CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison

Jeremy Irvin,[1,*] Pranav Rajpurkar,[1,*] Michael Ko,[1] Yifan Yu,[1]
Silviana Ciurea-Ilcus,[1] Chris Chute,[1] Henrik Marklund,[1] Behzad Haghgoo,[1]
Robyn Ball,[2] Katie Shpanskaya,[3] Jayne Seekins,[3] David A. Mong,[3]
Safwan S. Halabi,[3] Jesse K. Sandberg,[3] Ricky Jones,[3] David B. Larson,[3]
Curtis P. Langlotz,[3] Bhavik N. Patel,[3] Matthew P. Lungren,[3,†] Andrew Y. Ng[1,†]
[1]Department of Computer Science, Stanford University
[2]Department of Medicine, Stanford University
[3]Department of Radiology, Stanford University
[*]Equal contribution
[†]Equal contribution
{jirvin16, pranavsr}@cs.stanford.edu

Abstract

Large, labeled datasets have driven deep learning methods to achieve expert-level performance on a variety of medical imaging tasks. We present CheXpert, a large dataset that contains 224,316 chest radiographs of 65,240 patients. We design a labeler to automatically detect the presence of 14 observations in radiology reports, capturing uncertainties inherent in radiograph interpretation. We investigate different approaches to using the uncertainty labels for training convolutional neural networks that output the probability of these observations given the available frontal and lateral radiographs. On a validation set of 200 chest radiographic studies which were manually annotated by 3 board-certified radiologists, we find that different uncertainty approaches are useful for different pathologies. We then evaluate our best model on a test set composed of 500 chest radiographic studies annotated by a consensus of 5 board-certified radiologists, and compare the performance of our model to that of 3 additional radiologists in the detection of 5 selected pathologies. On Cardiomegaly, Edema, and Pleural Effusion, the model ROC and PR curves lie above all 3 radiologist operating points. We release the dataset to the public as a standard benchmark to evaluate performance of chest radiograph interpretation models.[1]

Introduction

Chest radiography is the most common imaging examination globally, critical for screening, diagnosis, and management of many life threatening diseases. Automated chest radiograph interpretation at the level of practicing radiologists could provide substantial benefit in many medical settings, from improved workflow prioritization and clinical decision support to large-scale screening and global population health initiatives. For progress, there is a need for labeled datasets that (1) are large, (2) have strong reference standards, and (3) provide expert human performance metrics for comparison.

[1]https://stanfordmlgroup.github.io/competitions/chexpert

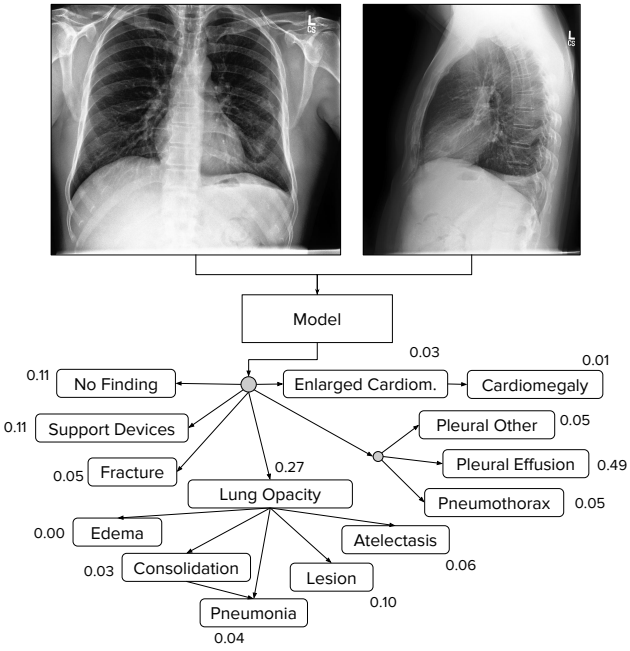


Figure 1: The CheXpert task is to predict the probability of different observations from multi-view chest radiographs.

In this work, we present CheXpert (**Ch**est e**Xpert**), a large dataset for chest radiograph interpretation. The dataset consists of 224,316 chest radiographs of 65,240 patients labeled for the presence of 14 common chest radiographic observations. We design a labeler that can extract observations from free-text radiology reports and capture uncertainties present in the reports by using an uncertainty label.

The CheXpert task is to predict the probability of 14 different observations from multi-view chest radiographs (see Figure 1). We pay particular attention to uncertainty labels in the dataset, and investigate different approaches towards incorporating those labels into the training process. We as-

Pathology	Positive (%)	Uncertain (%)	Negative (%)
No Finding	16627 (8.86)	0 (0.0)	171014 (91.14)
Enlarged Cardiom.	9020 (4.81)	10148 (5.41)	168473 (89.78)
Cardiomegaly	23002 (12.26)	6597 (3.52)	158042 (84.23)
Lung Lesion	6856 (3.65)	1071 (0.57)	179714 (95.78)
Lung Opacity	92669 (49.39)	4341 (2.31)	90631 (48.3)
Edema	48905 (26.06)	11571 (6.17)	127165 (67.77)
Consolidation	12730 (6.78)	23976 (12.78)	150935 (80.44)
Pneumonia	4576 (2.44)	15658 (8.34)	167407 (89.22)
Atelectasis	29333 (15.63)	29377 (15.66)	128931 (68.71)
Pneumothorax	17313 (9.23)	2663 (1.42)	167665 (89.35)
Pleural Effusion	75696 (40.34)	9419 (5.02)	102526 (54.64)
Pleural Other	2441 (1.3)	1771 (0.94)	183429 (97.76)
Fracture	7270 (3.87)	484 (0.26)	179887 (95.87)
Support Devices	105831 (56.4)	898 (0.48)	80912 (43.12)

Table 1: The CheXpert dataset consists of 14 labeled observations. We report the number of studies which contain these observations in the training set.

