Datasheet for 'Global Suicide Rates by Age, Region, Sex, and Income Group'*

Table 1 of the Health paper

Hyuk Jang

18 April 2024

This paper looks into the motivation, composition, collection process, preprocessing/cleaning/labeling, uses, distribution, and maintenance of the sum_sta table.

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 dataset was created to enable analysis of global suicide rates in 2019. The task in mind was to create a linear regression model by cleaning the dataset downloaded from WHO (Organization (2021b)) and (Organization (2021a)).
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 created by the author for his personal use for research
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.*
    - No one funded the creaton of the dataset
4. *Any other comments?*
    - NA

Composition

*Code and data are available at: Code and data are available at: https://github.com/anggimude/Health.

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 the suicide rate per 100,000 population depending on age group and region/income group/sex.

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

    - There are 117 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).*

    - It is not a sample of a larger set. The table has been created from the original dataset downloaded by mapping and averaging values.

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 consists of the age groups such as "15-24 years", "25-34 years", etc. All suicide rates depending on age is used to calculate the mean suicide rates for the regional, income, sex groups.

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

    - The primary target or label in this dataset is the suicide rate per 100,000 population. This is the key variable of interest, around which analysis such as correlation with age, sex, and socio-economic factors can be conducted. It is used for assessing and comparing the impact of these demographic and geographic dimensions on suicide rates.

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 missing information can include countries or years. Missing country is likely because there is no data available from the country. Years may be missing because the most recent available data for the age group data is 2019.

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.*

- Relationships in this dataset are derived from the various region, age, sex. This allows for comparative analysis (e.g., comparing suicide rates across age groups within the same region or income group), but direct links or interactions between instances is not present. The dataset is structured for statistical analysis rather than relational mapping.

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.*

    - Recommended validation testing for the model created using this dataset is the posterior predictive checks and credibility interval.

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

    - NA

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 because it is published data from 2019.

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 it doesn't contain confidential 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 it doesn't contain data that may be offensive, insulting, threatening.

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 age groups are age-standardized, 85+, 75-84, ..., 15-24. Sex is divided by male and female. Then income groups are defined by low income, lower-middle, upper-middle, high income. Regions are North America, South America, Asia, Europe, Africa, Oceania.

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 this is not possible

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.*

    - No it doesn't contain any of this.

16. *Any other comments?*

    - NA

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 was derived the WHO. Since these values are averages and aggregates, they are indirectly inferred from raw data like death certificates and hospital reports. The validation and verification of this data are generally handled at the national or institutional level before being reported to international databases.

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 collection mechanisms would involve national health surveillance systems, which might include manual data entry from hospitals, clinics, and registration offices. These mechanisms are validated through national health statistics procedures, which include checks for consistency, completeness, and accuracy across reporting periods and regions.

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 not a sample from a larger set.

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 WHO employees would have been involved in this process.

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 likely collected over the year of 2019 and beyond as data from around must be gathered and validated.

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.*

    - Data involving human subjects typically undergoes an ethical review, but since this dataset involves aggregated, anonymized public health data, individual ethical reviews might not apply.

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 data was obtained from secondary sources like government health statistics and not directly from individuals. These sources compile data from various healthcare providers and registry offices.

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.*

    - In cases of national health data collection, individuals are not directly notified as the data is collected through administrative records and is anonymized and aggregated.

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.*

    - Individual consent is not required for such collection and uses.

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).*

    - NA

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.*

    - NA

12. *Any other comments?*

    - NA

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.*

    - The raw data was cleaned and labeled to create the sum_sta dataset.

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 was downloadedd and saved in "~/Health/data/raw_data".

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

    - R Core Team (2023) was used to clean and label the dataset.

4. *Any other comments?*

    - To see detailed instructions and the code to cleaning the dataset please read "~/Health/scripts/02-data_cleaning.R".

Uses

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

    - It has been used in the Health.qmd for creating a graph based on the table and doing a multiple linear regression. Check the paper for further details.

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.*

    - https://github.com/anggimude/Health.

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

- It can be used for creating other models or further study can be done when merged with another dataset to see correlations between suicide rates and happiness or so forth.

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?*

    - NA

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 it will not.

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

    - It will be distributed on github through the link above. It does not have a DOI

3. *When will the dataset be distributed?*

    - April 18, 2024

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.*

    - No

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?*
    - NA

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The author will be doing this.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - hyuk.jang@mail.utoronto.ca
3. *Is there an erratum? If so, please provide a link or other access point.*
    - NA
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 it will not be updated.
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 applicable as the data is anonymized and obtained from governments and health institutes.
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.*
    - No updates are planned so older versions will not be available.
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.*

- They may download the repository as a zip and follow the instructions written in scripts to extend on it.

8. *Any other comments?*

    - NA

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.

Organization, World Health. 2021a. "GHO | by Category | Suicide Rate Estimates, Age-Standardized - Estimates by Country."

———. 2021b. "GHO | by Category | Suicide Rate Estimates, Crude, 10-Year Age Groups - Estimates by Country." https://doi.org/MINERVA URL TO PAGE.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.