Healthsheet for Health Datasets

The aim of this datasheet is to provide guidance towards healthcare dataset documentations, and improving transparency w.r.t. healthcare dataset uses. Healthsheet is built as a contextualized form of [datasheets for dataset questionnaires](https://arxiv.org/pdf/1803.09010.pdf). Every question that is answered is a step towards this goal. It is important to note that dataset generators and curators are providing this information to the community out of their own will, and to refrain from value judgements. The paper describing the development of the [healthsheet questionnaire can be found here](https://arxiv.org/pdf/2202.13028.pdf).

## General information

1. If the answer to any of the questions in the questionnaire is N/A, please describe why the answer is N/A (e.g: data not being available) provide a 2 sentence summary of this dataset.
2. Has the dataset been audited before? If yes, by whom and what are the results?

## Dataset versioning

| **Version:** A dataset will be considered to have a new version if there are major differences from a previous release. Some examples are a change in the number of patients/participants, or an increase in the data modalities covered.  **Sub-versions:** A sub-version tends to apply smaller scale changes to a given version. Some datasets in healthcare are released without labels and predefined tasks, or will be later labeled by researchers for specific tasks and problems, to form sub-versions of the dataset.  The following set of questions clarifies the information about the current (latest) version of the dataset. It is important to report the rationale for labeling the data in any of the versions and sub-versions that this datasheet addresses, funding resources, and motivations behind each released version of the dataset. |
| --- |

#### Please answer the questions in all sections, including the motivation questions presented below, for the current (latest) version that you are addressing.

1. Does the dataset get released as static versions or is it dynamically updated?
   1. If static, how many versions of the dataset exist?
   2. If dynamic, how frequently is the dataset updated?
2. Is this datasheet created for the original version of the dataset? If not, which version of the dataset is this datasheet for?
3. Are there any datasheets created for any versions of this dataset?
4. Does the current version/sub-version of the dataset come with predefined task(s), labels, and recommended data splits (e.g., for training, development/validation, testing)?

If yes, please provide a high-level description of the introduced tasks, data splits, and labeling, and explain the rationale behind them. Please provide the related links and references. If not, is there any resource (website, portal, etc.) to keep track of all defined tasks and/or associated label definitions? (please note that more detailed questions w.r.t labeling is provided in further sections)

1. If the dataset has multiple versions, and this datasheet represents one of them, answer the following questions:
   1. What are the characteristics that have been changed between different versions of the dataset?
   2. Explain the motivation/rationale for creating the current version of the dataset.
   3. Does this version have more subjects/patients represented in the data, or fewer?
   4. Does this version of the dataset have extended data or new data from the same patients as the older versions? Were any patients, data fields, or data points removed? If so, why?
   5. Do we expect more versions of the dataset to be released?
   6. Is this datasheet for a version of the dataset? If yes, does this sub-version of the dataset introduce a new task, labeling, and/or recommended data splits? If the answer to any of these questions is yes, explain the rationale behind it.
   7. Are you aware of any widespread version(s)/sub-version(s) of the dataset? If yes, what is the addressed task, or application that is addressed?

## Motivation

| Reasons and motivations behind creating the dataset, including but not limited to funding interests.  For any of the following questions, if a healthsheet has already been created for this dataset, then refer to those answers when filling in the below information. |
| --- |

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.
2. What are the applications that the dataset is meant to address? (e.g., administrative applications, software applications, research)
3. Are there any types of usage or applications that are discouraged from using this dataset? If so, why?
4. Who created this dataset (e.g., which team, research group), and on behalf of which entity (e.g., company, institution, organization)?
5. 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. If the funding institution differs from the research organization creating and managing the dataset, please state how.
6. What is the distribution of backgrounds and experience/expertise of the dataset curators/generators?

## Data Composition

| What is the dataset made of? What are the modalities, and schema involved in creating the preliminary version of the dataset or following versions and sub-versions?  **Instances:** Refers to the unit of interest. The unit might be different in the datasheet compared to the downstream use case: an instance might relate to a patient in the database, but will be used to provide predictions for specific events for that patient, treating each event as separate. |
| --- |

1. What do the instances that comprise the dataset represent (e.g., documents, images, people, countries)? Are there multiple types of instances? Please provide a description.
2. How many instances are there in total (of each type, if appropriate) (breakdown based on schema, provide data stats)?
3. How many patients / subjects does this dataset represent? Answer this for both the preliminary dataset and the current version of the dataset.
4. 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 (e.g., 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 (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable). Answer this question for the preliminary version and the current version of the dataset in question.