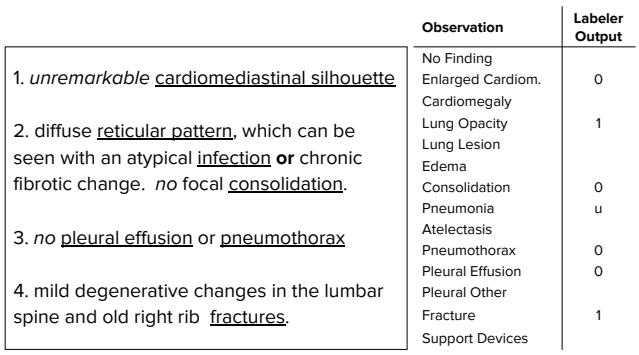


Figure 2: Output of the labeler when run on a report sampled from our dataset. In this case, the labeler correctly extracts all of the mentions in the report (underline) and classifies the uncertainties (bolded) and negations (italicized).

sess the performance of these uncertainty approaches on a validation set of 200 labeled studies, where ground truth is set by a consensus of 3 radiologists who annotated the set using the radiographs. We evaluate the approaches on 5 observations selected based on their clinical significance and prevalence in the dataset, and find that different uncertainty approaches are useful for different observations.

We compare the performance of our final model to 3 additional board certified radiologists on a test set of 500 studies on which the consensus of 5 separate board-certified radiologists serves as ground truth. We find that on 4 out of 5 pathologies, the model ROC and PR curves lie above at least 2 of 3 radiologist operating points. We make our dataset publicly available to encourage further development of models.

Dataset

CheXpert is a large public dataset for chest radiograph interpretation, consisting of 224,316 chest radiographs of 65,240 patients labeled for the presence of 14 observations as positive, negative, or uncertain. We report the prevalences of the labels for the different obsevations in Table 1.

Data Collection and Label Selection

We retrospectively collected chest radiographic studies from Stanford Hospital, performed between October 2002 and July 2017 in both inpatient and outpatient centers, along with their associated radiology reports. From these, we sampled a set of 1000 reports for manual review by a board-certified radiologist to determine feasibility for extraction of observations. We decided on 14 observations based on the prevalence in the reports and clinical relevance, conforming to the Fleischner Society's recommended glossary (Hansell et al. 2008) whenever applicable. "Pneumonia", despite being a clinical diagnosis, was included as a label in order to represent the images that suggested primary infection as the diagnosis. The "No Finding" observation was intended to capture the absence of all pathologies.

Label Extraction from Radiology Reports

We developed an automated rule-based labeler to extract observations from the free text radiology reports to be used as structured labels for the images. Our labeler is set up in three distinct stages: mention extraction, mention classification, and mention aggregation.

Mention Extraction The labeler extracts mentions from a list of observations from the *Impression* section of radiology reports, which summarizes the key findings in the radiographic study. A large list of phrases was manually curated by multiple board-certified radiologists to match various ways observations are mentioned in the reports.

Mention Classification After extracting mentions of observations, we aim to classify them as negative ("no evidence of pulmonary edema, pleural effusions or pneumothorax"), uncertain ("diffuse reticular pattern may represent mild interstitial pulmonary edema"), or positive ("moderate bilateral effusions and bibasilar opacities"). The 'uncertain' label can capture both the uncertainty of a radiologist in the diagnosis as well as ambiguity inherent in the report ("heart size is stable"). The mention classification stage is a 3-phase pipeline consisting of pre-negation uncertainty, negation, and post-negation uncertainty. Each phase consists of rules which are matched against the mention; if a match is found, then the mention is classified accordingly (as uncertain in the first or third phase, and as negative in the second phase). If a mention is not matched in any of the phases, it is classified as positive.

Rules for mention classification are designed on the universal dependency parse of the report. To obtain the universal dependency parse, we follow a procedure similar to Peng et al.(2018): first, the report is split and tokenized into sentences using NLTK (Bird, Klein, and Loper 2009); then, each sentence is parsed using the Bllip parser trained using David McClosky's biomedical model (Charniak and Johnson 2005; McClosky 2010); finally, the universal dependency graph of each sentence is computed using Stanford CoreNLP (De Marneffe et al. 2014).

