Supplementary Information

Benchmarking saliency methods for chest X-ray interpretation

Adriel Saporta MS MBA[1*], Xiaotong Gui MS[1*], Ashwin Agrawal MS[1*], Anuj Pareek MD PhD[2], Steven QH Truong MBA[3], Chanh DT Nguyen PhD[3,4], Van-Doan Ngo MD[5], Jayne Seekins DO[6], Francis G. Blankenberg MD[6], Andrew Y. Ng PhD[1], Matthew P. Lungren MD MPH[2†], Pranav Rajpurkar PhD[7†]

[1]Stanford University Department of Computer Science, USA
[2]Stanford University AIMI Center, USA
[3]VinBrain, Vietnam
[4]VinUniversity, Vietnam
[5]Vinmec International Hospital, Vietnam
[6]Stanford University Department of Radiology, USA
[7]Department of Biomedical Informatics, Harvard University, USA
[*]These authors contributed equally: Adriel Saporta, Alex Gui, Ashwin Agrawal
†These authors contributed equally: Matthew P. Lungren, Pranav Rajpurkar
Corresponding author: Pranav Rajpurkar (email: pranav_rajpurkar@hms.harvard.edu)

Supplementary information

We would like to localize the following **10 observations** on a set of chest X-rays (CXRs): Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Atelectasis, Pneumothorax, Pleural Effusion, and Support Devices.

You will be given the same set of CXRs on which you earlier drew segmentations for the above 10 observations. As a reminder, each CXR has 10 ground truth labels (0 or 1) for each of the above 10 observations, and you drew segmentations on each of the assigned CXRs only for the observations that were present as determined by the ground truth labels.

Now, for these same CXRs on which you drew the segmentations, we would like for you to select **the single most salient point** on the CXR for each positive observation. Which point is most salient will depend on the observation at hand, but the point will always lie inside the segmentation you drew. For example, for Pneumothorax, this point might be wherever the pathology is most pronounced. As another example, for Cardiomegaly, the most salient point will be the center of the heart. Below, we have included descriptions and examples of what we are expecting the most salient point to be for each of the 10 observations.

For some CXRs, there are multiple instances/segmentations for a given observation (for example, if there are multiple support devices, or if the patient has bilateral pleural effusion). Even when there are multiple segmentations, please only select **one single point** for that observation on the CXR; this point will necessarily lie inside *one* of the segmentations that you drew for that observation.

At the end of this exercise, there should be one point for each positive label for each CXR. For example, if a CXR has positive labels for Cardiomegaly, Lung Opacity, and Edema, then there should be three series of segmentations with three points for that CXR.

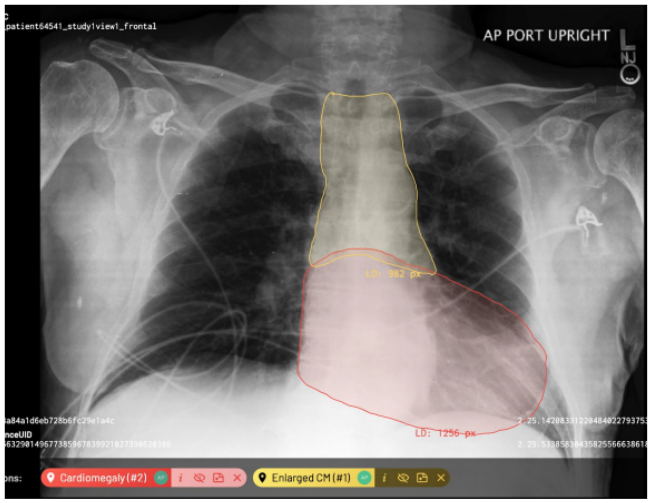
As before, we understand that you may not always agree with the ground truth labels. However, we ask that you please attempt to annotate all of the positive observations according to the ground truth labels. Also, please note that you do *not* need to select a point if the CXR has a positive label for No Finding (however, a CXR with a No Finding label may have a positive label for Support Devices, which does require a point!).

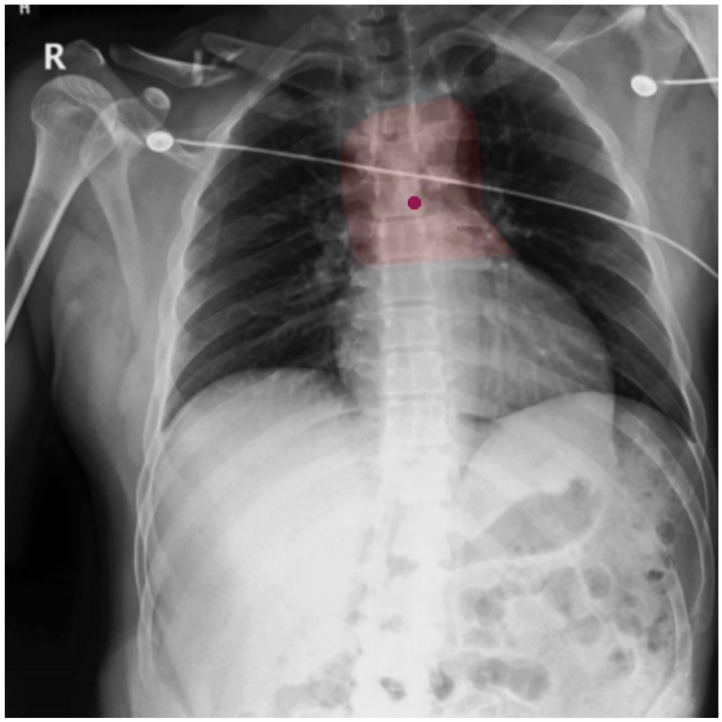
Figure S1. General instructions given to benchmark radiologists on how to select the most representative points on a CXR.

Enlarged Cardiomediastinum

Enlarged Cardiomediastinum just refers to Enlarged Mediastinum, because Cardiomegaly is a separate label.

Annotation of simultaneous Cardiomegaly and Enlarged Cardiomediastinum will look like this:

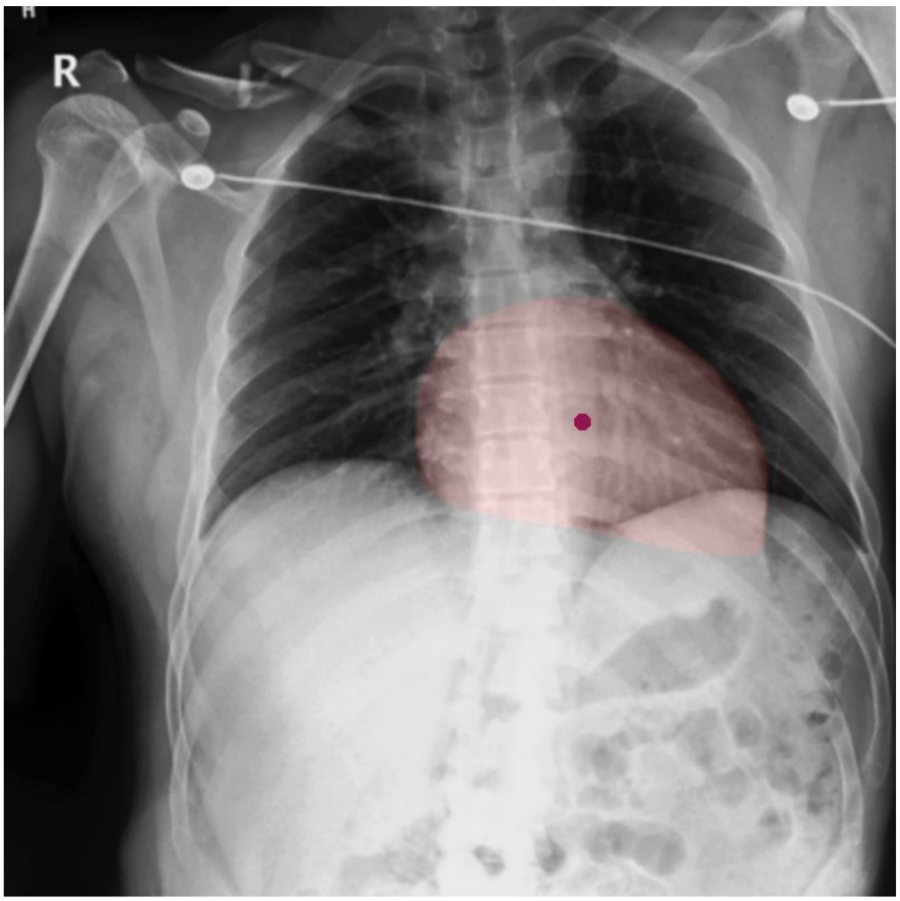




For Enlarged Cardiomediastinum, the most salient point will be in the center of the mediastinum.

