MedAug: Contrastive learning leveraging patient metadata improves representations for chest X-ray interpretation

Yen Nhi Truong Vu[1]* NTRUONGV@STANFORD.EDU
Richard Wang[1]* RCMWANG@STANFORD.EDU
Niranjan Balachandar[2]* NIRANJA9@STANFORD.EDU
Can Liu[1] CANLIU@STANFORD.EDU
Andrew Y. Ng[1] ANG@CS.STANFORD.EDU
Pranav Rajpurkar[1] PRANAVSR@CS.STANFORD.EDU

*Equal Contribution
[1] *Department of Computer Science, Stanford University*
[2] *School of Medicine, Stanford University*

Abstract

Self-supervised contrastive learning between pairs of multiple views of the same image has been shown to successfully leverage unlabeled data to produce meaningful visual representations for both natural and medical images. However, there has been limited work on determining how to select pairs for medical images, where availability of patient metadata can be leveraged to improve representations. In this work, we develop a method to select positive pairs coming from views of possibly different images through the use of patient metadata. We compare strategies for selecting positive pairs for chest X-ray interpretation including requiring them to be from the same patient, imaging study or laterality. We evaluate downstream task performance by fine-tuning the linear layer on 1% of the labeled dataset for pleural effusion classification. Our best performing positive pair selection strategy, which involves using images from the same patient from the same study across all lateralities, achieves a performance increase of 14.4% in mean AUC from the ImageNet pretrained baseline. Our controlled experiments show that the keys to improving downstream performance on disease classification are (1) using patient metadata to appropriately create positive pairs from different images with the same underlying pathologies, and (2) maximizing the number of different images used in query pairing. In addition, we explore leveraging patient metadata to select hard negative pairs for contrastive learning, but do not find improvement over baselines that do not use metadata. Our method is broadly applicable to medical image interpretation and allows flexibility for incorporating medical insights in choosing pairs for contrastive learning.

1. Introduction

Self-supervised contrastive learning has recently made significant strides in enabling the learning of meaningful visual representations through unlabeled data (Wu et al., 2018b; Hjelm et al., 2018; Chen et al., 2020b,a). In medical imaging, previous work has found

performance improvement when applying contrastive learning to chest X-ray interpretation (Sowrirajan et al., 2020; Sriram et al., 2021; Azizi et al., 2021), dermatology classification (Azizi et al. (2021)) and MRI segmentation (Chaitanya et al., 2020). Despite the early success of these applications, there is limited work on determining how to improve upon standard contrastive algorithms using medical information (Sowrirajan et al., 2020; Chaitanya et al., 2020; Zhang et al., 2020; Kiyasseh et al., 2020).

In contrastive learning, the selection of pairs controls the information contained in learned representations, as the loss function dictates that representations of positive pairs are pulled together while those of negative pairs are pushed apart (Oord et al., 2018). For natural images where there are no other types of annotations, positive pairs are created using different augmented views of the same image while negative pairs are views of different images (Chen et al., 2020a). Tian et al. (2020) argue that good positive pairs are those that contain minimal mutual information apart from common downstream task information. In the natural image setting, Tamkin et al. (2020) train a generative model which learns to produce multiple positive views from a single input. However, previous contrastive learning studies on medical imaging have not systematically investigated how to leverage patient metadata available in medical imaging datasets to select positive pairs that go beyond augmentations of the same image.

In this work, we propose a method to treat different images that share common properties found in patient metadata as positive pairs in the context of contrastive learning. We demonstrate the application of this method to a chest X-ray interpretation task. Similar to the concurrent work by Azizi et al. (2021), we experiment with requiring positive pairs to come from the same patient as these images likely share highly similar pathological features. However, our method incorporates these positive pairs with possibly different images directly as part of the view generation scheme in a single contrastive pretraining stage, as opposed to Azizi et al. (2021), which adds a second pretraining stage where a positive pair must be formed by two distinct images. Further, we go beyond the simple strategy of forming a positive pair using any two data points coming from the same patient as in Azizi et al. (2021) and Kiyasseh et al. (2020) and experiment with other metadata such as study number and laterality to identify a pair of images that are likely to have the same pathologies. Although study number has also been leveraged successfully in Sriram et al. (2021) to create a sequence of pretrained embeddings representating patient disease progression, our work differs in that we use this information specifically to choose positive pairs during the contrastive pretraining stage.

We conduct MoCo-pretraining (Chen et al., 2020b) using these different criteria and evaluate the quality of pretrained representations by freezing the base model and fine-tuning a linear layer using 1% of the labeled dataset for the task of pleural effusion classification. Our contributions are:

1. We develop a method, *MedAug*, to use patient metadata to select positive pairs in contrastive learning, and apply this method to chest X-rays for the downstream task of pleural effusion classification.