Category	Mention F1		Negation F1		Uncertain F1	
	NIH	Ours	NIH	Ours	NIH	Ours
Atelectasis	0.976	**0.998**	0.526	**0.833**	0.661	**0.936**
Cardiomegaly	0.647	**0.973**	0.000	**0.909**	0.211	**0.727**
Consolidation	0.996	**0.999**	0.879	**0.981**	0.438	**0.924**
Edema	0.978	**0.993**	0.873	**0.962**	0.535	**0.796**
Pleural Effusion	0.985	**0.996**	0.951	**0.971**	0.553	**0.707**
Pneumonia	0.660	**0.992**	0.703	**0.750**	0.250	**0.817**
Pneumothorax	0.993	**1.000**	0.971	**0.977**	0.167	**0.762**
Enlarged Cardiom.	N/A	0.935	N/A	0.959	N/A	0.854
Lung Lesion	N/A	0.896	N/A	0.900	N/A	0.857
Lung Opacity	N/A	0.966	N/A	0.914	N/A	0.286
Pleural Other	N/A	0.850	N/A	1.000	N/A	0.769
Fracture	N/A	0.975	N/A	0.807	N/A	0.800
Support Devices	N/A	0.933	N/A	0.720	N/A	N/A
No Finding	N/A	0.769	N/A	N/A	N/A	N/A
Macro-average	N/A	0.948	N/A	0.899	N/A	0.770
Micro-average	N/A	0.969	N/A	0.952	N/A	0.848

Table 2: Performance of the labeler of NIH and our labeler on the report evaluation set on tasks of mention extraction, uncertainty detection, and negation detection, as measured by the F1 score. The Macro-average and Micro-average rows are computed over all 14 observations.

Mention Aggregation We use the classification for each mention of observations to arrive at a final label for 14 observations that consist of 12 pathologies as well as the "Support Devices" and "No Finding" observations. Observations with at least one mention that is positively classified in the report is assigned a positive (1) label. An observation is assigned an uncertain (u) label if it has no positively classified mentions and at least one uncertain mention, and a negative label if there is at least one negatively classified mention. We assign ($blank$) if there is no mention of an observation. The "No Finding" observation is assigned a positive label (1) if there is no pathology classified as positive or uncertain. An example of the labeling system run on a report is shown in Figure 2.

Labeler Results

We evaluate the performance of the labeler and compare it to the performance of another automated radiology report labeler on a report evaluation set.

Report Evaluation Set

The report evaluation set consists of 1000 radiology reports from 1000 distinct randomly sampled patients that do not overlap with the patients whose studies were used to develop the labeler. Two board-certified radiologists without access to additional patient information annotated the reports to label whether each observation was mentioned as confidently present (1), confidently absent (0), uncertainly present (u), or not mentioned (blank), after curating a list of labeling conventions to adhere to. After both radiologists independently labeled each of the 1000 reports, disagreements were resolved by consensus discussion. The resulting annotations serve as ground truth on the report evaluation set.

Comparison to NIH labeler

On the radiology report evaluation set, we compare our labeler against the method employed in Peng et al.(2018) which was used to annotate another large dataset of chest radiographs using radiology reports (Wang et al. 2017). We evaluate labeler performance on three tasks: mention extraction, negation detection, and uncertainty detection. For the mention extraction task, we consider any assigned label (1, 0, or u) as positive and $blank$ as negative. On the negation detection task, we consider 0 labels as positive and all other labels as negative. On the uncertainty detection task, we consider u labels as positive and all other labels as negative. We report the F1 scores of the labeling algorithms for each of these tasks.

Table 2 shows the performance of the labeling methods. Across all observations and on all tasks, our labeling algorithm achieves a higher F1 score. On negation detection, our labeling algorithm significantly outperforms the NIH labeler on Atelectasis and Cardiomegaly, and achieves notably better performance on Consolidation and Pneumonia. On uncertainty detection, our labeler shows large gains over the NIH labeler, particularly on Cardiomegaly, Pneumonia, and Pneumothorax.

We note three key differences between our method and the method of Wang et al.(2017). First, we do not the use automatic mention extractors like MetaMap or DNorm, which we found produced weak extractions when applied to our collection of reports. Second, we incorporate several additional rules in order to capture the large variation in the ways negation and uncertainty are conveyed. Third, we split uncertainty classification of mentions into pre-negation and post-negation, which allowed us to resolve cases of uncertainty rules double matching with negation rules in the reports. For example, the following phrase "cannot exclude pneumothorax." conveys uncertainty in the presence of pneumothorax. Without the pre-negation stage, the 'pneumothorax' match is classified as negative due to the 'exclude XXX' rule. However, by applying the 'cannot exclude' rule in the pre-negation stage, this observation can be correctly classified as uncertain.

Model

We train models that take as input a single-view chest radiograph and output the probability of each of the 14 observations. When more than one view is available, the models output the maximum probability of the observations across the views.

Uncertainty Approaches

The training labels in the dataset for each observation are either 0 (negative), 1 (positive), or u (uncertain). We explore different approaches to using the uncertainty labels during the model training.

