npj Digital Medicine

https://doi.org/10.1038/s41746-026-02554-0

Article in Press

Barriers and opportunities of scaling ambient AI scribes for clinical documentation across diverse healthcare settings

Received: 21 October 2025

Accepted: 6 March 2026

Cite this article as: Ohde, J.W., Thompson, A., Liu, Z. *et al.* Barriers and opportunities of scaling ambient AI scribes for clinical documentation across diverse healthcare settings. *npj Digit. Med.* (2026). https://doi.org/10.1038/s41746-026-02554-0

Joshua W. Ohde, Arjun Thompson, Zhenghong Liu, Lauren M. Rost, Joshua D. Overgaard, Jonathan Yue En Tan, Shauna M. Overgaard, Raymond Francis R. Sarmiento, Yuhe Ke, Jonathan Chong Kai Liew, Jasmine Chiat Ling Ong & Nan Liu

Barriers and Opportunities of Scaling Ambient AI Scribes for Clinical Documentation Across Diverse Healthcare Settings

Joshua W. Ohde[1*], Arjun Thompson[2*], Zhenghong Liu[2], Lauren M. Rost[1], Joshua D. Overgaard[3], Jonathan Yue En Tan[4], Shauna M. Overgaard[1,5], Raymond Francis R. Sarmiento[6], Yuhe Ke[7,8,9], Jonathan Chong Kai Liew[7,10], Jasmine Chiat Ling Ong[7,11+], Nan Liu[7,12,13,14,15+]

1. Center for Digital Health, Mayo Clinic, Rochester, MN, USA
2. Department of Emergency Medicine, Singapore General Hospital, Singapore, Singapore
3. Department of Medicine, General Internal Medicine, Mayo Clinic, Rochester, MN, USA
4. Department of Future Health Systems, Singapore General Hospital, Singapore, Singapore
5. AI Validation & Stewardship Research Program, Mayo Clinic Health System, Rochester, MN, USA
6. Metro Pacific Health Tech Corporation, Pasig City, Philippines
7. Duke-NUS AI + Medical Sciences Initiative, Duke-NUS Medical School, Singapore, Singapore
8. Department of Anesthesiology, Singapore General Hospital, Singapore, Singapore
9. Data Science and Artificial Intelligence Lab, Singapore General Hospital, Singapore, Singapore
10. Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
11. Division of Pharmacy, Singapore General Hospital, Singapore, Singapore
12. Centre for Biomedical Data Science, Duke-NUS Medical School, Singapore, Singapore
13. Pre-hospital & Emergency Research Centre, Health Services Research and Population Health, Duke-NUS Medical School, Singapore, Singapore
14. NUS Artificial Intelligence Institute, National University of Singapore, Singapore, Singapore
15. Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA

*Contributed equally
+Corresponding authors

Correspondence to:
Jasmine Chiat Ling Ong, Singapore General Hospital,1 Hospital Blvd, SingHealth Towers, Singapore. Email: jasmine.ong.c.l@sgh.com.sg

Nan Liu, Centre for Biomedical Data Science, Duke-NUS Medical School, 8 College Road, Singapore 169857, Singapore. Email: liu.nan@duke-nus.edu.sg

ABSTRACT

Ambient AI scribes are reshaping clinical documentation and clinician-patient interactions. These tools were initially tested in low-acuity ambulatory settings. However, their deployment in diverse care settings raises new challenges. This perspective examines the clinical, technical, and ethical implications of ambient AI scribes in diverse settings. With thoughtful integration, ambient AI scribes can evolve into valuable assistive tools to clinicians. Responsible use of these tools can improve interactions with patients, enhance safety, reduce clinician burden, and improve care continuity.

BACKGROUND

Ambient intelligence refers to a digital environment that is aware of, adapts to, and interacts with individuals.[1] It combines user-centred interfaces capable of remote sensing, listening, interpreting, and performing tasks. The emergence of large language models (LLMs) has dramatically transformed the ambient listening landscape in healthcare. One of the most promising applications is the ambient artificial intelligence (AI) scribe, which has shifted how clinical documentation is generated. Ambient AI scribes are voice-enabled AI tools capable of recording, transcribing and summarizing patient-physician conversations. In ambulatory care, physicians devote excessive time to electronic health record (EHR) documentation[2,3], often extending their workday by several hours. The time spent on documentation beyond usual business hours is commonly referred to as "pyjama time".[4] This administrative burden is directly linked to clinician burnout.[5] Ambient AI scribes aim to alleviate this burden by passively capturing and transcribing healthcare professional-patient conversations in real time, thereby reducing documentation workload, enhancing clinicians' attention to the patient, and improving the overall quality of records.[6] Additional benefits may come as new features of ambient technology takes shape, including the ability to facilitate ordering, identification of discrepancies in the EHR, support decision making, and reduce coding burden.

The adoption of LLM-based ambient documentation tools has rapidly expanded in outpatient and primary care settings. Early studies report promising outcomes, including reductions in total EHR usage, time spent on notes, and after-hours documentation without compromising patient safety, experience, or documentation quality.[3,7,8] However, these findings are inconsistent. Other investigations suggests no significant improvements in provider productivity or patient experience, highlighting variability in performance across clinical settings and possibly individual scribing tools.[9,10]

In this perspective, we examine the challenges and potential opportunities of integrating ambient scribes into diverse healthcare settings, with a focus on in-patient, high acuity, pre- and out-of-hospital, and low-resource settings. We discuss workflow specific, technical, and ethical considerations and propose strategies for responsible and context-aware implementation.

