Datasheet for Dataset of Life Expectancy across countries*

Dingning Li

April 17, 2024

This datasheet documents the datasets used in a studying data of life expectancy in different countries. The dataset in the study is compiled by WHO, and is based on public data.It serves as a great resource for analyzing trends and patterns in life expectancy. The focus of the study is on understanding the impact of various factors such as adult mortality rates, prevalence of diseases, socio-economic status, and healthcare investment on the life expectancy of populations.

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 global life expectancy trends and identify the impact of health and socio-economic factors on longevity. Its compilation by WHO aimed to fill the knowledge gap for policy-makers and health researchers in understanding health outcomes across different countries.

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 World Health Organization

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 compilation was undertaken by the World Health Organization, an agency of the United Nations responsible for international public health. Specific funding details, such as grants or grantor names and numbers, if applicable, are not provided in the provided context.

*Code and data are available at: https://github.com/iamldn2002/life_expectancy

4. *Any other comments?*

None

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 in the dataset represent countries, with each entry detailing various health and socio-economic indicators that influence life expectancy. There are no multiple types of instances; rather, each country's data is a single instance characterized by factors such as adult mortality rates, disease prevalence, and schooling years.

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

The provided context does not include the specific number of instances in the dataset. To accurately determine the total count, one would need to review the dataset's documentation or the dataset itself. If the dataset consists of annual life expectancy data for countries from 2000 to 2016, the total instances would be the number of countries multiplied by the number of years covered.

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 likely represents a sample of instances from a larger set of global health data. As it is compiled by the World Health Organization, it is presumed to cover a comprehensive range of countries over the specified years, providing a snapshot rather than an exhaustive record of all global health data. The representativeness would typically be validated by the WHO's rigorous data collection and validation protocols, ensuring a broad geographic coverage and diversity that reflects global health trends. However, without specific details on the dataset's methodology, it is not possible to fully describe the representativeness or the reasons for any limitations in coverage.

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 in the dataset consists of processed and structured features rather than raw data. These features likely include quantifiable health indicators such as life expectancy at birth, adult mortality rates, prevalence of certain diseases, education levels (e.g., average years of schooling), and possibly socioeconomic factors (e.g., GDP per capita, development status). The data has been curated by the WHO from various sources, aggregated, cleaned, and standardized to allow for analysis and comparison across countries and over time.

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

Yes, the primary label or target associated with each instance in this dataset is the life expectancy value for a country in a given year.

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.*

If information is missing from individual instances, it could be due to factors such as incomplete data reporting by some countries, discrepancies in data collection methods, or changes in national or international data recording standards over time.

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.*

In this dataset, relationships between individual instances are not made explicit as each instance represents independent country-year data points focused on life expectancy and related health indicators. There are no links such as social network connections or relational data like user interactions that would interconnect these instances.

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 recommended data splits such as training, development/validation, and testing are mentioned for this dataset. Typically, for analytical and research-focused datasets like this one, which aim to analyze trends over time across countries, the splits might depend on the research objectives. For example, one could use data from earlier years as a training set and more recent years as a test set to validate models predicting future trends based on past data. However, the rationale for any splits would need to be determined based on the specific goals of the study or analysis.

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

The dataset, like many involving large-scale international data collection, might contain sources of noise and potential errors due to discrepancies in how data is reported by different countries, changes in data collection methodologies over time, or translation and transcription errors. Redundancies could also occur if data overlap or are reported in multiple ways by different agencies. Such issues are common in global datasets and usually require careful preprocessing and validation to ensure data quality and consistency.

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, while primarily self-contained, likely relies on external resources for its initial compilation. Given that it is based on public data compiled by the World Health Organization (WHO), the dataset may include data sourced from national health ministries, international health surveys, and other health-related databases. While the WHO strives for data accuracy and consistency, there are no guarantees that external resources will remain constant over time due to updates in data collection methodologies or revisions in historical data.

External Resources: Data may be sourced from national health databases, reports published by governments, and international bodies such as the United Nations and World Bank. These resources may change as new health data is collected and old data is revised.

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, the dataset does not contain confidential data; it comprises aggregated public health data from countries, focusing on indicators like life expectancy and health determinants without revealing individual-level or sensitive information.

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 does not contain data that might be viewed as offensive, insulting, threatening, or anxiety-inducing, as it consists of statistical measures related to health and socio-economic factors without any personal or sensitive content.

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 primarily focuses on overall life expectancy and health metrics by country and does not specify data by sub-populations such as age or gender in the provided description.

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.*

Individual identification is not possible with these datasets as they deal with aggregated data covering various countries.

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 datasets do not contain sensitive personal data.

16. *Any other comments?* None

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 indirectly derived from various public health databases and reports by governments and organizations, validated and standardized by the World Health Organization through rigorous data collection protocols and cross-verification with multiple data sources.

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?*

Not applicable since the dataset was compiled from various resources.

