RESEARCH MAY 11, 2026

# 57 HEALTHCARE PROFESSIONALS TOLD US WHAT THEY NEED FROM AI

By Sami Hassaan, Oscar Kavanagh

At Scale, we're building AI evaluations for healthcare and life sciences. To refine our understanding of what this work requires on the ground day-to-day, we surveyed 47 clinical care professionals (physicians, pharmacists, nurses across 18 specialties) and 10 who specialize in revenue cycle management and administration. We also conducted deep-dive interviews with practitioners in radiology, hematology, and emergency medicine.

We found three capability gaps that healthcare AI evaluations need to measure:

1. **Context assembly from fragmented systems**. Healthcare AI needs to retrieve the right evidence across notes, labs, medication histories, imaging reports, outside records, local guidelines, and patient-specific context. The hard part is reconciling incomplete, duplicated, or conflicting information into a reliable clinical picture.
2. **Hidden judgment inside "simple" tasks**. Many tasks look easy to specify: summarize the chart, check the lab, verify the medication, draft the referral. That makes them tempting to evaluate as retrieval, summarization, or rule-following tasks. But clinicians described these as judgment-heavy because the correct output depends on patient trajectory, baseline health, institutional rules, and risk profile. A model may pass the surface task while missing why this patient is an exception.
3. **Verifiable outputs**. 83% of clinical respondents said an AI hallucination in their domain could directly harm a patient if not caught. Healthcare AI needs to produce cited sources, flagged uncertainty, and reasoning that clinicians can check efficiently. In this setting, an output that cannot be verified is not useful, even if it sounds clinically fluent.

*(For details on survey structure and methodology, see the note at the end of this post.)*

## The Clinical Care Landscape We Studied

Our clinical respondents work across emergency medicine, surgery, radiology, oncology, cardiology, pediatrics, critical care, and other specialties, in hospitals, outpatient clinics, academic medical centers, private practices, and a VA hospital. Experience ranges from early-career to 20+ years.

## What Clinicians Ask AI

We asked each clinical respondent for three prompts they would give an AI assistant with full access to their systems. We got 141 prompts back, and most were built on the same foundation: pulling together data from multiple systems. Some stopped there, while others went further to cross-reference guidelines, draft clinical documents, or make clinical recommendations.

### Six Capabilities

To put structure on what these 141 prompts ask for, we coded each one against six capabilities. Most prompts combine several, with an average of about three per prompt.

[Table: see online version]

The top two capabilities form the dominant workflow shape: about 58% of prompts ask the model to both assemble data from multiple sources and produce a structured output. Clinical reasoning is rarely the whole task; in 89% of prompts that require domain judgment, it's paired with context assembly or artifact generation. The smaller categories still matter for evaluation; temporal reasoning, exception handling, and rule checking each appear often enough that benchmarks missing them will underrepresent the actual work.

### Example Prompts

Below are 10 prompts from the 141, chosen to span the capability space. Each is labeled with the capabilities it invokes.

[Table: see online version]

Even the prompts that touch on clinical reasoning (identifying candidates for antibiotic de-escalation, generating differentials) assume a physician will review the output and that the AI will have already pulled together the relevant patient data to reason over.

Part of the assembly problem is infrastructure: EHR interoperability, data standards, the fact that two hospitals ten minutes apart may not share records. No model fixes that, but even with full data access, the synthesis itself is a challenging AI capability problem: reconciling conflicting medication lists, tracking a patient's trajectory across dozens of encounters, and producing an output a clinician can verify in 30 seconds.

## Deceptive Complexity

We asked clinicians: "What's a task in your work that looks simple from the outside but is actually complex?" The common thread across their answers was that tasks that look mechanical from the outside require integrating patient context, clinical history, and professional judgment in real time.

**Lab interpretation:** An emergency medicine physician wrote that it "is not just checking what is red/out of range. It involves complex pattern recognition and interpretation to place results in the context of the overall patient presentation and suspected diagnosis." A potassium level that looks alarming might be expected after surgery. A "normal" creatinine might be dangerously abnormal for a patient whose baseline is much lower.

**Medication verification:** An oncology pharmacist described what screening a chemotherapy prescription actually involves: "determining whether the patient should safely receive treatment at that dose at that moment. I need to integrate all information about the patient," including labs, organ function, prior treatment cycles, cumulative toxicity, and drug interactions. To an outside observer, it looks like clicking "approve."

