What actually matters in clinical AI right now: a reality check for health tech investors

NOV 28, 2025

DISCLAIMER: The views and opinions expressed in this essay are solely my own do not reflect the views, opinions, or positions of my employer, Datavant, or any affiliates.

If you are interested in joining my generalist healthcare angel syndicate, reach o trey@onhealthcare.tech or send me a DM. Accredited investors only.

Abstract

This essay examines six recent publications that represent the most significant advances in clinical AI over the past six months. Rather than focusing on increm improvements in narrow tasks or benchmarks, these papers demonstrate AI syst functioning in real clinical workflows with measurable impact on physician time diagnostic accuracy, medication safety, and patient outcomes. The analysis revea three critical investment themes: the shift from demonstration studies to pragm randomized trials, the superiority of human-AI collaboration over full automatic and the emergence of sophisticated evaluation frameworks that will shape regul pathways. For angel investors evaluating health tech opportunities, understandi these methodological shifts is as important as understanding the underlying technology, because they define which companies will successfully navigate clini validation and health system adoption.

Introduction

The healthcare AI landscape has reached an inflection point where methodology matters more than model architecture. Over the past 18 months, we've seen a fundamental shift from papers showing that AI can do impressive things in cont settings to papers demonstrating that AI can improve actual clinical outcomes ir routine practice. This distinction is critical for investors because it separates technologies that will generate real revenue from technologies that will remain trapped in pilot purgatory.

The papers covered in this essay represent the current state of the art not becaus they achieve the highest benchmark scores or use the most sophisticated algorith but because they answer questions that health system decision-makers actually c about. Can this technology reduce physician burnout in real-world conditions? V catch medication errors that human pharmacists miss? Does it introduce system bias that will expose us to liability? Can we deploy it without regulatory friction These are the questions that determine whether a million-dollar pilot contract becomes a ten-million-dollar enterprise deployment, and these papers provide th evidence base that makes those decisions possible.

For angel investors, especially those without deep clinical backgrounds, understanding the methodological evolution of clinical AI research is essential because it tells you what kinds of companies will succeed over the next five years companies that are building products aligned with these emerging evidence stan will navigate procurement processes smoothly. The companies that are still pitch based on impressive demos and favorable case studies will struggle.

The ambient AI scribe trial from UCLA

The UCLA ambient AI scribe study published in NEJM AI represents the first la scale pragmatic randomized controlled trial of ambient documentation technolo embedded in routine clinical care. Previous evaluations of AI scribes have mostly observational studies where a hospital implements the technology and then meas changes in documentation time or physician satisfaction. Those studies consiste

show positive results, which is unsurprising given that any new technology depl with institutional support and physician opt-in is going to look good in the shor due to selection bias and Hawthorne effects.

What makes the UCLA study different is the rigor of the experimental design. T randomized 238 physicians across 14 specialties into three arms: usual documen without AI assistance, Microsoft DAX, or Nabla. The randomization happened a physician level rather than the encounter level, which matters because it control physician-specific factors like baseline documentation speed, typing proficiency, personal workflow preferences. They tracked approximately 72,000 encounters c the study period and measured documentation time using actual EHR log data extracted from system timestamps rather than self-reported estimates from phys surveys.

That methodological choice alone is worth emphasizing because clinician self-re about time usage are notoriously unreliable. When you ask physicians how long spend on documentation, you get numbers that are colored by how burned out t feel, how much they dislike the EHR, and whether they're trying to make a point administrators about workload. EHR log data doesn't have those biases. It just te you when the note was opened and when it was signed, adjusted for time spent c other activities within the EHR during that window.

The headline finding was that Nabla users reduced documentation time by approximately 41 seconds per note compared to 18 seconds in the control group. adjusting for baseline differences and potential confounders, this translated to a percent larger reduction in documentation time for Nabla versus control, and th difference reached statistical significance. Microsoft DAX showed similar direct effects but didn't quite hit significance thresholds, possibly because of sample si because the effect size was genuinely smaller. Both AI tools were associated with modest improvements in validated burnout and cognitive workload measures, ar 7 percent versus control, using the Stanford Professional Fulfillment Index and Task Load Index rather than homegrown survey instruments.

The safety profile is equally important to understand. Clinically relevant inaccu in AI-generated documentation were described as occasional, and only one mild patient safety event occurred during the entire study that was potentially attribu to AI-generated content. Patient acceptance was high, with fewer than 10 percer patients declining to have the AI scribe used during their visit. This is importan because one of the concerns with ambient AI is that patients might find it creep intrusive to have their conversations recorded and processed by an algorithm, bu empirical evidence suggests that concern is mostly theoretical.

For investors, the key signal here is not the exact magnitude of time savings, whi modest in absolute terms, but the fact that we now have RCT-level evidence demonstrating measurable workflow impact in routine conditions across multip specialties. This fundamentally changes the value proposition from theoretical efficiency gains supported by vendor-provided case studies to demonstrated out with quantified effect sizes published in a peer-reviewed journal. That distinctic matters enormously in enterprise sales cycles because it shifts the burden of pro Previously, health systems considering ambient AI deployment had to make deci based on limited evidence and vendor claims. Now they have a template for what evidence looks like, and they're going to start demanding similar evidence from vendors.

