Human-AI Collaboration in Healthcare

Nikhil Agarwal (MIT)

Health Economics Conference, TSE

June 19, 2025

AI in Healthcare

Rapid development of Artificial Intelligence (AI) tools for Healthcare

- Clinical decision support [diagnostic and treatment recommendations]
- Operational efficiency [ER triage, allocation of resources]
- Drug discovery [vaccine and gene therapy design]

...

- New Applications [personalized medicine, virtual assistants]

AI in Healthcare

Rapid development of Artificial Intelligence (AI) tools for Healthcare

- Clinical decision support [diagnostic and treatment recommendations]
- Operational efficiency [ER triage, allocation of resources]
- Drug discovery [vaccine and gene therapy design]

. . .

- New Applications [personalized medicine, virtual assistants]
- ✓ Interest in Medicine, CS, and Economics [Mullainathan & Obermeyer, 2021; Rajpurkar et al., 2017; Lakkaraju & Farronato, 2023; Goh et al., 2024 . . .]

AI in Healthcare

Rapid development of Artificial Intelligence (AI) tools for Healthcare

- Clinical decision support [diagnostic and treatment recommendations]
- Operational efficiency [ER triage, allocation of resources]
- Drug discovery [vaccine and gene therapy design]

. . .

- New Applications [personalized medicine, virtual assistants]
- ✓ Interest in Medicine, CS, and Economics [Mullainathan & Obermeyer, 2021; Rajpurkar et al., 2017; Lakkaraju & Farronato, 2023; Goh et al., 2024 . . .]

Classification problems are common in medicine

AI in Healthcare

Rapid development of Artificial Intelligence (AI) tools for Healthcare

- ▶ Clinical decision support [diagnostic and treatment recommendations]
- ▶ Operational efficiency [ER triage, allocation of resources]
- ▶ Drug discovery [vaccine and gene therapy design]

...

- ▶ New Applications [personalized medicine, virtual assistants]
- ✓ Interest in Medicine, CS, and Economics [Mullainathan & Obermeyer, 2021; Rajpurkar et al., 2017; Lakkaraju & Farronato, 2023; Goh et al., 2024 ...]

Classification problems are common in medicine

- ▶ Radiology is an iconic example:

> *"We should stop training radiologists now. It's just completely obvious that within five years, deep learning is going to do better than radiologists"*
>
> — Geoffrey Hinton (in 2016)
>
> [see also Obermeyer and Emmanuel, NEJM 2016]

Will AI Replace Radiologists?

"The right answer is: Radiologists who use AI will replace radiologists who don't."

— Curtis Langlotz (2019)

- Partial task automation [radiologists can diagnose the "long-tail" of diseases]
- Radiologists can master new imaging technology
- AI assistance can help radiologists

Will AI Replace Radiologists?

"The right answer is: Radiologists who use AI will replace radiologists who don't."

— Curtis Langlotz (2019)

- Partial task automation [radiologists can diagnose the "long-tail" of diseases]
- Radiologists can master new imaging technology
- AI assistance can help radiologists

"Focus is placed on the performance of the human-AI team"

– Joint statement by US FDA, and Canada and UK MHRA

- Approval of autonomous diagnostic AI is rare
- Presumption of human oversight, except for low-risk applications

Humans vs AI, or Collaboration?

Questions:

1. What are the relative strengths and weaknesses of humans and AI?

Humans vs AI, or Collaboration?

Questions:

1. What are the relative strengths and weaknesses of humans and AI?
2. How should we design human-AI collaboration?

Humans vs AI, or Collaboration?

Questions:

1. What are the relative strengths and weaknesses of humans and AI?
2. How should we design human-AI collaboration?

Humans' potential strengths in diagnostic imaging

1. Have access to valuable information (non-systematic) data
2. Diagnosing the "long-tail"

Humans vs AI, or Collaboration?

Questions:

1. What are the relative strengths and weaknesses of humans and AI?
2. How should we design human-AI collaboration?

Humans' potential strengths in diagnostic imaging

1. Have access to valuable information (non-systematic) data
2. Diagnosing the "long-tail"

Designing Human-AI Collaboration

- How do humans incorporate AI information?

