www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

Data Descriptor: A database of human exposomes and phenomes from the US National Health and Nutrition Examination Survey

Received: 06 June 2016
Accepted: 26 September 2016
Published: 25 October 2016

Chirag J. Patel[1], Nam Pho[1], Michael McDuffie[1], Jeremy Easton-Marks[1], Cartik Kothari[1], Isaac S. Kohane[1] & Paul Avillach[1]

The National Health and Nutrition Examination Survey (NHANES) is a population survey implemented by the Centers for Disease Control and Prevention (CDC) to monitor the health of the United States whose data is publicly available in hundreds of files. This Data Descriptor describes a single unified and universally accessible data file, merging across 255 separate files and stitching data across 4 surveys, encompassing 41,474 individuals and 1,191 variables. The variables consist of phenotype and environmental exposure information on each individual, specifically (1) demographic information, physical exam results (e.g., height, body mass index), laboratory results (e.g., cholesterol, glucose, and environmental exposures), and (4) questionnaire items. Second, the data descriptor describes a dictionary to enable analysts find variables by category and human-readable description. The datasets are available on DataDryad and a hands-on analytics tutorial is available on GitHub. Through a new big data platform, BD2K *Patient Centered Information Commons* (http://pic-sure.org), we provide a new way to browse the dataset via a web browser (https://nhanes.hms.harvard.edu) and provide application programming interface for programmatic access.

Design Type(s)	source-based data transformation objective
Measurement Type(s)	database creation objective
Technology Type(s)	computational method
Factor Type(s)	
Sample Characteristic(s)	Homo sapiens • United States of America

[1]Department of Biomedical Informatics, Harvard Medical School, 10 Shattuck St., Boston, Massachusetts 02115, USA. Correspondence and requests for materials should be addressed to C.J.P. (email: chirag_patel@hms.harvard.edu).

Background & Summary

United States health agencies, including the United States Centers for Disease Control and Prevention (CDC), have made a significant investment in monitoring the health of the population through surveys such as the National Health and Nutrition Examination Survey (NHANES). These datasets provide individual-level health-related measures in a large and representative sample of the United States (e.g., from 1999–2006, $N=41,474$). For example, these data are instrumental in providing prevalence of disease-related factors, such as diabetes and obesity (e.g., refs 1–3), drug use[4], and present reference intervals for child growth, such as head circumference. These data have helped to shape public health policy. For example, these data were used to demonstrate the effect of removal of lead from gasoline (a gross decrease since legislation). Many have used these data to create hypotheses regarding associations between biomarkers of environmental chemical factors and disease, such as diabetes and heart disease. We have used these data to perform the first 'environment-wide association studies' (EWAS)[5], linking >250 environmental biomarkers with disease phenotypes, such as diabetes[6,7], self-reported preterm birth[8], serum cholesterol levels[9], blood pressure[10], all-cause mortality[11], telomere length[12], and income[13].

The NHANES is a CDC program that began in the 1960 s and in the current day, bi-annually samples 15 counties of United States population (N ~ 5 K per year). Each year, the counties that are sampled change, ensuring a representative and diverse sampling. Specifically, NHANES uses a multistage and 'probability' sampling design. To provide reliable statistics, the NHANES 'over-samples' persons 60 and older, African Americans, and Hispanics and analysts.

The NHANES is designed to estimate major disease prevalence, such as diabetes, obesity, and cardiovascular disease in the United States. It is one of the only studies that combines simultaneously assessed self-reported questionnaires and physical measurements. Self-reported instruments include dietary questionnaires the estimate nutrient content of foods consumed around the time of survey and health and disease-related questionnaire. Second, the NHANES contains phenotypes such as blood pressure, pulse rate, respiratory capacity, height, weight, and tooth count in an effort to estimate the range and prevalence of these phenotypic measures. They are not used for medical diagnoses for the participants.

The *exposome* has been touted as the comprehensive battery of environmental exposures encountered in humans[14]. The CDC NHANES is one of the first population survey programs to have *exposome* measurements. The CDC samples urine, blood, and other human tissue to measure environmental exposure indicators of the exposome using gold standard mass spectrometry and immunological assays. Environmental exposure assays include, for example, lead, mercury, arsenic, pesticide metabolites, air pollution indicators, and plasticizing agents, all hypothesized to have some relationship with health. The NHANES has been instrumental in providing what and how many environmental chemicals are found in human tissue (e.g., ref. 15). Clinical and physiological phenotypes of the phenome include cholesterol (e.g., HDL-cholesterol, LDL-cholesterol, triglycerides), glucose, insulin, C-reactive protein (CRP), white blood counts, and other blood or urine based measures. All of the measures are taken simultaneously.

The NHANES raw datasets for surveys currently exist in >250 number of separate proprietary SAS-formatted files (e.g.: https://wwwn.cdc.gov/Nchs/Nhanes/1999-2000/DEMO.XPT). Description of each variable (e.g., a human-readable variable name and units of measurement) exist in a separate table embedded in an.html webpage (e.g.: https://wwwn.cdc.gov/nchs/nhanes/search/variablelist.aspx?Component=Demographics). All technical information about each variable, such as way it was measured, are also available on the NHANES website as a.html page. The NHANES has variables of many types, including biomarkers of environmental exposures, clinical markers, physiological measures, questionnaire items, that are continuous or categorical. Next, the NHANES consists of multiple 'survey waves' that represent a sampling for a 2-year period (e.g., 1999–2000 to 2005–2006 and beyond). Our data resource allows investigators move beyond examining a handful of variables to one that takes advantage of the multiple variables across a number of NHANES survey waves (e.g., akin to refs 11,16,17). Second, our data resource allows for quick evaluation of hypotheses before executing a formal scientific investigation. We are offering this integrated resource ready to analyze for free of cost, leveraging our previous experience.