Figure S2. Specific instructions given to benchmark radiologists on how to select the most representative point on a CXR for Enlarged Cardiomediastinum.

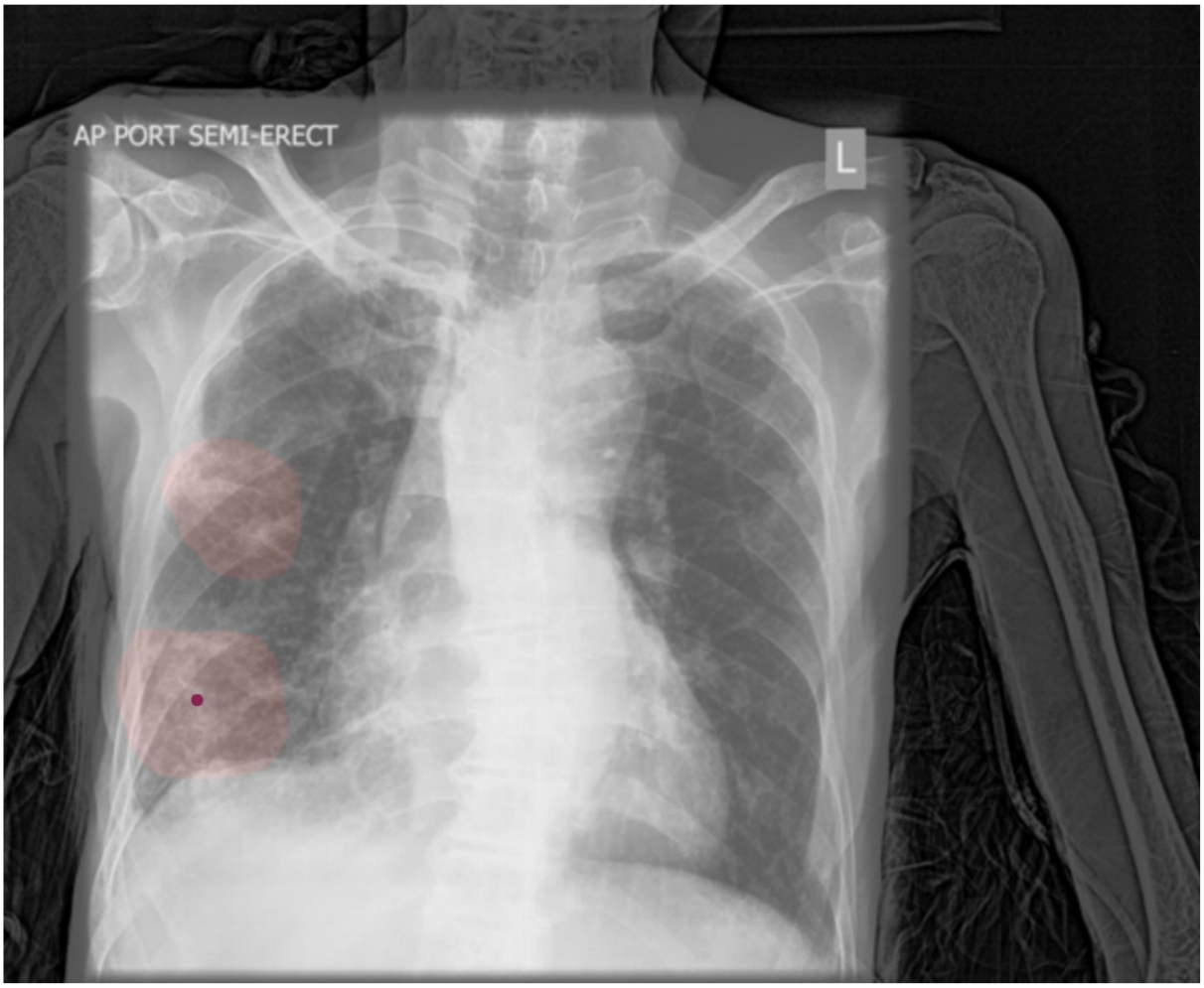
Cardiomegaly



For Cardiomegaly, the dot should always be in the center of the segmentation of the enlarged heart.

Figure S3. Specific instructions given to benchmark radiologists on how to select the most representative point on a CXR for Cardiomegaly.

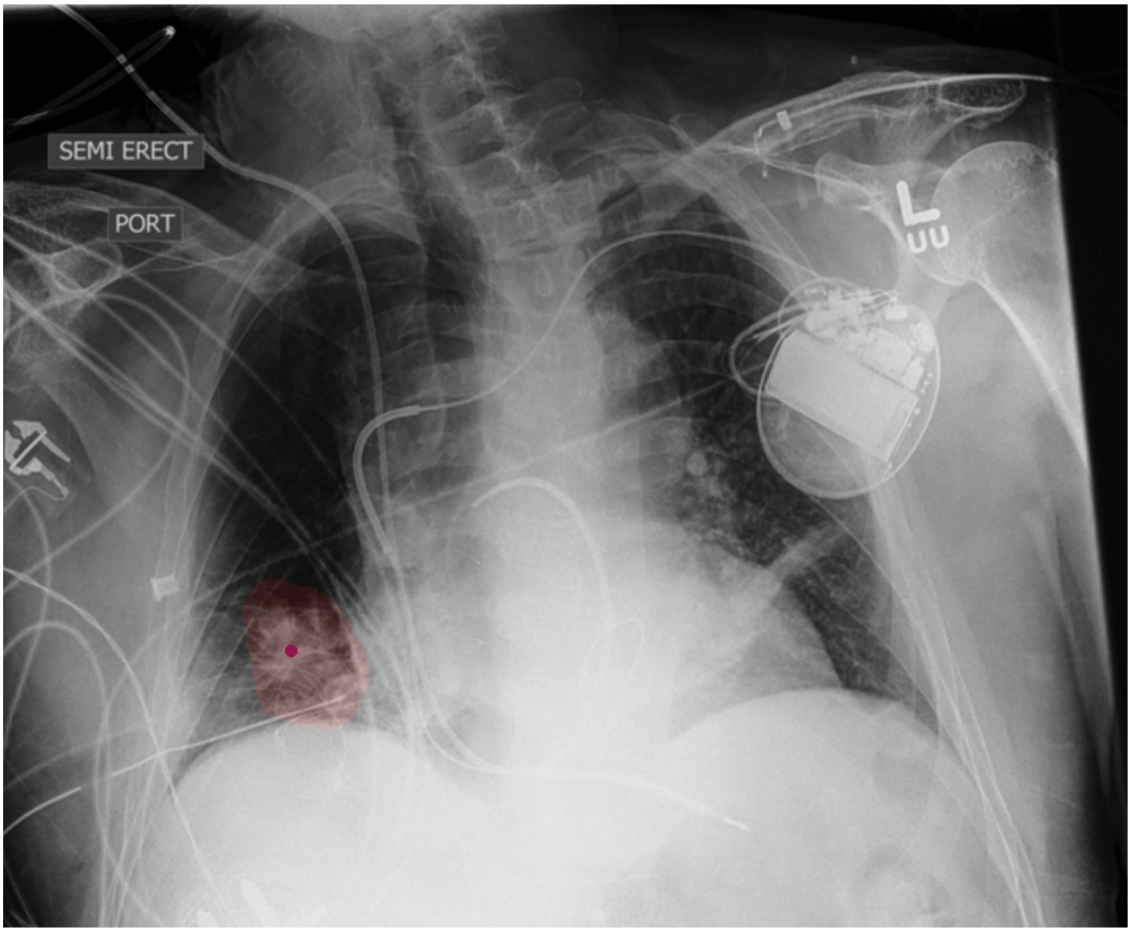
Lung Opacity



A Lung Opacity means a radiographic opacity that is not explained by the other types of opacities in our label set (Consolidation, Atelectasis and Lung Lesion). There may be more than one opacity. The most salient point should always lie in the center of the most prominent opacity.

