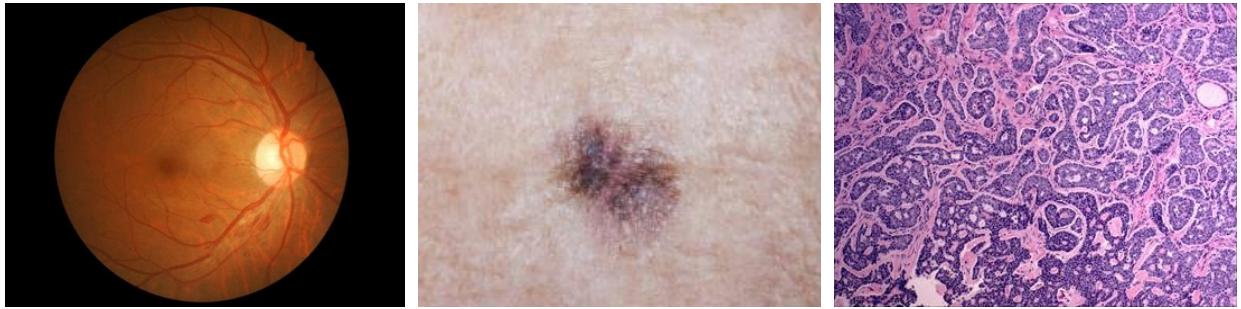
Deep Learning For Medical Image Interpretation

Pranav Rajpurkar
Computer Science Department
Stanford University

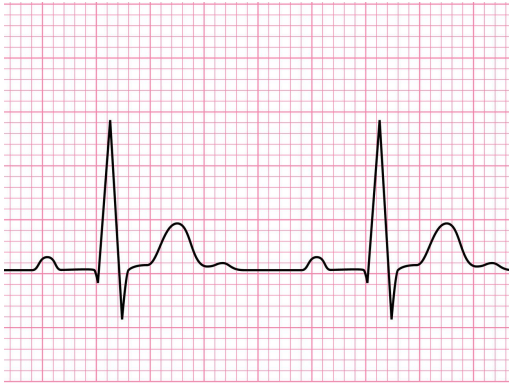
Today: How can we develop deep learning technologies that will be used routinely to improve clinical decision making?

Deep learning algorithms have driven successful application in medical imaging



Diabetic Retinopathy from retinal fundus photos (Gulshan et al., 2016)
Melanomas from photos of skin lesions (Esteva et al., 2017)
Lymph node metastases from H&E slides (Bejnordi et al., 2018)

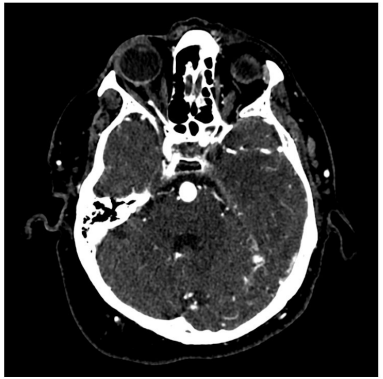
Developing DL that matches/improves clinical experts



Arrhythmia
FDA Cleared

CheXNeXt

MRNet



HeadXNet

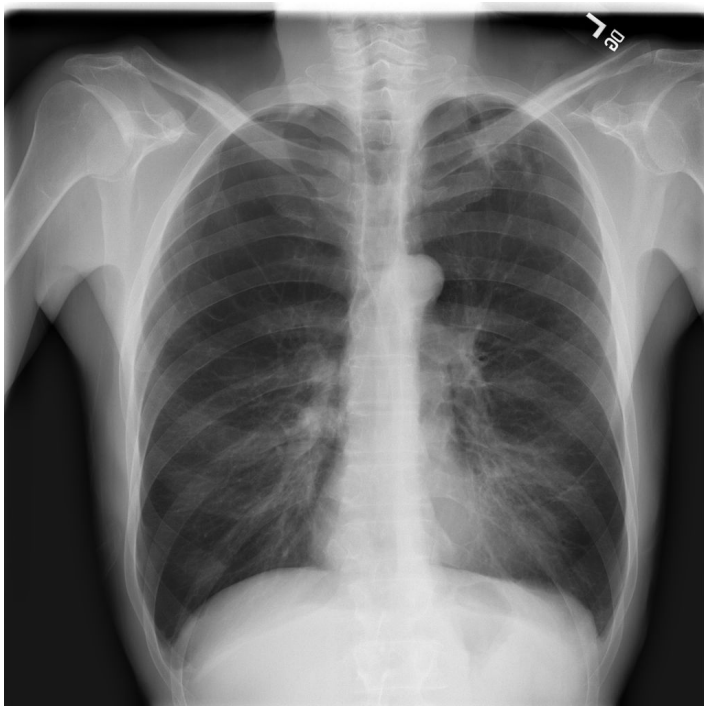
Hannun & Rajpurkar et al., Nature Medicine, 2019
Rajpurkar & Irvin et al., PLOS Medicine, 2018
Bien & Rajpurkar et al., PLOS Medicine, 2018
Park & Chute & Rajpurkar et al., JAMA Network Open. 2019

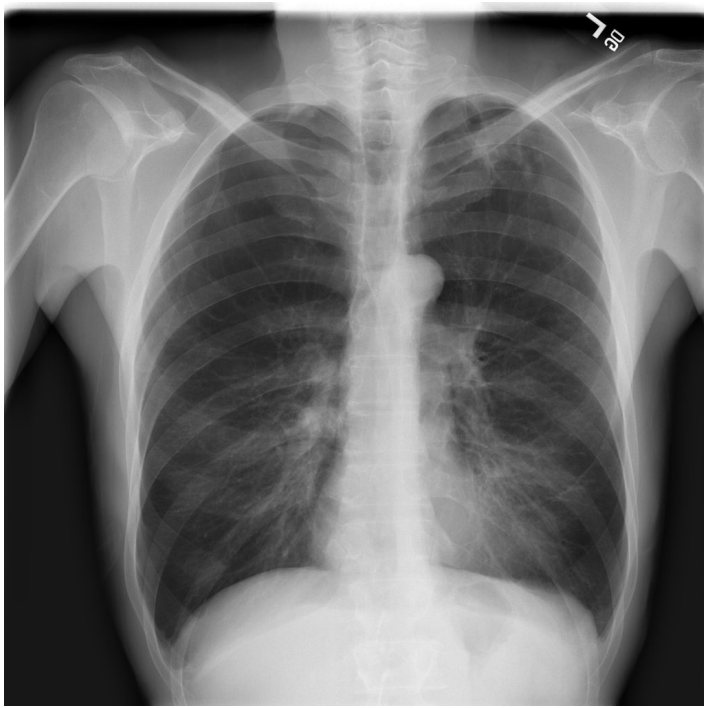
We can develop AI technologies to improve clinical decision-making with:

1. Clinically-informed ML techniques
2. Dataset Design
3. Human-AI Interaction

1. How can clinically-informed **ML techniques** improve medical AI?

Case study: chest x-ray interpretation

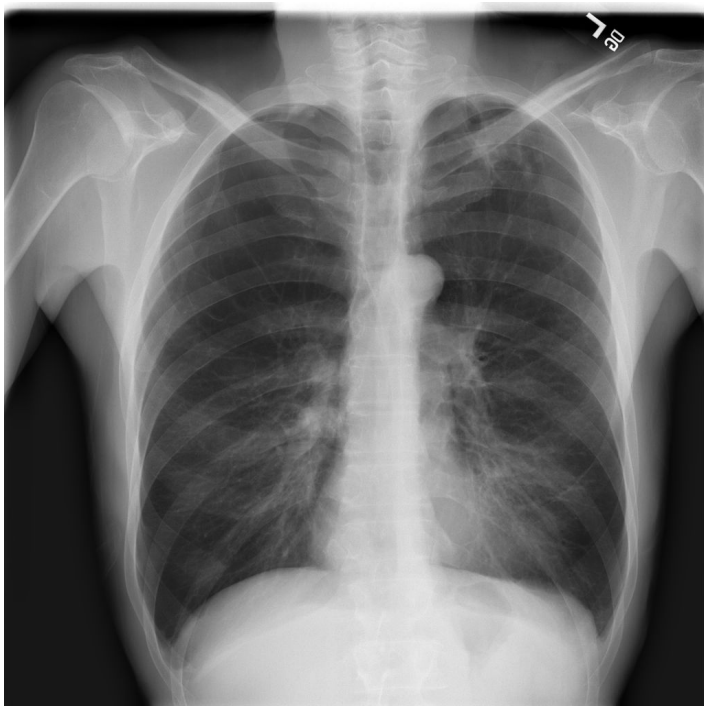




Sep 2017

100,000+ examples

100,000+ examples

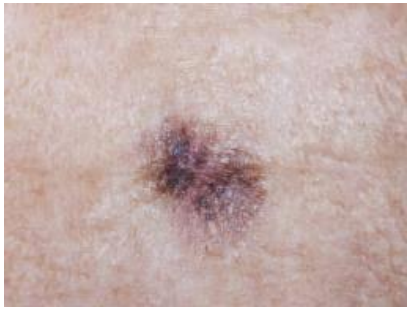


Sep 2017

100,000+ examples

Dec 2016

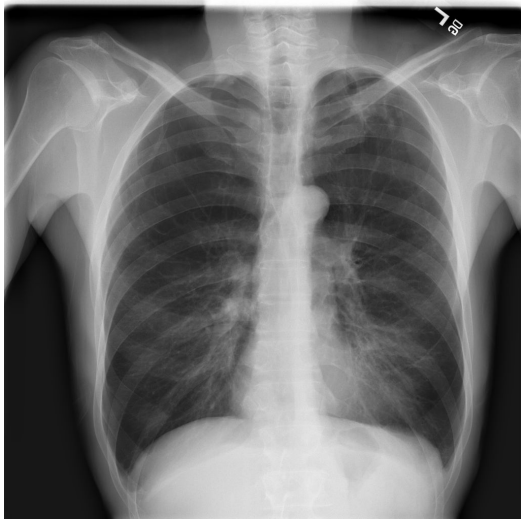
100,000+ examples



Jan 2017

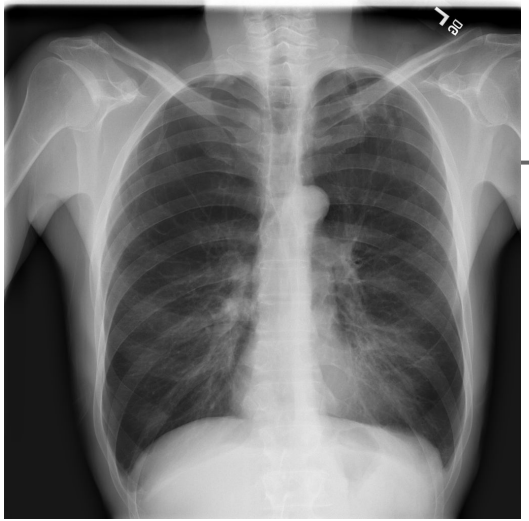
Diabetic Retinopathy from retinal fundus photos (Gulshan et al., 2016)
Melanomas from photos of skin lesions (Esteva et al., 2017)

Chest x-ray input can map to multiple pathologies



Atelectasis
Cardiomegaly
Consolidation
Edema
Effusion
Emphysema
Fibrosis
Hernia
Infiltration
Mass
Nodule
Pleural Thickening
Pneumonia
Pneumothorax

Developed a model to predict pathologies from chest x-ray



CheXNet

Search over:

- CNN architectures
- Class weighting strategies
- Data augmentation

Atelectasis
Cardiomegaly
Consolidation
Edema
Effusion
Emphysema
Fibrosis
Hernia
Infiltration
Mass
Nodule
Pleural Thickening
Pneumonia
Pneumothorax

CheXNet Achieved SOTA performance

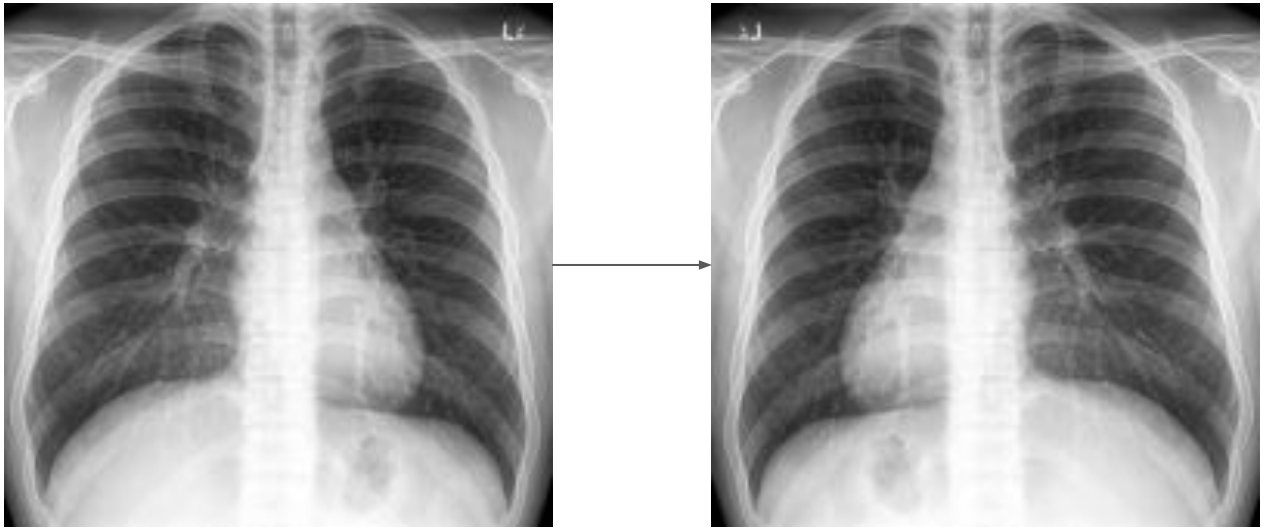
Pathology	Wang et al. (2017)	Yao et al. (2017)	CheXNet (ours)
Atelectasis	0.716	0.772	**0.8094**
Cardiomegaly	0.807	0.904	**0.9248**
Effusion	0.784	0.859	**0.8638**
Infiltration	0.609	0.695	**0.7345**
Mass	0.706	0.792	**0.8676**
Nodule	0.671	0.717	**0.7802**
Pneumonia	0.633	0.713	**0.7680**
Pneumothorax	0.806	0.841	**0.8887**
Consolidation	0.708	0.788	**0.7901**
Edema	0.835	0.882	**0.8878**
Emphysema	0.815	0.829	**0.9371**
Fibrosis	0.769	0.767	**0.8047**
Pleural Thickening	0.708	0.765	**0.8062**
Hernia	0.767	0.914	**0.9164**

What can SOTA be attributed to?