We also offer a way to access the dataset programmatically through an 'application programming interface' (API). We utilize *i2b2/tranSMART*, a data repository software platform used to implement BD2K *Patient Information Commons-Standardized Unification of Research Elements (PIC-SURE)* (http://pic-sure.org). The *Informatics for Integrating Biology and the Bedside* (*i2b2*) open-source software was developed to provide a federated informatics infrastructure to house/store, maintain, and analyze cohort data emerging from population-level datasets from around the nation for the purpose of driving biological discovery[18–20]. *i2b2* enables the cohesive analysis of heterogeneous phenotypic data. *tranSMART* is an open-sourced 'application layer' for *i2b2* (refs 21,22), providing a software add-ons to *i2b2* for user interfaces, data mapping, and loading cohort data. This software provides a means to assemble, query, and analyze disparate and heterogeneous cohort datasets, such as the NHANES. The PIC-SURE software technology provides an accessible representation of NHANES, facilitating ad hoc querying of the health measures of the US while providing an application programming interface (API) for consumption by external applications and scripts, such as statistical tools such as *R*.

In this data descriptor, we provide (1) a data descriptor for unified raw NHANES data, (2) sample starter analytic code, analytic compute environment in a *Docker* container, and guide to conduct analysis with the NHANES data, and (3) introduce the *PIC-SURE* enabled web application to browse and download the data through an 'application programming interface' (API). Further, we have provided a web video tutorial on the web application located here: https://vimeo.com/182576739.

We emphasize our data descriptor is an introduction for use of the NHANES dataset and that all analyses must be verified with data from CDC/NHANES directly. Furthermore, we also emphasize that the derived variables we include were suitable for our own analyses of NHANES and may not be suitable for hypotheses specific to other investigators. Therefore, we include all raw variables in our integrated dataset for investigators.

Methods

National Health and Nutrition Examination Surveys (NHANES) data

NHANES datasets are publicly accessible through the United States Centers of Disease Control and Prevention (US CDC)[23–26]. All NHANES participants have consented for their information to be used in research.

Figure 1 shows our procedure. We downloaded 255 total data files, encoded in proprietary SAS '.xpt' format, corresponding to participants surveyed in 1999–2000 (52 files), 2001–2002 (57 files), 2003–2004 (77 files), 2005–2006 (69 files) from the CDC NHANES website (Fig. 1a,b) which are hyperlinked to a CDC website in January 2014. We chose to focus on these surveys as they had the greatest number of variables available at the time of download. We will make future instances of merged NHANES available via DataDryad with additional Data Descriptors.

Each participant of the NHANES has a unique identifier; in other words, there is no overlap in participants in the 1999–2000, 2001–2002, 2003–2004, and 2005–2006 surveys. In total, these 255 files contain information on 41,474 distinct individuals representative of the United States population and 1,191 unique variables.

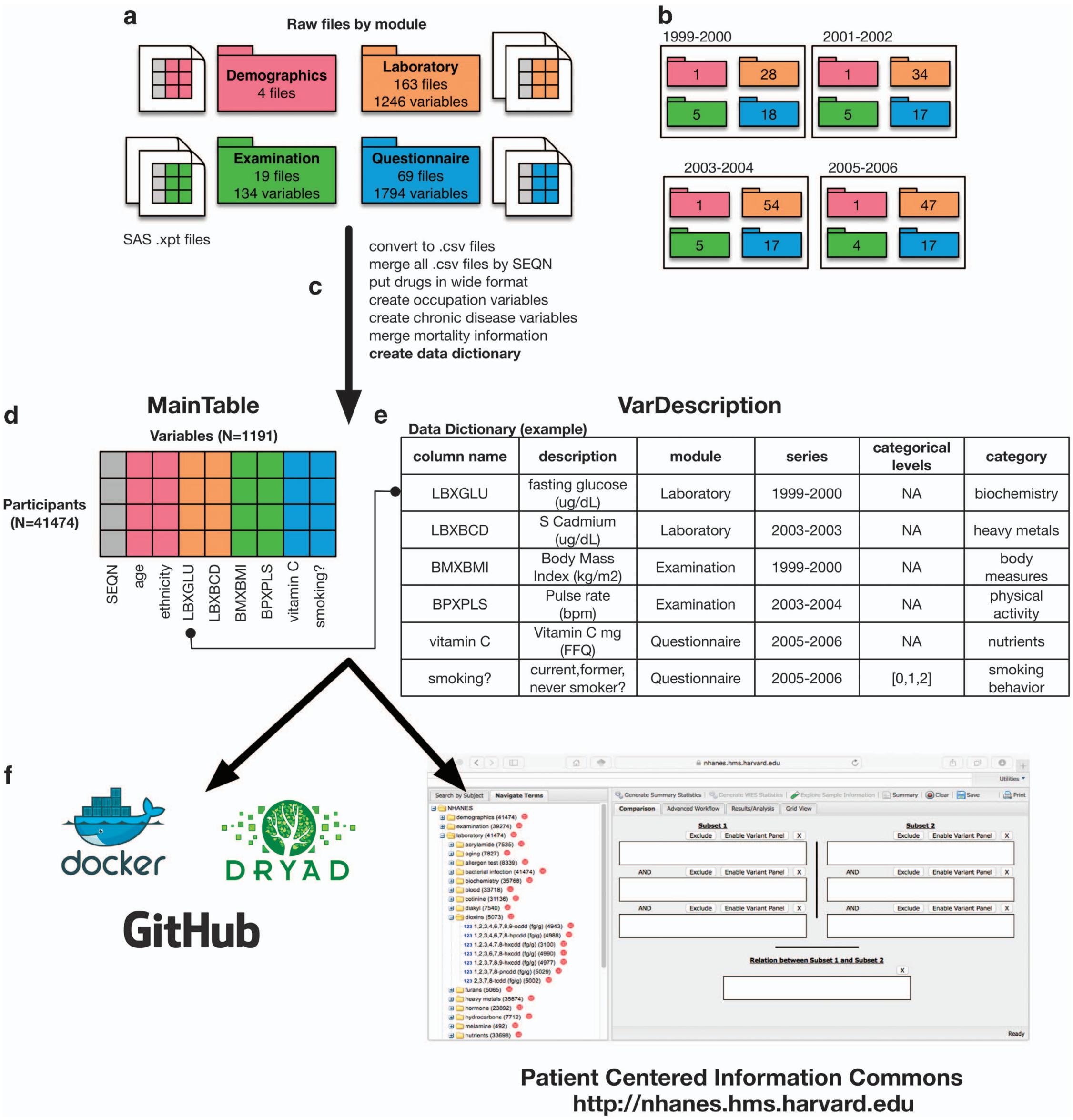
Each.xpt formatted data file consists of information structured in a 'N × M' form, in which N number of individuals make up every row and M number of columns of variables for each individual (Fig. 1a,b) and a participant identifier (called 'SEQN'), the primary key that joins the data files together (shown as a gray column, Fig. 1a). The CDC/NHANES have binned each file in 4 separate 'modules' that corresponded to (1) whether they contain demographic information (e.g., age, race/ethnicity, survey characteristics, income [Fig. 1a, red folder]), (2) laboratory measures (e.g., biomarker measurements assayed in biological tissue, such as serum or urine, depicted in orange [Fig. 1a, orange folder]), (3) physical examination (e.g., measurements such as body mass index, weight, height; [Fig. 1a, green folder]), or questionnaire (e.g., food-frequency questionnaire or health status questionnaire [Fig. 1a, blue folder]). Each of these categories, or 'modules', are called Demographics, Laboratory, Examination, and/or Questionnaire modules respectively.

