Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Deep Learning to Quantify Pulmonary Edema in Chest Radiographs

*Steven Horng, MD, MMSc** • *Ruizhi Liao, SM** • *Xin Wang, PhD* • *Sandeep Dalal, BA* • *Polina Golland, PhD* • *Seth J. Berkowitz, MD*

From the Department of Radiology, Beth Israel Deaconess Medical Center, Harvard Medical School, 330 Brookline Ave, Boston, MA 02215 (S.H., S.J.B.); Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Mass (R.L., P.G.); and Clinical Informatics Solutions and Services, Philips Research, Cambridge, Mass (X.W., S.D.). Received December 24, 2019; revision requested February 16, 2020; revision received December 7; accepted December 17. **Address correspondence to** S.H. (e-mail: *shorng@bidmc.harvard.edu*).

*S.H. and R.L. contributed equally to this work.

Supported by the National Institutes of Health (grant NIBIB NAC P41EB015902), Philips Research, Wistron, MIT Lincoln Laboratory, and MIT Deshpande Center.

Conflicts of interest are listed at the end of this article.

See also the commentary by Auffermann in this issue.

Radiology: Artificial Intelligence 2021; 3(2):e190228 • https://doi.org/10.1148/ryai.2021190228 • Content codes: AI CH

Purpose: To develop a machine learning model to classify the severity grades of pulmonary edema on chest radiographs.

Materials and Methods: In this retrospective study, 369 071 chest radiographs and associated radiology reports from 64 581 patients (mean age, 51.71 years; 54.51% women) from the MIMIC-CXR chest radiograph dataset were included. This dataset was split into patients with and without congestive heart failure (CHF). Pulmonary edema severity labels from the associated radiology reports were extracted from patients with CHF as four different ordinal levels: 0, no edema; 1, vascular congestion; 2, interstitial edema; and 3, alveolar edema. Deep learning models were developed using two approaches: a semisupervised model using a variational autoencoder and a pretrained supervised learning model using a dense neural network. Receiver operating characteristic curve analysis was performed on both models.

Results: The area under the receiver operating characteristic curve (AUC) for differentiating alveolar edema from no edema was 0.99 for the semisupervised model and 0.87 for the pretrained models. Performance of the algorithm was inversely related to the difficulty in categorizing milder states of pulmonary edema (shown as AUCs for semisupervised model and pretrained model, respectively): 2 versus 0, 0.88 and 0.81; 1 versus 0, 0.79 and 0.66; 3 versus 1, 0.93 and 0.82; 2 versus 1, 0.69 and 0.73; and 3 versus 2, 0.88 and 0.63.

Conclusion: Deep learning models were trained on a large chest radiograph dataset and could grade the severity of pulmonary edema on chest radiographs with high performance.

Supplemental material is available for this article.

Chest radiographs are commonly performed to assess pulmonary edema (1). The signs of pulmonary edema on chest radiographs have been known for over 50 years (2,3). The grading of pulmonary edema is based on well-known radiologic findings on chest radiographs (4–7). The symptom of dyspnea caused by pulmonary edema is the most common reason a patient with acute decompensated congestive heart failure (CHF) seeks care in the emergency department and is ultimately admitted to the hospital (89% of patients) (8–10). Clinical management decisions for patients with acutely decompensated CHF are often based on grades of pulmonary edema severity, rather than its mere absence or presence. Clinicians often monitor changes in pulmonary edema severity to assess the efficacy of therapy. Accurate monitoring of pulmonary edema is essential when competing clinical priorities complicate clinical management (additional information in Appendix E1 [supplement]).

While we focused on patients with CHF within this study, the quantification of pulmonary edema on chest radiographs is useful throughout clinical medicine. Pulmonary edema is a manifestation of volume status in sepsis and renal failure, just as in CHF. Managing volume status is critical in the treatment of sepsis, but large-scale research has been limited because of longitudinal data on volume status. Quantification of pulmonary edema on a chest radiograph could be used as a surrogate for volume status, which would rapidly advance research in sepsis and other disease processes in which volume status is critical.

Large-scale and common datasets have been the catalyst for the rise of machine learning today (11). In 2019, investigators released MIMIC-CXR, a large-scale publicly available chest radiograph dataset (12–15). This investigation builds on that prior work by developing a common, clinically meaningful machine learning task and evaluation framework with baseline performance metrics to benchmark future algorithmic developments in grading pulmonary edema severity from chest radiographs. We developed image models using two common machine learning approaches: a semisupervised learning model and a supervised learning model pretrained on a large common image dataset.

Abbreviations

AUC = area under the ROC curve, CHF = congestive heart failure, ROC = receiver operating characteristic

Summary

Deep learning models were developed to quantify the extent of pulmonary edema on chest radiographs; the dataset and code used in this study are publicly available.

Key Points

- The area under the receiver operating characteristic curve for differentiating alveolar edema from no edema was 0.99 for a semisupervised model using a variational autoencoder and 0.87 for a model developed by using a pretrained supervised learning model.
- Performance of the algorithm was inversely related to the difficulty in categorizing milder states of pulmonary edema.

Materials and Methods

Study Design

This was a retrospective cohort study. This study was approved by the Beth Israel Deaconess Medical Center Committee on Clinical Investigation with a waiver of informed consent. We collected 369 071 chest radiographs and their associated radiology reports from 64 581 patients from the MIMIC-CXR chest radiograph dataset (12–14). Each imaging study is associated with one or more images. We aimed to identify patients with CHF within the dataset to limit confounding labels from other disease processes. First, we limited our study to only frontal radiographs, excluding a total of 121 646 images. Of these frontal radiographs (n = 247 425), there were 17 857 images which were acquired during visits with an emergency department discharge diagnosis code consistent with CHF. In total, this resulted in 16 108 radiology reports and 1916 patients who were included who had CHF. As part of a prior study (16), we manually reviewed patient charts and found this method of cohorting patients with CHF had 100% sensitivity and specificity. The other 62 665 patients were classified as non-CHF, and data were used in the semisupervised training model. An enrollment diagram is shown in Figure 1.

