Radiology

Simplified Transfer Learning for Chest Radiography Models Using Less Data

Andrew B. Sellergren, BA • Christina Chen, MD • Zaid Nabulsi, MS • Yuanzhen Li, PhD • Aaron Maschinot, PhD • Aaron Sarna, BA • Jenny Huang, MS • Charles Lau, MD, MBA • Sreenivasa Raju Kalidindi, MD • Mozziyar Etemadi, MD • Florencia Garcia-Vicente, MS • David Melnick, BA • Yun Liu, PhD • Krish Eswaran, PhD • Daniel Tse, MD • Neeral Beladia, MS • Dilip Krishnan, PhD • Shravya Shetty, MS

From Google Health, Google, 3400 Hillview Ave, Palo Alto, CA 94304 (A.B.S., C.C., Z.N., Y. Liu, K.E., D.T., N.B., S.S.); Google Research, Cambridge, Mass (Y. Li, A.M., A.S., J.H., D.K.); Google via Advanced Clinical, Deerfield, Ill (C.L.); Apollo Radiology International, Hyderabad, India (S.R.K.); and Northwestern Medicine, Chicago, Ill (M.E., F.G.V., D.M.). Received October 14, 2021; revision requested November 11; revision received March 23, 2022; accepted April 14. **Address correspondence to** A.B.S. (email: *asellerg@google.com*).

Supported by Google.

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

Radiology 2022; 305:454–465 • https://doi.org/10.1148/radiol.212482 • Content codes: CH AI

Background: Developing deep learning models for radiology requires large data sets and substantial computational resources. Data set size limitations can be further exacerbated by distribution shifts, such as rapid changes in patient populations and standard of care during the COVID-19 pandemic. A common partial mitigation is transfer learning by pretraining a "generic network" on a large nonmedical data set and then fine-tuning on a task-specific radiology data set.

Purpose: To reduce data set size requirements for chest radiography deep learning models by using an advanced machine learning approach (supervised contrastive [SupCon] learning) to generate chest radiography networks.

Materials and Methods: SupCon helped generate chest radiography networks from 821 544 chest radiographs from India and the United States. The chest radiography networks were used as a starting point for further machine learning model development for 10 prediction tasks (eg, airspace opacity, fracture, tuberculosis, and COVID-19 outcomes) by using five data sets comprising 684 955 chest radiographs from India, the United States, and China. Three model development setups were tested (linear classifier, nonlinear classifier, and fine-tuning the full network) with different data set sizes from eight to 8^5.

Results: Across a majority of tasks, compared with transfer learning from a nonmedical data set, SupCon reduced label requirements up to 688-fold and improved the area under the receiver operating characteristic curve (AUC) at matching data set sizes. At the extreme low-data regimen, training small nonlinear models by using only 45 chest radiographs yielded an AUC of 0.95 (noninferior to radiologist performance) in classifying microbiology-confirmed tuberculosis in external validation. At a more moderate data regimen, training small nonlinear models by using only 528 chest radiographs yielded an AUC of 0.75 in predicting severe COVID-19 outcomes.

Conclusion: Supervised contrastive learning enabled performance comparable to state-of-the-art deep learning models in multiple clinical tasks by using as few as 45 images and is a promising method for predictive modeling with use of small data sets and for predicting outcomes in shifting patient populations.

Online supplemental material is available for this article.

Approximately 837 million chest radiographs are obtained annually worldwide for detecting, diagnosing, and managing cardiothoracic conditions; chest radiography is also considerably more accessible than CT in many parts of the world (1,2). Serious effort has been invested into developing deep learning models to detect chest radiography abnormalities (3–12). However, core challenges for model development include the need for extremely large, labeled training data sets and the ability to generalize to different populations and institutions (13–15).

Transfer learning, a machine learning approach that repurposes a model trained on one task for a different but related task, may reduce the need for large data sets. A common transfer learning workflow involves first pretraining a deep learning model on a generic source task (often using large nonmedical data sets) and then refining the model on a specific target medical task (using a medical data set) (Fig 1A; gray box). Although transfer learning is more effective when the source and target tasks are similar (eg, both medical), this would typically require tens of thousands of labeled medical images (16,17). Fortunately, this obstacle of lacking medical labels can be partially overcome by using self-supervised machine learning techniques that can make use of unlabeled data (18–21). Some of these concepts have been investigated in chest radiography (17,22).

In this article, we describe a system to facilitate modeling chest radiograph–specific tasks through a three-step training setup: generic image pretraining, chest radiograph–specific pretraining, and task-specific training. The first step uses large nonmedical image data sets for pretraining, similar to the more traditional transfer-learning approach described previously. The second step uses chest radiography data sets with scalable albeit noisy labels of abnormality from natural language processing of radiology reports, in combination with a supervised contrastive (SupCon)

Abbreviations

AUC = area under the receiver operating characteristic curve, SupCon = supervised contrastive

Summary

This method enabled prediction performance comparable to state-of-the-art deep learning models in multiple clinical tasks by using as few as 45 chest radiographs.

Key Results

- Compared with transfer learning from a nonmedical data set, this method reduced label requirements up to 688-fold and improved the area under the receiver operating characteristic curve (AUC) at matching data set sizes.
- Training small nonlinear models by using only 45 chest radiographs yielded an AUC of 0.95 (noninferior to radiologist performance) in classifying microbiology-confirmed tuberculosis in external validation and was better than fine-tuning the network directly (AUC, 0.8).
- Training small nonlinear models by using only 528 chest radiographs yielded an AUC of 0.75 in predicting severe COVID-19 outcomes; fine-tuning the entire network yielded an AUC of 0.74.

learning approach to build a chest radiography network. This chest radiography network converts chest radiographs into information-rich numerical vectors ("embeddings," which, depending on the specific network, are typically hundreds of thousands of numbers in length) that can be used to more easily train models for specific medical prediction tasks (eg, imaging findings, clinical condition, or patient outcome).

We evaluated our approach by measuring the data versus performance tradeoff under three training scenarios: *(a)* a linear model applied to frozen embeddings, *(b)* a nonlinear model applied to frozen embeddings, and *(c)* a nonlinear model produced by fine-tuning the entire network. Our results suggest that performant models can be trained from as few as 10–100 examples. To accelerate chest radiography modeling efforts with low data and computational requirements, we plan to release our model as a service along with scripts to train linear and nonlinear classifiers on top of them.

Materials and Methods

Experiment Design

Our primary goal was to evaluate whether our three-step training setup improves the final performance of prediction tasks. Our baseline consisted of the traditional two-step training setup that starts with pretraining from a large nonmedical data set to produce a generic preinitialized network (Fig 1A). Our three-step training setup included an additional pretraining step that uses a large chest radiography data set with radiology report–derived labels (abnormal or normal) to create a chest radiography network (Fig 1A). The last step in both setups is task-specific training, with the task being to predict an imaging finding, clinical condition, or patient outcome. Note that the chest radiography network pretraining uses noisy radiology report–derived labels, while the medical task–specific training as well as task-specific evaluations use clean labels from radiologist image reviews, molecular testing, or clinical outcomes.

Next, we evaluated how well the embeddings (from the traditional generic network or our chest radiography network) generated by means of these two-step and three-step training setups performed in linear classification, nonlinear classification, and as a starting point for fine-tuning of the full network (Fig 1A). Because the appropriateness of these training methods likely depends on both the volume of data available and the difficulty of the predictive task, we simulated different-sized training data sets by subsampling the training data sets (detailed in the following paragraphs).

In our experiments, we used SupCon to build two generic preinitialized networks from two well-known nonmedical image data sets: ImageNet, a data set containing natural images and commonly used to initialize machine learning models for other applications (with approximately 10^7 images), and JFT-300M, a larger data set consisting of weakly labeled natural images used in this work as an alternative to ImageNet initialization (with approximately 10^8 images) (23,24). For each data set, we used a different network architecture (a documented design of how the network's layers and connections are arranged)—the small modern neural network architecture (EfficientNet-B7) for the ImageNet data set and the larger architectures (ResNet-101 × 3 and ResNet-152 × 4) for the larger JFT-300M data set. Henceforth, experiments described as using the data sets "ImageNet" and "JFT-300M" will also describe their associated architectures, unless otherwise specified. This study, using de-identified retrospective data, was reviewed by the Advarra institutional review board (Columbia, Md), which determined that it was exempt from further review under Title 45, Code of Federal Regulations Part 46 *(https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html)*. This study was funded by Google.

