Datasheet for Decoding Sleep Disorders Through Self-Reported Patterns*

Terry Tu

April 18, 2024

This study uses logistic regression to analyze the relationship between self-reported sleep behaviors and the likelihood of reporting sleep troubles, using data from the NHANES 2017 to March 2020 Sleep Disorders dataset. We examine factors such as snoring frequency, daytime sleepiness, and sleep duration on weekdays and weekends. Findings indicate that excessive snoring and high daytime sleepiness are strongly associated with more reported sleep issues, while longer weekend sleep correlates with fewer reports. These results highlight the importance of good sleep hygiene for overall health and provide insights into how daily behaviors affect sleep quality.

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

1 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 dataset was created to examine the relationships between various sleep-related factors and reported sleep troubles. It aims to analyze the influence of sleep duration, sleep habits, and sleep disorders based on data collected prior to the COVID-19 pandemic disruptions.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The dataset was compiled by the National Center for Health Statistics (NCHS) as part of the National Health and Nutrition Examination Survey (NHANES).

*Code and data supporting this analysis are available at https://github.com/TEJMaster/Sleep-Disorder-Analysis

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 creation was funded by the Centers for Disease Control and Prevention (CDC), an agency within the Department of Health and Human Services (HHS) of the United States.

4. *Any other comments?*

    - The dataset includes important adaptations to ensure continuity and representativeness despite disruptions caused by the COVID-19 pandemic.

2 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 instance represents data from an individual participant aged 16 or older, detailing their sleep patterns and associated disorders as part of the NHANES 2017-March 2020 cycle.

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

    - The exact number of instances corresponds to the number of participants surveyed during the NHANES cycles from 2017 to March 2020. However, specific numbers can be derived from the NHANES dataset documentation.

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).*

    - This dataset is a nationally representative sample combining two NHANES cycles due to the incomplete data collection caused by the COVID-19 pandemic. Special weighting processes were applied to ensure representativeness.

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 includes structured data about the participant's usual sleep duration on weekdays and weekends, sleep habits, and reported sleep disorders, among other health-related information.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

    - The primary labels include sleep duration (weekdays and weekends) and the presence of sleep disorders, used to analyze patterns and correlations in sleep behavior.

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.*

    - Some instances may have missing data due to non-response or inconsistencies during data collection, which are handled according to NHANES data processing protocols.

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.*

    - No explicit relationships between instances are noted; each instance is treated as independent for the purposes of public health analysis.

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.*

    - No specific data splits are recommended by NHANES; researchers may determine splits based on their analytical needs.

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

    - Errors and noise might be present due to the self-reported nature of some data and the limitations in data collection methodologies, as noted in the NHANES documentation.

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 dataset is self-contained and does not rely on external resources for its primary use. However, additional details and guidelines can be accessed through the NHANES website.

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.*

    - No confidential data is included; all NHANES data is anonymized and made public for research purposes under strict ethical guidelines.

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.*

    - No, the dataset contains health-related information that is scientific and non-offensive in nature.

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.*

    - The dataset includes demographic variables such as age, gender, and ethnicity, allowing for detailed subgroup analyses within the population.

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.*

    - NHANES data is designed to be anonymous, with no direct or indirect identifiers included in the dataset.

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 dataset includes sensitive health data, which is handled according to the ethical guidelines and privacy policies mandated by the CDC and HHS.

16. *Any other comments?*

    - Researchers are advised to follow appropriate ethical guidelines when utilizing this dataset, particularly when dealing with sensitive health information.

3 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.

- Data was directly collected through in-person interviews and examinations using the Computer-Assisted Personal Interview (CAPI) system, ensuring high data quality and reliability. The interviews were conducted by trained personnel, and the collected data underwent quality checks to validate the accuracy.

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?*

   - The data was collected using standardized procedures and the CAPI system, which includes built-in consistency checks. Additional quality control measures included reviewing a percentage of the audio recordings of the interviews to validate data accuracy.

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 nationally representative sample designed to reflect the U.S. population, using stratified multistage sampling. The sample includes individuals from various demographics and geographic locations across the country.

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)?*

   - Data collection was performed by trained NHANES staff who are part of the NCHS. These staff members are trained healthcare professionals and data collectors employed by the government, and compensation is part of their regular salaries.

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 was collected from 2017 to March 2020, directly matching the timeframe of the dataset instances. The data reflects the health statuses and conditions of the respondents during this specific period before the COVID-19 pandemic.

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, the NHANES study protocols are reviewed by the NCHS Ethics Review Board (ERB) to ensure compliance with ethical standards and protection of participants.

