

TextRay: Mining Clinical Reports to Gain a Broad Understanding of Chest X-rays

Jonathan Laserson¹, Christine Dan Lantsman², Michal Cohen-Sfady¹, Itamar Tamir³, Eli Goz¹, Chen Brestel¹, Shir Bar⁴, Maya Atar⁵, and Eldad Elnekave¹

¹ Zebra Medical Vision LTD, Shefayim, Israel jonil@zebra-med.com

² Sheba Medical Center and Tel Aviv University, Israel

³ Rabin Medical Center, Israel

⁴ Technion, Israel Institute of Technology

⁵ Ben Gurion University, Israel

Abstract. The chest X-ray (CXR) is by far the most commonly performed radiological examination for screening and diagnosis of many **cardiac and pulmonary** diseases. There is an immense world-wide shortage of physicians capable of providing rapid and accurate interpretation of this study. A radiologist-driven analysis of over two million CXR reports generated an ontology including the 40 most prevalent pathologies on CXR. By manually tagging a relatively small set of sentences, we were able to construct **a training set of 959k** studies. A deep learning model was trained to predict the findings given the patient frontal and lateral scans. For 12 of the findings we compare the model performance against a team of radiologists and show that in most cases the radiologists agree on average more with the algorithm than with each other.

Keywords: radiology, chest x-ray, deep learning

1 Introduction

Chest X-rays (CXRs) are the most commonly performed radiology examination world-wide, with over **150 million obtained annually** in the United States alone. CXRs are a cornerstone of **acute triage** as well as longitudinal surveillance. Despite the ubiquity of the exam and its apparent technical simplicity, the chest x ray is widely regarded among radiologists as among the most difficult to master[1].

Due to a shortage in supply of radiologists, radiographic technicians are increasingly called upon to provide preliminary interpretations, particularly in Europe and Africa. In the US, non-radiology physicians often provide preliminary or definitive readings of CXRs, decreasing the waiting interval at the nontrivial expense of diagnostic accuracy.

Even among expert radiologists, clinically substantial errors are made in 3-6% of studies[1,2], with minor errors seen in 30% [3]. Accurate diagnosis of some entities is particularly challenging: early lung cancer for example is missed in 19-54% of cases, with similar sensitivity figures described for **pneumothorax**

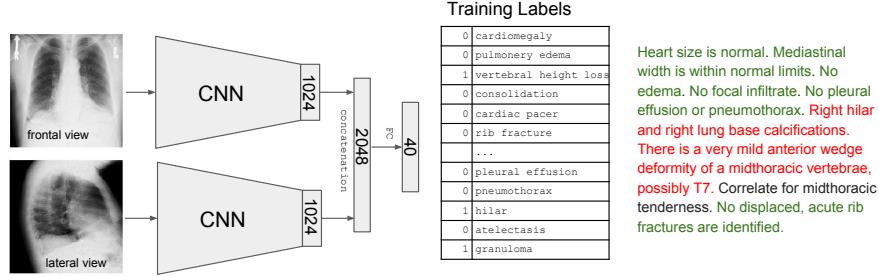


Fig. 1. TextRay Model Illustration. Frontal (PA) and lateral view images each go through a separate CNN. A fully-connected layer is applied on their concatenated feature vectors and emits the confidence for each finding. Training labels were extracted by analyzing the report sentences. Negative (green) and positive (red) sentences identified. Findings in positive sentences receive a positive training label. Negative or unmentioned findings receive a negative label.

and rib fracture detection. The likelihood for major diagnostic errors is directly correlated with both shift length and volume of examinations being read[4], a reminder that diagnostic accuracy varies substantially even at different times of the day for a given radiologist.

Hence there exists an immense unmet need and opportunity to provide immediate, consistent and expert-level insight into every CXR. In the present work we describe a novel methodology employed in this endeavor and we present the results achieved using a robust method of clinical validation.

2 Material and Methods

Data All Patient Health Information (PHI) was removed from the data prior to acquisition in compliance with HIPAA standards. We utilized a dataset of 2.1 million CXRs with their respective diagnostic reports. All postero-anterior (PA) CXR films of individuals aged 18 and above were procured. Corresponding lateral views were present in 85% of the CXR examinations and were included in the study data.

Textual Analysis A standardization process was employed whereby all CXR reports were reduced to a set of distinct canonical labels. First, a sentence boundary detection algorithm was applied to the 2.1M reports, yielding a pool of 827k unique sentences. Three expert radiologists and two medical students categorized the most occurring sentences with respect to their pertinence to CXR images.

