Accepted Manuscript online: 2024-03-05 Article published online: 2024-04-17

A Survey of Clinicians' Views of the Utility of Large Language Models

Matthew Spotnitz[1] Betina Idnay[1] Emily R. Gordon[1,2] Rebecca Shyu[1] Gongbo Zhang[1] Cong Liu[1] James J. Cimino[1,3] Chunhua Weng[1]

[1] Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, New York, United States
[2] Department of Dermatology, Vagelos College of Physicians and Surgeons, Columbia University Irving Medical Center, New York, New York, United States
[3] Department of Biomedical Informatics and Data Science, Informatics Institute, Heersink School of Medicine, University of Alabama at Birmingham, Birmingham, Alabama, United States

Address for correspondence Matthew Spotnitz, MD, MPH, Department of Biomedical Informatics, Columbia University Irving Medical Center, 630 West 168th Street New York, NY 10032, United States (e-mail: mes698@mail.harvard.edu).

Chunhua Weng, PhD, Department of Biomedical Informatics, Columbia University Irving Medical Center, 630 West 168th Street New York, NY 10032, United States (e-mail: cw2384@cumc.columbia.edu).

Appl Clin Inform 2024;15:306–312.

Abstract

Objectives Large language models (LLMs) like Generative pre-trained transformer (ChatGPT) are powerful algorithms that have been shown to produce human-like text from input data. Several potential clinical applications of this technology have been proposed and evaluated by biomedical informatics experts. However, few have surveyed health care providers for their opinions about whether the technology is fit for use.

Methods We distributed a validated mixed-methods survey to gauge practicing clinicians' comfort with LLMs for a breadth of tasks in clinical practice, research, and education, which were selected from the literature.

Results A total of 30 clinicians fully completed the survey. Of the 23 tasks, 16 were rated positively by more than 50% of the respondents. Based on our qualitative analysis, health care providers considered LLMs to have excellent synthesis skills and efficiency. However, our respondents had concerns that LLMs could generate false information and propagate training data bias.

Our survey respondents were most comfortable with scenarios that allow LLMs to function in an assistive role, like a physician extender or trainee.

Conclusion In a mixed-methods survey of clinicians about LLM use, health care providers were encouraging of having LLMs in health care for many tasks, and especially in assistive roles. There is a need for continued human-centered development of both LLMs and artificial intelligence in general.

Keywords

- artificial intelligence
- clinical decision support
- clinical informatics
- clinical information systems

Background and Significance

Large language models (LLMs), which are a type of artificial intelligence (AI), are designed to process and understand human language. They are usually trained on massive amounts of text. For example, ChatGPT is a very efficient LLM that has garnered a great deal of public attention for its capabilities since its recent introduction in late 2022.[1–3] The possible health care applications of LLMs are numerous. Representative examples include generating clinical

received
December 1, 2023
accepted after revision
February 15, 2024
accepted manuscript online
March 5, 2024

DOI https://doi.org/10.1055/a-2281-7092.
ISSN 1869-0327.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

documentation, personalized educational materials, and original scientific manuscripts.[4,5]

One well-known limitation of ChatGPT is its tendency for "hallucination," the generation of text that is perceived as convincing but is not accurate.[6] A second limitation is that ChatGPT can propagate bias that is intrinsic in the training data. These issues have raised concerns about the safety of LLM use in health care. Specifically, some researchers envision scenarios where ChatGPT could provide clinical care advice that is outdated, inaccurate, or incomplete.[7–10]

Determining the best uses of LLMs in health care has been the focus of recent studies. In a prior publication, clinicians with informatics expertise evaluated LLMs for clinical decision support and concluded that they may provide valuable assistance (Liu et al, 2023).[11] However, that study did not address the experience of novice LLM users. Furthermore, few studies have investigated health care provider comfort with LLMs or used both quantitative and qualitative methods. Those studies either asked general questions about the suitability of LLMs in different health care domains[12,13] or compared it to human performance on one health care delivery service.[14]

Presently, applications of LLMs are being developed at a rapid pace and could have widespread adoption within health care by novice and expert users alike. The most ethical and effective implementation of the technology must consider user requirements and concerns from representative stakeholders of this technology in the clinical setting. In this study, we surveyed diverse practicing clinicians about using LLMs for tasks in clinical practice, research, and education and summarized their perceptions of the potential and limitations of LLMs, to inform the development of clinically meaningful evaluation standards for LLMs to ensure their appropriate and ethical implementation in clinical settings.

