VIEWPOINT

AI IN MEDICINE

AI-Generated Clinical Summaries Require More Than Accuracy

Katherine E. Goodman, JD, PhD
Department of Epidemiology and Public Health, The University of Maryland School of Medicine, Baltimore; and The University of Maryland Institute for Health Computing, North Bethesda.

Paul H. Yi, MD
Department of Diagnostic Radiology and Nuclear Medicine, The University of Maryland School of Medicine, Baltimore.

Daniel J. Morgan, MD, MS
Department of Epidemiology and Public Health, The University of Maryland School of Medicine, Baltimore; and VA Maryland Healthcare System, Baltimore.

Viewpoint page 639

Supplemental content

Corresponding Author: Katherine E. Goodman, JD, PhD, Department of Epidemiology and Public Health, University of Maryland School of Medicine, 10 S Pine St, MSTF 257-A, Baltimore, MD 21201 (kgoodman@som.umaryland.edu).

Little more than a year after ChatGPT's public release, clinical applications of generative artificial intelligence and large language models (LLMs) are advancing rapidly. In the long term, LLMs may revolutionize much of clinical medicine, from patient diagnosis to treatment. In the short term, however, it is the everyday clinical tasks that LLMs will change most quickly and with the least scrutiny. Specifically, LLMs that summarize clinical notes, medications, and other forms of patient data are in advanced development and could soon reach patients without US Food and Drug Administration (FDA) oversight. Summarization, though, is not as simple as it seems, and variation in LLM-generated summaries could exert important and unpredictable effects on clinician decision-making.

Summarization Without FDA Oversight

Large language models that summarize clinical data represent a broad category. Simpler clinical documentation tools, which are already clinically available, create LLM-generated summaries from audio-recorded patient encounters. More sophisticated decision-support LLMs are under development that can summarize patient information from across the electronic health record (EHR). For example, LLMs could summarize a patient's recent visit notes and laboratory results to create an up-to-date clinical "snapshot" before an appointment. They could condense many lengthy radiology reports to an easily reviewable paragraph. Or LLMs could describe all of a patient's antibiotic exposure during the past year.

Current EHRs were built for documentation and billing and have inefficient information access and lengthy cut-and-pasted content. This poor design contributes to physician burnout and clinical errors.[1] If implemented well, LLM-generated summaries therefore offer impressive advantages and could eventually replace many point-and-click EHR interactions.

Yet there is also the potential for patient harm because LLMs performing summarization are unlikely to fall under FDA medical device oversight and could reach clinics without safety and efficacy safeguards. Indeed, FDA final guidance for clinical decision support software—published 2 months before ChatGPT's release—provides an unintentional "roadmap" for how LLMs could avoid FDA regulation.[2] Even LLMs performing sophisticated summarization tasks would not clearly qualify as devices because they provide general language-based outputs rather than specific predictions or numeric estimates of disease. With careful implementation, we expect that many LLMs summarizing clinical data could meet device-exemption criteria.[2]

"Accurate" Summaries Could Cause Harms

Currently, there are no comprehensive standards for LLM-generated clinical summaries beyond the general recognition that summaries should be consistently accurate and concise.[3] Yet there are many ways to accurately summarize clinical information. Variations in summary length, organization, and tone could all nudge clinician interpretations and subsequent decisions either intentionally or unintentionally. To illustrate these challenges concretely, we prompted ChatGPT-4 to summarize a small sample of deidentified clinical documents (**Figure**; eAppendix in the Supplement).

First, LLM-generated summaries are variable both because LLMs are probabilistic and because there is no "right" answer for precisely which information to include or how to order it. For example, running identical prompts on identical discharge documents, LLM summaries differed in the patient conditions listed and in the clinical history elements emphasized (Figure, A). These differences have important clinical implications because it is well documented that how information is organized and framed can change clinical decision-making.[4,5] Evaluating the impact of varied summaries on patient care requires clinical studies.