Three categories emerged: sentences that report the presence or absence of a finding, for example "the heart is enlarged", or "normal cardiac shadow", and could be used as labels; neutral sentences, which referenced information not

derived from or inherently related to the image itself, for example: "84 year old man with cough", "lung nodule follow up", or "comparison made to CT chest".

A third category of sentences could render the study unreliable for training due to ambiguity regarding the relationship of the text to the image, for example "no change in the appearance of the chest since yesterday".

After filtering out neutral and negative sentences using a few hand-crafted regular expressions, it was possible to fully cover 826k reports using just the 20k most prevalent positive sentences. The same expert radiologists reviewed each of these sentences and mapped them to an initial ontology of 60 findings which covered 99.99% of all positive sentence volume.

In making the final ontology, we focused on visual findings rather than clinical interpretations or diagnoses. We chose to merge some categories: osteoporosis was merged into *osteopenia*, twisted and uncoiled aorta into *abnormal aorta*, and bronchial markings into *interstitial markings*, since it is often impossible to differentiate these based on the image alone. Although visually distinct, all tubes and venous lines were consolidated into two respective categories. The resulting 40 categories are presented in Table 1.

Table 1. Number of studies with each finding in our data. 596k (62%) of the total 959k studies had no reported findings.

#	finding	total	%	#	finding	total	%
1	abnormal aorta	15,932	1.66	21	mass	633	0.07
2	aortic calcification	11,508	1.20	22	mediastinal widening	1,639	0.17
3	artificial valve	5,847	0.61	23	much bowel gas	441	0.05
4	atelectasis	5,492	0.57	24	nodule	553	0.06
5	bronchial wall thickening	2,773	0.29	25	orthopedic surgery	717	0.07
6	cardiac pacer	17,378	1.81	26	osteopenia	5,585	0.58
7	cardiomegaly	95,137	9.92	27	pleural effusion	16,688	1.74
8	central line	3,802	0.40	28	pleural thickening	8,164	0.85
9	consolidation	34,260	3.57	29	pneumothorax	741	0.08
10	costophrenic angle blunting	13,673	1.43	30	pulmonary edema	8,637	0.90
11	degenerative changes	18,545	1.93	31	rib fracture	4,607	0.48
12	elevated diaphragm	21,913	2.28	32	scoliosis	4,907	0.51
13	fibrotic changes	11,027	1.15	33	soft tissue calcifications	1,086	0.11
14	fracture	526	0.05	34	sternotomy wires	45,002	4.69
15	granuloma	1,475	0.15	35	surgical clips noted	8,147	0.85
16	hernia diaphragm	8,892	0.93	36	thickening of fissure	1,714	0.18
17	hilar prominence	10,407	1.08	37	trachea deviation	601	0.06
18	hyperinflation	37,319	3.89	38	transplant	5,180	0.54
19	interstitial markings	97,703	10.18	39	tube	2,025	0.21
20	kyphosis	5,531	0.58	40	vertebral height loss	1,212	0.13

Training Set Generation On completion of sentence labeling, we set out to design the appropriate training set. A conservative approach would only include studies whose report sentences were *fully-covered*, i.e. every potentially positive

sentence in them was manually reviewed and mapped to a finding. A more permissive *any-hit* approach would include any study with a recognized positive sentence in its report, ignoring other unrecognized sentences, with the risk that some of them also mention abnormalities that would be mislabeled as negatives.

The *fully-covered* approach yielded 596k normal studies (no positive findings), and 230k abnormal studies. The *any-hit* approach, while noisier, added 58% more abnormal studies, for a total of 363k. Hence our final training set had 826k studies in the *fully-covered* approach, and 959k studies in the *any-hit* approach.

Additionally, many radiologists will omit mention of normal structures in favor of brevity, thereby implying a negative label. This bias extends to many studies in which even mildly abnormal or senescent changes are omitted. For example, the same CXR may produce a single-line report of "No acute disease" by one radiologist and descriptions of cardiomegaly, and degenerative changes by another radiologist. Inherently, this omission bias introduces noise into the labeling process, particularly for findings which are not deemed critical, even in the more conservative *fully-covered* training set.

We decided to compare both approaches, and took the larger *any-hit* training set as our baseline. To the best of our knowledge, this is the largest training set ever assembled for chest X-ray, both in terms of the number of studies and the number of labels (see Table 1 for its composition). We partitioned the training set into *training*, *validation*, and *testing* (80%/10%/10% respectively), based on the (anonymized) patient identity. From the 10% of studies designated as *validation* we compiled a validation set of size 994 with at least 25 positives from each finding. We picked the model with lowest validation loss.

