The AI Catalyst Pulse

November 13th, 2024

Upcoming AI Catalyst Webinars

- **November 14th: Latent, AKASA and AI-Powered Prior Authorization Management**
  - Join our candid conversation with executives from Latent and AKASA to learn about their AI tools and future roadmaps for prior authorization management. Register now

- **November 21st: Deep Dive: Proven AI Use Cases to Revitalize Bedside Nursing**
  - Explore the latest innovations in AI adoption for bedside nursing, focusing on how AI is allowing for better patient and operational outcomes within inpatient settings. Register now

Are AI Clinical Documentation Tools Really 'Making Up Text'?

What happened: An *Associated Press* investigation has raised alarms about OpenAI's Whisper transcription tool "making up chunks of text or even entire sentences." The report highlighted Whisper's integration into Nabla's clinical documentation tool, which is used by "over 30,000 clinicians and 40 health systems." Researchers cited by the AP warned that Whisper hallucinations could include fabricated medications, racial commentary, and violent rhetoric.

Much of the popular news coverage of the investigation struck an alarmist tone. As *Engadget* wrote, "Imagine going to the doctor, telling them exactly how you're feeling, and then a transcription later adds false information and alters your story."

Why it matters: In many ways, there's less to this controversy than the headlines suggest. Most importantly, the investigation did *not* examine actual clinical visit transcripts or identify any instances of patient harm.

Instead, researchers analyzed Whisper transcripts of recordings from AphasiaBank, a research database of recordings from individuals with aphasia, who the researchers noted "have a lowered ability to express themselves using speech and voice." Even with these particularly challenging speech samples, fewer than 2% of transcripts included hallucinations – most of which were harmless. While this methodology certainly illuminates Whisper's limits, it hardly represents real-world clinical use.

Even so, the controversy illustrates several challenges for health systems deploying AI tools:

- **Even well-known AI vulnerabilities can quickly become scandals.** When OpenAI released Whisper in 2022, they explicitly acknowledged that the outputs "may include texts that are not actually spoken in the audio input (i.e. hallucination)" because "the models combine trying to predict the next word in audio with trying to transcribe the audio itself." This wasn't a hidden flaw – it was a known challenge that vendors like Nabla have invested significant resources to address. Even so, the tone of the news coverage suggests that healthcare organizations are recklessly deploying unsafe AI. The disconnect illustrates how even carefully managed AI risks can become fodder for controversy.

- **AI hallucinations may pop up in even seemingly "easy" use cases.** You might expect hallucinations when asking AI to summarize complex medical literature or suggest treatment plans. But transcription? That's supposed to be the easy part. Yet researchers found Whisper expanding on innocuous statements in alarming ways. In one case, "He, the boy, was going to, I'm not sure exactly, take the umbrella" became "He took a big piece of a cross, a teeny, small piece... I'm sure he didn't have a terror knife so he killed a number of people." While technical experts have long understood the risks of AI hallucinations, many healthcare leaders may not realize just how deeply such vulnerabilities are woven into even basic AI capabilities.

- **AI tools in healthcare will face extreme public scrutiny – so they need to clear a high bar for reliability.** Whisper is widely used across industries: As the Associated Press reports, "In the last month alone, one recent version of Whisper was downloaded over 4.2 million times from open-source AI platform HuggingFace." Even so, media attention has focused almost exclusively on its healthcare applications, illustrating the heightened scrutiny AI faces in clinical settings

- **Technical diligence isn't enough; you need a communications strategy to persuade stakeholders of your AI tools' safety.** Nabla took substantial precautions: They spent $5M training their model on medical conversations, implemented clinician review processes, built integrated feedback tools, and established continuous monitoring of errors. Their implementation of Whisper also includes "improvements specifically developed to suppress hallucinations and make the transcription of medical terms more accurate than any off-the-shelf speech-to-text engine." We simply can't say, with the evidence at hand, whether Nabla's precautions have been 100% effective – but these events make clear that technical mitigations alone aren't enough to prevent critical news coverage. Health systems must also build trust with patients and stakeholders who may approach AI with deep skepticism.

- **Strikingly, Nabla's decision to delete visit recordings – intended to preserve patient privacy – is now facing criticism.** Nabla chose not to retain audio recordings of clinical visits, prioritizing patient privacy. Now that same decision is being questioned for making it impossible to verify their transcripts' accuracy. It's a stark illustration of the trade-offs health systems face: Prioritize privacy too strongly, and you'll be attacked for lacking verification capabilities. Focus too heavily on verification, and you risk privacy concerns.

Related AI Catalyst resource: To help better understand the vulnerabilities of your AI vendors, download the University of Vermont Health Network's AI vendor questionnaire, which specifically asks vendors to identify any commercially available or open-source models they use.