The study also establishes important precedent for how these technologies shou evaluated going forward. Pragmatic randomized trials embedded in routine clin operations are feasible and they provide much more credible evidence than retrospective observational studies. Vendors that can point to similar trial evide that are actively partnering with academic medical centers to generate it will ha significant advantage in competitive procurements. Vendors that continue to rel pilot testimonials and before-after analyses are going to find themselves at a disadvantage as health system procurement teams become more sophisticated ab evidence standards.

The other piece that's worth understanding is what this study doesn't tell us. It powered to detect rare safety events, so we don't know if there are low-frequenc high-severity errors that might emerge at scale. The follow-up period was relati

short, so we don't know if the time savings persist over longer periods or if phys eventually revert to baseline as they find workarounds or as the novelty wears of study was conducted at a single large academic medical center, so we don't know the results generalize to community hospitals or small practices with different p populations and workflows. These aren't criticisms of the study, they're just limitations that define where the evidence ends and where judgment begins.

From an investment perspective, the companies operating in this space that look attractive are the ones that have strong clinical partnerships, that are actively generating peer-reviewed evidence, and that have a clear path to demonstrating in ways that align with how health systems make purchasing decisions. The ones are still selling primarily on user enthusiasm and anecdotal feedback are going t struggle as the market matures.

It's also worth noting that 40 seconds per note may not sound like much, but wh you multiply it across millions of encounters per year in a large health system, it translates to meaningful FTE savings or alternatively to more time available for patient care. The business case for ambient AI has always been strong in theory, now we have credible evidence that the theory holds up in practice, which mean adoption curves are likely to accelerate.

Medication safety and the case for human-AI collaboration

The medication safety paper published in Cell Reports Medicine is probably the methodologically important study from the past year on how to deploy large lan models in high-stakes clinical environments, and its importance stems specifical from what it doesn't recommend. The study evaluated LLM-based clinical decisi support for identifying drug-related problems across 16 specialties using 91 cur error scenarios spanning 40 vignettes. The research team built and validated five RAG-enhanced LLM variants, selected the best-performing model, and then test in three distinct deployment modes: LLM-only CDSS, pharmacist with LLM cop and pharmacist alone.

The copilot mode, where pharmacists had access to the LLM's recommendations alongside their own clinical judgment, achieved 61 percent accuracy with an F1 around 0.59. More significantly, for drug-related problems with serious harm potential, the copilot mode was 1.5 times more accurate than pharmacists worki alone. That's a substantial improvement in exactly the cases where accuracy matt most. The LLM-only mode, by comparison, underperformed the copilot configu by enough that you wouldn't want to deploy it in production even if regulatory frameworks and liability considerations allowed it, which they mostly don't.

This finding aligns with essentially every other rigorous evaluation of clinical AI conducted over the past two years: human-AI collaboration consistently outperfo either humans or AI operating independently. The pattern holds across different clinical domains, different model architectures, and different task complexities. you give clinicians access to high-quality AI recommendations and let them inte those recommendations with their own expertise and contextual knowledge, you better outcomes than you do from either component alone. This isn't a minor technical detail, it's a fundamental insight that should shape how every clinical A company thinks about product design and deployment strategy.

For investors, the implication is that you should be extremely skeptical of any he tech company pitching full automation of clinical decision-making. Not because underlying technology isn't impressive, it often is, but because the performance ceiling for collaboration is demonstrably higher than the performance ceiling fo automation, and health systems understand this now. The companies that will w clinical AI are the ones building for collaboration from day one, not the ones bu autonomous systems and then grudgingly adding human review when customers regulators push back.

The methodology here is also worth understanding in detail because it's likely to become a template that other researchers and companies follow. The research te used retrieval-augmented generation architecture rather than simply prompting base model with clinical scenarios. That matters enormously because it allows th system to pull in current drug interaction databases, formulary information, and clinical guidelines without requiring constant retraining of the underlying mode

RAG architectures separate the knowledge base from the reasoning engine, whic means you can keep the knowledge current without the computational expense a engineering overhead of continuous model retraining.

They also explicitly compared multiple deployment modes rather than just repor aggregate accuracy numbers. That comparison is critical because it directly addr the question that health system decision-makers actually care about: should we d this with human oversight or without it? The answer from the data is unambigu you should deploy it with human oversight, and specifically you should deploy it way that presents the AI's recommendations to trained clinicians who can evalua them in context and override them when necessary.

The paper includes a detailed flow diagram showing how the copilot system coul integrated into existing pharmacy workflows. That kind of implementation-focu documentation signals that the research team was thinking about real-world deployment from the beginning rather than just trying to publish impressive benchmark results. It's also exactly the kind of documentation that health syster need when they're evaluating whether to implement a new technology, because i helps them understand what changes to workflows, staffing, and infrastructure w required.

For companies building in this space, the implication is that your product roadn needs to assume human-in-the-loop deployment as the default architecture, not afterthought. The user interface design, the workflow integration points, the tra requirements, the feedback mechanisms for capturing clinician overrides and corrections, all of these need to be optimized for collaboration rather than automation. That's a harder product problem to solve than building a model tha generates good recommendations in isolation, but it's the problem that actually onto how health systems want to deploy this technology.

