OPEN

PRS GLOBAL OPEN

ORIGINAL ARTICLE
Technology

Evaluation of Artificial Intelligence–generated Responses to Common Plastic Surgery Questions

Libby R. Copeland-Halperin, MD*
Lauren O'Brien, BSN, RN†
Michelle Copeland, DMD, MD, FACS‡

Background: Artificial intelligence (AI) is increasingly used to answer questions, yet the accuracy and validity of current tools are uncertain. In contrast to internet queries, AI generates summary responses as definitive. The internet is rife with inaccuracies, and plastic surgery management guidelines evolve, making verifiable information important.

Methods: We posed 10 questions about breast implant-associated illness, anaplastic large lymphoma, and squamous carcinoma to Bing, using the "more balanced" option, and to ChatGPT. Answers were reviewed by two plastic surgeons for accuracy and fidelity to information on the Food and Drug Administration (FDA) and American Society of Plastic Surgeons (ASPS) websites. We also presented 10 multiple-choice questions from the 2022 plastic surgery in-service examination to Bing, using the "more precise" option, and ChatGPT. Questions were repeated three times over consecutive weeks, and answers were evaluated for accuracy and stability.

Results: Compared with answers from the FDA and ASPS, Bing and ChatGPT were accurate. Bing answered 10 of the 30 multiple-choice questions correctly, nine incorrectly, and did not answer 11. ChatGPT correctly answered 16 and incorrectly answered 14. In both parts, responses from Bing were shorter, less detailed, and referred to verified and unverified sources; ChatGPT did not provide citations.

Conclusions: These AI tools provided accurate information from the FDA and ASPS websites, but neither consistently answered questions requiring nuanced decision-making correctly. Advances in applications to plastic surgery will require algorithms that selectively identify, evaluate, and exclude information to enhance the accuracy, precision, validity, reliability, and utility of AI-generated responses.
(Plast Reconstr Surg Glob Open 2023; 11:e5226; doi: 10.1097/GOX.0000000000005226; Published online 30 August 2023.)

INTRODUCTION

Internet-based artificial intelligence (AI) software tools such as Microsoft Bing (Microsoft Corp, Redmond, Wash.) and OpenAI ChatGPT (Open AI, San Francisco, Calif.) are used by consumers, patients, and healthcare providers to address a wide range of topics.[1,2] The accuracy and validity of statements generated by AI websites based on internet-accessible resources is uncertain. Although general internet queries present information from a variety of sources that users can selectively pursue, current-generation AI engines typically generate a single summary response as a definitive answer. Furthermore, citations or sources underlying the generated response are provided inconsistently.

The internet is rife with erroneous and inaccurate information, and evidence and management guidelines in plastic surgery constantly evolve. Given the chance and potential consequences of misinformation, we compared statements generated by AI tools in response to common clinical plastic surgery questions with information available from reputable websites, such as those developed and managed by the US Food and Drug Administration (FDA) and American Society of Plastic Surgeons (ASPS). We then evaluated the ability of these chatbots to correctly answer multiple-choice questions

*From *Northwell Health, New York, N.Y.; †Michelle Copeland DMD MD PC, New York, N.Y.; and ‡Icahn School of Medicine at Mount Sinai, New York, N.Y.*

Received for publication May 31, 2023; accepted July 13, 2023.

DOI: 10.1097/GOX.0000000000005226

Disclosure statements are at the end of this article, following the correspondence information.

Related Digital Media are available in the full-text version of the article on www.PRSGlobalOpen.com.

from the 2022 plastic surgery residency in-service examination.