Ignoring A simple approach to handling uncertainty is to ignore the u labels during training, which serves as a baseline to compare approaches which explicitly incorporate the uncertainty labels. In this approach (called *U-Ignore*), we optimize the sum of the *masked* binary cross-entropy losses

	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion
U-Ignore	0.818 (0.759,0.877)	*0.828 (0.769,0.888)*	0.938 (0.905,0.970)	0.934 (0.893,0.975)	0.928 (0.894,0.962)
U-Zeros	*0.811 (0.751,0.872)*	0.840 (0.783,0.897)	0.932 (0.898,0.966)	0.929 (0.888,0.970)	0.931 (0.897,0.965)
U-Ones	**0.858 (0.806,0.910)**	0.832 (0.773,0.890)	0.899 (0.854,0.944)	0.941 (0.903,0.980)	0.934 (0.901,0.967)
U-SelfTrained	0.833 (0.776,0.890)	0.831 (0.770,0.891)	0.939 (0.908,0.971)	0.935 (0.896,0.974)	0.932 (0.899,0.966)
U-MultiClass	0.821 (0.763,0.879)	**0.854 (0.800,0.909)**	0.937 (0.905,0.969)	0.928 (0.887,0.968)	0.936 (0.904,0.967)

Table 3: AUROC scores on the validation set of the models trained using different approaches to using uncertainty labels. For each of the uncertainty approaches, we choose the best 10 checkpoints per run using the average ROC across the competition tasks. We run each model three times, and take the ensemble of the 30 generated checkpoints on the validation set.

over the observations, masking the loss for the observations which are marked as uncertain for the study. Formally, the loss for an example X is given by

$$L(X, y) = -\sum_{o} \mathbb{1}\{y_o \neq u\}[y_o \log p(Y_o = 1|X) + (1 - y_o) \log p(Y_o = 0|X)],$$

where X is the input image, y is the vector of labels of length 14 for the study, and the sum is taken over all 14 observations. Ignoring the uncertainty label is analogous to the list-wise (complete case) deletion method for imputation (Graham 2009), which is when all cases with a missing value are deleted. Such methods can produce biased models if the cases are not missing completely at random. In this dataset, uncertainty labels are quite prevalent for some observations: for Consolidation, the uncertainty label is almost twice as as prevalent (12.78%) as the positive label (6.78%), and thus this approach ignores a large proportion of labels, reducing the effective size of the dataset.

Binary Mapping We investigate whether the uncertain labels for any of the observations can be replaced by the 0 label or the 1 label. In this approach, we map all instances of u to 0 (*U-Zeroes* model), or all to 1 (*U-Ones* model).

These approaches are similar to zero imputation strategies in statistics, and mimic approaches in multi-label classification methods where missing examples are used as negative labels (Kolesov et al. 2014). If the uncertainty label does convey semantically useful information to the classifier, then we expect that this approach can distort the decision making of classifiers and degrade their performance.

Self-Training One framework for approaching uncertainty labels is to consider them as unlabeled examples, lending its way to semi-supervised learning (Zhu 2006). Most closely tied to our setting is *multi-label learning with missing labels* (MLML) (Wu et al. 2015), which aims to handle multi-label classification given training instances that have a partial annotation of their labels.

We investigate a self-training approach (*U-SelfTrained*) for using the uncertainty label. In this approach, we first train a model using the *U-Ignore* approach (that ignores the u labels during training) to convergence, and then use the model to make predictions that re-label each of the uncertainty labels with the probability prediction outputted by the model. We do not replace any instances of 1 or 0s. On these relabeled examples, we set up loss as the mean of the binary cross-entropy losses over the observations.

Our work follows the approach of (Yarowsky 1995), who train a classifier on labeled examples and then predict on unlabeled examples labeling them when the prediction is above a certain threshold, and repeating until convergence. (Radosavovic et al. 2017) build upon the self-training technique and remove the need for iteratively training models, predicting on transformed versions of the inputs instead of training multiple models, and output a target label for each unlabeled example; soft labels, which are continuous probability outputs rather than binary, have also been used (Hinton, Vinyals, and Dean 2015; Li et al. 2017a).

3-Class Classification We finally investigate treating the u label as its own class, rather than mapping it to a binary label, for each of the 14 observations. We hypothesize that with this approach, we can better incorporate information from the image by supervising uncertainty, allowing the network to find its own representation of uncertainty on different pathologies. In this approach (*U-MultiClass* model), for each observation, we output the probability of each of the 3 possible classes $\{p_0, p_1, p_u\} \in [0, 1], p_0 + p_1 + p_u = 1$. We set up the loss as the mean of the multi-class cross-entropy losses over the observations. At test time, for the probability of a particular observation, we output the probability of the positive label after applying a softmax restricted to the positive and negative classes.

Training Procedure

We follow the same architecture and training process for each of the uncertainty approaches. We experimented with several convolutional neural network architectures, specifically ResNet152, DenseNet121, Inception-v4, and SE-ResNeXt101, and found that the DenseNet121 architecture produced the best results. Thus we used DenseNet121 for all our experiments. Images are fed into the network with size 320×320 pixels. We use the Adam optimizer with default β-parameters of $\beta_1 = 0.9$, $\beta_2 = 0.999$ and learning rate 1×10^{-4} which is fixed for the duration of the training. Batches are sampled using a fixed batch size of 16 images. We train for 3 epochs, saving checkpoints every 4800 iterations.