Figure S4. Specific instructions given to benchmark radiologists on how to select the most representative point on a CXR for Lung Opacity.

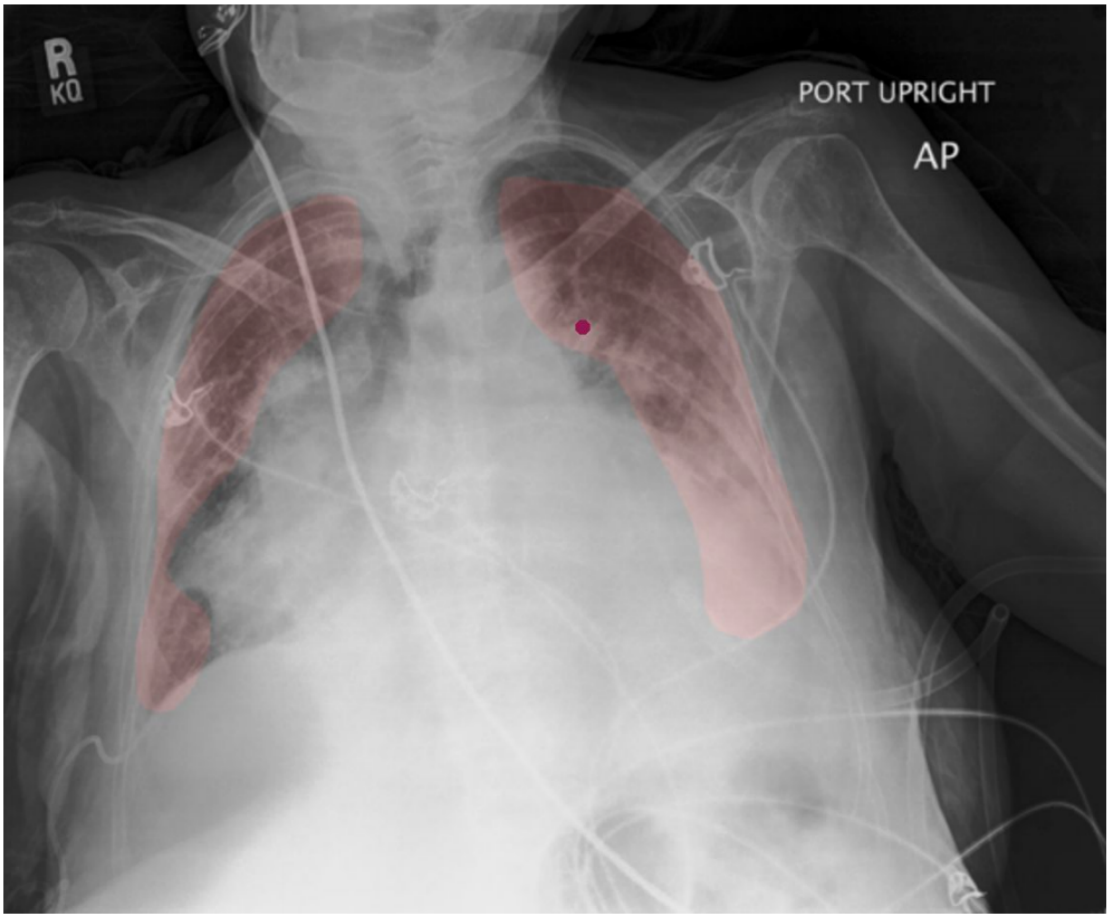
Lung Lesion



A Lung Lesion means a mass or nodule. There may be more than one lesion. The most salient point should lie inside the segmentation of the most pronounced lesion, and should be placed wherever that lesion is most pronounced.

Figure S5. Specific instructions given to benchmark radiologists on how to select the most representative point on a CXR for Lung Lesion.

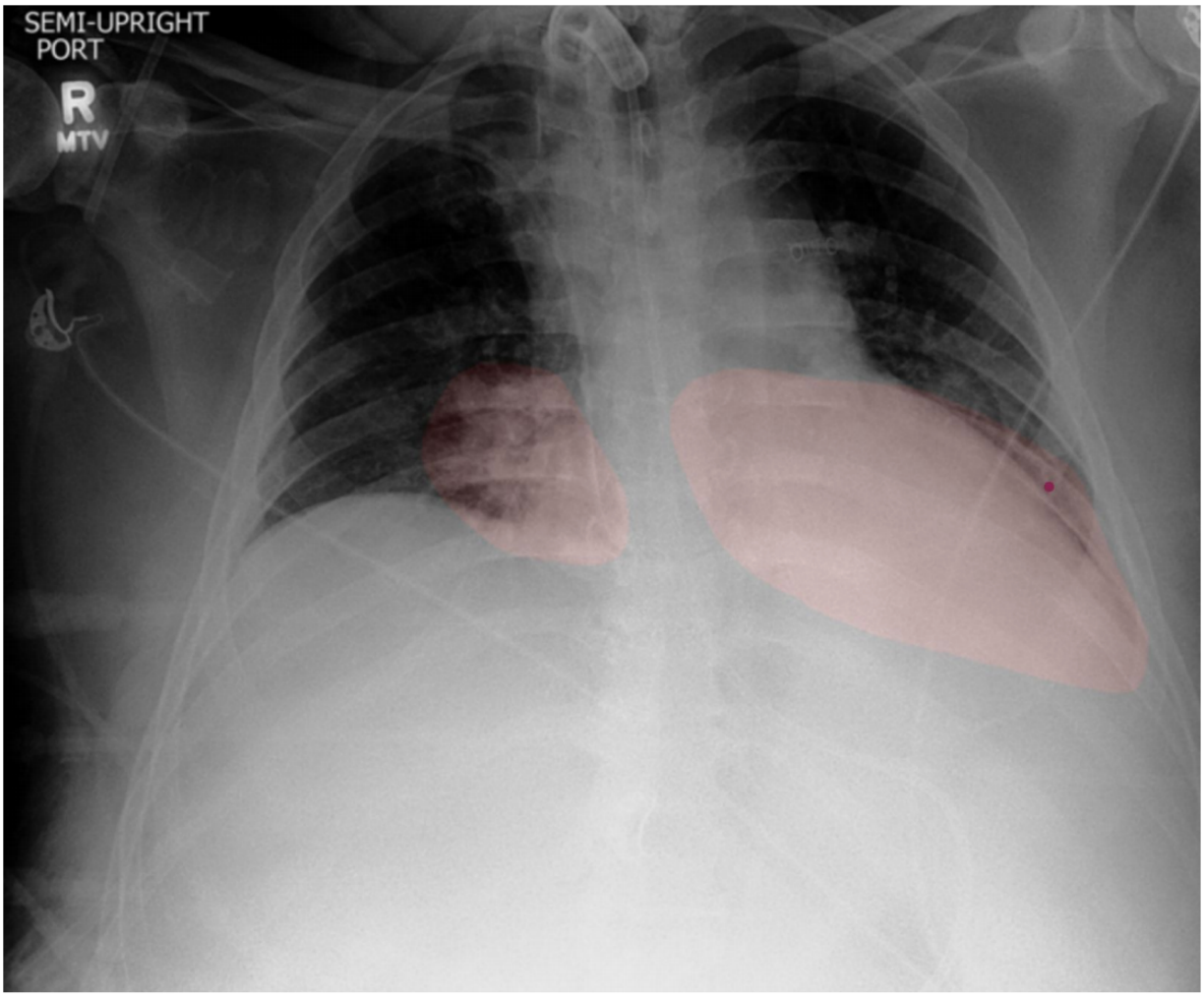
Edema



The most salient point should be placed wherever the visual features of edema are most pronounced. Please only select one side (right/left) where the visual features of edema are most pronounced.

Figure S6. Specific instructions given to benchmark radiologists on how to select the most representative point on a CXR for Edema.