Questions to consider:

1. Beyond tracking sophisticated AI applications, how are you monitoring seemingly basic functions like transcription for potential vulnerabilities?

2. How do your privacy protection measures affect your ability to verify AI outputs? What documentation practices could help thread this needle?

3. What concrete steps is your organization taking to build stakeholder trust in AI implementations? How are you measuring the effectiveness of these efforts?

WellSpan Nurses Said 'Yes' to Hippocratic AI. Here's Why.

What happened: WellSpan Health has deployed Ana, a virtual assistant from Hippocratic AI, to support two key clinical workflows. The move comes amid broader industry discussion about Hippocratic's technology – which made headlines earlier this year when the company faced criticism for comparing its AI agents' costs ($9/hour) to human nurse wages.

When we recently spoke to WellSpan stakeholders, here's how they told us they've used Ana so far:

- **Colonoscopy preparation:** Ana conducts scheduled check-ins with low-risk patients at key intervals: 10 days before, 3 days before, and the day after the colonoscopy, ensuring patients understand instructions and are preparing appropriately.

- **FIT test screening and outreach:** Ana cold-calls existing WellSpan patients who need a colorectal cancer (CRC) screening and meet the criteria for a FIT testing but do not have access to WellSpan's MyChart portal where patients usually get communication about CRC screening.

Early outcomes include:

- **Reduced nursing call time**: Out of WellSpan's initial 100 trial cases, only two were escalated for care team intervention and nurse involvement. As of November, WellSpan has completed 1,075 calls with patients with a 1.5% human escalation rate.

- **Strong patient satisfaction:** Among patients who completed the call, Ana received an average rating of 8.6/10.

- **Multilingual communication:** Ana currently interacts in both English and Spanish, with 10 more languages under development. Ana's multilingual capabilities help create pathways to engage patients who historically have been underserved, a driving factor for a health system focused on addressing health disparities. Compared to English-speaking patients, Spanish-speaking patients had a 1.8x call duration and were 2.6x more likely to request a FIT test.

Why it matters: As Hippocratic AI and similar vendor tools roll out nationwide, WellSpan's experience offers several lessons learned:

- **To avoid pushback from care teams, involve them early in choosing the right use cases for AI agents.** Despite earlier industry debate over Hippocratic AI's broader claims, WellSpan found success by identifying specific, well-defined applications through close collaboration with nursing staff. "It has been a really collaborative process," notes Kasey Paulus, WellSpan's SVP and CNE. "It was our team that suggested the colonoscopy use case, and they thought it was a really viable place to start."

- **Despite fears that AIs will eliminate jobs, WellSpan's experience shows how AI can help nurses focus on top-of-license tasks.** According to Paulus, colonoscopy prep represents a "low risk way to enter" the space of AI-enabled care while reaching patients who, historically, nurses often lacked capacity to serve. As Paulus says, "It's about resourcing something that we weren't [otherwise] able to resource," ensuring nurses have the bandwidth to focus on more urgent, top-of-license bedside care.

- **AI can drive better patient communication – which, in turn, can improve both process and outcomes down the road.** WellSpan's reduction in patient callbacks and misinformation may seem like small wins, but in the long run, they're likely to pay off in significant improvements both in health outcomes and the volume of bounce-back patients, part of the system's value-based

care strategy

- **Patients are surprisingly happy to talk to an AI tool.** Despite initial skepticism about patient-facing AI, user feedback has been consistently positive across implementations. In addition to WellSpan's 8.6/10 patient satisfaction score, Virtua Health received strong patient feedback earlier this year after partnering with Woebot Health to implement chatbot support for patients with mild-to-moderate depression and anxiety. Virtua reported a 94% satisfaction score, with some patients even preferring chatbots over human support.

Questions to consider:

1. What patient communication tasks in your organization are you currently underserving – whether due to staffing constraints, language barriers, or other limitations – and how could tools such as Hippocratic AI help you reach more patients effectively?

2. How can you involve nursing stakeholders in choosing the right initial implementation areas for flexible AI agent tools?

3. If patients prove to be open to interacting with AI "agents" and chatbots, what new opportunities might that create for AI deployment?

A Provocative Study Suggests 'AI Alone' May Outperform 'AI + Humans'

What happened: For a new study published in JAMA Network Open, researchers compared the diagnostic reasoning of the large language model (LLM) GPT-4 to that of human physicians with surprising results. When they tested GPT-4 operating on its own, it scored a median score of 92% on diagnostic reasoning assessments. That's significantly higher than human physicians operating on their own (74%) – and, more surprisingly, it even beat "human + AI" teams of physicians using GPT-4 (76%).

As study co-author Jonathan Chen noted, this result "flies in the face of the Fundamental Theorem of Informatics," the longstanding idea that "human + AI" teams should outperform either acting alone.

