Artificial Intelligence and the Medical Record in the Context of Social Security Disability Evaluations: A Workshop Medical Imaging

Adam E Flanders MD
Thomas Jefferson University

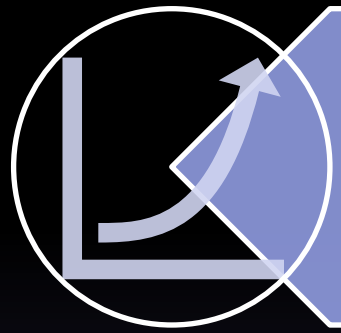
Jefferson®

No relevant disclosures

Outline

- State of the state – Imaging AI applications
- Limitations of medical imaging AI
- Future opportunities
- Warnings

Exponential Growth of Imaging Services



300% increase in demand for CT in 15 years

Majority of imaging is for patients >65 years age

Only 3% commensurate increase in Radiologists over same period.

Bruls, R.J.M. and Kwee, R.M., 2020

Current State of FDA Cleared AI Products

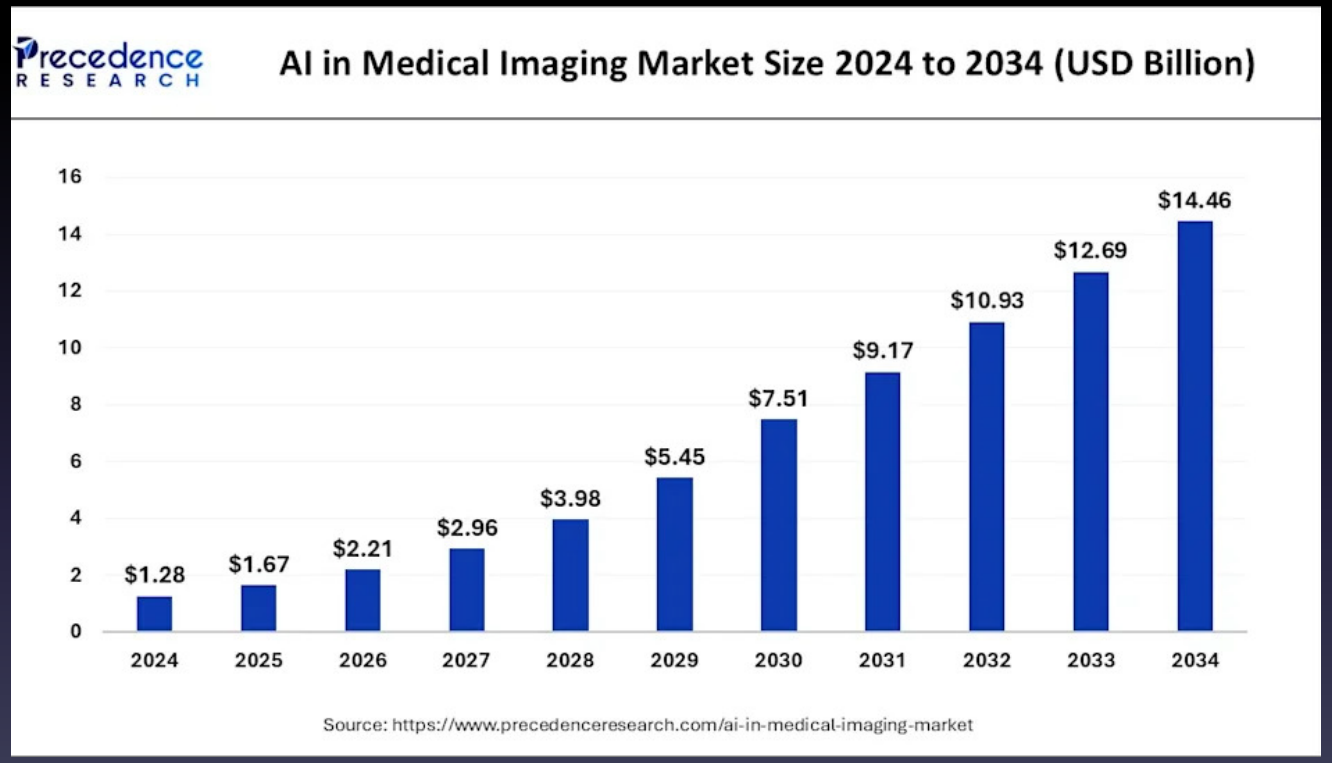
First imaging AI device cleared by FDA 28 years ago IR2 image checker M100 by R2)

Exponential growth from 2015 onward.

1,451 FDA cleared AI devices in the marketplace as of 2026

Radiology dominates market at 76% (1,104)

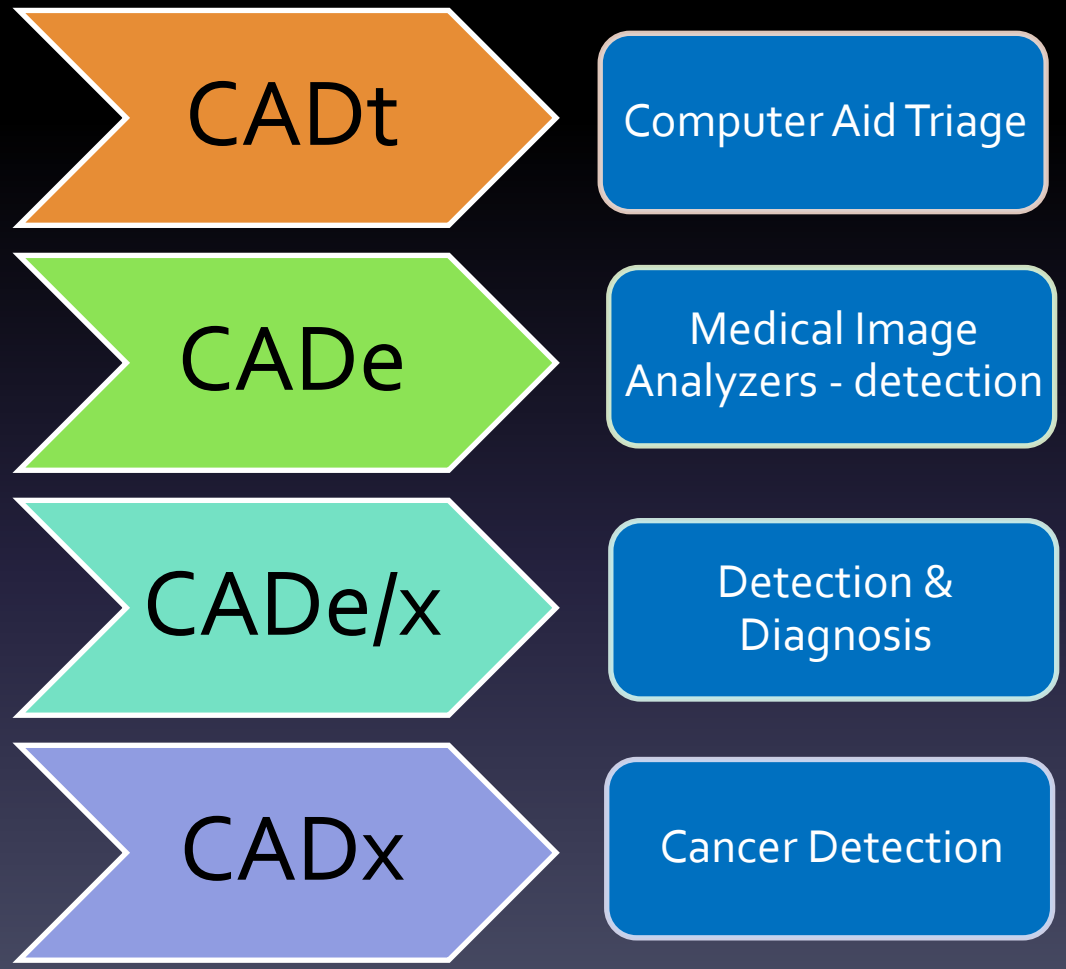
Exponential Growth of Imaging AI



- The global medical imaging market surpassed 49.61 billion USD in 2024
- 80.52 billion USD by 2034
- *The medical imaging AI market size was valued at 2.21 billion USD in 2026*
- *Is projected to reach around 14.46 billion USD by 2034*

https://www.precedenceresearch.com/sample/2380

FDA Classifications for Radiological Computer-Aided/Assisted Diagnostic (CAD) Devices



AI Imaging Categories – Detect, Classify, Triage

- Stroke
- Cerebral hemorrhage
- LVO
- Pulmonary Embolism
- DVT
- Aneurysm
- Lung nodules
- Bone Density/Fracture risk
- Fractures
- Breast masses
- CAC
- Cardiac physiology
- Etc. etc.....

406042006
E20560456

20.1 cm

Ser:5
Img: 168/ 255
512 x 512
Loc: -243.47 mm

ung lesions 6 | Calcification score

les 5 | Vesicular Shadow | Fibrosis 1

t all 5 lesions | Hide micronodules

Solid	3x3mm	LUL apicoposterior	L
GGN	6x5mm	LLL dorsal	L
GGN	6x5mm	RLL anterior basal	L
Solid	3x3mm	RML lateral	L
Solid	5x4mm	LLL lateral basal	L
	17.92mm³	-686 ~ -141HU	

Nodule density analysis

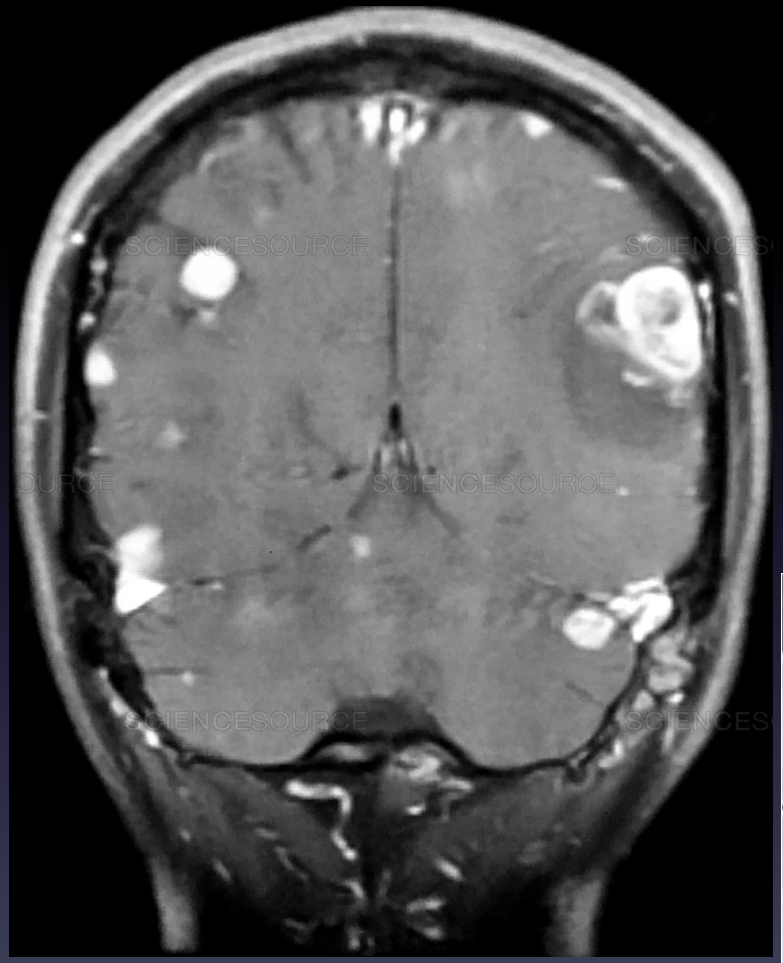
Guideline

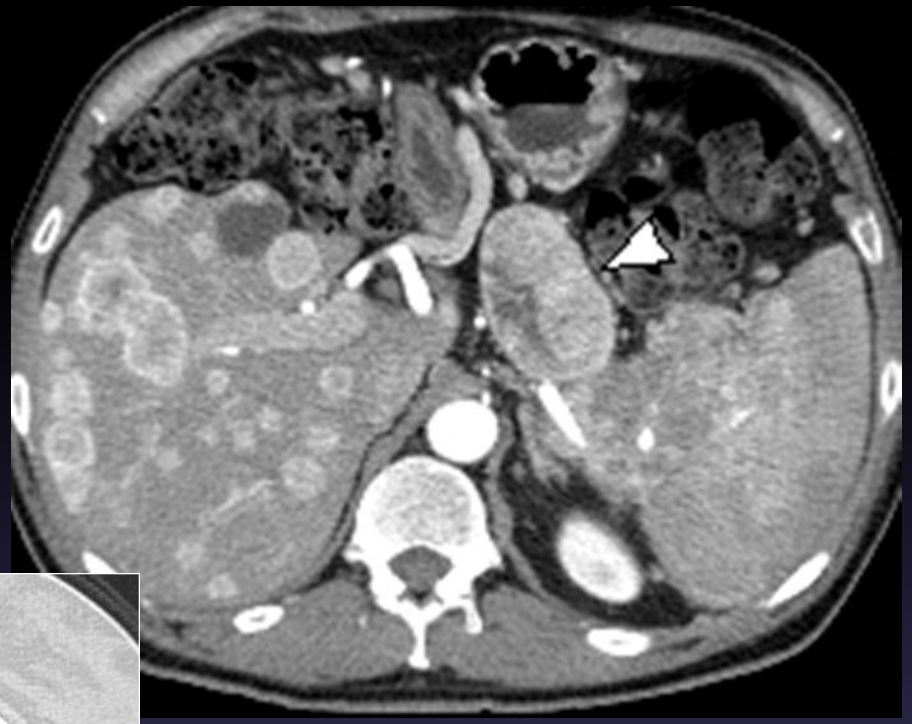
port Single nodule

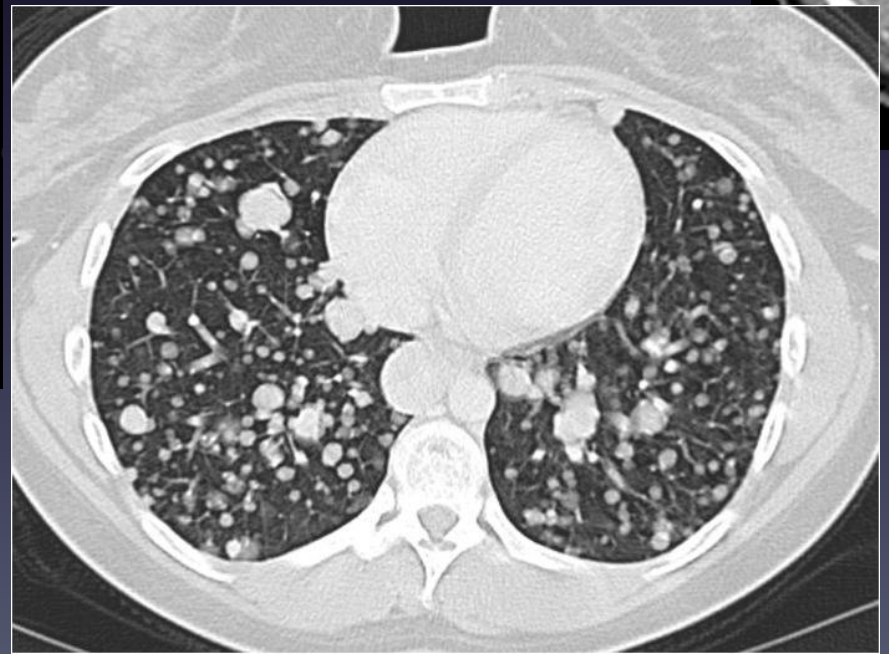
Limitations with Narrow AI Tools

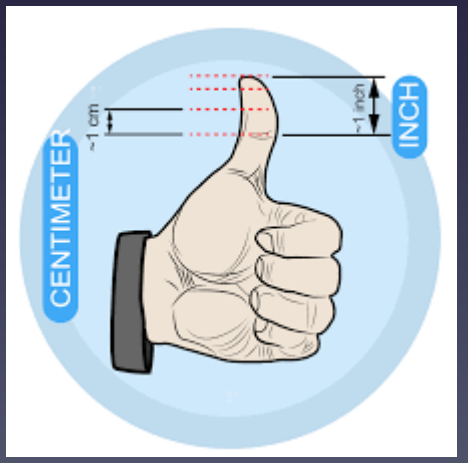
- Large operational expense to support annually.
- With few exceptions, not a billable service; difficult to demonstrate a true ROI or value.
- Tools are not well integrated into the diagnostic "cockpit" of the radiologist - so use/engagement is variable.
- Known performance drift with time that is not easily measured or acted on.
- Industry is not focused entirely on biggest challenges.

Focus on More Tedious Time Consuming Tasks

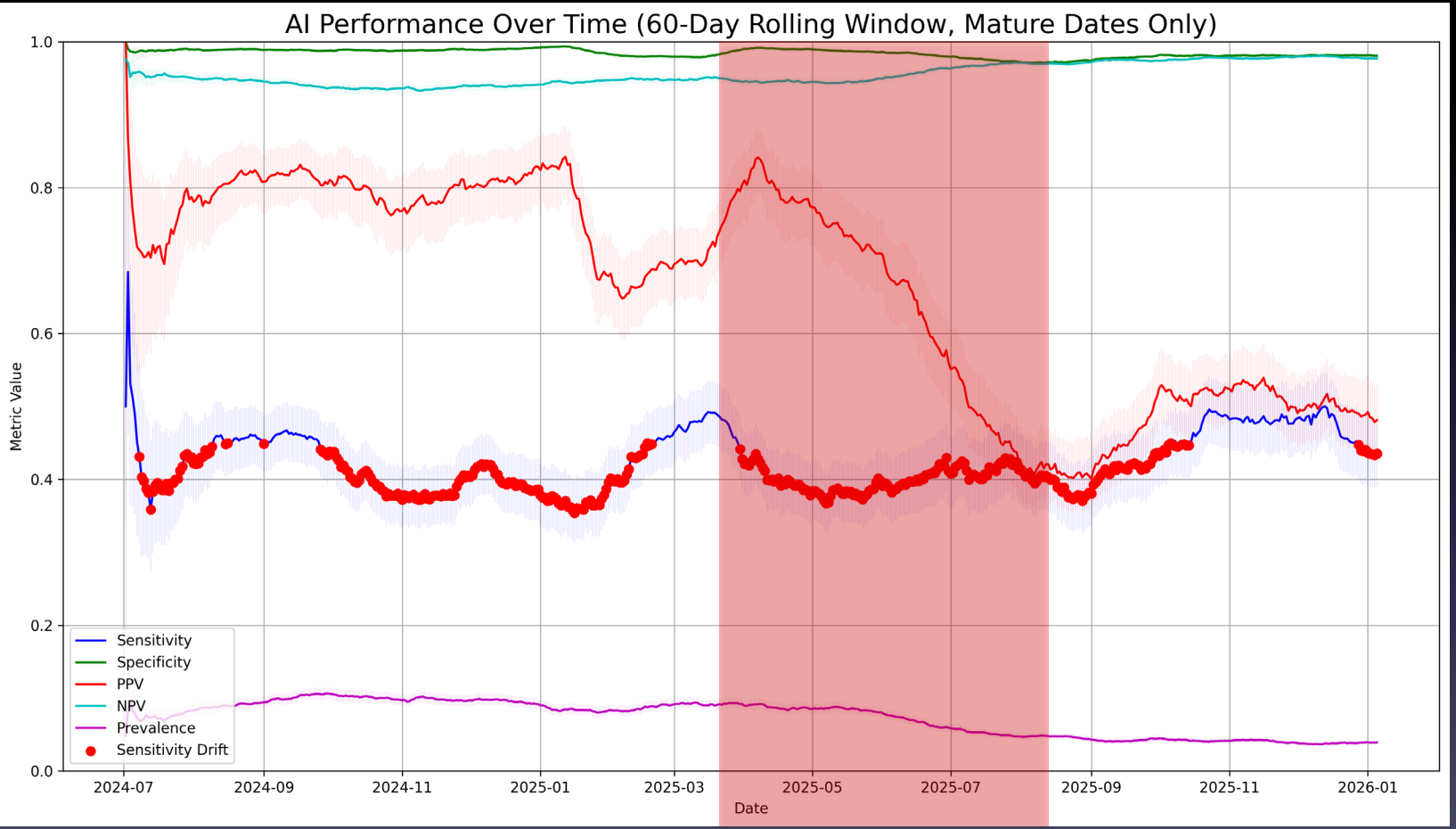








Measuring Temporal Drift in Performance



- Home-built tool (RADAR) for measuring continuous performance of a commercial ICH detection product.
- Performance rapidly dropped after expansion to other hospitals.
- Investigation is pending.

