Simulating Deployment of a Multi-{Modal, Task, Site} Chest X-Ray Model

Cara Van Uden
Stanford University
cvanuden@stanford.edu

Anuj Pareek
Stanford Univeristy
anujpare@stanford.edu

Andrew Gaut
Stanford Univeristy
agaut@stanford.edu

Abstract

With access to a multi-task chest X-ray (CXR) model deployed in the Intermountain healthcare system, we iterate on this model with new data and a multi-step pretraining method (multimodal contrastive self-supervised, followed by supervised). Our model achieves equal or better performance than supervised and image-only contrastive pretraining methods on three tasks: pneumonia (0.880 AUROC), multilobar pneumonia (0.883 AUROC), and pleural effusion (0.959 AUROC) in a set of real-world CXRs from multiple sites across the Intermountain healthcare system. We build a framework for retraining our model on a regular monthly cadence and, for a model "deployed" in 2020, simulate this retraining through 2021. We see meaningful improvement in model performance due to retraining, including an increase of 0.024 (from 0.881 to 0.905) for pneumonia AUROC on a 2021 test set. Finally, we analyze performance of our model using class activation maps (CAM) and confusion matrices.

1. Introduction

Pneumonia is a major cause of morbidity and mortality worldwide. Pneumonia diagnosis relies on a combination of clinical symptoms, blood tests, and chest X-rays (CXRs) read by radiologists [15]. Deep learning models able to quickly assess CXRs for pneumonia could reduce radiologist workload and improve patient treatment.

We previously worked with Intermountain Healthcare to deploy a deep learning model called CheXED [12], which is able to identify radiographic pneumonia and pleural effusions on CXRs. CheXED is a multi-task DenseNet pretrained on the Stanford CheXpert dataset and fine-tuned on an Intermountain dataset of CXR studies from over 6,500 patients collected between 2009 and 2015 [11]. This model was deployed at Intermountain emergency departments in 2020. In 2020 and 2021, we received over 10,000 further labeled CXR studies from Intermountain. In this work, we update and extend the original CheXED model in several ways.

To be clear on our task setup, the input to our model is an image of a patient CXR. We then use a Dense Convolutional Network (DenseNet121), which is the backbone of CheXED, in a multi-task setup to output a predicted probability of 1) pneumonia, 2) multilobar pneumonia (pneumonia in multiple parts of the lungs), and 3) pleural effusion (fluid build-up between the lungs and chest cavity).

There are several notable contributions in this work:

1. We update the previously deployed CheXED model with different pretraining methods, including novel self-supervised learning (SSL) and multimodal methods, and demonstrate improved performance across multiple tasks and multiple care sites.

2. We simulate CheXED model retraining at a regular cadence and demonstrate consistent model performance despite data drift.

3. We perform a detailed error analysis and demographic subgroup analysis of our model to prepare for deployment.

2. Related Work

Several prior studies have investigated the utility of deep learning for interpreting CXRs. In 2017, Rajpurkar et al. [18] trained a DenseNet [9] and achieved pneumonia-classification performance on par with radiologists. This landmark paper inspired others sto attempt to automate CXR interpretation using supervised deep learning on large labeled datasets [13,17,21]. Later, a large dataset of labelled CXRs called CheXpert was released along with a model which was the first to leverage uncertain labels and achieved state-of-the-art performance on several disease classification tasks [11]. The CheXpert dataset is used to pre-train deep learning-based CXR models such as CheXED [12], the deployed CXR model that we are investigating in this paper.

In addition to supervised pretraining, SSL pretraining has recently become popular for image tasks. After SSL pre-training methods like SimCLR [1], MoCo [8] and

BYOL [7] demonstrated success, researchers started investigating SSL pre-training for CXR models. One study showed that momentum contrastive pre-training resulted in significantly better pleural effusion classification performance measured by AUROC when compared to ImageNet pre-training [19]. In another related work [20], researchers cleverly leveraged metadata such as patient unique ID, CXR view, and study number to create additional pairs for contrastive pre-training and demonstrated further gains in classification performance.

Moving beyond image-only SSL, recent work such as CLIP [16], ConVIRT [23], and GLoRIA [10] have all focused on multimodal SSL. In particular, GLoRIA was able to achieve a significant performance jump, from an AUROC of 0.744 to 0.866 when comparing ImageNet pre-trained models to GloRIA models on CheXpert, using only 1% of CheXpert as labeled fine-tuning data.

3. Methods

The original CheXED is a 121-layer Densely Connected Convolutional Network architecture [12]. In this study, we continue to use this DenseNet121 architecture for our CNN backbone. All DenseNet121 models used in this study were first pretrained on ImageNet [4] before any further pretraining steps.

Note that we built directly off of a codebase from the Stanford Machine Learning Group when running our experiments. This codebase provides much of the basic functionality seen in CheXED [12]. However, the code and experiments for CXR preprocessing, pretraining method comparison, simulated retraining, model scale-up ablation, error analysis, demographic subgroup statistics and analyses are all novel to this work. All code for this project is available here.

3.1. Pretraining

3.1.1 Supervised

As our first supervised pretraining model, we simply used a DenseNet121 pretrained with supervision on ImageNet, which is readily available through PyTorch's torchvision model zoo [14]. The weights of all of the following models were initialized from this baseline ImageNet-pretrained model.

As our second supervised pretraining model, we used a DenseNet121 model pretrained with supervision on CheXpert. In the CheXpert pretraining task, the DenseNet121 model is trained to classify the absence or presence of 14 observations (including pneumonia and pleural effusion) on the CheXpert data set [11] containing more than 200,000 radiographs from Stanford Medical Center patients. This is the pretraining method used in the original CheXED model that is currently deployed.

As our third supervised pretraining model, we used a DenseNet121 model pretrained with supervision on a large set of publically available CXRs from CheXpert (Stanford), the RSNA Pneumonia Challenge, NIH chest X-ray8, PadChest (University of Alicante), and MIMIC-CXR (MIT). This pretrained model is available via torchxrayvision's model zoo [2, 3].

3.1.2 MoCo: Contrastive Image-Image

MoCo, or Momentum Contrast, is a type of self-supervised learning (SSL) [8]. Specifically, MoCo is a contrastive learning approach whose pretext task is maximizing agreement between different views of the same image (positive pairs) and to minimize agreement between different images (negative pairs). In MoCo, the model learns to encode the "query" image as similar to a moving queue of "key" images that are encoded with a learned "momentum encoder".

We build directly off work done by MoCo-CXR [19], who used MoCo pretraining to achieve improved performance after finetuning for tasks across two CXR datasets. For data augmentation, the authors of MoCo-CXR used random rotation (10 degrees) and horizontal flipping. This is a set of augmentations commonly used in training CXR models [11, 18]. As a side note, the original authors of MoCo-CXR chose to use MoCo instead of SimCLR [1] because MoCo requires far smaller batch sizes during pretraining than contrastive methods like SimCLR, making it computationally tractable to train on a single GPU [1].

We use the MoCo-CXR team's DenseNet121 model pretrained with MoCo on CheXpert. This model is available via MoCo-CXR.

3.1.3 GLoRIA: Contrastive Image-Text

Global-Local Representations for Images using Attention (GLoRIA) is a multimodal image-text model that we use to leverage radiologist reports that accompany CXRs [10]. Like MoCo, GLoRIA is a contrastive pretraining method, but it is multimodal: it pairs images and text instead of images and images. This model learns both global and local representations of its input by contrasting image representations (attention-weighted regions in CXR images) and text representations (words in the accompanying radiology reports). The authors of GLoRIA find that their method is highly performant and label-efficient for various downstream CXR classification tasks. For our downstream transfer learning, we use the learned weights from the global image encoder to initialize our DenseNet121.

