JAMA Internal Medicine | Original Investigation

Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum

John W. Ayers, PhD, MA; Adam Poliak, PhD; Mark Dredze, PhD; Eric C. Leas, PhD, MPH; Zechariah Zhu, BS; Jessica B. Kelley, MSN; Dennis J. Faix, MD; Aaron M. Goodman, MD; Christopher A. Longhurst, MD, MS; Michael Hogarth, MD; Davey M. Smith, MD, MAS

IMPORTANCE The rapid expansion of virtual health care has caused a surge in patient messages concomitant with more work and burnout among health care professionals. Artificial intelligence (AI) assistants could potentially aid in creating answers to patient questions by drafting responses that could be reviewed by clinicians.

OBJECTIVE To evaluate the ability of an AI chatbot assistant (ChatGPT), released in November 2022, to provide quality and empathetic responses to patient questions.

DESIGN, SETTING, AND PARTICIPANTS In this cross-sectional study, a public and nonidentifiable database of questions from a public social media forum (Reddit's r/AskDocs) was used to randomly draw 195 exchanges from October 2022 where a verified physician responded to a public question. Chatbot responses were generated by entering the original question into a fresh session (without prior questions having been asked in the session) on December 22 and 23, 2022. The original question along with anonymized and randomly ordered physician and chatbot responses were evaluated in triplicate by a team of licensed health care professionals. Evaluators chose "which response was better" and judged both "the quality of information provided" (*very poor*, *poor*, *acceptable*, *good*, or *very good*) and "the empathy or bedside manner provided" (*not empathetic*, *slightly empathetic*, *moderately empathetic*, *empathetic*, and *very empathetic*). Mean outcomes were ordered on a 1 to 5 scale and compared between chatbot and physicians.

RESULTS Of the 195 questions and responses, evaluators preferred chatbot responses to physician responses in 78.6% (95% CI, 75.0%-81.8%) of the 585 evaluations. Mean (IQR) physician responses were significantly shorter than chatbot responses (52 [17-62] words vs 211 [168-245] words; $t = 25.4$; $P < .001$). Chatbot responses were rated of significantly higher quality than physician responses ($t = 13.3$; $P < .001$). The proportion of responses rated as *good* or *very good* quality (≥ 4), for instance, was higher for chatbot than physicians (chatbot: 78.5%, 95% CI, 72.3%-84.1%; physicians: 22.1%, 95% CI, 16.4%-28.2%;). This amounted to 3.6 times higher prevalence of *good* or *very good* quality responses for the chatbot. Chatbot responses were also rated significantly more empathetic than physician responses ($t = 18.9$; $P < .001$). The proportion of responses rated *empathetic* or *very empathetic* (≥ 4) was higher for chatbot than for physicians (physicians: 4.6%, 95% CI, 2.1%-7.7%; chatbot: 45.1%, 95% CI, 38.5%-51.8%; physicians: 4.6%, 95% CI, 2.1%-7.7%). This amounted to 9.8 times higher prevalence of *empathetic* or *very empathetic* responses for the chatbot.

CONCLUSIONS In this cross-sectional study, a chatbot generated quality and empathetic responses to patient questions posed in an online forum. Further exploration of this technology is warranted in clinical settings, such as using chatbot to draft responses that physicians could then edit. Randomized trials could assess further if using AI assistants might improve responses, lower clinician burnout, and improve patient outcomes.

Invited Commentary page 596

Related article page 507

Supplemental content

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: John W. Ayers, PhD, MA, Qualcomm Institute, University of California San Diego, La Jolla, CA (ayers.john.w@gmail.com).

JAMA Intern Med. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838
Published online April 28, 2023. Corrected on May 8, 2023.

The COVID-19 pandemic hastened the adoption of virtual health care,[1] concomitant with a 1.6-fold increase in electronic patient messages, with each message adding 2.3 minutes of work in the electronic health record and more after-hours work.[2] Additional messaging volume predicts increased burnout for clinicians[3] with 62% of physicians, a record high, reporting at least 1 burnout symptom.[4] More messages also makes it more likely that patients' messages will go unanswered or get unhelpful responses.

Some patient messages are unsolicited questions seeking medical advice, which also take more skill and time to answer than generic messages (eg, scheduling an appointment, accessing test results). Current approaches to decreasing these message burdens include limiting notifications, billing for responses, or delegating responses to less trained support staff.[5] Unfortunately, these strategies can limit access to high-quality health care. For instance, when patients were told they might be billed for messaging, they sent fewer messages and had shorter back-and-forth exchanges with clinicians.[6] Artificial intelligence (AI) assistants are an unexplored resource for addressing the burden of messages. While some proprietary AI assistants show promise,[7] some public tools have failed to recognize even basic health concepts.[8,9]

ChatGPT[10] represents a new generation of AI technologies driven by advances in large language models.[11] ChatGPT reached 100 million users within 64 days of its November 30, 2022 release and is widely recognized for its ability to write near-human-quality text on a wide range of topics.[12] The system was not developed to provide health care, and its ability to help address patient questions is unexplored.[13] We tested ChatGPT's ability to respond with high-quality and empathetic answers to patients' health care questions, by comparing the chatbot responses with physicians' responses to questions posted on a public social media forum.

Key Points

Question Can an artificial intelligence chatbot assistant, provide responses to patient questions that are of comparable quality and empathy to those written by physicians?

