



# HHS Public Access

Author manuscript

*JVS Vasc Insights*. Author manuscript; available in PMC 2023 September 12.

Published in final edited form as:

*JVS Vasc Insights*. 2023 ; 1: . doi:10.1016/j.jvsvi.2023.100019.

## The potential of chatbots in chronic venous disease patient management

Anand Athavale, MD<sup>a</sup>, Jonathan Baier, BS<sup>b</sup>, Elsie Ross, MD, MSc<sup>a</sup>, Eri Fukaya, MD, PhD<sup>a</sup>

<sup>a</sup>Division of Vascular Surgery, Stanford University School of Medicine, Palo Alto

<sup>b</sup>NextNext LLC, Lovettsville.

### Abstract

**Objective:** Health care providers and recipients have been using artificial intelligence and its subfields, such as natural language processing and machine learning technologies, in the form of search engines to obtain medical information for some time now. Although a search engine returns a ranked list of webpages in response to a query and allows the user to obtain information from those links directly, ChatGPT has elevated the interface between humans with artificial intelligence by attempting to provide relevant information in a human-like textual conversation. This technology is being adopted rapidly and has enormous potential to impact various aspects of health care, including patient education, research, scientific writing, pre-visit/post-visit queries, documentation assistance, and more. The objective of this study is to assess whether chatbots could assist with answering patient questions and electronic health record inbox management.

**Methods:** We devised two questionnaires: (1) administrative and non-complex medical questions (based on actual inbox questions); and (2) complex medical questions on the topic of chronic venous disease. We graded the performance of publicly available chatbots regarding their potential to assist with electronic health record inbox management. The study was graded by an internist and a vascular medicine specialist independently.

**Results:** On administrative and non-complex medical questions, ChatGPT 4.0 performed better than ChatGPT 3.5. ChatGPT 4.0 received a grade of 1 on all the questions: 20 of 20 (100%). ChatGPT 3.5 received a grade of 1 on 14 of 20 questions (70%), grade 2 on 4 of 16 questions (20%), grade 3 on 0 questions (0%), and grade 4 on 2/20 questions (10%). On complex medical

---

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Correspondence: Eri Fukaya, MD, PhD, Division of Vascular Surgery, Stanford University School of Medicine, 780 Welch Road, Ste CJ350, Palo Alto, CA 94304 (efukaya@stanford.edu).

#### AUTHOR CONTRIBUTIONS

Conception and design: AA, JB, EF

Analysis and interpretation: AA, JB, ER, EF

Data collection: AA, JB, ER, EF

Writing the article: AA, JB, ER, EF

Critical revision of the article: AA, JB, ER, EF

Final approval of the article: AA, JB, ER, EF

Statistical analysis: Not applicable

Obtained funding: Not applicable

Overall responsibility: AA

Author conflict of interest: none.

The editors and reviewers of this article have no relevant financial relationships to disclose per the Journal policy that requires reviewers to decline review of any manuscript for which they may have a conflict of interest.

questions, ChatGPT 4.0 performed the best. ChatGPT 4.0 received a grade of 1 on 15 of 20 questions (75%), grade 2 on 2 of 20 questions (10%), grade 3 on 2 of 20 questions (10%), and grade 4 on 1 of 20 questions (5%). ChatGPT 3.5 received a grade of 1 on 9 of 20 questions (45%), grade 2 on 4 of 20 questions (20%), grade 3 on 4 of 20 questions (20%), and grade 4 on 3 of 20 questions (15%). Clinical Camel received a grade of 1 on 0 of 20 questions (0%), grade 2 on 5 of 20 questions (25%), grade 3 on 5 of 20 questions (25%), and grade 4 on 10 of 20 questions (50%).

**Conclusions:** Based on our interactions with ChatGPT regarding the topic of chronic venous disease, it is plausible that in the future, this technology may be used to assist with electronic health record inbox management and offload medical staff. However, for this technology to receive regulatory approval to be used for that purpose, it will require extensive supervised training by subject experts, have guardrails to prevent “hallucinations” and maintain confidentiality, and prove that it can perform at a level comparable to (if not better than) humans. (*JVS-Vascular Insights* 2023;1:100019.)