Methods

Study Design and Sampling

The survey instrument, which is shown in ►**Supplementary Appendix 1** (available in the online version only), was developed by two authors with both clinical and informatics experience (M.S., B.I.) and refined based on feedback from a third author (E.R.G.). It was implemented through Qualtrics (Qualtrics, Provo, UT), took approximately 15 minutes to complete, and gauged clinicians' perceptions on the appropriateness of using LLMs in clinical practice, research, and education.

The opening questions quantified participants' experience in clinical medicine and informatics with multiple choice answers. Next, there were questions that asked if the amount of LLM experience in health care within the past year had exceeded 50 hours. Then, there were questions that prompted rating the appropriateness of LLM use for 23 different tasks in clinical practice, research, and education on a 5-point Likert scale (i.e., Highly Appropriate to Highly Inappropriate). Those tasks represented a sample of proposed LLM uses that were synthesized from the literature and included, but were not limited to, optimizing alerts for clinical decision support, providing a differential diagnosis, writing a discharge summary, recommending treatment options, translating radiology reports into layperson language, writing scientific manuscripts, and generating personalized study plans for students or trainees among others.[2,6–8,15–25]

We distributed the proposed clinical practice tasks over two questions, and had one question for research tasks and one question for education tasks. The purpose of this section was to measure the appropriateness of these tasks by category and to determine if any individual tasks were negative or positive outliers. We hypothesized that perceptions about the strengths, limitations, and ethical concerns about LLMs could contribute to the ratings. Therefore, we had open-ended questions about each of those as well as an open-ended question about other possible uses of LLMs.

Data Collection

We recruited participants with an email invitation that was sent to a listserv of clinicians at Columbia University Irving Medical Center and by word of mouth. To be eligible, participants needed to be practicing clinicians affiliated with Columbia University within the past 12 months and were able to comprehend and communicate fluently in English. Respondents were compensated with a $20 Amazon Gift Card for completing the survey.

Data Analysis

We calculated descriptive statistics on the participants and tabulated their ratings for each question. Two independent reviewers performed an inductive thematic analysis with the narrative comments. Both of them performed independent coding of free text using NVivo (Version 14) with generation of themes. They met regularly for a total of three times and developed themes iteratively. Once a consensus was reached, the reviewers determined a final list of themes and applied them to the narrative comments. A third reviewer was available to resolve any discrepancies.

Results

We recruited practicing clinicians from internal medicine, otolaryngology, ophthalmology, pediatrics, urology, anesthesiology, neurosurgery, and general surgery. We distributed a prescreening survey to 350 clinicians, among whom 108 responded, and 30 were eligible and enrolled. All completed the survey. Their demographics are shown in ►**Table 1**.

Survey Ratings

Heat maps of the ratings for clinical, research, and education tasks are shown in ►**Tables 2–4**. Of the 23 tasks, 16 (69.6%) had positive ratings by at least half of the participants. The highest rated tasks were "assist with vaccine development by predicting the antigenicity of different proteins from genomic data" (25 positive ratings from 30 participants), "model the spread and transmission of an infectious disease" (25 positive ratings), "generate case studies for training purposes" (24 positive ratings), "monitor data for an

Table 1 Participant information

Survey characteristics	*N* (%)
Clinical Training	
1–2 years	12 (40)
3–5 years	3 (10)
> 5 years	15 (50)
Informatics Training	
None	26 (86.7)
1–2 years	3 (10)
3+ years	1 (3.3)
LLM use within the past 12 months	
< 50 hours	28 (93.3)
50+ hours	2 (6.7)

Abbreviations: LLM, large language model; *N*, number; %, percent.

emerging disease cluster" (24 positive ratings), and "generate alerts to improve compliance with clinical guidelines" (24 positive ratings). In contrast, 7 out of 23 tasks had positive ratings by fewer than half of the participants. Two of the tasks with the lowest number of positive ratings also had the highest number of negative ratings, which were "Respond to patient questions about a radiology report" (7 positive ratings, 16 negative ratings), and "Write an original scientific manuscript" (5 positive ratings, 20 negative ratings).

Thematic Analysis

We received 20 open-ended responses about LLM strengths, limitations, and ethical concerns, respectively. The responses about the limitations and ethical concerns of LLMs were very similar; therefore, we combined them for a total of 40 responses. There were 19 responses about additional uses of LLMs. The themes and corresponding examples are shown in ►**Table 5**. Some respondent answers addressed multiple themes and were mapped to each of them. The full responses to our open-ended questions are shown in ►**Supplementary Appendix 2** (available in the online version only).