2. Our best pretrained representation achieves a performance increase of 14.4% in mean AUC compared to the ImageNet pretrained baseline, showing that using patient metadata to select positive pairs from multiple images can notably improve representations.

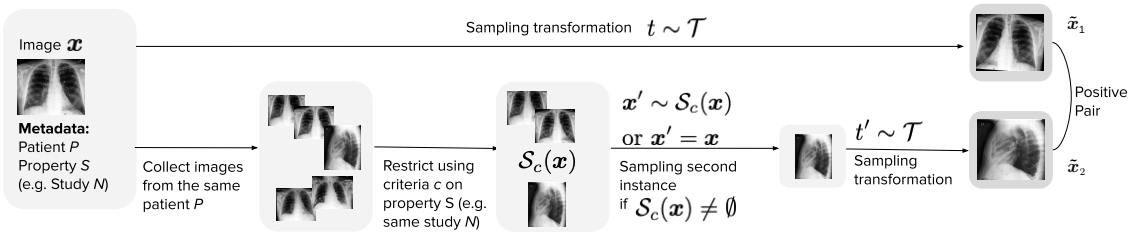

Figure 1: Selecting positive pairs for contrastive learning with patient metadata

3. We perform comparative empirical analysis with label information and show that (1) using positive pairs of different images from the same patients that share underlying pathologies improves pretrained representations, and (2) increasing the number of distinct images selected to form positive pairs per image query improves the quality of pretrained representations.

4. We perform an exploratory analysis of strategies to select negative pairs using patient metadata, and do not find improvement over the default strategy that does not use metadata.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our work presents methods to improve contrastive pretraining for medical image interpretation. We evaluate our methods on the task of chest X-ray classification, but our approach is generalizable to other medical imaging tasks. First, we demonstrate that leveraging patient metadata – specifically information about the patient, study, and laterality of the images – to identify pathological similarity and define positive pairs for contrastive learning improves pretrained representations. This can be applicable to tasks in medical image interpretation where such metadata is available. Second, we show that using two distinct images as positive pairs for contrastive learning provides better pretrained representations than only using two augmentations of the same image. Therefore, other contrastive learning methods for medical imaging may benefit from expanding the construction of positive pairs by using patient metadata to inform this expansion. Third, our analysis indicates that increasing the number of images available for selecting positive pairs improves pretrained representations. In general, in contrastive learning for medical imaging, using domain-specific information to increase the pool of images to select positive pairs may improve pretrained representations and downstream task performance.

2. Cohort

We use CheXpert, a large collection of de-identified chest X-ray images (Irvin et al., 2019). The dataset consists of 224,316 images from 65,240 patients labeled for the presence or absence of 14 radiological observations. We use these images for pretraining and random samples of 1% of these images for fine-tuning. The test set consists of 500 additional

labeled images from 500 studies not included in the training set. We perform fine-tuning experiments on the downstream task of pleural effusion classification, which was selected based on quality of ground truth labels, clinical importance and prevalence. We evaluate model performance using test set AUC after fine-tuning.

3. Methods

3.1. Selecting positive pairs for contrastive learning with patient metadata

Given an input image $\boldsymbol{x}$, encoder g, and a set of augmentations $\mathcal{T}$, most contrastive learning algorithms involve minimizing the InfoNCE loss

$$\mathcal{L}(\boldsymbol{x}) = -\log \frac{\exp[g(\tilde{\boldsymbol{x}}_1) \cdot g(\tilde{\boldsymbol{x}}_2)]}{\exp[g(\tilde{\boldsymbol{x}}_1) \cdot g(\tilde{\boldsymbol{x}}_2)] + \sum_{i=1}^{K} \exp[g(\tilde{\boldsymbol{x}}_1) \cdot g(\boldsymbol{z}_i)]}. \tag{1}$$

Here, the positive pair $(\tilde{\boldsymbol{x}}_1 = t_1(\boldsymbol{x}), \tilde{\boldsymbol{x}}_2 = t_2(\boldsymbol{x}))$ with $t_1, t_2 \in \mathcal{T}$ are augmentations of the input image $\boldsymbol{x}$, while the negative pairs $(\tilde{\boldsymbol{x}}_1, \boldsymbol{z}_i)$, $1 \leq i \leq K$ are pairs of augmentations of different images, with $\boldsymbol{z}_i$ coming from either a queue in the case of MoCo or the minibatch in the case of SimCLR. Recognizing that many augmentation strategies available for natural images are not applicable to medical images, Sowrirajan et al. (2020) restrict $\mathcal{T}$ to be the set of simple augmentations such as horizontal flipping and random rotation between -10 to 10 degrees. As a result, their method can be thought of as instance discrimination, as $\tilde{\boldsymbol{x}}_1$ and $\tilde{\boldsymbol{x}}_2$ must come from the same image input.