Findings In this cross-sectional study of 195 randomly drawn patient questions from a social media forum, a team of licensed health care professionals compared physician's and chatbot's responses to patient's questions asked publicly on a public social media forum. The chatbot responses were preferred over physician responses and rated significantly higher for both quality and empathy.

Meaning These results suggest that artificial intelligence assistants may be able to aid in drafting responses to patient questions.

Methods

Studying patient questions from health care systems using a chatbot was not possible in this cross-sectional study because, at the time, the AI was not compliant with the Health Insurance Portability and Accountability Act of 1996 (HIPAA) regulations. Deidentifying patient messages by removing unique information to make them HIPAA compliant could change the content enough to alter patient questions and affect the chatbot responses. Additionally, open science requires public data to enable research to build on and critique prior research.[14] Lastly, media reports suggest that physicians are already integrating chatbots into their practices without evidence. For reasons of need, practicality, and to empower the development of a rapidly available and shareable database of patient questions, we collected public and patient questions and physician responses posted to an online social media forum, Reddit's r/AskDocs.[15]

The online forum, r/AskDocs, is a subreddit with approximately 474 000 members where users can post medical questions and verified health care professional volunteers submit answers.[15] While anyone can respond to a question, subreddit moderators verify health care professionals' credentials and responses display the respondent's level of credential next to their response (eg, physician) and flag a question when it has already been answered. Background and use cases for data in this online forum are described by Nobles et al.[16]

All analyses adhered to Reddit's terms and conditions[17] and were determined by the University of California, San Diego, human research protections program to be exempt. Informed consent was not required because the data were public and did not contain identifiable information (45 CFR §46). Direct quotes from posts were summarized to protect patient's identities.[18] Actual quotes were used to obtain the chatbot responses.

Our study's target sample was 200, assuming 80% power to detect a 10 percentage point difference between physician and chatbot responses (45% vs 55%). The analytical sample ultimately contained 195 randomly drawn exchanges, ie, a unique member's question and unique physician's answer, during October 2022. The original question, including the title and text, was retained for analysis, and the physician response was retained as a benchmark response. Only physician responses were studied because we expected that physicians' responses are generally superior to those of other health care professionals or laypersons. When a physician replied more than once, we only considered the first response, although the results were nearly identical regardless of our decision to exclude or include follow-up physician responses (see eTable 1 in Supplement 1). On December 22 and 23, 2022, the original full text of the question was put into a fresh chatbot session, in which the session was free of prior questions asked that could bias the results (version GPT-3.5, OpenAI), and the chatbot response was saved.

The original question, physician response, and chatbot response were reviewed by 3 members of a team of licensed health care professionals working in pediatrics, geriatrics, internal medicine, oncology, infectious disease, and preventive medicine (J.B.K., D.J.F., A.M.G., M.H., D.M.S.). The evaluators were shown the entire patient's question, the physician's response, and chatbot response. Responses were randomly ordered, stripped of revealing information (eg, statements such as "I'm an artificial intelligence"), and labeled *response 1* or *response 2* to blind evaluators to the identity of the responders. The evaluators were instructed to read the entire patient question and both responses before answering questions about the interaction. First, evaluators were asked "which response [was] better" (ie, response 1 or response 2). Then, using Likert scales, evaluators

judged both "the quality of information provided" (*very poor*, *poor*, *acceptable*, *good*, *or very good*) and "the empathy or bedside manner provided" (*not empathetic*, *slightly empathetic*, *moderately empathetic*, *empathetic*, and *very empathetic*) of responses. Response options were translated into a 1 to 5 scale, where higher values indicated greater quality or empathy.

We relied on a crowd (or ensemble) scoring strategy,[19] where scores were averaged across evaluators for each exchange studied. This method is used when there is no ground truth in the outcome being studied, and the evaluated outcomes themselves are inherently subjective (eg, judging figure skating, National Institutes of Health grants, concept discovery). As a result, the mean score reflects evaluator consensus, and disagreements (or inherent ambiguity, uncertainty) between evaluators is reflected in the score variance (eg, the CIs will, in part, be conditional on evaluator agreement).[20]

We compared the number of words in physician and chatbot responses and reported the percentage of responses for which chatbot was preferred. Using 2-tailed *t* tests, we compared mean quality and empathy scores of physician responses with chatbot responses. Furthermore, we compared rates of responses above or below important thresholds, such as *less than adequate*, and computed prevalence ratios comparing the chatbot to physician responses. The significance threshold used was $P < .05$. All statistical analyses were performed in R statistical software, version 4.0.2 (R Project for Statistical Computing).

We also reported the Pearson correlation between quality and empathy scores. Assuming that in-clinic patient questions may be longer than those posted on the online forum, we also assessed the extent to which subsetting the data into longer replies authored by physicians (including those above the median or 75th percentile length) changed evaluator preferences and the quality or empathy ratings relative to the chatbot responses.

Results

The sample contained 195 randomly drawn exchanges with a unique member-patient's question and unique physician's answer. The mean (IQR) length of patient questions in words averaged 180 (94-223). Mean (IQR) physician responses were significantly shorter than the chatbot responses (52 [17-62] words vs 211 [168-245] words; $t = 25.4$; $P < .001$). A total of 182 (94%) of these exchanges consisted of a single message and only a single response from a physician. A remaining 13 (6%) exchanges consisted of a single message but with 2 separate physician responses. Second responses appeared incidental (eg, an additional response was given when a post had already been answered) (eTable 1 in Supplement 1).