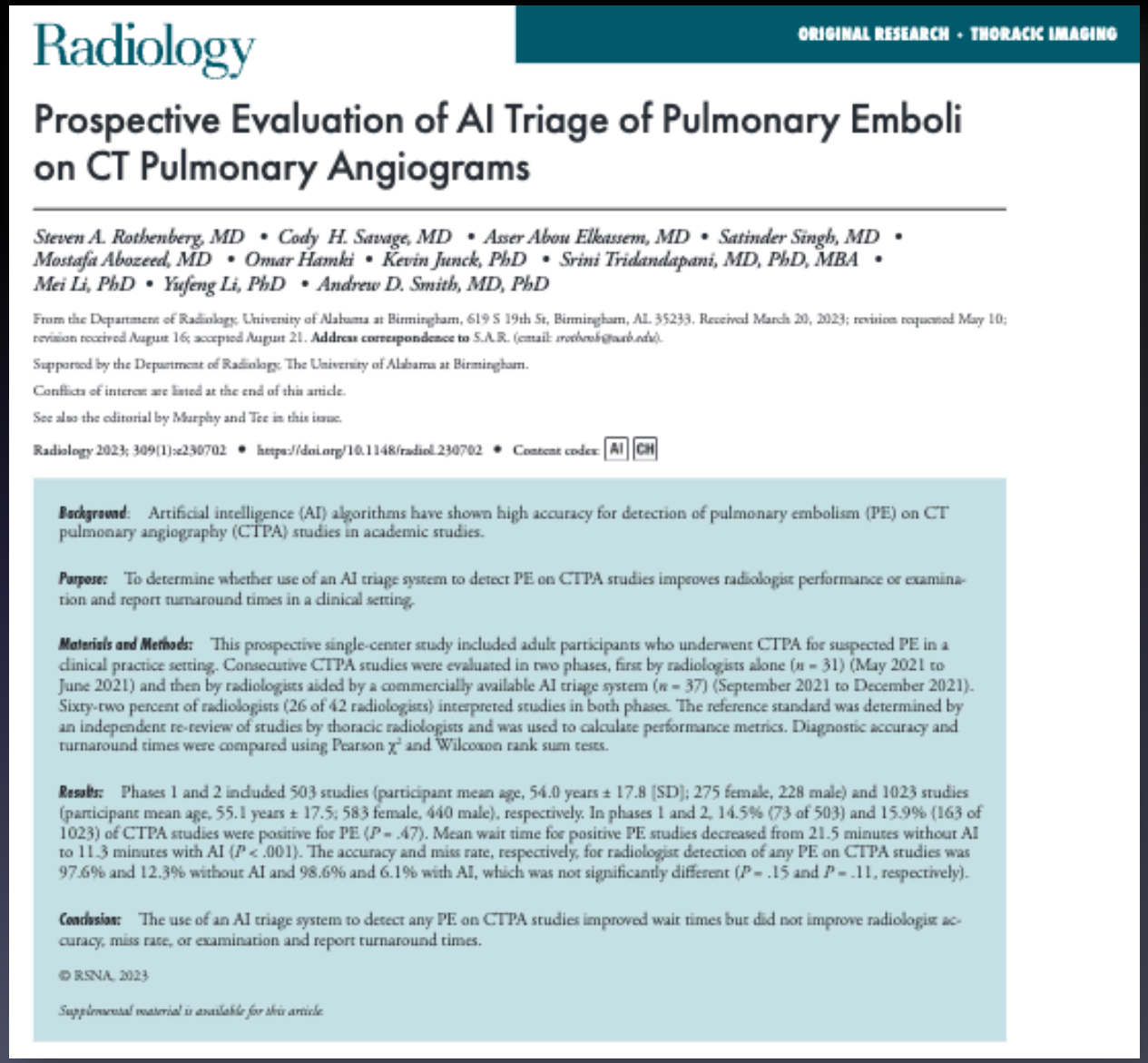
AI Increases Rad Efficiency

	ICH	PE	C-spine Fx	Rib Fx
Yale NewHaven Health — RSNA 2019	READ TAT 10% REDUCTION			
radiology partners — 2020 Pilot	READ TAT 11% REDUCTION	READ TAT 12% REDUCTION		
8 site evaluation > 500,000 exams — 2022 Pilot	READ TAT 9% REDUCTION	READ TAT 9% REDUCTION	READ TAT 8% REDUCTION	READ TAT 15% REDUCTION

Courtesy Nina Kottler, MD MS

PE AI Triage and Notification

- No improvements in radiologist accuracy, sensitivity, or specificity
- No improvements in report TAT or interpretation time
- May be useful with long TAT
- Modest improvements in wait times for positive exams (~5 minutes)



Radiology

ORIGINAL RESEARCH • THORACIC IMAGING

Prospective Evaluation of AI Triage of Pulmonary Emboli on CT Pulmonary Angiograms

Steven A. Rothenberg, MD • Cody H. Savage, MD • Asser Abou Elkassem, MD • Satinder Singh, MD • Mostafa Abozeed, MD • Omar Hamki • Kevin Junck, PhD • Srini Tridandapani, MD, PhD, MBA • Mei Li, PhD • Yufeng Li, PhD • Andrew D. Smith, MD, PhD

From the Department of Radiology, University of Alabama at Birmingham, 619 S 19th St, Birmingham, AL 35233. Received March 20, 2023; revision requested May 10; revision received August 16; accepted August 21. **Address correspondence to** S.A.R. (email: *srothenb@uab.edu*).

Supported by the Department of Radiology, The University of Alabama at Birmingham.

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

See also the editorial by Murphy and Tee in this issue.

Radiology 2023; 309(1):e230702 • https://doi.org/10.1148/radiol.230702 • Content codes: AI CH

Background: Artificial intelligence (AI) algorithms have shown high accuracy for detection of pulmonary embolism (PE) on CT pulmonary angiography (CTPA) studies in academic studies.

Purpose: To determine whether use of an AI triage system to detect PE on CTPA studies improves radiologist performance or examination and report turnaround times in a clinical setting.

Materials and Methods: This prospective single-center study included adult participants who underwent CTPA for suspected PE in a clinical practice setting. Consecutive CTPA studies were evaluated in two phases, first by radiologists alone ($n = 31$) (May 2021 to June 2021) and then by radiologists aided by a commercially available AI triage system ($n = 37$) (September 2021 to December 2021). Sixty-two percent of radiologists (26 of 42 radiologists) interpreted studies in both phases. The reference standard was determined by an independent re-review of studies by thoracic radiologists and was used to calculate performance metrics. Diagnostic accuracy and turnaround times were compared using Pearson χ^2 and Wilcoxon rank sum tests.

Results: Phases 1 and 2 included 503 studies (participant mean age, 54.0 years ± 17.8 [SD]; 275 female, 228 male) and 1023 studies (participant mean age, 55.1 years ± 17.5; 583 female, 440 male), respectively. In phases 1 and 2, 14.5% (73 of 503) and 15.9% (163 of 1023) of CTPA studies were positive for PE ($P = .47$). Mean wait time for positive PE studies decreased from 21.5 minutes without AI to 11.3 minutes with AI ($P < .001$). The accuracy and miss rate, respectively, for radiologist detection of any PE on CTPA studies was 97.6% and 12.3% without AI and 98.6% and 6.1% with AI, which was not significantly different ($P = .15$ and $P = .11$, respectively).

Conclusion: The use of an AI triage system to detect any PE on CTPA studies improved wait times but did not improve radiologist accuracy, miss rate, or examination and report turnaround times.

Supplemental material is available for this article.

Rothenberg et al 2023 - Radiology

Real-World Performance

- 82.2% sensitivity, 97.6% specificity, and 96.6% accuracy.
- Sensitivity was highest for acute (86.2%), large >10 mm (95.0%), and multi-compartment hemorrhages (93.6%),
- Lower for subacute (45.5%), chronic (54.8%), small ≤10 mm (74.8%), and single bleeds (76.0%).
- Performance reduced for OP (72.2%), where subtle hemorrhages more common.

npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02244-3

Real-world performance evaluation of a commercial deep learning model for intra

Findings show that the model performs reliably for acute and extensive ICH but is less sensitive to subtle or localized presentations, underscoring the need for ongoing real-world evaluation and targeted improvements to support safe clinical triage.

Mohammadre
Theodorus Da
Chad Robicha

r updates

Intracranial he
yet the real-w
retrospective
non-contrast
2023–April 20
radiology rep
manually annotated cases. The LLM achieved 90% accuracy (κ = 0.98) for ICH reclassification. Overall, the Aidoc model demonstrated 82.2% sensitivity, 97.6% specificity, and 96.6% accuracy. Sensitivity was highest for acute (86.2%), large >10 mm (95.0%), and multi-compartment hemorrhages (93.6%), but substantially lower for subacute (45.5%), chronic (54.8%), small ≤10 mm (74.8%), and single-compartment bleeds (76.0%). Performance was also reduced in the outpatient setting (72.2%), where subtle hemorrhages were more common, while remaining consistent across demographic subgroups. These findings show that the model performs reliably for acute and extensive ICH but is less sensitive to subtle or localized presentations, underscoring the need for ongoing real-world evaluation and targeted improvements to support safe clinical triage.

Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Economic Value of AI in Radiology: A Systematic Review

Isabel Molwitz, MD, MBA[1] • Inka Ristow, MD[1] • Jennifer Erley, MD[1] • Tugba Akinci D'Antonoli, MD[2,8,9] • Ali S. Tejani, MD[3] • Michail E. Klontzas, MD, PhD[4] • Merel Huisman, MD, PhD[5] • Gerhard Adam, MD[1] • Stephan Nüesch, MA[6] • Lisa Adams, MD[7]

Author affiliations, funding, and conflicts of interest are listed at the end of this article.

See also the commentary by Amindarolzarbi and Siegel in this issue.

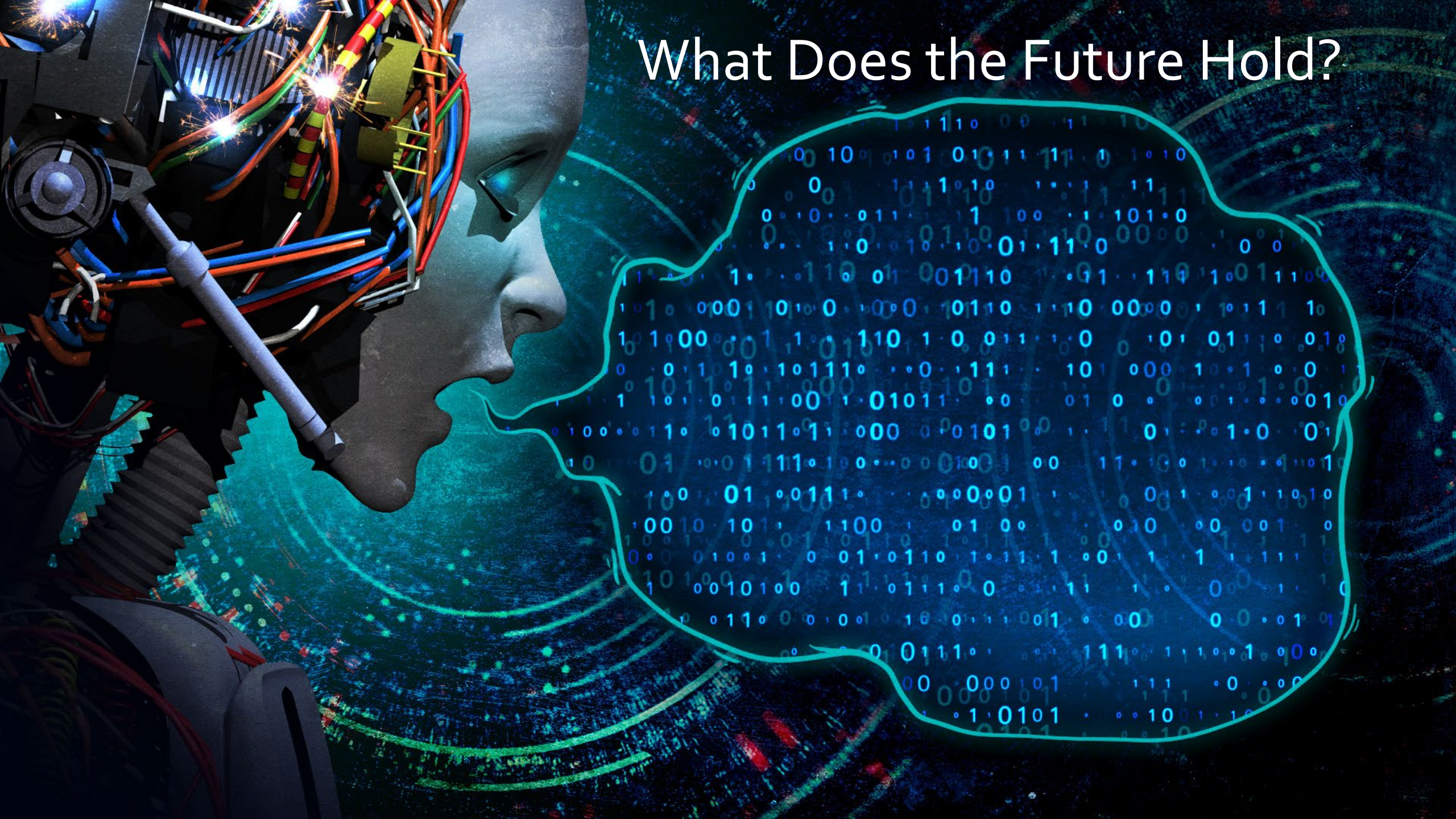
Radiology: Artificial Intelligence 2026; 8(1):e250090 • https://doi.org/10.1148/ryai.250090 • Content code: AI

- Many ROI studies (1,879) yet only a few (21) showed rigor.
- Cost savings or cost-effectiveness ratios in resource intensive tasks when accuracy matched human performance.
- AI showed value in settings where there are radiologist shortages.
- AI showed reduced costs by optimizing protocols
- AI showed increased revenue when follow up compliance was improved.

Rethinking the Last Mile Around AI

- Diagnostic cockpits have become complex
- Integration and user engagement remain challenging in current environments

What Does the Future Hold?



nature communications

Article

https://doi.org/10.1038/s41467-023-39631-x

Opportunistic detection of type 2 diabetes using deep learning from frontal chest radiographs

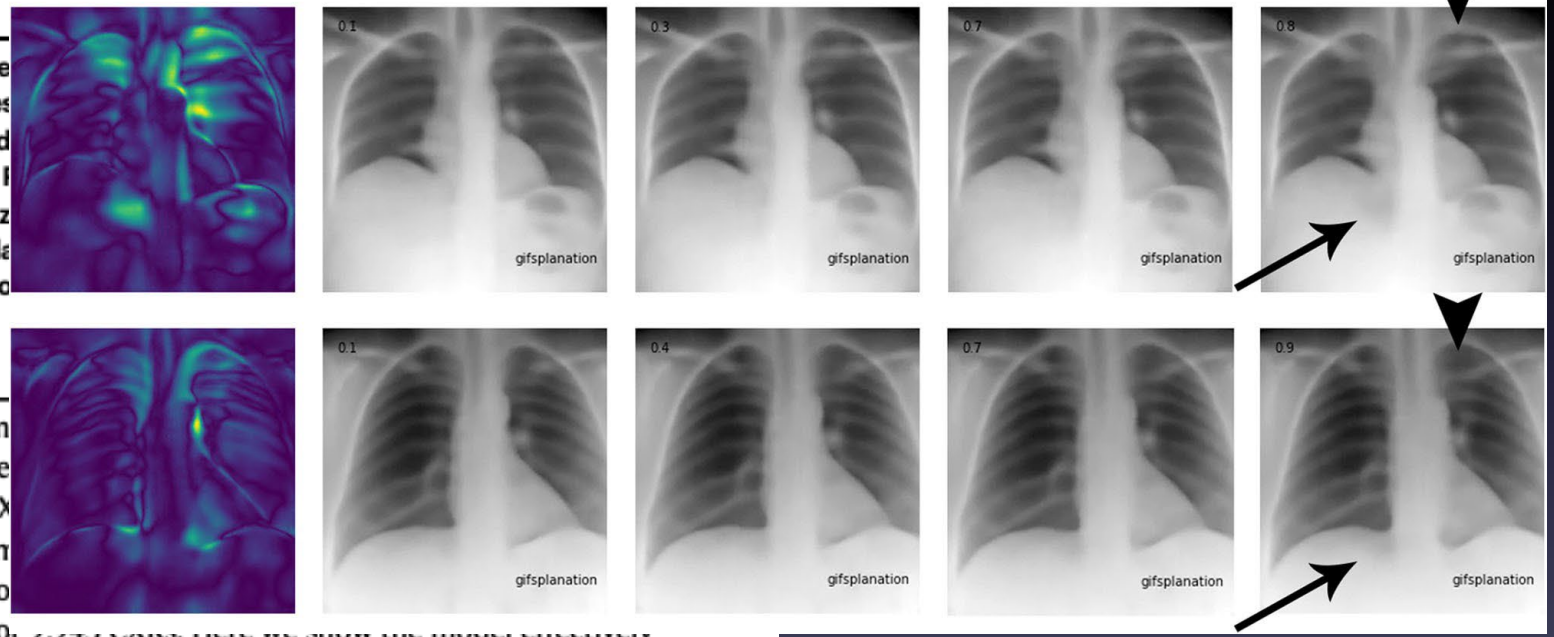
Received: 14 December 2022

Accepted: 19 June 2023

Published online: 07 July 2023

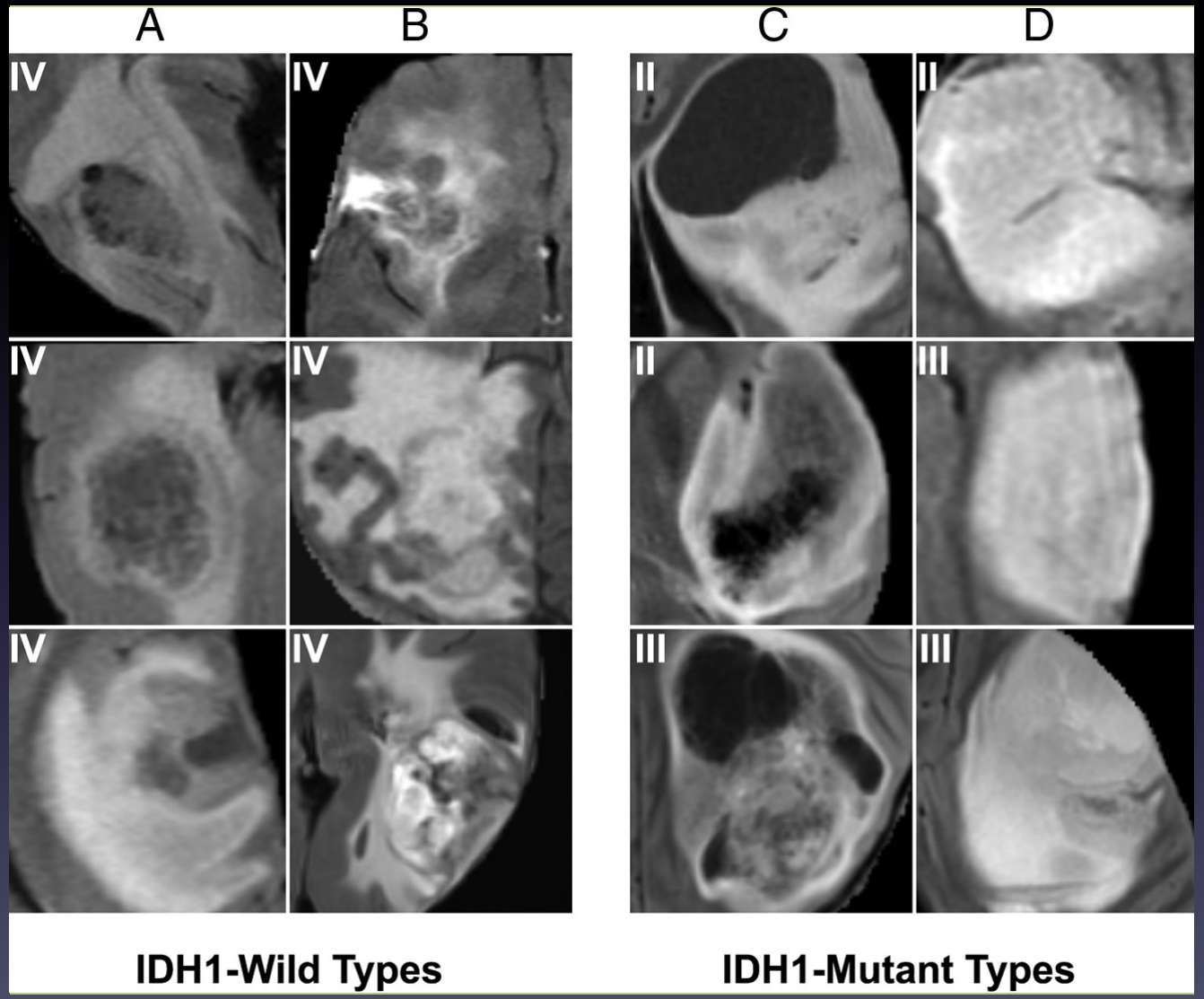
Check for updates

Ayis Pyrros[1,2,20] ✉, Ste
Zachary Zaiman[5], Kaes
Nasir Siddiqui[1], Melind
Jeanne M. Horowitz[9], I
Jorge Mario Rodríguez
Sanmi Koyejo[14], Ada
John W. Garrett[18], Jo
William Galanter[19]



Deep learning (DL) m
predict diseases and e
chest radiographs (CX
diabetes (T2D) by com
model, developed fro
prospective dataset o
detected T2D with a ROC AUC of 0.84 and a 16% prevalence. The algorithm flagged 1,381 cases (14%) as suspicious for T2D. External validation at a distinct institution yielded a ROC AUC of 0.77, with 5% of patients subsequently diagnosed with T2D. Explainable AI techniques revealed correlations between specific adiposity measures and high predictivity, suggesting CXRs' potential for enhanced T2D screening.