In this work, we propose MedAug, a method to use multiple images as a way to increase the number of positive pair choices. Beyond the disease labels, we can use patient metadata such as patient number, study number, laterality, patient historical record etc. to create appropriate positive pairs. Formally, we can use patient metadata to obtain an enhanced augmentation set dependent on $\boldsymbol{x}$ as follows

$$\mathcal{T}_{\text{enhanced}}(\boldsymbol{x}) = \begin{cases} \{t_i(\boldsymbol{x}') | t_i \in \mathcal{T}, \boldsymbol{x}' \in \mathcal{S}_c(\boldsymbol{x})\} & \text{if } \mathcal{S}_c(\boldsymbol{x}) \neq \emptyset \\ \mathcal{T}(\boldsymbol{x}) & \text{otherwise} \end{cases} \tag{2}$$

where $\mathcal{S}_c(\boldsymbol{x})$ is the set of all images satisfying some predefined criteria c in relation to the properties of $\boldsymbol{x}$. The criteria for using the metadata could be informed by clinical insights about the downstream task of interest.

We apply this method on chest X-ray interpretation and pretrain ResNet-18 models using MoCo v2 (Chen et al., 2020b) with learning rate of 10^{-4}, batch size of 16, temperature of 0.2, MLP projection, and 20 epochs. Since the downstream task is disease classification, we experiment with using $\mathcal{S}_{\text{same patient}}(\boldsymbol{x})$ since images from the same patient are likely to share high amount of pathological features. We also experiment with further applying criteria on study numbers as well as laterality. An example application of the method is illustrated in Figure 1. Given an image query, the method collects all images from the same patient as the query, chooses a subset of the collection using the given criteria, then sample an image from the subset. Finally, an augmentation of the image together with the augmentation of an initial query image form a positive pair.

3.2. Fine-tuning and evaluation

We evaluate the pretrained representations by (1) training a linear classifier on outputs of the frozen encoder using labeled data and (2) end-to-end fine-tuning. Pretrained checkpoints are selected with k-nearest neighbors algorithm (Wu et al., 2018a) based on Faiss similarity search and clustering library (Johnson et al., 2017). To simulate label scarcity encountered in medical contexts, we fine-tune using only 1% of the labeled dataset. The fine-tuning experiments are repeated on 5 randomly drawn 1% splits from the labeled dataset to provide an understanding of the model's performance variance. We report the mean AUC and standard deviation over these five 1% fine-tuning splits. Following Irvin et al. (2019), we use a learning rate of 3×10^{-5}, batch size of 16 and 95 epochs for training.

4. Experiments

4.1. Positive pair selection

Our formulation of using any set of images $\mathcal{S}_c(\boldsymbol{x})$ from the same patient to enhance the set of augmentations for contrastive learning provides the flexibility of experimenting with different criteria c for constraining $\mathcal{S}_c(\boldsymbol{x})$. We experiment with limiting $\mathcal{S}_c(\boldsymbol{x})$ using properties found in the metadata of the query $\boldsymbol{x}$. In particular, we focus on two properties:

Study number. The study number of an image associated with a particular patient reflects the session in which the image was taken. We experiment with three different criteria on study number:

1. All studies: no restriction on $\mathcal{S}_{\text{all studies}}(\boldsymbol{x})$ is dependent on the study number of $\boldsymbol{x}$
2. Same study: only images from the same study with $\boldsymbol{x}$ belong to $\mathcal{S}_{\text{same study}}(\boldsymbol{x})$
3. Distinct studies: only images with different study number from $\boldsymbol{x}$ belong to $\mathcal{S}_{\text{distinct studies}}(\boldsymbol{x})$

Laterality. Chest X-rays can be of either frontal (AP/PA) view or lateral view.

1. All lateralities: no restriction on $\mathcal{S}_{\text{all lateralities}}(\boldsymbol{x})$ is dependent on the laterality of $\boldsymbol{x}$
2. Same laterality: only images from the same laterality with $\boldsymbol{x}$ belong to $\mathcal{S}_{\text{same laterality}}(\boldsymbol{x})$
3. Distinct lateralities: only images with a different laterality from that of $\boldsymbol{x}$ belongs to $\mathcal{S}_{\text{distinct lateralities}}(\boldsymbol{x})$

Results. We report the results of experiments using these criteria in Table 1. Except from when $\mathcal{S}_c(\boldsymbol{x})$ includes images with different study numbers from $\boldsymbol{x}$, where there is a drop in performance, we see consistent large improvement from the baseline in Sowrirajan et al. (2020). The best result is obtained when using $\mathcal{S}_{\text{same study, all lateralities}}(\boldsymbol{x})$, the set of images from the same patient and same study as that of $\boldsymbol{x}$, regardless of laterality. Incorporating this augmentation strategy while holding other settings from Sowrirajan et al. (2020) constant results in respective gains of 0.029 (3.4%) and 0.021 (2.4%) in AUC for the linear model and end-to-end model. We also experiment with including random crop augmentation from MoCo v2 (Chen et al., 2020b), where the scaling is modified to be $[0.95, 1.0]$ in

Table 1: Except for criteria c that involve images from different studies, using images from the same patient to select positive pairs result in improved AUC in downstream pleural effusion classification.