The Current State of Ambient Scribes

Recent studies across ambulatory and specialty care settings demonstrate variable time savings after ambient scribe implementation, with benefits varying between organisations and individual clinicians. Mass General Brigham observed modest but consistent improvements, with median total EHR time

reduced by 5.6 minutes per appointment, with the biggest gains in specialty practices and heavy EHR users.[3] In contrast, Permanente Medical Group's implementation, the largest to date, revealed minimal time savings of only 18 seconds per appointment compared to non-users.[11] In contrast, Intermountain Health's matched cohort study reported no statistically significant productivity gains.[12] This heterogeneity may be explained by study design–related factors, including variation in the types of ambient AI scribes evaluated, insufficient sample sizes, and the absence of standardized protocols for outcome measurement. To facilitate cross-study interpretation, several practical reporting signals have emerged in recent work, such as clinician edit burden (time spent revising drafts or the share requiring substantial changes) and source attribution/provenance (linking key assertions to diarized speakers/time stamps).[7,13,14]

In contrast to time metrics, consistent improvements have been observed in clinician well-being and patient interaction quality. Significant reductions in burnout scores, mental effort and cognitive load was observed across different studies.[7,15] Notably, clinicians spend nearly a third of visit time looking at their monitor for documentation.[16] Scribe technology is fundamentally reshaping the clinician-patient dynamic; potentially supporting increased eye contact, active listening, and empathy that paradoxically makes clinical visits more human via technological assistance. Patient experience also appears enhanced, with the Permanente Medical Group reporting that 84% of clinicians felt ambient scribes had a positive impact on visit interactions, and 56% of patients reporting a positive impact on visit quality.[11] Overall, while time savings vary, the most consistent value proposition for ambient scribes lies in reduced cognitive load, enhanced patient engagement, and clinician well-being. This collectively translates into potential improved quality of care, greater clinician satisfaction and retention, and a more sustainable healthcare workforce. However, questions about long-term sustainability, uniform benefit among clinicians, potential liability risks, and optimal implementation strategies remain.[17,18]

Primary care is particularly well suited for ambient AI scribe use considering its documentation-intensive nature and large administrative burden. The 2022 International Health Policy Survey by the Commonwealth Fund revealed a global crisis in primary care, with increasing workloads, widespread burnout, and alarming workforce shortages across 10 high-income countries.[19] In the UK, 75% of younger general practitioners described their work as very or extremely stressful, the highest among surveyed nations. In 2023, Canada experienced a record number of unfilled primary care residency positions alongside a growing exodus of providers.[20] These were driven in part by the high administrative burden in this complex and multifaceted field.[21] Primary care physicians often encounter patients with multiple comorbidities, generating significant documentation demands, making it a prime candidate for support from ambient AI scribes. While comprehensive data on their use remains limited, early evidence suggests that these tools had one of the highest perceived impacts on primary care physicians through a perceived reduction of administrative burden and documentation time, enabling clinicians to dedicate more attention to patient care.[8,22]

The Expanding Role of Ambient Scribes

High Acuity Settings

The existing literature has concentrated primarily on ambient AI scribes within low-acuity ambulatory environments that are often quieter, more structured, and involve more predictable workflows making them more conducive to consistent audio capture and transcription.[12,13] In contrast, high-acuity environments, such as emergency departments (EDs), intensive care units (ICUs), and operating theatres (OTs) are defined by their noisy, dynamic, and clamant nature. Such settings raise the stakes for timely and accurate documentation, where lapses can carry serious clinical and medico-legal consequences.[23,24] Furthermore, documentation requirements and scope of practice in these complex clinical settings differs significantly from low-acuity structured ambulatory environments[25–27], which may affect the utility of ambient scribe tools therein. High-acuity practice settings require bi-directional, real-time clinical data exchange, without limiting the location or number of simultaneous users.[26] It is unknown if the benefits observed in ambulatory settings translate to such environments. Future deployment may require evaluating physician fit, workflow adaptation, and exam-room or hardware needs as ambient solutions mature. Current tools are also limited in their ability to integrate prior patient history into generated notes.

Emergency Department

Vendors have begun marketing ambient scribe solutions specifically for ED workflows[28] while their use and evidence in critical care and inpatient settings remains limited. Capturing the interactions of multidisciplinary teams, dealing with repeated interruption, and interference from noisy environments remain significant technical barriers.[1]

Critical Care Units and Operating Theatres

Despite these limitations, ambient intelligence is not new in critical care or surgery. Ambient intelligence can prevent data fatigue and support clinician workflow by filtering high-value, context-specific information from EHRs, as demonstrated by Mayo Clinic's ICU system.[1] Sensors coupled with computer vision are capable of automated identification of clinically important bedside clinical actions such as ICU preventive bundle elements and patient mobilization.[29] Ambient scribe technology introduces an additional data stream that could improve the completeness and richness of captured clinical information.

Pre-hospital Care

Pre-hospital care is typically delivered by emergency medical services (EMS), such as paramedics, emergency medical technicians, firefighters, and police officers. Care in the pre-hospital environment is often highly acute but differs greatly from the ED or ICU setting. Effective pre-hospital care can significantly improve patient outcomes.[30] Accurate and time-sensitive information captured by ambient intelligence technology could be passively captured by devices attached to EMS personnel and instantly relayed to hospitals or trauma centers during emergency transfers. Another opportunity is the utilization of ambient scribes to document information provided by the public when calling emergency communication centers for emergency assistance, providing real-time, detailed documentation to EMS while enroute, ensuring continuity of care between the public, EMS, and hospital staff. We acknowledge

the lack of integration with hospital EHR systems pose practical limitations to maximizing the utility of EMS documentation and ambient intelligence technology.

Virtual Care Settings

Beyond traditional clinical environments, ambient AI scribes are increasingly relevant in virtual healthcare settings, such as remote patient monitoring and virtual home care, which have expanded rapidly due to telemedicine's growth. In virtual consultations, clinicians face the dual challenge of managing both the visual and verbal components of patient interaction alongside accurate and timely documentation. Ambient scribes integrated into telehealth platforms can capture conversational data in real-time without disrupting the flow of patient-provider interaction, thus enhancing clinicians' attention to clinical nuances rather than note-taking. In addition, these systems facilitate comprehensive documentation of teleconsultations, which can often become fragmented due to technology-related distractions or the absence of dedicated medical scribes. The ability to automatically generate structured clinical notes directly from remote audio feeds supports continuity of care by providing rich, standardized data for follow-up or multidisciplinary review across different providers and settings. Moreover, ambient scribes can be configured to recognize and flag critical clinical terms or alerts during virtual visits, which can potentially improve patient safety by prompting clinicians to focus on urgent findings even in the absence of in-person cues. Policymakers and healthcare organizations must address ethical considerations unique to virtual care and telehealth settings, such as ensuring patient privacy in crowded or home environments, and obtaining explicit informed consent for the use of AI tools such as scribes in virtual care settings.