An Experiment on Human-AI Collaboration

- Largest experiment with radiologists' use of AI [Agarwal et.al., 2023; R&R ECMA]
  - 227 radiologists, approx 90 cases with X-rays
  - AI assistance from CheXperT [Irvin et al., 2019]
  - 2 x 2 design varying AI assistance and clinical history

Alex Moehring (Purdue) Pranav Rajpurkar (HMS) Tobias Salz (MIT)

An Experiment on Human-AI Collaboration

- Largest experiment with radiologists' use of AI [Agarwal et.al., 2023; R&R ECMA]
  - 227 radiologists, approx 90 cases with X-rays
  - AI assistance from CheXperT [Irvin et al., 2019]
  - 2 x 2 design varying AI assistance and clinical history

Alex Moehring (Purdue) Pranav Rajpurkar (HMS) Tobias Salz (MIT)

- Collaborators:
  - Radiologists at Mt. Sinai (NYC), Stanford, VINBrain
  - Three US teleradiology companies

Research Questions

1. **Today's Focus:** How should human-AI collaboration be designed? [Agarwal et.al., 2023; R&R ECMA]

   i. Measure predictive value of **contextual information**

   ii. Measure **biases in belief updating** relative to Bayesian benchmark

   iii. Solve **optimal collaboration** between humans and machines

$$\tau : s^A \rightarrow \{\text{Human, AI, Human+AI}\}$$

Research Questions

1. **Today's Focus:** How should human-AI collaboration be designed? [Agarwal et.al., 2023; R&R ECMA]
   i. Measure predictive value of **contextual information**
   ii. Measure **biases in belief updating** relative to Bayesian benchmark
   iii. Solve **optimal collaboration** between humans and machines

$$\tau : s^A \to \{\text{Human, AI, Human+AI}\}$$

2. Other Results:
   i. Which types of radiologists use AI assistance well? [Yu et al., 2024; Nature Medicine]
   ii. Are humans better at predicting the long tail? [Agarwal et al., 2024; AEA: P&P]
   iii. A public dataset [Moehring et al., 2025; Scientific Data]

Outline

Outline

Overview of the Experimental Design

2 x 2 (x 2) Design

Treatment Dimension 1: Access to AI prediction (AI)

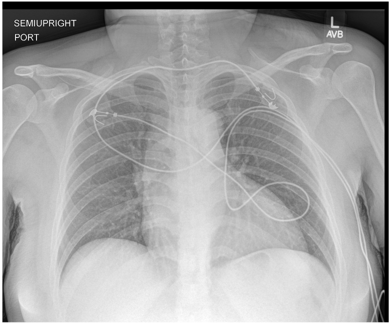
Treatment Dimension 2: Clinical History (CH)

(Treatment Dimension 3: Incentives for Accuracy [BSR: Hossain and Okui, 2013])

Radiologists **participate remotely** through tailormade interface

- Mimics clinical practice but generates structured quantifiable report
- In collaboration with radiologists at Stanford and Mt. Sinai (NYC)
- 324 historical cases from Stanford Healthcare System with Chest-X-ray and clinical history, manually reviewed for public release
- ✓ Structured data entry v. free text report

Interface



Zoom In | Zoom Out | Reset X-Ray | Full Screen

Contrast

Brightness

Airspace Opacity

AI Prediction: **12% (Very unlikely)**

Highly unlikely	Very unlikely	Unlikely	Possible	Likely	Highly likely

Probability of Airspace Opacity: 43%

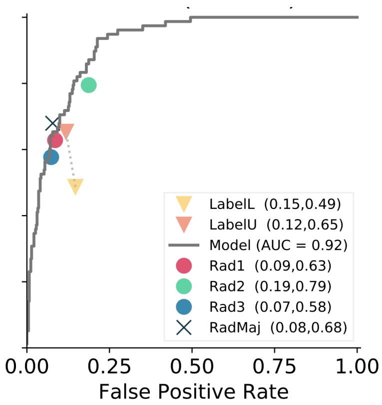
Size ○ Small ● Medium ○ Large ○ Very Large

Recommend follow up ● Yes ○ No

Treatment Dimension 1: AI Algorithm

CheXperT

- Trained on reports from $\geq 250,000$ chest X-rays
- Probabilities for 14 pathologies
- Performance matches board certified Stanford radiologists



$\rightarrow$ **AI treatments:** access to CheXperT's probability of disease presence.

