Datasheet for AI and medical datasets:

"**add dataset name here**"

This form is intended to assist with documentation for medical datasets and expands upon the work by Gebru *et al.* (1). Data documentation in this format has been conducted in recent years following the introduction of the concept in machine learning communities to enhance reproducibility and mitigate bias (2). We modified and added questions that are more relevant for medical research. The modification consists of replacing the terms and examples that are less familiar to medical experts with those that they may use more frequently. In addition, instructions, definitions, and correspondence are added to facilitate responding to the questions and versioning the documentation. The completed documentation could additionally serve as supplementary information for research articles and grant applications. Furthermore, it provides researchers with in-depth and practical information about the dataset and permits them to be more informed about which machine learning and statistical approaches are most relevant and appropriate to apply.

# Instructions:

Please first read the definitions and then continue with responding to the questions.

If a question is not relevant for the dataset, please respond by “not relevant, because …”.

If you do not have the information needed to respond to a question, please respond by “information not available”.

In response to “Any other comments?”, either respond with “no further comments”, or add any comment about that section that might not be covered by the questions.

Please do not modify the questions.

Please do not leave any question unanswered.

At the end of the form please specify the persons who have responded to each section.

If any reference was required, use a standard citation format such as Vancouver and insert the bibliography in the section for References.

After filling out the form, please rename it with this format: for example, if the dataset name is “Danish Cancer Cohort” name the file to “DAIMS\_DanishCancerCohort\_DDMMYYYY.docx” in which DDMMYYYY is the date the file is filled out. This is used for file versioning, as updates may be required in the future.

# Definitions

**Dataset:** a computer file containing the data. It could be in a format of a table like a CSV file (also known as tabular or structured data) or it could be in more complex form such as image or sequencing files.

**Instance:** a data point, record, or sample that can be described for each individual (e.g., a row in a tabular dataset)

**Feature:** in datasets, features are the same as variables, such as age or sex.

**data splits:** data can be split to train/development and test sets either randomly or based on criteria, for example if data come from multiple hospitals or hospital departments, etc.

# Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests.

## 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.

### Response:

## Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

### Response:

## 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.

### Response:

## Any other comments?

### Response:

# Composition

Most of these questions are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for specific tasks. The answers to some of these questions reveal information about compliance with the EU’s General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

## Are there multiple types of instances (e.g., text, images, videos, time series, voice recordings, sequencing files, complex/multidimensional signals)?

Please provide a description.

### Response:

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

### Response:

## Does the dataset contain all available instances or is it a subset (not necessarily random) of instances from a larger set?

If the dataset is a subset from a larger dataset, then what is the larger set?

### Response:

## What data does each instance consist of?

“Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

### Response:

## Is there a label, target, or outcome (e.g., mortality) associated with each instance?

If so, please provide a description.

### Response:

## Is any information missing from individual instances?

If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.

### Response:

## Are relationships between individual instances made explicit (e.g., family links, multiple instances from the same patients)?

If so, please describe how these relationships are made explicit.

### Response:

## Are there recommended data splits (e.g., training, development/validation, testing)?

If so, please provide a description of these splits, explaining the rationale behind them.

### Response:

## Are there any similar datasets?

For example, there could be a publicly available data similar to the dataset described here. If so, please mention. By “similar”, it means that there could be possibilities to use such datasets for further (external) validations.

### Response:

## Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, public databases, 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 (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

### Response:

## Does the dataset contain data that might be considered confidential (e.g., data that are protected by legal privilege or by doctor-patient confidentiality, data that include the content of individuals’ non-public communications)?

If so, please provide a description.

### Response:

## Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

If so, please describe why.

### Response:

## Does the dataset relate to only non-humans (e.g., animals, chemical substances)?

If yes, you may skip the remaining questions in this section.

### Response:

## Does the dataset identify any subpopulations (e.g., by age, gender)?

If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

### Response:

## Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

If so, please describe how.

