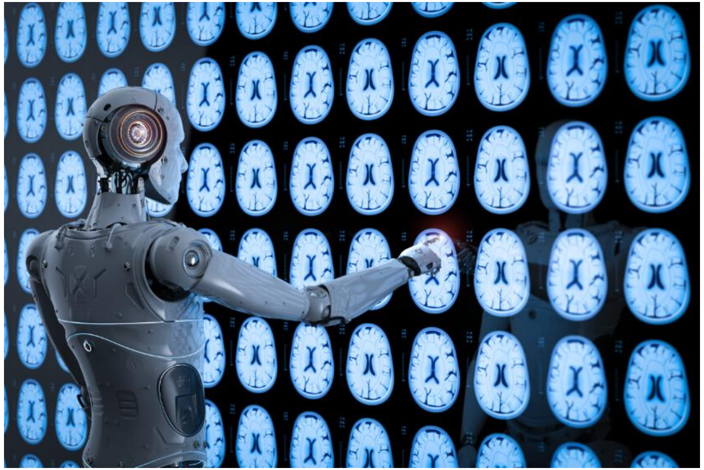
8th Nordic Course in Emergency Radiology
June 2-5, 2025 – Oslo, Norway

NORDICFORUM www.nordictraumarad.com
TRAUMA & EMERGENCY RADIOLOGY

AI in ER: Challenges of Implementation



4rai.com

Ferco Berger
Emergency & Trauma Radiologist

Sunnybrook, University of Toronto, Canada

fhberger@gmail.com

Sunnybrook
HEALTH SCIENCES CENTRE
when it matters MOST

Disclosure

- No financial disclosures

- I have to disclose that I don’t have a lot of experience using AI …
  … which has do a lot with challenges of implementation

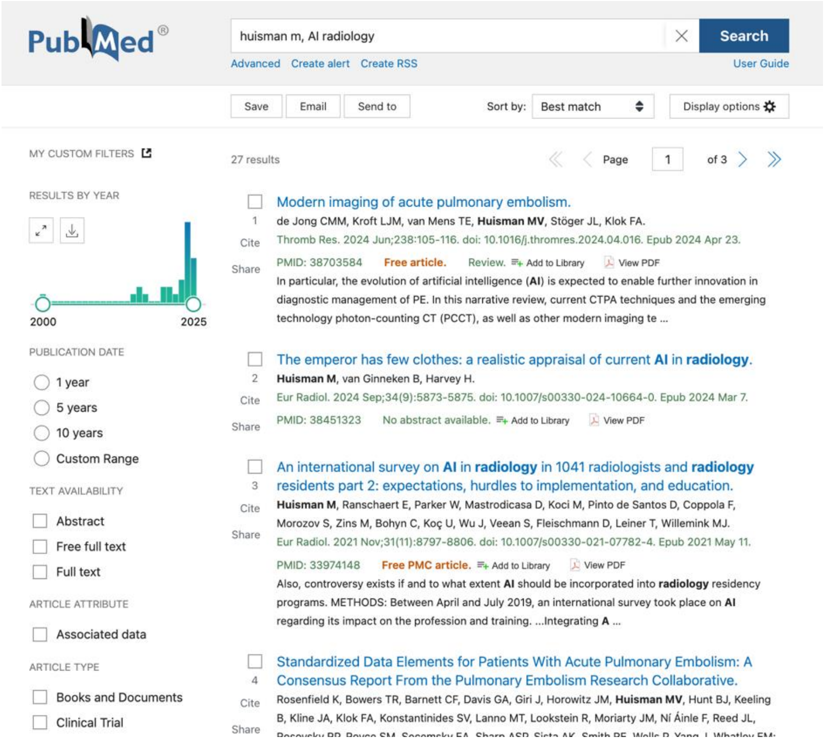
Thanks to U of T colleagues:

- Masoom Haider
- Ben Fine
- Errol Colak

Recommendation

Merel Huisman, Radiologist

Founded EuSoMII Young Club



My experience

RAPID in Stroke

Bone density in chest and MSK radiographs, built by colleague

Trying to introduce more AI tools but failed so far…

Goal here is to *raise awareness* around challenges

Content – challenges

- Local support
  - IT
  - medicolegal
- Selecting tool
  - accuracy
  - efficiency
  - turnaround time
  - quality / safety
- Validating tool
  - external / internal
  - choosing thresholds
  - F-neg/F-pos/Sens/Spec
- Bias & lack generalization
  - Diversity population, equipment, protocols, populations etc.
  - Test vs validation
  - Data imbalance, algorithm, oversight
- Pitfalls:
  - Errors
  - Alarm fatigue
  - Automation bias
    - AI overcall / AI miss, Complacency
- Continuous monitoring & Governance