Treatment Dimension 2: Clinical history

Provided information

- Vitals
- Demographic variables
- Indications
- Labs

Indication

30 years of age, Female, history of hypertension, abnormal EKG, abdominal pain, evaluate for cardiomegaly or mediastinal widening.

Vitals

Variable	Value
Weight	170 lbs
BP	243/166 mmHg
Temp	99.1F
Pulse	99.0 bpm
Age	30

Abnormal Labs All Labs

Variable	Value	Unit	Flag
ALT (SGPT), Ser/Plas	38.0	U/L	High
AST (SGOT), Ser/Plas	39.0	U/L	High
Eosinophil, Absolute	0.01	K/uL	Low

Diagnostic Standard

Diagnostic standard ω_i constructed using aggregate assessment of experts

- Five board certified chest radiologists from Mount Sinai Health Care System
- Follows the medical AI literature [Irving et al., 2019; McCluskey et al., 2021]

Definitive diagnostic test typically unavailable

- Selective labels problem when administered [e.g. Mullainathan and Obermeyer, 2022]

Diagnostic Standard

Diagnostic standard ω_i constructed using aggregate assessment of experts

- Five board certified chest radiologists from Mount Sinai Health Care System
- Follows the medical AI literature [Irving et al., 2019; McCluskey et al., 2021]

Definitive diagnostic test typically unavailable

- Selective labels problem when administered [e.g. Mullainathan and Obermeyer, 2022]

Baseline uses cutoff at $\bar{p} = 0.5$ [Wallsten and Diederich, 2001]

- Robust to log-odds averaging Definition
- Robustness to comparisons with $\bar{p}$

Experimental Design

Challenges:

- Compare w/ Bayesian benchmark $\rightarrow$ need linked assessments w/ and w/o AI
- Power $\rightarrow$ Expensive subject pool ($\approx \$10$ a case)

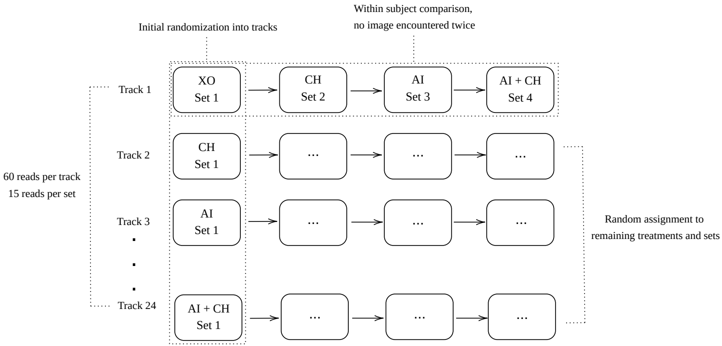
Approach: Hybrid design that collects both within and across subject data

1. All radiologists are exposed to all treatments
   - ✓ Enables within comparisons
   - ✓ Across-radiologist comparison based on first treatment
2. Subset of radiologists read the same case both with and without AI
   - ✓ Allows estimating and comparing with Bayesian benchmark
   - ✓ Two-week wash-out period to address memory

Primary Across Design

Simple across design with a within subject component

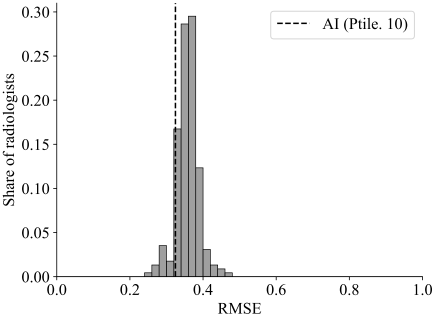
- ✓ Clear across design
- ✓ Within subject comparison hedges power
- ▶ Two variations targeted for estimating biases in belief updating



AI Performance

Radiologists and AI performance:

