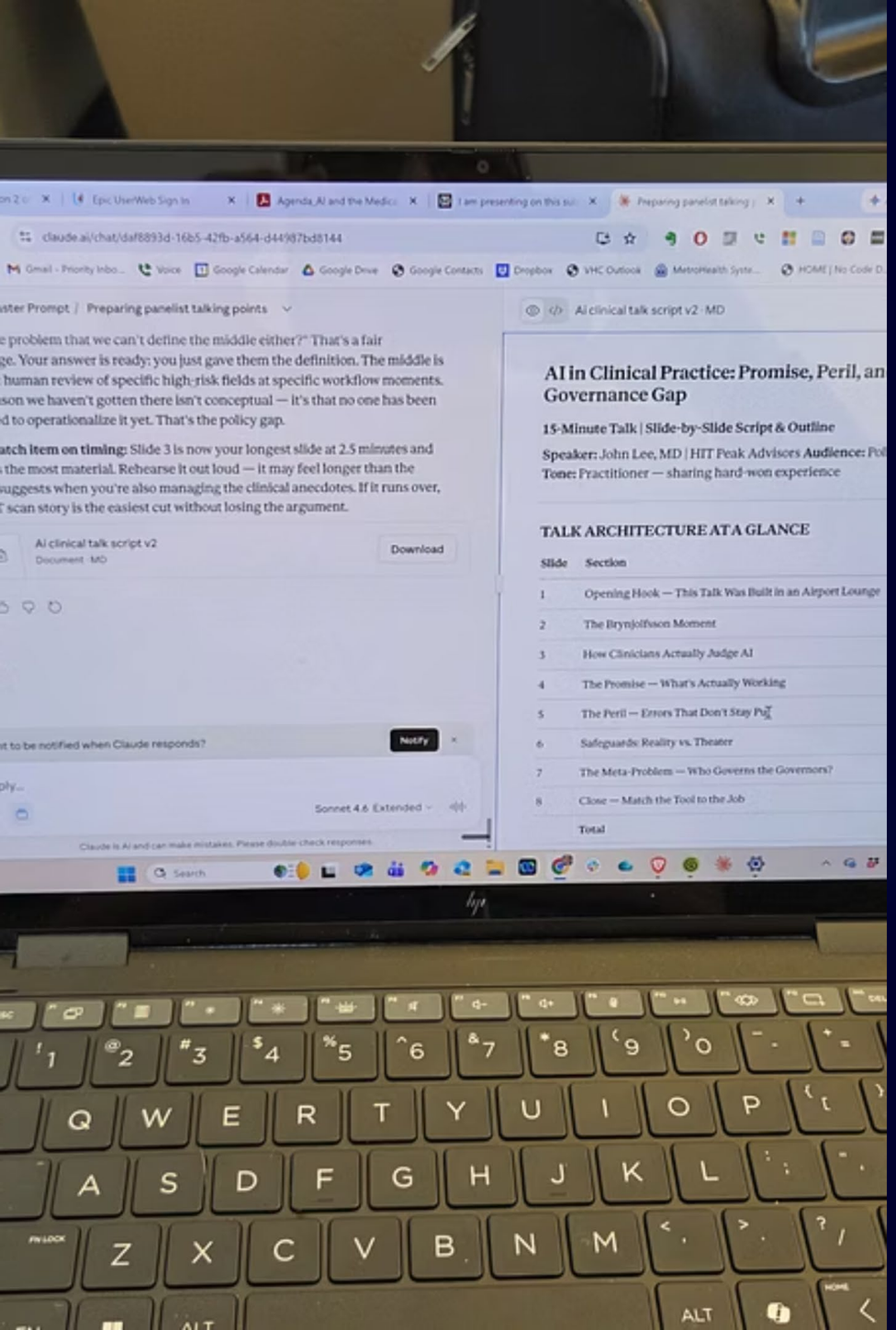


AI in Clinical Practice: Promise, Peril, and the Governance Gap

John Lee, MD | HIT Peak Advisors · Emergency Physician & Health IT Consultant

"This talk took 30 minutes. It used to take weeks. That's not a boast. That's the whole argument."

THE BRYNJOLFSSON MOMENT

We keep the broken workflows. We just electrify them.

1890s → 1920s

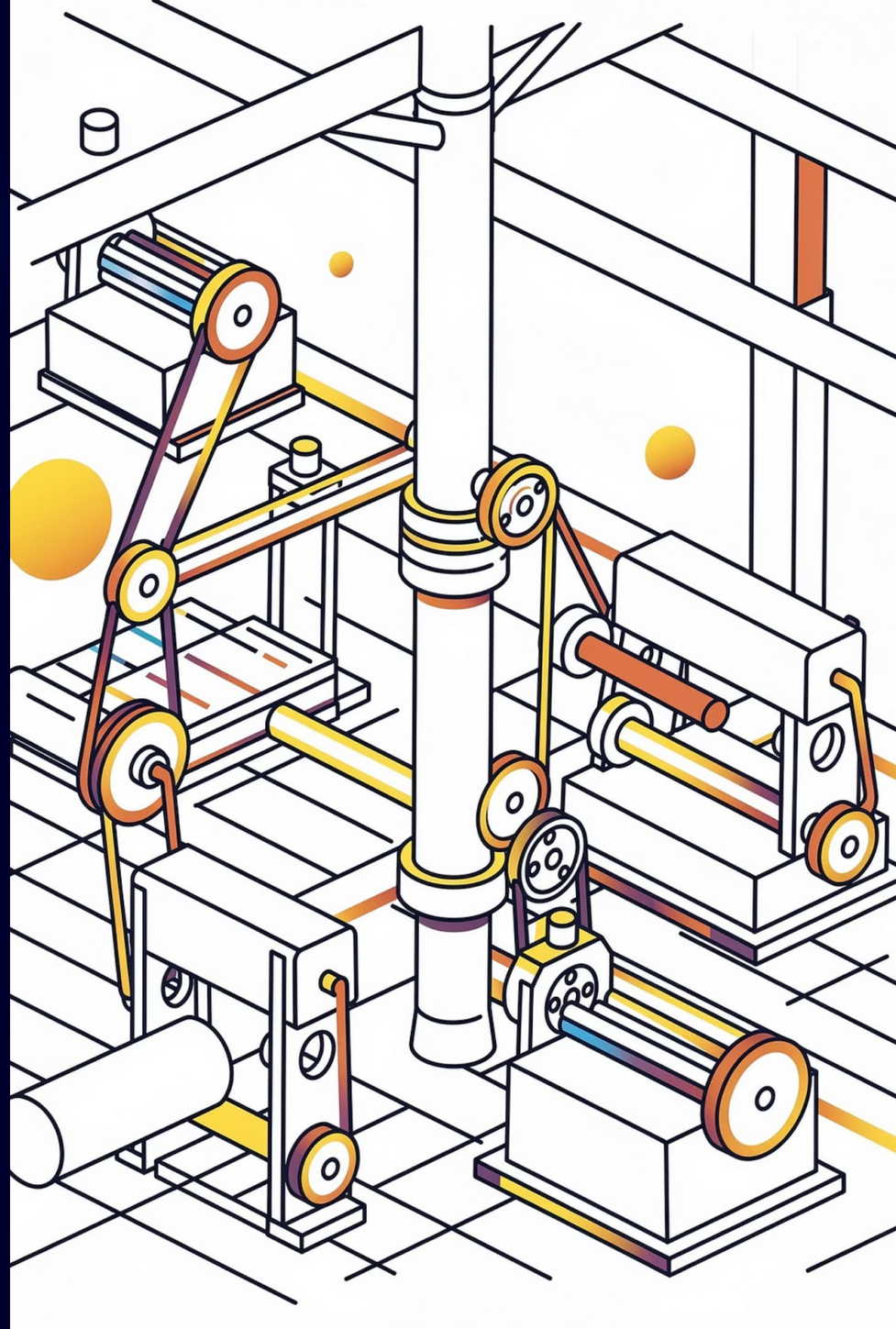
2nd Machine Age

Steam to Electric

Healthcare Today

New tools, old workflows

"Same floor plan. Wrong technology."The organizations that win will redesign the factory — not just rewire it.



CLINICAL JUDGMENT FRAMEWORK

It's not *"is it accurate?"* It's "what happens when it's wrong?"

Real Problem?

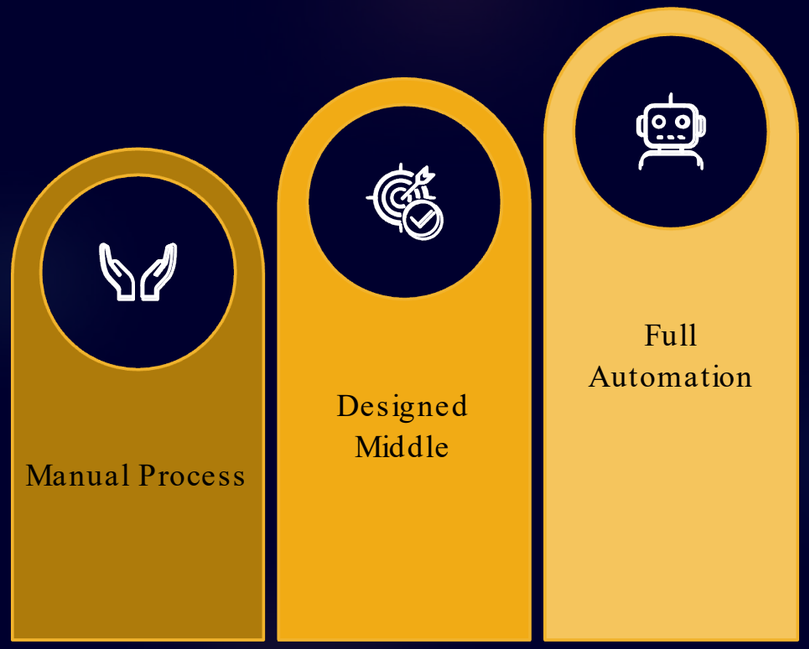
How Does It Fail?

