
- DMDEDUC3 - Education level - Children/Youth 6-19
- DMDEDUC2 - Education level - Adults 20+
- DMDMARTL - Marital status
- RIDEXPRG - Pregnancy status at exam



RIDAGEYR - Age in years at screening

Variable Name: RIDAGEYR

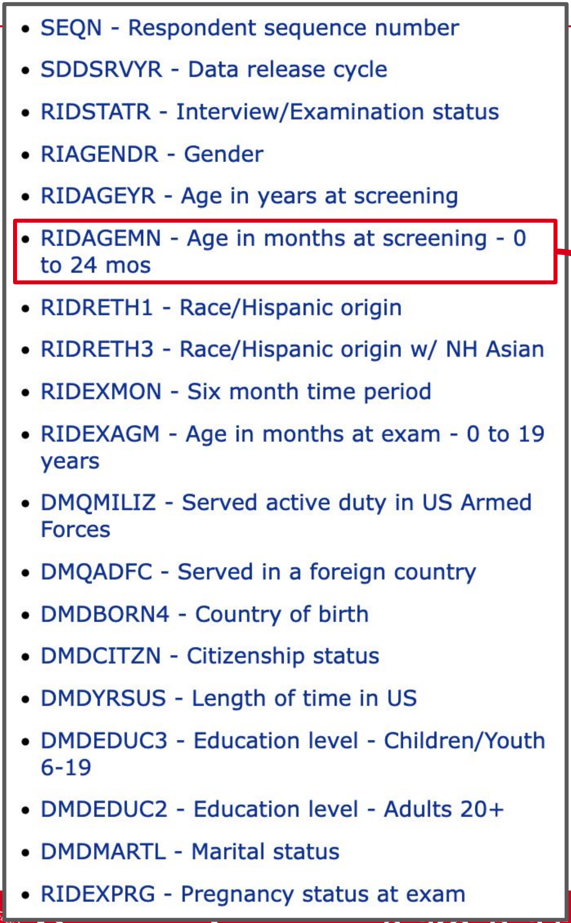
SAS Label: Age in years at screening

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

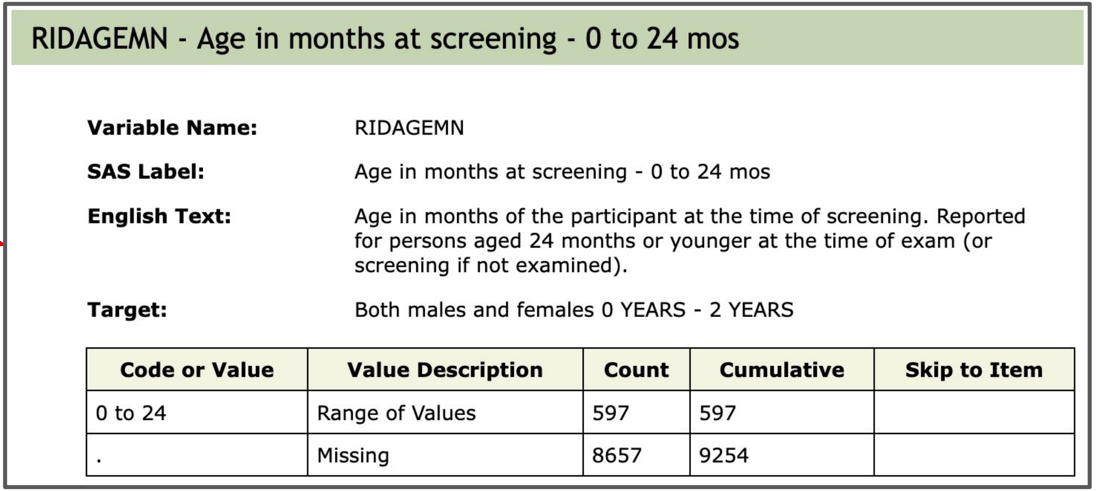
Target: Both males and females 0 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
0 to 79	Range of Values	8827	8827	
80	80 years of age and over	427	9254	
.	Missing	0	9254	

NHANES Data Dictionaries (2017-2018 cycle)



