Title

AI scribe functions in psychiatric practice: clinical oversight, consent, and regulation

Journal

The Lancet Psychiatry, 13(9)

ISSN

2215-0366

Authors

Roth, Alexander S

Ayers, Nolan B

Publication Date

2026-09-01

DOI

10.1016/s2215-0366(26)00201-4

Published as: Roth AS, Ayers NB. AI scribe functions in psychiatric practice: clinical oversight, consent, and regulation. Lancet Psychiatry 2026; 13: 804–10. https://doi.org/10.1016/S2215-0366(26)00201-4

Published article on The Lancet website: https://www.thelancet.com/journals/lanpsy/article/PIIS2215-0366(26)00201-4/fulltext

The Lancet Psychiatry homepage: https://www.thelancet.com/journals/lanpsy/home

---

Your AI scribe is doing three things at once: they are not the same

Alexander S Roth[1,2], **Nolan Ayers**[1]

[1]Department of Psychiatry and Psychology, Mayo Clinic, Rochester, MN, USA

[2]Department of Psychiatry and Behavioral Sciences, University of California, San Francisco, CA, USA

Correspondence to: Dr Alexander S Roth, Department of Psychiatry and Behavioral Sciences, University of California, San Francisco, CA 94143, USA; Alexander.Roth2@ucsf.edu

Summary

Ambient AI documentation tools have entered psychiatric practice bundling at least three functionally distinct operations into a single workflow, without sufficient attention to the differences among them. In this Personal View, we argue that AI-generated clinical narrative (history of present illness), AI-generated clinical observation (mental status examination), and AI-generated diagnostic reasoning (assessment and plan) are functionally distinct operations that draw on the same input but stand in qualitatively different relationships to it, with distinct failure modes and oversight demands. We describe why psychiatric practice exposes these distinctions with particular clarity and examine their implications for liability, informed consent, and regulation. Psychiatry is uniquely positioned to surface these

distinctions because of the therapeutic and legal weight of word choice, the unnarrated structure of the mental status exam, and the formulation-driven nature of diagnostic reasoning. We offer provisional practice considerations for clinicians currently using these tools and call for the field to build frameworks that reflect the different claims these tools' outputs make.

Introduction

Ambient AI documentation tools are now in routine use across psychiatric settings.1,2 Large language models generate clinical notes from session transcripts, promising to reduce administrative burden and return physician attention to the patient. The psychiatrist who can be fully present with the patient may gather richer clinical data and provide the patient with the experience of being seen, and presence is itself a therapeutic intervention.

In practice, an AI scribe produces, in a single note, a clinical narrative, a mental status examination, and an assessment and plan. This article, focused on current ambient, transcript-dominant documentation tools, makes three claims about these outputs. First, they correspond to functionally distinct operations (narrative synthesis, observational inference, and diagnostic reasoning) that differ in their relationship to the available data and in their characteristic failure modes. Second, bundling them into a single tool confounds efforts to evaluate each operation independently and creates the possibility of cascading errors across them. Third, psychiatry makes these distinctions especially stark because of the therapeutic and legal weight of word choice, the unnarrated structure of the mental status exam, and the formulation-driven nature of diagnostic reasoning.

The psychiatric note as clinical instrument

The psychiatric note is a clinical instrument that serves multiple audiences simultaneously. The traditional functions, communication among the clinical team, billing justification, and legal protection, have been well described in training for decades. The 21st Century Cures Act,3 which took effect in 2021, formally added a fourth audience by mandating that patients have direct, real-time access to their clinical notes. A fifth audience is often left unnamed, the physician themself. The act of composing a note is itself a form of clinical reasoning, a process through which the physician may make sense of the encounter and arrive at a formulation.4,5 Drafting the note is where clinical thinking often takes shape, and removing that step risks losing nuance for the sake of efficiency. Clinicians review what these tools draft, and that review is genuine oversight. Whether it engages the same cognitive work as composing is a different question, one this article returns to throughout. The psychiatric note therefore serves five functions at once. As Childers observed, psychiatrists now navigate a documentation environment in which how we use words is never negligible.6 Others have argued that the patient-facing note should be reconceived as a therapeutic

tool that extends the clinical encounter.7,8 AI documentation tools were introduced into this already-complex, multi-audience environment without attending to any of these tensions.9