Label Extraction and Validation

We extracted the pulmonary edema severity labels ("none," "vascular congestion," "interstitial edema," and "alveolar edema") from the reports using regular expressions with negation detection. The extracted labels were numerically coded as follows: 0, none; 1, vascular congestion; 2, interstitial edema; and 3, alveolar edema (Table 1). Examples of the grades are shown in Figure E1 (supplement). We were able to label 3028 radiology reports and thus 3354 frontal view radiographs from 1266 patients (Fig 1). Among the 1266 patients, 1180 patients still have some of their reports unlabeled. The other 650 patients with CHF had no labeled reports.

To validate our label extraction in radiology reports, we randomly selected 200 labeled reports (50 for each severity category from patients with CHF). A board-certified radiologist (S.J.B., 5 years of experience, interventional radiology) then manually labeled the 200 reports, blinded from our label extraction results. We reported the precision (positive predictive value) of the regular expression results for each category and each keyword, and sensitivity and specificity of each keyword.

We had three senior radiology residents and one attending radiologist (S.J.B.) manually label a set of 141 frontal view radiographs from 123 patients (from the unlabeled dataset of 650 patients with CHF), which had no patient overlap with the report-labeled set (Fig E2 [supplement]). These images were set aside as our test set. Each radiologist assessed the images independently, and we reported their interrater agreement (Fleiss κ). We used a modified Delphi consensus process, further described in Appendix E1 (supplement), to develop a consensus reference standard label.

Model Development

To establish a baseline performance benchmark for this clinical machine learning task and to address the challenge of limited pulmonary edema labels, we developed models using two common computer vision approaches: a semisupervised model using a variational autoencoder (17) and a pretrained supervised learning model using a dense neural network (18,19). Both approaches aim to address the challenge of limited pulmonary edema labels. The first approach (semisupervised model) takes advantage of the chest radiographs without pulmonary edema severity labels, which includes approximately 220 000 images (from individuals with and without CHF) and is domain specific. The second approach (pretrained supervised model) uses a large-scale common image dataset with common object labels (such as cats and dogs), which includes approximately 14 million images and leverages the image recognition capability from other domains.

To mitigate the imbalanced dataset size of each severity level, we employed weighted cross-entropy as the loss term for training both models. Data augmentation (including random translation and rotation) was performed during training to accommodate the variable patient positionings.

Semisupervised learning model development.—To take advantage of the large number of unlabeled chest radiographs, we developed a Bayesian model that included a variational autoencoder for learning a latent representation from the entire radiograph set (exclusive of the test set) trained jointly with a classifier that employs this representation for estimating edema severity. We first trained the variational autoencoder on both unlabeled and labeled images (exclusive of the test set), although the labels were not involved at this stage. The variational autoencoder learned to encode the chest radiographs into compact (low-dimensional) image feature representations by an encoder and learned to reconstruct the images from the feature representation by a decoder. We then took the trained encoder and concatenated it with an image classifier that estimates pulmonary edema severity. Finally, we trained this encoder with the classifier on labeled images in a supervised learning fashion. The use of this variational autoencoder architecture allowed us to leverage a large number of unlabeled images to train a model that learns the underlying

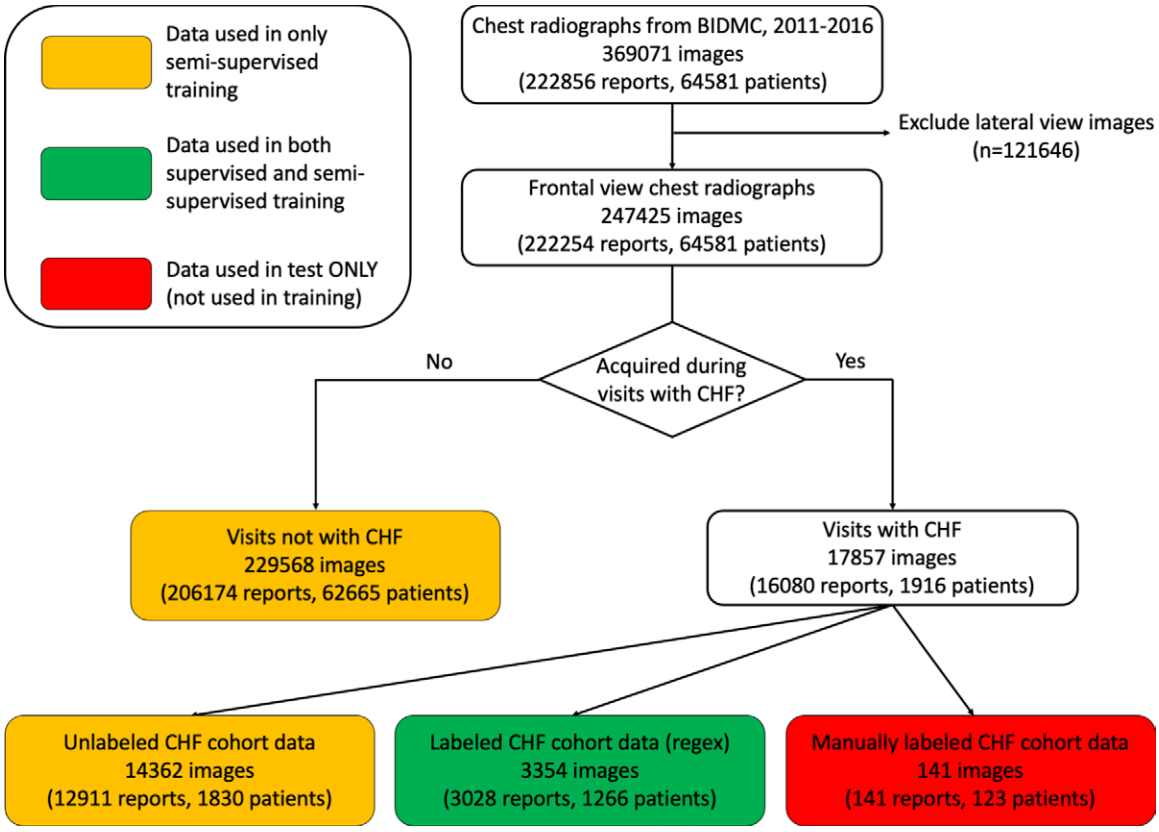

Figure 1: Cohort selection flowchart. A total of 369 071 chest radiographs and their associated radiology reports from 62 665 patients were collected. Images for this study were limited to frontal view radiographs (247 425). Of the 247 425 frontal view radiographs, 17 857 images were acquired during visits with a diagnosis consistent with congestive heart failure (CHF). In the CHF cohort, we were able to label 3028 radiology reports and thus 3354 frontal view radiographs from 1266 patients, using regular expressions (regex) on the reports. We also curated a test set of 141 radiographs that were manually labeled by radiologists (from the 650 unlabeled radiographs from patients with CHF). BIDMC = Beth Israel Deaconess Medical Center.