Chest Radiography Data Sets

The medical data sets used are detailed in Tables 1–3. The chest radiographs used to produce the chest radiography network included more than 700 000 images across five hospitals in five cities in India (hereafter, IND1 data set) (25), the ChestX-ray14 data set (5,26), and a hospital from Illinois in the United States (hereafter, US1 data set) (25). The data sets for training the task-specific models included IND1, ChestX-ray14, and CheXpert for the general chest radiography findings (described next). The tuberculosis setup attempted the classification both ways—training on tuberculosis data sets from the United States (hereafter, US2-TB) and evaluating on tuberculosis data sets from China (hereafter, CN-TB) and training on CN-TB and evaluating on US2-TB. The COVID-19 prediction task involved training on COVID-19 data sets from the United States (hereafter, US1-COV1) and evaluating on a separate site, US1-COV2. Independent external validation test data sets include US2-TB, CN-TB, and CheXpert.

Findings Studied

To ensure our conclusions were robust and not overly specific to certain findings, we conducted experiments across

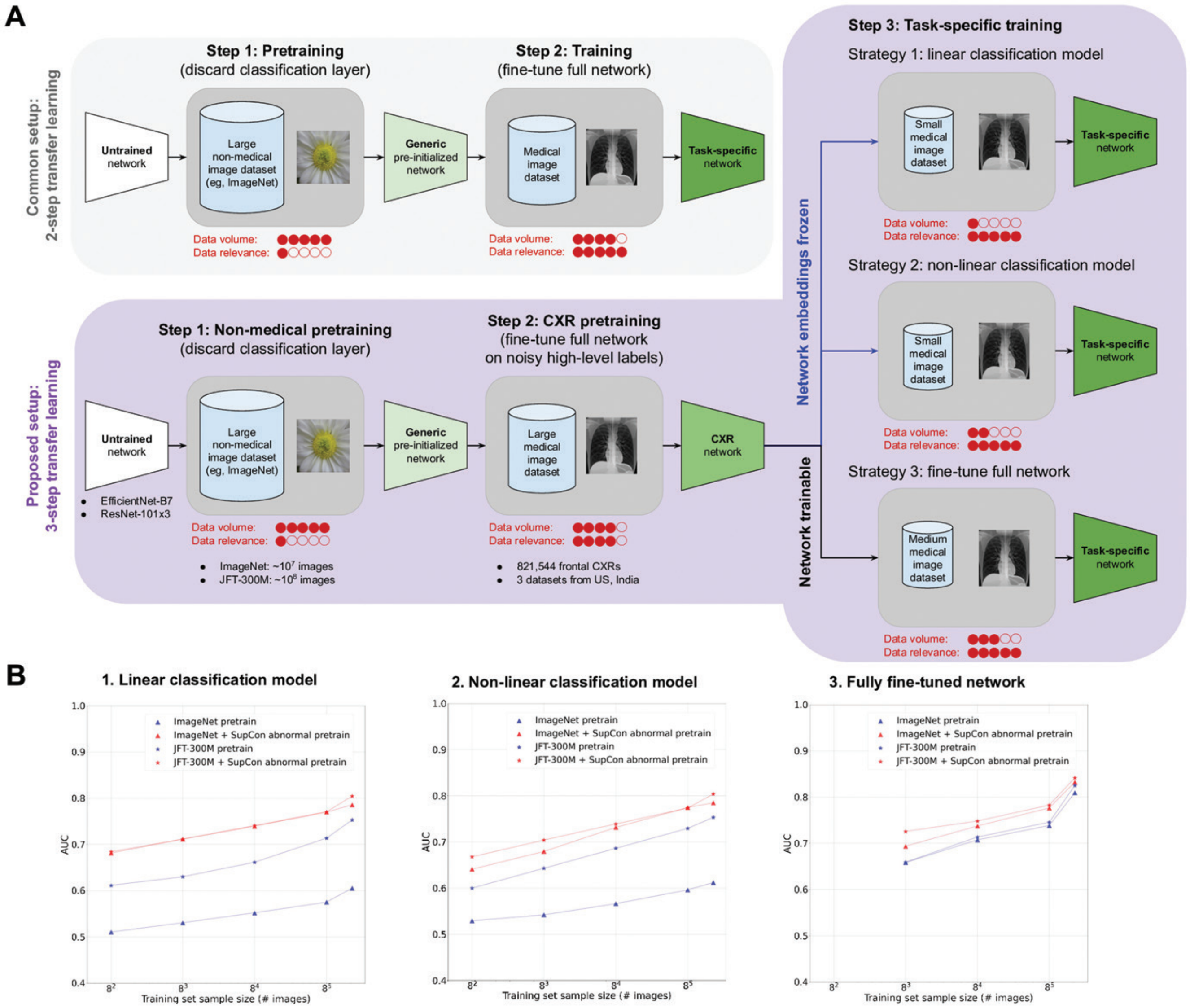

Figure 1: Baseline two-step training setup (generic network) is compared with our proposed three-step training setup that generates a chest radiography (CXR) network by using a supervised contrastive (SupCon) machine learning technique. **(A)** The chest radiography network was evaluated in three settings by using task-specific training data: by freezing the chest radiography network and training *(a)* a small linear model or *(b)* a nonlinear model (multilayer perceptron) and *(c)* by fine-tuning the entire network. Additional experiments around different types of pretraining are described in Figure E1 (online). **(B)** The graphs present the results for the three strategies described in **A**, averaged across multiple tasks (airspace opacity, fracture, pneumothorax, consolidation, pleural effusion, and pulmonary edema) on ChestX-ray14 data set; see Figure 2 for the same analysis per task. The final training set size is the largest size available for each task (at least 68 801 radiographs) and differs based on task (for additional details, refer to the Materials and Methods section). AUC = area under the receiver operating characteristic curve, ImageNet = data set containing natural images, JFT-300M = larger data set containing natural images.

multiple findings and/or diseases and clinical outcomes (Tables 1–3). These included multiple imaging findings found in a general clinical setting (six findings in the ChestX-ray14 data set [5,26] and five findings in the CheXpert data set [27]), microbiologically confirmed tuberculosis in two publicly available data sets from Montgomery County, Md, in the United States (US2-TB) and Shenzhen, China (CN-TB) (which have both tuberculosis-positive and tuberculosis-negative chest radiographs [28–32]), and five important clinical end points (four individual end points and one composite end point; see Table E1 [online]) for COVID-19 in a hospital in the United States (US1-COV2) (11). These findings are detailed in Appendix E1 (online). Ground truth tuberculosis status was provided with the two public data sets (US2-TB, CN-TB).

Chest Radiography Network Pretraining by Means of SupCon

Our second pretraining step produces a chest radiography network using SupCon (21). SupCon builds on the self-supervised learning technique, a simple framework for contrastive learning of visual representations, or SimCLR (18), which is designed to encourage the network to learn a good representation from unlabeled examples by leveraging the idea that crops of the same image ("A") are more similar than crops from different source images ("A" and "B"). SupCon extends this by leveraging the idea that images

Table 1: Baseline Characteristics for Data Sets Used in Chest Radiography Network Pretraining and Task-specific Training and Tuning

Characteristic	IND1	ChestX-ray14	US1	US1-COV1	US2-TB*	CN-TB*	CheXpert (Training)*
Used for chest radiology network pretraining	Yes	Yes	Yes	No	No	No	No
Used for task-specific training and fine-tuning	Yes	Yes	No	Yes	Yes	Yes	Yes
Data set origin	Five clusters of hospitals from five cities in India	NIH Clinical Center	Hospital in Illinois, United States	Hospital in Illinois, United States (same system as US1)	Montgomery, Md, United States	Shenzhen, China	Stanford hospital system, Calif, United States
No. of patients	348 335	23 152	9958	648	91	595	65 240
Age (y)[†]	35–58	35–59	50–72	42–67	27.5–51.5	26–43	NA
Sex							
Female	133 833 (38.5)	10 646 (46)	5267 (52.9)	329 (46.7)	47 (51.6)	190 (31.9)	NA
Male	214 334 (61.5)	12 506 (54)	4671 (47.1)	376 (53.3)	44 (48.4)	405 (68.1)	NA
Unknown	168 (<0.1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	NA

Note.—Unless otherwise stated, data are numbers of patients. Data in parentheses are percentages. CN-TB = tuberculosis data set from China, IND1 = data set from five hospitals in five cities in India, NA = not available, NIH = National Institutes of Health, US1 = data set from a hospital system from Illinois in the United States, US1-COV1 = COVID-19 data set from a hospital system from Illinois in the United States (first set of hospitals), US2-TB = tuberculosis data set from the United States.

