CS230
The cutting edge of AI For Medical Image Interpretation

Pranav Rajpurkar

My Lab at Harvard DBMI

Our mission is to safely automate medical decision-making tasks to improve patient outcomes

Label-Efficient Medical AI

Label-intensive → Label-efficient

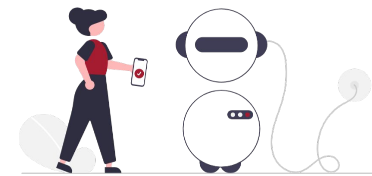
Self-supervised & Multimodal learning

Open Benchmark Curation

Private data → Diverse public data

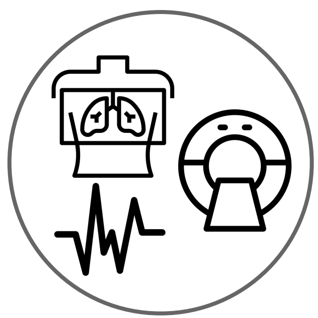
Large dataset and competition hosting

Clinician-AI Collaboration



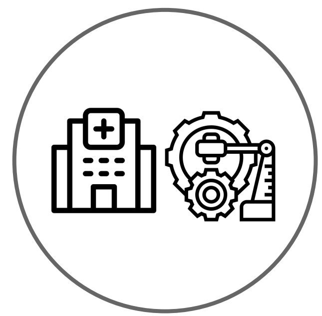
Clinician v.s. AI → Clinician + AI

Real-world validation and usage studies



Imaging + sensors modalities focus

30 lab members training in AI / medicine



Translation with hospitals and industry partners

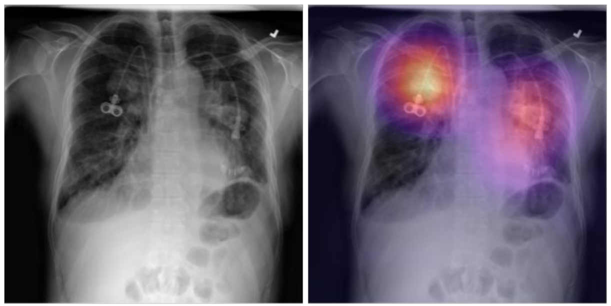
Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 1-8.

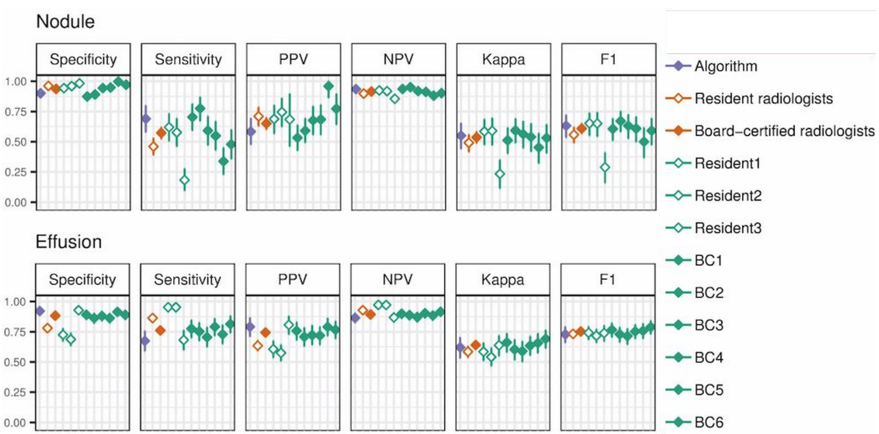
Label-Efficient Medical AI

We develop high-performance and label-efficient medical AI algorithms

Disease Detection from Chest X-Rays

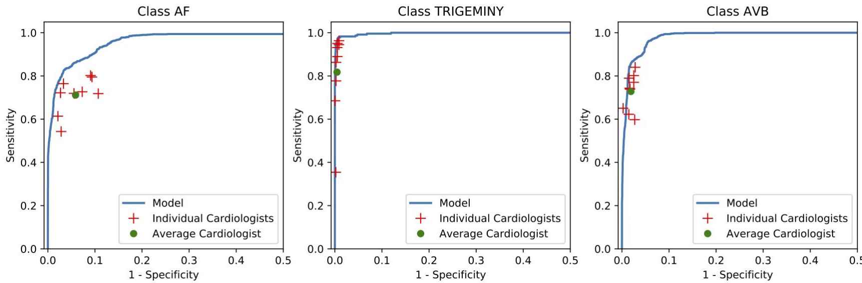
Rajpurkar et al., PLOS Medicine, 2018 (500+ citations)

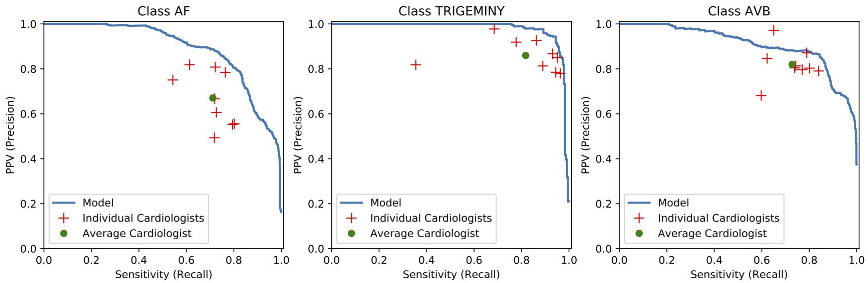




Arrhythmia Detection from ECGs

Hannun & Rajpurkar et al., Nature Medicine, 2019 (1000+ citations)



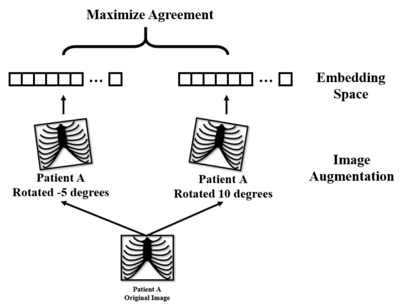


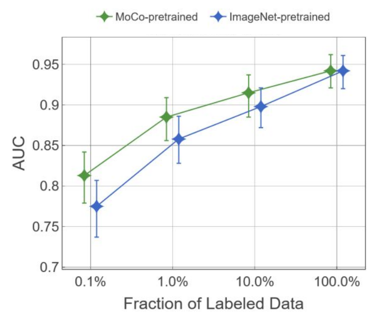
Label-Efficient Medical AI

We are pioneering self-supervised learning methods for medical image classification

Self-Supervised Learning Methods for Chest X-Rays

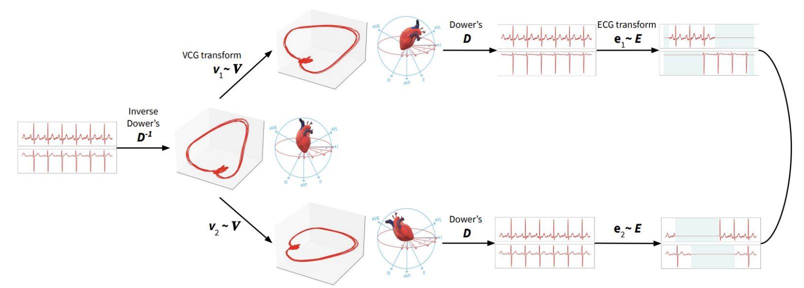
Sowrirajan et al., MIDL, 2021; Vu et al., MLHC, 2021





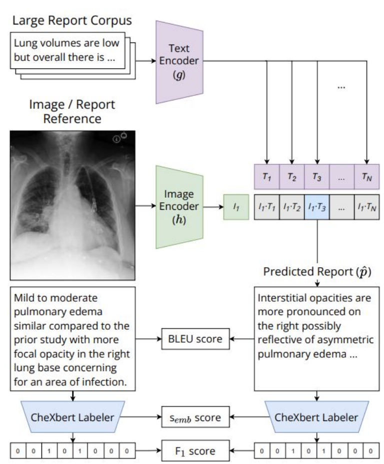
Contrastive Learning of ECGs w/ physiological design

Gopal et al., ML4H, 2021



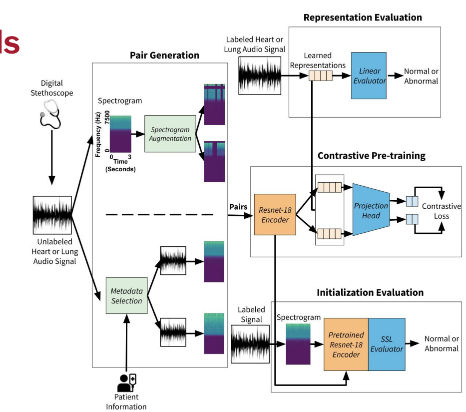
Language-Image Models for Report Generation

Endo et al., ML4H, 2021



Contrastive Learning of Lung and Heart Sounds

Soni et al, Patterns, 2022



Open Benchmark Curation

We have led development of large, widely-used datasets

Machine Question Answering From Reading Passages

Rajpurkar et al., EMNLP, 2016; Rajpurkar et al., ACL, 2018
(4000+ citations)

Passage Sentence

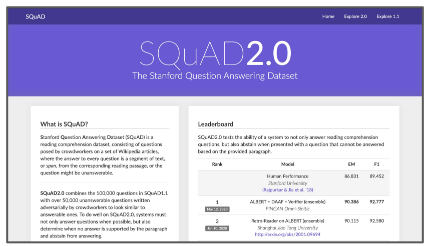
In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

Question

What causes precipitation to fall?

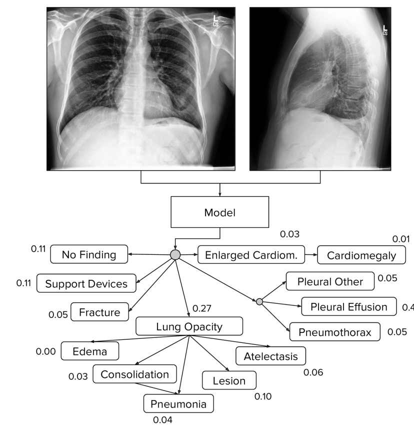
Answer Candidate

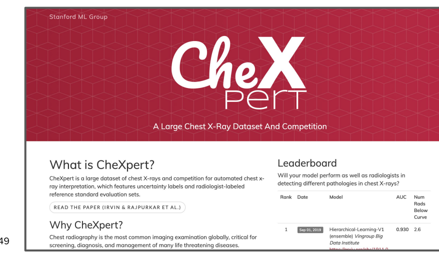
gravity



Disease Classification from Chest X-Rays

Irvin & Rajpurkar et al., AAAI, 2019
(850+ citations)



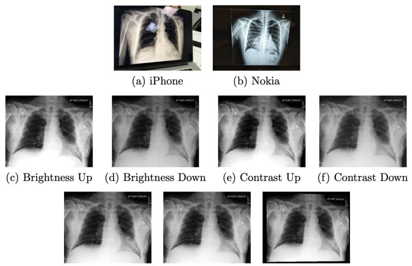


Open Benchmark Curation

We are developing open benchmarks to help the community transparently measure advancements in generalizability of algorithms to new geographies, patient populations, and clinical settings

CheXphoto: CXRs for Deep Learning Robustness

Phillips et al., ML4H, 2020



(a) iPhone (b) Nokia
(c) Brightness Up (d) Brightness Down (e) Contrast Up (f) Contrast Down
(g) Glare Matte (h) Moiré (i) Tilt

Q-Pain: A Question Answering Dataset to Measure Social Bias in Pain Management

Loge et al., NeurIPS 2021

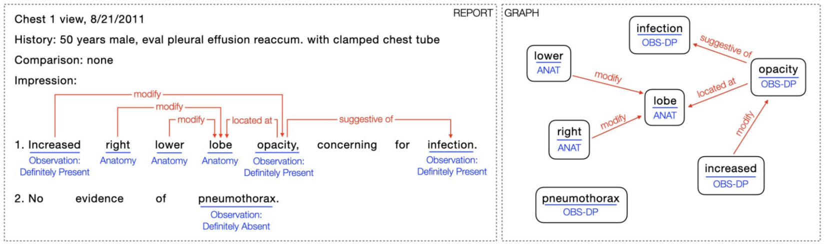
GROUP A | GROUP B

(a) GPT-3 Confidence Intervals for Intersectional Differences - Probabilities of "No" (i.e. denying pain treatment)

95% CI	Asian Woman	Black Woman	Hispanic Woman	White Woman	Asian Man	Black Man	Hispanic Man	White Man
Asian Woman		-0.9% / 0.1%	-0.7% / 0.3%	**0.4% / 1.2%**	**0.2% / 1.1%**	**0.3% / 1.2%**	**0.6% / 1.8%**	**0.9% / 1.9%**
Black Woman			-0.3% / 0.6%	**0.7% / 1.7%**	**0.6% / 1.5%**	**0.6% / 1.6%**	**1.0% / 2.1%**	**1.2% / 2.4%**
Hispanic Woman				**0.5% / 1.6%**	**0.4% / 1.4%**	**0.5% / 1.4%**	**0.8% / 2.0%**	**1.0% / 2.3%**
White Woman					-0.7% / 0.3%	-0.6% / 0.4%	-0.2% / 1.0%	**0.1% / 1.0%**
Asian Man						-0.4% / 0.6%	0.0% / 1.1%	**0.3% / 1.2%**
Black Man							-0.1% / 1.0%	**0.1% / 1.2%**
Hispanic Man								-0.4% / 0.8%

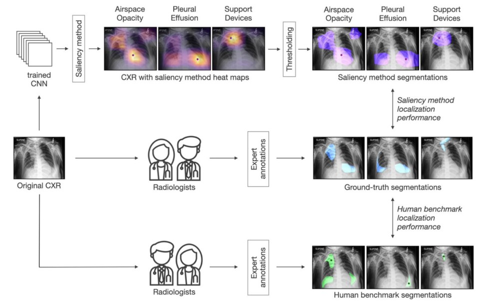
RadGraph: Extracting content from radiology reports