The ERB reviews all aspects of the NHANES, including data collection, consent processes, and the use and storage of data.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - Data was collected directly from the individuals through structured interviews and physical examinations conducted by NHANES personnel.

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.*

    - Participants were fully informed about the nature of the study, the types of data collected, and the use of the data. They were provided with detailed consent forms that explained the data collection process and their rights as participants, including privacy protections and the voluntary nature of their participation.

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.*

    - Yes, informed consent was obtained from all participants. The consent process involved participants reading the consent form and having the opportunity to ask questions. Consent was documented electronically in the CAPI system. The consent form detailed the participant's rights, the confidential handling of their data, and the purposes of the study.

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).*

    - Participants were informed that they could withdraw from the study at any time without any penalty or loss of benefits to which they were otherwise entitled. The NHANES staff provided contact information for participants to use if they chose to withdraw their consent.

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.*

    - An impact analysis was conducted to assess the potential risks associated with data collection and use. The analysis considered privacy risks, data security, and the impact of data collection on participants. Measures were implemented to mitigate

any identified risks, ensuring the protection of participant data and compliance with federal regulations.

12. *Any other comments?*

    - The NHANES team is committed to maintaining high ethical standards and transparency in its data collection and research processes. The study's methodology and results are regularly published in scientific journals and presented at conferences, contributing to public health knowledge and practice.

4 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.*

    - Yes, preprocessing included the cleaning of data to correct common entry errors, such as the confusion of AM and PM in time entries and the transposition of digits. Instances with extreme or implausible sleep duration times were either corrected or marked as missing based on additional information from interviewer comments.

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.*

    - The raw data, as collected, is maintained alongside the processed dataset to allow for validation of processing steps and to enable further analysis if required. Both versions of the dataset are available through the NHANES website.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - The software tools used for data processing, primarily statistical and data management software supported by the CDC, are standardized and not specifically available for public use. However, the methodologies and procedures used are detailed in the NHANES documentation.

4. *Any other comments?*

    - The preprocessing steps are designed to ensure that the data is reliable and suitable for analysis, with a focus on maintaining the integrity and usability of the data for health research.

5 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.*

    - The dataset is publicly available and can be accessed by researchers, policymakers, and the public. It is intended to support a wide range of health-related research and policy-making.

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

    - The dataset is distributed through the NHANES website, where users can download the data files directly. There is no DOI assigned to the dataset; however, each cycle of NHANES data is uniquely identified and can be cited using the specific cycle reference.

3. *When will the dataset be distributed?*

    - The dataset is already available for public access following its release in July 2021.

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 dataset is distributed under the terms of the Centers for Disease Control and Prevention's public use data policy, which allows for unrestricted use of the data for research and educational purposes. There are no fees associated with accessing the data.

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.*

    - There are no third-party IP restrictions on the data. The NHANES data is fully controlled and distributed by the CDC, ensuring open access.

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.*

    - No export controls or regulatory restrictions apply to the dataset. The data is intended for public use and is accessible worldwide.

7. *Any other comments?*

- The NHANES program is committed to providing high-quality, accessible data that can inform health research and policy. The program encourages the use of its data to advance public health knowledge and practices globally.

6 Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The 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)?*
    - The NCHS can be contacted through their website, where users can find contact information for various departments and services.
3. *Is there an erratum? If so, please provide a link or other access point.*
    - Updates, corrections, and errata related to the NHANES data are posted on the NHANES website. Users can access this information to ensure they are using the most accurate and up-to-date data.
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)?*
    - The NHANES data is updated periodically as new data collection cycles are completed. Updates are managed by the NCHS and are communicated through the NHANES website and through official CDC communications.
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.*
    - The data collected in NHANES is retained indefinitely as part of the historical and scientific record of population health. Data retention policies are governed by federal regulations and guidelines, which ensure the protection of participant information and the long-term availability of the data for research.
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 versions of the NHANES data are maintained and remain accessible to the public. This ensures that researchers can access historical data for longitudinal studies and comparisons across different time periods.

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.*

    - While external contributions to the dataset itself are not typically solicited, the NHANES program encourages researchers to use the data in their studies and to publish their findings. Contributions to the broader scientific community through research and publications based on NHANES data are welcomed and supported by the NCHS.

8. *Any other comments?*

    - The NHANES dataset is a critical resource for health research in the United States and globally. It provides invaluable insights into the health status, lifestyle, and risk factors of the U.S. population and serves as a foundation for public health policy and interventions.

References

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.