- Algorithm performs better than most radiologists in our sample



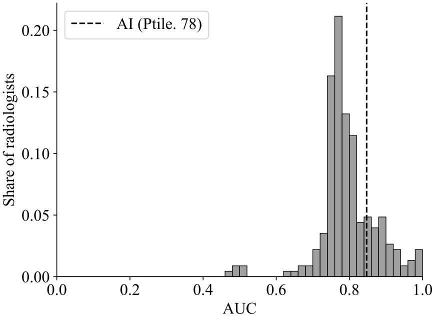
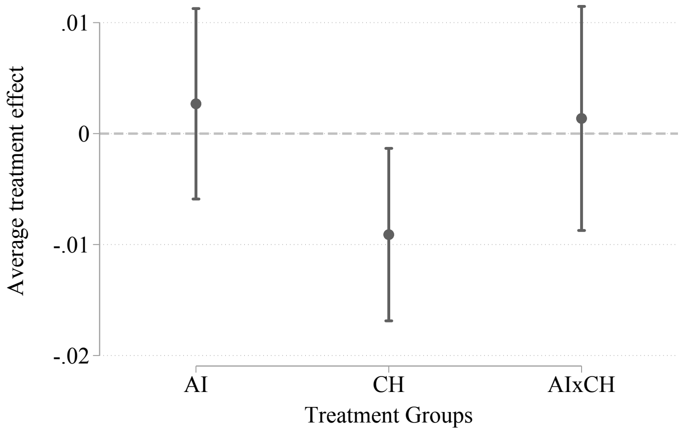


Image Only | Design 2 | Design 3 | Internal GT

Outline

Treatment Effect — Deviation from Diagnostic Standard



No effects by CH noted in endline surveys By CH Group

Table Design 2 Design 3 Internal GT

Treatment Effect — Deviation from AI

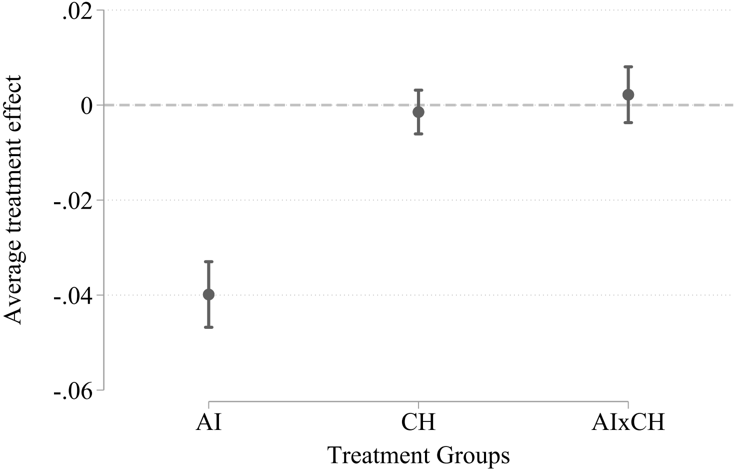
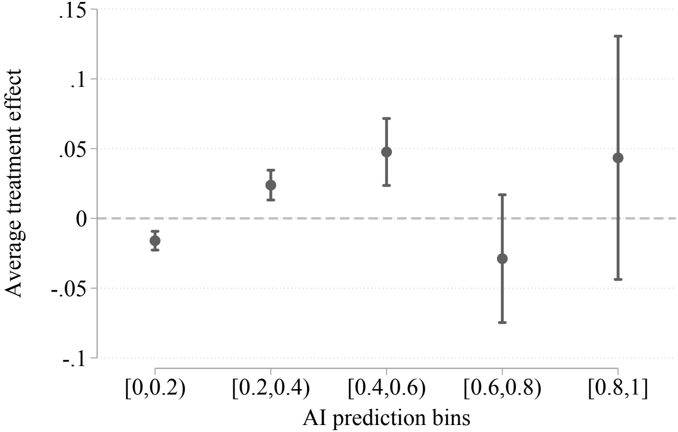


Table
Design 2
Design 3

Deviation from GT — CATE of AI



Regression Table | Pooled AI | Design 2 | Design 3 | Internal GT | From $\bar{p}$ | |AI - GT|

Outline

Biases in Belief Updating

Describe via [building on Grether 1980, 1992]:

$$\overbrace{\ln \frac{p(\omega=1|s_A,s_E)}{p(\omega=0|s_A,s_E)}}^{\text{Decision-relevant posterior log-odds}} = b \cdot \overbrace{\ln \frac{\pi(s_A|\omega=1,s_E)}{\pi(s_A|\omega=0,s_E)}}^{\text{Update from AI}} + \overbrace{\ln \frac{\pi(s_E|\omega=1)}{\pi(s_E|\omega=0)}}^{\text{Own-information log-odds}} + k$$