Table 1: Validation of Keyword Terms

Edema Severity	Keyword	No. of Reports	Precision (%)	Sensitivity (%)	Specificity (%)
"Overall"	N/A	200	92	N/A	N/A
None	No pulmonary edema	24	95.83	40.35	99.41
	No vascular congestion	18	94.44	29.82	99.41
	No fluid overload	2	100	3.51	100
	No acute cardiopulmonary process	13	92.31	21.05	99.41
Vascular congestion	Cephalization	24	75	33.96	96.55
	Mild pulmonary vascular congestion	24	91.67	41.51	98.85
	Mild hilar engorgement	2	100	3.77	100
	Mild vascular plethora	8	100	15.09	100
Interstitial edema	Interstitial opacities	15	93.33	20.90	99.38
	Kerley	19	100	28.36	100
	Interstitial edema	20	100	29.85	100
	Interstitial thickening	8	75	8.96	98.75
Alveolar edema	Alveolar infiltrates	16	100	32.00	100
	Severe pulmonary edema	33	90.91	60.00	98.87
	Perihilar infiltrates	1	100	2.00	100
	Hilar infiltrates	1	100	2.00	100

Note.—The total number of reports from all the keywords is more than 200 because some reports have more than one keyword. The low sensitivity and high specificity of each keyword indicate that no single keyword can represent the entire severity level but every keyword is specific to the severity level that it is supposed to belong to.

features of chest radiograph images. By training the variational autoencoder jointly with a classifier on the labeled images, we ensured it captured compact feature representations for scoring pulmonary edema severity. We also used data augmentation by random image translation, rotation, and cropping to a size of 2048 × 2048 during training to improve the robustness of the model. We used deep convolutional neural networks to implement the variational autoencoder and the classifier. The encoder of the variational autoencoder has eight residual blocks (5), the decoder has five deconvolution layers, and the classifier has four residual blocks followed by two fully connected layers.

We also varied the number of unlabeled chest radiographs used to train this semisupervised model to assess how the model performance changed with the amount of unlabeled data. We reported the average of the nine area under the receiver operating characteristic curve (AUC) values (as in Table 2) in Table E1 (supplement).

Table 2: AUC from the Semisupervised Model and the Pretrained Supervised Learning Model on the Test Set

Comparison	Semisupervised	Pretrained Supervised	*P* Value*
0 vs 1	0.79	0.66	.02
0 vs 2	0.88	0.81	.29
0 vs 3	0.99	0.87	.003
1 vs 2	0.69	0.73	.58
1 vs 3	0.93	0.82	.07
2 vs 3	0.88	0.63	.01
0 vs 1, 2, 3	0.85	0.74	.008
0, 1 vs 2, 3	0.88	0.81	.15
0, 1, 2 vs 3	0.96	0.82	.002

*Significance testing between the semisupervised model and the pretrained supervised model area under the curve (AUC) using DeLong method (*P* value of the hypothesis that they have the same performance). To account for multiple comparisons, a Bonferroni correction was used where a *P* value below .005 indicates a significant difference (α = .05/9 = .005). All the results are based on the predictions of the test set.

Pretrained model development.—In the second approach, we started with a neural network that had been pretrained to recognize common images (eg, cats and dogs) and then further tuned it to recognize the specific image features of chest radiographs for assessing pulmonary edema. Specifically, we used the densely connected convolutional neural networks (DenseNet) (6), and the model was pretrained on ImageNet (7). The DenseNet has four dense blocks (6), which consist of 6, 12, 24, and 16 convolutional layers, respectively. The four dense blocks are concatenated with a 2-by-2 averaging pooling layer between each two consecutive dense blocks. We kept the first three pretrained dense blocks for low-level image feature extraction, followed by one global average pooling layer, one dropout layer, and two fully connected layers. We then retrained this model on our labeled chest radiographs. We also used data augmentation by random image translation, rotation, and cropping to a size of 512 × 512 (for adjusting the image size in the ImageNet) during training to improve the robustness of the model.

Statistical Analysis

Study population means and 95% CIs were reported for age, and percentages were reported for sex and disposition. A Student *t* test was used to test for significance for age, and a Pearson χ^2 test was used for sex and disposition.

To understand how many and how frequently chest radiographs have been taken on our CHF cohort and non-CHF cohort, we calculated the number of images from each patient in our dataset and plotted the histograms of the numbers for the CHF cohort and for the non-CHF cohort. We also showed the distributions of time intervals between two consecutive chest radiographs obtained in a patient with CHF.

To evaluate the model, we performed fivefold cross-validation and randomly split the 3354 labeled images into fivefolds, ensuring that images from the same patients were allocated to the same fold. For each round, fourfolds were used for training and the remaining fold was held out for evaluation. Each model was trained five times independently to evaluate all fivefolds. During training, the validation fold was never seen by the model. We selected the best trained model among the five and tested it on the manually labeled image test set. The distribution of severity labels across folds and the test set is summarized in Table 3. The cross-validation results are summarized in Appendix E1 (supplement).

We plotted receiver operating characteristic (ROC) curves and reported the AUC for each pairwise comparison between severity labels on the test set. We then dichotomized the severity and reported three comparisons: *(a)* 0 versus 1,2,3; *(b)* 0,1 versus 2,3; and *(c)* 0,1,2 versus 3. We used the DeLong method to test for significance between AUCs between the semisupervised model and the pretrained model. To account for multiple comparisons, a Bonferroni correction was used with α = .05 ÷ 9 = .005.

Last, we show the confusion matrices for each of the models. To interpret the model predictions, we used Grad-CAM (gradient-weighted class activation mapping) to produce heatmaps to visualize the areas of the radiographs that were most informative for grading pulmonary edema severity. Grad-CAM computes the gradients of the model prediction with respect to the feature maps of the last convolutional layer in the model. The gradients are used to calculate the weighted average of the feature maps, and the weighted average map is displayed as a heatmap to visualize image regions that are important for the model prediction (20).

