Supplemental Online Content

Wu JT, Wong KCL, Gur Y, et al. Comparison of chest radiograph interpretations by artificial intelligence algorithm vs radiology residents. *JAMA Netw Open*. 2020;3(10):e2022779. doi:10.1001/jamanetworkopen.2020.22779

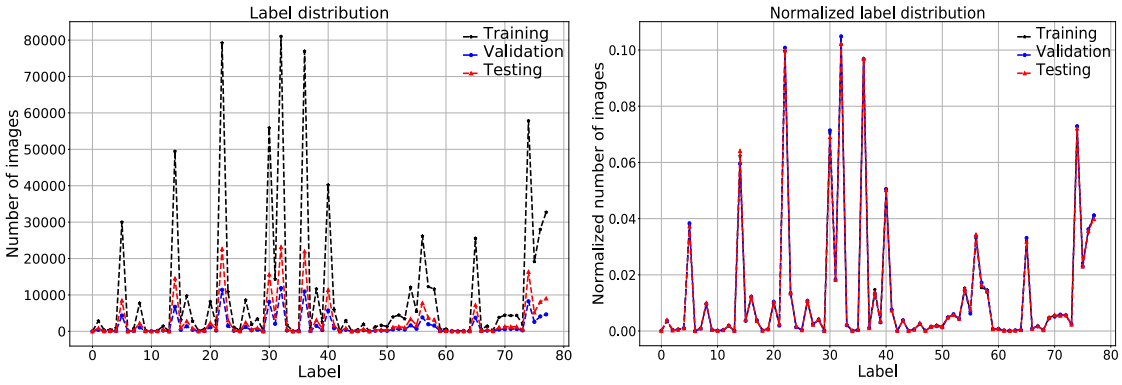
This supplemental material has been provided by the authors to give readers additional information about their work.

eFigure 1. Curation Process for Generating the List of Possible Findings in AP Chest Radiographs

eFigure 2. Vocabulary Expansion Process Used for the Chest Radiograph Lexicon Construction
The current candidate for expansion is the concept 'linear density'. The unsupervised learning algorithm analyzes textual reports such as the one shown in column 1. The proposed candidates are shown in column 3. The accepted and rejected candidates are used to propose better candidates in the next iteration.

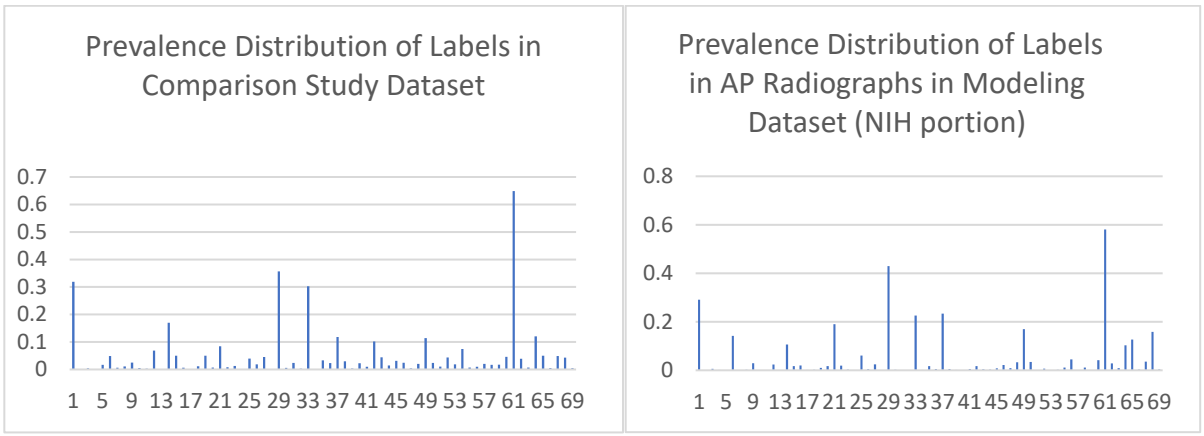
eFigure 3. Splitting Algorithm for Producing the Partitions for Training, Validation, and Testing in the Modeling Data Set

(a) Unnormalized distributions. (b) normalized distributions. The modeling dataset has both AP and PA images reflecting ambulatory and inpatient data across two hospital sources (NIH, MIMIC).



