Datasheet for the Children's Depression*

Boxuan Yi

November 30, 2024

Extract of the questions from TIMNIT GEBRU (2021).

The information provided below is sourced from:

https://www.census.gov/programs-surveys/nsch/data/datasets.html (US Census Bureau 2024a)

https://www.census.gov/programs-surveys/nsch/technical-documentation/complete-technical-documentation.html (US Census Bureau 2024b)

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.*
   - To collect information on factors related to the well-being of children, including access to and quality of health care, family interactions, parental health, school and after-school experiences, and neighborhood characteristics.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - The U.S. Census Bureau conducted the survey on behalf of the Health Resources and Services Administration's Maternal and Child Health Bureau.
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.*
   - Funding and direction for this survey was provided by the Health Resources and Services Administration's Maternal and Child Health Bureau (HRSA MCHB).
4. *Any other comments?*
   - No.

*Code and data are available at: https://github.com/Elaineyi1/Children_Health

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.*

   - The instances represent individuals, with each row containing information about one child, defined as being between 0 and 17 years old.

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

   - Each instance represents an individual child at an individual level.

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 dataset is a sample representing the broader population of children in the U.S. The 2023 NSCH surveyed approximately 385,000 addresses. One child from each household with children was selected, or subsampled, to participate in the topical questionnaire. To enhance the representativeness of the sample, the weights for each child were calculated using base weights and various adjustments.

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 row contains informationa about access to and quality of health care, family interactions, parental health, school and after-school experiences, and neighborhood characteristics.

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

   - In the topical/follow-up survey, children were categorized into three age groups. T1 included children aged 0 through 5, T2 included children aged 6 through 11, and T3 covered children aged 12 through 17. This classification helped structure the survey data to better understand and address the different developmental and social needs of children at various stages of childhood and adolescence.

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.*

   - No.

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. One child was selected from one household to conduct the topical survey.

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.

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

   - No. The team excluded those observations with errors or redundacies.

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.*

    - Yes, the dataset is self-contained.

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 dataset included many specific questions related to health conditions. The Census Bureau and HRSA MCHB tooke extraordinary measures to assure that the identity of survey subjects cannot be disclosed. All direct identifiers, as well as characteristics that might lead to identification, had been omitted from the dataset.

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.

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. Demographic information including age, race, and sex were used for weight adjustments.

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.

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.*

    - This survey was designed to investigate children well-being, so it included many specific questions related to health conditions and demographics.

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 2023 NSCH retained a two-phase data collection approach: (1) an initial household screener to assess the presence, basic demographic characteristics, and special health care needs status of any children in the home; and (2) a substantive topical questionnaire to be completed by a parent or caregiver of the selected child. Respondents had four options to respond to the survey and receive assistance including: Web Instrument (English and Spanish), Paper Instrument (English and Spanish), Telephone Questionnaire Assistance (TQA), and Email Questionnaire Assistance (EQA).

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?*

    - Data was collected through surveys. Respondents were usually parents.

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 2023 NSCH sampled approximately 385,000 addresses to participate in the survey. The sample frame uses administrative records-based flags to identify four mutually exclusive strata. Stratum 1A and 1B: Addresses directly linked to children through administrative records are placed in Stratum 1, which includes about 80% of households with children. Within Stratum 1, if the linked child is 5 years old or younger, the address is assigned to Stratum 1A; if the child is older, the address is assigned in Stratum 1B. Stratum 2a: Addresses that are probabilistically linked to children. Approximately 15% of these addresses are households with children. Stratum 2b: The remaining addresses. Less than 5% of these addresses are households with children. The sampling rates by strata in each state were optimized to maximize the number of households with children in each state without compromising the reliability of survey estimates. The sample was distributed across states to produce a roughly equal number of completed interviews per state. Fourteen states included an oversample to increase the number of interviews completed in those states.

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)?*

   - Not mentioned.

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.*

   - June 23, 2023 to January 19, 2024. Yes.

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.*

   - Not mentioned. The Census Bureau and HRSA MCHB tooke extraordinary measures to assure that the identity of survey subjects cannot be disclosed. All direct identifiers, as well as characteristics that might lead to identification, had been omitted from the dataset. Before releasing any statistics to the public, the Census Bureau reviews them to make sure none of the information or characteristics could identify someone.

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

   - Directly.

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.*

- Yes, all individuals included in the dataset were notified as they completed the questionnaire. This process ensured that the participants were aware of their involvement and the purpose of the survey.

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, respondents completed the questionnaire. This process ensured that the participants were aware of their involvement and the purpose of the survey.

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).*

    - Not mentioned.

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.*

    - Not mentioned.

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.*

    - Web and paper survey responses were cleaned for analysis, including unduplication of responses, edits for data quality, creating standardized and derived variables, and imputation of missing values.

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.*

    - Not mentioned. But users must use the data in this data set for statistical reporting and analysis only, make no use of the identity of any person discovered, inadvertently or otherwise, not link this data set with individually identifiable data from any other Census Bureau or non- Census Bureau data sets.

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.*

    - No

3. *What (other) tasks could the dataset be used for?*

    - Not mentioned, but users should not link this data set with individually identifiable data from any other Census Bureau or non-Census Bureau data sets.

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.*

    - No

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.*

- No.

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

   - Not mentioned.

3. *When will the dataset be distributed?*

   - Not mentioned.

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.*

   - Not mentioned.

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.*

   - Not mentioned.

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.*

   - Not mentioned.

7. *Any other comments?*

   - No

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

   - US Census Bureau.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

   - Contact the Policy Coordination Office toll-free at 1-800-923-8282.

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)?*

   - No, but the National Survey of Children's Health is conducted frequently.

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.*

   - Not mentioned.

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.*

   - The older datasets from 2016 to 2022 can still be found on the 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.*

   - No.

8. *Any other comments?*

   - No.

References

TIMNIT GEBRU, BRIANA VECCHIONE, JAMIE MORGENSTERN. 2021. “Datasheets for Datasets.” https://dl.acm.org/doi/pdf/10.1145/3458723.

US Census Bureau. 2024a. “NSCH Datasets.” https://www.census.gov/programs-surveys/nsch/data/datasets.html.

———. 2024b. “Technical Documentation Complete List.” https://www.census.gov/programs-surveys/nsch/technical-documentation/complete-technical-documentation.html.