Validation Results

We compare the performance of the different uncertainty approaches on a validation set on which the consensus of radi-

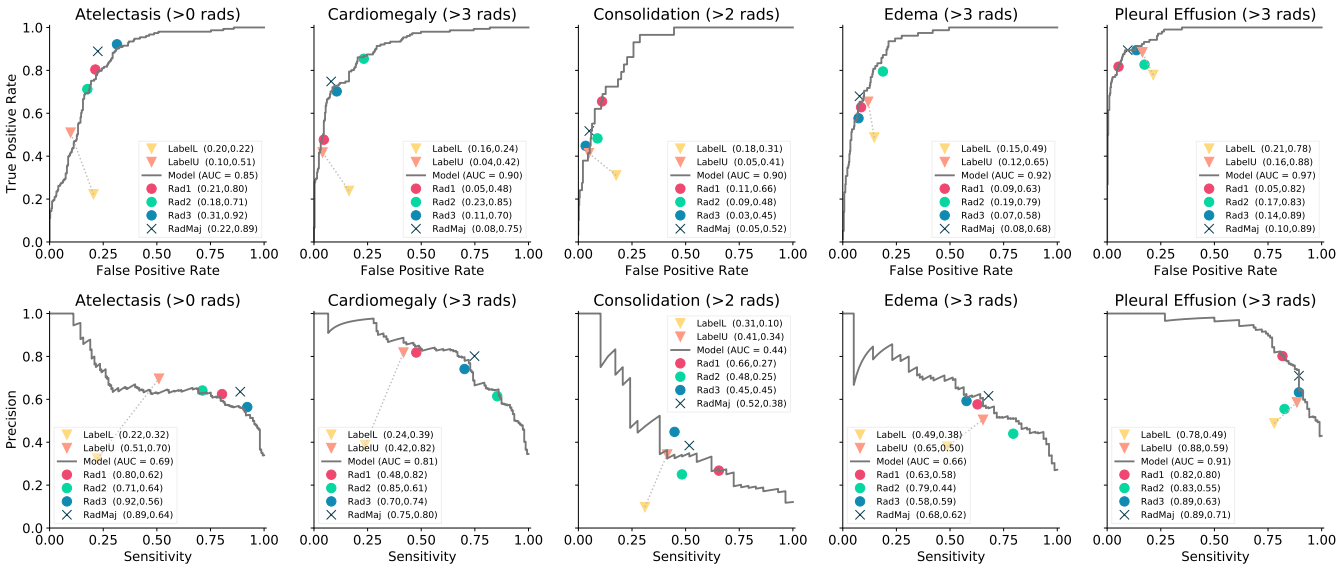


Figure 3: We compare the performance of 3 radiologists to the model against the test set ground truth in both the ROC and the PR space. We examine whether the radiologist operating points lie below the curves to determine if the model is superior to the radiologists. We also compute the lower (LabelL) and upper bounds (LabelU) of the performance of the labels extracted automatically from the radiology report using our labeling system against the test set ground truth.

ologist annotations serves as ground truth.

Validation Set

The validation set contains 200 studies from 200 patients randomly sampled from the full dataset with no patient overlap with the report evaluation set. Three board-certified radiologists individually annotated each of the studies in the validation set, classifying each observation into one of present, uncertain likely, uncertain unlikely, and absent. Their annotations were binarized such that all present and uncertain likely cases are treated as positive and all absent and uncertain unlikely cases are treated as negative. The majority vote of these binarized annotations is used to define a strong ground truth (Gulshan et al. 2016).

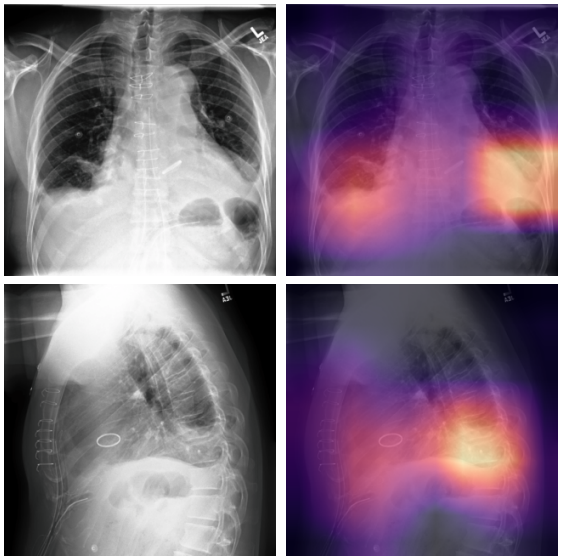
Comparison of Uncertainty Approaches

Procedure We evaluate the approaches using the area under the receiver operating characteristic curve (AUC) metric. We focus on the evaluation of 5 observations which we call the competition tasks, selected based of clinical importance and prevalence in the validation set: (a) Atelectasis, (b) Cardiomegaly, (c) Consolidation, (d) Edema, and (e) Pleural Effusion. We report the 95% two-sided confidence intervals of the AUC using the non-parametric method by DeLong (DeLong, DeLong, and Clarke-Pearson 1988; Sun and Xu 2014). For each pathology, we also test whether the AUC of the best-performing approach is significantly greater than the AUC of the worst-performing approach using the one-sided DeLongs test for two correlated ROC curves (DeLong, DeLong, and Clarke-Pearson 1988). We control for multiple hypothesis testing using the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995); an adjusted p-value < 0.05 indicates statistical significance.

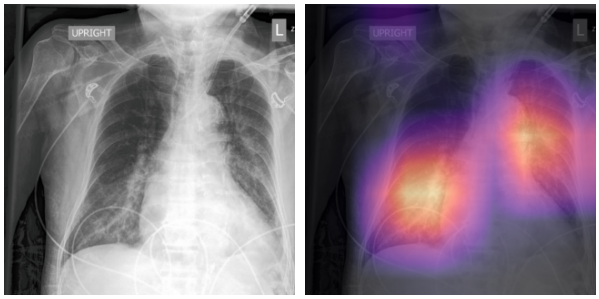
Model Selection For each of the uncertainty approaches, we choose the best 10 checkpoints per run using the average AUC across the competition tasks. We run each model three times, and take the ensemble of the 30 generated checkpoints on the validation set by computing the mean of the output probabilities over the 30 models.