The evaluators preferred the chatbot response to the physician responses 78.6% (95% CI, 75.0%-81.8%) of the 585 evaluations. Summaries of example questions and the corresponding physician and chatbot responses are shown in the **Table**.

Evaluators also rated chatbot responses significantly higher quality than physician responses ($t = 13.3$; $P < .001$). The mean rating for chatbot responses was better than *good* (4.13; 95% CI, 4.05-4.20), while on average, physicians' responses were rated 21% lower, corresponding to an *acceptable* response (3.26; 95% CI, 3.15-3.37) (**Figure**). The proportion of responses rated less than *acceptable* quality (<3) was higher for physician responses than for chatbot (physicians: 27.2%; 95% CI, 21.0%-33.3%; chatbot: 2.6%; 95% CI, 0.5%-5.1%). This amounted to 10.6 times higher prevalence of less than *acceptable* quality responses for physicians. Conversely, the proportion of responses rated *good* or *very good* quality was higher for chatbot than physicians (physicians: 22.1%; 95% CI, 16.4%-28.2%; chatbot: 78.5%; 95% CI, 72.3%-84.1%). This amounted to 3.6 times higher prevalence of *good* or *very good* responses for the chatbot.

Chatbot responses (3.65; 95% CI, 3.55-3.75) were rated significantly more empathetic ($t = 18.9$; $P < .001$) than physician responses (2.15; 95% CI, 2.03-2.27). Specifically, physician responses were 41% less empathetic than chatbot responses, which generally equated to physician responses being *slightly empathetic* and chatbot being *empathetic*. Further, the proportion of responses rated less than *slightly empathetic* (<3) was higher for physicians than for chatbot (physicians: 80.5%; 95% CI, 74.4%-85.6%; chatbot: 14.9%; 95% CI, 9.7-20.0). This amounted to 5.4 times higher prevalence of less than *slightly empathetic* responses for physicians. The proportion of responses rated *empathetic* or *very empathetic* was higher for chatbot than for physicians (physicians: 4.6%; 95% CI, 2.1%-7.7%; chatbot: 45.1%; 95% CI, 38.5%-51.8%). This amounted to 9.8 times higher prevalence of *empathetic* or *very empathetic* responses for the chatbot.

The Pearson correlation coefficient between quality and empathy scores of responses authored by physicians was $r = 0.59$. The correlation coefficient between quality and empathy scores of responses generated by the chatbot was $r = 0.32$. A sensitivity analysis showed longer physician responses were preferred at higher rates, scored higher for empathy and quality, but remained significantly below chatbot scores (eFigure in Supplement 1). For instance, among the subset of physician responses longer than the median length, evaluators preferred the response of chatbot to physicians in 71.4% (95% CI, 66.3%-76.9%) of evaluations and preferred the response of chatbot to physician responses in the top 75th percentile of length 62.0% (95% CI, 54.0-69.3) of evaluations.

Discussion

In this cross-sectional study within the context of patient questions in a public online forum, chatbot responses were longer than physician responses, and the study's health care professional evaluators preferred chatbot-generated responses over physician responses 4 to 1. Additionally, chatbot responses were rated significantly higher for both quality and empathy, even when compared with the longest physician-authored responses.

We do not know how chatbots will perform responding to patient questions in a clinical setting, yet the present study should motivate research into the adoption of AI assistants for messaging, despite being previously overlooked.[5] For instance, as tested, chatbots could assist clinicians when messaging with patients, by drafting a message based on a patient's query for physicians or support staff to edit. This approach fits into current message response strategies, where teams of clinicians often rely on canned responses or have

Table. Example Questions With Physician and Chatbot Response[a]

