Supplemental Online Content

Homayounieh F, Digumarthy S, Ebrahimian S, et al. An artificial intelligence–based chest x-ray model on human nodule detection accuracy from a multicenter study. *JAMA Netw Open.* 2021;4(12):e2141096. doi:10.1001/jamanetworkopen.2021.41096

This supplemental material has been provided by the authors to give readers additional information about their work.

eAppendix. AI Algorithm and Statistical Information

AI algorithm

The algorithm does not have the United States Food and Drug Administration (FDA) approval at the time of conducting this research study. The algorithm can detect pulmonary lesions (including subtypes of mass, nodule, and granuloma), pleural effusion, pneumothorax, atelectasis, and consolidation [1-4]. The algorithm localizes each detected finding with a bounding box and assigns a confidence score on a scale of 1 to 10 where 1 represents the lowest and 10 represents the highest probability for the presence of detected abnormality.

The detection system was trained in a multi-class setting, that is with a single parametric learning model that supported the joint classification and detection of abnormalities (such as pulmonary nodules, consolidation, atelectasis, pleural effusion, and pneumothorax). The architecture of this model has a multi-scale structure and is inspired from the network architecture described in a prior publication [1]. During training, there was an active and implicit transfer of class conditional information in the learning model. For inference, we selected the best performing model for lesion. Furthermore, we used backbone pretraining on ImageNet followed by pretraining on Chest-Xray14 [5]. A total of 7,776 cases (5,086 positive CXRs with lesions and 2,690 CXRs without lesions) were used for training. Data were acquired from multiple sites in the United States, Europe and Asia, acquired on the radiographic equipment from over 10 major vendors.

For the preprocessing, data augmentation and pretraining steps of the AI algorithm, we rescaled CXRs to an isotropic resolution of 1025x1025 pixels using letterboxing. Bilinear interpolation

was used for resampling, followed by a robust brightness/contrast normalization based on a linear remapping of the pixel values. Validation of trained model was performed on a set of 444 CXRs from two sites, with 138 cases positive for lesion and controls sampled randomly from sequential data to represent the population (average age of 60±18, 233 females and 211 males). This validation dataset comprises 186, 214 and 44 cases from FUJIFILM Corporation, Carestream Health and Agfa manufacturer, respectively.

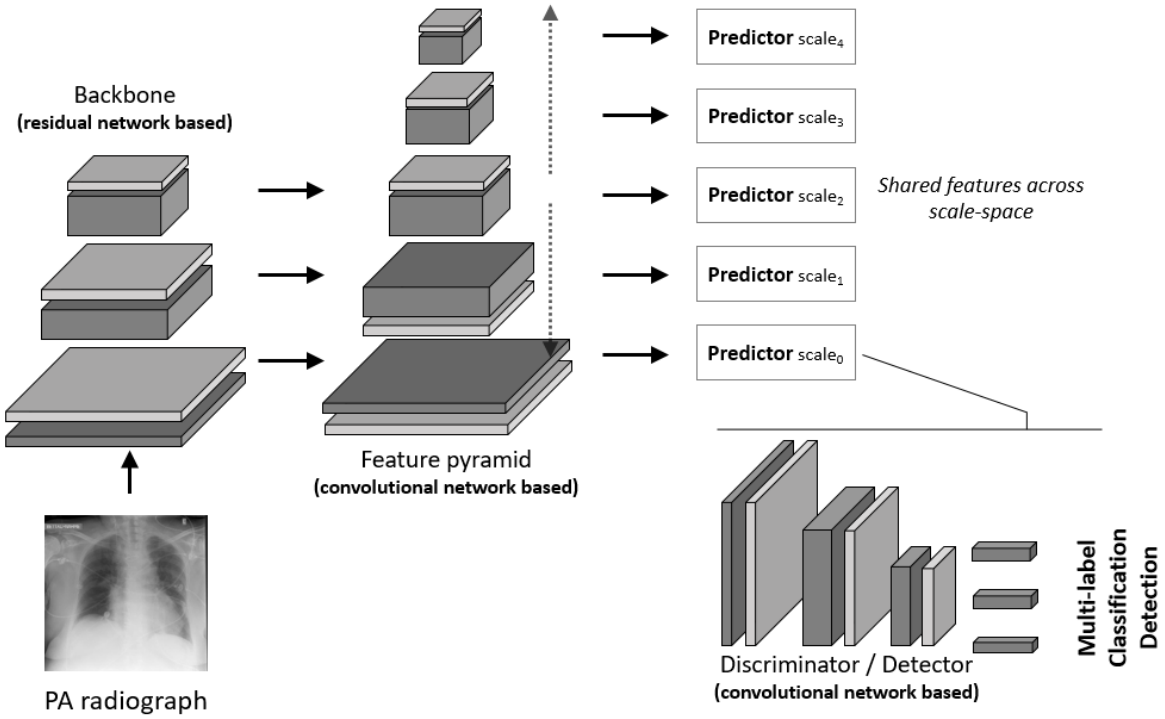
Additionally, an independent testing was performed on another validation dataset of 647 patients from three sources, with 158 positives and 489 negative cases.