METHODS

This was an exploratory study designed to assess the feasibility of a larger research initiative. Since no patient or protected health information was involved, institutional review board approval was not required. In Part 1, 10 questions on breast implant-associated anaplastic large lymphoma (BIA-ALCL), small lymphoma (BIA-SCC), and breast implant illness (BII) obtained from the FDA (website section, "Questions and Answers about Breast Implant-Associated Anaplastic Large Lymphoma," and ASPS section, "Breast Implant Safety: What Patients Need to Know," were posed to the Microsoft Bing and OpenAI ChatGPT chatbot engines.[3–6] (**See appendix, Supplemental Digital Content 1**, which displays ten questions and answers from FDA and ASPS websites about BIA-ALCL, BIA-SCC, and breast implant illness. Citations, **http://links.lww.com/PRSGO/C750**.)

For Bing, the "more balanced" response option was used, whereas ChatGPT offers only one response option. Questions were written as free text and not in PDF format. This process was repeated weekly over three consecutive weeks in April and May 2023, each time posing the questions in the same sequence.

The accuracy of the generated responses was independently evaluated by two plastic surgeons (M.C. and L.C.H.) for accuracy compared to the answers provided on the FDA and ASPS websites. Citations were analyzed for relevance to the clinical question and classified as arising from a peer-reviewed (eg, journal publication), academic (eg, medical institution or hospital, FDA, ASPS or The Esthetic Society website with references available), physician website, or other unverified source.

We then posed 10 multiple-choice questions from the 2022 plastic surgery resident in-service examination using the question stem, "Which of the following is the best answer to [insert question with possible answers]."[7] (**See appendix, Supplemental Digital Content 2**, which displays ten multiple-choice questions and answers from the 2022 plastic surgery resident in-service examination with Bing and ChatGPT answers, **http://links.lww.com/PRSGO/C751**.)

Questions were posed to Bing using the "more precise" response option. Questions were written as free text and not in PDF format. As before, this process was repeated weekly over three consecutive weeks in April and May 2023, each time posing the questions in the same sequence. Answers were reviewed for accuracy compared with the answer listed on the exam syllabus, comprehensiveness, and rationale. Bing citations were analyzed for relevance and categorized as in part 1, above. ChatGPT did not provide reference citations.

Takeaways

Question: Are internet-based artificial intelligence (AI)–generated responses to plastic surgery questions accurate and valid?

Findings: Compared with responses on the FDA and ASPS websites, Bing and ChatGPT were accurate. Neither AI platform consistently answered multiple-choice questions correctly. Bing responses were less detailed and referred to verified and unverified sources; ChatGPT did not provide citations.

Meaning: As AI tools are increasingly used, medical professionals should be aware of the utility and limitations of available platforms. Additional research should assess the ability of this technology to interpret and answer questions involving nuanced decision-making and clinical judgment.

RESULTS

Over three consecutive weeks, a total of 30 responses to the 10 FDA and ASPS website questions were obtained from Bing and ChatGPT and compared to answers taken directly from the FDA and ASPS websites.[8] Although there was variation in the statements generated in response to weekly sequential queries over 3 weeks, answers generated by Bing and ChatGPT were rated as accurate overall (30 Bing, 30 ChatGPT) (Table 1). However, there were inaccurate statements within four responses generated by ChatGPT to two questions. Specifically, ChatGPT described BIA-SCC as a "type of skin cancer" originating from skin epithelial cells in response to the same question over the three weeks instead of an "epithelial-based tumor that emanates from the breast implant capsule."[9] ChatGPT also responded that "there is some evidence… that the type of fill in a breast implant may play a role in the risk of developing" BIA-ALCL, which is contradictory to statements by the FDA and ASPS that "BIA-ALCL has been found with both silicone and saline implants…."[8] No references were provided by ChatGPT.

Bing responses were generally shorter and less detailed than those generated by ChatGPT, but citations with footnotes were generally provided as hyperlinks. All cited references were relevant to the question topic, though some were from unverified sources. Among the 102 cited sources (including duplicate citations), 4.9%

Table 1. Accuracy of Responses to Website-based Questions

	Responses Week 1		Responses Week 2		Responses Week 3	
	Overall Accurate (#)	Overall Not Accurate (#)	Overall Accurate (#)	Overall Not Accurate (#)	Overall Accurate (#)	Overall Not Accurate (#)
Bing	10	0	10	0	10	0
ChatGPT	10*	0	10*	0	10[a]	0

*ChatGPT incorrect about defining BIA-SCC as skin cancer in one response.