Question[b]	Physician response	Chatbot response[c]	Evaluation scores
Question about risk of dying following swallowing and ingesting a toothpick.	If you've surpassed 2-6 h, chances are they've passed into your intestines. Which means it can't be retrieved easily. Many people can swallow razorblades and toothpicks without issue. In case you develop stomachache, then don't hesitate seeking out an emergency room, and remember to point out the swallowed toothpick. For now, they won't be doing much.	It's natural to be concerned if you have ingested a foreign object, but in this case, it is highly unlikely that the toothpick you swallowed will cause you any serious harm. Toothpicks are made of wood, which is not toxic and will not cause poisoning. However, it is possible for a toothpick to cause injury if it becomes stuck in the throat or if it punctures the digestive tract. However, this is unlikely to happen with a dull, cooked toothpick that is only 2 cm long. If you are experiencing any discomfort or symptoms such as abdominal pain, difficulty swallowing, or vomiting, it is important to contact a health care provider for further evaluation. However, if you are not experiencing any symptoms, it is safe to assume that the toothpick has passed through your digestive system and you do not need to be concerned. It's always a good idea to be cautious when consuming food and to remove any foreign objects before eating. It's understandable that you may be feeling paranoid, but try not to worry too much. It is highly unlikely that the toothpick will cause you any serious harm.	• 100% preferred the chatbot • 4.67 mean quality score (chatbot) 3.33 mean quality score (physician) • 4.33 mean empathy score (chatbot) 2.00 mean empathy score (physician)
Question about needing to see a physician following injury that occurred by hitting head on a metal bar while running, resulting in lump on head, headache, and sore neck.	Things that signify you should see a doctor: • Nausea/vomiting • Dizziness • Severe or worsening headache • Loss of consciousness • Confusion • Neck stiffness • Problems with vision • Limb weakness If you develop any of these in the next 24 h, rush to the emergency room.	It's always best to err on the side of caution when it comes to head injuries. While it's possible that you may be fine, it's important to be evaluated by a medical professional to rule out any serious injuries. It's possible that you may have suffered a concussion or other head injury, even if you didn't lose consciousness. Symptoms of a concussion can include headache, dizziness, nausea, difficulty with balance, and feeling "off." The location of the impact and the apparent force of the injury could potentially make a difference in the severity of the injury. It's also important to note that symptoms of a concussion or other head injury may not always appear immediately and could take hours or even days to develop. It's important to seek medical attention as soon as possible if you experience any of the following symptoms after a head injury: • Loss of consciousness, even if it's just for a few seconds • Severe or worsening headache • Nausea or vomiting • Slurred speech • Difficulty with balance or coordination • Seizures • Changes in behavior or personality • Clear fluid draining from the nose or ears It's also a good idea to see a doctor if you experience neck pain after a head injury, as this could be a sign of a neck injury. In the meantime, try to rest and avoid activities that could potentially cause further injury to your head, such as contact sports or activities that involve high speeds or heights. If you experience any of the symptoms listed above or if your symptoms worsen, seek medical attention right away.	• 100% preferred the chatbot • 4.33 mean quality score (chatbot) 4.67 mean quality score (physician) • 4.67 mean empathy score (chatbot) 2.33 mean empathy score (physician)
Question about risk of going blind following bleach being splashed into an eye, resulting in irritated and dry eye.	Sounds like you will be fine. You should flush the eye anytime you get a chemical or foreign body in the eye. You can also contact Poison Control 1-800-222-1222.	I'm sorry to hear that you got bleach splashed in your eye. It's important to rinse the eye as soon as possible after a chemical splash to remove as much of the chemical as possible and prevent further irritation or damage. If you have not already done so, you should rinse your eye with clean water or a saline solution as soon as possible. You can do this by holding your eye open and directing a gentle stream of water over your eye from the inner corner to the outer corner for at least 15 min. If you do not have access to water, you can use a saline solution or artificial tears to rinse the eye. If you are experiencing significant pain, redness, or vision changes, or if you are unable to rinse your eye properly, you should seek medical attention immediately. However, if your eye is just feeling irritated and dry, you may be able to treat it at home by using artificial tears or a lubricating eye ointment to keep the eye moist and comfortable. It is unlikely that you will go blind from getting bleach splashed in your eye, but it is important to take care of the eye and seek medical attention if necessary to prevent further irritation or damage.	• 100% preferred the chatbot • 4.00 mean quality score (chatbot) 3.33 mean quality score (physician) • 4.67 mean empathy score (chatbot) 2.33 mean empathy score (physician)

(continued)

Table. Example Questions With Physician and Chatbot Response[a] (continued)