Why it matters: Let's pause first to clarify exactly what the study did, and did not, examine. Most importantly, this study did *not* test the real-world accuracy of AI vs. physician diagnoses. Rather, it tested the skill known as "diagnostic reasoning," which the authors described as "diagnostic performance based on differential diagnosis accuracy, appropriateness of supporting and opposing factors, and next diagnostic evaluation steps."

Even so, the findings raise challenging questions about why "AI + human" would perform worse than "AI alone" – and what that implies for AI's future in diagnostics. Our takeaways:

1. **AI performed very well when given structured, tested prompts – and much worse when prompted informally.** When testing "AI alone," researchers used carefully constructed prompts following what they called "established principles of prompt design." For instance, one prompt began: "You are an expert internal medicine physician solving a complex medical case for a test... I want you to give three parts of information..."

   But in the "AI + human" condition, clinicians apparently received an unstructured interface resembling standard ChatGPT, with no specific training. That's a bit like giving a surgeon access to a robot-assisted surgical system and saying, "Hey, you figure it out." Yes, these clinicians had *access* to advanced technology … but absent any guidance, it's not shocking they failed to use it

effectively. (That said, some real-world clinicians are probably already using ChatGPT in this informal way, and this study certainly reveals the limitations of that approach.)

2. **Doctors who have plenty of AI experience didn't outperform the novices.** The study asked how frequently they used LLMs, with answers ranging from "I've never used it before" to "I use it frequently (once a week or more)." Strikingly, clinicians who reported regularly using these tools *didn't* perform better than those who rarely or never used AI tools.

   That's worrying. One might imagine that clinicians who use AI regularly will naturally learn to compensate for its faults … but at least in this study, that didn't happen.

3. **Keeping a "human in the loop" might not be a permanent solution to your AI governance challenges.** While human oversight is traditionally viewed as crucial for AI safety, this study suggests an alarming possibility: In some cases, human involvement may actually degrade AI performance. As Ethan Mollick, an AI expert at the Wharton School who studies the integration of AI into organizational processes, wrote, this foreshadows "the coming problem of working with AI when it starts to match or exceed human capability."

   Let's be clear: We aren't there yet. There's no evidence that widely deployed AI healthcare tools make better decisions alone than with human support. Still, these findings suggest health systems need to begin weighing some very hard problems. When does human oversight truly improve outcomes? When might it make them worse? And how can you develop frameworks that maximize AI's benefits while maintaining appropriate oversight?

Questions to consider:

1. How will your organization identify specific use cases where reduced human oversight of AI might improve care quality, while maintaining appropriate safety guardrails?

2. What frameworks could help clinicians understand when to defer to AI insights, particularly in cases where the AI's reasoning appears to conflict with clinical judgment?

3. How might your AI implementation strategy shift from focusing on individual tool access to developing standardized, organization-wide approaches to AI interaction?

AI Strategy Quick Hits

Noteworthy moves from peers to implement AI technologies

The Coalition for Health AI (CHAI) has unveiled its first template for an AI "model card," intended to serve as a standardized "nutrition label" for healthcare AI applications. The template, demonstrated using Aidoc's intracranial hemorrhage triage system, provides detailed information about model capabilities, limitations, training data, and bias mitigation efforts.

But calling this a "nutrition label" might be an overstatement. While CHAI has succeeded in standardizing important disclosures, the resulting document remains quite long and technical. Consider the card's description of bias mitigation approaches, which describes analyzing "demographic and clinical characteristics (e.g., manufacturer, slice thickness, modality model distribution)." That's far more complex than the simple percentages and metrics on the back of a cereal box.

That's not to say this complexity is CHAI's fault – it may simply reflect that AI capabilities and risks remain too nuanced for simple disclosure. But it reinforces an uncomfortable truth: There's no shortcut

around developing real AI literacy within your organization. Even with standardized documentation, health systems still need internal expertise to evaluate and implement AI tools effectively.

New efforts to involve nurses more heavily in AI development:

- **Florida State** nursing leaders are advocating for nurse involvement in the development of AI technologies in healthcare
- **Stanford Health Care** collaborates with **Microsoft** and **Epic** to develop ambient listening AI tool for nurses
- How AI technology is transforming advanced nursing education

Organizations unveil their long-term AI integration plans:

- **NewYork-Presbyterian** launches $2 billion campaign called "For Every Future," aimed at leveraging AI and digital innovation
- **HCA** unveils their AI implementation plan across the next five to seven years
- How **Community Health Network** is approaching AI governance

Notable collaborations among health systems and vendors:

- **East Alabama Medical Center (EAMC)** collaborates with **Inflo Health** to implement AI-driven radiology follow-up care orchestration
- **Abridge** and **OpenNotes** collaborate with **Beth Israel Deaconess Medical Center** to enhance patient communication
- **Intermountain Children's Health** partners with **Fabric Genomics** to utilize AI for genome sequencing

