
MIMIC-CXR: A LARGE PUBLICLY AVAILABLE DATABASE OF LABELED CHEST RADIOGRAPHS

Alistair E. W. Johnson^{1*}, Tom J. Pollard¹, Seth Berkowitz², Nathaniel R. Greenbaum³, Matthew P. Lungren⁴, Chih-ying Deng⁵, Roger G. Mark¹, Steven Horng³

¹ Institute of Medical Engineering & Science, Massachusetts Institute of Technology, Cambridge, MA, USA

² Department of Radiology, Beth Israel Deaconess Medical Center, Boston, MA, USA

³ Department of Emergency Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA

⁴ Department of Radiology, Stanford University, Palo Alto, CA, USA

⁵ Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

Correspondence to: aewj@mit.edu

ABSTRACT

Chest radiography is an extremely powerful imaging modality, allowing for a detailed inspection of a patient's thorax, but requiring specialized training for proper interpretation. With the advent of high performance general purpose computer vision algorithms, the accurate automated analysis of chest radiographs is becoming increasingly of interest to researchers. However, a key challenge in the development of these techniques is the lack of sufficient data. Here we describe MIMIC-CXR, a large dataset containing 371,920 chest x-rays associated with 227,943 imaging studies sourced from the Beth Israel Deaconess Medical Center between 2011 - 2016. Images are provided with 14 labels derived from a natural language processing tool applied to the corresponding free-text radiology reports. A single radiology report is written for each imaging study, and imaging studies usually pertain to two x-rays: a frontal view and a lateral view. All images have been de-identified to protect patient privacy. The dataset is made freely available to facilitate and encourage wide range of research in medical computer vision.

Keywords healthcare · radiology · computer vision

1 Introduction

Chest radiography is a common imaging modality used to assess the thorax and the most common medical imaging study in the world. Chest radiographs are used to identify acute and chronic cardiopulmonary conditions, verify that devices such as pacemakers, central lines, and tubes are correctly positioned, and to assist in related medical workups. In the U.S., the number of radiologists as a percentage of the physician workforce is decreasing [1], and the geographic distribution of radiologists favors larger, more urban counties [2]. Delays and backlogs in timely medical imaging interpretation have demonstrably reduced care quality in such large health organizations as the U.K. National Health Service [3] and the U.S. Department of Veterans Affairs [4]. The situation is even worse in resource-poor areas, where radiology services are extremely scarce. As of 2015, only 11 radiologists served the 12 million people of Rwanda [5], while the entire country of Liberia, with a population of four million, had only two practicing radiologists [6]. Accurate automated analysis of radiographs has the potential to improve the efficiency of radiologist workflow and extend expertise to under-served regions.

The combination of burgeoning datasets with increasingly sophisticated algorithms has resulted in a number of significant advances in other application areas of computer vision [7, 8]. A key requirement in the application of these advances to automated chest radiograph analysis is sufficient data. Over time, progressively larger databases have been made available. The Japanese Society of Radiological Technology (JSRT) Database contains 247 images with labels of chest nodules as confirmed by subsequent computed tomography (CT) [9]. Notably, the dataset is provided with annotations segmenting the lungs and heart. The Open-I Indiana University Chest X-ray dataset contains 8,121 images associated with 3,996 de-identified radiology reports [10]. More recently, the NIH released ChestX-ray14 (originally ChestX-ray8), a collection of 112,120 frontal chest radiographs from 30,805 distinct patients with 14 binary labels indicating existence pathology or lack of pathology [11].

Here we present MIMIC Chest X-ray (MIMIC-CXR), a large publicly available dataset of chest radiographs. The dataset contains 371,920 images corresponding to 224,548 radiographic studies performed at the Beth Israel Deaconess Medical Center in Boston, MA. The dataset is de-identified to satisfy the US Health Insurance Portability and Accountability Act of 1996 (HIPAA) Safe Harbor requirements. Protected health information (PHI) has been removed. Randomly generated identifiers are used to group distinct reports and patients.

Images are provided with one or more labels derived from the corresponding free-text radiology report using an open source labeler [12]¹. Labels significantly overlap with those of the NIH ChestX-ray14 dataset and cover a variety of common pathologies in a U.S. urban population. The dataset is intended to support a wide body of research in medical imaging including image understanding and decision support.