The three operations this article distinguishes map onto these functions unevenly, and they are not organised along a single gradient of risk; each presents a different kind of problem. Narrative synthesis, which reorganises content the patient actually said, bears most directly on the therapeutic, communicative, and legal weight of the words chosen. Observational inference, which describes features the recording may not reliably capture, bears on the evidentiary integrity of the record for every audience that relies on it. Diagnostic reasoning, which can shift the clinician from formulating independently to reviewing what the tool has generated, bears especially on the fifth audience, the physician's own cognitive process. The sections that follow take each in turn (table 1).

Three operations, one tool

We use the term operations to denote functionally distinct documentation tasks, defined by the relationship between what each section of the note asserts and the data available to the tool, not by any claim about the tool's internal computation. That computation is not open to inspection by the clinician who signs the note. A commercial tool may generate the entire note in a single pass of one language model, assemble it from separate generation calls, or route sections through different models entirely, and models with equivalent benchmark behaviour have been shown to differ substantially in their internal organisation.10 Multiple internal configurations can produce identical notes. Whether learned mechanisms within any of these systems differentiate transcribing from inferring from reasoning is an open empirical question for mechanistic interpretability research.11 What can be classified, regardless of internal configuration, is the evidentiary status of each section's claims. Different sections of the note make different claims to knowledge, but the tool presents them with identical documentary authority.

In current implementations, the data that all three operations receive consists primarily of the encounter recording, supplemented by limited chart context that will likely broaden as these tools develop. All three operations are missing the same things, including broader sensory context, comprehensive chart familiarity, and the clinical experience the treating psychiatrist brings to this specific encounter. A clinical narrative organising reported speech can largely be verified against the recording. A structured description of mental functioning at the time of the encounter often cannot, because many of its domains require sensory data the recording does not contain. A transcript-driven diagnostic formulation can be evaluated against the recording but lacks the longitudinal context that would validate or challenge that formulation.

Nor are these workflows always single-pass. Clinicians may dictate addenda after the encounter, upload their own notes, or keep taking notes alongside the tool, and the tool may fold that material into its draft. Iterative input of this kind narrows the gap between what the encounter contained and what the tool receives, because the clinician can supply observations the recording lacks. It does not change what kind of claim the generated content makes, and the dynamics of reviewing a draft the tool finalised first persist across these variations.

Separate from any gap in the tool's data, the concern about anchoring and cognitive outsourcing applies across all three operations and persists regardless of how much the tool's data improves, because it is about the reviewing physician's process, not the tool's input. The risk is most acute for diagnostic reasoning, where the formulation is abstract and difficult to independently verify.

Table 1: Three AI documentation operations distinguished by different kinds of problems

	Narrative synthesis (HPI)	**Observational inference (MSE)**	**Diagnostic reasoning (A/P)**
Type of output	Organisation and synthesis of reported content	Structured description of mental functioning at the time of the encounter	Diagnostic conclusions and treatment decisions
Main evidentiar	Transcript of patient speech	Varies by domain, from transcript-	Encounter transcript; chart

y basis		derivable to entirely unsupported	or other longitudinal elements largely not yet incorporated
Dominant failure mode	Omission, distortion, weaker therapeutic nuance	Inference presented as direct observation; findings generated without sufficient basis	Cognitive outsourcing, anchoring, premature closure
Clinician's chief review task	Evaluate across factual, therapeutic, communicative, legal, and cognitive dimensions	Confirm against direct observation	Recognise when the tool's framing anchors or colours the assessment

***HPI**=history of present illness. **MSE**=mental status examination. **A/P**=assessment and plan.*

Narrative (history of present illness)

An AI that organises what the patient said into an HPI is performing sophisticated transcription and synthesis. Clinicians who use AI-generated narrative must remain actively engaged in shaping the output. But the informational basis is sound at its core. The psychiatrist's task is to evaluate the output not only for factual accuracy but for the therapeutic, communicative, legal, and cognitive dimensions described above.