- DenseNet, pretrained and all-layers finetuned
- Weighted multi-label CE Loss on each category
- Random horizontal flipping, and learning rate decay

- **Early signal that DenseNet121 perform really well on medical imaging tasks.**
- **Combination of architecture and training strategy re-discovered in other datasets.**

The hidden cost of data-augmentation (Normal or Dextrocardia)

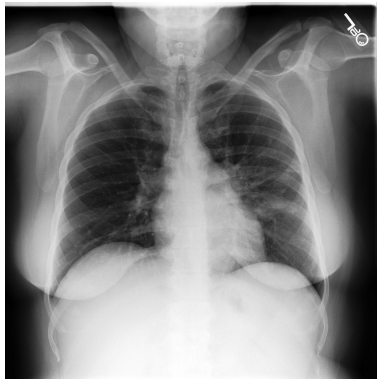


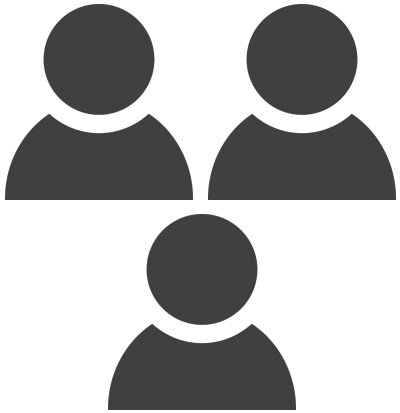
Normal

Dextrocardia

How to set up performance vs radiologists?

Test Set



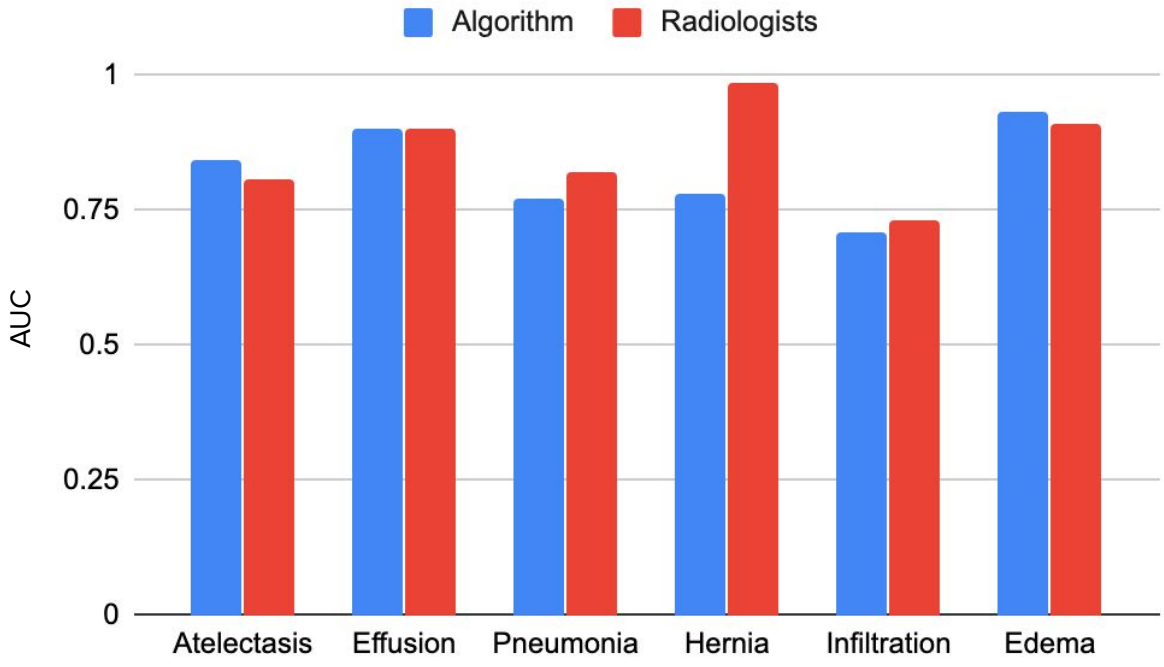


X 9

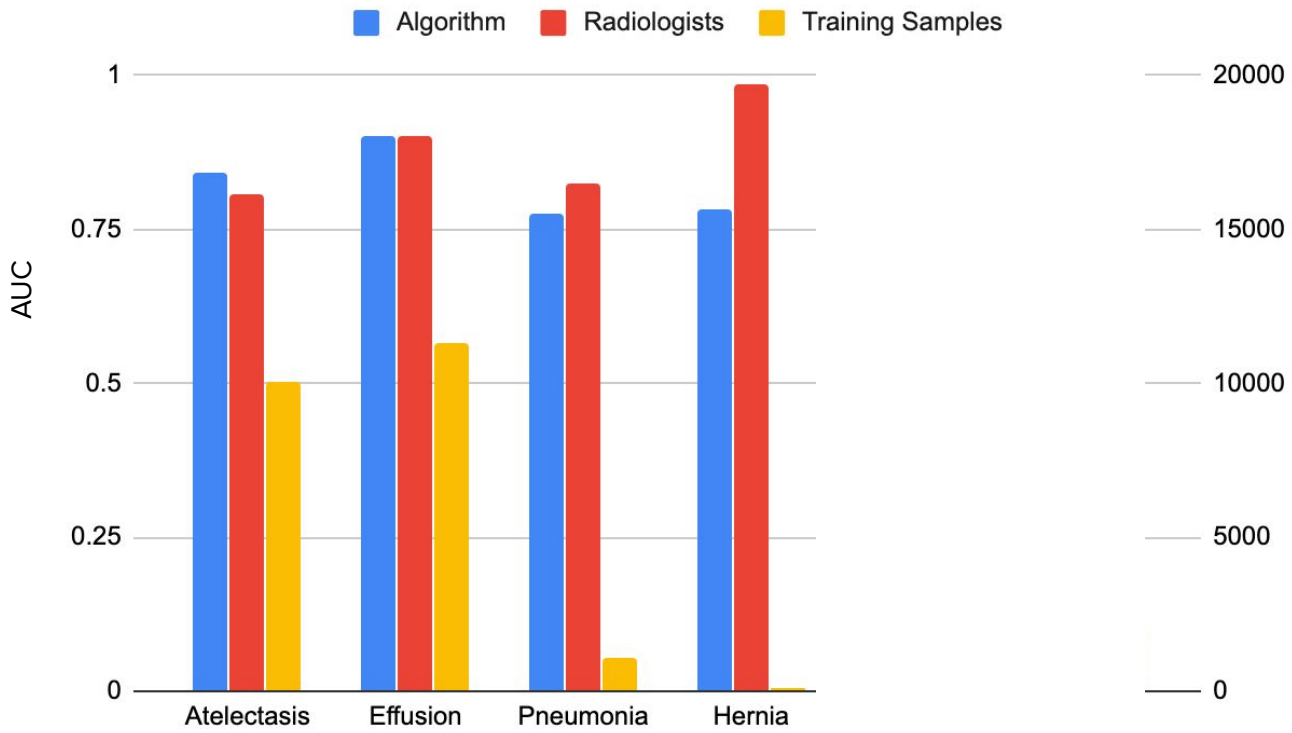
Human Performance

Diabetic Retinopathy from retinal fundus photos (Gulshan et al., 2016)
Melanomas from photos of skin lesions (Esteva et al., 2017)

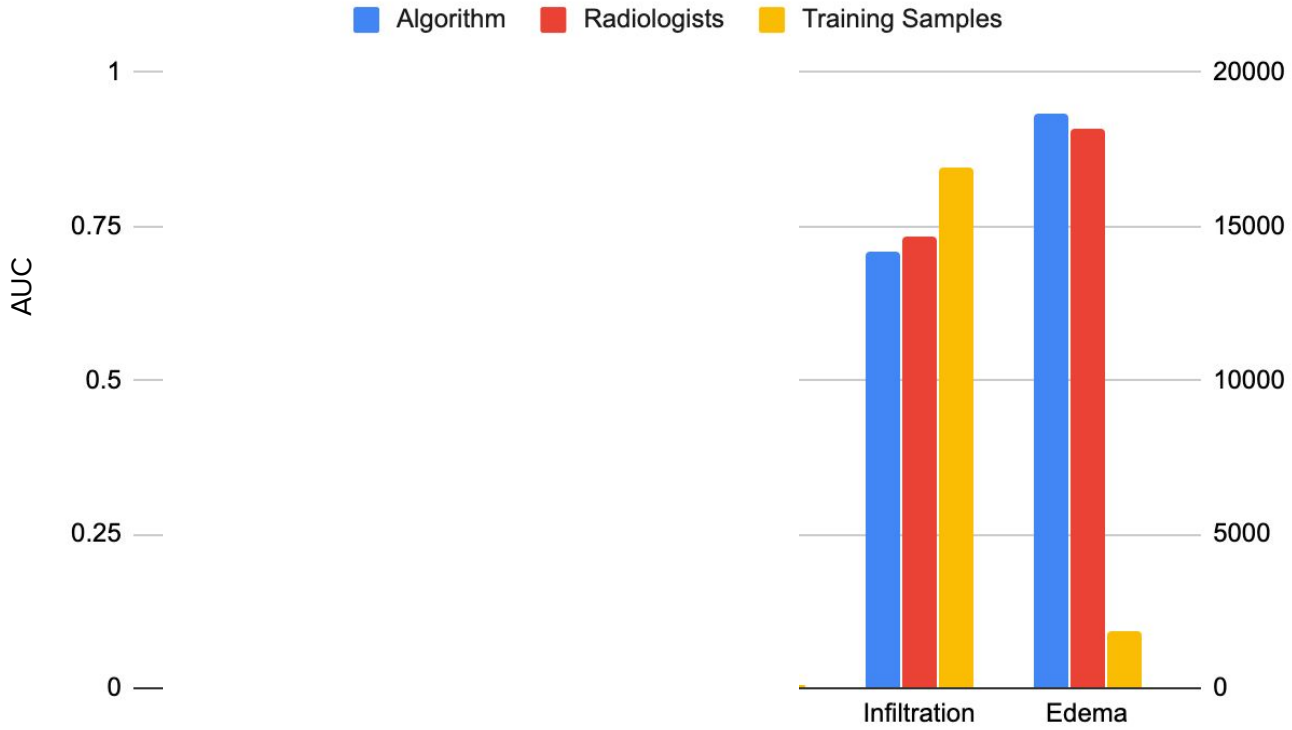
Algorithm performs comparably on some, poorer on others. Why?