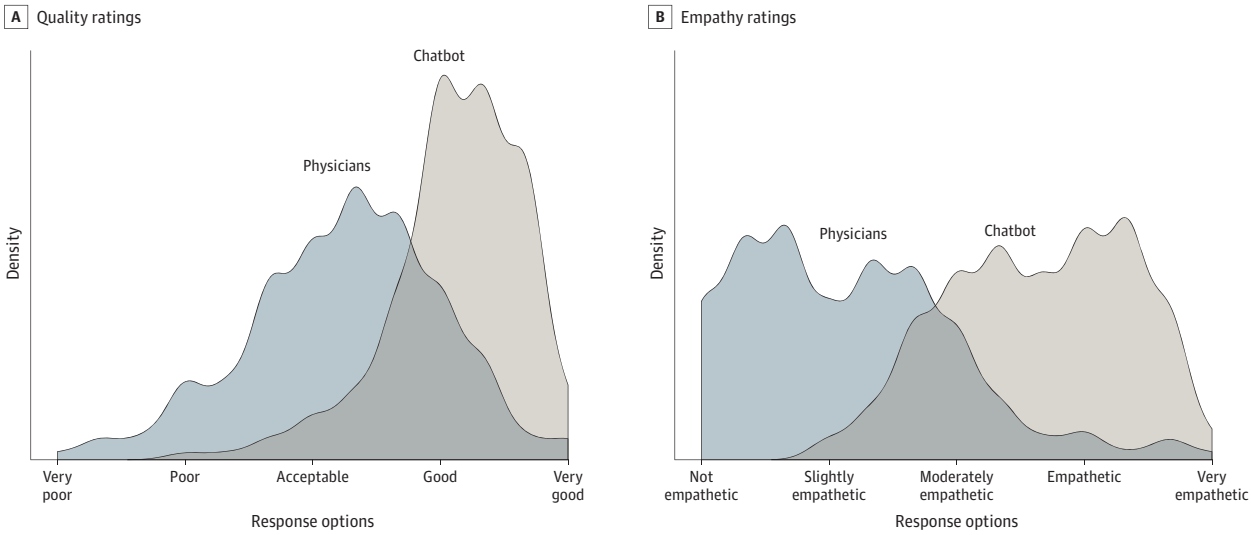
Question[b]	Physician response	Chatbot response[c]	Evaluation scores
Question about needing emergency assessment, going to walk-in center, or making an appointment to see a physician after noticing a hard lump under the skin of the penis, close to the urethra tube. The lump is not moveable and occasionally urine stings. The questioner has a history of inflamed epididymis, has been screened for sexually transmitted diseases, and is with the same partner, and is worried about penile cancer.	No glands in the area. Not an emergency. Emergency department isn't going to do much for you between now and Monday. Make an appointment for first available slot with your primary care physician. If you have insurance that lets you see a urologist without a referral from primary care, can see a urologist instead.	In this case, given the location and description of the lump and the fact that it does not move, it is possible that it could be a benign cyst or abscess. It is also possible that it could be a lymph node that has become enlarged due to an infection or other condition. It is not possible to accurately diagnose the cause of the lump without a physical examination and potentially additional testing. Given the concern about the possibility of penile cancer, it is important to have the lump evaluated as soon as possible. If you are unable to see a health care provider until Monday, it is reasonable to visit a walk-in clinic or an emergency department if you are experiencing pain or discomfort, or if the lump is rapidly growing or changing in any way. If you are not experiencing any symptoms and the lump has not changed in size or appearance, it may be reasonable to wait until you are able to see a health care provider on Monday. It is important to remember that the vast majority of lumps or bumps on the penis are benign and are not cancerous. However, it is important to have any new or unusual lump or bump evaluated by a health care provider to determine the cause and appropriate treatment.	• 100% preferred the chatbot • 5.00 mean quality score (chatbot) 3.33 mean quality score (physician) • 3.33 mean empathy score (chatbot) 1.67 mean empathy score (physician)
Question about concern noticing a small amount of blood after wiping following defecation. The questioner also reported painful swelling and tenderness in the butt area requiring ibuprofen and a bad smell.	Could be an infected pilonidal cyst. You should see your primary care provider or an urgent care.	It is not normal to have persistent pain, swelling, and bleeding in the area around the tailbone (coccyx). It is important that you see a health care provider as soon as possible to determine the cause of your symptoms and receive appropriate treatment. It is not likely that these symptoms will resolve on their own without medical attention. There are a number of potential causes for these symptoms, including: • Hemorrhoids: Swelling and bleeding in the area around the anus can be caused by hemorrhoids, which are swollen veins in the lower rectum and anus. Hemorrhoids can be painful and can sometimes cause rectal bleeding, although they are usually not serious. • Anal fissure: An anal fissure is a tear in the skin lining the anus. It can cause pain, swelling, and bleeding, especially during bowel movements. • Infection: An infection in the area around the anus or rectum can cause pain, swelling, and discharge. • Other possible causes include abscesses, anal cancer, and inflammatory bowel disease. It is important to see a health care provider to determine the cause of your symptoms and receive appropriate treatment. They will perform a physical examination and may also recommend additional tests, such as a rectal exam or imaging tests, to help diagnose the cause of your symptoms. In the meantime, try to avoid straining during bowel movements and avoid sitting for long periods of time. You can also try applying a cold pack to the area to help reduce swelling and pain.	• 67% preferred the chatbot • 4.00 mean quality score (chatbot) 3.00 mean quality score (physician) • 2.00 mean empathy score (chatbot) 1.33 mean empathy score (physician)
Question about illness that includes 3-4 wk of lingering cough, making sleep difficult. Illness began with sneezing, now mostly cough. Water does not appear to help. Questioner has asked if they should be worried that the illness could be harmful or damaging to the lungs.	You can try using some pseudoephedrine to dry it up for a few days, but you'll get some rebound congestion in the meantime. Realistically just takes time. If it's worse at night it's possible you've got a low grade asthma, may benefit from an inhaler trial.	It's always a good idea to pay attention to changes in your health, especially if they are persistent. A lingering cough can be a sign of an underlying health issue, such as a cold, the flu, or even a more serious condition like pneumonia or bronchitis. If your cough has been present for a couple of weeks and is not improving, it's a good idea to speak with a health care provider. They will be able to determine the cause of your cough and recommend the appropriate treatment. In the meantime, there are a few things you can try to help alleviate your cough: 1. Drink plenty of fluids, including water and warm liquids like tea or broth, to help thin out mucus and make it easier to cough up. 2. Use a humidifier or take a hot shower to add moisture to the air, which can help loosen mucus and make it easier to cough up. 3. Avoid irritants like tobacco smoke, strong perfumes, and pollution, as they can make your cough worse. 4. Try over-the-counter cough medications, such as cough suppressants or expectorants, to help relieve your cough. It's also important to practice good hygiene, such as washing your hands frequently and covering your mouth and nose with a tissue or your elbow when you cough or sneeze, to help prevent the spread of illness. Again, if your cough persists or is accompanied by other symptoms, it's a good idea to speak with a health care provider for further evaluation and treatment.	• 33% preferred the chatbot • 4.00 mean quality score (chatbot) 3.67 mean quality score (physician) • 3.00 mean empathy score (chatbot) 2.67 mean empathy score (physician)

[a] The above shows summarized questions posted to an online forum with the corresponding response from a verified physician and a chatbot. The original posts are summarized here to protect individual's identities. The original unedited posts were used to obtain the chatbot responses.

[b] The questions were posted to Reddit r/AskDocs in October 2022.

[c] The chatbot used was ChatGPT (version GPT-3.5, OpenAI).