In total, we downloaded 4 Demographics data files, 163 Laboratory data files, 19 Examination files, and 69 Questionnaire data files. Fig. 1b and Table 1 depicts the total files for each NHANES module for the 1999–2000, 2001–2002, 2003–2004, and 2005–2006 datasets.

After downloading all 255.xpt files, we executed a number of data processing steps. First, all.xpt files were converted into.csv files using using the 'foreign' R package[27], preserving the original 'N × M' form of the data. Next, we created some derived variables to ease potential downstream analyses, including (1) occupation (1 variable), (2) chronic disease (40 variables), and (3) pharmaceutical drug use (100 variables) (Fig. 1c).

We coded occupation as variables that correspond to (1) white-collar and professional jobs that are coded as white-collar and semi-routine (e.g., technicians), blue-collar and high-skill (e.g., mechanics, construction trades, and military), blue-collar and semi-routine (e.g., personal services, farm workers) as previously described in our previous EWAS[28]. Labor force participation was defined as working at a job or business or having a job or business within the last two weeks, not including work around the house.

We defined presence of 6 types of chronic diseases, including diabetes (1 variable), coronary disease (1 variable), hypertension (1 variable), asthma (1 variable), rheumatoid arthritis, osteoarthritis, and 30 site-specific cancers. We coded diabetes as present (as an integer 1) if the participant had a fasting blood glucose greater than 125 mg/dl (as per American Diabetes Association [ADA]) threshold for diabetes diagnosis or if the participant answered 'yes' to the question, 'Other than during pregnancy, have you ever been told by a doctor or health professional that {you have/{he/she/SP} has} diabetes or sugar diabetes?'. If the participant did not have both of those characteristics, he/she were coded as 0 (ref. 29). Similarly, we defined presence of hypertension as 1 if the participant had a systolic over diastolic blood pressure greater than 130 over 90 or answered 'yes' to the question, 'Have you ever been told by a doctor or other health professional that you had hypertension, also called high blood pressure' and 0 otherwise. We defined presence of coronary disease as 1 if the participant answered 'yes' to the question, 'Has a doctor or other health professional ever told you that you had coronary (kor-o-nare-ee) heart disease?' and 0 otherwise. The NHANES also contains coding for site-specific cancers. First, participants were asked whether a doctor has 'ever told you you have cancer?'. If the participant replies yes to a question, a followup question is administered, 'what type of cancer do you have' and the participant can answer from a set of 27 cancers, such as breast, skin, lung, colon, bladder, kidney, and other type of cancers. We turned these into 27 separate variables that are coded 1 if the site-specific cancer is present, 0 otherwise.



convert to .csv files
merge all .csv files by SEQN
put drugs in wide format
create occupation variables
create chronic disease variables
merge mortality information
create data dictionary

column name	description	module	series	categorical levels	category
LBXGLU	fasting glucose (ug/dL)	Laboratory	1999-2000	NA	biochemistry
LBXBCD	S Cadmium (ug/dL)	Laboratory	2003-2003	NA	heavy metals
BMXBMI	Body Mass Index (kg/m2)	Examination	1999-2000	NA	body measures
BPXPLS	Pulse rate (bpm)	Examination	2003-2004	NA	physical activity
vitamin C	Vitamin C mg (FFQ)	Questionnaire	2005-2006	NA	nutrients
smoking?	current,former, never smoker?	Questionnaire	2005-2006	[0,1,2]	smoking behavior

Figure 1. Methods overview for creating the unified NHANES dataset. (**a**) Each SAS-formatted (.xpt) data file provided by the CDC/NHANES are binned by 'module' (represented by folders), including Demographics (4 files), Laboratory (163 files), Examination (19 files), and Questionnaire (69 files). Participant identifiers to merge data files across modules are depicted as gray colums. (**b**) File number breakdown by survey year and module. (**c**) We processed the data to create new variables, added pharmaceutical drug information, and added mortality information. (**d**) We merged all 255 files by the patient identifier to create a large unified table ('MainTable') consisting of 41 K participants and 1191 unique variables. (**e**) We created a data dictionary that contains human readable variable descriptions and other meta-data, such as variable category and the levels of the variable if categorical. (**f**) Data is accessible via DataDryad and browsable through the PIC-SURE website (https://nhanes.hms.harvard.edu). Data and a Usage Guide is available on GitHub. Rstudio analytics environment with dataset, xwas R library, and user guides packaged as a Docker hub container (chiragjp/nhanes_scidata).

Third, we extracted pharmaceutical drug use for each participant. The CDC used a Master Drug Database (MDDB), a proprietary but comprehensive database of all prescription and some nonprescription drug products available in the U.S. drug market. The CDC NHANES interviewer asked participants whether they were taking a drug in the past month, and if they were, what drugs they were taking. The CDC NHANES interviewer matched each drug to an MDDB identifier and drug description (e.g., METFORMIN or ALBUTEROL). Second, the CDC NHANES interviewer—if the interview was

Description	NHANES module	Number of variables	Number of files
Physical examination	Examination	134	19
Laboratory assay (serum and urine) results	Laboratory	1,246	163
Self-reported questionnaire items	Questionnaire	1,794	69
Demographic attributes and cause of death in 2006	Demographics and mortality	28	4

Table 1. Number of variables and files per NHANES module.

occurring at the participant's home—verified possession of the prescription drug container. Each participant could report taking more than one drug. There were 626, 668, 667, and 692 unique drugs found by the CDC interviewers in the 1999–2000, 2001–2002, 2003–200, and 2005–2006 cohorts respectively. To keep the merged data table (Fig. 1d) of tractable size, we chose to focus on the top 100 drugs that were most prevalent in the population. We coded a participant was on a drug if (1) they reported use of a drug and (2) whether the interviewer verified the container was present.