The other technical detail that's important is the use of RAG architecture. If you evaluating a company that's building clinical decision support tools on top of a b language model without retrieval augmentation, that should raise questions abo how they're handling hallucinations, outdated information, and knowledge grou

RAG isn't a magic solution to these problems, but it's increasingly becoming tab stakes for production clinical AI systems because it provides a principled way to ground model outputs in curated knowledge sources.

The challenge with RAG is that you need high-quality, structured, continuously updated knowledge bases to retrieve from, and building or licensing those is expensive and time-consuming. For medication safety specifically, you need drug interaction databases, pharmacokinetic data, formulary restrictions, institutiona protocols, and ideally patient-specific information about allergies, renal functio hepatic function, and concurrent medications. Integrating all of that into a RAG pipeline that runs fast enough to be useful in clinical workflows is non-trivial engineering work.

Companies that have solved those integration challenges or that have partnershi giving them access to high-quality structured medical knowledge are going to ha significant advantage. Companies that are trying to rely purely on the knowledge embedded in pretrained models are going to struggle with accuracy, safety, and reliability in production environments.

It's also worth noting that this study used curated vignettes rather than real pati data from live clinical workflows. Vignettes allow for careful control of the scen and ensure that you're testing specific kinds of drug-related problems, but they fully capture the messiness of real clinical data where information is incomplete contradictory, or buried in unstructured text. The performance numbers from th study should be viewed as an upper bound on what you'd expect in production deployment, not as a guaranteed baseline. Companies making claims based on vignette studies without real-world validation should be viewed with appropriat skepticism.

The broader message for investors is that medication safety is one of the highest applications for clinical AI because medication errors are common, expensive, a often preventable. The literature suggests that somewhere between 5 and 10 per of hospitalized patients experience adverse drug events, and a substantial fractio those are preventable. If you can build a system that reduces preventable adverse

events by even 20 or 30 percent, the value proposition is immediately obvious to hospital quality and safety teams. But you can only capture that value if you buil system correctly, which means human-AI collaboration, RAG architecture, robus knowledge bases, and continuous monitoring infrastructure.

Reasoning models in critical care diagnostics

The DeepSeek-R1 study published in Critical Care represents one of the first prospective evaluations demonstrating that reasoning-style large language mode meaningfully improve diagnostic performance in high-acuity clinical settings. T study recruited 32 critical care residents from six tertiary hospitals and randomi them to diagnosing 48 diagnostically challenging critical illness cases either witl without access to the model's output. These were NEJM-style case challenges an similar complex cases published after the model's training cutoff date, which eliminates the possibility of training data leakage where the model has seen the during pretraining.

The model alone, operating without human input, achieved 60 percent top-1 diagnostic accuracy across the 48 cases. That's impressive given that these are specifically selected to be diagnostically challenging and that they span multiple organ systems and disease processes. For context, human critical care residents without AI assistance achieved 27 percent top-1 accuracy on the same cases. Wh residents had access to the AI's output, their accuracy jumped to 58 percent, nea matching the model's solo performance. Diagnostic time roughly halved, droppi from a median of 1920 seconds to 972 seconds when residents used AI assistance

Those are substantial improvements in both accuracy and efficiency, and import they occurred with subspecialty residents who already have significant domain expertise, not with medical students or general practitioners who might be expe to benefit more from decision support. The magnitude of the effect suggests tha reasoning models are providing genuine diagnostic insight rather than just servi sophisticated search engines or differential diagnosis generators.

For investors, this study is significant for several reasons. First, it provides empi evidence that reasoning models, the ones that explicitly generate step-by-step th processes before producing final answers, are genuinely better at complex diagn tasks than standard instruction-tuned models that go directly from input to outp That matters because reasoning models are computationally more expensive to r They require more tokens, more processing time, and therefore more infrastruct cost per query. You need evidence that the extra computational expense translat meaningful performance gains that justify the cost, and this study provides that evidence at least for diagnostic reasoning in critical care.

Second, the magnitude of the improvement is large enough that it's difficult to dismiss as a marginal gain. Going from 27 percent to 58 percent diagnostic accu on challenging cases is the kind of delta that changes clinical outcomes, not just workflow efficiency. Diagnostic errors in critical care settings are high-stakes be patients are physiologically unstable and delays or errors in diagnosis can lead t rapid deterioration. If you can deploy a system like this in ICUs and it prevents small number of diagnostic errors per year, the value proposition becomes immediately obvious to hospital administrators, quality improvement teams, and intensivists.

Third, the study followed TRIPOD-LLM reporting guidelines and conducted ex consistency checks on model output, which signals where evaluation norms are heading for clinical AI research. TRIPOD-LLM is an extension of the TRIPOD framework for reporting prediction model studies, specifically adapted for LLM based systems. It requires documentation of the prompt engineering process, reporting of consistency across multiple runs with identical inputs, separation o performance evaluation from prompt optimization, and transparency about mod selection and hyperparameters. As more journals adopt these standards, compan that can demonstrate adherence will have advantages in publication, credibility, regulatory pathways.

The limitation, which the authors explicitly acknowledge, is that this remains a vignette-based study rather than a live deployment with actual ICU patients in r time. Vignettes can't fully capture the messiness of real clinical data: the incomp

information that arrives in fragments over hours, the contradictory lab values th don't make physiological sense, the patients who can't communicate their sympt clearly, the family members providing unreliable histories. All of those real-wor complexities can degrade model performance relative to what you see in clean vignettes.