The limitations of these tools are particularly apparent in the subtleties of psychiatric interviewing. Consider the choice between documenting "cluster B traits" and "borderline personality disorder". A psychiatrist selecting the former might be making a deliberate therapeutic choice, conveying actionable diagnostic information to the next provider while being mindful of where the patient is in their own process of understanding their diagnosis, or anticipating that the formulation may evolve over subsequent encounters. Neither term is inherently preferable; each has its place. The issue is not that the substitution is uncorrectable. A clinician who made the deliberate choice can restore it on review, and often will. But the substitution is not flagged as a substitution. It reads as an ordinary, plausible note, and catching it depends on recalling a judgment that was never spoken aloud, with nothing in the transcript to prompt the recollection. Verification of automated output is effortful and imperfect, particularly under the time pressure that motivates adoption of these

tools.12,13 The more precise-sounding phrasing may be the less appropriate one, and nothing in the note will mark it as a choice.

The stakes of word choice become especially acute in safety documentation. There is a significant difference between “patient denies suicidal ideation” and “patient does not endorse suicidal ideation”. The first implies the clinician asked directly and the patient actively negated. The second documents that the clinician had suicide risk on their radar without specifying whether a direct inquiry occurred. A clinician might select the latter phrasing intentionally, for a range of reasons such as when safety has already been assessed through other means and direct inquiry in a particular encounter is not the priority. To be clear, this flexibility in documentation should never be confused with flexibility in safety assessment itself. When safety is the overriding clinical variable, the psychiatrist should not defer the question.

The phrasing “does not endorse” threads multiple functions of the note simultaneously, addressing liability, clinical communication, the therapeutic relationship, and the physician’s own record of their clinical reasoning. An AI scribe operating on the transcript is likely to omit safety language regarding suicidality entirely when the topic was not explicitly discussed, because the AI documents what was said. The clinician documents both what was discussed and what was considered. Many clinicians will add the missing language themselves. What changes is the nature of the task. The burden shifts from composing the language to noticing its absence, and omissions in fluent text may be harder to detect than errors of commission during review of automated output.12

Observation (mental status examination)

Current AI documentation tools routinely generate a mental status examination from the session recording,1 but not all MSE components can be reliably produced this way.

The reason is structural. In a medical or surgical encounter, AI scribes transcribe the physician's narrated findings (“lungs clear bilaterally, abdomen soft, nontender”). In psychiatry, the mental status examination is typically not narrated aloud during the encounter, in part because narrating findings could disrupt rapport and shift the texture of the encounter in ways a psychiatric interview may not tolerate. The AI is therefore not transcribing stated findings but inferring unstated observations from indirect evidence.

Two distinct problems follow, and they have different remedies. The first is a problem of data availability. Circumstantial or tangential thought process can be reasonably identified from a transcript alone. But features such as flattened prosody, tremulous voice, or interruptibility require analysis of the acoustic signal itself rather than the transcribed words. Most current tools

operate primarily on transcription, not waveform analysis, meaning that even audio-accessible MSE components might be less reliably captured than they appear. Emerging research on vocal biomarkers suggests that signal-level analysis has genuine diagnostic potential,14 but this work remains early-stage and largely separate from clinical documentation tools. At the far end, findings such as eye contact, kinetics, or olfactory ones like an unwashed stench or alcohol on the breath are not accessible through audio or text at all. This first problem is contingent on what the tool can capture. Multimodal systems that add video or other sensory channels would narrow it, and such capabilities are advancing.

