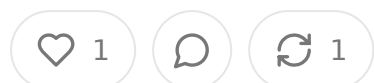


The AI Scribe Gold Rush: What This Lancet Systematic Review Tells Us About Betting on Ambient Documentation

NOV 16, 2025 • PAID



Share

Disclaimer: The thoughts and opinions expressed in this essay are my own and do not represent the views of my employer, Datavant, or any other organization with which I am affiliated.

If you are interested in joining my generalist healthcare angel syndicate, reach out at treyrawles@gmail.com or send me a DM. Accredited investors only.

Abstract

This essay analyzes a July 2025 systematic review published in eBioMedicine examining AI-powered voice-to-text technology (AIVT) for clinical documentation in primary care and outpatient settings. The review synthesized nine studies involving 524 healthcare professionals, 616 patients, and 1,069 consultations to assess AIVT impact across seven quality domains: effectiveness, efficiency, safety, patient-centredness, timeliness, equity, and integration. Key findings relevant to investors include:

- All studies assessing effectiveness, patient-centredness, and efficiency reported improvements
- Documentation speed increased 2.7x for history-taking compared to manual methods

- Safety concerns emerged in 50 percent of studies examining transcription accuracy
- Integration with EHR systems showed feasibility but limited real-world validation
- Severe generalizability issues due to controlled settings and homogeneous populations
- Market predominantly US-focused with limited international validation
- Publication bias likely understates safety risks and implementation challenges

For angel investors, this review highlights both the enormous TAM and persistent technical challenges in the AI scribing market. The technology clearly works in controlled settings but faces significant hurdles in diverse real-world deployment. Companies that can solve the accuracy problem at scale while maintaining patient safety will capture substantial value, but the path from pilot to production remains treacherous.

Table of Contents

The Market Opportunity and Why Everyone's Suddenly Building AI Scribes

What This Systematic Review Actually Found and Why It Matters

The Bull Case: Why AI Scribes Could Be Absolutely Massive

The Bear Case: Why Most AI Scribing Companies Will Probably Fail

Safety Isn't Just a Regulatory Concern, It's an Existential Threat

The EHR Integration Problem Nobody Wants to Talk About

Geographic and Demographic Limitations That Should Worry You

What This Means for Your Portfolio Construction

Due Diligence Red Flags When Evaluating AI Scribing Startups

The Market Opportunity and Why Everyone's Suddenly Building AI Scribes

Look, I get it. Every health tech angel investor I know has seen at least five AI scribe pitches in the past year. The thesis writes itself, right? Physicians spend two hours on documentation for every hour of patient care. Burnout rates are hitting 50 percent. Medicare's cutting reimbursement while demanding more detailed notes. And now we've got LLMs that can actually understand medical conversations and generate coherent clinical documentation. It's the perfect storm for a venture-backable market opportunity.

But here's where it gets interesting. This systematic review out of Imperial College London, published in *eBioMedicine* in July 2025, is one of the first rigorous academic attempts to actually measure whether AI scribes deliver on their promises. They searched five major databases, screened nearly 2,000 papers, and ultimately included nine studies that met their quality criteria. The studies collectively examined 52 healthcare professionals, 616 patients, and 1,069 medical consultations. That's not a massive sample size, but it's enough to draw some meaningful conclusions about where this technology actually stands versus where the pitch decks say it stands.

The review used the Institute of Medicine's quality framework, which breaks down healthcare quality into seven dimensions: effectiveness, efficiency, safety, patient centredness, timeliness, equity, and integration. This framework is actually pretty useful for investors because it maps almost directly to the value propositions AI scribing companies pitch. Effectiveness translates to "does the documentation quality match or exceed manual documentation?" Efficiency maps to "does this actually save physicians time?" Safety is "will this thing kill someone by transcribing medication incorrectly?" And so on.

What makes this review particularly relevant for investors is the timing. All nine included studies were published between 2018 and 2024, with five coming out in 2023-2024. This means we're looking at recent implementations using modern NLP

and LLM technologies, not legacy speech recognition systems from the Dragon NaturallySpeaking era. Seven of the nine studies were conducted in the US, which tells you something about where the market opportunity is concentrated but also should raise questions about international scalability.