Automation Bias

What does it do to my *thinking?*

Accountability

Who is responsible when it's wrong? "Human in the loop"



THE PROMISE

The real wins are in the *unglamorous middle.*

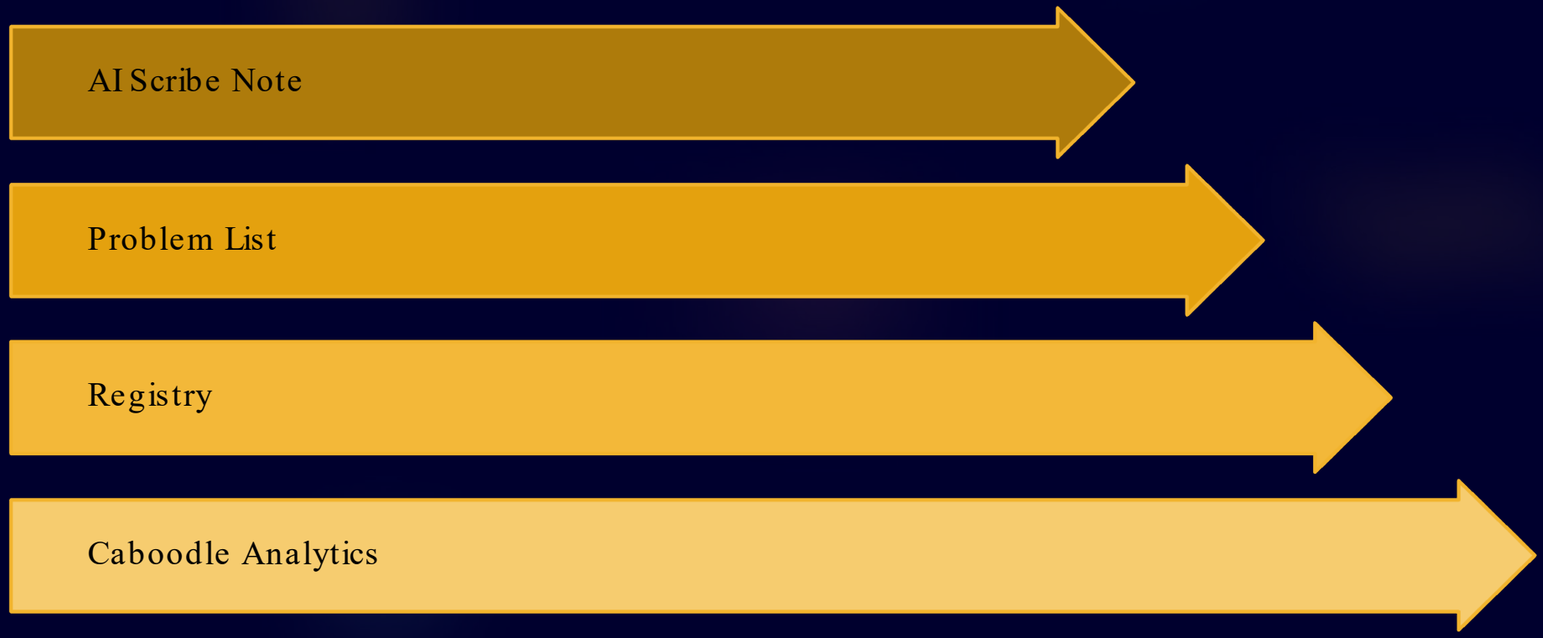
Administrative Burden

Early Warning — When Redesigned

Structured Data, Right Moment

THE PERIL

An AI error in one note is *not one error.*



The Scribe Gap

Multi-vendor study: none of 6 AI scribes were error-free. Omissions and Comissions

Laundered History

Next AI summarization reads prior notes — treats the error as **confirmed clinical history**.

AI scribe omits a symptom. Physician signs in 5 seconds between patients. That omission is now in the legal record — and in every system downstream.

SAFEGUARDS

Reality vs. Theater

What We Say We Do	What Actually Happens
Pre-deployment validation	Vendor data on vendor-selected populations
Human in the loop	Physician reviewing in 5 seconds, managing 5 other things
Post-deployment monitoring	Dashboards no one has staff to act on
Governance committee	Volume » Pace & Capacity

"The note is fluent. Professional. Complete-looking. The error is in what's missing."

Automation bias is the hidden risk: confident output meets an overloaded clinician. What is the practical "Human in the loop“?

GOVERNANCE

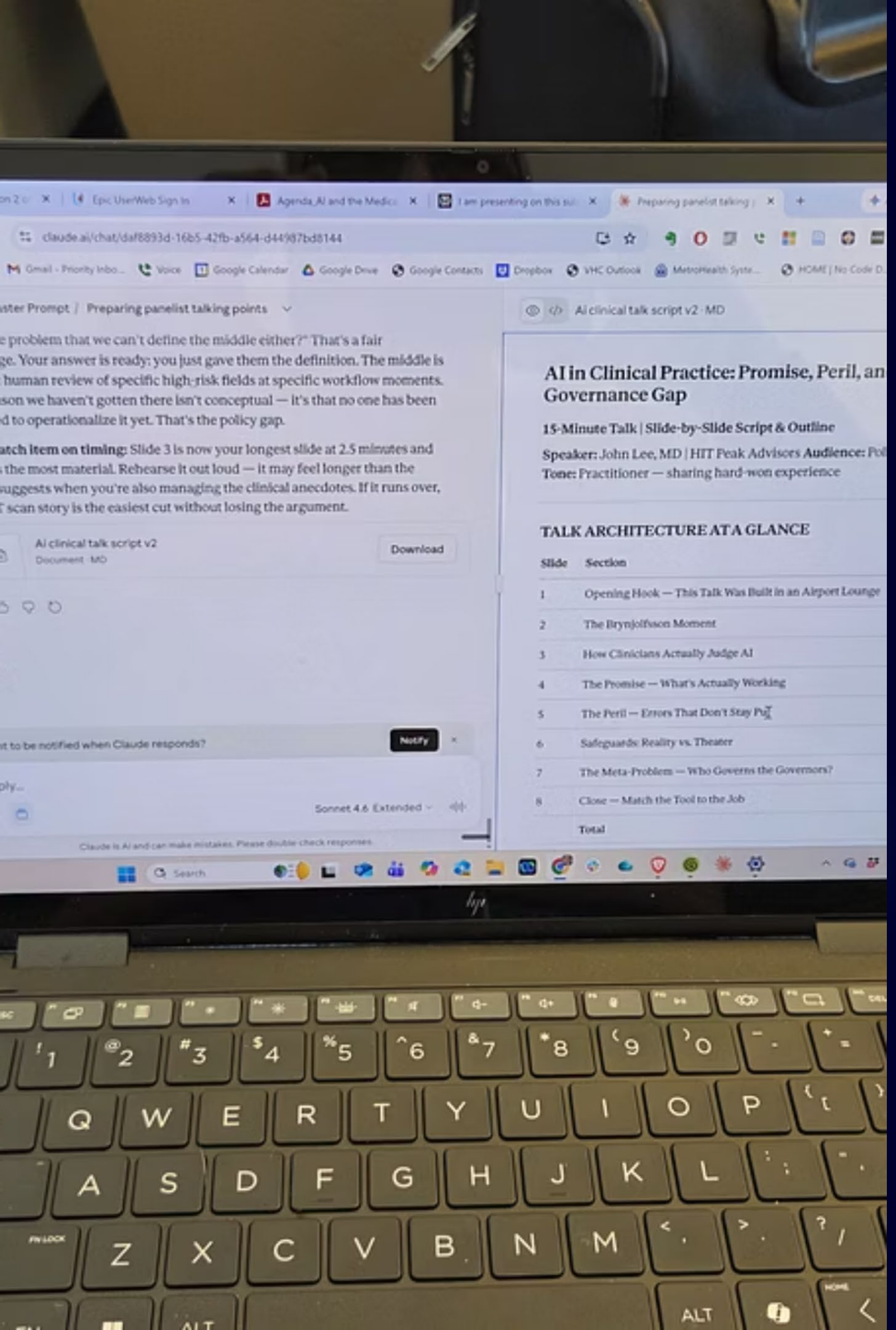
We are trying to govern a tidal wave *with a paper process.*

Scale Mismatch

What's Required

Use AI to Govern AI?

The policy question is not *"was AI used?"* — it's "was there a system to know when it was failing?"



CLOSE

The factories that won didn't just get new power. They rebuilt.

01

Right Tool, Right Task

02

The Brynjolfsson Lesson

03

The Path Is Visible

04

What's Missing

A policy mandate that as nimble as the technology.

Cautiously optimistic. Eyes open.

Built with Claude. Reviewed, edited, and delivered by a human. Accountable end to end.Thank you.

AI and the Clinical Note: Real-World Oversight

Jinoos Yazdany, MD MPH

Alice Betts Endowed Professor of Medicine

Chief of Rheumatology, San Francisco General Hospital, UCSF

Executive Director of AI Monitoring at UCSF

She had been on tacrolimus.

The side effects were serious.

We made a clear decision:

Stop the drug.

The AI scribe documented:
"Continue tacrolimus."

What is IMPACC?