1. What data modality does each patient data consist of? If the data is hierarchical, provide the modality details for all levels (e.g: text, image, physiological signal). Break down in all levels and specify the modalities and devices.
2. What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.
3. 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).
4. Are relationships between individual instances made explicit? (e.g., They are all part of the same clinical trial, or a patient has multiple hospital visits and each visit is one instance)? If so, please describe how these relationships are made explicit.

1. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. (e.g., losing data due to battery failure, or in survey data subjects skip the question, radiological sources of noise).
2. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, other datasets)? If it links to or relies on external resources,
   1. are there guarantees that they will exist, and remain constant, over time;
   2. 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);
   3. 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.
3. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications that is confidential)? If so, please provide a description.
4. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise pose any safety risk (such as psychological safety and anxiety)? If so, please describe why.

1. Does the dataset relate to people? If not, you may skip the remaining questions in this section.
2. If the dataset has been de-identified, were any measures taken to avoid the re-identification of individuals? Examples of such measures: removing patients with rare pathologies or shifting time stamps.
3. 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.

### Devices and Contextual Attributes in Data Collection

1. For data that requires a device or equipment for collection or the context of the experiment, answer the following additional questions or provide relevant information based on the device or context that is used (for example)
   1. If there was an MRI machine used, what is the MRI machine and model used?
   2. If heart rate was measured what is the device for heart rate variation that is used?
   3. If cortisol measurement is reported at multi site, provide details,
   4. If smartphones were used to collect the data, provide the names of models.
   5. And so on,..

### Challenge tests and confounding factors

1. Which factors in the data might limit the generalization of potentially derived models? Is this information available as auxiliary labels for challenge tests? For instance:
   1. Number and diversity of devices included in the dataset.
   2. Data recording specificities, e.g., the view for a chest x-ray image.
   3. Number and diversity of recording sites included in the dataset.
   4. Distribution shifts over time.
2. What confounding factors might be present in the data?
   1. Interactions between demographic or historically marginalized groups and data recordings, e.g., were women patients recorded in one site, and men in another?
   2. Interactions between the labels and data recordings, e.g. were healthy patients recorded on one device and diseased patients on another?

### Collection and use of demographic information

1. Does the dataset identify any demographic sub-populations (e.g., by age, gender, sex, ethnicity)?  
   If yes,
   1. The reasons that these categories were assessed also should be described in the datasheet.
   2. How was this information acquired? Please describe who identified these categories and the source of the classifications used (e.g: self-report or selection, investigator observed, database, electronic health record, survey instrument).
   3. If patients’ demographic data is included, are patients aware / did they consent to the collection and use of their demographic information?
   4. In some cases, there have been biologically proven associations between demographics and the outcome. Are you aware of similar associations in the tasks covered by this dataset? Should users be wary of specific proxies or associations when using the dataset? If yes, please provide a link to the study, or publication.
   5. Is there any mechanism for updating some of this demographic information after its initial collection? For example, if someone wants to change their gender information, what are the mechanisms to do so?
   6. Provide a description of the respective distributions of each subgroup population within the dataset.
2. If no,
   1. Is there any regulation that prevents demographic data collection in your study (for example, the country that the data is collected in)?
   2. Are you employing methods to reduce the disparity of error rate between different demographic subgroups when demographic labels are unavailable? Please describe.