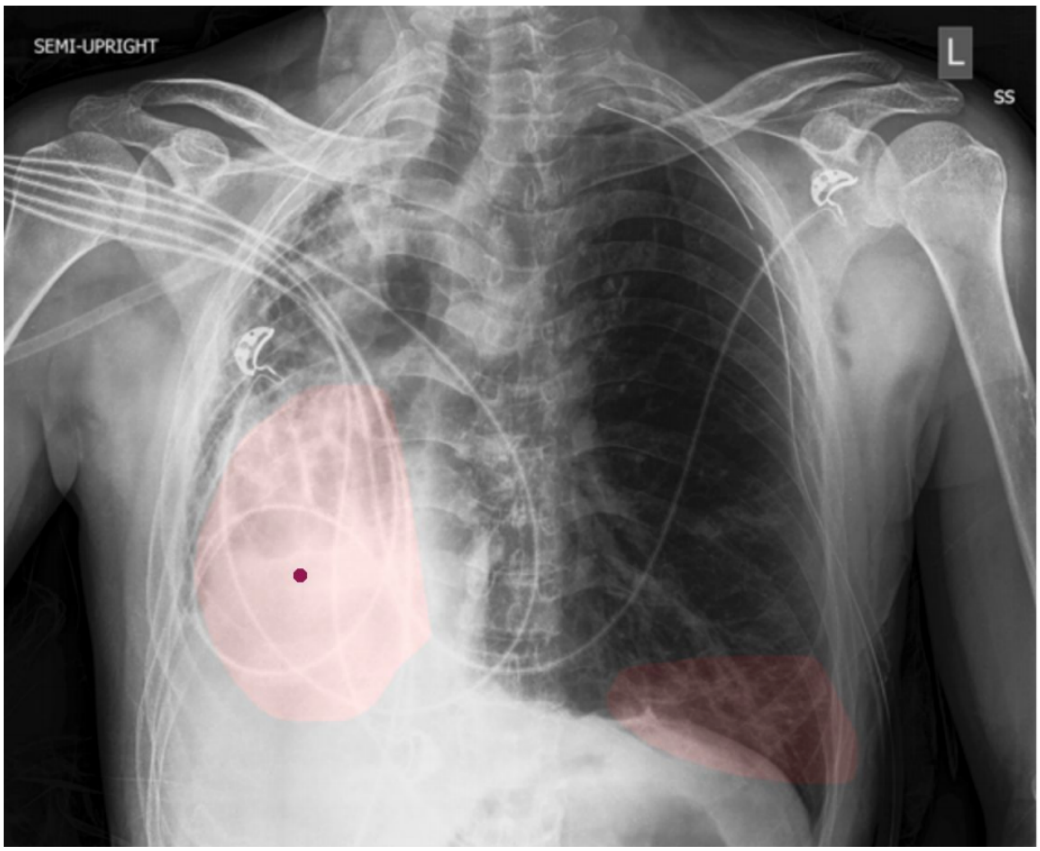
Consolidation



There may be more than one consolidation. The most salient point should lie inside the segmentation of the most pronounced consolidation, and should be placed wherever that consolidation is most pronounced. If the patient has diffuse consolidation throughout the entire lung field, aim for the area where the diffuse consolidation is more pronounced.

Figure S7. Specific instructions given to benchmark radiologists on how to select the most representative point on a CXR for Consolidation.

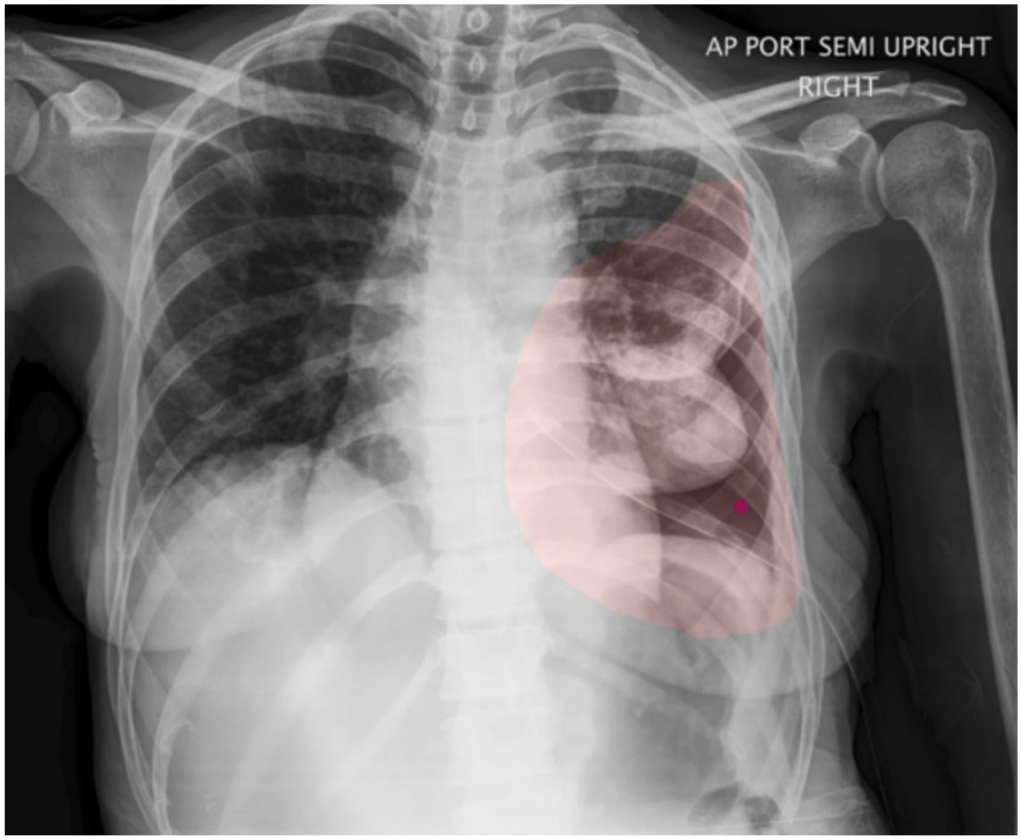
Atelectasis



There may be multiple instances of Atelectasis. The most salient point should lie wherever the Atelectasis is most pronounced. As always, please only select one point, although there may be more than one segmentation.

Figure S8. Specific instructions given to benchmark radiologists on how to select the most representative point on a CXR for Atelectasis.