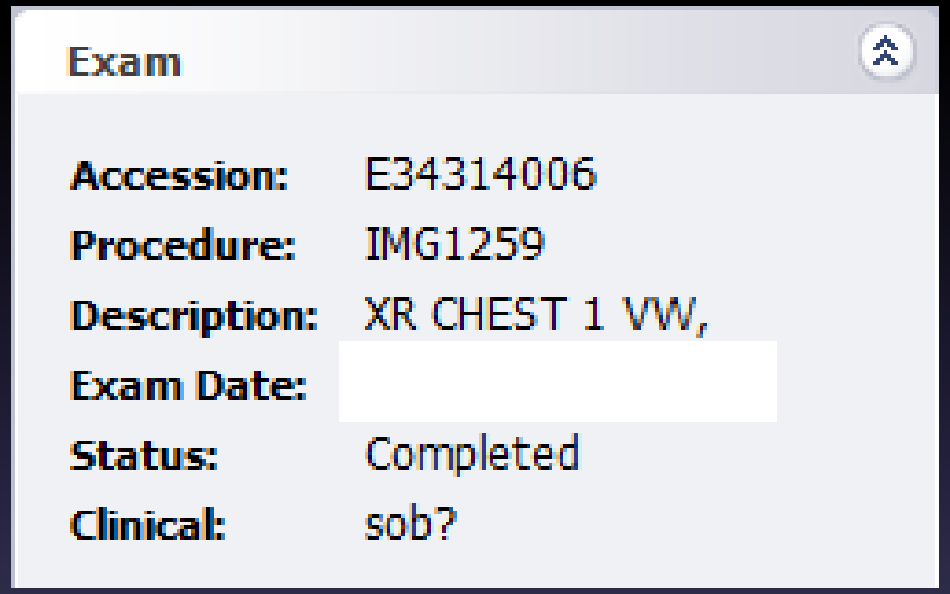
Deep-Learning Convolutional Neural Networks Accurately Classify Genetic Mutations in Gliomas

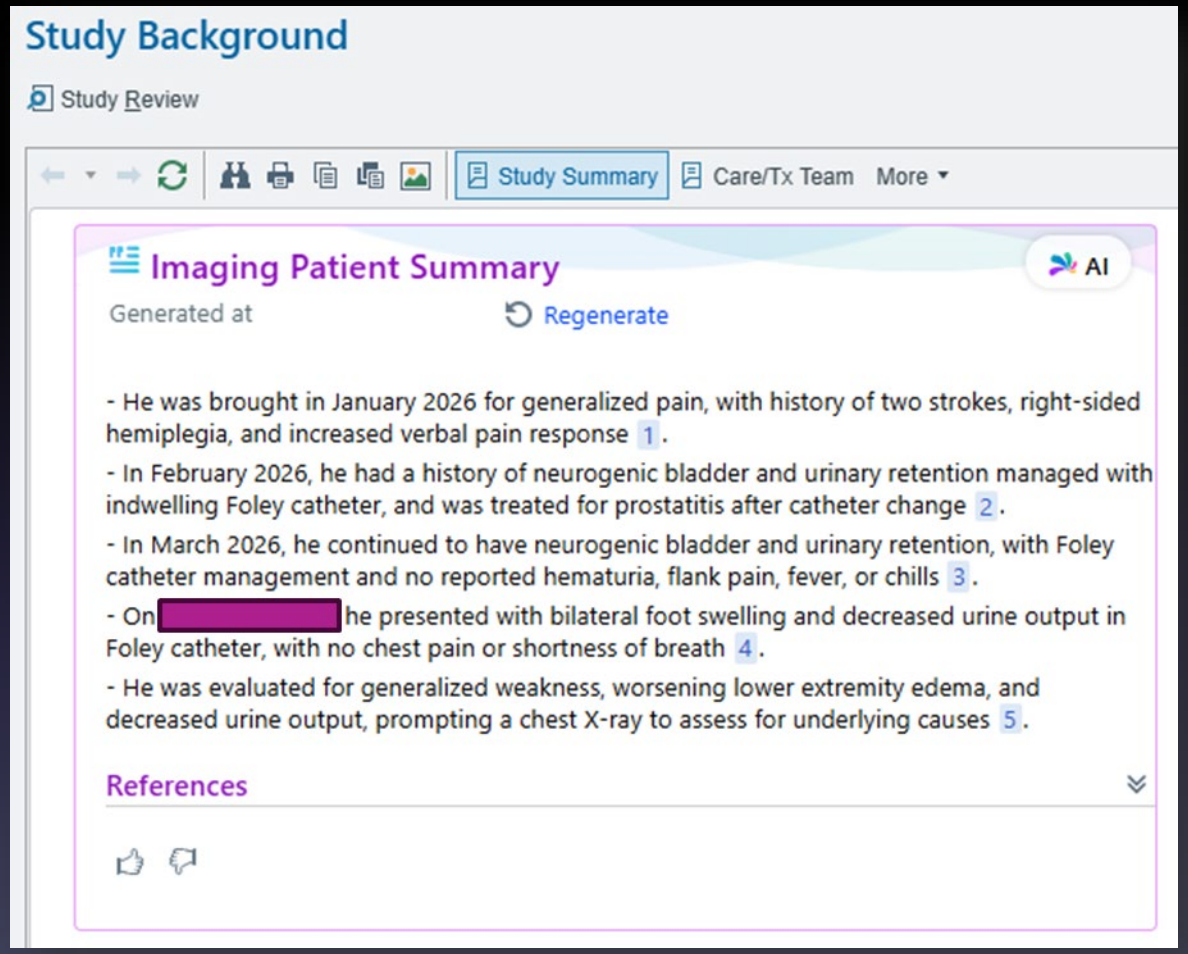


Chang P etal. AJNR 2018 May


LLM

EMR Data Abstraction Tools



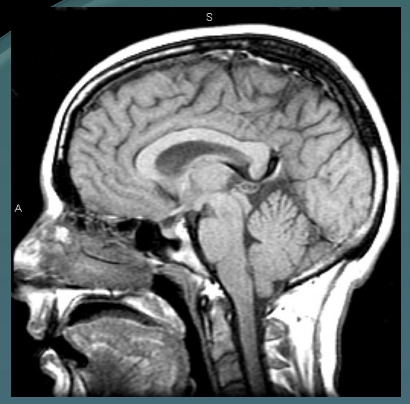


Courtesy Steven Rothenberg MD

LLMs Facilitate Reporting Efficiencies

- ✓ Improve readability – grammer
- ✓ Report correction (gender / laterality)
- ✓ Auto-generated impressions (custom)
- ✓ Automatic report structuring / organization
- ✓ Automatic concept extraction (registry / data collection)
- ✓ Completion checking (current / past finding)
- ✓ Longitudinal history checking (current against past)
- ✓ Actionable finding detection (notification)
- ✓ Translation to other languages
- ✓ Translation to match education level
- ✓ Interactive reports
- ✓ Auto-generated report drafts

VLM Contrastive Learning with Text & Computer Vision



Images

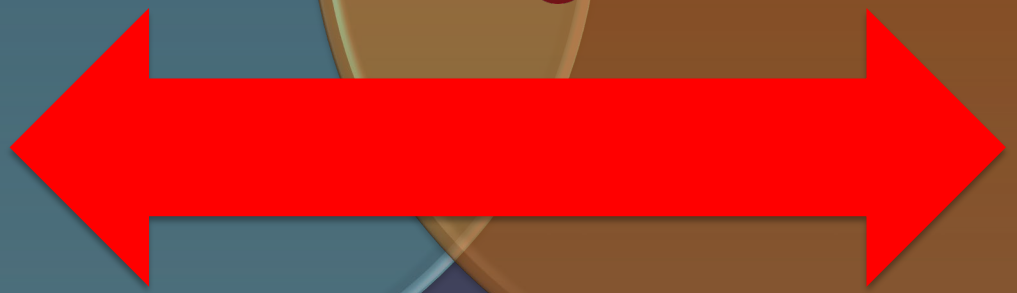
DEPARTMENT OF RADIOLOGY: MRI REPORT
Patient: DOE, JANE
DOB: 01/01/1985
Exam Date: April 5, 2026
Accession #: 8675309
Referring Physician: Dr. Aris Totle

EXAMINATION
MRI BRAIN WITHOUT AND WITH IV CONTRAST
CLINICAL INDICATION
Chronic tension-type headaches and occasional episodes of localized vertigo. No known history of malignancy.
TECHNIQUE
Multiplanar, multisequence magnetic resonance imaging was performed through the brain before and after the administration of 15 mL of Gadavist intravenous contrast.
Sequences obtained: T1, T2, FLAIR, DWI/ADC, and post-contrast T1-weighted images.

FINDINGS
Brain Parenchyma: There is no evidence of acute intracranial hemorrhage, large territorial infarct, or suspicious mass effect. The gray-white matter differentiation is preserved.
White Matter: Scattered, punctate T2/FLAIR hyperintensities are noted within the subcortical white matter of the frontal lobes. These are non-specific but are most commonly associated with chronic microvascular changes or migraines in a patient of this age.
Ventricular System: The ventricles and sulci are prominent for the patient's age but remain within the upper limits of normal. There is no evidence of hydrocephalus.
Extra-axial Spaces: No abnormal extra-axial fluid collections are identified.
Post-Contrast: No areas of abnormal enhancement are seen. Specifically, there is no evidence of internal auditory canal (IAC) enhancement to explain the patient's vertigo.
Incidental Finding: A 5 mm benign-appearing pineal cyst is noted, which is a common incidental finding with no clinical significance.

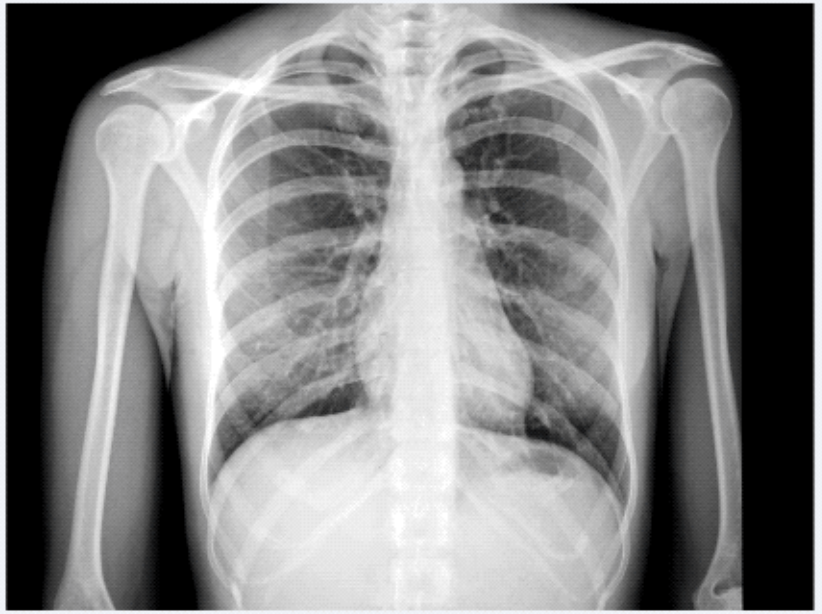
IMPRESSION
Unremarkable MRI of the brain with and without contrast.
Small, non-specific white matter hyperintensities, likely representing mild chronic microvascular changes.
Incidental 5 mm pineal cyst, stable in appearance.
Radiologist: *Dr. A. I. Collaborator, MD Board Certified Radiologist*

Report Text



Automated Report Drafting is the New Reality

Can you write me a report analyzing this chest X-ray?



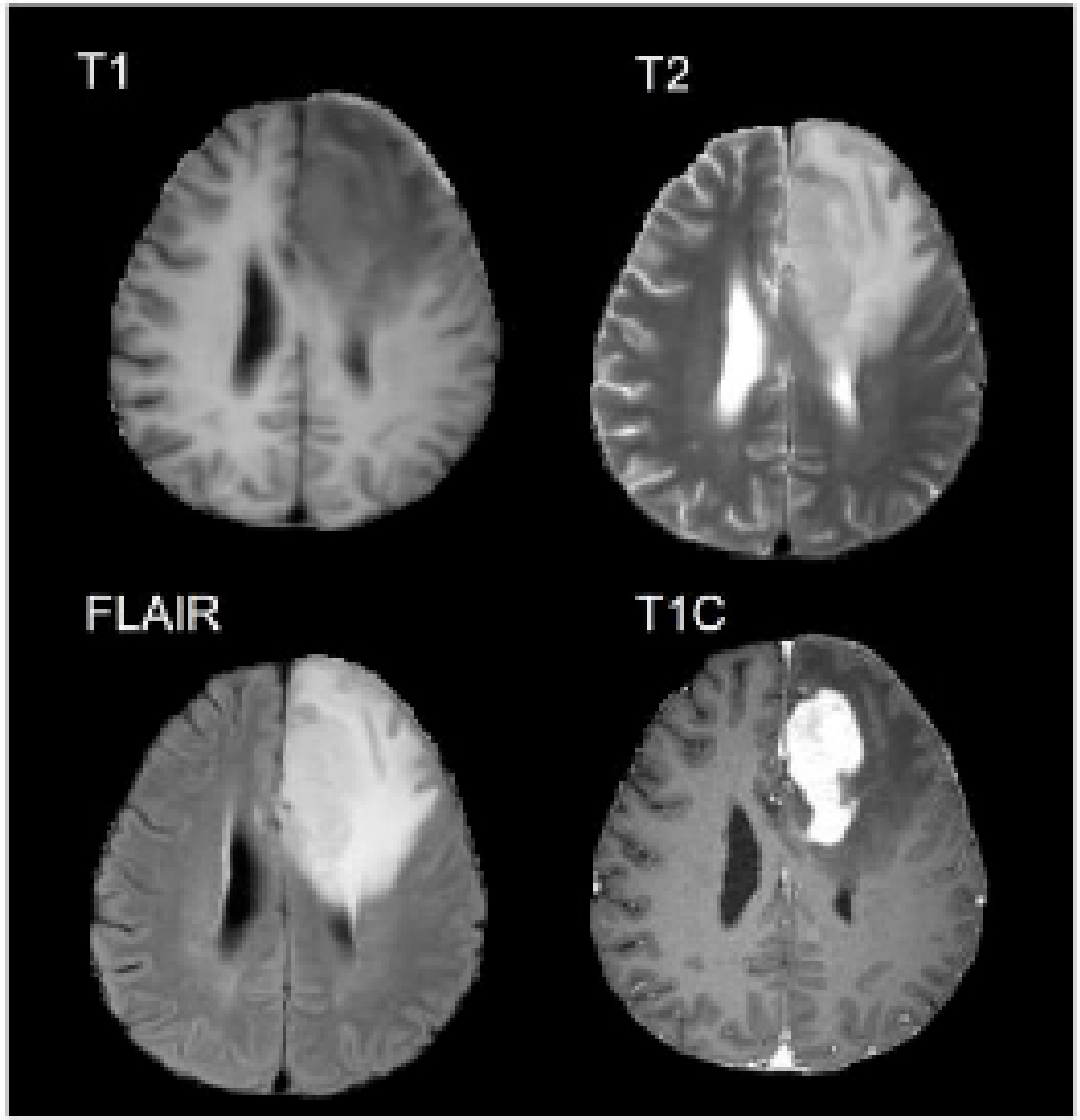
Findings:

- Devices: None.
- Lungs: No pneumothorax. No substantial pleural effusion. Lungs appear clear.
- Cardiomediastinal: Normal heart size. Mediastinal contours within normal limits.
- Other: No acute skeletal abnormality.

Impression:
No active disease seen in chest.

Enter a question here

Automated Brain Tumor MR Report Drafting



Original Report

A patient presented with a homogeneously enhancing mass located at the left frontal lobe of the brain, which is accompanied by perilesional edema, and there is no internal hemorrhage within the mass.

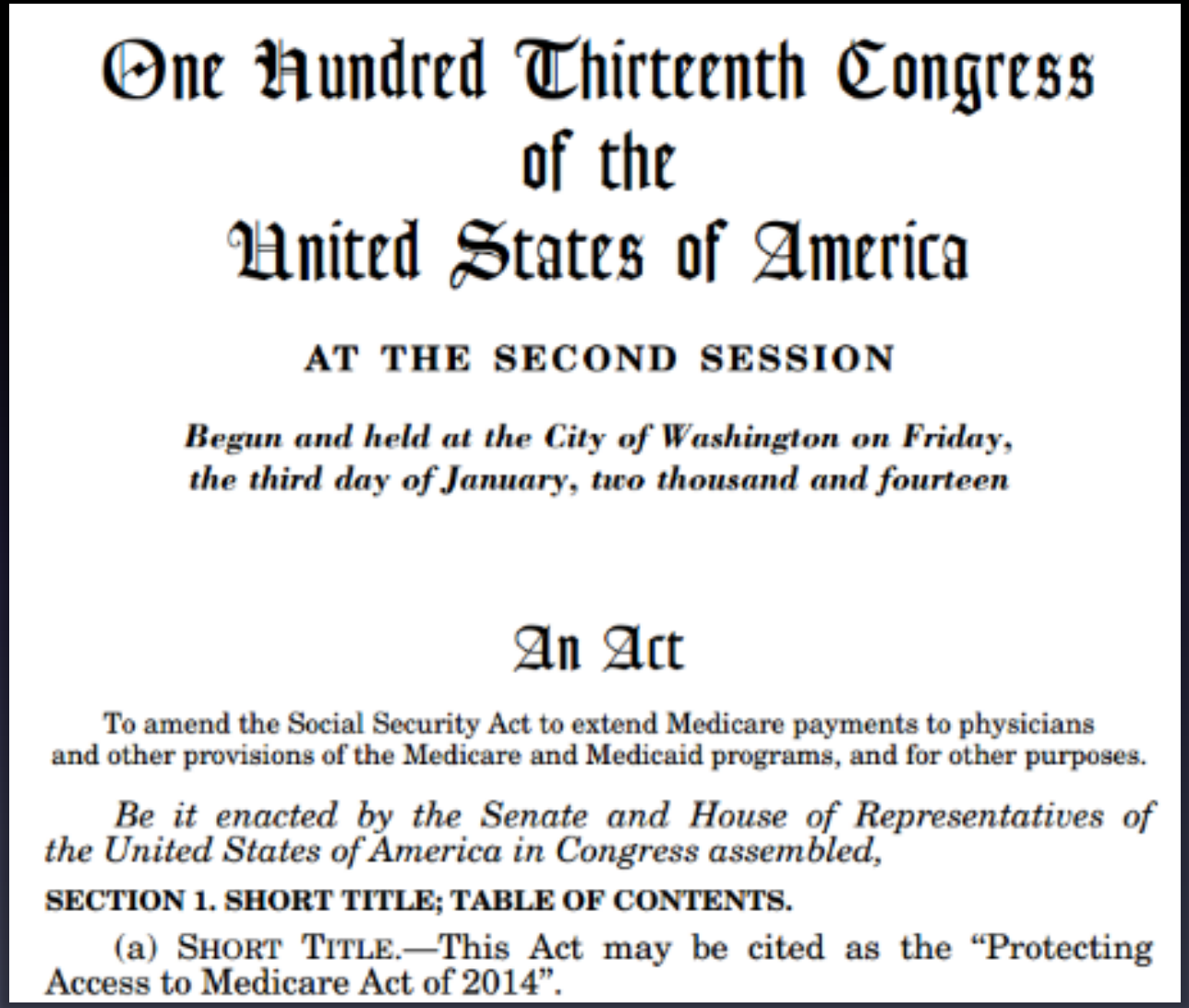
RRG

A patient has a 5.5-cm enhancing mass located at the left frontal lobe. The mass is accompanied by marked peritumoral edema. There is also a 1.5-cm enhancing nodule located at the right frontal lobe.

npj Digit. Med. (2026).https://doi.org/10.1038/s41746-026-02581-x

PAMA Legislation - CDS

- The Protecting Access to Medicare Act (PAMA) of 2014, Section 218(b), amended Section 1834(q) of the Act to establish the AUC program to increase the use of AUC for advanced diagnostic imaging services provided to Medicare patients.
- Now "on hold" for administrative and logistical reasons.
- Never proven to reduce overutilization of expensive imaging services.
- AI Opportunity - Relook at this with AI.