2.1 Model

Our model, called TextRay, is illustrated in Fig. 1. We start by applying a CNN (DenseNet121[5]) on the Lateral and PA views (**separately**). We removed the last fully connected layer from each CNN and concatenated their outputs (just after the average pooling layer). We then applied our own fully-connected layer resulting in $K = 40$ outputs, one for each finding, followed by a **sigmoid activation**. Hence, our model treats each study as a bag of findings, reporting the confidence for each one. We used the mean of the **binary cross-entropy losses** as our main loss function:

$$loss = \frac{1}{K} \sum_{k=1}^K y_k \log(p_k) + (1 - y_k) \log(1 - p_k)$$

where p_k is the value of the k -th output unit and y_k is the binary label for the k -th finding.

Our model receives two inputs of size 299x299. When lateral view was unavailable, we fed the network with random noise instead. Each X-Ray image (up to 3000x3000 pixels in raw format) was zero-mean-normalized, rescaled to a size of $330(1 + a) \times 330(1 + b)$, and rotated c degrees. A random patch of 299x299

was taken as input. For **training augmentation**, we sampled a, b uniformly from ± 0.09 and c from ± 9 , randomly flipping each image horizontally. For balance, we replaced the PA view with random noise in 5% of the samples. For test we used $a = b = c = 0$ and took the central patch as input, without flipping.

We trained on two 1080Ti GPUs, putting each CNN on a different GPU. We used the built-in Keras 2.1.3 implementation of DenseNet121 over Tensorflow 1.4. We used the Adam optimizer with Keras default parameters, and a batch size of 32. We sorted the studies in two queues, normals and abnormalities, and filled each batch with 95% abnormal studies on average. An epoch was defined as 150 batches. We started with a learning rate of 0.001 and multiplied it by 0.75 if validation loss hadn't improved for 30 epochs. We trained for 2000 epochs.

2.2 Evaluation Sets

We chose 12 of the 40 findings and prepared evaluation sets for them, using studies from the *test* partition. Most sets focused on a single finding except *cardiomegaly*, *hilar prominence*, and *pulmonary edema*, which were lumped together as they are commonly seen in the setting of congestive heart failure. In each set, the studies were derived from two pools: *pos-pool* are studies that the reports indicated as positive for that finding. These studies were obtained by a manual textual search for terms indicative for each finding, independently of our sentence-tagging operation; *neg-pool* are randomly sampled studies, which are mostly negative for any finding (see Table 2 for the sets composition).

Each set was evaluated by three expert radiologists. In each set, the radiologist reviewed the shuffled studies and indicated the presence or absence of the relevant finding, using a web-based software operated on a desktop. The radiologists were shown both PA and Lateral view in their original resolutions.

We considered the report as a fourth expert opinion. To measure the accuracy of the label-extraction process, we cross referenced the report opinion with the training set labels. The positive labels in the training set were accurately mentioned in the report; frequently, positive findings mentioned in the reports were mislabeled as negatives, (see Table S5) as would be expected in the *any-hit* training set, but this was also observed to lesser degree even in the *fully-covered* set.

3 Results

We performed pairwise analysis of the radiologist agreement following the procedure in [6], except we used the **agreement rate** between two taggers (e.g. accuracy) instead of the F1 score, because (a) it also measures agreement on the negatives; and (b) it is easier to interpret. The *average agreement rate* (AAR) for a radiologist (or a model) is the average of the agreement rates achieved against the other two (three for a model) radiologists. The *avg. radiologist rate* is the mean of the three radiologists' AARs. We used the bootstrap method ($n = 10000$) to obtain 95% confidence intervals over the difference between TextRay and the average radiologist agreement rates. As TextRay's threshold for each finding, we used the one that maximized the AAR on the validation set.

