RadGraph Datasheet: Purpose & Documentation

Authors: Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, Pranav Rajpurkar

Links

For the purpose of the NeurIPS 2021 Datasets and Benchmarks Track reviewers:

Dataset:
https://physionet.org/RY12LgHwMe2SjTv9V7ahWkiepmgpTJc8tDFzXvpJ8U9fqFvT6DkTuSwCSwXwRluu/
Passphrase: 1bGUQ9ay6zgNnolEItWM

Our dataset with documentation and associated code will be hosted and maintained on PhysioNet under the following license: PhysioNet Credentialed Health Data License 1.5.0. It can be accessed at the following link: https://doi.org/10.13026/hm87-5p47.

Author Statement

We hereby assume responsibility for the dataset. Our dataset can be accessed using the following license: PhysioNet Credentialed Health Data License 1.5.0.

The following Datasheet framework was modeled after Datasheets for Datasets.

Dataset 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.

> This dataset was created for the task of information (entity and relation) extraction from radiology reports. Existing radiology report datasets either (1) are limited to class annotations, failing to capture fine-grained information such as entities and relations, or (2) adopt information extraction schemas with limited coverage and generalizability. To address this gap, we create a dataset of radiology reports with dense annotations for both entities and relations that extracts a broader range of information from the radiology text using a new information extraction schema designed for report coverage and generalizability.

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

This dataset was created by a team of researchers at Stanford University within the Stanford Machine Learning Group (https://stanfordmlgroup.github.io/) under Professor Andrew Ng.

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.

No funding or associated grant. No conflict of interest.

Dataset Composition

What do the instances that comprise the dataset represent? How many instances are there in total?

Our dataset consists of a development dataset and an inference dataset.

Each instance in the development dataset represents the text of a single radiology report densely annotated for entities and relations by board-certified radiologists. Each instance in the inference dataset represents the text of a single radiology report annotated for entities and relations by the RadGraph Benchmark model.

The development dataset is represented in the following three files: (1) `train.json` containing annotations obtained from board-certified radiologists for 425 radiology reports from the MIMIC-CXR dataset, (2) `dev.json` containing annotations obtained from board-certified radiologists for 75 radiology reports from the MIMIC-CXR dataset, and (3) `test.json` containing two independent sets of annotations obtained from board-certified radiologists for 100 radiology reports, 50 from the MIMIC-CXR dataset and 50 from the CheXpert dataset.

The inference dataset is represented in the following two files: (1) `MIMIC-CXR_graphs.json` containing annotations obtained from the RadGraph Benchmark model for 220,763 radiology reports from the MIMIC-CXR dataset, and (2) `CheXpert_graphs.json` containing annotations obtained from the RadGraph Benchmark model for 500 radiology reports from the CheXpert dataset.

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

Reports in the development dataset, which includes the train and dev sets, and test datasets were taken from randomly sampled patients in the MIMIC-CXR and CheXpert datasets. Train, dev, and the MIMIC-CXR test set came from the MIMIC-CXR dataset. The CheXpert test set came from the CheXpert dataset. Patients associated with reports

in the train set and the dev set do not overlap. Similarly, patients associated with reports in the test dataset and development dataset (train and dev sets) do not overlap.

Reports in the inference dataset were sampled from the MIMIC-CXR dataset and the CheXpert dataset. For the MIMIC-CXR portion of the inference dataset, we run inference on all the 227,835 MIMIC-CXR reports using RadGraph Benchmark and use 220,763 of the initial 227,835 reports for which RadGraph Benchmark produced results that directly followed the RadGraph information extraction schema. Since RadGraph Benchmark uses a joint entity and relation extraction approach, the model sometimes made mispredictions drawing relations between non-entities. We choose to ignore any report containing such a misprediction for our inference dataset. For the CheXpert portion of the inference dataset, we randomly sample 500 reports from the CheXpert train set (train.csv in the CheXpert release).

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description. Is there a label or target associated with each instance? If so, please provide a description.

The dataset is formatted as a series of json files that can be found, along with a detailed description, at https://doi.org/10.13026/hm87-5p47. Each instance of the dataset is a dictionary that holds the text for a single radiology report, along with the source of the report (MIMIC-CXR or CheXpert), which data split the report belongs to, annotations for entities in the report, and annotations for relations in the report.