The second problem would survive the first one's solution. Whatever data the tool draws on, the findings it generates for these fields are probabilistic inferences, yet the note presents them in a section whose format has always signified direct clinical observation. The issue is not that the inferences are always wrong but that they are presented with the documentary authority of examined fact, that the record gives the reader no way to distinguish what was observed from what was generated, and that the cases where the correlation breaks down are often the most clinically significant. Human MSE documentation has its own well-documented reliability limitations,15 but current AI tools generate observational findings from data that cannot confirm them, introducing a novel kind of error into the record. The remedy for the first problem is richer capture. The remedy for the second is provenance labelling, an explicit record of which content was transcribed, which was inferred, and from what. The novel error lives in the second problem and persists after the first is solved, because findings generated from richer sensor data remain machine inferences rather than clinician attestations, and they may become harder to question as they become more plausible.

Reasoning (assessment and plan)

An AI that generates an assessment and plan is performing clinical reasoning and presenting it as documentation. In standard workflows, many tools capture the assessment and plan from the physician’s verbal summary to the patient at the end of the encounter rather than generating reasoning de novo. But in practice, practitioners might abbreviate or omit portions of their verbal reasoning, leaving the AI to fill inferential gaps from the transcript. The result is a spectrum between transcription of stated clinical reasoning and autonomous generation of unstated reasoning. This does not mean that AI-generated reasoning is inherently inferior. A well-constructed AI-generated formulation could surface diagnostic considerations the clinician had not prioritised. The problem is not that the tool reasons, nor that review is itself flawed, but that reviewing AI-generated reasoning differs from the process of generating one’s own.12 For the physician, writing the assessment can itself be part of how the formulation forms, particularly when the act of writing surfaces tensions not fully resolved

during the encounter; when the tool produces the words first, that cognitive work does not necessarily happen in the same way.4,5 That proposition is grounded in the broader literature on writing and clinical reasoning but has not been tested in AI-assisted psychiatric documentation, and we advance it as an inference rather than a finding.

Imagine a patient who describes mood swings during a session. An AI tool might generate an assessment favouring bipolar disorder. A psychiatrist with longitudinal knowledge of the same patient, and with access to the affective and behavioural qualities of the encounter that the recording does not capture, might weigh the presentation differently, recognising a pattern more consistent with emotional dysregulation in the context of characteropathy, or mood instability driven by substance use. The AI's formulation is not unreasonable given the available data, but it reflects reasoning from limited aspects of one encounter. The risk is not hypothetical. When investigators presented a leading medical language model with an oncology case in which staging information was deliberately insufficient, it returned a definitive stage in every trial under direct prompting, while another model appropriately flagged the uncertainty.10

What distinguishes this from other forms of clinical decision support is how naturally the output integrates into the workflow. A pop-up alert or a differential diagnosis generator presents itself as a separate input to be evaluated. An AI scribe embeds its reasoning directly into the documentation workflow, in the same format and location as clinician-generated content, making the cognitive outsourcing easier to overlook even for the physician performing it. This is not a claim that clinicians abandon review. Awareness of responsibility and liability cuts against wholesale outsourcing, and physicians using these tools describe substantial time spent editing their drafts.13 The narrower claim is about sequence. When the tool's formulation is accessed before the clinician's own has set, review begins from the tool's framing rather than the clinician's.

The bundling problem

Bundling these three operations confounds any effort to evaluate any single one.16 More critically, it creates the possibility of cascading cross-category failures. Consider a patient with worsening depression who presents with substantial neurovegetation that the psychiatrist would have noted on direct observation. An AI-generated MSE, drawing on the patient's coherent verbal responses, populates the record with "normal psychomotor activity". The AI then generates an assessment consistent with outpatient depression management. The clinician reads a note in which the observation and the plan cohere, and is anchored to a clinical picture that does not reflect what they would have seen had they documented their own observations first. They might find it convincing enough to accept or overweigh. The failure is not in any single operation but in the interaction between an inferred

observation and a downstream assessment, each of which might have been caught independently but which together produce a coherent, plausible, and wrong clinical narrative. This failure mode does not depend on any particular internal configuration. The check that would catch the error, comparison of the inferred observation against what the clinician actually saw, requires data the tool was never given, so no internal architecture, however arranged, can perform it. The most a tool could do is mark the inference as an inference, the provenance distinction described above. And where the draft arrives as a finished note, as it typically does, no human review intervenes between the generation of the observation and the reasoning that absorbs it. To the extent a single undifferentiated process produces both sections, as is plausible for current systems, the path from the inferred observation to the assessment is more direct still.