Baseline models	**Linear**	**End-to-end**
ImageNet baseline	0.766 ± 0.009	0.858 ± 0.011
MoCo v2 baseline (Sowrirajan et al., 2020)	0.847 ± 0.007	0.881 ± 0.017
MoCo v2 baseline with random crop scale	0.864 ± 0.005	0.890 ± 0.026
Criteria c for creating $\mathcal{S}_c(\boldsymbol{x})$	**Linear**	**End-to-end**
Same patient, same study, same laterality	0.862 ± 0.004	0.894 ± 0.013
Same patient, same study, distinct lateralities	0.865 ± 0.008	0.897 ± 0.008
Same patient, same study	0.876 ± 0.013	0.902 ± 0.007
Same patient, all studies	0.859 ± 0.006	0.877 ± 0.012
Same patient, distinct studies	0.848 ± 0.007	0.874 ± 0.013
Same patient, same study with random crop scale	$\mathbf{0.883 \pm 0.005}$	$\mathbf{0.906 \pm 0.015}$

order to avoid cropping out areas of interest in the lungs. Adding this augmentation to the same patient, same study strategy, we obtain our best pretrained model, which achieves a linear fine-tuning AUC of 0.883 and an end-to-end fine-tuning AUC of 0.906 on the test set, significantly outperforming previous baselines. Results for repeated experiments across other CheXpert competition tasks are included in Appendix A.

4.2. Comparative Empirical Analysis

We perform comparative analysis with labels to understand how different criteria on patient metadata affect downstream performance results seen in Table 1.

4.2.1. All studies v.s. same study

We hypothesize that the drop in transfer performance when moving from using images with the same study number to using images regardless of study number is because $\mathcal{S}_{\text{all studies}}(\boldsymbol{x})$ may contain images of a different disease pathology than that seen in $\boldsymbol{x}$. As a result, the model is asked to push the representation of a diseased image close to the representation of a non-diseased image, causing poor downstream performance. To test this hypothesis, we carry out an oracle experiment with $\mathcal{S}_{\text{all studies, same label}}(\boldsymbol{x})$, the set of images from the same patient and with the same downstream label as that of $\boldsymbol{x}$, regardless of study number.

Results. Table 2 shows that the model pretrained with $\mathcal{S}_{\text{all studies, same label}}(\boldsymbol{x})$ achieves a respective improvement of 0.034 and 0.022 in AUC over $\mathcal{S}_{\text{all studies}}(\boldsymbol{x})$ strategy for the linear model and end-to-end model. This experiment supports our hypothesis that positive pairs from images with different downstream labels hurt performance.

Table 2: Experiment with and without using downstream labels shows that positive pairs with different labels hurt downstream classification performance.

Criteria c for creating $\mathcal{S}_c(\boldsymbol{x})$	**Linear**	**End-to-end**
Same patient, all studies	0.859 ± 0.006	0.877 ± 0.012
Same patient, all studies, same disease label as $\boldsymbol{x}$	$\mathbf{0.893 \pm 0.009}$	$\mathbf{0.899 \pm 0.010}$

4.2.2. All studies v.s. distinct studies

There is a further performance drop when moving from using images across all studies of the same patient to images with a different study number from the current query image (Table 1). This finding may also support our hypothesis because there is a larger proportion of positive pairs of different disease pathologies in pairs of images from strictly different studies. To make sure this result holds independently of the different number of available images to form pair per query, we repeated these experiments while forcing $|\mathcal{S}_{\text{same study, all lateralities}}(\boldsymbol{x})| = |\mathcal{S}_{\text{same study, same laterality}}(\boldsymbol{x})|$ via random subset pre-selection. Further, we only use distinct images as a pair, i.e. skipping any $\boldsymbol{x}$ with $\mathcal{S}_c(\boldsymbol{x}) = \emptyset$ in (2) in order to remove any possible contribution from positive pairs formed from the same image.

Results. Table 3 shows the same patient, all studies strategy (AUC = 0.848) outperforms the same patient, distinct studies strategy (AUC = 0.792) even when the size of $\mathcal{S}_c(\boldsymbol{x})$ is controlled. This supports the hypothesis that a higher proportion of positive pairs with different disease pathologies hurts downstream task performance. Downstream performance from the $\mathcal{S}_{\text{distinct studies}}(\boldsymbol{x})$ is likely lower than that of $\mathcal{S}_{\text{all studies}}(\boldsymbol{x})$ because there is a higher proportion of positive pairs with different disease labels in $\mathcal{S}_{\text{distinct studies}}(\boldsymbol{x})$. Figure 2 shows that there is almost 9% of $\boldsymbol{x}$ where $\mathcal{S}_{\text{distinct studies}}(\boldsymbol{x})$ contains only images with a different disease label from $\boldsymbol{x}$, whereas this scenario does not appear for $\mathcal{S}_{\text{all studies}}(\boldsymbol{x})$.