That said, the performance gap between residents with and without AI assistanc large enough that even if real-world results are 20 or 30 percent lower than vign results, you'd still expect meaningful clinical benefit. That makes this a strong candidate for follow-on implementation trials, and I'd expect to see those studie starting within the next 12 to 18 months at academic medical centers that have t research infrastructure and institutional review board processes to support then

From a technical perspective, the study demonstrates that reasoning models can effective without fine-tuning on medical data, which is significant because fine-t is expensive and requires large labeled datasets that often don't exist for speciali clinical tasks. If you can get good performance from clever prompting and chain thought reasoning with a base model, that substantially lowers the barrier to deploying these systems across different clinical domains.

For companies building diagnostic decision support tools, the implication is tha reasoning models should be seriously considered for complex diagnostic tasks w the cognitive process involves synthesizing information from multiple sources, generating differential diagnoses, and weighing evidence for and against differe possibilities. Those are exactly the kinds of tasks where explicit reasoning chain likely to outperform black-box predictions. The tradeoff is computational cost, b inference costs continue to decline, that tradeoff becomes more favorable over ti

The other important signal from this study is that diagnostic decision support is of the highest-value applications for clinical AI because diagnostic errors are bo common and consequential. Estimates suggest that around 10 to 15 percent of diagnoses in acute care settings are wrong or delayed, and those errors contribut substantial morbidity, mortality, and cost. If you can build a system that reduces diagnostic error rates by even a modest percentage, the value creation is enormo

But you have to get the deployment model right, which means human-in-the-loo clear presentation of reasoning and uncertainty, and integration into clinical workflows that don't create alert fatigue or cognitive burden.

It's worth noting that the critical care setting is particularly well-suited for this of AI deployment because intensivists are already accustomed to working with decision support tools, they have time to review AI-generated recommendations because critical care workflows involve frequent reassessment, and the patient population is monitored continuously so errors can be caught relatively quickly. Those characteristics make ICUs a natural testbed for diagnostic AI, and succes deployments in ICU settings will likely pave the way for broader adoption in oth acute care environments.

Multi-task clinical workflows and the future of triage systems

The npj Digital Medicine paper from Gaber and colleagues takes a different methodological approach than the other studies covered here, and that differenc instructive for understanding where clinical AI evaluation is heading. Instead of testing a single narrowly defined task like generating documentation or identify drug interactions, they built Claude-based workflows that simultaneously predi Emergency Severity Index triage level, mapped patients to appropriate specialty referrals, and suggested likely diagnoses. They used 2,000 cases derived from MI ICU data with structured vitals, history of present illness, and demographic information, then benchmarked several Claude-family models including a RAG-assisted Claude 3.5 Sonnet configuration.

The study is slightly older than the strict six-month window but it's too importa exclude because it represents a fundamental shift in how clinical AI systems are conceptualized and evaluated. Most clinical AI research focuses on single endpo can you predict sepsis, can you identify diabetic retinopathy on fundus photos, c you generate an accurate radiology report. That narrow focus makes sense for research purposes because it's easier to isolate what the model is doing well or p

when you're measuring performance on a single well-defined task. But it doesn't reflect how these systems will actually be used in clinical practice, where you do want three separate tools that each do one thing, you want a single integrated interface that helps with triage, referral, and preliminary diagnosis as part of a coherent clinical workflow.

For investors, this distinction matters because it changes how you think about product scope, competitive positioning, and defensibility. If the future of clinica integrated multi-task systems rather than point solutions, then companies that h built narrow single-purpose tools are going to struggle unless they can expand tl capabilities quickly and coherently. Conversely, companies that are already thinl about multi-task architectures and can demonstrate that their models handle mu clinical reasoning steps in a coordinated way will have significant advantages in enterprise sales.

The study found that RAG-enhanced models outperformed base models across a three tasks, which aligns with the pattern we've seen in the medication safety re and elsewhere. Multi-task clinical reasoning requires access to current clinical guidelines, institutional protocols, specialist availability, local epidemiology, anc other contextual information that can't be baked into a pretrained model. RAG provides a principled architectural approach for grounding model outputs in tha kind of structured contextual knowledge without requiring continuous retrainir

The implication is that the barrier to entry for building competitive clinical AI products is rising because you need not just a good foundation model but also th infrastructure to integrate it with real-time data sources in ways that are fast, re secure, and clinically validated. Small companies with limited engineering resou are going to struggle to build and maintain that infrastructure, especially as hea systems start demanding evidence that the RAG pipelines are pulling from high-quality sources and that the retrieval mechanisms are robust to missing or contradictory data.

This paper also functions as a blueprint for product development in ways that ar somewhat unusual for academic research. It includes detailed workflow diagram

showing how the system processes inputs and generates outputs, specific implementation details about prompt structure and RAG configuration, and exp evaluation methodology that other teams could replicate. Any competent engine team could use this paper as a starting point for building a similar system, whicl means the value in companies working on this problem is less about the core technology, which is increasingly commoditized, and more about execution, clin partnerships, regulatory strategy, data access, and go-to-market.