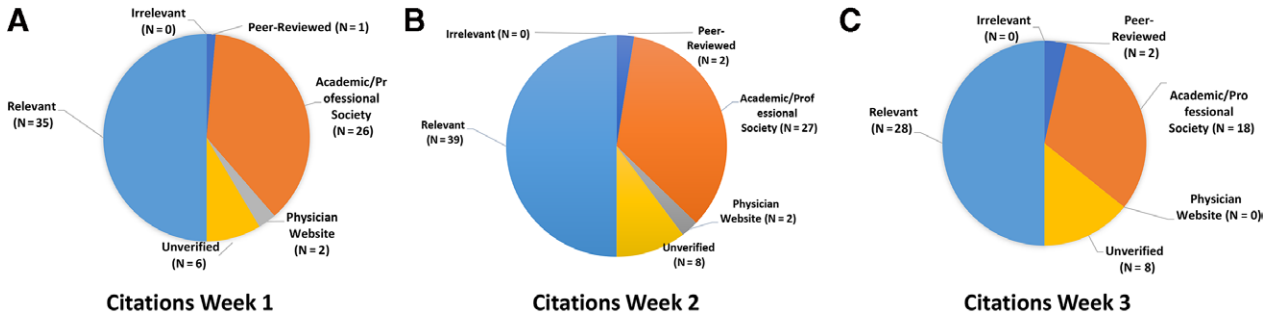

Fig. 1. Responses to website-based questions: categorization of Bing citations. A, Week 1. B, Week 2. C, Week 3.

(five references) derived from peer-reviewed journals, 69.6% (71 responses) from academic sources, 3.9% (four responses) from physician websites, and 21.6% (22 responses) from other sources (Fig. 1).

Ten multiple-choice questions from the in-service examination were posed to Bing and ChatGPT over three consecutive weeks, generating 30 responses. Overall, Bing correctly answered 10 (three, four, and three each week; 33.3% overall), incorrectly answered nine (three, six, and zero each week; 30% overall), and was unable to answer 11 (four, zero, and seven each week; 36.7% overall). ChatGPT generated responses to all questions and correctly answered 16 (five, six, and five each week; 53.3% overall) and incorrectly answered 14 (five, four, and five each week; 46.7% overall) (Fig. 2).

As in part 1, Bing responses were generally shorter and less detailed than those generated by ChatGPT. ChatGPT more often explained why other answer choices were incorrect, while Bing did not, although reasoning was not always accurate. Similarly, a correct rationale was not always provided, even when the correct answer was chosen. Among all questions, Bing provided a factually correct explanation for nine answers and no factually incorrect explanations. Among questions Bing answered incorrectly, review of the response indicated misunderstanding of the focus of the question for three responses and misunderstanding or misapplication of available data for nine responses. Among all responses generated by ChatGPT, 11 included accurate explanation/rationale (including once providing a contradictory statement, but ultimately answering the question correctly). Among questions that ChatGPT incorrectly answered, four appeared due to misunderstanding the focus of the question, whereas 14 were due to misunderstanding/

