Utilizing AI to Enhance Health Care Quality & Safety for Diagnosis

Promise & Perils

David E. Newman-Toker, MD, PhD

Professor of Neurology, Otolaryngology, & Emergency Medicine
Professor of Epidemiology and Health Policy & Management
Director, Armstrong Institute Center for Diagnostic Excellence
Johns Hopkins Medicine & Bloomberg School of Public Health

AI for Clinical Diagnosis – Disclosures for Dr. Newman-Toker

1. **Grant & Contract Support…**
   - **US Federal Grants & Contracts**: NIH (NIDCD U01 DC013778, NINDS U01NS080824, NCATS U24 TR001609), AHRQ (R18 HS026640, R01HS27614, EPC 503-4262, R18 HS029350)
   - **US Foundation Grants & Contracts**: Gordon & Betty Moore Foundation, American Heart Association, Coverys Foundation, AARP, SIDM
   - **US Industry Grants & Contracts**: Natus-Otometrics
2. **Equipment Support (research video-oculography [VOG] devices)…**
   - Autronics-Interacoustics
   - Natus-Otometrics (licensing JHU decision support technology, related research grant as principal investigator)
3. **Related Roles…**
   - Past President / Former Board Member, Society to Improve Diagnosis in Medicine (SIDM) (unpaid)
   - Director, Johns Hopkins Armstrong Institute Center for Diagnostic Excellence (salary support for effort)
4. **Career focus on 'Diagnosis' (academic conflict of interest)**

AI for Clinical Diagnosis – Learning Objectives

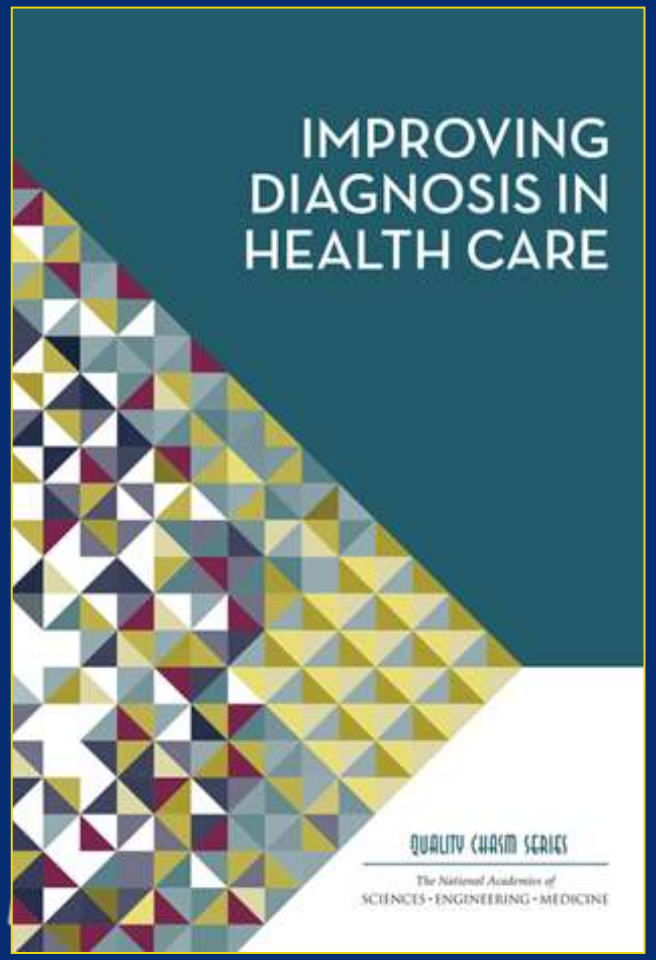
1. Summarize public health burden & financial impact of diagnostic errors and misdiagnosis-related harms.
2. Discuss potential pitfalls of applying artificial intelligence (AI) for clinical diagnosis without adequate guardrails.
3. Describe prerequisites and systems of care essential to deploying AI to achieve diagnostic excellence.

AI for Clinical Diagnosis – Lecture Outline

1. Diagnostic Errors & Harms
2. AI for Clinical Diagnosis
3. The Path of Diagnostic Decay
4. The Path of Diagnostic Excellence
5. Key Takeaways
6. Questions & Answers

Diagnostic Errors & Harms

Improving Diagnosis in Healthcare



"The delivery of healthcare has proceeded for decades with a blind spot: Diagnostic errors…"

"…most people will experience at least one diagnostic error in their lifetime, sometimes with devastating consequences."

"Improving the diagnostic process is not only possible, but it also represents a moral, professional, and public health imperative."