Content – challenges

- Local support
  - IT
  - medicolegal
- Selecting tool
  - accuracy
  - efficiency
  - turnaround time
  - quality / safety
- Validating tool
  - external / internal
  - choosing thresholds
  - F-neg/F-pos/Sens/Spec

- Bias & lack generalization
  - Diversity population, equipment, protocols, populations etc.
  - Test vs validation
  - Data imbalance, human oversight
- Pitfalls:
  - Errors
  - Alarm fatigue
  - Automation bias
    - AI overcall / AI miss, Complacency
- Continuous monitoring & Governance

fhberger@gmail.com

Narrow vs General AI

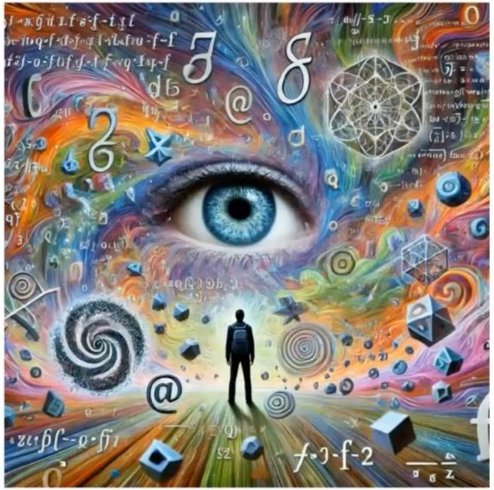
	Narrow	General
Level 0: No AI	calculator software; compiler	human-in-the-loop computing Mechanical Turk
Level 1: Emerging equal to or somewhat better than an un- skilled human	simple rule based systems 1970s	ChatGPT (OpenAI, 2023) Gemini (Pichai & Hassabis, 2023)
Level 2: Competent at least 50th percentile of skilled adults	Watson (IBM)	none
Level 3: Expert at least 90th percentile of skilled adults	Spelling & grammar checkers Dall-E 2 (Ramesh et al., 2022)	none
Level 4: Virtuoso at least 99th percentile of skilled adults	Deep Blue (Campbell et al., 2002) Al-phaGo (Silver et al., 2016; 2017)	none
Level 5: Superhuman outperforms 100% of humans	AlphaFold (Jumper et al.,2021) AlphaZero (Silver et al., 2018) StockFish (Stockfish, 2023)	none

Narrow = focussed task
General = like human mind

Courtesy: Prof. Masoom Haider, University of Toronto

Narrow vs General AI

Hallucinations and Reasoning



How many r's in elderberry?

- The word elderberry contains two 'r's.

You are hallucinating. The answer is wrong.

- You're right, my mistake. There is only one 'r' in "elderberry". Thanks for catching that.

2023 - ChatGPT v4
Level 1 general?

Courtesy: Prof. Masoom Haider, University of Toronto

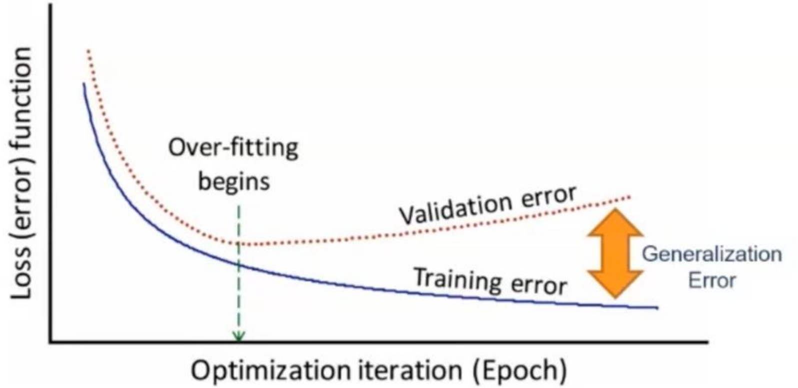
Local support

- Need buy-in:
  - radiologist / imaging colleagues
  - referring teams
  - administration: cost, philosophy, medicolegal aspects (who is liable?)
  - patient?