Table 3: Experiments where we force positive pairs to come from different images and control the size of $\mathcal{S}_c(\boldsymbol{x})$ shows that higher proportion of pairs with different downstream labels contribute to lower downstream performance.

Criteria c for creating $\mathcal{S}(\boldsymbol{x})$	**Linear**	**End-to-end**
Same patient, distinct studies	0.792 ± 0.007	0.841 ± 0.013
Same patient, all studies (size controlled)	$\mathbf{0.848 \pm 0.009}$	$\mathbf{0.863 \pm 0.010}$

4.2.3. All lateralities v.s. distinct lateralities v.s. same laterality

First, we hypothesize that the drop in performance from the all lateralities to the same laterality strategy could be due to $\mathcal{S}_{\text{same study, same laterality}}(\boldsymbol{x})$ having smaller size. To test this, we carry out an experiment in which $\mathcal{S}_{\text{same study, all lateralities}}(\boldsymbol{x})$ is constrained by

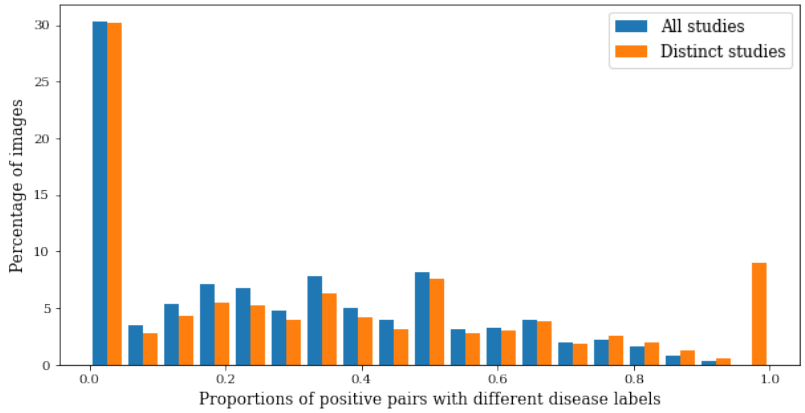

Figure 2: Histogram showing the distribution of the proportions of positive pairs with different disease labels in $\mathcal{S}_{\text{distinct studies}}(\boldsymbol{x})$ versus $\mathcal{S}_{\text{all studies}}(\boldsymbol{x})$.

$|\mathcal{S}_{\text{same study, same laterality}}(\boldsymbol{x})|$, the number of images from the same study and has the same laterality as $\boldsymbol{x}$.

Table 4: Experiments with all lateralities where we control the size of $S_{\text{same study, all lateralities}}$ show that the size of $\mathcal{S}_c(\boldsymbol{x})$ affects downstream performance.

Criteria c for creating $\mathcal{S}_c(\boldsymbol{x})$	**Linear**	**End-to-end**
Same patient, same study, same laterality	0.862 ± 0.004	0.894 ± 0.013
Same patient, same study, all lateralities (size controlled)	0.860 ± 0.004	0.899 ± 0.011
Same patient, same study, all lateralities (no control)	$\mathbf{0.876 \pm 0.013}$	$\mathbf{0.902 \pm 0.007}$

Table 5: Experiments to compare same v.s. distinct lateralities with size restriction on $\mathcal{S}_c(\boldsymbol{x})$ shows no significant difference.

Criteria c for creating $\mathcal{S}_c(\boldsymbol{x})$	**Linear**	**End-to-end**
Same patient, same study, same laterality	0.856 ± 0.016	0.878 ± 0.015
Same patient, same study, distinct lateralities	$\mathbf{0.866 \pm 0.015}$	$\mathbf{0.882 \pm 0.017}$

Our second hypothesis is that mutual information in images with different lateralities is lower, which benefits retaining only information important to the downstream task, as shown in Tian et al. (2020). We test this by training two models on images that include at least one counterpart from the other laterality. We pretrain one model with $\mathcal{S}_{\text{same study, same laterality}}(\boldsymbol{x})$ containing only images with the same laterality as $\boldsymbol{x}$, and

the other model with $\mathcal{S}_{\text{same study, distinct lateralities}}(\boldsymbol{x})$ containing only images with different laterality from $\boldsymbol{x}$. To prevent the effect of different sizes of $\mathcal{S}_c(\boldsymbol{x})$, we force that $|\mathcal{S}_{\text{same study, same laterality}}(\boldsymbol{x})| = |\mathcal{S}_{\text{same study, distinct lateralities}}(\boldsymbol{x})|$ via random subset pre-selection.