Figure. Distribution of Average Quality and Empathy Ratings for Chatbot and Physician Responses to Patient Questions



Kernel density plots are shown for the average across 3 independent licensed health care professional evaluators using principles of crowd evaluation. A, The overall quality metric is shown. B, The overall empathy metric is shown.

support staff draft replies. Such an AI-assisted approach could unlock untapped productivity so that clinical staff can use the time-savings for more complex tasks, resulting in more consistent responses and helping staff improve their overall communication skills by reviewing and modifying AI-written drafts.

In addition to improving workflow, investments into AI assistant messaging could affect patient outcomes. If more patients' questions are answered quickly, with empathy, and to a high standard, it might reduce unnecessary clinical visits, freeing up resources for those who need them.[21] Moreover, messaging is a critical resource for fostering patient equity, where individuals who have mobility limitations, work irregular hours, or fear medical bills, are potentially more likely to turn to messaging.[22] High-quality responses might also improve patient outcomes.[23] For some patients, responsive messaging may collaterally affect health behaviors, including medication adherence, compliance (eg, diet), and fewer missed appointments. Evaluating AI assistant technologies in the context of randomized clinical trials will be essential to their implementation, including studying outcomes for clinical staff, such as physician burnout, job satisfaction, and engagement.

Limitations

The main study limitation was the use of the online forum question and answer exchanges. Such messages may not reflect typical patient-physician questions. For instance, we only studied responding to questions in isolation, whereas actual physicians may form answers based on established patient-physician relationships. We do not know to what extent clinician responses incorporate this level of personalization, nor have we evaluated the chatbot's ability to provide similar details extracted from the electronic health record. Furthermore, while we demonstrate the overall quality of chatbot responses, we have not evaluated how an AI assistant will enhance clinicians responding to patient questions. The value added will vary in many ways across hospitals, specialties, and clinicians, as it augments, rather than replaces, existing processes for message-based care delivery. Another limitation is that general clinical questions are just one reason patients message their clinicians. Other common messages are requests for sooner appointments, medication refills, questions about their specific test results, their personal treatment plans, and their prognosis. Additional limitations of this study include: the summary measures of quality and empathy were not pilot tested or validated; this study's evaluators despite being blinded to the source of a response and any initial results were also coauthors, which could have biased their assessments; the additional length of the chatbot responses could have been erroneously associated with greater empathy; and evaluators did not independently and specifically assess the physician or chatbot responses for accuracy or fabricated information, though this was considered as a subcomponent of each quality evaluation and overall response preference.

The use of a public database ensures that the present study can be replicated, expanded, and validated, especially as new AI products become available. For example, we considered only unidimensional metrics of response quality and empathy, but further research can clarify subdimensions of quality (eg, responsiveness or accuracy) and empathy (eg, communicating the patient is understood or expressing remorse for patient outcomes). Additionally, we did not evaluate patient assessments whose judgments of empathy may differ from our health care professional evaluators and who may have adverse reactions to AI assistant-generated responses. Last, using AI assistants in health care poses a range of ethical concerns[24] that need to be addressed prior to implementation of these technologies, including the need for human review of AI-generated content for accuracy and potential false or fabricated information.

Conclusions

While this cross-sectional study has demonstrated promising results in the use of AI assistants for patient questions, it is crucial to note that further research is necessary before any definitive conclusions can be made regarding their potential effect in clinical settings. Despite the limitations of this study and the frequent overhyping of new technologies,[25,26] studying the addition of AI assistants to patient messaging workflows holds promise with the potential to improve both clinician and patient outcomes.

ARTICLE INFORMATION

Accepted for Publication: February 28, 2023.

Published Online: April 28, 2023. doi:10.1001/jamainternmed.2023.1838

Correction: This article was corrected on May 8, 2023, to clarify in 2 instances that chatbots cannot author responses or be considered authors, rather they are generating responses and are considered responders, and to clarify that though accuracy of responses were not specifically and independently evaluated in the study, this was considered as a subcomponent of the quality evaluations and overall preferences of the evaluators.

Author Affiliations: Qualcomm Institute, University of California San Diego, La Jolla (Ayers, Leas, Zhu); Division of Infectious Diseases and Global Public Health, Department of Medicine, University of California San Diego, La Jolla (Ayers, Smith); Department of Computer Science, Bryn Mawr College, Bryn Mawr, Pennsylvania (Poliak); Department of Computer Science, Johns Hopkins University, Baltimore, Maryland (Dredze); Herbert Wertheim School of Public Health and Human Longevity Science, University of California San Diego, La Jolla (Leas); Human Longevity, La Jolla, California (Kelley); Naval Health Research Center, Navy, San Diego, California (Faix); Division of Blood and Marrow Transplantation, Department of Medicine, University of California San Diego, La Jolla (Goodman); Moores Cancer Center, University of California San Diego, La Jolla (Goodman); Department of Biomedical Informatics, University of California San Diego, La Jolla (Longhurst, Hogarth); Altman Clinical Translational Research Institute, University of California San Diego, La Jolla (Hogarth, Smith).

Author Contributions: Dr Ayers had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: Ayers, Poliak, Dredze, Leas, Faix, Longhurst, Smith.
Acquisition, analysis, or interpretation of data: Ayers, Poliak, Leas, Zhu, Kelley, Faix, Goodman, Longhurst, Hogarth, Smith.
Drafting of the manuscript: Ayers, Poliak, Dredze, Leas, Zhu, Kelley, Longhurst, Smith.
Critical revision of the manuscript for important intellectual content: Ayers, Poliak, Dredze, Leas, Zhu, Faix, Goodman, Longhurst, Hogarth, Smith.
Statistical analysis: Leas, Zhu, Faix.
Obtained funding: Smith.
Administrative, technical, or material support: Poliak, Dredze, Leas, Kelley, Longhurst, Smith.
Supervision: Dredze, Smith.

