**MIMIC-CXR-JPG - chest radiographs with structured labels**

**Alistair Johnson ,  matt lungren ,  Yifan Peng ,  Zhiyong Lu ,  Roger Mark ,  Seth Berkowitz ,  Steven Horng**

Published: Nov. 14, 2019. Version: 2.0.0

**When using this resource, please cite:**

Johnson, A., lungren, m., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., Horng, S. (2019). MIMIC-CXR-JPG - chest radiographs with structured labels (version 2.0.0). *PhysioNet*. <https://doi.org/10.13026/8360-t248>

**Additionally, please cite the original publication:**

[Johnson AE, Pollard TJ, Berkowitz S, Greenbaum NR, Lungren MP, Deng CY, Mark RG, Horng S. MIMIC-CXR: A large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042. 2019 Jan 21.](https://arxiv.org/abs/1901.07042)

**Please include the standard citation for PhysioNet:**

Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals (2003). Circulation. 101(23):e215-e220.

**Abstract**

The MIMIC Chest X-ray JPG (MIMIC-CXR-JPG) Database v2.0.0 is a large publicly available dataset of chest radiographs in JPG format with structured labels derived from free-text radiology reports. The MIMIC-CXR-JPG dataset is wholly derived from [MIMIC-CXR](https://physionet.org/content/mimic-cxr/), providing JPG format files derived from the DICOM images and structured labels derived from the free-text reports. The aim of MIMIC-CXR-JPG is to provide a convenient processed version of MIMIC-CXR, as well as to provide a standard reference for data splits and image labels. The dataset contains 377,110 JPG format images and structured labels derived from the 227,827 free-text radiology reports associated with these images. 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. The dataset is intended to support a wide body of research in medicine including image understanding, natural language processing, and decision support.

**Background**

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 [MIMIC-CXR](https://physionet.org/content/mimic-cxr/) database aimed to galvanize research around automated analysis of chest radiographs. The chest radiographs in [MIMIC-CXR](https://physionet.org/content/mimic-cxr/) are published in DICOM format, which is commonly used in clinical practice. DICOM is a well defined binary file format which stores a large amount of meta-data with the pixel values of the image. Unfortunately, due to the complexity of the application domain (radiology), the DICOM file format can be difficult to comprehend, creating an undesirable barrier for those traditionally outside of the medical domain. Outside of radiology, digital images tend to be stored using one of a number of more common general purpose formats. One particularly common format, JPG, achieves significant savings in image storage size using a lossy compression algorithm. While the loss of information is undesirable, the benefits of a reduced image storage size are many and so the JPG image format remains popular among computer vision researchers.

The primary goal of the MIMIC-CXR-JPG database is to provide a standard reference for JPG images derived from the DICOM files. This is particularly important as DICOMs contain higher pixel depth than can be perceived by the human eye, and thus a design decision must be made in converting the 16-bit depth raw images into 12-bit depth images in JPG format. Furthermore, a number of image pixel normalization strategies are employed in computer vision, and providing the most common approach as a reference database saves researchers time and makes it easier to compare derivative works.

The MIMIC-CXR-JPG database also provides structured labels for the provided JPG images derived from the associated free-text radiology report. While other researchers can derive structured labels from the free-text radiology reports themselves, providing labels here ensures their derivation is consistent across distinct researchers.

**Methods**

The source data, [MIMIC-CXR](https://physionet.org/content/mimic-cxr/), contains DICOM format images with free-text radiology reports. The images and free-text reports were processed independently. Creation of MIMIC-CXR-JPG involved three steps: (1) conversion of the DICOMs into JPG, (2) extraction of structured labels from free-text radiology reports associated with each image, and (3) creation of meta-data files providing further information regarding the images.

**Chest radiographs**

Chest radiographs were converted from DICOM to a compressed JPG format. First, the image pixels were extracted from the DICOM file using the pydicom library [10]. Pixel values were normalized to the range [0, 255] by subtracting the lowest value in the image, dividing by the highest value in the shifted image, truncating values, and converting the result to an unsigned integer. The DICOM field PhotometricInterpretation was used to determine whether the pixel values were inverted, and if necessary images were inverted such that air in the image appears white (highest pixel value), while the outside of the patient's body appears black (lowest pixel value). The OpenCV library was then used to histogram equalize the image with the intention of enhancing contrast. Histogram equalization involves shifting pixel values towards 0 or towards 255 such that all pixel values 0 through 255 have approximately equal frequency. Images were then converted to JPG files using OpenCV with a quality factor of 95.

**Labeling of the radiology reports**

Radiology reports in [MIMIC-CXR](https://physionet.org/content/mimic-cxr/) are semi-structured, with radiologists documenting their interpretations in titled sections. The structure of the reports is generally consistent due to the use of standardized templates, though occasional amendments to the template results in a slight drift  
over time. Inter-reporter variability in report structure also exists as the template is 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. Custom code was written in Python 3.7 to extract the findings and impression sections for labeling by open-source tools. 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 two methods, described in turn.