National Academy of Medicine, Improving Diagnosis in Healthcare 2015

Diagnostic Errors – Public Health Imperative

Most Common
Most Catastrophic
Most Costly

Newman-Toker et al., BMJ Quality & Safety 2023

Diagnostic Errors – "Big Three" Causes of Serious Harm

Vascular

Infection

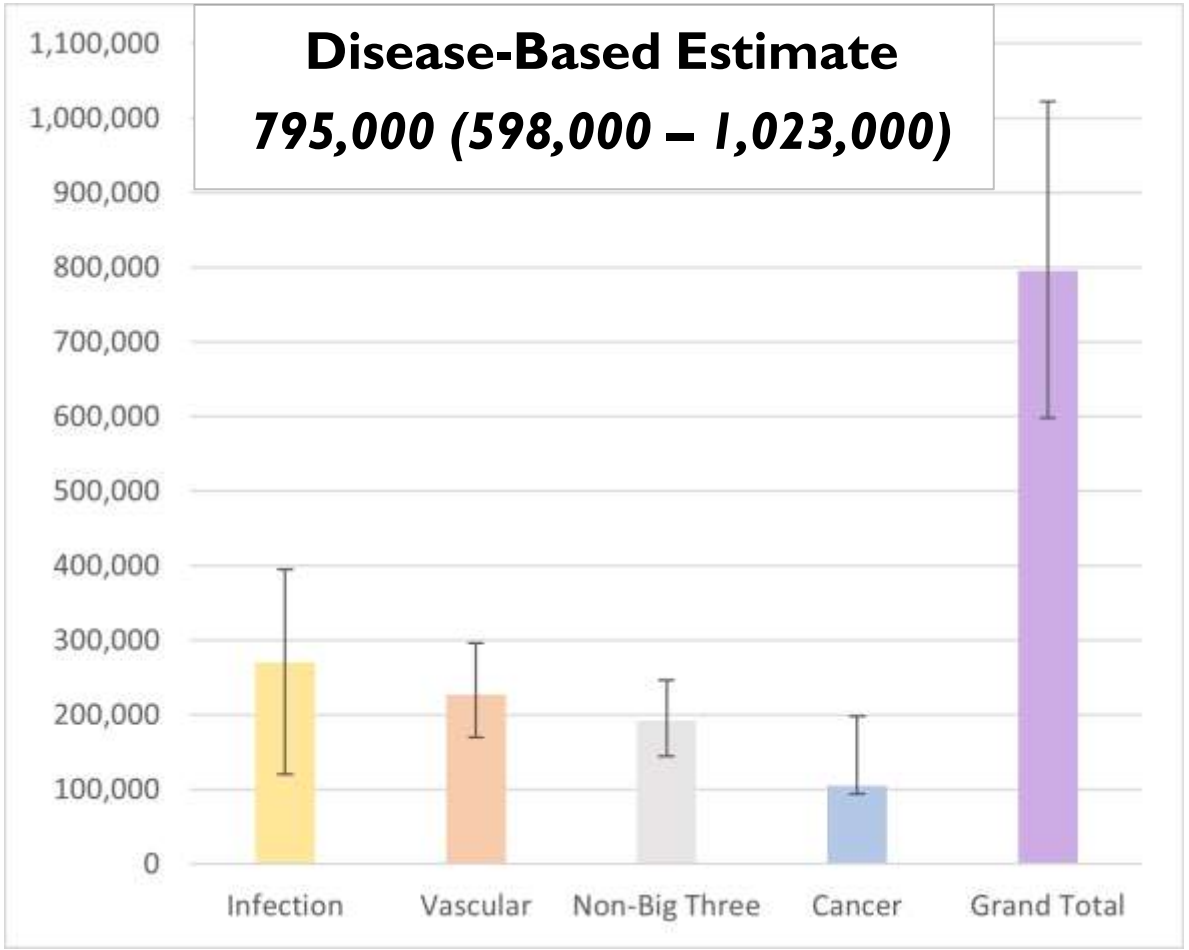
Cancer

Our prior work showed that the "Big Three" account for 75% of serious harms in both malpractice claims & clinical studies of diagnostic error.

And that 34% of serious harms in the ED are attributable to missed neurologic emergencies

