Artificial Intelligence in Health Insurance and Care Delivery

Keisuke Nakagawa, M.D.
Director of Strategic Impact, Jacobs Center for Health Innovation, UC San Diego Health

Christina Silcox, PhD
Research Director, Duke-Margolis Institute for Health Policy

July 15, 2025

AI in Healthcare

Opportunities and Responsible Innovation

Keisuke Nakagawa, M.D.

Director of Strategic Impact, Jacobs Center for Health Innovation,
UC San Diego Health

> The **AI you're using today** will be the **dumbest AI** you'll use in your lifetime.

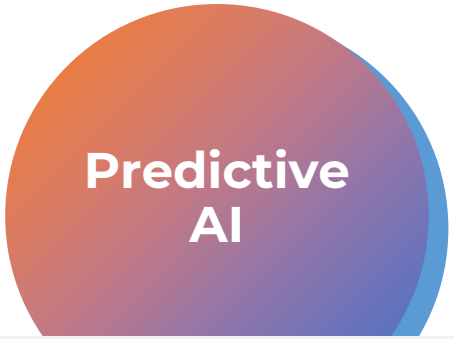
– **Sam Altman**
OpenAI

Why AI?

We want AI to tell us
what we would already know
if we were at everyone's bedside and in their chart all at once.

- We cannot be at every patient's bedside
- We cannot review every chart
- We cannot intervene on every patient

Broad Classes of Health AI

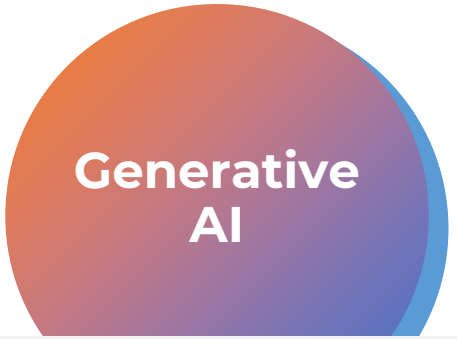


Usually trained with our data

Used to predict risk of events/counts:

- Risk of sepsis in next 6 hrs
- Hospital census

Appropriate when: Need information to make a decision (that re-allocates resources)



Usually not trained with our data

Used to generate text, images, audio

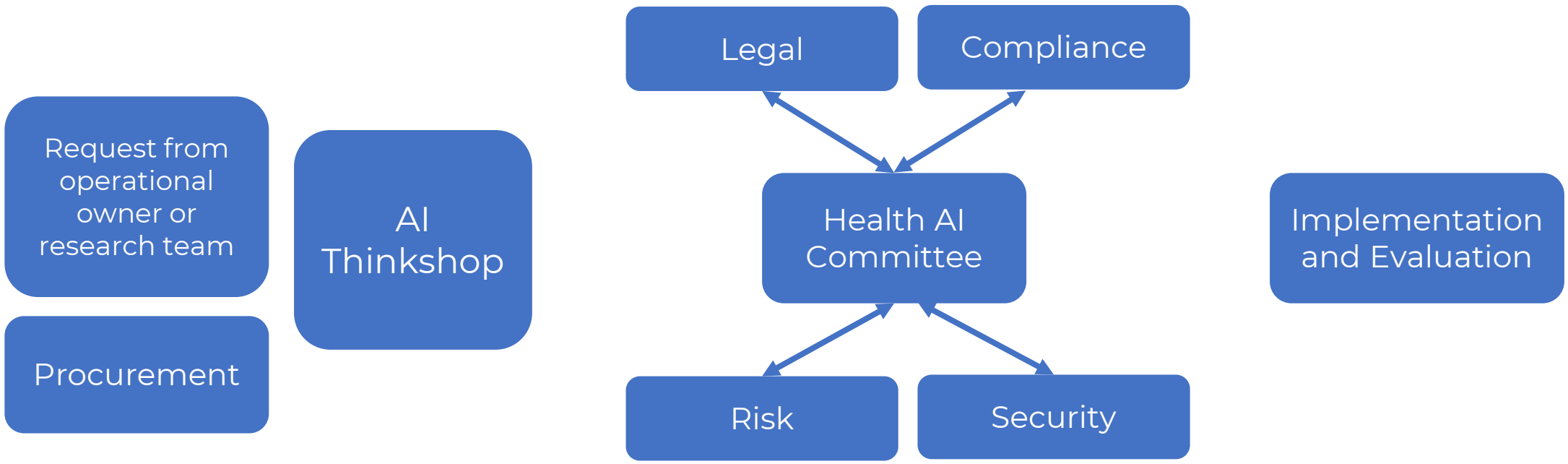
- Large language models (LLMs)
- Vision-language models

Appropriate when: Need to extract, summarize information, or answer questions.