- SEQN - Respondent sequence number
- SDDSRVYR - Data release cycle
- RIDSTATR - Interview/Examination status
- RIAGENDR - Gender
- RIDAGEYR - Age in years at screening
- RIDAGEMN - Age in months at screening - 0 to 24 mos
- RIDRETH1 - Race/Hispanic origin
- RIDRETH3 - Race/Hispanic origin w/ NH Asian
- RIDEXMON - Six month time period
- RIDEXAGM - Age in months at exam - 0 to 19 years
- DMQMILIZ - Served active duty in US Armed Forces
- DMQADFC - Served in a foreign country
- DMDBORN4 - Country of birth
- DMDCITZN - Citizenship status
- DMDYRSUS - Length of time in US
- DMDEDUC3 - Education level - Children/Youth 6-19
- DMDEDUC2 - Education level - Adults 20+
- DMDMARTL - Marital status
- RIDEXPRG - Pregnancy status at exam



RIDAGEMN - Age in months at screening - 0 to 24 mos

Variable Name: RIDAGEMN

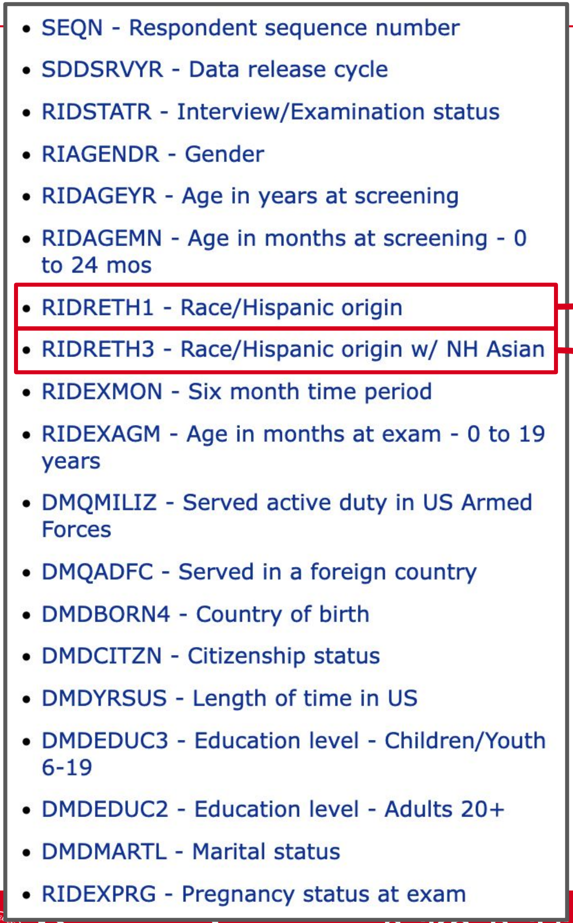
SAS Label: Age in months at screening - 0 to 24 mos

English Text: Age in months of the participant at the time of screening. Reported for persons aged 24 months or younger at the time of exam (or screening if not examined).