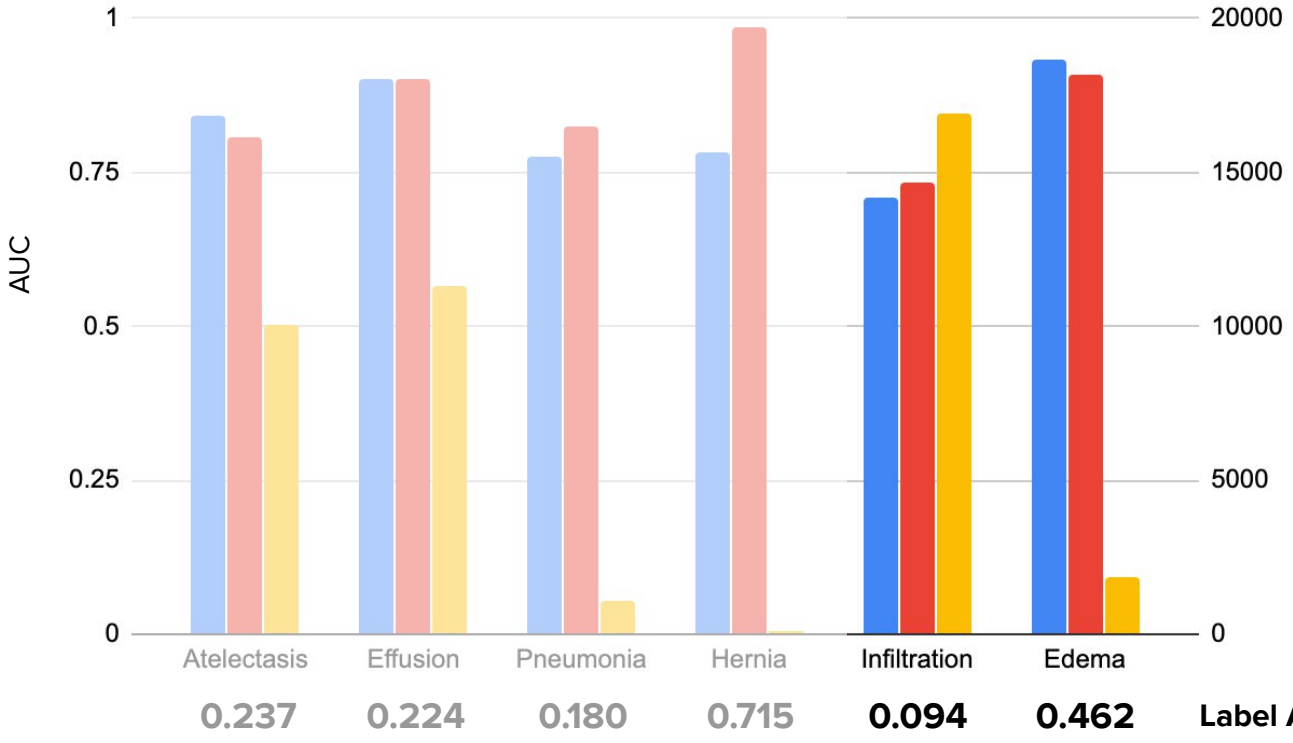
Having more prevalent categories helps



But there are exceptions. Why?

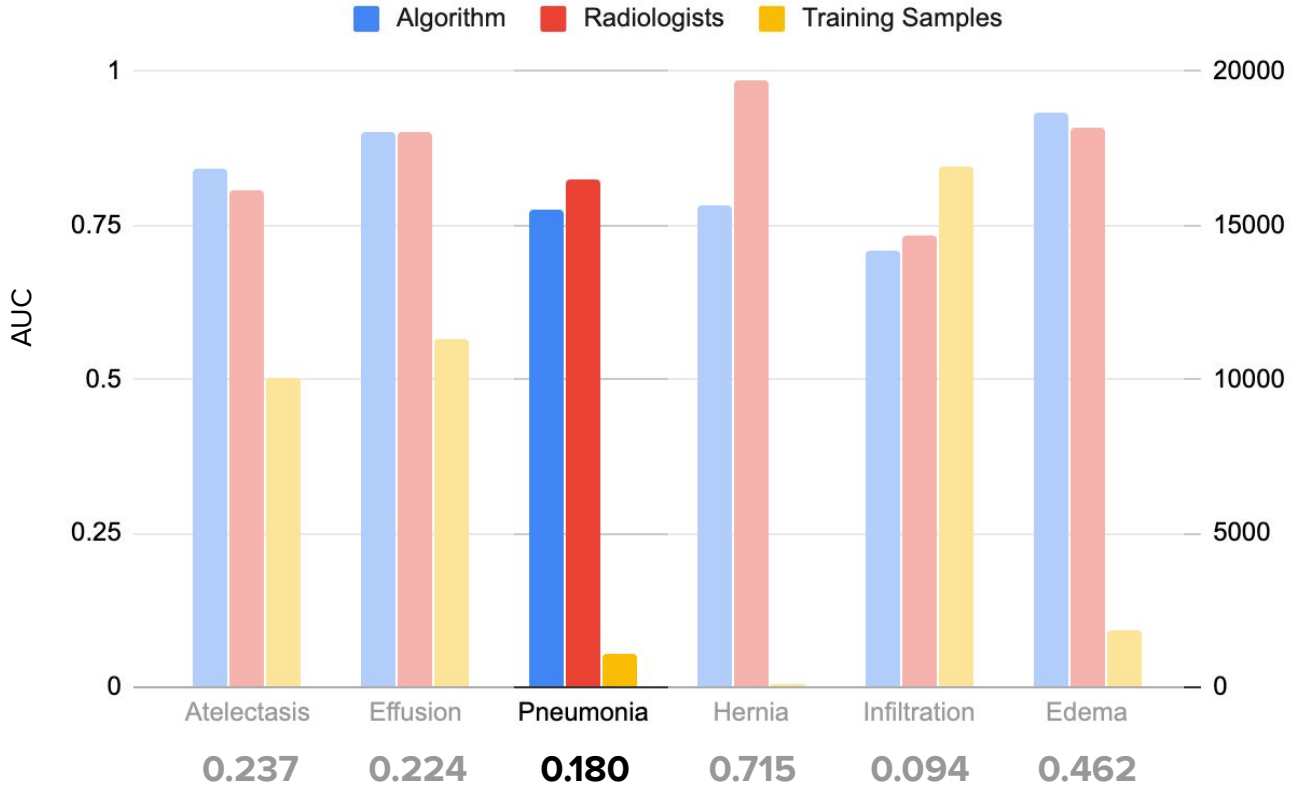


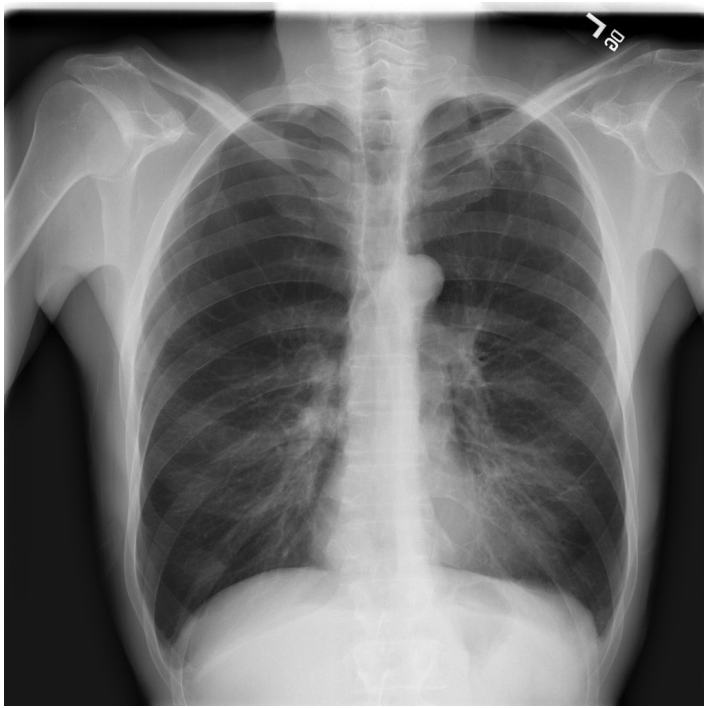
Labeling Agreement With the Ground truth Can Explain



Label Agreement with Ground Truth

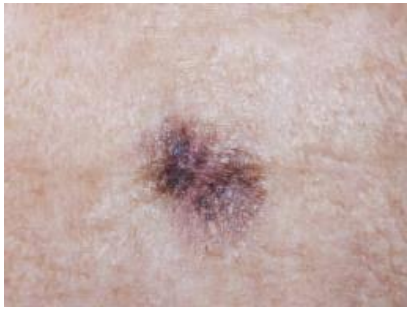
Can clinical insights help achieve radiologist-level performance on pneumonia?





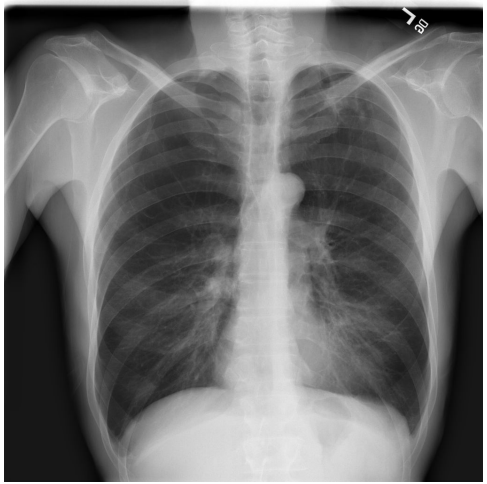
Automatically labeled using NLP on radiology reports

Manual Expert labeled



Manual Expert labeled

Image is automatically labeled using NLP on corresponding radiology report



Low lung volumes. Mild pulmonary edema. Patchy bibasilar atelectasis with infection in the right lung base not excluded in the correct clinical setting.

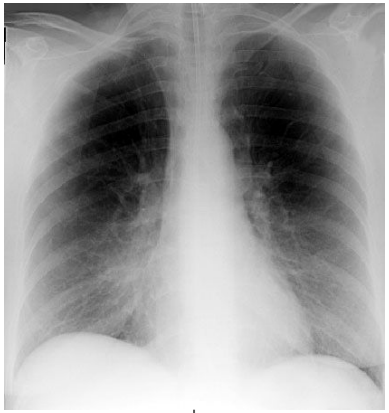
Report labeling NLP System

Atelectasis
Cardiomegaly
Consolidation
Edema
Effusion
Emphysema
Fibrosis
Hernia
Infiltration
Mass
Nodule
Pleural Thickening
Pneumonia
Pneumothorax

Peng Y et al. NegBio. AMIA 2018 Informatics Summit. 2018.
Wang X et al. ChestX-ray8. CVPR 2017.

NLP System uses rules on the dependency parse to label whether pneumonia present

Low lung volumes. Mild pulmonary edema. Patchy bibasilar atelectasis with infection in the right lung base not excluded in the correct clinical setting.



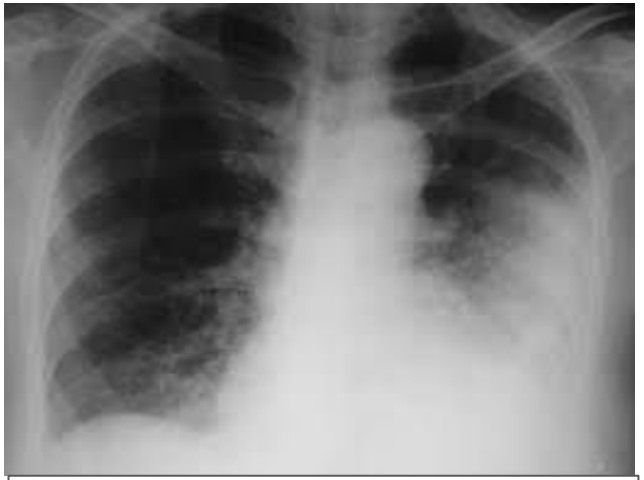
Report labeling NLP System

Is pneumonia present / absent / uncertain ?

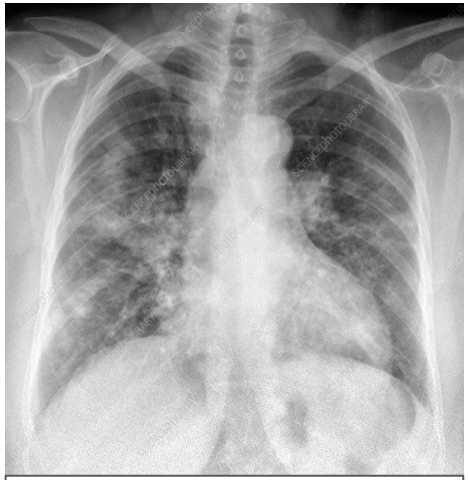
Labeled 1 if present
0 if absent / uncertain

Peng Y et al. NegBio. AMIA 2018 Informatics Summit. 2018.
Wang X et al. ChestX-ray8. CVPR 2017.

But pneumonia mostly expressed with uncertainty!



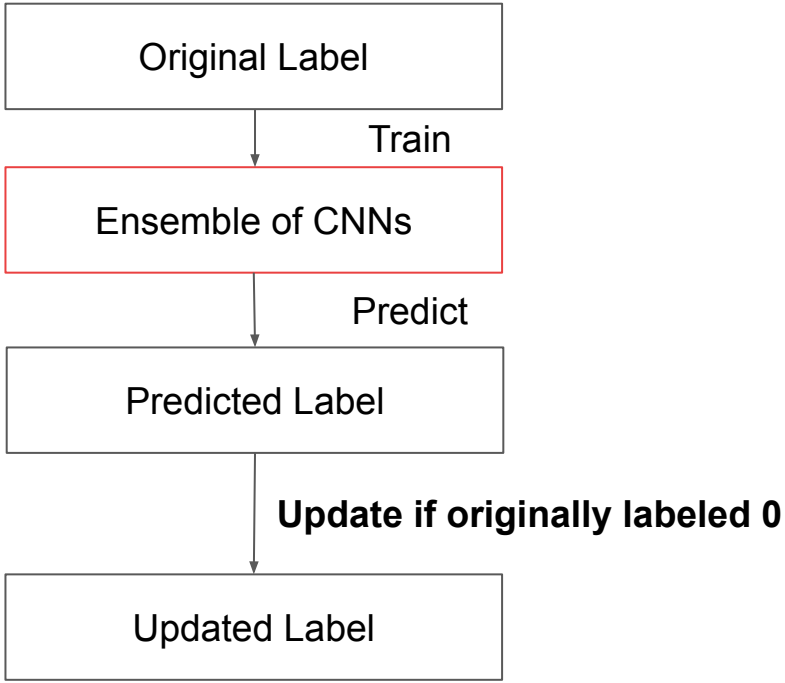
Labeled 0

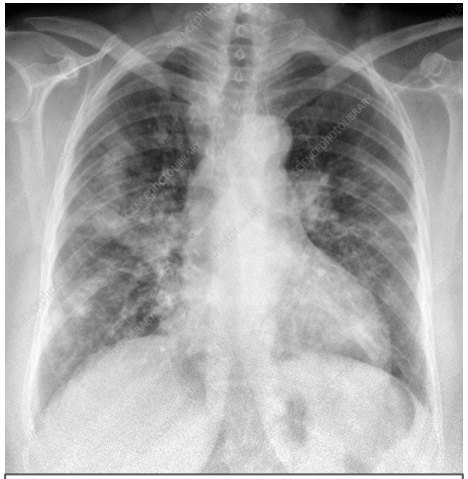


Labeled 0

Peng Y et al. NegBio. AMIA 2018 Informatics Summit. 2018.
Wang X et al. ChestX-ray8. CVPR 2017.