(a) (b)

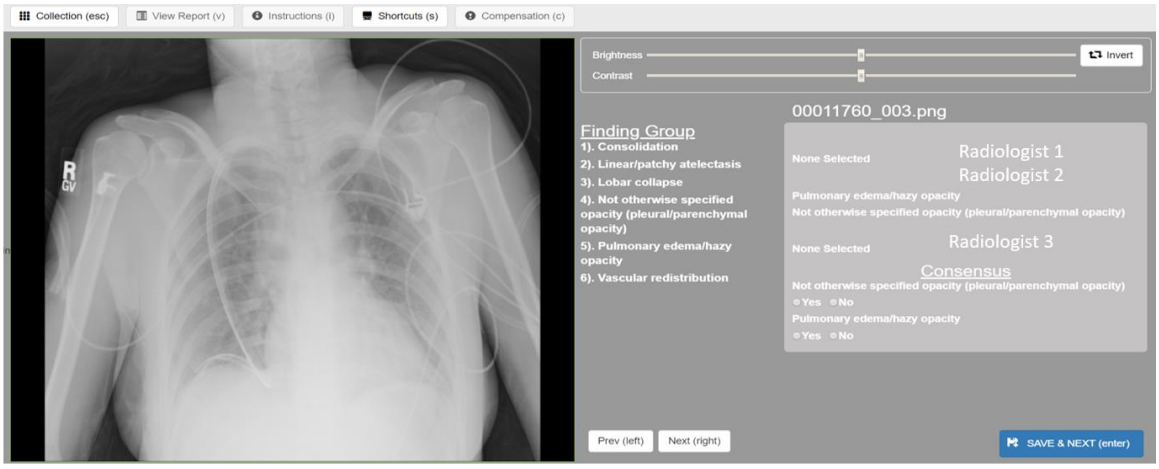
eFigure 4. Prevalence Distribution of the Labels in the Comparison Study Data Set
(a) The prevalence distribution of finding labels in the comparison study dataset. (b) The prevalence label distribution of AP chest radiographs in the NIH portion of the modeling dataset.



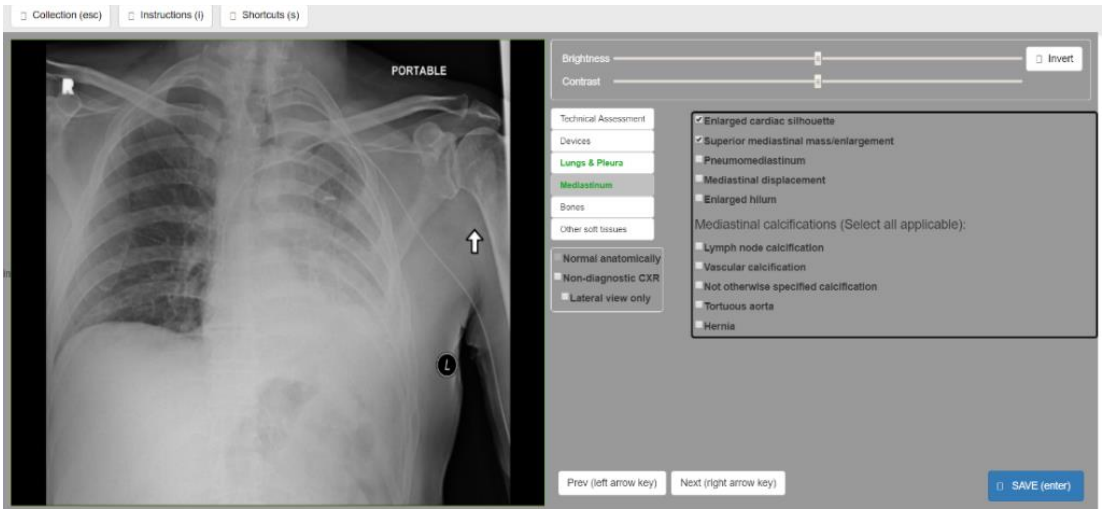
(a) (b)

eFigure 5. User Interface Used by Radiologists for Building Consensus After Independent Read Discrepancies Were Catalogued