What This Systematic Review Actually Found and Why It Matters

Let's start with the good news because there's actually quite a bit of it. On effectiveness, all nine studies reported positive results. The AI-generated documentation was comparable to manual documentation quality in most cases, in some scenarios actually exceeded it. Healthcare professionals consistently report that AIVT allowed them to focus more on patient interaction rather than screen time, which addresses one of the core value propositions these companies sell.

Efficiency gains were even more impressive. Five studies examined efficiency, and five found significant improvements. Wang et al. found that digital scribes were 2.17 times faster for history-taking documentation compared to manual typing and 3.12 times faster for physical examination documentation compared to typing and dictation respectively. That's not marginal improvement, that's order of magnitude stuff. Owens et al. reported that average documentation time per patient dropped from 5.9 minutes pre-implementation to 4.1 minutes post-implementation, a 28.8 percent reduction. Documentation time outside scheduled hours decreased 11.8 percent, saving physicians an average of 4 minutes per day in after-hours work.

The patient-centredness findings are probably the most compelling from a market adoption perspective. All six studies that examined this dimension reported positive results. Patients felt their providers were more focused on them during visits. In one study, 80.9 percent of patients reported feeling the conversation was more personal when AIVT was used. That's huge because it means the technology isn't just helping physicians with their workflow problem, it's actually improving the patient experience, which is what healthcare systems ultimately care about when they're deciding whether to pay for this stuff.

Now here's where things get dicey. On safety, the results were mixed at best and concerning at worst. Of six studies examining safety, three raised significant concerns. Goss et al. found that 75.5 percent of clinicians reported fewer than 10 errors per transcription, but here's the kicker: 19.6 percent said half or more of those errors were clinically significant. Think about that math for a second. If you're making errors per note and half are clinically significant, you're potentially introducing patient safety risks per encounter. At scale, that's terrifying.

Kodish-Wachs et al. got into the technical weeds and measured Word Error Rate across eight different AIVT systems. WER ranged from 35 percent at best to 65 percent at worst, with some scenarios hitting 86 percent error rates. The real problem was medication names. These systems consistently struggled with drug nomenclature which is literally the highest-stakes piece of clinical documentation. Get a medication name wrong and you're not just creating administrative friction, you're potentially killing someone.

Tran et al. examined something really specific but important: non-lexical conversational sounds. You know, the "mm-hmm" and "uh-huh" sounds that carry clinical meaning in patient-provider interactions. Both Google and Amazon's AI systems had error rates above 94 percent for clinically relevant non-lexical sounds. That might seem like a niche problem until you realize that these sounds often signal patient agreement, understanding, or concern, which clinicians use to guide their documentation and clinical decision-making.

The Bull Case: Why AI Scribes Could Be Absolutely Massive

Despite the safety concerns, the bull case for AI scribing is pretty straightforward and honestly quite compelling. The TAM is enormous. There are roughly 1 million physicians in the US, and if you can save each one even 30 minutes per day at a blended reimbursement rate of, say, 200 dollars per hour, you're creating 100 dollars per physician per day in value. That's 25,000 dollars per physician per year at 250

working days. Multiply that by 1 million physicians and you've got a 25 billion dollar annual value creation opportunity just in the US.

The actual pricing for AI scribing services ranges from about 200 to 500 dollars per month per provider depending on volume and features. At 300 dollars per month that's a 3.6 billion dollar annual market at full penetration. But here's the thing: value creation is so much higher than the pricing that there's massive room for market expansion beyond just physicians. NPs, PAs, mental health providers, even pharmacists doing MTM visits could all benefit from this technology. Suddenly, TAM is 5 to 10 billion dollars annually just in the US.

The switching costs and network effects are also pretty favorable for whoever gets embedded first. Once a physician gets used to a particular AIVT system and it integrates with their EHR, the friction to switch is substantial. You've got workflow adaptation, muscle memory, integration complexity, and the very real risk that a new system might be worse. This creates potential winner-take-most dynamics in specific provider segments or health systems.