* External data set not used to build the chest radiography network.

[†] Data are IQRs.

Table 2: Characteristics of Images Used for Chest Radiography Network Pretraining and Task-specific Training and Tuning

Characteristic	IND1	ChestX-ray14	US1	US1-COV1	US2-TB*	CN-TB*	CheXpert (Training)*
No. of images	485 082	71 292	147 011	705	91	595	223 414
Chest view							
Anteroposterior	79 958 (16.5)	26 103 (36.6)	101 476 (69)	684 (97)	0 (0)	0 (0)	NA
Posteroanterior	625 735 (83.5)	45 189 (63.4)	21 411 (14.6)	20 (2.8)	91 (100)	595 (100)	NA
Unknown	0 (0)	0 (0)	24 124 (16.4)	1 (0.1)	0 (0)	0 (0)	…
Airspace opacity	43 629 (9)	2003 (2.8)	14 620 (9.9)[†]	157 (22.3)[†]	NA	NA	94 211 (42.2)
Fracture	5200 (1.1)	391 (0.5)	4969 (3.4)[†]	73 (10.4)[†]	NA	NA	7436 (3.3)
Pneumothorax	1657 (0.3)	2109 (3)	7199 (4.9)[†]	0 (0)[†]	NA	NA	17 693 (7.9)
Consolidation	15 144 (3.1)	2741 (3.8)	6250 (4.3)[†]	21 (3)[†]	NA	NA	12 983 (5.8)
Pleural effusion	1228 (0.3)	6974 (9.8)	33 141 (22.5)[†]	29 (4.1)[†]	NA	NA	76 899 (34.4)
Pulmonary edema	1136 (0.2)	1082 (1.5)	48 661 (33.1)[†]	69 (9.8)[†]	NA	NA	49 675 (22.2)
Atelectasis	15 929 (3.3)	6536 (9.2)	33 797 (23)[†]	153 (21.7)[†]	NA	NA	29 720 (13.3)
Cardiomegaly	1115 (0.2)	1382 (1.9)	16 762 (11.4)[†]	45 (6.4)[†]	NA	NA	23 385 (10.5)
Tuberculosis	1090 (0.2)[‡]	NA	NA	NA	37 (40.7)	304 (51.1)	NA
Severe COVID-19	NA	NA	NA	186 (26.4)	NA	NA	NA

Note.—Unless otherwise stated, data are numbers of images. Data in parentheses are percentages. CN-TB = tuberculosis data set from China, IND1 = data set from five hospitals in five cities in India, US1 = data set from a hospital system from Illinois in the United States, NA = not available, US1-COV1 = COVID-19 data set from a hospital system from Illinois in the United States (first set of hospitals), US2-TB = tuberculosis data set from the United States.

* External data set not used to build chest radiography network.

[†] Estimated from radiology reports.

[‡] Estimated from radiology reports; tuberculosis is endemic in India, so this may be an underestimate.

Table 3: Baseline Characteristics and Image Characteristics for Task-specific Test Data Sets

Characteristic	ChestX-ray14	US2-TB*	CN-TB*	US1-COV2*	CheXpert (Validation)*
Data set origin	NIH Clinical Center	Montgomery, Md, United States	Shenzhen, China	Hospital in Illinois, United States (same system as US1)	Stanford hospital system, Calif, United States
No. of patients	2797	138	67	294	234
Age (y)[†]	33–59	28–52	26–44	44–63	NA
Sex					
Female	1240 (44.3)	74 (53.6)	23 (34.3)	141 (48)	NA
Male	1557 (55.7)	63 (45.7)	44 (65.7)	153 (52)	NA
Unknown	0 (0)	1 (0.7)	0 (0)	0 (0)	NA
No. of images	25 594	138	67	294	234
Chest view					
Anteroposterior	14 502 (56.6)	0 (0.0)	0 (0.0)	192 (65.3)	NA
Posteroanterior	11 092 (43.3)	138 (100)	67 (100)	9 (3.1)	NA
Unknown	0 (0)	0 (0)	0 (0)	93 (31.6)	NA
Airspace opacity	1161 (4.5)	NA	NA	102 (34.7)[‡]	117 (50)
Fracture	73 (0.3)	NA	NA	6 (2)[‡]	0 (0)
Pneumothorax	2665 (10.4)	NA	NA	0 (0)[‡]	7 (3)
Consolidation	1815 (7.1)	NA	NA	8 (2.7)[‡]	32 (13.7)
Pleural effusion	4658 (18.2)	NA	NA	9 (3.1)[‡]	64 (27.4)
Pulmonary edema	925 (3.6)	NA	NA	52 (17.7)[‡]	42 (17.9)
Atelectasis	3279 (12.8)	NA	NA	39 (13.3)[‡]	75 (32.1)
Cardiomegaly	1069 (4.2)	NA	NA	18 (6.1)[‡]	66 (28.2)
Tuberculosis	NA	58 (42)	32 (48)	NA	NA
Severe COVID-19	NA	NA	NA	71 (24.1)	NA

Note.— Unless otherwise stated, data are numbers of patients or numbers of images. Data in parentheses are percentages. CN-TB = tuberculosis data set from China, IND1 = data set from five hospitals in five cities in India, NA = not available, NIH = National Institutes of Health, US1 = data set from a hospital system from Illinois in the United States, US1-COV2 = COVID-19 data set from a hospital system from Illinois in the United States (second set of hospitals), US2-TB = tuberculosis data set from the United States.

* External data set not used to build chest radiography network.

[†] Data are IQRs.

[‡] Estimated from radiology reports.

of the same class (eg, class 0) are more similar than images from different classes (eg, class 0 and class 1). Here, we apply this idea to leveraging noisy labels of whether a chest radiograph contains abnormal findings or not. These noisy labels were extracted from natural language processing of radiologist reports (please see Majkowska et al [25] for details on IND1 and Table E2 [online] for details on US1) and combining more specific labels (eg, ChestX-ray14). This chest radiography network converts chest radiographs into high-dimensional numerical vectors (embeddings). Intuitively, the embedding of an image is a numerical representation that summarizes the information contained in that image, such that a simple model can use this "summary" as input for new prediction tasks. Because these simple models (eg, a small linear or nonlinear model) are smaller than the full network, they often require much less data for training.

Task-specific Model Training

We explored three types of task-specific model development: *(a)* a linear model applied to frozen embeddings, *(b)* a nonlinear model applied to frozen embeddings, and *(c)* a model produced by fine-tuning the entire network (including the embeddings). The first two models use static frozen embeddings created by using the chest radiography network to transform each chest radiograph in our task-specific data sets to an embedding. The linear model consists of training a single-layer linear probe, which makes one prediction for each embedding. The nonlinear model is similar but consists of a multilayer perceptron instead of a single layer. The third approach starts from the same pretrained chest radiography network but adds a custom classification layer and subsequently fine-tunes the entire network for each prediction task, which enables the embeddings to be refined for the specific prediction tasks.

Evaluation

To evaluate the model's sensitivity to data set size and understand generalization across multiple tasks, we subsampled the training data set to five sizes spanning a logarithmic scale of 64, 512, 4096, 32 768, and 68 801 or 674 533. The maximum training set sample size was based on the availability of labels across multiple data sets; for example, airspace opacity, fracture, and pneumothorax were available in both ChestX-ray14 and IND1 (and subsampled up to 674 533), whereas consolidation, pleural effusion, and pulmonary edema were only available in ChestX-ray14 (and subsampled up to 68 801). Sampling was stratified to