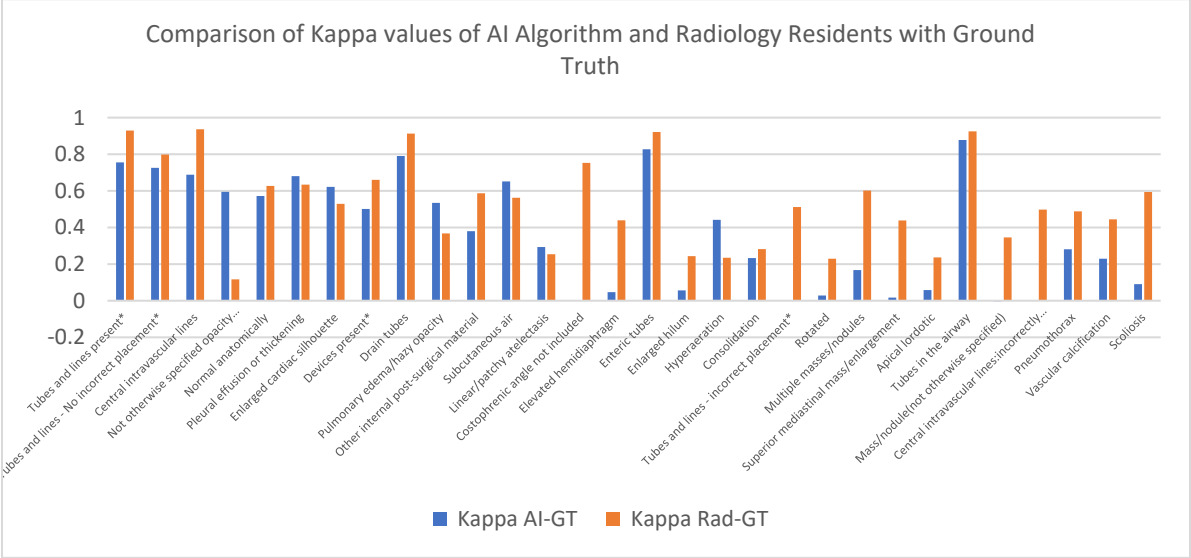
The discrepant labels resulting from the independent reads of the radiologists (stage 1) are resolved through a video conference discussion to build consensus in stage 2 whose interface is shown in the figure.



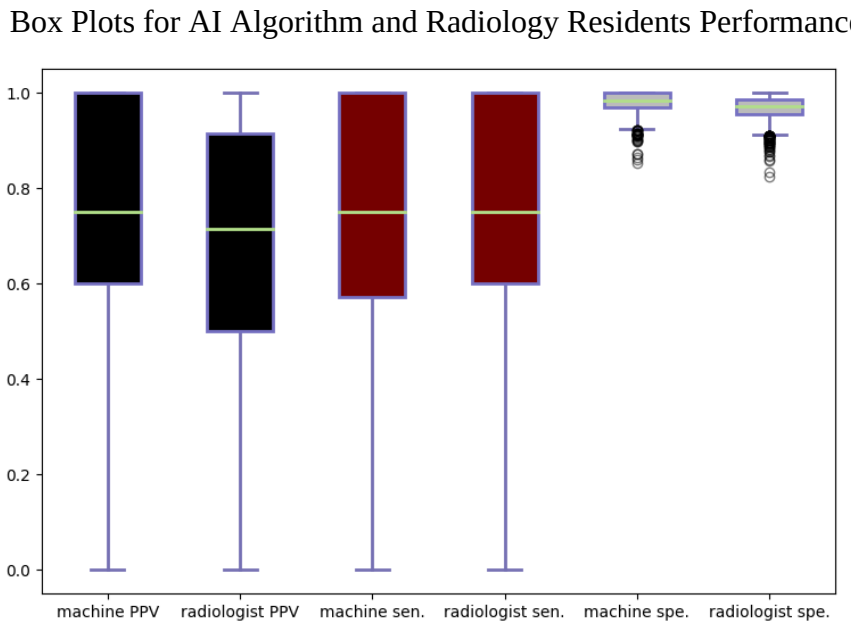
eFigure 6. Web-Based User Interface Used for Collecting the Reads from Radiology Residents on the Comparative Study Data Set