### Keywords

Artificial intelligence; ChatGPT 3.5; ChatGPT 4.0; Chronic venous disease; Clinical Camel; Electronic health record inbox management; Generative AI

---

The world is experiencing a revolutionary technology of generative artificial intelligence (AI), where we are awed by the extraordinary andromorphic responses provided by large language models (LLMs) such as ChatGPT. Over the next decade, it is inevitable that generative AI will disrupt what we have believed to be the norm in almost everything, including our practices in medicine. Although we hope to take advantage of this extraordinary technology, we must also realize that there is an urgent need for discussions and regulations around this to provide safe and sound clinical care. We want to reap the benefits while mitigating risk, and to do so, we must start by thinking and having conversations about this. This includes considering our core philosophy about how we provide care, and how we should constantly update our knowledge of this technology so that we can adapt to and face a future where we make this incredible technology work for us.

Generative AI is a set of algorithms that can generate realistic content, including text, images, music, and video from data that this is trained on. Within this framework, a chatbot is a computer program designed to simulate human-like conversations in the language of the user (natural language). ChatGPT is a sophisticated chatbot with eloquent and intelligent “conversational skills,” which appears to understand the context of a conversation, can mimic celebrities, write essays, poems, and even have philosophical discussions. This is done by using natural language processing, a subfield of AI, to combine the fields of linguistics and computer science to understand, interpret, and generate human-like responses. Natural language processing uses machine learning, LLM, and reinforcement learning with human feedback to develop this capability. Machine learning algorithms can identify patterns, relationships, and structures within the language data by processing large amounts of data. LLMs learn to predict the next word in a sequence based on the context of the previous words. The models assign probabilities to different word combinations and refine their predictions as they process more data. In doing so, these models are able to recreate human-like responses to prompts.

Electronic health record (EHR) inbox management has been reported to contribute to physician dissatisfaction and burnout.<sup>1,2</sup> These are patient portal messages and messages from other clinicians and staff, the pharmacy, laboratory, and other medical departments. It is often the recommended practice to reply to messages within two business days, which leads to much of this work being done by the physician outside of scheduled work hours. The amount of time spent answering these messages has increased and led to high cynicism and emotional exhaustion for care providers.<sup>1</sup> Realizing the convenience of sending messages with no additional fee for its use, patients have a low threshold to utilize this. Although some may require complex medical decision-making or personalized responses, many of these are standard frequently asked questions. However, its response can take up a significant amount of personnel resources, whether this is by a nurse, advanced practitioner, or physician. The workload is often inadequately captured and is a cause of significant exhaustion for the clinical providers.

Microsoft Corporation and Epic have recently announced their collaboration to integrate generative AI into EHR to improve productivity and decrease the administrative burden. One of the conceivable use cases for an AI-powered chatbot is EHR inbox management to decrease the burden of clinical providers that are currently handling this. In this article, we delve into the basics of ChatGPT and its processes, medical use cases with caveats, and the future landscape in medicine for such programs. We test the performance of various chatbots on administrative, non-complex medical and complex medical questions.

## METHODS

We devised two questionnaires. The first set included 20 questions on non-complex medical and administrative matters (Table I). These questions were based on actual patient messages. The second set included 20 complex medical questions requiring subject matter expertise in chronic venous disease (Table II). The questions were classified as non-complex if a medical assistant or a nurse would be expected to provide an adequate response and complex if the question would have to be escalated to the physician/subject matter expert. The questions were written by two physicians and one person without a medical background. The study was graded by an internist and a vascular medicine specialist independently. In case of a disagreement, the grade awarded by the vascular medicine specialist took precedence. The two versions of general purpose chatbots, ChatGPT 3.5 and ChatGPT 4.0, were asked administrative and non-complex medical questions. Complex medical questions were asked of the two versions of ChatGPT and the research preview version of a health care-focused chatbot, Clinical Camel, which is an ongoing project by the WangLab at the University of Toronto for developing an open-source health care-focused chatbot based on Large Language Model Meta AI (LLaMA), designed by Meta. The responses were graded as: (1) Appropriate and complete; (2) Appropriate but incomplete; (3) Neither appropriate nor inappropriate; and (4) Wrong/inappropriate.