- Bayesian with correct beliefs $\implies b = 1$

Biases in Belief Updating

Describe via [building on Grether 1980, 1992]:

$$\overbrace{\ln\frac{p(\omega=1|s_A,s_E)}{p(\omega=0|s_A,s_E)}}^{\text{Decision-relevant posterior log-odds}} = b\cdot\overbrace{\ln\frac{\pi(s_A|\omega=1,s_E)}{\pi(s_A|\omega=0,s_E)}}^{\text{Update from AI}}+\overbrace{\ln\frac{\pi(s_E|\omega=1)}{\pi(s_E|\omega=0)}}^{\text{Own-information log-odds}}+k$$

- Bayesian with correct beliefs $\implies b=1$

Terminology:

- Automation bias/neglect: $b \lessgtr 1$
- Neglect signal dependence: Update term doesn't condition on s_E

Biases in Belief Updating

Analysis in the paper

1. Theoretical
   i. AI improves performance if only automation neglect is at play
   ii. Optimal delegation problem sensitive to signal distributions in other cases

Biases in Belief Updating

Analysis in the paper

1. Theoretical
   i. AI improves performance if only automation neglect is at play
   ii. Optimal delegation problem sensitive to signal distributions in other cases

2. Empirical methods
   i. Solve challenges in estimating empirical analog in observational setting
   ii. Develop model selection method to identify type of bias

Biases in Belief Updating

Analysis in the paper

1. Theoretical
   i. AI improves performance if only automation neglect is at play
   ii. Optimal delegation problem sensitive to signal distributions in other cases

2. Empirical methods
   i. Solve challenges in estimating empirical analog in observational setting
   ii. Develop model selection method to identify type of bias

3. Results
   i. Two biases: Automation neglect and signal dependence neglect
   ii. Selected model replicates treatment effect patterns

Biases in Belief Updating

Analysis in the paper

1. Theoretical
   i. AI improves performance if only automation neglect is at play
   ii. Optimal delegation problem sensitive to signal distributions in other cases
2. Empirical methods
   i. Solve challenges in estimating empirical analog in observational setting
   ii. Develop model selection method to identify type of bias
3. Results
   i. Two biases: Automation neglect and signal dependence neglect
   ii. Selected model replicates treatment effect patterns

✓ Potential gains from human-AI collaboration undercut by biases

Optimal Delegation Problem

Optimal delegation solution $\tau^*(s_{A,i}) \in \{$Full Auto, No AI, AI assist$\}$ to

$$\min_{\tau \in \{H, H+AI, AI\}} \overbrace{mV_\tau(s_{A,i})}^{\text{Decision Loss in \$}} + \overbrace{wC_\tau(s_{A,i})}^{\text{Effort cost in \$}}$$

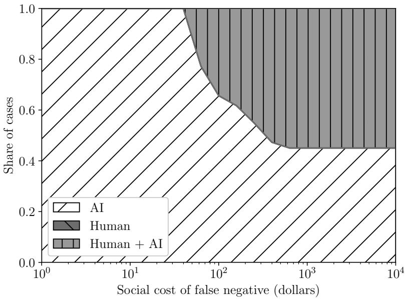
- Measure $C(\cdot)$ in minutes from experiment
- Opportunity cost of radiologist time $w = \$4$ per minute

Unknowns

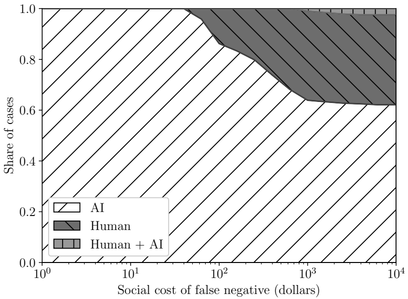
- m – calculate frontier of $V_{i\tau^*}$ and $C_{i\tau^*}$
- $V_{ir\tau}$ – experiment allows estimating (central) c_{rel} for each pathology

Delegation Solution

Bayesians



Humans



$\rightarrow$ Humans are more likely to work alongside AI than with AI [Goh et al., 2024; Agarwal. Moehring, Wolitzky, 2025]

- Potential benefits from training $\rightarrow$ See Bayesian solution

Outline

Which radiologists benefit from AI assistance?