eFigure 7. Extent of Agreement With the Ground Truth for AI Algorithm and Radiology Residents on Labels in the Comparison Study Data Set With at Least 2.5% Prevalence
*Some finding labels are ontological abstractions of underlying labels (marked with an *).*



eFigure 8. Preliminary Read Performance Differences of Radiology Residents and the AI Algorithm



eTable 1. Finding Label Extraction From Reports Through Text Analytics
Column 1 shows the original text, and Column 2 lists the detected findings (both positive and negative).

Sample text from actual radiology reports	**NLP Produced Findings labels with negative (NO) or positive context (YES)**
Lines and tubes: None. Cardiomediastinal silhouette: Cardiomegaly. Lungs and pleura: perihilar and lower lobe ground glass opacity. query small effusions. no pneumothorax. Bones and soft tissues: No acute osseous abnormality. prior median sternotomy. surgical clips along the right axilla. Impression: cardiomegaly with moderate-severe pulmonary edema, likely from heart failure.	- Enlarged cardiac silhouette **(YES)** - Not otherwise specified opacity, Pleural/parenchymal opacity **(YES)** - Pleural effusion or thickening **(YES)** - Pneumothorax **(NO)** - Pulmonary edema/hazy opacity **(YES)** - Fracture (**NO**)
Mild peri-hilar opacities atelectasis or edema. Right small pleural effusion. No pneumothorax. ETT In appropriate position. CRT-D. mild cardiomegaly. Calcified density overlying the right shoulder maybe degenerative or post-traumatic.",	- Linear/patchy atelectasis **(YES)** - Not otherwise specified opacity, Pleural/parenchymal opacity **(YES)** - Pulmonary edema/hazy opacity **(YES)** - Pleural effusion or thickening **(YES)** - Pneumothorax **(NO)** - Enlarged cardiac silhouette **(YES)** - Not otherwise specified calcification **(YES)**

eTable 2. Performance of AI Algorithm vs Radiology Residents Across Labels With at Least 2.5% Prevalence in the Comparison Study Data Set

Finding label	Number of images in the comparison study dataset	Interpret difficulty	AUC in Comparison Study Dataset	DL Label-based PPV	DL Label-based sensitivity	DL Label-based specificity	Rads Label-based PPV	Rads Label-based sensitivity	Rads label-based specificity
Central intravascular lines	1296	medium	0.865	0.864	0.935	0.729	0.976	0.979	0.956
Not otherwise specified opacity (pleural/parenchymal opacity)	713	low	0.787	0.695	0.818	0.801	0.719	0.122	0.974
Normal anatomically	637	medium	0.932	0.608	0.936	0.718	0.748	0.744	0.882
Pleural effusion or thickening	604	low	0.940	0.729	0.849	0.863	0.862	0.629	0.956
Enlarged cardiac silhouette	339	low	0.902	0.621	0.785	0.902	0.631	0.581	0.931
Drain tubes	240	low	0.927	0.746	0.904	0.958	0.888	0.963	0.984
Pulmonary edema/hazy opacity	236	low	0.936	0.504	0.737	0.903	0.397	0.525	0.893
Other internal post-surgical material	228	low	0.997	0.503	0.395	0.950	0.540	0.794	0.913
Subcutaneous air	203	low	0.817	0.671	0.704	0.961	0.946	0.429	0.997
Linear/patchy atelectasis	168	low	0.997	0.322	0.405	0.922	0.235	0.661	0.802
Costophrenic angle not included	149	medium	0.836	1.000	0.000	1.000	0.829	0.718	0.988
Elevated hemidiaphragm	137	low	0.976	0.444	0.029	0.997	0.508	0.445	0.968
Enteric tubes	100	high	0.978	0.832	0.840	0.991	0.921	0.930	0.996
Enlarged hilum	100	high	0.583	0.250	0.040	0.994	0.355	0.220	0.979
Hyperaeration	100	low	0.917	0.440	0.510	0.966	0.450	0.180	0.988
Consolidation	97	low	0.869	0.205	0.464	0.908	0.232	0.577	0.903
Rotated	92	low	0.513	0.167	0.022	0.995	0.196	0.489	0.903
Multiple masses/nodules	91	low	0.744	0.264	0.154	0.980	0.540	0.736	0.970
Superior mediastinal mass/enlargement	88	high	0.757	0.200	0.011	0.998	0.471	0.455	0.976
Apical lordotic	87	high	0.72	0.190	0.046	0.991	0.421	0.184	0.988
Tubes in the airway	86	low	0.773	0.884	0.884	0.995	0.951	0.907	0.998
Mass/nodule (not otherwise specified)	79	high	0.944	0.000	0.000	0.998	0.348	0.405	0.969
Central intravascular lines - incorrectly positioned	78	low	0.753	1.000	0.000	1.000	0.365	0.949	0.933
Pneumothorax	66	high	0.682	0.250	0.409	0.958	0.463	0.561	0.978
Vascular calcification	62	low	0.611	0.476	0.161	0.994	0.769	0.323	0.997
Scoliosis	59	low	0.666	0.600	0.051	0.999	0.587	0.627	0.987