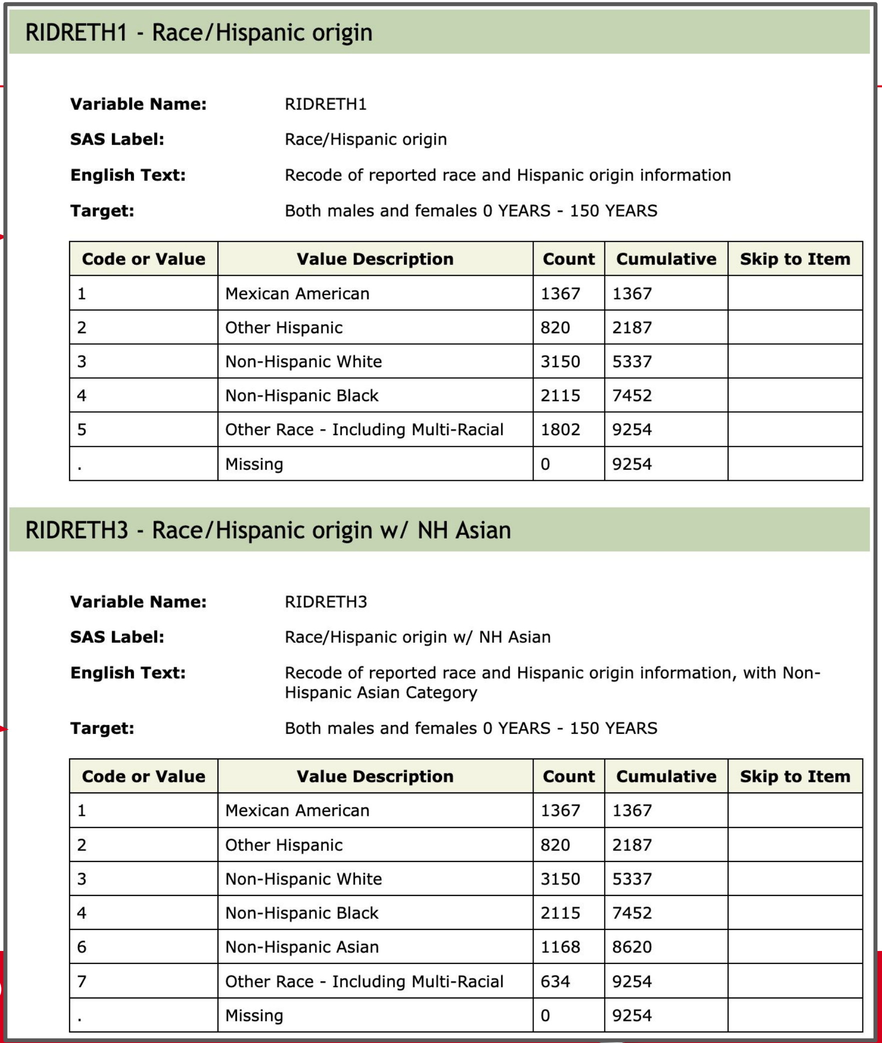
Target: Both males and females 0 YEARS - 2 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
0 to 24	Range of Values	597	597	
.	Missing	8657	9254	

NHANES Data Dictionaries (2017-2018 cycle)



- SEQN - Respondent sequence number
- SDDSRVYR - Data release cycle
- RIDSTATR - Interview/Examination status
- RIAGENDR - Gender
- RIDAGEYR - Age in years at screening
- RIDAGEMN - Age in months at screening - 0 to 24 mos
- RIDRETH1 - Race/Hispanic origin
- RIDRETH3 - Race/Hispanic origin w/ NH Asian
- RIDEXMON - Six month time period
- RIDEXAGM - Age in months at exam - 0 to 19 years
- DMQMILIZ - Served active duty in US Armed Forces
- DMQADFC - Served in a foreign country
- DMDBORN4 - Country of birth
- DMDCITZN - Citizenship status
- DMDYRSUS - Length of time in US
- DMDEDUC3 - Education level - Children/Youth 6-19
- DMDEDUC2 - Education level - Adults 20+
- DMDMARTL - Marital status
- RIDEXPRG - Pregnancy status at exam



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	1367	1367	
2	Other Hispanic	820	2187	
3	Non-Hispanic White	3150	5337	
4	Non-Hispanic Black	2115	7452	
5	Other Race - Including Multi-Racial	1802	9254	
.	Missing	0	9254	