Yu, Moehring, Banerjee, Agarwal, Salz, Rajpurkar, *Nature Medicine*, 2024

Hypothesis: Large benefits from personalized delegation

- ✓ Predict which radiologists do better with AI

Which radiologists benefit from AI assistance?

Yu, Moehring, Banerjee, Agarwal, Salz, Rajpurkar, *Nature Medicine*, 2024

Hypothesis: Large benefits from personalized delegation

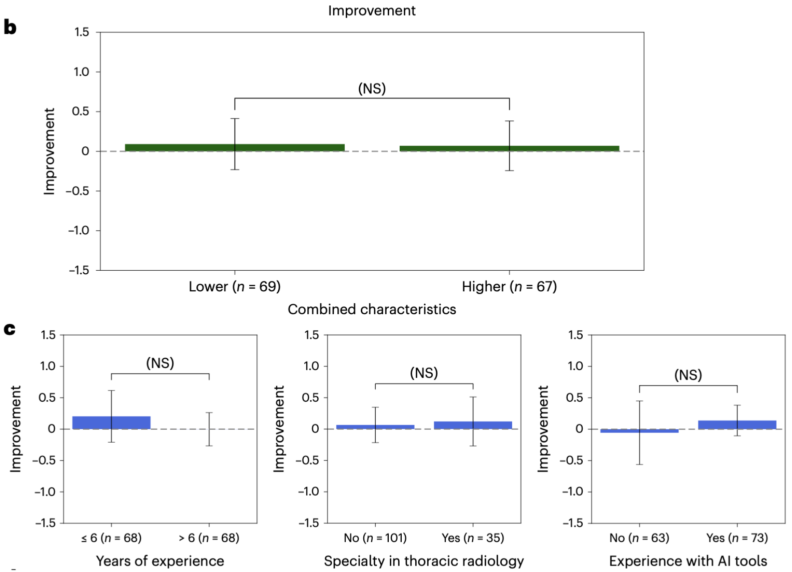
✓ Predict which radiologists do better with AI

Experiment collects data on

- Experience
- Prior experience with AI
- Board certifications and subspecialty

Caveat: 227 radiologists

(Un-)Predictability of Benefits from AI



Is AI an equalizer?

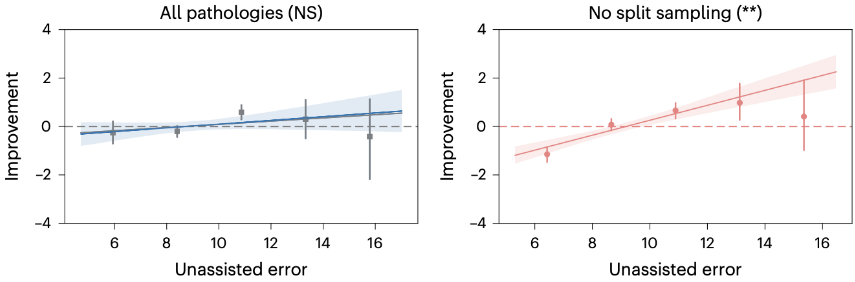
- Do lower-skilled radiologists benefit more? [e.g. Noy and Zhang, 2023]

$$Y_i(\text{AI}) - Y_i(\text{No AI}) = \beta Y_i(\text{No AI}) + \varepsilon_i$$

Is AI an equalizer?

- Do lower-skilled radiologists benefit more? [e.g. Noy and Zhang, 2023]

$$Y_i(\text{AI}) - Y_i(\text{No AI}) = \beta Y_i(\text{No AI}) + \varepsilon_i$$



- Measurement error in $Y_i(\text{No AI})$ biases β $\rightarrow$ Mean reversion
- **Split sample** measure of $Y_i(\text{No AI})$ finds **no relationship**

Outline

The Long Tail Hypothesis

Agarwal, Huang, Moehring, Rajpurkar, Salz, Yu, *AEA: P&P*, 2024

Supervised deep learning requires large labeled training datasets [see LeCun, Bengio, Hinton, 2015, for a review]