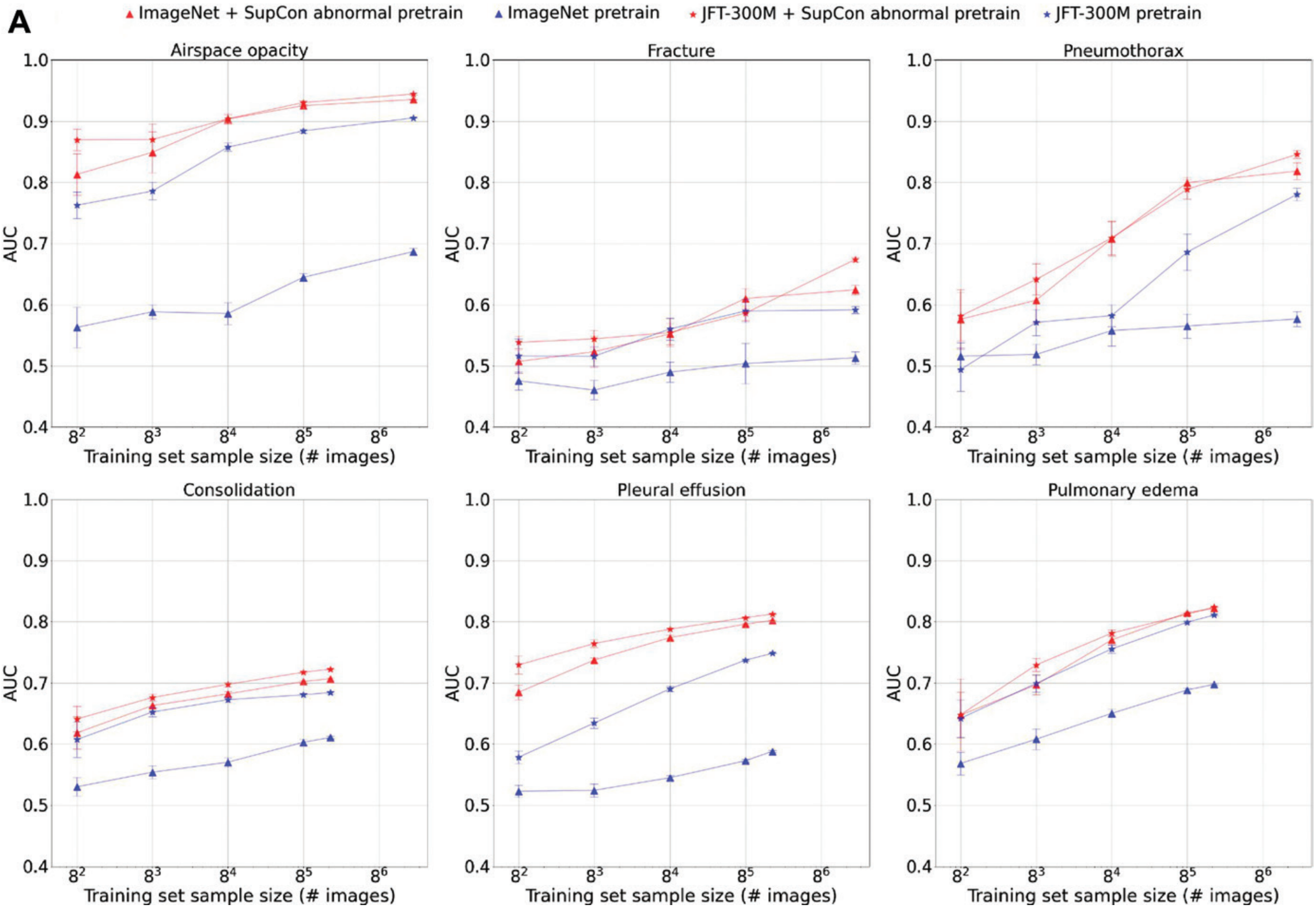

Figure 2: Graphs show effect of using the chest radiography network from our three-step training setup for task-specific prediction on ChestX-ray14 data set. Results are from Figure 1B and were sectioned on a per-finding basis. **(A)** Results with nonlinear model using frozen embeddings. Results for the linear model are similar though generally slightly lower than that for the nonlinear models. AUC = area under the receiver operating characteristic curve, ImageNet = data set containing natural images, JFT-300M = larger data set containing natural images, SupCon = supervised contrastive *(Fig 2 continues)*.

maintain the positive class ratio and to ensure that all subsamples contained both classes. Note that the full network refinement approach was not attempted for the smallest data set size (n = 64) due to the small size. Because all prediction tasks were binary, we used the area under the receiver operating characteristic curve (AUC) for all evaluations. Comparator radiologists for tuberculosis were India-based consultants with radiology certifications who had experience in reading tuberculosis images. Additional details on statistical analysis are presented in Appendix E1 (online).

Visualizing the Embeddings at Each Step

To better understand how the embeddings change at each step, we leveraged t-distributed stochastic neighbor embedding, or t-SNE (33), a widely used technique for visualizing high-dimensional data. Although the units are not necessarily quantitatively interpretable, the qualitative observation of how data points are distributed can give an idea of how examples (chest radiographs) from different classes are spread out in the high-dimensional space. We use this to look at the separation between positive and negative labels at each of our three steps (generic network, chest radiography network, task-specific network).

Code Availability

Supervised contrastive training code is available at *https://github.com/google-research/google-research/tree/master/supcon*.

A service will be made available for researchers to obtain embeddings generated by our chest radiography model by uploading chest radiographs. Code to train linear and nonlinear classifiers by using these generated embeddings will also be made available.

Results

Our experiments revolved around understanding whether adding a second pretraining step to produce a chest radiography network improves the final performance for prediction tasks and under what data set size regimen this occurs. The following sections will describe our results for the findings or end points across five data sets and their associated findings or end points.

General Clinical Findings: ChestX-Ray14

In the ChestX-ray14 data set, on average across six findings, SupCon substantially and consistently improved accuracy of models developed across a range of training data set sizes for both

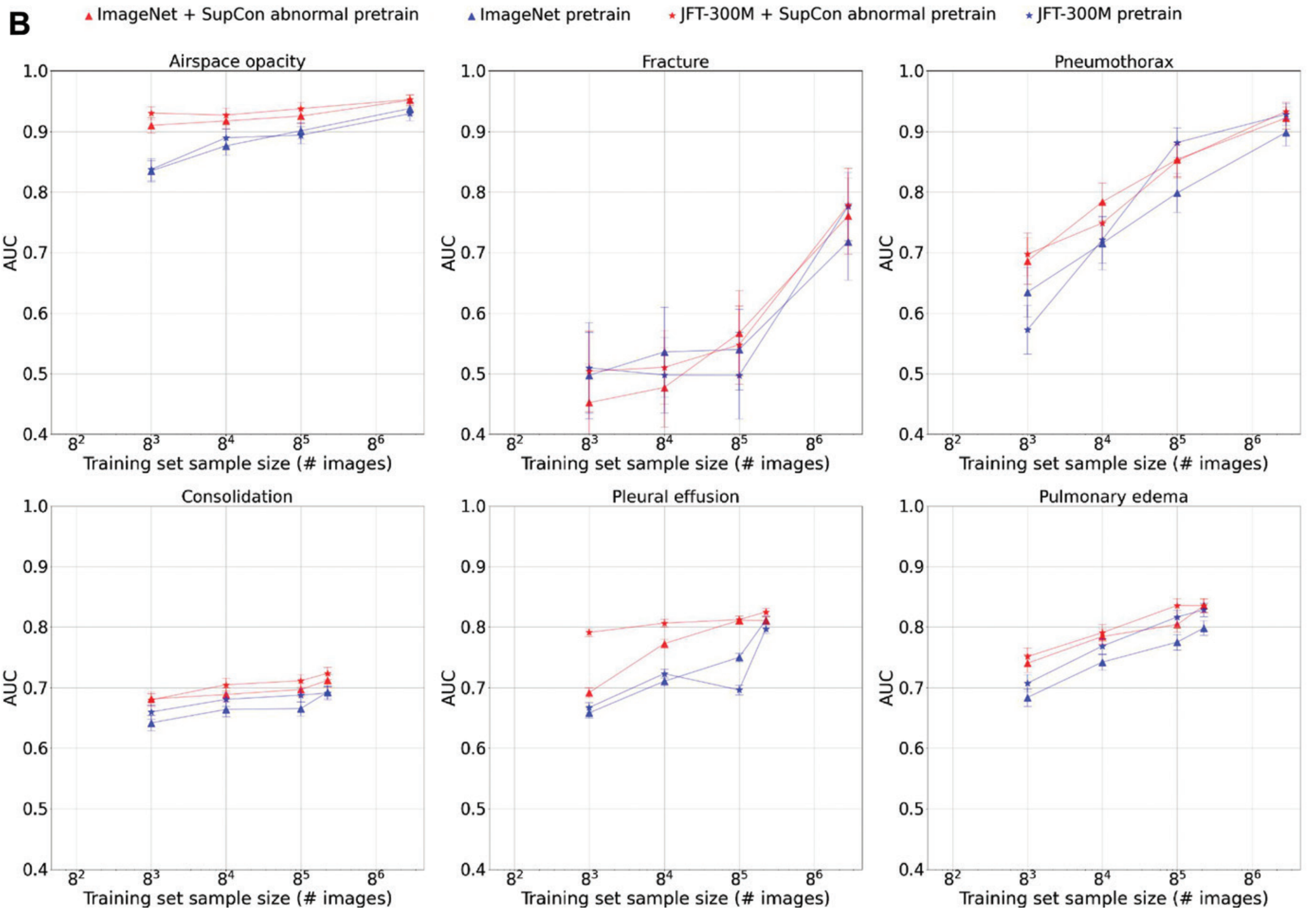

Figure 2 *(continued)*: **(B)** Results with fine-tuning the full network. AUC = area under the receiver operating characteristic curve, ImageNet = data set containing natural images, JFT-300M = larger data set containing natural images, SupCon = supervised contrastive.