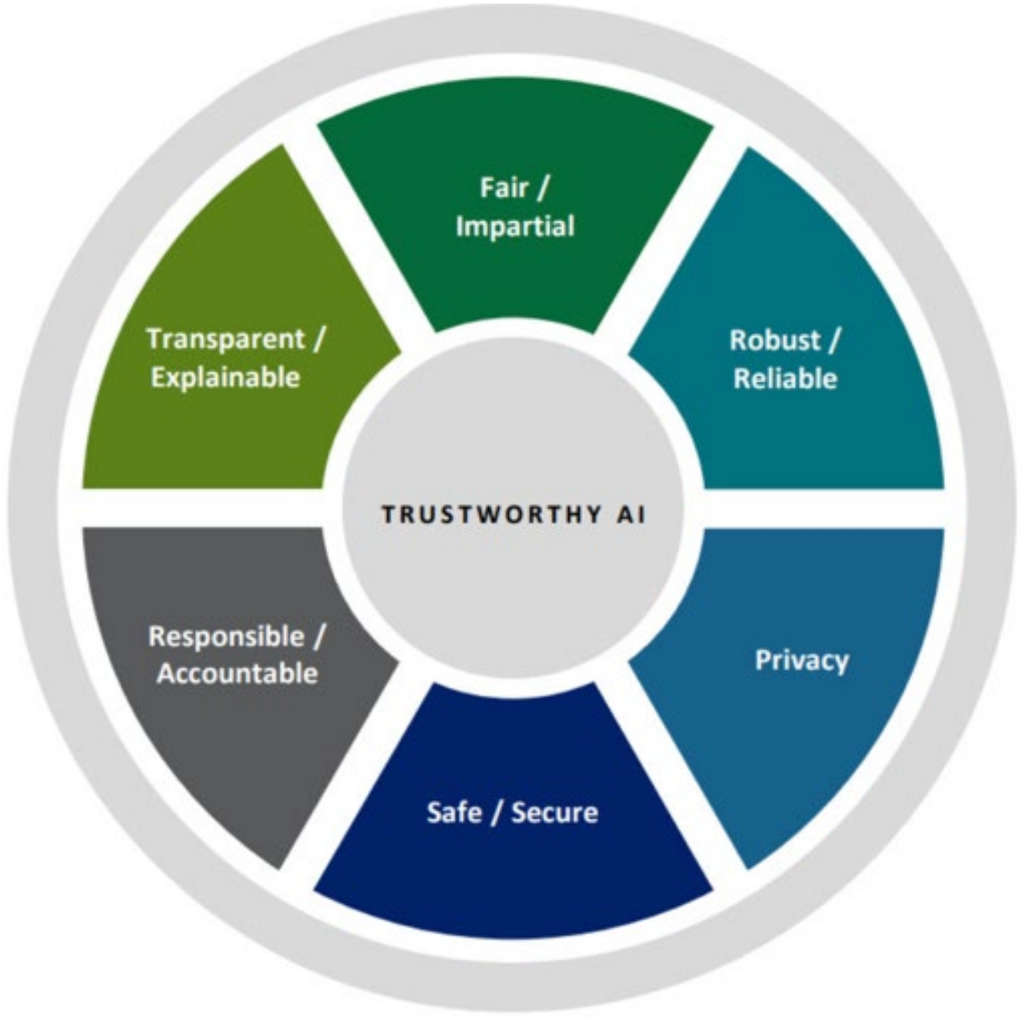
IMPACC = Impact Monitoring Platform for AI in Clinical Care

A platform to support longitudinal **monitoring of AI impact** & to advance AI monitoring **research**

"This is the first partnership between UCSF and UCSF Health on AI monitoring. Together, we are uniquely positioned to create the first effective model platform across health systems in the United States that will offer real-time visibility into AI tool performance and clinical impact."

Suresh Gunasekaran
President and Chief Executive Officer, UCSF Health

IMPACC Framework for AI Monitoring



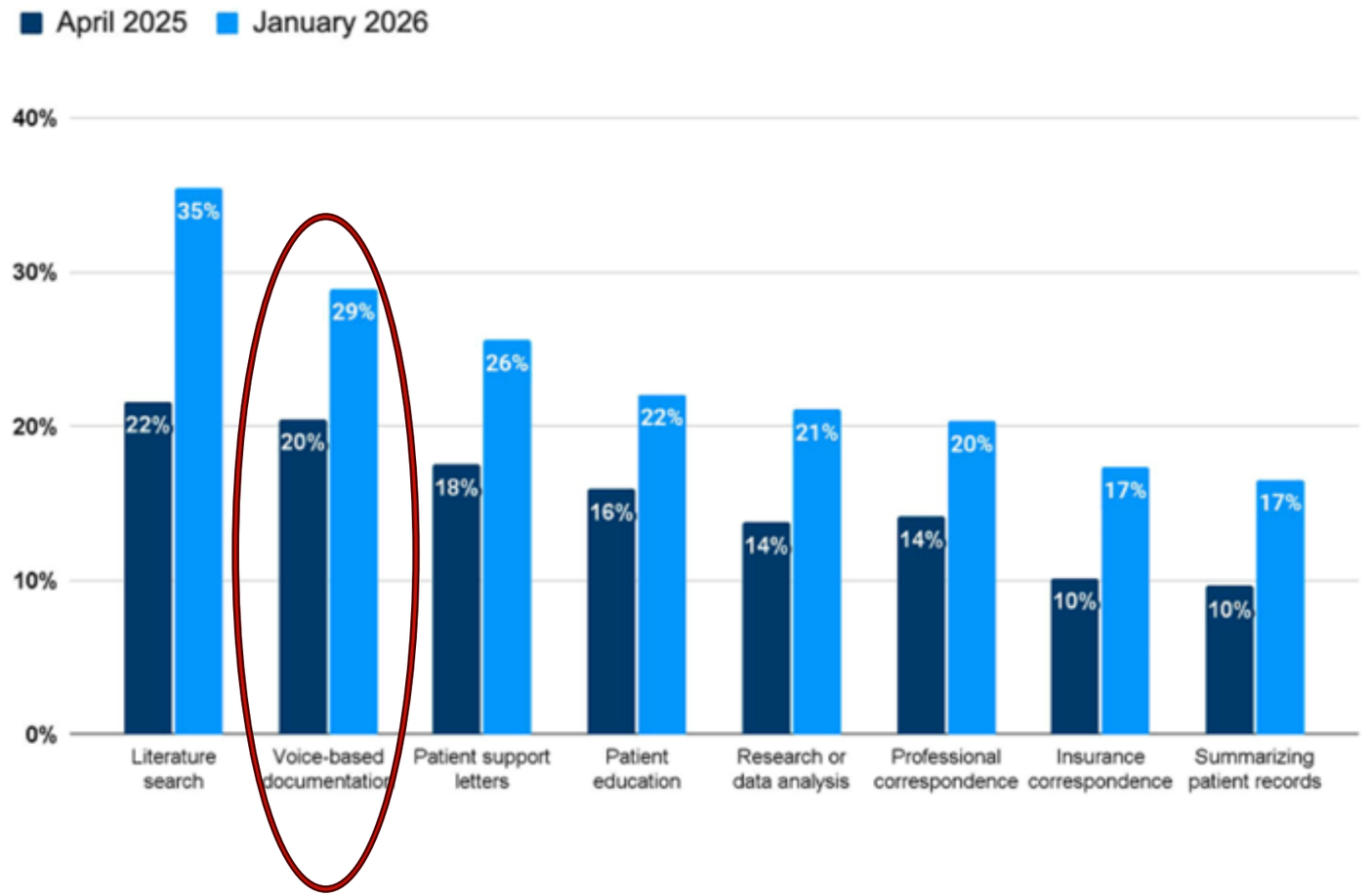
Trustworthy AI (TAI) is a framework to mitigate the various forms of risk arising from AI. Created by the US Department of Health and Human Services in 2021, the framework offers detailed guidance to ensure that health-related AI is implemented in a way that is ethical, effective and secure.

Usefulness and Impact

- Utilization
- Workflow Impact
- Automation Bias
  Deskilling
- Healthcare Professional Experience
- Patient Experience

AI Scribes Are Everywhere and Retention is HIGH.

How Physicians Are Using AI in Their Practice*†



*Physician respondents selected all that applied.

†Survey dates are March–April 2025 and November 2025–January 2026.

https://www.doximity.com/reports/state-of-ai-medicine-report/2026

JAMA Network | Open™

Research Letter | Health Informatics

Ambient Artificial Intelligence Scribes and Physician Financial Productivity

A Jay Holmgren, PhD, MHI; Cynthia L. Fenton, MD; Robert Thombley, BS; Hossein Soleimani, PhD; Rhiannon Croci, BSN, RN-BC; Orianna DeMasi, PhD; Maria E. Byron, MD; Sara G. Murray, MD, MAS; Julia R. Adler-Milstein, PhD; Jinoos Yazdany, MD, MPH