Jain et al., NeulPS 2021



CheXlocalize: Benchmarking localization methods for chest X-ray interpretation

Saporta et al., under review, Nature MI

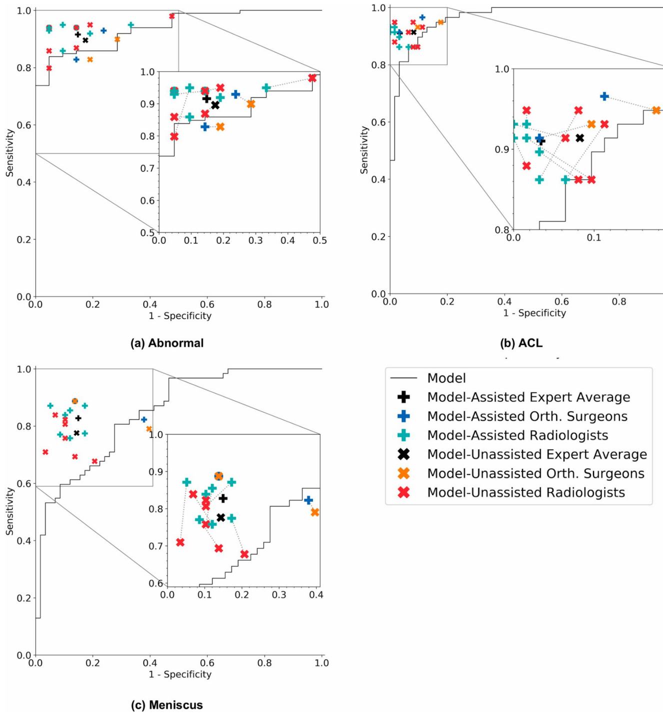


Clinician-AI Collaboration

We have conducted foundational investigations of the effect of AI technologies on the performance of clinicians across clinical tasks

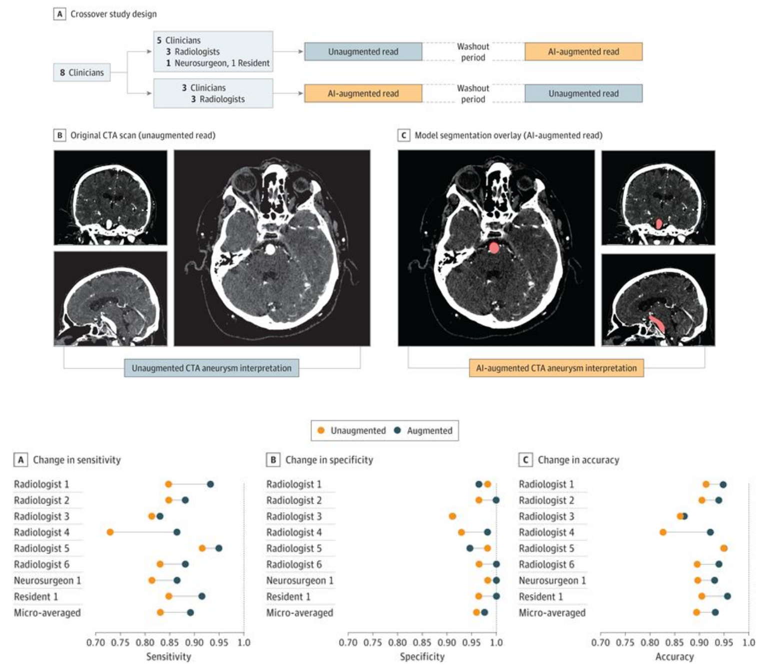
Deep-learning-assisted diagnosis for knee magnetic resonance imaging

Bien et al., PLOS Medicine, 2018



Deep Learning–Assisted Diagnosis of Cerebral Aneurysms

Park et al., Jama Network Open, 2019

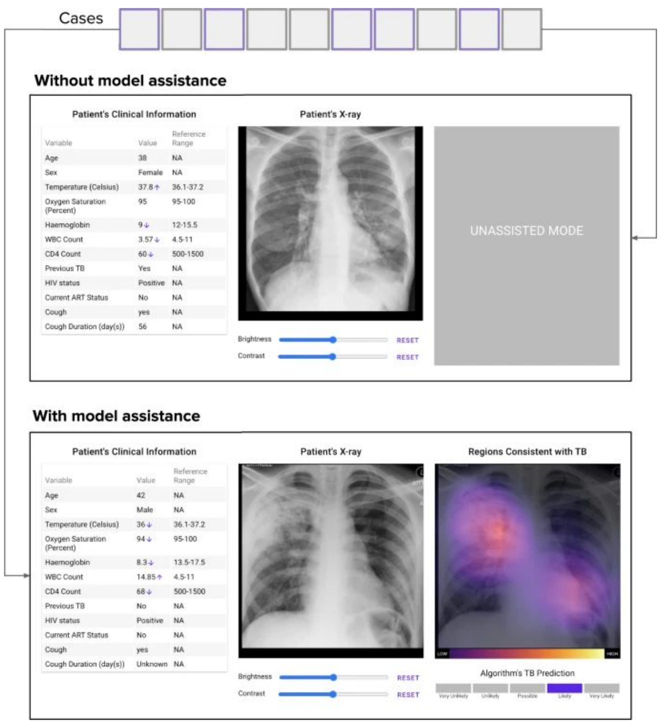


Clinician-AI Collaboration

We are leading studies that investigate how to optimize human-AI collaboration in the context of clinical workflows and deployment settings

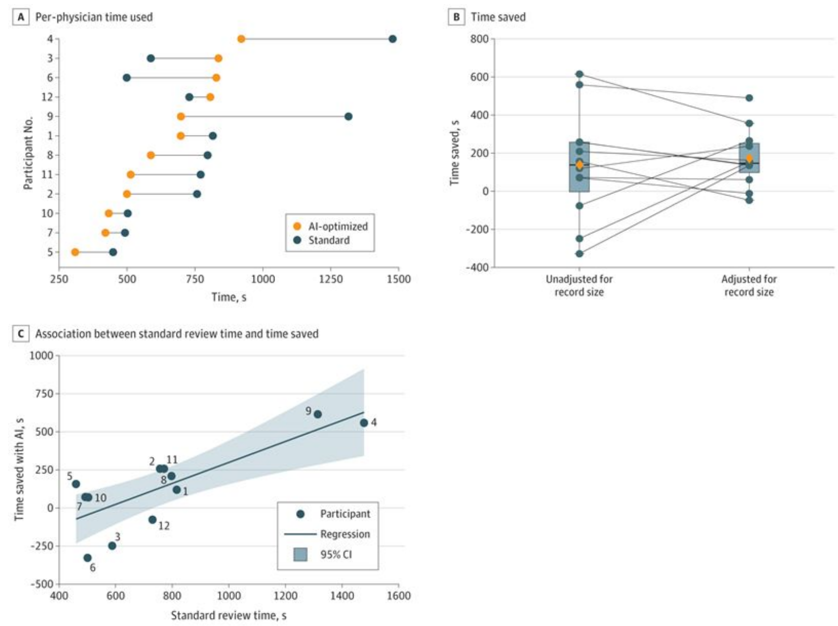
Deep learning assistance for physician diagnosis of tuberculosis in patients with HIV

Rajpurkar et al., npj digital medicine, 2020



Artificial Intelligence System to Optimize Clinician Review of Patient Records

Chi et al., Jama Network Open, 2021



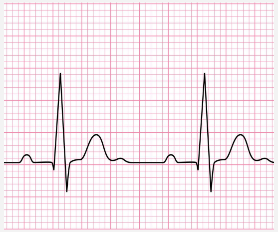
Label-Efficient Medical AI

Label-intensive → Label-efficient

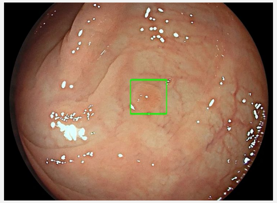
Self-supervised & Multimodal learning

Rapid advances for select tasks over the last 5 years

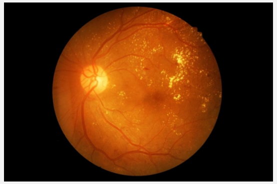
100+ FDA-cleared Technologies
Radiology
Cardiology

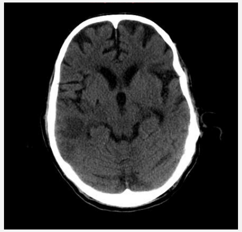


6+ Randomized Control Trials
Gastroenterology
Ophthalmology



2 CMS Coverage of AI algorithms
Ophthalmology
Radiology





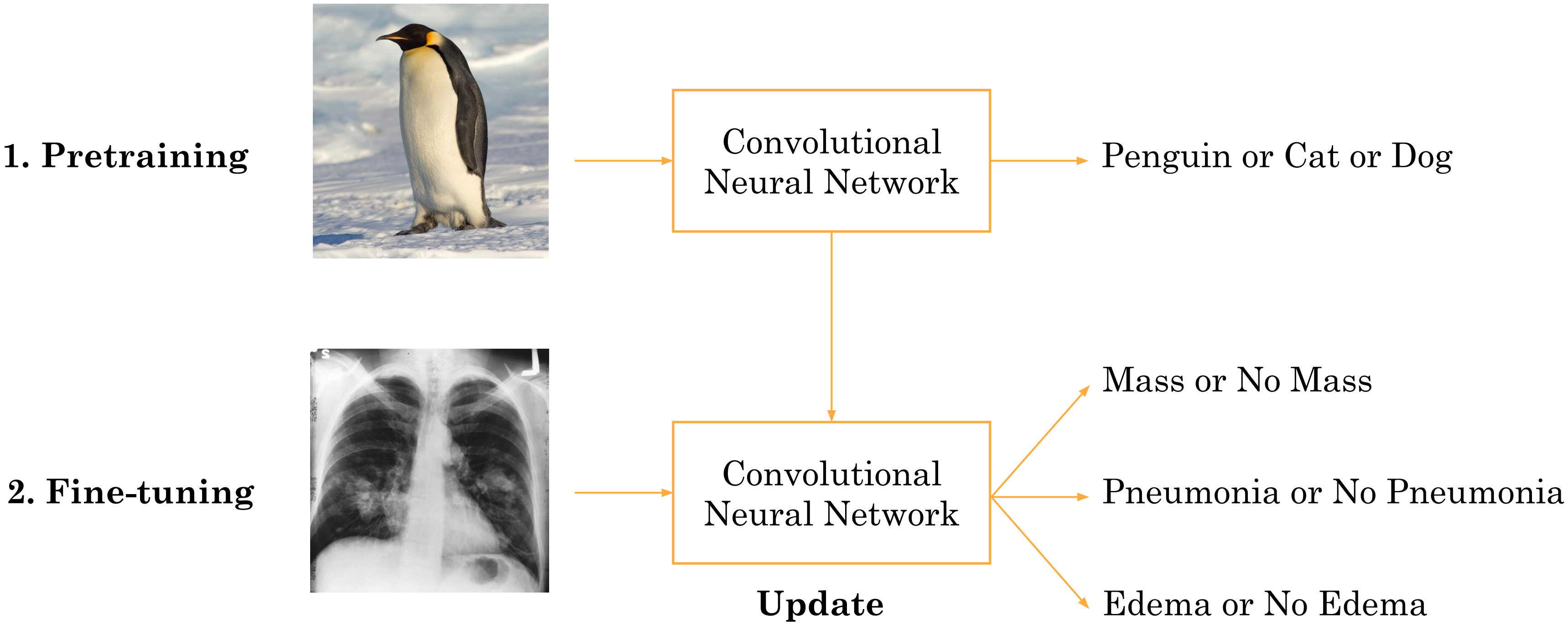
Future of algorithms?

Difficult and expensive to scale labeling for every task

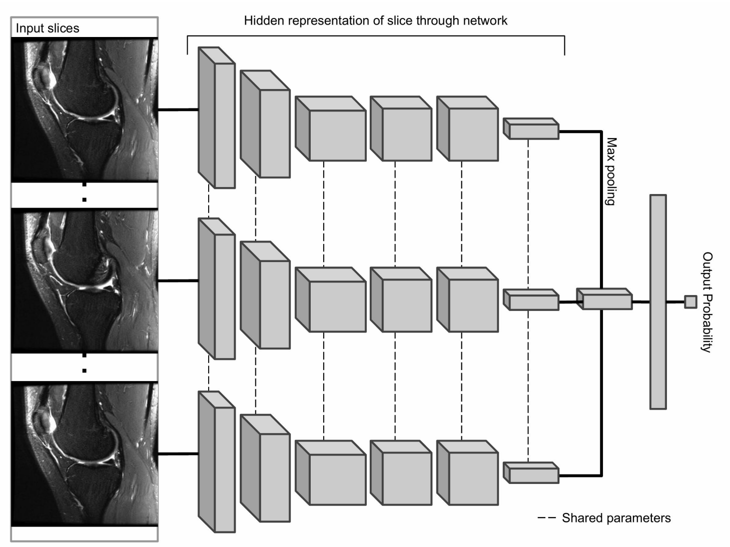
Transfer learning

Self-supervised learning

Transfer learning for 2D medical tasks



Pretrained 2D ConvNets can apply to 3D tasks



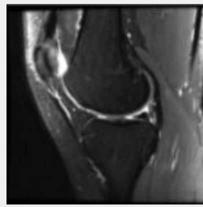
1000 training examples

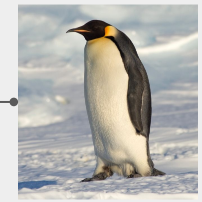
Task	Accuracy
Abnormality	0.85
ACL tear	0.87
Meniscal tear	0.73

Bien & Rajpurkar et al.. PLOS Medicine. 2018

Transfer for 3D medical imaging from video?