## RESULTS

On administrative and non-complex medical questions, ChatGPT 4.0 performed better than ChatGPT 3.5. ChatGPT 4.0 received a grade of 1 on all the questions: 20 of 20 (100%).

ChatGPT 3.5 received a grade of 1 on 14 of 20 (70%), grade 2 on 4 of 16 (20%), grade 3 on 0 (0%), and grade 4 on 2/20 (10%) questions.

On complex medical questions, ChatGPT 4.0 performed the best. ChatGPT 4.0 received a grade of 1 on 15 of 20 (75%), grade 2 on 2 of 20 (10%), grade 3 on 2 of 20 (10%), and grade 4 on 1 of 20 (5%) questions. ChatGPT 3.5 received a grade of 1 on 9 of 20 (45%), grade 2 on 4 of 20 (20%), grade 3 on 4 of 20 (20%), and grade 4 on 3 of 20 (15%) questions. Clinical Camel received a grade of 1 on 0 of 20 (0%), grade 2 on 5 of 20 (25%), grade 3 on 5 of 20 (25%), and grade 4 on 10 of 20 (50%) questions. Chatbot responses to the questions and the awarded grades are available in Supplementary Tables 1–5 (online only).

## DISCUSSION

The objective of this experimental observation was to understand the basics of ChatGPT and its processes, as well as its potential medical use cases and the future landscape in medicine for such programs. Regarding the answers provided, overall, ChatGPT 4.0 performed the best on administrative, non-complex medical and complex medical questions. It is notable that the general-purpose chatbot performed better than Clinical Camel, which is a health care-focused chatbot that was trained on a “mixture of user-shared conversations and synthetic conversations designed to encode highquality clinical data from curated clinical articles.” The models of the chatbots studied are not publicly disclosed; hence, we cannot determine the specific differences. We surmise that the differences in performance may result from better unsupervised and supervised training algorithms and a larger user base providing feedback for ChatGPT. Our results indicated that the fine-tuning of the model is critically important and may have a larger impact on its overall performance than how the chatbot is specialized. Once this technology can be made Health Insurance Portability and Accountability Act-compliant, has supervised training by subject experts, and has guard rails against fabricated information, its integration with EHR can potentially assist with task such as inbox management.

Additionally, access to the medical literature, including those beyond paywalls, can make it an effective tool to include information based on literature review. ChatGPT has a pre-training non-supervised phase and a fine-tuning phase. In the non-supervised phase, a model analyzes the input data and tries to learn patterns and structures from input data without any labels. It discovers patterns and relationships in data on its own. The developers feed this model massive amounts of information. LLMs owe much of their language aptitude to generative pre-trained transformers (GPT). Transformers not only track relationships between words in a sentence but also store which portion of the sentence requires more attention. This “attention” is a proxy for storing context and thus producing better results. This model takes a piece of text as input and represents each word or token as a unique numerical vector. These vectors capture the meaning and context of the words in the text. It then uses a feature called the “attention mechanism” to determine how important each word is in relation to others in the text. By analyzing the relationships between words and their context, the transformer model learns to understand language patterns and attempts to predict the most likely next word or phrase in a specific context to generate responses. Hence the name Chat Generative Pre-training Transformer (ChatGPT). This process can be

scaled as it is not limited by the human ability to label data sets or anticipate prompts. The initial foundation (GPT 3) is trained on a large collection of data from various web sources, including general web crawling (largest weight), Reddit outbound links, online book collections, and Wikipedia. It is important to note that additional data is used for the reinforcement step discussed below, but that data and the foundational data from GPT 4 is not publicly disclosed.