The architecture of the AI algorithm is depicted in **Error! Reference source not found.** and comprised of an early feature extractor acting as the candidate generator in an abstract feature space, followed by a discriminator sub-network used to compute probabilities on whether the abnormality is present or not (in an image sub-region of interest). The architecture was fully convolutional and processed the entire image content in one single shot, while analyzing its content on multiple levels of scales. As such, the architecture is capable of implicitly capturing both global as well as local comorbidities present in the image.

For the classification and detection of abnormalities, the input is an image rescaled to an isotropic resolution of 1025x1025 pixels using letterboxing. The output is a set of boxes indicating the location of the aforementioned abnormalities, each associated with a label and a probability.

Training was conducted in one end-to-end stage and in a multi-class setting. The loss function is based on summation of three elements: 1) a classification loss based on the focal loss described in detail in [6]; 2) a bounding box coordinate regression loss based on an intersection-over-union

based metric; and 3) a center-ness loss designed to reduce outlier detections which is based on a weighted binary cross entropy loss. A batch-size of 8 is used for training and validation sets are used to track the system performance during training and perform early stopping.



Supplemental Figure 1: Architecture used for classification and detection

Statistical considerations

To decrease a potential bias in detection of only one finding (nodules) and simulate routine CXR interpretation (where other findings can be present besides nodules), each reader recorded the presence of four radiographic findings (not analyzed due to small sample of these findings) in addition to pulmonary nodules. The full dynamic range of the confidence level (probability of the presence of a finding) was used for a better estimation of the reader performance in terms of ROC AUC. The AUC is calculated on a per-case level. The analyzed

metrics for assessing performance change between the two reading modes included sensitivity, specificity, and accuracy. To investigate how AI-aided interpretation affects the human detection accuracy at a case level, the performance of 9 readers in two reading modes together with the ROC curve of AI was plotted (Figure 4 in supplemental material). AUC is a scale-invariant measure of performance which describes how well predictions are ranked rather than their absolute values. It is also classification-threshold-invariant that measures the quality of the model's predictions regardless of the chosen threshold.

Initial data analysis demonstrated a clustering of data in high specificity regions within a specificity range of 0.8 to 1 (corresponding to the false-positive-rate interval of 0 to 0.2), associated with different level of sensitivity. Since most data were present in the high specificity (0.8-1 range) and missing data outside this interval, we assessed the sensitivity, specificity, and the AUC (partial AUC [7]) associated within this effective interval (false positive rate from 0 to 0.2) to obtain reliable measurements of reader performance, with areas outside the effective interval being mainly driven by extrapolation artifacts [7]. ROC curves represent sensitivities and specificities at all possible thresholds of confidence scores to define a positive test result. Fitting a ROC curve across all specificities values requires extrapolation beyond the study data with the result that the area underneath the extrapolated curve dominates ROC AUC rather than being driven only by observed data, a task accomplished with partial AUCs in the effective interval [7]

As suggested [7, 8], we estimated 95% confidence interval for the output variables for interpreting results because in addition to evaluating any chance occurrence, it reflects the degree of variability as well as the sample size. The statistical significance of a result can be determined either from the p-value (whether it is less than 5% or not) or the confidence

interval based on whether the null value lies within the 95% confidence interval [7]. To estimate the confidence interval (CI) of the mean, bootstrap resampling was applied repeated for 100 times to the 100 cases selected in this study, which resulted in 100 observations of sensitivity, specificity, accuracy and partial AUC for each reader and corresponding 100 observations of the mean of above measurements for the nine radiologists. The total AUC of all readers was estimated as averaged ROC curve which is calculated statistically (it is not a numerical averaging). The readers true positive rates (TPR) were averaged for corresponding false positive rates (FPR) to obtain the final average ROC curve. To assess the effect of AI-aided interpretation beyond the two radiologists with disproportionately higher improvement, we performed a separate statistical analysis for the remaining 7 radiologists and applied Mann-Whitney U rank test on two independent samples to estimate statistical difference ($p<0.05$ was set as statistically significant difference). We did not perform any analysis into variability in diagnostic accuracy from a pre-specified level to the exploratory study.