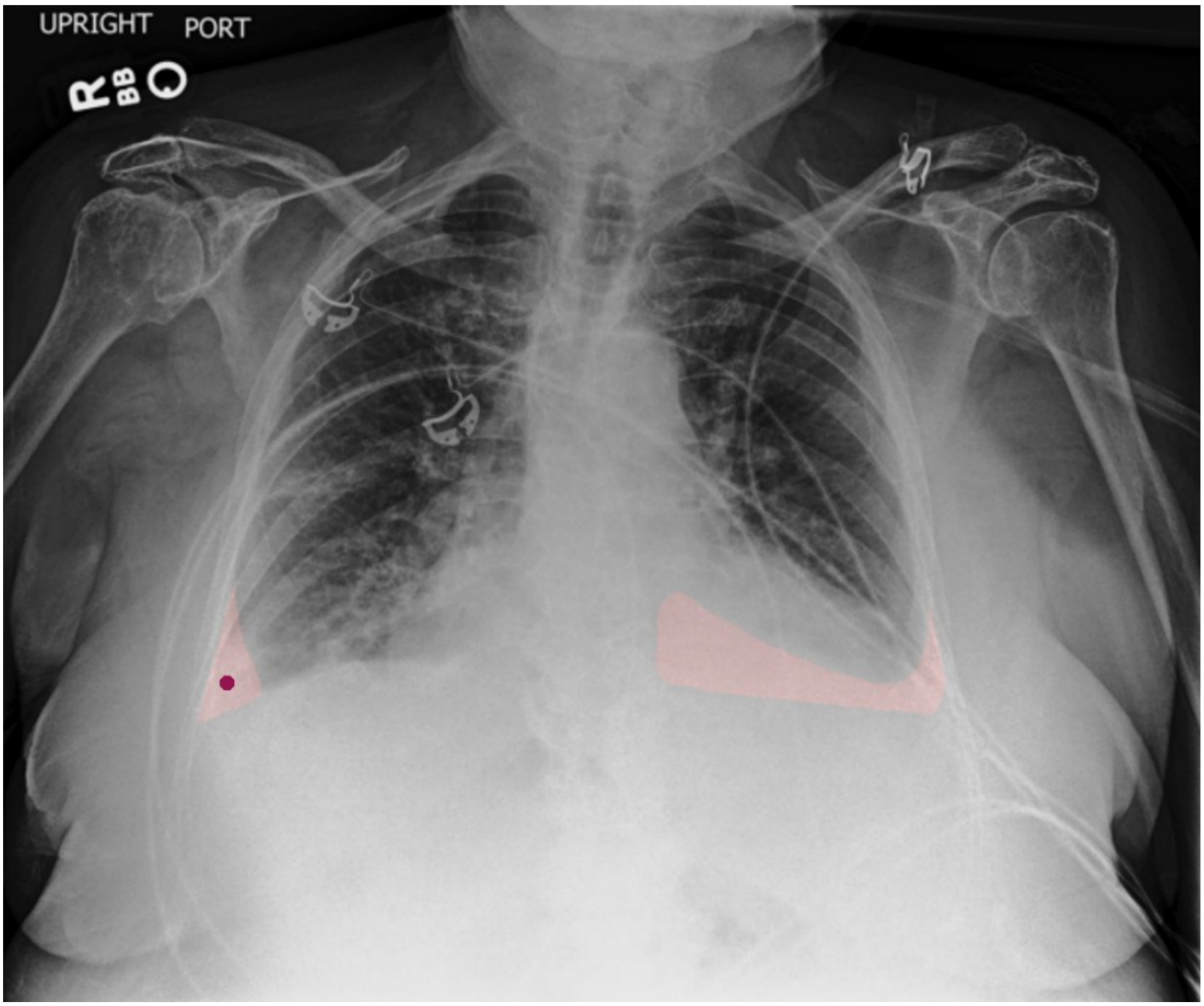
Pneumothorax



The single point should lie wherever the Pneumothorax is most pronounced.

Figure S9. Specific instructions given to benchmark radiologists on how to select the most representative point on a CXR for Pneumothorax.

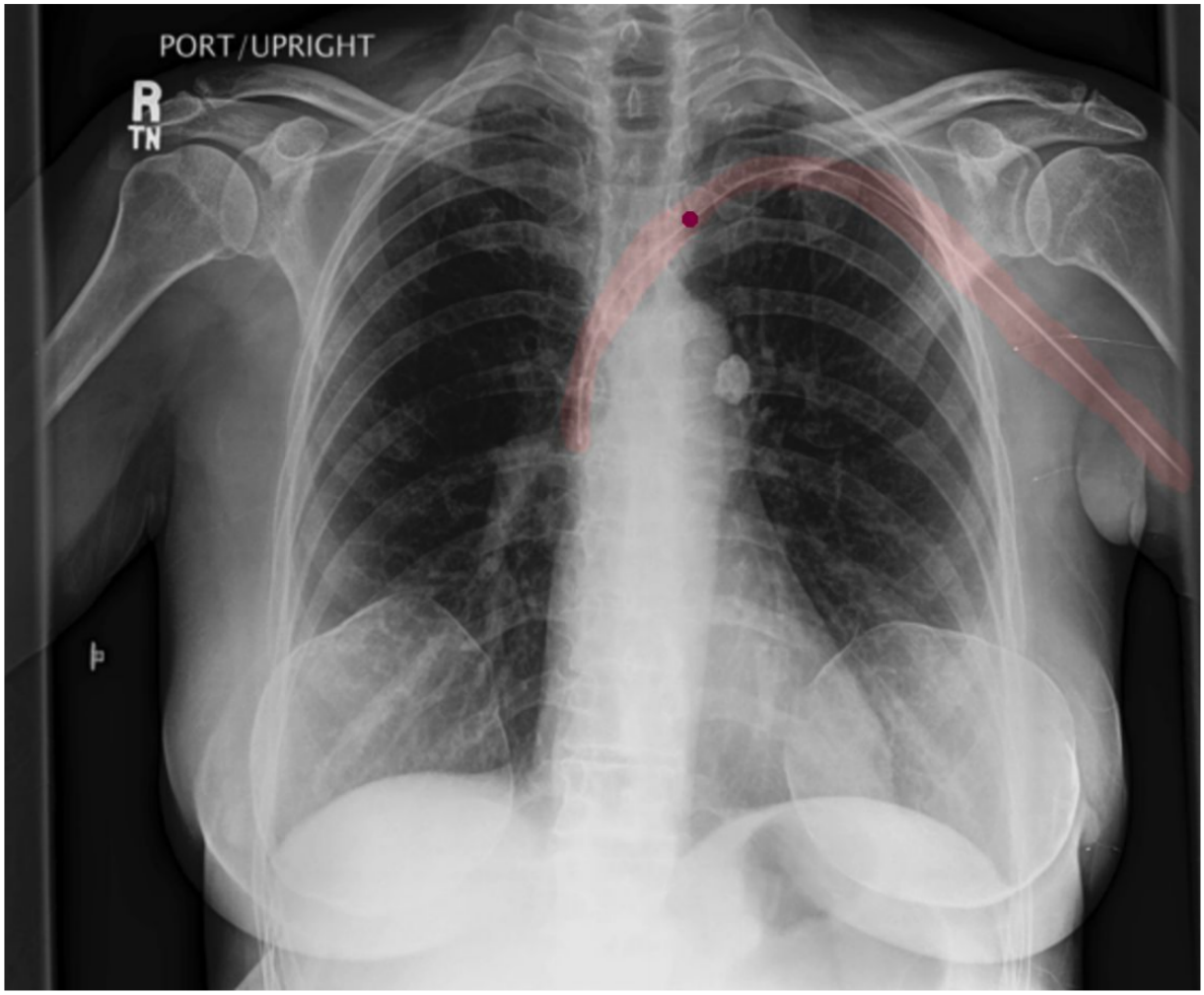
Pleural Effusion



The most salient point should lie in the center of the pleural effusion. If the patient has bilateral pleural effusion, then please place the point in the center of whichever pleural effusion is more pronounced.

Figure S10. Specific instructions given to benchmark radiologists on how to select the most representative point on a CXR for Pleural Effusion.

Support Devices



Support Devices are implanted or invasive devices such as pacemakers, PICC/central catheters, chest tubes, endotracheal tubes, feeding tubes and stents. ECG lead wires or stickers placed externally on the patient do not require labeling. If there is a single support device, please select the point at the estimated center of the support device. If there are several support devices, please select a point on the support device that you feel is most prominent. Please only consider the parts of the support device that are either inside (i.e. pacemaker, tube) or on (i.e. venous port) the patient. Do not place the point on a part of the support device that is completely outside the patient body (i.e. chest tube outside the thorax).

Figure S11. Specific instructions given to benchmark radiologists on how to select the most representative point on a CXR for Support Devices.

Table S1. Hit rate: Coefficients from regressions on model assurance.

Pathology	**CXRs (*n*)**	**Linear regression coefficient**	**Spearman correlation coefficient**
Airspace Opacity	381	0.498***	0.16**
Atelectasis	296	0.443**	0.126
Cardiomegaly	229	0.195*	0.185*
Consolidation	120	0.082	0.199
Edema	124	0.195	0.132
Enlarged Cardiomediastinum	668	0.548**	0.253***
Lung Lesion	50	0.54	0.453
Pleural Effusion	159	0.654***	0.278**
Pneumothorax	11	0.21	0.142
Support Devices	327	-0.058	-0.029
All pathologies	2365	-0.411***	-0.239***
* p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001			

Table S2. Dataset summary statistics.