Discussion

LLMs promise to transform health care. A human-centered approach is critical to ensure ethical and effective implementation of this powerful technology in clinical settings. This was the first study of clinical practitioners, who were mostly novice LLM users and from diverse health care domains, to rate tasks that may be improved by LLMs.

The fundamental theorem of biomedical informatics is user augmentation so that "a person working in partnership with an information resource is better than that same person unassisted."[26] Similarly, the clinicians who we studied were encouraging of having LLMs as their assistants. The tasks that

Table 2 Heatmap of ratings for large language model uses in clinical practice tasks (orange = lowest; yellow = highest)

Task	Highly Inappropriate	Inappropriate	Neutral	Appropriate	Highly Appropriate
Generate alerts to improve compliance with clinical guidelines	1	2	3	19	5
Provide a differential diagnosis	0	2	9	16	3
Describe how to perform a procedure	0	7	6	16	1
Translate radiology reports into layperson language	2	2	5	15	6
Synthesize and present patient data from the electronic health record for clinical decision support	2	5	4	15	4
Write discharge summaries	2	6	2	13	7
Suggest patient management or treatment options	2	5	8	13	2
Check for inaccuracies in a radiology report, and notify providers of them	2	1	10	11	6
Report current information on a topic for clinical decision support	3	1	14	10	2
Write radiology reports	3	8	12	6	1
Respond to patient questions about a radiology report	2	14	7	6	1

Table 3 Heatmap of ratings for large language model uses in research tasks (orange = lowest; yellow = highest)

Task	Highly Inappropriate	Inappropriate	Neutral	Appropriate	Highly Appropriate
Suggest how to interpret a dataset	1	2	7	18	2
Assist with vaccine development by predicting the antigenicity of different proteins from genomic data	1	0	4	16	9
Model the spread and transmission of an infectious disease	1	1	3	16	9
Monitor data for an emerging disease cluster	1	2	3	15	9
Generate programming code	0	0	10	14	6
Write a literature review for a research publication	5	8	8	8	1
Write an original scientific manuscript	10	10	5	3	2

Table 4 Heatmap of ratings for large language model uses in education tasks (orange = lowest; yellow = highest)

Task	Highly Inappropriate	Inappropriate	Neutral	Appropriate	Highly Appropriate
Generate interactive simulations for training purposes	0	1	6	17	6
Generate case studies for training purposes	0	1	5	15	9
Write quizzes and self-assessments for students or trainees	0	2	6	15	7
Generate a personalized study plan for a medical student	0	1	8	15	6
Write personalized patient education texts for students or trainees	3	3	12	8	4

leveraged LLMs for supportive roles were rated the highest. In the qualitative analysis, emerging themes were that LLMs were highly skilled at different tasks; however, there were ethical concerns about using the technology. Supportive LLM roles may have been more popular, because in those scenarios clinicians could correct for false information that the algorithms might generate.

Therefore, we expect that clinicians would prefer to have LLMs function more like trainees or physician extenders than attending physicians. LLMs could assist clinicians by drafting notes and reports, making suggestions for patient triage, extracting patient information from charts, and identifying discrepancies from standard patient care. Since LLMs are very skilled at processing large amount of data, they could help monitor patients in critical care and perioperative settings. Also, they could help translate medical information between languages, or from technical jargon into layperson language. The contributions of LLMs to these tasks could be reviewed by a clinician.

However, the notion of allowing LLMs to function without supervision in clinical practice raises ethical concerns. They have a propensity to produce false information and propagate data bias, which could lead to incorrect medical decisions. Furthermore, LLMs lack human empathy, which could be a source of mistrust with patients. Instead, patients are more likely to trust medical advice from a clinician because of the human connection. Overall, we believe that clinicians would prefer to have LLMs assist them instead of replace their practice.

Our study participants were encouraging of LLM assistance in the research and education domains as well. In research, the processing power of LLMs would allow them to help with a range of statistical analyses. Also, their linguistic capabilities could translate ideas across human and programming languages. Those skills could be especially useful in large research networks, which consist of individuals from different countries and who have different programming skills. However, having LLMs author an original manuscript

Table 5 Summary of narrative comments about perceived advantages, ethical concerns, and clinical applications of large language models with representative examples