For GLoRIA pretraining, we used hyperparameters that were previously found to result in the best downstream transfer task performance after pretraining [10]. This includes using a batch size of 48 to fit on a single GPU, an Adam optimizer with weight decay of 1e-6 and betas=(0.5,

0.999), a starting learning rate of 5e-5 that decreases when validation loss has plateaued for more than 5 epochs (patience of 5), and early stopping upon plateaued validation loss (patience of 10 epochs and maximum of 50 training epochs). We performed data augmentation during pretraining: random image crops to size 224x224 and image normalization with mean=(0.5, 0.5, 0.5) and std=(0.5, 0.5, 0.5). Specifically, this model was trained using PyTorch Lightning [6].

We worked with the GLoRIA team to pretrain a DenseNet121 with GLoRIA on CheXpert. This model is now publically available, along with a ResNet18 and ResNet50 pretrained on CheXpert, in the GLoRIA model zoo.

3.1.4 GLoRIA++: Contrastive Image-Text and Supervised

The Intermountain set of tasks is very similar to the CheXpert set of tasks. Specifically, both pneumonia and pleural effusion are CheXpert tasks, and multilobar pneumonia is closely related to the baseline pneumonia task. To take advantage of this similarity, we further fine-tune the GLoRIA model with supervision on 10% of CheXpert, using the same setup as the baseline model pretrained with supervision on CheXpert. For brevity, we call this pretraining method GLoRIA++. To summarize, GLoRIA++ has the following pretraining steps: 1) pretrain with supervision on ImageNet, 2) pretrain with GLoRIA on 100% of CheXpert, 3) pretrain with supervision on 10% of CheXpert. In future work, we plan to test model performance after pretraining with supervision on 1% and 100% of CheXpert as well.

For supervised pretraining using the GLoRIA pretrained model, we again used hyperparameters that were previously found to result in the best downstream transfer task performance [10]. This includes using BinaryCrossEntropy (BCE) loss as our loss function, a batch size of 64 to fit on a single GPU, an Adam optimizer with weight decay of 1e-6 and betas=(0.5, 0.999), a starting learning rate of 1e-4 that decreases when validation loss has plateaued for more than 5 epochs (patience of 5), and early stopping upon plateaued validation loss (patience of 10 epochs and maximum of 50 training epochs). We performed data augmentation during pretraining: random image crops to size 224x224 and image normalization with mean=(0.5, 0.5, 0.5) and std=(0.5, 0.5, 0.5). Specifically, this model was trained using PyTorch Lightning [6].

This model is now publically available in the GLoRIA model zoo.

3.2. Fine-tuning with CheXED

The network was trained to classify a chest radiographic study as (1) negative, uncertain, or positive for radiographic pneumonia; (2) unilobar or multilobar for the possible pneumonia studies; and (3) negative or positive for pleural effusion.

Statistics	Train	Validation	Test
CXR Studies, N	15774	467	467
Years, N (%)			
2009-2015	6551 (41.5)	-	-
2020	5490 (34.8)	-	-
2021	3733 (23.7)	467 (100)	467 (100)
Medical Facility			
Emergency	11081 (70.2)	467 (100)	467 (100)
Urgent Care	4693 (29.8)	-	-
Labels, N (%)			
Pneumonia			
Positive	5372 (34.1)	70 (15.0)	71 (15.2)
Uncertain	3872 (24.5)	86 (18.4)	82 (17.6)
Negative	6530 (41.4)	311 (66.6)	314 (67.2)
No. of Lobes			
Single Lobe	5381 (58.2)	42 (26.9)	48 (31.4)
Multilobar	3773 (40.8)	107 (68.6)	98 (64.1)
Unknown	90 (1.0)	7 (4.5)	7 (4.6)
Pleural Effusion			
Positive	1390 (8.8)	24 (5.1)	20 (4.3)
Negative	14384 (91.2)	443 (94.9)	447 (95.7)

Table 1. The table displays statistics and distribution of ground truth labels for the training, validation and test data. The training data was a mixture of data collected in 2009-15, 2020 and 2021, while the validation and test data was all from 2021.

To fine-tune the pretrained models for the Intermountain dataset, we split our dataset into a time-based train/valid/test split by year. To use binary classification metrics, the radiologist labels were binarized such that all unlikely and uncertain-unlikely cases were considered negative and all likely and uncertain-likely cases were considered positive. The CheXED operating points for each finding were set at the equal error rate thresholds, corresponding to the thresholds which led to equal false-positive and false-negative rates on the validation set.

Since our multi-task setup covers 3 tasks (1 with 3 classes, and 2 with 2 classes), we use a loss function specific to our setup, which for brevity we call "Intermountain loss". We modified the final fully connected layer of the DenseNet121 architecture to produce a 5-dimensional output, where 3 elements are used for pneumonia and the final 2 are used for pleural effusion and multilobar pneumonia. The networks were optimized to minimize a sum of 3 cross entropy losses (3-class cross entropy for pneumonia when using the 3-class model, binary cross entropy in all other cases). The loss for multilobar is only computed when the label for pneumonia is positive.

For supervised pretraining using the various pretrained

models, we used simple hyperparameters that were previously found to result in the best downstream transfer task performance [12]. This includes using the "Intermountain loss" described above as our loss function, a batch size of 32 to fit on a single GPU, an Adam optimizer with weight decay of 0 and betas=(0.9, 0.999), a starting constant learning rate of 1e-4, and training for 5 epochs. We performed data augmentation during training: resizing images to 320 for the shorter side of the image, random image crops to size 320x320, and image normalization with ImageNet's mean and standard deviation.

To generate a prediction for a new study, CheXED was run on all available views (frontal and lateral) in the study and the maximum probability for each finding was taken as the predicted output for the whole study. The model was developed using PyTorch v1.10.0 [14] and all code is available here.

To examine the effect of dataset size on performance after fine-tuning for each of our pretraining methods, we also perform finetuning on 1, 5, 10, and 50% of the full dataset for each pretraining method. Fine-tuning on small datasets can be unstable; results may change dramatically given a new split of data [22]. To account for this, we run 10 "replicates" for each dataset size, where each replicate will train and evaluate on a randomly sampled training set and the same validation set.

3.3. Simulated Retraining

To achieve better performance and avoid data drift over time, we simulate monthly retraining/fine-tuning of our deployed model.[1] For each month, we split that month's CXRs into a time-based train/valid/test split of 80%/10%/10%. We will deal with three models: A, the model deployed at the beginning of 2021 with no fine-tuning on any 2021 data; B, the model deployed at the beginning of the month with no fine-tuning on the current month's train set; and C, which is model B after fine-tuning on the current month's train set. Then, for each of the models A, B, and C, we will evaluate their performance on this month's test set. We repeat this for every month for which we have data (September and October).

3.4. Error Analysis

For our 2021 test set, we generate class activation maps (CAM) for each CXR using the approach described in [24]. We then split these CAMs into true positives, true negatives, false positives, and false negatives. A radiology resident (and co-author of this work!) checked these CAMs to find meaningful patterns in model outputs. We also investigate the confusion matrices for the currently deployed model and our updated models to investigate data drift across time (by seeing if thresholds calculated on previous data translate well to newer data) and ensure that our classification threshold is calibrated correctly for the updated models.