frozen embedding scenarios (linear classification and nonlinear classification). When the entire network was fine-tuned, an improvement was also observed, but the baselines were substantially higher and the gains were smaller (Fig 1B, comparing gap between red vs blue lines). These trends were generally consistent for both ImageNet and JFT-300M, although with a much higher baseline performance for the JFT-300M (which has more parameters and was pretrained on a larger data set; see Fig 1B, comparing blue lines with triangles vs stars as points). With the application of SupCon, the performance converged (Fig 1B, comparing both red lines).

When radiologists dove deeper into the individual findings for the nonlinear model (Fig 2A) and for the fine-tuned network (Fig 2B), the observations remained generally similar for five findings, although with widely varying performances across findings. Pretraining with SupCon demonstrated statistically superior performance compared with not using SupCon ($P < .01$) (Tables E3 and E4 [online] for linear classification and Tables E5 and E6 [online] for nonlinear classification, Tables E7 and E8 [online] for fine-tuning the network, and Table E9 [online] for a comparison between using SupCon vs not). The exception to the other five findings was fractures, for which none of the approaches tested performed particularly well at low training-set sizes.

General Clinical Findings: CheXpert

We next benchmarked SupCon on the publicly available CheXpert data set (an external data set not used for pretraining), with a slightly different set of five findings (Fig 3) and a larger architecture for the JFT-300M pretraining, ResNet-152 × 4 (which generally showed better performance than ResNet-101 × 3). The observations were generally similar to those seen in ChestX-ray14 for four of five findings. The exception was cardiomegaly, where the use of SupCon actually was associated with a lower performance at the smallest data set sizes (n = 64 and 512), but which reversed to similarly show higher performance with SupCon (n = 4096, 32 768, and 224 316). When these results were compared to that of the original CheXpert model (Irvin et al [27]), SupCon enabled comparable performance (as assessed by performance within the CIs) at 8^4 images for atelectasis, cardiomegaly, and pleural effusion and 8^5 images for pulmonary edema.

Tuberculosis

When evaluated for its ability to develop tuberculosis-detection models (using two external data sets), SupCon showed similar trends as general clinical findings on the CheXpert data set, although the maximum data set sizes were smaller, and the delta (difference) was more striking (Fig 4). This held true whether

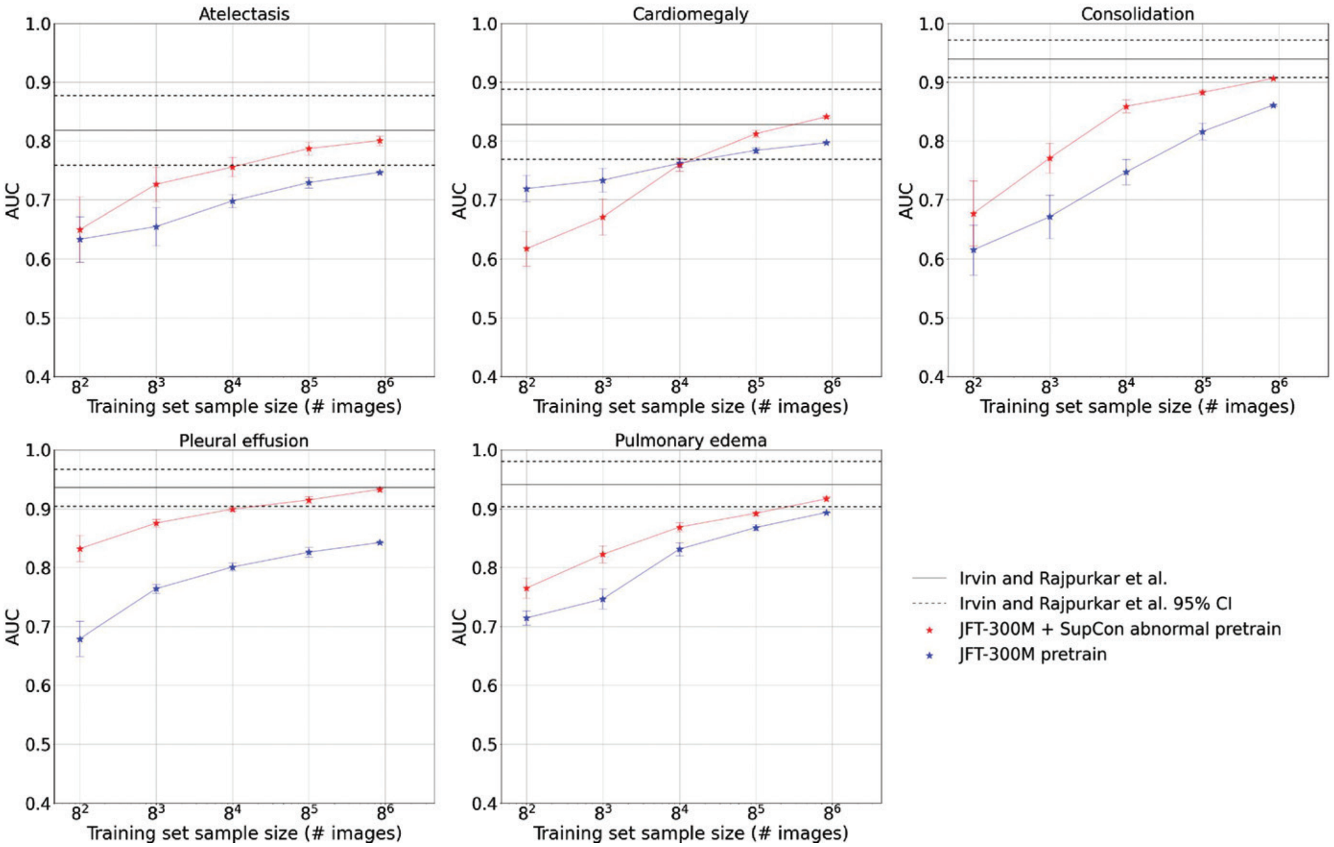

Figure 3: Graphs show effect of using the chest radiography network from our three-step training setup with nonlinear classifiers on task-specific findings in the CheXpert data set. Solid and dotted horizontal lines indicate performance of the original CheXpert model (Irvin et al [27]) on all available training data (224 000 images); area under the receiver operating characteristic curves (AUC) for the nonlinear models trained on 1% and 10% of the train set (data points at 8^4 and 8^5) approached published performance of the original CheXpert model for atelectasis, cardiomegaly, pleural effusion, and pulmonary edema. JFT-300M = larger data set containing natural images, SupCon = supervised contrastive.

training on US2-TB and testing on CN-TB or vice versa. We showed that the naive approach of fine-tuning the entire network on tuberculosis labels performed less well than our nonlinear models on top of frozen embeddings (Fig 4A). The resultant models (all 10 trained on random subsamples of the training data) attained noninferiority to India-based radiologists in detecting tuberculosis when using just 45 training images, and the AUC reached 0.92 with eight training images (Fig 4B). At all training set sizes (including two and eight images), the model trained on US2-TB had a receiver operating curve comparable to the performance of radiologists (40%–60% sensitivity at near-perfect specificity) in the CN-TB data set.

COVID-19 Clinical Outcomes

Finally, we evaluated the ability of our model to predict four key clinical outcomes for patients with COVID-19, as well as a composite end point encompassing all four outcomes. Despite the relatively small training data set containing 12–173 outcomes, the performance of a nonlinear model (JFT-300M) rose rapidly with increasing sample size, attaining an AUC of 0.75 with use of just 528 examples (Fig 5). We showed that our nonlinear classifier trained on frozen embeddings outperformed a model that was fine-tuned on the entire data set.