eTable 3. Comparative Finding Label Recognition Performance Between Radiologists and AI Algorithm

AI Outperformed radiologists	**Similar Performance of AI and Radiologists**	**Radiologists outperformed AI**
Not otherwise specified opacity (pleural/parenchymal opacity)	Tubes and lines present	Scoliosis
Pleural effusion or thickening	Tubes in the airway	Enlarged hilum
Enlarged cardiac silhouette	Enteric tubes	Rotated
Pulmonary edema/hazy opacity	Drain tubes	Costophrenic angle not included
Subcutaneous air	tubesandlines - no incorrect placement	Elevated hemidiaphragm
Hyperaeration	Device present	Superior mediastinal mass/enlargement
	Normal anatomically	Apical Lordotic
	Other internal post-surgical material	Mass/nodule (not otherwise specified)
	Pneumothorax	Central vascular lines – incorrectly positioned
	Linear/patchy atelectasis	Multiple masses and nodules
	Central intravascular lines	Tubes and lines – incorrect placement
	Consolidation	
	Vascular calcification	

eTable 4. Variation in Read Performance Across Radiology Residents

Method	Number of Images	Number of findings	Average image-based PPV	Average image-based sensitivity	Average image-based specificity
Resident 1	399	72	0.594 [0.567, 0.621]	0.688 [0.662,0.716]	0.958 [0.955, 0.962]
Resident 2	399	72	0.722 [0.697,0.748]	0.743 [0.719,0.768]	0.975 [0.972, 0.977]
Resident 3	400	72	0.704 [0.678,0.731]	0.729 [0.704,0.754]	0.971 [0.968,0.974]
Resident 4	400	72	0.648 [0.623,0.674]	0.685 [0.659,0.711]	0.967 [0.964,0.969]
Resident 5	400	72	0.743 [0.714,0.766]	75.45 [0.729,0.780]	0.975 [0.972,0.977]

eAppendix 1. Splitting Algorithm for Model Training

Here we provide additional details on the algorithm used for splitting the model dataset into training, validation and testing datasets. The goal of the splitting algorithm was two-fold: (a) ensuring the low incidence labels are still present in the training set in adequate numbers so that the model can be trained for these labels. (b) Ensuring the label distributions in the split datasets to be in a similar proportion to the original prevalence distribution so as to create the least sampling bias for testing. The splitting algorithm works by first sorting the distribution of labels by their frequencies of occurrences. Starting from the least frequent label, it then iteratively determines the size of the training, test, and validate sets of patients containing the target label so as to maintain the desired ratio of 70%,10%,20% for training, validation and test datasets. Once the number of patients in each split is determined per label, the assignment of the patients (and hence their images) is random. Note that since each image has multiple labels, each such split assignment per label covers other possible labels present in these images also maintaining their relative frequencies in the resulting distribution. The detailed algorithm is given below