From a statistical point of view, our study highlights the challenge associated with evaluation of change in reader confidence with and without AI-aided interpretation. Although we used a narrow 10-point scale of confidence as opposed to a wider -scale (such as 100-point), the readers did not use the entire scale to record their confidence. This might have resulted from the fact that radiologists do not use such scalar approach in their routine clinical practice while interpreting CXRs or other radiologic exam. In our experience, most radiologists describe imaging findings as present or absent (rather than definitely or certainly present or absent) or as probable, possible or suspicious (used interchangeably rather than in a structured categories of scaled confidence). In such circumstance, our study highlights use of partial

AUCs within the effective interval to assess impact of AI algorithms on reader confidence in either ruling in or ruling out radiologic findings.

References:

1. Tian Z, Shen C, Chen H, He T. Fcos: Fully convolutional one-stage object detection. InProceedings of the IEEE/CVF international conference on computer vision 2019 (pp. 9627-9636).
2. Gündel S, Setio AAA, Ghesu FC, et al. Robust classification from noisy labels: Integrating additional knowledge for chest radiography abnormality assessment. Med Image Anal. 2021;72:102087. doi:10.1016/j.media.2021.102087
3. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal Loss for Dense Object Detection. IEEE Trans Pattern Anal Mach Intell. 2020;42(2):318-327. doi:10.1109/TPAMI.2018.2858826
4. Rueckel J, Huemmer C, Fieselmann A, et al. Pneumothorax detection in chest radiographs: optimizing artificial intelligence system for accuracy and confounding bias reduction using in-image annotations in algorithm training. Eur Radiol. 2021;31(10):7888-7900. doi:10.1007/s00330-021-07833-w
5. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., 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), pp. 3462–3471.
6. Lin, Tsung-Yi, et al. "Focal loss for dense object detection." Proceedings of the IEEE international conference on computer vision. 2017.

7. Mallett S, Halligan S, Collins GS, Altman DG. Exploration of analysis methods for diagnostic imaging tests: problems with ROC AUC and confidence scores in CT colonography. PLoS One. 2014;9(10):e107633. Published 2014 Oct 29. doi:10.1371/journal.pone.0107633
8. Adedokun BO. P - value and confidence intervals - facts and farces. Ann Ib Postgrad Med. 2008;6(1):33-34. doi:10.4314/aipm.v6i1.64041

eTable 1. Vendor distribution of different radiographic equipment in our study.

Vendor	**Number of CXRs**
Carestream Health	27
Fujifilm Corporation	23
GE Medical System	22
Varian Medical Systems	13
Fuji Photo film	6
The Eastman Kodak Company	4
Siemens Healthineers AG	2
Philips Medical Systems	2
DeJarnette Research Systems	1

eTable 2. Partial AUC values under effective interval partial AUCs with in FPR interval 0-0.2 for unaided and AI-aided interpretation modes (AUC: area under the curve; FPR: false positive rate)

Readers	**Unaided mode** (Mean ± standard error)	**AI-aided mode** (Mean ± standard error)	**Delta** (Mean ± standard error)
J1	0.794 (0.038)	0.805 (0.037)	0.012 (0.040)
S1	0.750 (0.036)	0.758 (0.036)	0.008 (0.079)
J2	0.715 (0.043)	0.8485 (0.034)	0.132 (0.060)
S2	0.673 (0.050)	0.784 (0.037)	0.111 (0.051)
J4	0.557 (0.047)	0.767 (0.038)	0.214 (0.075)
S3	0.714 (0.044)	0.721 (0.046)	0.007 (0.054)
J5	0.747 (0.045)	0.751 (0.042)	0.004 (0.057)
J6	0.735 (0.038)	0.740 (0.039)	0.006 (0.039)
S4	0.766 (0.039)	0.778 (0.044)	0.012 (0.057)
Mean (95% CI)	0.717 (0.640,0.798)	0.772 (0.702,0.836)	0.052 (-0.014,0.120)

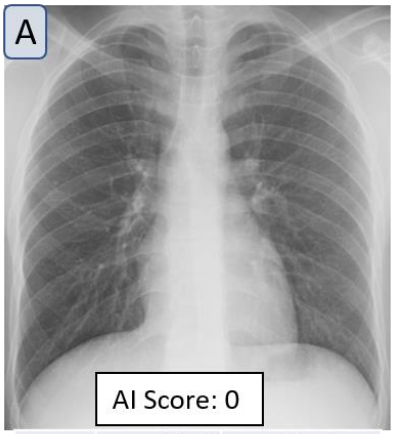
eTable 3. Summary of stratified readers performance for detection of easy and challenging pulmonary nodules versus control cases without nodules. While the specificity remained constant, the sensitivity of detection improved for both nodule types with AI-aided CXR interpretation as compared to unaided interpretation.