Key idea is to relabel the negatives using the vision algorithm

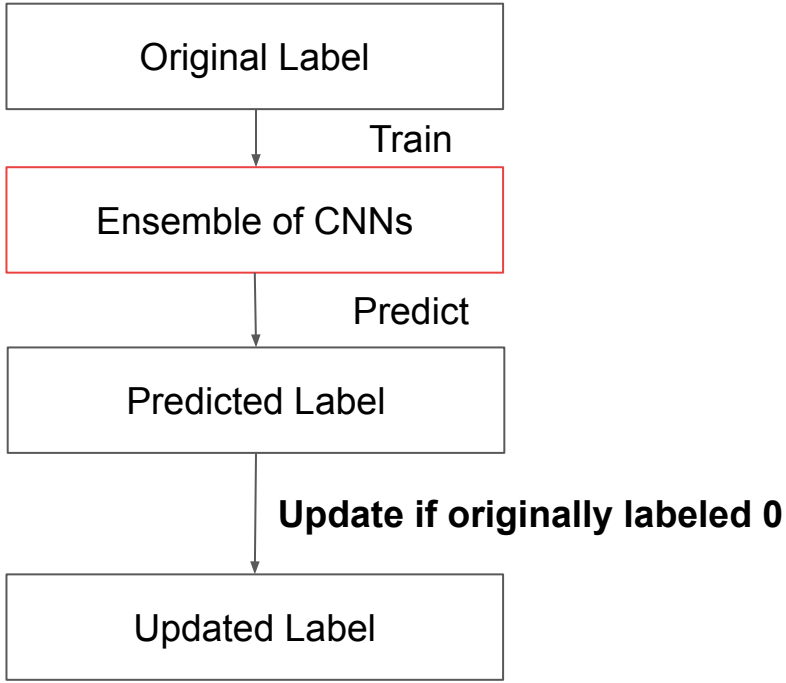




Original Labeled 0
Now 1

Yarowsky et al. Unsupervised word sense disambiguation rivaling supervised methods. ACL. 1995.

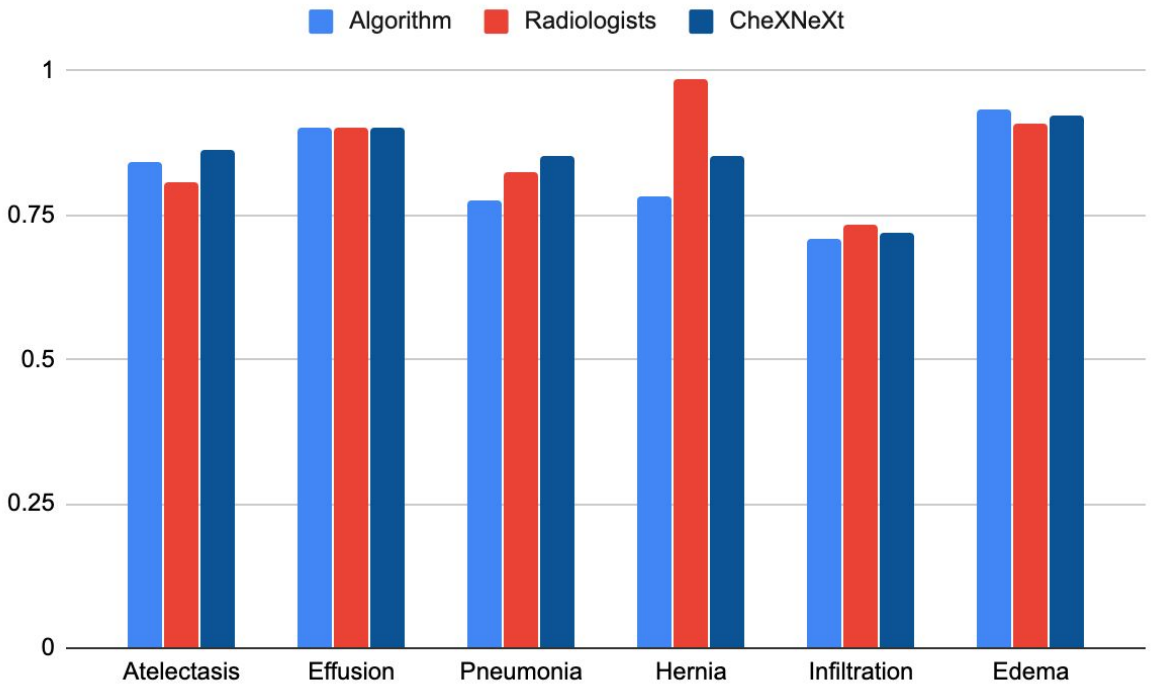
Increases prevalence of pneumonia. But does relabeled dataset improve model?



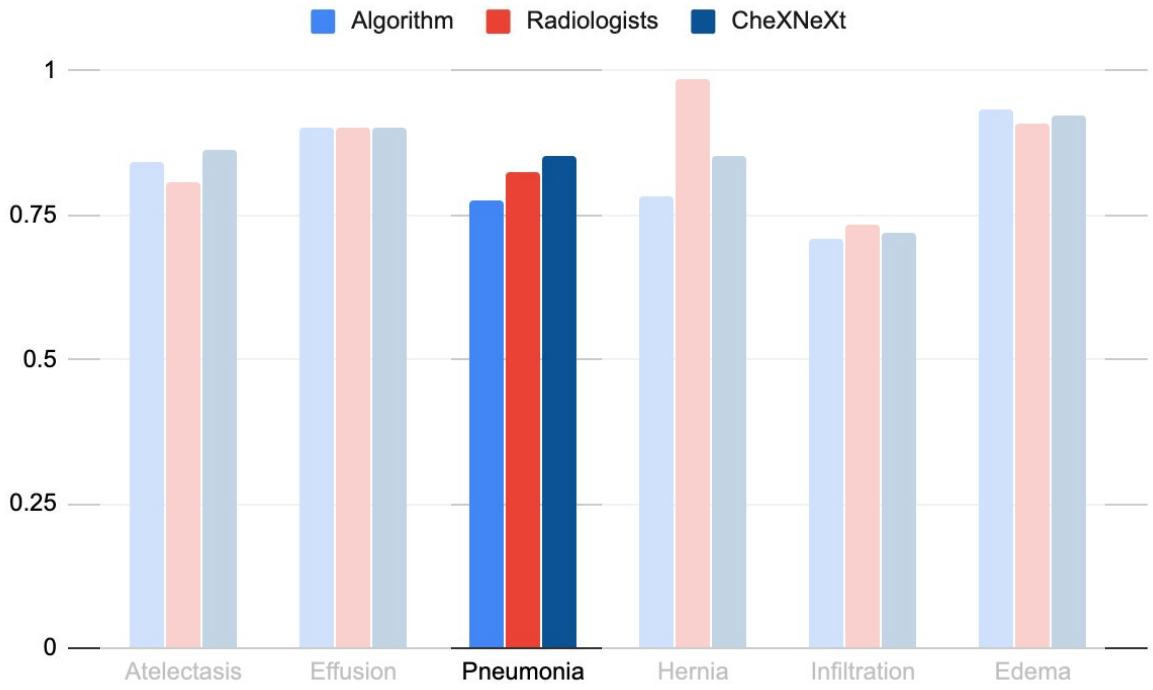
Original Pneumonia # Examples (%)	1107 (1.1)
Updated Pneumonia # Examples (%)	9838 (10.0)

Yarowsky et al. Unsupervised word sense disambiguation rivaling supervised methods. ACL. 1995.

Performance improves on 11 of 14 categories

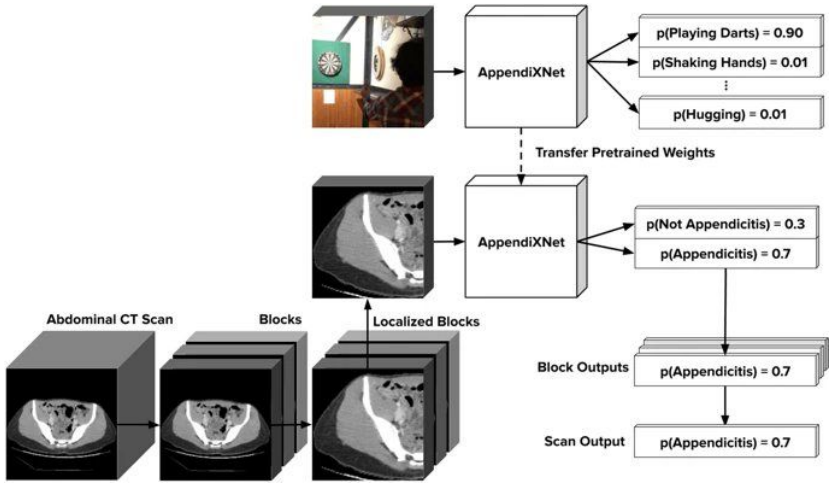


Now radiologist level performance on pneumonia achieved

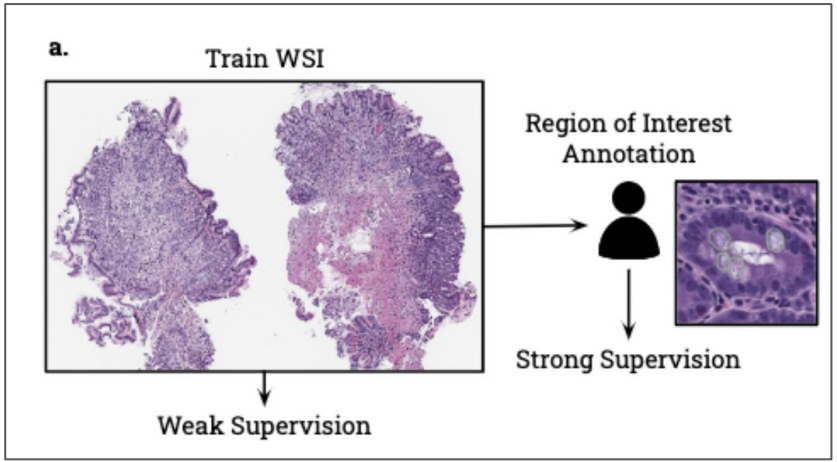


The CheXNeXt algorithm performed as well as the radiologists for 10 of 14 pathologies and performed better than the radiologists on 1 pathology.

Other examples of clinically informed ML



Can pre-training on videos improve the performance of 3D convolutional neural network on appendicitis?

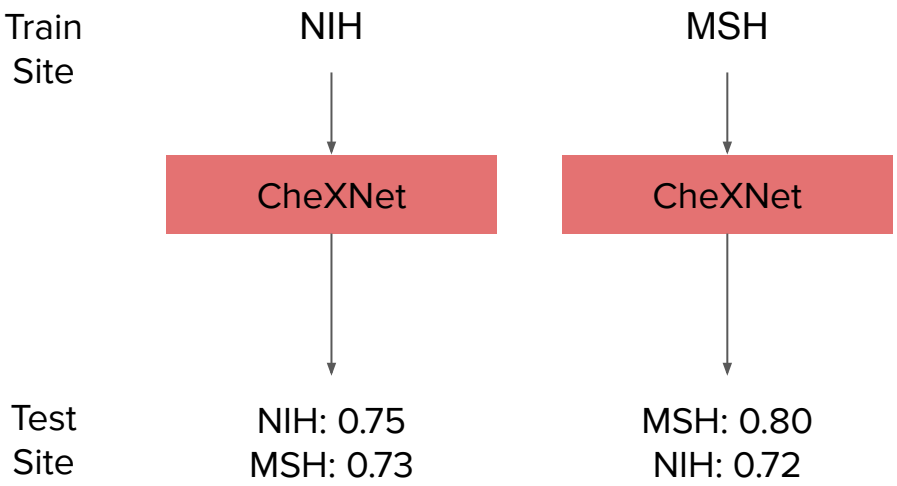


How does the collection of expert region of interest annotations affect the performance of algorithm at different magnification levels?