NegBio is an open-source rule based tool for negation and uncertain detection in radiology reports [9]. NegBio takes as input a sentence with pre-tagged mentions of medical findings, and determines whether a specific finding is negative or uncertain. Unlike previous methods, NegBio uses hand-crafted heuristic rules to search the syntactic structure (dependency graph) of each sentence in the report to determine if a mention is covered by a negated cue (e.g., “no evidence of”). If so, this mention will be marked as negative. NegBio also detects uncertain mentions of medical findings, a common occurrence in radiology reports. More detail is provided in the NegBio article [9]. The output of NegBio was saved to a CSV file with one row per study and one column per finding.

CheXpert is an open-source rule based tool that is built on NegBio. It 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) [8]. 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 [8]. The output of CheXpert was saved to a CSV file with one row per study and one column per finding.

NegBio was run using the custom mention patterns defined by the CheXpert tool [8]. Note that these patterns are different than those used by NegBio to create the labels for the NIH ChestX-ray8/ChestX-ray14 dataset [10]. When CheXpert or NegBio were unable to derive a label no label is generated and no row appears for the study in the CSV. Therefore the two label CSV files will contain a strict subset of studies in MIMIC-CXR.

**Data Description**

**Overview**

MIMIC-CXR-JPG v2.0.0 contains:

* A set of 10 folders, each with ~6,500 sub-folders corresponding to all the JPG format images for an individual patient.
* mimic-cxr-2.0.0-metadata.csv.gz - a compressed CSV file providing useful metadata for the images including view position, patient orientation, and an anonymized date of image acquisition time allowing chronological ordering of the images.
* mimic-cxr-2.0.0-split.csv.gz - a compressed CSV file providing recommended train/validation/test data splits.
* mimic-cxr-2.0.0-chexpert.csv.gz - a compressed CSV file listing all studies with labels generated by the CheXpert labeler.
* mimic-cxr-2.0.0-negbio.csv.gz - a compressed CSV file listing all studies with labels generated by the NegBio labeler.

Images are provided in individual folders. An example of the folder structure for a single patient's images is as follows:

files/

p10/

p10000032/

s50414267/

02aa804e-bde0afdd-112c0b34-7bc16630-4e384014.jpg

174413ec-4ec4c1f7-34ea26b7-c5f994f8-79ef1962.jpg

s53189527/

2a2277a9-b0ded155-c0de8eb9-c124d10e-82c5caab.jpg

e084de3b-be89b11e-20fe3f9f-9c8d8dfe-4cfd202c.jpg

s53911762/

68b5c4b1-227d0485-9cc38c3f-7b84ab51-4b472714.jpg

fffabebf-74fd3a1f-673b6b41-96ec0ac9-2ab69818.jpg

s56699142/

ea030e7a-2e3b1346-bc518786-7a8fd698-f673b44c.jpg

Above, we have a single patient, p10000032. Since the first three characters of the folder name are p10, the patient folder is in the p10/ folder. This patient has four radiographic studies: s50414267, s53189527, s53911762, and s56699142. These study identifiers are completely random, and their order has no implications for the chronological order of the actual studies. Each study has two chest x-rays associated with it, except s56699142, which only has one study. The free-text radiology report corresponding to each study and the original DICOM format images are available in the [MIMIC-CXR](https://doi.org/10.13026/C2JT1Q) database.

**Metadata files**

The mimic-cxr-2.0.0-metadata.csv.gz file contains useful meta-data derived from the original DICOM files in [MIMIC-CXR](https://doi.org/10.13026/C2JT1Q). The columns are:

* dicom\_id - An identifier for the DICOM file. The stem of each JPG image filename is equal to the dicom\_id.
* PerformedProcedureStepDescription - The type of study performed ("CHEST (PA AND LAT)", "CHEST (PORTABLE AP)", etc).
* ViewPosition - The orientation in which the chest radiograph was taken ("AP", "PA", "LATERAL", etc).
* Rows - The height of the image in pixels.
* Columns - The width of the image in pixels.
* StudyDate - An anonymized date for the radiographic study. All images from the same study will have the same date and time. Dates are anonymized, but chronologically consistent for each patient. Intervals between two scans have not been modified during de-identification.
* StudyTime - The time of the study in hours, minutes, seconds, and fractional seconds. The time of the study was not modified during de-identification.
* ProcedureCodeSequence\_CodeMeaning - The human readable description of the coded procedure (e.g. "CHEST (PA AND LAT)". Descriptions follow Simon-Leeming codes [11].
* ViewCodeSequence\_CodeMeaning - The human readable description of the coded view orientation for the image (e.g. "postero-anterior", "antero-posterior", "lateral").
* PatientOrientationCodeSequence\_CodeMeaning - The human readable description of the patient orientation during the image acquisition. Three values are possible: "Erect", "Recumbent", or a null value (missing).