Table 2. Evaluation Sets. The number of studies taken from the *pos-pool* (finding is positive in report) and *neg-pool* (random sample) are indicated, along with the average agreement rate (AAR) of the 3 radiologists (rads) assigned to each set vs. the report. The AAR between our model and the rads (column *textray*) is compared against the AAR between any radiologist and the other rads (*avg. rad.*). Confidence intervals are computed over the difference ($\Delta = \text{textray} - \text{avg. rad.}$).

finding	pool		avg. agreement w/ rads		Δ (CI)	
	pos	neg	report	avg. rad	textray	textray vs. rads
pulmonary edema	128	482	0.613	0.639	0.730	+0.09 (0.07, 0.11)
elevated diaphragm	202	77	0.731	0.675	0.754	+0.08 (0.05, 0.10)
abnormal aorta	198	80	0.736	0.693	0.771	+0.08 (0.05, 0.11)
hyperinflation	95	80	0.678	0.619	0.657	+0.04 (-0.02, 0.10)
vertebral height loss	126	55	0.781	0.742	0.757	+0.02 (-0.02, 0.06)
atelectasis	201	78	0.778	0.756	0.767	+0.01 (-0.03, 0.04)
cardiomegaly	238	372	0.755	0.861	0.866	+0.01 (-0.02, 0.03)
pleural effusion	207	73	0.905	0.893	0.896	+0.00 (-0.02, 0.03)
consolidation	194	78	0.690	0.730	0.707	-0.02 (-0.07, 0.02)
pneumothorax	111	124	0.830	0.855	0.823	-0.03 (-0.08, 0.01)
rib fracture	183	76	0.683	0.799	0.745	-0.05 (-0.10, -0.01)
hilar prominence	184	426	0.552	0.797	0.736	-0.06 (-0.09, -0.03)

Table 2 shows that TextRay is on par with human radiologists (within the 95% CI) on 10 out of 12 findings, with the exception of *rib fracture* and *hilar prominence*. On some findings (*elevated diaphragm*, *abnormal aorta*, and *pulmonary edema*), radiologists agree significantly more with our algorithm than with each other (e.g. the CI does not include 0). Table 2 also shows the average agreement of the radiologists with the report. Here as well, this agreement is often higher than the average agreement among the radiologists themselves. This provides evidence that the noise added by using the reports as labels is no larger than the noise added by training a radiologist to do the tagging.

Using our text-based labels as ground-truth, TextRay’s performance was then tested over all 40 findings. To create the test set, a random sample of 5,000 studies was chosen from the *test* partition. Then, more studies were added from the partition until each finding had at least 100 positive cases, for a total of 7,030 studies. The ROC plots are shown in supp. Fig. 4, with their AUCs ranging between 0.7 and 1.0 (average 0.892). At the top of the chart, artificial objects (i.e. pacers, lines, tubes, wires, and implants) are detected with AUCs approaching 1.0, much better than all diseases.

Fig. 2 shows the area under the ROC curve (AUC) achieved by our model compared to a variant that was trained only with the PA view of each study (the approach used in [6,7]). We see that in most findings, the performance is similar, but *vertebral height loss*, *consolidation*, *rib fracture*, and *kyphosis* stand out as findings in which the lateral view improved detection. These findings are expected from a clinical radiographic perspective.

For comparison, we also trained a variant of TextRay with the *fully-covered* training set, but it achieved significantly lower results in almost all findings (see

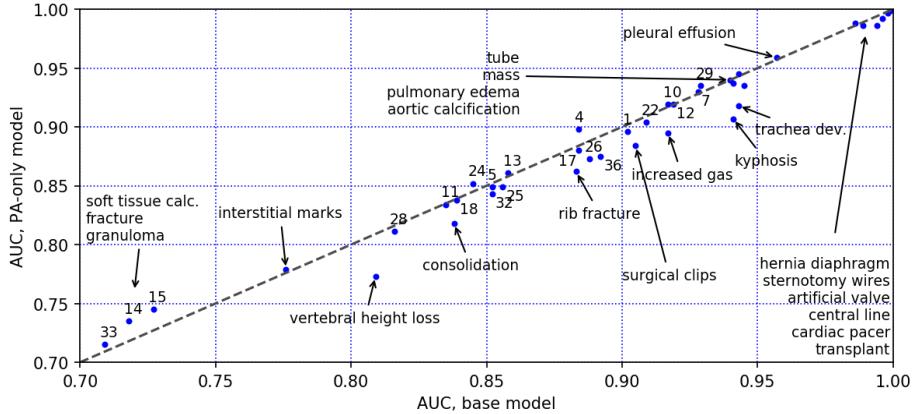


Fig. 2. Area under the ROC curve (AUC) of our base model vs. the PA-only variant over 40 chest X-ray findings. The numbers refer to the index of Table 1. A cluster of labels should be mapped left-to-right.

supp. Table 6), suggesting that the additional abnormal studies in the *any-hit* set, more than compensated for the higher label noise. Finally, we draw heat maps based on the procedure presented in [7], and present them in Supp. Fig. 5.