The CDC also ascertained cause and time of death (mortality) information for a subset of the participants in 2006 by linking eligible participants to the National Death Index. We incorporated this data into our data merge ($n = 11,429$ participants). The variables that describe the mortality information include ELIGSTAT (whether the person was eligible for death linkage), MORTSTAT (whether the participant was deceased), PERMTH_INT (time to death from interview or time to linkage if participant is living [censored data]), PERMTH_EXM (time to death from examination or time to linkage if participant if living), DIABETES (if the cause of death was diabetes), HYPERTEN (if the cause of death was hypertension, and HIPFRACT (if the cause of death was hip fracture).

Finally, we combined the 255 files together into single data file by merging by the patient identifier ('SEQN') (Fig. 1d). This merge resulted in one consolidated and analysis-ready data file representing a grand total of 1,191 variables on 41,474 participants.

Creation of a digital handbook: annotating and categorizing the NHANES datasets

The CDC NHANES have provided a.html formatted codebook (e.g.: https://wwwn.cdc.gov/Nchs/Nhanes/Search/variablelist.aspx?Component = Laboratory&CycleBeginYear = 1999) that consists of variable name (column in the.xpt file) and a human-readable description of each variable. For example, the variable with names RIDAGEYR or LBXGLU is described as 'Age in Years' and 'fasting serum glucose [mg ul^{-1}]' respectively. These descriptions include the variable units, such as 'ug/mL' (inferred as a continuous variable), or 'positive'/'negative' (a binary variable) of each variable.

We have extended the CDC NHANES data description methodology in the following ways (Fig. 1e) to facilitate analysis and data browsing. Specifically, we have created a data dictionary that contains the name of the variable, a human readable description of the variable, what 'module' a variable belongs to, what survey the variable was measured (e.g., 1999–2000). Second, we have binned each variable into categories that offer more specificity than the CDC NHANES 'module' characterization. We make available the data dictionary (Fig. 1e) along with the data set (Data Citation 1). A summary of the number of variables per category, the median sample size for the variables in the category, and the demographic representation (percent female and race/ethnicity available for each variable) in Table 2. The entire data dictionary is available as Table 3 (available online only) (Data Citation 1 and Table 3 (available online only)).

These categories aide in the filtering and querying of variables with common types, such as 'nutrients', 'body measures', 'pharmaceutical drug', 'viral infection', and 'pesticides'. Second, we have created a column that denotes the categorical levels for variables that are categorical or binary. For example, 'Are you a past, current, or never smoker?' is a variable that has three levels, one representing 'never smoker', 'current smoker', and 'past smoker'; these categories are captured in a column called 'categorical levels'.

Browsing and accessing the data through BD2K Patient-Centered Information Commons (PIC)

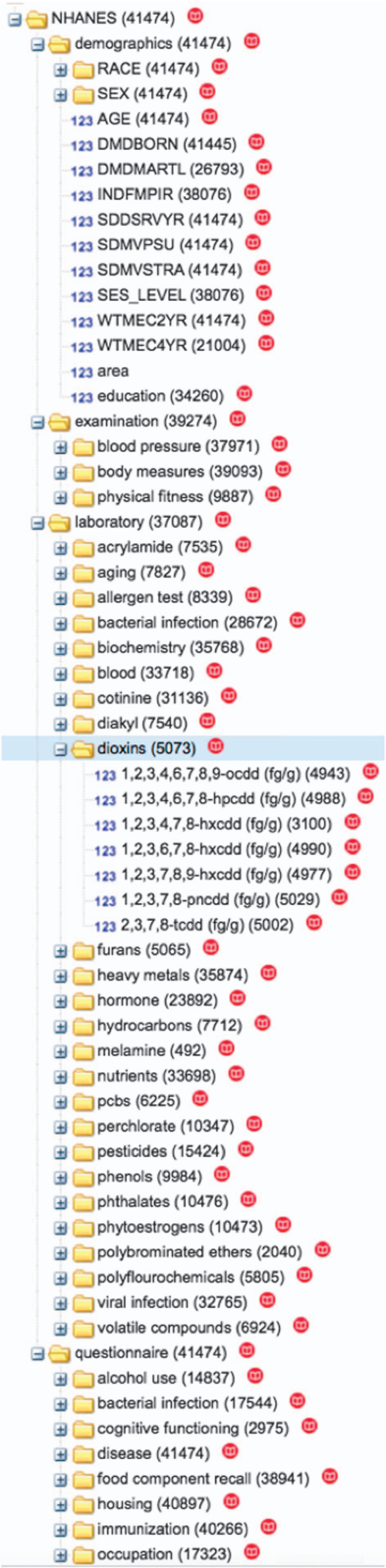
We leveraged the Patient-Centered Information Commons (PIC, for an overview, see: http://pic-sure.org)) platform is leveraged to (1) enable interactive web browsing of the NHANES data (see: https://nhanes.hms.harvard.edu) and (2) access data through an application programming interface (API). PICs are built using the i2b2/tranSMART software stack. Data is organized into a hierarchy resembling a directory structure to facilitate browsing (Figs 2 and 3). Raw data can be also queried using a drag and drop interface (Fig. 3). With the NHANES, we organized each of the 1171 variables into a multi-level hierarchy that was ordered by the module (*i.e.*, 'Laboratory', 'Examination', 'Demographics', and 'Questionnaire') and category (*i.e.*, 'pesticides', 'body measures', etc, see Table 2). To display this NHANES data hierarchy in our user interface we created a Metadata mapping file located here: https://github.com/hms-dbmi/public-data-deployments/blob/master/NHANES/nhanes_9906.map and used this mapping file to integrate the data file.

The merged dataset ('MainTable') and data dictionary ('VarDescription') (Fig. 1d,e) are made available in DataDryad (Fig. 1f). A Usage Guide and.Rdata files are provided for download in GitHub (Fig. 1f). Finally, all data are browsable at https://nhanes.hms.harvard.edu.