Liability, consent, and regulatory implications

Under current malpractice frameworks, the physician who signs a note bears full responsibility for its contents.17 If a patient's safety assessment is insufficiently documented and the patient subsequently dies by suicide, liability turns on whether the documentation reflects what a reasonable psychiatrist would have done, regardless of what actually occurred in the room. A psychiatrist who conducted a thorough safety assessment but signed an AI-generated note that failed to reflect that assessment might face liability not because the clinical care fell short but because the documentation does not demonstrate otherwise.

Current consent practices at many institutions focus narrowly on audio recording for documentation purposes, without capturing the full scope of what AI systems do with the recorded content.18 A patient who consents to having their visit recorded "to assist with documentation" might not understand that the same system is generating diagnostic impressions, populating a mental status exam, or proposing treatment plans. The analogy to a supervising physician working with a resident is instructive but imperfect. Supervision evaluates the resident's reasoning interactively. The resident presents reasoning rather than only a conclusion, can be questioned in real time, and operates within a recognised framework of graduated autonomy. Some tools may already invite a version of this, allowing the clinician to question the draft conversationally. The reply, however, is itself generated content, with the same evidentiary status as the draft it explains, not testimony from an accountable trainee within a longitudinal supervisory relationship.19 Most decisive for consent, patients are generally aware of resident involvement and are often unaware of the tool's inferential role. Consent matters here not simply because AI was used, but because AI use might alter the quality of the clinician's independent judgment through cognitive offloading and anchoring in ways that the patient cannot evaluate from the outside. Commentary on ambient listening has similarly argued that consent processes must move beyond

narrow disclosure of audio recording toward meaningful acknowledgment of AI’s inferential role.20 A recent single-site evaluation of 103 patients and 18 clinicians found that patients reported substantially different comfort levels depending on the AI's role, with 63% comfortable with AI for routine note generation, 43% for AI-supported treatment planning, and 30% for AI-supported diagnosis.21 That study combined narrative and observational documentation within routine note generation, so the gradient supports the broader claim that patients differentiate among AI functions rather than the three operations distinguished here, though the direction is consistent with this section's argument.

If these three operations carry different informational foundations and risk profiles, they might warrant different regulatory scrutiny.22 In particular, AI-generated diagnostic reasoning that more heavily influences clinical decision-making might meet criteria for classification as a medical device, a question that current regulatory frameworks have not yet addressed in the context of documentation tools.

A call for governance must define its object. A complete framework linking the mechanistic layer of these systems, their outputs, and the user's conception of them is beyond the scope of this Personal View, but the object we propose sits at the output layer. What can be evaluated today, irrespective of how a given tool is configured internally, is the relationship between what each section of the note asserts and the evidence available to the tool, together with the provenance of each assertion. A maturing regulatory regime could compel access to a tool's internals; what no mandate can yet supply is the science to interpret them. Mechanistic interpretability research has begun to trace how large language models compute their outputs,23 and these methods are reaching medicine,24 but the field is young, its findings do not yet transfer reliably across architectures,10 and current explanation methods cannot bear clinical or regulatory weight on their own.25 If that science matures, mechanism-level oversight could complement what we propose here; until then, oversight of documentation tools should rest on what is verifiable at the output layer, which is where the three operations are defined.