### Pre-processing / de-identification

1. Was there any pre-processing for the de-identification of the patients? Provide the answer for the preliminary and the current version of the dataset
2. Was there any pre-processing for cleaning the data? Provide the answer for the preliminary and the current version of the dataset
3. Was the “raw” data (post de-identification) saved in addition to the preprocessed/cleaned data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.
4. Were instances excluded from the dataset at the time of preprocessing? If so, why? For example, instances related to patients under 18 might be discarded.
5. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Answer this question for both the preliminary dataset and the current version of the dataset

### Labeling and subjectivity of labeling

| **Labeling:**  In medical domains, researchers usually take a dataset and appropriate it for a defined task. researchers may have their own guidance. It is important to know what the incentive of the original creators was, if there was a guideline or there is a guideline for the current version or sub-versions of the dataset? |
| --- |

1. Is there an explicit label or target associated with each data instance? Please respond for both the preliminary dataset and the current version.
   1. If yes:
      1. What are the labels provided?
      2. Who performed the labeling? For example, was the labeling done by a clinician, ML researcher, university or hospital?
   2. What labeling strategy was used?
      1. Gold standard label available in the data (e.g. cancers validated by biopsies)
      2. Proxy label computed from available data:
         1. Which label definition was used? (e.g. Acute Kidney Injury has multiple definitions)
         2. Which tables and features were considered to compute the label?
      3. Which proportion of the data has gold standard labels?
   3. Human-labeled data
      1. How many labellers were considered?
      2. What is the demographic of the labellers? (countries of residence, of origin, number of years of experience, age, gender, race, ethnicity, …)
      3. What guidelines did they follow?
      4. How many labellers provide a label per instance?

If multiple labellers per instance:

* + - 1. What is the rater agreement? How was disagreement handled?
      2. Are all labels provided, or summaries (e.g. maximum vote)?
    1. Is there any subjective source of information that may lead to inconsistencies in the responses? (e.g: multiple people answering a survey having different interpretation of scales, multiple clinicians using scores, or notes)
    2. On average, how much time was required to annotate each instance?
    3. Were the raters compensated for their time? If so, by whom and what amount? What was the compensation strategy (e.g. fixed number of cases, compensated per hour, per cases per hour)?

1. What are the human level performances in the applications that the dataset is supposed to address?
2. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
3. Is there any guideline that the future researchers are recommended to follow when creating new labels / defining new tasks?
4. Are there recommended data splits (e.g., training, development/validation, testing)? Are there units of data to consider, whatever the task? If so, please provide a description of these splits, explaining the rationale behind them. Please provide the answer for both the preliminary dataset and the current version or any sub-version that is widely used.

## **Collection Process**

1. Were any REB/IRB approval (e.g., by an institutional review board or research ethics board) received? 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.
2. How was the data associated with each instance acquired? Was the data directly observable (e.g., medical images, labs or vitals), reported by subjects (e.g., survey responses, pain levels, itching/burning sensations), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
3. 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? Provide the answer for all modalities and collected data. Has this information been changed through the process? If so, explain why.
4. Who was involved in the data collection process (e.g., patients, clinicians, doctors, ML researchers, hospital staff, vendors, etc) and how were they compensated (e.g., how much were contributors paid)?
5. 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.
6. Does the dataset relate to people? If not, you may skip the remaining questions in this section.
7. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., hospitals, app company)?
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.
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.
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).
11. In which countries was the data collected?
12. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a [data protection impact analysis](https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/)) 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.

### Accessibility in data collection

1. Is there any language-based communication with patients (e.g: English, French)? If yes, describe the choices of language(s) for communication. (for example, if there is an app used for communication, what are the language options?)
2. What are the accessibility measurements and what aspects were considered when the study was designed and implemented?
3. If data is part of a clinical study, what are the inclusion criteria?

## Uses

1. Has the dataset been used for any tasks already? If so, please provide a description.
2. Does using the dataset require the citation of the paper or any other forms of acknowledgement? If yes, is it easily accessible through google scholar or other repositories
3. 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. (besides Google scholar)
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 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?
5. Are there tasks for which the dataset should not be used? If so, please provide a description. (for example, dataset creators could recommend against using the dataset for considering immigration cases, as part of insurance policies)

## Dataset Distribution

1. 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.
2. How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
3. When was/will the dataset be distributed?
4. Assuming the dataset is available, will it be/is the dataset 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.
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.
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.

## Maintenance

1. Who will be supporting/hosting/maintaining the dataset?
2. How can the owner/curator/manager of the dataset be contacted (e.g., email address, forms, etc.)?
3. Is there an erratum? If so, please provide a link or other access point.

1. 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)?
2. 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.
3. Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how and for how long. If not, please describe how its obsolescence will be communicated to users.
4. 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 other users? If so, please provide a description.