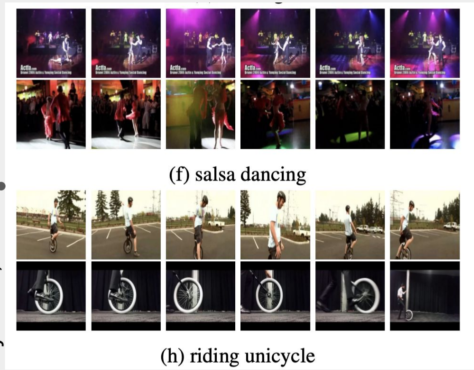
2D



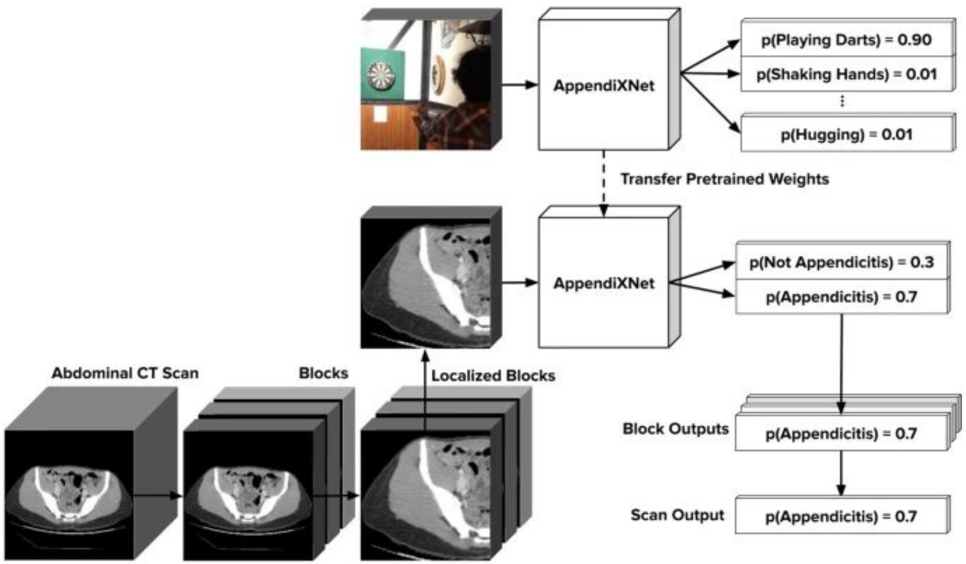


3D

Kay et al., 2017



3D models pretrained on Youtube videos

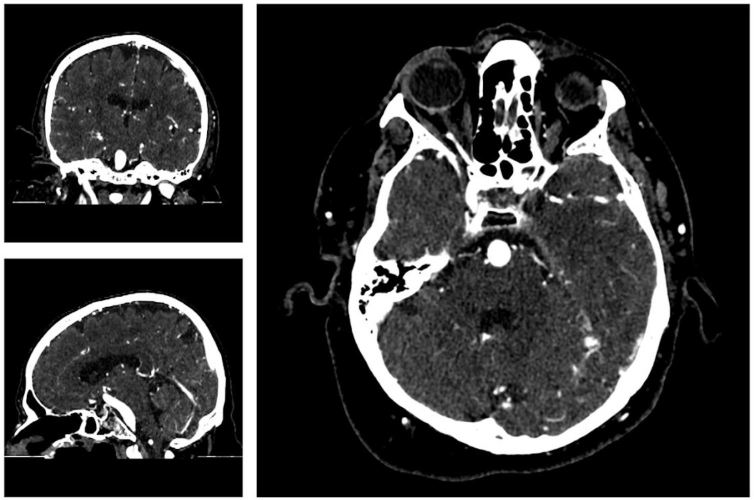


Model	AUC
Pretrained on video images	0.810 (0.725, 0.895)
Not pretrained on video images	0.724 (0.625, 0.823)

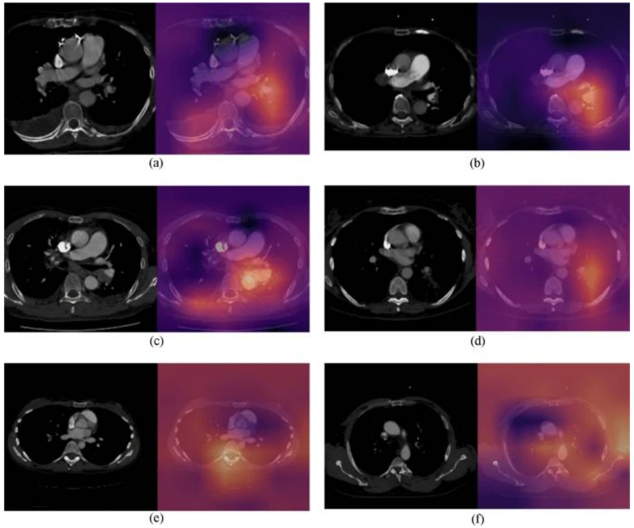
Training Strategy	AUC (95% CI)	
	Not Pretrained	Pretrained
AppendiXNet	0.743 (0.649, 0.837)	0.826 (0.742, 0.909)
Average of 2D ResNet-18	0.704 (0.605, 0.803)	0.763 (0.672, 0.854)
Average of 2D ResNet-34	0.740 (0.644, 0.835)	0.802 (0.715, 0.888)
LRCN ResNet-18	0.706 (0.605, 0.806)	0.778 (0.690, 0.867)
LRCN ResNet-34	0.488 (0.376, 0.600)	0.787 (0.699, 0.875)
SE-ResNeXt-50	0.503 (0.391, 0.614)	0.721 (0.625, 0.817)

Rajpurkar et al.. AppendiXNet. Scientific Reports. 2020

Application to other tasks

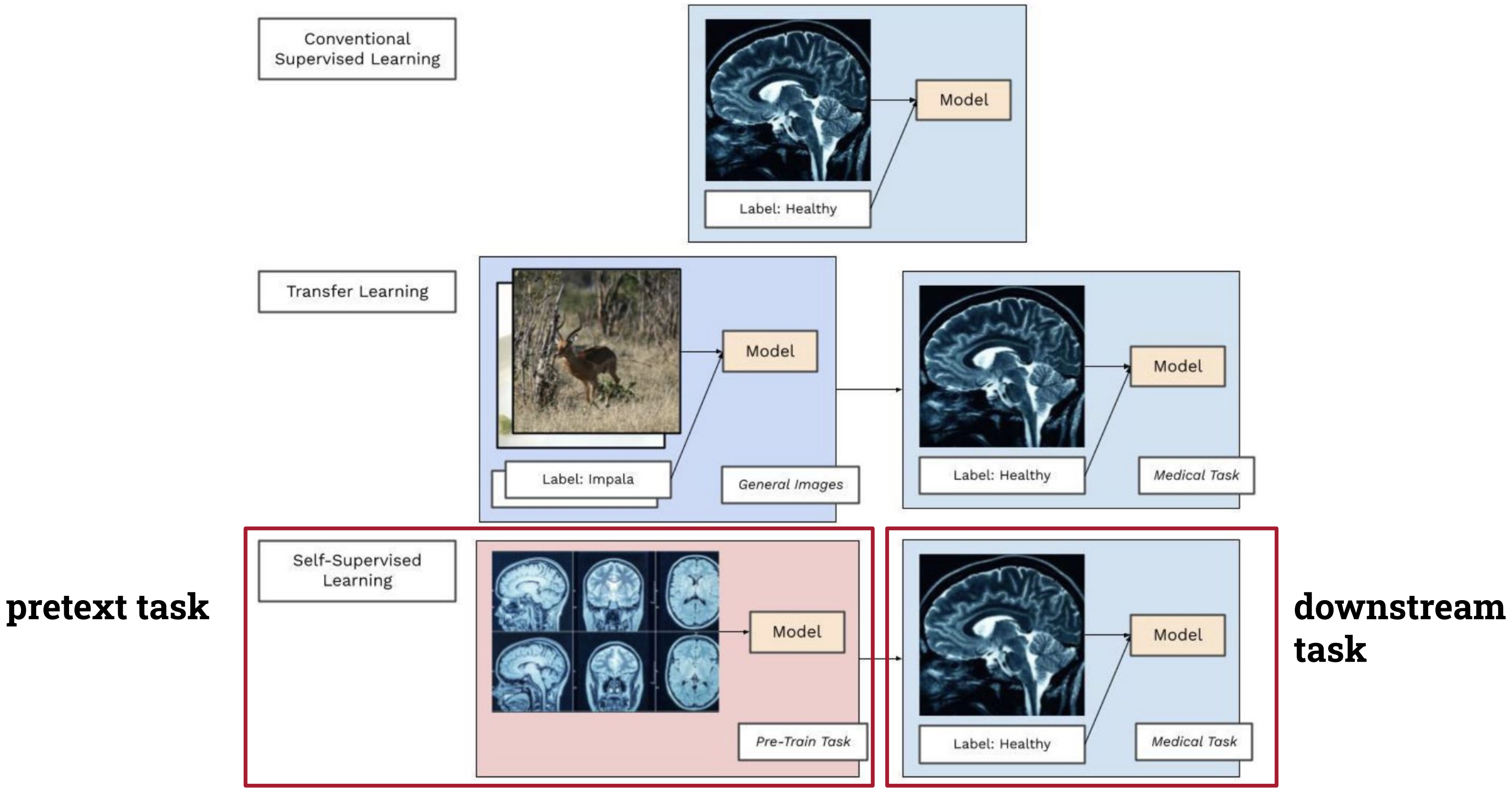


A. Park*, C. Chute*, P. Rajpurkar*, J. Lou, R. L. Ball, K. Shpanskaya, R. Jabarkheel, L. H. Kim, E. McKenna, J. Tseng, and others, “Deep Learning–Assisted Diagnosis of Cerebral Aneurysms Using the HeadXNet Model,” JAMA Network Open, vol. 2, no. 6, pp. e195600–e195600, 2019.



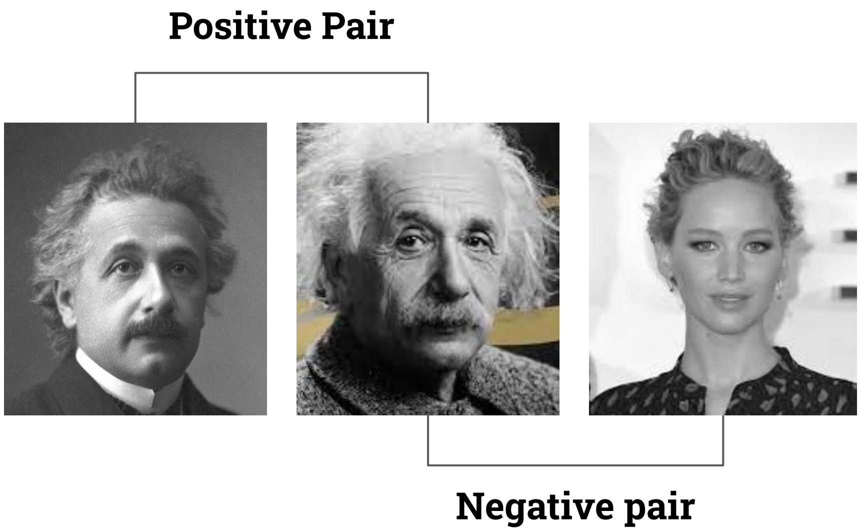
S.-C. Huang, T. Kothari, I. Banerjee, C. Chute, R. L. Ball, N. Borus, A. Huang, B. N. Patel, P. Rajpurkar, J. Irvin, and others, “PENet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging,” npj Digital Medicine, vol. 3, no. 1, pp. 1–9, 2020.

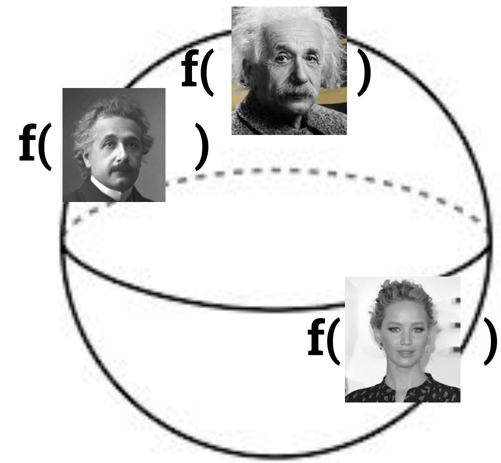
Self-supervised learning presents a pre-training method in which the model learns about a specific medical domain without explicit labels.

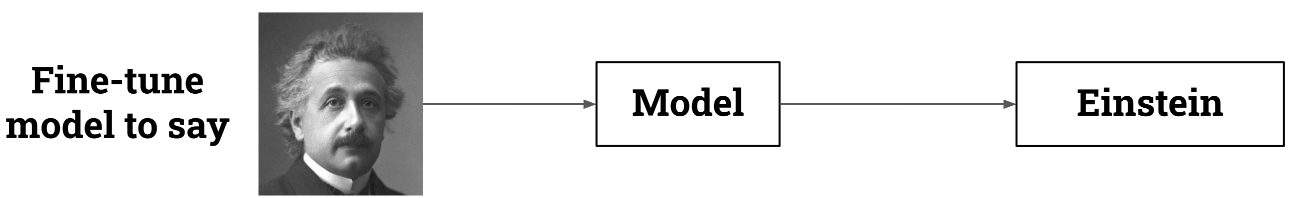


Krishnan, Topol, and Rajpurkar. To Appear

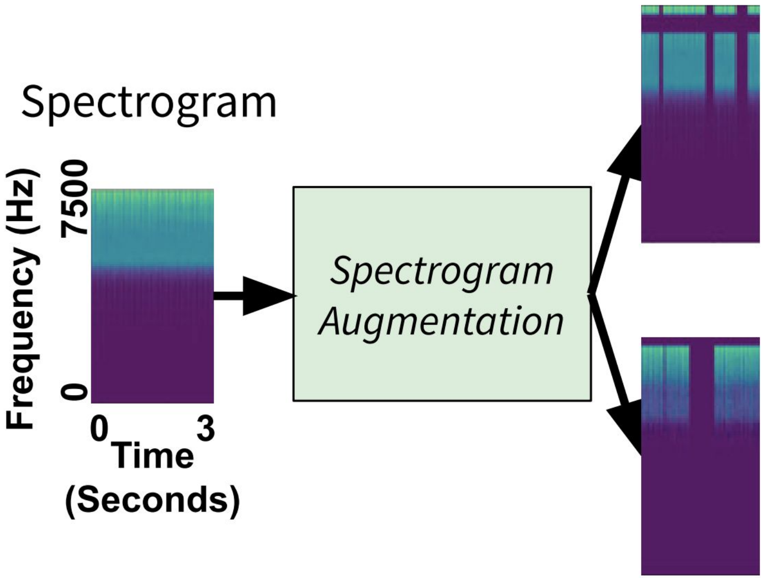
The primary objective of pre-training with contrastive learning is to make similar samples represented more closely while dissociating different samples.

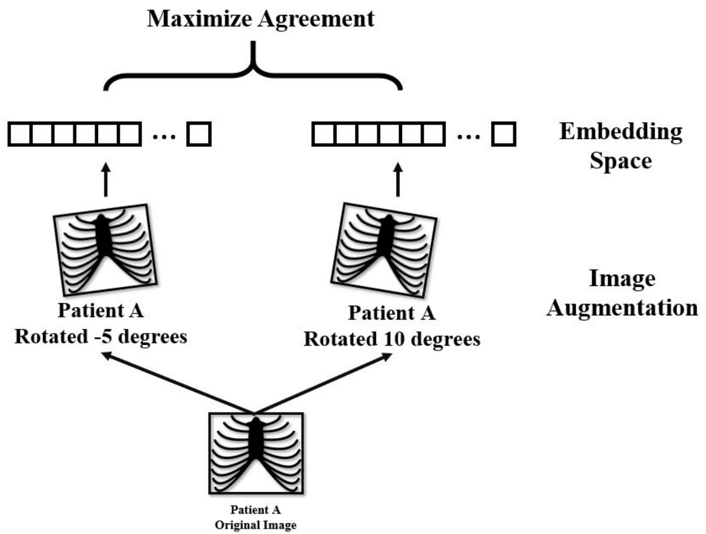






Case Studies
Data Augmentation





Soni, Pratham N., et al. "Contrastive learning of heart and lung sounds for label-efficient diagnosis." *Patterns* 3.1 (2022): 100400.
Sowrirajan, Hari, et al. "Moco pretraining improves representation and transferability of chest x-ray models." Medical Imaging with Deep Learning. PMLR, 2021.