Results. Table 4 shows that once we control for the size of $\mathcal{S}_c(\boldsymbol{x})$, there is no significant difference between using images from the same laterality (AUC = 0.862) or from all lateralities (AUC = 0.860). However, the model pretraining with all images from all lateralities achieves much larger downstream AUC of 0.876. Thus, it supports our first hypothesis that the size of $\mathcal{S}_c(\boldsymbol{x})$ influences pretrained representation quality. Table 5 shows that once we control for the size of $\mathcal{S}_c(\boldsymbol{x})$, the model pretrained with images from different lateralities only gain 0.010 AUC in linear fine-tuning performance and a non-significant 0.004 in end-to-end performance. This experiment shows that the effect of mutual information from different lateralities on pretrained representation quality is less pronounced.

4.3. Negative pair selection

We explore strategies using metadata in the CheXpert dataset to define negative pairs. Similar to our method of defining positive pairs, we take advantage of metadata available in the dataset to select the negative pairs. However, unlike positive pair selection, where only a single pair is required for each image, an image has to pair with the entire queue to select negative pairs. This property makes selecting negative pairs from the same patient as done in selecting positive pairs not suitable because only a small number of images are available for a patient. We instead use a more general property – laterality – across the patients to define negative pairs to retain sufficient negative pairs in the loss function (1). Similarly, other metadata such as age and sex may be exploited for the same purpose.

The default negative pair selection strategy is to select all keys from the queue that are not views of the query image. However, we hypothesize that negative pairs with the same laterality are "hard" negative pairs that are more difficult to distinguish and provide more accurate pretrained representations for the downstream task. We describe our four strategies briefly as follows and in more detail in Appendix B. Our first strategy is to only select images from the queue with the same laterality as the query to create negative pairs. Our second strategy is to reweight the negative logits based on laterality so in effect queries with each laterality (frontal and lateral) equally contribute to the loss and the queue size remains fixed as in the original MoCo approach. Following a similar idea in Kalantidis et al. (2020), our third strategy is to sample a portion of negative pairs with the same laterality for each query and append them to the queue for loss computation. Our fourth strategy is to create synthetic negatives for additional hard negative pairs. Unlike Kalantidis et al. (2020), we do not determine hardness of negative pairs based on similarities of representations. Instead, we use existing metadata (image laterality) to approximate hardness of an negative pair. We evaluate the performance of each of these negative pair strategies combined with the positive pair strategy of "same patient, same study, all lateralities".

Results. Results are given in Table 6. The default negative pair selection strategy (AUC = 0.876) is not outperformed by any of the metadata-exploiting negative pair selection strategies including same laterality only (AUC = 0.872), same laterality reweighted (AUC = 0.864), same laterality appended (AUC = 0.875) and same laterality synthetic (AUC =

0.870). Thus, our exploratory analysis does not indicate sufficient evidence for performance improvement using strategies that incorporate metadata, but further experiments with other metadata sources may be required to further understand this relationship.

Table 6: Experiments with the default negative pair definition (different images) and various negative pair selection strategies.

Negative Pairs Strategy	**Linear**
Default	$\mathbf{0.876 \pm 0.013}$
Same Laterality only	0.872 ± 0.011
Same Laterality (reweighted)	0.864 ± 0.006
Same Laterality (appended)	0.875 ± 0.006
Same Laterality (synthetic)	0.870 ± 0.004

5. Discussion

We introduce MedAug, a method to use patient metadata to select positive pairs for contrastive learning, and demonstrate the utility of this method on a chest X-ray interpretation task.

Can we improve performance by leveraging metadata to choose positive pairs? Yes. Our best pretrained strategy with multiple images from the same patient and same study obtains an increase of 3.4% in linear fine-tuning AUC in comparison to the instance discrimination approach implemented in Sowrirajan et al. (2020). A similar result has been shown by Kiyasseh et al. (2020) for ECG signal interpretation. Azizi et al. (2021) also found improvement in dermatology classification when applying a second contrastive pretraining stage where strictly distinct images from the same patient are selected as positive pairs.

Unlike previous work, our empirical analysis on using images from all studies and distinct studies shows that simply choosing images from the same patient may hurt downstream performance. We show that using appropriate metadata such as study number to select positive pairs that share underlying disease information is needed to obtain the best representation for the downstream task of disease classification. For future studies, it is of interest to experiment with other metadata such as age group, medical history, etc. and how they can inform on tasks other than disease classification.

Our analysis using different criteria on laterality shows that the number of images selected to form positive pairs plays an important role, while the effect of mutual information is less clear. Given time and resources, it would be informative to experiment with how the maximum number of distinct images chosen per query affect downstream performance.