The technology trajectory is strongly favorable. The studies in this review mostly examined pre-GPT-4 technology. We're now in a world where frontier LLMs have vastly better language understanding, contextual reasoning, and ability to handle ambiguity. The accuracy problems that showed up in these 2023-2024 studies are being actively addressed by the next generation of models. You can make a reasonable argument that many of these findings are already outdated because the underlying technology has improved so dramatically in the past 12-18 months.

There's also a regulatory tailwind that's easy to overlook. CMS and private payers are increasingly focused on documentation quality and specificity for risk adjustment and quality reporting. Physicians who under-document hurt their reimbursement. AI scribes can potentially capture more comprehensive clinical information, leading to better documentation completeness and higher reimbursement. That's not the primary value prop most companies lead with, but it's real money for provider organizations.

The Bear Case: Why Most AI Scribing Companies Will Probably Fail

Okay, now let's talk about why most of the companies in this space are going to get acquired, or become zombie companies limping along at sub-5 million ARR. The systematic review actually illuminates several fundamental challenges that make investors cautious.

First, the generalizability problem is severe. Seven of the nine studies were conducted in the US. Most used highly educated, native English-speaking participants. Several excluded entire patient populations like pediatrics and psychiatry. One study focused exclusively on diabetes patients. Another was conducted in controlled, quiet environments with high-quality microphones. This matters because real-world clinical environments are chaotic, multicultural, and unpredictable. A system that works great in a quiet primary care office in suburban Massachusetts might completely fall apart in a busy urban emergency department where half the patients speak English as a second language.

Four of the nine studies used simulated consultations rather than real patient encounters. That's a massive red flag for anyone who's done real-world healthcare deployments. Simulated consultations are cleaner, more structured, and less complex than actual clinical encounters. The performance metrics from simulated studies almost certainly overstate real-world effectiveness.

Second, the safety signal is genuinely concerning and probably understated due to publication bias. The review authors explicitly note that studies reporting safety may be less likely to get published, leading to an incomplete understanding of the challenges. Of the six studies that examined safety, three raised significant concerns. That's a 50 percent hit rate on safety problems in a sample that's probably biased toward positive results. The real-world safety profile is likely worse.

The medication transcription problem is particularly troubling because it's a solvable problem in other contexts. Pharmacy systems have been dealing with sound-alike names for decades. The fact that AIVT systems in 2023-2024 were still struggling

medication nomenclature suggests either inadequate training data or insufficient domain customization. Either way, it's a fundamental execution challenge that can result in patient harm and massive liability exposure.

Third, the EHR integration challenge is more complex than most pitch decks acknowledge. Four studies reported seamless integration with EHRs, but these were largely in controlled or pilot settings. Real-world EHR integration is a nightmare. You've got hundreds of EHR systems, each with different data models, APIs, and workflows. Epic alone has dozens of different implementation configurations across different health systems. Building AIVT that works well with Epic at Mass General doesn't mean it'll work well with Epic at Cleveland Clinic.

The review found that integration success was often dependent on whether the AIVT system could automatically populate the right fields in the EHR without requiring physicians to do manual copy-paste or field mapping. That's a much harder technical problem than just generating good clinical notes. You need deep integration with EHR data models, understanding of clinical workflows, and often custom development for each health system. That doesn't scale like a SaaS product. It scales like a service business, which is not what venture investors want to hear.

Fourth, the market concentration in the US should worry anyone thinking about building a global health tech company. Only two of nine studies were conducted outside the US, one in Bangladesh and one in the Philippines. The US healthcare market is massive but also highly competitive and increasingly consolidated. If your entire TAM is US-based and you're competing with well-funded startups plus inevitable expansions from Microsoft, Google, and Amazon into this space, your path to venture-scale returns gets narrow quickly.

