Editorial

AI Scribes: Are We Measuring What Matters?

Enrico Coiera, PhD; David Fraile-Navarro, PhD

Australian Institute of Health Innovation, Macquarie University, Sydney, NSW, Australia

Corresponding Author:

Enrico Coiera, PhD
Australian Institute of Health Innovation, Macquarie University
Level 6, 75 Talavera Road
Sydney, NSW 2109
Australia
Phone: 61 2 9850 2403
Email: enrico.coiera@mq.edu.au

Related Article:
Comment in: https://medinform.jmir.org/2026/1/e103162/

Abstract

Artificial intelligence (AI) scribes, software that can convert speech into concise clinical documents, have achieved remarkable clinical adoption at a pace rarely seen for digital technologies in health care. The reasons for this are understandable: the technology works well enough, it addresses a genuine pain point for clinicians, and it has largely sidestepped regulatory requirements. In many ways, clinical adoption of AI scribes has also occurred well ahead of robust evidence of their safety and efficacy. The papers in this theme issue demonstrate real progress in the technology and evidence of its benefit: documentation times are reported to decrease when using scribes, clinicians report feeling less burdened, and the notes produced are often of reasonable quality. Yet as we survey the emerging evidence base, there remains one outstanding and urgent unanswered question: Are AI scribes safe? We need to know the clinical outcomes achievable when scribes are used compared to other forms of note taking.

JMIR Med Inform 2026;14:e89337; doi: 10.2196/89337

Keywords: artificial intelligence; AI; digital scribes; medical scribes; electronic documentation; patient safety; error prevention

Beyond Efficiency

Early evaluation of any technology naturally focuses on its primary promise, in the case of scribes [1], reducing documentation burden. The studies in this collection confirm that artificial intelligence (AI) scribes deliver on this front. Kanaparthy and colleagues' [2] rapid review found a general trend toward reduced self-reported documentation time and improvements in clinician satisfaction. The comparative analysis by Ha and colleagues [3] demonstrates that current commercial systems can generate reasonable-quality SOAP (subjective, objective, assessment, and plan) notes in about a minute after a 15-minute encounter. These are meaningful findings. Having shown that scribes likely save time in the settings they have so far been evaluated in (eg, primary care and outpatient settings), we can turn to harder questions of safety, clinical reasoning, and wider system-level effects.

Toward Systematic Harm Measurement

How safe are clinical scribes? Do they make errors, what type of errors might see when using digital scribes, and are these errors clinically consequential? What are the causes of errors, and what then are the harm mitigations we need to put into place?

Several papers in this collection begin this work. Ha and colleagues [3] highlight that none of the systems they evaluated are error free. Biro and colleagues [4] have developed and validated an instrument specifically designed

to assess the accuracy and safety of AI scribe outputs, an essential foundation since, without standardized measurement, we cannot compare across systems or track changes over time. Their early work confirms that AI scribes do make errors, and some have patient safety implications.

Digital technologies have the capacity to both reduce human error and generate new error classes [5]. This raises another question worthy of further investigation: Are digital scribe errors equivalent to human errors, or do they have different risk profiles? A human might lose attention momentarily; a digital scribe will not fatigue but may misrecognize words through speech recognition errors. An AI might confidently fabricate a medication or symptom that was never mentioned or omit clinically significant details in the pursuit of conciseness, each reshaping the clinical narrative in different ways.

Equally, we should recognize that the status quo carries its own safety risks: clinician burnout and cognitive overload contribute to errors that scribes may help reduce. A complete safety evaluation must weigh new risks introduced by AI against existing risks that the technology may mitigate and consider what analogous error-detection and correction mechanisms we need to build for AI-generated documentation.

Clinical Reasoning When Using a Digital Scribe: An Underexplored Frontier

Perhaps the richest opportunity for future research relates to the quality of clinical care when documentation is outsourced to AI. Note-taking is not merely administrative work. When a clinician summarizes their thoughts into a document, they are actively processing information, prioritizing what matters, and forming and testing hypotheses in real time. The clinical note is not just a record of the consultation; it is a cognitive artifact that supports clinical reasoning [6].

What happens to this human sense-making process when documentation is delegated to an AI? We may be freeing the attention of clinicians to be more present with patients, or we may be altering how they would normally think. The Y-KNOT (Your-Knowledgeable Navigator of Treatment) implementation study offers an intriguing signal: expert ratings suggested that while most AI-generated drafts were rated positively, around 1 in 6 preanesthetic assessments were judged to have a negative impact on clinical decision-making [7]. As we document faster, are we also documenting differently? Does that difference matter for patient care? If so, can we mitigate the potential for harm, for example, through clinical training or redesigning the user interaction with a scribe to bring clinicians back into the document-and-reason loop?

Ecosystem Perspectives

AI scribes sit at a critical position within a clinical workflow. They determine what gets recorded and how it is structured, which in turn shapes what downstream systems, including other AI tools, will see and act on. In this sense, scribes are a gateway technology, creating data layers that propagate through the health system. As we have argued elsewhere in the context of generative AI more broadly, it is helpful to view these technologies through an ecosystem lens that emphasizes system-level properties over isolated components [8]. With this perspective, we can evaluate the broader scribe ecosystem against system-level dimensions: resilience (how does care adapt when the scribe fails?), sustainability (what happens when cloud-based systems change or disappear?), and service interactions (does optimizing documentation affect other aspects of care?).