Is any information missing from individual instances?

No, all of the relevant information has been provided.

Are relationships between individual instances made explicit?

Individual instances are independent from one another. Each instance is labeled with its dataset source.

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.

Yes, our development dataset has been divided into a train and dev split (see above for details). The dev set consists of 15% of the development dataset. For our test dataset, we sample 50 radiology reports from the MIMIC-CXR dataset and 50 radiology reports from the CheXpert dataset in order to test generalization of approaches across institutions. Patients associated with reports in the train set and dev set do not overlap, and patients associated with reports in the MIMIC-CXR test dataset and the development dataset do not overlap.

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

Because our development and test sets were annotated by radiologists, there will be some inherent noise and subjectivity in their annotation decisions. Annotators can make mistakes or disagree with one another. We describe annotation disagreements in Section 6.2 of our paper.

Is the dataset self-contained, or does it link to or otherwise rely on external resources?

Our dataset requires credentialing on PhysioNet, where it is hosted, to access. The development dataset is self-contained, as it contains report text along with annotations. The inference dataset contains report text along with annotations, but it can also be extended by using the mappings to chest radiographs in the MIMIC-CXR and CheXpert datasets. The RadGraph dataset requires the same credentialing as the MIMIC-CXR dataset, meaning that there are no barriers to accessing MIMIC-CXR chest radiographs if one can access the RadGraph dataset. Accessing the CheXpert chest radiographs requires separate credentialing. All of the datasets described above are freely available upon approval.

Dataset Confidentiality

Does the dataset contain data that might be considered confidential?

Each report in our dataset is de-identified according to the US Health Insurance Portability Act (HIPAA). MIMIC-CXR reports had already been its authors, who replaced protected health information (PHI) in reports with three consecutive underscores. We de-identified CheXpert reports using an automated, transformer-based de-identification algorithm followed by manual review of each report. PHI was replaced with fake PHI following a hiding-in-plain-sight (HIPS) approach. The de-identification of the CheXpert reports was confirmed by manual review.

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 relate to people?

Yes, the dataset contains patient reports.

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.

> Our dataset does not directly identify any subpopulations since we do not include structured information related to age and gender in our dataset. However, this information can be found by mapping the report to its source in the MIMIC-CXR and CheXpert datasets. However, in text format, the words “man” and “woman” may be referenced in addition to ages below 89 on the CheXpert test set, following the same distribution reported in the CheXpert and MIMIC-CXR datasets given.

Is it possible to identify individuals either directly or indirectly?

> The dataset has been de-identified according to the US Health Insurance Portability Act (HIPAA). Individuals cannot be identified either directly or indirectly.

Does the dataset contain data that might be considered sensitive in any way?

> No, each report in our dataset is de-identified according to the US Health Insurance Portability Act (HIPAA).

Collection Process

How was the data associated with each instance acquired?

> Each radiology report is taken from one of two publicly available datasets: MIMIC-CXR or CheXpert. The annotations for our development set were obtained from board-certified radiologists, each with at least eight years of experience.

What mechanisms or procedures were used to collect the data?

> We ran three labeling pilots, which included around 15 reports each, to train our radiologists and iteratively improve our schema based on their feedback. To ensure high quality annotations, we do not include any of the annotations obtained during pilot labeling initiatives in our released dataset. The radiologists used a text labeling platform Datasaur.ai to directly annotate the free-text reports.

If the dataset is a sample from a larger set, what was the sampling strategy?

> The dataset is sampled from two large publicly available datasets: MIMIC-CXR and CheXpert. For the development dataset, the reports were sampled by patient as described in the previous section, ensuring that (1) patients associated with reports in the train set and dev set do not overlap, and (2) patients associated with reports in the

MIMIC-CXR test dataset and the development dataset do not overlap. For the inference dataset, we use all reports from the MIMIC-CXR dataset (except for those with mispredictions as described in the previous section), and we randomly sample 500 reports from the CheXpert train dataset.

Who was involved in the data collection process and how were they compensated?