[1] Originally, we had planned to perform this study with retraining for every month, every three months, and every six months; however, all the emergency department data we have from 2021 is from September and October (with 992 datapoints from September and 3676 from October). Therefore, we were limited to performing this study with simulated retraining each month for those two months only.

3.5. Demographic Subgroup Analysis

We have demographic data for all of our 2021 patients. In this work, we investigate our model's performance for the pneumonia task (the main task used in the Intermountain healthcare system from this model) on different demographic subgroups of age, race, ethnicity, sex, insurance payer type, and location of hospital care. As a first step for this, we calculate the model's AUROC, AUPRC, and F1-score when tested on every population within each demographic subgroup type.

4. Dataset

In this work, we train, validate, and test on a set of 16,708 CXRs from several emergency department and urgent care sites in Intermountain's healthcare system. In particular, we later specifically analyze model performance across the five emergency department sites. We handle four separate datasets from Intermountain. These four datasets were collected between 2009 and 2021 and contain frontal-view and lateral-view CXRs and associated labels for pneumonia, pleural effusion, and number of lobes affected by pneumonia. For all of our data from 2021, we additionally have demographic data for each patient and CheXED predicted labels for pneumonia, pleural effusion, and number of lobes affected by pneumonia. The datasets are summarized in Table 1.

To preprocess this data, we first converted all CXR views from dicoms to images. We performed minimal data preprocessing, and leveraged radiologist expertise, to map labels across all of our datasets to the same values for CheXED. As an example, some datasets used "yes" to indicate a patient was positive for pneumonia, while others used "probable" to indicate the same. We performed this data preprocessing for all four input datasets and all three tasks. We then matched each CXR to its associated metadata and labels using its accession number.

When finetuning our models, we split the Intermountain dataset into a time-based train/valid/test split by year, and by percentages within years (train set was 100% of data from '09-'20 and 80% of data from 2021, valid set was 10% of data from 2021, and test set was 10% of data from 2021. This resulted in a split of 15,774/467/467 CXR studies. We trained on emergency department and urgent care data (72% of studies were from the emergency department) from '09-'21 and evaluated on emergency department data from 2021.

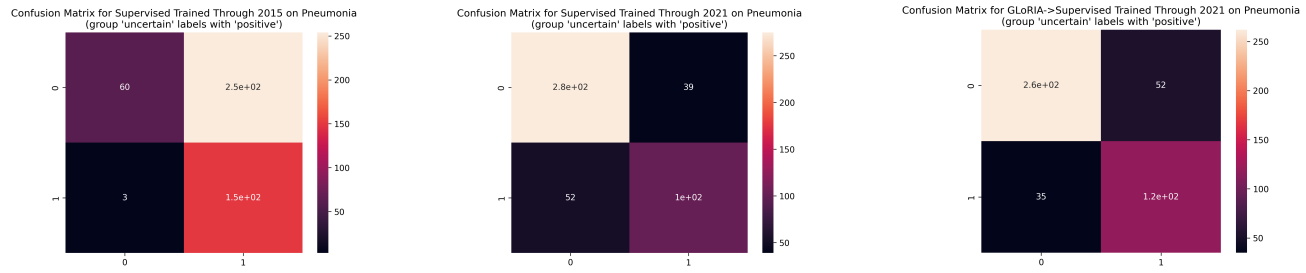

Figure 1. Confusion matrices for pneumonia, with y-axis showing ground truth and x-axis showing model predictions. The left uses the currently deployed model (trained and validated on data through 2015, tested on data from 2021), the middle uses the same pretraining method as the currently deployed model, but with updated data (trained and validated on data through 2021, tested on data from 2021), and the right uses the updated GLoRIA++ model (trained and validated on data through 2021, tested on data from 2021).

5. Results and Discussion

Our primary metrics are Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision Recall Curve (AUPRC). We use these because they are less affected by class imbalance than metrics like accuracy, which is important for our highly imbalanced dataset (see Table 1). We additionally chose these metrics because they are operating point threshold-agnostic. We combine these threshold-agnostic metrics with the F1-score metric (a metric that nicely summarizes precision and recall but depends on the model's threshold between class 0 and 1) to more closely examine our model's performance at the chosen threshold (we describe how we choose this threshold in Section 3.2). We also gather and briefly discuss precision and recall metrics at this chosen operating point threshold.

5.1. Comparing Pretraining Methods for Fine-tuning with CheXED

The AUPRC and AUROC attained by each of our models on the three tasks can be found in Table 2. We find that GLoRIA++ is consistently either the best model (pneumonia) or a close second-best model (second to GLoRIA for number of lobes affected by pneumonia, and second to the currently deployed CheXED model for pleural effusion, which we comment on in Appendix Section C) across the three tasks. The ROC and PRC curves for GLoRIA++ for the pneumonia task are in Figure 2.

In particular, we want to highlight the tradeoffs in task performance between GLoRIA and GLoRIA++. Remember that GLoRIA++ builds off of GLoRIA and only has additional supervised pretraining. We notice that GLoRIA++ performs better than GLoRIA on the two Intermountain tasks that are shared with CheXpert - pleural effusion and pneumonia/consolidation. However, GLoRIA performs better than GLoRIA++ on multilobar pneumonia, which is not a task explicitly included in CheXpert. This supports our in initial reasoning behind GLoRIA++: GLoRIA appears to learn a more general representation of CXRs, while GLoRIA++ learns slightly more task specificity. For this specific application, we want to leverage the task similarity between CheXpert and Intermountain. However, there are also clear use cases, such as model development for any of the wide range of tasks that are not included in CheXpert, for leveraging a purely self-supervised approach such as GLoRIA.

5.2. Simulated Retraining

We find that retraining every month results in performance improvements in an end-of-retraining-period test set for pneumonia, but slight decreases in performance for multilobar pneumonia and pleural effusion (Table 3). Performance increase may be due to data drift caused by an uptick of Covid-19 cases in 2021, which the re-trained model is robust to. This result suggests future models may be able to be retrained regularly (e.g. monthly) and achieve good performance despite data drift.

5.3. Error Analysis

A 3rd year radiology resident (AP) provided qualitative error analysis for models by reviewing CAMs across true positives, true negatives, false positives, and false negatives as classified by the GLoRIA models. We find meaningful patterns in the model's mistakes (see Figure 3).

False positives were most often a result of other chest diseases or chest radiographic features that look similar to pneumonia. For example, we found that pleural calcifications can lead to a false positive classification (Figure 3). In other cases, false positives were a result of complicated chest X-rays acquired with suboptimal patient positioning, often with several overlapping diseases such as pleural effusion and cardiomegaly. *False negatives* were most often a result of the pneumonia only having very subtle radiographic features; i.e. the disease was difficult to detect even for a radiologist. In other cases, obscuring diseases or medical devices such as chest tubes and pacemakers could also lead to false negatives. For example, we found that