4 Discussion

The extraction of labels from full CXR reports has been recognized as essential for efficient and robust CNN training on large datasets. Shin et al.[8] extracted labels from the 3,955 CXR reports in the OpenI dataset, using the MeSH system[9]. The ChestX-ray14 dataset released by Wang et al.[7] contains 112k PA images loosely labeled using a combination of NLP and hand-crafted rules. Rajpurkar et al.’s [6] team of four radiologists reported a high degree of disagreement with the provided ChestX-ray14 labels in general, although they demonstrate the ability to achieve expert-level prediction for the presence of pneumonia after training upon a DenseNet121 CNN.

Utilizing several public datasets with image labels and reports provided, Jing et al.[10] built a system that can generate a natural appearing radiology report using a hierarchical RNN. The high-level RNN generates sentence embeddings that seed low-level RNNs that produce the words of each sentence. As part of their report generation, they also produce tags representing the clinical finding present in the image. Interestingly, the model trained using these tags and the text of the reports did not predict the tags better than the model that was trained just using the tags. The ultimate accuracy of the system however remains poorly defined due to lack of clinical radiologic validation.

To the best of our knowledge, the present study is the first to utilize extensive radiology expertise for both multi-label generation and visual validation of algorithmic results. Study labels were generated bottom-up via ontology-based

methodology which was rooted in the text rather than pre-existing categories or tags (i.e. MeSH). We trained upon the largest dataset of CXRs described to date, achieving results on twelve distinct visual findings which are on par with inter-radiologist agreement and in some cases, better.

5 Conclusion

In this work we attempt to broadly cover all findings radiologists usually report when reviewing a PA and Lateral chest X-ray. Since a relatively small set of sentences is heavily re-used in CXR reports, we were able to generate organic labels for millions of reports by examining and indexing twenty thousand individual sentences. This massive amount of data allowed us to obtain radiology-level detection performance on various of findings using a single model, in essence distilling the insight of millions of radiographic interpretations into software code. Application of a similar technique upon AP chest X-ray scans, musculoskeletal and abdominal radiographies is currently ongoing.

References

1. Robinson, P.J., Wilson, D., Coral, A., Murphy, A., Verow, P.: Variation between experienced observers in the interpretation of accident and emergency radiographs. *The British journal of radiology* **72**(856) (4 1999) 323–30
2. Brady, A., Laoide, R., McCarthy, P., McDermott, R.: Discrepancy and error in radiology: concepts, causes and consequences. *The Ulster medical journal* **81**(1) (1 2012) 3–9
3. Bruno, M.A., Walker, E.A., Abujudeh, H.H.: Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction. *RadioGraphics* **35**(6) (10 2015) 1668–1676
4. Hanna, T.N., Lamoureux, C., Krupinski, E.A., Weber, S., Johnson, J.O.: Effect of Shift, Schedule, and Volume on Interpretive Accuracy: A Retrospective Analysis of 2.9 Million Radiologic Examinations. *Radiology* (11 2017) 170555
5. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2017)
6. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M.P., Ng, A.Y.: CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. (11 2017)
7. Wang, X., Peng, Y., Lu, L., ..., Z.L..o.C.V., 2017, u.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *openaccess.thecvf.com*
8. Shin, H.C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., Summers, R.M.: Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In: *Computer Vision and Pattern Recognition (CVPR)*. (June 2016)
9. Demner-Fushman, D., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R.: Annotation of Chest Radiology Reports for Indexing and Retrieval. (2015) 99–111
10. Jing, B., Xie, P., Xing, E.: On the Automatic Generation of Medical Imaging Reports. (11 2017)

Supplementary Material

Table 3. Most frequent positive sentences and their occurrences in reports.

sentence	#reports	sentence	#reports
The heart is enlarged	39,245	Twisted aorta	6,771
The heart is widened	20,270	Infiltrate?	6,540
Enlarged heart	14,689	Increased lung volume	6,494
Chronic bronchial changes	9,515	After sternotomy	6,268
Enhanced interstitial markings in the lungs	9,216	Interstitial changes in the lungs	5,303
Permanent cardiac pacer	6,881	Hyperinflation	5,064