2 Chest radiographs

Chest radiographs were sourced from the hospital picture archiving and communication system (PACS) in Digital Imaging and Communications in Medicine (DICOM) format. All studies between 2011 - 2015 were queried. Images were linked to corresponding radiology reports using the hospital’s radiology information system. Images sometimes contain “burned in” annotations: areas where pixel values have been modified after image acquisition in order to display text. Annotations contain relevant information including: image orientation, anatomical position of the subject, timestamp of image capture, and so on. The resulting image, with annotations encoded within the pixel themselves, is then transferred from the modality to PACS. Since the annotations are applied at the modality, it is impossible to recover the original image without annotations. As all patient PHI must be removed to satisfy HIPAA Safe Harbor, images were de-identified using a custom algorithm which removed dates and patient identifiers, but retained radiologically relevant information such as orientation. The algorithm applied an ensemble of image preprocessing and optical character recognition approaches to detect text within an image. Text was identified due to its significant contrast with the background, and due to its consistent pixel value within an image. Suspected PHI was removed by setting all pixel values in a bounding box encompassing the PHI to black. Subsequent to deidentification, we manually reviewed 6,900 radiographs for PHI. Each image was reviewed by two independent annotators. 180 images were identified for a secondary consensus review; none of which ultimately had PHI. The most common causes for annotators to request consensus review were: (1) existence of a support device such as a pacemaker, (2) text identifying in-hospital location (e.g. “MICU”), and (3) obscure text relating to radiograph technique (e.g. “prt rr slot 11”).

After de-identification, images were exported in the JPEG standard format. Pixel data were normalized to the unit interval, and bit-depth was subsequently scaled to 8-bit (0-255). If necessary, image intensity values were inverted to ensure the image transitioned from dark to bright as pixel value increased. Histogram equalization was then applied, and the image was written out in the compressed JPEG format with a quality value of 95.

Note that, aside from de-identification and conversion to JPEG, no filtering or processing of the images was performed. Consequently, images exhibit a number of phenomena common in daily practice. The quality of images varies, both in terms of technique and in terms of patient orientation (e.g. not all patients are healthy enough to stand for a posterior-anterior orientated radiograph). Images may omit anatomy present in a standard chest radiograph, or have objects that obstruct important anatomy. Finally, images may have post-processing applied at the modality such as collimation to improve image pre-processing or rotation to correct orientation. Figure 1 presents a selection of images from the dataset exhibiting challenges to automated processing.



Figure 1: Images which exhibit variation in MIMIC-CXR. From left to right: (1) poor patient positioning, (2) black box obscuring potential PHI, (3) collimation to improve pre-processing, and (4) incorrect image orientation.

¹<https://github.com/stanfordmlgroup/chexpert-labeler>

3 Labeling of the reports

Radiology reports at the source hospital are semi-structured, with radiologists documenting their interpretations in titled sections. The structure of these reports are generally consistent through the use of standardized documentation templates, though can drift over time as the template changed. There can also be some inter-reporter variability as the structure of the reports are not enforced by the user interface and can be overridden by the user. The two primary sections of interest are *findings*; a natural language description of the important aspects in the image, and *impression*; a short summary of the most immediately relevant findings. Labels for the images were derived from either the impression section, the findings section (if impression was not present), or the final section of the report (if neither impression nor findings sections were present). Of the total 227,943 reports, 82.4% had an impression section, 12.5% had a findings section, and 5.1% did not have an impression or findings section.

Labels were determined using the open source CheXpert labeler [12]. CheXpert is a rule based classifier which proceeds in three stages: (1) extraction, (2) classification, and (3) aggregation. In the extraction stage, all mentions of a label are identified, including alternate spellings, synonyms, and abbreviations (e.g. for pneumothorax, the words “pneumothoraces” and “ptx” would also be captured). Mentions are then classified as positive, uncertain, or negative using local context. Finally, aggregation is necessary as there may be multiple mentions of a label. Priority is given to positive mentions, followed by uncertain mentions, and lastly negative mentions. If a positive mention exists, then the label is positive. Conversely, if a negative and uncertain mention exist, the label is uncertain. These stages are used to define all labels except “No Finding”, which is only positive if all other labels except “Support Devices” are negative or unmentioned. More detail is provided in the CheXpert article [12]. Example reports with labels are shown in Table 1. Table 2 shows the frequency of various labels in the reports in the majority subset of the images.

Table 1: Example radiology reports with labels. Labels in italics are *negated*. Labels with a ^u are uncertain.

| Section | Report | Label |
|------------|---|---|
| Impression | No evidence of acute cardiopulmonary process. | No Finding |
| Findings | The left lung is relatively well aerated and clear. The right hemithorax is markedly opacified with volume loss, circumferential pleural thickening and pleural fluid with near complete opacification of the right lung with right basal pleural catheter noted. Hydropneumothorax previously seen is not as well evaluated on this not fully upright film. Cardiac contours are somewhat obscured but unremarkable. | <i>No Cardiomegaly</i> <i>No Enlarged Cardiomediastinum</i> Pneumothorax ^u Airspace Opacity Pleural Effusion Pleural Other Support Devices |
| Other | Cardiac size is top normal. Right supraclavicular central catheter tip is in opacities are new. Compared to prior study CT chest study, these could be due to atelectasis, but superimposed infection cannot be excluded. | <i>No Cardiomegaly</i> <i>No Pneumothorax</i> <i>No Pleural Effusion</i> Pneumonia ^u Atelectasis ^u Airspace Opacity Support Devices |