Task	Pretrain Method	Pretrain Dataset	Train Dataset	AUROC	AUPRC	F1 Score
Pneumonia	Supervised	CheXpert	'09-'15	0.801	0.690	0.665
	Supervised	ImageNet	'09-'21	0.860	0.779	0.713
	Supervised	CheXpert	'09-'21	0.861	0.783	0.706
	Supervised	All CXRs	'09-'21	0.852	0.766	0.700
	MoCo	CheXpert	'09-'21	0.863	0.790	0.709
	GLoRIA	CheXpert	'09-'21	0.863	0.795	0.722
	GLoRIA++	CheXpert	'09-'21	**0.880**	**0.810**	**0.738**
Effusion	Supervised	CheXpert	'09-'15	**0.961**	**0.696**	**0.780**
	Supervised	ImageNet	'09-'21	0.954	0.597	0.634
	Supervised	CheXpert	'09-'21	0.953	0.596	0.732
	Supervised	All CXRs	'09-'21	0.926	0.466	0.595
	MoCo	CheXpert	'09-'21	0.947	0.540	0.600
	GLoRIA	CheXpert	'09-'21	0.955	0.589	0.708
	GLoRIA++	CheXpert	'09-'21	0.959	0.647	0.683
NumLobes	Supervised	CheXpert	'09-'15	0.865	0.938	0.851
	Supervised	ImageNet	'09-'21	0.873	0.943	0.869
	Supervised	CheXpert	'09-'21	0.881	0.945	0.869
	Supervised	All CXRs	'09-'21	0.866	0.945	0.867
	MoCo	CheXpert	'09-'21	0.874	0.946	0.866
	GLoRIA	CheXpert	'09-'21	**0.900**	**0.955**	**0.894**
	GLoRIA++	CheXpert	'09-'21	0.883	0.949	0.876

Table 2. Evaluation results for all three tasks using the DenseNet121 model evaluated on a 2021 test set for all pretraining methods and pretraining datasets, after finetuning on data from either '09-'15 (currently deployed model) or '09-'21 (proposed new models). Within each task, top scores are bolded and second place scores are underlined. Scores for pneumonia and multilobar pneumonia are computed after pneumonia "uncertain" labels are grouped with positive labels.

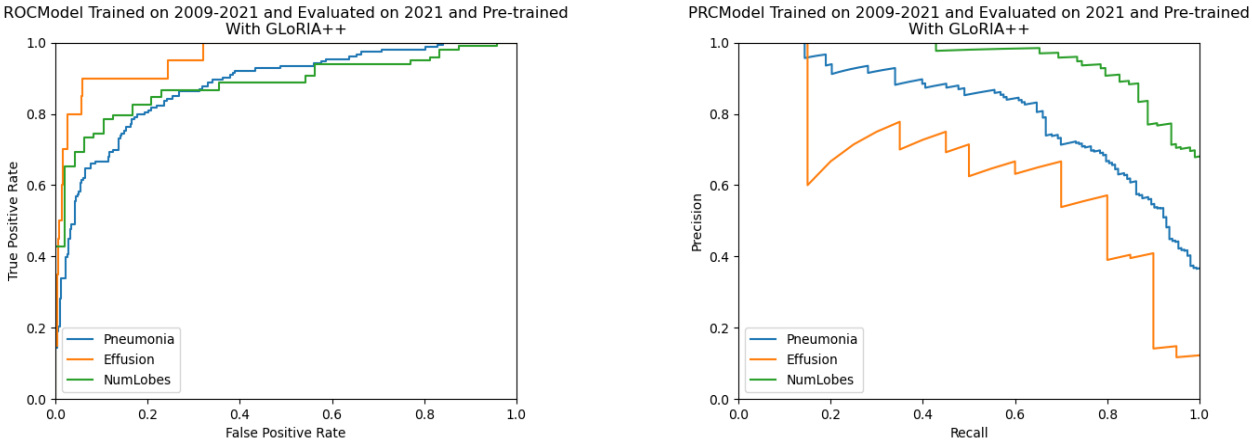

Figure 2. ROC and PRC curves for the DenseNet121 CheXpert-GLoRIA++ model after evaluation on the test set for 2021. Labels of "uncertain" for pneumonia are grouped with positive labels.

pulmonary atelectasis (which can hide pneumonia features) lead to a false negative classification (Figure 3).

We additionally investigated CAMs for hard examples of chest X-rays. We define hard examples as cases where there was disagreement between the original supervised CheXED model and the attending radiologist. The CAMs show that there is a clear tendency for the original CheXED to be "overly sensitive" and make false positive mistakes if the chest X-ray is slighty different from the standard normal pattern (see Figure 4). For example, CAMs from the original CheXED model highlights vessel crowding as pneumonia. Similarly medical devices such as chest tubes and cardiac electrodes are easily mistaken for pneumonia by the original CheXED model (Figure 4). However, the GloRIA++ model is more robust to such small deviations in chest X-rays and does not incorrectly classify such examples as pneumonia. Furthermore, the CAMs do not excessively highlight any focal areas in these hard examples. The confusion matrices also show that going from original CheXED to GloRIA++ greatly reduces the number of false

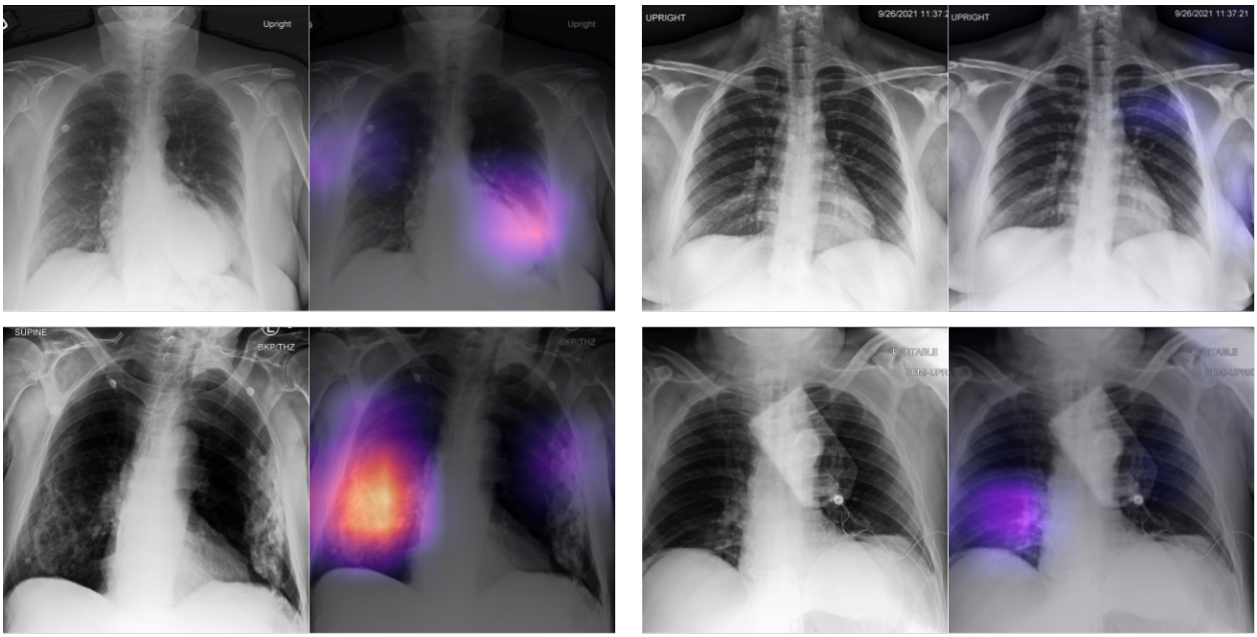

Figure 3. The figure shows selected TP, TN, FP and FN examples of class activation maps for pneumonia using the updated GLoRIA++ model. *Top Left*: True positive. There is a silhouette sign at the left cardiac border as a result of pulmonary consolidation in the left lung lingula segment. The model classifies the example correctly as pneumonia, and the CAMs highlight the abnormality correctly. *Top Right*: True negative. The lung fields are clear bilaterally with no sign of pneumonia. The CAMs do not highlight any focal area consistent with a true negative example. *Bottom Left*: False positive. Extensive pleural calcifications are seen in lung fields bilaterally. The CAMs show that the model incorrectly classifies these calcifications as pneumonia. *Bottom Right*: False negative. There is atelectasis of the left lower lobe with a faint air bronchogram consistent with consolidating pneumonia. The CAMs do not highlight the area with pneumonia, possibly because the model was unable to locate pneumonia-like features within the atelectic lung tissue.