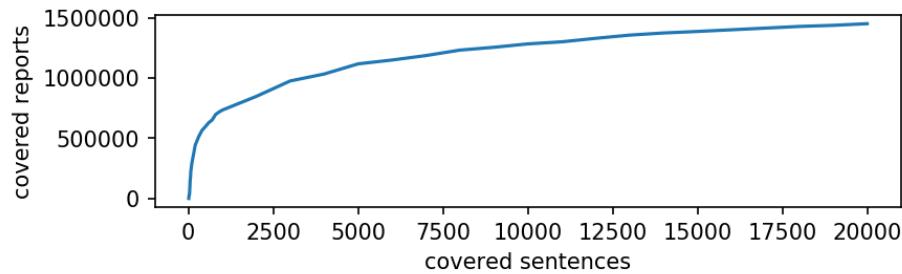


Fig. 3. Number of reports fully-covered by tagged sentences as a function of the number of tagged sentences (assuming we tag the most commons ones).

finding	AAR of each radiologist						rad. average
	A	B	C	D	E	F	
abnormal aorta	0.75	0.72		0.6			0.69
atelectasis		0.77	0.73	0.77			0.76
cardiomegaly		0.85	0.86		0.87		0.86
elevated diaphragm			0.67	0.6	0.75		0.68
hilar prominence		0.79	0.79		0.82		0.80
hyperinflation	0.65		0.69	0.52			0.62
consolidation	0.77	0.72	0.7				0.73
pleural effusion	0.91		0.88	0.9			0.89
pneumothorax			0.87	0.84	0.85		0.86
pulmonary edema		0.71	0.61		0.6		0.64
rib fracture			0.8	0.79	0.81		0.80
vertebral height loss	0.75	0.8		0.68			0.74

Table 4. Evaluation sets and their assigned taggers (marked with letters A-G). The taggers A-G are attending radiologists with 40, 6, 5, 5, 2, 2, and 2 years of experience, respectively. The numbers indicate the average agreement rate (AAR) of each radiologist vs. the other two radiologists in the set.

finding	% pos studies included		% pos findings correctly labeled	
	fully-covered	any-hit	fully-covered	any-hit
abnormal aorta	44.9	81.8	97.8	87.7
atelectasis	13.9	64.2	78.6	47.3
cardiomegaly	55.0	95.0	100.0	99.1
elevated diaphragm	42.1	76.7	95.3	81.9
hilar prominence	43.5	84.2	96.2	86.5
hyperinflation	56.8	85.3	90.7	91.4
consolidation	24.2	50.5	87.2	58.2
pleural effusion	34.3	80.7	94.4	70.1
pneumothorax	19.8	56.8	50.0	39.7
pulmonary edema	57.0	93.8	86.3	75.0
rib fracture	34.4	63.4	82.5	61.2
vertebral height loss	20.6	64.3	73.1	39.5

Table 5. Estimation of label noise in two training sets. The *fully-covered* training set only includes a study if all its potentially positive sentences were parsed and mapped to their respective findings. The *any-hit* training set includes a study if at least one sentence was parsed and mapped to a finding, even if the rest of the positive sentences were not parsed and their findings are unknown. The *any-hit* training set includes more positive studies for every finding (left 2 columns), but a larger portion of those positive findings is erroneously labeled as negative (right 2 columns).