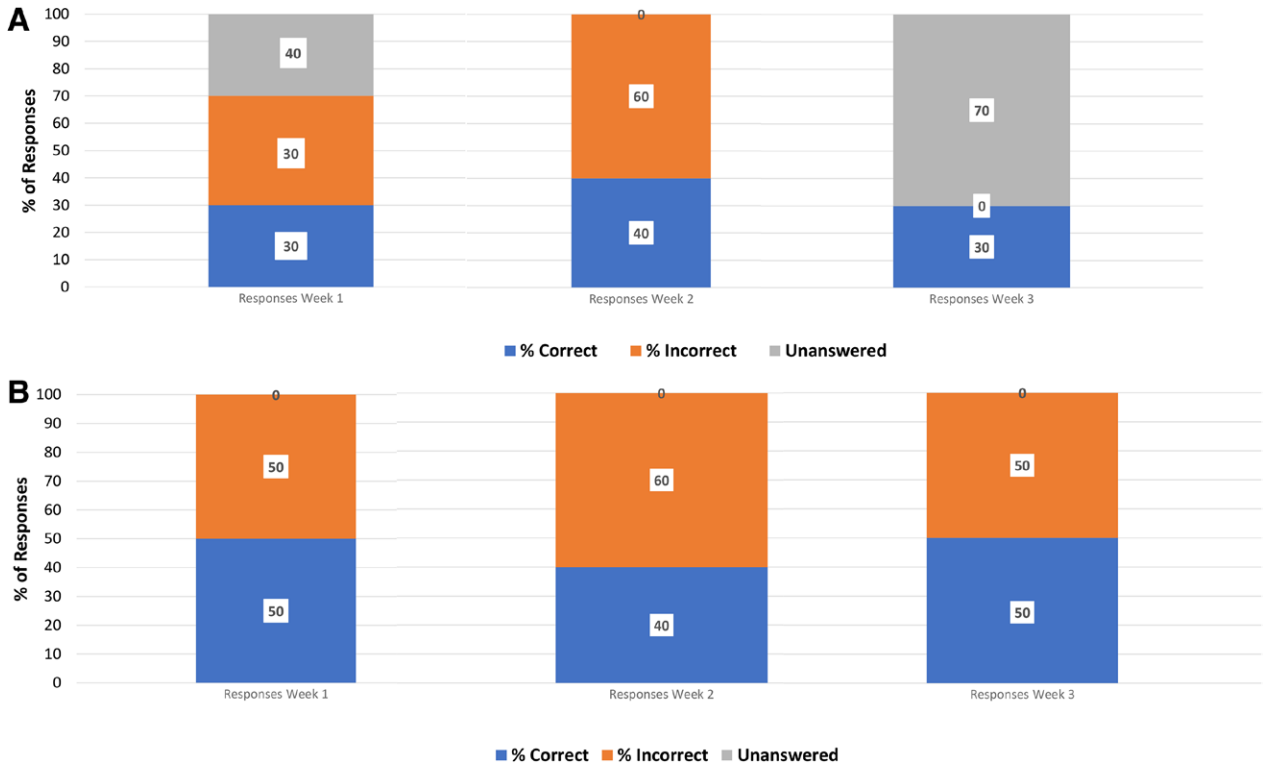

Fig. 2. Responses to multiple-choice questions. A, Accuracy of Bing Responses. B, Accuracy of ChatGPT responses.

Table 2. Multiple-choice Question and Answer: Categorization of Bing and ChatGPT Explanations

		Bing	ChatGPT
Responses week 1	Correct explanation (#)	3	4*
	Misunderstood question (#)	3	1†
	Misunderstood/applied data (#)	0	6*
	Factually incorrect explanation (#)	0	1†
	No explanation	4	0
Responses week 2	Correct explanation (#)	4	4
	Misunderstood question (#)	0	2†
	Misunderstood/applied data (#)	6	3
	Factually incorrect explanation (#)		1†
	No explanation	0	1
Responses week 3	Correct explanation (#)	2	3
	Misunderstood question (#)	0	1†
	Misunderstood/applied data (#)	0	5
	Factually incorrect explanation (#)	0	1†
	No explanation	8	1

*Duplicate when both correct explanation and incorrect/contradictory explanation provided within the same response.
†Duplicate when misunderstood focus of question and provided factually incorrect explanation.

misapplying data (including responses that were incorrect for more than one reason). ChatGPT provided three factually incorrect statements (implying that radiation from CT scans is associated with an increased risk of BIA-ALCL) and twice provided correct responses without explanation (Table 2).