One Hundred Thirteenth Congress
of the
United States of America

AT THE SECOND SESSION

Begun and held at the City of Washington on Friday, the third day of January, two thousand and fourteen

An Act

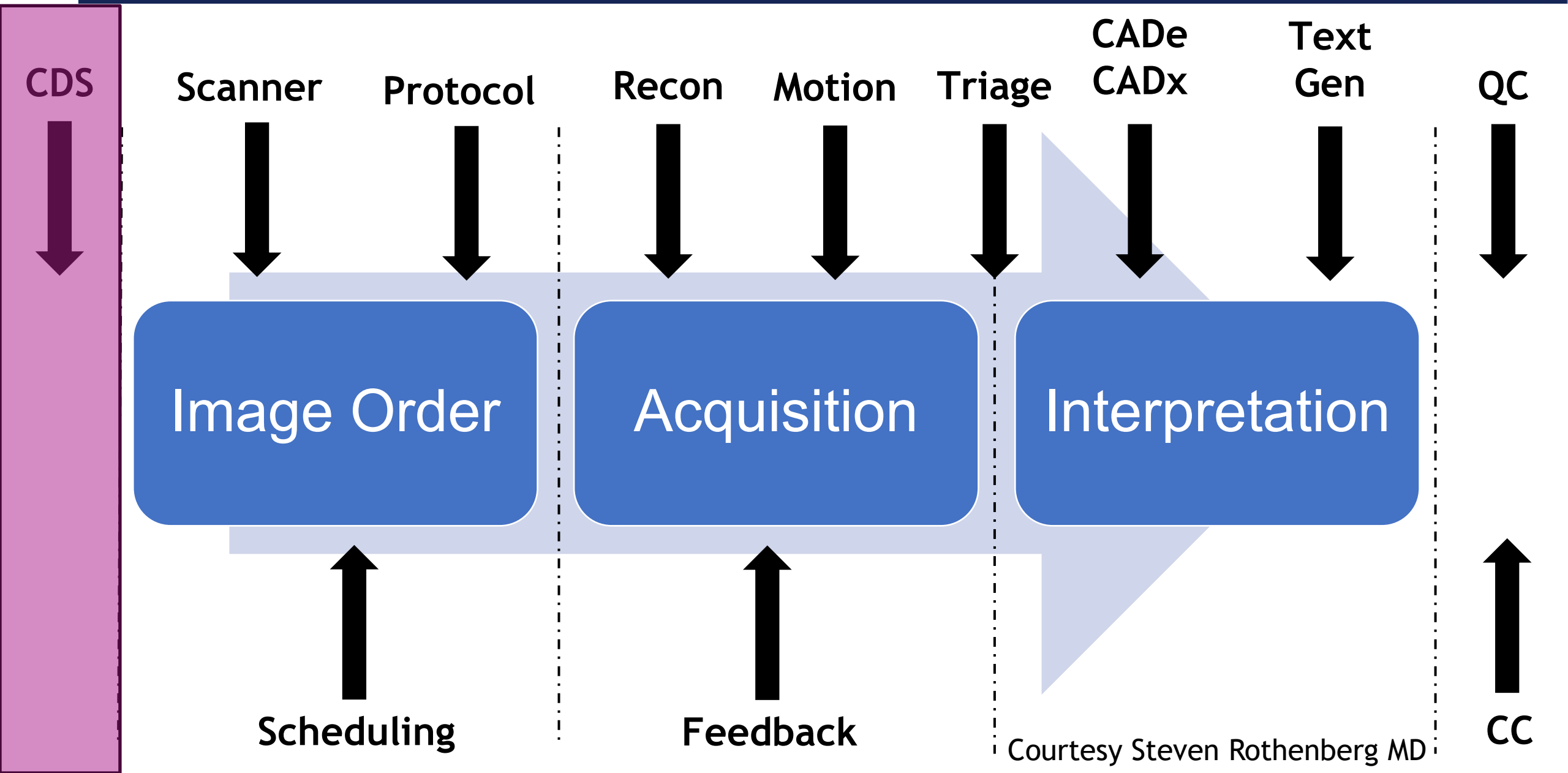
To amend the Social Security Act to extend Medicare payments to physicians and other provisions of the Medicare and Medicaid programs, and for other purposes.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE; TABLE OF CONTENTS.

(a) SHORT TITLE.—This Act may be cited as the "Protecting Access to Medicare Act of 2014".

Workflow: Lifecycle of an Imaging Exam



Exam: MRI Lumbar Spine without Contrast
Clinical Indication: Low back pain with bilateral lower extrem

Technique:
Multiplanar, multis
contrast.

Findings

Alignment:
Mild straightening

Vertebral Bodies:
Vertebral body hei
most pronounced
fracture or suspicio

Conus Medullaris:
Terminates at L1 a

Paraspinal Soft Tis
Unremarkable.

Level-by-Level Evaluation

T12–L1: No significant disc bu

L1–L2: Mild disc desiccation
or foraminal stenosis.

L2–L3: Disc desiccation with mil
bilateral facet arthropathy and li
narrowing. Mild bilateral forami

L3–L4: Moderate disc height loss with broad-based disc bulge and superimposed small

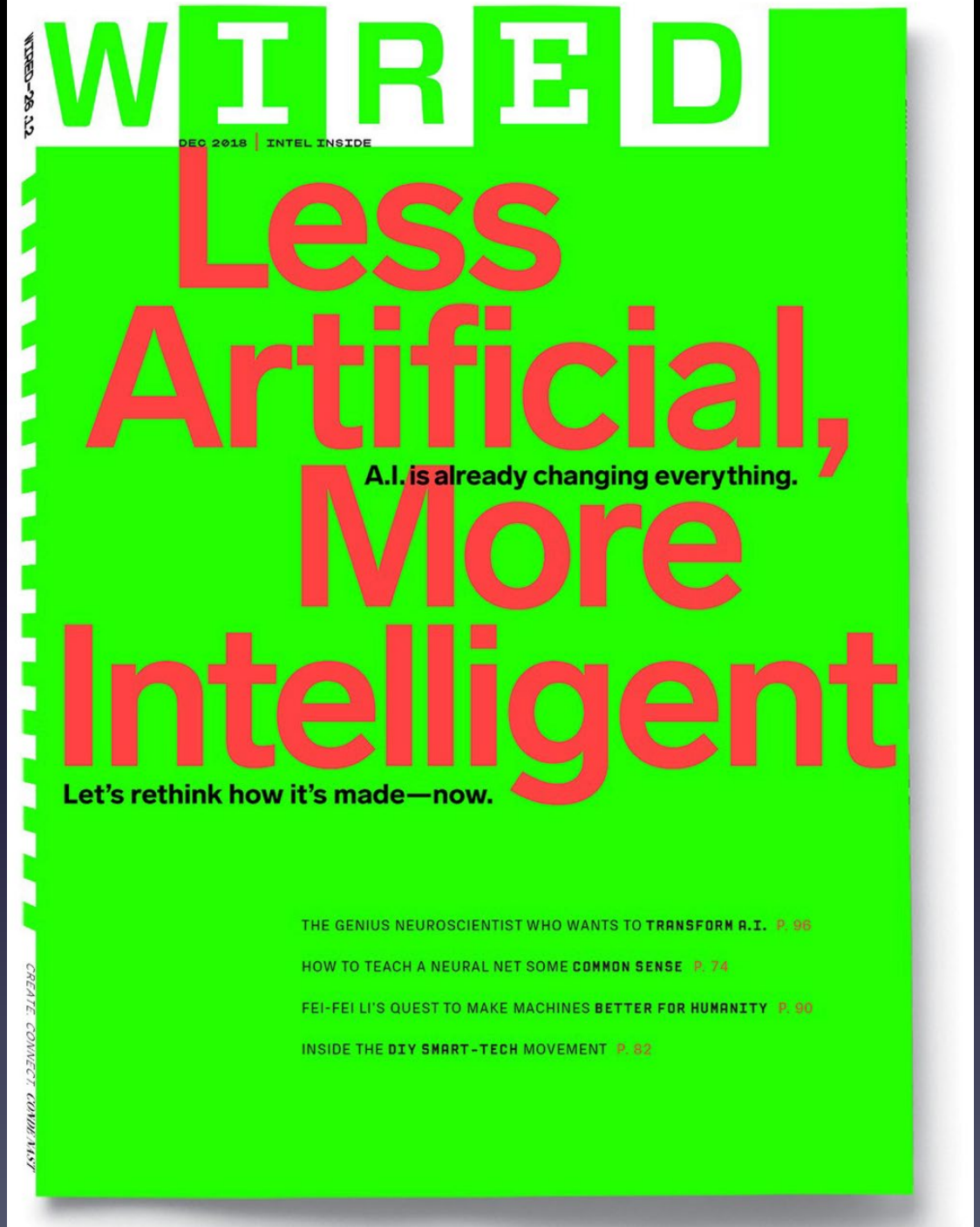
Disc Level Radiomic Features

Feature Category	Metric	Value	Norm Range	Interpretation
First-order	Mean T2 intensity	↓ 1.8 SD	±1 SD	Reduced hydration
Texture (GLCM entropy)	4.92			Increased structural heterogeneity
Textu				
short-				
emph				
Shape				

Sarcopenia Measures

Metric	Value	Age/Sex Norm	Interpretation
Total Paraspinal CSA	18.4 cm²	26.1 ± 3.2	↓ 2.4 SD
Multifidus CSA	7.1 cm²	11.2 ± 1.5	Markedly reduced
Erector Spinae CSA	11.3 cm²	14.9 ± 2.0	Reduced
Muscle Fat Fraction	32%	<15%	Elevated
Muscle T2 Heterogeneity (entropy)	↑ 2.1 SD	—	Structural degeneration

L3–L4	82 mm²	12th %ile	Moderate
L4–L5	58 mm²	4th %ile	Severe


WIRED
DEC 2018 | INTEL INSIDE
Less Artificial, More Intelligent
A.I. is already changing everything.
Let's rethink how it's made—now.
THE GENIUS NEUROSCIENTIST WHO WANTS TO TRANSFORM A.I. P. 96
HOW TO TEACH A NEURAL NET SOME COMMON SENSE P. 74
FEI-FEI LI'S QUEST TO MAKE MACHINES BETTER FOR HUMANITY P. 90
INSIDE THE DIY SMART-TECH MOVEMENT P. 82
CREATE. CONNECT. COMMUNITY.

THE HILL

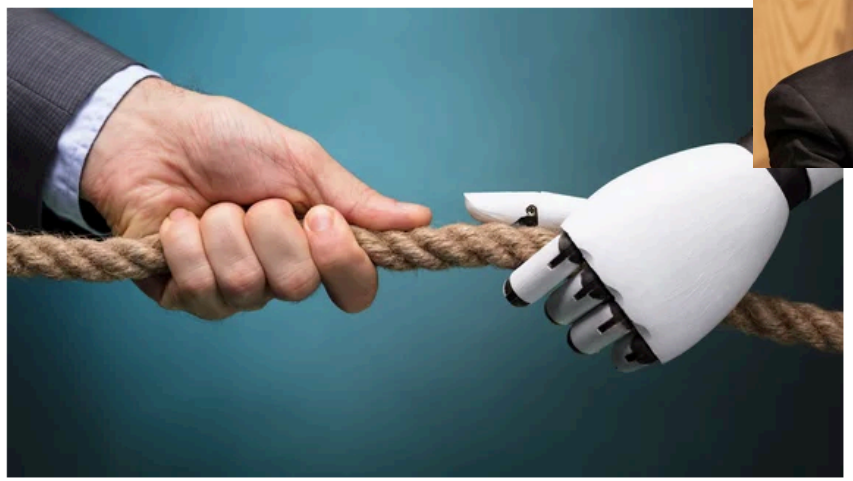
Artificial intelligence needs guardrails

BY EUGENE SANTOS JR., OPINION CONTRIBUTOR — 06/03/19 06:30 PM EDT

THE VIEWS EXPRESSED BY CONTRIBUTORS ARE THEIR OWN AND NOT THE VIEW OF THE HILL

CEO of America's largest public hospital system says he's ready to replace radiologists with AI

Marty Stempniak | *March 31, 2026* | *Radiology Business* | *Artifici*



The chief executive of America's largest public hospital system says he is prepared to sta
reg

Mi
du
me
ma

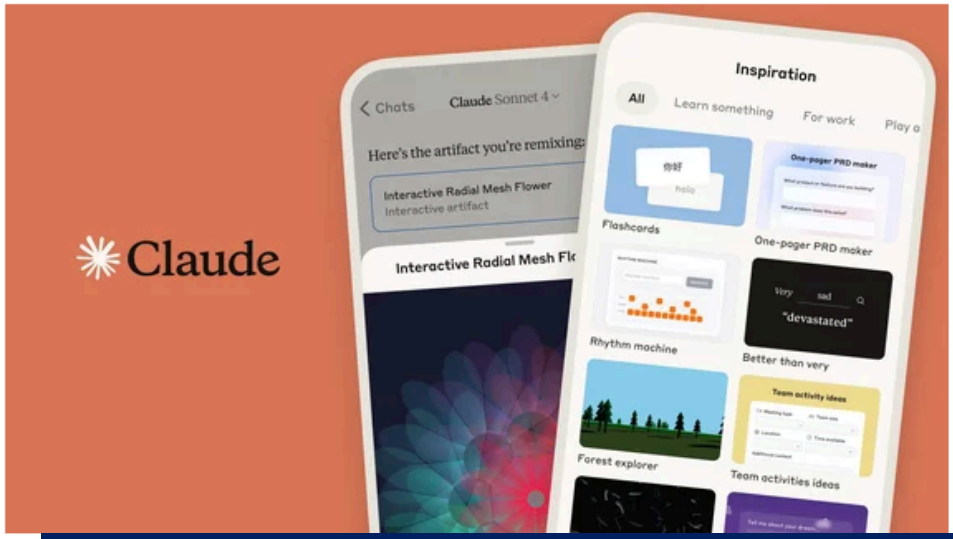
Th
wh
Th

"We could replace a great deal of radiologists with AI at this moment, if we are ready to do the regulatory challenge"

"We could replace a great deal of radiologists with AI at this moment, if we are ready to do the regulatory challenge," Katz said at the forum, held on March 25.

Radiologists criticize Anthropic CEO's recent comments about the specialty

Marty Stempniak | *March 09, 2026* | *Radiology Business* | *Artificial Intelligence*



So

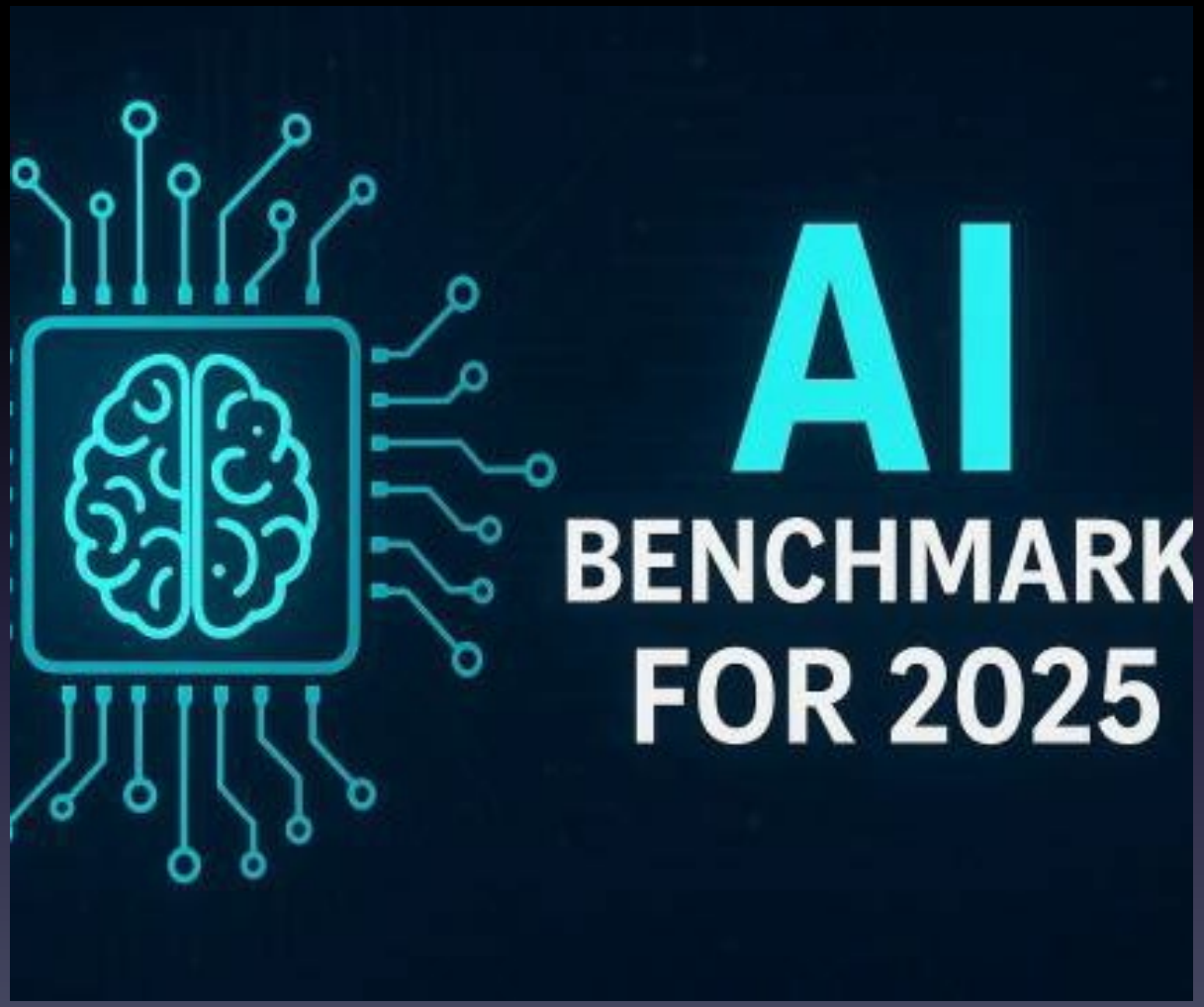
Ra
ab

Da
lar
po

"There's this story of, like—I think it was Geoff Hinton—predicting that AI will replace radiologists. And indeed, AI has gotten better than radiologists at, you know, doing scans, right?"

hourlong conversation focused on a range of topics under the umbrella of "The AI Tsunami is Here."

AI Benchmarking

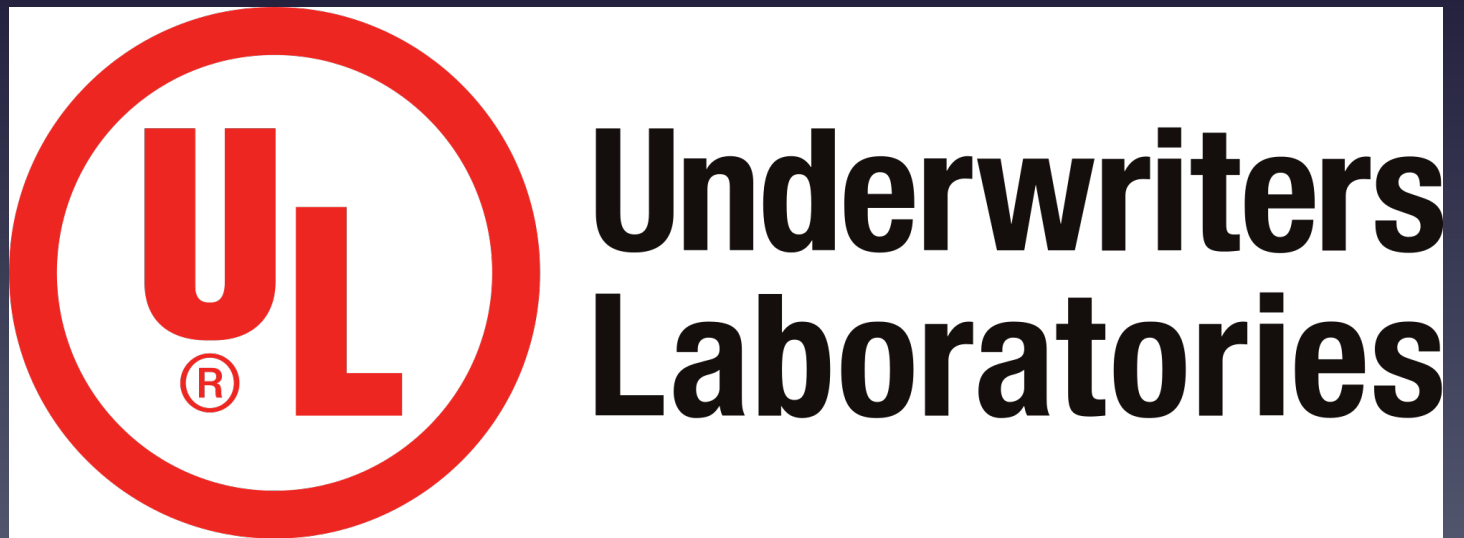


- AI benchmarking is an overloaded term – different meaning.
- Domain experts should establish quality thresholds to compare AI tools performance for specific tasks.
- RSNA & ACR have been working towards developing AI imaging benchmark datasets and performance registries.