Introduction

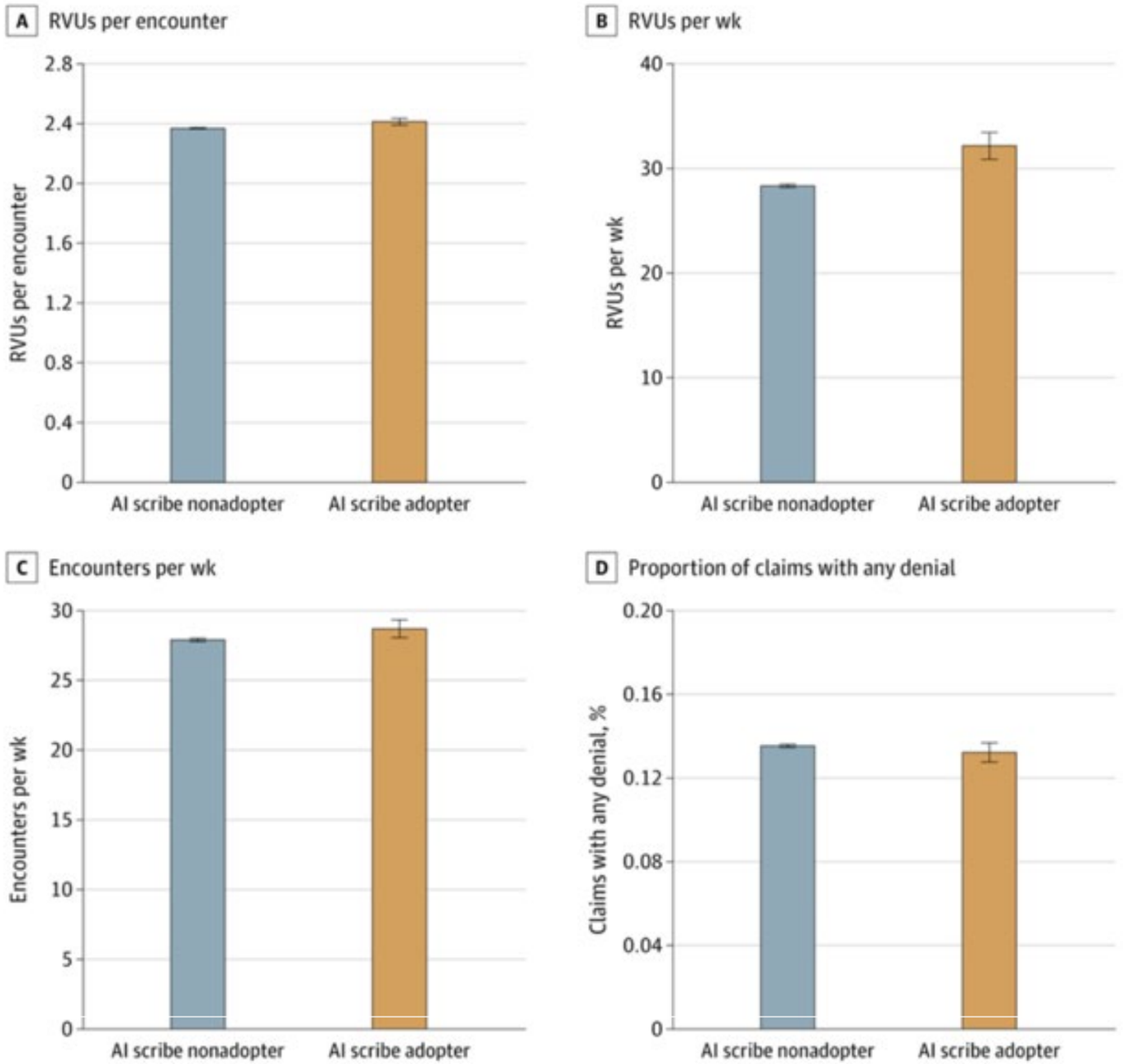
Adoption of ambient artificial intelligence (AI) scribes, which generate clinical documentation from audio recordings and are associated with reduced documentation time and burnout,[1-3] is increasing. Little is known regarding AI scribes and revenue changes. Without requirements to see more patients, increased relative value units (RVUs) or decreased claim denials may help mitigate the expense of AI scribes.[4] Understanding their financial implications also informs policymakers of potential increases in health care spending. We examined whether ambient AI scribe adoption is associated with changes in RVUs, ambulatory visit volume, and claim denials.

+ Invited Commentary

+ Supplemental content

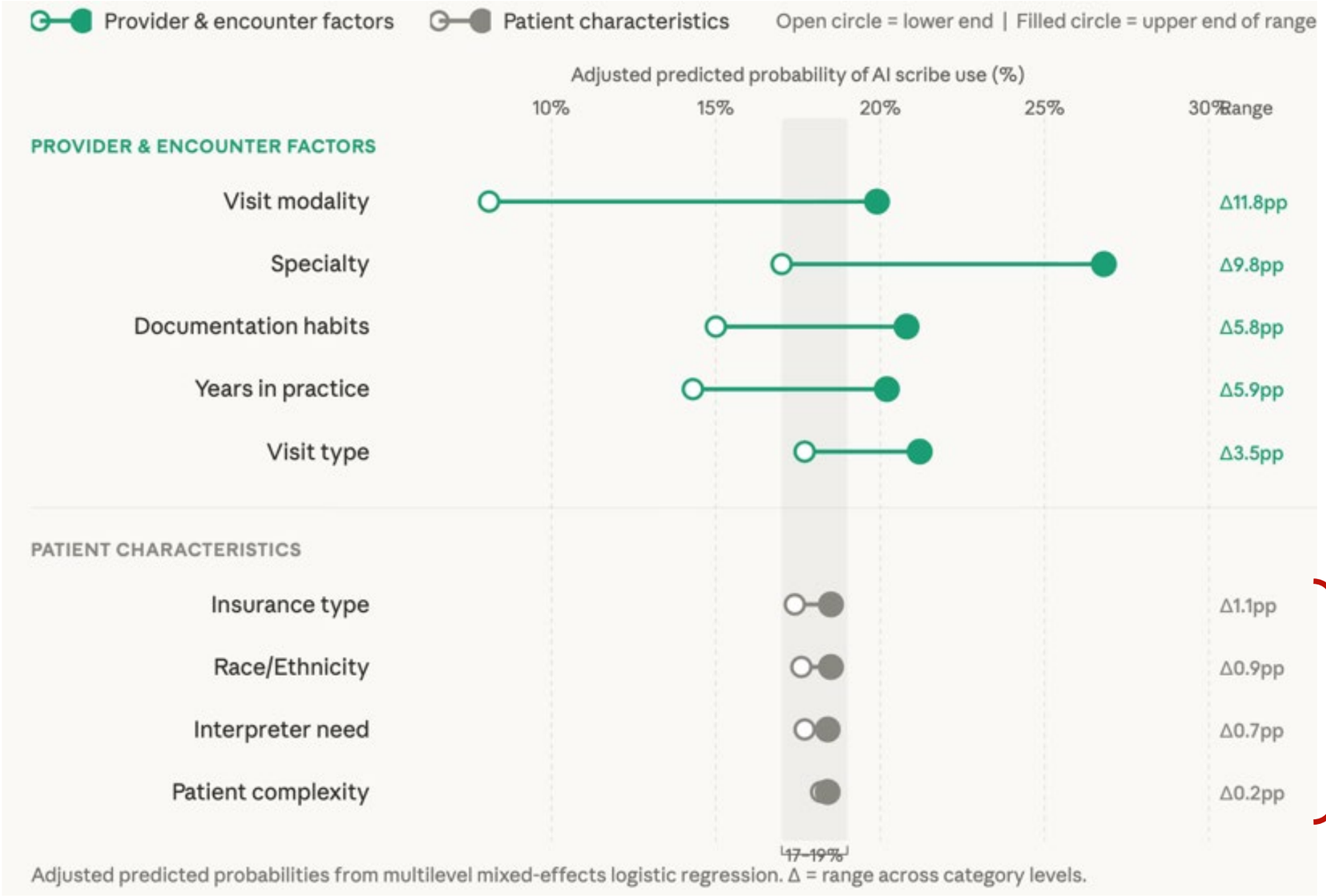
Author affiliations and article information are listed at the end of this article.

Figure 1. Association Between Physician Financial Productivity and Artificial Intelligence (AI) Scribe Adoption



Holmgren et al, *JAMA Network Open*

AI Scribes Use Does Not Vary by Patient



Ostrer et al., Abstract Academy Health to be presented in June 2026.

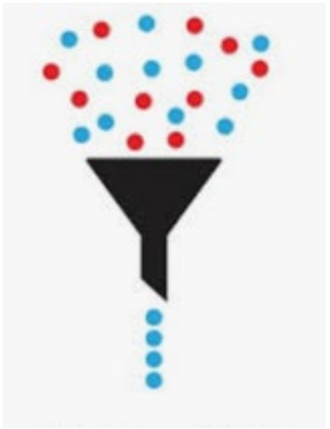
The Note is Changing – the History of Present Illness