Among the 125 sources cited by Bing (including duplicate sources), 72% (90 citations) were relevant. 36% (24 citations) referred to peer-reviewed publications, 47.2% (59 citations) to academic sources, 4% (five citations) were physician websites, and 12.8% (16 citations) were unverified sources. ChatGPT did not provide citations (Fig. 3).

DISCUSSION

User-friendly AI software has become widely available, and numerous studies have focused on its potential roles in medical writing and research. There has been less attention to the accuracy and validity of statements generated by AI websites based on internet-accessible resources. In one report, ChatGPT generated "interpretable responses" to common cancer questions and "appeared to minimize the likelihood of alarm" compared with Google searches on the same topics.[10] In another study, ChatGPT provided accurate information about common cancer misconceptions compared to information available on the National Cancer Institute website.[11] Our study evaluated two AI engines, ChatGPT and Bing, and focused on topics in plastic surgery.

Both Bing and ChatGPT are based on large language models. ChatGPT uses reinforcement learning with human feedback, a method OpenAI describes as using "human demonstrations and preference comparisons to guide the model toward desired behavior," and was trained on data from the internet available before 2021, but is not actively connected to the internet. Microsoft's Bing is based on ChatGPT and "consolidates reliable sources across the web to give you a single, summarized answer" through a technology known as Prometheus.[1,12]

Search engines like Google generate output based upon sources of information that appear relevant to the search terms. The user must then review the various sources and decide which to use. In contrast, AI chatbots present a single answer as authoritative, distilled from the various sources of often accurate, but sometimes misleading or incomplete information, and it may not be possible to discern the source.[13] In our study, Bing provided shortened hyperlink footnote references immediately following the majority of responses, but ChatGPT did not. Overall, the majority of sources cited were academic or professional societies. However, many of the references cited were unverified sources or irrelevant to the question topic. For instance, citing sources about general management of PIP joint contracture when a question asked specifically about management of a Dupuytren-induced PIP joint contracture.

Beyond citing a relevant source, the capacity of AI engines to interpret information is a recognized limitation. Responses are subject to bias resulting from the training data and algorithms that are not necessarily specified. For instance, disparities in healthcare research on minority groups will impact responses.[14]

Both Bing and ChatGPT include disclosures that the information provided may be "incomplete, inaccurate, or inappropriate," and ChatGPT acknowledges it has "limited knowledge of … events after 2021," and may "hallucinate outputs" or make up facts.[1,15] This was apparent in some responses to both direct website questions and answers,

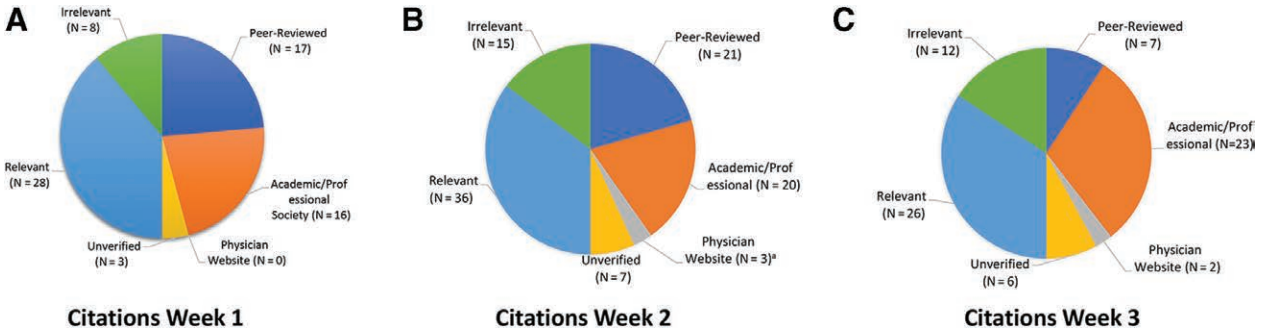

Fig. 3. Multiple-choice question and answer: categorization of Bing citations. [a]One duplicate reference within the same response. A, Week 1. B, Week 2. C, Week 3.

as well as the multiple-choice question and answers, such as when ChatGPT incorrectly suggested CT scans cause BIA-ALCL.