RSNA® Radiological Society of North America

American College of Radiology™

Independent Testing



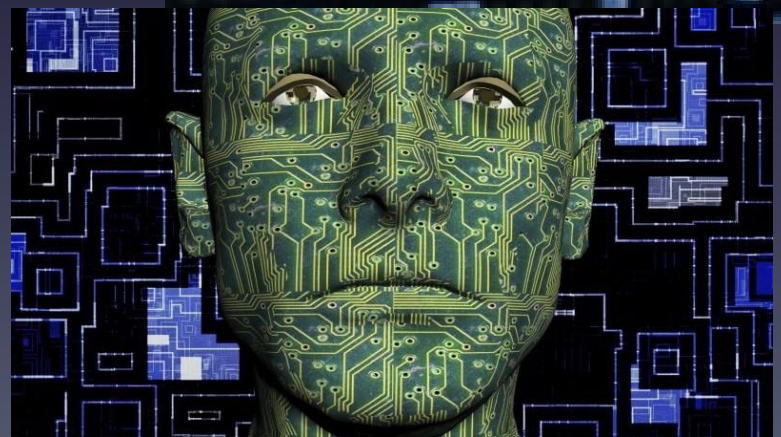
Photoshop -> change the background to white

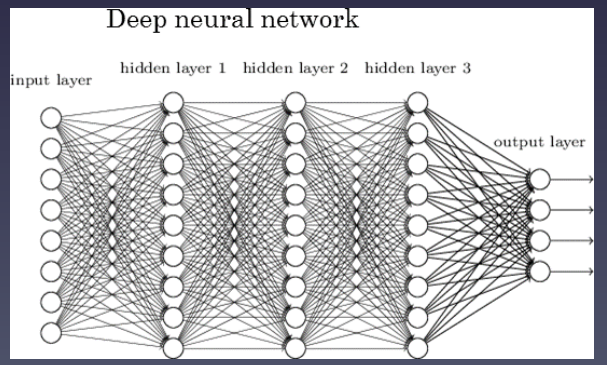
Courtesy John Eng, MD

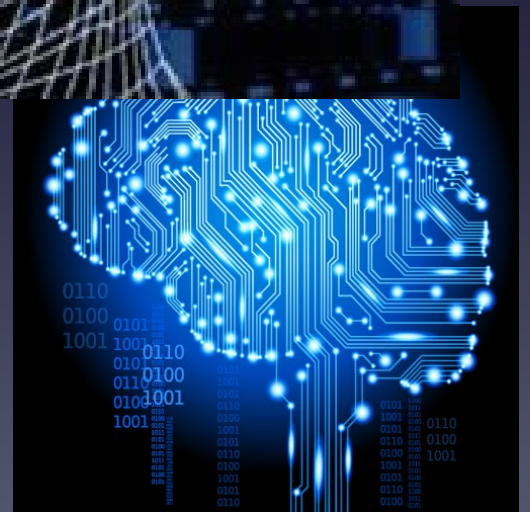
Summary

- Dependency on medical imaging continues to rise at an alarming rate.
- AI is looked upon to help with the demand & short resources.
- Health of the population is a high stakes proposition
- AI provides a set of tools that may improve efficiency and accuracy in healthcare – autonomous use without supervision should be approached with caution.
- Continuous monitoring of tools is paramount to ensure patient safety.

THANK YOU




Deep neural network
input layer
hidden layer 1
hidden layer 2
hidden layer 3
output layer



Benefits and Risks In AI

For Clinical Imaging and Diagnosis

Peter McCaffrey, MD, MS, FCAP
VP, Chief Digital and AI Officer
Director, UTMB AI Center
University of Texas Medical Branch

*Old Red AD 2123 courtesy of MidJourney v5

Learning Objectives

- Explore benefits, risks, and evidence of AI interpretation of diagnostic information such as imaging.
- Consider indicators used by the medical community to identify or suggest errors or misinterpretations in the output of AI systems or software.
- How AI helps, or could help, guide medical decision-making regarding the need for additional diagnostic or laboratory testing.

The Promise of AI in Diagnostics

- Alleviate Burnout and Manage Volume.
- Enhance What Can Be Seen.
- Expand the Breadth of Use for Key Tools

AI creates a richer, faster medical record but those who adjudicate over it must understand how that information was generated

Opportunistic Screening

Existing Diagnostic Methods on Classical Modalities

Simplifying obstetric sonography with AI

Researchers developed an AI-enabled, battery-operated tool that can be operated by clinicians with no sonography experience – and that measures gestational age as accurately as high-specification ultrasound.

JOURNAL ARTICLE

Hypertrophic cardiomyopathy detection with artificial intelligence electrocardiography in international cohorts: an external validation study

Circulation
Volume 147, Issue 9, 28 February 2023; Pages 703-714
https://doi.org/10.1161/CIRCULATIONAHA.122.062746

American Heart Association.

ORIGINAL RESEARCH ARTICLE

Incidental Coronary Artery Calcium: Opportunistic Screening of Previous Nongated Chest Computed Tomography Scans to Improve Statin Rates (NOTIFY-1 Project)

New Diagnostic Methods on Classical Modalities

PLOS Digit Health. 2024 Dec 23;3(12):e0000698. doi: 10.1371/journal.pdig.0000698

Measurement of breast artery calcification using an artificial intelligence detection model and its association with major adverse cardiovascular events

Technol Cancer Res Treat. 2021 Nov 20;20:15330338211058352. doi: 10.1177/15330338211058352

Colorectal Cancer Detected by Machine Learning Models Using Conventional Laboratory Test Data

nature > nature communications > articles > article

Article | Open access | Published: 07 July 2023

Opportunistic detection of type 2 diabetes using deep learning from frontal chest radiographs

New Diagnostic Methods on New Modalities

nature > scientific reports > articles > article

Article | Open access | Published: 03 June 2024

Depression recognition using voice-based pre-training model

AI speech analysis predicted progression of cognitive impairment to Alzheimer's with over 78% accuracy

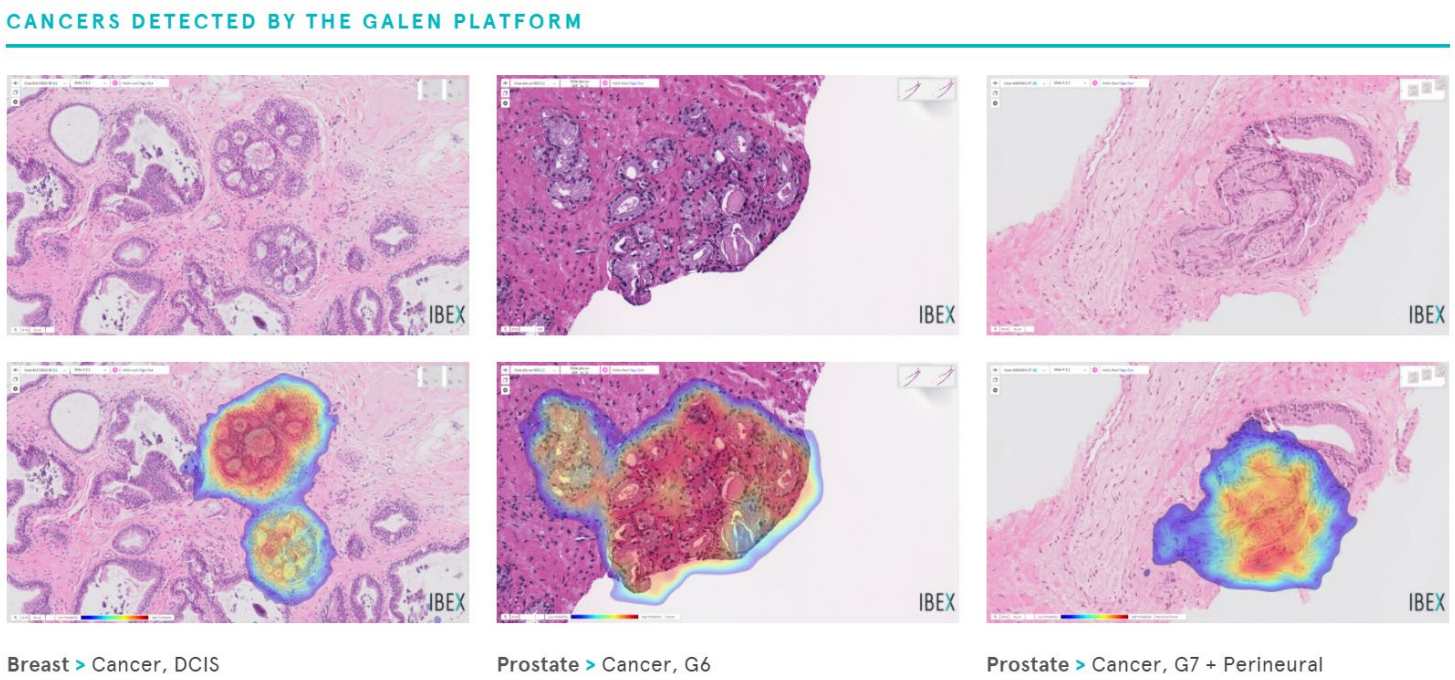
RESEARCH HIGHLIGHTS
January 2, 2025

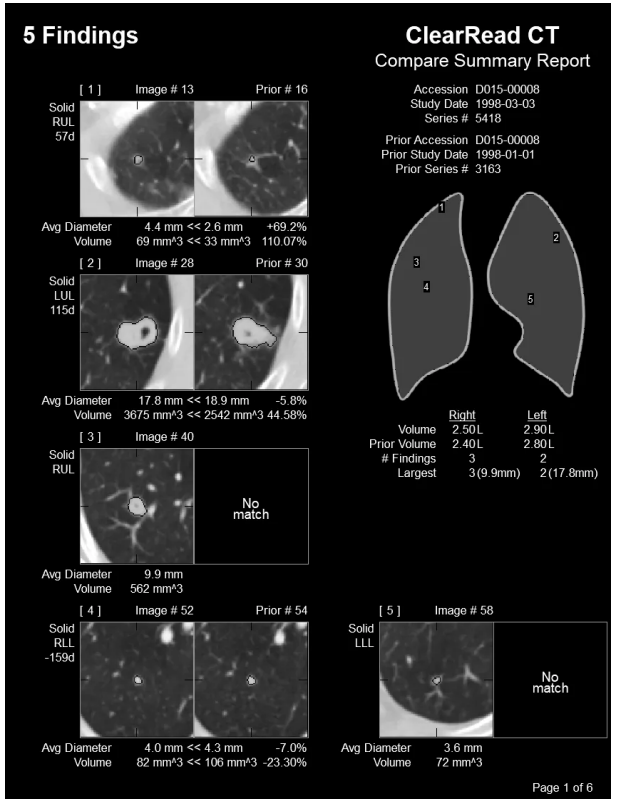
J Med Internet Res. 2018 Mar 26;20(3):e89. doi: 10.2196/jmir.9462.

Detecting Motor Impairment in Early Parkinson's Disease via Natural Typing Interaction With Keyboards: Validation of the neuroQWERTY Approach in an Uncontrolled At-Home Setting

AI Modernizing Cancer Detection

AI is central to prostate and lung cancer workflows at UTMB

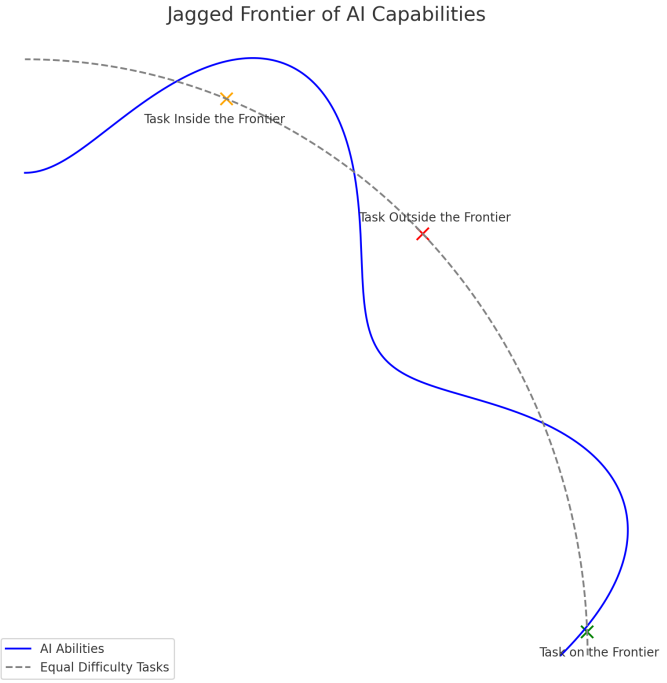




A Reality Check on AI

- AI models can Degrade in Performance Over Time
- Data Shift and Model Drift as Soft Failures
- Data Anomalies and Hard Failures

AI capability is a "Jagged Frontier"



https://www.hbs.edu/faculty/Pages/item.aspx?num=64700

AI Error Distorts the Medical Record

- AI performance metrics do not reflect the real clinical consequences of many different errors.
- Missed findings can cause serious harm, while other errors may simply create extra review work.
- AI tools may populate or shape records before a physician fully reviews the case.
- AI can miss context that matters most for assessing functional capacity.
- Diagnostic records may appear complete while still containing uncorrected errors or omissions.

Quality, Trustworthiness, and Accountability

- Accountability Dilemma: If an AI tool directly or indirectly leads to a misdiagnosis in the medical record, who is at fault? Conversely, if a physician overrides or ignores AI and makes an error, are they liable?
- Calibration Drift: If an AI model's predicted probabilities of a disease no longer match actual clinical outcomes, how does this skew an adjudicator's assessment of a claimant's disability risk
- Defining "Good Enough": There is an ongoing need to establish standardized benchmarks for acceptable AI accuracy compared to human specialists.

› BMJ Qual Saf. 2024 Jan 19;33(2):109-120. doi: 10.1136/bmjqs-2021-014130.

Burden of serious harms from diagnostic error in the USA

David E Newman-Toker [1 2], Najlla Nassery [3], Adam C Schaffer [4 5], Chihwen Winnie Yu-Moe [5], Gwendolyn D Clemens [6], Zheyu Wang [6 7], Yuxin Zhu [8 6], Ali S Saber Tehrani [8], Mehdi Fanai [8], Ahmed Hassoon [8 2], Dana Siegal [9 10]

Affiliations + expand

PMID: 37460118 PMCID: PMC10792094 DOI: 10.1136/bmjqs-2021-014130

Thank You

Appendix

Information Everywhere

Over 8.5 million consultations per month

More than 10,000 hospitals and medical centers

40% of US physicians log in daily

OpenEvidence®

BIAS IN IMAGING AI

What SSA should assume when the record doesn't say so

Judy Wawira Gichoya

William and Kay Casarella Professor of Radiology

Department of Radiology and Imaging Sciences

Emory University School of Medicine

NASEM Workshop · Session 3B · April 7, 2026

HITILAB

EMORY UNIVERSITY

Why this matters, in one number

Radiology is the test case for every other specialty

1,357 FDA-cleared AI-enabled medical devices as of December 2025.

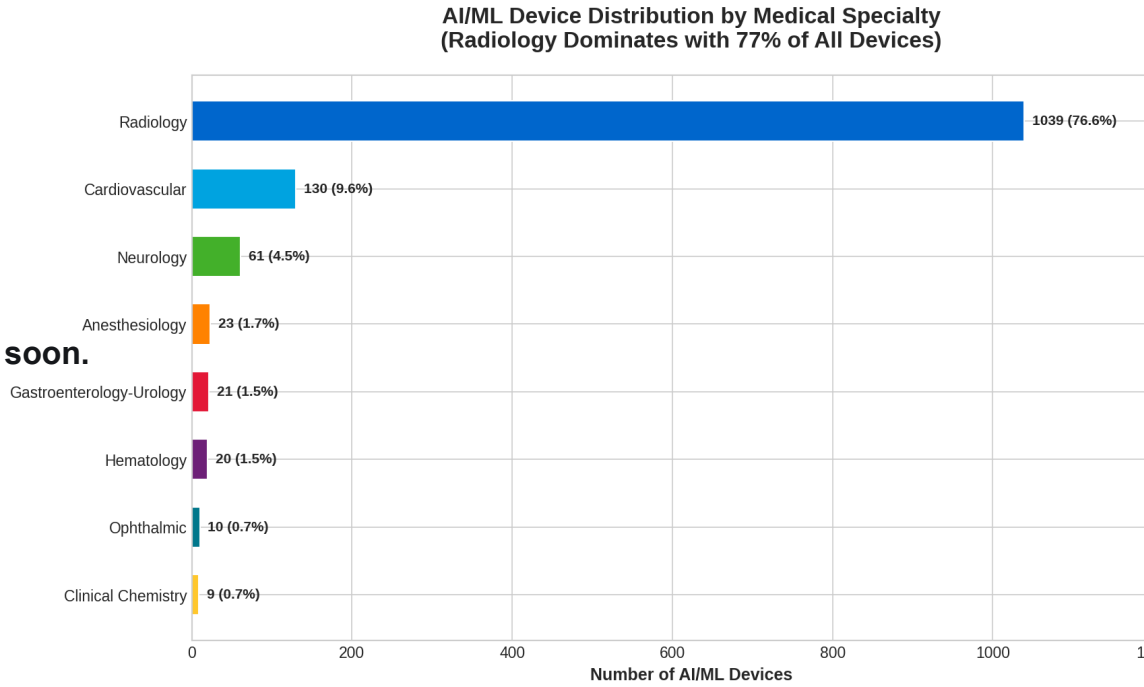
1039 are radiology — the single largest concentration across all medical specialties.

Deployed performance does not match trial performance.

Workflow incompatibilities, automation bias, and differential subgroup performance repeatedly documented in real-world evaluations.

Whatever bias patterns we see in radiology first, SSA will be reading in every specialty soon.

Thesis: imaging AI bias is real, measurable, and almost always invisible in the record SSA sees.



FDA Center for Devices and Radiological Health, 2025; Zhang K, Khosravi B, Vahdati S, Erickson BJ. Radiology 2024;310(1):e230242.

Bias starts before the AI model — not all patients get imaged the same

The upstream disparity the record never shows

Access to advanced imaging is already unequal.

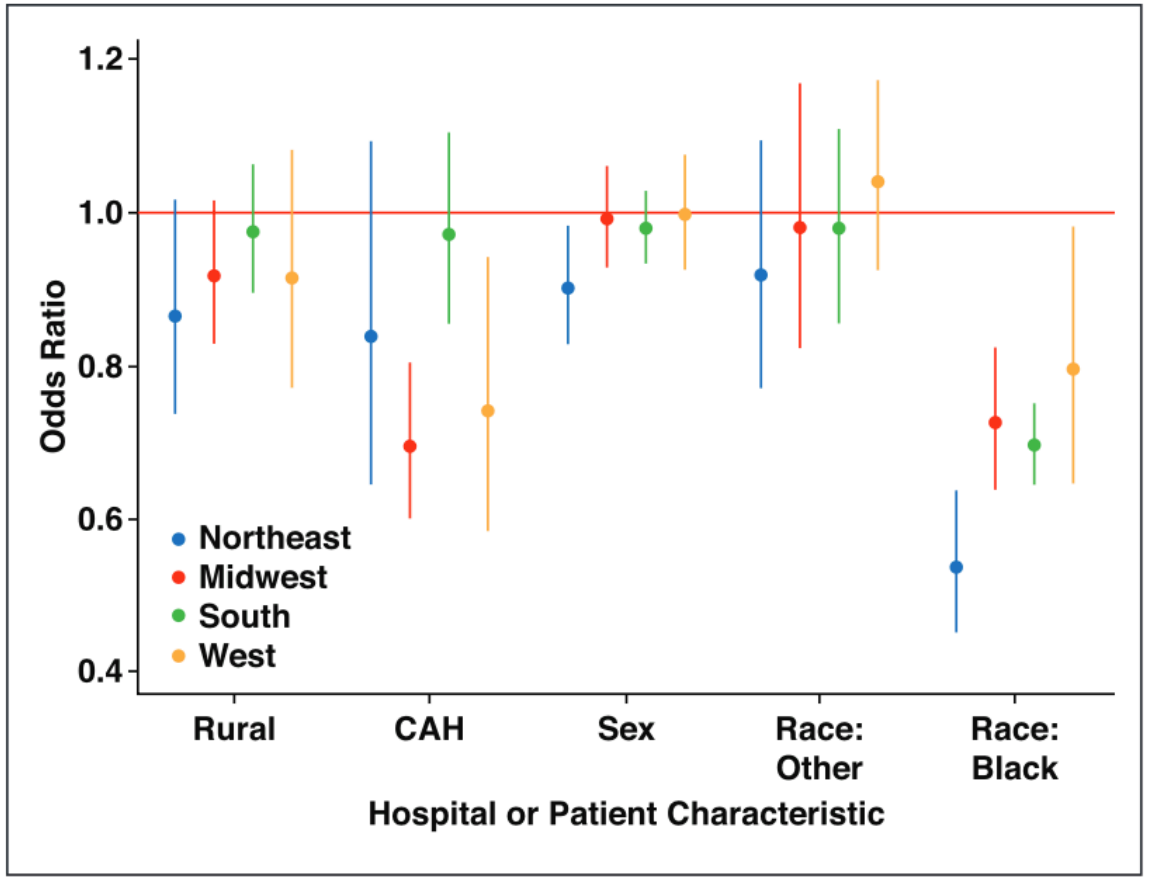
- Medicaid and uninsured patients receive MRI/CT/PET at 52–67% the rate of privately insured patients for comparable presentations (Ross et al., JACR 2020; Smith-Bindman, JAMA 2019).
- Access to mpMRI/DBT/baseline screening

Rural and non-English-speaking patients sit further down the gradient.