Visualizing the Embeddings at Each Step

We use the example case of airspace opacity to show that the SupCon embeddings from the chest radiography network produced a better visual separation than the generic pretrained network and a visually comparable separation to the fully fine-tuned task-specific network (Fig 6).

Discussion

This study presented a three-step training setup that involved generating a chest radiography network (compared with the standard approach of a generic network) to accelerate the building of task-specific deep-learning models. Our main takeaways are as follows: *(a)* A simple natural language processing of radiology reports could scalably generate weak labels for the supervised contrastive learning approach used for building the chest radiography network (21); *(b)* in small data regimens the resultant embeddings can improve the task-specific classification performance substantially, by as much as an absolute area under the receiver operating characteristic curve of 0.1–0.2; *(c)* phrased differently, we were able to achieve similar performance with three- to 688-fold less data; and *(d)* the gains were less prominent in the large data regimen.

We found that while using our chest radiography network in a frozen manner and training simple linear and nonlinear

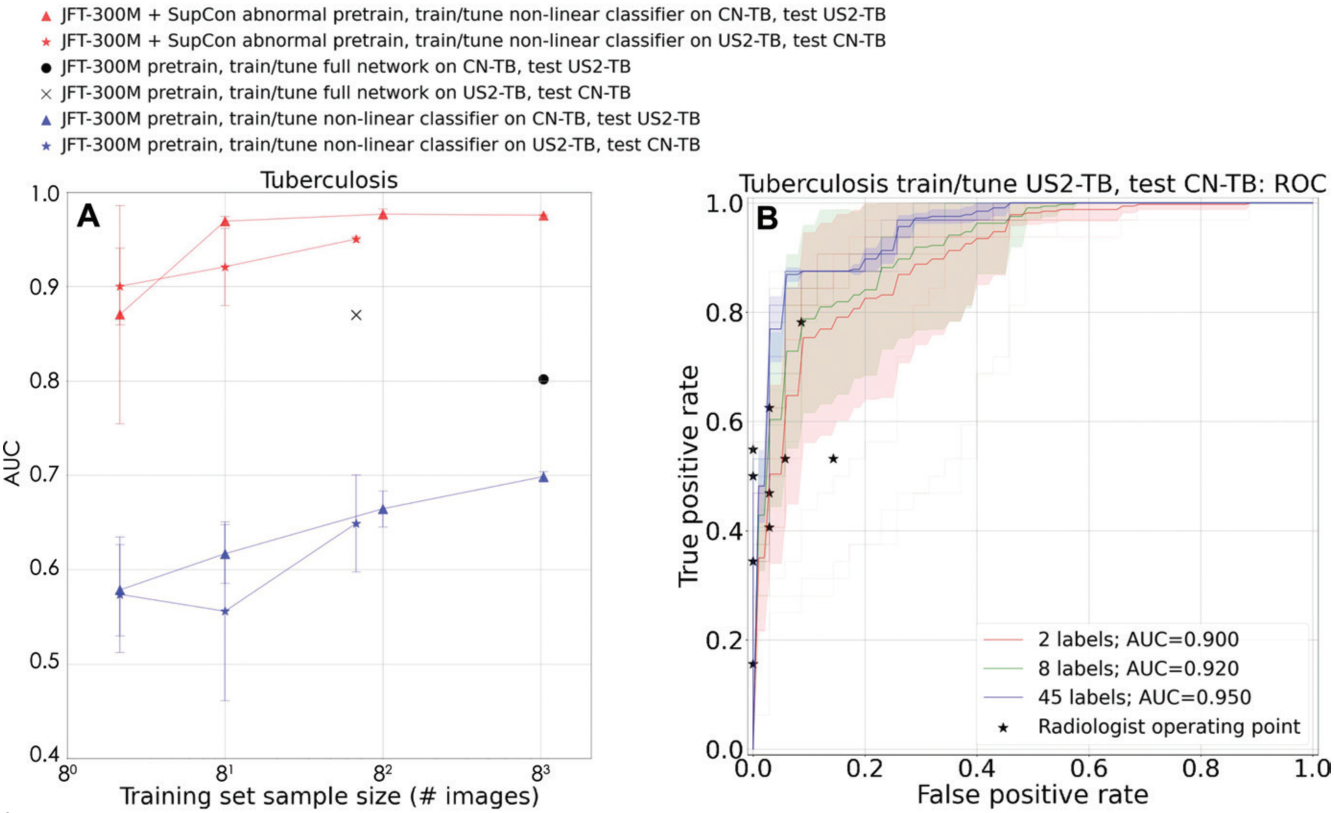

Figure 4: Graphs show effect of using the chest radiography network from our three-step training setup with nonlinear classifiers for tuberculosis detection. **(A)** Graph shows the improvement in performance with increasing training examples. Radiologist noninferiority was achieved with orders of magnitude of fewer training examples; a nonlinear model trained on 45 images from the tuberculosis data set from the United States (US2-TB) was noninferior to 10 India-based radiologists ($P < .01$) on the external validation data set (tuberculosis data set from China [CN-TB]). The × shows fine-tuning the entire network on US2-TB with testing on CN-TB (area under the receiver operating characteristic curve [AUC] = 0.87). The circle shows fine-tuning the entire network on CN-TB with testing on US2-TB (AUC = 0.80). **(B)** Graph shows the performance and CIs with increasing numbers of training examples. Lightly shaded areas represent the 95% CIs of models trained on different random subsets of US2-TB, with the dark lines corresponding to the mean. India-based radiologist consultants had an average of 6 years of experience (range, 3–9 years). JFT-300M = larger data set containing natural images, ROC = receiver operating characteristic, SupCon = supervised contrastive.

models, performance increased when using SupCon on most task-specific predictions in a data set that was also used for pretraining (ChestX-ray14), as well as an external data set not used for pretraining (CheXpert). This improvement generalized to identification of medical disease not labeled in pretraining (tuberculosis) and to prediction of downstream clinical outcomes (COVID-19 clinical outcomes). While the benefits seen for these two tasks were large, the maximum training data set size was only in the hundreds. Additionally, we found that when data sets were small, as in the case of our tuberculosis data sets, fine-tuning the entire network (without our proposed pretraining method) was less robust than our nonlinear models on top of frozen embeddings. Finally, we found that the choice of architecture and generic pretraining data set (ImageNet and JFT-300M) made less of a difference in performance than the gains from using SupCon; both architectures performed similarly when combined with SupCon.

Although our observation that SupCon substantially improved performance was generally highly consistent across architectures, data sets, and tasks, there were two tasks that showed less robust improvements: cardiomegaly and fracture. While the cardiomegaly trendline for SupCon had lower performance with low sample sizes (Fig 3), it improved more rapidly and surpassed the control at around 8^4 examples. As for fracture, the relatively lackluster performance for both SupCon and the control was potentially due to the small number of positives in the training set. Both of these may also have “label noise” in the training data set arising from variable accounting for breath depth or nominal cardiac size in anteroposterior versus posteroanterior images, and some fractures were potentially mischaracterized in the reference standard (34). It is notable that these are both nonpulmonary tasks. By contrast, many of the abnormalities used for pretraining are more relevant to identifying opacities in the lungs and may be less sensitive to these issues. Future work is needed to balance different types of manifestations of disease in chest radiographs to potentially balance the embeddings’ ability to generalize to nonpulmonary findings.