Typical Pre-scribe HPI for my patient

She reports two discoid lesions -one on her scalp and one on her eyebrow

Some joint pain - it is clearly activity associated

Her blood pressure has been normal at home



Typical post-scribe HPI for my patient

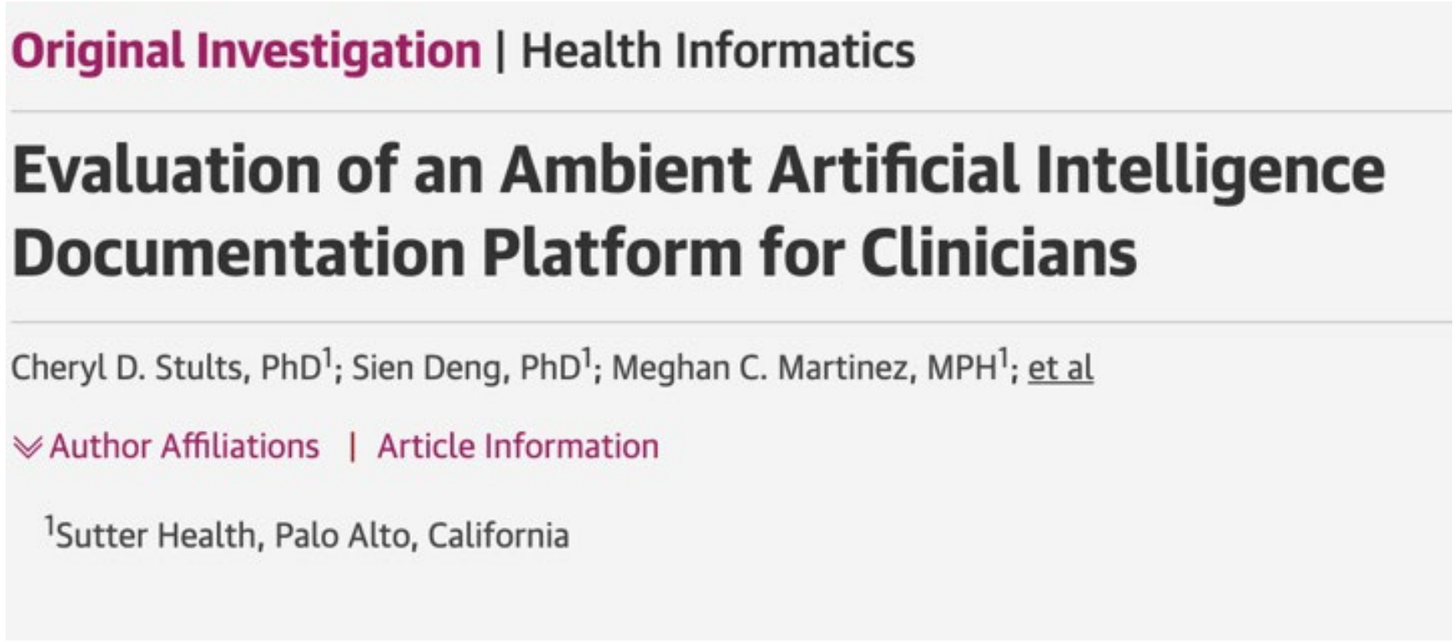
Patient with lupus reports recent travel to Asia, where she visited family. During her trip, she experienced increased fatigue after attempting to reduce her Cellcept dose to 500 mg daily; she subsequently resumed her prior dose of 1000 mg daily, which improved her symptoms. She continues hydroxychloroquine and gabapentin as previously prescribed. No new joint pain or swelling reported.

She notes a change in her left toe nail color to black after returning from Asia, which she attributes to possible trauma from walking. She also reports a dark spot under her right fingernail, which she is treating with antifungal medication. No new rashes. Hair loss is minimal and stable, with some shedding during washing but not excessive. She applies steroid cream to a mild facial rash as needed.

She continues to follow dietary and lifestyle recommendations from her Chinese medicine provider, including adjustments to sleep schedule, meal timing, and food preparation methods. No recent hospitalizations or medication changes aside from the temporary Cellcept dose adjustment. Blood pressure has been elevated in the clinic but remains controlled at home; she monitors with multiple devices and is currently on amlodipine 5 mg daily. She reports occasional dizziness with higher doses of amlodipine.

Review of Systems: Denies chest pain, shortness of breath, or new joint pain. No fevers, night sweats, or weight loss. No oral ulcers or photosensitivity. No new neurological symptoms.

Note length is increasing



Original Investigation | Health Informatics

Evaluation of an Ambient Artificial Intelligence Documentation Platform for Clinicians

Cheryl D. Stults, PhD[1]; Sien Deng, PhD[1]; Meghan C. Martinez, MPH[1]; et al

Author Affiliations | Article Information

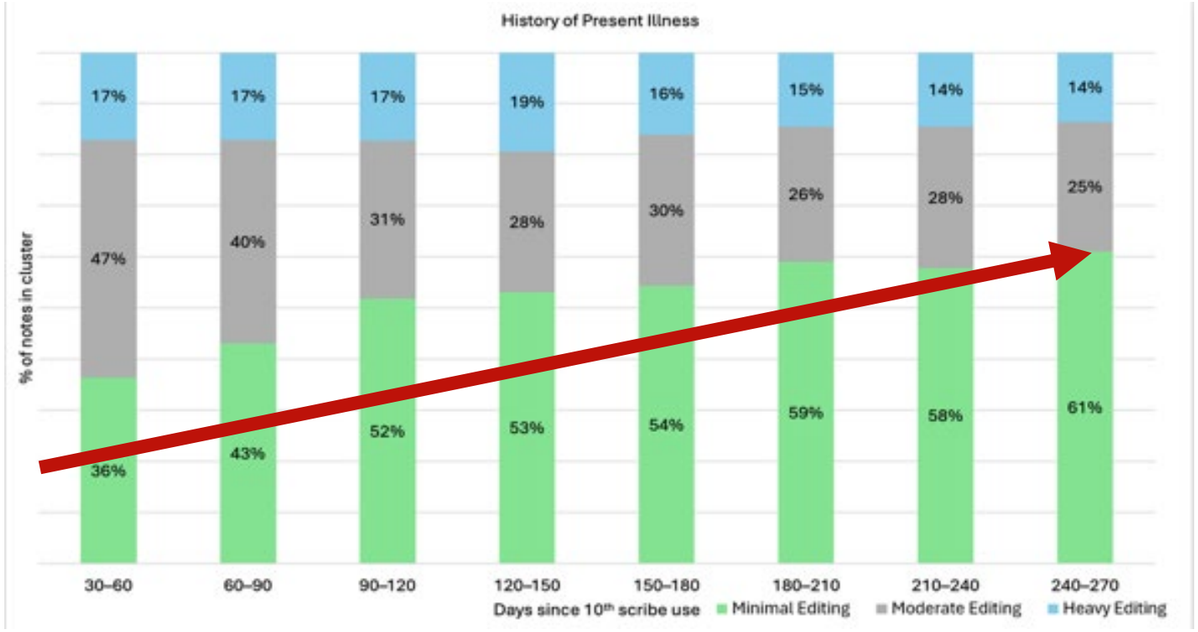
[1]Sutter Health, Palo Alto, California

Mean progress note length increased from 5,683 → 5,961 characters

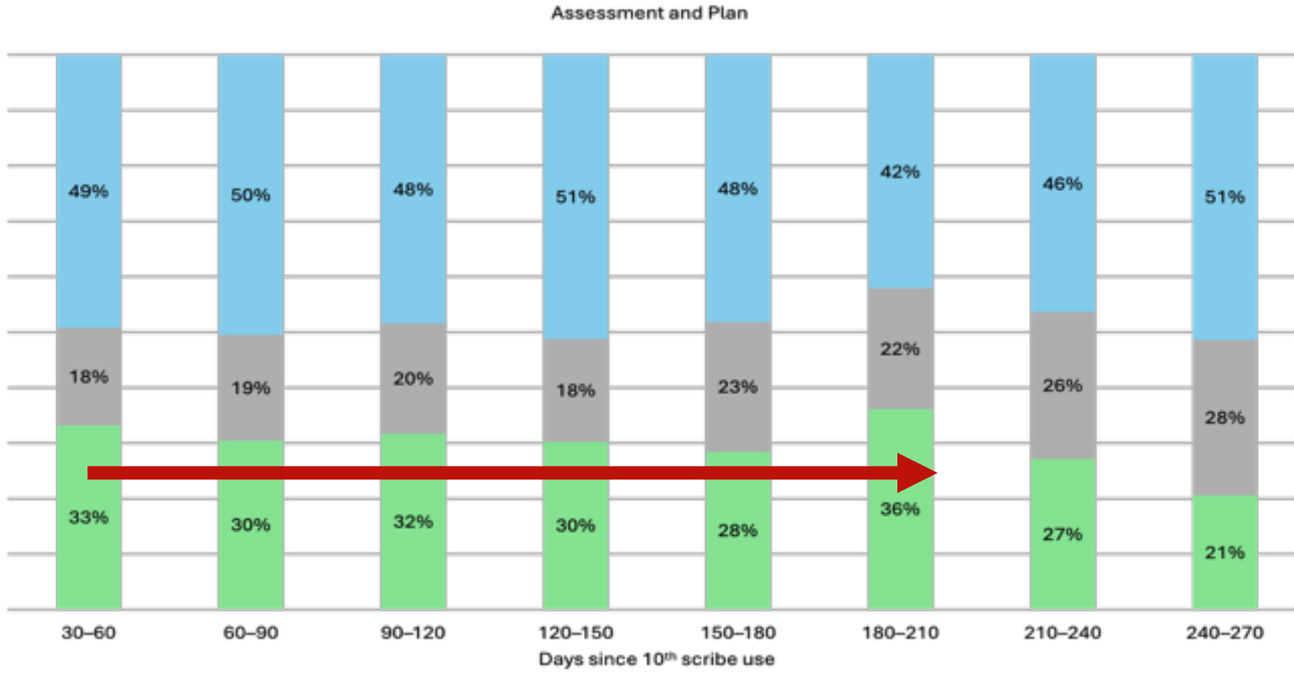
→ +278 characters (~5%)

Is editing behavior changing over time?

History of Present Illness



Assessment and Plan



Garcia-Agundez et al, FDA CERSI conference 2026. https://cersisummit.com/wp-content/uploads/2026/01/2026-Summit-Poster-List.pdf

AI Scribe Safety: Physician in-app feedback

Cluster	Freq.	Manual Summary	LLM Summary
HPI errors	41 (23.7%)	Incorrect details, specifically in HPI: misplaced diagnostic information, incorrect medication, missing details	Misplaced exam findings, incomplete structure, and blurred lines between HPI and other sections Critical information is often omitted or fabricated within the HPI
Medication errors	34 (19.7%)	Errors in medication dosage and spelling	Incorrect medication names, dosages, taper instructions, or titration schedules Frustration over locked After Visit Summary (AVS) sections that prevent corrections
Speaker details and attribution	28 (16.2%)	Incorrect details from appointment: speaker misattribution, missing symptom discussion, exam results	Statements are often misattributed between patients, caregivers, or providers Details (comorbid conditions, social history, family dynamics) are omitted/inaccurately captured
Sleep-specific	27 (15.6%)	Feedback specific to sleep medicine setting: requests for information like sleep/wake times to be transcribed	Sleep schedules, lifestyle factors, and compliance details are missing in summaries
Pronouns and personalization	17 (9.8%)	Mostly positive feedback; requests to use patient names, correct name misspellings	Misuse of pronouns and failure to use preferred or personalized names Request more conversational language
Surgery-specific	15 (8.7%)	Errors in capturing surgical/treatment discussions, such as surgery risks	Risks, benefits, and alternative surgical options are poorly documented/omitted Surgical plans include fabricated/inaccurate information

Table 1: *Quantitative results: Summaries of BERTopic clusters produced on safety-filtered feedback. LLM summaries have been edited for length; see Table 4 in Appendix A for full LLM outputs.*

We are building LLMs to tag critical safety events

https://arxiv.org/html/2512.04118v1

Loss of nuance in some sections

Pre-scribe A&P

#SLE with lupus nephritis

Her proteinuria is down significantly, and her only symptom is occasional swelling of areas of her face and scalp. The appearance today on her L forehead is concerning for panniculitis, although the time course she describes (getting better in days) unusual. She still has significant serologic activity.

