

# Datasheet for NHANES dataset’\*

Sakura Hu

December 14, 2024

Extract of the questions from Gebru et al. (2021).

## Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

- The National Health and Nutrition Examination Survey (NHANES) was created to assess the health and nutritional status of adults and children in the United States. The survey combines both interviews and physical examinations to gather comprehensive health data (Pruim 2015). The primary task of NHANES is to provide vital statistics on the nation’s health, enabling public health experts, policymakers, and researchers to track trends in nutrition, health behaviors, and the prevalence of chronic diseases. The resampled version of NHANES data, which is utilized in this analysis, has been adapted specifically for educational purposes (Pruim 2015). This version of the data corrects for the oversampling of certain populations to ensure a more accurate representation of the U.S. population. The aim of this adjustment is to make the dataset suitable for educational use, enabling students and researchers to analyze and learn from it while avoiding potential biases introduced by the original survey’s complex sampling methods (Pruim 2015). This resampling version helps simplify the dataset and facilitates learning without compromising the integrity of the data collection process.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

- The NHANES dataset was created by the National Center for Health Statistics (NCHS), which is part of the Centers for Disease Control and Prevention (CDC). NCHS is responsible for producing and maintaining vital health statistics in the United States. The resampled data was assembled by Michelle Dalrymple of Cashmere High School and Chris Wild of the University of Auckland, New Zealand for educational purposes (Pruim 2015).

---

\*Code and data are available at: <https://github.com/xyccw/BMI>.

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

- The dataset itself is a public health initiative and does not have a specific grant associated with its creation. Instead, it is part of the ongoing federal program aimed at monitoring the health and nutritional status of the U.S. population.

4. *Any other comments?*

- No

## Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

- Each instances in the National Health and Nutrition Examination Survey (NHANES) dataset represent individuals (people) from the non-institutionalized civilian population of the United States (Pruim 2015).

2. *How many instances are there in total (of each type, if appropriate)?*

- The NHANES raw data contains 20,293 instances, while the NHANES resampled data contains 10,000 instances.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

- The NHANES dataset is a sample of the larger set, which is the non-institutionalized civilian resident population of the United States. The dataset does not include all possible instances of this population, but instead employs a complex survey design that uses multistage probability sampling to select participants. The larger set includes all individuals in the U.S. who live in households, excluding institutionalized populations like those in prisons or long-term care facilities.

The sample is designed to be representative of the larger U.S. population, but with oversampling of certain subpopulations such as racial minorities, low-income individuals, and the elderly. These oversampling strategies are employed to ensure that there is sufficient data to generate reliable estimates for these groups, which may be underrepresented in a simple random sample. For example, the NHANES oversamples non-Hispanic Black,

Hispanic, low-income White persons, and individuals aged 80 and older (Curtin et al. 2013).

To make the sample representative of the entire U.S. population, weights are applied to account for the oversampling and to adjust for differential probabilities of selection. This allows the data to be treated as if it were a simple random sample, and the weighted results can be generalized to the broader population. The representativeness of the sample is verified through the use of these weights and by ensuring that the sample covers various geographic regions of the U.S., including both metropolitan and non-metropolitan areas (Curtin et al. 2013). Additionally, the NHANES sampling design has been rigorously tested and adjusted to ensure that it accurately reflects the population in terms of demographic factors (Curtin et al. 2013).

4. *What data does each instance consist of? “Raw” data (for example, unprocessed text or images) or features? In either case, please provide a description.*
  - Each instance of the raw data consists of 76 variables. Variables include study variables such as participant identifier, demographic variables such as age, gender, race, as well as physical measurements such as weight and height. It also includes health data like sleep hours, cholesterol level, and lifestyle factors like smoking and alcohol consumption.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
  - There is no single, explicit label or target associated with each instance, as the dataset contains a wide range of variables that may be used depending on the specific analysis.
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
  - The NHANES dataset contains missing data, which can be categorized as either unit nonresponse or component/item nonresponse. Unit nonresponse occurs when an individual selected to participate in the survey does not respond during either the interview or examination phase (Johnson et al. 2013). For example, during the 2009-2010 cycle, of the 13,272 individuals eligible for participation, only 10,537 completed the household interview, leading to an interview nonresponse rate of 21% (Johnson et al. 2013). Furthermore, of the interviewed participants, only 10,253 participated in the Medical Examination Center (MEC) examination, contributing to an additional 2% of nonresponse during the examination phase (Johnson et al. 2013). This nonresponse is addressed by applying sample weights that adjust for the nonresponse (Johnson et al. 2013). Component/item nonresponse refers to missing data from specific survey components, such as not all participants opting