The MIMIC-derived case methodology is both a strength and a limitation. MIMI data is detailed, well-structured, and includes rich clinical information from ICU patients at a major academic medical center. That makes it excellent for testing whether models can perform complex reasoning tasks when given complete information. But ICU patients are sicker and more complex than typical emerge department patients, the documentation is more thorough, and the clinical cont quite different from what you'd see in a community hospital ED at 2 AM on a Saturday night. So the performance numbers from this study should be interpre evidence that the approach works in principle, not as a guarantee of what you'd production deployment across diverse clinical settings.

For companies targeting the emergency department triage and decision support market, this study provides a roadmap for what good looks like in terms of mult evaluation and RAG integration. It also highlights that triage systems need to dc than just assign severity scores, they need to route patients to appropriate specia and provide preliminary diagnostic hypotheses that can guide initial workup. Sir task triage models that only predict ESI level are solving an incomplete problem health systems are going to prefer integrated solutions that address the full worl

The other insight from this work is that evaluation frameworks need to move be single-task accuracy metrics to assess how well models perform across multiple related tasks simultaneously. If you're building a triage system that predicts seve specialty, and diagnosis, you need to evaluate not just whether each individual prediction is accurate but whether the combination of predictions is clinically coherent. A model that predicts high severity, routes to cardiology, but suggests

dermatologic diagnosis is doing something wrong even if each individual predic
might be defensible in isolation.

That kind of coherence checking is difficult to formalize but it's critical for real
world deployment. Companies that are thinking seriously about multi-task clini
need to build evaluation frameworks that can detect incoherent or internally
contradictory outputs, and they need to have mitigation strategies for when thos
occur. That might mean adding consistency checks in the inference pipeline, usi
multi-step reasoning where later steps can catch errors in earlier steps, or flaggi
low-confidence cases for human review.

Sociodemographic bias and why fairne testing matters for enterprise adoptio

The Nature Medicine bias paper from Mount Sinai is arguably the most widely c
clinical AI paper from the past year and its influence extends well beyond the
academic research community into health system governance processes, vendor
procurement requirements, and regulatory policy discussions. The study genera
over 1.7 million AI recommendations by taking 1,000 emergency department vig
and replicating each one with 32 distinct sociodemographic profiles across nine
different large language models. They then systematically analyzed patterns in t
priority, diagnostic imaging recommendations, treatment aggressiveness, and m
health escalation decisions to identify whether recommendations differed based
patient demographics rather than clinical presentation.

What they found is that some models escalated care, especially mental health
evaluation, primarily based on sociodemographic characteristics rather than clir
status. High-income profiles received recommendations for advanced imaging m
frequently than low-income profiles despite identical symptom presentations. L
income profiles were more likely to receive recommendations that no further tes
was needed. These aren't small effects occurring at the margins of the distributi
they're systematic patterns that showed up consistently across multiple models a
multiple clinical scenarios.

The study also found that demographic bias wasn't uniform across all clinical situations. Some scenarios showed minimal bias while others showed substantial disparities, and the magnitude and direction of bias varied across different mode That heterogeneity is important because it suggests that bias isn't simply a prop of training data that affects all models equally, it's also influenced by model architecture, fine-tuning approaches, and prompt engineering. That means bias potentially mitigable through careful model design and evaluation, but it also m you can't assume that a model is fair just because it was trained on diverse data.

For investors, this paper is critically important not because it suggests you shoul avoid companies working on clinical AI, that would be a profoundly misguided takeaway, but because it defines what due diligence questions you need to be ask and what red flags you need to watch for. If a company is building clinical decisi support tools and they don't have a clear articulation of how they're testing for sociodemographic bias, that's a significant red flag. If they're relying exclusively aggregate accuracy metrics without any subgroup analysis, that's a red flag. If th can't describe what their fairness testing framework looks like or what mitigatic strategies they have in place for disparities they identify, that's a red flag.

The paper proposes an AI assurance framework that involves systematic stress te of LLMs across sociodemographic axes before deployment. This is rapidly becon standard expectation in enterprise procurement processes, especially at large academic medical centers and integrated delivery networks that have sophisticat health equity programs and institutional commitments to reducing disparities. Companies that can demonstrate they've conducted this kind of stress testing ar have documented mitigation strategies for any disparities they identified will ha significant competitive advantages in sales cycles.

The liability and risk management implications are also substantial. If a health s deploys a clinical AI tool that systematically undertreats or inappropriately esca care for certain demographic groups, they're exposed to discrimination claims u civil rights law, medical malpractice liability for patients who are harmed by inappropriate recommendations, regulatory scrutiny from CMS and state health departments, and reputational damage that can affect patient volumes and comn

trust. Hospital risk management and legal teams are increasingly aware of these exposures, which means they're requiring vendors to provide detailed evidence c fairness testing as part of the contracting process.

Companies that can't provide that evidence aren't getting past the contracting st regardless of how impressive their core technology is. This isn't theoretical, I'm procurement processes at major health systems where fairness documentation is mandatory requirement in the RFP, and vendors without it are eliminated before technical evaluation even begins. That's a fundamental shift in how clinical AI is being purchased, and it happened remarkably quickly once this paper and a few others like it demonstrated the scope of the bias problem.