Task	SRT?	Train Dataset	AUROC	AUPRC	F1 Score
Pn.	No	'09-'15	0.827	0.701	0.678
	No	'09-'20	0.881	0.768	0.726
	Yes	'09-'20	**0.906**	**0.802**	**0.780**
Eff.	No	'09-'15	**0.987**	**0.817**	0.744
	No	'09-'20	**0.987**	0.776	**0.813**
	Yes	'09-'20	0.982	0.712	0.757
M.L.	No	'09-'15	0.864	**0.949**	0.870
	No	'09-'20	**0.886**	0.885	**0.886**
	Yes	'09-'20	0.864	0.945	0.871

Table 3. Evaluation results on a 2021 test set for the three tasks (Pneumonia, Effusion, and MultiLobar) and three models: the model currently deployed, the "best" model we found on training through 2020, and this "best" model after simulated retraining through 2021.

positives.

We also looked at confusion matrices on the 2021 test set for the currently deployed model, a model using the same pretraining method as the currently deployed model, but with updated data, and the new GLoRIA++ model. We find an apparent distribution shift in pneumonia since 2015 (the currently deployed model was trained on data through 2015). The currently deployed model has an threshold for classification that results in most examples being classified in the positive class. Specifically, the precision at this threshold is 0.371, while the recall at this threshold is 0.980. Retraining the same model with new data (precision at threshold of 0.721 and recall at threshold of 0.660) and training a GLoRIA++ model with the new data (precision at threshold of 0.694 and recall at threshold of 0.771) both result in an improvement.

6. Conclusion and Future Work

In this work, we develop a multi-task CXR model that leverages multimodal and supervised pretraining to predict pneumonia, multilobar pneumonia, and pleural effusion with performance that closely matches or beats both a currently deployed CXR model. Crucially, we find that our model performs equivalently well for pneumonia across different emergency department care sites. We simulate what deployment and retraining/finetuning of a model trained on data through 2020 would have looked like during 2021 and show that pneumonia prediction performance improves with finetuning on a regular monthly cadence. Finally, we find meaningful patterns in our model's false nega-

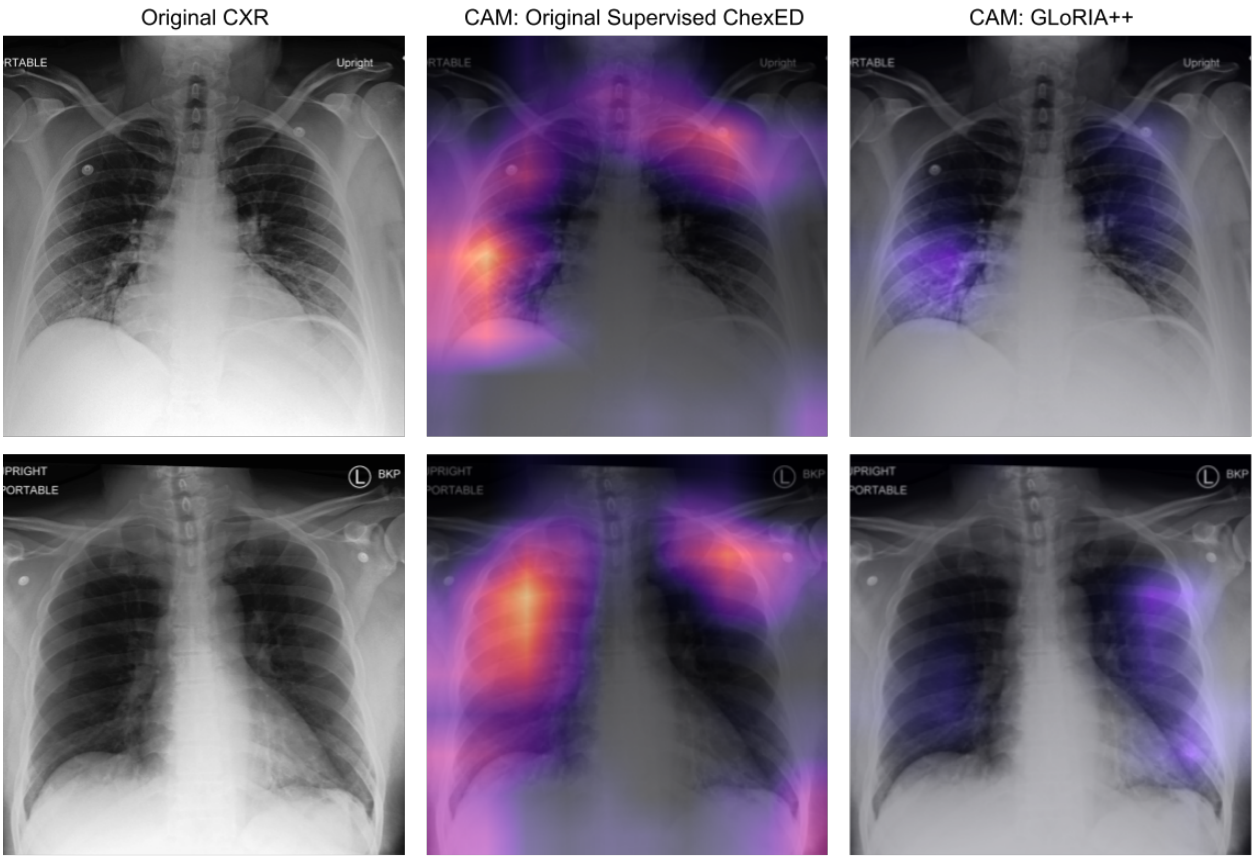

Figure 4. The figure shows examples of hard pneumonia cases which were classified incorrectly by the original CheXED, but classified correctly by GLoRIA++. *Top row*: False positive. A normal chest X-ray acquired suboptimally in the expiratory breathing phase which leads to vessel crowding in basal lung fields. CAMs from original CheXED show that this vessel crowding in the right lung resulted in an incorrect classification of pneumonia, but the GLoRIA++ model correctly classified the chest X-ray as negative for pneumonia and CAMs did not excessively highlight any focal lung areas. *Bottom row*: False positive. A completely normal chest X-ray but with cardiac electrodes on the shoulders. The CAMs from the orignal CheXED highlight the cardiac electrodes which are mistaken for pneumonia. The GLoRIA++ model does not excessively highlight any areas on this normal chest X-ray and classifies it correctly as negative for pneumonia.

tives, false positives, and "hard examples". We analyze our model's performance across demographic subgroups and notice performance gaps (see E); mitigating this is a major focus of our future work before model deployment.

Before deployment, we have a few steps for future work:

1. We want to further explore the task-representation-specificity relationship between GLoRIA and GLoRIA++. As a first step, we plan to pretrain additonal GLoRIA++ models on 1% and 100% of CheXpert, in addition to the 10% model described here.

2. We have about 20,000 additional Intermountain CXRs that lack labels but do have radiology reports. We plan to pretrain a new GLoRIA and GLoRIA++ model using 1) CheXpert and 2) this CXR-report data, and then finetune on our most recent Intermountain CXR data.