Data Availability

All underlying data, labels, and code are available at *https://github.com/RayRuizhiLiao/mimic_cxr_edema*.

Table 3: Distribution of Severity Labels across Folds and Test Set

Test Set/Fold	0, None	1, Vascular Congestion	2, Interstitial Edema	3, Alveolar Edema	Total Images
Unlabeled (n = 63 149)	…	…	…	…	229 519
Labeled-regular expressions (cross validation)					
Fold 1 (n = 254)	260	130	189	27	606
Fold 2 (n = 253)	296	150	215	31	692
Fold 3 (n = 253)	269	130	236	26	661
Fold 4 (n = 253)	292	153	194	38	677
Fold 5 (n = 253)	302	153	237	26	718
Subtotal (n = 1266)	1419 (42.13%)	716 (21.35%)	1071 (31.93%)	148 (4.41%)	3354 (100%)
Labeled-manual (test) (n = 123)	61 (43.26%)	44 (31.21%)	20 (14.18%)	16 (11.35%)	141 (100%)

Note.—Regular expressions indicate the retrieval of text from the original radiology reports within the database.

Results

Patient and Chest Radiograph Characteristics

We analyzed the chest radiograph distributions in our CHF cohort (1916 patients) and non-CHF cohort (62 665 patients). The histograms for number of chest radiographs and interval time are shown in Figure E3 (supplement). The mean number of chest radiographs taken per patient with CHF was 14 (median, nine; range, 1–153) and per patient with no CHF was five (median, three; range, 1–174). For patients with CHF, the mean interval time between each two consecutive chest radiograph orders from the same patient was 71 days (median, 7 days; range 0.13–1545 days). A total of 21.53% of patients had interval times within 1 day, while 66.08% had interval times within 30 days. Additional information on radiographs and patients is shown in Table 4.

Validation of Outcome Measures

The precision values (positive predictive value) of the regular expression results (ie, extracting pulmonary edema severity labels from the radiology reports within the dataset) for "none," "vascular congestion," "interstitial edema," and "alveolar edema" based on the manual review results were 96%, 84%, 94%, and 94%, respectively. The overall precision was 92%. The precision, sensitivity, and specificity for each keyword are summarized in Table 1.

After independent labeling, discussion, and voting, the interrater agreement (Fleiss κ) among the three radiology residents was 0.97 (more details in Figure E2 [supplement]). Our modified Delphi process yielded consensus labels for all 141 images.

ROC Curve Analysis

The ROC curves of the two models on the test set are shown in Figure 2. As expected, both models performed well on the task of distinguishing images between level 0 and level 3 and on the task of classifying between level 3 and the rest. The AUC for differentiating alveolar edema (score 3) from no edema (score 0) was 0.99 and 0.87 for semisupervised and pretrained models, respectively. Performance of the algorithm was inversely related to the difficulty in categorizing milder states of pulmonary edema (shown as the AUC for the semisupervised and pretrained model, respectively, for differentiating the following categories): 2 versus 0, 0.88 and 0.81; 1 versus 0, 0.79 and 0.66; 3 versus 1, 0.93 and 0.82; 2 versus 1, 0.69 and 0.73; and 3 versus 2, 0.88 and 0.63. The ROC curves from the cross-validation are shown in Figure E4 (supplement).

The AUCs of the two models on the test set are reported in Table 2. Seven out of the nine Delong test significance values were higher than .005, which means that the two models did not have significantly different AUCs. The AUCs of the cross-validation results are reported in Table E2 (supplement).

Confusion Matrix Analysis

We computed a confusion matrix for each of the models on the test set (Fig 3). Each image was placed in a cell by the true severity level from consensus score and the predicted severity level from the image model. In each cell, we reported the fraction of the predicted severity level in the actual severity level. Both models performed better in predicting level 0 and level 3 compared with predicting level 1 and level 2. The confusion matrices from the cross-validation are summarized in Figure E5 (supplement).

Predicted Edema Severity in Bar Charts

We plotted bar charts of predicted edema severity versus true edema severity on the test set (Fig 4). Both plots show the linear trend of predicted edema severity with ground truth edema severity. Overlap of error bars graphically depicts the challenges in discriminating less severe stages of pulmonary edema. Pulmonary edema severity exists on a continuous spectrum and future work on this will be discussed in the following section. Similar bar charts from the cross-validation are reported in Figure E6 (supplement).

Model Interpretation

We used Grad-CAM to visualize the regions in a radiograph that are important for the model prediction. Figure 5 dem-

Table 4: Patient Demographics and Characteristics

	CHF (n = 1916)				
Parameter	Labeled (n = 1266)	Unlabeled (n = 650)	Total (n = 1916)	Non-CHF (n = 62 665)	P Value
Age (y)*	73 (72.0, 74.1)	75.8 (75.2, 76.4)	75.1 (74.5, 75.6)	51.0 (50.9, 51.1)	< .001
Women (%)	51.8	51.3	51.4	54.6	.001
Disposition (%)					< .001
Admit	91.5	93.6	92.8	35.6	
Discharge	8.2	5.9	6.5	59.6	
AMA	0.0	0.2	0.2	0.3	
Cardiac catheterization	0.0	0.1	0.0	0.1	
Eloped	0.0	0.0	0.0	1.1	
Died	0.0	0.2	0.1	0.1	
Labor & Delivery	0.0	0.0	0.0	0.0	
LWBS	0.2	0.0	0.0	1.1	
OR	0.2	0.1	0.1	0.7	
Transfer	0.0	0.0	0.2	1.4	
No. of chest radiographs†			9 (1–153)	3 (1–174)	
Interval (d)†			7.09 (0.13–1545)		

Note.—Data are percentages. Interval indicates the interval between two consecutive chest radiographs from the same patient. AMA = against medical advice, CHF = congestive heart failure, LWBS = leave without being seen, OR = operating room.
* Age is mean, with 95% CIs in parentheses.
† Number of chest radiographs per patient and the interval time between two chest radiographs are shown as median with range in parentheses.

onstrates two sample images from the two models. We also manually reviewed the test data set in an attempt to classify the failure modes of both the semisupervised and pretrained models (Table E3 [supplement]).