to complete a particular health examination (e.g., a blood pressure measurement). Additionally, NHANES uses specific codes for missing values, such as a period (.) for numeric variables and blank spaces for character variables (Johnson et al. 2013). Responses like “refused” or “don’t know” are also treated as missing and are assigned specific codes (e.g., 7, 77, or 777 for “refused” and 9, 99, or 999 for “don’t know”) (Johnson et al. 2013). These should also be treated as missing data to avoid distorting analysis results.

7. *Are relationships between individual instances made explicit (for example, users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

- There are no relationships between individual instances made explicitly in the NHANES dataset.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

- there are no formal recommendations for typical splits such as training, development/validation, and testing. However, given the structure and design of NHANES, there are important considerations for subsetting and combining data that influence how data is prepared for analysis. Since NHANES uses a complex, multi-stage sampling design, any analysis must account for the sampling weights and stratification to avoid bias (Johnson et al. 2013).

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

- One prominent issue is nonresponse bias, which occurs when certain individuals or groups fail to participate in the survey or specific components of the survey (e.g., medical examinations). This nonresponse can introduce bias if the missing data is systematically related to key variables. For example, older individuals or those with lower income may be less likely to respond, resulting in underrepresentation of these groups. NHANES mitigates this through the use of sampling weights that adjust for nonresponse, but analysts must still carefully account for this in their analysis. Another source of noise is measurement error. NHANES collects data through a combination of self-reported surveys and physical examinations. Self-reported data, such as dietary intake, physical activity, or smoking habits, are prone to recall bias and social desirability bias. These errors can introduce variability and reduce the reliability of the data. On the other hand, while data from physical examinations and laboratory tests are generally more reliable, they are not immune to technical or procedural errors during measurement or data entry.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
  - The NHANES dataset itself is self-contained and does not link to or otherwise rely on external resources.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
  - The NHANES dataset does not contain data that might be considered confidential in the publicly available versions.
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
  - The NHANES dataset does not contain data that, if viewed directly, would generally be considered offensive, insulting, or threatening. However, it includes sensitive health-related information such as data on weight, body mass index (BMI), chronic diseases, substance use (e.g., smoking, alcohol, and drug use), mental health, sexual history, and socioeconomic status. While this data is anonymized and presented in an aggregated form, its content could potentially cause discomfort, anxiety, or distress to certain individuals, particularly if the topics touch on personal or stigmatized experiences.
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
  - Yes, the NHANES dataset identifies several subpopulations based on demographic and socioeconomic characteristics.
  - Age is recorded in years and categorized into decades. The dataset includes participants across a wide age range, with individuals aged 80 and over intentionally oversampled to ensure adequate representation of this subgroup (Curtin et al. 2013).
  - Race is categorized as Mexican, Hispanic, White, Black, Asian, or Other. Non-Hispanic Black persons and Hispanic persons are intentionally oversampled to improve the reliability and precision of estimates for these subpopulations (Curtin et al. 2013).

- Income is categorized using a poverty ratio, which reflects a range of socioeconomic statuses, with specific attention given to low-income white groups (Curtin et al. 2013).
  - Gender is categorized as male and female, with the dataset typically having an approximately equal distribution of males and females (Curtin et al. 2013).
14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
- No, it is not possible to identify individuals from the dataset. Data were edited such that all direct personal identifiers, as well as any characteristics that could lead to identification, were omitted from the data sets (Zipf et al. 2013).
15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
- The NHANES dataset contains data that could be considered sensitive in several ways. This includes data on race and ethnicity, which are categorized as Mexican, Hispanic, White, Black, Asian, or Other. While these categories are used for public health research purposes, their inclusion may be viewed as sensitive because such data could potentially be misused or misinterpreted outside its intended context.

The dataset also includes socioeconomic data, such as household income levels and poverty ratios. These variables reveal participants' financial statuses, which can be considered private and sensitive. Additionally, health-related data, including medical diagnoses, physical measurements (e.g., body mass index, blood pressure), and laboratory test results, comprise a significant portion of the dataset. These health metrics may reveal personal and potentially stigmatizing information about an individual's physical or mental health.