Safety Isn't Just a Regulatory Concern, It's an Existential Threat

Let's dig deeper into the safety issue because I think this is where most investors are underestimating the risk. The review found that clinician-reported error rates were dramatically higher, but even in the best-case scenarios, you're talking about multiple e

per note. When 19.6 percent of clinicians say that half or more of the errors are clinically significant, that's not a quality problem, that's a patient safety crisis waiting to happen.

Here's the thing about patient safety in healthcare: it only takes one bad outcome to destroy a company. One patient dies because an AI scribe transcribed "metformin" instead of "metoprolol" and suddenly you're in a national news cycle, facing regulatory scrutiny, dealing with wrongful death litigation, and watching your existing customers flee. The liability exposure is asymmetric and existential.

The review makes it clear that these systems require human oversight. Physicians need to review and edit the AI-generated notes before they're finalized in the medical record. But here's the problem: if physicians are spending significant time reviewing and correcting AI-generated notes, you've just eroded most of the efficiency gains that justify the product in the first place. Goss et al. found that 21.2 percent of clinicians reported spending 25 percent or more of their documentation time on editing AI-generated notes. That's not a marginal tax, that's a major workflow burden.

There's also a more subtle safety concern that the review touches on but doesn't explore: the risk of automation complacency. When physicians get used to AI-generated notes being mostly correct, they start to skim rather than carefully review. That's when errors slip through. We've seen this pattern in other domains where humans oversee automated systems. The automation works great until it doesn't, and humans are terrible at catching the edge cases.

From an investment perspective, the safety challenges create a significant regulatory overhang. The FDA has been relatively hands-off with clinical documentation tools because they've traditionally been classified as administrative rather than clinical. As these tools become more sophisticated and start to influence clinical decision making, that classification could change. If AI scribes get pulled into FDA oversight as clinical decision support tools, suddenly you're looking at a multi-year, multi-million-dollar regulatory pathway that most early-stage startups can't afford.

The EHR Integration Problem Nobody Wants to Talk About

The review found that four studies reported seamless AIVT integration with EHR systems, which sounds great until you look at the details. These were often custom implementations, pilot programs, or partnerships with specific EHR vendors. The real-world EHR integration landscape is vastly more complex.

Epic dominates the US hospital market with something like 30 percent market share but there are hundreds of different Epic implementations, each configured differently. Cerner, now owned by Oracle, has its own fragmentation issues. Smaller EHRs like Athenahealth, eClinicalWorks, and NextGen each have their own APIs, data models, and integration requirements. And that's just the US market.

Here's why this matters for investors: EHR integration complexity creates a ceiling on how fast these companies can scale. If you need custom development work for each health system or each EHR implementation, your gross margins look more like a consulting business than a SaaS business. Your sales cycles get longer because you need to do technical due diligence on each customer's EHR configuration. Your customer success costs are higher because you need ongoing support for integration maintenance as EHRs update their systems.

The review doesn't really address the economic implications of integration complexity, but anyone who's worked in health IT knows this is where promising technologies go to die. You can have the best AI models in the world, but if physicians have to manually copy-paste your output into their EHR or if your tool doesn't integrate with their clinical workflow, adoption will be minimal.

There's also the competitive dynamic with EHR vendors themselves. Epic, Oracle, and others are all working on their own AI documentation capabilities. They have distribution advantages, existing customer relationships, and deep integration with their own systems that third-party vendors can't match. If Epic launches a native scribing feature that's even 80 percent as good as your standalone product, why

a health system pay you separately when they're already paying Epic millions of dollars annually?

Geographic and Demographic Limitations That Should Worry You

The review's findings on equity and generalizability are particularly troubling from a market development perspective. Most studies involved native English speakers, in controlled settings, with specific patient populations excluded. This creates enormous question marks about how these systems perform in diverse, real-world populations.

Healthcare in the US is increasingly multilingual and multicultural. In many urban markets, 30 to 40 percent of patients speak a primary language other than English. Even when patients speak English, accents and dialects vary enormously. An AIVT system trained primarily on standard American English might struggle with Southern dialects, AAVE, or non-native English speakers with heavy accents.

