**Inspect Datasheet: Purpose & Documentation**

*Authors: Shih-Cheng Huang, Zepeng Huo, Ethan Steinberg, Chia-Chun Chiang, Matthew P. Lungren, Curtis P. Langlotz, Serena Yeung, Nigam H. Shah, Jason Alan Fries*

**Motivation\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

* The dataset was created to enable research on multimodal fusion for medical diagnosis (identifying which diseases a patient currently has) and prognoses (predicting the risk of a patient developing a disease in the future). – i.e., given a patient’s electronic health records and medical image, predict whether the patient have certain medical conditions or will eventually develop a disease.
* While there is a growing trend for medical AI models to fuse information from multiple sources of modalities for diagnosis, there is a lack of studies that apply multimodal learning in context to prognosis and outcome prediction. This is largely due to the constraints of publicly available medical datasets, which suffer from one of the following limitations: 1) insufficient scale, 2) limited multimodality, or 3) limited diagnosis.
* While our dataset focuses on a cohort of patient that had a CT pulmonary angiogram (CTPA) to diagnose for Pulmonary Embolism, the methods developed using this dataset should enable multimodal fusion for any patient cohort with any types of medical images.

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

* The dataset was created by Shih-Cheng Huang, Zepeng Huo, Ethan Steinberg and Jason Alan Fries at Stanford University on behalf of the Artificial Intelligence for Medicine and Imaging (AIMI) Center.

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 was provided from National Heart, Lung, And Blood Institute of the National Institutes of Health Awards R01HL155410 and R01HL144555 as well as the National Library of Medicine of the National Institutes of Health Award R01LM012966. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health
* Additionally, compute funding is partially covered by Stanford’s Human-centered AI (HAI) google cloud compute credits.

**Composition\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

* The instances are specific timepoint of a patient’s Electronic Health Record (EHR). Specifically, instances are timepoint that the patient received a CTPA.

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

* There are a total of 23,248 instances from 19,402 patients.

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).

* We identify a total of 155,950 cases with the CT pulmonary angiography (CTPA) procedure code from the Stanford University Medical Center (SUMC) database (2000-2021). Through a comprehensive cohort definition protocol, encompassing stages of random sampling, rigorous data cleaning, and adherence to specified criteria, we refined our cohort o 23,248 studies. Please refer to Supplemental Materials B for more information.

What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

* The instances itself includes the CTPA (as DICOM files), a de-identified DICOM header information for that CTPA, the impression section for that CTPA, the patient’s timeline and featurized timeline until their CTPA, and diagnosis and outcome labels based on that timepoint. For more details, please refer to Supplemental Materials.
* Structured electronic health record (EHR) data consists of CSV files containing de-identified, tab-separated values for patient medical codes. Medical codes are symbols drawn from controlled vocabularies (e.g., SNOMED, ICD, LOINC) for encoding information about clinical events occurring in a patients medical history.

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

* Each instance also includes diagnostic (PE) and prognosis (mortality, re-admission and Pulmonary Hypertension (PH)) labels. The labels for PE is either positive and negative. The labels for prognosis are either positive, negative or censored. For more details, please refer to Supplemental Materials Section D.

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, e.g., redacted text.

* For a given CTPA study, we pick the series with the largest slice thickness between 1.0-3.0mm. This is because the CTPA series for a give study are typically repeated, just with a different slices thickness
* Only a subset of all DICOM headers for the CTPA are included to protect patient privacy (see Supplemental Materials Section C 2.2)
* Currently, we can only release the impression section of a patient’s radiology report. This is the section written by a radiologist to directly describe the medical findings observed in the corresponding CT scan.

Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.

* We use an anonymized patient identifier to link individual instances to a particular patient.

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.

* We define canonical training, validation, and test splits that comprise 80%, 5% and 15% of the dataset respectively.

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

* Care was taken to ensure that multiple patients are not mapped to the same instance
* EHR timelines consist of medical codes assigned at different time points during a patient’s encounter with Stanford Health Care. To remove possible data leakage when training machine learning models, timestamps for all structured billing codes (which are typically assigned at future time points and are not available in real-time during a patient visit) are re-assigned to the end of a patient’s hospital visit.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., 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 (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 dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

* The dataset is entirely self-contained.

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’ nonpublic communications)? If so, please provide a description.

* The dataset contains patient medical information. Care was taken de-identify the patient ensure our dataset is compliant to HIPPA’s policies and preserve patient privacy.
* In addition, all data released is reviewed by a reviewer to ensure no patient privacy.

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.

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.

* The cohort is based on patients from the Stanford Healthcare Center with the CT pulmonary angiography (CTPA) procedure code
* Description of how this subpopulation can be found in Supplemental Materials B
* A breakdown of the cohort’s demographic information can be found in Manuscript Table 1.

