Datasheet for 'COVerAGE-DB' COVID-19 Dataset*

A summary of characteristics of the data and how it was used

Bruce Zhang

November 26, 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 dataset was created to analyze COVID-19 trends across different regions, sexes, and age groups. It allows for investigating relationships between predictors like tests, deaths, cases, and computed metrics like case fatality rates (CFR). The goal was to enable public health insights and predictive modeling.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The dataset originates from publicly available sources compiled and structured by organizations such as the COVerAGE-DB project, which aggregates global COVID-19 data by demographic breakdowns.
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 sources include research initiatives and public health institutions supporting COVID-19 data dissemination, though specifics depend on the data contributors (e.g., COVerAGE-DB).
4. *Any other comments?*
    - The overall motivation of the data collection was pushed by the large social and political say that COVID-19 had and the importance to understand the pandemic overall.

*Code and data are available at: LINK.

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.*
    - Instances represent demographic and epidemiological data for COVID-19, including age, region, sex, number of tests, cases, deaths, and derived metrics (e.g., CFR).
2. *How many instances are there in total (of each type, if appropriate)?*
    - The dataset contains thousands of rows representing data aggregated at various levels (e.g., country-region-date).
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 structured sample derived from larger repositories (e.g., national or regional health databases). While comprehensive, it may not include every instance globally.
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 raw counts (e.g., cases, deaths), demographic variables (age, region, sex), and calculated fields like CFR. Data preprocessing includes cleaning for missing or invalid entries.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
    - While not directly labeled, "high risk" categories were derived during analysis based on the CFR and median thresholds.
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.*
    - Certain entries might have missing data for specific regions or demographics due to reporting inconsistencies. There were also some problematic datapoints where the number of deaths for a specific demographic exceeded the number of cases.
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.*

- Instances are grouped by age, sex, and region but do not explicitly represent relationships between different demographic groups. The specific indicators of demographics are not necessarily correlated with each other.

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.*

    - The dataset can be split temporally (e.g., pre- and post-vaccine availability), geographically (e.g., urban vs. rural, e.g. by country), or by other demographic characteristics to test specific hypotheses. These can lead to different trends and patterns based on what the focus is for the analysis.

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

    - The dataset may include errors due to underreporting, misclassification, or inconsistencies in data collection. An example is datapoints having more number of deaths than number of cases.

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 relies on external reporting sources such as health departments and international organizations like WHO, with potential limitations in long-term availability.

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 contains aggregated data and does not identify individuals, minimizing confidentiality concerns.

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.*

    - While the dataset itself is not offensive, findings could raise concerns about inequities or vulnerabilities in specific populations.

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.*

- Sub-populations are identified by country, age group, sex, and region.

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.*

    - The dataset does not allow for individual identification as it is aggregated.

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 involves public health data, which can be sensitive but is anonymized and aggregated.

16. *Any other comments?*

    - The data is collected from government and healthcare sources anonymously. The source of the data is generally trustworthy and does not violate privacy or confidenciality.

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 collected from government health agencies, WHO reports, and academic repositories. Preprocessing ensured standardization.

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?*

    - Collection involved web scraping, manual curation, and automated aggregation pipelines (e.g., through APIs).

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 near-comprehensive aggregation but may have regional gaps due to reporting biases or lack of publicly available data.

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 involved government officials, researchers, and volunteers in health agencies.

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.*

    - Data spans the COVID-19 pandemic timeline, with regular updates. Specifically, the data covers the year of 2022.

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.*

    - The dataset used in this analysis was aggregated from publicly available data provided by health organizations, governments, and international agencies (e.g., WHO). As the data is aggregated and anonymized, ethical reviews specific to this dataset were not conducted by the authors of this analysis. However, ethical guidelines from the original sources, such as health agencies, likely applied during the data collection and reporting process.

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 third-party sources, including official health department reports, COVID-19 dashboards, and repositories such as the COVerAGE-DB and WHO's publicly available data.

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.*

    - As the data is aggregated and anonymized, individual notification was not applicable. Data collection practices adhered to reporting standards established by the respective health authorities or governments.

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.*

- The dataset does not involve individual-level data but aggregated public health statistics. Consent was not required since the data does not contain personally identifiable information (PII).

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 applicable. The dataset is anonymized and aggregated, and does not involve individual 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.*

    - Given that the dataset is anonymized and aggregated, a formal data protection impact analysis specific to this analysis was not conducted. However, the original data providers (e.g., WHO, government health organizations) are likely to have conducted such assessments to ensure compliance with data protection laws and ethical standards.

12. *Any other comments?*

    - The data accessed was through COVerAGE-DB, which is a third-party that collects data from government and medical institutes on COVID-19. The details on the sources and methods of collection of the data were limited.

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.*

    - Preprocessing using R (R Core Team 2023) included: Removal of invalid entries (e.g., deaths exceeding cases), creation of age intervals, calculation of metrics like CFR, and filtering for specific regions, sexes, and age groups.