All authors of the paper, in addition to the board certified radiologists, were directly or indirectly involved in the data collection process. No direct compensation was provided for data labeling efforts. Datasaur.ai generously provided our team with free access to their labeling platform.

Over what timeframe was the data collected / created?

The dataset was created between March 2021 and June 2021.

Were any ethical review processes conducted (e.g., by an institutional review board)?

No ethical review process was required because we did require any study participants. The original datasets (MIMIC-CXR and CheXpert) had previously collected the data for research purposes.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources?

We obtained it via other sources. The radiology reports were taken from two large, previously collected datasets: MIMIC-CXR and CheXpert.

Were the individuals in question notified about the data collection?

No, the data was collected prior to our use by the MIMIC-CXR and CheXpert team.

Did the individuals in question consent to the collection and use of their data?

No, the data was collected prior to our use by the MIMIC-CXR and CheXpert team.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

Not applicable.

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?

Not applicable.

Dataset Preprocessing, Cleaning, and 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 remainder of the questions in this section.

> We processed each radiology report into a sequence of space-delimited tokens, where punctuation like commas and semicolons were separated from words to support entity recognition.

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 raw radiology reports processed from the MIMIC-CXR dataset are still available in their raw form on the MIMIC-CXR dataset page: https://physionet.org/content/mimic-cxr/2.0.0/. The raw radiology reports processed from the CheXpert dataset are not available in their "raw" form because the de-identification process used removed all formatting from the report. As such, CheXpert reports in their "raw" form cannot currently be released as they contain personal health information (PHI).

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

> Yes, the software used to preprocess the instances is available as part of our dataset release in the following file: https://physionet.org/content/radgraph/1.0.0/models/inference.py. Note that credentialing or a review link with password (provided above) is required to view this file.

Dataset Use Cases

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

> The dataset has only been used by the authors for training models for named entity extraction and relation extraction and then using a trained model to label reports for entities and relations.

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.

Not applicable.

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

Our dataset could be used for various purposes in the healthcare domain. One use case is to develop NLP models for entity and relation extraction in radiology using our development dataset. A second use case is to develop multi-modal models in radiology using our inference dataset, which enables linkage of full-text radiology reports, knowledge graphs (entities / relations as per our schema), and associated chest radiographs from the MIMIC-CXR and CheXpert datasets.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

Yes, we make the decision to separate punctuation from words in radiology reports to support the task of entity recognition. However, the original reports for almost all data (excluding the CheXpert data due to the de-identification processing mentioned above) can be found via the unique identifier provided for the report.

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

To avoid any potential harm to patients, researchers training models on our data should take into account potential distribution shifts that may occur when they apply their models to other datasets with different patient populations.

Dataset 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.

Use of the dataset is free to all researchers after signing a data use agreement which stipulates, among other items, that (1) the user will not share the data, (2) the user will make no attempt to re-identify individuals, and (3) any publication that makes use of the data will also make the relevant code available.

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

Our dataset with documentation and associated code will be hosted and distributed via PhysioNet under the following license: PhysioNet Credentialed Health Data License 1.5.0. PhysioNet provides a wget command to download the data as well as a download button. It has the following DOI: https://doi.org/10.13026/hm87-5p47.

When will the dataset be distributed?

Our dataset is already available on PhysioNet. It was released on June 3, 2021.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

Yes, it will be distributed under the following license: PhysioNet Credentialed Health Data License 1.5.0.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

No IP restrictions apply to the dataset.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

No regulatory restrictions apply apart from the licensed use mentioned above.

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset? How can the owner/curator/manager of the dataset be contacted?

The dataset is being hosted by PhysioNet. It is being maintained by the authors of the paper. The corresponding author can be contacted at: Saahil Jain (saahil.jain@cs.stanford.edu).

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

No, this is the first version of the dataset.

Will the dataset be updated?

If the dataset is updated in future, the older version will still be supported and kept around for consistency.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances?

There are no applicable limits on the retention of the dataset.

Will older versions of the dataset continue to be supported/hosted/maintained?

Yes, each older version of the dataset will continue to be hosted on PhysioNet. New versions will have new release numbers, but old versions will still be available and easily accessible.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

Others may do so and should contact the original authors about incorporating fixes/extensions.