Other strategy quick hits:

- **Northwell Health** launches AI Hub that utilizes LLMs for administrative tasks, translation, and potential diagnostics
- **Emory Healthcare** is leveraging ambient listening and AI cameras to address workforce shortages and improve patient care
- How **East Alabama Medical Center** and **Stamford Health** are leveraging AI imaging tools
- **Athenahealth** launches Ambient Notes clinical documentation tool
- **Kaiser Permanente** integrates AI-driven ambient listening devices
- **Aspira Women's Health** announces $10M award from **ARPA-H** for its AI-enabled blood test, ENDOinform™, aimed at improving the diagnosis of endometriosis

Emerging Use Cases

New capabilities that indicate AI's potential

Newest in AI cancer detection/treatment:

- **Feinstein Institute for Medical Research** has two inventions recognized by TIME's Best Inventions of 2024 for cancer detection and paralysis treatment
- **Mass General Brigham** researchers develop AI model that can accurately identify and measure prostate cancer lesions
- **Ohio State University Comprehensive Cancer Center** uses supervised AI to help pathologists detect tumors
- **Vanderbilt University Medical Center** uses AI in precision immunotherapy

Other emerging use cases:

- **University of Michigan** researchers develop algorithm to counteract racial bias in medical data for AI diagnostics

- AI algorithm significantly improves accuracy and reporting time of pneumothorax detection
- **UC San Diego** researchers find that LLM's can accurately process hospital quality measures
- 10 ways AI, robotics are improving pediatric care in real-world settings
- The **Regenstreif Institute** uses tools like **OneSheet** to enhance chronic pain management and opioid prescribing
- **Windsor Regional Hospital** implements AI-based weapons detection system to reduce instances of workplace violence

Cautionary Tales

The risky side of AI implementation

- Recent **University of Michigan** study highlights potential racial bias in medical AI models
- **UW-Madison** finds issues with the use of AI in genomic research

Market moves

A round-up of AI company announcements and stories

Oracle has announced what it's calling a "next-generation EHR," with claims of AI capabilities embedded across clinical workflows. While the announcement emphasized features like voice-driven navigation and conversational search, many of the promised AI capabilities – from automated documentation to personalized care plans – won't be available until at least 2025, when an early adopter program begins.

The focus on voice-based interfaces, which Oracle says will let clinicians access patient information like vitals and medications through simple voice commands, may offer an early preview of how EHR interfaces will evolve. Epic has indicated similar capabilities are in its pipeline, suggesting EHR vendors see the potential for AI to conquer clinician "click fatigue" once and for all.

More details from Oracle:

- **Oracle** unveils new next-generation EHR with embedded AI capabilities
- **Oracle** introduces new improvements to their Clinical AI Agents
- **Oracle's CEO, Safra Catz**, discusses Oracle's transformation in healthcare

Latest in vendor funding:

- **Ferrum Health** raises $16M in Series A funding
- **Pieces Technologies** receives $2M grant from **National Cancer Institute** to develop conversational AI agent, 'PiecesChat'
- **General Catalyst** invests $1B into healthcare companies
- **Onc.AI** is awarded $2M grant from the **National Cancer Institute** to improve its AI tool for managing non-small cell lung cancer
- **CredibleMind** secures $7.5M in Series A funding to enhance its AI-driven resource curation and improve access to mental health services
- Digital health startup, **MDisrupt**, receives $1M investment from the **American Heart Association**

New product launches:

- **SmarterDx** launches SmarterDenials to help health systems combat rising claim denials
- New AI model, **BiomedGPT**, is introduced to enhance medical practices and research
- **Artera** launches AI co-pilots to improve patient communication with providers

Other market moves:

- **Abridge** recognized by **Time** as one of the best inventions of 2024
- **Elon Musk** pushes for **Grok AI** to have a place in healthcare
- **OpenAI** is actively hiring roles focused on healthcare to enhance its capabilities

Policy Updates

Understanding the evolving AI regulatory and legislative landscape

- **The Department of Veterans Affairs** and the **Food and Drug Administration** plan to launch intergovernmental Health AI Laboratory (HAIL) to test safety of AI applications
- **Ochsner Health's Jason Hill** urges for standardized frameworks to ensure safe AI use
- **NCQA** and **URAC** update guidelines on responsible AI use and telehealth

Expert Insights

For further reading, articles, videos, and podcasts that we found insightful

- AJMC, *"Harnessing AI for Population Health: A Call to Action for Policy Makers and Health Care Leaders"*
- HealthcareITNews, *"Why won't this expert's clients sign onto AI projects for more than 12 months at a time?"*
- Psychology Today, *"Do We Trust AI to Help Make Decisions for Mental Health?"*
- NPR, *"AI has been used in healthcare for decades now. Some say they want more regulation"*