RIDRETH3 - Race/Hispanic origin w/ NH Asian

Variable Name: RIDRETH3

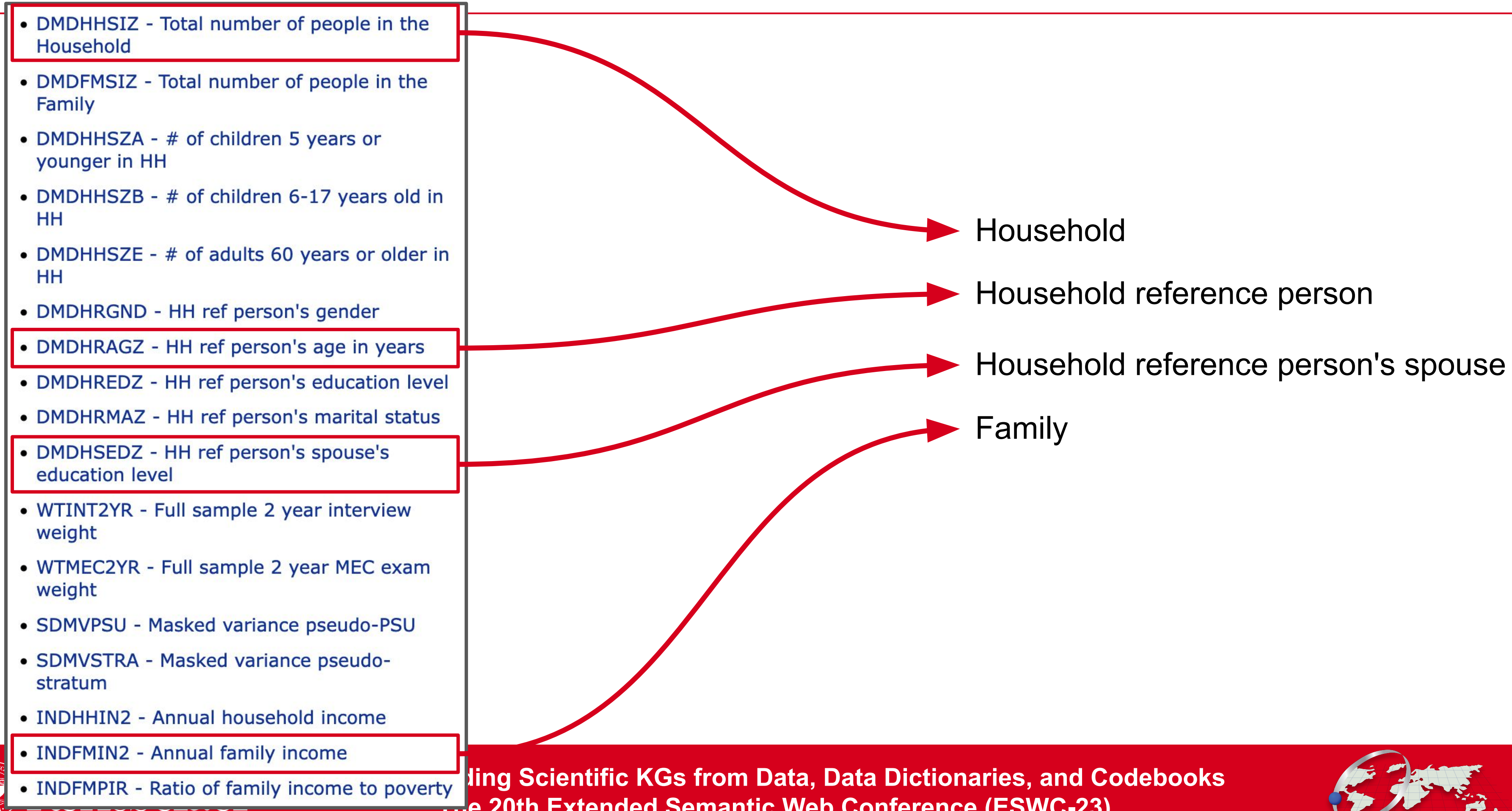
SAS Label: Race/Hispanic origin w/ NH Asian

English Text: Recode of reported race and Hispanic origin information, with Non-Hispanic Asian Category

Target: Both males and females 0 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Mexican American	1367	1367	
2	Other Hispanic	820	2187	
3	Non-Hispanic White	3150	5337	
4	Non-Hispanic Black	2115	7452	
6	Non-Hispanic Asian	1168	8620	
7	Other Race - Including Multi-Racial	634	9254	
.	Missing	0	9254	

NHANES Data Dictionaries (2017-2018 cycle)



- DMDHHSIZ - Total number of people in the Household
- DMDFMSIZ - Total number of people in the Family
- DMDHHSZA - # of children 5 years or younger in HH
- DMDHHSZB - # of children 6-17 years old in HH
- DMDHHSZE - # of adults 60 years or older in HH
- DMDHRGND - HH ref person's gender
- DMDHRAGZ - HH ref person's age in years
- DMDHREDZ - HH ref person's education level
- DMDHRMAZ - HH ref person's marital status
- DMDHSEDZ - HH ref person's spouse's education level
- WTINT2YR - Full sample 2 year interview weight
- WTMEC2YR - Full sample 2 year MEC exam weight
- SDMVPSU - Masked variance pseudo-PSU
- SDMVSTRA - Masked variance pseudo-stratum
- INDHHIN2 - Annual household income
- INDFMIN2 - Annual family income
- INDFMPIR - Ratio of family income to poverty