Notably, we have shown that our chest radiography network can perform well with as few as hundreds of task-specific training examples. By using the COVID-19 pandemic as an example, as the affected demographic and severity of illness changed over time (perhaps affected by virus variants, improved medical interventions, availability and use of vaccines, and changing viral transmission patterns), model updates may be more easily achieved by means of these data-efficient techniques. This chest radiography network may also be useful in the study of less

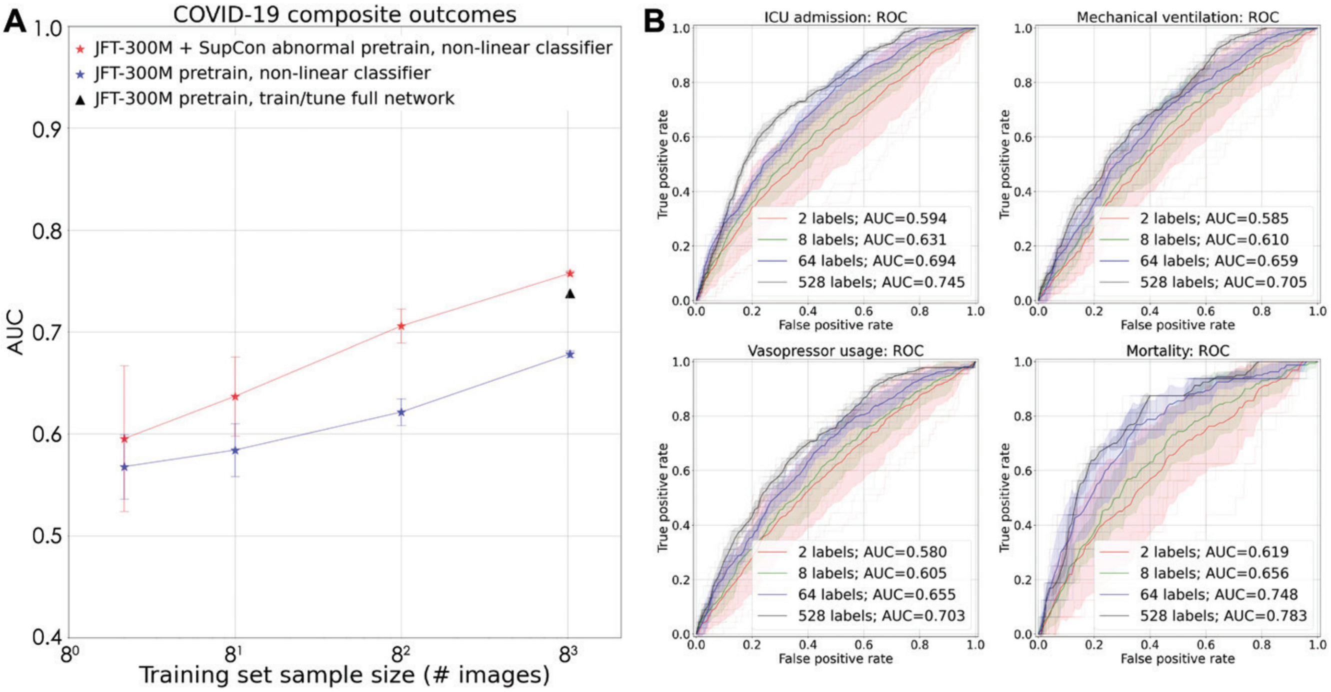

Figure 5: Graphs show effect of the use of the chest radiography network from our three-step training setup with nonlinear classifiers to predict COVID-19 outcomes. **(A)** Graph shows the improvement in performance with increasing training examples. The performance increases rapidly despite having only hundreds of training examples (with a quarter experiencing the outcomes). A nonlinear classifier trained (with a fraction of the compute resources) on frozen embeddings of the whole data set (528 images) also outperformed a network pretrained on JFT-300M and fully fine-tuned on the whole data set (triangle, area under the receiver operating characteristic curve [AUC] = 0.74). **(B)** Graphs show the performance and CIs of individual outcomes with increasing number of training examples. Lightly shaded areas represent the 95% CIs of models trained on different random subsets of the COVID-19 data from a hospital system in the United States, with the dark lines corresponding to the mean. ICU = intensive care unit, JFT-300M = larger data set containing natural images, ROC = receiver operating characteristic, SupCon = supervised contrastive.

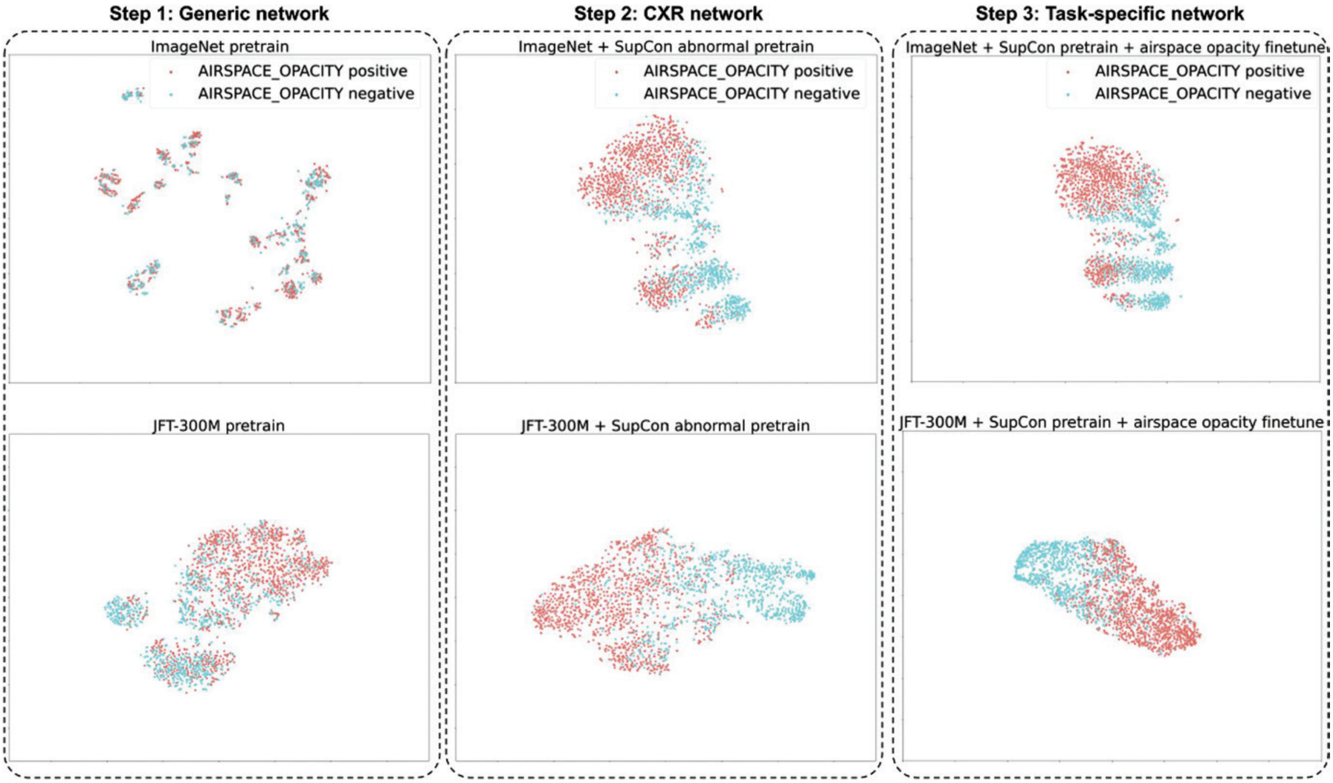

Figure 6: The *t*-distributed stochastic neighbor embedding visualizations of the embeddings at each step in our three-step training setup. The supervised contrastive (SupCon) embeddings produced a better visual separation of the classes (middle) than the generic pretrained network (left) and as good a separation as a fully fine-tuned network (right). Note that this visualization technique leverages highly nonlinear axes, so neither axis can be assigned readily interpretable units. CXR = chest radiography, ImageNet = data set containing natural images, JFT-300M = larger data set containing natural images.

common diseases, which is also often limited by small data. Yet another area where the approach presented may be of value lies in institutions and teams desiring to develop and study custom models for their local task of interest and patient population and, thus, operating in the regimen of small data.

Our study has limitations. First, although both anteroposterior and posteroanterior chest radiographs were well represented in the data sets used to develop the chest radiography models, information was lacking about the availability of digital radiography versus film-screen radiography and whether there were also camera-taken photographs of chest radiographs viewed on a screen or view box. Such nonstandard chest radiograph formats may be a useful case of interest in resource-scarce environments, and generalization will need to be further explored in future work. While we have included data from multiple countries, caution should be taken before generalization into new health care settings, environments, and tasks, and further testing on additional data sets (35) may be useful. Also, while we have described a method to improve model performance on clinical tasks, this does not necessarily translate into improved clinical outcomes. Future work will also be needed to understand generalization with respect to chest radiography imaging quality.

The main goal of this study was to evaluate an approach to substantially overcome the data limitations in developing application-specific chest radiography classification models. Our results suggest that pretraining on scalably extractable noisy labels allows us to provide generalizable embeddings that substantially improve predictive performance across a wide range of data sets and prediction tasks with as few as tens of examples. This unlocks the ability to rapidly train chest radiography models on smaller data sets or when data are scarce. We will be releasing a service to enable use of the trained embeddings (see code availability section) and look forward to follow-up work leveraging these embeddings to develop new artificial intelligence technologies.