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.

* We make sure that our data follows HIPPA’s policies to prevent individuals to be identified.

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

* Our dataset includes the patient’s health and demographic information, including race/ethnicity and biological sex.

Any other comments?

* None.

**Collection Process\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

How was the data associated with each instance acquired? Was 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 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 collection process is described in Supplemental Materials B
* The code for data collection can be found here:

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

* CTPAs are collected using a CT machine. The CTPAs are downloaded from Stanford’s PACS system.
* The radiology reports are written by the radiologist that interprets the CTPA exam. All reports are also downloaded from Stanford’s PACS system.
* EHR data is collected by Epic Systems’ EHR software. Quality control is addressed by Stanford Health Care and Epic internal review/auditing processes.

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

* The dataset is a subset of all patients that received a CTPA exam between 2000-2021 at Stanford university. We use a probabilistic approach to sample the subset.

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)?

* The data are naturally collected in the hospital.

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.

* The timeframe for the data is between year 2000-2021

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.

* Release of INSPECT was approved by the Stanford University Institutional Review Board (IRB), given data privacy review via a standardized workflow conducted by the Center for Artificial Intelligence in Medicine and Imaging (AIMI) and the University Privacy Office.

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

* As mentioned above, we extract the data using STARR.

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.

* The data was extract from Stanford Healthcare’s database. All included patients from SHC signed a privacy notice, which informs them that their records may be used for research purposes given approval by the IRB with study procedures in place to protect patient confidentiality.

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.

* Our study was approved by the Stanford University Administrative Panel on Human Subjects Research, protocol #24883, and included a waiver of consent. All included patients from SHC signed a privacy notice, which informs them that their records may be used for research purposes given approval by the IRB.

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).

* N/A.

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.

* N/A.

Any other comments?

* N/A.

**Preprocessing/cleaning/labeling\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Was any preprocessing/cleaning/labeling of the data done (e.g., 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.

* Please refer to Supplemental Materials for a detailed descriptions of the preprocessing/cleaning/labeling process.

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.

* The released dataset is properly anonymized and reviewed. The “raw” data was not provided to ensure patient privacy.

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

* It is included in the supplement of our submission to NeurIPS

Any other comments?

* No.

**Uses\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

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

* A subset of the dataset were released as part of RadFusion and RSNA RESPECT (https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/rsna-pe-detection-challenge-2020) .
* Table 1 in the main manuscript includes the count and percentage of data included in the RadFusion and RESPECT dataset.

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.

* RadFusion: <https://arxiv.org/abs/2111.11665>
* RSNA RESPECT: <https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/rsna-pe-detection-challenge-2020>

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

* The dataset could be used for anything related to multimodal medical diagnosis and prognosis
* Since we provided the entire patient timeline and the radiology reports for CTPA, users may create their own labeling function to labels to predict.
* We also provide the time to certain events, which allows users to train a Time-to-event (TTE) model.

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 (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

* We provide patient demographics to allow users to evaluate model bias.
* We restrict usage to academic research use cases (i.e., these data cannot be used for providing medical advice or decision making).

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

* We want to emphasize the our dataset is licensed for research use only, and should not be applied for any product or clinical use.

**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.

* Yes, the dataset is publicly available on the internet.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

* The dataset is distributed under AIMI’s webpage at <https://stanfordaimi.azurewebsites.net/>
* The dataset does not have a DOI.

When will the dataset be distributed?

* The dataset will be distributed after manual review of PHI. We have a preview subset of the dataset ready for initial review and are currently reviewing the entire dataset for release in the coming months.

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.

* We share data and trained model weights under Data Use Agreement (DUA) for non-commercial, research use.

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.

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

Any other comments?

* None.

**Maintenance\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Who will be supporting/hosting/maintaining the dataset?

* Shih-Cheng Huang, Zepeng Huo, Ethan Steinberg and Jason Alan Fries are supporting/maintaining the dataset.

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

* Please refer to the manuscript for our email address.

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

* No

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 dataset consumers (e.g., mailing list, GitHub)?

* The dataset will be updated on the AIMI website as needed, on a regular release basis, depending on dataset errors or other corrections identified by users of the dataset.

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

* N/A.

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.

* We do not preserve older versions of the dataset but will provide diffs and changelogs so that users can replicate modifiers as needed. If specific data instances are deleted, e.g., if a specific patient’s record is excluded in the future, from our research STARR database, we will remove these patients from releases.

If others want to extend/augment/build on/contribute to the dataset, 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.

* No, the dataset will only be extended, augmented, build on and contributed by the authors of this paper.

Any other comments?

* None.