Category	Number of variables	Surveys	Median (N)	Female (%)	White (%)	Black (%)	Mexican (%)	Other his (%)	Other eth (%)
acrylamide	2	3	7189.50	0.51	0.41	0.26	0.25	0.03	0.04
aging	1	1;2	7827.00	0.52	0.51	0.17	0.24	0.05	0.03
alcohol use	4	1;2;3;4	11141.50	0.46	0.54	0.17	0.21	0.04	0.03
allergen test	20	4	7796.50	0.51	0.40	0.26	0.26	0.03	0.05
bacterial infection	48	1;2;3;4	742.00	0.50	0.43	0.26	0.24	0.04	0.04
biochemistry	56	1;2;3;4	26038.00	0.51	0.43	0.23	0.26	0.04	0.04
blood	20	1;2;3;4	33661.00	0.51	0.39	0.25	0.28	0.04	0.04
blood pressure	4	1;2;3;4	26036.00	0.51	0.39	0.25	0.28	0.04	0.04
body measures	19	1;2;3;4	27259.00	0.48	0.40	0.25	0.27	0.04	0.04
cognitive functioning	2	1;2	2975.00	0.52	0.61	0.14	0.19	0.04	0.02
cotinine	1	1;2;3;4	31136.00	0.51	0.40	0.25	0.27	0.04	0.04
diakyl	7	1;2;3	7422.00	0.52	0.39	0.25	0.27	0.04	0.04
dioxins	7	1;2;3	4988.00	0.52	0.44	0.21	0.26	0.04	0.04
disease	40	1;2;3;4	18526.50	0.53	0.48	0.21	0.23	0.04	0.04
food component recall	162	1;2;3;4	16412.00	0.51	0.39	0.25	0.26	0.04	0.04
furans	10	1;2;3	4980.00	0.52	0.44	0.21	0.26	0.04	0.04
heavy metals	31	1;2;3;4	10081.00	0.50	0.39	0.26	0.27	0.04	0.04
hormone	8	1;2;3;4	9473.00	0.52	0.42	0.22	0.26	0.04	0.04
housing	9	1;2;3;4	35087.00	0.51	0.39	0.25	0.28	0.04	0.05
hydrocarbons	23	1;2;3	7209.00	0.52	0.41	0.25	0.27	0.04	0.04
immunization	3	1;2;3;4	35305.00	0.52	0.40	0.25	0.27	0.04	0.04
melamine	1	3	492.00	0.53	0.42	0.27	0.23	0.04	0.04
nutrients	31	1;2;3;4	22880.00	0.51	0.42	0.25	0.25	0.04	0.04
occupation	21	1;2;3;4	769.00	0.27	0.46	0.20	0.27	0.05	0.04
pcbs	38	1;2;3	6049.00	0.52	0.43	0.22	0.27	0.04	0.04
perchlorate	7	3;4	5479.50	0.51	0.41	0.26	0.25	0.03	0.05
pesticides	66	1;2;3;4	4999.00	0.52	0.39	0.25	0.27	0.04	0.04
pharmaceutical	221	1;2;3;4	20456.00	0.51	0.39	0.25	0.28	0.04	0.04
phenols	7	3;4	5065.00	0.51	0.42	0.26	0.25	0.03	0.05
phthalates	15	1;2;3;4	10476.00	0.51	0.40	0.25	0.27	0.04	0.04
physical fitness	15	1;2;3;4	8688.00	0.48	0.34	0.26	0.32	0.04	0.04
phytoestrogens	6	1;2;3;4	10453.50	0.51	0.40	0.25	0.27	0.04	0.04
polybrominated ethers	12	3	1999.50	0.51	0.45	0.24	0.24	0.03	0.04
polyflourochemicals	12	1;3;4	5805.00	0.51	0.42	0.24	0.27	0.04	0.03
sexual behavior	2	1;2;3;4	5178.00	0.00	0.48	0.21	0.22	0.04	0.04
smoking behavior	30	1;2;3;4	7479.50	0.48	0.51	0.20	0.20	0.04	0.04
smoking family	8	1;2;3;4	7668.00	0.49	0.43	0.35	0.15	0.04	0.04
social support	3	1;2;3;4	9937.00	0.51	0.56	0.19	0.18	0.03	0.03
street drug	24	1;2;3;4	600.00	0.39	0.54	0.20	0.20	0.04	0.04
sun exposure	1	3;4	2444.00	0.50	0.70	0.06	0.17	0.02	0.04
supplement use	85	1;2;3;4	41366.00	0.51	0.39	0.25	0.28	0.04	0.04
viral infection	18	1;2;3;4	15400.00	0.51	0.40	0.24	0.27	0.04	0.04
volatile compounds	51	1;2;3;4	5573.00	0.53	0.45	0.24	0.23	0.04	0.05

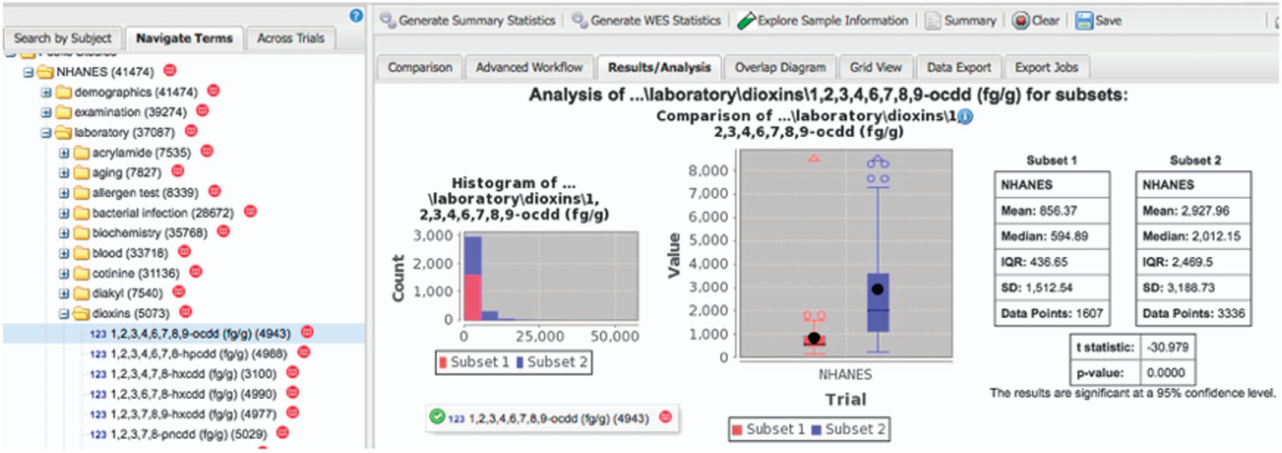
Table 2. Categories of variables, the number of variables, surveys represented (1 = 1999–2000, 2 = 2001–2002,3 = 2003–2004,4 = 2005–2006) number of raw data files, sample size, and demographic distribution. Sample sizes are the median participants available for the variables in the respective categories (e.g., the median sample size available for all the alcohol use variables is 11,141.5). ‘Other His’ denotes ‘Other Hispanic’. ‘Other Eth’ denotes ‘Other race/ethnicity’.