One thing that sometimes gets lost in discussions about AI bias is that this isn't purely a social justice issue, although it absolutely is that, it's also a business risk issue and a clinical safety issue. Biased recommendations lead to worse patient outcomes for the affected populations, which leads to higher readmission rates, performance on quality metrics, lower patient satisfaction scores, and potentiall lower reimbursement under value-based care contracts. Health systems care abo equity because it's the right thing to do, but they also care about it because biase care is expensive care that produces bad outcomes and creates legal exposure.

The methodology of the study is worth understanding because it's likely to beco template for how fairness evaluation is done going forward. Creating synthetic vignettes with systematically varied demographic profiles allows you to isolate th effect of demographics on model recommendations while holding clinical presentation constant. That's difficult to do with real patient data because clinic presentation and sociodemographic characteristics are often correlated in ways t are hard to disentangle. A patient's ZIP code might be associated with different disease prevalence, different access to preventive care, different health literacy, a different clinical trajectories, not just with discriminatory treatment recommendations.

Synthetic vignettes let you ask the counterfactual question: if this exact patient v this exact clinical presentation had a different race or income level or insurance

status, would the model make different recommendations? That's a powerful toc detecting bias, and it's something that companies should be doing routinely as p their model validation process. The challenge is that creating high-quality synth vignettes requires clinical expertise to ensure they're realistic and that the demographic variations you're testing are clinically appropriate.

For companies building clinical AI products, the actionable implications are tha need to build fairness testing into your development and validation workflows fi the beginning, not as an afterthought once someone raises concerns. That means maintaining demographic metadata in your training and validation datasets, conducting routine subgroup analyses, implementing bias detection algorithms, having clear escalation processes when disparities are identified. It also means b transparent with customers about what fairness testing you've done, what you fc and what mitigation strategies you implemented.

The companies that will be successful in clinical AI over the next five years are t ones that treat fairness as a core product requirement rather than a compliance checkbox. That requires investment in diverse training data, sophisticated evalu frameworks, ongoing monitoring infrastructure, and partnerships with health ec researchers who can help identify blind spots and validate approaches. It's not cl and it's not easy, but it's increasingly non-negotiable for enterprise adoption.

Foundation models in radiology and th case for vertical specialization

The Radiology review from Tavakoli and colleagues is less of a primary research contribution and more of an agenda-setting synthesis document, but it's worth c attention because Radiology is one of the most influential specialty journals and reviews published there tend to shape how academic radiology departments, ima AI companies, and healthcare organizations think about technology adoption priorities. The paper synthesizes emerging evidence on foundation models and generative architectures in radiology with explicit focus on clinical readiness an workflow integration rather than just algorithmic performance metrics.

The central argument is that subspecialty-specific foundation models are likely t reach clinical readiness substantially faster than broad generalist models attemp to cover all of radiology or all of clinical medicine. That conclusion is grounded several observations about radiology as a domain: it has relatively well-defined ta with standardized imaging protocols, structured reporting requirements that cr clean training data, quantifiable performance metrics that correlate with clinica utility, and established regulatory pathways for imaging AI through FDA's radiological device framework.

For investors, this review reinforces a thesis that's been gaining traction over the 18 months: vertical specialization matters more in clinical AI than it does in con AI applications. In consumer contexts, there's a strong argument for building generalist models that can handle diverse and unpredictable user needs across m domains. In clinical applications, especially in procedural specialties like radiol pathology, and dermatology, you're generally better off building models that are deeply optimized for narrow sets of tasks with rich domain-specific training dat tight integration into clinical workflows.

That doesn't mean there's no role for generalist clinical models. Ambient documentation tools are essentially generalist models and they're performing we because documentation is a relatively general task that doesn't require deep dom specific knowledge. But when you're building tools that directly impact diagnos treatment decisions, depth matters more than breadth, and partnerships with subspecialists who can provide training data, validate outputs, and guide workfl integration become critical competitive advantages.

The companies that are winning radiology AI contracts at major health systems n now are overwhelmingly the ones that have focused on specific imaging modaliti specific clinical indications: chest CT for lung nodule detection, mammography breast cancer screening, brain MRI for stroke assessment. They have deep partnerships with radiologists in those subspecialties, subspecialty-specific trair datasets that reflect current imaging protocols, integration with PACS and RIS systems that radiologists actually use, and clear value propositions around reduc turnaround time or improving diagnostic accuracy for specific high-value tasks.

Contrast that with companies trying to build general-purpose radiology AI that across all modalities and all anatomical regions. Those companies face much har training data challenges because they need representative data across everything much harder validation challenges because performance requirements differ substantially across subspecialties, much harder go-to-market challenges becaus they're trying to be everything to everyone, and much harder competition becaus they're fighting specialist competitors in every niche simultaneously.

The review also emphasizes the critical importance of continuous monitoring an failure mode analysis for foundation models deployed in clinical settings. Found models are powerful but they're also opaque and they can fail in unexpected way especially when they encounter input distributions that differ from their trainin data. A model trained primarily on outpatient imaging might perform poorly on patients where image quality is compromised by portable equipment, patient positioning constraints, and physiological instability. A model trained on acader medical center data might struggle with community hospital imaging where pro and equipment vary.