Furthermore, the dataset includes biometric data such as weight, height, and blood test results. While NHANES does not include genetic data or government identification numbers, the presence of health and demographic variables may still pose confidentiality risks if improperly handled. However, strict measures are in place to protect participant privacy, including anonymization, ethical review processes, and federal confidentiality protections (Zipf et al. 2013).

16. *Any other comments?*

- No

## Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

- The data associated with each instance in the NHANES dataset was primarily acquired through both directly observable measurements and self-reported survey responses. For instance, health data such as blood pressure, body mass index (BMI), and laboratory test results were directly measured during health examinations conducted at mobile examination centers (MECs) (Zipf et al. 2013). These direct observations are objective and typically subject to standardized procedures to ensure accuracy and consistency.

In contrast, demographic and lifestyle information (e.g., age, gender, race, smoking habits, physical activity) were self-reported by participants through structured interviews (Zipf et al. 2013). This data is subjective and relies on the accuracy and honesty of the respondents.

To enhance the reliability of the dataset and minimize biases, self-reported data were verified and validated through multiple approaches. First, cross-validation with objective measures was performed whenever possible (Zipf et al. 2013). Standardized questionnaires, designed based on validated instruments and periodically reviewed, were administered by trained interviewers to ensure consistency in data collection (Zipf et al. 2013). Built-in logic checks identified implausible or inconsistent responses during interviews, prompting clarifications or follow-up questions (Zipf et al. 2013). Proxy reporting was allowed when participants, such as children, were unable to respond directly.

Sample weights were applied to correct for unequal probabilities of selection and nonresponse (Zipf et al. 2013). For example, oversampling of subpopulations, including racial minorities and low-income groups, was performed, and sample weights adjusted for these techniques. This comprehensive approach to verification and adjustment reduces bias and enhances the validity of the dataset.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

- Health Examinations and Laboratory Tests: Data such as blood pressure, height, weight, BMI, and other health measurements were collected during in-person visits to mobile examination centers (MECs). These centers are equipped with specialized medical equipment to directly measure participants' health metrics. The collection of these data points is validated by well-established clinical protocols and the use of calibrated equipment to ensure the accuracy of the measurements (Zipf et al. 2013).

- Survey Interviews: Participants also answered detailed surveys that included demographic, socioeconomic, and lifestyle-related questions. This data was gathered through structured interviews conducted by trained personnel using electronic devices such as tablets (Zipf et al. 2013). The interviewers followed strict protocols to ensure uniformity across interviews and minimize errors (Zipf et al. 2013). The responses provided by participants were validated by cross-referencing answers within the survey and ensuring consistency across different sections (Zipf et al. 2013).
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
- The dataset is a sample drawn from the larger U.S. civilian noninstitutionalized population, and the sampling strategy is specifically designed to ensure the sample is representative of the broader population.
  - Sample selection for NHANES followed 4 stages:
  - Primary Sampling Units (PSUs): In the first stage, primary sampling units, which are typically counties or groups of contiguous counties, are selected with probability proportional to size (PPS) (Zipf et al. 2013). This means that PSUs with larger populations have a higher probability of being selected (Zipf et al. 2013). This approach helps ensure that the sample reflects the distribution of the population, with oversampling occurring for certain groups of interest, such as racial minorities, to improve the reliability of estimates for these subgroups (Zipf et al. 2013).
  - Segmentation: After selecting the PSUs, the next stage involves dividing the selected PSUs into smaller segments, typically geographic units such as city blocks (Zipf et al. 2013). These segments are then chosen using PPS to ensure an adequate representation of different population groups (Zipf et al. 2013).
  - Dwelling Units (DUs): The third stage involves selecting dwelling units (households) within the chosen segments (Zipf et al. 2013). In areas with higher concentrations of specific groups, such as low-income or minority populations, these households are selected at a higher rate to ensure proper representation of these groups (Zipf et al. 2013).
  - Selection of Individuals: In the final stage, individuals are randomly selected from within the households (Zipf et al. 2013). The selection is done within designated age, sex, and race/ethnicity subdomains to achieve a balance across these demographic characteristics (Zipf et al. 2013). This step ensures that the sample includes a diverse range of individuals and that subgroups, such as the elderly or certain racial/ethnic populations, are adequately represented (Zipf et al. 2013).
  - To ensure that specific demographic subgroups are sufficiently represented, NHANES uses oversampling techniques. For example, non-Hispanic black, Hispanic, and low-income white individuals, as well as those aged 80 and older, are oversampled (Zipf et al. 2013). This oversampling allows for more precise estimates for these groups, which may otherwise be underrepresented in a simple random sample.