Different Types of AI Technologies in Healthcare

Type	Description	Healthcare Example
Machine Learning (ML)	Learns patterns from historical data to make predictions	Predicting risk of hospital readmission or sepsis
Natural Language Processing (NLP)	Interprets human language	Extracting structured data from clinical notes (unstructured)
Computer Vision (CV)	Interprets visual information (images, videos)	Detecting lung nodules on CT scans
Large Language Models (LLMs)	Trained on massive corpora to generate human-like text	Summarizing chart notes for clinicians; Pre-drafting messages to patients
Multi-Modal	Processes multiple types of data (e.g., text, images, vitals)	AI surgical monitoring combining video, sensor data, and transcripts

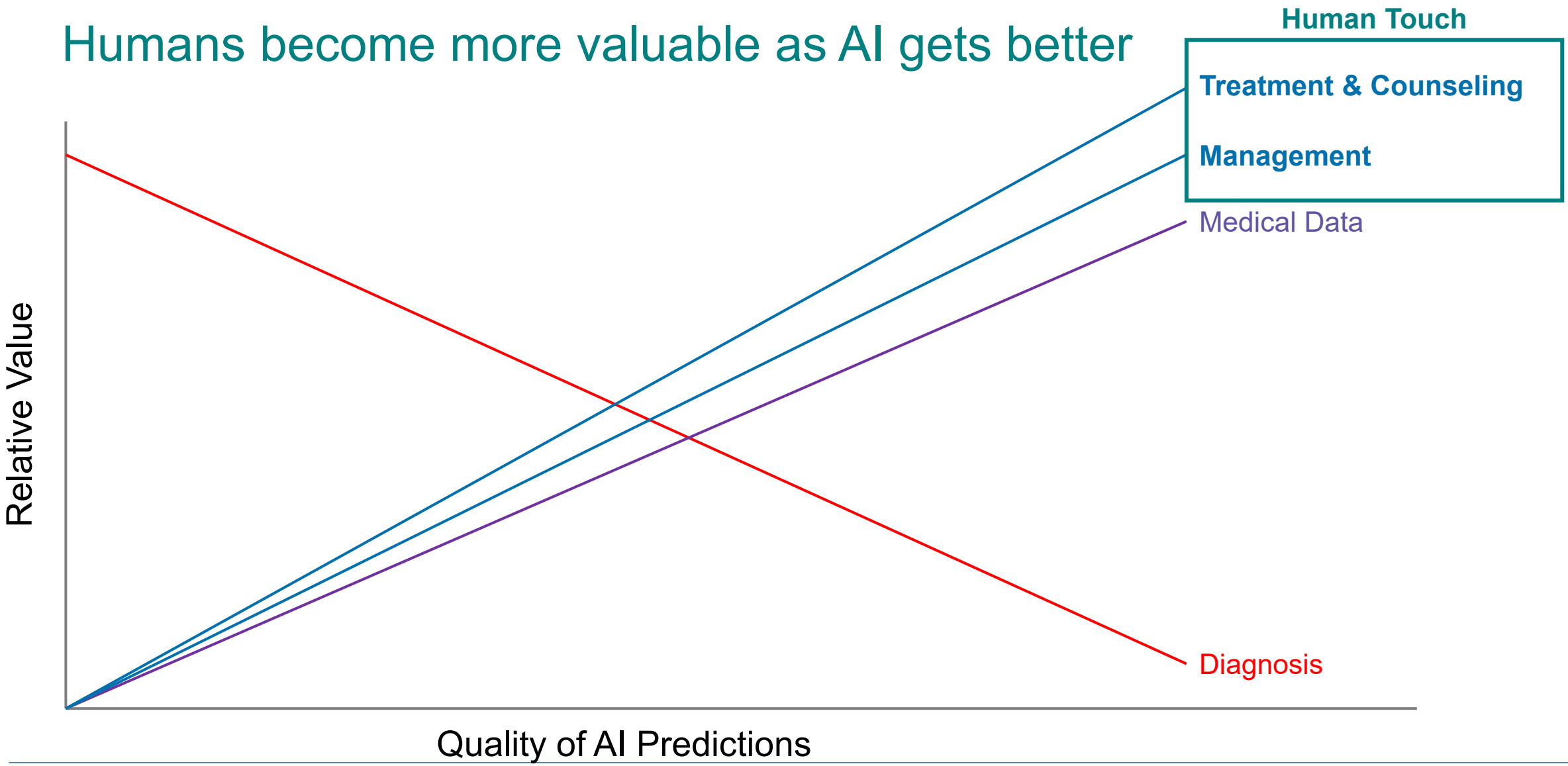
AI Governance: Could we? Should we?