The patients in Leiserowitz and colleagues' [9] survey were generally open to an AI scribe when it was framed as supporting clinician focus. Their study also showed one interaction worth monitoring: a meaningful proportion of patients indicated they might withhold sensitive information if an always-listening device was present. Understanding these broader system effects will require looking beyond the scribe itself to the clinical environment it inhabits.

Regulatory Evolution

There are several reasons why we still lack safety evidence. AI scribes have proliferated [10] in part because they sit outside traditional medical device classification [11]. Many commercial scribes skirt the software as a medical device definition of a decision support system and so have evaded regulation. With no regulatory demands for robust safety evaluation, there is little commercial value, it seems, in publishing commercial safety data. The evidence in this collection suggests that we may need new regulatory thinking, not because current systems are demonstrably unsafe, but because they are different. These systems do influence clinical decisions, and unlike traditional medical devices, they are not static; the large language model underlying a scribe may be updated or retrained over time. Regulatory frameworks designed for deterministic, frozen technologies may need to evolve alongside the technology itself. The question is not whether to regulate but how to do so in ways that preserve innovation while ensuring ongoing safety. Potential mechanisms might include postmarket surveillance requirements, mandatory incident reporting for generative medical AI, or periodic re-evaluation as underlying models are updated.

Conclusion

The papers in this theme issue represent important progress. AI scribes appear to deliver on their core promise of reducing documentation burden, and the field is developing increasingly rigorous evaluation approaches. Building on this foundation, we see an opportunity to expand what

we measure: systematic assessment of errors and harms, investigation of effects on clinical reasoning, and attention to ecosystem-level dynamics. Documentation burden is real, and technologies that address it are welcome. The next chapter of research can help ensure that the time saved translates into better care.

Funding

This work was supported by the National Health and Medical Research Council (NHMRC) Centre for Research Excellence in Digital Health. EC and DF-N are supported by an NHMRC investigator grant (GNT2008645). The funders had no role in this manuscript. A large language model (Claude, Anthropic) was used to assist with editing this manuscript.

Authors' Contributions

Conceptualization: EC
Writing – original draft: EC, DF-N
Writing – review & editing: EC, DF-N

Conflicts of Interest

None declared.

References

1. Coiera E, Kocaballi B, Halamka J, Laranjo L. The digital scribe. NPJ Digit Med. 2018;1(1):58. [doi: 10.1038/s41746-018-0066-9] [Medline: 31304337]
2. Kanaparthy NS, Villuendas-Rey Y, Bakare T, et al. Real-world evidence synthesis of digital scribes: rapid review. JMIR AI. Oct 10, 2025;4:e76743. [doi: 10.2196/76743] [Medline: 41071988]
3. Ha E, Choon-Kon-Yune I, Murray L, et al. Evaluating the usability, technical performance, and accuracy of artificial intelligence (AI) scribes for primary care: a competitive analysis. JMIR Hum Factors. Jul 23, 2025;12:e71434. [doi: 10.2196/71434] [Medline: 40700466]
4. Biro J, Handley JL, Cobb NK, et al. Accuracy and safety of AI-enabled scribe technology: instrument validation study. J Med Internet Res. Jan 27, 2025;27:e64993. [doi: 10.2196/64993] [Medline: 39869899]
5. Magrabi F, Ong MS, Runciman W, Coiera E. An analysis of computer-related patient safety incidents to inform the development of a classification. J Am Med Inform Assoc. 2010;17(6):663-670. [doi: 10.1136/jamia.2009.002444] [Medline: 20962128]
6. Leung TI, Coristine AJ, Benis A. AI scribes in health care: balancing transformative potential with responsible integration. JMIR Med Inform. Aug 1, 2025;13:e80898. [doi: 10.2196/80898] [Medline: 40749188]
7. Kim J, Lee SY, You SC, et al. A bilingual on-premises AI agent for clinical drafting: implementation report (Y-KNOT project). JMIR Med Inform. Nov 24, 2025;13:e76848. [Medline: 41284981]
8. Coiera E, Fraile-Navarro D. AI as an ecosystem — ensuring generative AI is safe and effective. NEJM AI. Aug 22, 2024;1(9). [doi: 10.1056/AIp2400611]
9. Leiserowitz G, Mansfield J, MacDonald S, Jost M. Patient attitudes toward ambient voice technology to support an AI scribe program. JMIR Med Inform. Nov 27, 2025;13:e77901. [doi: 10.2196/77901] [Medline: 41308194]
10. Henry TA. 2 in 3 physicians are using health AI—up 78% from 2023. American Medical Association. Feb 26, 2025. URL: https://www.ama-assn.org/practice-management/digital-health/2-3-physicians-are-using-health-ai-78-2023 [Accessed 2026-01-28]
11. Digital scribes. Australian Therapeutic Goods Administration. 2025. URL: https://www.tga.gov.au/products/medical-devices/software-and-artificial-intelligence/manufacturing/artificial-intelligence-ai-and-medical-device-software/digital-scribes [Accessed 2025-12-08]

Abbreviations

AI: artificial intelligence
SOAP: subjective, objective, assessment, and plan
Y-KNOT: Your-Knowledgeable Navigator of Treatment

Edited by Andrew Coristine; This is a non–peer-reviewed article; submitted 10.Dec.2025; final revised version received 31.Dec.2025; accepted 14.Jan.2026; published 06.Feb.2026

Please cite as:
Coiera E, Fraile-Navarro D
AI Scribes: Are We Measuring What Matters?
JMIR Med Inform 2026;14:e89337

URL: https://medinform.jmir.org/2026/1/e89337
doi: 10.2196/89337

© Enrico Coiera, David Fraile-Navarro. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 06.Feb.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.