Rajpurkar & Irvin & Park et al. Scientific Reports. 2020

2. How can **dataset design** improve medical AI?

Although reproducible training procedure, drop on external institution

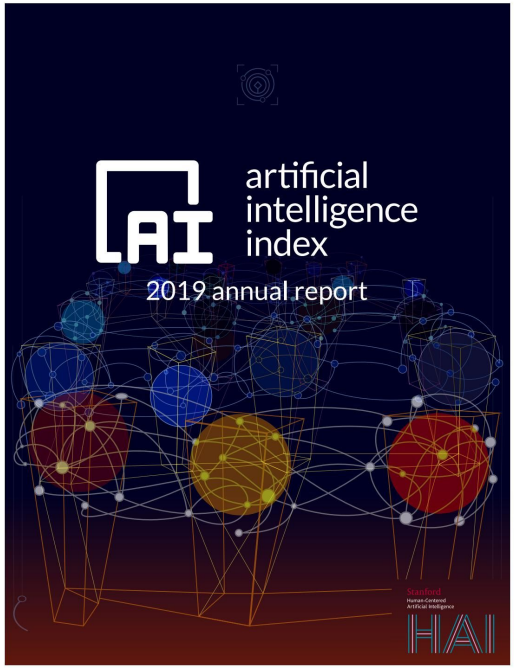


Zech & Badgeley et al., PLOS Medicine, 2018

Do we just need a bigger, cleaner dataset for cross institution generalization?

Lessons from building Stanford Question Answering Dataset (SQuAD)

A large reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles



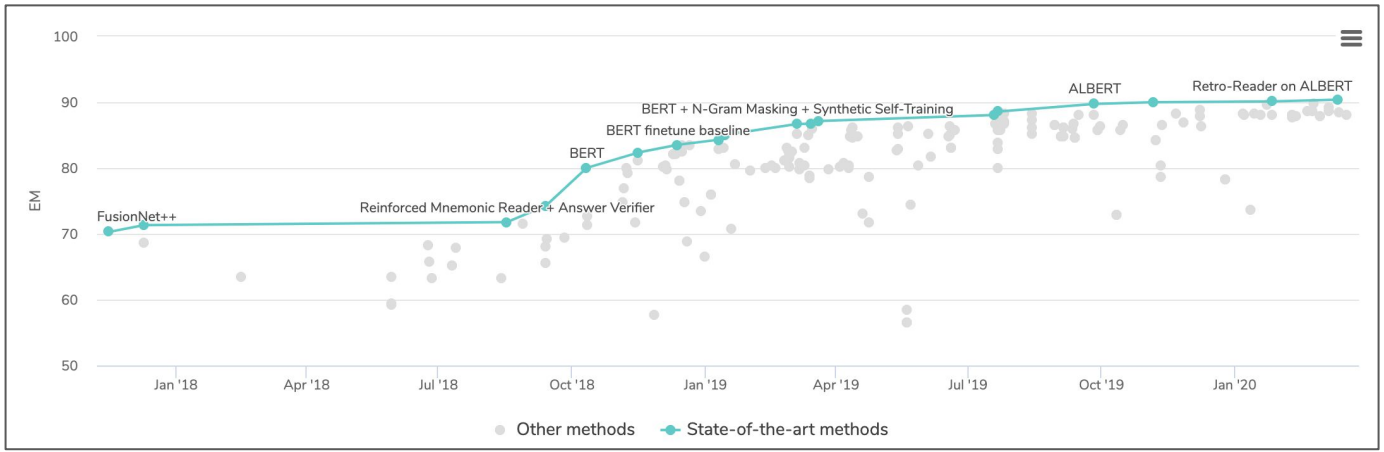
Rajpurkar et al., EMNLP, 2016
Rajpurkar & Jia et al., ACL, 2018

Drove advancements in deep learning models from top institutions and companies

Google AI

Microsoft

facebook Artificial Intelligence



Advancements part of our search engine. Driven by scale and high label quality?

"**BERT** will help Search better understand one in 10 searches in the U.S. in English, and we'll bring this to more languages and locales over time." -- Google AI Blog

"Microsoft is already applying earlier versions of the models that were submitted for the SQuAD dataset leaderboard in its **Bing search** engine, and the company is working on applying it to more complex problems." -- Microsoft AI Blog

Has been a driving force for advancements in language representations, network architectures, and transfer to other tasks.

→ Methods it drove required scale and high label quality.

Rajpurkar et al., EMNLP, 2016, *1410 citations*
Rajpurkar & Jia et al., ACL, 2018, *299 citations*

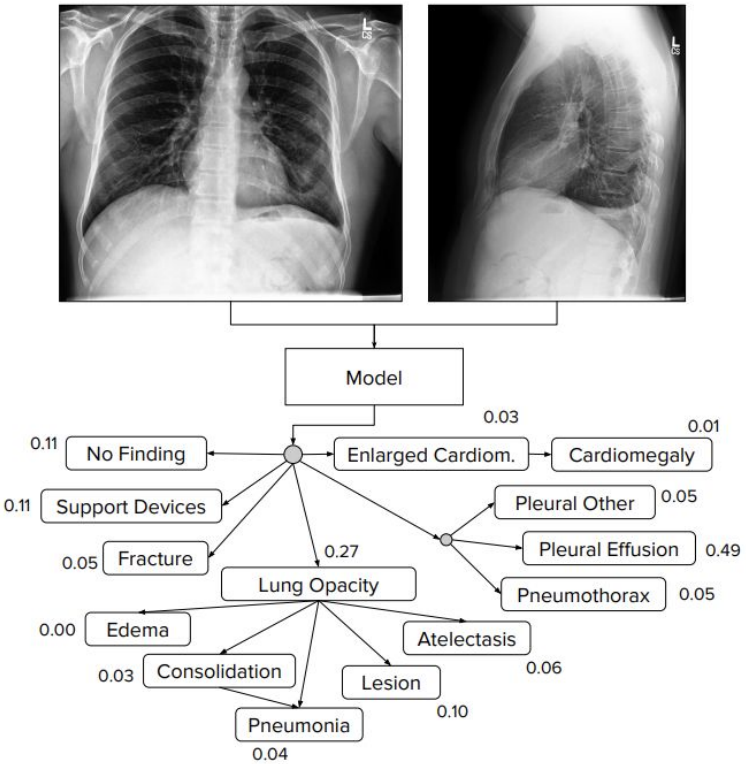
CheXpert

200k Chest X-Rays

Richer / Cleaner Labels

Strong Val+Test Ground Truth

Expert Comparison



Irvin & Rajpurkar et al. 2019, AAAI

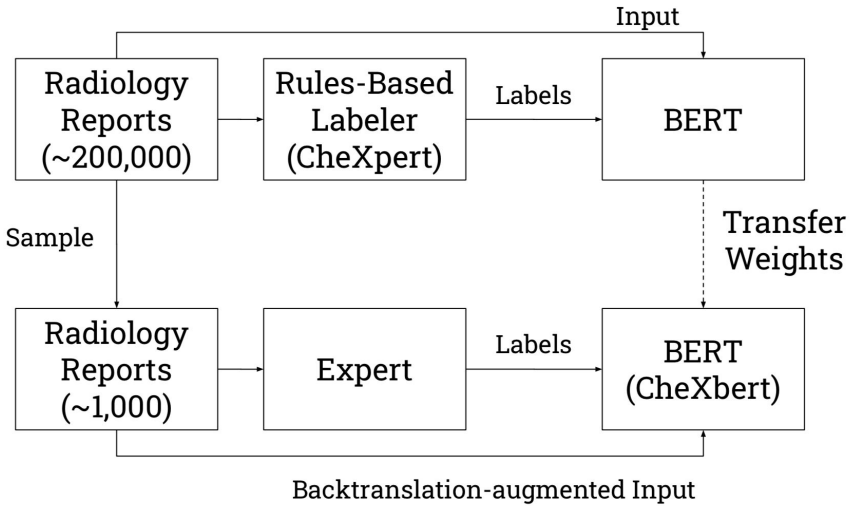
Improve on NegBio's NLP Tool to achieve SOTA with CheXpert labeler

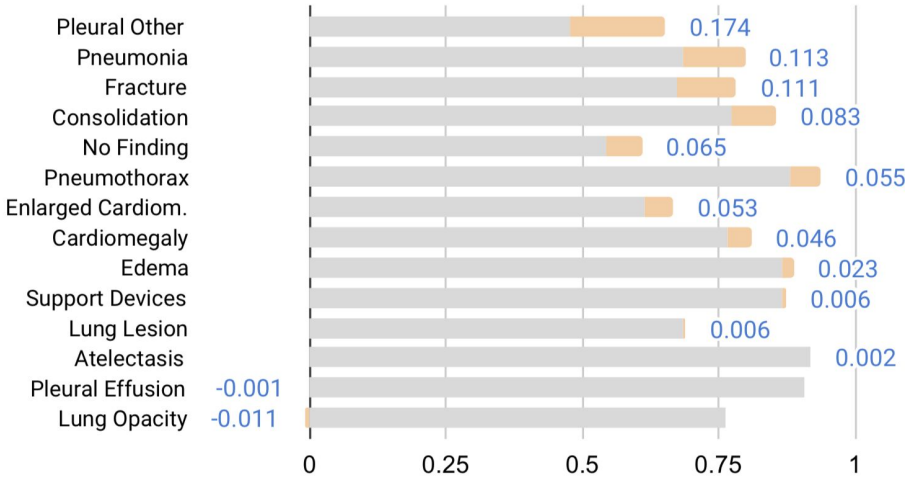
- Comprehensive error analysis to identify gaps in rules.
- Open-sourced, and now used to label some of largest available datasets.

	Mention		Negation		Uncertainty	
Category	NegBio	CheXpert	NegBio	CheXpert	NegBio	CheXpert
Atelectasis	0.930	**0.998**	**0.727**	0.400	0.379	**0.835**
Cardiomegaly	0.596	**0.954**	0.043	**0.830**	0.000	**0.333**
Consolidation	0.966	**0.986**	0.917	**0.958**	0.235	**0.486**
Edema	0.855	**0.996**	0.701	**0.878**	0.214	**0.742**
Pleural Effusion	0.971	**0.987**	0.873	**0.947**	0.368	**0.500**
Pneumonia	0.836	**0.981**	0.750	**0.785**	0.388	**0.674**
Pneumothorax	0.983	**0.998**	**0.951**	0.948	0.182	**0.286**

Johnson AE et al., Scientific Data. 2019

Newer labeler uses BERT on existing feature-engineered systems with expert annotations





Development time lowered from months to days.

Smit & Jain & Rajpurkar et al., 2020

Explicitly retain the uncertainty output

1. *unremarkable* cardiomediastinal silhouette

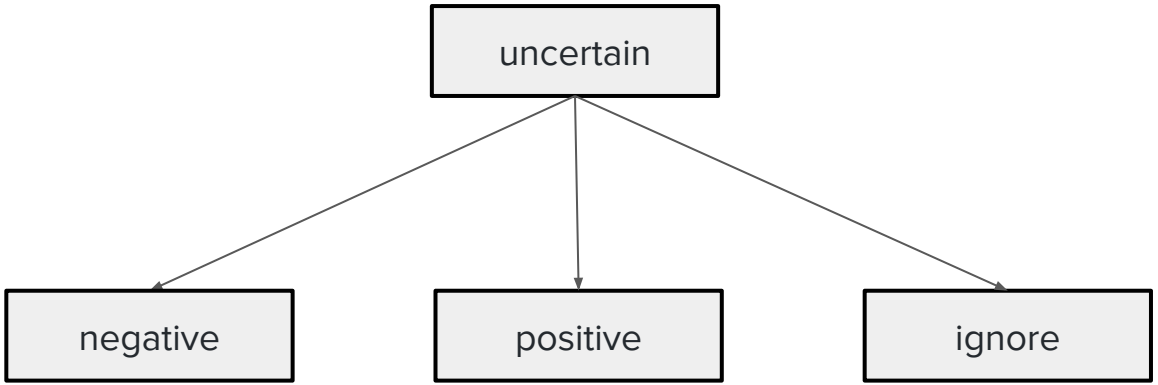
2. diffuse reticular pattern, which can be seen with an atypical infection **or** chronic fibrotic change. *no* focal consolidation.

3. *no* pleural effusion or pneumothorax

4. mild degenerative changes in the lumbar spine and old right rib fractures.