Results The validation AUCs achieved by the different approaches to using the uncertainty labels are shown in Table 3. There are a few significant differences between the performance of the uncertainty approaches. On Atelectasis, the *U-Ones model* (AUC=0.858) significantly outperforms ($p = 0.03$) the *U-Zeros model* (AUC=0.811). On Cardiomegaly, we observe that the *U-MultiClass* model (AUC=0.854) performs significantly better ($p < 0.01$) than the *U-Ignore* model (AUC=0.828). On Consolidation, Edema and Pleural Effusion, we do not find the best models to be significantly better than the worst.

Analysis We find that ignoring the uncertainty label is not an effective approach to handling uncertainty in the dataset, and is particularly ineffective on Cardiomegaly. Most of the uncertain Cardiomegaly cases are borderline cases such as "minimal cardiac enlargement", which if ignored, would likely cause the model to perform poorly on cases which are difficult to distinguish. However, explicitly supervising the model to distinguish between borderline and non-borderline cases (as in the *U-MultiClass* approach) could enable the model to better disambiguate the borderline cases. Moreover, assignment of the Cardiomegaly label when the heart is mentioned in the impression are difficult to categorize in many cases, particularly for common mentions such as "un-



(a) Frontal and lateral radiographs of the chest in a patient with bilateral pleural effusions; the model localizes the effusions on both the frontal (top) and lateral (bottom) views, with predicted probabilities $p = 0.936$ and $p = 0.939$ on the frontal and lateral views respectively.



(b) Single frontal radiograph of the chest demonstrates bilateral mid and lower lung interstitial predominant opacities and cardiomegaly most consistent with cardiogenic pulmonary edema. The model accurately classifies the edema by assigning a probability of $p = 0.824$ and correctly localizes the pulmonary edema. Two independent radiologist readers misclassified this examination as negative or uncertain unlikely for edema.

Figure 4: The final model localizes findings in radiographs using Gradient-weighted Class Activation Mappings. The interpretation of the radiographs in the subcaptions is provided by a board-certified radiologist.

changed appearance of the heart" or "stable cardiac contours" either of which could be used in both enlarged and non-enlarged cases. These cases were classified as uncertain by the labeler, and therefore the binary assignment of 0s and 1s in this setting fails to achieve optimal performance as there is insufficient information conveyed by these modifications.

In the detection of Atelectasis, the *U-Ones* approach performs the best, hinting that the uncertainty label for this observation is effectively utilized when treated as positive. We expect that phrases such as "possible atelectasis" or "may be atelectasis," were meant to describe the most likely findings in the image, rather than convey uncertainty, which supports the good performance of *U-Ones* on this pathology. We suspect a similar explanation for the high performance of *U-Ones* on Edema, where uncertain phrases like "possible mild pulmonary edema" in fact convey likely findings. In contrast, the *U-Ones* approach performs worst on the Consolidation label, whereas the *U-Zeros* approach performs the best. We also note that Atelectasis and Consolidation are often mentioned together in radiology reports. For example, the phrase "findings may represent atelectasis versus consolidation" is very common. In these cases, our labeler assigns uncertain for both observations, but we find that in the ground truth panel review that many of these sorts of uncertainty cases are often instead resolved as Atelectasis-positive and Consolidation-negative.

Test Results

We compare the performance of our final model to radiologists on a test set. We selected the final model based on the best performing ensemble on each competition task on the validation set: *U-Ones* for Atelectasis and Edema, *U-MultiClass* for Cardiomegaly and Pleural Effusion, and *U-SelfTrained* for Consolidation.

Test Set

The test set consists of 500 studies from 500 patients randomly sampled from the 1000 studies in the report test set. Eight board-certified radiologists individually annotated each of the studies in the test set following the same procedure and post-processing as described for the validation set. The majority vote of 5 radiologist annotations serves as a strong ground truth: 3 of these radiologists were the same as those who annotated the validation set and the other 2 were randomly sampled. The remaining 3 radiologist annotations were used to benchmark radiologist performance.

Comparison to Radiologists

Procedure For each of the 3 individual radiologists and for their majority vote, we compute sensitivity (recall), specificity, and precision against the test set ground truth. To compare the model to radiologists, we plot the radiologist operating points with the model on both the ROC and Precision-Recall (PR) space. We examine whether the radiologist operating points lie below the curves to determine if the model is superior to the radiologists. We also compute the performance of the labels extracted automatically

from the radiology report using our labeling system against the test set ground truth. We convert the uncertainty labels to binary labels by computing the upper bound of the labels performance (by assigning the uncertain labels to the ground truth values) and the lower bound of the labels (by assigning the uncertain labels to the opposite of the ground truth values), and plot the two operating points on the curves, denoted *LabelU* and *LabelL* respectively. We also measure calibration of the model before and after applying post-processing calibration techniques, namely isotonic regression (Zadrozny and Elkan 2002) and Platt scaling (Platt and others 1999), using the scaled Brier score (Steyerberg 2008).

