Datasheet for the Prenatal Mental Health Dataset*

Boxuan Yi

31 March 2024

Extract of the questions from TIMNIT GEBRU (2021).

The information provided below is sourced from:

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10339202/ (Catherine Lebel 2023)
- https://osf.io/ha5dp/ (Gerald Giesbrecht 2023).

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 survey aimed to explore the associations among exposure to objective hardship caused by the pandemic, perceived stress and psychological distress in pregnant individuals, as well as developmental outcomes in their offspring. Although there is evidence that disasters increase symptoms of mental illness, research on the mental health consequences of pandemics is limited.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - Authors are Gerald Giesbrecht, Catherine Lebel, and Lianne Tomfohr-Madsen from the University of Calgary.
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.*
    - This project was supported by funds from the Owerko Center at the Alberta Children's Hospital Research Institute.
4. *Any other comments?*

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

- 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.*

    - The instances represent individuals, categorized into three types: ID, pregnant participants, and birth outcomes.

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

    - There are 3 instances and 16 variables in the dataset. One variable is the ID, nine variables are information about pregnant participants, and six contain birth outcomes.

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).*

    - Several surveys and follow-ups were conducted over time, but only one finalized dataset is available for open access, which is the dataset currently being utilized. Surveys were released at different stages. Apart from the variables included in the dataset being used, there are additional variables such as the health information of the new born.

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.*

    - The ID instance contains the unique ID assigned to each participant. The pregnant individuals instance contains demographic information and mental health data. The birth outcome instance include birth length, birthweight, and whether the new born was admitted to the NICU.

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

    - No.

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.*

- Other than the variables included in the dataset being used, there are additional variables such as the health information of the new born, which are not open-access and hence missing.

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.*

   - Yes, the pregnant individuals were the mothers who gave birth to the newborns, and the birth outcomes are data specific to these newborns.

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 researchers excluded those observations with errors.

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. However, the data could be compared to pre-pandemic datasets to explore the impacts of the pandemic.

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, and all the participants provided permission for the use of their data.

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, education, and household income was collected. These data could help understand how age and socioeconomic status may influence an individual's level of depression and anxiety level during the pandemic.

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 prenatal mental health, so health data was included. The use of language, whether English or French, is also collected. This may indicate whether the participant is from Quebec or not.

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.*

    - Data collection was conducted via online surveys. All the responses were self-reported or parent-reported.

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 online questionnaires using REDCap. Participants were recruited via advertisements distributed through pregnancy organizations and care providers, social media, and via paid ads on Facebook and Instagram.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

- Participants were recruited via advertisements distributed through pregnancy organizations and care providers, social media, and via paid ads on Facebook and Instagram, with extra paid-ads focusing on underrepresentation groups.

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.*

    - April 2020 to April 2021. 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.*

    - No. All participants in this study provided informed consent and signed the electronic contract form. Research was approved by the University of Calgary conjoint health research ethics board (CHREB; REB20-0500) and conduced in accordance with the Declaration of Helsinki.

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

    - The dataset was available on the Open Science Framework at https://osf.io/ha5dp/

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.*

    - All participants provided informed consent for research and signed the electronic consent form before starting 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.*

    - All participants signed the electronic consent form before proceeding to the first questionnaire.

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.*

    - Cases that seem ineligible, with insufficient information, duplicate entries, or nonsense information were removed. The researchers also looked at the time taken to complete the survey. All participants took at least 9 minutes to complete the surveys with sensible answers, suggesting that these were legitimate responses.

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 task aimed to understand the relationships among exposure to objective hardship caused by the pandemic, perceived stress in pregnant individuals, and developmental outcomes in their children.

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?*
    - To investigate how mothers' mental health affect the babies.
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.*
    - It permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Research Protocols, is properly cited.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - This is open-access distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/)
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.*

   - 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.*

   - No

7. *Any other comments?*

   - No

Maintenance

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

   - The corresponding author, Gerald F Giesbrecht, PhD.

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

   - The corresponding author, Gerald F Giesbrecht, PhD, provided the email ggiesbre@ucalgary.ca and phone number (+1)403-441-8469.

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 team claimed that 'the Pregnancy during the COVID-19 Pandemic (PdP) cohort study is ongoing, and we plan to release further data on pregnancy and child development as it is collected and cleaned'. However, I did not see any updates.

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.*

   - All the information gathered was anonymous, and participants provided consent for information use in the study.

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 Pregnancy during the COVID-19 Pandemic cohort study was ongoing, and the team planed to release further data on pregnancy and child development as it was collected and cleaned.

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, but the data could be compared to other datasets from the COVID-19 pandemic to establish generalizability, or to pre-pandemic datasets to determine the extent of changes during the COVID-19 pandemic.

8. *Any other comments?*

    - No.

References

Catherine Lebel, Gerald Giesbrecht, Lianne Tomfohr-Madsen. 2023. "Prenatal Mental Health Data and Birth Outcomes in the Pregnancy During the COVID-19 Pandemic Dataset." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10339202/.

Gerald Giesbrecht, Lianne Tomfohr-Madsen, Catherine Lebel. 2023. "Pregnancy During the COVID-19 Pandemic Study." https://osf.io/ha5dp/.

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