Newman-Toker et al., Diagnosis (Berl.) 2018; Newman-Toker et al., AHRQ 2022

Total Serious Misdiagnosis-Related Harms in the U.S. per Year



Serious Harms ~795,000/yr

- ~425,000 disabilities
- ~370,000 deaths

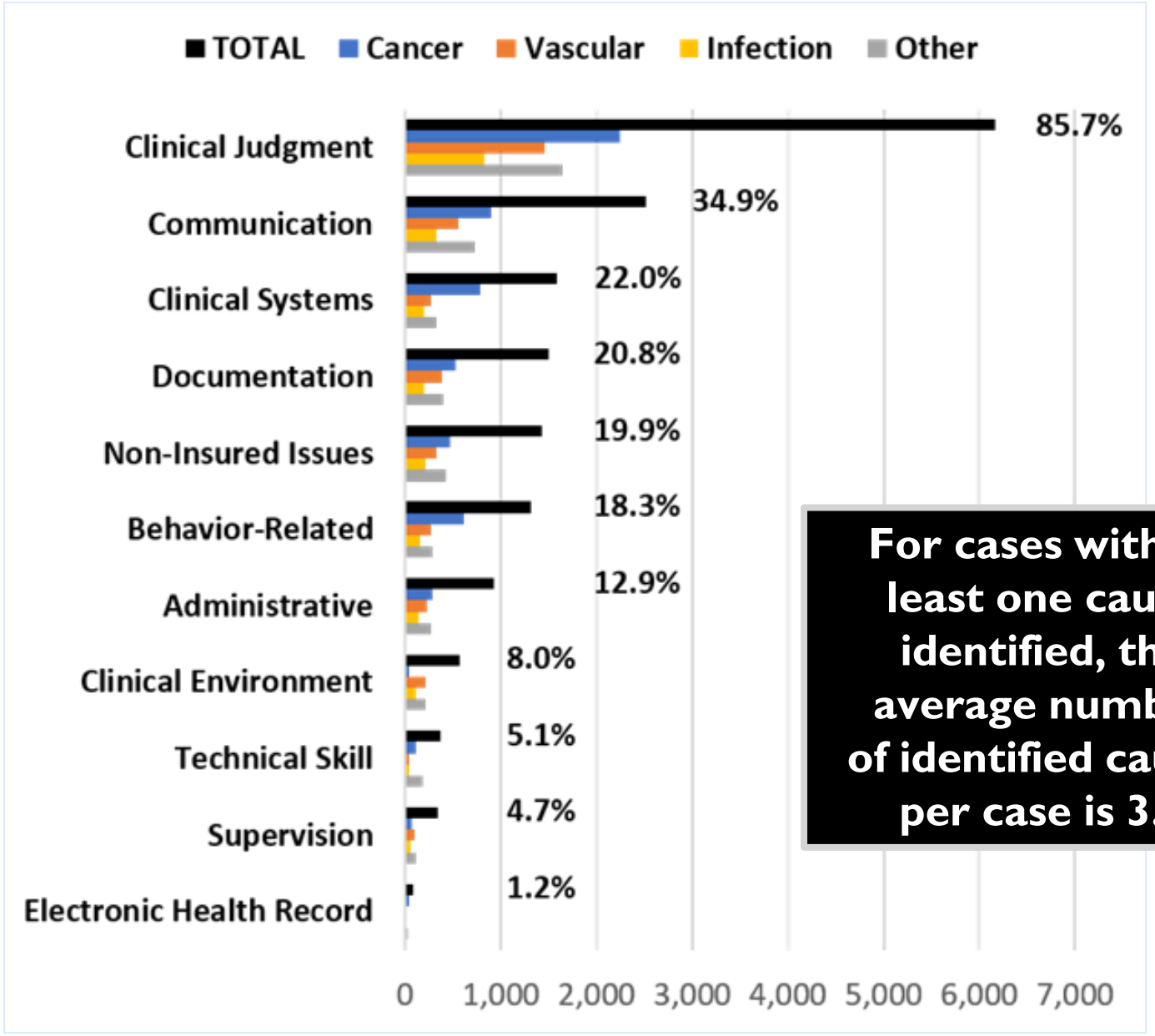
Serious Harms Breakdown

- 34% Infection – 271,000
- 29% Vascular – 228,000
- 24% Non-Big 3 – 192,000
- 13% Cancer – 105,000

N.B. – Data are for U.S. in 2014, so likely represent ~5% underestimates for 2024

Newman-Toker et al., BMJ Quality & Safety 2023

Top Causes of Diagnostic Errors



For cases with at least one cause identified, the average number of identified causes per case is 3.7

Newman-Toker et al., Diagnosis (Berl.) 2019

Diagnostic Errors – Subtlety is a Huge Risk Factor

Obviousness predicts correct diagnosis

Subtlety predicts incorrect diagnosis

Points to Gaps in Expertise

a) low prevalence (pre-test probability / base rate)
b) degree of difficulty (atypical, non-specific, red herrings, "wrong" demographic group, bigger problems)
c) training background, knowledge/familiarity/expertise

Newman-Toker et al., AHRQ Report on Diagnostic Errors in the Emergency Department 2022

AI for Clinical Diagnosis

Artificial Intelligence Defined

Artificial Intelligence is the branch of computer science concerned with endowing computers with the ability to simulate intelligent human behavior.

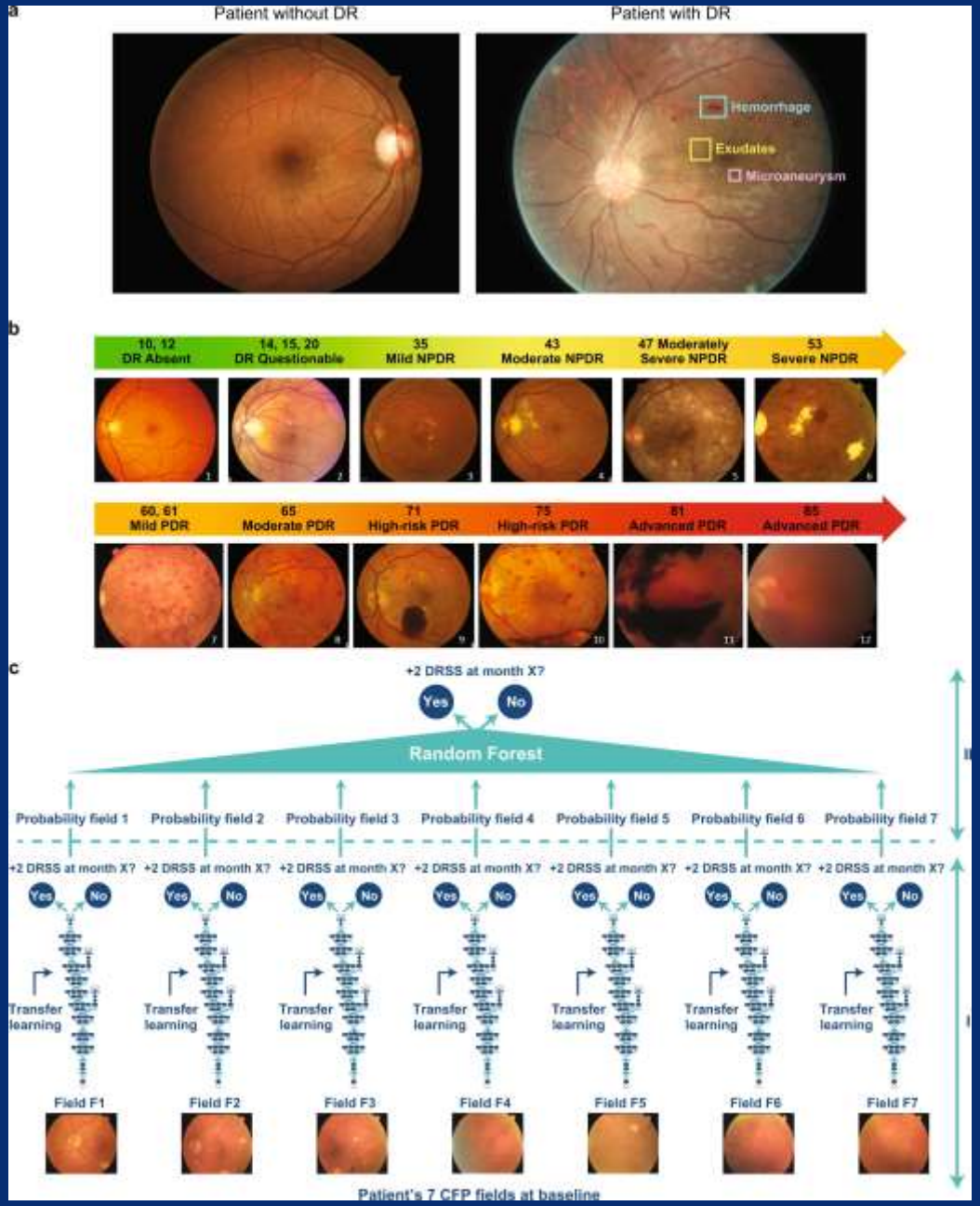
Shortliffe & Cimino. Biomedical informatics: computer applications in health care and biomedicine, 2006