Low- and Middle-Income Countries

Ambient AI scribes hold considerable promise for transforming healthcare documentation, especially in low-resource settings where healthcare systems face unique challenges such as limited infrastructure, a lack of medical training facilities, workforce shortages, and linguistic diversity. In these environments, ambient AI scribes can relieve overextended healthcare professionals, allowing them to dedicate more time to direct patient care.[11,31,32] By automating the documentation process in busy outpatient clinics or emergency wards in extremely low-resource settings, ambient AI scribes can help improve data accuracy and completeness, which are frequently compromised due to high patient volumes and low physician-to-population ratio. According to the World Health Organization, there will be an estimated global shortfall of 11 million health workers by 2030, with the greatest burden concentrated in low- and middle-income countries (LMICs).[33] In these contexts, the integration of ambient scribes holds considerable potential to mitigate the impact of workforce shortages and strengthen healthcare delivery capacity.

The capability of ambient scribes to adapt to multiple local languages and dialects through customized natural language processing models also addresses significant communication barriers that commonly impair clinical documentation quality, particularly in LMICs where multiple local languages and dialects are spoken interchangeably during a conversation. While training a model on every distinct language or dialect within a country is not practical, focusing on the key languages spoken across a nation may improve documentation completeness and accuracy. For example, Papua New Guinea has the most living languages spoken among any country with over 840 different languages. Three to six principal languages

are spoken throughout, with their distribution generally determined by geography and demographic factors, including age. An ambient scribe trained on this subgroup of languages could improve healthcare communication between this unique language composition.

Crucially, the low-resource context often involves intermittent or unreliable internet connectivity. Therefore, ambient scribe systems designed with on-device or edge-computing capabilities ensure continuous functionality without dependency on cloud connectivity, which is indispensable for consistent record-keeping and clinical decision support. It is important to note that patient privacy protections may be particularly vulnerable in LMICs, where legal, regulatory, and technical frameworks remain underdeveloped. In addition to infrastructural gaps, limited digital literacy and low awareness of data-protection requirements further amplify the risk of privacy violations. Leveraging ambient scribe technology thoughtfully will allow healthcare systems in resource-constrained and virtual contexts to enhance documentation quality, optimize clinician workflow, and ultimately improve patient outcomes.

Patient Outcomes and Experience

Studies of ambient scribes have mainly focused on clinician efficiency, well-being, time savings, and costs, with no evidence yet of improved clinical or patient-centered outcomes. While identifying potential errors and improving documentation completeness—particularly in high-acuity or time sensitive environments—might lead to reduced errors, reduced harm, and improved outcomes, this remains to be proven. Further research is needed. Conversely, patient experience represents a measurable and pertinent outcome as well, especially in consideration of clinician undivided attention during an appointment.

Challenges to Scaling and Implementation

Workflow Integration and Social Context

While one-on-one conversations between clinicians and patients are transcribed well, the integration of data from multiple sources asynchronously and source attribution currently prove challenging for an AI scribe.[18] Current AI scribe solutions struggle to integrate information across multiple source systems because of limited interoperability with disparate platforms and electronic health record systems. Another concern is “note bloat”, medical records that become excessively long or redundant, may occur when AI scribe systems capture excess verbatim without sufficient summarization.[7] Whilst reducing the documentation burden during the clinical encounter, this phenomenon may increase the duration clinicians spend reviewing prior documented clinical notes compared to traditional dictated summaries. Downstream workflow inefficiencies may arise when AI-generated clinical notes contain excessive non-essential information due to inadequate summarization. Such issues may stem from poorly optimized user prompts or from limitations in the AI scribe’s clinical contextual understanding. However, increased note length is not unique to AI ambient scribes.[34] Innovations such as personal microphones trained to identify the specific voice signature of the provider may allow for multiple providers to input data into the same AI model, with only one combined summary is produced. However, resultant workflow changes would require evaluation. Many AI scribe solutions include additional plug-in functionalities, such as pre-visit note generation. Similarly, lengthy AI-generated briefs may potentially distract

clinicians, rather than highlight critical information during pre-visit preparation. Whether these features enhance or hinder effective use of AI scribe solution ultimately depends on the performance of individual features, and effective technical orchestration of the full solution.

Integrating data across time presents another challenge. Unlike ambulatory settings where the patient usually encounters the provider once, in high acuity, pre-hospital, and low-resource settings, patients often have multiple shorter touch points, resulting in repeated documentation in the same day. Having to reengage the ambient scribe multiple times can reduce efficiency. However, the tool might also help to ameliorate the effect of having to recall patient details over several hours with other patients seen in the interval. A tool that can refer to existing documentation to generate context relevant notes would be ideal for such situations. Finally, clinicians, paramedics, and other healthcare staff often use jargon, acronyms, and shorthand that is unique to their clinical setting, profession, and social context. Ambient scribes need to adapt to local vernacular by contextual training and customization.

Evaluation Challenges

Existing evaluation frameworks for ambient AI scribes are largely designed for outpatient environments and focus on linguistic accuracy or physician satisfaction. These are insufficient for many diverse care settings where minor errors or omissions can have serious consequences. High-quality domain-specific benchmarks[35] which incorporate accuracy, clinical relevance, and potential for harm are urgently needed. Moreover, study designs that randomize clinicians to use or not use AI scribes will naturally result in the Hawthorne effect, limiting their ability to measure accuracy outcomes. Conversely, studies that recruit clinicians for survey feedback may result in a bias towards early adopters.