Can we improve performance by leveraging metadata to choose hard negative pairs? Not necessarily. We perform an exploratory analysis of strategies to leverage patient metadata to select negative pairs, and do not find them to outperform the baseline.

In closing, our work demonstrates the potential benefits of incorporating patient metadata into self-supervised contrastive learning for medical images, and can be extended to a broader set of tasks (Rajpurkar et al., 2020; Uyumazturk et al., 2019).

Limitations In the CheXpert dataset, any two images from the same patient and study will always have the same set of ground truth pathology labels. For medical image datasets with different constraints regarding patient metadata, it remains future work to determine the positive pair selection strategies that are clinically relevant and produce good pretrained representations. Our approach is not applicable to datasets lacking patient metadata altogether. For datasets with limited data per patient, future work could be to cluster data using the images and available metadata into larger groups, and define positive pairs based on cluster assignments. While the results show that our metadata-based contrastive learning methods are generalizable across all CheXpert competition tasks, it remains future work to assess performance in other datasets.

Furthermore, our strategies for negative pair selection did not improve pretrained representations. Our strategies leveraged information regarding image laterality. However, future work is required to whether negative pair selection strategies using other metadata such as image view (anteroposterior or posteroanterior), patient age or patient sex, or strategies using similarity metrics can improve negative pair selection.

References

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. *arXiv preprint arXiv:2101.05224*, 2021.

Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *arXiv preprint arXiv:2006.10511*, 2020.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020a.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

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.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning, 2020.

Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals. *arXiv preprint arXiv:2005.13249*, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Pranav Rajpurkar, Allison Park, Jeremy Irvin, Chris Chute, Michael Bereket, Domenico Mastrodicasa, Curtis P Langlotz, Matthew P Lungren, Andrew Y Ng, and Bhavik N Patel. Appendixnet: Deep learning for diagnosis of appendicitis from a small dataset of ct exams using video pretraining. *Scientific reports*, 10(1):1–7, 2020.

Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. *arXiv preprint arXiv:2010.05352*, 2020.

A Sriram, M Muckley, K Sinha, F Shamout, J Pineau, KJ Geras, L Azour, Y Aphinyanaphongs, N Yakubova, and W Moore. Covid-19 prognosis via self-supervised representation learning and multi-image prediction. 2021.

Alex Tamkin, Mike Wu, and Noah Goodman. Viewmaker networks: Learning views for unsupervised representation learning, 2020.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020.

Bora Uyumazturk, Amirhossein Kiani, Pranav Rajpurkar, Alex Wang, Robyn L. Ball, Rebecca Gao, Yifan Yu, Erik Jones, Curtis P. Langlotz, Brock Martin, Gerald J. Berry, Michael G. Ozawa, Florette K. Hazard, Ryanne A. Brown, Simon B. Chen, Mona Wood, Libby S. Allard, Lourdes Ylagan, Andrew Y. Ng, and Jeanne Shen. Deep learning for the digital pathologic diagnosis of cholangiocarcinoma and hepatocellular carcinoma: Evaluating the impact of a web-based diagnostic assistant, 2019.

Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination, 2018a.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018b.

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 preprint arXiv:2010.00747*, 2020.

Appendix A. Repetition Across Tasks

We repeated each linear fine-tuning experiment for the other 4 CheXpert competition classification tasks: Atelectasis, Cardiomegaly, Consolidation, and Edema. The results are given in Table 7. For all 5 CheXpert competition classification tasks (see Table 1 for Pleural Effusion), we found that our best metadata-based approach to contrastive learning, which was the "same patient, same study" positive pair selection criterion, resulted in higher downstream classification performance than did the ImageNet baseline and MoCo v2 baselines.

Table 7: Linear fine-tuning AUCs for baseline methods and positive pair selection criteria repeated across the tasks of Atelectasis, Cardiomegaly, Consolidation, and Edema classification.

Baseline models	**Atelectasis**	**Cardiomegaly**	**Consolidation**	**Edema**
ImageNet baseline	0.612	0.634	0.615	0.810
MoCo v2 baseline (Sowrirajan et al., 2020)	0.671	0.735	0.699	0.847
MoCo v2 baseline with random crop scale	0.708	0.730	0.693	0.858
Criteria c for creating $\mathcal{S}_c(\boldsymbol{x})$	**Atelectasis**	**Cardiomegaly**	**Consolidation**	**Edema**
Same patient, same study, same laterality	0.738	0.718	0.672	0.866
Same patient, same study, distinct lateralities	0.712	0.800	0.703	0.883
Same patient, same study	0.762	0.785	0.721	0.889
Same patient, all studies	0.727	0.786	0.805	0.879
Same patient, distinct studies	0.676	0.757	0.760	0.846
Same patient, same study with random crop scale	0.721	0.779	0.801	0.866