Discussion

We employed two different machine learning techniques to quantify pulmonary edema. The semisupervised approach learned from all the radiographs in the training set. The pretrained image model learned from a large common image set and the labeled radiographs. Both approaches aimed to address the challenge of limited labels extracted from the radiology reports. Both approaches had similar performance statistically in terms of AUC on most pairwise classification comparisons (seven of nine). On the other two comparisons (two of nine), the semisupervised approach outperformed the pretrained approach. The semisupervised approach may have given better results because it learned from approximately 220 000 chest radiographs and was thus tailored to the image feature extraction of chest radiographs.

The semisupervised model was rarely off by two levels of pulmonary edema and never disagreed by three levels from the consensus label. However, there were examples in which the pretrained model predicted alveolar edema or no pulmonary edema when the consensus label was on the opposite end of the spectrum. More work is needed to improve the explainability of the model to understand these failure modes which are clearly critical before such a model could be deployed in clinical practice. Importantly, however, the manual review showed several examples where the models were able to correctly assess the absence of pulmonary edema despite the presence of severe cardiomegaly and pleural effusions.

The results of these algorithms provided a performance benchmark for future work. We have shown that it is feasible to automatically classify four levels of pulmonary edema on chest radiographs. Understandably, the performance of the algorithm mirrors the challenge of distinguishing these disease states for radiologists. The differentiation of alveolar edema from no pulmonary edema (level 3 vs 0) is an easier task than distinguishing interstitial edema from pulmonary vascular congestion (level 2 vs 1). Even among radiologists, there is substantial variability in the assessment of pulmonary edema. More machine learning approaches should be explored for this clinical task in future work.

Our work expanded on prior studies by employing machine learning algorithms to automatically and quantitatively assess the severity of pulmonary edema from chest radiographs. Prior work has shown the ability of convolutional neural networks to detect pulmonary edema among several other pathologic conditions that may be visualized on chest radiographs (21–23). Neural networks have been validated in large datasets to achieve expert level identification of findings in chest radiographs (24). Their AUCs in detecting the presence of pulmonary edema range from 0.83 to 0.88. By treating pulmonary edema as a single pathologic condition, it is difficult to draw direct comparison to our work which considered pulmonary edema as a spectrum of findings. A conservative comparison would be to compare prior work to our

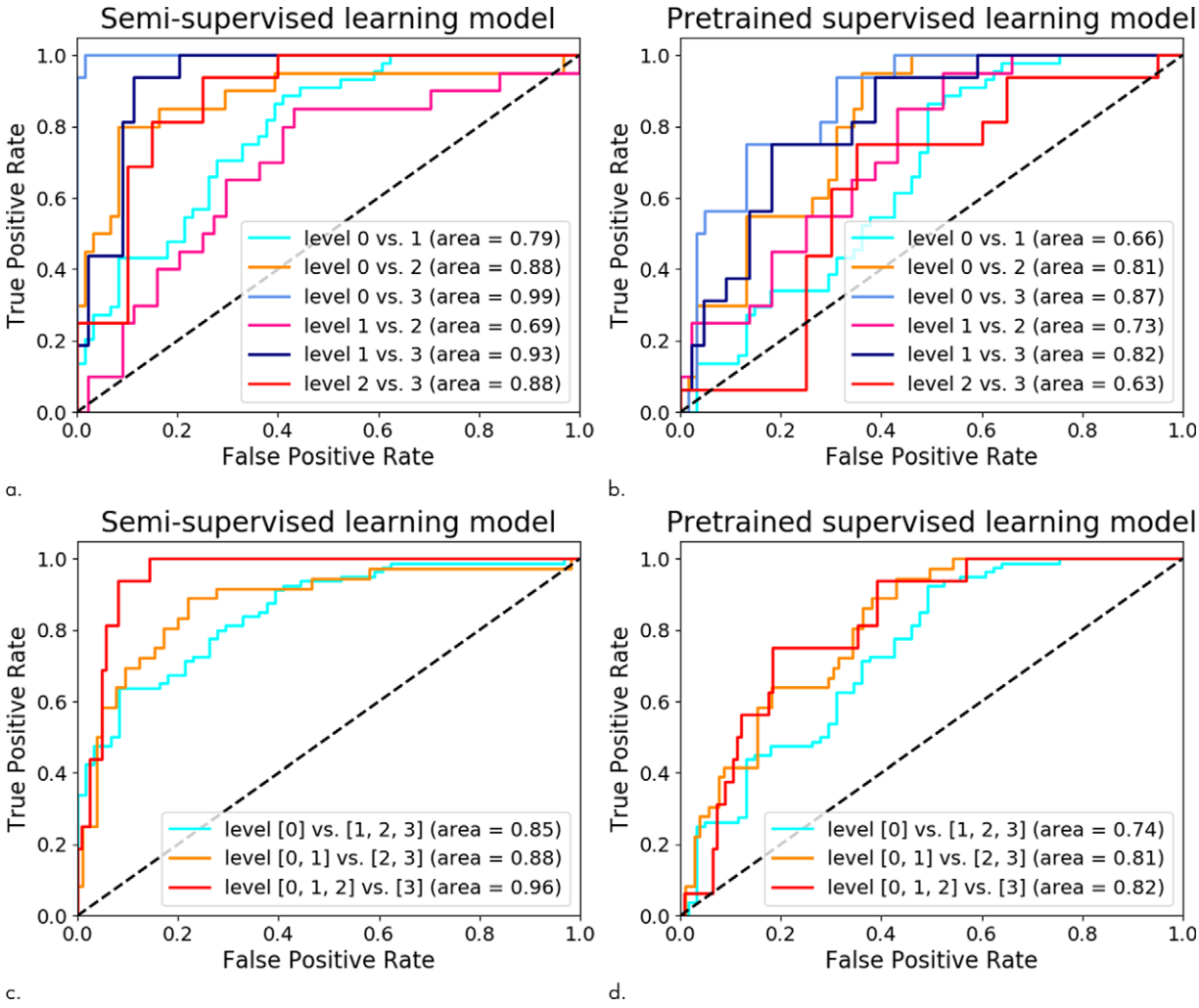

Figure 2: Receiver operating characteristic (ROC) curves of the semisupervised learning model and the pretrained supervised learning model. All the curves are based on the predictions of the test set. **(a, b)** ROC curves for six pairwise comparisons. **(c, d)** ROC curves for three dichotomized severity comparisons. All the curves are based on the predictions of the test set.