Easy-to-detect nodules vs. Control				
	Sensitivity		Specificity	
Readers	Unaided	Aided	Unaided	Aided
J1	0.61	0.78	0.98	0.96
S1	0.74	0.70	0.86	0.92
J2	0.43	0.78	0.96	0.96
S2	0.65	0.78	0.88	0.92
J4	0.74	0.78	0.82	0.96
S3	0.35	0.43	0.94	0.96
J5	0.57	0.70	0.94	0.94
J6	0.43	0.43	0.98	0.98
S4	0.48	0.83	0.98	0.94
Avg.	0.55	0.69	0.93	0.95

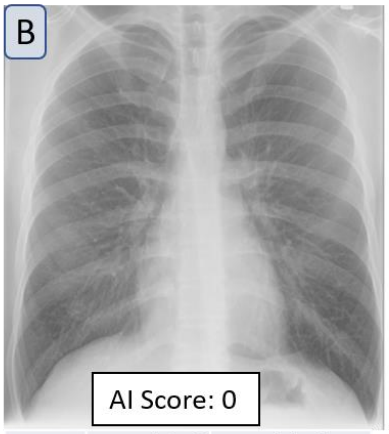
Challenging nodules vs. Control				
	Sensitivity		Specificity	
Readers	Unaided	Aided	Unaided	Aided
J1	0.30	0.40	0.98	0.96
S1	0.70	0.45	0.86	0.92
J2	0.15	0.55	0.96	0.96
S2	0.50	0.50	0.88	0.92
J4	0.40	0.45	0.82	0.96
S3	0.25	0.20	0.94	0.96
J5	0.30	0.45	0.94	0.94
J6	0.20	0.20	0.98	0.98
S4	0.15	0.55	0.98	0.94
Avg.	0.33	0.42	0.93	0.95

eTable 4. Sensitivity, specificity, accuracy and partial AUC within different reader groups in unaided and AI-aided interpretation modes (CI: confidence interval)

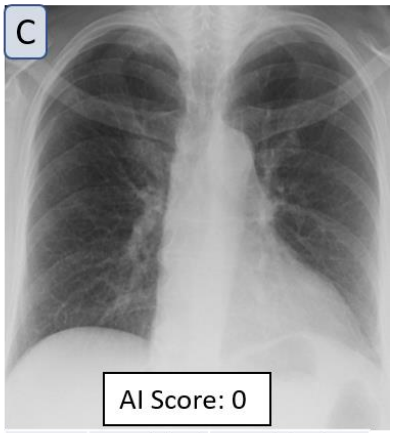
		Junior Radiologists (Mean with 95% CI)	**Senior Radiologists (Mean with 95% CI)**
Sensitivity	Unaided	0.400 (0.295,0.505)	0.510 (0.423,0.602)
	AI-aided	0.516 (0.439,0.624)	0.600 (0.5,0.689)
	Delta	*0.116 (0.04,0.189)*	*0.090 (0.005,0.166)*
Specificity	Unaided	0.948 (0.911,0.99)	0.900 (0.869,0.939)
	AI-aided	0.961 (0.921,0.994)	0.940 (0.88,0.985)
	Delta	*0.040 (-0.026,0.045)*	*0.040 (-0.02,0.085)*
Accuracy	Unaided	0.674 (0.606,0.772)	0.705 (0.641,0.77)
	AI-aided	0.770 (0.68,0.826)	0.770 (0.704,0.829)
	Delta	*0.064 (0.018,0.106)*	*0.065 (0.01,0.114)*
Partial AUC	Unaided	0.726 (0.655,0.807)	0.705 (0.63,0.785)
	AI-aided	0.780 (0.716,0.853)	0.763 (0.681,0.828)
	Delta	*0.064 (-0.002,0.107)*	*0.058 (-0.038,0.128)*