Upon completion of pre-training in the non-supervised phase, the model is refined with reinforcement learning with human feedback in the supervised phase. Humans rank responses from the model at this point and create a reward model to improve responses further. This reward model will continue a loop as more responses are generated. During this phase, the model is also trained on a paired input-output set called a “labeled dataset.” For example, the model can be trained for the input “What is CEAP 4 class disease?” The appropriate response is “Changes in skin and subcutaneous tissue secondary to chronic venous disease.” Although this process allows for tight quality control, it requires subject matter expertise and large labeled datasets. There is a limit to questions that human trainers can anticipate in a finite amount of time. Hence, this cannot be scaled to generate responses to any conceivable prompt. These two training phases function in tandem to provide convincing human-like conversations.

Although GPT 4.0 performed well on both tasks, it is essential to understand that it is a prediction model and does not necessarily understand medical concepts. Thus, the responses may reflect the inherent biases of the data it was trained on, or the biases introduced with human feedback. It is also very difficult to find the sources of information it uses and figure out how the algorithm processes the data.

An obvious use case for this chatbot is as a search engine by health care providers or recipients to obtain medical information and in medical education. Although they appear similar, it is essential to distinguish the “generative chatbot” from a “search engine,” as a chatbot has been trained to “reason.” In the best-case scenario, ChatGPT can provide accurate and thorough information in various languages and at varying degrees of complexity. This has the potential to disseminate critical information in a manner easily understood by the user and help bridge some of the communication gaps that can reduce the quality of health care delivery. ChatGPT 3.5 and 4.0 were trained on the text that was freely available on the internet until September 2021, and the training data did not include information beyond paywalls. It is conceivable that the future versions will have better access to information and that the creators of a chatbot can enter into an arrangement with publishers that enables access to scientific papers and use this for training data. It may then be prompted to summarize published research to the user in the language and level of complexity of choice. As things stand now, a lack of training on recent data and a lack of access to published data beyond paywalls is a significant limitation in its use to obtain accurate and relevant medical information. ChatGPT generates text in response to a prompt without a way to determine the source, as the links are often outdated or dysfunctional (ChatGPT 3.5) or inability to provide links (ChatGPT 4.0). Therefore, the veracity of the generated text cannot be ascertained.

It is unclear how ChatGPT handles conflicting information, as the parent company, OpenAI, has not published it publicly. When ChatGPT 3.5 was asked about how it handles conflicting information, it responded that it assessed the credibility of sources by analyzing their reputation and expertise in a field as determined by the author's credentials, previous work, and the quality of the publishing platform. It stated that it also cross-referenced a source with other sources to see if they corroborated or contradicted each other; if multiple sources with high credibility and expertise agreed on a particular fact, ChatGPT was more likely to consider that fact as accurate. This response is possibly a "hallucination." ChatGPT 4.0 responded that it had no mechanism to rank the credibility of sources in real-time. Given a plethora of information ranging from accurate, outdated, partially correct information presented with a high degree of confidence and repetition, personal opinions, untested hypotheses, alternative medicine, and deliberate misinformation, it is essential to have a transparent and objective method of assessing credibility with reproducible results.

Another significant limitation is the propensity for "hallucinations." It refers to generating syntactically correct but factually incorrect responses that seem plausible. The following example is an interaction with ChatGPT 3.5 quoted verbatim.

“Q: Is congenital heart disease related to chronic venous disease? Provide citations and links

A: Congenital heart disease (CHD) and chronic venous disease (CVD) are two distinct medical conditions that affect different parts of the cardiovascular system. CHD is a condition in which the heart and blood vessels that supply the heart are malformed or damaged from birth, while CVD is a condition that affects the veins and blood flow in the legs.

There is no direct relationship between CHD and CVD. However, some studies suggest that there may be an association between CHD and an increased risk of CVD. A study published in the Journal of the American Heart Association found that adults with CHD had a higher risk of developing CVD than those without CHD.<sup>i</sup> Another study published in the European Journal of Preventive Cardiology found that adults with CHD had a higher prevalence of CVD risk factors, such as obesity and diabetes, which may contribute to the increased risk of CVD.<sup>ii</sup>

It is important to note that these studies do not establish a causal relationship between CHD and CVD. Further research is needed to better understand the potential link between these two conditions”.