Second, even subtle differences between prompts can influence outputs.[6] In particular, LLMs can exhibit "sycophancy" bias.[7] Like the behavior of an eager personal assistant, sycophancy occurs when LLMs tailor responses to perceived user expectations. In the clinical context, sycophantic summaries could accentuate or otherwise emphasize facts that comport with clinicians' preexisting suspicions, risking a confirmation bias that could increase diagnostic error. For example, when prompted to summarize previous admissions for a hypothetical patient, summaries varied in clinically meaningful ways, depending on whether there was concern for myocardial infarction or pneumonia (Figure, B).

Third, even summaries that appear generally accurate could include small errors with important clinical influence. These errors are less like full-blown hallucinations than mental glitches, but they could induce faulty decision-making when they complete a clinical narrative or mental heuristic. For example, a chest radiography report noted indications of chills and nonproductive cough, but our LLM summary added "fever" (Figure, C). Including "fever," although a 1-word mistake, completes an illness script that could lead a physician toward a pneumonia diagnosis and initiation of antibiotics when they might not have reached that conclusion otherwise.

Recommendations

Absent statutory changes from Congress, the FDA will not have clear legal authority to regulate most LLMs generating clinical summaries. However, regulatory clarifications, coupled with robust voluntary actions, will go a long way toward protecting patients while preserving LLMs' benefits.

Figure. Summarization Considerations for Large Language Model (LLM)–Generated Clinical Summaries Beyond Accuracy

Summarization concern	Summary output (abbreviated)
A. Variability Variation across summaries due to random variability (large language models [LLMs] are probabilistic) and the many "right" ways to summarize information **Inputs:** Deidentified discharge summaries for a hypothetical patient **Prompt (abridged):** Summarize recent hospital visits for patient in emergency department. Provide a succinct, clinically relevant, and accurate summary. (Executed twice, a few hours apart)	**First run:** *"... 63-year-old male with schizophrenia, COPD, diabetes, and a history of urinary retention, ... Both admissions involved management of his chronic conditions and significant antibiotic use."* **Second run:** *"... 63-year-old male, had two recent hospital admissions... He received ceftriaxone and azithromycin for likely community-acquired pneumonia, and IV fluids for dehydration and orthostatic hypotension."* Summaries varied across otherwise identical runs, including in their organization, phrasing, and inclusion or exclusion of specific clinical details.
B. Sycophancy A form of bias in which the LLM tailors the summary output to perceived user expectations embedded in the prompt **Inputs:** Deidentified discharge summaries for a hypothetical patient **Prompt (abridged):** Summarize recent hospital visits for patient in emergency department with suspected myocardial infarction or pneumonia. Provide a succinct, clinically relevant, and accurate summary.	**Suspected myocardial infarction:** *"... 63-year-old male, has a history of paranoid schizophrenia, type 2 diabetes mellitus (T2DM), chronic obstructive pulmonary disease (COPD), coronary artery disease (CAD), hypertension, and hyperlipidemia... Both admissions involved complexity due to his mental health condition and multiple comorbidities."* **Suspected pneumonia:** *"... 63-year-old male, was admitted due to fever and altered mental status. He has a history of paranoid schizophrenia, COPD, type 2 diabetes mellitus, and urinary retention... Overall, his medical history includes chronic conditions like COPD, type 2 diabetes, and paranoid schizophrenia, along with repeated urinary tract infections and episodes suggesting pneumonia."* LLM emphasized patient's cardiac history or infection history from underlying discharge summaries.
C. "Complete-the-narrative" errors A small but clinically meaningful error (eg, 1-word addition) that completes a clinical narrative or illness script **Inputs:** Deidentified chest radiography report **Prompt (abridged):** Summarize the patient's radiology report in 2-3 sentences.	*"The patient's recent radiology report, indicating **fever**, chills, and a nonproductive cough in the context of known fibrotic lung disease, shows: ..."* "Fever" was added to summary by LLM, although not in original radiology report.