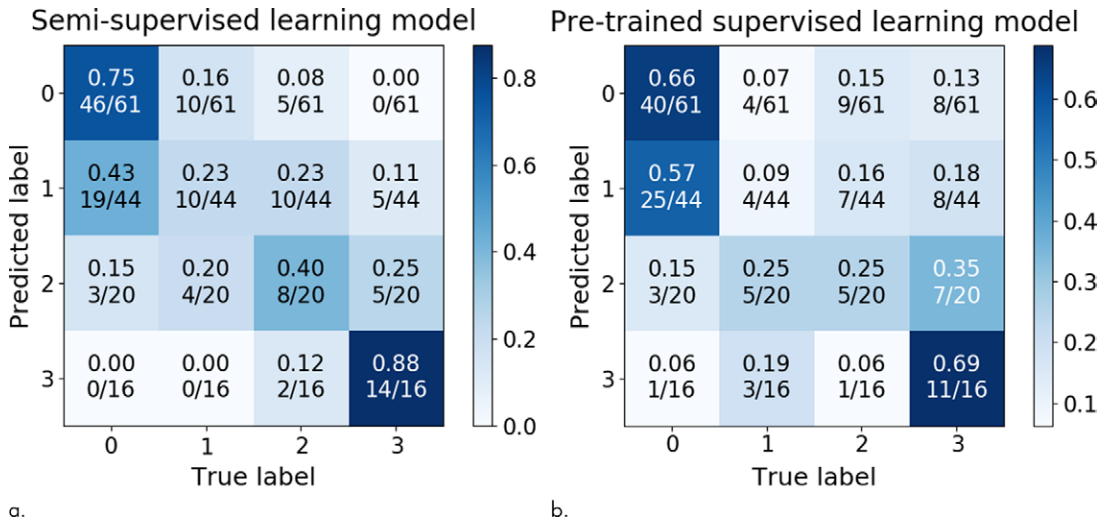

Figure 3: Confusion matrices from the **(a)** semisupervised learning model and the **(b)** pretrained supervised learning model. The denominator of each fraction number is the number of images that the algorithm predicts of the corresponding row, and the numerator is the number of images that belongs to the corresponding column. The quadratic-weighted κ values of the semisupervised learning model and the pretrained supervised learning model are 0.70 and 0.41. All the results are based on the predictions of the test set.

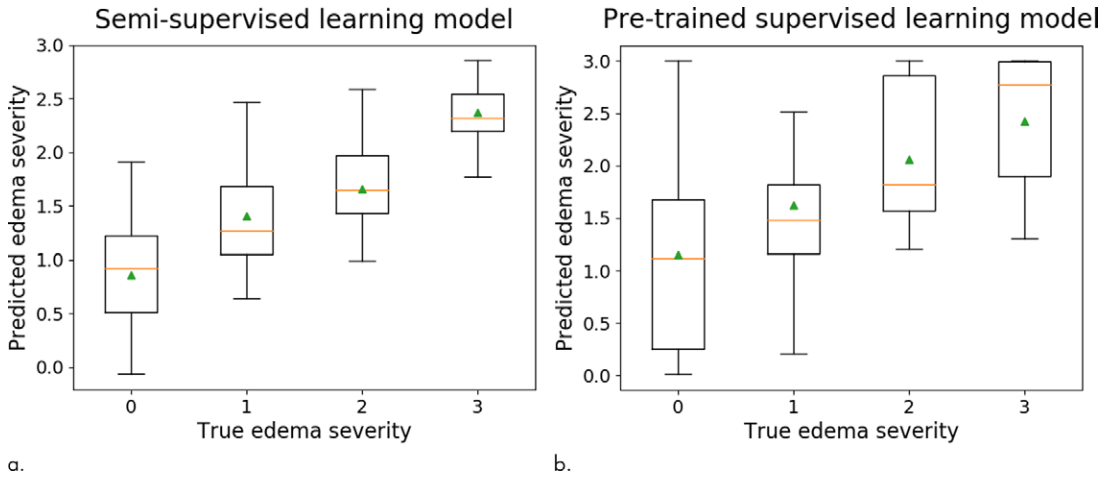

Figure 4: Predicted edema severity scores versus true edema severity labels from the **(a)** semisupervised learning model and the **(b)** pretrained supervised learning model. The box extends from the lower to upper quartile values of the distribution, with the orange line at the median and the green triangle at the mean. The whiskers extend from the box to show the range of the data. All the results are based on the predictions of the test set.

model's ability to distinguish no edema and pulmonary vascular congestion from interstitial and alveolar edema (levels 0,1 vs 2,3) which have AUCs of 0.81 (pretrained) and 0.88 (semisupervised). Although their test sets are based on labels extracted from radiology reports, our test set labels were annotated and had consensus reached by four radiologists. Others have trained neural networks on B-type natriuretic peptide values to produce a quantitative assessment of CHF (25). However, B-type natriuretic peptide increases nonlinearly with worsening CHF and exhibits marked interpatient variability. A B-type natriuretic peptide of 1000 in one patient could represent an acute exacerbation, while being the baseline for another patient, making B-type natriuretic peptide a poor surrogate outcome measure for acute pulmonary edema. The grading of pulmonary edema severity relies on much more subtle radiologic findings (image features). The clinical management of patients with pulmonary edema requires comparisons of serial examinations and understanding serial trends. Accurate, reproducible, and rapid quantification of pulmonary edema is of paramount value to clinicians caring for these patients.

There were limitations in our study. Extracting labels from clinical radiology reports allowed us to quickly obtain a reasonable amount of labeled data, but is inferior to data labeled for a specific purpose. Not only is there poor interreader agreement among radiologists for pulmonary edema detection (26), but radiologists may use different languages to describe a similar pathophysiologic state. In future work, we will explore joint modeling of chest radiographs and radiology reports and aim to mitigate the bias introduced by simply employing regular expressions.

Pulmonary edema exists on a continuous spectrum of severity. By discretizing our data into four classes, we have potentially lost valuable information and contaminated the categories. The category of severe edema in our dataset contained all images containing alveolar edema, even though this varies wildly in clinical practice. In practice, it is challenging to quantify pulmonary edema at a more granular level. Comparisons between images are easier and more reproducible. Future work could leverage pairs of images to quantify edema on a continuous scale.