Sample Size	
No. Studies	500
No. CXRs	668
Pathology	**CXRs (*n*)**
Airspace Opacity	309
Atelectasis	177
Cardiomegaly	175
Consolidation	35
Edema	83
Enlarged Cardiomediastinum	297
Lung Lesion	14
Pleural Effusion	120
Pneumothorax	10
Support Devices	314
No Pathology Identified	169

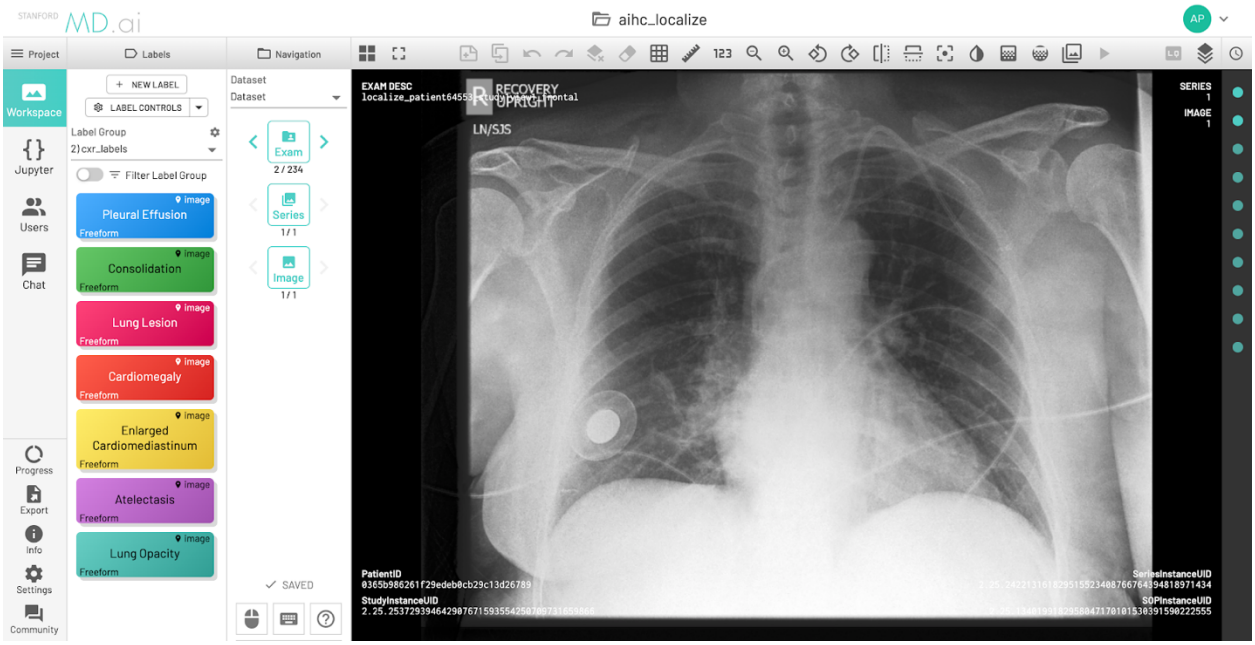


Figure S12. Screenshot of MD.ai project page where two radiologists drew reference segmentations.

Page from which radiologists could scroll through each CXR exam they were meant to segment. On the left side of the screen, a column of colored tabs, one for each of the 10 pathologies of interest.

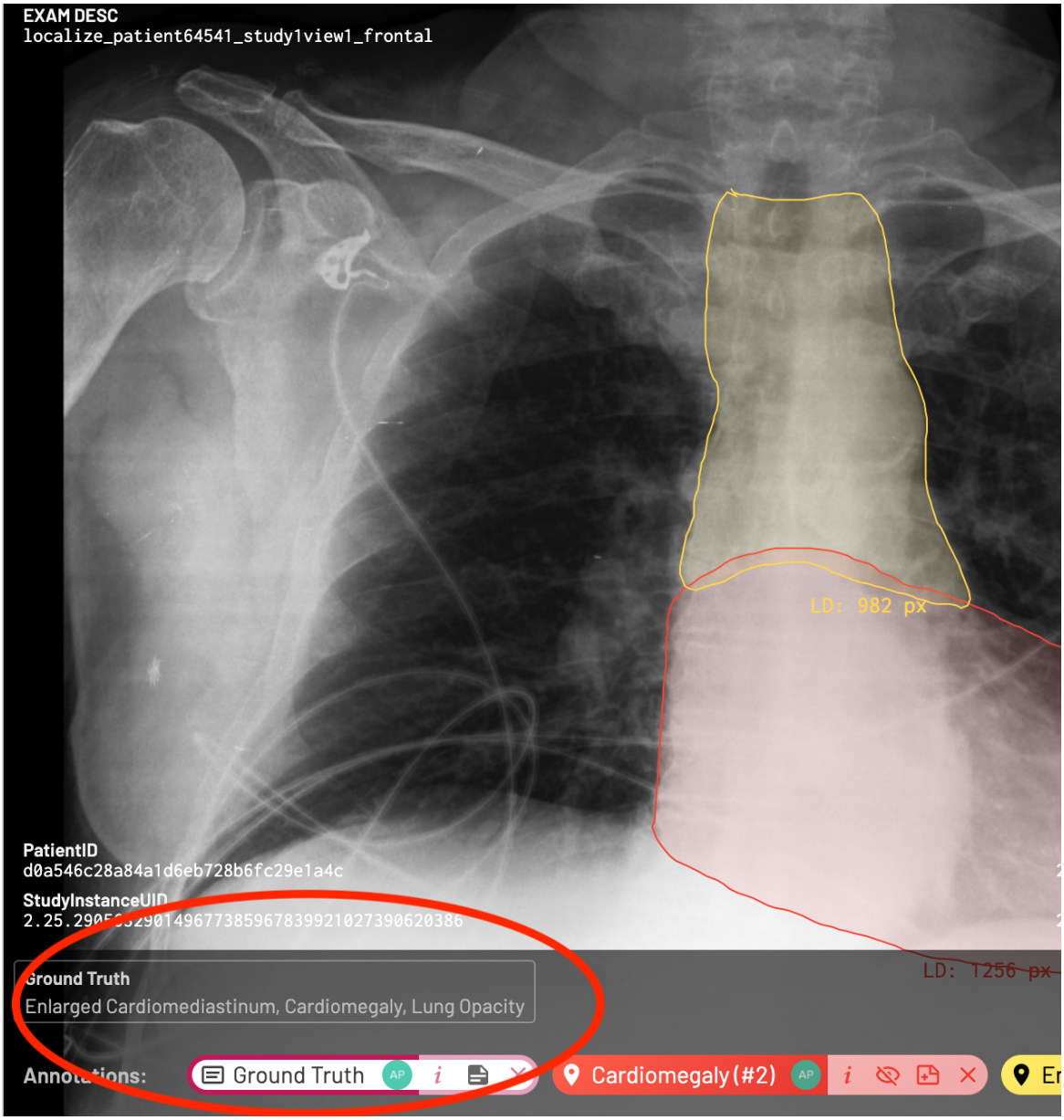


Figure S13. Screenshot of MD.ai ground truth labels.

The radiologists had access to the ground truth labels for each CXR, which were displayed at the bottom left of the CXR image.