Criteria important for LLM-generated summaries of patient data in the electronic health record beyond accuracy are shown. Output summaries illustrate these respective concerns with real summaries of deidentified discharge summaries and radiologic reports generated by ChatGPT-4 (abbreviated for space; generated December 2023). Note that the FDA has interpreted clinical decision support software involved in "time-critical" decision-making as a regulated device function, which could possibly include LLM generation of a clinical summary. See the eAppendix in the Supplement for unabridged input documents, prompts, and output summaries.

First, we need comprehensive standards for LLM-generated summaries, with domains that extend beyond accuracy and that include stress-testing for sycophancy and small but clinically important errors. These standards should reflect scientific and clinical consensus, with input beyond the few large technology companies developing health care LLMs. Second, LLMs performing clinical summarization are ultimately clinical aids. Regardless of current FDA regulation, we believe that they should be clinically tested to quantify clinical harms and benefits before widespread deployment. This testing carries minimal risk and could be performed as quality improvement in a learning health system. Third, the highest-risk—but likely most useful—summarization LLMs will permit more open-ended clinician prompting, and we encourage the FDA to clarify regulatory criteria preemptively. These clarifications should specify that some prompts (eg, "summarize my patient's history relevant to risk of heart failure") cause LLMs to function as medical devices despite semantically restricting to summarization. The FDA could offer these statements in new guidance or as updates to existing guidance to recognize that the world has changed meaningfully since the clinical decision support guidance's original release in late 2022.

Large language models summarizing clinical data promise powerful opportunities to streamline information-gathering from the EHR. But by dealing in language, they also bring unique risks that are not clearly covered by existing FDA regulatory safeguards. As summarization tools speed closer to clinical practice, transparent development of standards for LLM-generated clinical summaries, paired with pragmatic clinical studies, will be critical to the safe and prudent rollout of these technologies. We encourage the FDA to clarify its oversight before summarization becomes a part of routine patient care.

ARTICLE INFORMATION

Published Online: January 29, 2024. doi:10.1001/jama.2024.0555

Conflict of Interest Disclosures: Dr Morgan reported receiving grants from NIH, AHRQ, CDC, and the VA outside the submitted work. No other disclosures were reported.

Funding/Support: Dr Goodman's work was supported by an AHRQ career development award (K01HS028363).

Role of the Funder/Sponsor: AHRQ had no role in the preparation, review, or approval of the manuscript and decision to submit the manuscript for publication.

Additional Information: We used OpenAI's ChatGPT-4 to create clinical summaries from deidentified clinical data to provide examples for the figure and text.

REFERENCES

1. Vaughn VM, Linder JA. Thoughtless design of the electronic health record drives overuse, but purposeful design can nudge improved patient care. *BMJ Qual Saf*. 2018;27(8):583-586. doi:10.1136/bmjqs-2017-007578

2. Clinical decision support software: guidance for industry and Food and Drug Administration staff. September 28, 2022. Accessed January 6, 2024. https://www.fda.gov/media/109618/download

3. Van Veen D, Van Uden C, Blankemeier L, et al. Clinical text summarization: adapting large language models can outperform human experts. Preprint posted online September 14, 2023. Accessed January 5, 2024. *arXiv*. doi:10.21203/rs.3.rs-3483777/v1

4. Ly DP, Shekelle PG, Song Z. Evidence for anchoring bias during physician decision-making. *JAMA Intern Med*. 2023;183(8):818-823. doi:10.1001/jamainternmed.2023.2366

5. Bui TC, Krieger HA, Blumenthal-Barby JS. Framing effects on physicians' judgment and decision making. *Psychol Rep*. 2015;117(2):508-522. doi:10.2466/13.PR0.117c20z0

6. Chuang YN, Tang R, Jiang X, Hu X. SPeC: a soft prompt-based calibration on performance variability of large language model in clinical notes summarization. Preprint posted online March 23, 2023. Accessed January 4, 2024. *arXiv*. doi:10.48550/arXiv.2303.13035

7. Sharma M, Tong M, Korbak T, et al. Towards understanding sycophancy in language models. Preprint posted online October 20, 2023. Accessed January 6, 2024. *arXiv*. doi:10.48550/arXiv.2310.13548