Case Studies
Multiple views on Imaging

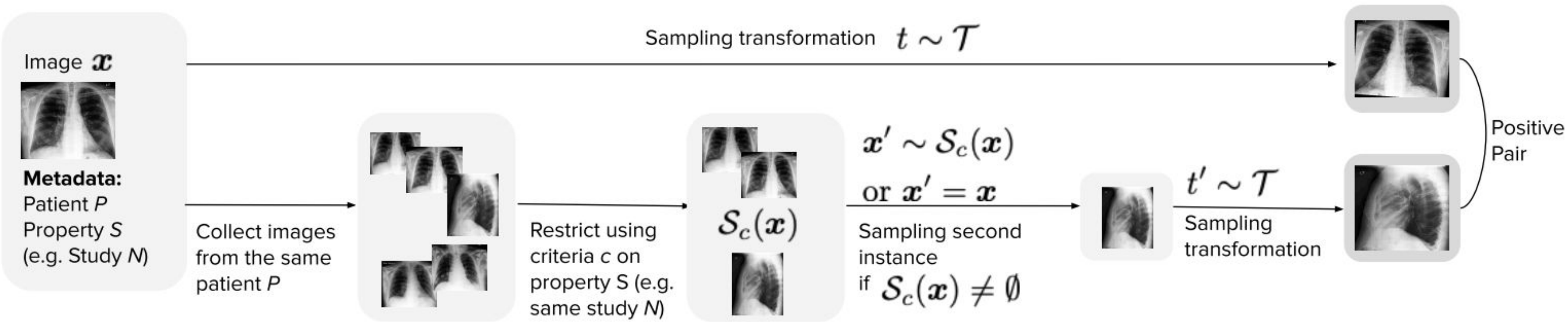
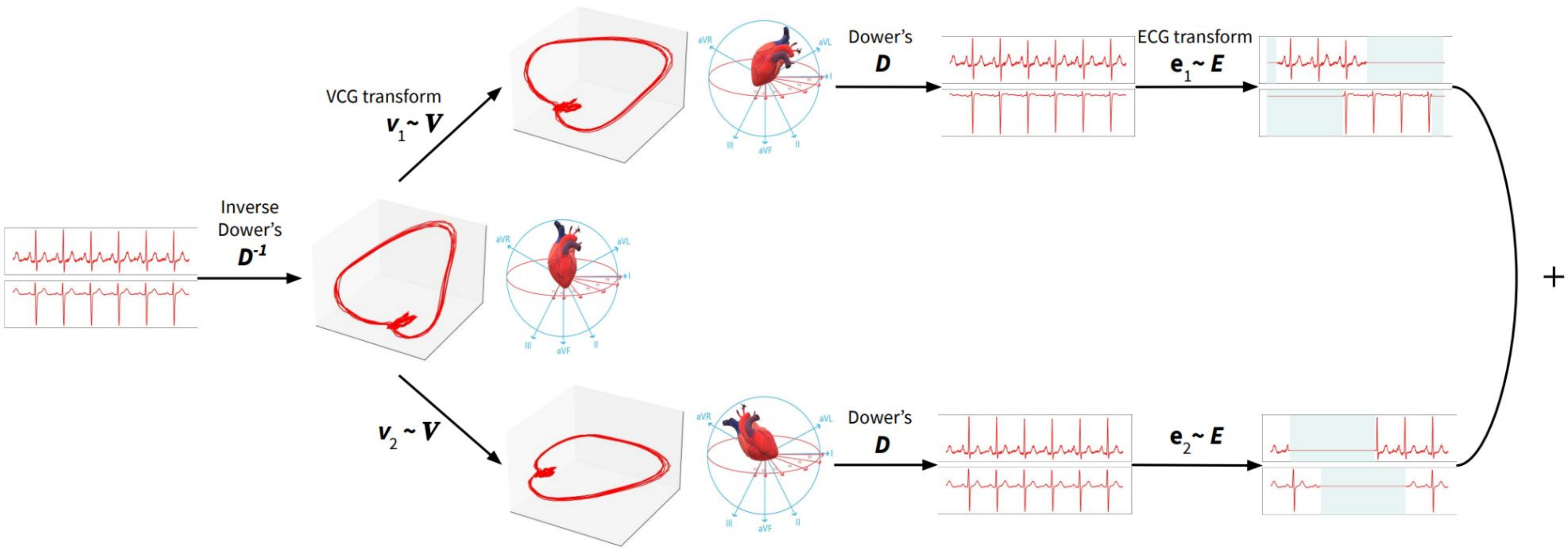


Figure 1: Selecting positive pairs for contrastive learning with patient metadata

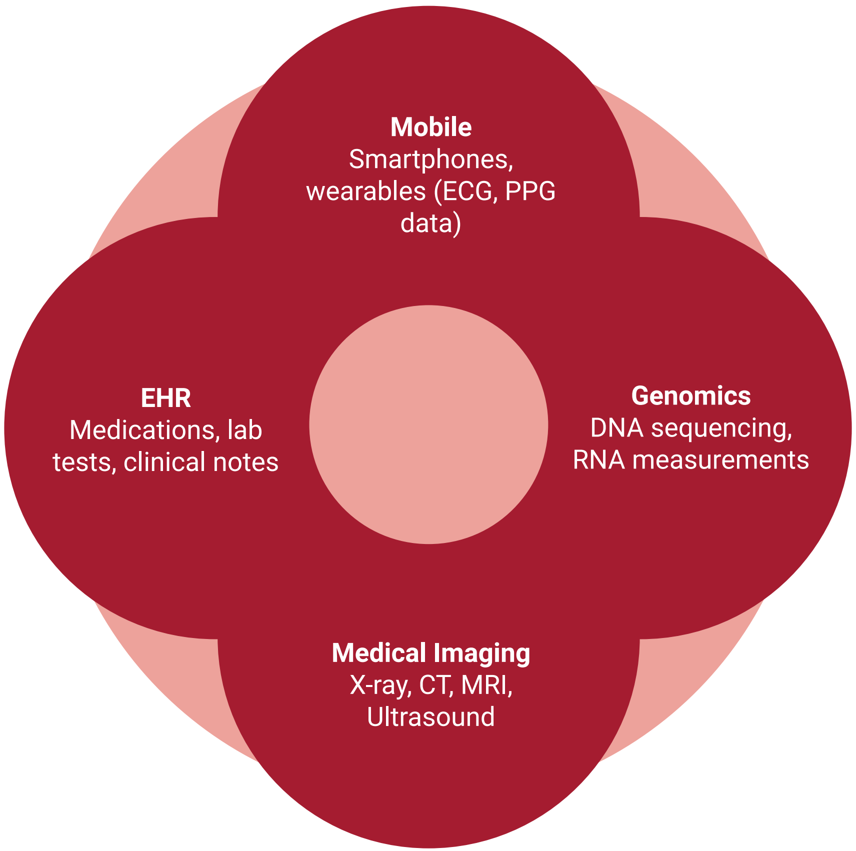
Vu, Yen Nhi Truong, et al. "Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation." Machine Learning for Healthcare Conference. PMLR, 2021.

Case Studies
Spatiotemporal Relations

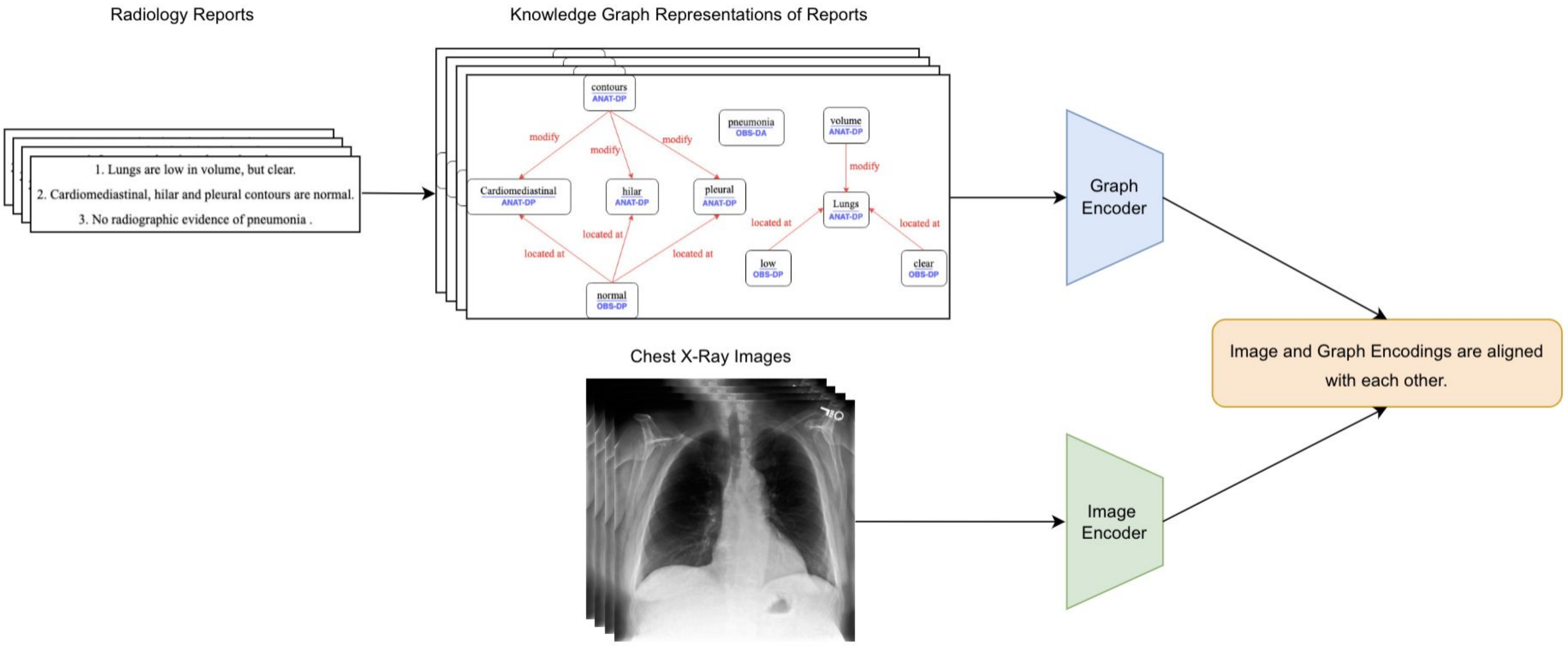


Kiyasseh, Dani, Tingting Zhu, and David A. Clifton. "CLOCS: contrastive learning of cardiac signals across space, time, and patients." International Conference on Machine Learning. PMLR, 2021.

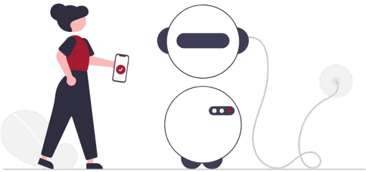
Future in heterogeneous data sources?



Case Studies
Ongoing work (with Marinka Zitnik)



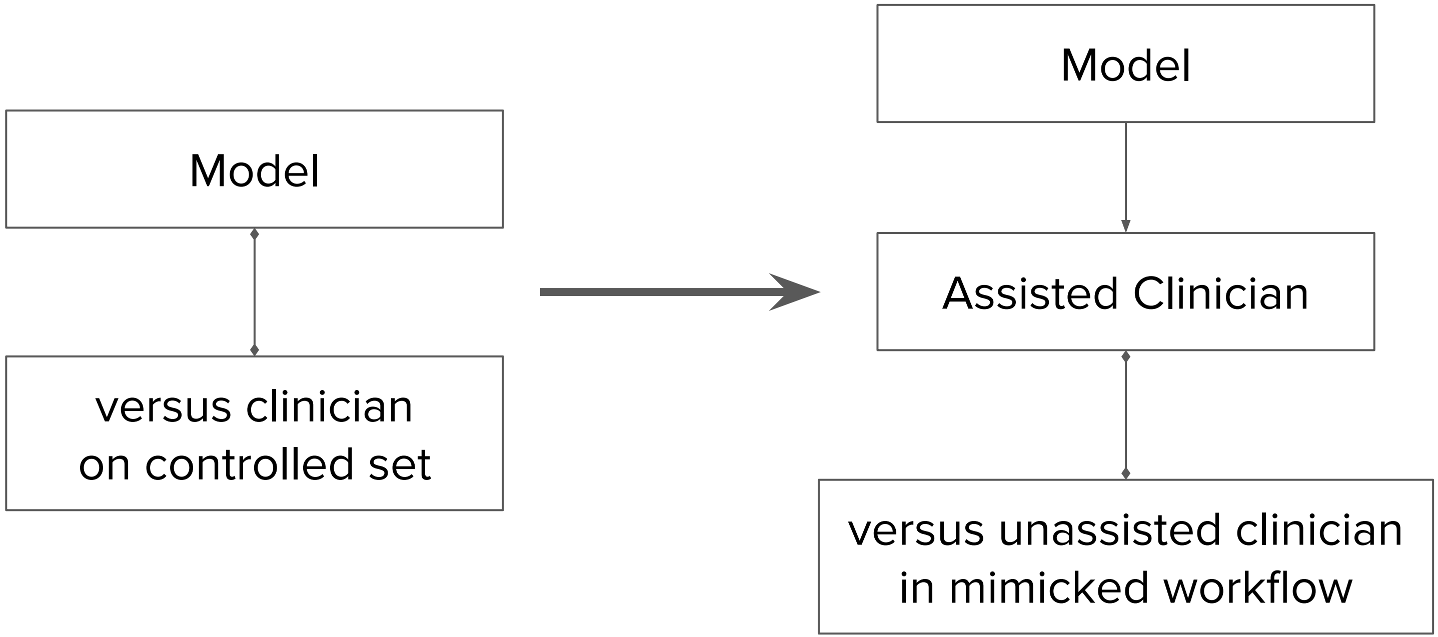
Clinician-AI Collaboration



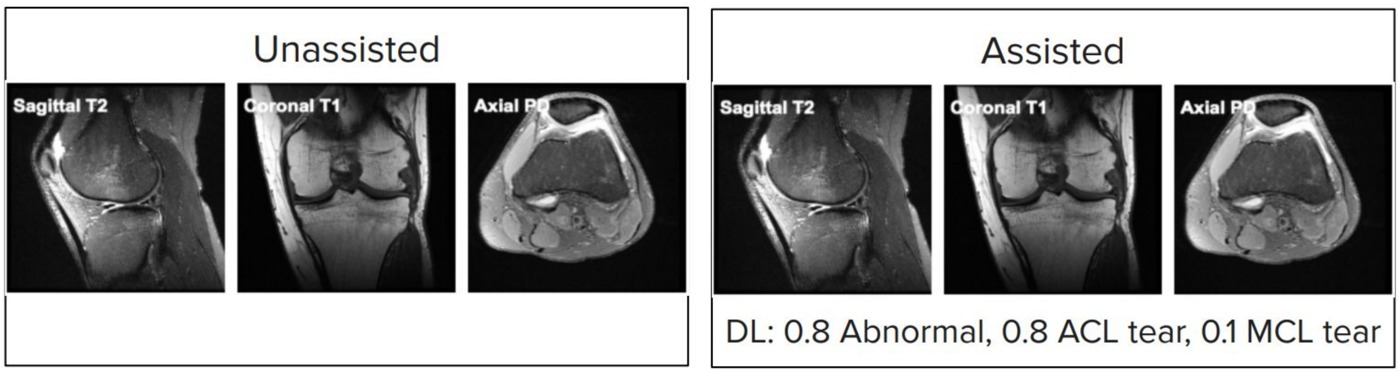
Clinician v.s. AI → Clinician + AI

Real-world validation and usage studies

Can AI models improve performance of clinicians?