Advantages ($n = 20$)	Ethical concerns ($n = 40$)	Recommended clinical applications for using LLMs ($n = 19$)
Aptitude for specific tasks ($n = 10$) *Ability to generate first drafts with low effort* *It can also help students and providers come up with a differential diagnosis* *Theoretically could reduce paperwork/administrative work* *Ability to write code for novice programmers*	False information ($n = 15$) *Hallucination, fabrication, reinforcement of assumptions and biases* *With the confabulation/hallucination issue, does not allow for the uncertainty that is almost always present in medicine* *Its propensity to make up information*	Drafting documentation ($n = 8$) *Note templates/drafts, especially for routine and predictable things like procedures*
Synthesis ability ($n = 9$) *Synthesize large amounts of data quickly* *Good at synthesizing information in a clear concise fashion*	Worsens patient care ($n = 14$) *This technology if unchecked at a patient care level may have serious implications of harm to patients* *Major concern about inappropriate use by lay public to self-diagnose* *Also worry about who gets care from health care workers versus from direct-to-patient LLM which could be less personalized, and initially less validated and trustworthy*	Decision support ($n = 5$) *Anything providing recommendations to patients or providers* *Flagging concerning trends (VS, laboratory values) earlier; providing guidance in managing chronic conditions*
Efficiency ($n = 8$) *Saves time and improves efficiency*	Data bias ($n = 12$) *Results are only as good as the datasets that are fed into the LLM* *Given the fast pace of evidence in health care, can be trained on old evidence* *Poor data quality leads to poor answers* *Replicates existing biases*	Patient communication ($n = 4$) *Drafting replies to patient messages in the outpatient inbox that are modeled off of the provider's communication style*
Accuracy ($n = 5$) *Fairly accurate and provide higher quality, more personalized information than most patient-facing information available on the internet*	Human oversight critical ($n = 7$) *They should not replace informed decision-making for patients or clinical decision-making for doctors completely*	–
Accessibility ($n = 4$) *Translating medical documents into plain English*	Impersonal ($n = 6$) *Worry about who gets care from health care workers versus from direct-to-patient LLM which could be less personalized*	–
–	Legal concerns ($n = 5$) *Gray area of ethical/legal limitations*	–
–	Privacy ($n = 3$) *Worries of patient confidentiality*	–
–	Worsens clinicians ($n = 2$) *If we become reliant on LLMs, we may lose opportunities to practice interpreting/synthesizing data ourselves*	–

Abbreviation: LLM, large language model; VS, vital signs.

instead of a researcher would raise similar ethical concerns to allowing LLMs to function as an autonomous clinician. The education tasks raised the fewest ethical concerns, perhaps because students have regular supervision and a smaller role in direct patient care than clinicians or researchers.

Our sampling method followed a defined set of recruitment criteria and enrolled a total of 30 practicing clinicians who completed the survey for this study. While a larger number of respondents would have been desirable, ours covered a variety of clinical domains and provided valuable, original insights regarding the ethical and reliable uses of LLMs in clinical settings. Given the unusually rapid evolution of LLM technology, this early study is timely and makes meaningful contributions by including the voices of key stakeholders of implementing LLMs for clinical tasks.

A limitation of our study, and a potential source of sampling bias, is that only a relatively small number of participants from a single medical center were recruited

by convenience sampling. Also, we used self-reported data as key elements of our analysis. These data may have introduced biases due to varying accuracy in self-reports and varying awareness of the problems by reporting individuals. Despite these limitations, we have developed an instrument that is capable of discerning different opinions about LLM use. We hope our findings can be taken into consideration by developers as the field continues its rapid evolution. As further progress is made, and clinicians have more significant experience with this technology, subsequent studies can build on our methods and experience to study larger sample sizes at multiple institutions to gain additional insights for future directions.

Future studies with a larger and more diverse sample will be warranted to ensure the generalizability of the results and allow for stratification by variables that could affect perceptions of LLM use, such as age, duration of clinical training, provider specialty, and experience with the technology. Those perceptions could be tracked longitudinally to gauge how they change over time. A more robust study about participants' general knowledge of LLMs and AI would strengthen future studies. Specifically, gauging to what extent participants understand how an AI algorithm is able to work, predict, learn, and generate responses, would be a valuable part of an analysis. Furthermore, comparing the perceptions regarding different LLMs, and how LLM-generated errors compare with human errors, may provide a more balanced view of the technology.

Our study found that health care providers would prefer to have LLMs assist than replace them. That finding has implications for future development and implementation of LLMs in clinical practice, research, and education. Studying active clinicians with novice LLM experience helped identify that preference. Therefore, for optimal development and implementation of LLMs in health care, continued human centered development is critical.

Conclusion

Clinicians are generally supportive of the use of LLMs for many tasks in clinical practice, research, and education, especially where LLMs play a supportive role to humans. Continued human centered development of the technology is critical.