The lack of harmonized governance standards and clinically aligned evaluation benchmarks limit cross-study comparison and slows evidence-based adoption. A systematic review of AI-powered documentation tools used in healthcare examined 11 studies and found diverse spread of evaluation tools used for documentation quality.[36] These range from expert opinion Likert scale to semi-quantitative scales such as the Physician Documentation Quality Instrument (PDQI-9) tool. Efforts to develop international consensus for technical performance, documentation quality, patient safety, and ethical compliance are needed to accelerate progress. Measurement considerations that have gained traction include diarization-aware provenance, burden of clinician edits, documentation quality, and subgroup-sensitive reliability.[14,37,38]

Consensus guidelines and checklists can be laid out by experts and academics on the expected accuracy and testing approaches in each individual setting. These would lay the foundations for organisations to decide on whether each individual scribe is adoptable by conducting a guided evaluation and being aware of the metrics to be measured. Examples include the SCRIBE evaluation framework[14] and Stanford's MedHelm[39] collection of evaluation datasets for note summary. In addition, it is important to acknowledge that substantial differences exist between in silico contexts and the clinical environment.[40] Translational trials that include a silent testing phase, wherein a model's performance is evaluated prospectively in its clinical context without impacting patient care or workflow as the model output is not shown to end-users, may be warranted as part of the evaluation process before full implementation.[41,42]

International multi-centre collaboration is needed more than ever as the equity gap widens between high- and low-resource organizations, and to uncover potential biases across different geographical regions and practice settings. Organizations should work together to facilitate knowledge exchange, highlight the challenges learned by early adopters, promote technological advancements, and standardized implementation strategies and clear guidelines for ambient scribe integration in different settings including high acuity workflow environments. These metrics could be compared pre- and post-implementation of LLM evaluators to provide a measure of effectiveness. These metrics are important to monitor and tailor for each setting considering their unique workflows.

Technical Challenges: Audio differentiating and sensitivity

High acuity and pre-hospital settings present acoustic challenges. Noise coming from alarms, code announcements, multiple speakers, vehicle traffic, and proximity of patients threatens to compromise ambient scribe effectiveness and accuracy. Variables like these have not been fully evaluated in simulated or real-world settings, potentially exposing patients to unforeseen risks and additional clerical work for providers to edit unorganized and artifact-injected scribe notes. Even in a quiet and controlled environment available AI scribe products are unable to consistently identify or distinguish between multiple speakers. Additionally, most products record using a single recording device, like the physician's phone. Emergency situations can easily involve simultaneous conversations between multiple physicians, and allied health staff, and patient companions, limiting the ability of a single device to accurately capture and distinguish all conversations and their relevance to the clinical scenario.

Ambient scribes that can differentiate amongst relevant and irrelevant sounds would clear the path for widespread implementation by incorporation of voice recognition in addition to identification and exclusion of irrelevant content. While usual omni directional microphones can be very sensitive, other kinds of microphones with different polar patterns, specifically figure of 8, might allow for the ambient scribe to accurately pick up the conversation between a single provider and patient pair, while not picking up other noises. Portable, high-fidelity hardware may be necessary for successful implementation, with most systems being exclusively targeted at specific mobile operating systems such as iOS or Android. Testing in various environments would be highly valuable for the field as AI scribes expand into new use cases, operating systems, and settings.

Nuanced patient history, pertinent social history, pre-hospital event details, and contextual factors are often filtered out by currently available technologies.[43] In acute situations these details are critical to the downstream care decisions when a patient transitions from one high acuity setting to another, step-down to a general in-patient care, or discharge. In low-resource settings a patient's history may be completely undocumented but highly relevant to the current medical situation.

Ethical and Regulatory Considerations

Various ethical concerns and regulatory conundrums have surfaced following deployment of ambient AI scribes in clinical environments. Many ethical considerations common to the use of generative AI or LLMs in other contexts, such as model bias and automation bias; hallucinations and potential for

misinformation; lack of transparency in training data; as well as the legal constructs and implications of liability when errors occur.[44] Within clinical workflows, ethical conundrums extend beyond this to include transparency, privacy, fairness, and accountability. Interestingly, these tools are branded as "ambient", giving the impression that they are passive and perhaps misguiding as it does not clearly inform patients that their conversations are recorded and saved, often in cloud infrastructure outside the organization's EHR.

Data Retention, Security, and Governance

Ambient scribing tools may generate multiple data artifacts beyond the finalized clinical note. These include raw audio recordings and conversation transcripts, as well as draft outputs, prompts entered by users and provenance logs. However, governance of these data streams is often unspecified. Formal guidance have been developed in some countries, for example, the NHS England has provided a guidance to the use of AI-enabled ambient scribing products. In the absence of explicit guidance, health systems need to define what to retain, duration of retention and access controls, with clear distinction between transient processing data and records incorporated into the medical records that are legally binding. For example, raw audio and intermediate transcripts may be deleted after clinician verification, while finalized notes and minimal provenance metadata are retained to support auditability and dispute resolution in accordance with medical record retention requirements and data protection laws.

Beyond local data and cybersecurity governance, cross-border governance further complicate deployment, as cloud-based processing introduces multi-vendor data handling chains and expands cyberattack surfaces. Security and governance frameworks will need to specify role-based access and restrictions, data encryption, and establish breach response protocols. When sensitive patient information are processed or stored across jurisdictions, clear contractual agreements are needed between vendors and health systems to mitigate risks associated with cross-border data flows.

Model Bias and Automation Bias

LLM-based systems have been shown to reproduce if not amplify biases present in the training data sets. Underrepresented groups may be excluded or misunderstood if ambient AI scribes are not trained on diverse linguistic patterns, accents, and dialects. Unfortunately, proprietary models rarely reveal their training and validation data for bias and fairness analysis.[37] If clinicians are exhibiting automation bias, displaying excessive trust of the tool, the issue could be further compounded. This may be more likely in high-pressure and fast-paced environments. Cautious evaluation is required by adopters of the technology to avoid perpetration of these biases. Future analyses should also include a qualitative component to elicit the experiences of physicians and patients with the AI scribe based on sociodemographic features. Physician–patient communication is often of poorer quality for patients of underrepresented backgrounds, creating potential gaps in outcomes.