Acknowledgments: The authors thank the members of the Google Health Radiology and labeling software teams for software infrastructure support, logistical support, and assistance in data labeling. For the ChestX-ray14 data set, we thank the NIH Clinical Center for making it publicly available. Sincere appreciation also goes to the radiologists who enabled this work with their image interpretation and annotation efforts throughout the study, Jonny Wong, BA, for coordinating the imaging annotation work, and Akinori Mitani, MD, and Craig H. Mermel, MD, PhD, for providing feedback on the manuscript.

Author contributions: Guarantors of integrity of entire study, **A.B.S., C.C., D.K., S.S.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **A.B.S., C.C., Z.N., Y. Li, A.M., C.L., M.E., Y. Liu, D.K.**; clinical studies, **C.C., C.L., S.R.K., M.E., S.S.**; experimental studies, **A.B.S., Z.N., Y. Li, A.M., A.S., C.L., M.E., F.G.V., D.T., D.K., S.S.**; statistical analysis, **A.B.S., Y. Li, M.E., F.G.V., D.M., Y. Liu**; and manuscript editing, **A.B.S., C.C., Z.N., Y. Li, A.M., J.H., S.R.K., M.E., Y. Liu, K.E., D.T., N.B., D.K., S.S.**

Data sharing: Data generated or analyzed during the study are available from the corresponding author by request.

Disclosures of conflicts of interest: **A.B.S.** Support for attending meetings and/or travel from Google; patents planned, issued, or pending with Google; stock with Google; receipt of equipment, materials, drugs, medical writing, gifts, or other services from Google. **C.C.** Stock with Google. **Z.N.** Stock with Google. **Y. Li** No relevant relationships. **A.M.** Stock with Google. **A.S.** Patents planned, issued, or pending with Google; stock with Google. **J.H.** Stock with Google. **C.L.** Radiologist consultant fees paid by Google. **S.R.K.** No relevant relationships. **M.E.** No relevant relationships. **F.G.V.** No relevant relationships. **D.M.** No relevant relationships. **Y. Liu** Stock with Google. **K.E.** Stock with Google. **D.T.** Work-related travel paid by Alphabet; patents planned, issued, or pending with Google; stock with Google. **N.B.** Stock with Google. **D.K.** Support for attending meetings and/or travel provided by Google; patent application planned; stock with Google. **S.S.** Patent pending; stock with Google.

References

1. Ngoya PS, Muhogora WE, Pitcher RD. Defining the diagnostic divide: an analysis of registered radiological equipment resources in a low-income African country. Pan Afr Med J 2016;25:99.
2. United Nations Scientific Committee on the Effects of Atomic Radiation. Sources and effects of ionizing radiation. United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) Reports. 2008. Accessed July 1, 2021.
3. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv [cs.CV]. 2017. http://arxiv.org/abs/1711.05225.
4. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 2018;15(11):e1002686.
5. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. arXiv [cs.CV]. 2017. http://arxiv.org/abs/1705.02315.
6. Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology 2017;284(2):574–582.
7. Tang YX, Tang YB, Peng Y, et al. Automated abnormality classification of chest radiographs using deep convolutional neural networks. NPJ Digit Med 2020;3(1):70.
8. Guendel S, Grbic S, Georgescu B, et al. Learning to recognize Abnormalities in Chest X-Rays with Location-Aware Dense Networks. arXiv [cs.CV]. 2018. http://arxiv.org/abs/1803.04565.
9. Kuo PC, Tsai CC, López DM, et al. Recalibration of deep learning models for abnormality detection in smartphone-captured chest radiograph. NPJ Digit Med 2021;4(1):25.
10. Sze-To A, Riasatian A, Tizhoosh HR. Searching for pneumothorax in x-ray images using autoencoded deep features. Sci Rep 2021;11(1):9817.
11. Nabulsi Z, Sellergren A, Jamshy S, et al. Deep learning for distinguishing normal versus abnormal chest radiographs and generalization to two unseen diseases tuberculosis and COVID-19. Sci Rep 2021;11(1):15523.
12. Candemir S, Nguyen XV, Folio LR, Prevedello LM. Training Strategies for Radiology Deep Learning Models in Data-limited Scenarios. Radiol Artif Intell 2021;3(6):e210014.
13. Taori R, Dave A, Shankar V, Carlini N, Recht B, Schmidt L. Measuring robustness to natural distribution shifts in image classification. arXiv [cs.LG]. 2020. https://nicholas.carlini.com/papers/2020_neurips_naturalrobustness.pdf. Accessed April 12, 2021.
14. Rajpurkar P, Joshi A, Pareek A, Ng AY, Lungren MP. CheXternal: Generalization of Deep Learning Models for Chest X-ray Interpretation to Photos of Chest X-rays and External Clinical Settings. arXiv [eess.IV]. 2021. http://arxiv.org/abs/2102.08660.
15. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell 2021;3:610–619.
16. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? arXiv [cs.LG]. 2014. http://arxiv.org/abs/1411.1792.
17. Sowrirajan H, Yang J, Ng AY, Rajpurkar P. MoCo-CXR: MoCo Pretraining Improves Representation and Transferability of Chest X-ray Models. arXiv [cs.CV]. 2020. http://arxiv.org/abs/2010.05352.
18. Chen T, Kornblith S, Norouzi M, Hinton G. A Simple Framework for Contrastive Learning of Visual Representations. arXiv [cs.LG]. 2020. http://arxiv.org/abs/2002.05709.
19. Azizi S, Mustafa B, Ryan F, et al. Big Self-Supervised Models Advance Medical Image Classification. arXiv [eess.IV]. 2021. http://arxiv.org/abs/2101.05224.
20. Lu Y, Jha A, Huo Y. Contrastive Learning Meets Transfer Learning: A Case Study In Medical Image Analysis. arXiv [cs.CV]. 2021. http://arxiv.org/abs/2103.03166.
21. Khosla P, Teterwak P, Wang C, et al. Supervised Contrastive Learning. arXiv [csLG]. 2020; https://arxiv.org/abs/2004.11362.

22. Gazda M, Gazda J, Plavka J, Drotar P. Self-supervised deep convolutional neural network for chest X-ray classification. arXiv [eess.IV]. 2021. http://arxiv.org/abs/2103.03055.
23. EfficientNet Github. https://github.com/tensorflow/tpu/tree/master/models/official/efficientnet. Accessed January 1, 2021.
24. Kolesnikov A, Beyer L, Zhai X, et al. Big Transfer (BiT): General Visual Representation Learning. arXiv [cs.CV]. 2019. http://arxiv.org/abs/1912.11370.
25. Majkowska A, Mittal S, Steiner DF, et al. Chest Radiograph Interpretation with Deep Learning Models: Assessment with Radiologist-adjudicated Reference Standards and Population-adjusted Evaluation. Radiology 2020;294(2):421–431.
26. Summers RM. NIH Chest X-ray Dataset of 14 Common Thorax Disease Categories. https://nihcc.app.box.com/v/ChestXray-NIHCC/file/220660789610. Accessed May 1, 2019.
27. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. arXiv [cs.CV]. 2019. http://arxiv.org/abs/1901.07031.
28. Jaeger S, Karargyris A, Candemir S, et al. Automatic tuberculosis screening using chest radiographs. IEEE Trans Med Imaging 2014;33(2):233–245.
29. Jaeger S, Candemir S, Antani S, Wáng YXJ, Lu PX, Thoma G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quant Imaging Med Surg 2014;4(6):475–477.
30. Candemir S, Jaeger S, Palaniappan K, et al. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. IEEE Trans Med Imaging 2014;33(2):577–590.
31. Wong A, Lee JRH, Rahmat-Khah H, Sabri A, Alaref A. TB-Net: A Tailored, Self-Attention Deep Convolutional Neural Network Design For Detection of Tuberculosis Cases From Chest X-Ray Images. 2021.
32. National Library of Medicine. National Institutes of Health. Bethesda, MD, USA. https://lhncbc.nlm.nih.gov/LHC-downloads/dataset.html. Accessed February 3, 2022.
33. van der Maaten L, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res 2008;9:2579–2605.
34. Duggan GE, Reicher JJ, Liu Y, Tse D, Shetty S. Improving reference standards for validation of AI-based radiography. Br J Radiol 2021;94(1123):20210435.
35. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019;6(1):317.