Algorithm 1: Label distribution-preserving data split

Data: Label distribution – unique patient IDs (PIDs) for each label.
Input : p_{train}, p_{valid}, $p_{test} \in [0, 1]$ – percentages for training, validation, and testing.
Output: Distribution-preserved splits of PIDs for training (S_{train}), validation (S_{valid}), and testing (S_{test}).

```
begin
    L ← labels sorted in ascending order in terms of the number of PIDs
    S_train ← ∅, S_valid ← ∅, S_test ← ∅
    for l ∈ L do
        N_l ← unique PIDs of label l
        n_l ← number of PIDs of label l
        for i ∈ {train, valid, test} do
            n_li ← n_l × p_i                          // Desired number of PIDs
            for pid ∈ N_l do
                if pid ∈ S_i then  // Remove PID if already assigned
                    N_l ← N_l − pid
                    n_li ← n_li − 1
                end
            end
        end
        for i ∈ {train, valid, test} do // Distribute the rest of N_l
            Randomly assign n_li PIDs from N_l to S_i
        end
    end
end
```

We illustrate by an example. Suppose there are 100 images to be split and 2 possible labels (L1, L2) and assume each image comes from a unique patient for the purposes of this illustration. Suppose we have the following distribution.

Total images	Images with L1 and L2	Images with L1 only	Images with L2 only	Frequency of L1	Frequency of L2	L1:L2	Original L1:L2 Ratio
100	60	15	25	75	85	75:85	0.88

To now split this in the ratios by starting with the lowest frequency and using the 70-10-20% ratios for train-validate-test, we get

Iteration	Frequency	Label to split on	Images split for	Remaining Images to split	Train Images	Validate Images	Test Images
1	75= (60+15)	L1	L1 common with L2 (60 portion)	100	42	6	12
1		L1	L1 (15 portion)	40	11	1	3
1	85 = (60+25)	L1	L2 common with L1 (60 portion)	40	42	6	12
2		L2	L2 (25 portion)	25	18	2	5

Total images	Frequency of L1: L2 in Training split	Ratio	Frequency of L1:L2 in Validation split	Ratio	Frequency of L1:L2 in Test split	Ratio
100	53:60	0.88	7:8	0.875	15:17	0.88

As can be seen, the prevalence ratio has been maintained in the resulting splits. At the same time, the lower prevalence label (L1) has at least 53 training samples sufficient for training. A random sampling may not have ensured this since the overall dataset size is still small in this case (100 images), particularly when there are more labels per image.

eAppendix 2. Method of Threshold Selection for Finding Labels

Thresholding is required to convert the real-number prediction scores of the AI model to the binary scores of positives and negatives. Let θ be a vector that contains all label thresholds. To compute the optimal thresholds, an objective function based on the image-based F1 score is used:

$$L(\theta) = -\ln\left(\frac{1}{n}\sum_{i=1}^{n} \mathrm{F1}_i(\theta)\right)$$

with $F1_i$ the F1 score of image i and n the number of images. The F1 score is the harmonic mean of PPV and sensitivity, which is computed as:

$$\mathrm{F1} = \frac{2\mathrm{TP} + \epsilon}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \epsilon}$$

where TP, FP, and FN are the true positives, false positives, and false negatives, respectively, computed between the ground truth and the binary AI scores after thresholding by θ. $\epsilon = 10^{-7}$ is used to handle the 0/0 situation when there are no positives in both prediction and ground truth. The optimal θ can be computed by minimizing $L(\theta)$ through an optimization algorithm. The derivative-free global optimization algorithm, ESCH, is used as it provided the best results in our tested algorithms[1]. By focusing on the positive occurrences of findings per image and minimizing $L(\theta)$ we ensure that the network prediction has as few false positives while still enabling the detection of relevant findings.