Overall, I would prioritize reducing her steroid exposure and ensuring her Myfortic dose is maximized. Ideally, we could try completely tapering off methylprednisolone and see if she can tolerate a slightly higher dose of Myfortic without GI side effects. I would recommend the following:
-see if she can tolerate Myfortic 360 mg qAM and 720 mg qPM; if GI symptoms return, she can go back down to 360 mg BID
-reduce prednisone to 5 mg per day for one month; then see if she can taper off
-if her facial swollen areas return, would recommend derm eval to help us understand if this is panniculitis or something else
-continue other medications (belimumab + HCQ)
-continue to follow B lymphocyte subsets (CD19) to monitor for B cell repletion
-check serum IgG to assess for hypogamm post-ritux
-consider starting ACE inhibitor to reduce remaining proteinuria.

Post-scribe A & P

Systemic lupus erythematosus

- Increase mycophenolic acid to 2 tablets in the morning and 1 tablet at night
- - Decrease pr isone to 5 m ily
- - Continue hydroxy e 200 mg daily - - Continue Benlysta i s once weekly
- - Continue iro plement
- - Ask for a local referral to an ophthalmologist for a yearly eye exam while taking hydroxychloroquine

I do not use the scribe A&P at all – writing IS part of my process for critical thinking in rheumatology

Loss of Clinical Nuance

Integrating AI Scribes into Medical Education: Guardrails for Preserving Clinical Reasoning

Jane Abernethy, MD, MBE[1], Anna Shah, MD[2], Belinda Chen, MD[1], Stasia Reynolds, MD[1], Scott M Wright, MD[1], and Paul O'Rourke, MD, MPH[1]

[1]Johns Hopkins Bayview Medical Center, Division of General Internal Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA; [2]Division of General Internal Medicine, University of South Florida, Tampa, FL, USA

"Subtle threads of reasoning — why a specific test was chosen, how a diagnosis was prioritized—were often missing, even when the AI-generated text appeared coherent."

AI scribes increase symptom documentation but decrease clinical action

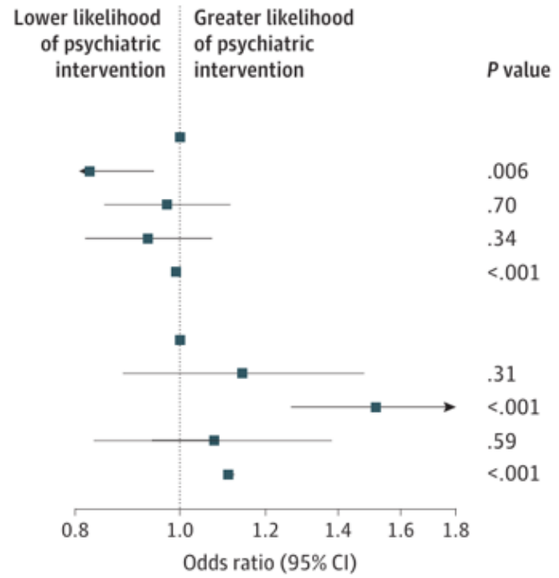
JAMA Psychiatry | Original Investigation | AI IN PSYCHIATRY

Psychiatric Documentation and Management in Primary Care With Artificial Intelligence Scribe Use

Victor M. Castro, MS; Thomas H. McCoy, MD; Pilar Verhaak, BS; Anudeepa Ramachandiran, BA; Roy H. Perlis, MD, MSc

Figure 2. Multiple Logistic Regression Model of Psychiatric Intervention at Visit

Variable	Visits, No.	Odds ratio (95% CI)	P value
Visit documentation			
No scribe	5075	1 [Reference]	
Ambient AI scribe	5076	0.83 (0.72-0.95)	.006
Human virtual scribe	5075	0.97 (0.85-1.11)	.70
No scribe (prior year)	5076	0.94 (0.82-1.07)	.34
Age at visit, y	20302	0.99 (0.99-0.99)	<.001
Race			
Asian	2274	1 [Reference]	
Black	1333	1.14 (0.88-1.48)	.31
White	14420	1.51 (1.27-1.81)	<.001
Other	2275	1.07 (0.83-1.38)	.59
PHQ-9 score	20302	1.11 (1.10-1.12)	<.001



Key Takeaways about AI scribe notes

Uptake is rapid and reflects real (if modest) benefits: time savings, higher RVUs, burnout reduction

AI-scribed notes are longer but may be clinically less nuanced; reasoning may be lost

Patient voice captured

Automation bias is a concern

Safety and transparency gaps persist: omissions, medication errors, speaker misattribution


Thank you!
Jinoos.Yazdany@ucsf.edu

References

- Castro VM, McCoy TH, Verhaak P, Ramachandiran A, Perlis RH. Psychiatric Documentation and Management in Primary Care With Artificial Intelligence Scribe Use. JAMA Psychiatry. 2026;83(3):281–286. doi:10.1001/jamapsychiatry.2025.4303
- Dai, J., Huang, A., Nasrallah, C., Croci, R., Soleimani, H., Pollet, S. J., Adler-Milstein, J., Murray, S. G., Yazdany, J., & Chen, I. Y. (2025). Patient safety risks from AI scribes: Signals from end-user feedback. arXiv. https://doi.org/10.48550/arXiv.2512.04118
- Doximity State of AI in Medicine Report 2026. https://www.doximity.com/reports/state-of-ai-medicine-report/2026
- Holmgren AJ, Fenton CL, Thombley R, et al. Ambient Artificial Intelligence Scribes and Physician Financial Productivity. JAMA Netw Open. 2026;9(1):e2553233. doi:10.1001/jamanetworkopen.2025.53233
- Stults CD, Deng S, Martinez MC, et al. Evaluation of an Ambient Artificial Intelligence Documentation Platform for Clinicians. JAMA Netw Open. 2025;8(5):e258614. doi:10.1001/jamanetworkopen.2025.8614

AI in the EHR

Protecting People, Organizations, and Ecosystems

Jodyn Platt, PhD, MPH

UNIVERSITY OF MICHIGAN

TIERRA
TRUST, INNOVATION & ETHICS
RESEARCH FOR RESPONSIBLE AI

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

NASEM

April 6-7, 2026

Washington, D.C.


STORY

MIT Technology Review

JUNE 11, 2025

Inside Amsterdam's High-Stakes Experiment To Create Fair Welfare AI


AP

WORLD U.S. POLITICS SPORTS ENTERTAINMENT BUSINESS SCIENCE FACT CHECK ODDITIES

LIVE: Trump administration | Gary 'Mani' Mounfield dies at 63 | Stock market today | Dick Cheney funeral | UPS plane crash

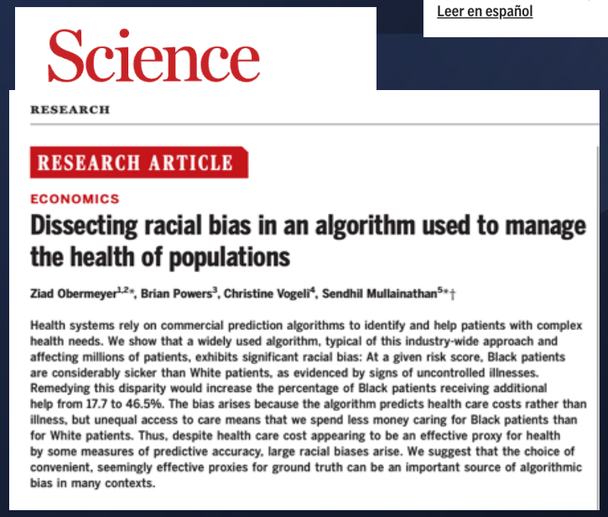
BUSINESS

Study says AI chatbots need to fix suicide response, as family sues over ChatGPT role in boy's death