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.*

    - Yes. The raw data can be found at https://osf.io/43ucn

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

- Yes. The software used was R programming, available at https://www.R-project.org/

4. *Any other comments?*

   - No

Uses

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

   - Yes, the dataset has been used for multiple tasks, including the analysis of COVID-19 trends by demographic factors such as age, sex, and region. Researchers have employed it to assess the impact of COVID-19 across different populations and inform policy-making. For example, studies have used this dataset to identify high-risk groups and evaluate disparities in healthcare outcomes. The dataset has also been cited in repositories such as GitHub for further analytical work and data visualizations.

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.*

   - The COVerAGE-DB has a GitHub repository that provides access to the data and associated analyses: https://github.com/timriffe/covid_age (Riffe and colleagues 2021). The repository contains scripts for data processing and examples of research using the dataset. Additional references to papers and systems utilizing the dataset may be linked within this repository or cited in its documentation.

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

   - The dataset could support analyses of vaccine impact, regional healthcare disparities, and demographic-specific risk factors.

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?*

   - Yes, users should be aware of potential reporting biases and inconsistencies in the dataset. As the dataset relies on publicly available COVID-19 data from various countries, differences in testing rates, reporting methods, and definitions of cases and deaths may introduce variability. Additionally, certain regions may have incomplete or missing data, potentially impacting the reliability of demographic analyses. To mitigate risks, users should consider cross-referencing data with other sources and explicitly noting any limitations due to reporting discrepancies.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

    - The dataset's accuracy is contingent on reporting consistency and may not reflect real-time dynamics.

6. *Any other comments?*

    - The dataset is a valuable resource for understanding the demographic impacts of COVID-19, but its use is constrained by the dynamic nature of the pandemic and the evolving accuracy of data reporting. Users should approach the dataset with a clear understanding of its limitations, ensuring that any conclusions drawn are contextualized appropriately within these constraints.

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 dataset is made publicly available through platforms like the Open Science Framework (OSF) (Open Science Framework (OSF) 2022) and GitHub. These platforms enable third parties, including researchers, policymakers, and healthcare organizations, to access the dataset.

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 GitHub as downloadable files and through OSF. The COVerAGE-DB repository does not currently have a DOI, but OSF assigns unique identifiers to its hosted files.

3. *When will the dataset be distributed?*

    - The dataset has already been distributed and is updated periodically. Users can access the latest versions on GitHub and OSF.

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 open-access terms, and users are expected to cite the COVerAGE-DB project when utilizing the data. Licensing terms are not explicitly mentioned on GitHub or OSF but generally follow the principles of open-source data sharing for academic and research purposes.

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 explicit IP-based restrictions are mentioned. The data is freely available for academic and research use, and there are no fees or additional restrictions imposed.

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. It is shared openly across international platforms for use by researchers globally.

7. *Any other comments?*

    - No

Maintenance

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

    - Maintenance is typically overseen by the organization providing the dataset, so COVerAGE-DB and Open Science Framework (OSF) maintains the dataset.

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

    - This information is not provided

3. *Is there an erratum? If so, please provide a link or other access point.*

    - Currently, no official erratum is mentioned. However, updates or corrections to the dataset are communicated via the GitHub repository or the Open Science Framework (OSF).

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)?*

    - Datasets are updated periodically to reflect new data or correct errors. It was last updated on November 29th of 2023. The intervals of update and the contributors are not mentioned.

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 dataset anonymizes data and does not contain personally identifiable information. Retention policies depend on the original data sources, and contributors are expected to comply with privacy and data protection laws of their respective countries.

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 dataset are hosted on the Open Science Framework (OSF) and GitHub. New updates and changes are documented in release notes or change logs, ensuring transparency. Obsolescence, if it occurs, would likely be communicated through these platforms.

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.*

    - Yes, contributors can collaborate via the GitHub repository. Pull requests and issues can be submitted for dataset augmentation or suggestions. Contributions are reviewed by the dataset maintainers to ensure consistency and accuracy. The GitHub repository also provides guidelines for contribution.

8. *Any other comments?*

    - The dataset is a collaborative effort relying on contributions from multiple global sources. While it serves as a comprehensive resource for analyzing COVID-19 trends, its utility depends on the accuracy of the underlying data sources. Users are encouraged to exercise caution when making causal inferences and consider cross-referencing data with other sources.

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. https://doi.org/10.1145/3458723.

Open Science Framework (OSF). 2022. "COVID-19 Cases, Deaths, and Tests by Age, Gender, and Region." Open Science Framework (OSF). https://osf.io/43ucn.

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

Riffe, Tim, and colleagues. 2021. "COVerAGE-DB: A Database of COVID-19 Cases and Deaths by Age and Sex." https://github.com/timriffe/covid_age.