3. We recently received an additional set of 4,500 labeled CXRs for Intermountain urgent care sites. We plan to re-do the core experiments using this new data and select an updated deployment candidate model.

4. We want to extend the "scale-up" ablations performed in D and make recommendations to our Intermountain collaborators about how many CXR studies they should regularly annotate for model retraining.

5. Mitigating performance gaps (summarized in E) across demographic subgroups. We are actively investigating this and will address it before model deployment.

After deployment in Intermountain emergency department and urgent care sites, we plan to run a clinical trial of this model.

A. Contributions & Acknowledgements

Note that we built directly off of a codebase from the Stanford Machine Learning Group when running our experiments. This codebase provides much of the basic functionality seen in CheXED [12]. However, the code and experiments for CXR preprocessing, pretraining method comparison, simulated retraining, model scale-up ablation, error analysis, demographic subgroup statistics and analyses are all novel to this work. All code for this project is available here.

We use pretrained model weights for ImageNet-supervised, CheXpert-supervised, CXR-supervised, and CheXpert-MoCo; we link those in the relevant Methods section. We additionally pretrain and release two new models in collaboration with Mars Huang from the GLoRIA team, which are publically available here.

A few specific thank yous: to Mars Huang for the original GLoRIA codebase, help with training the GLoRIA models, and for uploading our models to the GLoRIA model zoo; to Jeremy Irvin for providing the CheXED repository that we started with and which had lots of core functionality for us to use; and to Jason Carr, Nat Dean, and other folks at Intermountain Healthcare for their amazing medical expertise and CXR dataset.

CVU preprocessed CXR data, wrangled supervised (ImageNet, CheXpert, and "all CXR") and MoCo pretrained models, worked with GLoRIA team to pretrain DenseNet121 GLoRIA model (previous GLoRIA models only used ResNet), pretrained DenseNet121 GLoRIA++ model, set up and ran CheXED fine-tuning experiments across pretraining methods, set up and ran model "scale-up" ablation experiments, generated confusion matrices and CAMs, performed demographic subgroup analyses, and contributed to relevant sections of the paper and poster.

AP generated data statistics and demographics for the patient cohort, made tables, wrote relevant sections of the paper and poster including the *Related Work* section, made the CAMs investigation, made the CAMs figures, error analysis, analysis of model prediction, analysis of medical reasons for possible data distribution shifts and aided in clinical motivation, and radiological and medical questions relevant to the paper.

AG wrote code for simulated retraining, conducted the simulated retraining experiments, aggregated the results, wrote code for and generated all the ROC and PRC figures, formatted and helped put together tables 4-7, and contributed to other relevant sections of the poster and paper.

B. How to Handle Uncertain Labels: Discarding or Grouping?

We want to highlight that we do see comparatively lower performance across the board for pneumonia and multilobar pneumonia - for AUROC, AUPRC, and F1-scores - as compared to other previously-reported CheXpert-pretrained models (and especially CheXED) [12]. This can be partially attributed to how we binarize our "uncertain" labels for pneumonia - previous work such as CheXED ignored these uncertain labels when generating metrics, but here we group these uncertain labels with the positive class. This also affects multilobar pneumonia evaluation, as a CXR is only evaluated for multilobar pneumonia if it is ground-truth positive for pneumonia. This is due to agreement between us and our Intermountain collaborators that this method for handling "uncertains" is more clinically- and deployment-relevant. Note that the previous method, though easier than our task, is relevant. It is an excellent way to ensure that the model is correctly and confidently classifying the CXR studies radiologists are also most certain about. In light of this, we also present the equivalent of Table 2 for this other method; these results for pneumonia and multilobar pneumonia can be found in Table 4.

C. Pleural Effusion and Differences Between Site Types

We see that the original deployed CheXED model performs slightly better on the pleural effusion task than GloRIA++ and the other models trained on data through 2021. A likely explanation for this slight performance drop (or rather, a failure of the new models to easily outperform the older model) is due to a distribution shift in training data. The original CheXED was trained on 2009-2015 data, which contained only emergency department chest X-rays which are almost exclusively supine Anterior-Posterior (AP) chest X-rays. Detecting smaller pleural effusions on supine chest X-rays is harder [5] because the fluid redistributes to the dependent parts of the pleural cavity. The fluid then is no longer visible as an opaque "meniscus" adjacent to the diaphragm as in erect PA chest X-rays. Since the older model was trained on the harder supine examples of pleural effusion, it performs better on the test set, compared to the GloRIA++ model which included 29.8% urgent care clinic training data with easier upright PA chest X-rays.

D. "Scale-up" Ablations

We also "scaled up" models using each pretraining method on 1, 5, 10, and 100% of the Intermountain finetuning CXRs (see Figure 5). AUROC and AUPRC on all tasks for supervised-, GLoRIA-, and GLoRIA++-CheXpert pretraining is consistently higher than AUROC and AUPRC of models with other pretraining methods on ImageNet, CheXpert, and a superset of CXRs. Noteably, across most pretraining methods (but most noticeably supervised, GLoRIA, and GLoRIA++), we approach "100%"-data performance after finetuning on 5-10% of available Intermoun-