Clinical Relevance Statement

We studied health care providers about the best uses of LLMs in health care. The clinicians who we studied were encouraging of having LLMs assist them for a range of tasks. The results of our work have implications for implementation of LLMs in health care.

Multiple Choice Questions

1. Which of the following are ethical concerns about LLM use?
   a. Efficiency
   b. Confabulation or hallucination
   c. Ability to synthesize information
   d. Capacity to make technical language accessible

   Answer: **b**. Confabulation or hallucination can cause the LLMs to generate false information, which can lead to incorrect medical decisions. The other answer choices are advantages of the technology.

2. What is the fundamental theorem of biomedical informatics?
   a. An information resource is better without assistance from a person.
   b. An information resource working in partnership with a person is better than an information resource unassisted.
   c. A person working in partnership with an information resource is better than that same person unassisted.
   d. A person is better without an information resource.

   Answer: **c**. The fundamental theorem of biomedical informatics states that people are more effective when partnered with an information resource. The alternatives, which are to have no cooperation with information resources and people, or to have people assist information resources, are less effective.

Protection of Human Subjects

The study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects, and was reviewed by Columbia University Irving Medical Center Institutional Review Board (AAAU7954).

Funding

This work was supported by National Library of Medicine (NLM) grants R01LM014344 and R01LM009886 to C.W., National Human Genome Institute grant R01HG012655 to C.L., and by National Center for Advancing Clinical and Translational Science grant UL1TR001873 to Columbia University Irving Medical Center. B.I. and R.S. acknowledge support from NLM grant T15LM007079.

Conflict of Interest

None declared.

References

1 Sezgin E, Sirrianni J, Linwood SL. Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the US health care system: outlook of generative pretrained transformer 3 (GPT-3) as a service model. JMIR Med Inform 2022; 10(02):e32875
2 Elkassem AA, Smith AD. Potential use cases for ChatGPT in radiology reporting. AJR Am J Roentgenol 2023;221(03):373–376
3 Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res 2023;25:e48659
4 Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023;47(01):33
5 Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthc Pap 2023;11(06):867

6 Athaluri SA, Manthena SV, Kesapragada VSRKM, Yarlagadda V, Dave T, Duddumpudi RTS. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus 2023;15(04):e37432
7 Peng Y, Rousseau JF, Shortliffe EH, Weng C. AI-generated text may have a role in evidence-based medicine. Nat Med 2023;29(07): 1593–1594
8 Tang L, Sun Z, Idnay B, et al. Evaluating large language models on medical evidence summarization. NPJ Digit Med 2023;6(01):158
9 Deik A. Potential benefits and perils of incorporating ChatGPT to the movement disorders clinic. J Mov Disord 2023;16(02): 158–162
10 Shashavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors 2023;10:e47564
11 Liu S, Wright AP, Patterson BL, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023;30(07):1237–1245
12 Hosseini M, Gao CA, Leibovitz DM, et al. An exploratory survey about using ChatGPT in education, healthcare, and research. medRxiv 2023.03.31.23287979
13 Choudhury A, Shamszare H. Investigating the impact of user trust on the adoption and use of ChatGPT: survey analysis. J Med Internet Res 2023;25:e47184
14 Dash D, Rahul T, Banda JM, et al. Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery. 2023. arXiv:2304.13714
15 Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. Int J Environ Res Public Health 2023;20(04):3378
16 Cheng K, Li Z, He Y, et al. Potential use of artificial intelligence in infectious disease: take ChatGPT as an example. Ann Biomed Eng 2023;51(06):1130–1135
17 Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023;5(03):e107–e108
18 Galido PV, Butala S, Chakerian M, Agustines D. A case study demonstrating applications of ChatGPT in the clinical management of treatment-resistant schizophrenia. Cureus 2023;15(04):e38166
19 Sharma S, Pajai S, Prasad R, et al. A critical review of ChatGPT as a potential substitute for diabetes educators. Cureus 2023;15(05): e38380
20 Macdonald C, Adeloye D, Sheikh A, Rudan I. Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. J Glob Health 2023;13:01003
21 Májovský M, Černý M, Kasal M, Komarc M, Netuka D. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. J Med Internet Res 2023;25:e46924
22 Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023;9:e48291
23 Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ 2023;9: e46885
24 Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The advent of generative language models in medical education. JMIR Med Educ 2023;9:e48163
25 Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023; 330(01):78–80
26 Friedman CPA. A "fundamental theorem" of biomedical informatics. J Am Med Inform Assoc 2009;16(02):169–170