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 likely represents a systematic or stratified sample from a larger set of global health data, where each country provides annual health statistics, chosen to ensure broad geographic and temporal coverage, rather than random sampling, to comprehensively represent global life expectancy trends over the specified years.

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 data collection was conducted by the World Health Organization and its network of national public health authorities, rather than through crowdworkers or contractors, and typically involves salaried employees or governmental officials within the participating countries.

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 collection covers several years, as indicated by the datasets (e.g., 2000-2016 for the data breach incidents).

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.*

There is no specific mention of ethical review processes.

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

Not applicable

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.*

Individual notification is not applicable.

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 relevant for this type of data, as it does not involve personal data collection from individuals.

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).*

Consent mechanisms are not applicable for this dataset as it consists of aggregated public health data without individual-specific information, thus eliminating the need for personal 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.*

No documentation is provided indicating that a data protection impact analysis has been conducted specifically for this dataset on life expectancy, which primarily aggregates anonymized public health data from countries without involving personal data subjects.

12. *Any other comments?*

None

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 and cleaning of the dataset likely included the aggregation of health indicators from various sources, normalization of these indicators to ensure consistency, and the handling of missing values either by imputation or exclusion, depending on the context and availability of data. Additionally, socioeconomic status indicators such as country development classifications may have been standardized to facilitate comparative analysis across different regions and time periods.

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.*

There is no mention of whether the raw data is saved alongside the processed 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.*

Details about the specific software or tools used for preprocessing, cleaning, or labeling are not provided.

4. *Any other comments?*

None

Uses

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

Yes, the dataset has been used for analyzing global trends in life expectancy, assessing the impact of health policies, and studying the relationships between various health determinants and life expectancy across different countries.

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.*

There isn't specific information provided about a repository linking to papers or systems that use this dataset. However, datasets compiled by the World Health Organization (WHO) are frequently cited in health research and can typically be found in scholarly articles and reports. For direct access to research and data, including papers that may utilize this life expectancy dataset, one can often find relevant resources on the WHO website or through academic databases like PubMed or Google Scholar.

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

The dataset could be used for a variety of other tasks including:

Epidemiological Studies: To investigate the correlation between specific diseases and life expectancy across different regions. Healthcare Resource Allocation: To help governments and organizations optimize resource distribution by identifying areas with lower life expectancy and potentially greater healthcare needs. Socio-economic Research: To explore how factors like education, income levels, and employment affect health outcomes across different populations.

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?*

The composition and collection methods of the dataset might have inherent biases or limitations that could impact its future uses, particularly in these ways: Geographical Coverage and Data Completeness, Historical Bias, Standardization Issues, Social and Economic Factors.

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

This dataset should not be used for making individual-level predictions or decisions, as it is aggregated at the country level and lacks the granularity needed for individual or localized analysis. Additionally, it should not be used for clinical or medical diagnostics without supplementary individual-specific data, as the broad, population-level indicators may not accurately reflect individual health conditions or outcomes.

6. *Any other comments?*

None

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.*

As the dataset is compiled by the World Health Organization (WHO) and based on publicly available data, it is typically distributed widely to support global health research and policy-making. It is accessible to researchers, policymakers, and the public through the WHO's data portals and other public health databases to encourage transparency and collaboration in global health initiatives.

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 typically distributed through the World Health Organization's official website, where it can be accessed directly or through linked data portals, often in formats such as CSV or via an API for more dynamic access. Currently, there is no specific mention of a digital object identifier (DOI) for this dataset, but it may be indexed or referenced through WHO publications and reports which themselves are documented and can be cited.

3. *When will the dataset be distributed?*

until to 2016

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, being compiled by WHO, is generally available under the Creative Commons Attribution (CC BY) license, which allows for free use, sharing, adaptation, distribution, and reproduction in any medium or format, as long as proper credit is given to the WHO as the source.

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 is no mention of any third-party IP restrictions on these datasets.

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.*

There are no export controls or regulatory restrictions mentioned for this dataset.

7. *Any other comments?* None

Maintenance

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

WHO, who create and update the data

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

The owner, curator, or manager of the dataset can typically be contacted through the official contact information provided on the World Health Organization (WHO) website, where the dataset is hosted or associated with.

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

There is no mention of an erratum.

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)?*

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.*

Since the dataset is aggregated and anonymized, focusing on country-level health statistics rather than individual-level data, there are no applicable limits on the retention of the data associated with individual instances, nor are there requirements for deleting data after a fixed period.

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 ongoing support, hosting, and maintenance of older versions of the dataset are typically communicated through version control systems or archival repositories, ensuring continued access and transparency for dataset consumers.

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.*

The WHO may provide mechanisms for researchers and organizations to contribute additional data or insights to the dataset, with contributions potentially subject to validation and verification processes to ensure data quality and consistency before dissemination to dataset consumers.

8. *Any other comments?* None

Reference

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.