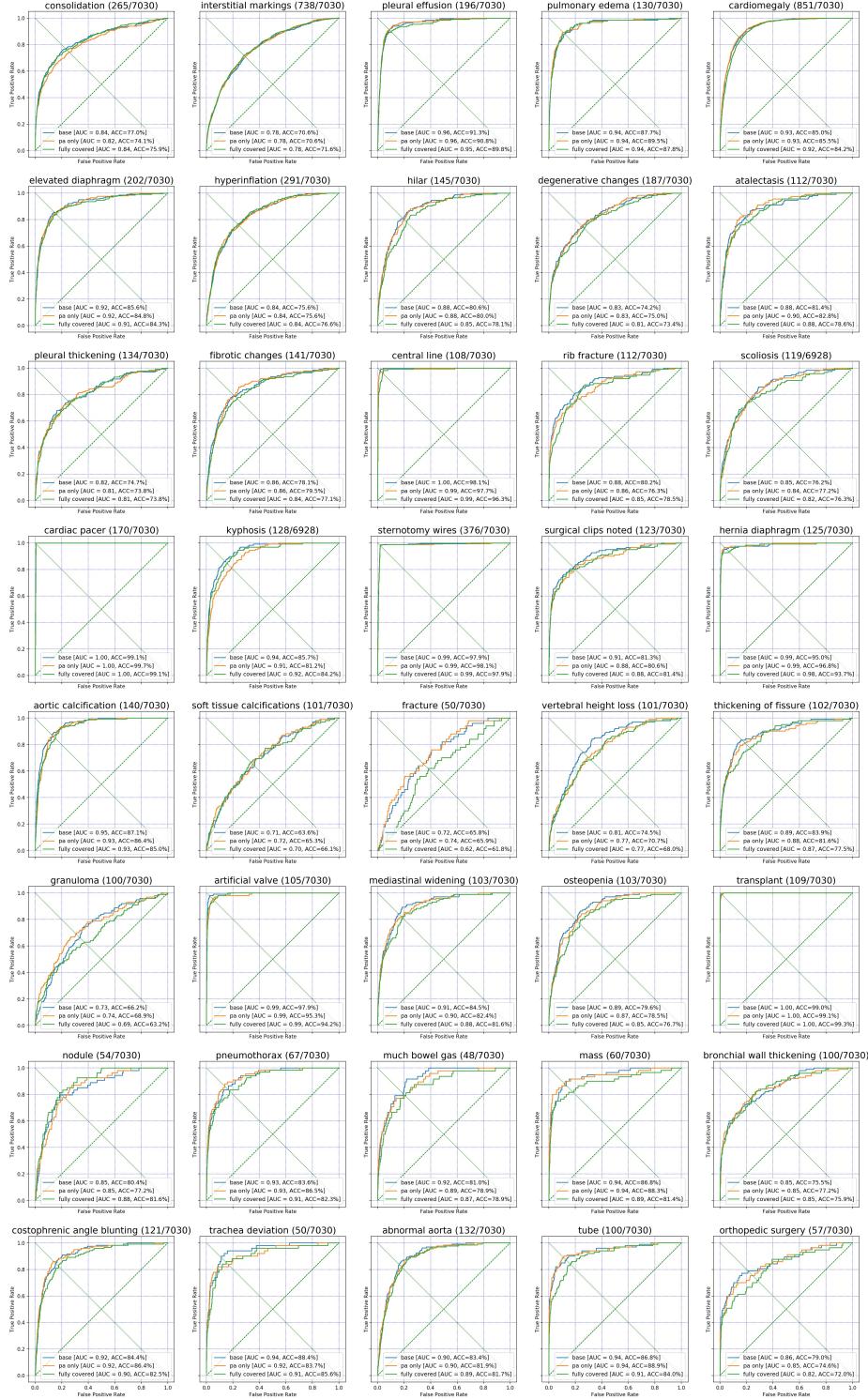


Fig. 4. ROC plots of our base model and its two variants on 40 chest X-ray findings. The title for each plot indicates the number positives in the test set of 7,030 studies. Indicated on each plot is the AUC and the accuracy when sensitivity=specificity.

Table 6. Model performance per finding, measured in AUC. PA: model trained only with PA view. FC: model trained only on studies with *fully-covered* reports.

#	finding	base	PA	FC	#	finding (cont.)	base	PA	FC
1	abnormal aorta	0.902	0.896	0.887	21	mass	0.941	0.937	0.894
2	aortic calcification	0.945	0.935	0.930	22	mediastinal widening	0.909	0.904	0.885
3	artificial valve	0.994	0.986	0.989	23	increased bowel gas	0.917	0.895	0.867
4	atelectasis	0.884	0.898	0.877	24	nodule	0.845	0.852	0.882
5	bronchial wall thick	0.852	0.849	0.852	25	orthopedic surgery	0.856	0.849	0.817
6	cardiac pacer	0.998	0.997	0.997	26	osteopenia	0.888	0.873	0.846
7	cardiomegaly	0.928	0.930	0.918	27	pleural effusion	0.957	0.959	0.949
8	central line	0.996	0.992	0.990	28	pleural thickening	0.816	0.811	0.811
9	consolidation	0.838	0.818	0.838	29	pneumothorax	0.929	0.935	0.910
10	costoph. angle blunt.	0.917	0.919	0.896	30	pulmonary edema	0.943	0.945	0.944
11	degenerative changes	0.835	0.834	0.812	31	rib fracture	0.883	0.862	0.855
12	elevated diaphragm	0.919	0.919	0.908	32	scoliosis	0.852	0.843	0.824
13	fibrotic changes	0.858	0.861	0.839	33	soft tissue calc.	0.709	0.715	0.703
14	fracture	0.718	0.735	0.616	34	sternotomy wires	0.989	0.986	0.988
15	granuloma	0.727	0.745	0.689	35	surgical clips noted	0.905	0.884	0.883
16	hernia diaphragm	0.986	0.988	0.984	36	thickening of fissure	0.892	0.875	0.870
17	hilar prominence	0.884	0.880	0.855	37	trachea deviation	0.943	0.918	0.908
18	hyperinflation	0.839	0.838	0.844	38	transplant	0.999	0.999	0.999
19	interstitial markings	0.776	0.779	0.781	39	tube	0.940	0.940	0.911
20	kyphosis	0.941	0.907	0.925	40	vertebral height loss	0.809	0.773	0.766