The diagnosis of pulmonary edema is often challenging because of the possibility of other competing diagnoses that have overlapping radiographic findings. For example, multifocal pneumonia can be confused with alveolar pulmonary edema, and chronic interstitial edema can be misinterpreted as interstitial pulmonary edema. To minimize this bias, we restricted our labeled data to a cohort of patients diagnosed with CHF. In this work, we purposely ignored image findings such as cardiomegaly and pleural effusions that are correlated with pulmonary edema and often used by radiologists when making the diagnosis. In future work, we plan to leverage multitask training to jointly learn these associated features. By incorporating multiple image observations in the model training, an algorithm would approximate the clinical gestalt that a radiologist has when considering the etiology of pulmonary opacities. By separating the features of pulmonary edema from features that are associated with CHF, however, our model was not biased against detecting noncardiogenic pulmonary edema.

Last, we compared our results only to the chest radiograph rather than some other reference standard of pulmonary edema. In clinical practice, the chest radiograph is usually considered the reference standard to measure pulmonary edema. Pulmonary capillary wedge pressure might be more accurate, but is extremely invasive, and performed only on a small fraction of patients; therefore, it would be impractical to be used as a reference standard.

Accurate grading of pulmonary edema on chest radiographs is a clinically important task. The models developed in this study were capable of classifying edema grades on chest radiographs.

Acknowledgments: The authors thank Alistair Johnson, James L. Smith, Stanley Y. Kim, and Amalie C. Thavikulwat for helping with the data curation.

Author contributions: Guarantors of integrity of entire study, S.H., R.L., P.G., S.J.B.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved,

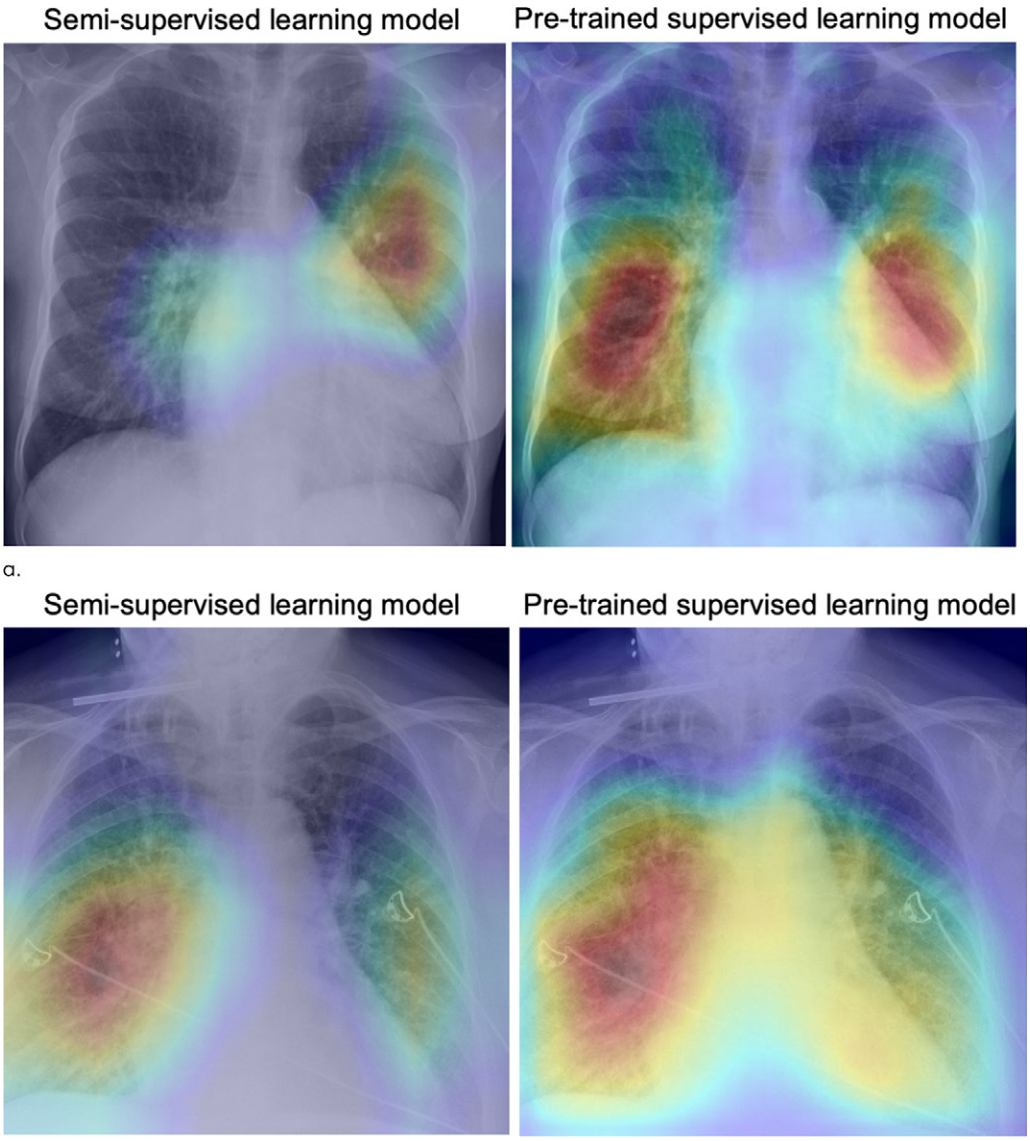

Figure 5: Grad-CAM heatmaps that highlight important regions for the model prediction. **(a)** A sample radiograph that is labeled as "vascular congestion" (level 1). **(b)** A sample radiograph that is labeled as "alveolar edema" (level 3).

all authors; literature research, S.H., R.L., X.W., P.G., S.J.B.; clinical studies, S.H., S.J.B.; experimental studies, R.L., X.W., P.G., S.J.B.; statistical analysis, R.L., X.W., P.G.; and manuscript editing, all authors.