The names of the columns (aside from dicom\_id) are defined as the *Keyword* from their corresponding DICOM data element, e.g. ViewPosition (0018, 5101). Column names for metadata derived from length-1 sequences are presented as *KeywordSequence\_KeywordSubitem*, e.g. PatientOrientationCodeSequence\_CodeMeaning is sourced from the DICOM standard Patient Orientation Code Sequence (0054, 0410), under Code Meaning (0008, 0104).

The mimic-cxr-2.0.0-split.csv.gz file contains:

* dicom\_id - An identifier for the DICOM file. The stem of each JPG image filename is equal to the dicom\_id.
* study\_id - An integer unique for an individual study (i.e. an individual radiology report with one or more.
* subject\_id - An integer unique for an individual patient.
* split - a string field indicating the data split for this file, one of 'train', 'validate', or 'test'.

The split file is intended to provide a reference dataset split for studies using MIMIC-CXR-JPG.

**Structured labels**

The mimic-cxr-2.0.0-chexpert.csv.gz and mimic-cxr-2.0.0-negbio.csv.gz files are compressed comma delimited value files. A total of 227,827 studies are assigned a label by CheXpert and NegBio. Eight studies could not be labeled due to a lack of a findings or impression section. The first three columns are:

* subject\_id - An integer unique for an individual patient
* study\_id - An integer unique for an individual study (i.e. an individual radiology report with one or more images associated with it)

The remaining columns are labels as presented in the CheXpert article [8]:

* Atelectasis
* Cardiomegaly
* Consolidation
* Edema
* Enlarged Cardiomediastinum
* Fracture
* Lung Lesion
* Lung Opacity
* Pleural Effusion
* Pneumonia
* Pneumothorax
* Pleural Other
* Support Devices
* No Finding

Note that "No Finding" is the absence of any of the 13 descriptive labels *and*a check that the text does not mention a specified set of other common findings beyond those covered by the descriptive labels. Thus, it is possible for a study in the CheXpert set to have no labels assigned. For example, study 57,321,224 has the following findings/impression text: "Hyperinflation.  No evidence of acute disease.". Normally this would be assigned a label of "No Finding", but the use of "hyperinflation" suppresses the labeling of no finding. For details see the CheXpert article [8], and the list of phrases are publicly available in their code repository (phrases/mention/no\_finding.txt). There are 2,414 studies which do not have a label assigned by CheXpert. Conversely, all studies present in the provided files have been assigned a label by NegBio.

Each label column contains one of four values: 1.0, -1.0, 0.0, or missing. These labels have the following interpretation:

* 1.0 - The label was positively mentioned in the associated study, and is present in one or more of the corresponding images
  + e.g. "A large pleural effusion"
* 0.0 - The label was negatively mentioned in the associated study, and therefore should not be present in any of the corresponding images
  + e.g. "No pneumothorax."
* -1.0 - The label was either: (1) mentioned with uncertainty in the report, and therefore may or may not be present to some degree in the corresponding image, or (2) mentioned with ambiguous language in the report and it is unclear if the pathology exists or not
  + Explicit uncertainty: "The cardiac size cannot be evaluated."
  + Ambiguous language: "The cardiac contours are stable."
* Missing (empty element) - No mention of the label was made in the report

**Usage Notes**

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.

Code to generate MIMIC-CXR-JPG from the source data, [MIMIC-CXR](https://doi.org/10.13026/C2JT1Q), is available online at the [MIMIC-CXR Code Repository.](https://doi.org/10.5281/zenodo.2591653)

**Release Notes**

**MIMIC-CXR-JPG v2.0.0**

MIMIC-CXR-JPG includes JPG formatted image files and structured labels extracted with publicly available natural language processing tools. The images are identical to MIMIC-CXR v2.0.0, but have been transformed into a more compressed file format.

**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 Beth Israel Deaconess Medical Center 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.

**Conflicts of Interest**

Philips Healthcare supported the creation of this resource.

**References**

1. Ray Smith. An overview of the tesseract OCR engine. In Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), volume 2, pages 629–633. IEEE, 2007.
2. 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.
3. 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.
4. Sarah Bastawrous and Benjamin Carney. Improving patient safety: Avoiding unread imaging exams in the nationalva enterprise electronic health record. Journal of digital imaging, 30(3):309–313, 2017.
5. Abi Rimmer. Radiologist shortage leaves patient care at risk, warns royal college. BMJ: British Medical Journal(Online), 359, 2017.
6. Andrew B Rosenkrantz, Wenyi Wang, Danny R Hughes, and Richard Duszak Jr. A county-level analysis of theus radiologist workforce: physician supply and subspecialty characteristics. Journal of the American College of Radiology, 15(4):601–606, 2018.
7. 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.
8. Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund,Behzad Haghgoo, 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.
9. Peng Y, Wang X, Lu L, Bagheri M, Summers R, Lu Z. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. AMIA Summits on Translational Science Proceedings. 2018;2018:188.
10. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. InProceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 2097-2106).
11. Simon, M., Leeming, B. W., Bleich, H. L., Reiffen, B., Byrd, J., Blair, D., & Shimm, D. (1974). Computerized radiology reporting using coded language. Radiology, 113(2), 343-349.