Observation	Labeler Output
No Finding	
Enlarged Cardiom.	0
Cardiomegaly	
Lung Opacity	1
Lung Lesion	
Edema	
Consolidation	0
Pneumonia	u
Atelectasis	
Pneumothorax	0
Pleural Effusion	0
Pleural Other	
Fracture	1
Support Devices	

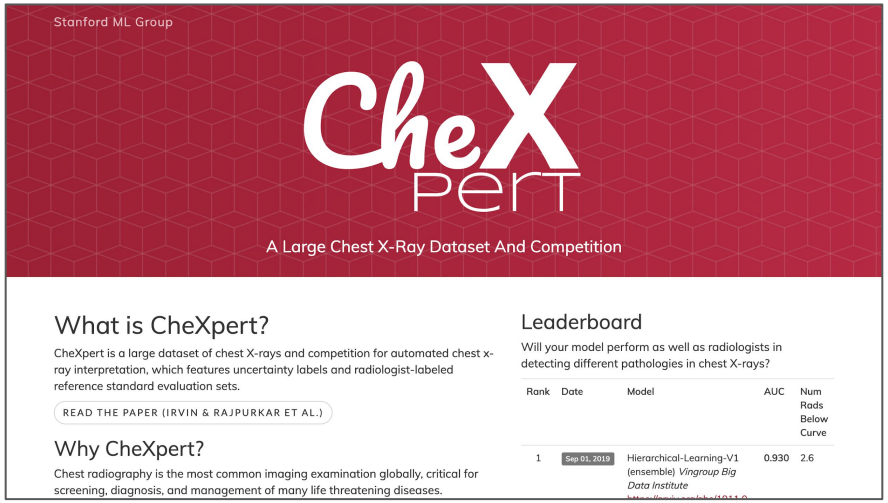
Using uncertainty output better than treating as negative for many classes



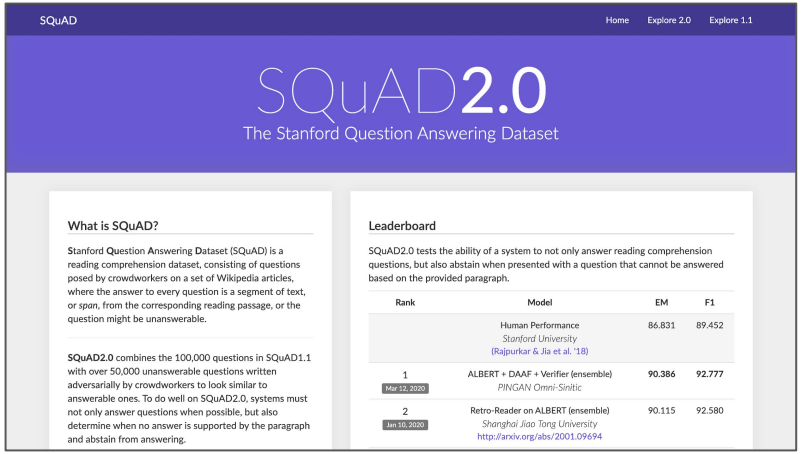
Best Label for	Cardiomegaly	Atelectasis, Edema, Effusion	Consolidation

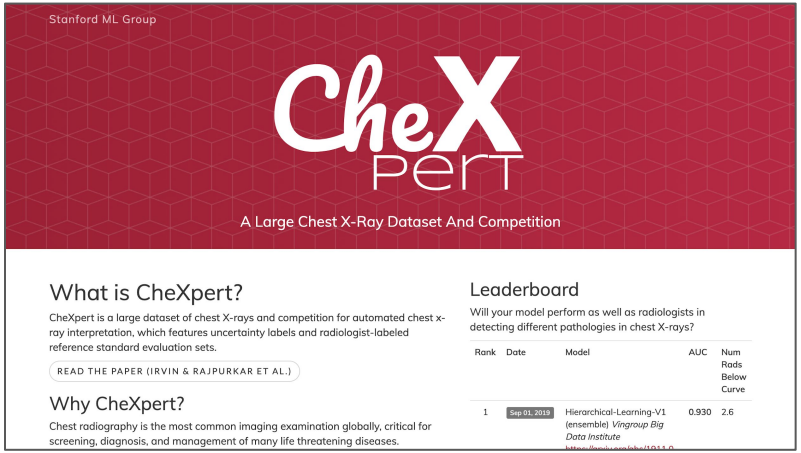
Current SOTA on Chexpert explicitly uses label smoothing regularization.

Released as open competition to the world



Public leaderboard where teams submit on a hidden test set





Rajpurkar & Irvin et al. 2018, MIDL
Bien & Rajpurkar et al, PLOS Medicine

Top models have significantly outperformed baseline. Are they truly more general?

3250 Users
130 Teams Competing

73	Jan 23, 2019	Stanford Baseline (ensemble) *Stanford University* https://arxiv.org/abs/1901.07031	0.907

Rank	Date	Model	AUC
1	Sep 01, 2019	Hierarchical-Learning-V1 (ensemble) *Vingroup Big Data Institute* https://arxiv.org/abs/1911.06475	0.930
2	Oct 15, 2019	Conditional-Training-LSR *ensemble*	0.929
3	Dec 04, 2019	Hierarchical-Learning-V4 (ensemble) *Vingroup Big Data Institute* https://arxiv.org/abs/1911.06475	0.929
4	Oct 10, 2019	YWW(ensemble) *JF&NNU* https://github.com/jfhealthcare/Chexpert	0.929

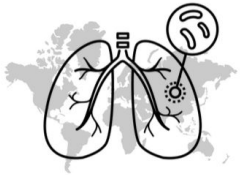
We investigated generalization performance of top leaderboard models

CheXpedition

Top 10 Chest X-Ray Models

14 observations, digital x-rays, single institution data

Translation Challenges

1.  Generalization to **TB Detection**

2.  Generalization to **Smartphone Photos**

3.  Generalization to **External Institution**

Rajpurkar & Joshi et al., preprint. 2020

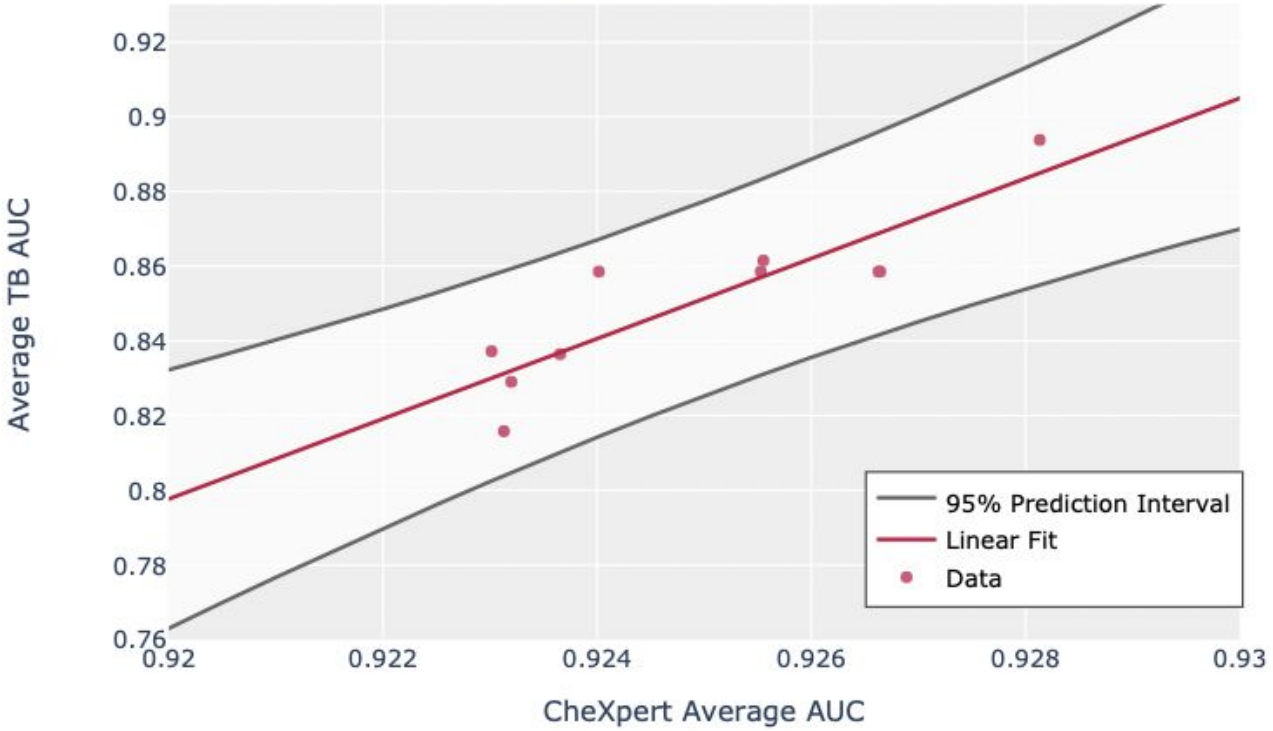
Could models detect TB, a label not included in model training?

- Global Health Relevance
- Consolidation Proxy

Yes! Performance competitive with when models are directly trained on those datasets.

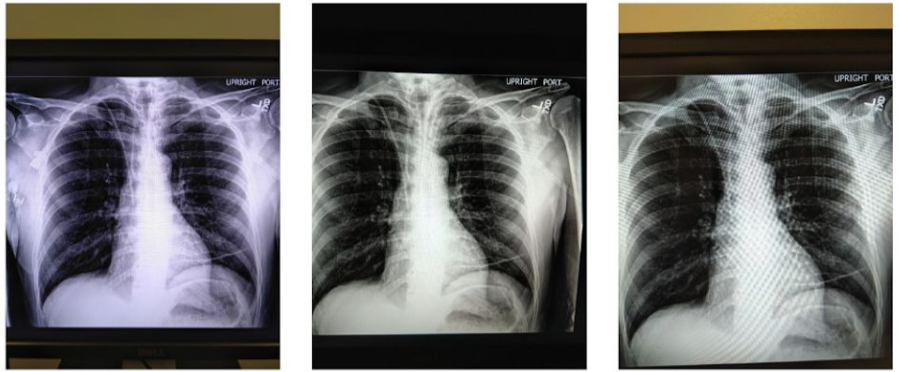
- Avg AUC on 2 TB test datasets: 0.82 and 0.89.

Avg performance across tasks predicts performance

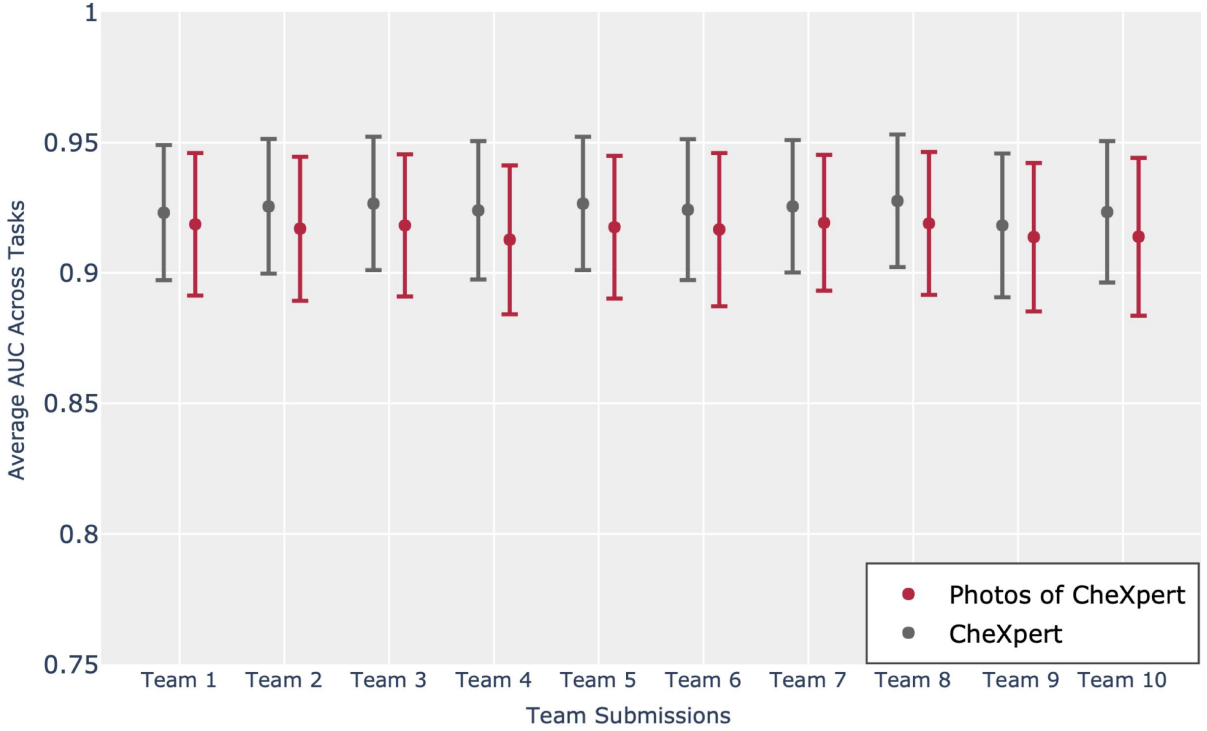


Could models detect diseases on smartphone photos of chest x-rays?

- Appealing solution to scaled deployment.
- Performance drop not thoroughly investigated in medical imaging.



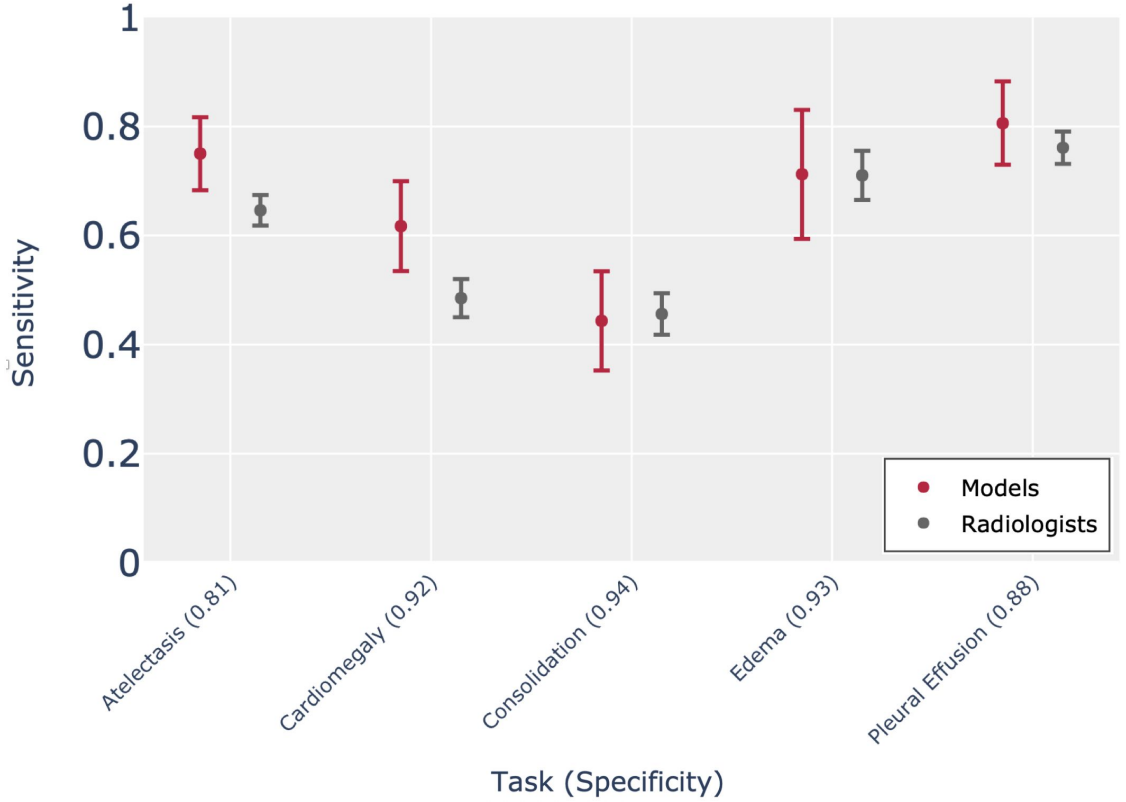
Yes! < 0.01 drop in performance.