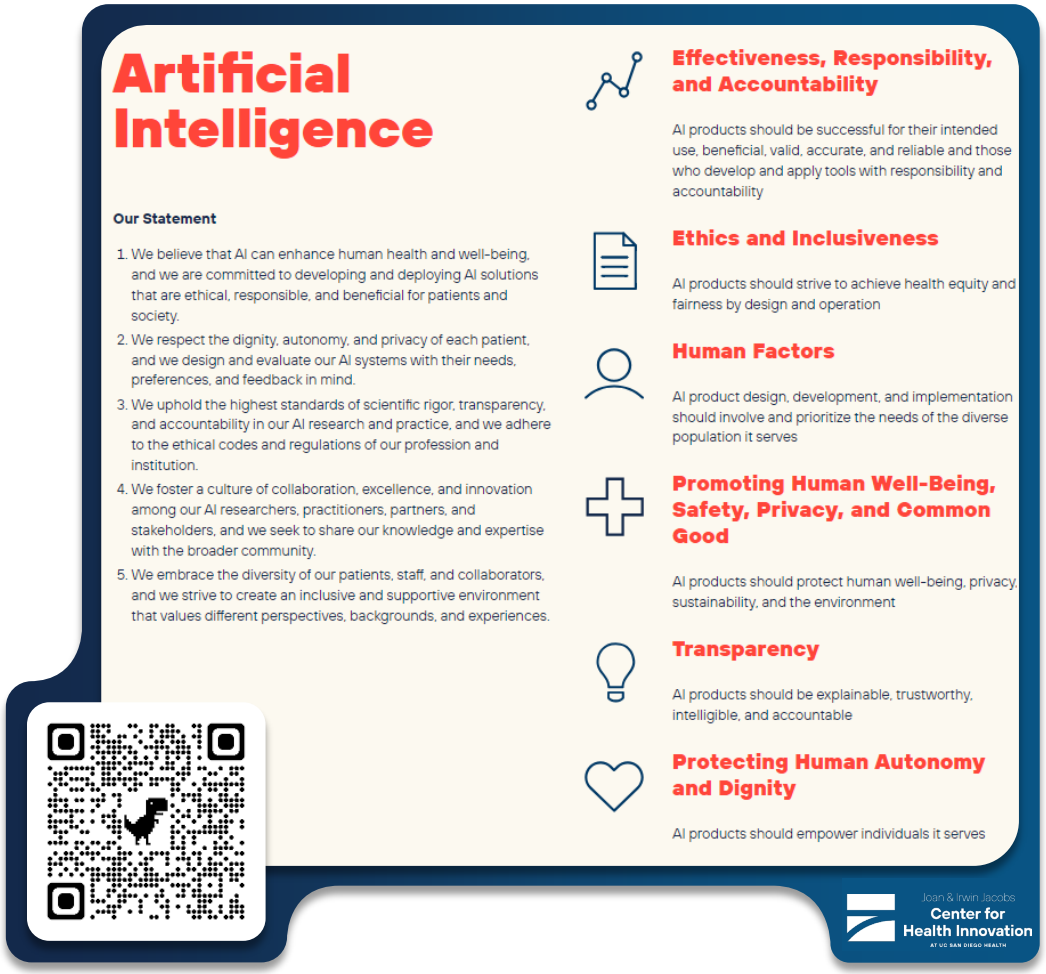
Is it feasible?

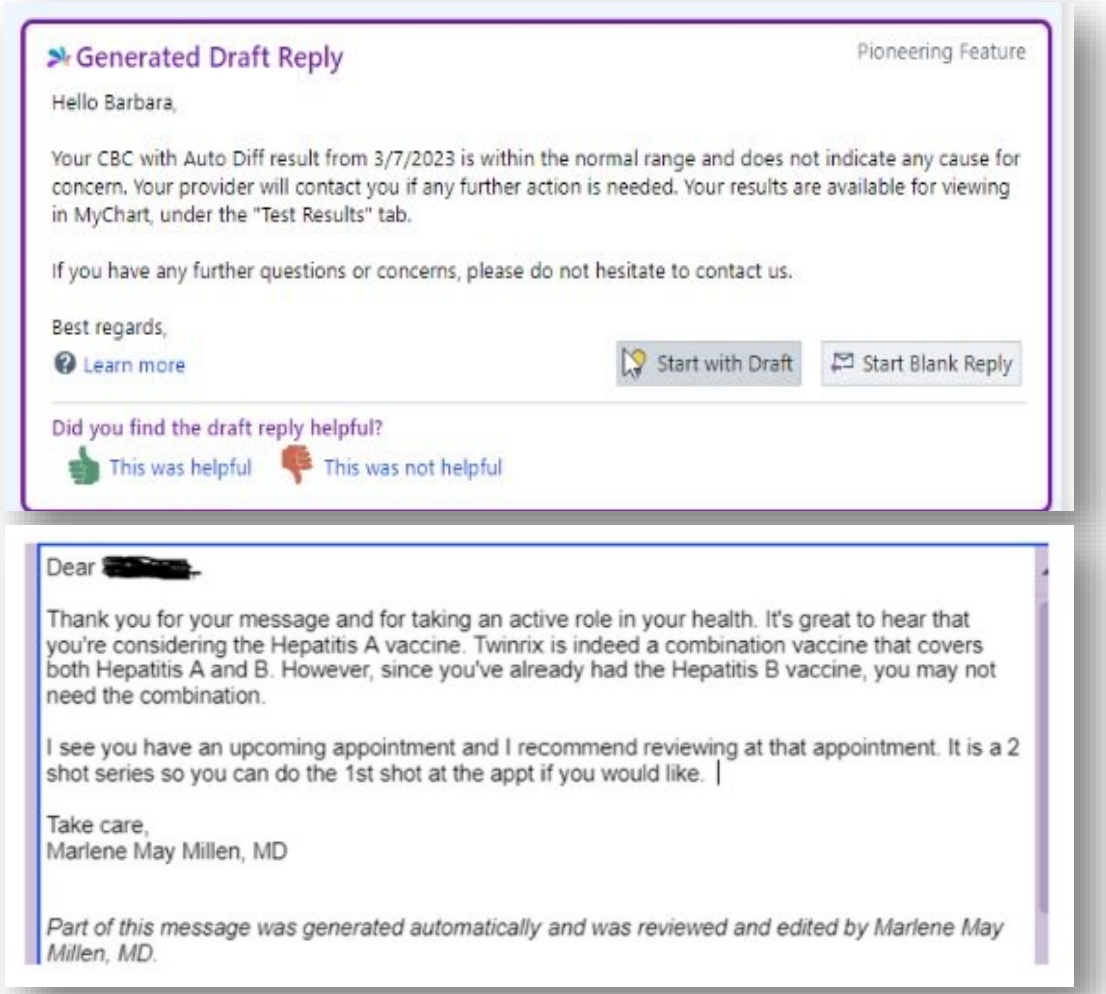
Is it the right thing to do? What safeguards do we need?