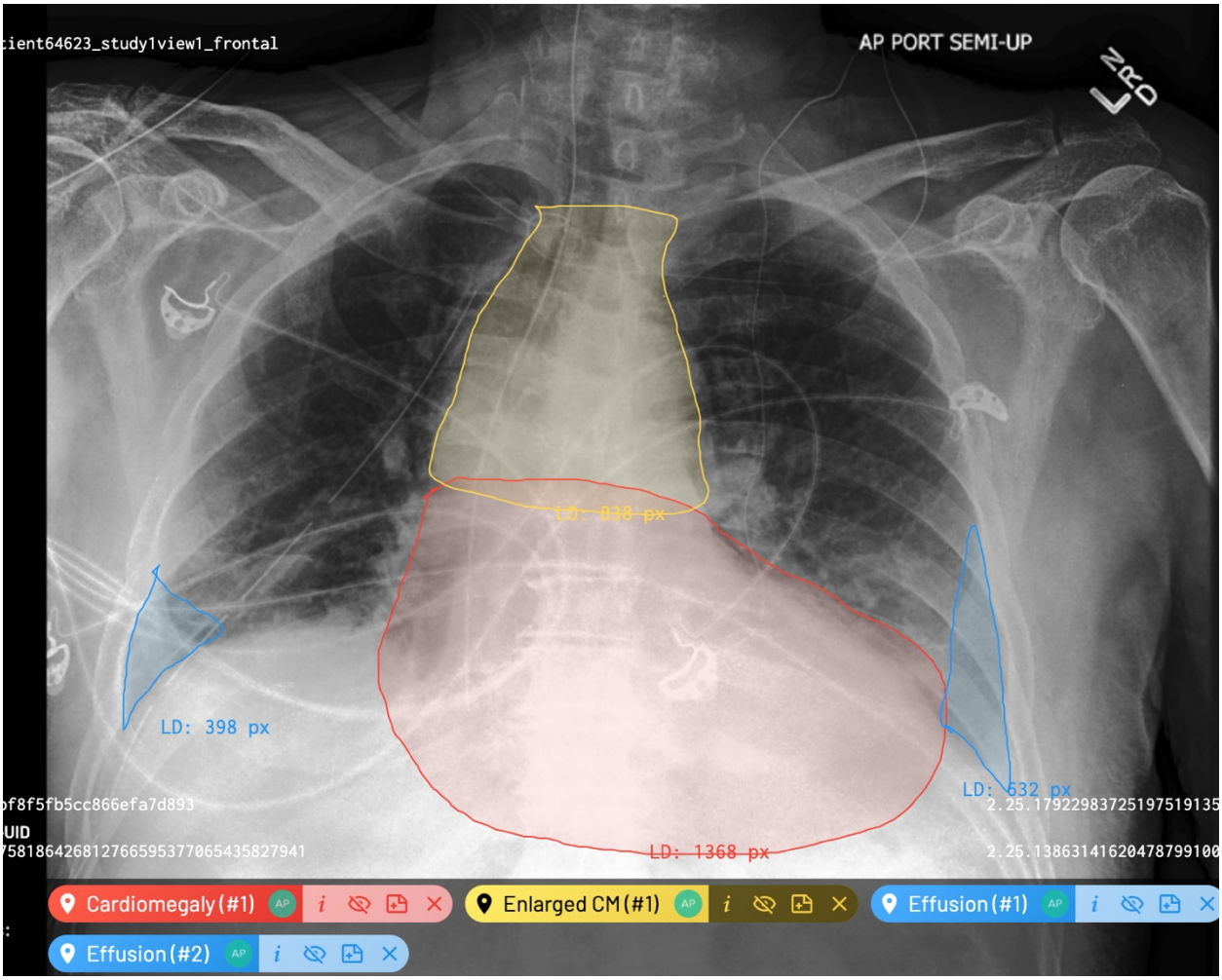


Figure S14. Screenshot of MD.ai segmentations for multiple pathologies on a single CXR.

Example CXR on MD.ai on which radiologists segmented Cardiomegaly, Enlarged Cardiomediastinum, and Pleural Effusion. After clicking on a pathology label from a sidebar on the left to activate the annotation mode, radiologists were able to draw directly on the CXR using free-hand contouring. After one pathology was segmented, the radiologist then segmented the next pathology in the same way. We asked the radiologists to strike a good balance between efficiency and accuracy.

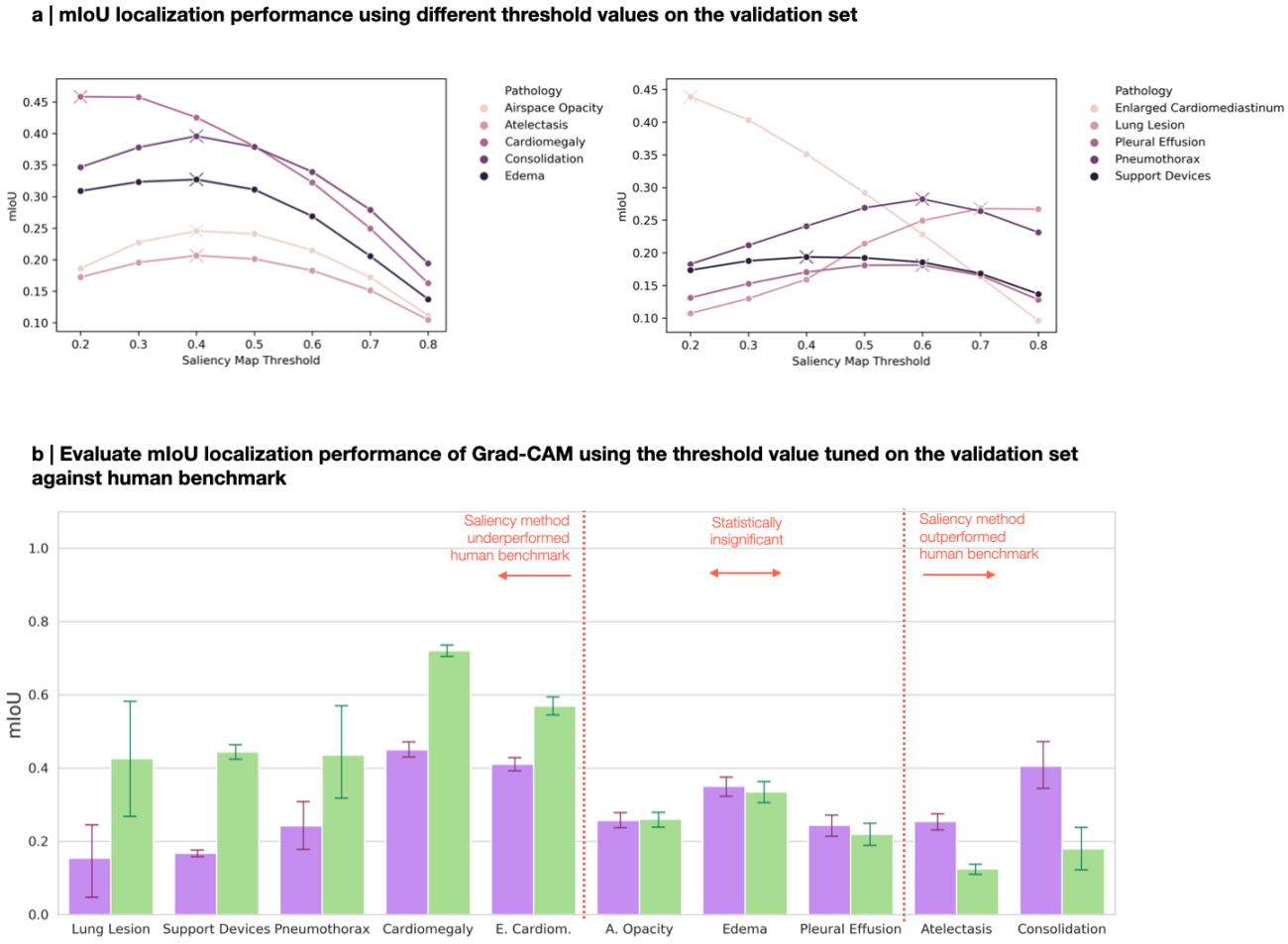


Figure S15. Sensitivity analysis of mIoU localization performance using different saliency map thresholding values.

We reported mIoU localization performance using different saliency map threshold values. a, We first applied max-min normalizations to the Grad-CAM saliency maps so that each value gets transformed into a decimal between 0 and 1. We then passed in a range of threshold values from 0.2 to 0.8 to create binary segmentations and plotted the mIoU score per pathology under each threshold on the validation set. The threshold that gives the max mIoU for each pathology is marked in "X". b, We compared the mIoU localization performance between Grad-CAM using the best thresholds tuned on the validation set and the human benchmark. The result shows the same conclusion as Section 2 in the main text (Figure 2, threshold using Otsu's method). This suggests that our result is robust to different values used for saliency map thresholding.

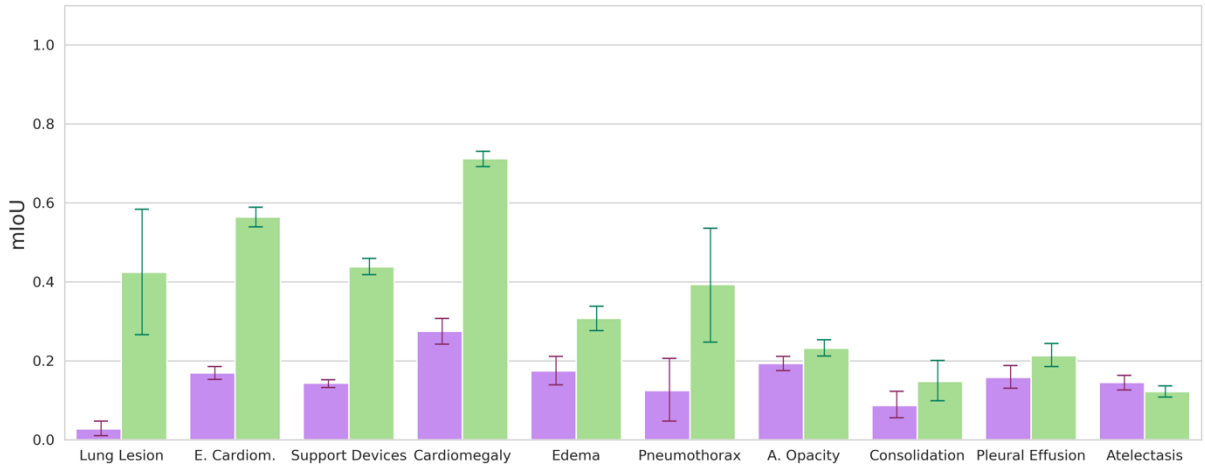


Figure S16. mIoU localization performance using the full dataset.