BY MATT O'BRIEN AND BARBARA ORTUTAY

Updated 2:24 PM EST, August 26, 2025

Leer en español


Science

RESEARCH

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

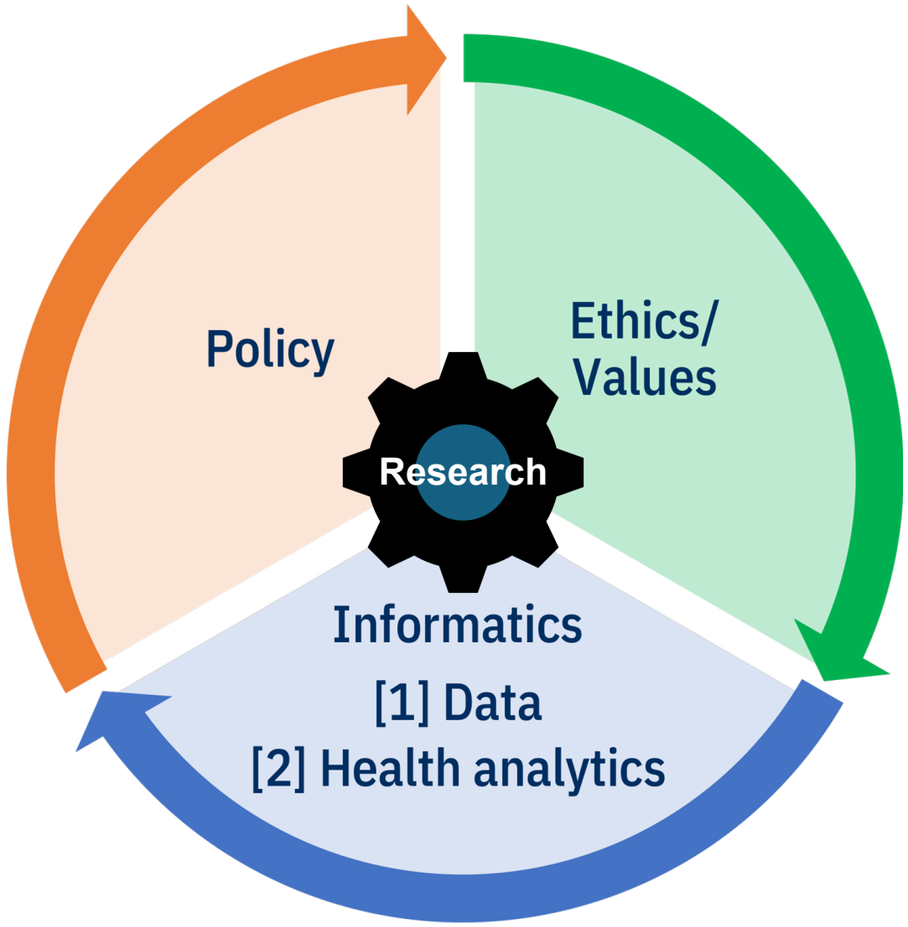
- Broaden the risk lens
  - NIST AI Risk Management Framework
- Present evidence-based insights from patients, clinicians, organizations
- Consider socioecological approaches for feedback to promote adaptable and learning systems

About our work

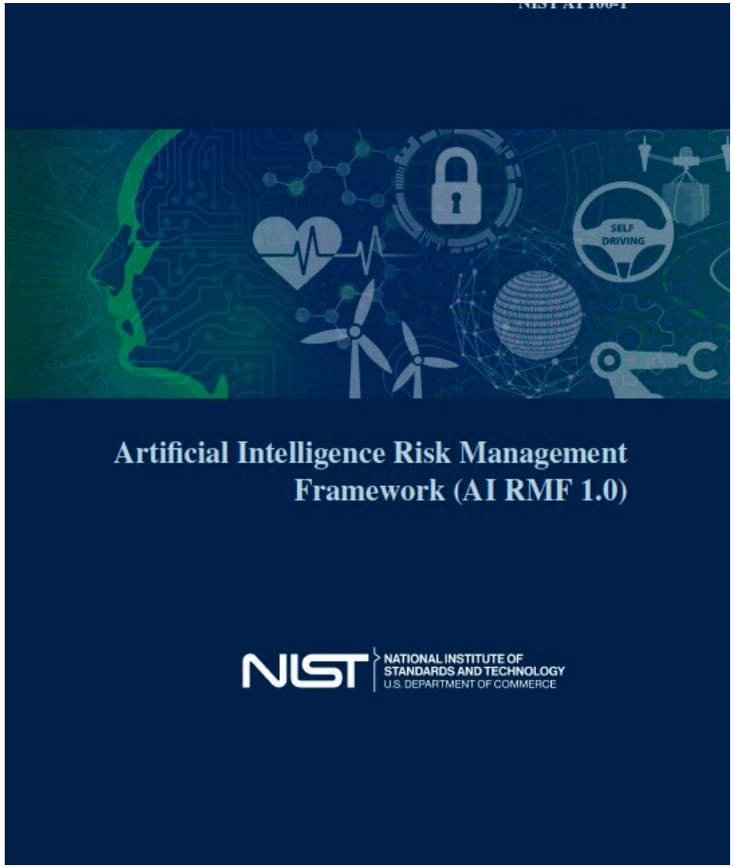
TIERRA
TRUST, INNOVATION & ETHICS
RESEARCH FOR RESPONSIBLE AI

We conduct **research**, understand **best practices,** develop **roadmaps** and evaluate and recommend **policies** regarding the responsible development and applications of AI in healthcare settings

Funding: University of Michigan Department of Learning Health Sciences, Institute for Health Policy and Innovation, ABIM Foundation, NIH (5R01EB030492, 5R01 CA214829)



Risk is ecological, not just technical

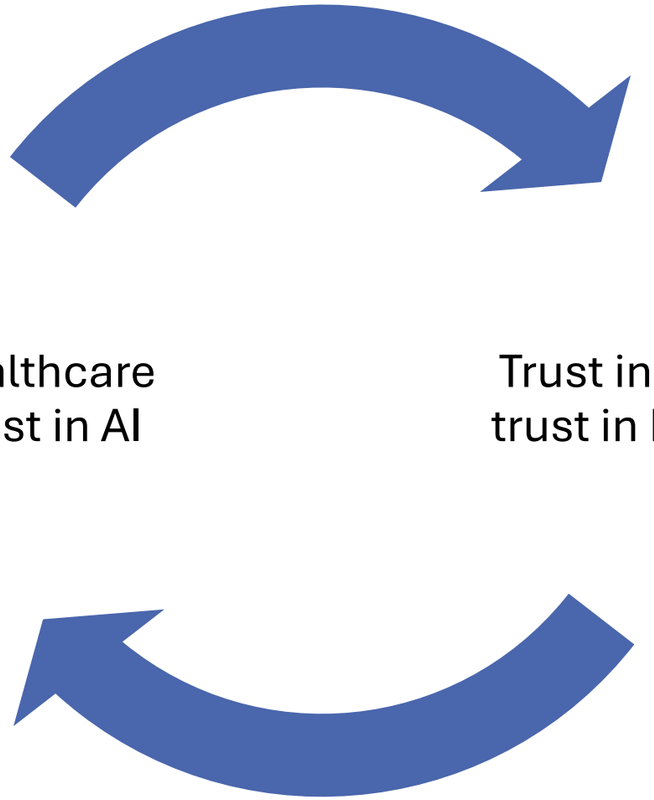


- Risks to people
  - Individuals (patients, clinicians)
  - Communities/ Groups
  - Society
- Risks to organizations
- Risks to ecosystems
  - Interrelated systems
  - Environment

What evaluation should assess

- Accountable and transparent
- Valid and reliable
- Safe
- Secure and resilient
- Explainable and interpretable
- Privacy-enhanced
- Fair with harmful bias managed

Ecosystems



Trust in healthcare shapes trust in AI

Trust in AI shapes trust in healthcare

Trust in health systems to use AI responsibly or not to harm patients is low[1]

Patient trust in AI is associated with patient trust in clinicians and systems[2]

Patients are concerned about the **loss of human connection and diminishing role of the provider**[3]

[1]Nong P, Platt J. Patients' Trust in Health Systems to Use Artificial Intelligence. JAMA Netw Open. 2025 Feb 3;8(2):e2460628.

[2]Nong P, Ji M. Expectations of healthcare AI and the role of trust: understanding patient views on how AI will impact cost, access, and patient-provider relationships. Journal of the American Medical Informatics Association. 2025 May;32(5):795-9.

[3] Ryan KA, Sielaff ML, Saleem D, Richardson J, Tan S, Hamasha R, Nong P, Kardia SL, Romanov V, Hammad A, Platt J. Community perspectives on health AI: hopes, concerns and implications for health systems and trustworthy AI. AI and Ethics. 2026 Apr;6(2):176.

Guardrails do not meet patient priorities

Patient concerns

- **Greater discomfort** with use of **administrative** applications as compared to **clinical applications**[1]

- Discrimination (-)
- Clear privacy policies (+)
- Having health insurance (+)

What patients want

- 95.2% of people **want to be notified about the use of AI in healthcare**[2]
  - For 62.7% it's very important

- Patients want to know about
  - Privacy protections[3]
  - Bias and fairness[3]
  - Safety and efficacy[3]

[1]Nong P, Adler-Milstein J, Platt J. How patients distinguish between clinical and administrative predictive models in health care. The American Journal of managed care. 2024 Jan;30(1):31.

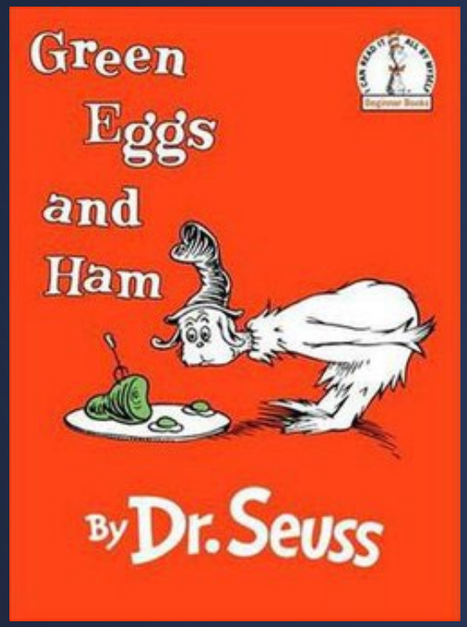
[2]Platt J, Nong P, Carmona G, Kardia S. Public Attitudes Toward Notification of Use of Artificial Intelligence in Health Care. JAMA Network Open. 2024 Dec 2;7(12):e2450102-.

