BJU Int 2025 doi:10.1111/bju.16784

BJUI
BJU International

Research Letter

A pilot study using ambient artificial intelligence scribes in clinical documentation in a urology outpatient clinic

The increasing complexity of electronic health records (EHRs) has significantly contributed to physician workload, with clinical documentation now cited as a major contributor to burnout, job dissatisfaction, and reduced patient engagement [1]. Physicians even report spending less than 30% of their work hours on direct patient care due to administrative tasks [2]. The use of medical scribes has shown benefits in reducing this burden [3]; however, manual scribes require staffing and infrastructure. Recent advances in large language models (LLMs) and real-time transcription have led to the development of ambient artificial intelligence (AI) scribes – digital tools that autonomously capture clinical encounters and generate structured notes [4].

To assess the feasibility and quality of such tools, we piloted an ambient AI tool that was developed using Microsoft's Azure OpenAI platform. Speech recognition capabilities were provided via Azure's Speech Service, which returned transcripts in real time, with speaker diarisation to distinguish among participants. Personally identifiable information (PII) entities, such as national identification numbers, telephone numbers and email addresses, were spoken into the transcript, identified with Azure Language Service, and masked to ensure data privacy. De-identified transcripts were sent to Azure OpenAI via API Management (APIM), hosted in the national health technology agency's secure GPT environment, ensuring that all interactions complied with national healthcare data privacy regulations.

The GPT-4o model was utilised to reformat and process transcripts into clinical consultation notes. A prompt was developed (File S1) that instructed GPT-4o to assume the role of a physician assistant given the task to transcribe and summarise a clinical consultation transcript for entry into the patient's EHRs. Safety elements included specific instructions to include only information provided in the transcript.

In this pilot study, we compared AI-generated notes against those authored by clinicians during five anonymised real-world patient encounters in the outpatient clinic of a urology department in a large tertiary centre. Each consultation was transcribed by the AI scribe and separately documented by the attending clinician. The resulting notes were then evaluated by eight senior urologists (11 to >20 years of experience), who were blinded to the note authorship. Evaluations used a modified version of the Physician Documentation Quality Instrument (PDQI-9), with domains including accuracy, thoroughness, usefulness, organisation, comprehensibility, succinctness, synthesis, and internal consistency. Scores ranged from 1 (poor) to 5 (excellent).

A total of 80 evaluations were received. The mean modified PDQI-9 score was 33.25 (SD 4.058). AI-generated notes had comparable mean scores across most domains (Table 1). Specifically, they performed similarly to human-written notes in accuracy (4.22 vs 4.20), thoroughness (4.22 vs 4.00), usefulness (4.30 vs 4.15), organisation (4.35 vs 4.13), comprehensibility (4.30 vs 4.25), and synthesis (4.08 vs 3.70), although none of these differences reached statistical significance ($P > 0.05$). Clinician-authored notes had marginally higher scores in succinctness (4.22 vs 4.00) and internal consistency (4.22 vs 4.15). Overall, the total PDQI-9 scores for AI and human documentation were comparable (33.63 vs 32.88; $P = 0.412$). AI-generated notes were preferred in 55% of evaluator ratings, suggesting non-inferiority in documentation quality.

These findings support the growing amount of literature on the potential for LLMs to streamline clinical workflows and alleviate the clerical burden on healthcare providers. By automating documentation, ambient AI scribes may help restore focus to patient care, reduce physician fatigue, and potentially improve clinic throughput [5]. For high-volume specialties such as urology, such tools could help streamline operations, assisting clinicians in managing their caseloads more effectively.

Despite these encouraging results, real-world adoption of AI scribes comes with certain caveats. Current models still require human oversight, particularly in complex cases where nuanced clinical decision-making, contextual understanding, and specialised terminology are crucial. Over-reliance on AI-generated documentation may lead to errors, misinterpretations, or omission of clinically relevant details. This will require manual corrections, negating the perceived efficiency gains. Additionally, the effectiveness of AI scribes depends on the quality and diversity of training data. At present, the majority of models are trained on general medical conversations rather than specialty-specific dialogues, potentially leading to biases or inaccurate transcription of specialty-related terminology, procedural nuances, and patient-specific concerns.

As ambient AI scribes continuously process confidential patient information and conversations, patient privacy is an area of concern. Safeguards must be in place to ensure compliance with regional privacy laws to prevent unauthorised access or breaches of sensitive patient

Table 1 Scores for AI- and human-written consultation notes by attribute.

Attribute	AI scribe		Clinician		*P*
	Mean (SD)	Range (min.–max.)	Mean (SD)	Range (min.–max.)	
Accurate	4.22 (0.620)	2.0 (3.0–5.0)	4.20 (0.464)	2.0 (3.0–5.0)	0.839
Thorough	4.22 (0.768)	2.0 (3.0–5.0)	4.00 (0.555)	2.0 (3.0–5.0)	0.137
Useful	4.30 (0.648)	2.0 (3.0–5.0)	4.15 (0.483)	2.0 (3.0–5.0)	0.244
Organised	4.35 (0.580)	2.0 (3.0–5.0)	4.13 (0.516)	2.0 (3.0–5.0)	0.070
Comprehensible	4.30 (0.648)	2.0 (3.0–5.0)	4.25 (0.439)	1.0 (4.0–5.0)	0.687
Succinct	4.00 (0.961)	4.0 (1.0–5.0)	4.22 (0.577)	2.0 (3.0–5.0)	0.208
Synthesised	4.08 (0.797)	3.0 (2.0–5.0)	3.70 (1.018)	4.0 (1.0–5.0)	0.070
Internally consistent	4.15 (0.770)	3.0 (2.0–5.0)	4.22 (0.423)	1.0 (4.0–5.0)	0.591
Total	33.63 (4.759)	16.0 (24.0–40.0)	32.88 (3.228)	13.0 (27.0–40.0)	0.412

AI, artificial intelligence.

information [6]. The LLM tool should be designed to only capture data pertinent to clinical documentation and exclude irrelevant or personal patient information. Robust data security protocols are also needed to prevent data loss to adversarial attacks, particularly when a larger number of patients are involved. Patients must be made aware of AI's role in their documentation and consent to the use of such ambient AI tools [7].