1. C. H. da Silva Santos, M. S. Gonçalves and HEH-F. Designing Novel Photonic Devices by Bio-Inspired Computing. *IEEE Photonics Technol Lett*. 2010;22(15):1177-1179.

eAppendix 3. Measuring Deep Learning Model Performances for Multilabel Reads

In this appendix, we give further clarification on the choice of the performance measure used in the comparative study on machine and resident physician read performance.

Conventional approach to measuring performance:

Conventional approach is to report the performance on a per label basis and using the positive occurrence of a label:

Consider a set of images I $= \{I_1, I_2, .. I_K\}$ and a label L_i:
$P(L_i)$= Number of real positive cases in the data. So for a single label case, this implies the number of images in the dataset that are assigned this label L_i.
$N(L_i)$= Number of real negative cases in the data. So for a single label case, this implies the number of images in the dataset that are not assigned this label L_i.

Then $TP(L_i)$=The number of *images* for which the machine also predicts the label L_i *and the actual label is also* L_i.

Then $TN(L_i)$=The number of *images* for which the machine does not predict the label L_i *and the actual label is also not* L_i.

And $FP(L_i)$=The number of *images* for which the machine predicts the label L_i *and the actual label is not* L_i.

Then label-based positive predictive value (PPV) or precision is defined per label L_i as

$$PPV(L_i)=TP(L_i)/(TP(L_i)+FP(L_i)) \quad (1)$$

Label-based sensitivity is defined as

$$Sensitivity(L_i)=TP(L_i)/P(L_i) \quad (2)$$

And label-based specificity is defined as

$$\text{Specificity}(L_i)=TN(L_i)/N(L_i) \quad (3)$$

Image-based approach to measuring performance:

For the radiology read problem, since we have to maintain high precision and recall for each image we are reading, we used the image-based positive predictive value (precision) and sensitivity (recall) by redefining the terms as follows:

Consider again the set of images I $= \{I_1, I_2, .. I_K\}$ and the set of labels $L = \{L_1, L_2, .. L_M\}$.

$Let\ P(I_i)$= Number of labels actually occurring in the image I_i .

$Let\ N(I_i)$= Number of labels from the set L that are not occurring in the image I_i .
Thus $N(I_i) = L - P(I_i)$

Let $TP(I_i)$=The number of *labels* selected by the machine (or residents) for image I_i which belong to $P(I_i)$

Let $TN(I_i)$=The number of *labels not* selected by the machine (or residents) for image I_i which belong to $N(I_i)$

And $FP(I_i)$=The number of *labels* selected by the machine (or residents) for image I_i which belong to *L* but not $P(I_i)$

Then Image-based positive predictive value (PPV) or image-based precision is defined per image as

$$PPV(I_i)=TP(I_i)/(TP(I_i)+FP(I_i)) \quad (5)$$

Image-based sensitivity is defined as

$$Sensitivity(I_i)=TP(I_i)/P(I_i) \quad (6)$$

Image-based specificity is defined as

$$Specificity(I_i)=TN(I_i)/N(I_i) \quad (7)$$

Now averaging across the K images, we get

$$Average\ PPV(I)=\frac{1}{K}\sum_{i=1}^{K} PPV(I_i) \quad (8)$$

$$Average\ sensitivity(I)=\frac{1}{K}\sum_{i=1}^{K} Sensitivity(I_i) \quad (9)$$

$$Average\ specificity(I)=\frac{1}{K}\sum_{i=1}^{K} Specificity(I_i) \quad (10)$$

The above equations 8, 9, and 10 were used in the paper and ANOVA test for comparing the performance used the above formulas for computing image-based specificity and sensitivity respectively. As can be seen, the above measure is less sensitive to prevalence of labels as the PPV and sensitivity are measured per image by normalizing with respect to the respective prevalence within the image itself. It is also a more appropriate measure for the preliminary read use case, where the goal is to flag as few incorrect findings per image while still not missing many relevant findings. Optimizing on individual label's sensitivity or specificity would introduce false positives that can cumulatively impact the per image decision making sufficiently for a large of images, with the net effect of reducing the overall preliminary read quality (in terms of misses or overcalls).