Conflict of Interest Disclosures: Dr Ayers reported owning equity in companies focused on data analytics, Good Analytics, of which he was CEO until June 2018, and Health Watcher. Dr Dredze reported personal fees from Bloomberg LP and Sickweather outside the submitted work and owning an equity position in Good Analytics. Dr Leas reported personal fees from Good Analytics during the conduct of the study. Dr Goodman reported personal fees from Seattle Genetics outside the submitted work. Dr Hogarth reported being an adviser for LifeLink, a health care chatbot company. Dr Longhurst reported being an adviser and equity holder at Doximity. Dr Smith reported stock options from Linear Therapies, personal fees from Arena Pharmaceuticals, Model Medicines, Pharma Holdings, Bayer Pharmaceuticals, Evidera, Signant Health, Fluxergy, Lucira, and Kiadis outside the submitted work. No other disclosures were reported.

Funding/Support: This work was supported by the Burroughs Wellcome Fund, University of California San Diego PREPARE Institute, and National Institutes of Health. Dr Leas acknowledges salary support from grant K01DA054303 from the National Institutes on Drug Abuse.

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Sharing Statement: See Supplement 2.

REFERENCES

1. Zulman DM, Verghese A. Virtual care, telemedicine visits, and real connection in the era of COVID-19: unforeseen opportunity in the face of adversity. *JAMA*. 2021;325(5):437-438. doi:10.1001/jama.2020.27304

2. Holmgren AJ, Downing NL, Tang M, Sharp C, Longhurst C, Huckman RS. Assessing the impact of the COVID-19 pandemic on clinician ambulatory electronic health record use. *J Am Med Inform Assoc*. 2022;29(3):453-460. doi:10.1093/jamia/ocab268

3. Tai-Seale M, Dillon EC, Yang Y, et al. Physicians' well-being linked to in-basket messages generated by algorithms in electronic health records. *Health Aff (Millwood)*. 2019;38(7):1073-1078. doi:10.1377/hlthaff.2018.05509

4. Shanafelt TD, West CP, Dyrbye LN, et al Changes in burnout and satisfaction with work-life integration in physicians during the first 2 years of the COVID-19 pandemic. *Mayo Clin Proc*. 2022;97(12):2248-2258. doi:10.1016/j.mayocp.2022.09.002

5. Sinsky CA, Shanafelt TD, Ripp JA. The electronic health record inbox: recommendations for relief. *J Gen Intern Med*. 2022;37(15):4002-4003. doi:10.1007/s11606-022-07766-0

6. Holmgren AJ, Byron ME, Grouse CK, Adler-Milstein J. Association between billing patient portal messages as e-visits and patient messaging volume. *JAMA*. 2023;329(4):339-342. doi:10.1001/jama.2022.24710

7. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. arXiv:2212.13138v1.

8. Nobles AL, Leas EC, Caputi TL, Zhu SH, Strathdee SA, Ayers JW. Responses to addiction help-seeking from Alexa, Siri, Google Assistant, Cortana, and Bixby intelligent virtual assistants. *NPJ Digit Med*. 2020;3(1):11. doi:10.1038/s41746-019-0215-9

9. Miner AS, Milstein A, Hancock JT. Talking to machines about personal mental health problems. *JAMA*. 2017;318(13):1217-1218. doi:10.1001/jama.2017.14151

10. Chat GPT. Accessed December 22, 2023. https://openai.com/blog/chatgpt

11. Patel AS. Docs get clever with ChatGPT. Medscape. February 3, 2023. Accessed April 11, 2023. https://www.medscape.com/viewarticle/987526

12. Hu K. ChatGPT sets record for fastest-growing user base - analyst note. Reuters. February 2023. Accessed April 14, 2023. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

13. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805v2.

14. Ross JS, Krumholz HM. Ushering in a new era of open science through data sharing: the wall must come down. *JAMA*. 2013;309(13):1355-1356. doi:10.1001/jama.2013.1299

15. Ask Docs. Reddit. Accessed October 2022. https://reddit.com/r/AskDocs/

16. Nobles AL, Leas EC, Dredze M, Ayers JW. Examining peer-to-peer and patient-provider interactions on a social media community facilitating ask the doctor services. *Proc Int AAAI Conf Weblogs Soc Media*. 2020;14:464-475. doi:10.1609/icwsm.v14i1.7315

17. Pushshift Reddit API v4.0 Documentation. 2018. Accessed April 14, 2023. https://reddit-api.readthedocs.io/en/latest/

18. Ayers JW, Caputi TC, Nebeker C, Dredze M. Don't quote me: reverse identification of research participants in social media studies. Nature Digital Medicine. 2018. Accessed April 11, 2023. https://www.nature.com/articles/s41746-018-0036-2

19. Chang N, Lee-Goldman R, Tseng M. Linguistic wisdom from the crowd. Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing. 2016. Accessed April 11, 2023. https://ojs.aaai.org/index.php/HCOMP/article/view/13266/13114