4 Validation of labels

A random set of reports were selected for validation of the CheXpert labeler. Stratified sampling was used to ensure adequate capture of the various pathologies. A total of 687 reports were reviewed by a board certified radiologist with 8 years experience (ML) and manually labeled according to the 14 categories in CheXpert. The labeling process followed guidelines set forth by the authors of the CheXpert labeler and described therein [12].

*NegBio*², the open source algorithm used to create the labels in the ChestX-ray14 dataset, was the basis for the development of CheXpert and is used as a comparator [11, 13].

The two label algorithms were evaluated in three tasks: mention extraction, negation detection, and uncertainty detection. For the mention extraction task, any assigned label (positive, negative, or uncertain) is considered a positive prediction, while blank (no mention) is considered a negative prediction. For negation detection, negated labels are positive while all other labels are negative. Finally, for uncertainty detection, uncertain labels are positive while all other labels are

²<https://github.com/ncbi-nlp/NegBio>

Table 2: Frequency of labels in MIMIC-CXR on the training subset of 369,188 images, corresponding to 222,952 unique radiologic studies.

| Label | Negative | Positive | Uncertain |
|--------------------------|-----------------|------------------|----------------|
| No Finding | 0 (0.00%) | 140,463 (38.05%) | 0 (0.00%) |
| Enlarged Cardiomeastinum | 9,802 (2.66%) | 9,850 (2.67%) | 12,411 (3.36%) |
| Cardiomegaly | 25,384 (6.88%) | 63,243 (17.13%) | 8,471 (2.29%) |
| Airspace Opacity | 4,931 (1.34%) | 74,658 (20.22%) | 5,802 (1.57%) |
| Lung Lesion | 1,727 (0.47%) | 10,785 (2.92%) | 1,960 (0.53%) |
| Edema | 40,090 (10.86%) | 35,797 (9.70%) | 18,884 (5.12%) |
| Consolidation | 14,048 (3.81%) | 14,326 (3.88%) | 6,496 (1.76%) |
| Pneumonia | 42,185 (11.43%) | 26,199 (7.10%) | 28,431 (7.70%) |
| Atelectasis | 2,208 (0.60%) | 64,033 (17.34%) | 15,132 (4.10%) |
| Pneumothorax | 57,667 (15.62%) | 14,363 (3.89%) | 1,591 (0.43%) |
| Pleural Effusion | 43,632 (11.82%) | 75,332 (20.40%) | 8,451 (2.29%) |
| Pleural Other | 253 (0.07%) | 3,319 (0.90%) | 1,382 (0.37%) |
| Fracture | 1,705 (0.46%) | 7,593 (2.06%) | 1,078 (0.29%) |
| Support Devices | 4,943 (1.34%) | 83,061 (22.50%) | 361 (0.10%) |

negative. The harmonic mean of the sensitivity and positive predictive value, referred to as the F1 score, was calculated for each group independently. Table 3 shows the performance across labels. Performance is provided with “macro” averages (mean across the table) and “micro” averages (mean accounting for the prevalence of each class).

Table 3: Performance of NegBio and CheXpert on 687 manually labeled reports.

| Category | Mention | | Negation | | Uncertainty | |
|--------------------------|---------|--------------|--------------|--------------|-------------|--------------|
| | NegBio | CheXpert | NegBio | CheXpert | NegBio | CheXpert |
| Atelectasis | 0.930 | 0.998 | 0.727 | 0.400 | 0.379 | 0.835 |
| Cardiomegaly | 0.596 | 0.954 | 0.043 | 0.830 | 0.000 | 0.333 |
| Consolidation | 0.966 | 0.986 | 0.917 | 0.958 | 0.235 | 0.486 |
| Edema | 0.855 | 0.996 | 0.701 | 0.878 | 0.214 | 0.742 |
| Pleural Effusion | 0.971 | 0.987 | 0.873 | 0.947 | 0.368 | 0.500 |
| Pneumonia | 0.836 | 0.981 | 0.750 | 0.785 | 0.388 | 0.674 |
| Pneumothorax | 0.983 | 0.998 | 0.951 | 0.948 | 0.182 | 0.286 |
| Enlarged Cardiomeastinum | | 0.761 | | 0.679 | | 0.697 |
| Lung Lesion | | 0.855 | | 0.500 | | 0.143 |
| Airspace Opacity | | 0.802 | | 0.421 | | |
| Pleural Other | | 0.592 | | 0.000 | | |
| Fracture | | 0.904 | | 0.000 | | 0.000 |
| Support Devices | | 0.880 | | 0.000 | | |
| No Finding | | 0.543 | | | | |
| Macro-average | | 0.874 | | 0.565 | | 0.470 |
| Micro-average | | 0.930 | | 0.846 | | 0.628 |