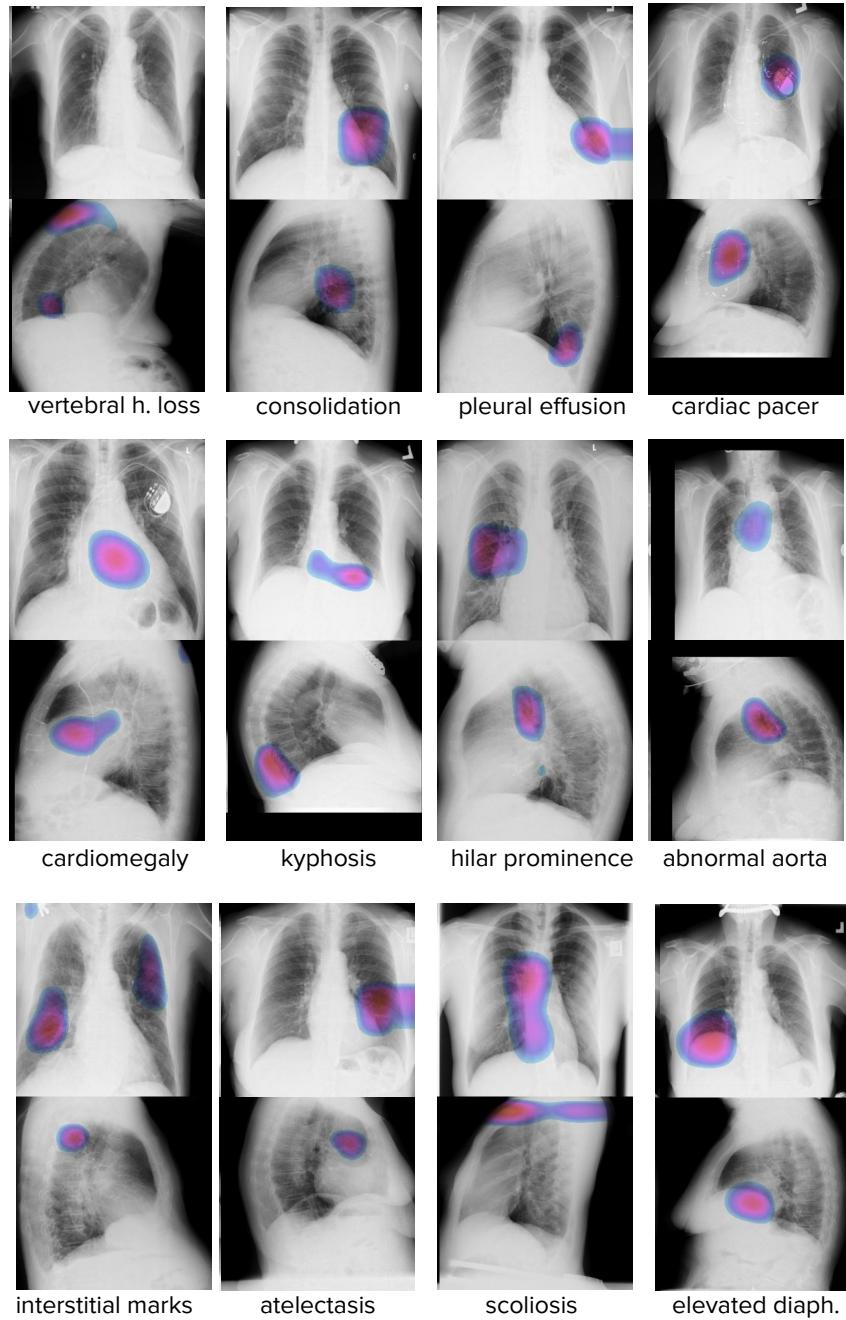


Fig. 5. Heat maps for 12 positive findings on a selected set of studies. For each finding, a heat map is generated as a linear combination over the 1024 feature maps calculated for each view. The weight given to feature-map i when generating a heat map for finding j is W_{ij} , where W are the weights of the last fully-connected layer of the network.