### Response:

## Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial 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.

### Response:

## Any other comments?

### Response:

# Collection process

The answers to questions here may provide information that allow others to reconstruct the dataset without having access to it.

## How was the data associated with each instance acquired?

Were the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data were reported by subjects or indirectly inferred/derived from other data, were the data validated/verified? If so, please describe how.

### Response:

## What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

How were these mechanisms or procedures validated?

### Response:

## If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

### Response:

## Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

### Response:

## Over what timeframe was the data collected?

Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

### Response:

## Were any ethical review processes conducted (e.g., 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.

### Response:

## Does the dataset relate to only to non-humans (e.g., animals, chemical substances)?

If yes, you may skip the remainder of the questions in this section.

### Response:

## Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

### Response:

## 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.

### Response:

## 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.

### Response:

## 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).

### Response:

## Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., 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.

### Response:

## Any other comments?

### Response:

# Preprocessing, cleaning, labeling

The questions in this section are intended to provide dataset consumers with the information they need to determine whether the “raw” data has been processed in ways that are compatible with their chosen tasks.

## Please fill out the following checklist regarding data cleaning and standardization (only for structured/tabular data).

|  |  |  |  |
| --- | --- | --- | --- |
| Mark if true | Mark if unsure | Item | Comment (if not done) |
|  |  | 1. Wide format: Each row is an instance, and each column is a variable (in addition to patient ID and target/outcome variable) |  |
|  |  | 1. Each patient has a unique identifier (ID) |  |
|  |  | 1. No Unicode character1 |  |
|  |  | 1. No duplicate rows and columns |  |
|  |  | 1. First column is patient ID |  |
|  |  | 1. Last column is the outcome variable or the main outcome in the case of multiple outcomes |  |
|  |  | 1. No further separator (e.g., "," for numbers with four digits or longer such as 1,050,099) and no extra characters "()","[]", "<>", "//", "||" and "{}". The only acceptable separator is the decimal point |  |
|  |  | 1. All missing entries are indicated by the same entry (e.g., either "missing" or an empty field for not available) |  |
|  |  | 1. No instance has all-missing variables |  |
|  |  | 1. A data dictionary (codebook) is provided that defines all variables, their types (continuous, categorical or date and time) and units (e.g., kg) |  |
|  |  | 1. The actual data values for continuous variables are within the range listed in data dictionary |  |
|  |  | 1. Categorical variables have all their categories listed in the data dictionary |  |
|  |  | 1. Rare categories in categorical variables are grouped |  |
|  |  | 1. Perfectly collinear variables are removed |  |
|  |  | 1. Irrelevant observations that could skew the results or cause bias (e.g., outliers or extreme values that do not reflect normal conditions) are removed |  |
|  |  | 1. Date and time are formatted as described in the data dictionary (for example 05-11-2022 12:23 DD-MM-YYYY CET) |  |
|  |  | 1. Numerical entries have “.” as decimal separator (i.e., 1.2 not 1,2) |  |
|  |  | 1. Non-English entries are translated to English. Non-Latin scripts are transformed to Latin scripts |  |
|  |  | 1. The terms used in the dataset follow international standards2,3,4 |  |
|  |  | 1. All entries for each variable follow the same standard (e.g., older entries may have different standards or definitions) |  |
|  |  | 1. Erroneous data are corrected or removed (e.g., a BMI value of 2000) |  |
|  |  | 1. Informative missingness is properly encoded (e.g., “not tested” when it was deemed unnecessary to test) |  |
|  |  | 1. Variables irrelevant for the study are removed (i.e., non-generalizable variables that could not relate to the outcome, e.g., billing ID or personal contact information) |  |
|  |  | 1. No sensitive data including name, address, or identity number of the participants (patients) are included |  |

<https://en.wikipedia.org/wiki/List_of_Unicode_characters>

2 <https://en.wikipedia.org/wiki/MEDCIN>

3 <https://en.wikipedia.org/wiki/International_Classification_of_Diseases>

4 <https://en.wikipedia.org/wiki/SNOMED_CT>

## Was any other preprocessing, cleaning, labeling of the data done (e.g., removal of human DNA or low-quality reads from sequencing data, marking regions of interest in image data, tumor/infection characterization by software or visual inspection, discretization or bucketing, tokenization, part-of-speech tagging, feature extraction)?