Transparency and Regulatory Oversight

There is ongoing uncertainty regarding how ambient AI scribes should be classified within current regulatory frameworks, particularly whether they meet the criteria of medical devices. Pure transcription

tools are less likely to influence clinical decision making and hence unlikely to be considered medical devices.[44] Tools such as ambient AI scribes which are capable of summarization and decision support, however, do have the potential to alter what and how information is communicated and influence clinical decision-making, which raises important questions about oversight, safety standards, and accountability. The National Health Services (NHS) is the first to release a guidance on the use of AI-enabled ambient scribing products in health and care settings, suggesting that Ambient AI scribes with summarization capability require regulatory scrutiny.[45] Regulatory frameworks such as the European Union's Medical Device Regulation and US Food and Drug Administration's Software as a Medical Device (SaMD) have yet to provide clear, harmonized guidance on classification and risk categorization.[46]

Patient Privacy and Consent

Significant privacy risk is introduced with passive recording of clinical encounters. Introducing a third-party vendor with access to stored encounter audio and text may cause concern by patients and clinicians about who has access to that information. As previously mentioned, the question of patient awareness that these devices are recording is not fully understood. Further, patients may not wish to have sensitive conversations recorded, such as those regarding genetic counselling, sexual and reproductive health, substance abuse, domestic violence, crime, and personal values. Studies in clinical ethics and early adopter reports suggest that automatic ambient recording may not always be appropriate in such contexts: Patients may feel inhibited in their disclosures, or intimidated to ask their provider to pause the device, if AI is recording the full conversation in the background, potentially undermining the trust and openness needed for high-quality care.[47] Protocols should define when ambient listening is appropriate, ensure patients are informed and provide consent to recording, and establish alternative processes for those who decline, ensuring their care remains timely and unaffected.[48]

In jurisdictions requiring two-party consent, the use of ambient scribes must be disclosed and agreed upon explicitly. In ambulatory settings, verbal consent that allows recording of the consultation for ambient scribe usage may be sought. However, this is often impractical in situations where patients are unable to provide informed consent, such as being unconscious or cognitively impaired.

Fairness and Inclusion

Other major ethical concerns include bias, fairness and inclusion in Ambient AI scribe outputs. LLM-based systems applied to healthcare tasks have been shown to propagate and perpetuate biases in their outputs.[37] If the underlying models were not trained on a diverse range of patient populations and dialects, the scribe may systematically misinterpret or omit content for certain groups. Scrutiny of training data may provide insights on the limits of LLM models.

Liability and Accountability

It remains unclear whether the clinician, institution, or AI developer is responsible for documentation errors when ambient scribes are used. Hence institutions must define accountability structures and ensure clinicians carefully review generated text. Additionally, vendors should design systems that

encourage review and validation of documentation and make the provenance of information transparent. [49] Audit trails and real-time monitoring will be essential.

Institutions are discouraged from creating dual documentation. It remains unclear whether the audio recordings and transcription existing within the tool represent such "shadow" records, Raising potential conflicts between versions and complicating compliance.[44]

Responsible Scaling of Ambient AI Scribes

Validation Across Settings

Despite sharing many similarities, high acuity settings are extremely diverse across organizations, varying by location, physical layout, staffing capabilities, local resources, and patient type.[50,51] These differences may impact the effectiveness and safety profile of ambient AI scribes. Implementation planning, user engagement, and post-deployment monitoring is critical to success and risk management under different settings. Testing of proprietary models across different practice settings should be conducted with the goal of ensuring fairness across different patient demographics.

Currently available, proprietary AI scribes are not easy to adopt or adapt to dynamic user workflows that vary between settings. Early user engagement to tailor interfaces, workflows, and outputs to the realities of clinical practice are necessary. Adaptive training and feedback loops between users and developers will be essential in refining these tools.

Monitoring and Governance

Monitoring in practice encompasses performance, equity, and instances of 'unsafe acceptance' - the uncorrected use of a scribe-generated element in pre-identified reduced-reliability contexts later judged incorrect, incomplete, or misleading.[14,37,39,52] Continued monitoring of performance drift and unintended adverse events is a critical step in ensuring safety of full-scale implementation, however current evaluation methods are heavily reliant on human expert evaluation are not scalable. One potential strategy to circumvent this limitation is to train LLM-based evaluators to assess correctness, relevance, faithfulness, task completion, etc. within each high-acuity speciality to augment expert review. LLM evaluators could flag notes for manual inspection when issues are identified. Such tools need to be tailored to context, undergo regular retraining on new data, and overseen by interdisciplinary governance bodies.

Addressing the Challenges

The adoption of ambient AI scribes within clinical practice introduces a broad spectrum of technical, operational, ethical, and equity-related challenges. Figure 1 summarizes current and proposed applications, challenges to implementation, ethical and regulatory concerns. Their integration thus requires comprehensive and systematic approach to ensure safety and effectiveness. From a technical perspective, issues such as ambient noise interference, errors in speaker attribution, limited multilingual support, and hardware or network instability substantially affect transcription accuracy. Operational

challenges are compounded by fragmented clinical workflows, variable documentation practices, and repeated use, necessitating site-specific adaptation which tends to counteract efforts toward standardization and scalability.