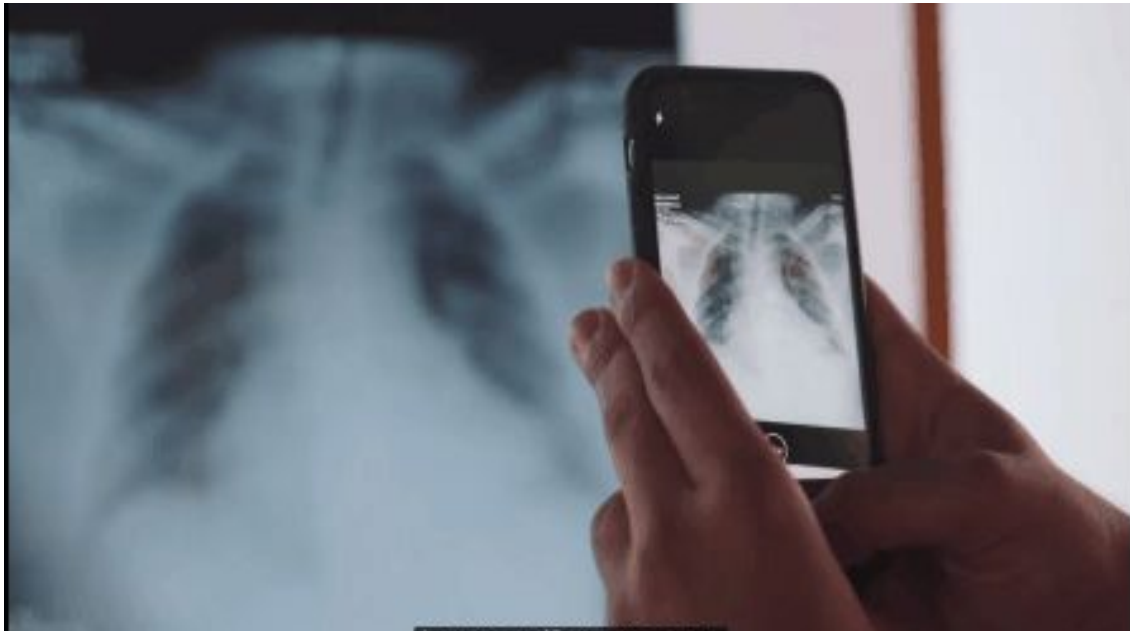


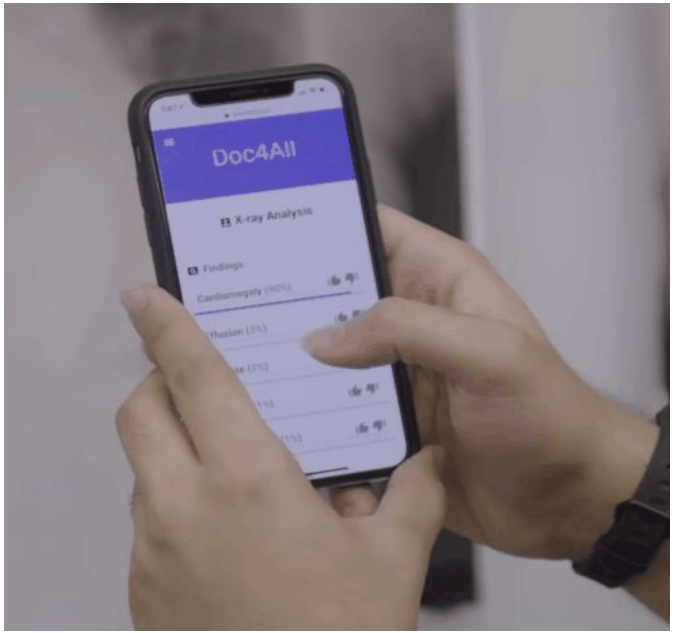
Can models achieve radiologist level performance on external institution w/ zero fine-tuning?

If so, first demonstration of that effect!

Yes! At radiologist specificity, higher model sensitivity for 4 of 5 tasks



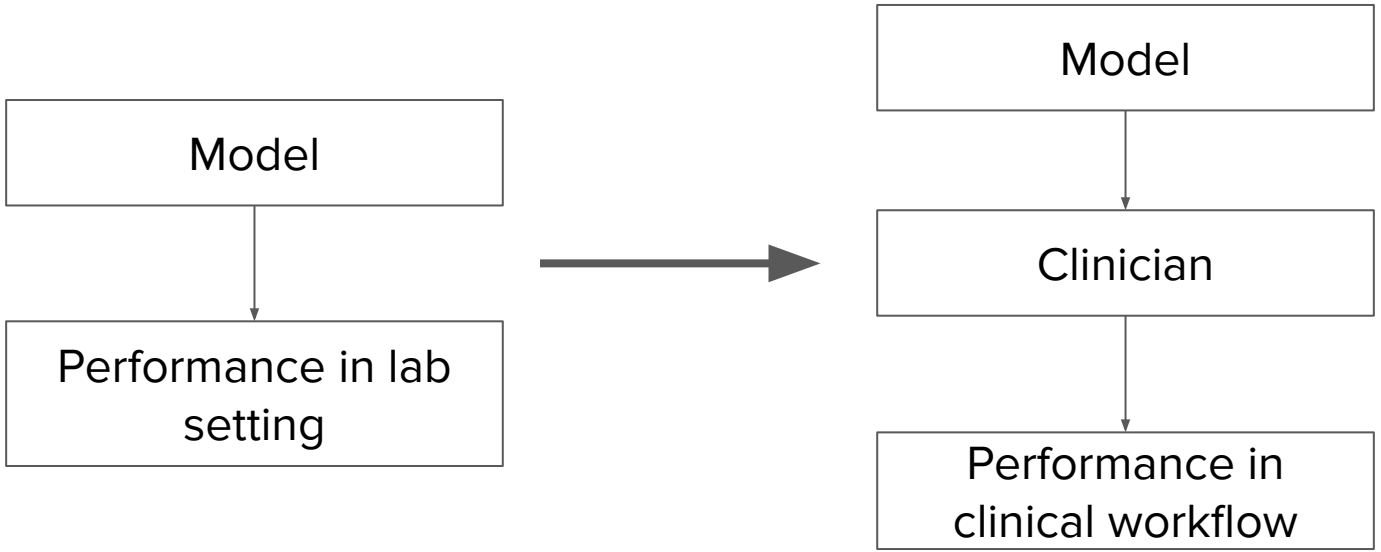




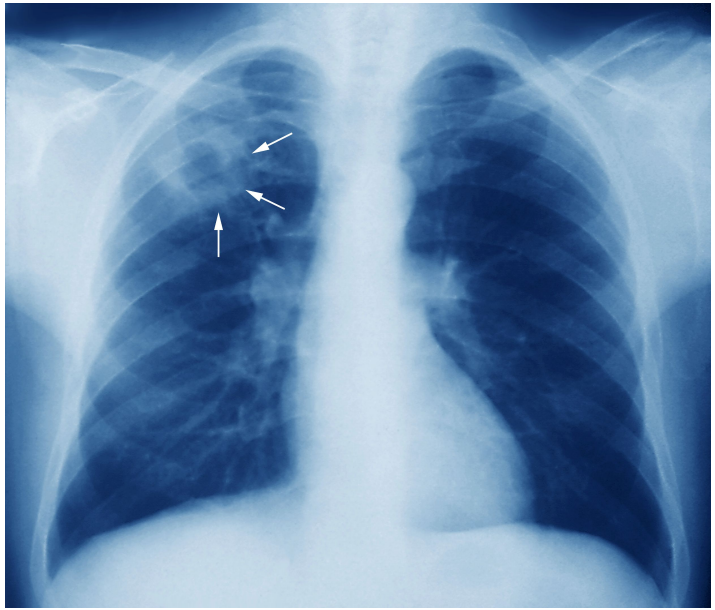
Link1 Link2 Link3

3. How can **Human-AI interaction** be designed for clinical decision-making?

Would AI algorithms necessarily improve clinicians on diagnostic tasks?

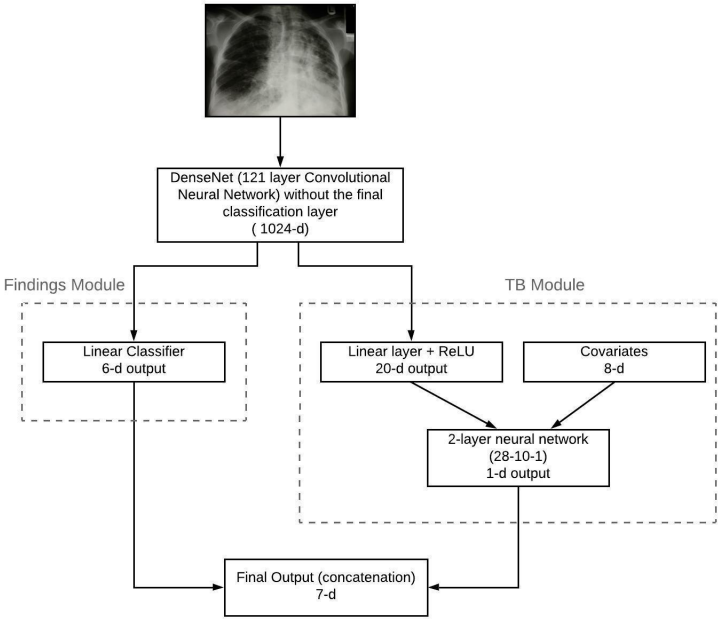


Can DL improve physician performance on TB detection in HIV+ patients?



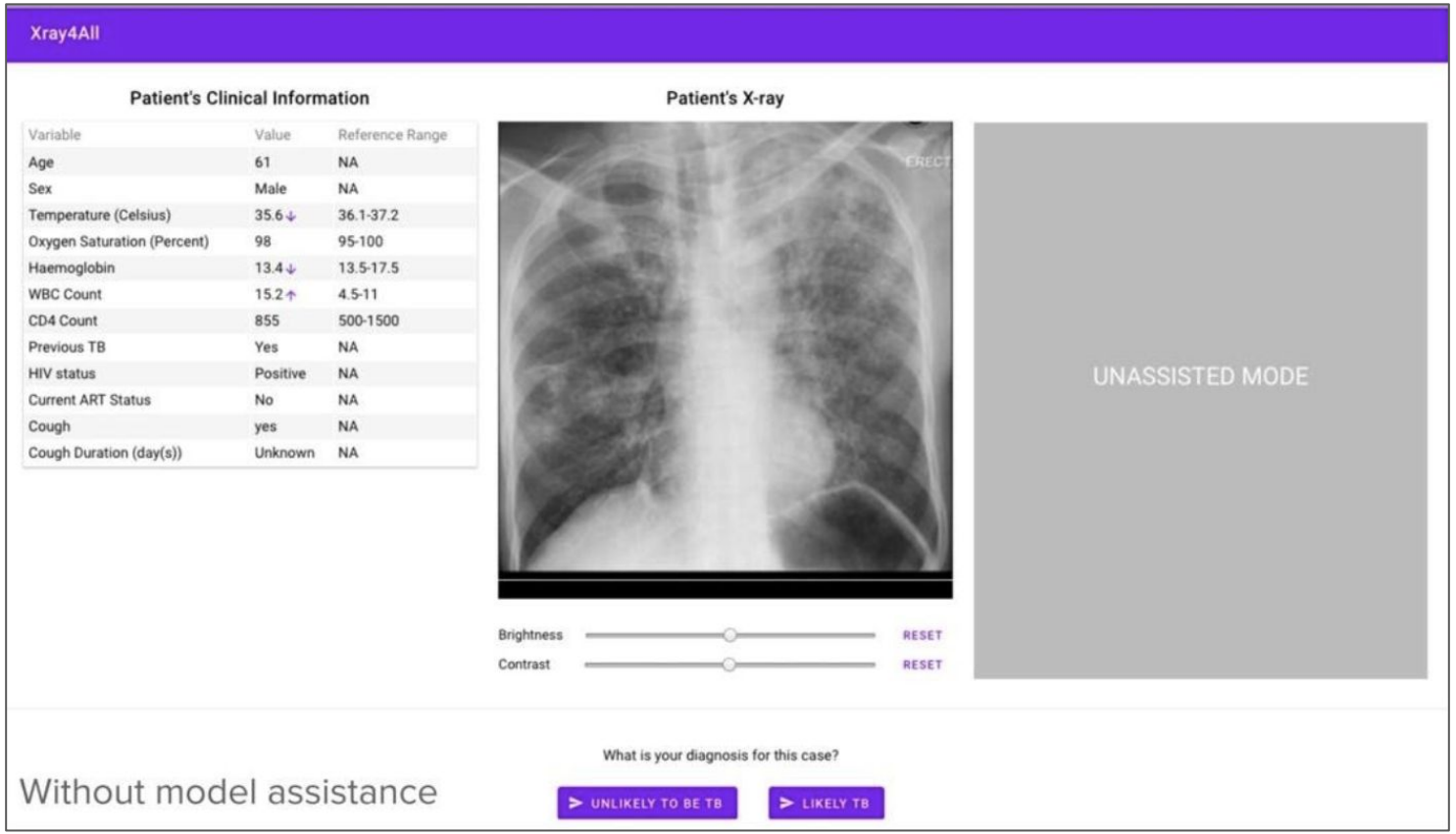
Rajpurkar & O'Connell & Schechter et al., *in review*