- Travel distance, interpreter availability, after-hours MRI scarcity, and ED triage patterns all push the same populations to lower-modality workups.

The populations least likely to be imaged are the same ones biased models under-read when they are.

- Two disparities compound in the same direction. A 'normal' file for an under-imaged claimant may simply mean nobody ordered the scan.



Odds ratio of CT/MRI imaging in the ER

Ross AB et al. JACR 2020. Smith-Bindman R et al. JAMA 2019;322(9):843. Onega T et al. Cancer 2008. Schrager JD et al. Health Aff 2019.

AI encodes demographics directly from pixels

Not a dataset artifact — an architectural property

Deep models predict self-reported race from chest X-ray, CT, and mammography.

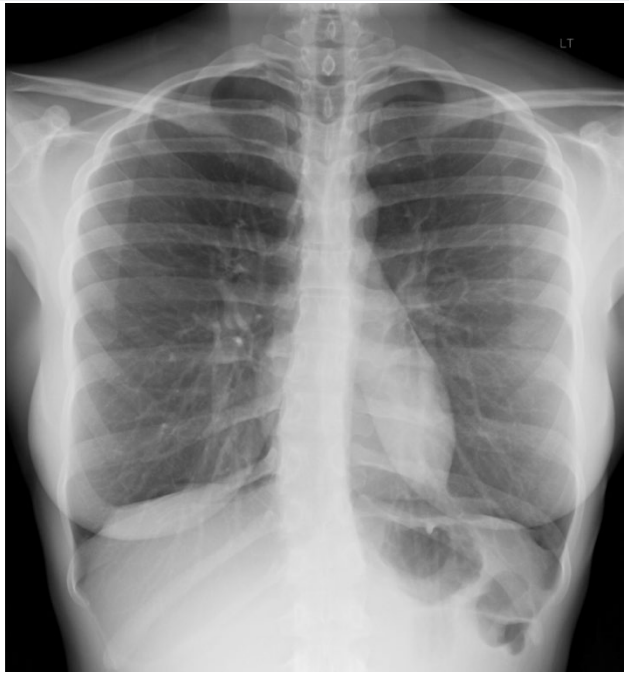
Performance surpasses experienced radiologists, across modalities and architectures. Gichoya et al., Lancet Digital Health 2022.

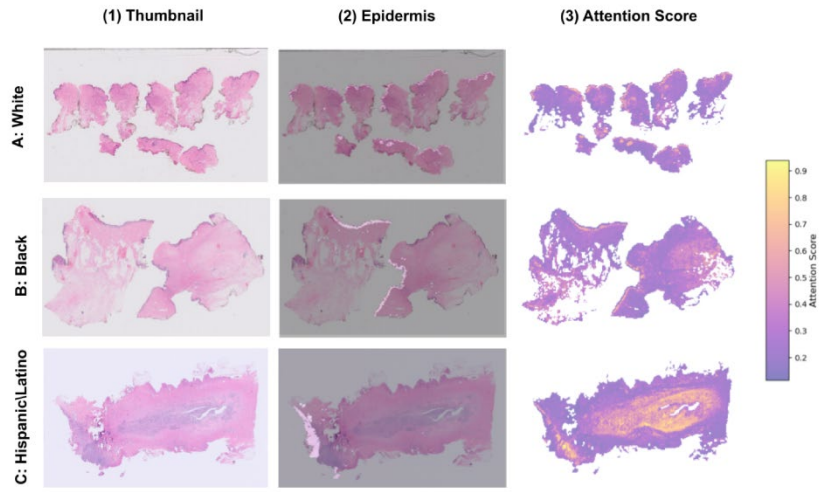
Extended to sex, age, and future healthcare expenditure.

Sohn 2022 (chest radiographs → future expenditure); Oura 2023 (knee radiographs → sex). The phenomenon is not task-specific.

Shows up in CNNs, vision transformers, and linear probes.

Not a bug of any one architecture. The image itself carries the signal.





Gichoya JW et al. Lancet Digit Health 2022;4(6):e406–e414.
Chen, Shengjia, et al. "Predicting Patient Self-reported Race From Skin Histological Images with Deep Learning." MICCAI Workshop on Fairness of AI in Medical Imaging. Cham: Springer Nature Switzerland, 2025.

ORIGINAL ARTICLE DATA SCIENCE | VOLUME 19, ISSUE 1, P184-191, JANUARY 01, 2022

Detecting Racial/Ethnic Health Disparities Using Deep Learning From Frontal Chest Radiography

Ayis Pyrros, MD • Jorge Mario Rodríguez-Fernández, MD • Stephen M. Borstelmann, MD •
Judy Wawira Gichoya, MD • Jeanne M. Horowitz, MD • Brian Fornelli, MS • Nasir Siddiqui, MD •
Yury Velichko, PhD • Oluwasanmi Koyejo, PhD • William Galanter, MD, PhD • Sh

DOI: https://doi.org/10.1016/j.jacr.2021.09.010 • Check for updates

Chest radiography as a biomarker of ageing: artificial intelligence-based, multi-institutional model development and validation in Japan

Yasuhito Mitsuyama, Toshimasa Matsumoto, Hiroyuki Tatekawa, Shannon L Walston, Tatsuo Kimura, Akira Yamamoto, Toshio Watanabe, Yukio Miki, Daiju Ueda

> JACC Cardiovasc Imaging. 2021 Nov;14(11):2226-2236. doi: 10.1016/j.jcmg.2021.01.008. Epub 2021 Mar 17.

Deep Learning to Estimate Biological Age From Chest Radiographs

Vineet K Raghu [1], Jakob Weiss [2], Udo Hoffmann [3], Hugo J W L Aerts [4], Michael T Lu [3]

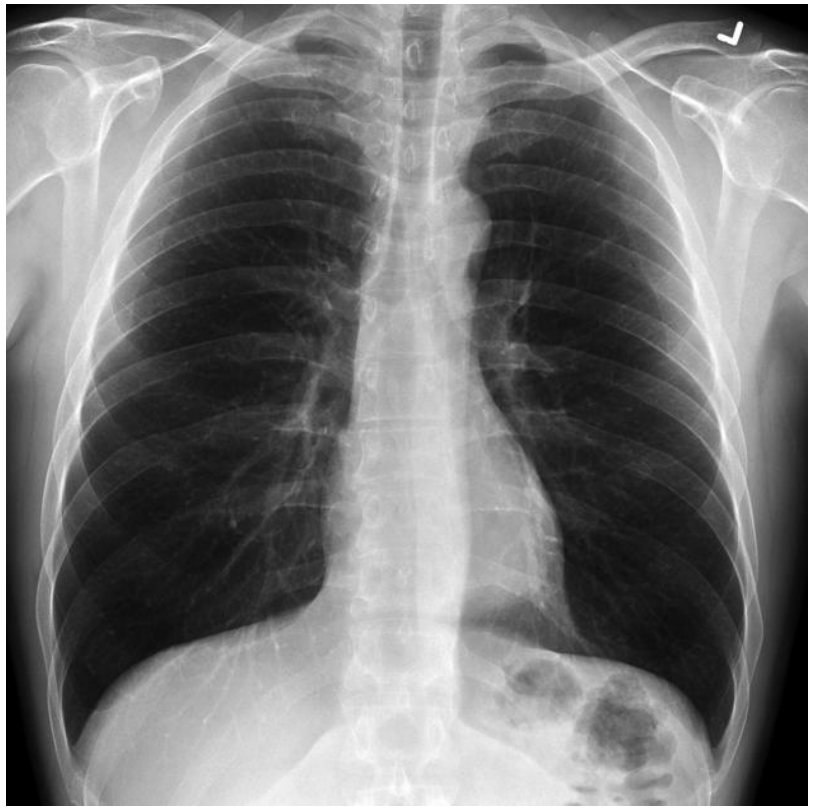
> articles > article

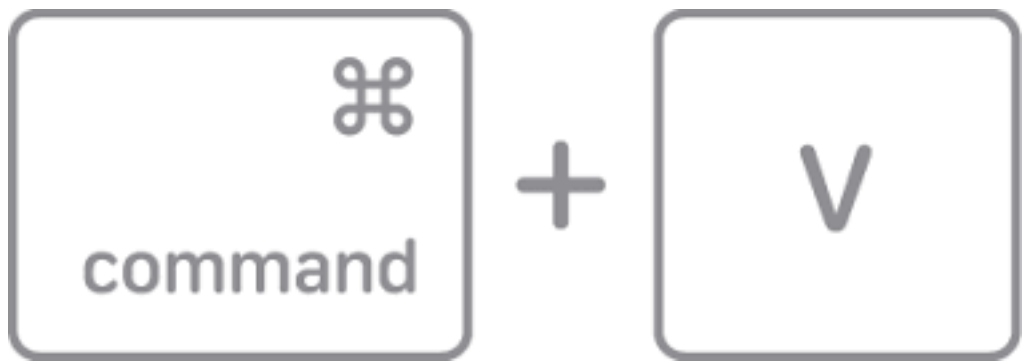
blished: 18 May 2022

Prediction of future healthcare expenses of patients from chest radiographs using deep learning: a pilot study

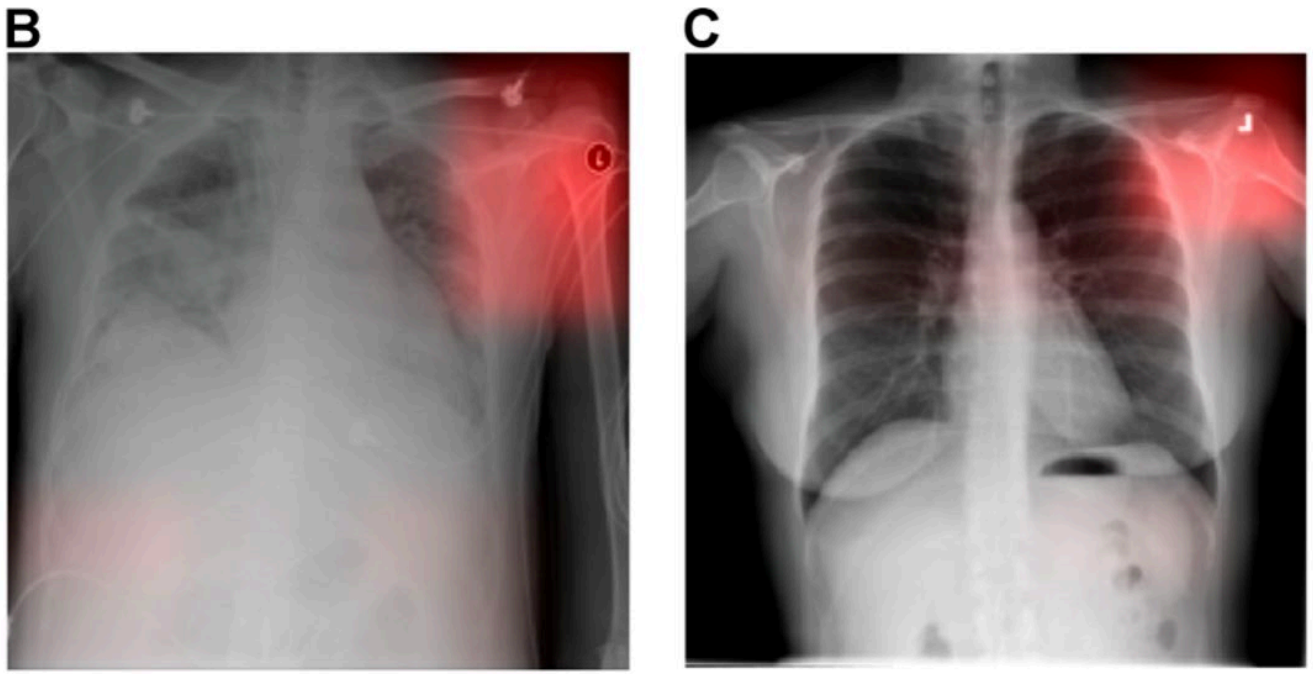
Jae Ho Sohn, Yixin Chen, Dmytro Lituiev, Jaewon Yang, Karen Ordovas, Dexter Hadley, Thienkhai H. Vu, Benjamin L. Franc & Youngho Seo

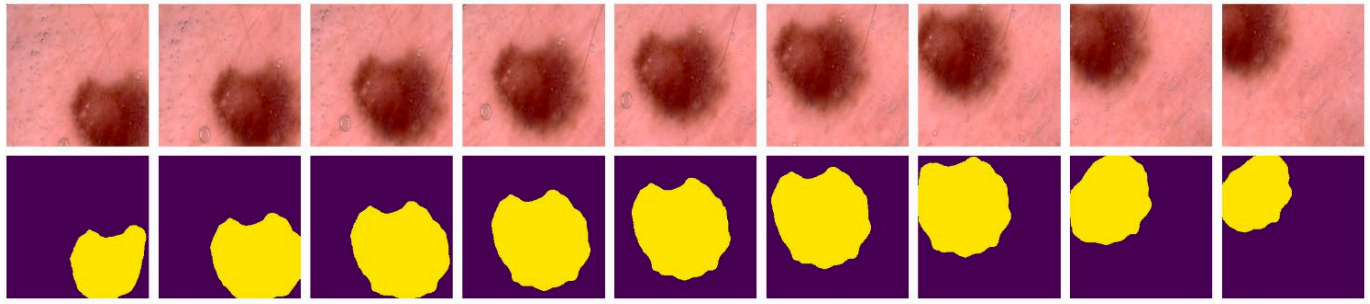
Judy is “Black”/ Kenyan, F, 60 yrs (CXR age = 78 yrs), SDI 45, ICD codes – COPD, CHF, 15,000 USD

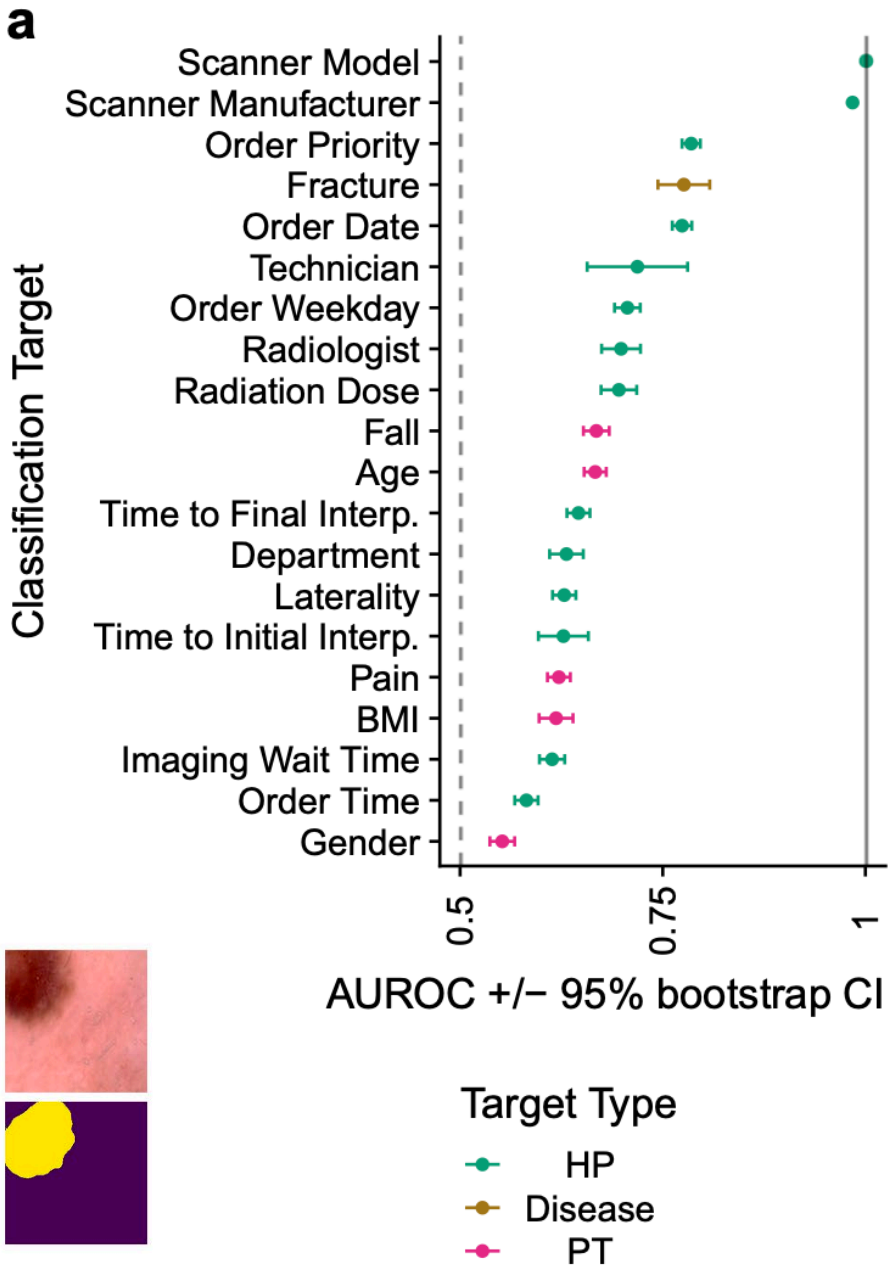



⌘
command
+
V

Shortcuts in medical imaging







Zech, John R. et al "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study." PLoS medicine 15, no. 11 (2018): e1002683.
Lin, Manxi, et al. "Shortcut Learning in Medical Image Segmentation." *arXiv preprint arXiv:2403.06748* (2024).
Badgeley, M.A. L. *et al.* Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digit. Med.* 2, 31 (2019).

How bias gets locked into the file

Shortcut learning + automation bias and deskilling

Shortcut learning.