We have provided two additional resources for individuals to learn about the resource. The first is a tutorial of the web application located at Vimeo (https://vimeo.com/182576739). This web application shows users how to count the number of variables and number of participants (by age, sex, and



- ⊟ NHANES (41474)
  - ⊟ demographics (41474)
    - ⊞ RACE (41474)
    - ⊞ SEX (41474)
    - 123 AGE (41474)
    - 123 DMDBORN (41445)
    - 123 DMDMARTL (26793)
    - 123 INDFMPIR (38076)
    - 123 SDDSRVYR (41474)
    - 123 SDMVPSU (41474)
    - 123 SDMVSTRA (41474)
    - 123 SES_LEVEL (38076)
    - 123 WTMEC2YR (41474)
    - 123 WTMEC4YR (21004)
    - 123 area
    - 123 education (34260)
  - ⊟ examination (39274)
    - ⊞ blood pressure (37971)
    - ⊞ body measures (39093)
    - ⊞ physical fitness (9887)
  - ⊟ laboratory (37087)
    - ⊞ acrylamide (7535)
    - ⊞ aging (7827)
    - ⊞ allergen test (8339)
    - ⊞ bacterial infection (28672)
    - ⊞ biochemistry (35768)
    - ⊞ blood (33718)
    - ⊞ cotinine (31136)
    - ⊞ diakyl (7540)
    - ⊟ dioxins (5073)
      - 123 1,2,3,4,6,7,8,9-ocdd (fg/g) (4943)
      - 123 1,2,3,4,6,7,8-hpcdd (fg/g) (4988)
      - 123 1,2,3,4,7,8-hxcdd (fg/g) (3100)
      - 123 1,2,3,6,7,8-hxcdd (fg/g) (4990)
      - 123 1,2,3,7,8,9-hxcdd (fg/g) (4977)
      - 123 1,2,3,7,8-pncdd (fg/g) (5029)
      - 123 2,3,7,8-tcdd (fg/g) (5002)
    - ⊞ furans (5065)
    - ⊞ heavy metals (35874)
    - ⊞ hormone (23892)
    - ⊞ hydrocarbons (7712)
    - ⊞ melamine (492)
    - ⊞ nutrients (33698)
    - ⊞ pcbs (6225)
    - ⊞ perchlorate (10347)
    - ⊞ pesticides (15424)
    - ⊞ phenols (9984)
    - ⊞ phthalates (10476)
    - ⊞ phytoestrogens (10473)
    - ⊞ polybrominated ethers (2040)
    - ⊞ polyflourochemicals (5805)
    - ⊞ viral infection (32765)
    - ⊞ volatile compounds (6924)
  - ⊟ questionnaire (41474)
    - ⊞ alcohol use (14837)
    - ⊞ bacterial infection (17544)
    - ⊞ cognitive functioning (2975)
    - ⊞ disease (41474)
    - ⊞ food component recall (38941)
    - ⊞ housing (40897)
    - ⊞ immunization (40266)
    - ⊞ occupation (17323)

Figure 2. Screenshot of NHANES data hierarchy displayed in the PIC data browser tool. Variables are shown with sample sizes. Highlighted in the screen shot are all laboratory measures of dioxins, a type of environmental exposure assayed in serum.



NHANES (41474)
- demographics (41474)
- examination (39274)
- laboratory (37087)
  - acrylamide (7535)
  - aging (7827)
  - allergen test (8339)
  - bacterial infection (28672)
  - biochemistry (35768)
  - blood (33718)
  - cotinine (31136)
  - diakyl (7540)
  - dioxins (5073)
    - 123 1,2,3,4,6,7,8,9-ocdd (fg/g) (4943)
    - 123 1,2,3,4,6,7,8-hpcdd (fg/g) (4988)
    - 123 1,2,3,4,7,8-hxcdd (fg/g) (3100)
    - 123 1,2,3,6,7,8-hxcdd (fg/g) (4990)
    - 123 1,2,3,7,8,9-hxcdd (fg/g) (4977)
    - 123 1,2,3,7,8-pncdd (fg/g) (5029)

Analysis of ...\laboratory\dioxins\1,2,3,4,6,7,8,9-ocdd (fg/g) for subsets:

Comparison of ...\laboratory\dioxins\1,2,3,4,6,7,8,9-ocdd (fg/g)

123 1,2,3,4,6,7,8,9-ocdd (fg/g) (4943)

Subset 1	Subset 2
NHANES	NHANES
Mean: 856.37	Mean: 2,927.96
Median: 594.89	Median: 2,012.15
IQR: 436.65	IQR: 2,469.5
SD: 1,512.54	SD: 3,188.73
Data Points: 1607	Data Points: 3336

t statistic:	-30.979
p-value:	0.0000

The results are significant at a 95% confidence level.

Figure 3. Screenshot of Drag-and-drop example to explore the NHANES datasets (left) using the PIC cohort browser tool. A comparison of raw Dioxin (1-9-ocdd) levels by age groups. Red, age <25 y; blue, age ≥25 y.

race/ethnicity) that we believe will aid in planning analyses of the data. Second, we have built an online course (http://www.chiragjpgroup.org/exposome-analytics-course/) to guide users step-by-step through an investigation our group recently published (Patel *et al.*, 2016).

We plan to assess how frequently our data descriptor and data resources are being utilized by the scientific community through traditional means (e.g., number of citations to this descriptor), but also through by counting the number of unique visitors to the Vimeo video website, the web application (http://nhanes.hms.harvard.edu), and through feedback from course materials.