- IT support (department and institution):
  - Privacy issues – can patient information go to cloud?
    Or only local / regional / national?
  - Embedding in PACS and install on servers?
  - Patient records incorporated?

Validation AI tool – external

- AI tools work better on test data sets originating from training facility
- External validation with data sets from other institutions not often done, but crucial to test reliability of tool
- AI tools mostly perform less well on external data sets with different patient population



Razavi S, 2021 - Environmental Modelling & Software

Validation AI tool – FDA approval

FDA approval low bar?

Study on 151 FDA summaries of approved AI tools, lacking information:

Number of patients	in 54%	Sensitivity	in 29%
Patient demographics	in 4%	Specificity	in 27%
Geographical location	in 25%	Source reference standard	in 52%
Model / machine specifications	in 5.3%		

Khunte M et al. Cin Radiol 3023

fhberger@gmail.com

FDA Approval

Comment

https://doi.org/10.1038/s41591-024-03203-3

Not all AI health tools with regulatory authorization are clinically validated

Sammy Chouffani El Fassi, Adonis Abdullah, Ying Fang, Sarabesh Natarajan, Awab Bin Masroor, Naya Kayali, Simran Prakash & Gail E. Henderson

Check for updates

Devices that lack adequate clinical validation pose risks for patient care. A new validation standard is proposed to evaluate FDA authorization as an indication of clinical effectiveness in medical AI.

Advances in artificial intelligence (AI) are beginning to revolutionize healthcare. AI algorithms attempt various combinations of statistical equations to find patterns in data that solve real-world problems. AI-powered devices can detect cancers and strokes on radiology scans, accurately predict the onset of disease and dose insulin. However, the implementation of medical AI devices has led to concerns about patient harm, liability, patient privacy, device accuracy, scientific acceptability and lack of explainability, sometimes called the 'black box' problem[1–5].

These concerns underscore the importance of the validation of AI technologies. Patients and providers need a gold-standard indicator of efficacy and safety for medical AI devices. Such a standard would build public trust and increase the rate of device adoption by end users. As the chief legal regulatory body for medical devices in the USA, the Food and Drug Administration (FDA) currently authorizes AI software as medical devices (SaMD)[6]. However, for the public to accept FDA authorization

Table 1 | Classification of clinical validation methods for AI devices

Term	Definition
Clinical validation	Device tested with real patient data to evaluate safety and effectiveness
Prospective validation	Device tested after implementation in patient care and/or data collected after study begins
RCT	Experimental group that uses device and control group that does not use device are compared after randomized assignment
Retrospective validation	Device tested before implementation in patient care and/or data collected before study begins

patient care and thus provide stronger evidence for clinical validation. Randomized controlled trials (RCTs), a type of prospective study, use random assignment to control for confounding variables, thus isolating the therapeutic effect of the device[9]. Given the differing quality of scientific evidence generated by retrospective studies versus prospective studies, including RCTs, such distinctions should be made.

El Fassi SC et al. Nature Medicine 2024

JAMA Network | Open

Invited Commentary | Ethics

Discrepancies Between Clearance Summaries and Marketing Materials of Software-Enabled Medical Devices Cleared by the US Food and Drug Administration

Nigam H. Shah, MBBS, PhD; Michelle M. Mello, PhD, JD, MPhil

This study by Clark and colleague[1] examines discrepancies between statements made by developers of software-enabled medical devices in 510(k) applications for US Food and Drug Administration (FDA) clearance and statements subsequently made in marketing materials for the same devices. Among 119 recently cleared devices, approximately 1 in 8 were found to have marketing statements at odds with representations made in FDA applications, and another 7% were considered arguably discrepant. Discrepant cases had marketing materials that claimed or suggested that the device had artificial