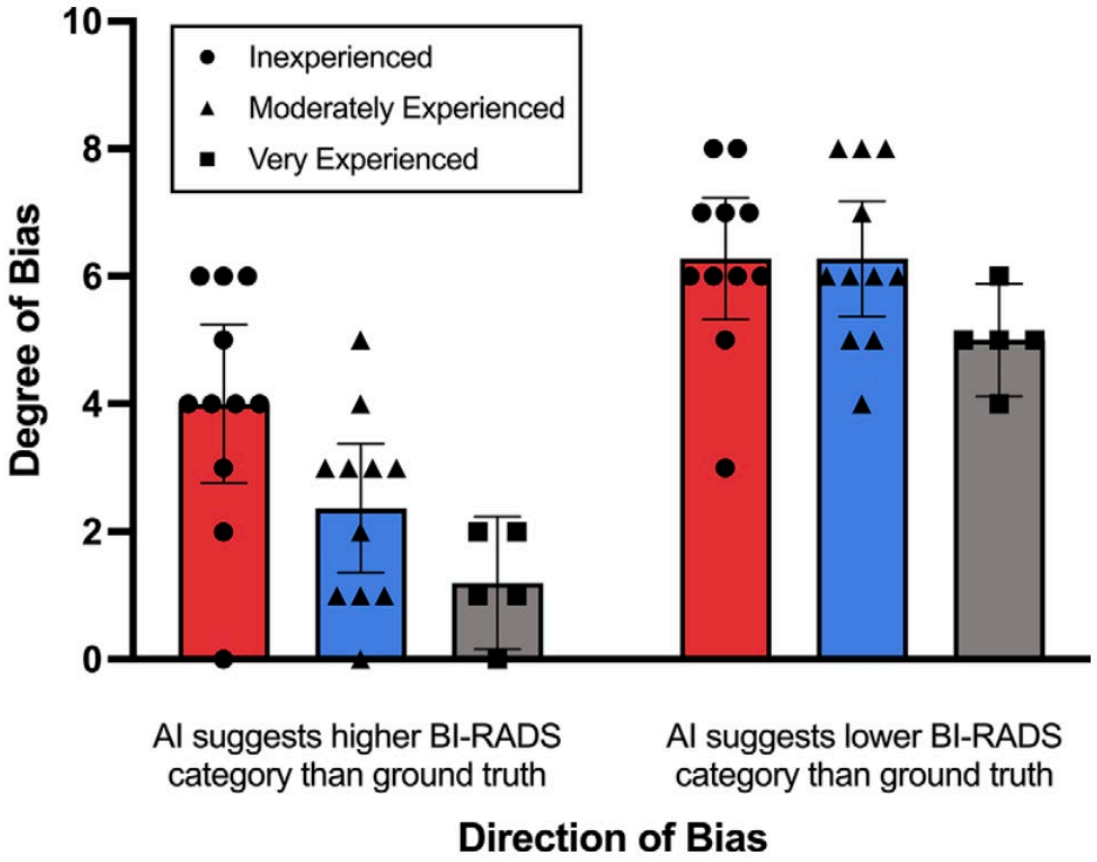
Models exploit hospital tokens, laterality markers, equipment signatures, acquisition year. CE-marked / FDA-cleared knee OA tools change predictions on horizontal image flip (Banerjee et al., JACR 2023).

Automation bias.

Biased AI raised error rates 0.69 → 2.21 and persisted on new cases (Vicente & Matute). Dratsch mammography crossover: inexperienced radiologists dropped 80% → 20% following incorrect AI.

Deskilling.

Gastroenterology: 20% decrease in adenoma detection on unassisted colonoscopy after routine AI exposure — the first documented case of AI-induced deskilling affecting patient outcomes.



Automation bias in mammography

Banerjee I et al. JACR 2023;20(9):842–851. Dratsch T et al., mammography crossover. Vicente & Matute; gastro deskilling 2024.

Best demographic predictors = most unfair models

Encoding is not neutral — it competes with pathology

The finding.

Models with the highest demographic prediction accuracy also show the largest fairness gaps in clinical performance. Yang, Zhang, Gichoya, Katabi, Ghassemi. Nat Med 2024.

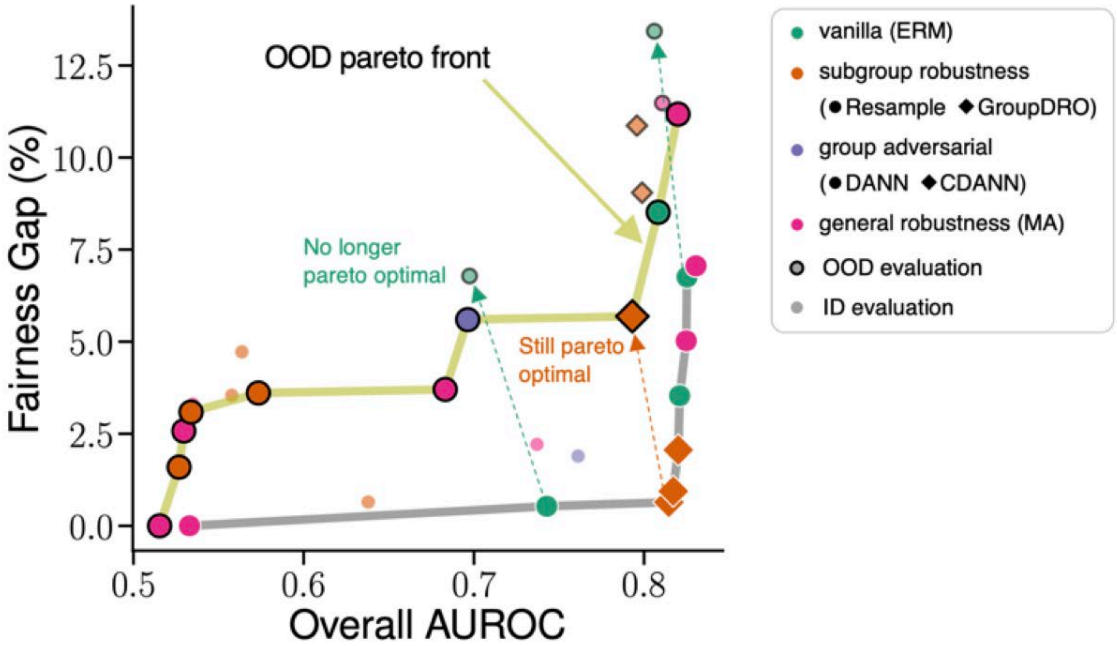
The mechanism.

Demographic shortcuts compete with pathology features. The same representations that l a model guess who the patient is make it worse at telling the doctor whether they are sick.

The policy implication.

Not 'remove bias' — mathematically impossible under differing base rates. It's 'deploy with awareness of bias you cannot fully remove.' and remember Fairness guarantees DO NOT travel

OOD evaluation →



What the gaps actually look like after AI deployment

Three modalities, three numbers

Chest X-ray.

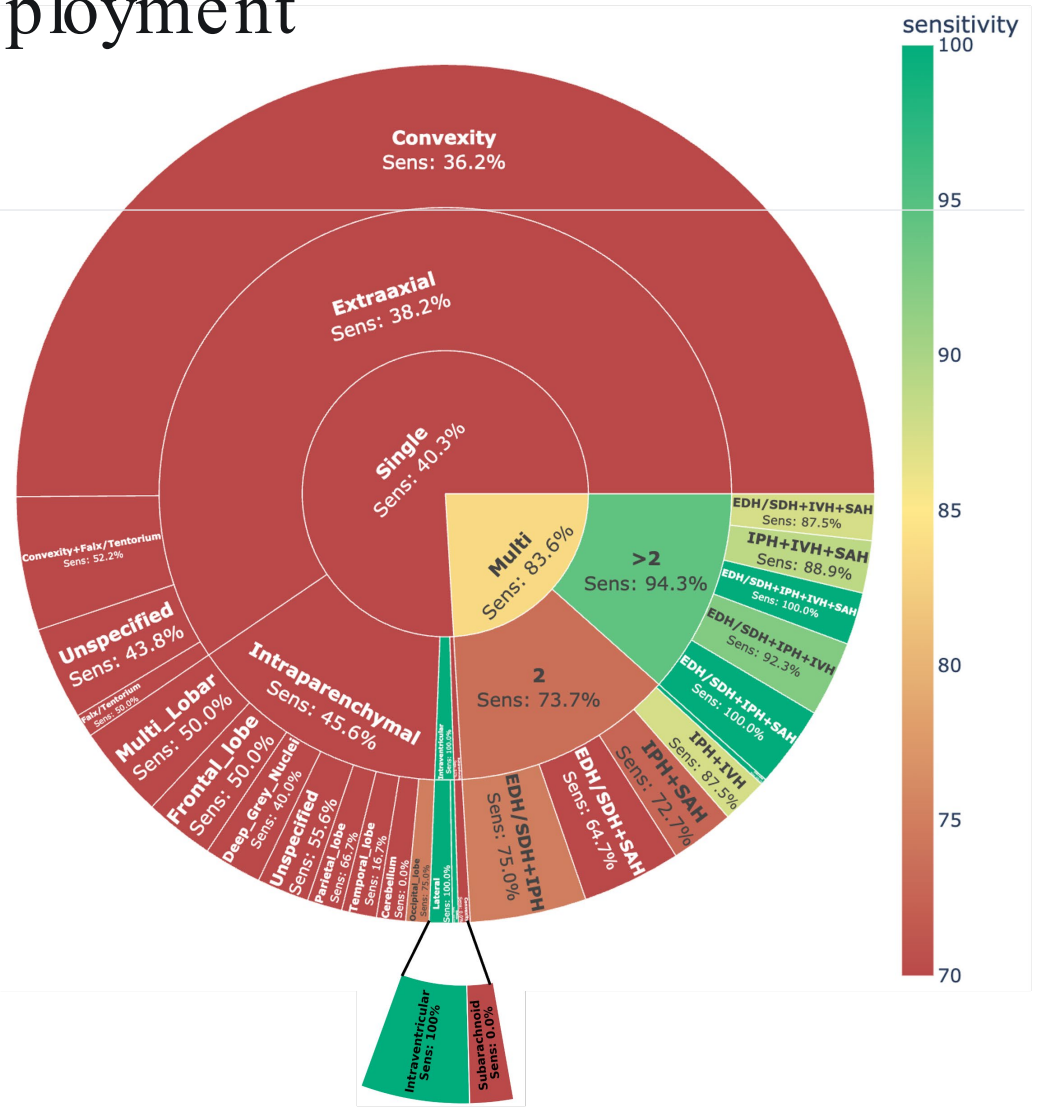
Systematic underdiagnosis of Black, female, Hispanic, and Medicaid patients. Age gaps reach 30% between elderly and young adults (Seyyed-Kalantari et al., Nat Med 2021).

Dermatology.

73% vs 60.5% accuracy — light-skinned vs dark-skinned subjects on commonly benchmarked commercial models.

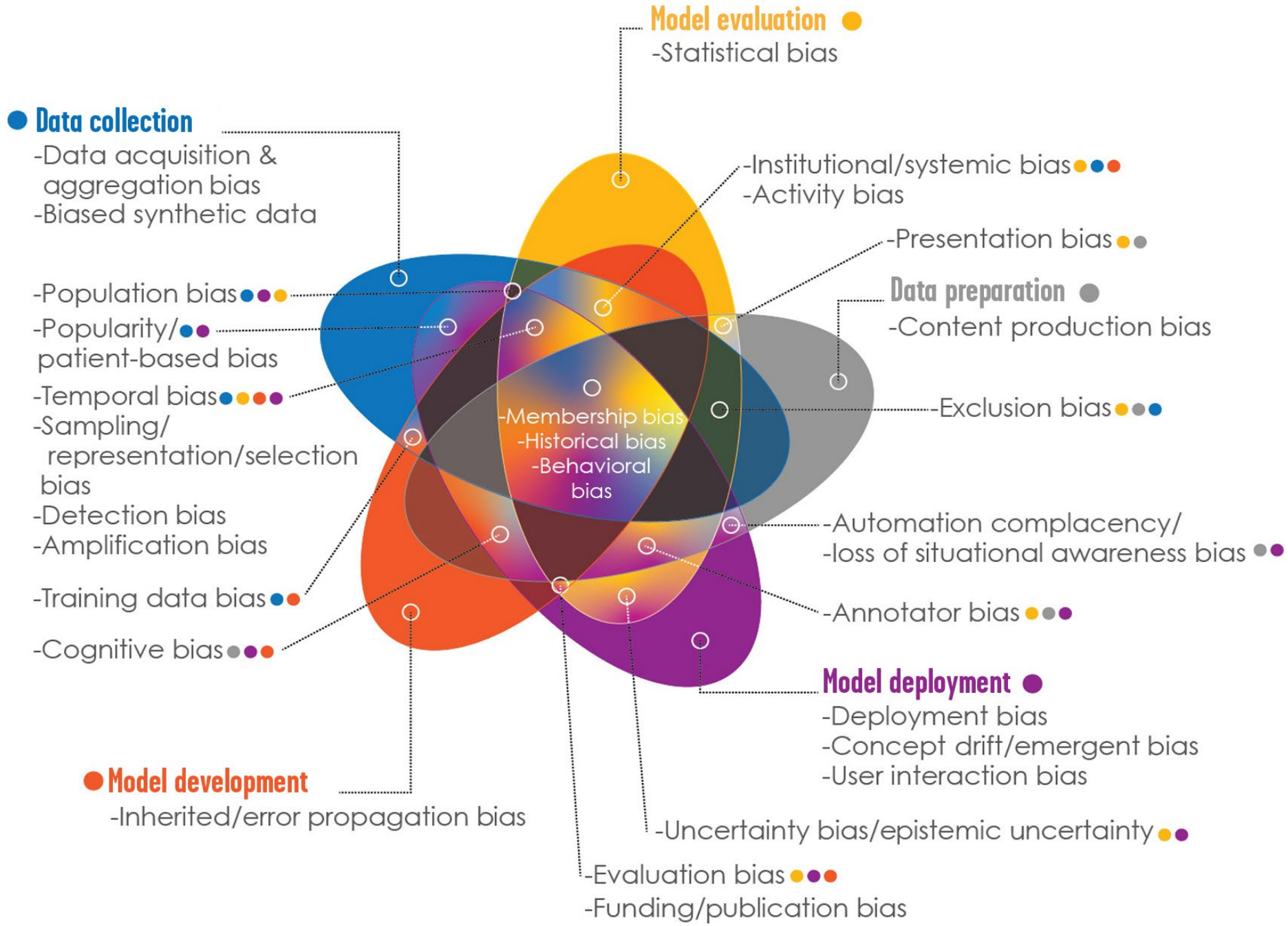
Cardiac MRI (UK Biobank).

10% relative performance drop for Black and female patients on segmentation. Intersectional gaps compound — young Black Medicaid patients in ED settings carry the highest CXR underdiagnosis.



Subgroup Evaluation of an ICH algorithm

Seyyed-Kalantari L et al. Nat Med 2021. Vision-language foundation models extend this across 48 radiographic findings (chapter §New Frontiers)
Chavoshi, Mohammadreza, et al. "Real-world performance evaluation of a commercial deep learning model for intracranial hemorrhage detection." npj Digital Medicine (2025).



"Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment," J. Med. Imag. 10(6) 061104 (26 April 2023)

The record is stripped of every bias signal

And the provenance, where it exists, is empty

What the adjudicator gets.

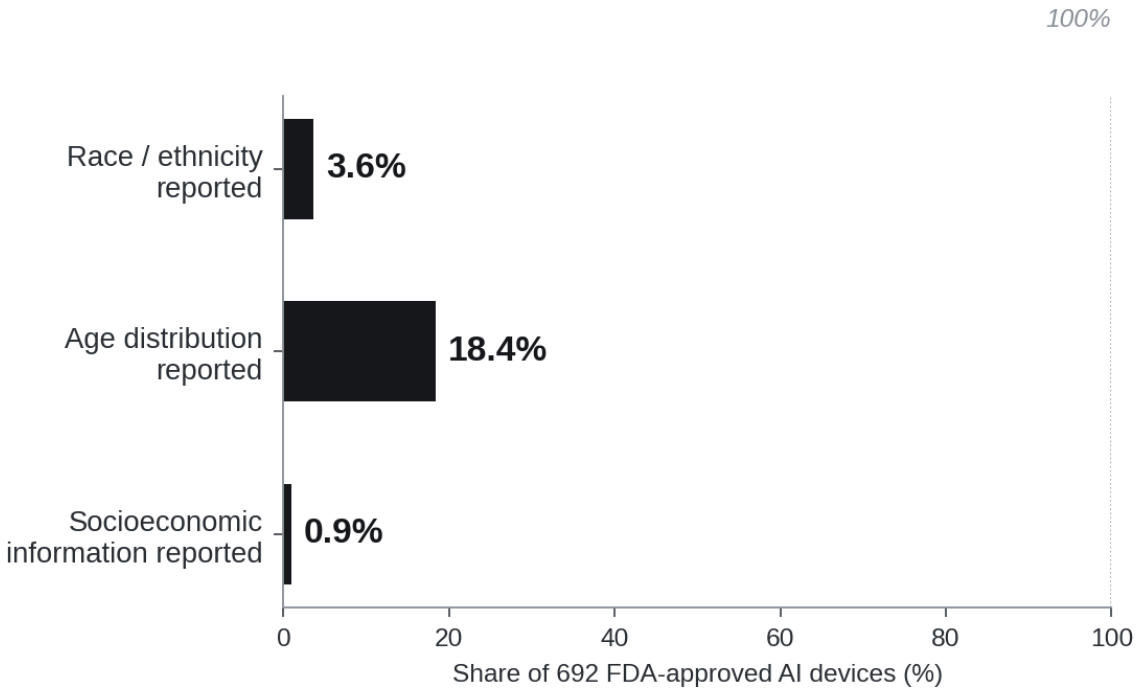
Radiology report text only. No image, no model ID, no version, no confidence, no subgroup performance. USCDI v3 carries diagnostic imaging reports but not images. FHIR DiagnosticReport has no 'AI-assisted' field.

What the provenance would say.

Of 692 FDA-approved AI devices: 3.6% report race/ethnicity, 99.1% provide no socioeconomic data, 81.6% fail to report age distributions. Fewer than 15% of published radiology AI papers report comprehensive sociodemographic data (Driessen, Gichoya et al., JACR 2023).

Even perfect disclosure would not tell SSA whether the model works for the claimant.

'Ask for provenance' is not enough when the provenance is empty.



FDA-approved AI device transparency gap. n = 692 devices.

Driessen, Rebecca, et al. "Sociodemographic variables reporting in human radiology artificial intelligence research." Journal of the American College of Radiology 20.6 (2023): 554-560.

Equity implications for disability evaluation

The bias is coherent and asymmetric

Under-imaged claimants look 'less objective.'

Access disparities upstream — rural, uninsured, non-English-speaking, behavioral-health — mean fewer imaging findings in the file. The Blue Book asks for objective evidence the claimant never got the chance to generate.

Biased models under-describe severity for the same people.

The populations least likely to be imaged are the populations most likely to be under-read when they are imaged. Two disparities compound in the same direction.

Language in the note tilts supportability and consistency.

Black patients are 2.54× more likely to have negative descriptors ('noncompliant,' 'refused') in their notes (Sun et al., Health Affairs 2022). Every piece of this goes into the 20 CFR 404.1520c analysis.

Sun M, Oliwa T, Peek ME, Tung EL. Health Aff 2022;41(2):203–211. Obermeyer Z et al. Science 2019;366:447.

Skepticism signals SSA can use tomorrow

Practical heuristics that don't depend on provenance disclosure

Sparse imaging for a severe presentation.

Trigger a consultative examination instead of inferring absence of disease.

Missing modality for the impairment.

No MRI for neuro, no EMG for radiculopathy, no echo for unexplained dyspnea — that is a record gap, not a negative finding.

Short, templated, single-reader reports.

Without a peer-review trail, treat as lower-weight evidence under the supportability prong.

Discordance between patient-reported function and imaging narrative.

The narrative may reflect shortcut features, not the patient's actual disease state.

Stigmatizing or dismissive language in notes.

Flag for review under consistency.

Site-of-care and equipment markers known to correlate with shortcut learning.

If the only imaging comes from a facility with templated AI-assisted reads, weight accordingly.

What to demand, what to leave with

Three asks, three takeaways

Three things to demand.

FHIR provenance extension for AI on imaging reports.

Model, version, confidence, training-cohort demographics.

Subgroup-performance disclosure – post real-world monitoring.

Ganapathi et al., Nat Med 2022. This is where Shantanu's FDA lane picks it up.

Routine SSA outcome audits by claimant demographics × imaging-modality coverage.

Detection, not prediction. Appeal pathway for AI-influenced findings.

Three things to leave with.

Bias in imaging AI is architectural.

Models encode demographics from pixels whether you ask them to or not.