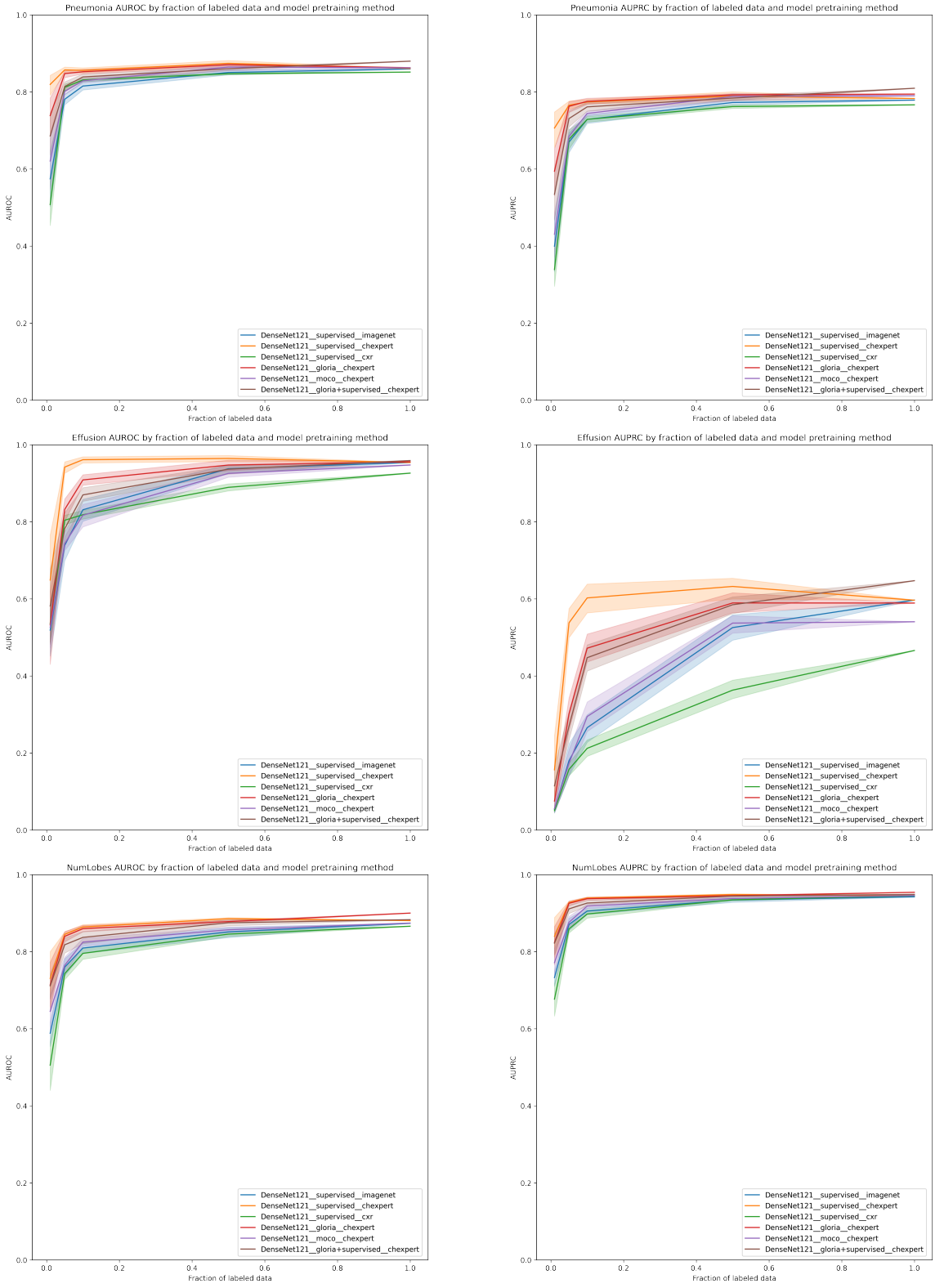

Figure 5. AUROC and AUPRC for the pneumonia, pleural effusion, and number of affected lobes tasks, for each of the foundation models pretrained with different methods and trained on different percentages of the Intermountain CXR data. 95% confidence intervals are generated from 10 replicates of each model and data subset. Labels of ”uncertain” for pneumonia are grouped with positive labels.

Task	Pretrain Method	Pretrain Dataset	Train Dataset	AUROC	AUPRC	F1 Score
Pneumonia	Supervised	CheXpert	'09-'15	0.945	0.785	0.773
	Supervised	ImageNet	'09-'21	0.960	0.869	0.839
	Supervised	CheXpert	'09-'21	0.966	0.884	0.848
	Supervised	All CXRs	'09-'21	0.941	0.837	0.847
	MoCo	CheXpert	'09-'21	**0.971**	0.895	**0.875**
	GLoRIA	CheXpert	'09-'21	0.968	**0.900**	0.850
	GLoRIA++	CheXpert	'09-'21	0.967	0.894	0.857
Effusion	Supervised	CheXpert	'09-'15	**0.961**	**0.696**	**0.780**
	Supervised	ImageNet	'09-'21	0.954	0.597	0.634
	Supervised	CheXpert	'09-'21	0.953	0.596	0.732
	Supervised	All CXRs	'09-'21	0.926	0.466	0.595
	MoCo	CheXpert	'09-'21	0.947	0.540	0.600
	GLoRIA	CheXpert	'09-'21	0.955	0.589	0.708
	GLoRIA++	CheXpert	'09-'21	0.959	0.647	0.683
NumLobes	Supervised	CheXpert	'09-'15	0.902	0.977	0.941
	Supervised	ImageNet	'09-'21	0.908	0.978	0.950
	Supervised	CheXpert	'09-'21	0.898	0.976	0.934
	Supervised	All CXRs	'09-'21	**0.938**	**0.988**	0.941
	MoCo	CheXpert	'09-'21	0.927	0.984	0.949
	GLoRIA	CheXpert	'09-'21	0.937	0.984	**0.966**
	GLoRIA++	CheXpert	'09-'21	0.934	0.985	0.958

Table 4. Evaluation results for all three tasks using the DenseNet121 model evaluated on a 2021 test set for all pretraining methods and pretraining datasets, after finetuning on data from either '09-'15 (currently deployed model) or '09-'21 (proposed new models). Within each task, top scores are bolded and second place scores are underlined. Scores for pneumonia and multilobar pneumonia are computed after pneumonia "uncertain" labels are ignored.

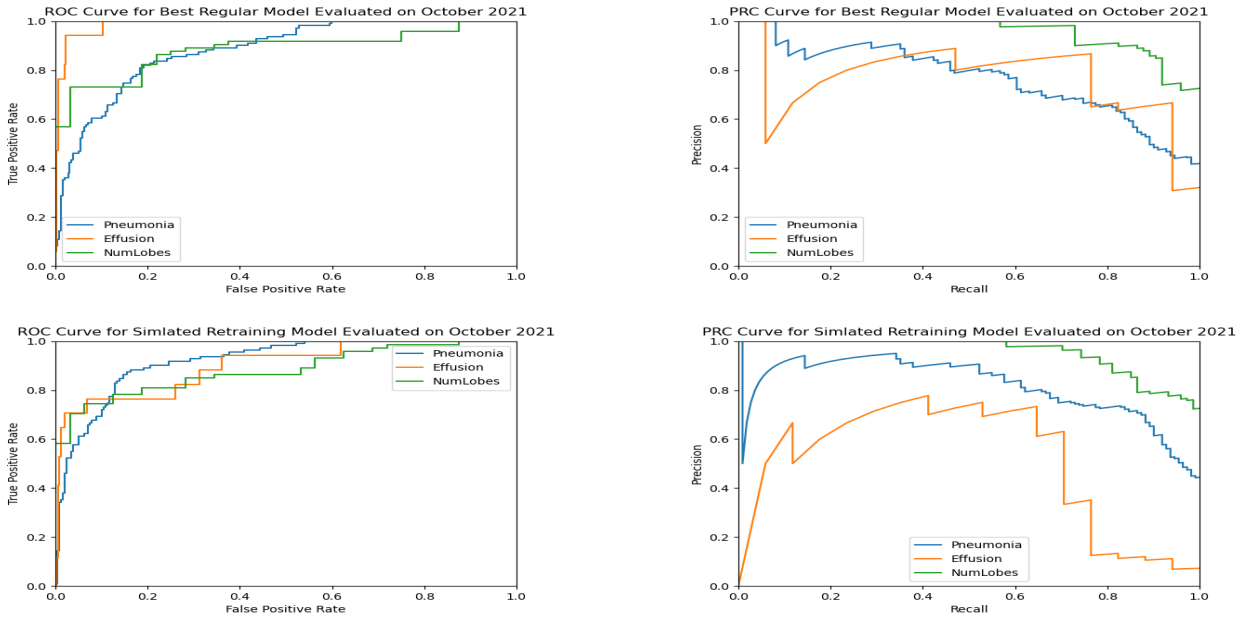

Figure 6. ROC and PRC curves for the best model trained on data up until 2020 (DenseNet121 GLoRIA++-pretrained on CheXpert) (top) and this model after simulated retraining (bottom), both evaluated on the test set for simulated retraining for October 2021. Labels of "uncertain" for pneumonia are grouped with positive labels.

tain CXR data. Combined with the results from simulated retraining, this can help inform data gathering and labeling in the future; for instance, we could ask our Intermountain collaborators to annotate just 5-10% of CXR studies and retrain on these studies each month.