The fact that only two of nine studies were conducted outside the US should concern anyone thinking about international expansion. Healthcare systems, clinical workflows, documentation requirements, and languages vary dramatically across countries. An AIVT system optimized for US primary care won't necessarily work in UK NHS trusts, German hospitals, or Australian general practice. The customization required for international markets could be substantial.

There's also the patient population exclusion problem. Several studies explicitly excluded pediatric and psychiatric consultations because they're harder to simulate, but they present unique challenges. But pediatrics and psychiatry represent substantial portions of the healthcare market. If your AIVT system doesn't work well for these specialties, you've just cut your TAM by 20 to 30 percent. And if the technology struggles with complex, nuanced conversations, which psychiatric consultations involve, that raises questions about how it'll handle other complex clinical scenarios like end-of-life discussions, delivering bad news, or navigating disagreements between patients and families.

What This Means for Your Portfolio Construction

If you're an angel investor building a healthcare portfolio, the AI scribing opportunity is tempting but requires careful positioning. Here's how I'm thinking about it: this is a category where you probably want exposure but need to be extremely selective about which companies you back.

The market opportunity is real and substantial. The technology clearly works in controlled settings and delivers measurable value. But the path from pilot to production to scale is fraught with technical, operational, and regulatory challenges. Most companies in this space will fail not because the core technology doesn't work but because they can't solve the integration, safety, and generalizability problem at venture scale.

From a portfolio construction perspective, I'd suggest a barbell approach. Either invest very early at seed or even pre-seed in teams with deep domain expertise and novel technical approaches, or wait until Series B or C when companies have demonstrated real-world traction, solved the safety problems, and built defensible distribution. The Series A stage in this category is particularly dangerous because companies have proven the technology works in pilots but haven't yet validated it works at scale across diverse settings.

The other consideration is diversification across different market segments. AI scribing for primary care is different from AI scribing for specialty care, which is different from AI scribing for mental health, which is different from AI scribing for telemedicine. Each segment has unique requirements, workflows, and competitive dynamics. A portfolio approach might involve backing one company in each segment rather than concentrating in a single horizontal player trying to serve all market segments.

There's also the question of timing. The systematic review makes it clear that the technology is still maturing. Safety concerns, accuracy issues, and generalizability problems are real and unsolved. You could make an argument that it's too early to invest because the technology isn't ready for prime time, or you could argue that

exactly the right time because the companies that solve these problems over the 24 to 36 months will capture enormous value. I lean toward the latter view, but with the caveat that you need to be backing teams that actually understand these problems and have credible plans to solve them.

Due Diligence Red Flags When Evaluating AI Scribing Startups

Based on this systematic review and my own experience evaluating health tech companies, here are the red flags that should make you pause or pass on an AI scribing investment:

If the company's pilot data is all from simulated consultations or controlled environments, that's a major concern. Real-world performance will be meaningfully worse. Ask to see data from actual clinical settings with actual patients and actual workflow integration. If they can't show you that, they're too early.

If the company doesn't have a clear plan for handling transcription errors and patient safety, that's disqualifying. This isn't a nice-to-have feature, it's fundamental to the product. Ask about their error detection mechanisms, their human-in-the-loop processes, and their liability insurance. If they're hand-waving about safety, walk away.

If the company's customer references are all single-physician practices or small clinics, that tells you they haven't solved the enterprise integration challenges. Ask about their largest deployment, their EHR integration capabilities, and their implementation timeline for new customers. If implementations take 6 to 12 months and require significant custom development, that's a scaling problem.

If the company's demo only works in quiet environments with clear audio, that's a red flag. They haven't solved the robustness problem. Real clinical environments are noisy. To see demos with background noise, multiple speakers, accents, and interruptions, if the system falls apart in those conditions, it's not ready for real-world deployment.

If the company doesn't have any clinicians on the founding team or in senior leadership, that's a yellow flag. Building healthcare AI requires deep clinical domain expertise. You need to understand clinical workflows, documentation requirements, medical terminology, and patient safety. Pure technologists often underestimate complexity of the clinical domain.