AI for Clinical Diagnosis – "The Holy Grail"

The most complex cognitive task in medicine is the act of diagnosing the cause of a patient's symptoms.

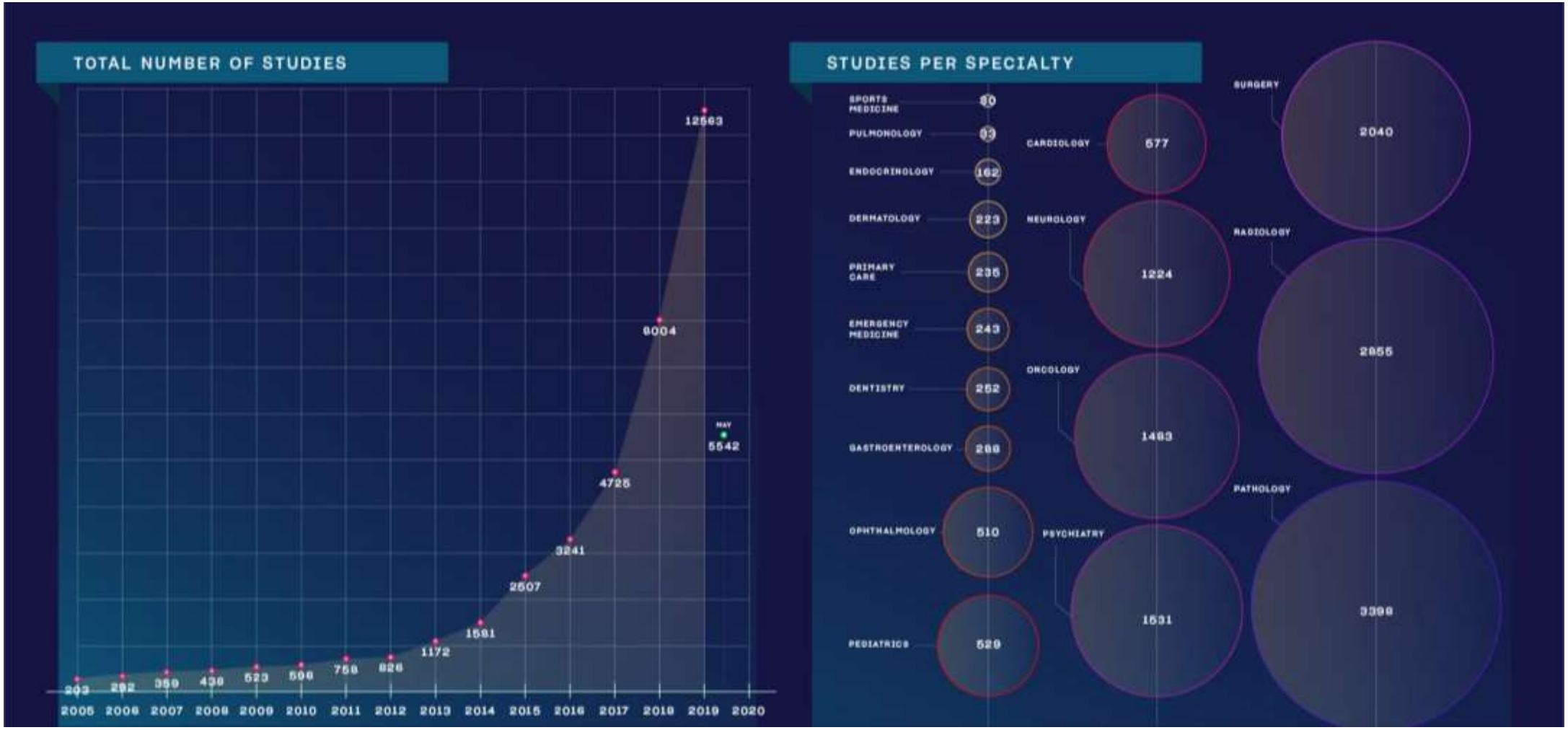
Newman-Toker & Sharfstein, JAMA Health Forum, 2024