Only the true negatives (there was no human benchmark segmentations and no saliency map segmentation) were excluded from the metric calculation. To control for false positives, we further ensure that the final binary segmentation is consistent with model probability output by applying another layer of thresholding such that the segmentation mask produced all zeros if the predicted probability was below a chosen level. The probability threshold is searched on the interval of [0,0.8] with steps of 0.1. The exact value is determined per pathology by maximizing the mIoU on the validation set. We found that on the full dataset, for seven of the 10 pathologies, the saliency method pipeline had a significantly lower mIoU than the human benchmark.

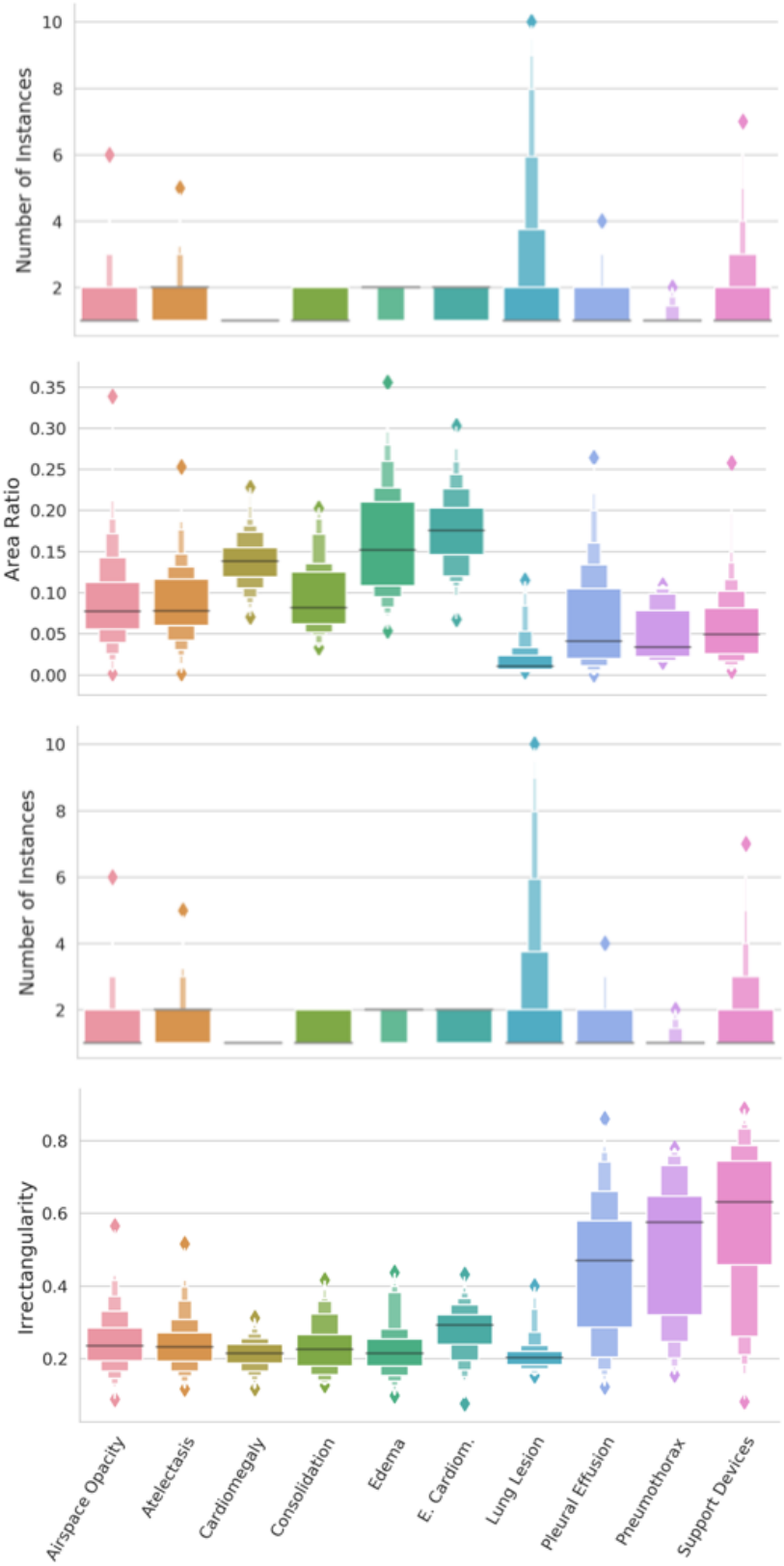

Number of Instances
10
8
6
4
2
Area Ratio
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Number of Instances
10
8
6
4
2
Irrectangularity
0.8
0.6
0.4
0.2
Airspace Opacity
Atelectasis
Cardiomegaly
Consolidation
Edema
E. Cardiom.
Lung Lesion
Pleural Effusion
Pneumothorax
Support Devices

Figure S17. Distribution of the four pathological characteristics across all 10 pathologies.

The black horizontal line in each box indicates the median feature value for that pathology, and each successive level outward contains half of the remaining data. The height of the box indicates the range of feature value in the quantile.

Table S3. Classification performance on test set.

Table S3 | Classification performance on test set			
Pathology	**DenseNet121**	**ResNet152**	**Inception-v4**
Airspace Opacity	0.926	0.923	0.912
Atelectasis	0.833	0.810	0.803
Cardiomegaly	0.885	0.879	0.863
Consolidation	0.868	0.862	0.876
Edema	0.915	0.901	0.897
Enlarged Cardiomediastinum	0.583	0.609	0.578
Lung Lesion	0.912	0.900	0.869
Pleural Effusion	0.965	0.960	0.955
Pneumothorax	0.993	0.990	0.983
Support Devices	0.969	0.968	0.954

Table S4. Percentage decrease from human benchmark mIoU to saliency method pipeline mIoU.

Table S4 | Percentage decrease from expert mIoU to AI mIoU for each pathology			
Pathology	**Expert mIoU**	**AI mIoU**	**% decrease**
Lung Lesion	0.426	0.101	76.2 (59.1, 87.5)
Support Devices	0.444	0.163	63.3 (60.8, 65.8)
Pneumothorax	0.435	0.213	51.0 (14.6, 69.5)
Cardiomegaly	0.72	0.452	37.2 (34.0, 40.4)
Enlarged Cardiomediastinum	0.569	0.379	33.4 (29.0, 37.4)
Airspace Opacity	0.26	0.248	4.6 (-5.8, 14.6)
Pleural Effusion	0.219	0.235	-6.8 (-25.6, 13.3)
Edema	0.335	0.362	-7.4 (-16.4, 2.6)
Atelectasis	0.124	0.254	-51.2 (-57.3, -43.9)
Consolidation	0.179	0.408	-56.1 (-69.4, -42.7)
Average	0.383	0.281	26.6 (18.1, 35.0)

Table S5. Percentage decrease from human benchmark hit rate to saliency method pipeline hit rate.

Table S5 | Percentage decrease from expert hit percentage to AI hit percentage for each pathology			
Pathology	**Expert hit %**	**AI hit %**	**% decrease**
Lung Lesion	0.85	0.29	65.9 (35.3, 91.7)
Support Devices	0.933	0.355	62.0 (56.2, 67.5)
Pneumothorax	1.0	0.392	60.8 (27.3, 92.3)
Atelectasis	0.87	0.501	42.4 (0.331, 0.51)
Pleural Effusion	0.718	0.507	29.4 (14.3, 42.5)
Enlarged Cardiomediastinum	0.957	0.818	14.5 (9.6, 19.2)
Airspace Opacity	0.559	0.498	10.9 (-2, 23.1)
Cardiomegaly	0.972	0.903	7.1 (2.1, 11.8)
Edema	0.769	0.746	3.0 (-11.7, 18.5)
Consolidation	0.51	0.738	-44.7 (-56.5, 0.5)
Average	0.82	0.58	29.4 (15.0, 43.2)