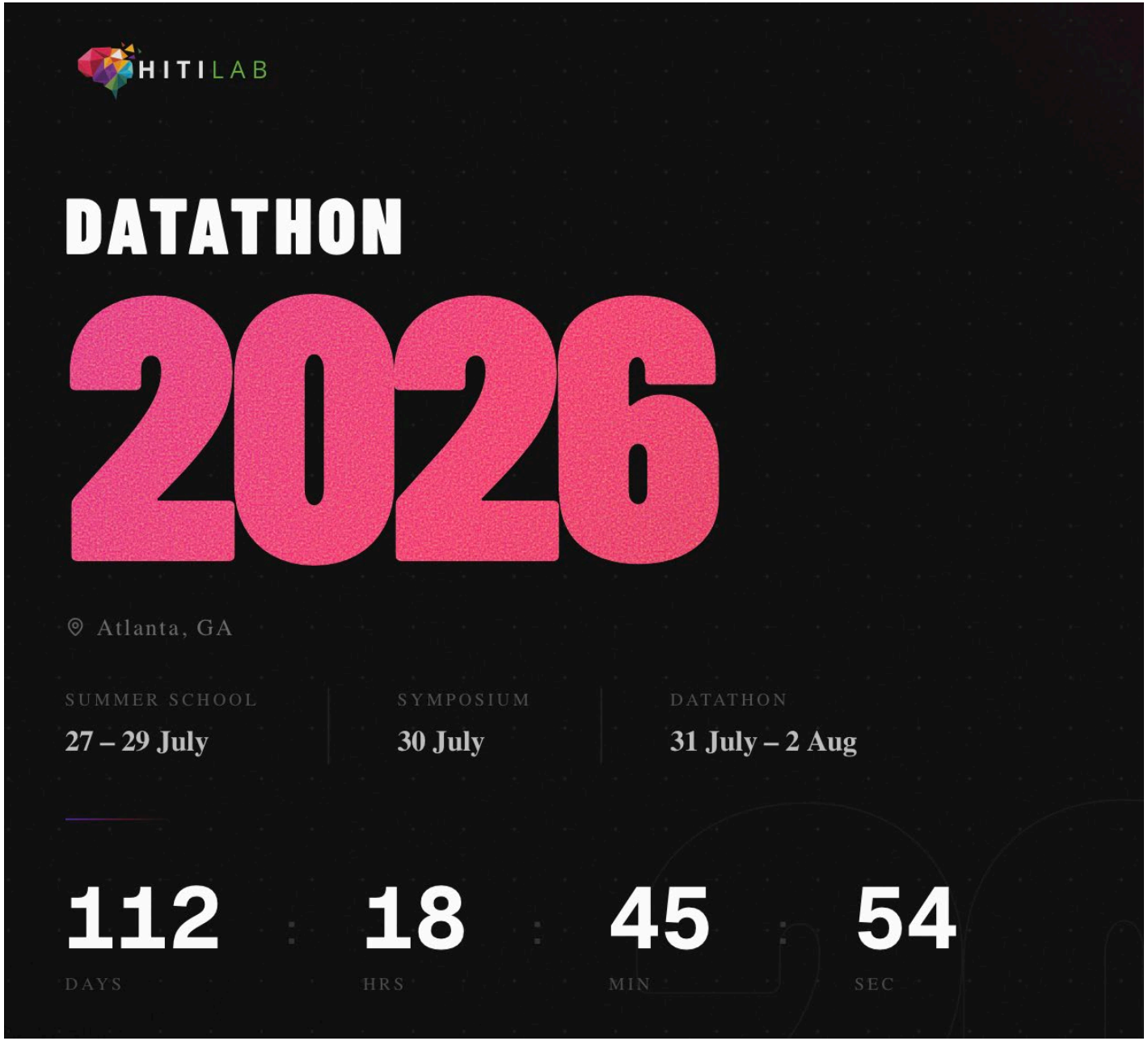
The best demographic predictors are the most unfair models.

'Pick a better model' is not a strategy.

Skepticism signals work today.

Build audit and appeal into the pipeline now, before deployment scales.

Thank you. Ganapathi S et al., STANDING Together. Nat Med 2022. Happy to continue in the panel.



https://datathon.org

Thank you

Questions & discussion

Judy Wawira Gichoya
Emory University School of Medicine · HITI Lab
judywawira@emory.edu

Healthcare Innovation and Translational Informatics (HITI) Lab at Emory

From Narrative to Navigation

The Impact of Ambient and Agentic AI on Medical Documentation

David Dorr, MD, MS

Chief Research Information Officer and Strategic Lead, AI, OHSU

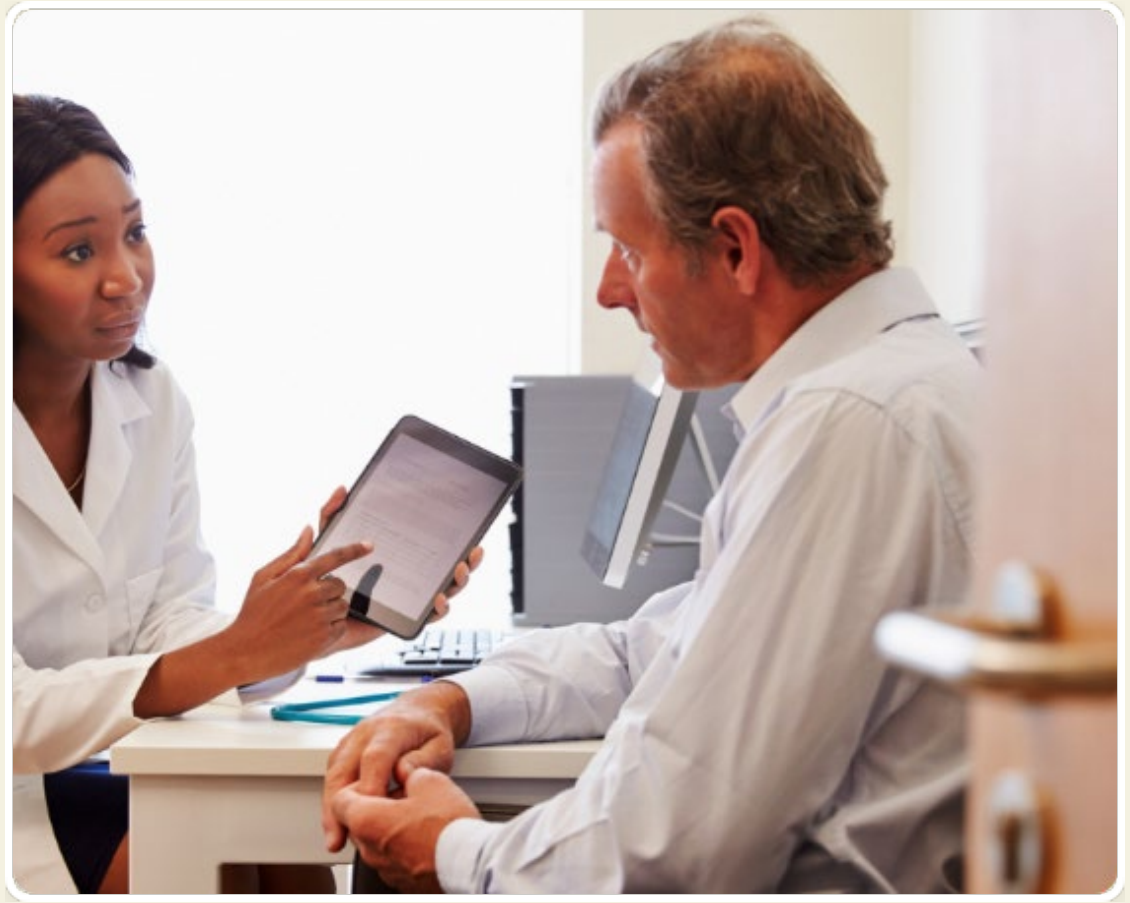
Workshop: AI and the Medical Record in the Context of SSA Disability Evaluations

OHSU
Center for
AI-ENABLED LEARNING
HEALTH SCIENCE

The Documentation Revolution

Ambient Scribe AI Adoption

- **Rapid Deployment:** Currently active in over 70% of major health systems.
- **The Shift:** Moving from manual typing to "Conversational Capture."
- **Pros:** Higher provider eye contact; comprehensive narrative.
- **Risks:** "Over-documentation" and "Summary Bias" favoring clinical signals over functional limits.



Source: Evaluation of Ambient AI Scribes in Primary Care (2024)

The AI- Driven Patient

Self- Diagnosis 2.0

Patients are using Large Language Models (LLMs) like ChatGPT and Gemini to interpret complex symptoms before reaching the clinic.

Patient- Led Encounters

AI provides patients with "scripts" to request specific diagnostic tests or specialty referrals, moving reports away from "raw" presentation.

'The medical may become a curated narrative of AI- driven self- advocacy."

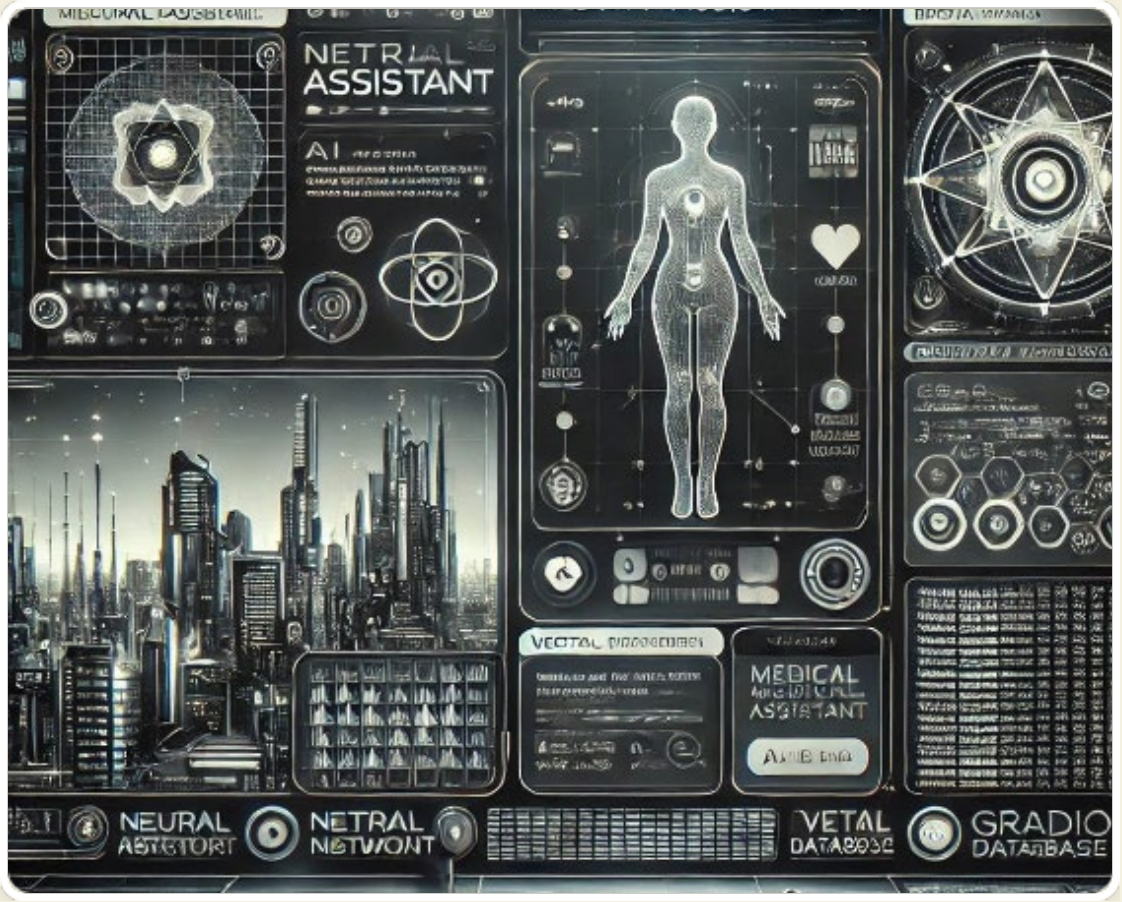
Agentic Navigation & Care Teams

From Tools to Agents

AI shifting from passive "checklists" to active "agents" that guide clinicians through complex workflows.

The Clinical Compass

Agents "nudge" care teams for completeness (e.g., "Patient meets criteria for X, order test Y").



The "Completeness" Paradox

The Interpretation Trap

Data gaps are now **systemic**, not random. A "complete" record may reflect technological access rather than clinical severity.

- **Digital Divide:** AI-robust documentation in urban centers vs. sparse records in rural safety-net clinics.
- **Signaling Disparity:** "Missing data" is becoming a primary signal of social and technological inequality.

Sources of AI Bias in Documentation

From Patient → AI Scribe → EHR → SSA Reviewer

Errors Extant in the AI Chart

Hallucinations: AI "inventing" physical exam findings that "fit" the diagnosis but were never actually performed.

Cloning: The rapid propagation of AI-generated errors across multiple visits and providers.

The "Veneer of Accuracy": Highly professional-looking notes that mask underlying clinical inaccuracies.

Call for Transparency

1. **Welcome Transparency:** Require "AI Provenance" tags to identify synthesized vs. human-heard content.

2. **Context is King:** Distinguish between AI quantification and AI narrative synthesis.

3. **The Human Anchor:** Maintain "Clinical Corroboration" to ensure longitudinal functional reality matches the note.

Key Resources for SSA

Organization / Author	Resource Title	Key Focus
Dorr, D. A., et al. (2024)	Evaluation of Ambient AI Scribes	Primary Care Deployment Data
CHAI	Responsible AI Framework	Coalition for Health AI Standards
National Academy of Medicine	AI in Health Care	Hope, Hype, Promise, and Peril
FSMB	AI in Documentation Guidelines	State Medical Board Standards

Thank You

U.S. FOOD & DRUG ADMINISTRATION

FDA Considerations for Regulating AI-Enabled Radiological Devices

Robert Ochs, Ph.D.
Director
OHT8: Office of Radiological Health
Office of Product Evaluation and Quality
Center for Devices and Radiological Health
U.S. Food and Drug Administration

Risk-Based Regulatory Approach

- Some software functions do **not** meet the definition of a medical device *(i.e., FDA does not regulate)*

- **Low-risk** medical devices are generally exempt from FDA premarket review

- **Moderate** to **high-risk** medical devices undergo premarket review of their safety and effectiveness – some are also subject to post-market studies

Authorized AI-Enabled Devices

- Public AI-Enabled Medical Devices List
  - 1430 AI-enabled devices (through March 4, 2026)

- Radiological AI examples:
  - **Image segmentation** and **quantitative** measurements
  - **Image acquisition**
  - Aids for **image triage** *(alerting & prioritizing potentially time-sensitive cases for review)*
  - Aids for **image interpretation** *(e.g., marking an image and assessing the likelihood of disease)*

AI-Enabled Medical Devices List

Devices are listed in reverse chronological order by Date of Final Decision. To change the sort order, click the arrows in the column headings.

Use the Submission Number link to display the approval, authorization, or clearance information for the device in the appropriate FDA database. The database page will include a link to the FDA's publicly available information.

Download a CSV File | Download an Excel File | Save as XML File*

*To save the XML file, right click and save the file to your computer and open in the appropriate program.

Search: Show 50 entries

Date of Final Decision	Submission Number	Device	Company	Panel (lead)	Primary Product Code
05/30/2025	K251408	BriefCase-Triage	Aidoc Medical, Ltd.	Radiology	QAS
05/30/2025	K250236	Swoop® Portable MR Imaging® System (V2)	Hyperfine, Inc.	Radiology	LNH
05/30/2025	K243863	Opulus™ Lymphoma Precision	Roche Molecular System	Radiology	QIH
05/30/2025	K243005	AudaxCeph Cephalogram Analysis Software	Audax d.o.o.	Radiology	QIH
05/30/2025	K242830	LensHooke X3 PRO Semen Quality Analyzer; LensHooke X3 PRO SE Semen Quality Analyzer	Bonraybio Co., LTD.	Hematology	POV
05/30/2025	DEN240047	Allix5	Clairity, Inc.	Radiology	SEZ
05/29/2025	K250543	Voluson™ Performance 16; Voluson™ Performance 18	GE Medical Systems Ultr	Radiology	IYN
05/28/2025	K243378	Rapid MLS	iSchemaview Inc.	Radiology	QIH
05/28/2025	K250427	TAIMedImg DeepMets	Taiwan Medical Imaging	Radiology	QKB
05/28/2025	K250367	CoLumboX	Smart Soft Healthcare AD	Radiology	QIH
05/27/2025	K250932	DeepRhythmAI	Medicalgorithmics S.A.	Cardiovascular	DQK
05/23/2025	K250005	Clever One	Ewoosoft Co., Ltd	Radiology	QIH
05/23/2025	K243989	Second Opinion® 3D	Pearl, Inc.	Radiology	QIH
05/23/2025	K243937	Accuro 3S	Rivanna Medical, Inc.	Radiology	IYO
05/23/2025	K242594	DEEPECHO	DeepEcho	Radiology	IYN
05/21/2025	K251276	Swoop® Portable MR Imaging® System	Hyperfine, Inc.	Radiology	LNH

Benefits & Risks

Benefits

- Accuracy
- Efficiency
- Consistency
- Workflow
- Insights

Risks

- Diagnostic errors
- Lack of generalizability
- Lack of transparency
- Automation bias
- Performance changes

Intended Use

- Understanding the intended use is essential for the premarket review

Segment lung anatomy

≠

Detect lung nodules > 5mm

≠

Output likelihood of lung cancer

Premarket Review

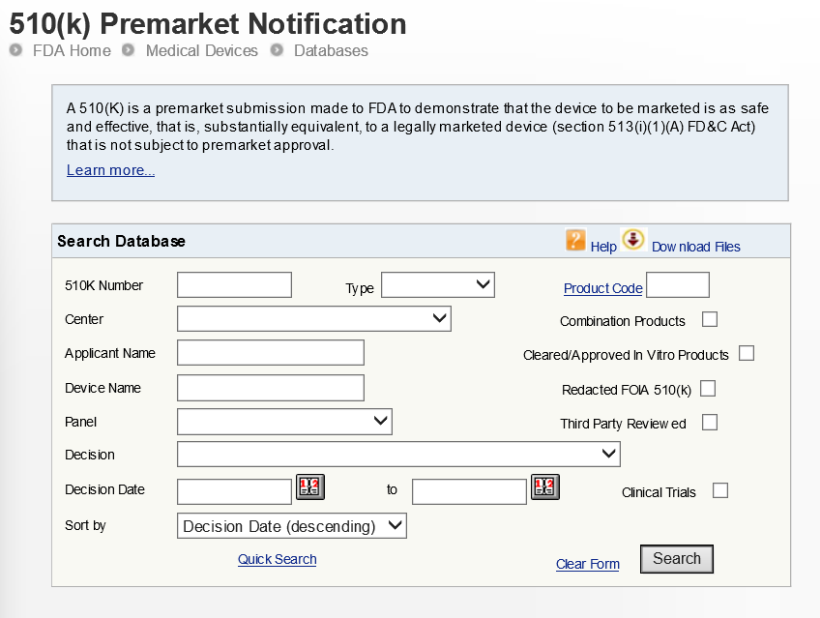
- Intended Use
- Design
- Performance testing
- Risk mitigations
- Labeling / user manual(s)
- Software & cybersecurity
- Predetermined change control plans *(if submitted)*
- Post-market monitoring plans *(if applicable)*

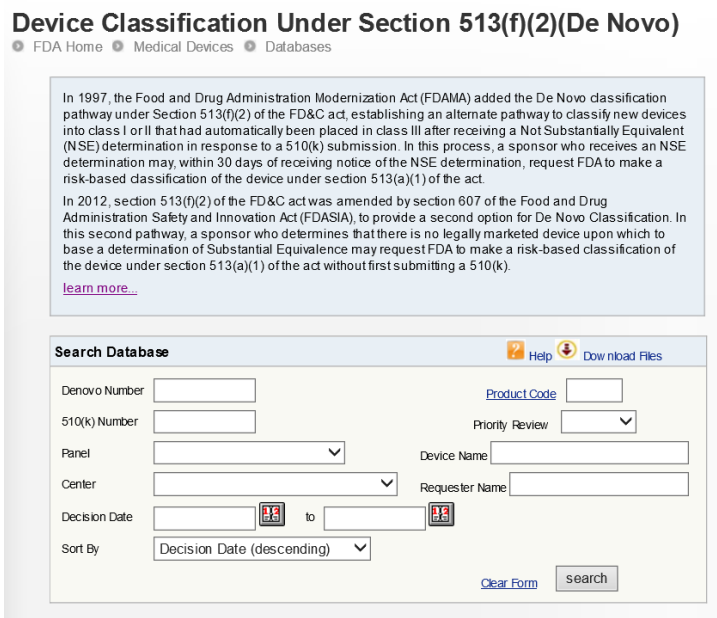
Good Machine Learning Practice for Medical Device Development: Guiding Principles

- Clinical Study Participants and Data Sets Are Representative of the Intended Patient Population
- Training Data Sets Are Independent of Test Sets
- Testing Demonstrates Device Performance during Clinically Relevant Conditions
- Users Are Provided Clear, Essential Information
- Deployed Models Are Monitored for Performance and Re-training Risks Are Managed
- …

FDA Public Databases

- Information on authorized devices will depend on the submission type
- At a minimum - the intended use and a summary of the premarket testing





Summary

- FDA takes a risk-based approach to the oversight of AI-enabled medical devices
- Moderate to high-risk medical devices undergo premarket review of their safety and effectiveness – some are also subject to post-market studies
- FDA public databases can provide information on a device's intended use and premarket testing
- Best practices, including transparency to users, can promote the generalizability and benefits of AI while reducing the risks

References

- Overview of Device Regulation | FDA
- Public databases
  - 510(k) Premarket Notification (most devices)
  - Device Classification Under Section 513(f)(2)(De Novo)
  - Premarket Approval (PMA)
- Artificial Intelligence in Software as a Medical Device | FDA
  - Good Machine Learning Practice for Medical Device Development: Guiding Principles
  - Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles | FDA
- Clinical Decision Support Software | FDA
- Policy for Device Software Functions and Mobile Medical Applications | FDA
- Cybersecurity | FDA