AI for Diagnosis – Promise



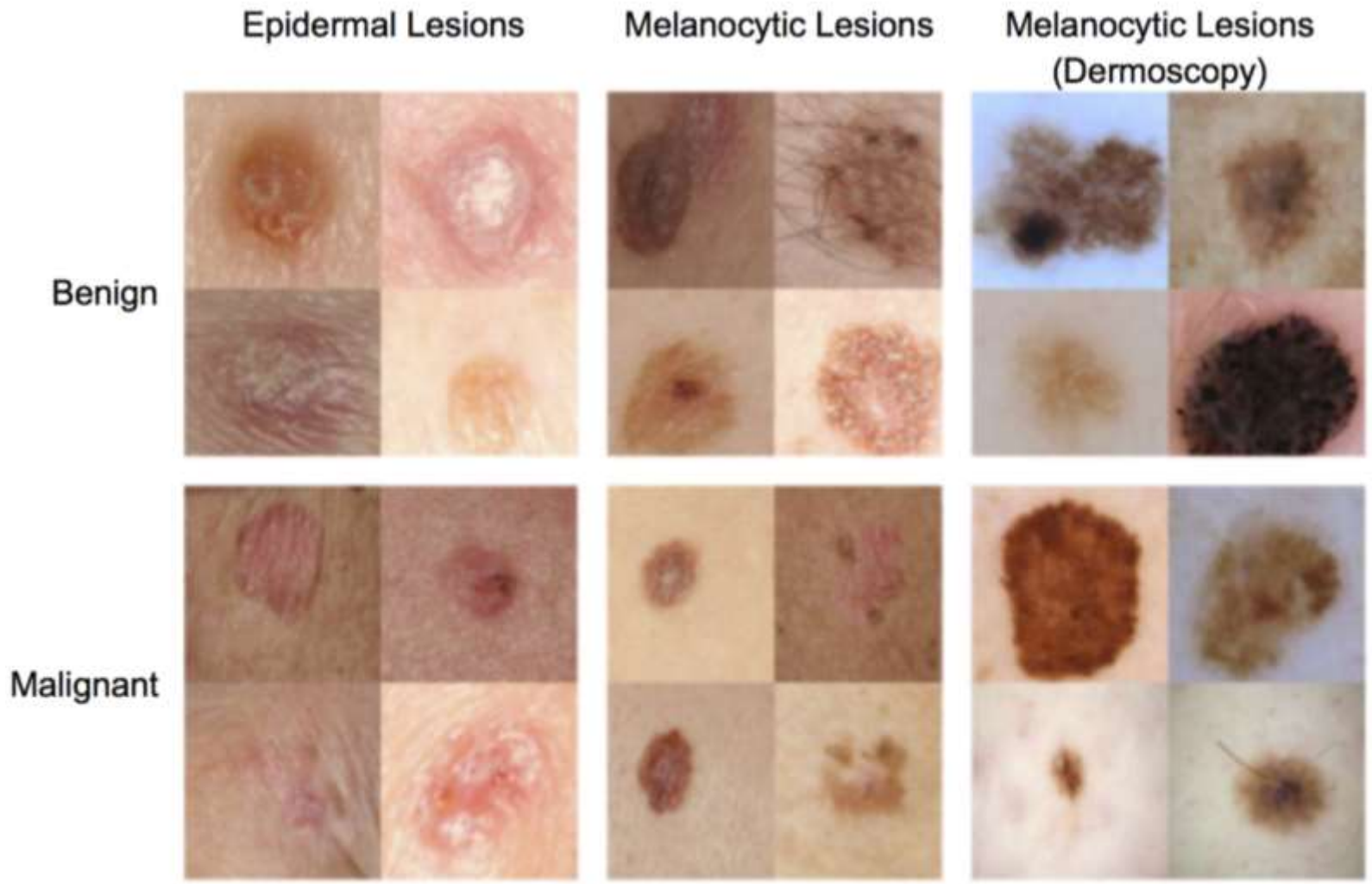
Arcadu, Nature Digital Medicine 2019

Medical A.I. Studies by Specialty



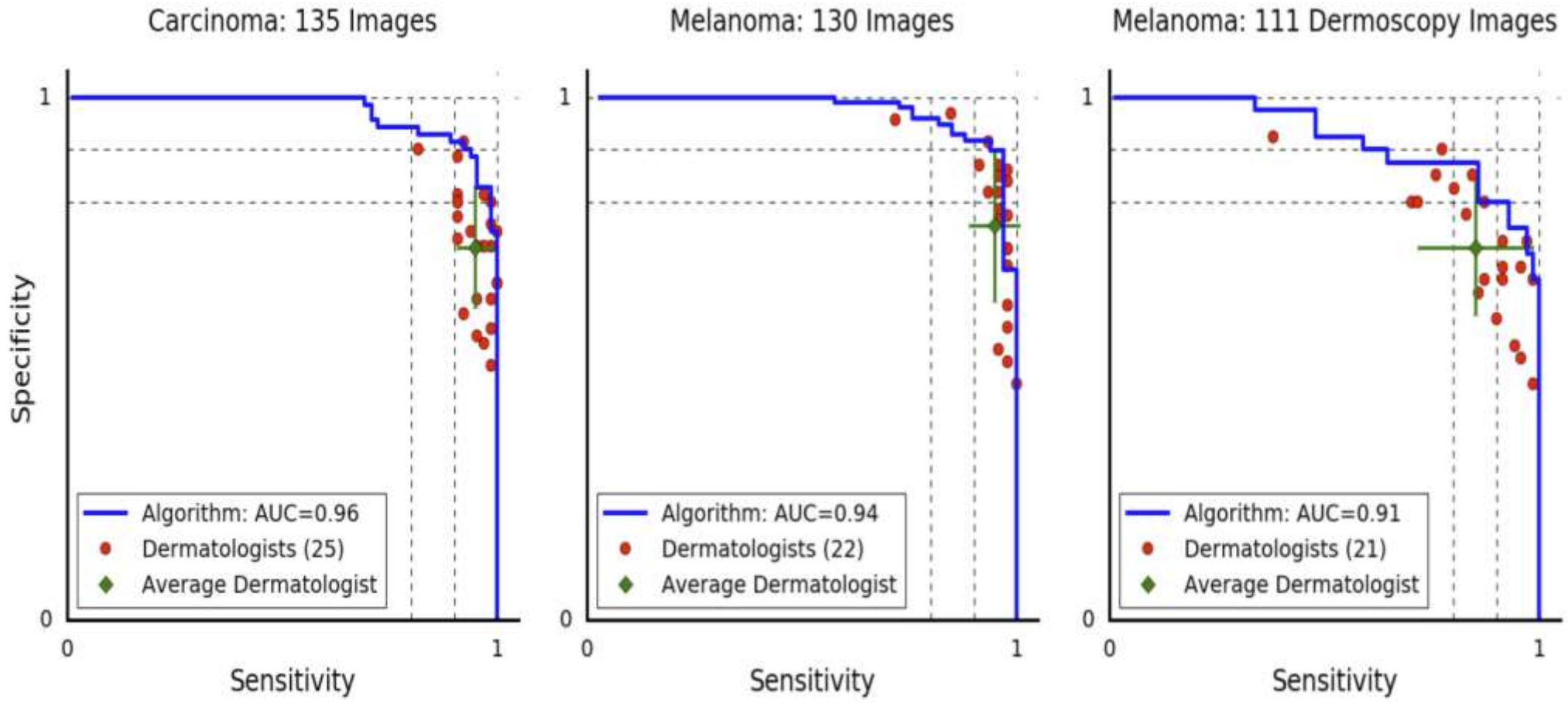
Meskó & Görög, Nature Digital Medicine, 2020