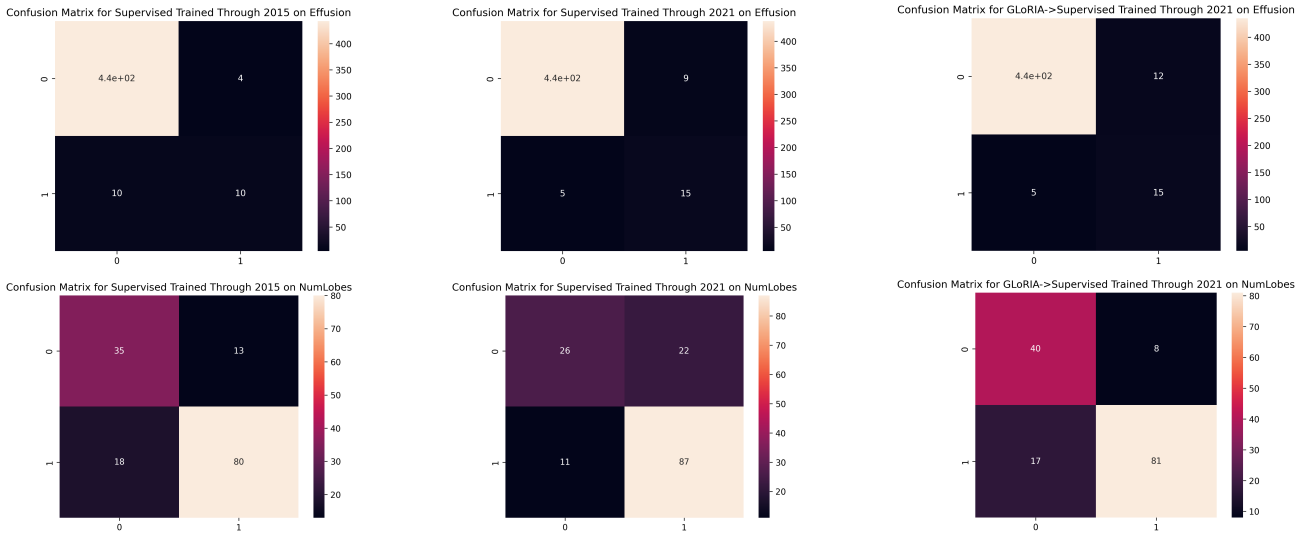

Figure 7. Confusion matrices for effusion and multilobar pneumonia, with y-axis showing ground truth and x-axis showing model predictions. The left uses the currently deployed model (trained and validated on data through 2015, tested on data from 2021), the middle uses the same pretraining method as the currently deployed model, but with updated data (trained and validated on data through 2021, tested on data from 2021), and the right uses the new GLoRIA++ model (trained and validated on data through 2021, tested on data from 2021).

Statistics	**2021 Data**
Patients, N	4651
Age, mean ± SD	50.8 ±22.8
Female, N (%)	2418 (52.0)
Race, N (%)	
White	3835 (82.5)
Hawaiian or Pacific	144 (3.0)
Black or African American	107 (2.3)
Asian	64 (1.4)
Other or Unavailable	501 (10.8)

Table 5. Demographic information was only available for the data collected from emergency departments in 2021. The table displays key demographic statistics for the 2021 patient population.

E. Demographic Subgroup Analysis

Finally, we analyze our model's performance across different demographic subgroups, mostly focusing on AUROC. These results are summarized for pneumonia in Table 6. We find that our model performs equivalently well across the five different emergency department care sites. We see drops (or inconsistencies, such as 1.0 AUROC) in performance for different race demographic subgroups, particularly for Black or African American, Pacific Islander or Native Hawaiian, Asian, and Native American subgroups. In our dataset, 83% of our patients are white; this demographic imbalance likely contributes to our performance gap. Additionally, we see drops in performance for patients with Medicare or who paid for care themselves. We see slight inconsistencies in performance by age, ethnicity, and sex (see Table 6 for details)[2]. Note that we could not gener-

Demographic	**Subgroup Name**	**AUROC**	**AUPRC**	**F1-Score**
Race	White	0.873	0.801	0.715
	Haw./Pac. Isl.	0.741	0.767	0.800
	Black	0.833	0.900	0.667
	Asian	1.000	1.000	0.889
	Nat. Am.	1.000	1.000	1.000
Latino/Spanish	No	0.868	0.79	0.717
	Yes	0.901	0.88	0.767
Gender	Male	0.887	0.865	0.773
	Female	0.868	0.735	0.680
Age	<19	1.0	1.0	0.5
	19-45	0.924	0.894	0.786
	45-65	0.856	0.855	0.744
	65-80	0.795	0.703	0.632
	80+	0.841	0.801	0.790
Insurance Plan	Medicare	0.795	0.757	0.676
	Medicaid	0.891	0.777	0.681
	Tricare	0.933	0.927	0.750
	Other Ins.	0.944	0.896	0.821
	Self Pay	0.756	0.719	0.750
Hospital Name	IM	0.866	0.799	0.773
	MK	0.898	0.866	0.743
	RV	0.891	0.809	0.681
	AV	0.852	0.694	0.710
	LD	0.910	0.831	0.571

Table 6. Performance on pneumonia task for different types of demographic subgroups.

[2]Please note that we received only male/female labels for sex, and we received no information about patient gender.

ate similar metrics for the other two tasks as there were not always patients from a demographic subgroup in both the negative and positive classes, particularly for pleural effusion. Before we deploy this model, analyzing and fixing these performance gaps in depth and gathering more data for these demographic subgroups is essential.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2

[2] Joseph Paul Cohen, Mohammad Hashir, Rupert Brooks, and Hadrien Bertrand. On the limits of cross-domain generalization in automated x-ray prediction. In *Medical Imaging with Deep Learning*, 2020. 2

[3] Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. TorchXRayVision: A library of chest X-ray datasets and models. *https://github.com/mlmed/torchxrayvision*, 2020. 2

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[5] SA Emamian, M-A Kaasbøl, JF Olsen, and JF Pedersen. Accuracy of the diagnosis of pleural effusion on supine chest x-ray. *European radiology*, 7(1):57–60, 1997. 9

[6] William Falcon and The PyTorch Lightning team. PyTorch Lightning, 3 2019. 3

[7] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 2

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1, 2

[9] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1

[10] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021. 2, 3

[11] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019. 1, 2

[12] Jeremy A Irvin, Anuj Pareek, Jin Long, Pranav Rajpurkar, David Ken-Ming Eng, Nishith Khandwala, Peter J Haug, Al Jephson, Karen E Conner, Benjamin H Gordon, et al. Chexed: Comparison of a deep learning model to a clinical decision support system for pneumonia in the emergency

department. *Journal of thoracic imaging*, 37(3):162–167, 2022. 1, 2, 4, 9

[13] Paras Lakhani and Baskaran Sundaram. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2):574–582, 2017. 1

[14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 2, 4

[15] Elena Prina, Otavio T Ranzani, and Antoni Torres. Community-acquired pneumonia. *The Lancet*, 386(9998):1097–1108, 2015. 1

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and et al. Learning transferable visual models from natural language supervision. *arXiv.org*, Feb 2021. 2

[17] Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, 15(11):e1002686, 2018. 1

[18] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017. 1, 2

[19] Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning*, pages 728–744. PMLR, 2021. 2

[20] Yen Nhi Truong Vu, Richard Wang, Niranjan Balachandar, Can Liu, Andrew Y Ng, and Pranav Rajpurkar. Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation. In *Machine Learning for Healthcare Conference*, pages 755–769. PMLR, 2021. 2

[21] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018. 1

[22] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. Revisiting few-sample bert fine-tuning. *arXiv preprint arXiv:2006.05987*, 2020. 4

[23] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv.org*, Oct 2020. 2

[24] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 4