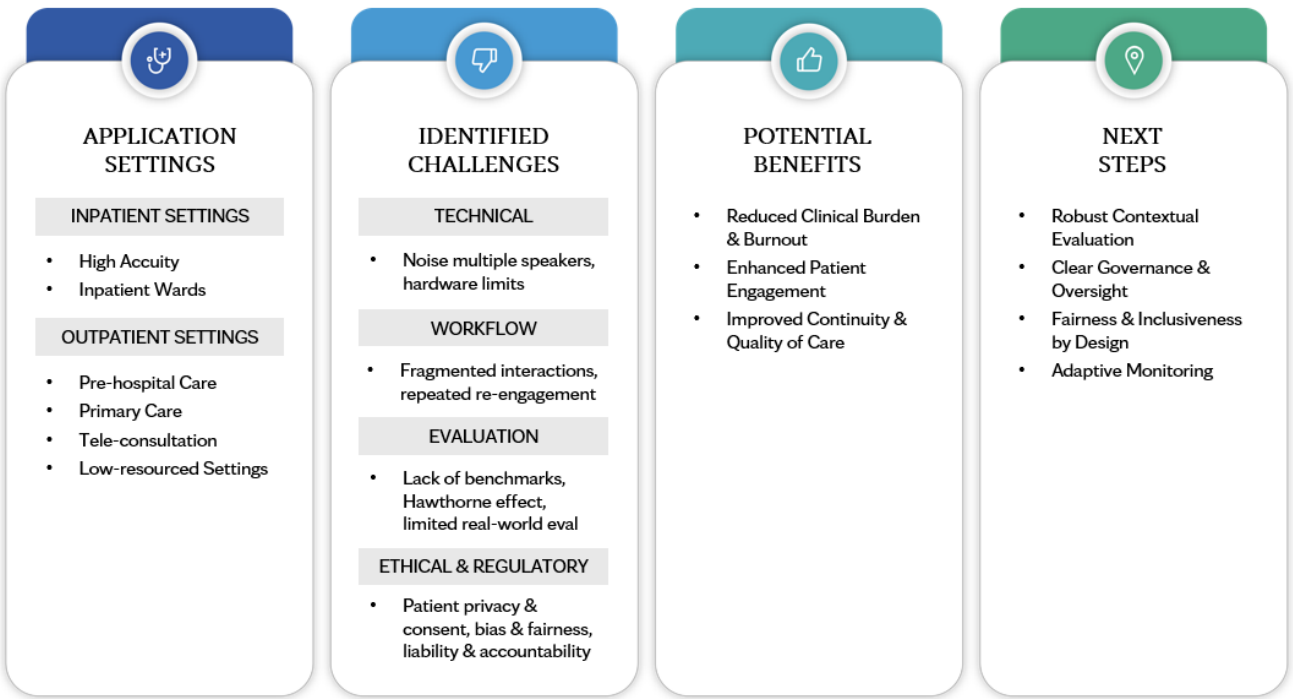

Figure 1: Application settings of ambient AI scribes, challenges to implementation, evaluation, and ethics, potential benefits, and next steps to ensure safe and trustworthy adoption. Created by Stephanie C. Bernthal, M.S.

The evaluation of these systems is further complicated by a lack of robust, validated metrics for measuring their performance in clinical use, safety, and reliability. Frameworks such as those highlighted in Figure 2 may not include the review of unnecessary verbose documentation, or the inclusion of irrelevant information from conversations. The phenomenon of "note bloat", coupled with undetected hallucinated content may result in a progressive degradation in the quality of clinical documents. Health systems should consider including qualitative metrics that captures conciseness usability of documents, beyond simply time savings.

Meanwhile the absence of formalized standards for real-time monitoring hinders the necessary continuous quality assurance – adding ethical and regulatory complexities to include ambiguity surrounding SaMD classification, legal liability concern, automation bias, and necessary protection of patient privacy. As it stands, ensuring the equitable implementation of AI scribes remains a significant issue, given that data disparities and differences in technological infrastructure pose substantial obstacles to widespread, inclusive adoption, particularly in under-resourced healthcare environments.

Given the speed of adoption, scalable evaluation and monitoring frameworks for ambient AI scribes in clinical applications is critically needed. Conventional validation approaches, such as small-scale pilot studies or retrospective reviews, are insufficient to capture the dynamic and evolving nature of these

systems. For this reason, innovative frameworks have emerged (Figure 2). One example is the SCRIBE framework that incorporates human, automated, and LLM evaluators, coupled with simulation testing to assess multiple criteria.[14] CRAFT-MD, is another scenario-based evaluation framework utilizing an AI-agent in the loop.[38] This allows model performance to be tested under controlled, clinically relevant contexts that mimic real-world variability.

FRAMEWORK	KEY FEATURES	APPLICABILITY	EXISTING GAPS
SCRIBE	Human + LLM evaluator mix, simulations	High for multiple settings	Needs more LMIC evaluation
MEDHELM	Benchmark datasets for note summary	High for research	Limited real-world integration
CRAFT-MD	Scenario-based AI-in-the-loop testing	Strong for acute care	Early stage, needs wider adoption
RE-AIM	Implementation outcomes (reach, adoption, etc.)	Strong for deployment evaluation	Doesn't directly assess linguistic accuracy

Figure 2: Evaluation and monitoring frameworks for ambient AI scribes, their key features, applicability in practice, and existing gaps. Created by Stephanie C. Bernthal, M.S.

Beyond validating performance, the systematic use of implementation frameworks supports the effective integration of ambient scribes across diverse clinical environments. Implementation science toolkits, such as the Reach, Efficacy, Adoption, Implementation, and Maintenance (RE-AIM) developed by Glasgow, Vogt, and Boles and the Implementation Outcomes Framework from Proctor and colleagues are some examples that can be used .[53,54] Regardless of how ambient scribes are evaluated or implemented, understanding the context in which these tools will be used, who are the end-users, and the nuances of each environment will limit the potential for patient harm, clinician burden, ethical issues, and liability concerns.

Finally, ambient scribing may have important implications for medical education. Core components of clinical reasoning, synthesis and reflection may be reduced through partial automation brought about by ambient scribing. For trainees and medical residents, increasing reliance on AI-generated drafting could limit opportunities to develop skills in clinical summarization and assessment if notes are accepted uncritically. To mitigate this risk, medical education may encompass tiered approaches whereby trainees are required to independently generate clinical documents without the aid of AI tools (AI-free phase) before they are allowed to use AI aids.