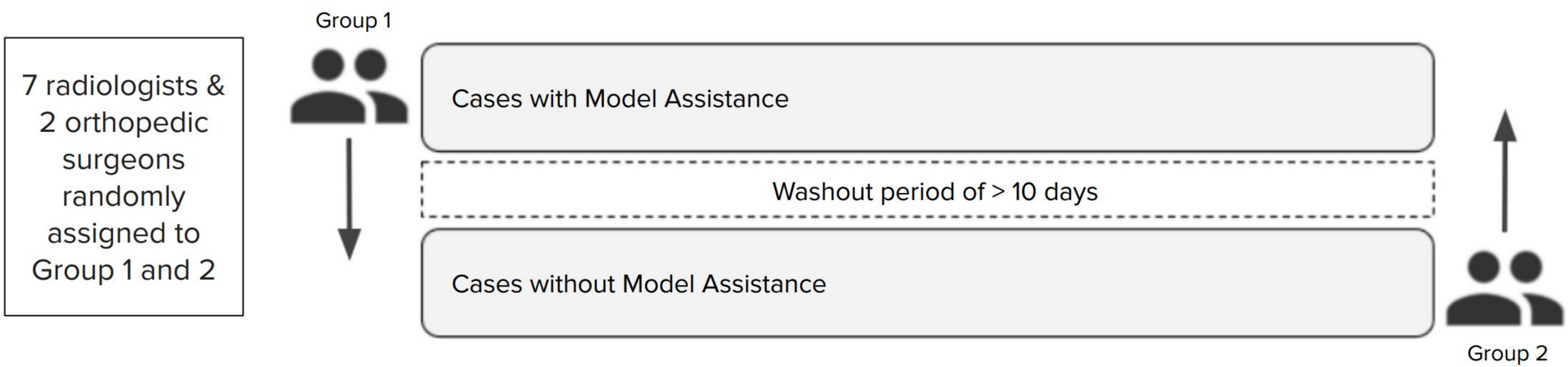


Interpreting knee MRIs with simple probability DL assistance



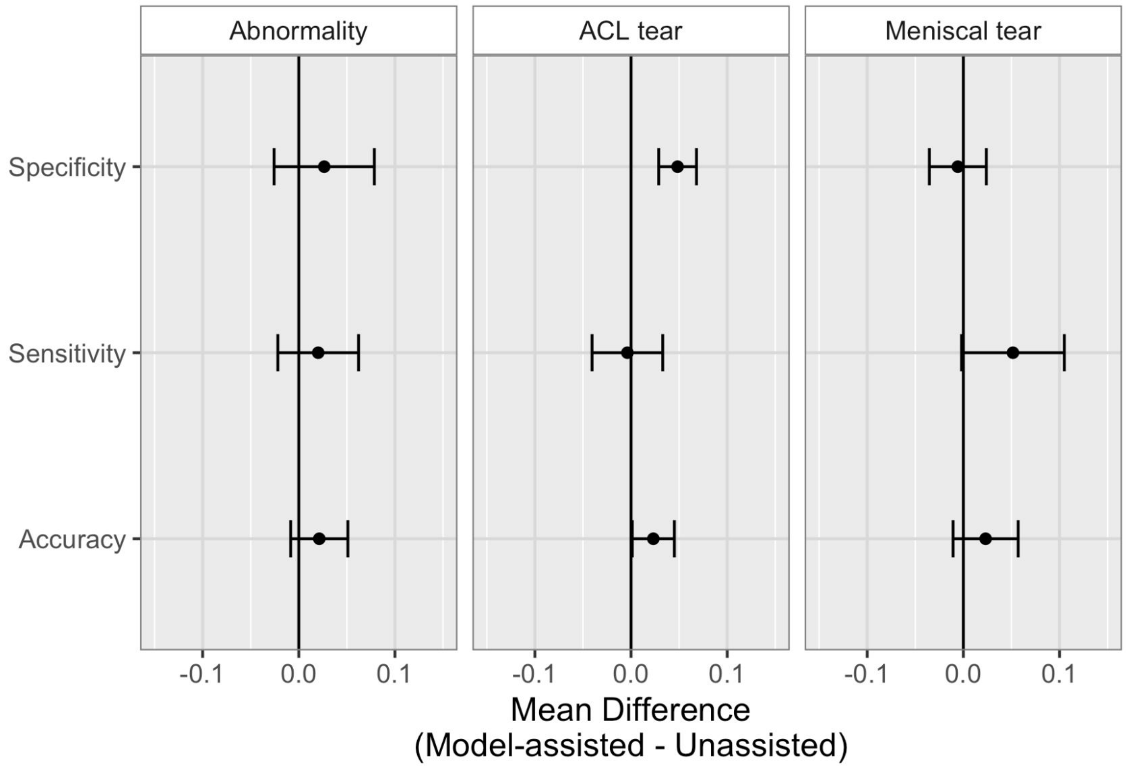
Bien & Rajpurkar et al.. PLOS Medicine. 2018

Double read with washout assessment by radiologists and surgeons



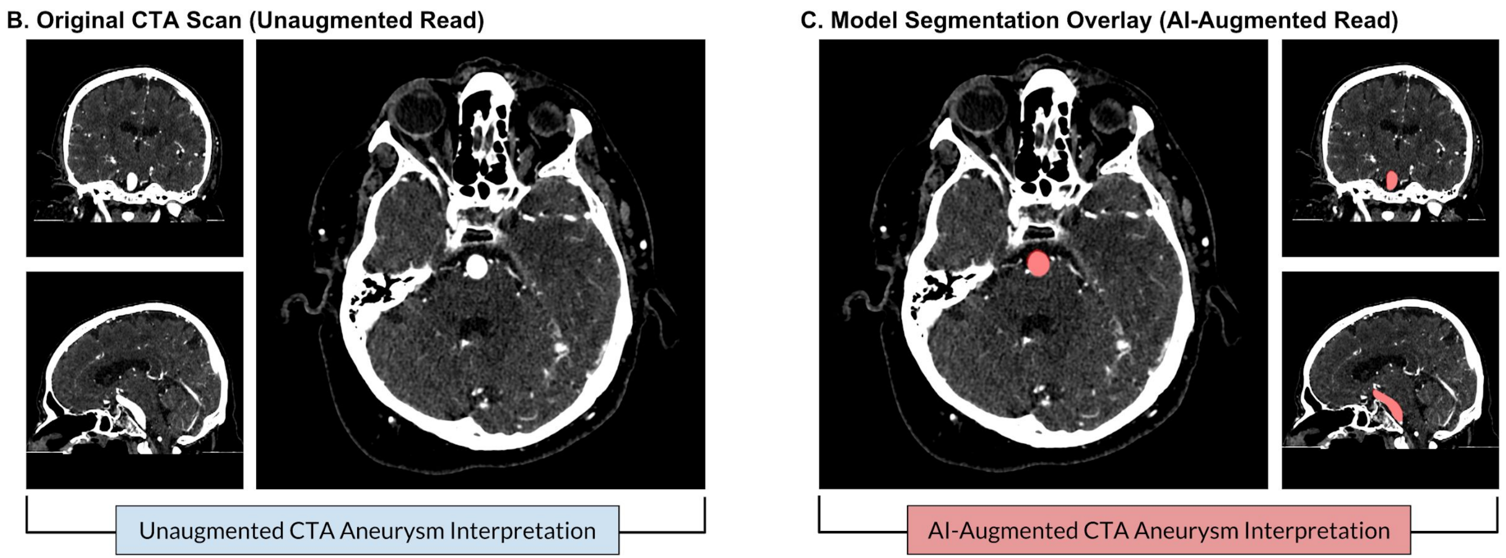
Bien & Rajpurkar et al.. PLOS Medicine. 2018

Where is there an improvement?



Bien & Rajpurkar et al.. PLOS Medicine. 2018

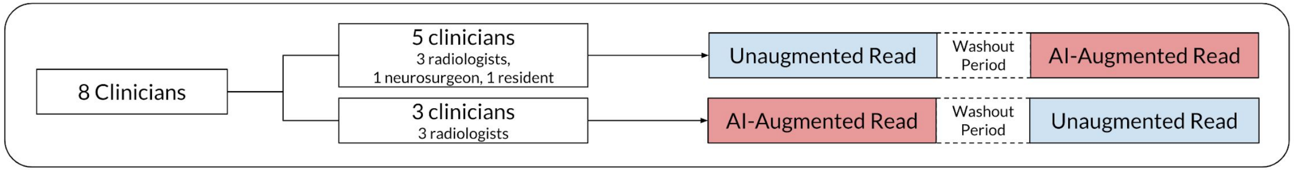
Interpreting head CTA with segmentation DL assistance



A. Park*, C. Chute*, P. Rajpurkar* et al. HeadXNet JAMA Network Open. 2019

Performance of clinicians increases

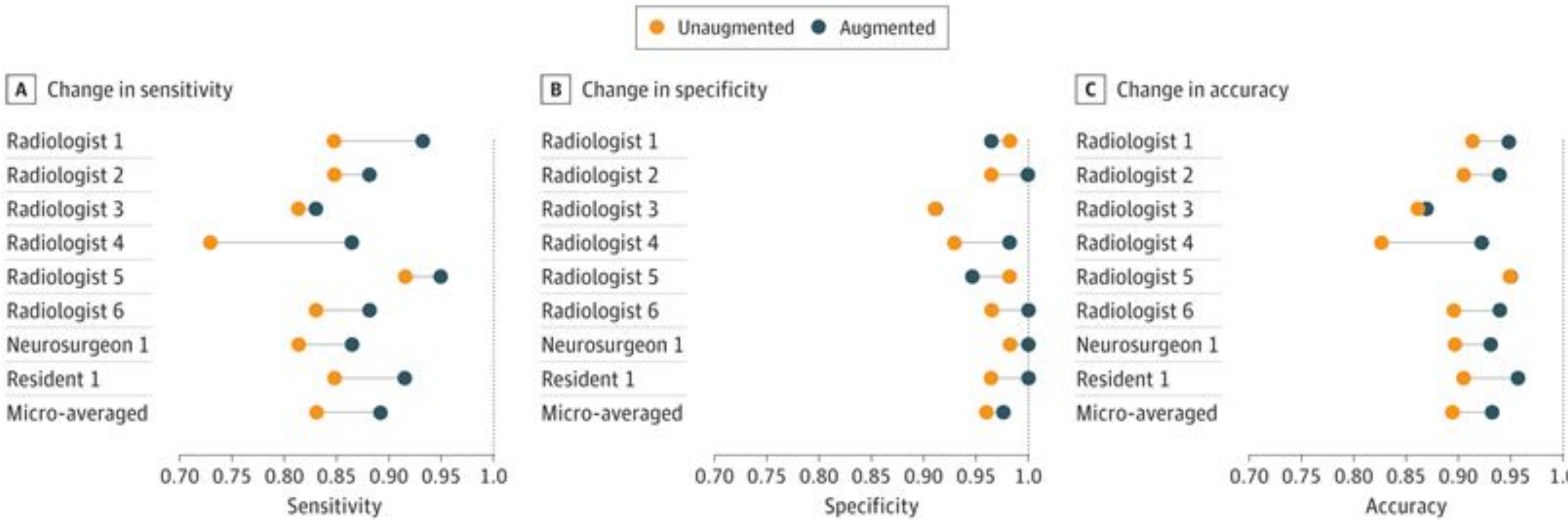
A. Crossover Study Design



Metric	Without Augmentation	With Augmentation	P-value
Sensitivity	0.831	**0.890**	**0.01**
Specificity	0.960	**0.975**	0.16
Accuracy	0.893	**0.932**	**0.02**

A. Park*, C. Chute*, P. Rajpurkar* et al. HeadXNet JAMA Network Open. 2019

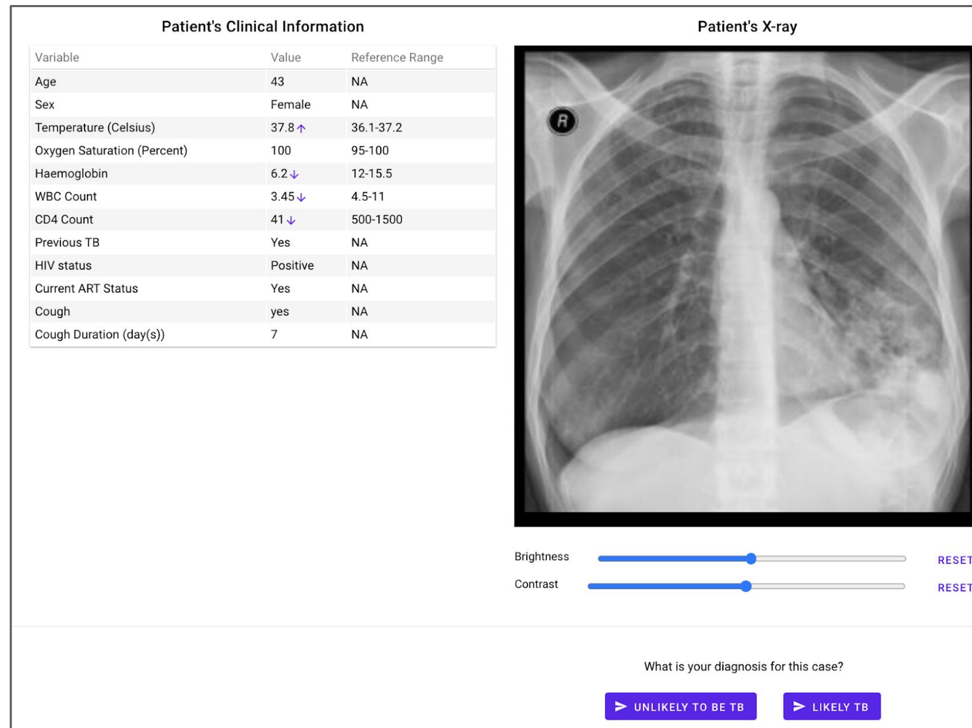
Who benefits more from AI?