[3]Sielaff ML, Platt J, Tan S, Ryan KA, Nong P, Kardia SL. Building Trust: Public Priorities for Health Care AI Labeling. The American journal of managed care. 2026 Jan 1;32(1):e18.

Validity and reliability: Is “mostly right” OK?

AI iteration of pain with lifting the leg:

[Patient] reports no pain when lifting her leg while getting in and out of a car. She reports no pain when lifting her leg while getting on and off a bus. She reports no pain when lifting her leg while getting on and off a motorcycle. She reports no pain when lifting her leg while getting on and off a scooter. She reports no pain when lifting her leg while getting on and off a skateboard. She reports no pain when lifting her leg while getting on and off a surfboard. She reports no pain when lifting her leg while getting on and off a snowboard. She reports no pain when lifting her leg while getting on and off a kayak. She reports no pain when lifting her leg while getting on and off a canoe. She reports no pain when lifting her leg while getting on and off a raft. She reports no pain when lifting her leg while getting on and off a paddleboard. She reports no pain when lifting her leg while getting on and off a rowboat. She reports no pain when lifting her leg while getting on and off a sailboat. She reports no pain when lifting her leg while getting on and off a yacht. She reports no pain when lifting her leg while getting on and off a cruise ship. She reports no pain when lifting her leg while getting on and off a ferry. She reports no pain when lifting her leg while getting on and off a barge. She reports no pain when lifting her leg while getting on and off a tugboat. She reports no pain when lifting her leg while getting on and off a fishing boat. She reports no pain when lifting her leg while getting on and off a speedboat. She reports no pain when lifting her leg while getting on and off a jet ski. She reports no pain when lifting her leg while getting on and off a water taxi. She reports no pain when lifting her leg while getting on and off a gondola. She reports no pain when lifting her leg while getting on and off a cable car. She reports no pain when lifting her leg while getting on and off a funicular. She reports no pain when lifting her leg while getting on and off a monorail. She reports no pain when lifting her leg while getting on and off a tram. She reports no pain when lifting her leg while getting on and off a trolley. She reports no pain when lifting her leg while getting on and off a streetcar. She reports no pain when lifting her leg while getting on and off a light rail. She reports no pain when lifting her leg while getting on and off a commuter train. She reports no pain when lifting her leg while getting on and off a high-speed train. She reports no pain when lifting her leg while getting on and off a bullet train. She reports no pain when lifting her leg while getting on and off a maglev train. She reports no pain when lifting her leg while getting on and off a freight train. She reports no pain when lifting her leg while getting on and off a cargo train. She reports no pain when lifting her leg while getting on and off a passenger train. She ...

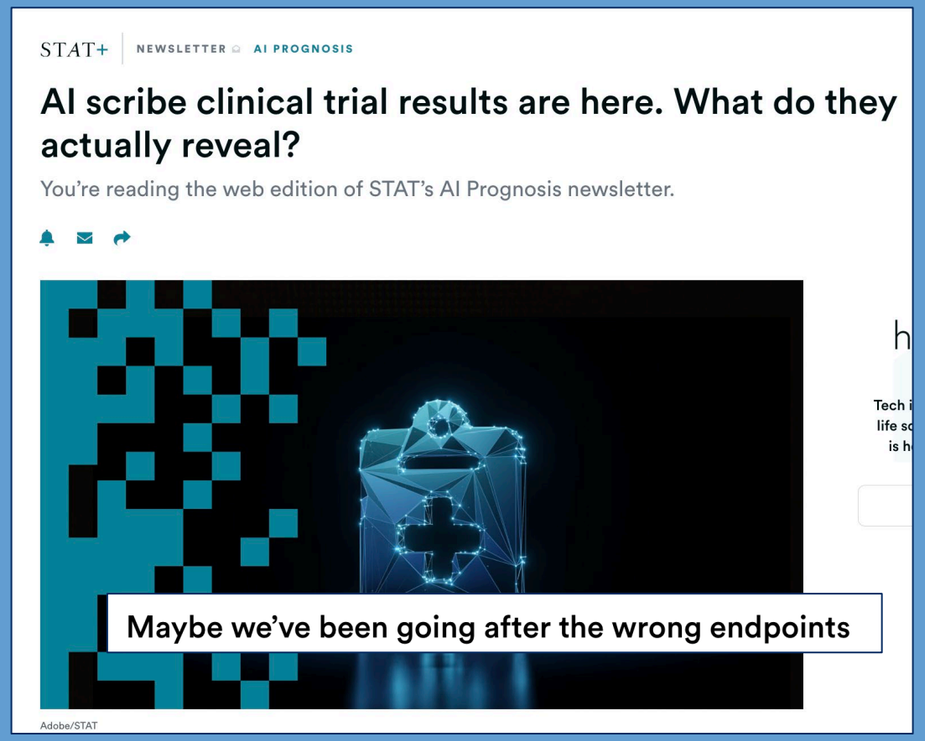


”Any hallucinations in the medical record is too many”

(physician)

Credit: Michael Burns, MM – CMIO-AI

Uncertain value



What we see

- Modest time savings
- Increases in visit volumes
- Increases in RVUs

Rotenstein LS, Holmgren AJ, Thombley R, et al. Changes in Clinician Time Expenditure and Visit Quantity With Adoption of Artificial Intelligence–Powered Scribes: A Multisite Study. *JAMA.* Published online April 01, 2026. doi:10.1001/jama.2026.2253

Sasseville, M., Yousefi, F., Ouellet, S., Naye, F., Stefan, T., Carnovale, V., Bergeron, F., Ling, L., Gheorghiu, B., Hagens, S. and Gareau-Lajoie, S., 2025, June. The Impact of AI Scribes on Streamlining Clinical Documentation: A Systematic Review. In *Healthcare* (Vol. 13, No. 12, p. 1447).

Holmgren AJ, Fenton CL, Thombley R, et al. Ambient Artificial Intelligence Scribes and Physician Financial Productivity. *JAMA Netw Open.* 2026;9(1):e2553233. doi:10.1001/jamanetworkopen.2025.53233

What we don't yet account for

- Variability in Ambient tech
- Impact on quality?
- Impact on health outcomes?
- Patient experience?
- Performance across patient populations?

Shah SJ, Garcia P. Ambient AI Scribes—What Is the Return on Investment? *JAMA Netw Open.* 2026;9(1):e2553238. doi:10.1001/jamanetworkopen.2025.53238

Tierney AA, Lee K, Liu VX. Ambient AI Scribes and the Quintuple Aim: What Is Counted—and What Matters. *JAMA.* Published online April 01, 2026. doi:10.1001/jama.2026.3529

Adoption of AI outpaces evaluation

- 65% of hospitals use AI
- 79% EHR-based AI
- 65% evaluate for accuracy
- 44% evaluate for bias

HEALTH INFORMATION TECHNOLOGY

By Paige Nong, Julia Adler-Milstein, Nate C. Apathy, A. Jay Holmgren, and Jordan Everson

DOI: 10.1377/hlthaff.2024.00842
HEALTH AFFAIRS 44, NO. 1 (2025): 90–98

Current Use And Evaluation Of Artificial Intelligence And Predictive Models In US Hospitals

Paige Nong (nong0016@umn.edu), University of Minnesota, Minneapolis, Minnesota.

Julia Adler-Milstein, University of California San Francisco, San Francisco, California.

Nate C. Apathy, University of Maryland, College Park, Maryland.

A. Jay Holmgren, University of California San Francisco.

Jordan Everson, Office of the Assistant Secretary for Technology Policy, Washington, D.C.

ABSTRACT **Effective evaluation and governance of predictive models used in health care, particularly those driven by artificial intelligence (AI) and machine learning, are needed to ensure that models are fair, appropriate, valid, effective, and safe, or FAVES. We analyzed data from the 2023 American Hospital Association Annual Survey Information Technology Supplement to identify how AI and predictive models are used and evaluated for accuracy and bias in hospitals. Hospitals use AI and predictive models to predict health trajectories or risks for inpatients, identify high-risk outpatients to inform follow-up care, monitor health, recommend treatments, simplify or automate billing procedures, and facilitate scheduling. We found that 65 percent of US hospitals used predictive models, and 79 percent of those used models from their electronic health record developer. Sixty-one percent of hospitals that used models evaluated them for accuracy using data from their health system (local evaluation), but only 44 percent reported local evaluation for bias. Hospitals that developed their own predictive models, had high operating margins, and were health system members were more likely to report local evaluation. Policy and programs that provide technical support, tools to assess FAVES principles, and educational resources would help ensure that all hospitals can use predictive models safely and prevent a new organizational digital divide in AI.**

Evaluation capacity varies

AI Digital divide

Disparity in capacity between organizations or health systems that can safely and effectively use AI while others cannot

Make existing population health disparities worse

Entrench inequalities in care delivery at the organizational level

Worsen gaps in access to high-quality care for individual patients

Nong P, Adler-Milstein J, Apathy NC, Holmgren AJ, Everson J. Current use and evaluation of artificial intelligence and predictive models In US hospitals. *Health Affairs*. 2025;44(1):90-98

Nong P, Maurer E, Dwivedi R. The urgency of centering safety-net organizations in AI governance. NPJ Digit Med. 2025 Feb 21;8(1):117.

Ecologies of risk

- Developing feedback mechanisms for adaptable and learning systems
- Risk = [Magnitude/ Severity] * Likelihood
- Iterative processes and clear milestones

“When a measure becomes a target, it ceases to be a good measure”
(Strathern/ Goodhart)

It’s about people

- Align policy with risk defined by people
- Adequate human oversight of technology to monitor for financial, health, and other consequences for patients
- Investment in people’s expertise and learning
- People are responsible for quality
- AI in healthcare is a team sport

AI disclosure: AI used in this presentation to edit for clarity