+ Related article

Author affiliations and artic listed at the end of this artic

Shah NH, Mello MM. JAMA 2023

Validation AI tool – internal

C-spine fracture triage:
Discrepancy of local performance with FDA documentation

	N =	Prevalence	Accuracy	Sensitivity	Specificity	PPV	NPV
FDA	186	50%	---	**91.7%** (82.7 – 96.9%)	**88.6%** (81.2 – 93.8%)	**47.2%** (31.3 – 57.5%)	**99.0%** (98.3 – 99.8%)
Small	665	21.5%	**92.3%** (90.0 – 94.2%%)	**76.2%** (68.4 – 82.9%)	**96.7%** (94.8 – 89.1%)	**86.5%** (79.9 – 91.2%)	**93.7%** (91.7 – 95.2%)
Voter	1,904	9.1%	---	**54.9%** (45.7 – 63.9%)	**94.1%** (92.9 – 95.1%)	**38.7%** (33.1 – 44.7%)	**96.8%** (96.2 – 97.4%)

Small JE et al. Am J Neuroradiol. 2021
Voter AF et al. Am J Neuroradiol. 2021

Degradation AI tools

- Over time, performance of AI tool may decrease due to:
  - Change in patient profile / demographics
  - Change in practice patterns
  - Change in imaging equipment

Trusting automation

theweek.com

Home | Tech

FEATURES

8 drivers who blindly followed their GPS into disaster

Take note: The machine does not always know where it's going

Newsletter sign up

Image posted on Facebook by Mike Czarny

ndtv.com

telegraph.co.uk

News

Blindly following your car's GPS can be deadly

Of all the grisly rumours that you hear, the ill-fated GPS directions story is sadly one that is all too true

Lorraine Sommerfeld

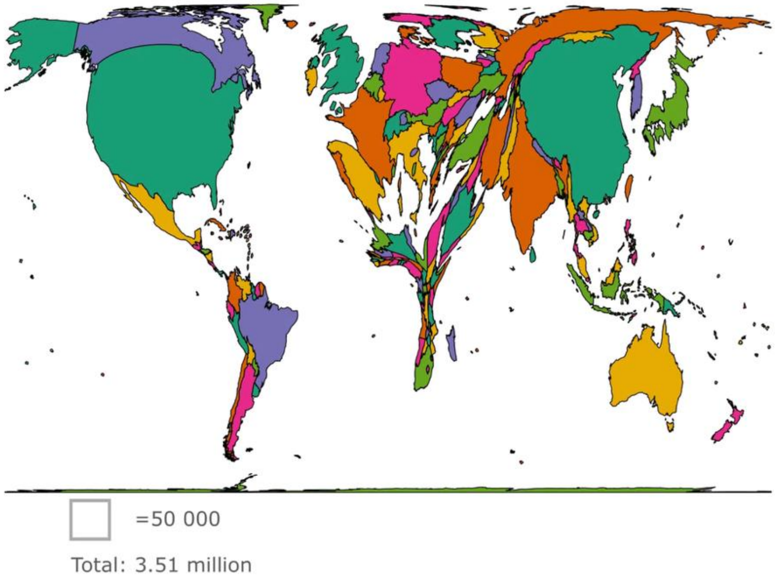
Published May 09, 2016 • Last updated Jun 12, 2020 • 5 minute read

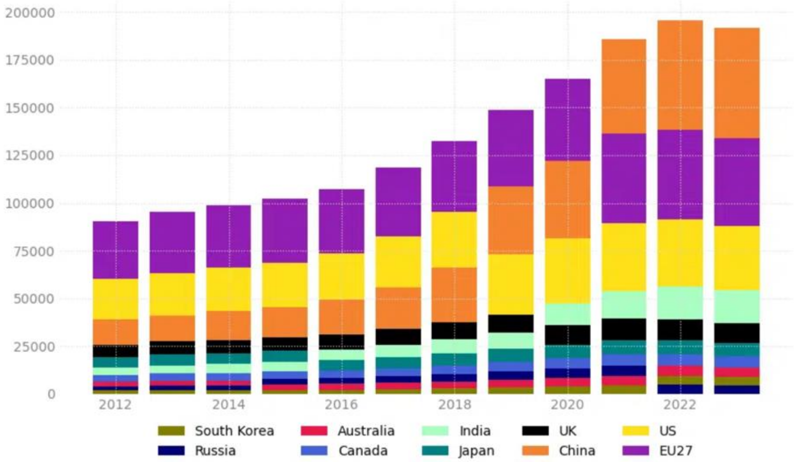
driving.ca

Potential bias AI tools

Distribution of AI Research Publications

Geographic distribution





https://www.digital-science.com/tldr/article/research-on-artificial-intelligence-the-global-divides/