If the company's differentiation is purely "we have better AI models," be skeptical. Model quality matters, but it's increasingly commoditized as foundation models improve. Sustainable differentiation comes from clinical domain expertise, EHR integration capabilities, workflow optimization, and distribution. Ask what they have beyond model performance.

The Companies That Might Actually Win

I'm not going to name specific companies here because this isn't investment advice and I don't want to inadvertently pump or dump anyone's portfolio companies. I will describe the characteristics of companies that I think have the best shot at winning in this market based on what the systematic review reveals.

Winners will have deep clinical domain expertise built into their product and team. They'll understand that transcription accuracy is necessary but insufficient. The value is in understanding clinical documentation requirements, anticipating what information physicians need to capture for quality reporting and billing, and structuring that information in ways that integrate seamlessly with clinical workflows.

Winners will solve the safety problem through a combination of better AI models and thoughtful human-in-the-loop workflows. They won't try to eliminate human oversight entirely because that's dangerous and probably impossible with current technology. Instead, they'll design systems that make human review efficient and effective, highlighting areas of uncertainty, flagging potential errors, and making it easy for physicians to correct mistakes.

Winners will build deep EHR integrations that go beyond simple copy-paste. They will understand Epic's data model, Cerner's APIs, and Athenahealth's workflows. They

invest in the unglamorous work of building and maintaining integrations with d of EHR systems. This is expensive and time-consuming, but it's what separates products that get pilot tested from products that get deployed at scale.

Winners will demonstrate their systems work across diverse patient populations clinical settings, and geographic markets. They'll have data showing their accuracy holds up with accented English, noisy environments, complex consultations, and specialized medical terminology. They'll be honest about where their system struggles and have clear plans to address those gaps.

Winners will have thoughtful go-to-market strategies that account for the complexity of healthcare sales. They'll understand that selling into health systems requires navigating IT, clinical leadership, compliance, and procurement. They'll have customer success capabilities that can support implementation and ongoing optimization. They'll think about distribution through EHR partnerships, GPO relationships, or other channels that provide leverage.

Winners will also probably have raised sufficient capital to survive the trough of sorrow between pilot success and scaled deployment. Based on this review, it's clear that getting AIVT to work well in controlled settings is achievable, but getting it to work reliably across diverse real-world settings is much harder. That transition requires capital, time, and patience. Companies that are undercapitalized or facing near-term funding pressure will be tempted to overpromise and underdeliver, which is how you end up with patient safety incidents and reputational damage.

The other characteristic of potential winners is strategic clarity about whether to build a horizontal platform or focusing on specific verticals. The review suggests that different clinical specialties and settings have meaningfully different requirements. A horizontal approach requires solving for the lowest common denominator and may result in a product that's mediocre across all settings. A vertical approach allows for deeper optimization but limits TAM. The companies that will win have made a clear strategic choice and executed against it rather than trying to build things to all people.

Finally, winners will probably have some level of academic validation and clinical credibility. The systematic review itself is evidence that the academic community is taking AIVT seriously and trying to rigorously evaluate its impact. Companies that engage with academic medical centers, publish their results in peer-reviewed journals, and welcome independent evaluation will build credibility that helps with both customer acquisition and regulatory navigation. Companies that avoid scrutiny or only share cherry-picked results will struggle to build trust with risk-averse healthcare systems.

Look, the bottom line is this: AI scribing is a real market opportunity with genuine potential to create value for physicians, patients, and investors. But it's also a market with significant technical, operational, and regulatory challenges that most companies won't successfully navigate. The systematic review provides valuable data on what works, what doesn't, and where the gaps remain. For angel investors, that means being selective, asking hard questions, and recognizing that the companies that win will be those that solve not just the AI problem but also the integration, safety, and scalability problems that plague most healthcare AI deployments.

The TAM is big enough that there will probably be multiple winners, but there will also be dozens of failures. Your job as an investor is to identify which is which in the market. This review gives you a framework for doing that due diligence on it.



1 Like • 1 Restack

[← Previous](#)

[Next →](#)

Discussion about this post

Comments

Restacks



Write a comment...