Radiology departments that deploy these systems need robust infrastructure for tracking performance over time, flagging cases where model confidence is low o where recommendations differ substantially from radiologist interpretations, an feeding that information back into continuous improvement processes. That monitoring infrastructure is expensive to build and maintain, requiring dedicate MLOps engineering, clinical oversight, data pipelines, and analytics capabilities

This creates another barrier to entry that favors larger, better-capitalized compa over smaller startups. You can't just train a model and ship it to customers, you the entire operational infrastructure to support continuous monitoring, perform tracking, version control, regulatory compliance, and customer support. Small companies with limited engineering resources and runway are going to struggle build and maintain that infrastructure at the quality level that enterprise health customers require.

For imaging AI specifically, there are additional technical challenges around integration with existing radiology IT infrastructure. PACS systems, RIS systems reporting workflows vary substantially across institutions and many are built on legacy technology stacks that make integration difficult. AI tools that require ma data export and import won't get adopted because radiologists won't tolerate workflow friction that slows down their reading. Tools that integrate seamlessly existing workflows and present results within the radiologist's normal workspac much higher adoption rates.

That means successful radiology AI companies need not just strong models but a strong integration engineering, partnerships with major PACS vendors, and deep understanding of radiology workflows and pain points. Those capabilities are ha build and they take time to develop, which creates meaningful moats for compar that have them.

The review also touches on the role of generative AI in radiology beyond just discriminative tasks like lesion detection. Generative models can potentially hel report generation, protocol optimization, synthetic data generation for training, even image reconstruction from undersampled acquisitions. Those applications still mostly in research phases but they represent substantial potential value if th can be validated and deployed safely.

For investors evaluating radiology AI companies, the key questions to ask are: W specific imaging modality and clinical indication are you focused on? Who are yo radiology subspecialist partners and advisors? What training data do you have ac to and how representative is it of real-world practice variation? How does your t integrate into existing PACS and RIS workflows? What's your regulatory pathwa where are you in that process? What's your monitoring and performance trackin infrastructure? How do you handle model updates and version control? Those questions will quickly separate companies that have thought deeply about real-w deployment from companies that are still in the research demo phase.

What this means for health tech investors going forward

If you're evaluating health tech companies working on clinical AI in 2025 and be the six papers covered in this essay point to several themes that should inform yo diligence process and portfolio construction strategy. These aren't abstract acad observations, they're practical implications that directly affect which companies succeed in capturing enterprise value and which will remain trapped in pilot purgatory.

First, evidentiary standards are rising rapidly and methodology matters as much technology. Companies that are partnering with academic medical centers to con pragmatic trials or that can point to peer-reviewed publications in respected jou are going to have substantially easier paths to health system adoption than comp still relying on pilot testimonials and vendor-provided case studies. The UCLA ambient scribe trial established a new benchmark for what good evidence looks this domain, and procurement teams at sophisticated health systems are going to demanding similar quality evidence from vendors.

This creates a bifurcation in the market where well-funded companies with stro clinical research partnerships can generate the evidence needed to win enterpris contracts, while smaller companies without those resources struggle to different themselves from competitors who have better evidence. For angel investors, this suggests you should be looking for companies that have strategic relationships w academic medical centers or integrated delivery networks that can serve as both development partners and validation sites. Companies that are purely product-focused without clinical research capabilities are going to find themselves at a disadvantage.

Second, human-AI collaboration is the deployment model that actually works in stakes clinical environments, not full automation. This has profound implicatior product design, user experience, training requirements, workflow integration, an to-market strategy. If a company is pitching autonomous clinical decision-makir they're either naive about regulatory and liability constraints or they're delibera

overselling what the technology can do and what customers want. Neither is a gc
sign.

The medication safety study and the critical care diagnostics study both demons
that collaborative systems where AI augments human judgment outperform eith
component working alone. That's not a temporary limitation that will disappear
better models, it's a fundamental insight about how complex cognitive tasks wor
domains with high stakes and incomplete information. For investors, this means
should be evaluating whether companies have designed their products for
collaboration from the beginning. Look at the user interface: does it present AI
recommendations in ways that support human judgment rather than replacing it
Look at the workflow: are there clear mechanisms for clinician override and feec
Look at the training materials: do they emphasize critical evaluation of AI outpu
rather than blind trust?

Third, fairness testing and bias mitigation are rapidly moving from nice-to-have
features to mandatory requirements for enterprise procurement. The Mount Sin
bias paper has had remarkable influence on how health systems think about clin
AI governance, and that influence is translating directly into RFP requirements
contract negotiations. Companies that can't demonstrate they've conducted syst
fairness testing across demographic subgroups are getting eliminated from
competitive procurements before technical evaluation begins.

This is creating a new category of must-have capability that smaller companies c
lack. Fairness testing requires diverse training and validation datasets with
demographic metadata, sophisticated evaluation frameworks, partnerships with
health equity researchers, and ongoing monitoring infrastructure. That's expens
build and maintain, which favors larger, better-capitalized companies. For angel
investors, this suggests you should be asking detailed questions about fairness te
methodologies, what disparities have been identified, and what mitigation strate
are in place. Companies that can't answer those questions clearly are accumulati
regulatory and liability risk.