Household

Household reference person

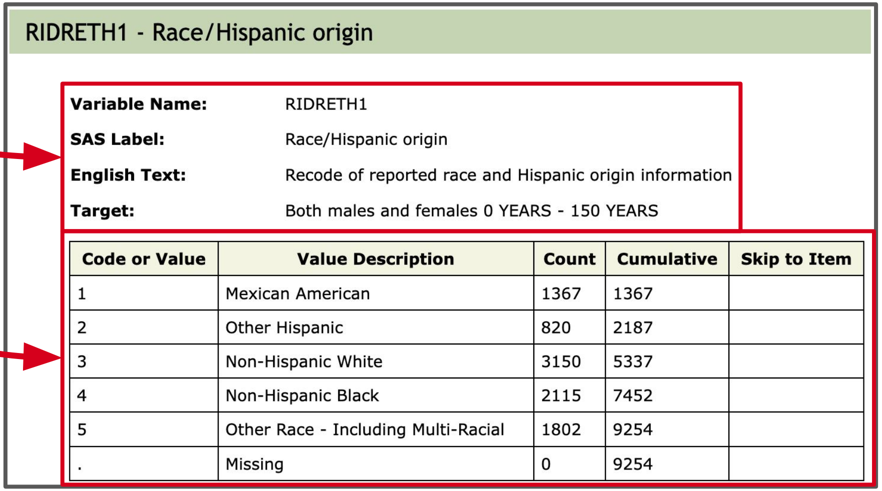
Household reference person's spouse

Family

NHANES Data Dictionaries (2017-2018 cycle)

Data Dictionary

Codebook



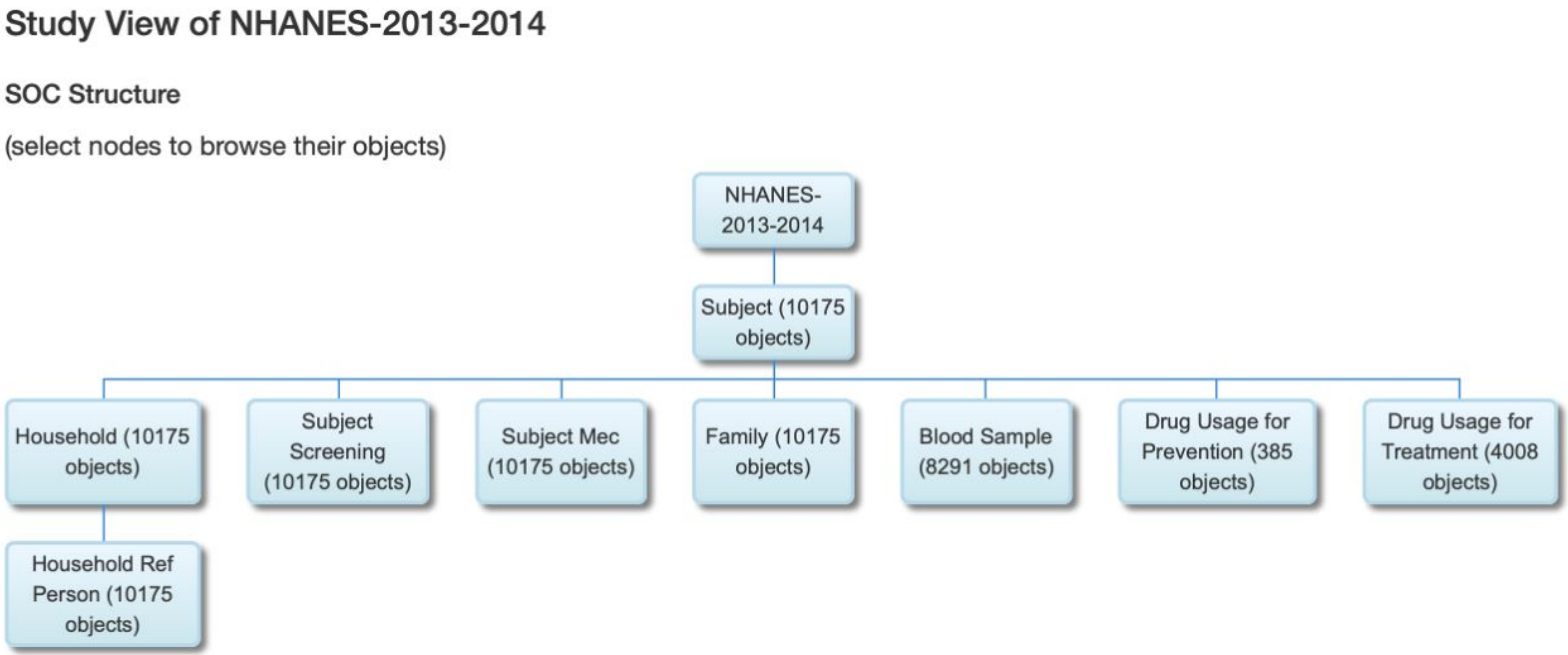
RIDRETH1 - Race/Hispanic origin

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	1367	1367	
2	Other Hispanic	820	2187	
3	Non-Hispanic White	3150	5337	
4	Non-Hispanic Black	2115	7452	
5	Other Race - Including Multi-Racial	1802	9254	
.	Missing	0	9254	