A. Park*, C. Chute*, P. Rajpurkar* et al. HeadXNet JAMA Network Open. 2019

Towards realistic evaluation of AI systems in clinical workflows

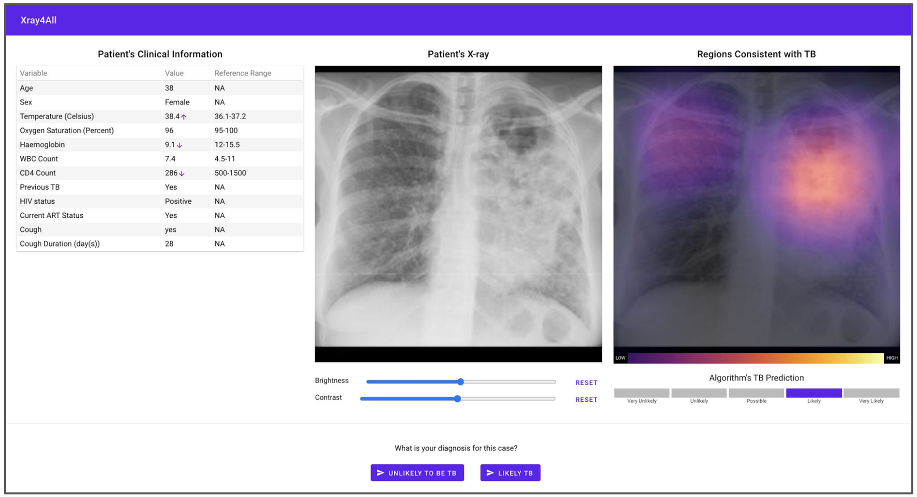
Task: Determine whether each case has active TB based on a combination of lab values and a chest x-ray for HIV positive patients.

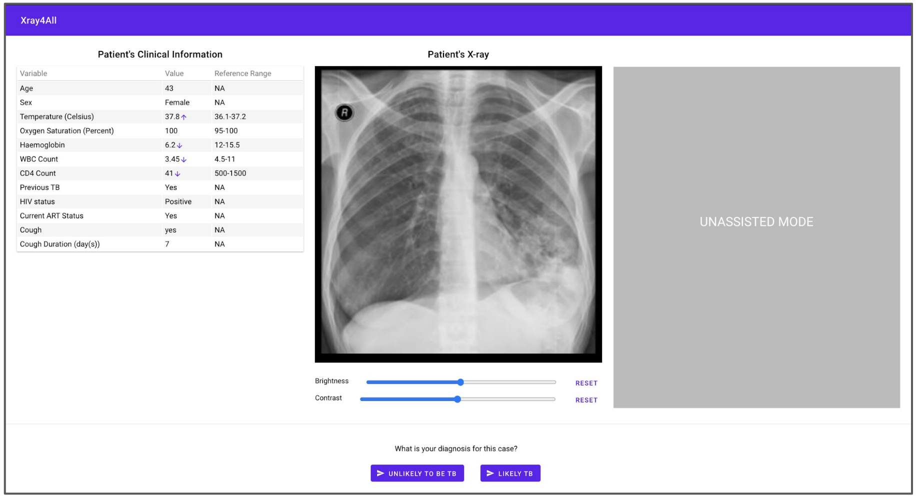


Variable	Value	Reference Range
Age	43	NA
Sex	Female	NA
Temperature (Celsius)	37.8 ↑	36.1-37.2
Oxygen Saturation (Percent)	100	95-100
Haemoglobin	6.2 ↓	12-15.5
WBC Count	3.45 ↓	4.5-11
CD4 Count	41 ↓	500-1500
Previous TB	Yes	NA
HIV status	Positive	NA
Current ART Status	Yes	NA
Cough	yes	NA
Cough Duration (day(s))	7	NA

P. Rajpurkar, C. O'Connell et al.. CheXaid. npj Digital Medicine. 2020.

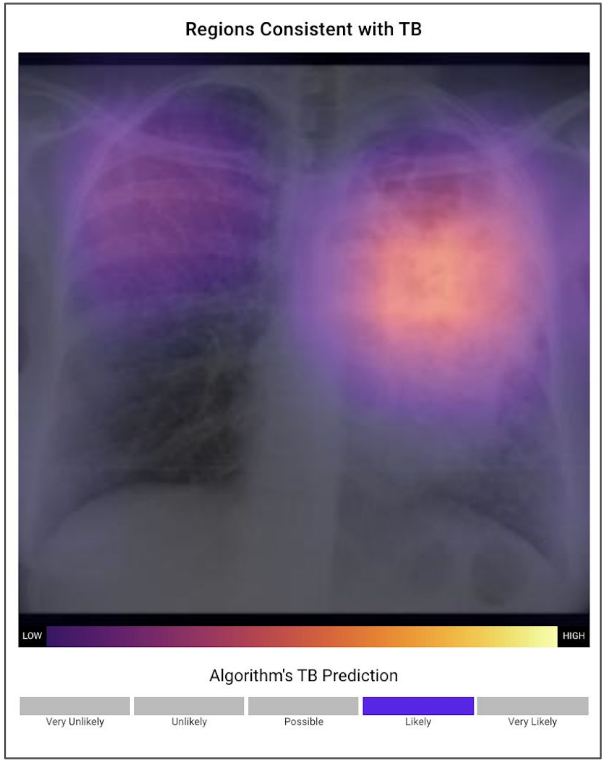
Within-subjects, intermodal study

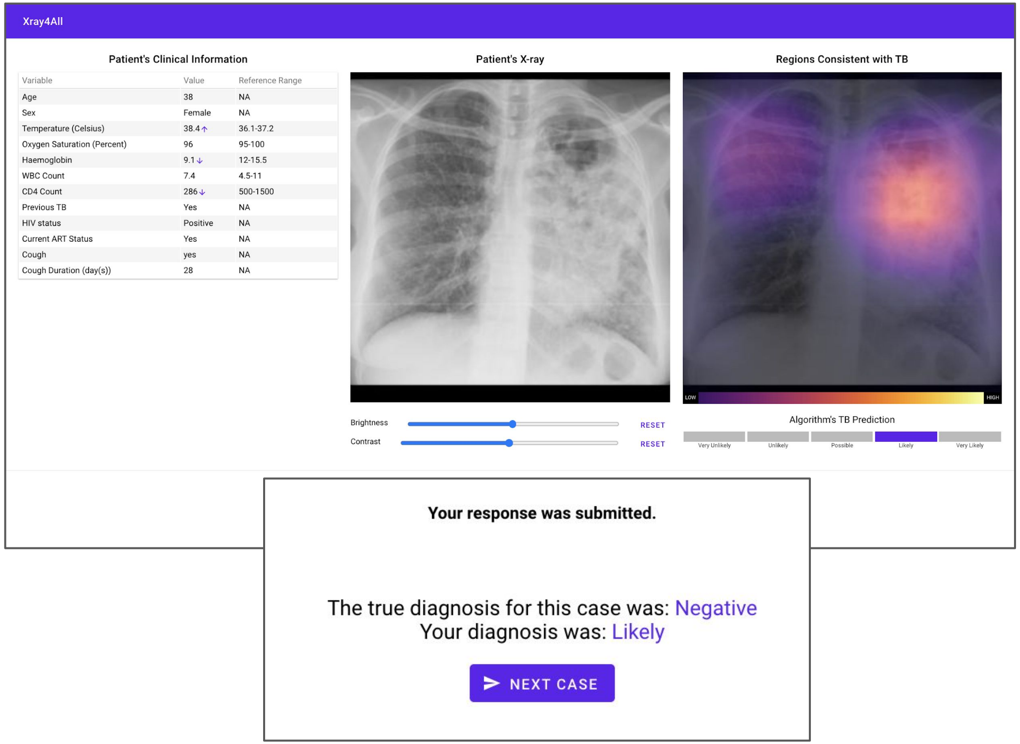




P. Rajpurkar, C. O'Connell et al.. CheXaid. npj Digital Medicine. 2020.

Assistance format and training sessions





P. Rajpurkar, C. O'Connell et al.. CheXaid. npj Digital Medicine. 2020.

Clinicians Assisted with model are more accurate than unassisted

	Accuracy	Sensitivity	Specificity
Clinicians Assisted	0.653	**0.728**	0.609
Clinicians Unassisted	0.602	0.704	0.521

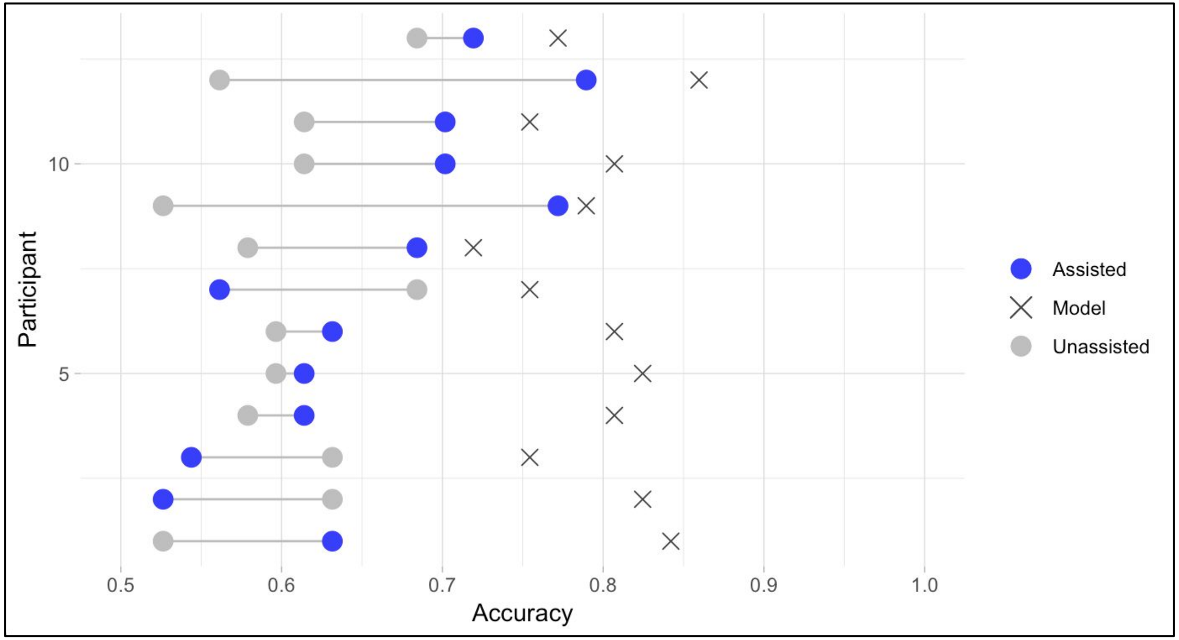
P. Rajpurkar, C. O'Connell et al.. CheXaid. npj Digital Medicine. 2020.

Stand-alone algorithm more accurate than clinicians with assistance

	Accuracy	Sensitivity	Specificity
Clinicians Assisted	0.653	**0.728**	0.609
Clinicians Unassisted	0.602	0.704	0.521
Stand-Alone Algorithm	**0.794**	0.671	**0.871**

P. Rajpurkar, C. O'Connell et al.. CheXaid. npj Digital Medicine. 2020.

Improvement was not consistent across the physicians



Each cross represents the stand-alone algorithm's performance on test data that was assigned as assisted cases for the correspondent physician.

Mistrust in algorithm output or overconfidence in own diagnosis?

P. Rajpurkar, C. O'Connell et al.. CheXaid. npj Digital Medicine. 2020.

Expert-level AI ❯ improved clinician performance in workflow is misguided. Future?