Fourth, vertical specialization is increasingly important in clinical AI. The radio review makes this point explicitly, but it applies more broadly across clinical do Subspecialty-focused companies with deep clinical partnerships, narrow task fo and rich domain-specific training data are generally going to outperform genera approaches for diagnostic and treatment decision support applications. The exc is workflow tools like ambient documentation where the task is general enough specialization doesn't confer major advantages.

For portfolio construction, this suggests a barbell strategy might make sense: in a few horizontal workflow tools that can deploy broadly across specialties, and i in multiple vertical-specific diagnostic or decision support tools that go deep in particular clinical domains. Avoid companies in the middle that are trying to be somewhat specialized but not specialized enough to build real moats against foc competitors.

Fifth, the infrastructure requirements for deploying and maintaining productio clinical AI systems are substantial and growing. Companies need MLOps capabi continuous monitoring infrastructure, integration engineering, regulatory exper clinical validation partnerships, fairness testing frameworks, and customer supp operations. Small teams without significant engineering capacity and adequate r are going to struggle to build all of that, which suggests the market will consoli toward a smaller number of well-capitalized players who can afford to do things correctly.

This has implications for entry valuation and capital requirements. Clinical AI companies need more runway and more engineering headcount than equivalent consumer AI companies because the operational infrastructure requirements ar demanding and the sales cycles are longer. Seed and Series A rounds that would adequate for consumer AI products may be insufficient for clinical AI products, investors need to underwrite accordingly.

Sixth, retrieval-augmented generation is becoming table stakes for clinical decis support applications. Multiple studies covered here demonstrate that RAG architectures outperform base models for tasks that require current medical

knowledge, institutional protocols, or patient-specific information. Companies t
are trying to build clinical AI products on top of foundation models without ret
augmentation are going to struggle with hallucinations, knowledge currency, an
grounding.

But RAG introduces its own complexities. You need high-quality structured knowledge bases to retrieve from, you need fast and reliable retrieval mechanisn you need security controls to prevent data leakage, and you need monitoring to c when retrieval is pulling in incorrect or contradictory information. Companies t have solved those challenges or that have partnerships giving them access to structured medical knowledge sources have meaningful advantages. When you'r evaluating companies, ask detailed questions about their RAG architecture: wha knowledge sources are they retrieving from, how often are those sources update do they handle retrieval failures, how do they ensure retrieved information is rel to the query.

Seventh and finally, regulatory pathways and reimbursement mechanisms are sti evolving rapidly for clinical AI. FDA has established frameworks for certain categories of imaging AI and is developing frameworks for clinical decision supp tools, but there's still substantial uncertainty about requirements for different ty of systems. CMS has been cautious about creating specific reimbursement codes AI-augmented services, which means most clinical AI is being paid for through institutional budgets rather than fee-for-service reimbursement.

Companies that have clear regulatory strategies and that are actively engaging w FDA and CMS are better positioned than companies that are ignoring regulatory questions until they become urgent. For investors, this means doing careful dilig on regulatory pathway and timeline, understanding what predicate devices or substantial equivalence claims exist, and having realistic expectations about how regulatory clearance and reimbursement establishment will take.

The broader pattern across all these papers is a shift from proof-of-concept rese demonstrating that AI can perform clinical tasks to implementation science examining whether AI improves outcomes when deployed in real clinical

environments. That's a critical transition because it separates technologies that v in controlled conditions from technologies that work in the messy reality of actu healthcare delivery. The companies that understand this shift and are building products optimized for real-world deployment rather than benchmark performa are the ones that will capture the value being created in clinical AI over the next years.

For angel investors, especially those without deep healthcare backgrounds, the implication is that you need to do more operational and go-to-market diligence a less purely technical diligence than you might for consumer AI companies. The models are increasingly commoditized, the differentiation is in clinical partnersl evidence generation, workflow integration, fairness testing, regulatory strategy, a operational infrastructure. Those are areas where domain expertise and executio matter more than algorithmic innovation.

The companies that will win in clinical AI are not necessarily the ones with the r impressive models or the highest benchmark scores. They're the ones that have figured out how to generate credible clinical evidence, build collaborative rather autonomous systems, test systematically for bias, specialize deeply in valuable cl domains, build robust operational infrastructure, and navigate complex regulatc and reimbursement pathways. Those capabilities take time, capital, and domain expertise to build, which creates meaningful barriers to entry and opportunities well-positioned companies to establish durable competitive advantages.

3 Likes

← Previous

Nex

Discussion about this post

Comments Restacks

Write a comment...

Stuart Miller Haverin about... Nov 28

I like the summary Trey. It speaks volumes to what GOOD human product manage
known for years. Understanding your market and the problem being sought to sol
critical. Your proposed portfolio arrangement for VC’s makes a lot of sense too wit
balance across your ‘barbell’. Sadly I see a number jumping in and placing multip
on one type of technology.

Finally, I will repeat the assertion I have been making for some time.

Institutions and the IT and Finance leadership need to establish infrastructure and infrastructure management to organize, control, validate and apply performance security checks against the AI tools landscape. The lack of creating this control ar monitoring system will be the undoing of some healthcare organizations, across b payer and provider sectors. Management and Governance will become the contro risk profile from the architectures and behaviors of any of these tools can be deva and fatal.

LIKE REPLY