- Few annotated examples of rare cases even in very large datasets

The Long Tail Hypothesis

Agarwal, Huang, Moehring, Rajpurkar, Salz, Yu, *AEA: P&P*, 2024

Supervised deep learning requires large labeled training datasets [see LeCun, Bengio, Hinton, 2015, for a review]

- Few annotated examples of rare cases even in very large datasets

Humans may be able to learn from limited examples [e.g. Kühl et al, 2020; Malaviya et al., 2022]

- Training data used in supervised learning outstrips human experience
  - CheXpert model is trained on $\approx 220,000$ radiographs
  - ✓ Assuming three mins per case, a human review would take > 6.5 years of FTE work

Zero-shot learning algorithms attempt to bridge this gap

- Self-supervised, mimics human inputs and outputs
- Do not require annotated labels

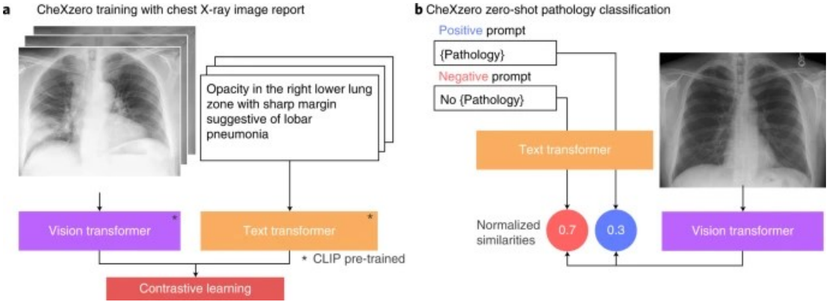
CheXpert vs CheXzero

CheXpert is a supervised learning algorithm

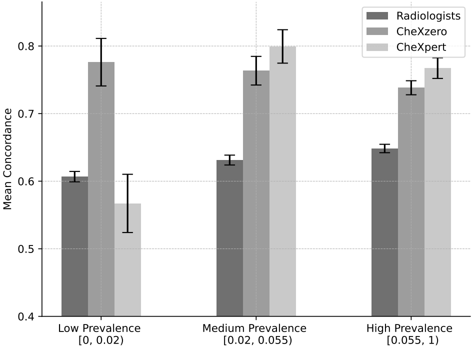
- Predicts 12 binary labels

CheXzero is self-supervised that uses text reports [Tiu et al., 2022]

- Predictions based on comparing a positive and a negative prompt

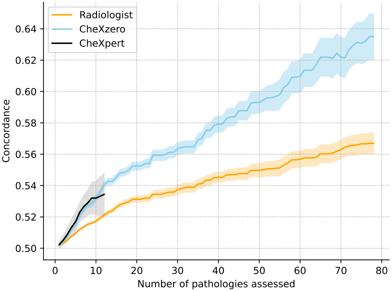


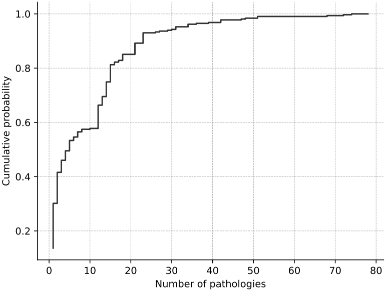
Performance by Prevalence



- CheXpert is substantially more accurate when prevalence is high
- CheXzero and radiologists have more similar performance across prevalence

The Long Tail





- Zero-shot algorithms match or surpass human performance throughout

Concluding: Human-AI in Healthcare

Main findings in Radiology:

1. Biased updating undercuts human-AI collaboration $\rightarrow$ Human or AI
2. AI capabilities continue to improve

Concluding: Human-AI in Healthcare

Main findings in Radiology:

1. Biased updating undercuts human-AI collaboration $\rightarrow$ Human or AI
2. AI capabilities continue to improve

Humans do more than classification in **Healthcare:**

- Example: Diagnosis versus treatment
- Where are there complementarities?

Concluding: Human-AI in Healthcare

Main findings in Radiology:

1. Biased updating undercuts human-AI collaboration $\rightarrow$ Human or AI
2. AI capabilities continue to improve

Humans do more than classification in **Healthcare:**

- Example: Diagnosis versus treatment
- Where are there complementarities?

Beyond Healthcare:

- Organizational incentives
- Training humans to use AI
- Specialization and complementarities

Thank You

email: agarwaln@mit.edu