Humans become more valuable as AI gets better

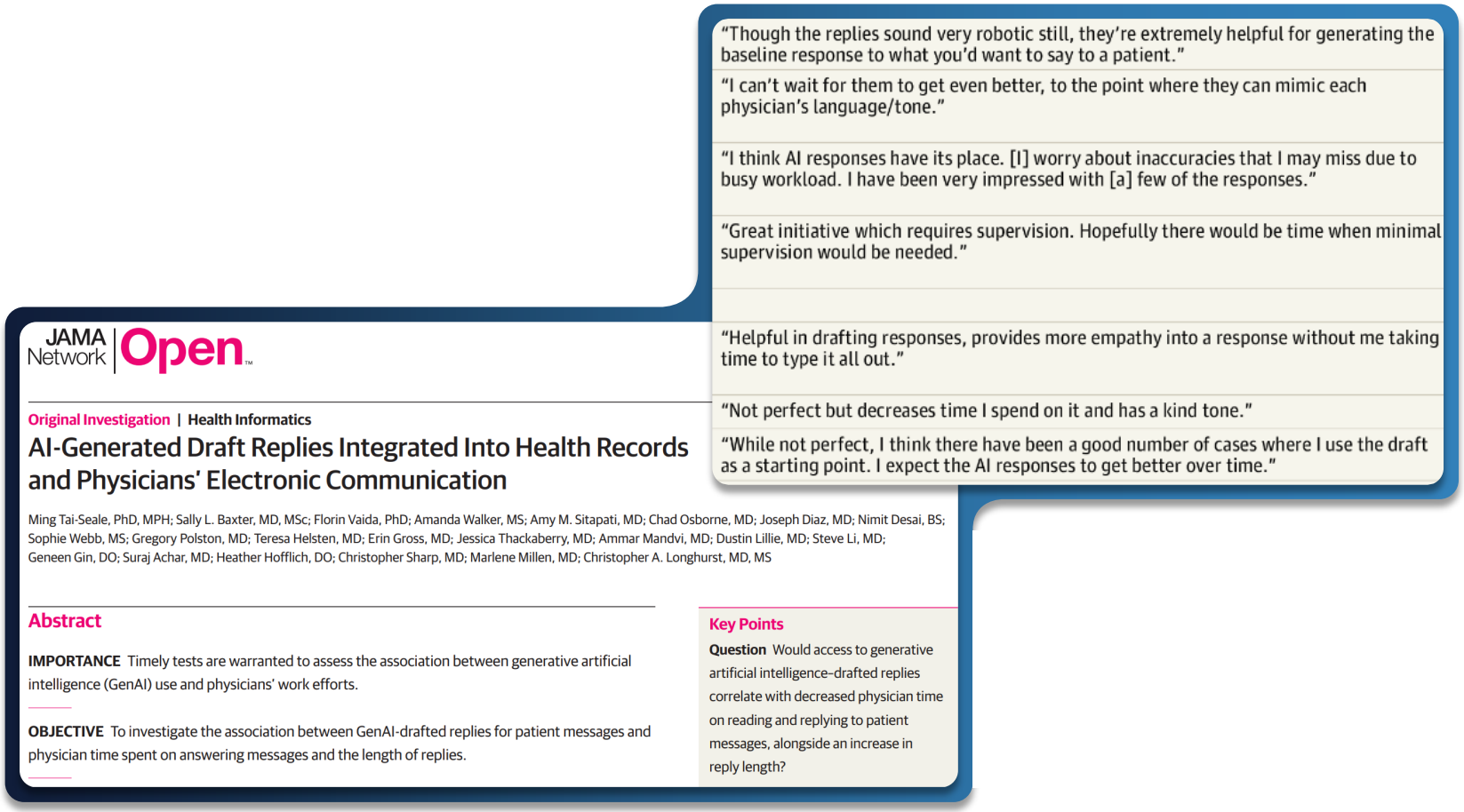


AI Principles in Action at UC San Diego Health





In early experience, AI does not save time when drafting replies to patients.