X-ray pretraining and clinical variables both improve performance

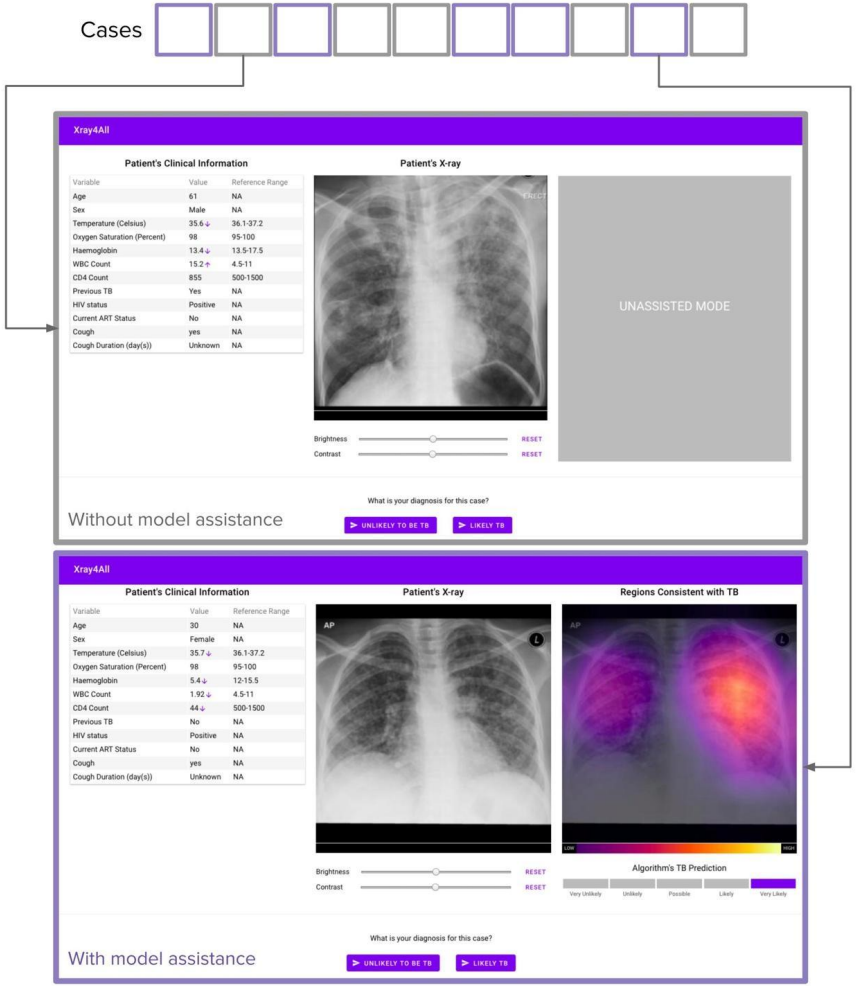


Strategy	AUC (95% CI)
With CheXpert pre-training and inclusion of clinical variables	0.83 (0.75, 0.91)
No CheXpert pre-training	0.714 (0.618, 0.810)
No inclusion of Clinical Variables	0.57 (0.46, 0.68)

Built out platform for deployment



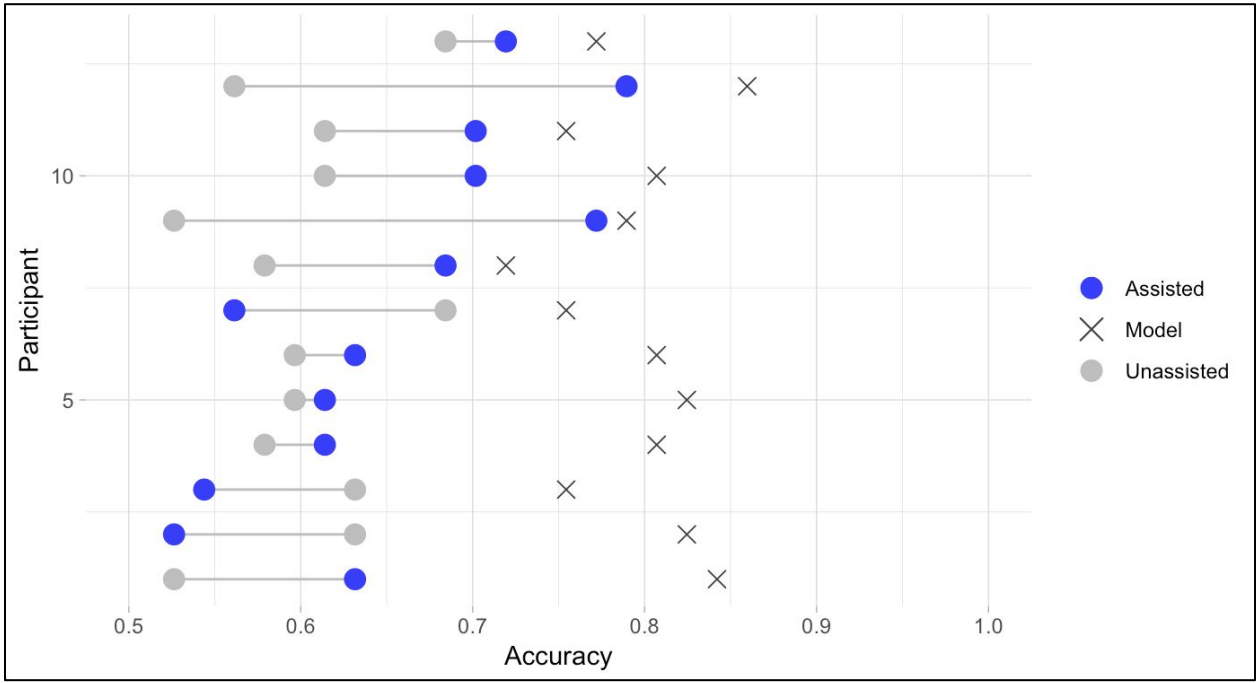
Investigated impacted of the AI assistant in a within-subjects design w/ 13 physicians



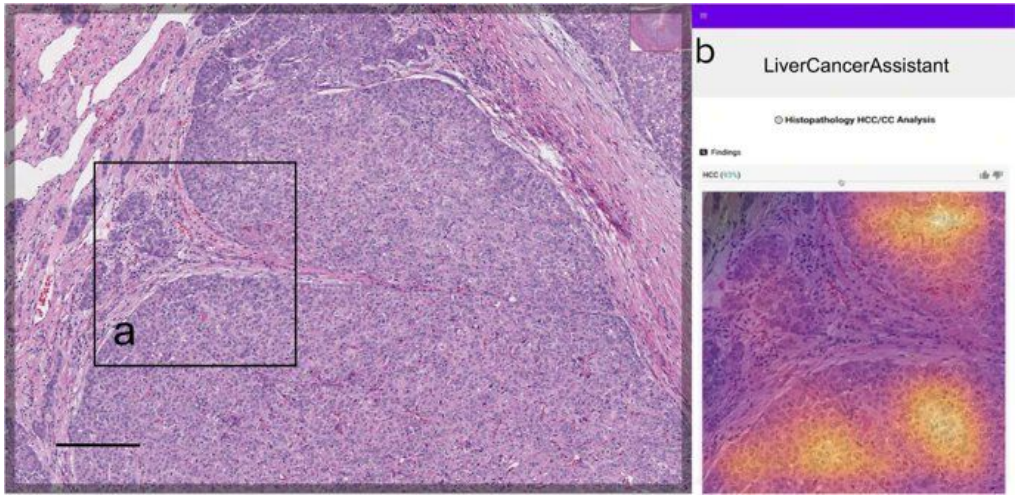
AI Improved Physician Performance but...

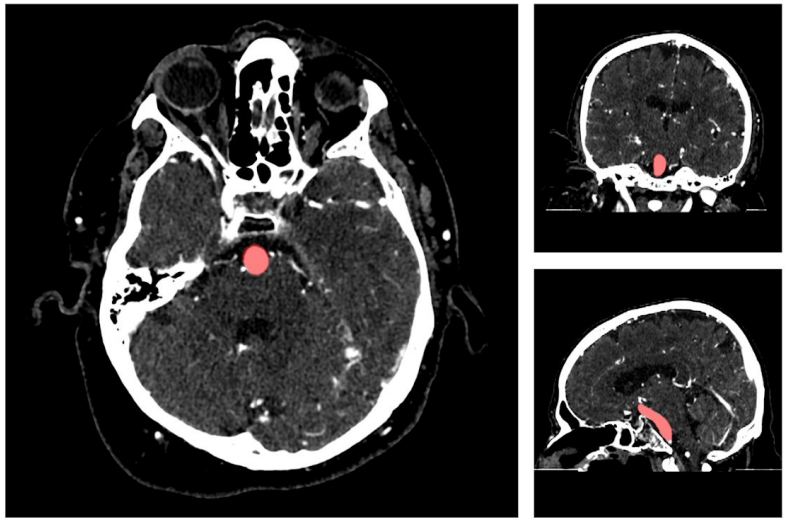
Performance	Accuracy (95% CI)
Physicians without assistance	0.60 (95% CI 0.57, 0.63)
Physicians with algorithm assistance	0.65 (95% CI 0.60, 0.70)
Algorithm by itself	0.79 (95% CI 0.77, 0.82)

Improvement is not consistent



Some studies across AI + Imaging





AI-Augmented CTA Aneurysm Interpretation

Takeaway: expert-level AI -> improved clinician performance is misguided

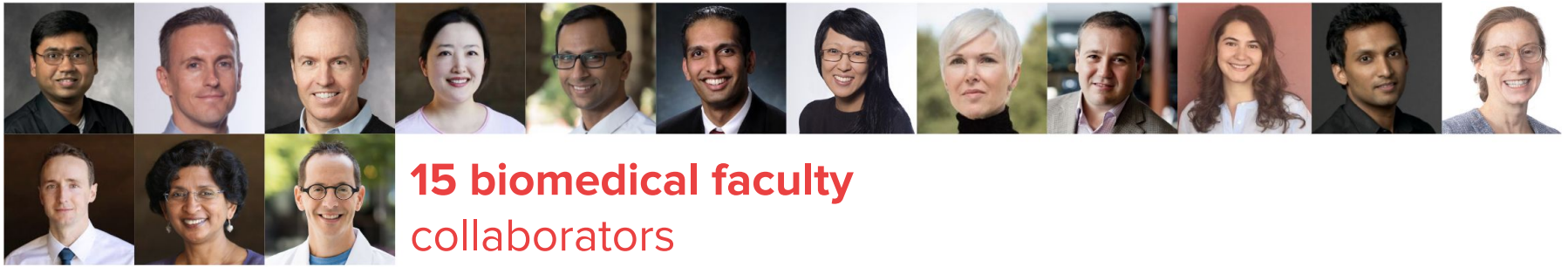
Critical considerations include:

- experience levels
- clinician interaction
- case difficulty
- automation bias.

We can develop AI technologies to improve clinical decision-making with:

1. Clinically-informed ML techniques
2. Dataset Design
3. Human-AI Interaction

How can I get involved?



15 biomedical faculty
collaborators

Help researchers keep up with the cutting edge of AI + Healthcare research

2000+ Weekly Readers

With Dr. Eric Topol

http://doctorpenguin.com

Help equip interested individuals learn tools to contribute

https://www.coursera.org/specializations/ai-for-medicine

With Dr. Andrew Ng

AI For Healthcare Bootcamp

6-month training program for undergrad + masters students to do research in AI + healthcare.

9 quarters. 80 bootcampers.

15 biomedical faculty collaborators

https://stanfordmlgroup.github.io/programs/aihc-bootcamp/

Apply

https://stanfordmlgroup.github.io/programs/aihc-bootcamp/

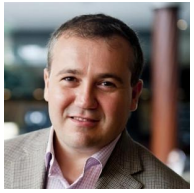
Lot of exciting opportunities

To build AI technologies that will be used routinely to improve clinical decision making.











Thanks!