Experience levels	Clinician interaction
Case difficulty	Automation bias

Open Benchmark Curation

Private data → Diverse public data

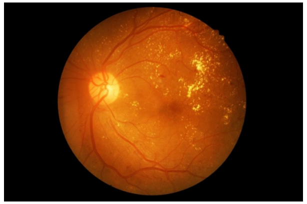
Large dataset and competition hosting

Medical Dataset Curation is hard

Partnerships with hospitals

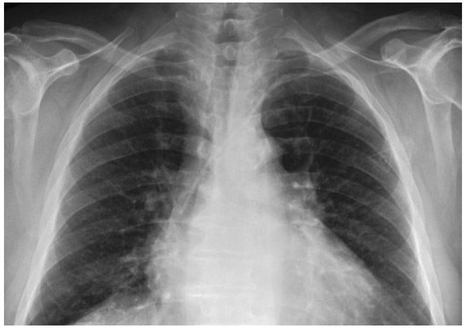
IT Frameworks for
de-identifying and pulling

Expensive manual



Medical data annotation can be creative

Release of NIH Chest X-Ray14 with 100,000+ examples



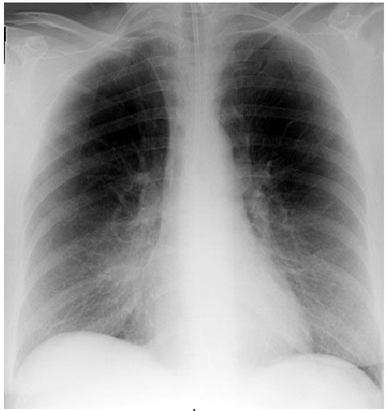
Wang et al. Chest X-Ray14. CVPR. 2017

Medical data annotation can be creative

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

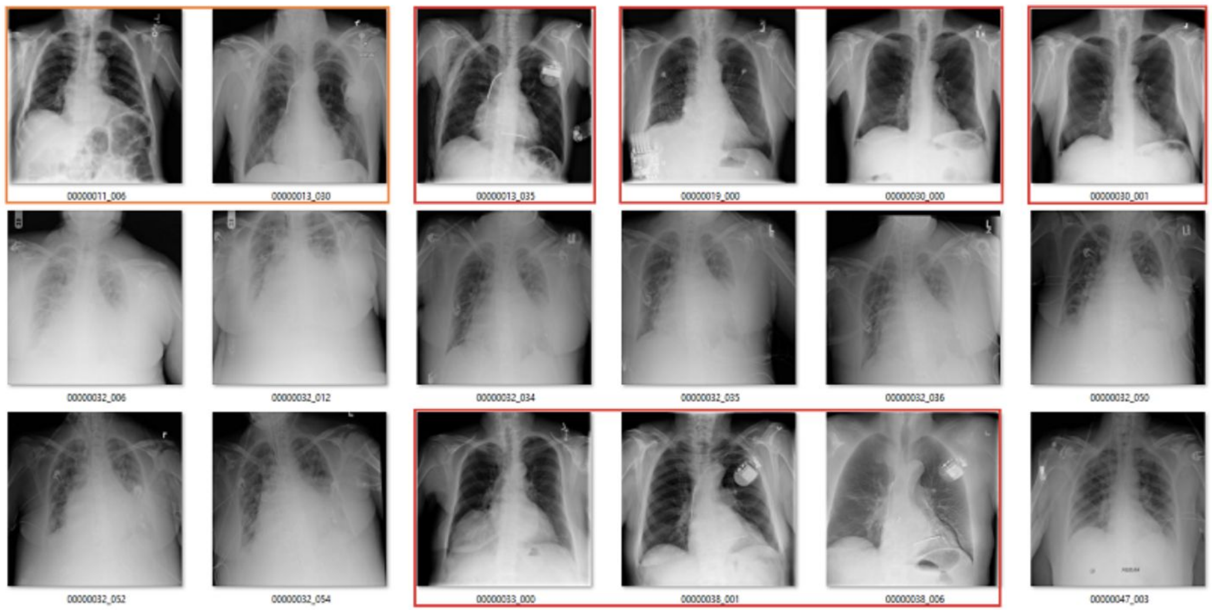
Report labeling NLP System

Is disease present / absent / **uncertain** ?



Labeled 1 if present
Labeled **0** if absent / uncertain

Criticized for high label noise



Oakden-Rayner. Exploring the ChestXray14 dataset: problems. 2017

Improved over previous best labeler through error analysis

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	

Mentions in the report (red) and classification of the uncertainties (purple) and negations (blue).

Irvin & Rajpurkar et al.. CheXpert. AAAI. 2019.

Across all observations and on all tasks, CheXpert labeler achieves a higher F1 score

	Mention F1		Negation F1		Uncertain F1	
Category	NIH	Ours	NIH	Ours	NIH	Ours
Atelectasis	0.976	**0.998**	0.526	**0.833**	0.661	**0.936**
Cardiomegaly	0.647	**0.973**	0.000	**0.909**	0.211	**0.727**
Consolidation	0.996	**0.999**	0.879	**0.981**	0.438	**0.924**
Edema	0.978	**0.993**	0.873	**0.962**	0.535	**0.796**
Pleural Effusion	0.985	**0.996**	0.951	**0.971**	0.553	**0.707**
Pneumonia	0.660	**0.992**	0.703	**0.750**	0.250	**0.817**
Pneumothorax	0.993	**1.000**	0.971	**0.977**	0.167	**0.762**

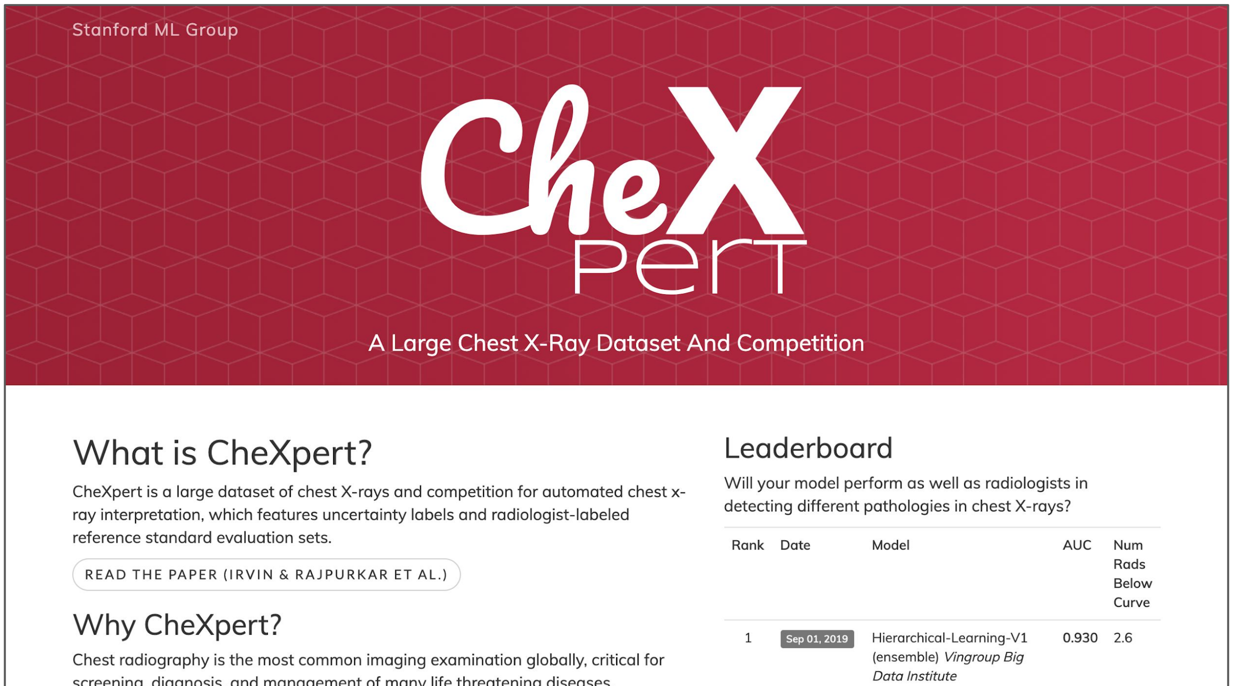
Irvin & Rajpurkar et al.. CheXpert. AAAI. 2019.

Open-sourced labeler & used to label MIMIC-CXR (MIT/Harvard/BIDMC)

	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**

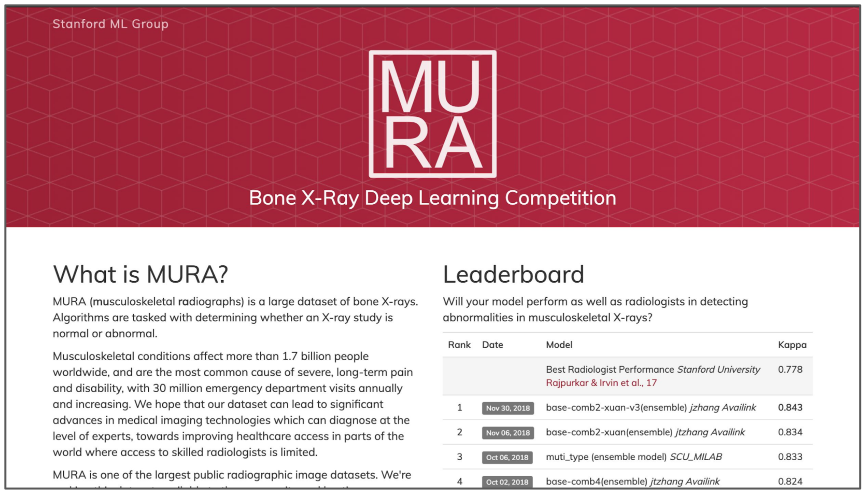
Johnson AE et al., MIMIC-CXR, a de-identifed publicly available database of chest radiographs with free-text reports. Scientific Data. 2019

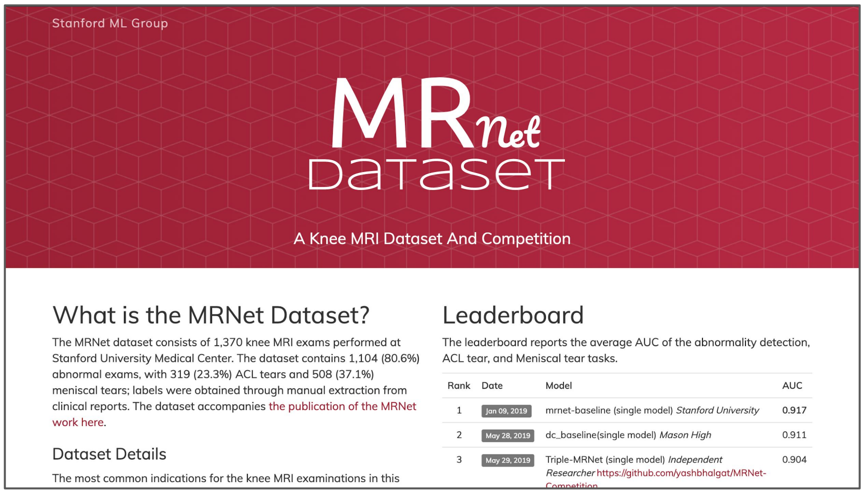
Release one of the largest datasets and competition to the world



Irvin & Rajpurkar et al.. CheXpert. AAAI. 2019.

Other publicly released medical benchmarks





MURA. Rajpurkar et al., MIDL 2018
MRNet. Bien et al., PLOS Medicine 2018

Datasets has been widely used for development and analysis of algorithms

Google AI

facebook Artificial Intelligence

Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis

Agostina J. Larrazabal[a,1], Nicolás Nieto[a,b,1], Victoria Peterson[b,c], Diego H. Milone[a], and Enzo Ferrante[a,2]

[a]Research Institute for Signals, Systems and Computational Intelligence sinc(i), Universidad Nacional del Litoral–Consejo Nacional de Investigaciones Científicas y Técnicas CONICET, Santa Fe CP3000, Argentina; [b]Instituto de Matemática Aplicada del Litoral, Universidad Nacional del Litoral–Consejo Nacional de Investigaciones Científicas y Técnicas, Santa Fe CP3000, Argentina; and [c]Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Oro Verde CP3100, Argentina

CheXclusion: Fairness gaps in deep chest X-ray classifiers

Laleh Seyyed-Kalantari[1,2*], Guanxiong Liu[1,2], Matthew McDermott[3], Irene Y. Chen[3], Maryzeh Ghassemi[1,2]

[1]*Computer Science, University of Toronto, Toronto, Ontario, Canada*

[2]*Vector Institute, Toronto, Ontario, Canada*

[3]*Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA USA*

https://stanfordmlgroup.github.io/competitions/chexpert/

Datasets have advanced modelling insights

5,800 Users
180 Teams Competing

Large improvement in performance attributed to incorporation of **hierarchy** and **uncertainty labels**.

Rank	Date	Model	AUC	Num Rads Below Curve
1	Aug 31, 2020	DeepAUC-v1 *ensemble*	**0.930**	2.8
2	Sep 01, 2019	Hierarchical-Learning-V1 (ensemble) *Vingroup Big Data Institute* https://arxiv.org/abs/1911.06475	0.930	2.6
3	Oct 15, 2019	Conditional-Training-LSR *ensemble*	0.929	2.6

88	Jan 23, 2019	Stanford Baseline (ensemble) *Stanford University* https://arxiv.org/abs/1901.07031	0.907	1.8

Datasets are toy task setups

Poor coverage of diseases

Poor heterogeneity in patients

Poor generalization across clinical workflows

No use of clinical context

No use of priors

Datasets are toy task setups

Poor coverage of diseases

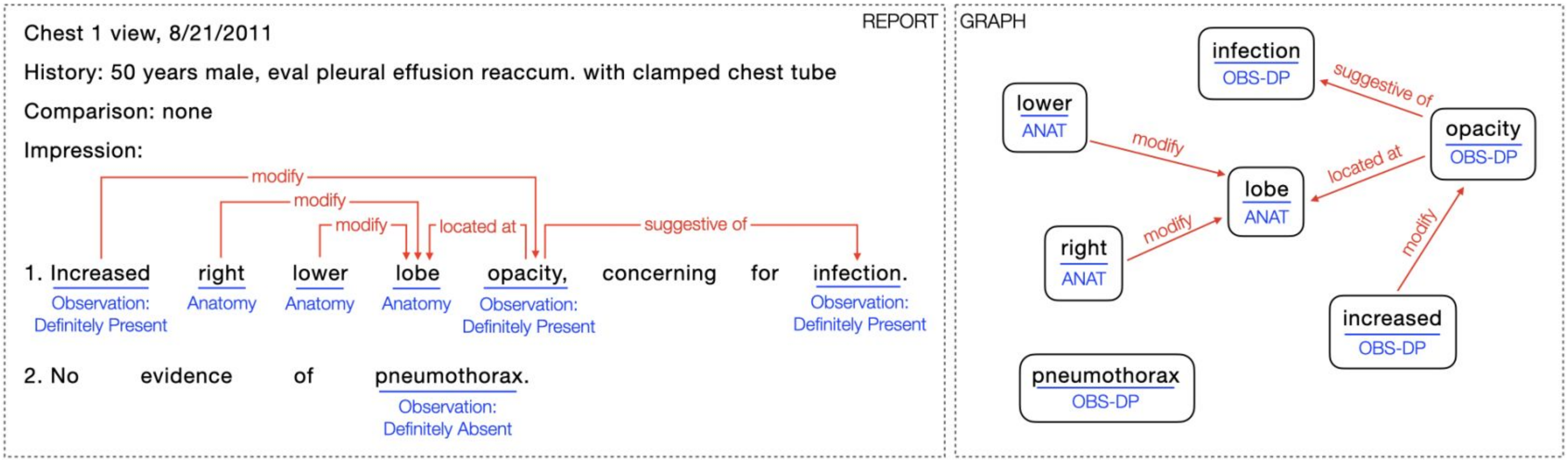
Poor heterogeneity in patients

Poor generalization across clinical workflows

No use of clinical context

No use of priors

Expanding scope of tasks



Publicly release a dataset, which contains annotations automatically generated by RadGraph Benchmark for 200,000+ reports, consisting of over 6 million entities and 4 million relations.