- Due to the complex sampling design, sample weights are applied to adjust for unequal probabilities of selection and to account for nonresponse (Zipf et al. 2013). These weights allow analysts to produce nationally representative estimates, as they correct for the oversampling of certain subgroups and for any nonresponse bias.
4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
    - The data collection for the NHANES dataset was carried out by a team of professionals and experts under the guidance of the National Center for Health Statistics (NCHS), which is part of the Centers for Disease Control and Prevention (CDC). The team responsible for data collection consisted of field interviewers, medical personnel, and technicians who worked in the mobile examination centers (MECs) as well as in the field during household interviews. They were compensated with salaries or hourly wages based on their roles and responsibilities.
  5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
    - The data used in this paper was collected as part of the 2009-2010 NHANES survey cycle, and it matches the creation timeframe of the data associated with the instances.
  6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - Yes, ethical review processes were conducted for the NHANES dataset. The NHANES protocol was developed and reviewed to be in compliance with the U.S. Department of Health and Human Services (HHS) Policy for Protection of Human Research Subjects, as outlined in 45 CFR part 46, which can be accessed at [the official website of the U.S. Department of health and human services](#).
  7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
    - I did not collect the data, I obtain it through the NHANES package.
  8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

- Before being approached by NHANES staff, participants receive an advance letter. This letter outlines the purpose of the survey, the types of data to be collected, and reassures individuals about the confidentiality of their information. Example text in the letter includes details about why the participant was chosen and how their data will contribute to public health research.

**NHANES - Attachments to Supporting Statement - Attachment 4**

***Advance Letter***

You or a member of your family may have a chance to take part in an important national health survey.

The National Center for Health Statistics, a part of the Centers for Disease Control and Prevention, is responsible for this survey—the National Health and Nutrition Examination Survey. This survey teaches us about the health and diet of people in the United States. Over the years, this survey has led to changes in the foods we eat and the health care we receive.

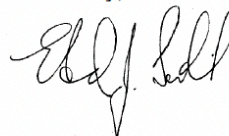
In the next few weeks, one of our health representatives may visit your home. She or he will show you official Centers for Disease Control (CDC) identification and ask some short, easy questions about you and other family members. This interview will take only a few minutes of your time. The purpose of these questions is to see if you or a member of your family will be asked to participate in the survey. Answering the questions is completely voluntary, and you may choose not to answer any questions. I assure you there will be no penalties or loss of benefits of any kind from refusing to answer.

If you or other household members are chosen, we hope you will participate in the survey. You will be among the many people in towns and cities across the country that help us increase our knowledge about the health of people in the United States.

This survey is a Federal program authorized by the Public Health Service Act. All of your answers will be kept in strict confidence. We will use this information only for statistical research and reports. Your answers will be added to others, so no one can identify which are yours.

Thank you in advance for helping.

Sincerely,

A handwritten signature in black ink, appearing to read 'Ed J. Sondik', written in a cursive style.

Edward J. Sondik, Ph.D.  
Director

P.S. If you have questions call Mr. George Zipf of my staff at 1-800-452-6115. The call is free, and we would be happy to answer your questions.

Figure 1: advance

9. *Did the individuals in question consent to the collection and use of their data? If so,*

*please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

- Before any data collection begins, participants are required to sign a written informed consent form. These forms clearly state that participation is voluntary, it provide assurances about privacy and confidentiality, highlighting that data will be used only for research purposes, and offer participants the opportunity to ask questions and withdraw at any point without any repercussions.

First	Middle	Last
-------	--------	------

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
  - yes, NHANES emphasizes that participation is entirely voluntary. Participants are explicitly told that they can refuse to participate in any part of the survey or withdraw from the study entirely at any point during the data collection process without facing any negative consequences. This information is shown at [participant FAQ](#).
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
  - No.
12. *Any other comments?*
  - No.

### Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
  - NHANES uses specific codes for “refused” and “don’t know” responses, assigning values such as 7, 77, or 777 for refusals, and 9, 99, or 999 for “don’t know” responses (Johnson et al. 2013).
2. *Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.*
  - No
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
  - No
4. *Any other comments?*
  - No

### Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*



- Yes, the NHANES dataset has been used extensively in a variety of tasks related to public health research and analysis. Since its inception, the dataset has been used to assess the health and nutritional status of the U.S. population, providing valuable insights into a wide range of health indicators, including obesity, diabetes, cardiovascular diseases, and other chronic conditions. Researchers have used NHANES data to analyze the prevalence of different health conditions, identify risk factors, evaluate the impact of lifestyle choices (such as diet, physical activity, and smoking) on health outcomes, and inform public health policies.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
- Some of the products can be found [here](#).
3. *What (other) tasks could the dataset be used for?*
- The NHANES dataset offers a wide range of potential uses beyond its current applications in health research. It can be used for longitudinal studies to track health trends, predictive modeling for disease prevention, and analysis of health disparities related to socioeconomic factors. Additionally, the dataset is valuable for exploring the relationship between diet and chronic disease, studying environmental health impacts, and evaluating medication use. It can also be applied in machine learning for disease prediction and in public health interventions aimed at changing health behaviors. Overall, NHANES supports diverse research areas, from epidemiology to policy simulation.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
- No
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
- The NHANES dataset should not be used for any tasks that involve attempting to identify individuals or establishments, as doing so would violate confidentiality agreements and ethical guidelines. Specifically, the dataset should not be used for tasks such as re-identifying individuals, linking it with other datasets that contain personally identifiable information, or attempting to assess the effectiveness of privacy protection methodologies. The data is intended solely for statistical reporting and analysis, and any attempt to discern the identity of participants or establishments could lead to legal, ethical, and privacy concerns. Users must also ensure

that the dataset is not used in ways that compromise the confidentiality of the individuals or establishments involved.

6. *Any other comments?*

- No

## Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

- Yes, the NHANES dataset is publicly available and can be distributed to third parties. It is released by the National Center for Health Statistics (NCHS) under the Centers for Disease Control and Prevention (CDC) for public use. The dataset is accessible via the CDC's NHANES website, where anyone can download it for statistical analysis and research purposes. However, there are specific guidelines on how the data should be used, including the restriction that it should only be used for statistical reporting and analysis and should not be linked to individually identifiable data.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

- The data can be downloaded directly from the NHANES page on the CDC official website.

3. *When will the dataset be distributed?*

- The NHANES 2009-2010 data were made available after the data collection period ended.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

- The NHANES dataset is provided under strict confidentiality protections, and data use is limited to statistical analysis. There are severe legal penalties for violating the confidentiality and data use conditions. Users must agree to these terms before using the dataset, ensuring compliance with the applicable laws and ethical guidelines.

– Data Use Agreement can be found at [this website](#)

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*



- No
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
- There are no export controls or regulatory restrictions like those typically imposed on military or sensitive commercial data. The NHANES dataset is governed by federal confidentiality laws, and users are required to adhere to data use restrictions as outlined in the Data Use Agreement.
7. *Any other comments?*
- No

## Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
- The NHANES dataset is supported, hosted, and maintained by the National Center for Health Statistics (NCHS), which is part of the Centers for Disease Control and Prevention (CDC).
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
- You can contact NHANES staff by calling 1 (800) 232-4636.
3. *Is there an erratum? If so, please provide a link or other access point.*
- No
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
- Updated will be informed in the CDC official website.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
- There is no limitation on this.
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
- Older version of dataset is hosted by CDC, and any updated will be informed in the CDC official website.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

- If someone wants to contribute, they are welcome to contact NHANES staffs.

8. *Any other comments?*

- No

## References

- Curtin, Lester R, Leila K Mohadjer, Suzanne M Dohrmann, et al. 2013. “National Health and Nutrition Examination Survey: Sample Design, 2007–2010.” *Vital Health Stat 2* 160.
- Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. “Datasheets for Datasets.” *Communications of the ACM* 64 (12): 86–92.
- Johnson, Claudia L, Rupa Paulose-Ram, Cynthia L Ogden, et al. 2013. “National Health and Nutrition Examination Survey: Analytic Guidelines, 1999–2010.” *Vital Health Stat 2* 161.
- Pruim, Randall. 2015. *NHANES: Data from the US National Health and Nutrition Examination Study*. <https://CRAN.R-project.org/package=NHANES>.
- Zipf, George, Michele Chiappa, Kimberly S Porter, et al. 2013. “National Health and Nutrition Examination Survey: Plan and Operations, 1999–2010.” *Vital Health Stat 1* 56.