Results Figure 3 illustrates these plots on all competition tasks. The model achieves the best AUC on Pleural Effusion (0.97), and the worst on Atelectasis (0.85). The AUC of all other observations are at least 0.9. The model achieves the best AUPRC on Pleural Effusion (0.91) and the worst on Consolidation (0.44). On Cardiomegaly, Edema, and Pleural Effusion, the model achieves higher performance than all 3 radiologists but not their majority vote. On Consolidation, model performance exceeds 2 of the 3 radiologists, and on Atelectasis, all 3 radiologists perform better than the model. On all competition tasks, the lower bound of the report labels lies below the model curves. On all tasks besides Atelectasis, the upper bound of the report label lies on or below the model operating curves. On most of the tasks, the upper bound of the labeler performs comparably to the radiologists. The average scaled Brier score of the model before post-processing calibration is 0.110, after isotonic regression is 0.107, and after platt scaling is 0.101.

Limitations We acknowledge two limitations to performing this comparison. First, neither the radiologists nor the model had access to patient history or previous examinations, which has been shown to decrease diagnostic performance in chest radiograph interpretation (Potchen et al. 1979; Berbaum, Franken, and Smith 1985). Second, no statistical test was performed to assess whether the difference between the performance of the model and the radiologists is statistically significant.

Visualization

We visualize the areas of the radiograph which the model predicts to be most indicative of each observation using Gradient-weighted Class Activation Mappings (Grad-CAMs) (Selvaraju et al. 2016). Grad-CAMs use the gradient of an output class into the final convolutional layer to produce a low resolution map which highlights portions of the image which are important in the detection of the output class. Specifically, we construct the map by using the gradient of the final linear layer as the weights and performing a weighted sum of the final feature maps using those weights. We upscale the resulting map to the dimensions of the original image and overlay the map on the image. Some examples of the Grad-CAMs are illustrated in Figure 4.

Existing Chest Radiograph Datasets

One of the main obstacles in the development of chest radiograph interpretation models has been the lack of datasets with strong radiologist-annotated groundtruth and expert scores against which researchers can compare their models. There are few chest radiographic imaging datasets that are publicly available, but none of them have test sets with strong ground truth or radiologist performances. The Indiana Network for Patient Care hosts the OpenI dataset (Demner-Fushman et al. 2015) consisting of 7,470 frontal-view radiographs and radiology reports which have been labeled with key findings by human annotators . The National Cancer Institute hosts the PLCO Lung dataset (Gohagan et al. 2000) of chest radiographs obtained during a study on lung cancer screening . The dataset contains 185,421 full resolution images, but due to the nature of the collection process, it is has a low prevalence of clinically important pathologies such as Pneumothorax, Consolidation, Effusion, and Cardiomegaly. The MIMIC-CXR dataset (Rubin et al. 2018) has been recently announced but is not yet publicly available.

The most commonly used benchmark for developing chest radiograph interpretation models has been the ChestX-ray14 dataset (Wang et al. 2017). Due to the introduction of this large dataset, substantial progress has been made towards developing automated chest radiograph interpretation models (Yao et al. 2017; Rajpurkar et al. 2017; Li et al. 2017b; Kumar, Grewal, and Srivastava 2018; Wang et al. 2018; Guan et al. 2018; Yao et al. 2018). However, using the NIH dataset as a benchmark on which to compare models is problematic as the labels in the test set are extracted from reports using an automatic labeler. The CheXpert dataset that we introduce features radiologist-labeled validation and test sets which serve as strong reference standards, as well as expert scores to allow for robust evaluation of different algorithms.

Conclusion

We present a large dataset of chest radiographs called CheXpert, which features uncertainty labels and radiologist-labeled reference standard evaluation sets. We investigate a few different approaches to handling uncertainty and validate them on the evaluation sets. On a test set with a strong ground truth, we find that our best model outperforms at least 2 of the 3 radiologists in the detection of 4 clinically relevant pathologies. We hope that the dataset will help development and validation of chest radiograph interpretation models towards improving healthcare access and delivery worldwide.

Acknowledgements

We would like to thank Luke Oakden-Rayner, Yifan Peng, and Susan C. Weber for their help in this work.

References

[Benjamini and Hochberg 1995] Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* 289–300.

[Berbaum, Franken, and Smith 1985] Berbaum, K.; Franken, J. E.; and Smith, W. 1985. The effect of comparison films upon resident interpretation of pediatric chest radiographs. *Investigative radiology* 20:124–128.

[Bird, Klein, and Loper 2009] Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

[Charniak and Johnson 2005] Charniak, E., and Johnson, M. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, 173–180. Association for Computational Linguistics.

[De Marneffe et al. 2014] De Marneffe, M.-C.; Dozat, T.; Silveira, N.; Haverinen, K.; Ginter, F.; Nivre, J.; and Manning, C. D. 2014. Universal stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, 4585–4592.

[DeLong, DeLong, and Clarke-Pearson 1988] DeLong, E. R.; DeLong, D. M.; and Clarke-Pearson, D. L. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 837–845.

[Demner-Fushman et al. 2015] Demner-Fushman, D.; Kohli, M. D.; Rosenman, M. B.; Shooshan, S. E.; Rodriguez, L.; Antani, S.; Thoma, G. R.; and McDonald, C. J. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association* 23(2):304–310.