Discussion

The core argument of this paper is an observation about current tools, not a speculation about future ones. Transcript-dominant AI documentation tools in psychiatry bundle at least three functionally distinct operations, and the differences among them matter for clinical oversight, consent, and regulation. The claims about anchoring effects and cognitive offloading are inferences from established research on automation bias12,26 applied to a new context.

Early evidence is consistent with these concerns. AI-scribed primary care notes have been found to contain greater documentation of psychiatric

symptoms yet to be associated with a lower likelihood of psychiatric intervention,27 suggesting that more thorough documentation does not necessarily translate into more attentive care. The first randomised trial of commercial ambient scribes documented modest reductions in documentation time and improvements in burnout, alongside reports of occasional clinically significant inaccuracies.28 Qualitative work with physicians using these tools likewise emphasises the editing the drafts require and the time it consumes.13 A UK proof-of-concept in a child and adolescent mental health service reported substantial time savings and near-total patient acceptance, though it did not assess note accuracy or downstream clinical decision-making.29 Recent commentary has argued that the current evidence base reflects outcomes easier to count rather than patient experience, population health, equity, and clinician wellbeing.30 The epistemic distinctions among what the tool transcribes, infers, and generates, and their implications for clinical accuracy and patient safety, remain similarly unmeasured.

Concern extends beyond documentation quality to the formation of trainees themselves, since writing notes is part of how clinicians develop diagnostic reasoning. Proposed responses include preserving particular note sections as clinician-generated,31 protecting the trainee’s role as active author rather than passive editor,32 and building faculty supervision frameworks that prevent deskilling.33 Other research on public attitudes toward automated systems suggests that tolerance for errors by machines is substantially lower than for equivalent human errors,34 a finding with implications for consent and institutional risk management.

These observations suggest provisional considerations for those who build, deploy, and work with the current generation of AI documentation tools, summarised in the panel. They are organised by audience because the three operations assign different work to different actors. In today's transcript-dominant environment, the clinician's task is to protect direct observation and their own formulation. The developer's task is to make explicit what kind of claim each part of the note makes. The organisation's task is to align consent and validation with the tools' full scope of function.

Panel: Considerations for current ambient, transcript-dominant AI documentation tools

For clinicians:

- Where clinically appropriate, verbalise the assessment and plan to the patient at the close of the encounter, making it more likely that the tool transcribes stated reasoning rather than generates unstated reasoning.

- As soon as possible after the encounter, record observations the recording cannot contain, such as appearance, eye contact, psychomotor findings, and odour, ideally before reading the AI draft.
- Where feasible, form and note an independent impression before reading the AI-generated assessment, so that more formulation precedes review.
- Review and edit the draft while unvoiced observations and judgments remain retrievable.
- Treat AI-generated mental status content as inference until confirmed against direct observation.
- Review AI-generated history for therapeutic appropriateness, clinical nuance, and the legal weight of word choice, not only factual accuracy.
- Search the draft for what is absent as well as what is wrong; omissions in fluent text may be harder to notice than errors.

For tool developers:

- Provide labels for the provenance of every field, distinguishing transcribed content from inferred content and recording the basis for each inference.
- Do not populate by default mental status fields the available input cannot support; suppress them or flag them for clinician entry.
- Allow each operation to be enabled, reviewed, and evaluated independently, so that adopting AI-generated narrative does not require accepting AI-generated observation or assessment.
- Support capture of the clinician's own observations during or soon after the encounter, so that direct findings enter the note rather than being lost.
- Offer configurations that preserve designated sections, particularly the assessment, as clinician-generated.31

For healthcare organisations:

- Ensure consent processes reflect the full scope of the tool's function, including the generation of observations and diagnostic impressions, not only audio recording for documentation.
- Validate each tool as deployed; equivalent benchmark performance does not imply equivalent behaviour in use.10

These considerations reflect the authors' clinical reasoning rather than validated guidelines.