Appendix B. Negative Pairs

Following the loss function in equation (1), we denote the exponential sum of the negative pairs by G

$$\mathcal{L}(\boldsymbol{x}) = -\log \frac{\exp[g(\tilde{\boldsymbol{x}}_1) \cdot g(\tilde{\boldsymbol{x}}_2)]}{\exp[g(\tilde{\boldsymbol{x}}_1) \cdot g(\tilde{\boldsymbol{x}}_2)] + G\left(\tilde{\boldsymbol{x}}_1, \boldsymbol{z}_i\right)}. \tag{3}$$

where

$$G\left(\tilde{\boldsymbol{x}}_1, \boldsymbol{z}_i\right) = \sum_{\boldsymbol{z}_i \in Q} \exp[g(\tilde{\boldsymbol{x}}_1)) \cdot g(\boldsymbol{z}_i))] \tag{4}$$

We follow the MoCo setup and denote Q as the queue. Let $\mathcal{S}(\boldsymbol{x})$ be the set of image representations in Q that have the same laterality as $\boldsymbol{x}$. We use the symbol $\|$ to denote list concatenation. We describe each of our negative pair selection strategies as follows:

1. (Same laterality only) For each query, we select keys in the queue that have the same laterality as the query. Specifically, we replace G in equation (4) by G^l

$$G^l\left(\tilde{\boldsymbol{x}}_1, \boldsymbol{z}_i\right) = \sum_{\boldsymbol{z}_i \in \mathcal{S}(\boldsymbol{x})} \exp[g(\tilde{\boldsymbol{x}}_1)) \cdot g(\boldsymbol{z}_i))] \tag{5}$$

2. (Same laterality reweighted) The first strategy excluded the keys in the queue that have different laterality from the query. Here we set a target hard negative weight and reweight each *exp* term to achieve the target weight. Let

$$G^w\left(\tilde{\boldsymbol{x}}_1, \boldsymbol{z}_i\right) = \sum_{\boldsymbol{z}_i \in \mathcal{S}(\boldsymbol{x})} w_i^s \exp[g(\tilde{\boldsymbol{x}}_1)) \cdot g(\boldsymbol{z}_i))] + \sum_{\boldsymbol{z}_i \in \mathcal{S}(x)^c} w_i^d \exp[g(\tilde{\boldsymbol{x}}_1)) \cdot g(\boldsymbol{z}_i))] \tag{6}$$

where t is the target hard negative weight and $r = \frac{|\mathcal{S}(\boldsymbol{x})|}{|Q|}$ is the proportion of the negative keys in the queue that have the same laterality as $\boldsymbol{x}$. Then $w_i^d = \frac{1-t}{1-r}$ and $w_i^s = \frac{t}{r}$ for all i. In our experiments, we set $t = 0.1$ to allocate 90% of the weight to hard negatives. This allows us to include all negative pairs in the contrastive loss but place emphasis on hard negative pairs with the same laterality.

3. (Same laterality appended) For each query, we select a random sample of the keys that have the same laterality and append them to the existing queue

$$Q = [z_1, z_2, ..., z_K]$$

where K is the queue size. Let

$$A = \{z_{i_1}, z_{i_2}, ..., z_{i_m}\} \subset S\left(x\right) \tag{7}$$

be the random sample of keys with the same laterality as the query. The new queue is

$$Q^a = Q \parallel A$$

and

$$G^a\left(\tilde{\boldsymbol{x}}_1, \boldsymbol{z}_i\right) = \sum_{\boldsymbol{z}_i \in Q^a} \exp[g(\tilde{\boldsymbol{x}}_1)) \cdot g(\boldsymbol{z}_i))]$$

replaces G in equation (4).

4. (Same laterality synthetic) For each query, in addition to appending samples of the keys from $S(\boldsymbol{x})$, we use the samples to generate synthetic keys and append them to the queue. We randomly sample m pairs $(\boldsymbol{s}_i, \boldsymbol{s}_j) \in A = \{\boldsymbol{z}_{i_1}, \boldsymbol{z}_{i_2}, ..., \boldsymbol{z}_{i_m}\}$ and call this set of pairs B.

   For each pair $(\boldsymbol{s}_i, \boldsymbol{s}_j) \in B$, we uniformly sample a number u between 0 and 1 and let

$$h = u \cdot \boldsymbol{s}_i + (1 - u) \cdot \boldsymbol{s}_j$$

   A synthetic image representation is defined as the normalized vector $\frac{h}{\|h\|}$. Let H be the set of these m synthetic image representations and

$$Q^h = Q \parallel B \parallel H$$

   is the new queue. G in equation (1) is replaced by

$$G^h\left(\tilde{\boldsymbol{x}}_1, \boldsymbol{z}_i\right) = \sum_{z_i \in Q^h} \exp[g(\tilde{\boldsymbol{x}}_1)) \cdot g(\boldsymbol{z}_i))]$$

   Note that unlike Kalantidis et al. (2020), we only construct synthetic images once.