[Gohagan et al. 2000] Gohagan, J. K.; Prorok, P. C.; Hayes, R. B.; and Kramer, B.-S. 2000. The prostate, lung, colorectal and ovarian (plco) cancer screening trial of the national cancer institute: history, organization, and status. *Controlled clinical trials* 21(6):251S–272S.

[Graham 2009] Graham, J. W. 2009. Missing data analysis: Making it work in the real world. *Annual review of psychology* 60:549–576.

[Guan et al. 2018] Guan, Q.; Huang, Y.; Zhong, Z.; Zheng, Z.; Zheng, L.; and Yang, Y. 2018. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. *arXiv preprint arXiv:1801.09927*.

[Gulshan et al. 2016] Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M. C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* 316(22):2402–2410.

[Hansell et al. 2008] Hansell, D. M.; Bankier, A. A.; MacMahon, H.; McLoud, T. C.; Muller, N. L.; and Remy, J. 2008. Fleischner society: glossary of terms for thoracic imaging. *Radiology* 246(3):697–722.

[Hinton, Vinyals, and Dean 2015] Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

[Kolesov et al. 2014] Kolesov, A.; Kamyshenkov, D.; Litovchenko, M.; Smekalova, E.; Golovizin, A.; and Zhavoronkov, A. 2014. On multilabel classification methods of incompletely labeled biomedical text data. *Computational and mathematical methods in medicine* 2014.

[Kumar, Grewal, and Srivastava 2018] Kumar, P.; Grewal, M.; and Srivastava, M. M. 2018. Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs. In *International Conference Image Analysis and Recognition*, 546–552. Springer.

[Li et al. 2017a] Li, Y.; Yang, J.; Song, Y.; Cao, L.; Luo, J.; and Li, L.-J. 2017a. Learning from noisy labels with distillation. In *ICCV*, 1928–1936.

[Li et al. 2017b] Li, Z.; Wang, C.; Han, M.; Xue, Y.; Wei, W.; Li, L.-J.; and Li, F.-F. 2017b. Thoracic disease identification and localization with limited supervision. *arXiv preprint arXiv:1711.06373*.

[McClosky 2010] McClosky, D. 2010. Any domain parsing: automatic domain adaptation for natural language parsing.

[Peng et al. 2018] Peng, Y.; Wang, X.; Lu, L.; Bagheri, M.; Summers, R.; and Lu, Z. 2018. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings* 2017:188.

[Platt and others 1999] Platt, J., et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10(3):61–74.

[Potchen et al. 1979] Potchen, E.; Gard, J.; Lazar, P.; Lahaie, P.; and Andary, M. 1979. Effect of clinical history data on chest film interpretation-direction or distraction. In *Investigative Radiology*, volume 14, 404–404. LIPPINCOTT-RAVEN PUBL 227 EAST WASHINGTON SQ, PHILADELPHIA, PA 19106.

[Radosavovic et al. 2017] Radosavovic, I.; Dollár, P.; Girshick, R.; Gkioxari, G.; and He, K. 2017. Data distillation: Towards omni-supervised learning. *arXiv preprint arXiv:1712.04440*.

[Rajpurkar et al. 2017] Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; Lungren, M. P.; and Ng, A. Y. 2017. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv:1711.05225 [cs, stat]*. arXiv: 1711.05225.

[Rubin et al. 2018] Rubin, J.; Sanghavi, D.; Zhao, C.; Lee, K.; Qadir, A.; and Xu-Wilson, M. 2018. Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks. *arXiv preprint arXiv:1804.07839*.

[Selvaraju et al. 2016] Selvaraju, R. R.; Das, A.; Vedantam, R.; Cogswell, M.; Parikh, D.; and Batra, D. 2016. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR, abs/1610.02391* 7.

[Steyerberg 2008] Steyerberg, E. W. 2008. *Clinical prediction models: a practical approach to development, validation, and updating*. Springer Science & Business Media.

[Sun and Xu 2014] Sun, X., and Xu, W. 2014. Fast implementation of delongs algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters* 21(11):1389–1393.

[Wang et al. 2017] Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3462–3471. Honolulu, HI: IEEE.

[Wang et al. 2018] Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; and Summers, R. M. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9049–9058.

[Wu et al. 2015] Wu, B.; Lyu, S.; Hu, B.-G.; and Ji, Q. 2015. Multi-label learning with missing labels for image annotation and facial action unit recognition. *Pattern Recognition* 48(7):2279–2289.

[Yao et al. 2017] Yao, L.; Poblenz, E.; Dagunts, D.; Covington, B.; Bernard, D.; and Lyman, K. 2017. Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv preprint arXiv:1710.10501*.

[Yao et al. 2018] Yao, L.; Prosky, J.; Poblenz, E.; Covington, B.; and Lyman, K. 2018. Weakly supervised medical diagnosis and localization from multiple resolutions. *arXiv preprint arXiv:1803.07703*.

[Yarowsky 1995] Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the*

33rd annual meeting on Association for Computational Linguistics, 189–196. Association for Computational Linguistics.

[Zadrozny and Elkan 2002] Zadrozny, B., and Elkan, C. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 694–699. ACM.

[Zhu 2006] Zhu, X. 2006. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison* 2(3):4.