	Unaided	Aided
J6	5	3
S3	6	0
J5	5	0
S1	10	0
J2	4	0
J4	9	0

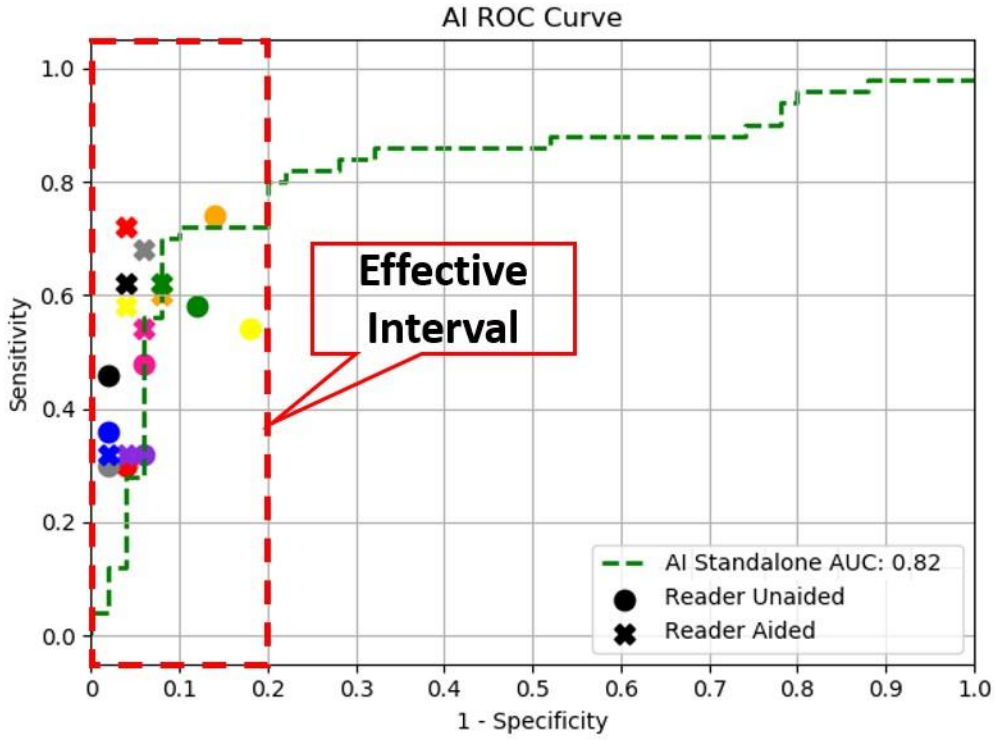


	Unaided	Aided
S1	7	0
S2	5	0



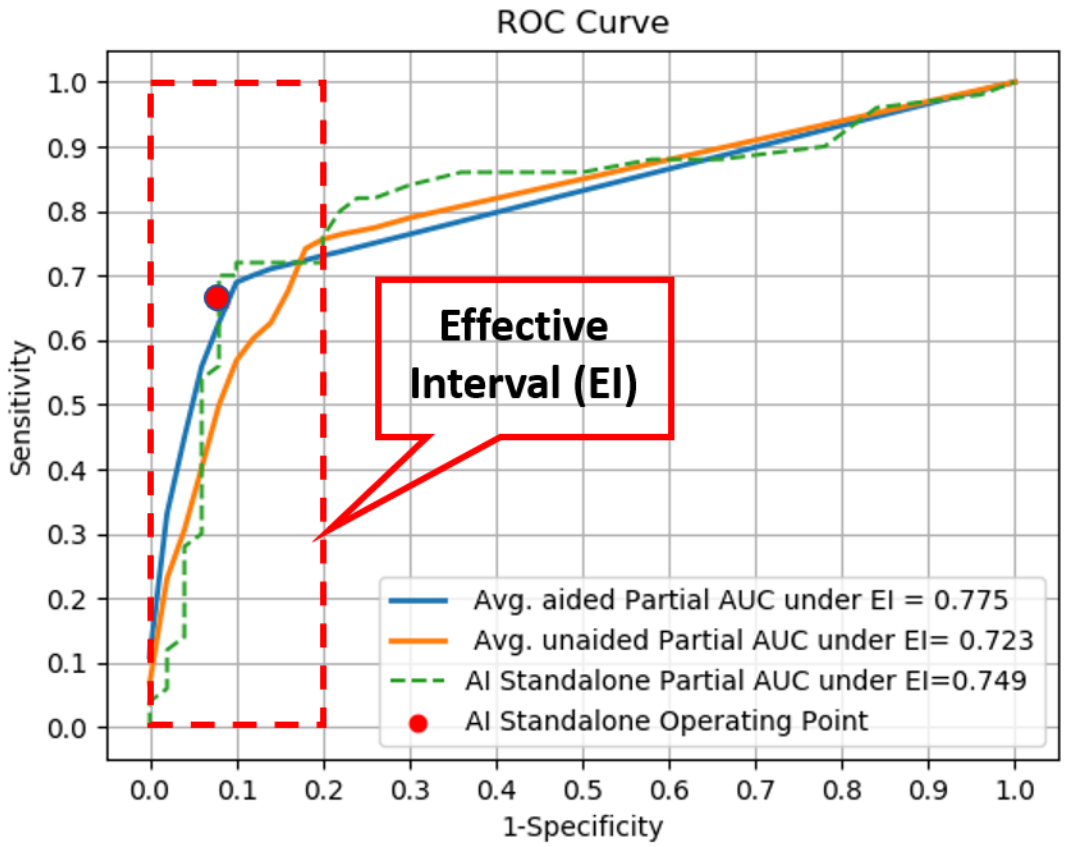
	Unaided	Aided
S3	5	0
J5	6	0
J1	4	0

eFigure 1. Deidentified CXRs From 3 Adult Patients Without Pulmonary Nodules (true negative per ground truth). For CXR A, AI decreased confidence of detecting false positive nodule for one junior radiologist and helped two senior and three junior radiologists avoid calling a false positive nodule. In CXR B, two senior radiologists did not detect a false positive nodule on AI-aided interpretation that they stated as present on unaided interpretation. AI helped one senior and two junior radiologists avoid calling a false positive nodule in CXR C that they reported on unaided interpretation.



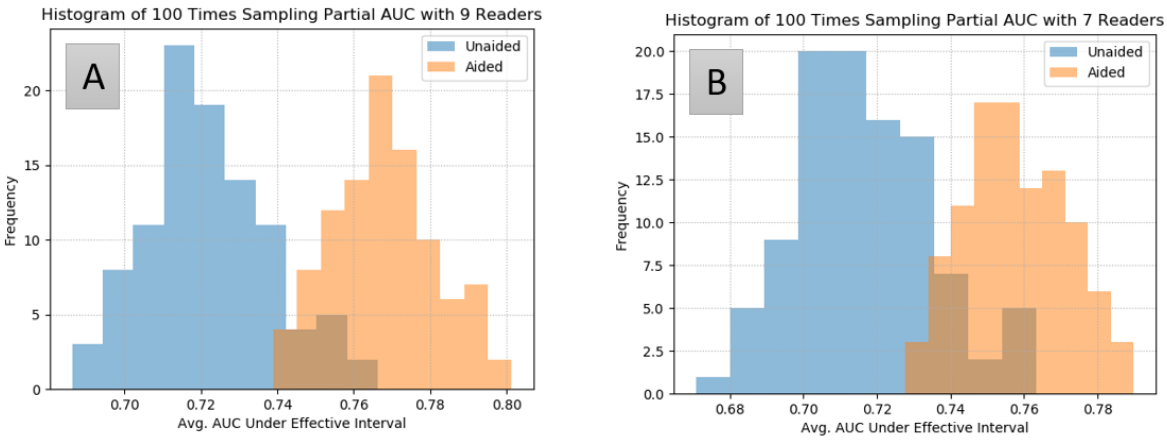
eFigure 2. AI ROC Curve Specificity and Sensitivity

Since all radiologists operated at a high specificity (0.8-1) which corresponds to low false positive rate (0-0.2), we deemed the 0-0.2 false positive rate as the effective interval and limited evaluation of changes in sensitivity and specificity of radiologists to this effective interval.



eFigure 3. ROC curve demonstrates averaged partial AUCs for unaided and AI-aided detection of pulmonary nodules on the included CXRs with in the effective interval

The AI standalone partial AUC represents partial AUC of standalone AI algorithm. The corresponding AI standalone sensitivity and specificity (the red point on the curve) are 0.66 and 0.92, respectively.



eFigure 4. Histograms depicting distribution of partial AUCs in aided and unaided modes over a 100 times sampling for different readers groups. **A:** All 9 readers ($p<0.001$) for improved partial AUCs with aided interpretation vs. unaided mode. **B:** Persistent and significant improvement with AI-aided interpretation for seven radiologists ($p<0.001$) after excluding the two radiologists (J2 and S4) with disproportionately higher improvement in aided mode.