Additionally, AI algorithms and LLMs are constantly evolving. The use of retrieval-augmented generative techniques has demonstrated promise as a convenient way to provide LLMs with domain-specific knowledge [8], which is in contrast to the resource-intensive methods of fine-tuning or pre-training models from the ground up. These provide further avenues by which ambient AI scribes may be further refined.

This pilot study had several limitations. First, the small sample size of only five patient encounters hampered statistical evaluation and was insufficiently powered to detect statistically significant differences in evaluation scores as measured on the modified PDQI-9. The selection of five patient encounters was based on the pilot nature of the study. While limited, this sample was sufficient for a preliminary proof-of-concept and allowed detailed qualitative review by experienced urologists. Nevertheless, larger ongoing studies will be necessary to validate these findings. Second, the study was conducted in the highly specific setting of urology outpatient clinics, and in a single language, limiting its generalisability to other specialties and care settings. Third, we did not measure other outcomes such as patient satisfaction or physician acceptance and usability of these tools, or their long-term impact on physician burnout. Lastly, the lack of formal training on AI scribes for the physicians involved in this analysis may have influenced their assessment of AI-generated notes, particularly in areas where AI structuring differs from traditional clinician-authored documentation.

Moving forward, one of the most significant promises of ambient AI scribes is their ability to convert free-text clinical conversations into structured EHR data. However, this process presents technical and clinical challenges that must be critically examined, including but not limited to the complexity of unstructured clinical notes, limitations related to contextual understanding unless the AI model is specifically trained in specialty-specific lexicons, or even errors in coding diagnoses, procedures or medication reconciliation, which can lead to billing discrepancies, misinterpretation of care plans, or patient safety concerns. In addition, current evaluation methods often involve urologists reviewing AI-generated documentation *post hoc*, rather than assessing its performance in real-time patient encounters. Future studies will need to include real-time validation in order to better determine the true clinical impact, reliability, and adoption feasibility of AI-powered documentation tools in urology.

In summary, this pilot study provides early evidence that ambient AI scribes can generate high-quality clinical notes that are comparable to those written by human clinicians. With appropriate oversight and iterative refinement, such tools hold promise for reducing documentation burden and enhancing clinical efficiency in outpatient settings.

Disclosure of Interests

The authors have no conflicts of interest to declare. No funding was received for this work.

Informed Consent

An informed consent waiver was granted by our institutional review board (CIRB 2024/3377).

Julene Hui Wun Ong[1], Joshua Yi Min Tung[1,2], Gerald Gui Ren Sng[2], Daniel Yan Zheng Lim[2], Xinyan Yang[1], Iffat Bin Mohamad Rafi[3], Michelle Zhen Yuan Loke[4], Ajeet Kumar[4], Jonathan Yue En Tan[3], Eileen Yi Lee Lew[4], Mark Kok Hong Heng[4] and Henry Sun Sien Ho[1]

[1]Department of Urology, [2]Data Science and Artificial Intelligence Laboratory, [3]Artificial Intelligence & Analytics Office, and [4]Office of Insights & Analytics, Singapore General Hospital, Singapore, Singapore

References

1 **Adler-Milstein J**, **Zhao W**, **Willard-Grace R**, **Knox M**, **Grumbach K**. Electronic health records and burnout: time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. *J Am Med Inform Assoc* 2020; 27: 531–8

2 **Sinsky C**, **Colligan L**, **Li L et al.** Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Ann Intern Med* 2016; 165: 753–60

3 **Mishra P**, **Kiang JC**, **Grant RW**. Association of Medical Scribes in primary care with physician workflow and patient experience. *JAMA Intern Med* 2018; 178: 1467–72

4 **Omiye JA**, **Gui H**, **Rezaei SJ**, **Zou J**, **Daneshjou R**. Large language models in medicine: the potentials and pitfalls: a narrative review. *Ann Intern Med* 2024; 177: 210–20

5 **Yang X**, **Chen A**, **PourNejatian N et al.** A large language model for electronic health records. *NPJ Digit Med* 2022; 5: 194

6 **Labkoff S**, **Oladimeji B**, **Kannry J et al.** Toward a responsible future: recommendations for AI-enabled clinical decision support. *J Am Med Inform Assoc* 2024; 31: 2730–9

7 **Ning Y**, **Teixayavong S**, **Shang Y et al.** Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist. *Lancet Digit Health* 2024; 6: e848–56

8 **Sahiner B**, **Chen W**, **Samala RK**, **Petrick N**. Data drift in medical machine learning: implications and potential remedies. *Br J Radiol* 2023; 96: 20220878

Correspondence: Julene Ong Hui Wun, Urology Centre, Singapore General Hospital, 16 College Road, Block 4 Level 1, Singapore 169854, Singapore.
E-mail: julene.ong@mohh.com.sg

Abbreviations: AI, artificial intelligence; EHR, electronic health record; LLM, large language model; PDQI-9, Physician Documentation Quality Instrument.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

File S1. Ambient AI scribe prompt.