Disclosures of Conflicts of Interest: S.H. Activities related to the present article: institution received grant from Philips Healthcare (sponsor had time to review the manuscript for patentable content but otherwise had no oversight of any portion of the study; funding source had no role in the design of this study, execution, analysis, or interpretation); research project was initiated by S.H. and P.G. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **R.L.** Activities related to the present article: institution received research grant from Philips Research and include research members from Philips Research; The research project was initiated by S.H. and P.G. The funding source had no role in the design of this study, execution, analysis, or interpretation. The funding source had an opportunity to preview the submission for potential patentable intellectual property, but otherwise had no role in the decision to submit results. Members from Philips Research participated in the algorithm design as stated in the contributorship statement. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **X.W.** disclosed no relevant relationships. **S.D.** disclosed no relevant relationships. **P.G.** Activities related to the present article: institution received grant from Philips Healthcare (sponsor had time to review the manuscript for patentable content but otherwise had no oversight of any portion of the study; funding source had no role in the design of this study, execution, analysis, or interpretation); research project was initiated by S.H. and P.G. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **S.J.B.** Activities related to the present article: institution received grant from Philips Research (Work performed under research grant from Philips that supported investigators and graduate student. Investigators remained completely independent.) Activities not related to the present article: author is consultant and board member for Change Healthcare. Other relationships: disclosed no relevant relationships.

References

1. Mahdyoon H, Klein R, Eyler W, Lakier JB, Chakko SC, Gheorghiade M. Radiographic pulmonary congestion in end-stage congestive heart failure. Am J Cardiol 1989;63(9):625–627.
2. Logue RB, Rogers JV, Gay BB. Subtle roentgenographic signs of left heart failure. Am Heart J 1963;65(4):464–473.
3. Harrison MO, Conte PJ, Heitzman ER. Radiological detection of clinically occult cardiac failure following myocardial infarction. Br J Radiol 1971;44(520):265–272.

4. Milne EN. Correlation of physiologic findings with chest roentgenology. Radiol Clin North Am 1973;11(1):17–47.
5. Van de Water JM, Sheh JM, O'Connor NE, Miller IT, Milne EN. Pulmonary extravascular water volume: measurement and significance in critically ill patients. J Trauma 1970;10(6):440–449.
6. Noble WH, Sieniewicz DJ. Radiological changes in controlled hypervolaemic pulmonary oedema in dogs. Can Anaesth Soc J 1975;22(2):171–185.
7. Snashall PD, Keyes SJ, Morgan BM, et al. The radiographic detection of acute pulmonary oedema. A comparison of radiographic appearances, densitometry and lung water in dogs. Br J Radiol 1981;54(640):277–288.
8. Gheorghiade M, Follath F, Ponikowski P, et al. Assessing and grading congestion in acute heart failure: a scientific statement from the acute heart failure committee of the heart failure association of the European Society of Cardiology and endorsed by the European Society of Intensive Care Medicine. Eur J Heart Fail 2010;12(5):423–433.
9. Hunt SA, Abraham WT, Chin MH, et al. 2009 Focused update incorporated into the ACC/AHA 2005 Guidelines for the Diagnosis and Management of Heart Failure in Adults: A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines Developed in Collaboration With the International Society for Heart and Lung Transplantation. J Am Coll Cardiol 2009;53(15):e1–e90.
10. Adams KF Jr, Fonarow GC, Emerman CL, et al. Characteristics and outcomes of patients hospitalized for heart failure in the United States: rationale, design, and preliminary observations from the first 100,000 cases in the Acute Decompensated Heart Failure National Registry (ADHERE). Am Heart J 2005;149(2):209–216.
11. Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, June 20–25, 2009. Piscataway, NJ: IEEE, 2009; 248–255.
12. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. CoRR. 2019; abs/1901.07042. [preprint] https://arxiv.org/abs/1901.07042. Posted January 21, 2019. Accessed October 2019
13. Johnson A, Pollard T, Mark R, Berkowitz S, Horng S. The MIMIC-CXR Database. https://physionet.org/content/mimic-cxr/2.0.0/. Published September 19, 2019. Accessed October 2019.
14. Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101(23):E215–E220.
15. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019;6(1):317.
16. Zhao CY, Xu-Wilson M, Gangireddy SR, Horng S. Predicting Disposition Decision, Mortality, and Readmission for Acute Heart Failure Patients in the Emergency Department Using Vital Sign, Laboratory, Echocardiographic, and Other Clinical Data. Circulation 2018;138(Suppl_1):A14287. https://www.ahajournals.org/doi/10.1161/circ.138.suppl_1.14287.
17. Liao R, Rubin J, Lam G, et al. Semi-supervised learning for quantification of pulmonary edema in chest x-ray images. arXiv arXiv:1902.10785. [preprint] https://arxiv.org/abs/1902.10785. Posted February 27, 2019. Accessed October 2019.
18. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, July 21–26, 2017. Piscataway, NJ: IEEE; 2017; 2261–2269.
19. Wang X, Schwab E, Rubin J, et al. Pulmonary Edema Severity Estimation in Chest Radiographs Using Deep Learning. https://www.semanticscholar.org/paper/Pulmonary-Edema-Severity-Estimation-in-Chest-Using-Wang-Schwab/2c21aac9e5758236aadec72713c1beff41fe5d75. Published 2019. Accessed October 2019.
20. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, October 22–29, 2017. Piscataway, NJ: IEEE; 2017; 618–626.
21. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. 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), Honolulu, HI, July 21–26, 2017. Piscataway, NJ: IEEE; 2017; 3462–3471.
22. Dunnmon JA, Yi D, Langlotz CP, Ré C, Rubin DL, Lungren MP. Assessment of convolutional neural networks for automated classification of chest radiographs. Radiology 2019;290(2):537–544.
23. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv. [preprint] https://arxiv.org/abs/1711.05225. Posted November 14, 2017. Accessed October 2019.
24. Majkowska A, Mittal S, Steiner DF, et al. Chest Radiograph Interpretation with Deep Learning Models: Assessment with Radiologist-adjudicated Reference Standards and Population-adjusted Evaluation. Radiology 2020;294(2):421–431.
25. Seah JCY, Tang JSN, Kitchen A, Gaillard F, Dixon AF. Chest radiographs in congestive heart failure: visualizing neural network learning. Radiology 2019;290(2):514–522.
26. Hammon M, Dankerl P, Voit-Höhne HL, et al. Improving diagnostic accuracy in assessing pulmonary edema on bedside chest radiographs using a standardized scoring approach. BMC Anesthesiol 2014;14(1):94.