Code availability

We demonstrate 3 use-cases in using the integrated NHANES datasets in a R markdown source file (see 'Usage Notes'). Code is available on GitHub here: https://github.com/chiragjp/nhanes_scidata. One other example using our API access is available here: https://github.com/hms-dbmi/R-IRCT/blob/master/Example_NHANES.Rmd

I2b2/TranSMART software stack. Code to implement a PIC is open-sourced and available here: https://github.com/hms-dbmi/HMS-DBMI-transmartApp

Data Records

Data record 1: Integrated NHANES dataset and data dictionary in.csv format.

The integrated NHANES dataset and a data dictionary is available online at Dryad (Data Citation 1) as a .zip file which includes 3 .csv formatted files. The first file ('data file') contains each individual (as rows) surveyed in 1999–2006 with all of their measurements (as columns) ('MainTable', Fig. 1d). The second file contains a data dictionary file which contains the name of the variable as represented in the data file, a human readable description of the variable, the categories that the variable belongs to), and the levels of the categories (if a categorical variable) (Fig. 1e). The third file is a dictionary specifically for demographic information, such as describing the columns for age, sex, race/ethnicity, whether the participant was born in the US, education level, income level, and mortality information. Also, to facilitate analyses using the *R* programming language, we have provided a 4th file that contains all the files described above as a *R* data object in.Rdata format.

Technical Validation

The raw data contained herein are from the CDC NHANES. The CDC NHANES have performed extensive technical validation of their data described elsewhere (e.g., refs 30,31).

Usage Notes

The NHANES utilizes a 'multistage survey sampled' study design to ensure minority subgroups (e.g., Blacks, Mexican-American, elderly, pre-adolescents) of the population are appropriately represented in the dataset[32] and to optimize sampling resources. Therefore, statistical analyses need to take into account the structure of the sampling into account to provide accurate estimates of the population, such as means, standard errors, and correlations[33].

To demonstrate how to properly analyse NHANES data, we provide a *R* markdown files in our GitHub repository (https://github.com/chiragjp/nhanes_scidata) to re-create several relevant analyses.

Conducting an 'environment-wide association analysis' in all-cause mortality in NHANES

Previously, we conducted a data-driven search of environmental exposure factors associated with all-cause mortality known as an 'environment-wide association study'[28]. In the guide (https://github.com/chiragjp/nhanes_scidata/blob/master/User_Guide.Rmd), we describe how to associate one of the top findings, serum cadmium, with all-cause mortality using survey-weighted Cox proportional hazards regression.

Distribution of serum lead in in children: Accessing the NHANES in PIC-SURE API

In this guide (https://github.com/chiragjp/nhanes_scidata/blob/master/User_Guide_PIC.Rmd), we demonstrate how to access the NHANES data programmatically through the PIC-SURE API. In our example, we show how to query the API to estimate the quartiles of serum lead in the US population of all ages and aged under 18.

Redistributable analytics environment in Docker

The issue of reproducibility, replicability, and scalability in computational scientific research has been raised on multiple occasions[34,35]. We promote a reproducible practice by packaging the curated NHANES data (Data Citation 1) with an analytics environment comprised of R-3.3.0 (ref. 36) and the Rstudio-0.99.902 (ref. 37) web interface in addition to a custom R library for regression studies in a Docker container[38]. The packaged environment is publically available on Docker hub (https://hub.docker.com/r/chiragjp/nhanes_scidata/) and can be consistently deployed across local or cloud-based environments. We have provided these materials as a hands-on short course available here: http://www.chiragjpgroup.org/exposome-analytics-course/

References

1. Skinner, A. C., Perrin, E. M., Moss, L. A. & Skelton, J. A. Cardiometabolic Risks and Severity of Obesity in Children and Young Adults. *N. Engl. J. Med.* **373,** 1307–1317 (2015).
2. Menke, A., Casagrande, S., Geiss, L. & Cowie, C. C. Prevalence of and Trends in Diabetes Among Adults in the United States, 1988-2012. *JAMA* **314,** 1021–1029 (2015).
3. Ogden, C. L., Carroll, M. D., Kit, B. K. & Flegal, K. M. Prevalence of Childhood and Adult Obesity in the United States, 2011-2012. *JAMA* **311,** 806–814 (2014).
4. Kantor, E. D., Rehm, C. D., Haas, J. S., Chan, A. T. & Giovannucci, E. L. Trends in Prescription Drug Use Among Adults in the United States From 1999-2012. *JAMA* **314,** 1818–1830 (2015).
5. Patel, C. J. & Ioannidis, J. P. A. Studying the elusive environment in large scale. *J. Am. Med. Assoc.* **311,** 2173–2174 (2014).
6. Patel, C. J., Bhattacharya, J. & Butte, A. J. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS ONE* **5,** e10746 (2010).
7. Patel, C. J., Chen, R., Kodama, K., Ioannidis, J. P. A. & Butte, A. J. Systematic identification of interaction effects between genome- and environment-wide associations in type 2 diabetes mellitus. *Hum. Genet.* **132,** 495–508 (2013).
8. Patel, C. J. *et al.* Investigation of maternal environmental exposures in association with self-reported preterm birth. *Reprod. Toxicol.* **45,** 1–29 (2013).
9. Patel, C. J., Cullen, M. R., Ioannidis, J. P. A. & Butte, A. J. Systematic evaluation of environmental factors: persistent pollutants and nutrients correlated with serum lipid levels. *Int. J. Epidemiol.* **41,** 828–843 (2012).
10. Tzoulaki, I. *et al.* A Nutrient-Wide Association Study on Blood Pressure. *Circulation* **126,** 2456–2464 (2012).
11. Patel, C. J. *et al.* Systematic evaluation of environmental and behavioural factors associated with all-cause mortality in the United States National Health and Nutrition Examination Survey. *Int. J. Epidemiol.* **42,** 1795–1810 (2013).
12. Patel, C. J., Manrai, A. K., Corona, E. & Kohane, I. S. Systematic correlation of environmental exposure and physiological and self-reported behaviour factors with leukocyte telomere length. *Int. J. Epidemiol.* doi: 10.1093/ije/dyw043 (2016).
13. Patel, C. J., Ioannidis, J. P. A., Cullen, M. R. & Rehkopf, D. H. Systematic assessment of the correlations of household income with infectious, biochemical, physiological, and environmental factors in the United States, 1999-2006. *Am. J. Epidemiol.* **181,** 171–179 (2015).
14. Rappaport, S. M. & Smith, M. T. Environment and Disease Risks. *Science* **330,** 460–461 (2010).
15. Rappaport, S. M., Barupal, D. K., Wishart, D., Vineis, P. & Scalbert, A. The Blood Exposome and Its Role in Discovering Causes of Disease. *Environ. Health Perspect.* **122,** 769–774 (2014).
16. Bell, S. M. & Edwards, S. W. Identification and Prioritization of Relationships between Environmental Stressors and Adverse Human Health Impacts. *Environ. Health Perspect.* **123,** 1193–1199 (2015).
17. Park, S. K., Tao, Y., Meeker, J. D., Harlow, S. D. & Mukherjee, B. Environmental Risk Score as a New Tool to Examine Multi-Pollutants in Epidemiologic Research: An Example from the NHANES Study Using Serum Lipid Levels. *PLoS ONE* **9,** e98632 (2014).
18. Kohane, I. S., Churchill, S. E. & Murphy, S. N. A translational engine at the national scale: informatics for integrating biology and the bedside. *J. Am. Med. Inform. Assoc* **19,** 181–185 (2012).
19. Murphy, S. N. *et al.* Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J. Am. Med. Inform. Assoc* **17,** 124–130 (2010).
20. Murphy, S. N. *et al.* Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. *AMIA Annu. Symp. Proc* 548–552 (2007).
21. Athey, B. D., Braxenthaler, M., Haas, M. & Guo, Y. tranSMART: An Open Source and Community-Driven Informatics and Data Sharing Platform for Clinical and Translational Research. *AMIA Jt Summits Transl Sci Proc* **2013,** 6–8 (2013).
22. Canuel, V., Rance, B., Avillach, P., Degoulet, P. & Burgun, A. Translational research platforms integrating clinical and omics data: a review of publicly available solutions. *Brief. Bioinform.* **16,** 280–290 (2015).
23. Centers for Disease Control and Prevention (CDC) & National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data, 1999-2000. Available at http://www.cdc.gov/nchs/nhanes/nhanes99_00.htm.
24. Centers for Disease Control and Prevention (CDC) & National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data, 2001-2002. Available at http://www.cdc.gov/nchs/nhanes/nhanes01-02.htm.
25. Centers for Disease Control and Prevention (CDC) & National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data, 2003-2004. Available at http://www.cdc.gov/nchs/nhanes/nhanes2003-2004/nhanes03_04.htm.

