Research Letter

Accuracy and Safety of AI-Enabled Scribe Technology: Instrument Validation Study

Joshua Biro[1], PhD; Jessica L Handley[1], MA; Nathan K Cobb[2,3], MD; Varsha Kottamasu[2], MHA; Jeffrey Collins[2], MHS; Seth Krevat[1,3], MD; Raj M Ratwani[1,3], PhD

[1]National Center for Human Factors in Healthcare, MedStar Health Research Institute, Washington, DC, United States

[2]MedStar Health Institute for Innovation, Washington, DC, United States

[3]Georgetown University School of Medicine, Washington, DC, United States

Corresponding Author:
Joshua Biro, PhD
National Center for Human Factors in Healthcare
MedStar Health Research Institute
3007 Tilden St NW
Washington, DC, 20008
United States
Phone: 1 3015423073
Email: joshua.m.biro@medstar.net

Abstract

Artificial intelligence–enabled ambient digital scribes may have many potential benefits, yet results from our study indicate that there are errors that must be evaluated to mitigate safety risks.

(J Med Internet Res 2025;27:e64993) doi: 10.2196/64993

KEYWORDS

artificial intelligence; AI; patient safety; ambient digital scribe; AI-enabled scribe technology; AI scribe technology; scribe technology; accuracy; safety; ambient scribe; digital scribe; patient-clinician; patient-clinician communication; doctor-patient relationship; doctor-patient communication; patient engagement; patient safety; dialogue script; scribe

Introduction

Generative artificial intelligence (AI)–enabled ambient digital scribe (ADS) technology uses the patient-clinician conversation to generate clinical documentation; it has the potential to improve patient engagement and reduce clinician burden [1,2]. These technologies are becoming more prevalent, especially in ambulatory care settings, yet there is little known about documentation accuracy and the types of errors that may stem from ADS use [3]. Error-prone ADS technology may have serious patient safety consequences [4]. We evaluated 2 popular commercially available ADS products in a simulated setting to systematically identify the frequency and pattern of documentation errors.

Methods

Ethical Considerations

This study was approved by the MedStar Health Institutional Review Board (00007789) to cover secondary analysis of existing patient data without additional consent. All data were deidentified. Participants did not receive any form of compensation.

Recording and Simulation

Recordings of 11 real outpatient encounters from a range of service lines (otolaryngology, cardiology, rheumatology, family medicine, pediatrics, endocrinology, internal medicine, gastroenterology, oncology, and urgent care) were transcribed by automated software and then deidentified and edited by a senior physician (NKC) for clarity to create 11 unique dialogue scripts. The dialogue scripts were used to evaluate 2 commercial ADS products. For each script, a researcher (JB or VK) simulating the patient and a medical resident simulating the physician read from the script while the ADS products were in use. Each script was read by 2 different residents per ADS product, yielding 22 draft notes per product and 44 draft notes in total across products. The residents reviewed the draft notes to identify errors. Each error was independently categorized by 2 reviewers (JB or JLH) as either an omission, addition, wrong output, or irrelevant or misplaced text, as defined in Table 1. Disagreements were discussed to reach consensus.

Results

There were 127 errors (mean 2.9, SD 2.7 errors per draft note) in 31 of 44 (70%) draft notes. ADS product A resulted in 66 errors (mean 3, SD 2.7 per draft note) and product B resulted in 61 errors (mean 2.8, SD 2.7 per draft note). Error frequency by error type and product is detailed in Table 1, with omission errors being the most frequent across products. Error types significantly differed between the 2 ADS products (Fisher exact test: *P*=.002).

Table 1. Frequency counts, percentages, definitions, and examples of ambient digital scribe (ADS) error types.

Error type	Errors by ADS product, n (%)		Definition	Example
	Product A (n=66)	Product B (n=61)		
Omission	55 (83)	33 (54)	Model leaves out key information from its response	"[N]o laterality mentioned in ears in physical exam section"
Addition	3 (4)	7 (11)	Model adds inappropriate or irrelevant information	"Patient doesn't refer to any flare ups in awhile, but the note shared that patient was using X medication to help with flare ups in HPI"
Wrong output	4 (6)	6 (10)	Model provides an incorrect response	"[A]ssociated the wrong test with the contrast"
Irrelevant or misplaced text	4 (6)	15 (25)	Model output is technically correct but not appropriate in clinical context	"[C]aptured all the supplemental information (asthma, mammogram, etc.) and harped on the steroid injections which doesn't matter"

Discussion

While ADS technologies may have potential benefits, there are frequent errors in the generated note. Across both products, errors of omission were the most common; this error type may be the most difficult for clinicians to identify since the identification process requires memory recall of details from the patient encounter. If clinicians review their documentation after several patient encounters, recalling omitted details may be challenging. It may be easier to identify errors such as additions and wrong outputs since this relies on recognition of an issue in the text being presented to the clinician. Notably, there was a different pattern of errors between the two products.