NHANES Knowledge Graph components (so far…)

- Cycles: 2009-2010, 2011-2012, 2013-2014, 2015-2016, 2017-2018



Study View of NHANES-2013-2014

SOC Structure

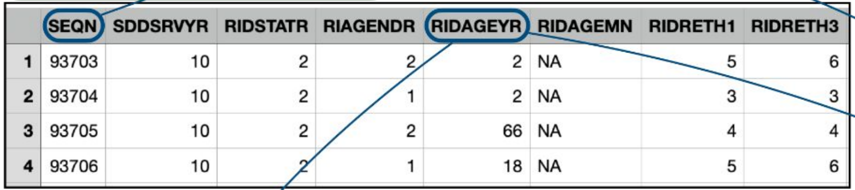
(select nodes to browse their objects)

NHANES-2013-2014
Subject (10175 objects)
Household (10175 objects)
Subject Screening (10175 objects)
Subject Mec (10175 objects)
Family (10175 objects)
Blood Sample (8291 objects)
Drug Usage for Prevention (385 objects)
Drug Usage for Treatment (4008 objects)
Household Ref Person (10175 objects)

Original Content

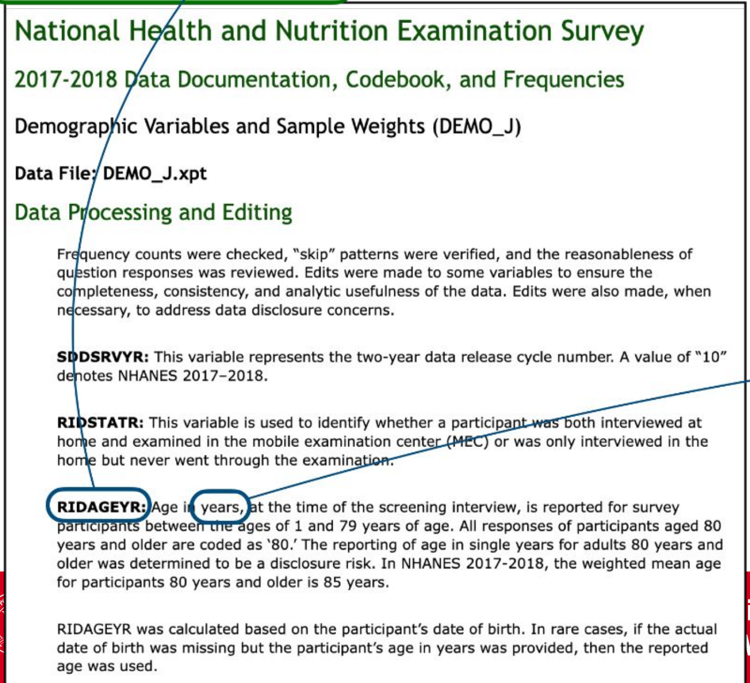
Semantic Content

NHANES Raw Data (Demographics)



	SEQN	SDDSRVYR	RIDSTATR	RIAGENDR	RIDAGEYR	RIDAGEMN	RIDRETH1	RIDRETH3
1	93703	10	2	2	2	NA	5	6
2	93704	10	2	1	2	NA	3	3
3	93705	10	2	2	66	NA	4	4
4	93706	10	2	1	18	NA	5	6

NHANES Demographics Dataset Documentation



National Health and Nutrition Examination Survey

2017-2018 Data Documentation, Codebook, and Frequencies

Demographic Variables and Sample Weights (DEMO_J)

Data File: DEMO_J.xpt

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.