5 Training, validation, and test sets

To ensure consistent evaluation of models, we have organized the data into training, validation, and test sets. The test set contains all studies for patients who had at least one report labelled in our manual review. We are not publicly releasing the test set. The validation set contains a random set of 500 patients and all of their associated studies. This set is made publicly available in a separate ‘valid’ folder. Finally, all remaining studies are made available in the training set. Table 4 provides summary information for studies in the three datasets. Note the enrichment of findings in the test set caused by the stratified sampling done to ensure sufficient coverage of all pathologies.

Table 4: Summary of the images split into training, validation, and test sets.

| Dataset | Training | Validation | Test |
|-----------------------------------|----------------|--------------|--------------|
| Number of images | 369188 | 2732 | 5239 |
| Frontal | 248285 (67.3%) | 1759 (64.4%) | 3708 (70.8%) |
| Lateral | 120756 (32.7%) | 972 (35.6%) | 1524 (29.1%) |
| Other | 147 (0.0%) | 1 (0.0%) | 7 (0.1%) |
| Number of reports with a finding | 222952 | 1596 | 3326 |
| | 149150 (66.9%) | 1004 (62.9%) | 2753 (82.8%) |
| Number of patients with a finding | 64588 | 500 | 295 |
| | 38470 (46.9%) | 293 (46.0%) | 286 (60.6%) |

6 Data availability

All data is made available on PhysioNet³ [14]. Use of the dataset is free to all researchers after signing of 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 reidentify individuals, and (3) any publication which makes use of the data will also make the relevant code available.

Future updates are planned for MIMIC-CXR. In particular, the original DICOM files with free-text radiology reports are planned for release. Due to the added sensitivity of this dataset, access will require completion of a training course in human subjects research, as is the process for MIMIC-III [15] and eICU-CRD [16].

7 Conclusions

MIMIC-CXR is a large, publicly available dataset of chest radiographs from over 170,000 studies performed at the BIDMC. The dataset contains labels for a number of common pathologies and will provide a benchmark for a number of medically relevant computer vision tasks.

Acknowledgements

We would like to acknowledge the Stanford Machine Learning Group and the Stanford AIMI center for their help in running the chexpert labeler and for their insight into the work; in particular we would like to thank Jeremy Irvin and Pranav Rajpurkar. We would also like to acknowledge the BIDMC for their continued collaboration.

This work was supported by grant NIH-R01-EB017205 from the National Institutes of Health. The MIT Laboratory for Computational Physiology received funding from Philips Healthcare to create the database described in this paper.

References

- [1] Andrew B Rosenkrantz, Danny R Hughes, and Richard Duszak Jr. The us radiologist workforce: an analysis of temporal and geographic variation by using large national datasets. *Radiology*, 279(1):175–184, 2015.
- [2] Andrew B Rosenkrantz, Wenyi Wang, Danny R Hughes, and Richard Duszak Jr. A county-level analysis of the us radiologist workforce: physician supply and subspecialty characteristics. *Journal of the American College of Radiology*, 15(4):601–606, 2018.
- [3] Abi Rimmer. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ: British Medical Journal (Online)*, 359, 2017.
- [4] Sarah Bastawrous and Benjamin Carney. Improving patient safety: Avoiding unread imaging exams in the national va enterprise electronic health record. *Journal of digital imaging*, 30(3):309–313, 2017.
- [5] David A Rosman, Jean Jacques Nshizirungu, Emmanuel Rudakemwa, Crispin Moshi, Jean de Dieu Tuyisenge, Etienne Uwimana, and Louise Kalisa. Imaging in the land of 1000 hills: Rwanda radiology country report. *Journal of Global Radiology*, 1(1):5, 2015.

³<https://www.physionet.org/physiobank/database/mimiccxr/>

- [6] Farah S Ali, Samantha G Harrington, Stephen B Kennedy, and Sarwat Hussain. Diagnostic radiology in liberia: a country report. *Journal of Global Radiology*, 1(2):6, 2015.
- [7] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.
- [8] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 843–852. IEEE, 2017.
- [9] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists’ detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1):71–74, 2000.
- [10] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2015.
- [11] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3462–3471. IEEE, 2017.
- [12] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghighi, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.
- [13] Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2017:188, 2018.
- [14] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.
- [15] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.
- [16] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5, 2018.