"Though the replies sound very robotic still, they're extremely helpful for generating the baseline response to what you'd want to say to a patient."

"I can't wait for them to get even better, to the point where they can mimic each physician's language/tone."

"I think AI responses have its place. [I] worry about inaccuracies that I may miss due to busy workload. I have been very impressed with [a] few of the responses."

"Great initiative which requires supervision. Hopefully there would be time when minimal supervision would be needed."

"Helpful in drafting responses, provides more empathy into a response without me taking time to type it all out."

"Not perfect but decreases time I spend on it and has a kind tone."

"While not perfect, I think there have been a good number of cases where I use the draft as a starting point. I expect the AI responses to get better over time."

JAMA Network Open

Original Investigation | Health Informatics

AI-Generated Draft Replies Integrated Into Health Records and Physicians' Electronic Communication

Ming Tai-Seale, PhD, MPH; Sally L. Baxter, MD, MSc; Florin Vaida, PhD; Amanda Walker, MS; Amy M. Sitapati, MD; Chad Osborne, MD; Joseph Diaz, MD; Nimit Desai, BS; Sophie Webb, MS; Gregory Polston, MD; Teresa Helsten, MD; Erin Gross, MD; Jessica Thackaberry, MD; Ammar Mandvi, MD; Dustin Lillie, MD; Steve Li, MD; Geneen Gin, DO; Suraj Achar, MD; Heather Hofflich, DO; Christopher Sharp, MD; Marlene Millen, MD; Christopher A. Longhurst, MD, MS

Abstract

IMPORTANCE Timely tests are warranted to assess the association between generative artificial intelligence (GenAI) use and physicians' work efforts.

OBJECTIVE To investigate the association between GenAI-drafted replies for patient messages and physician time spent on answering messages and the length of replies.

Key Points

Question Would access to generative artificial intelligence–drafted replies correlate with decreased physician time on reading and replying to patient messages, alongside an increase in reply length?

Use of AI resulted in:

21.8% increase in read time

17.9% increase in reply length

Tai-Seale, M., Baxter, S. L., Vaida, F., Walker, A., Sitapati, A. M., Osborne, C., ... & Longhurst, C. A. (2024). AI-generated draft replies integrated into health records and physicians' electronic communication. *JAMA Network Open*, *7*(4), e246565-e246565.

> The **AI you're using today** will be the **dumbest AI** you'll use in your lifetime.

– **Sam Altman**
OpenAI

A Vision for the Future: Hybrid Clinician-AI Teams

PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES MEDICAL SCIENCES OPEN ACCESS

Human–AI collectives most accurately diagnose clinical vignettes

Nikolas Zöller[a,1], Julian Berger[a], Irving Lin[b], Nathan Fu[b], Jayanth Komarneni[b], Gioele Barabucci[c], Kyle Laskowski[b], Victor Shia[d], Benjamin Harack[e], Eugene A. Chu[f], Vito Trianni[g], Ralf H. J. M. Kurvers[a,h,1,2], and Stefan M. Herzog[a,1,2]

Affiliations are included on p. 9.

Edited by Susan Fiske, Princeton University, Jamaica, VT; received December 19, 2024; accepted May 13, 2025

AI systems, particularly large language models (LLMs), are increasingly being employed in high-stakes decisions that impact both individuals and society at large, often without adequate safeguards to ensure safety, quality, and equity. Yet LLMs hallucinate, lack common sense, and are biased—shortcomings that may reflect LLMs' inherent limitations and thus may not be remedied by more sophisticated architectures, more data, or more human feedback. Relying solely on LLMs for complex, high-stakes decisions is therefore problematic. Here, we present a hybrid collective intelligence system that mitigates these risks by leveraging the complementary strengths of human experience and the vast information processed by LLMs. We apply our method to open-ended medical diagnostics, combining 40,762 differential diagnoses made by physicians with the diagnoses of five state-of-the art LLMs across 2,133 text-based medical case vignettes. We show that hybrid collectives of physicians and LLMs outperform both single physicians and physician collectives, as well as single LLMs and LLM ensembles. This result holds across a range of medical specialties and professional experience and can be attributed to humans' and LLMs' complementary contributions that lead to different kinds of errors. Our approach highlights the potential for collective human and machine intelligence to improve accuracy in complex, open-ended domains like medical diagnostics.

medical diagnostics | collective intelligence | large language models | health informatics | AI

Diagnostic errors are among the most pressing issues in medical practice (1–3), causing an estimated 795,000 deaths and permanent disabilities in the United States alone each year (4). Reducing diagnostic errors—without incurring substantially higher costs—is essential to improve patient outcomes worldwide. This challenge has motivated a recent surge in diagnostic technologies within the field of health informatics, which exploit AI to interpret medical records, tests, and images (5, 6). Deep learning approaches in medical imaging have shown great promise. Notable examples include mammography interpretation, cardiac function assessment, and lung cancer screening, some of which have progressed beyond the testing phase and entered clinical practice (7–9).