Conclusion

Ambient AI scribes hold promise in transforming clinical documentation and relieving cognitive and administrative burden as an assistive tool to clinicians. Yet, their success under different care setting hinges not just on technical sophistication but on ethical design, inclusive evaluation, and governance clarity. To address these challenges, stakeholders must adopt a systems-level approach grounded in

contextual validation, inclusive design, and robust governance. With knowledge of current limitations and careful integration, ambient AI scribes can evolve from passive transcription tools into trusted partners in the delivery of complex care across all care settings.

Author contributions

J.W.O. and A.T. contributed equally. Initial conceptualization: J.W.O., A.T., Z.L., J.Y.E.T., and J.C.L.O. Drafting of the first manuscript: J.W.O., A.T., Z.L., L.M.R., J.D.O., J.Y.E.T., S.M.O., and J.C.L.O. Critical revision of the manuscript: J.W.O., A.T., Z.L., L.M.R., J.D.O., J.Y.E.T., S.M.O., R.F.R.S., Y.K., J.C.K.L., J.C.L.O., and N.L. All authors have read, reviewed, and approved of the final manuscript.

Competing interests

Author J.C.L.O is an associate editor of npj digital medicine. N.L. is an editorial board member of npj digital medicine. J.C.L.O and N.L. were not involved in the journal's review of, or decisions related to, this manuscript.

Acknowledgements

We sincerely thank Stephanie C. Bernthal, M.S. for designing the figures in this manuscript. This work was supported by the Duke-NUS Signature Research Programme funded by the Ministry of Health, Singapore. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Health. This work is supported in part by generous funds from Dalio Philanthropies. The content expressed in this publication are those of the authors and does not necessarily represent the official views of Dalio Philanthropies.

References

1. Nahar, J. K. & Kachnowski, S. Current and Potential Applications of Ambient Artificial Intelligence. *Mayo Clin. Proc. Digit. Health* **1**, 241–246 (2023).
2. Arndt, B. G. *et al.* Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. *Ann. Fam. Med.* **15**, 419–426 (2017).
3. Rotenstein, L. *et al.* Virtual Scribes and Physician Time Spent on Electronic Health Records. *JAMA Netw. Open* **7**, e2413140 (2024).
4. Saag, H. S., Shah, K., Jones, S. A., Testa, P. A. & Horwitz, L. I. Pajama Time: Working After Work in the Electronic Health Record. *J. Gen. Intern. Med.* **34**, 1695–1696 (2019).
5. Budd, J. Burnout Related to Electronic Health Record Use in Primary Care. *J. Prim. Care Community Health* **14**, 21501319231166921 (2023).
6. Shah, S. J. *et al.* Physician Perspectives on Ambient AI Scribes. *JAMA Netw. Open* **8**, e251904 (2025).
7. Duggan, M. J. *et al.* Clinician Experiences With Ambient Scribe Technology to Assist With Documentation Burden and Efficiency. *JAMA Netw. Open* **8**, e2460637 (2025).
8. Tierney, A. A. *et al.* Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation. *NEJM Catal.* **5**, CAT.23.0404 (2024).
9. Misurac, J., Knake, L. A. & Blum, J. M. The Effect of Ambient Artificial Intelligence Notes on Provider Burnout. *Appl. Clin. Inform.* **16**, 252–258 (2025).
10. Stults, C. D. *et al.* Evaluation of an Ambient Artificial Intelligence Documentation Platform for Clinicians. *JAMA Netw. Open* **8**, e258614 (2025).

11. Tierney, A. A. *et al.* Ambient Artificial Intelligence Scribes: Learnings after 1 Year and over 2.5 Million Uses. *NEJM Catal.* **6**, CAT.25.0040 (2025).
12. Haberle, T. *et al.* The impact of nuance DAX ambient listening AI documentation: a cohort study. *J. Am. Med. Inform. Assoc. JAMIA* **31**, 975–979 (2024).
13. Ha, E. *et al.* Evaluating the Usability, Technical Performance, and Accuracy of Artificial Intelligence Scribes for Primary Care: Competitive Analysis. *JMIR Hum. Factors* **12**, e71434 (2025).
14. Wang, H. *et al.* An evaluation framework for ambient digital scribing tools in clinical applications. *NPJ Digit. Med.* **8**, 358 (2025).
15. Shah, S. J. *et al.* Ambient artificial intelligence scribes: physician burnout and perspectives on usability and documentation burden. *J. Am. Med. Inform. Assoc.* **32**, 375–380 (2025).
16. Street, R. L. *et al.* Provider interaction with the electronic health record: The effects on patient-centered communication in medical encounters. *Patient Educ. Couns.* **96**, 315–319 (2014).
17. Gerke, S., Simon, D. A. & Roman, B. R. Liability Risks of Ambient Clinical Workflows With Artificial Intelligence for Clinicians, Hospitals, and Manufacturers. *JCO Oncol. Pract.* **0**, OP-24-01060 (2025).
18. Leung, T. I., Coristine, A. J. & Benis, A. AI Scribes in Health Care: Balancing Transformative Potential With Responsible Integration. *JMIR Med. Inform.* **13**, e80898 (2025).
19. Lawson, E. The Global Primary Care Crisis. *Br. J. Gen. Pract.* **73**, 3.
20. Duong, D. & Vogel, L. Ontario, Quebec and Alberta lead record family medicine residency vacancies. *CMAJ* **195**, E557–E558 (2023).
21. Mishra, P., Kiang, J. C. & Grant, R. W. Association of Medical Scribes in Primary Care With Physician Workflow and Patient Experience. *JAMA Intern. Med.* **178**, 1467–1472 (2018).
22. Ma, S. P. *et al.* Ambient artificial intelligence scribes: utilization and impact on documentation time. *J. Am. Med. Inform. Assoc. JAMIA* **32**, 381–385 (2025).
23. Sanderson, A. L. & Burns, J. P. Clinical Documentation for Intensivists: The Impact of Diagnosis Documentation. *Crit. Care Med.* **48**, 579–587 (2020).
24. Gkiala, A. Assessing the Correct Documentation of Time and Physician Information on Medical Records in the Emergency Department of Queen's Hospital: An Audit and Re-audit. *Cureus* **14**, e33000 (2022).
25. Blome, A., Yu, D., Lu, X. & Schreyer, K. E. Pitfalls of Extensive Documentation in the Emergency Department. *Ochsner J.* **20**, 299–302 (2020).
26. Davidson, S. J., Zwemer, F. L., Nathanson, L. A., Sable, K. N. & Khan, A. N. G. A. Where's the beef? The promise and the reality of clinical documentation. *Acad. Emerg. Med. Off. J. Soc. Acad. Emerg. Med.* **11**, 1127–1134 (2004).
27. Nates, J. L. *et al.* ICU Admission, Discharge, and Triage Guidelines: A Framework to Enhance Clinical Operations, Development of Institutional Policies, and Further Research. *Crit. Care Med.* **44**, 1553–1602 (2016).
28. ScribeAmerica to Attend the ACEP Scientific Assembly 2023. *ScribeAmerica* https://www.scribeamerica.com/press_release/scribeamerica-to-attend-the-acep-scientific-assembly-2023/.
29. Dai, W. *et al.* Developing ICU Clinical Behavioral Atlas Using Ambient Intelligence and Computer Vision. *NEJM AI* **2**, AIoa2400590 (2025).
30. Martin-Gill, C. *et al.* 2022 Systematic Review of Evidence-Based Guidelines for Prehospital Care. *Prehosp. Emerg. Care* **27**, 131–143 (2023).
31. Sasseville, M. *et al.* The Impact of AI Scribes on Streamlining Clinical Documentation: A Systematic Review. *Healthcare* **13**, 1447 (2025).
32. Bongurala, A. R., Save, D., Virmani, A. & Kashyap, R. Transforming Health Care With Artificial Intelligence: Redefining Medical Documentation. *Mayo Clin. Proc. Digit. Health* **2**, 342–347 (2024).
33. Health workforce. https://www.who.int/health-topics/health-workforce.