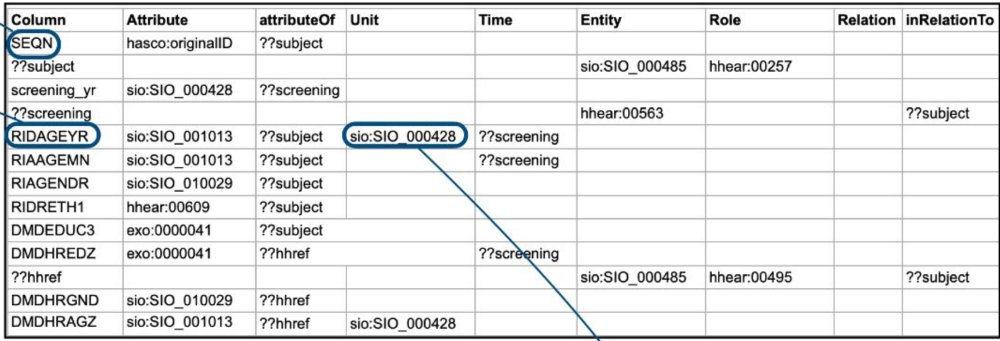
SDDSRVYR: This variable represents the two-year data release cycle number. A value of "10" denotes NHANES 2017–2018.

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. All responses of participants aged 80 years and older are coded as '80.' The reporting of age in single years for adults 80 years and older was determined to be a disclosure risk. In NHANES 2017-2018, the weighted mean age for participants 80 years and older is 85 years.

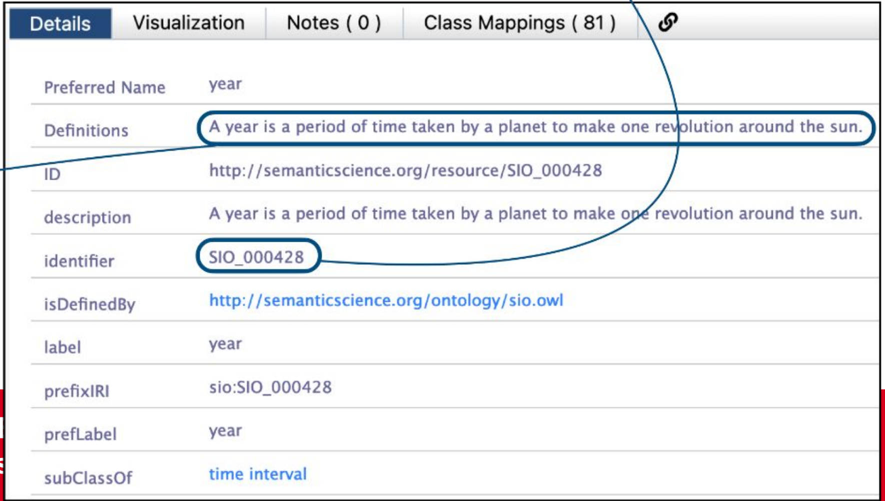
RIDAGEYR was calculated based on the participant's date of birth. In rare cases, if the actual date of birth was missing but the participant's age in years was provided, then the reported age was used.

NHANES Demographics Semantic Data Dictionary



Column	Attribute	attributeOf	Unit	Time	Entity	Role	Relation	inRelationTo
SEQN	hasco:originalID	??subject						
??subject					sio:SIO_000485	hhear:00257		
screening_yr	sio:SIO_000428	??screening						
??screening					hhear:00563			??subject
RIDAGEYR	sio:SIO_001013	??subject	sio:SIO_000428	??screening				
RIAAGEMN	sio:SIO_001013	??subject		??screening				
RIAGENDR	sio:SIO_010029	??subject						
RIDRETH1	hhear:00609	??subject						
DMDEDUC3	exo:0000041	??subject						
DMDHREDZ	exo:0000041	??hhref		??screening				
??hhref					sio:SIO_000485	hhear:00495		??subject
DMDHRGND	sio:SIO_010029	??hhref						
DMDHRAGZ	sio:SIO_001013	??hhref	sio:SIO_000428					

Community developed ontologies



Preferred Name	year
Definitions	A year is a period of time taken by a planet to make one revolution around the sun.
ID	http://semanticscience.org/resource/SIO_000428
description	A year is a period of time taken by a planet to make one revolution around the sun.
identifier	SIO_000428
isDefinedBy	http://semanticscience.org/ontology/sio.owl
label	year
prefixIRI	sio:SIO_000428
prefLabel	year
subClassOf	time interval