26. Centers for Disease Control and Prevention (CDC) & National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data, 2005-2006. Available at http://www.cdc.gov/nchs/nhanes/nhanes2005-2006/nhanes05_06.htm.
27. Lumley, T. *survey: analysis of complex survey samples*, version 3.30 (2014).
28. Patel, C. J. *et al.* Systematic evaluation of environmental and behavioural factors associated with all-cause mortality in the United States National Health and Nutrition Examination Survey. *Int. J. Epidemiol.* **42,** 1795–1810 (2014).
29. Cowie, C. C. *et al.* Prevalence of diabetes and impaired fasting glucose in adults in the U.S. population: National Health And Nutrition Examination Survey 1999-2002. *Diabetes Care* **29,** 1263–1268 (2006).
30. National Centers for Health Statistics Centers for Disease Control and Prevention. *National Health and Nutrition Examination Survey Operations Manuals* (2015). Available at http://www.cdc.gov/nchs/nhanes/nhanes1999-2000/manuals99_00.htm. Accessed on 20 May 2016.
31. National Centers for Health Statistics, US Centers for Disease Control and Prevention. *National Health and Nutrition Examination Survey Laboratory Methods* (2010). Available at http://www.cdc.gov/nchs/nhanes/nhanes1999-2000/lab_methods_99_00.htm. Accessed on 20 May 2016.
32. National Centers for Health Statistics. *The National Health and Nutrition Examination Survey: Sample Design, 1999–2006* (US Centers for Disease Control and Prevention, 2012).
33. National Centers for Health Statistics. *National Health and Nutrition Examination Survey: Analytic Guidelines, 2011-2012* (US Centers for Disease Control and Prevention, 2013).
34. Dudley, J. T. & Butte, A. J. In silico research in the era of cloud computing. *Nat. Biotechnol.* **28,** 1181–1185 (2010).
35. Leek, J. T. & Peng, R. D. Opinion: Reproducible research can still be wrong: adopting a prevention approach. *Proc. Natl. Acad. Sci. USA* **112,** 1645–1646 (2015).
36. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2015).
37. Rstudio Team. *RStudio: integrated development for R*, version 0.99.92, https://www.rstudio.com/ (2016).
38. Kacamarga, M. F., Pardamean, B. & Wijaya, H. in *Intelligence in the Era of Big Data* Vol. 516 (eds Intan, R., Chi, C.-H., Palit, H. N. & Santoso, L. W.) 439–445 (Springer Berlin Heidelberg, 2015).

Data Citation

1. Patel, C. J. *Dryad Digital Repository* http://dx.doi.org/10.5061/dryad.d5h62 (2016).

Acknowledgements

This project is conducted with support from NIH Big Data to Knowledge (BD2K) grant (U54 HG007963). We thank Cassandra Perry of Harvard DBMI for producing and narrating the video tutorial. C.J.P. is additionally supported by a NIH/NIEHS R00 ES23054 R21 ES025052 and a gift from Agilent Technologies, Inc.

Author Contributions

C.J.P. downloaded and processed the NHANES data, wrote the Usage Notes R markdown guide, and wrote the manuscript. N.P. wrote the Usage Notes R markdown manuscript, wrote the manuscript. M.M. participated in developing the i2b2/tranSMART application and loaded NHANES data. J.E.M. developed BD2K PIC-SURE RESTful API. C.K. participated in loading NHANES data in i2b2/tranSMART. I.S.K. architected the i2b2 software infrastructure and wrote the manuscript. P.A. architected the extension to the i2b2/tranSMART platform to accommodate NHANES data, architected BD2K PIC-SURE API, and wrote the manuscript.

Additional Information

Table 3 is only available in the online version of this paper.

Competing financial interests: The authors declare no competing financial interests.

How to cite: Patel, C. J. *et al.* A database of human exposomes and phenomes from the US National Health and Nutrition Examination Survey. *Sci. Data* 3:160096 doi: 10.1038/sdata.2016.96 (2016).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.