In the above example, the articles referenced to support the association between CHD and CVD are purely fictional; however, they are presented plausibly. On the other hand, for the same question, ChatGPT 4.0 did not fabricate an association.

However, these limitations will not prevent its use in its current form. Even for the most technology-reclusive person, a reference to a ChatGPT result seems inevitable, be it in the context of a classroom, clinic, literature search/ writing, or a research meeting.

Another possible area of concern is privacy. A “bug” in ChatGPT apparently allowed some users to view the titles of others' conversations. This bug was claimed to have been fixed

by the company. However, the risk of a data breach remains. Additionally, there should be explicit clarity regarding the monetization policy of the data generated by user prompts.

Once privacy-related issues have been resolved, a chatbot like ChatGPT can assist with writing and screening medical records and provide real-time suggestions to make medical records more accurate and easily interpretable by coders. Stonko et al have suggested that LLMs may be able to integrate data from a patient's prior medical and surgical history, algorithms that have diagnosed and triaged the patient and technology used for intraoperative guidance to draft an operative note.<sup>3</sup>

ChatGPT can be used in literature searches with the abovementioned limitations (time cutoff, access to information, hallucinations). It can also assist with certain aspects of scientific writing; however, its use for scientific writing has been controversial. The use has varied from article outline/structure creation, editing, and proofreading, to having ChatGPT write the entire article. This raises issues about accountability, plagiarism, accuracy, and reproducibility.

Various publishing houses have different policies about the use of ChatGPT in scientific writing.

Providing medical care to a patient requires a complex interplay of acquiring: (1) administrative information; (2) clinical information and history; (3) diagnostic testing and imaging; (4) assessment of the patient; (5) medical decision-making; (6) creating a treatment plan and its execution; (7) providing patient education; and (8) coordination of care. With the growing capacity of generative AI, we may see that some of these tasks may be taken over. At some point, AI may have more data points and knowledge for decision-making than the human provider. If so, can the algorithms developed in AI be much better than one that comes from the human brain? Will there be a future where the AIs may do the medical decision-making part, and the providers will be executing on that? Medicine is both a science and an art. The art of medicine lies in human skills and our ability to evaluate things in context. This includes the ability to read the patient's facial expression and body language and understand their cultural/religious/spiritual background to connect with the patient to have conversations that draw out valuable information. This leads to the ability to have empathy, establish relationships, alleviate suffering, and share emotions, resulting in the delivery of humane care. It may be more of these skill sets that cannot be readily taken over by AI.

## CONCLUSION

Based on our interactions with ChatGPT regarding the topic of chronic venous disease, it is plausible that in the future, this technology may be used to assist with EHR inbox management and offload medical staff. However, for this technology to receive regulatory approval to be used for that purpose, it will require extensive supervised training by subject experts, have guardrails to prevent hallucinations and maintain confidentiality, and prove that it can perform at a level comparable to (if not better than) humans.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

### REFERENCES

1. Adler-Milstein J, Zhao W, Willard-Grace R, Knox M, Grumbach K. Electronic health records and burnout: time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. *J Am Med Inform Assoc* 2020;27:531–8. [PubMed: 32016375]
2. Gardner RL, Cooper E, Haskell J, et al. Physician stress and burnout: the impact of health information technology. *J Am Med Inform Assoc* 2019;26:106–14. [PubMed: 30517663]
3. Stonko DP, Morrison JJ, Hicks CW. A review of mature machine learning and artificial intelligence enabled Applications in Aortic Surgery. *JVS-Vascular Insights* 2023:100016.

### REFERENCES

- i. Mondellini G, Cervellin G, Maffeo D, Lippi G. Increased risk of chronic venous disease in adults with congenital heart disease: a population-based study. *J Am Heart Assoc* 2019;8:e012481.
- ii. Waldmann V, Jünger C, Katalinic A, Knirsch W, Apitz C, Michel-Behnke I. Prevalence and risk factors for chronic venous disease in adults with congenital heart disease. *Eur J Prev Cardiol* 2017;24:1178–84.



**ARTICLE HIGHLIGHTS**

- **Type of Research:** Practice management
- **Key Findings:** ChatGPT 4.0 answered administrative and non-complex medical questions well, and this technology has the potential to assist in various tasks including electronic health record inbox management. ChatGPT 4.0 performed better than other chatbots in answering complex medical questions.
- **Take Home Message:** ChatGPT generally performs well in answering administrative, non-complex medical and complex medical questions. With supervised training by subject experts, safeguards for privacy, and guardrails against “hallucination,” this technology can integrate with electronic health records to assist with patient communication and education.

**Table I.****Administrative and non-complex medical questions**

- 
- 1) I developed a wound on my foot yesterday by stepping on a rock, I do not have a fever, what should I do?
  - 2) I have an old wound on my foot, yesterday I developed a fever. What should I do?
  - 3) I need to set up an appointment with my doctor, should I contact the office front desk or the doctor?
  - 4) I need to re-schedule my doctor's appointment, who should I call?
  - 5) My medications are running out, what should I do?
  - 6) How long after endovenous thermal ablation can I take a shower?
  - 7) How long after endovenous thermal ablation do I need to come for follow-up?
  - 8) What is the diagnostic code for varicose vein condition?
  - 9) I have Blue Cross Blue Shield insurance. Does it cover Rivaroxaban?
  - 10) I have a question about my DVT. Is walking encouraged or discouraged? Can I walk gently on a treadmill?
  - 11) What can cause leg swelling?
  - 12) How did I get varicose veins?
  - 13) How do I prevent my varicose veins from becoming larger?
  - 14) In my labs, the platelet count is slightly high, what does it mean?
  - 15) Can you draft a letter on behalf of Dr Smith for Mr Wilson addressed to his employer that he was in the clinic and missed work in the morning because of this appointment?
  - 16) Should I hold my aspirin before endovenous thermal ablation?
  - 17) What should I do with my swollen leg?
  - 18) What is venous reflux?
  - 19) My leg hurts after sclerotherapy. Is this normal?
  - 20) Does insurance cover my vein procedure?
- 

*DVT*, Deep vein thrombosis.

**Table II.**

## Complex medical questions

- 
- 1) When do you need to have a venous ablation for varicose veins?
  - 2) Is high ligation and stripping surgery better than endovenous thermal ablation for the treatment of chronic venous insufficiency?
  - 3) How do you heal a venous leg ulcer?
  - 4) Can you do a venous ablation if you have deep vein thrombosis?
  - 5) I have a venous leg ulcer. What is my CEAP?
  - 6) My venous leg ulcer healed, but I have pigmentation in my leg with swelling. I do not have venous reflux. What is my CEAP?
  - 7) I have pain that I need to take medication for but no swelling or wounds. I do not have cellulitis and I wear compression socks all the time. What is my VCSS?
  - 8) Does having chronic venous insufficiency increase risk of DVT??
  - 9) What is tumescent anesthesia?
  - 10) What are the signs and symptoms of chronic venous insufficiency?
  - 11) What tests can diagnose chronic venous insufficiency?
  - 12) What are the treatment options for chronic venous insufficiency?
  - 13) What is the best treatment for chronic venous insufficiency?
  - 14) Is radiofrequency ablation better than endovenous laser ablation for the treatment of varicose veins?
  - 15) Is cyanoacrylate closure better than endovenous thermal ablation for the treatment of chronic venous insufficiency?
  - 16) What are the complications of endovenous thermal ablation for the treatment of chronic venous insufficiency?
  - 17) I have PAD and chronic venous insufficiency, can compression stockings worsen my leg ulcers?
  - 18) I have very large incompetent veins, >15 mm in diameter, what is the best treatment option?
  - 19) Should you do venous ablation for below knee great saphenous vein reflux?
  - 20) Is endovenous thermal ablation better than sclerotherapy for the treatment of chronic venous insufficiency?
- 

*CEAP*, Clinical, etiological, anatomical, and pathological classification; *DVT*, deep vein thrombosis; *PAD*, peripheral artery disease; *VCSS*, Venous Clinical Severity Score.