Convolutional Neural Networks vs. Dermatologists



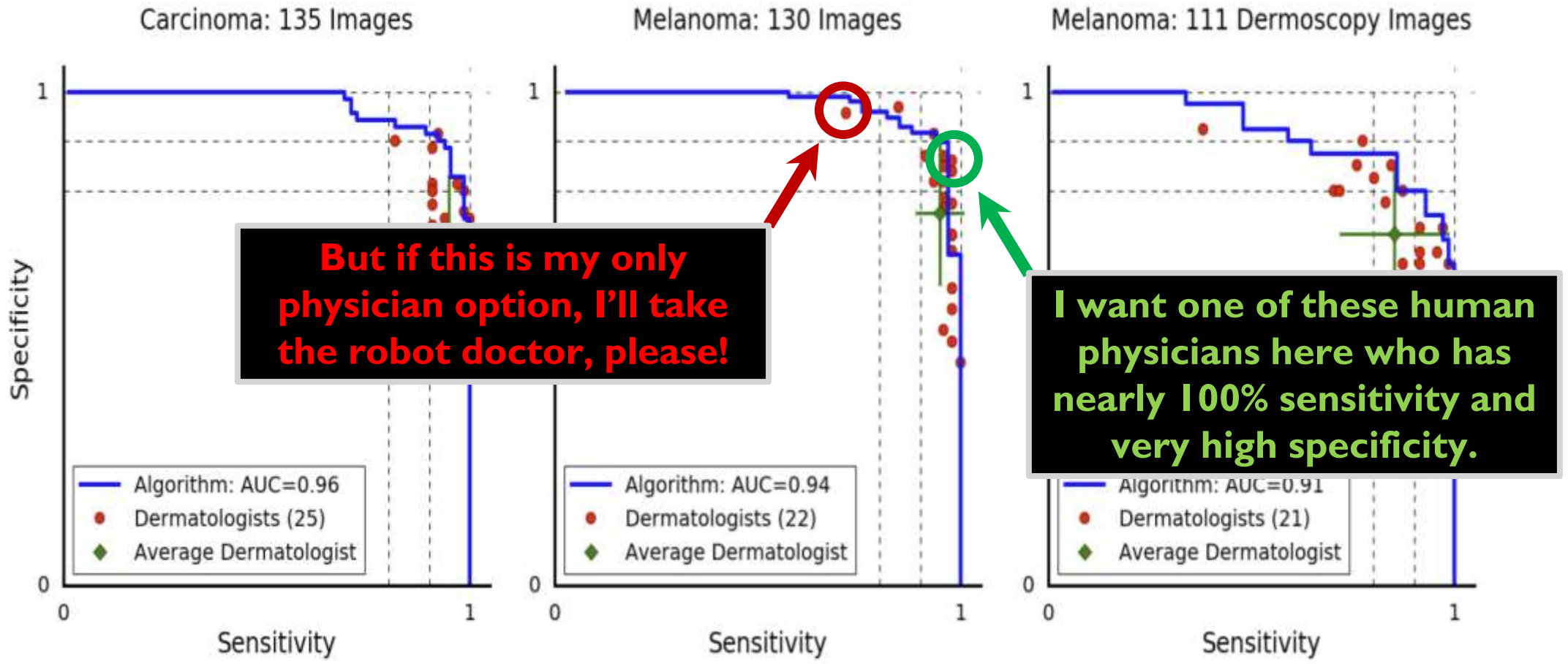
Esteva et al., Nature, 2017

Convolutional Neural Networks vs. Dermatologists



Esteva et al., Nature, 2017

Convolutional Neural Networks vs. Dermatologists



Esteva et al., Nature, 2017

AI for Diagnosis – Perils

How IBM's Watson Went From the Future of Health Care to Sold Off for Parts

BY LIZZIE O'LEARY

JAN 31, 2022 • 9:00 AM

https://slate.com/technology/2022/01/ibm-watson-health-failure-artificial-intelligence.html

The 'GIGO' Principle

Garbage In…

Garbage Out

https://www.cliparts101.com/free_clipart/59231/Pig_in_Trash_Can

"Looking Where the Light Is" for Data to Train Algorithms

Many in tech sector believe in "BIG DATA" over "RIGHT DATA"