There are limitations to this study. The ADS technologies were evaluated against a limited number of patient cases in a controlled environment that did not fully represent clinical workflows. In addition, the cases were read by 2 researchers acting as patients, and it is likely that both clinicians and patients would have more variability in language, tone, volume, and many other characteristics that could impact ADS accuracy.

It is imperative that ADS technologies be evaluated in a realistic clinical setting (either in situ or in a representative simulation) to determine the frequency and types of errors so that appropriate risk mitigation and safety plans can be developed. Developing methods to capture AI-related safety issues was a component of President Joe Biden's "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" [5], and robust processes for AI safety are needed. In the absence of a standardized evaluation framework, health care facilities currently bear the burden of testing and reporting these results in the United States. It is to be noted that, effective August 2024, the European Union Artificial Intelligence Act legally requires developers of AI-based systems to evaluate the safety of their products [6]. While a step in the right direction, the underlying vendor algorithms are often proprietary, opaque, and the subject of continuous innovation; thus, there is still a need for independent ongoing testing to confirm vendor claims of safety. Future work should develop a robust, standardized, and repeatable ADS evaluation framework to facilitate efficient knowledge sharing in this fast-paced, decentralized system.

Acknowledgments

We would like to acknowledge Dr Sahithi Reddy and Dr James Mickler for their assistance in executing this work.

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

JB contributed to study design, data acquisition, data analysis, interpretation, drafting, and reviewing the manuscript. JLH contributed to study design, interpretation, drafting, and critically reviewing the manuscript. NKC contributed to study conception, design, data acquisition, interpretation, drafting, and critically reviewing the manuscript. JC contributed to study conception, data acquisition, and critically reviewing the manuscript. SK contributed to study design, data analysis, interpretation, and critically reviewing the manuscript. VK contributed to data acquisition, interpretation, and critically reviewing the manuscript. RMR contributed to study conception, study design, interpretation, and critically reviewing the manuscript.

Conflicts of Interest

None declared.

References

1. Tierney AA, Gayre G, Hoberman B, Mattern B, Ballesca M, Kipnis P, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. NEJM Catalyst. Feb 21, 2024;5(3):692-694. [doi: 10.1056/CAT.23.0404]
2. van Buchem MM, Kant IMJ, King L, Kazmaier J, Steyerberg EW, Bauer MP. Impact of a digital scribe system on clinical documentation time and quality: usability study. JMIR AI. Sep 23, 2024;3:e60020. [FREE Full text] [doi: 10.2196/60020] [Medline: 39312397]
3. Seth P, Carretas R, Rudzicz F. The utility and implications of ambient scribes in primary care. JMIR AI. Oct 04, 2024;3:e57673. [FREE Full text] [doi: 10.2196/57673] [Medline: 39365655]
4. van Buchem MM, Boosman H, Bauer MP, Kant IMJ, Cammel SA, Steyerberg EW. The digital scribe in clinical practice: a scoping review and research agenda. NPJ Digit Med. Mar 26, 2021;4(1):57. [FREE Full text] [doi: 10.1038/s41746-021-00432-5] [Medline: 33772070]
5. Biden JR. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. The White House. 2023. URL: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ [accessed 2025-01-07]
6. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828. 2024/1689, 32024R1689. European Union. Jun 13, 2024. URL: https://artificialintelligenceact.eu/ai-act-explorer/ [accessed 2024-09-27]

Abbreviations

ADS: ambient digital scribe
AI: artificial intelligence

Edited by A Coristine; submitted 01.08.24; peer-reviewed by D Sharma, F Warg; comments to author 24.09.24; revised version received 03.10.24; accepted 09.12.24; published 27.01.25

Please cite as:
Biro J, Handley JL, Cobb NK, Kottamasu V, Collins J, Krevat S, Ratwani RM
Accuracy and Safety of AI-Enabled Scribe Technology: Instrument Validation Study
J Med Internet Res 2025;27:e64993
URL: https://www.jmir.org/2025/1/e64993
doi: 10.2196/64993
PMID: