Title

Advancing Toward Clinical Deployment of AI-Generated Discharge Summaries—Beyond the Bench

Journal

JAMA Network Open, 8(8)

ISSN

2574-3805

Authors

Subramanian, Charumathi Raghu
Rosner, Benjamin I

Publication Date

2025-08-01

DOI

10.1001/jamanetworkopen.2025.26350

JAMA Network Open

Invited Commentary | Health Informatics

Advancing Toward Clinical Deployment of AI-Generated Discharge Summaries—Beyond the Bench

Charumathi Raghu Subramanian, MD; Benjamin I. Rosner, MD, PhD

+ Related article

Author affiliations and article information are listed at the end of this article.

The hospital discharge summary is a crucial document in a patient's safe transition of care to the community, serving as the primary communication tool for a diverse array of individuals, including primary care physicians, skilled nursing facility clinicians, consulting physicians, and increasingly, patients themselves. However, generating the hospital course (HC) (ie, the narrative portion of the discharge summary) can be time consuming and detract from face-to-face clinical care. With the rapid advancement of large language models (LLMs) in health care, there is growing interest in their potential to draft HCs for physician review and editing. While artificial intelligence (AI) ambient scribing technology is witnessing widespread adoption in the outpatient setting, summarizing a full set of hospital encounter notes is a substantially more complex task.

To date, few studies have evaluated the quality of LLM-drafted HCs using real patient data rather than synthetic data or well-curated vignettes.[1] However, there has been a recent shift toward assessing the quality of LLM-generated HCs with real patient hospital data.[2,3] In *JAMA Network Open*, Small et al[4] have taken the next, often overlooked step in advancing LLM-drafted HCs toward use in real clinical care, not only by having LLMs draft the HCs, but by simulating the way in which these drafted HCs would enter the clinician workflow for editing and sign-off. Their study involved an electronic health record (EHR)-embedded LLM HC summarization tool developed by an EHR vendor.[5] In what the authors describe as an operational prepilot program, they randomly selected 100 general internal medicine hospital encounters from December 2023, with lengths of stay between 4 and 8 days, using the tool to generate HCs. Ten resident physicians were then tasked, in a blinded manner, with editing both LLM- and clinician-generated HCs within certain time constraints to enhance them according to a 4Cs quality standard (complete, concise, cohesive, and confabulation-free) developed by the research team. Subsequently, hospitalist attending physicians, also blinded to the authorship, evaluated the edited HCs. The LLM-generated and clinician-generated HCs were then compared head-to-head.

The findings from Small et al[4] showed that LLM HCs required less editing compared with clinician HCs (32% of characters vs 45%) and exhibited a smaller degree of semantic change following editing by residents (2% vs 5%). While the overall quality score did not differ significantly between the LLM and clinician HCs, the LLM HCs were rated as more complete, while also containing more confabulations than the clinician HCs—a trend observed in other recent studies examining LLM-generated discharge summaries.[2] However, it's also important to interpret these findings in the context of the time constraints imposed on the residents; 10 minutes for review of the EHR, and 3 minutes for editing. Some degree of time constraint in the study design is appropriate and essential to emulate the clinical conditions that clinicians will face reviewing and editing LLM HCs. While the actual amount of time that clinicians ought to spend in these activities remains to be determined, it is possible the 10-minute review window, coupled with the review of unfamiliar patients, may have limited residents' capacity to detect subtle errors.

Any study on LLMs ultimately raises the question of whether they are ready for use in real health care settings. By mimicking a clinical, time-constrained scenario where physicians edit a running summary for HCs, Small et al[4] make substantial headway from a clinician workflow perspective in answering this question. The authors also worked with the EHR vendor to modify the tool so that statements within LLM HC drafts contained linked citations to their sources within the clinical notes,

a key feature that would enable clinicians under the demands of real clinical practice to more readily find and verify them.

One of the major challenges in piloting LLM tools in health care is the evaluation process. To date, there has been no standard evaluation framework for AI-generated summaries,[6] placing the responsibility on health systems to develop and validate rubrics themselves. In this context, the authors offer 2 significant contributions to the health care community: (1) a 4Cs HC quality standard and (2) a method for quantifying the degree of human editing required for both LLM- and clinician-generated HCs. More recently, the Provider Documentation Summarization Quality Instrument (PDSQI-9),[7] was developed as a validated rubric for LLM summarizations that captures several other important dimensions including citability and lack of stigmatizing language that might also prove to be valuable in assessing LLM summarization performance moving forward. Ultimately, any broadly applicable quality and safety rubric will need to find its way into use in a manner that can scale beyond the limitation of human reviewers.

As these summarization tools begin to be deployed in health care settings, multisite longitudinal studies will be valuable to understand not only the impact of LLM-drafted HCs on clinician workflow (eg, do LLM-drafted HCs in fact reduce documentation time and burden while improving HC quality and clinician satisfaction?) but on how the sustained use of AI for documentation influences clinician behavior. Will the quality of documentation and summarization improve as the AI learns from human users, or will humans start to emulate the AI tool's behavioral patterns, including the same errors and biases?[8] As the authors note, LLM HCs continued to contain confabulations even after editing by humans, and this suggests that LLM drafts could influence humans to make more documentation errors than the status quo. The well-known fallacies of human vigilance[9] need to be considered in the long-term plans for monitoring these tools. Beyond predeployment evaluation frameworks, health systems will need to plan for ongoing monitoring of AI quality and safety and for AI-human interactions that could reflect human overreliance on technology, deskilling, or review complacency. Ultimately, Small et al[4] bring us closer to the immediate opportunity that LLMs hold for clinical care, while laying the groundwork to address challenges that lay ahead once they are deployed.

ARTICLE INFORMATION

Published: August 13, 2025. doi:10.1001/jamanetworkopen.2025.26350

Corresponding Author: Charumathi Raghu Subramanian, MD, University of California San Francisco, 10 Koret Way, San Francisco, CA 94143 (charumathi.raghusubramanian@ucsf.edu).

Author Affiliations: University of California, San Francisco, San Francisco, California (Raghu Subramanian); Division of Hospital Medicine, University of California, San Francisco, San Francisco, California (Rosner).

Conflict of Interest Disclosures: Dr Raghu Subramanian reported receiving personal fees from Lind, Suki AI, Ambience Healthcare Inc, and Evidently outside the submitted work. No other disclosures were reported.

REFERENCES

1. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA*. 2025;333(4):319-328. doi:10.1001/jama.2024.21700

2. Williams CYK, Subramanian CR, Ali SS, et al. Physician- and large language model-generated hospital discharge summaries. *JAMA Intern Med*. 2025;185(7):e250821. doi:10.1001/jamainternmed.2025.0821

3. Ganzinger M, Kunz N, Fuchs P, et al. Automated generation of discharge summaries: leveraging large language models with clinical data. *Sci Rep*. 2025;15(1):16466. doi:10.1038/s41598-025-01618-7

4. Small WR, Austrian J, O'Donnell L, et al. Hospital course summarization by an electronic health record–based large language model. *JAMA Netw Open*. 2025;8(8):e2526339. doi:10.1001/jamanetworkopen.2025.26339

5. Epic. AI for clinicians. Accessed June 2, 2025. https://www.epic.com/software/ai-clinicians/

6. Gebauer S. Benchmarking and datasets for ambient clinical documentation: a scoping review of existing frameworks and metrics for AI-assisted medical note generation. *medRxiv*. Preprint posted online January 29, 2025. doi:10.1101/2025.01.29.25320859

7. Croxford E, Gao Y, Pellegrino N, et al. Development and validation of the provider documentation summarization quality instrument for large language models. *J Am Med Inform Assoc*. 2025;32(6):1050-1060. doi: 10.1093/jamia/ocaf068

8. Glickman M, Sharot T. How human-AI feedback loops alter human perceptual, emotional and social judgements. *Nat Hum Behav*. 2025;9(2):345-359. doi:10.1038/s41562-024-02077-2

9. Adler-Milstein J, Redelmeier DA, Wachter RM. The limits of clinician vigilance as an AI safety bulwark. *JAMA*. 2024;331(14):1173-1174. doi:10.1001/jama.2024.3620