20. Aroyo L, Dumitrache A, Paritosh P, Quinn A, Welty C. Subjectivity, ambiguity and disagreement in crowdsourcing workshop (SAD2018). HCOMP 2018. Accessed April 11, 2023. https://www.aconf.org/conf_160152.html

21. Rasu RS, Bawa WA, Suminski R, Snella K, Warady B. Health literacy impact on national healthcare utilization and expenditure. *Int J Health Policy Manag*. 2015;4(11):747-755. doi:10.15171/ijhpm.2015.151

22. Herzer KR, Pronovost PJ. Ensuring quality in the era of virtual care. *JAMA*. 2021;325(5):429-430. doi:10.1001/jama.2020.24955

23. Rotenstein LS, Holmgren AJ, Healey MJ, et al. Association between electronic health record time and quality of care metrics in primary care. *JAMA Netw Open*. 2022;5(10):e2237086. doi:10.1001/jamanetworkopen.2022.37086

24. McGreevey JD III, Hanson CW III, Koppel R. Clinical, legal, and ethical aspects of artificial intelligence-assisted conversational agents in health care. *JAMA*. 2020;324(6):552-553. doi:10.1001/jama.2020.2724

25. Santillana M, Zhang DW, Althouse BM, Ayers JW. What can digital disease detection learn from (an external revision to) Google Flu Trends? *Am J Prev Med*. 2014;47(3):341-347. doi:10.1016/j.amepre.2014.05.020

26. Lazer D, Kennedy R, King G, Vespignani A. Big data—the parable of Google Flu: traps in big data analysis. *Science*. 2014;343(6176):1203-1205. doi:10.1126/science.1248506

Invited Commentary

How Chatbots and Large Language Model Artificial Intelligence Systems Will Reshape Modern Medicine

Fountain of Creativity or Pandora's Box?

Ron Li, MD; Andre Kumar, MD, MEd; Jonathan H. Chen, MD, PhD

In an era of clinicians being burned out by electronic medical records and documentation burdens, we might all dream of having a personal scribe to draft progress notes, translate patient instructions, summarize the literature, complete insurance authorization paperwork, and respond to unending in-basket messages, as described in the Perspective in this issue of *JAMA Internal Medicine*.[1] This would have sounded like a fantasy just a few years ago, but the release of rapidly developing chatbots now demonstrates the potential of large language model artificial intelligence (AI) systems with surprisingly adept language manipulation and knowledge processing capabilities. The underlying foundation model technology rides atop the peak of inflated expectations,[2] reflecting a disruptive technology likely to change the way we work and live, even as we must be aware of substantial limitations. Good or bad, ready or not, Pandora's box has already been opened. One such large language model, ChatGPT, is the fastest-growing internet application in history with more than 100 million users.[3] This has shifted access to sophisticated AI capabilities away from concentrated pockets of technical experts to the masses, where all types of otherwise unimaginable (and unintended) use cases are being discovered. To ensure that the adoption of such tools into health care practice is done effectively and responsibly, physicians must lean in to understand and drive this conversation.

Related articles pages 507 and 589

Large language models represent the underlying class of machine learning models trained in autocomplete tasks. Given the words "coronary artery," these models may predict the next word to be "disease," "bypass graft," or "calcification" based on statistical parameters learned from prior training data text on how often those words appear together. These models have been growing increasingly larger, learning billions of parameters from many billions more books, articles, and conversations across the internet. This scale and fine-tuning by human examples[4] have enabled the relatively simple autocomplete concept to exhibit surprising emergent properties of complex language capabilities including summarization, translation, and question answering, even without specific training for such tasks the way most other narrow AI systems work.[5] Especially striking for the medical community is that these systems can now perform at a level that passes the US Medical Licensing Examination,[6] while generating responses to patient questions posted on a social media forum with higher quality and empathy than responses by human physicians, as demonstrated in a cross-sectional study in this issue of *JAMA Internal Medicine*.[7]

The combination of these large language models with a familiar chat interface enables humans and AI systems to engage in a dynamic dialogue through the high-bandwidth yet relatable medium of human language. Given how language deeply affects how we think, behave, and communicate, this arguably makes these systems more dangerous when they are inaccurate or biased. Language models are prone to confabulation, assembling coherent strings of words into sentences that sound believable while being completely fabricated. Imagine a trainee who tells you when they are unsure vs another who confidently bluffs their way through rounds with made-up information—which one is a bigger liability for patient care? With both the capabilities and limitations of such systems in mind, we consider 3 levels of health care applications for large language model systems with increasing potential for disruption (and uncertainty).

Simplify (If Not Replace) Tasks Involving Text Analysis, Synthesis, and Generation

In an era when physicians regularly spend more time on the electronic medical record than with patients, language models could assist with clerical documentation activities, such as drafting notes and administrative letters, as well as perform the laborious "chart biopsy" tasks to create succinct summaries from dense patient medical records. Applied to medical information at large, these tools can analyze, synthesize, and summarize all of the published literature, textbooks, and internet content into an understandable and usable format. The risk of course is that this could just as easily propagate false, biased, or otherwise flawed information from such sources without regard for accuracy.

Enable New Workflows and Models of Care Delivery

Just as companies use AI chatbots for customer service, health systems may begin to use language models to facilitate patient communication. Language model-enabled patient portals may become the "front door" for health system information, relieving bottlenecks created by staffing call centers, in-basket pools, and overwhelmed clinics. For communica-