Limitations

This article is a Personal View informed by clinical observation and selective review of the emerging literature, not a systematic review. It assumes full fidelity between the spoken word and the transcript the AI ingests; transcription errors introduce an additional layer of distortion not addressed here. The claims about reviewing versus generating clinical content, while grounded in the broader literature on cognitive effort and anchoring,12,26 have not been tested specifically in the context of AI-assisted psychiatric documentation. These arguments also presume that individual clinician reasoning remains the appropriate baseline against which AI-assisted workflows should be evaluated; whether that assumption holds as AI capability develops is itself a question the field may need to revisit. The spectrum within the MSE (which components are reasonably extractable from audio and which are not) would benefit from further empirical investigation.

Conclusion

AI documentation tools in psychiatry are not doing one thing. They are doing at least three, and the three are not the same. Narrative synthesis, observational inference, and diagnostic reasoning differ in how each reaches beyond the available data, what can go wrong, and what the psychiatrist must do to ensure the output serves all five functions of the note. The field should stop accepting AI documentation as one undifferentiated category and start building the frameworks that distinguish what these tools transcribe, what they infer, and what they generate.

Contributors

ASR conceived the framework, drafted the manuscript, contributed the clinical examples, and revised subsequent versions. NA contributed to conceptual development and provided critical revision of the manuscript for intellectual content. Both authors approved the final version for submission.

Declaration of interests

We declare no competing interests.

Acknowledgments

No external funding was received for this work. Companion articles address a clinical heuristic for problematic AI-patient relationships (Ayers & Roth, in preparation) and broader conceptual intersections between psychiatry and artificial intelligence (Roth, Ayers, & Breitinger, in preparation).

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Claude (Anthropic), Gemini (Google), and ChatGPT (OpenAI) to assist with the organisation and preparation of the manuscript. After using these tools, the authors reviewed, verified, and edited the content as needed and take full responsibility for the content of the publication.

References

1 Buckley P, Wang Y, Gopalan P. Artificial intelligence scribes in psychiatry. *Focus (Am Psychiatr Publ)* 2025; **23**: 44–48.

2 Tierney AA, Gayre G, Hoberman B, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catal Innov Care Deliv* 2024; **5**: 3.

3 21st Century Cures Act, Pub L No. 114-255, 130 Stat 1033 (2016).

4 Bowker D, Torti J, Goldszmidt M. Documentation as composing: how medical students and residents use writing to think and learn. *Adv Health Sci Educ Theory Pract* 2023; **28**: 453–75.

5 Richardson KM, Cristiano JA, Schafer KR, Shen E, Burns CA. Writing is thinking: implementation and evaluation of an internal medicine residency clinical reasoning and documentation curriculum. *Med Sci Educ* 2022; **32**: 773–77.

6 Childers C. Navigating psychiatric documentation in the era of open notes. *Psychiatr News* 2024; **59**: 11.

7 Blease C, O'Neill S, Walker J, Hägglund M, Torous J. Open notes become law: a challenge for mental health practice. *Psychiatr Serv* 2021; **72**: 750–52.

8 Smith HS, Stavig M, McCann R, Moskovich A, Merwin R. "Let's talk about your note": using open notes as an acceptance and commitment therapy based intervention in mental health care. *Front Psychiatry* 2021; **12**: 703959.

9 Leung TI, Coristine AJ, Benis A, et al. AI scribes in health care: balancing transformative potential with responsible integration. *JMIR Med Inform* 2025; **13**: e80898.

10 Modi M, Krull JE, Johnson D, et al. Understanding clinical reasoning variability in medical large language models: a mechanistic interpretability study. medRxiv 2026; published online Feb 13. https://doi.org/10.64898/2026.01.26.26344845 (preprint).

11 Sharkey L, Chughtai B, Batson J, et al. Open problems in mechanistic interpretability. arXiv 2025; published online Jan 27. https://arxiv.org/abs/2501.16496 (preprint).

12 Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *J Am Med Inform Assoc* 2017; **24**: 423–31.