Recent years have also witnessed the rise of AI foundation models, especially large language models (LLMs), which show remarkable abilities to process natural language, providing accurate answers to questions in almost any domain, including medicine (10–12). However, a recent meta-analysis (13) found that physicians often outperform LLMs, and that LLMs differ vastly in performance, also between medical specialties. While LLMs' performance in the medical domain keeps improving (12), their deployment in clinical

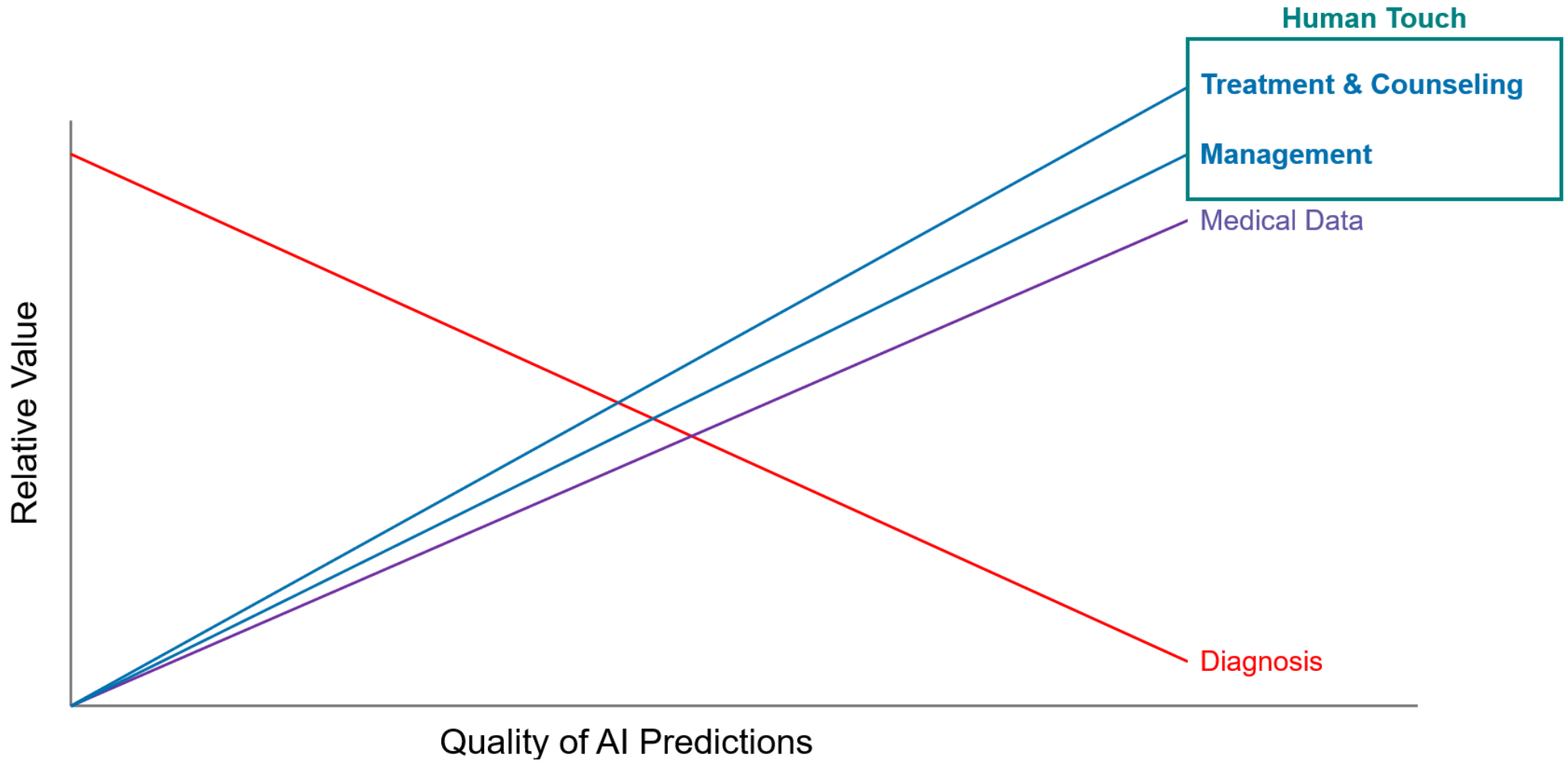
Significance

Large language models (LLMs) have great potential for high-stakes applications such as medical diagnostics but face challenges including hallucinations, biases, and lack of common sense. We address these limitations through a hybrid human–AI system that combines physicians' expertise with LLMs to generate accurate differential medical diagnoses. Analyzing over 2,000 text-based medical case vignettes, hybrid collectives outperform individual physicians, standalone LLMs, and groups composed solely of physicians or LLMs, by leveraging complementary strengths while mitigating their distinct weaknesses. Our findings underscore the transformative potential of human–AI collaboration to enhance decision-making in complex, open-ended domains, paving the way for safer, more equitable applications of AI in medicine and beyond.

- AI ensemble outperformed about 85% of individual clinicians
- Hybrid teams (humans + AI) consistently outperformed all other configurations, even improving over pure human or pure AI collectives
- Hybrid human–AI systems could enhance patient safety and health outcomes

Zöller, N., Berger, J., Lin, I., Fu, N., Komarneni, J., Barabucci, G., ... & Herzog, S. M. (2025). Human–AI collectives most accurately diagnose clinical vignettes. *Proceedings of the National Academy of Sciences*, *122*(24), e2426153122.