https://twitter.com/ChenxinLi2/status/1769403046805561500

Spectrum Bias & Tail Comparisons in Algorithm Development

One cause for unhealthy initial optimism for AI-ML algorithms

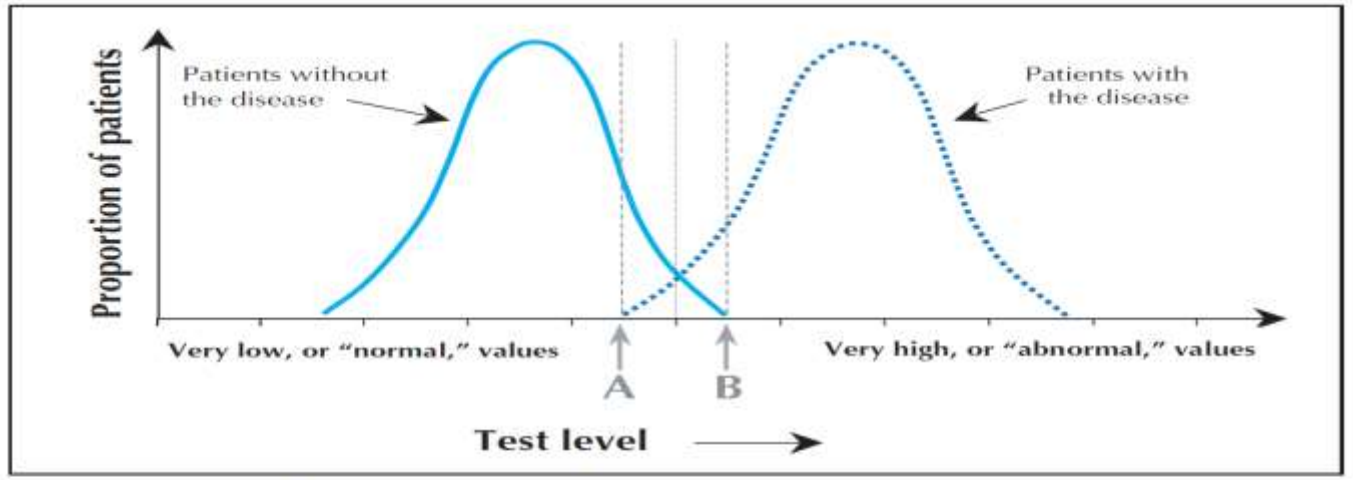


Fig. 1: Spectrum bias. Test performance when differentiating normal volunteers (disease-negative) from severely ill patients (disease-positive). For test results above point B, all patients have the condition, and for all test results below point A, no patients have the condition. The distance between A and B shows the extent of the overlap of test results between the 2 groups.

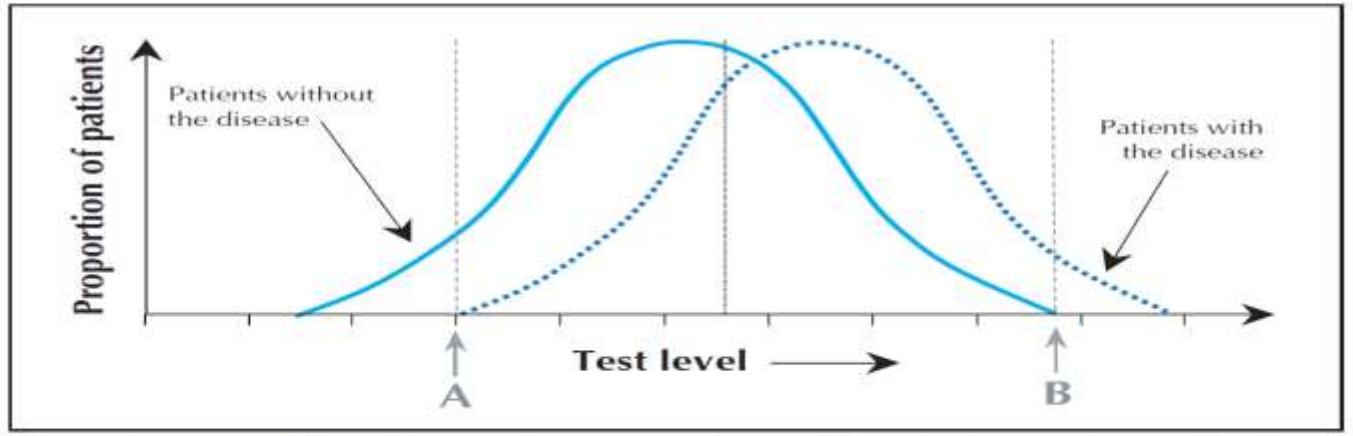


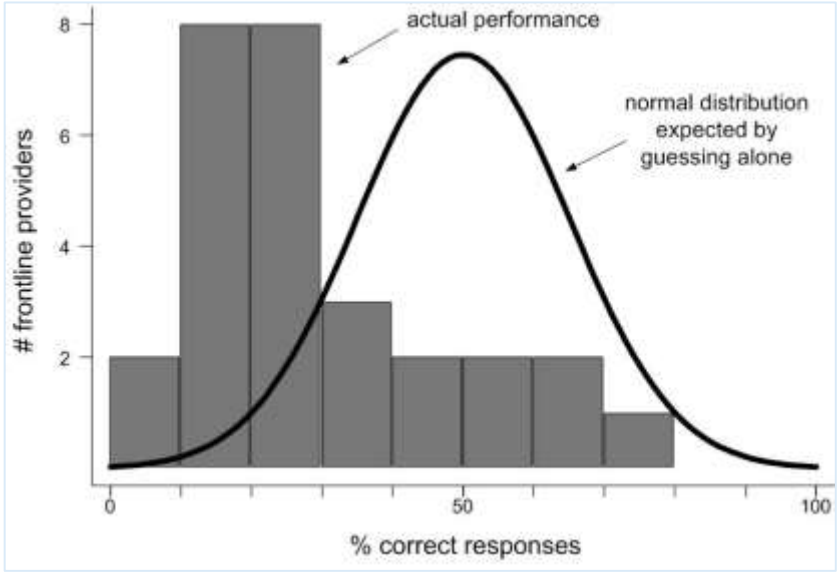
Fig. 2: Spectrum bias. Test performance when differentiating patients who have the disease from those who don't when both groups appear to have the target disease. The distance between A and B is now much wider.