AI research bias within USA

Table. US Patient Cohorts Used for Training Clinical Machine Learning Algorithms, by State[a]

States	No. of studies
California	22
Massachusetts	15
New York	14
Pennsylvania	5
Maryland	4
Colorado	2
Connecticut	2
New Hampshire	2
North Carolina	2
Indiana	1
Michigan	1
Minnesota	1
Ohio	1
Texas	1
Vermont	1
Wisconsin	1

3 states primarily contribute the AI publications in the USA

Kaushal A et al. JAMA 2020

Bias – hidden stratification


Author Manuscript

HHS Public Access

Author manuscript

Proc ACM Conf Health Inference Learn (2020). Author manuscript; available in PMC 2020 November 13.

Published in final edited form as:
Proc ACM Conf Health Inference Learn (2020). 2020 April ; 2020: 151–159. doi:10.1145/3368555.3384468.

Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging

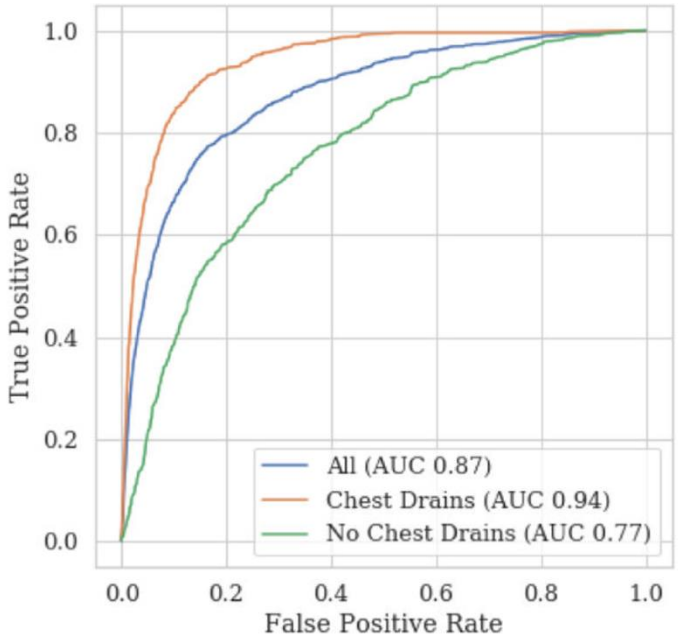
Luke Oakden-Rayner*,
Australian Institute for Machine Learning, University of Adelaide, Adelaide, Australia

Jared Dunnmon*,
Department of Computer Science, Stanford University, Stanford, California, USA

Gustavo Carneiro,
Australian Institute for Machine Learning, University of Adelaide, Adelaide, Australia

Christopher Ré
Department of Computer Science, Stanford University, Stanford, California, USA

Author Manus



Pneumothorax detection better since inclusion of CXRs with presence of chest tube

Automation Bias

3 major errors of automation bias:

- *Commission* (AI overcall)
  - AI says Pos, but is Neg
  - Rad reports Abnormal
- *Omission* (AI miss)
  - AI says Neg, but is Pos
  - Rad reports Normal
- *Complacency*

  as time goes on, too much trust...

Automation Bias

Users with greater trust in automation are less likely to detect issues

Excessive trust over time

Greater dependence when:

- High workload
- difficult tasks
- multi-tasking

= Radiologist

Issues: possible missed diagnoses, erosion of expertise, false security

Conversely – Nay-sayers

- Preference for human interpretation
- Averse to trusting algorithm
- One error will exaggerate distrust
- Think lesser of people using AI

Solutions for AI bias and distrust

- Start with pilot
- Balance AI and human input
- Prevent blind signing of AI results
- Continuous training
- Foster critical thinking
- Have rounds or other feedback loop

Governance best practice

- Have a multidisciplinary team of Radiologists, IT and admin
- Develop policies for AI use (ethical, legal, operational risks)
- Implement protocols for continuous monitoring and auditing
- Establish feedback loop for AI performance and improvements

Key considerations

- Evaluate AI tools with *local* data
- Expect decreased performance compared to promise vendor / FDA doc
- Understand the issues around biases, including automation bias
- Monitor performance after deployment

fhberger@gmail.com