**Specialty referrals:** A surgeon described how a referral requires "the synthesis of significant volume of patient information including images, bloods and patient notes into a 30 second handover. It then requires preparation for any possible questions from the referring specialty." Multiple physicians added that each department has its own referral system, varying by institution.

If a benchmark treats lab interpretation as a lookup task, medication verification as a rule check, or referral writing as template filling, it will pass models that fail at the real work.

## Two Specialties Up Close

### 1. Radiology

A radiologist needs prior scans (sometimes 8-10 for a complex case), the clinical indication for the study, the patient's surgical history, and local protocol rules. Scan requests frequently arrive with minimal clinical context, sometimes as sparse as "abdominal pain, query cause." When that happens, a scan that should take 5 minutes to interpret turns into a 30-minute investigation: the radiologist has to go into the electronic patient record, piece together the relevant clinical picture, and figure out why the scan was ordered before they can interpret it meaningfully.

In our deep-dive interview with a radiologist working within the British NHS (National Health Service), the gap between assumptions about radiology and the actual workflow became clear quickly. He described a typical case pattern: a patient presents with a seizure. CT leads to MRI, which reveals a brain tumor, followed by surgery, then multiple follow-up scans tracking post-operative changes. By the time the patient returns with new symptoms, understanding the current scan requires reviewing the entire longitudinal sequence. As he put it: "The most important scan in radiology is the previous scan."

Radiology benchmarks that test image-to-report in isolation are testing one component of a multi-step workflow. A model can be excellent at image interpretation and still fail the clinical workflow if it can't assemble the pre-read context or reference prior studies.

**Nuance for other specialties**: Radiology and pathology are also the exception when it comes to raw image interpretation. 83% of clinical respondents work in modalities that involve images, but outside these two specialties, their AI prompts almost never asked the model to read pixels; they asked for findings and impressions extracted from PACS or pathology reports, combined with labs, notes, and history. A radiology benchmark needs DICOM images. A discharge, pre-rounding, or medication benchmark mostly needs the imaging *reports*, not the images themselves.

### 2. Emergency Medicine

In the Emergency Department, the objective is to rule out the most dangerous diagnoses, even when they're statistically unlikely. An experienced emergency medicine physician develops this skill through years of seeing thousands of routine presentations until deviations, even subtle ones, trigger deeper investigation.

Our deep-dive with an emergency medicine physician at a VA (Department of Veterans Affairs) hospital revealed a different evaluation gap. She noted that current models tend to optimize for the most probable diagnosis; a model that correctly identifies chest pain as likely musculoskeletal but fails to rule out aortic dissection has gotten the statistics right and the medicine wrong.

She also described a practical reality that benchmark designers need to account for: patients often have poor health literacy ("I take the white pill" without knowing the medication name), records from outside hospitals may arrive as faxed scans or not at all, and electronic information retrieval succeeds only about 75-80% of the time. The rest requires phone calls, family interviews, or manual chart review.

This isn't unique to the ED. A hematologist we interviewed described the same pattern in outpatient care: a single patient's records routinely arrive as a mix of electronic discrete data, PDFs, and scanned faxes, even within the same EHR vendor where institution-level customization makes interfaces and data formats vary widely.

## Documentation as Patient-State Synthesis

70% of clinical respondents named documentation as their most tedious workflow when asked unprompted, which most AI "scribe" products treat it as a transcription problem. However, a clinical note is a compressed representation of patient state, clinical reasoning, treatment plan, handoff context, legal record, and billing support, all at once. One physician described how ward round notes require capturing senior clinicians' insights in real time, where "important details slip through the cracks." A surgeon described how operative notes require specific phrasing to correspond to CPT billing codes, turning documentation into a simultaneous medical, legal, and financial exercise.

The real capability being tested is maintaining an accurate, source-grounded representation of clinical state over time, one that serves multiple audiences with different requirements simultaneously.

## Verification as Trust Model

We asked clinical respondents what the most likely consequence of AI hallucination would be in their domain.

- 39 of 47 selected "Critical safety risk: could directly harm a patient if not caught."
- 38 of 47 said they would need the AI to cite specific source documents.
- 38 of 47 said they would cross-reference multiple sources themselves.
- Only 4 of 47 said they would not act on AI-generated clinical recommendations at all.