Jain, S et al. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. NeurIPS datasets 2021

Datasets are toy task setups

Poor coverage of diseases

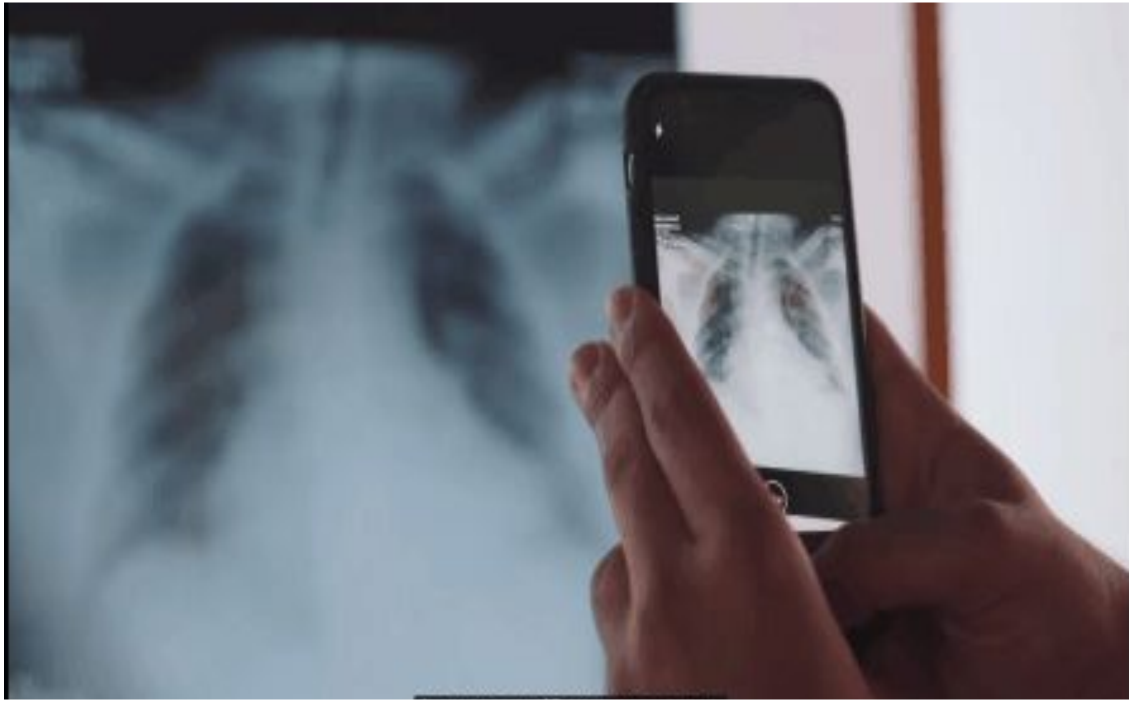
Poor heterogeneity in patients

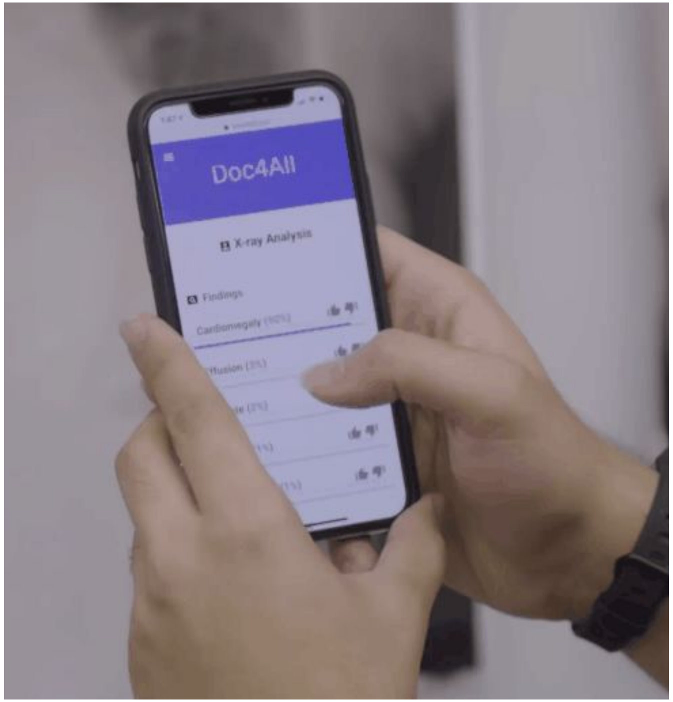
Poor generalization across clinical workflows

No use of clinical context

No use of priors

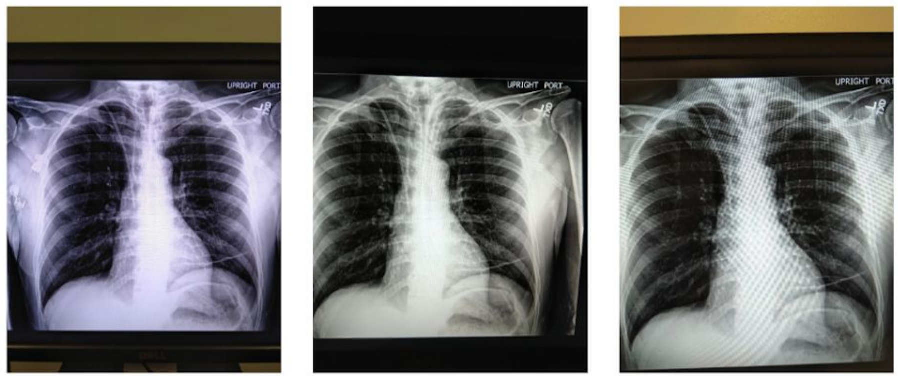
Deployment difference across clinical workflows





Increase access, especially in non-digital workflows

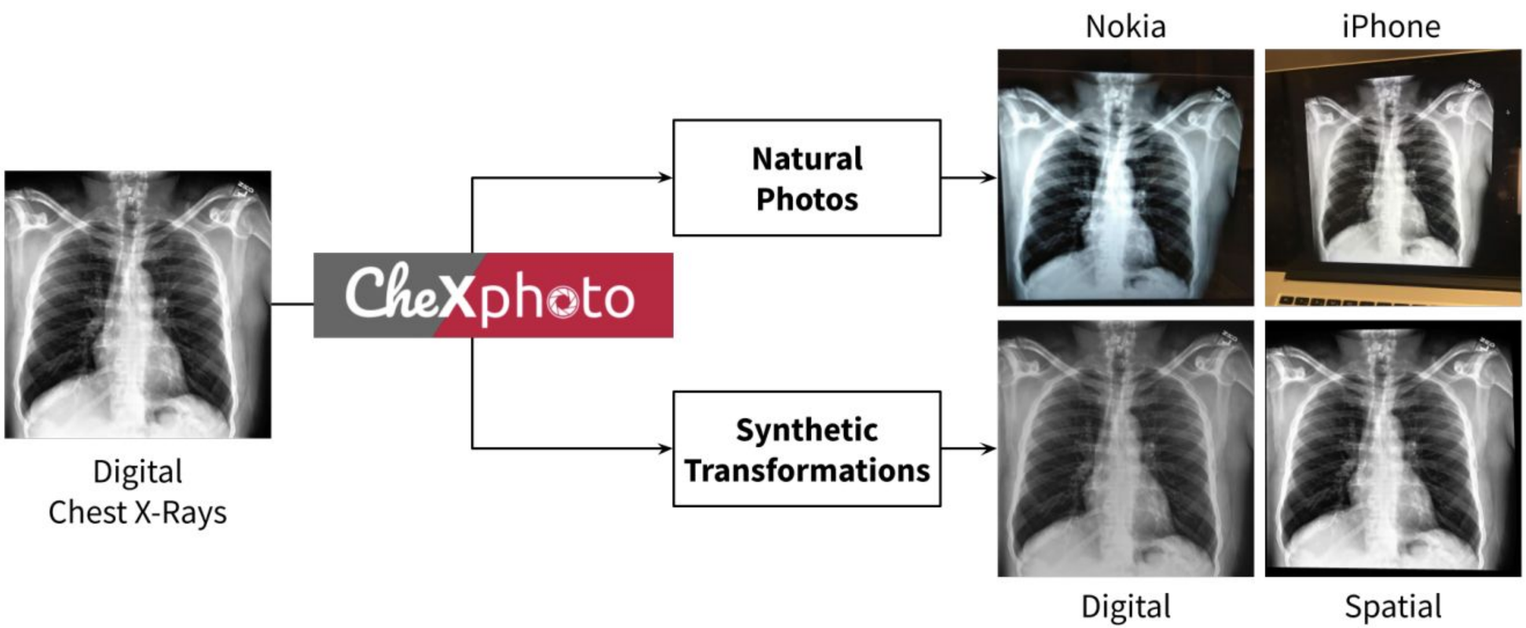
Clinical workflow integrations
Often require new modules



	Comparison	Result
AUC	Photos	0.856 (0.840, 0.869)
	Standard	0.871 (0.855, 0.863)
AUC	Standard-Photos	0.016 (0.012, 0.019)

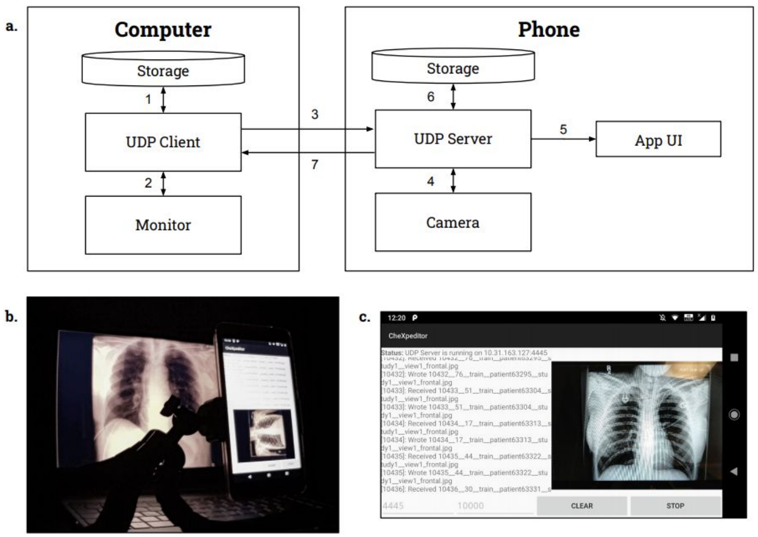
P. Rajpurkar et al., CheXternal. ACM CHIL 2021.

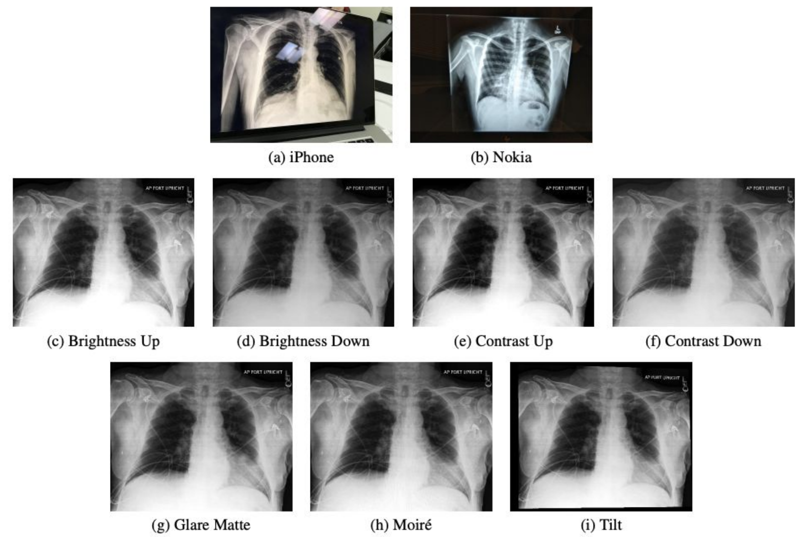
New modules often require new datasets



N. A. Phillips* & P. Rajpurkar* et al. CheXphoto. PMLR. 2020.

Training set of natural photos and synthetic transformations





Test set also includes images from Vietnam deployment setting

N. A. Phillips* & P. Rajpurkar* et al. CheXphoto. PMLR. 2020.

Open dataset and competition release!

Stanford ML Group

CheXphoto

A Perturbed Chest X-Ray Dataset And Competition

What is CheXphoto?

CheXphoto is a competition for x-ray interpretation based on a new dataset of naturally and synthetically perturbed chest x-rays hosted by Stanford and VinBrain.

READ THE PAPER (PHILLIPS, RAJPURKAR & SABINI ET AL.)

Why CheXphoto?

Chest radiography is the most common imaging examination globally, and is critical for screening, diagnosis, and management of many life threatening diseases. Most chest x-ray algorithms have been developed and validated on digital x-rays, while the vast majority of developing regions use films. An appealing solution to

Leaderboard

Will your model perform as well as radiologists in detecting different pathologies in chest X-rays?

We have launched as of August 18, 2021.

Rank	Date	Model	AUC Film	AUC Digital
1	Oct 01, 2021	LBC-v2 (ensemble) *Macao Polytechnic Institute*	0.850	0.89

https://stanfordmlgroup.github.io/competitions/chexphoto/

Community Outreach Efforts

We aim to equip the community to play an active role in the medical AI transformation

MOOC

AI For Medicine

Three-Course Series

52,000 students enrolled

Podcast

The AI Health

Podcast

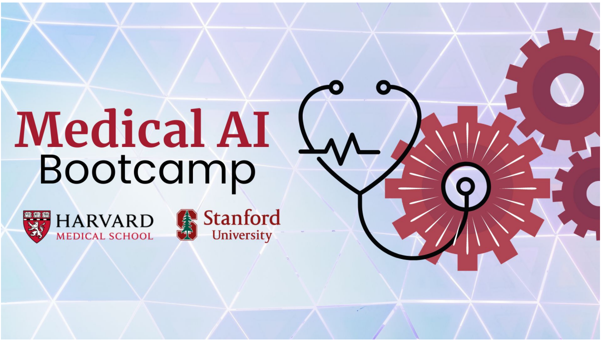
1000 weekly listeners; 2 seasons

Newsletter

Doctor Penguin AI Health

Newsletter

5000 weekly readers



Medical AI Bootcamp

A Harvard-Stanford Program for closely mentored research at the intersection of AI and Medicine. Over 6 months, graduate and undergraduate students receive training to work on high-impact research problems in small interdisciplinary teams.

http://medical-ai-bootcamp.hms.harvard.edu/