13 Shah SJ, Crowell T, Jeong Y, et al. Physician perspectives on ambient AI scribes. JAMA Netw Open 2025; 8: e251904.

14 Anmella G, De Prisco M, Joyce JB, et al. Automated speech analysis in bipolar disorder: the CALIBER study protocol and preliminary results. *J Clin Med* 2024; **13**: 4997.

15 Blaabjerg ES, Hemmingsen RPA, Høegh E, Wang AG, Gefke M, Arnfred S. Variability between psychiatrists on domains of the mental status examination. *Nord J Psychiatry* 2020; **74**: 287–92.

16 Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med* 2022; **28**: 924–33.

17 Mello MM, Guha N. Understanding liability risk from using health care artificial intelligence tools. *N Engl J Med* 2024; **390**: 271–78.

18 Mello MM, Char D, Xu SH. Ethical obligations to inform patients about use of AI tools. *JAMA* 2025; **334**: 767–70.

19 Turpin M, Michael J, Perez E, Bowman SR. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. Advances in Neural Information Processing Systems 36 (NeurIPS 2023). https://arxiv.org/abs/2305.04388.

20 Cohen IG, Ritzman J, Cahill RF. Ambient listening: legal and ethical issues. *JAMA Netw Open* 2025; **8**: e2460642.

21 Lawrence K, Kuram VS, Levine DL, Sharif S, Polet C, Malhotra K, et al. Informed consent for ambient documentation using generative AI in ambulatory care. *JAMA Netw Open* 2025; **8**: e2522400.

22 US Food and Drug Administration. Artificial intelligence and machine learning in software as a medical device. Updated January 2025. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device (accessed April 16, 2026).

23 Lindsey J, Gurnee W, Ameisen E, et al. On the biology of a large language model. Transformer Circuits Thread, 2025. https://transformer-circuits.pub/2025/attribution-graphs/biology.html (accessed June 9, 2026).

24 Metzger A, Patil S, Sugarmann LR, Karabacak M, Margetis K. Application of sparse autoencoders to enhance mechanistic interpretability of large language models in medicine. JMIR AI 2026; 5: e81134.

25 Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021; 3: e745–50.

26 Croskerry P. From mindless to mindful practice: cognitive bias and clinical decision making. *N Engl J Med* 2013; **368**: 2445–48.

27 Castro VM, McCoy TH, Verhaak P, Ramachandiran AK, Perlis RH. Psychiatric documentation and management in primary care with artificial intelligence scribe use. *JAMA Psychiatry* 2026; **83**: 281–86.

28 Lukac PJ, Turner W, Vangala S, et al. Ambient AI scribes in clinical practice: a randomized trial. *NEJM AI* 2025; **2**: 10.1056/AIoa2501000.

29 Stanton N, Aziz A, Jakhra S, et al. Evaluating artificial intelligence ambient voice technology as a documentation assistant in psychiatry: proof-of-concept study. *BJPsych Bull* 2025; published online December 1. doi:10.1192/bjb.2025.10186.

30 Tierney AA, Lee K, Liu VX. Ambient AI scribes and the quintuple aim: what is counted and what matters. *JAMA* 2026; published online April 1. doi:10.1001/jama.2026.3529.

31 Kramer EN, Velicu V, Ross B, Trinh NH, Ambrose AJ. Artificial intelligence scribes in psychiatry training. *Acad Psychiatry* 2026; published online. doi:10.1007/s40596-026-02316-w.

32 Abernethy J, Shah A, Chen B, Reynolds S, Wright SM, O'Rourke P. Integrating AI scribes into medical education: guardrails for preserving clinical reasoning. *J Gen Intern Med* 2026; published online February 2. doi:10.1007/s11606-025-10149-w.

33 Gin B, Boscardin C, Abdulnour R-E. Educational strategies for clinical supervision of artificial intelligence use. *N Engl J Med* 2025; **393**: 786–97.

34 Pozharliev R, De Angelis M, Rossi D. Do not put the blame on me: asymmetric responses to service outcome with autonomous vehicles versus human agents. *J Consum Behav* 2023; **22**: 647–59.