34. Rule, A. *et al.* Comparing Scribed and Non-scribed Outpatient Progress Notes. *AMIA. Annu. Symp. Proc.* **2021**, 1059–1068 (2022).
35. Reuel, A. *et al.* BetterBench: assessing AI benchmarks, uncovering issues, and establishing best practices. in *Proceedings of the 38th International Conference on Neural Information Processing Systems* vol. 37 21763–21813 (Curran Associates Inc., Red Hook, NY, USA, 2025).
36. Bracken, A. *et al.* Artificial Intelligence (AI) - Powered Documentation Systems in Healthcare: A Systematic Review. *J. Med. Syst.* **49**, 28 (2025).
37. Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V. & Daneshjou, R. Large language models propagate race-based medicine. *Npj Digit. Med.* **6**, 1–4 (2023).
38. Johri, S. *et al.* An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat. Med.* **31**, 77–86 (2025).
39. Bedi, S. *et al.* MedHELM: Holistic Evaluation of Large Language Models for Medical Tasks. Preprint at https://doi.org/10.48550/arXiv.2505.23802 (2025).
40. McCradden, M. D. *et al.* CANAIRI: the Collaboration for Translational Artificial Intelligence Trials in healthcare. *Nat. Med.* **31**, 9–11 (2025).
41. Kwong, J. C. C. *et al.* The silent trial - the bridge between bench-to-bedside clinical AI applications. *Front. Digit. Health* **4**, (2022).
42. Ma, R. *et al.* Multimodal machine learning enables AI chatbot to diagnose ophthalmic diseases and provide high-quality medical responses. *Npj Digit. Med.* **8**, 64 (2025).
43. Schiff, G. D. AI-Driven Clinical Documentation — Driving Out the Chitchat? *N. Engl. J. Med.* **392**, 1877–1879 (2025).
44. Cohen, I. G., Ritzman, J. & Cahill, R. F. Ambient Listening—Legal and Ethical Issues. *JAMA Netw. Open* **8**, e2460642 (2025).
45. England, N. H. S. NHS England » Guidance on the use of AI-enabled ambient scribing products in health and care settings. https://www.england.nhs.uk/long-read/guidance-on-the-use-of-ai-enabled-ambient-scribing-products-in-health-and-care-settings/.
46. Aboy, M., Minssen, T. & Vayena, E. Navigating the EU AI Act: implications for regulated digital medical products. *NPJ Digit. Med.* **7**, 237 (2024).
47. Robertson, C. *et al.* Diverse patients' attitudes towards Artificial Intelligence (AI) in diagnosis. *PLOS Digit. Health* **2**, e0000237 (2023).
48. Svarovsky, T. Having Difficult Conversations: The Advanced Practitioner's Role. *J. Adv. Pract. Oncol.* **4**, 47–52 (2013).
49. Felzmann, H., Fosch-Villaronga, E., Lutz, C. & Tamò-Larrieux, A. Towards Transparency by Design for Artificial Intelligence. *Sci. Eng. Ethics* **26**, 3333–3361 (2020).
50. Joseph, A. *et al.* Minor flow disruptions, traffic-related factors and their effect on major flow disruptions in the operating room. *BMJ Qual. Saf.* **28**, 276–283 (2019).
51. Malhotra, S., Jordan, D. & Patel, V. L. Workflow Modeling in Critical Care: Piecing your own Puzzle. *AMIA. Annu. Symp. Proc.* **2005**, 480–484 (2005).
52. Lyell, D. & Coiera, E. Automation bias and verification complexity: a systematic review. *J. Am. Med. Inform. Assoc.* **24**, 423–431 (2017).
53. Proctor, E. *et al.* Outcomes for Implementation Research: Conceptual Distinctions, Measurement Challenges, and Research Agenda. *Adm. Policy Ment. Health* **38**, 65–76 (2011).
54. Glasgow, R. E., Vogt, T. M. & Boles, S. M. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *Am. J. Public Health* **89**, 1322–1327 (1999).