When we asked how respondents would actually combine these checks in practice, 45% said they'd glance at the answer with a citation backing it, and 32% said they'd want citations plus cross-referencing multiple sources

themselves. Only 6% said a glance would be enough on its own. Citation is the floor; for roughly a third of workflows, cross-source consistency is required on top.

The failure modes clinicians described went far beyond generic hallucination concerns. They described institutional gaps (applying generic guidelines when local protocols or physician preferences change the correct action) and the board-exam gap (one physician: "AI companies think just because LLMs pass board questions they can also write correct plans. Having a correct plan requires situational awareness, understanding of hospital and patient context").

Clinicians apply the same verification standard to AI that they'd apply to a junior colleague: show your work, cite your sources, and expect it to be checked.

## Beyond the Bedside: Revenue Cycle & Admin

We also ran a smaller revenue cycle and administration survey. Because that sample was smaller, we treat it as directional rather than definitive. But it pointed to a similar evaluation problem outside bedside care: healthcare agents need to reconcile fragmented systems, apply context-specific rules, and produce outputs humans can verify.

In revenue cycle workflows, denials management stood out as the central pain point. Resolving a denial can require pulling together clinical documentation, payer-specific rules, claims data, contract terms, portal information, and appeal deadlines before any action is useful. The risk is different from clinical care: financial loss, compliance exposure, delayed claims, or patient billing harm rather than direct patient safety, but the evaluation lesson is similar. A healthcare agent that cannot operate across messy data sources and show its work will fail in administrative workflows too.

## What This Means for Evaluation Design

If you're building benchmarks, training data, or agentic environments for healthcare AI:

**Test information assembly.** Can the model gather the right context from multiple sources, reconcile inconsistencies, and produce an accurate, verifiable synthesis? Our respondents' prompts suggest this is where the most clinical time goes.

**Build multi-system environments.** Clinical work involves EHRs, lab systems, imaging reports, pharmacy records, institutional guidelines, and external records simultaneously. Single-document benchmarks miss the real difficulty.

**Include deceptively simple tasks.** Lab interpretation, medication verification, referral synthesis, and documentation all look like they should be straightforward. Including them in benchmarks exposes failures that knowledge-focused evaluations miss.

**Test for safety, not just accuracy.** Clinical evaluations should test whether the model identifies dangerous possibilities, reasons about trajectories, respects institutional context, and grounds its outputs in source evidence.

**Test handling of missing and conflicting information.** In practice, records are incomplete, values conflict across systems, and key context is absent. Evaluations should test whether the model flags what it couldn't find, identifies contradictions, and degrades gracefully rather than filling gaps with plausible-sounding fabrications.

**Make verification a first-class dimension.** If a clinician can't verify the output, the output isn't useful. Evaluation rubrics should measure whether the model cites sources, exposes uncertainty, and produces work that a human expert can efficiently review.

## What's Next

Healthcare and life sciences are among the highest-impact applications of frontier AI. Clinical workflows directly affect patient outcomes at scale. Drug discovery determines which therapies exist at all. In both domains, our approach is to build evaluations that directly measure what experts actually do. We start by talking to clinicians, researchers, and operators about what blocks their work, then engineer evaluation tasks and training environments faithful to those workflows. We work on this from both directions: with frontier labs on the evaluation and training side, and with healthcare organizations on the deployment side.

This survey is one input into that broader effort. We'll share more over coming weeks, including specific evaluation tasks for both domains and what we've found about how today's frontier models perform on them.

## Appendix

### Methodology

- *We surveyed 57 healthcare professionals: 47 in clinical care (29 physicians, 11 pharmacists, 7 others across 18 specialties) and 10 in revenue cycle and administration.*
- *The survey collected open-ended responses first (pain points, delegation preferences, deceptive complexity) before showing any predefined workflows, then asked respondents to write three realistic AI prompts, rate 12 clinical workflows on time and importance, and describe expected AI failure modes and verification methods.*
- *Surveys were supplemented by 60-minute deep-dive interviews with practitioners in radiology, hematology, and emergency medicine.*
- *The 141 AI prompts were hand-coded against a six-capability rubric. Each prompt could receive multiple labels; the average prompt received about three.*
- *Respondents were recruited via expert panel and represent a convenience sample, not a statistically representative population. The strength of this work is qualitative depth, not population-level generalization.*
- *Data collected January-February 2026*