Enhancing the human touch in medicine with AI



Balancing Innovation and Risks in the Use of AI by Health Insurers

Christina Silcox, PhD

Research Director, Duke-Margolis Institute for Health Policy

How Health Plans Use AI Today

Business uses
- IT
- Employee support
- Member services

Actuarial modeling
- Projecting future claims costs
- Innovative policy design

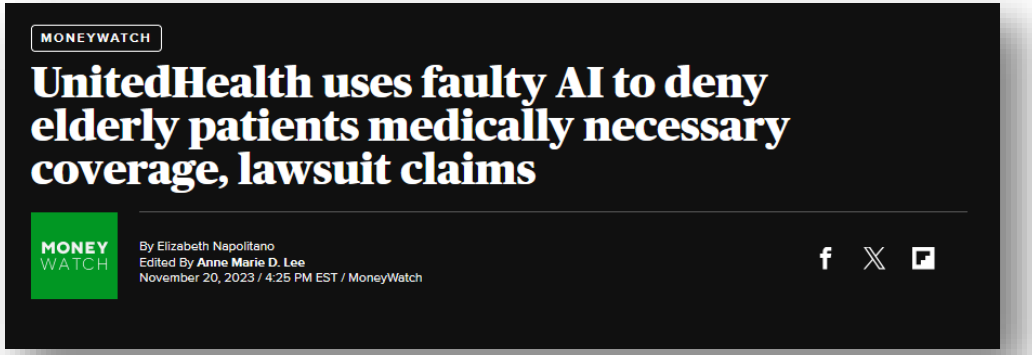
Fraud detection
- Real-time anomaly detection
- Pattern detection to flag unusual coding or abrupt changes

Population Health
- Risk stratification
- Targeting preventive health resources
- Identifying social services needs

Claims Processing
- Automated claims processing
- Prior authorization
- Utilization management

AI Risks: The Usual Suspects

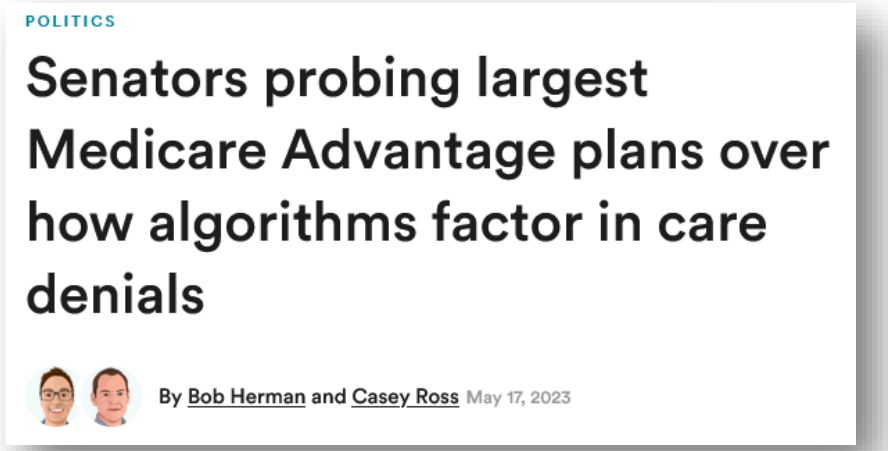
- Accuracy
  - Lack of transparency and regulation
    - Unknown tools
    - Unknown performance
    - Unknown subpopulation performance
    - Performance drift
  - Hallucinations
- Generalizability
- No explanation of output


MONEYWATCH

UnitedHealth uses faulty AI to deny elderly patients medically necessary coverage, lawsuit claims

MONEY WATCH

By Elizabeth Napolitano
Edited By Anne Marie D. Lee
November 20, 2023 / 4:25 PM EST / MoneyWatch


POLITICS

Senators probing largest Medicare Advantage plans over how algorithms factor in care denials

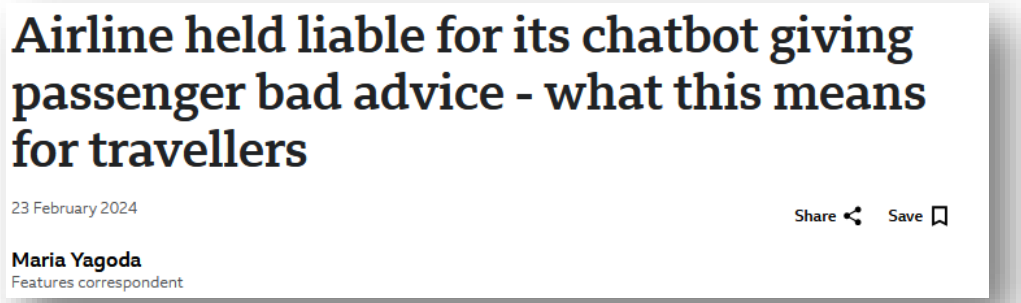
By Bob Herman and Casey Ross May 17, 2023

What Else Should We Be Watching?