If so, please provide a description.

### Response:

## Are there any considerations for measurement/observational error in the processes involved to prepare the dataset?

It could be for example, variables in the dataset with known levels/percentages of measurement error provided by the measurement device/apparatus.

### Response:

## 

## Are there any remaining cleaning or preprocessing required in the dataset?

For example, there could be a percentage of samples waiting for cleaning or preprocessing. There might also be multiple tables of data in the dataset, and they could be merged but have not merged yet.

### Response:

## Was the “raw” data saved in addition to the preprocessed, cleaned, labeled data (e.g., to support unanticipated future uses)?

If so, please provide a link or other access point to the “raw” data.

### Response:

## Is the software used to preprocess, clean, and label the instances available?

If so, please provide a link or other access point.

### Response:

## Does the dataset include any synthetic or imputed data?

If so, please describe it.

### Response:

## Are there variables in the dataset that are equivalents of the outcome (e.g., heart disease and coronary artery disease) or could be considered as secondary outcomes (outcomes of outcomes such as cardiac failure/arrest and death)?

If so, please describe it.

### Response:

## Any other comments?

### Response:

# Uses

These questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

## Has the dataset already been used for a particular task?

For example, it could be statistical analysis (e.g., logistic regression, survival analysis). If so, please mention if there has been any conclusion or inference about the dataset.

### Response:

## Has the dataset been used for any other tasks?

If so, please provide a description.

### Response:

## 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.

### Response:

## What (other) tasks could the dataset be used for?

### Response:

## 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 future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

### Response:

## Are there tasks for which the dataset should not be used?

If so, please provide a description.

### Response:

## Any other comments?

### Response:

# Distribution

## Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

If so, please provide a description.

### Response:

## How will the dataset be distributed (e.g., tarball on website, API, GitHub)?

Does the dataset have a digital object identifier (DOI)?

### Response:

## When will the dataset be distributed?

### Response:

## 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.

### Response:

## 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.

### Response:

## 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.

### Response:

## Any other comments?

### Response:

# Maintenance

These questions are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers.

## Who is supporting/hosting/maintaining the dataset?

### Response:

## How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

### Response:

## Is there an erratum?

If so, please provide a link or other access point.

### Response:

## Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

### Response:

## If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were 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.

### Response:

## 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 users.

### Response:

## If others want to extend, augment, build on, or 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 other users? If so, please provide a description.

### Response:

## Any other comments?

### Response:

# Correspondence

Please specify the authors who filled out the form in the following table. It will be used to contact relevant authors for future enquiries. The information provided in this documentation is confirmed by the authors contributed to each section.

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Email | Affiliation | Sections\* |
|  |  |  |  |
|  |  |  |  |

\*Section: 1) Motivation, 2) Composition, 3) Collection process, 4) Preprocessing/cleaning/labeling 5) Uses, 6) Distribution, 7) Maintenance. Only mention by numbers separated by comma, for example: 1, 4, 6.

# References

1. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H, et al. Datasheets for Datasets. 2018 Mar 23; Available from: http://arxiv.org/abs/1803.09010

2. Paetzold JC, McGinnis J, Shit S, Ezhov I, Büschl P, Prabhakar C, et al. Whole Brain Vessel Graphs: A Dataset and Benchmark for Graph Learning and Neuroscience (VesselGraph). 2021 Aug 30; Available from: http://arxiv.org/abs/2108.13233