Montori et al., CMAJ 2005

The Edge Case Problem for Diagnostic Errors

Data are selectively missing or biased in diagnostic error cases

Frontline providers harbor misconceptions about the bedside evaluation of dizzy patients



Gaps in Neuro-Vestibular Eye Movement Assessment by Clinical Neurologists in the AVERT Randomized Trial

Results

Neurology consultations were obtained in 53.1% (n=69/130).Complete documentation was infrequent (HINTS 28.9%, positional 4.3%) and complete, accurate documentation rare (HINTS 5.7%, positional 0.0%). Missing HINTS documentation included the head impulse test (HIT) in 49.2% (28.5% had abnormal HIT), nystagmus in 27.5% (57.8% had pathologic nystagmus), and test of skew in 62.3% (1.6% had skew present). Missing positional exam documentation (95.6%) included numerous cases where positional nystagmus was present by VOG (Dix-Hallpike 41.4%, supine roll 43.9%). We observed mismatched findings in 37.1% of HIT, 22% of nystagmus, 11.5% of skew, 35.7% of Dix-Hallpike, and 33.7% of supine roll exams documented by neurologists.

Newman-Toker et al., Acta Otolaryngol 2008; Bastani et al., submitted to Bárány 2024

The Path of Diagnostic Decay

AI for Clinical Diagnosis – The Path of Diagnostic Decay

This path takes us towards a future in which AI replicates human diagnosis at slightly (?) lower accuracy/quality and eliminates the need for many physicians (or replaces them with those of lesser skill and lower wage), making healthcare cheaper but worse for patients.

This is a future in which AI systems are deployed in a way that deskills clinicians, making it impossible for humans to fathom (or the system to know) whether the AI diagnoses are correct or not. This future will be easy to achieve but is probably not a future we want.

AI for Clinical Diagnosis – The Path of Diagnostic Decay

- Diagnostic AI penetrates consumer healthcare via "toys" with hard-to-find legal disclaimers (e.g., symptom checkers).
- Diagnostic AI penetrates institutional healthcare via unregulated "workflow" tools, where diagnosis is inadvertently impacted.
- Diagnostic AI penetrates "quality improvement" processes under minimal oversight with the intent to "gradually improve care" using AI algorithms, but plateau due to lack of "RIGHT data" streams.
- Clinicians become progressively reliant & deskilled.

The Path of Diagnostic Excellence

AI for Clinical Diagnosis – The Path of Diagnostic Excellence

This path takes us towards a future in which AI helps us achieve a world where diagnoses are accurate, timely, and effectively communicated to patients, avoiding both diagnostic error and overdiagnosis… a future in which diagnostic processes are patient-centered, evidence-based, prompt, efficient, safe, and equitable.

This will only be achieved by deliberate attempts to create well-designed, trustworthy AI systems. More specifically, it requires focusing major efforts on creating the necessary gold-standard data sets to train AI systems for clinical diagnosis, conducting high-quality research studies on diagnostic outcomes (not merely accuracy), and deploying monitoring/learning health system programs that are ever vigilant about accuracy, health outcomes, and bias/discrimination. This future also ensures that human diagnostic skills are bolstered rather than decaying.

AI for Clinical Diagnosis – Rigorous Training Data Essential

- Datasets for "visual diagnosis" based solely on interpretation of medical images (as in radiology, ophthalmology, and dermatology) are already available or being developed and systems trained.
- However, comparable initiatives for the bulk of clinical diagnostic encounters across care settings are either nascent or do not exist.
- Training diagnostically accurate AI systems requires high quality data at the front end (patient demographics, symptoms, signs, and laboratory and radiographic findings) and back end (accurate final diagnoses, treatment effects, and morbid or mortal outcomes).

Newman-Toker & Sharfstein, JAMA Health Forum, 2024

AI for Clinical Diagnosis – Success Factors for AI Tools

- Solve important diagnostic problems
- Build modular, context-specific solutions
- Frame tools around symptom-disease pairs
- Emphasize ergonomics, clinical workflow
- Tackle adaptive barriers head on, early
- Remember the 'fundamental theorem'
  - Human & computer in the proper roles
  - Human & computer in the proper relationship

AI for Diagnosis is more about Diagnosis than it is about AI.

The tech is plenty good already. It's the rest we need to get right.

AI for Clinical Diagnosis – Friedman's Fundamental Theorem

Figure 1. A "Fundamental Theorem" of informatics.

NOT…

Figure 2. What informatics is not.

Friedman, JAMIA 2009

Veteran's Admin Framework for Trustworthy AI

1, 2, & 3 are mission critical for improving diagnosis using AI

Veteran's Administration (https://department.va.gov/ai/trustworthy-ai/)

Key Takeaways

Key Takeaways – Diagnostic Errors & Harms

- An estimated 800,000 Americans suffer death or permanent disability each year from diagnostic errors at an estimated societal cost of more than $200 billion.
- Three disease categories (vascular events, infections, and cancers) account for 75% of the serious harms from diagnostic error. Missed stroke causes the most harm across clinical settings.
- Most diagnostic errors are associated with issues in bedside diagnostic reasoning and appear to disproportionately reflect failures of expertise, which has the potential to be addressed by AI.

Key Takeaways – The Path of Diagnostic Excellence

- The next decade must focus on constructing gold standard data sets for diagnosis—the promise of AI will not be realized without quantifying bedside evaluations in a symptom-oriented framework.
- AI systems must be held to a high diagnostic standard—they must be demonstrated scientifically to improve safety and quality over current care and then monitored closely over time.
- The impact of AI on human clinical diagnostic skills must be monitored and managed—clinical deployment of AI should be explicitly designed to enhance rather than reduce clinician skills by applying educational and human factors science (e.g., AI-human peer review/competition).