- Liability/Accountability
  - Who is responsible when AI makes a mistake?
- Increasing administrative requirements

Alternatively

Being too cautious to realize the full potential of AI to improve patient outcomes, patient experience, and lower costs



Airline held liable for its chatbot giving passenger bad advice - what this means for travellers

23 February 2024

Share Save

Maria Yagoda
Features correspondent

Battle of the bots: As payers use AI to drive denials higher, providers fight back

As denial rates climb to record highs, driven in part by AI-powered robots, health systems are starting to stand their ground.

Jeni Williams

March 28, 2024 3:39 pm

How to Make AI Safer and More Trustworthy

What can be done?

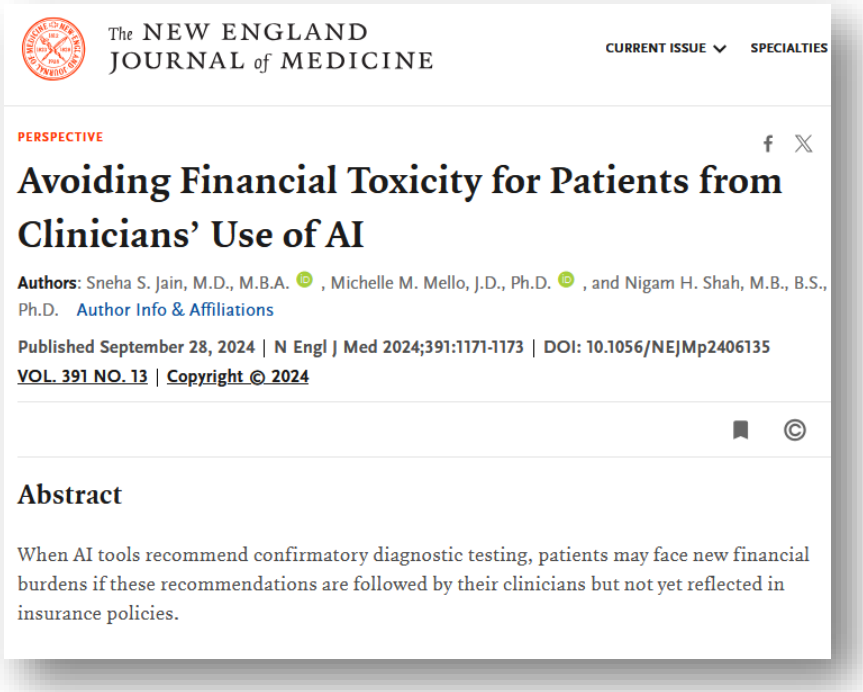
- High quality governance
  - Risk assessment, testing, and monitoring
  - Based on consensus standards
  - Independent testing
- Disclosure and Transparency
- Require explainability for certain uses
- Require “human in the loop” for certain uses

What’s the challenge?

- Lack of consensus standards
  - There are emerging best practices
- Lack of third-party evaluators
- Testing results are not reliable over time due to performance drift
- Still learning about AI, especially Generative AI
- Still learning how humans interact with AI over time

How Do We Pay for AI?

- Payment for AI tools themselves
  - Some coding and coverage for AI-enabled clinical tools (FDA authorized medical devices)
  - Indirect provider benefit
    - Increased provider efficiencies
    - Decreased costs within value-based care
- Payment for AI-enabled diagnosis and treatment recommendations



The NEW ENGLAND JOURNAL of MEDICINE

CURRENT ISSUE SPECIALTIES

PERSPECTIVE

Avoiding Financial Toxicity for Patients from Clinicians' Use of AI

Authors: Sneha S. Jain, M.D., M.B.A. , Michelle M. Mello, J.D., Ph.D. , and Nigam H. Shah, M.B., B.S., Ph.D. Author Info & Affiliations

Published September 28, 2024 | N Engl J Med 2024;391:1171-1173 | DOI: 10.1056/NEJMp2406135
VOL. 391 NO. 13 | Copyright © 2024

Abstract

When AI tools recommend confirmatory diagnostic testing, patients may face new financial burdens if these recommendations are followed by their clinicians but not yet reflected in insurance policies.

Policy Landscape

- Federal
  - Administration's focus on innovation and encouraging AI integration into agency work
  - Existing Regulations
    - Prohibits tools with **biased performance** or cause **biased outcomes**
    - Some tools require **pre-market FDA review** (clinical tools only)
    - Some tools require **transparency** (limited set of tools integrated into EHRs)
    - Medicare Advantage claims tools must comply with all applicable CMS coverage rules (**explainability, human in the loop**)
  - Proposed "pause" on state and local laws on AI (as of 6/27)
- California
  - Medical necessity determinations must be ultimately made by a licensed human (**human in the loop**)

Challenge

- Defining AI / scope of policy
- AI is evolving significantly faster than typical policymaking

Requirements on disclosure, transparency, governance, risk mitigation plans are helpful

- More specific policies often need to be tailored to specific uses
- Need to assign accountability to the stakeholder that can most affect the issue

Questions?