In our study, both Bing and ChatGPT performed poorly when responding to multiple-choice questions, correctly answering 33.3% and 53.3%, overall, respectively. These chatbots performed better with "fact-based" questions amenable to internet queries (eg, "The periodontal ligament anchors the teeth to the surrounding alveolar bone through attachment to which of the following structures?"). They performed relatively poorly when answering questions requiring nuanced decision-making, and Bing was unable to generate responses to several questions, although this improved over the three weeks, suggesting that Bing learns over time from its data bank or from the user. Notably, in these instances, it advised users to speak with a medical professional. ChatGPT also frequently directed users to consult a medical professional for additional recommendations.

In general, Bing generated shorter, more direct responses, while ChatGPT explained why other answer choices were incorrect, albeit often providing flawed explanations. In response to a question about management of ductal carcinoma in situ identified during reduction mammoplasty, both Bing and ChatGPT recognized the diagnosis of ductal carcinoma in situ, negative margins, and the indication for adjuvant radiation therapy following partial mastectomy, but did not recommend appropriate treatment following breast reduction, calling instead for observation. This points to the challenges facing the design of these platforms to understand and apply more complex, patient-specific information.

This exploratory study has several limitations. We used only English language questions posed verbatim to the AI engines to minimize bias or confounding by phrasing simplified questions and comparing answers. Future research should investigate the ability of AI to interpret and answer questions requiring more nuanced decision-making and clinical judgment. Free versions of the Bing and ChatGPT AI platforms were used for this study, as they are readily accessible to the public. It is possible that newer or more advanced versions would generate different results. The accuracy and reliability of commonly employed AI services likely vary across topics or from medical to nonmedical topics, and accuracy will likely improve as the technology evolves. Additionally, while the machine-generated responses in our study were subjected to evaluation by plastic surgeons familiar with the subject matter to determine accuracy, future studies should evaluate how nonmedical lay users and patients evaluate and interpret the information.

CONCLUSIONS

As AI tools are increasingly used, medical professionals should be aware of the utility and limitations of available platforms. Although the answers to clinical queries generated by Microsoft's Bing and Chat GPT generally aligned with information available on the FDA and ASPS websites, they were less successful at answering more complex multiple-choice questions. Additional research is needed to assess the ability of this technology to interpret and answer questions involving more nuanced decision-making and clinical judgment. Furthermore, advances in applications of AI to plastic surgery will require algorithms that identify and exclude erroneous or inaccurate information to enhance the accuracy of AI-generated responses. In medicine and surgery, as in other disciplines, professional societies and individuals must continuously evaluate the accuracy of machine-generated information, adopt standards that ensure the accuracy and timeliness of information, discourage misinformation, and direct users to appropriate resources for further information. Due to the constantly evolving and nuanced and individual decision-making required in plastic surgery, it is possible that plastic surgery will be less adaptable to AI than other areas of medicine. Regulatory bodies and medical societies should be prepared to raise awareness about the potential uses, as well as limitations, of these platforms for patients and healthcare professionals alike.

Libby R. Copeland-Halperin, MD
Northwell Health
1001 Fifth Avenue
NY, NY 10028
E-mail: lcopelandhalperin@drcopeland.com

DISCLOSURE

The authors have no financial interest to declare in relation to the content of this article.

REFERENCES

1. The New Bing—Learn More. Available at https://www.bing.com/new. Accessed April 24, 2023.
2. Schulman J, Zoph B, Kim C, et al. Introducing ChatGPT. Available at https://openai.com/blog/chatgpt. Accessed April 24, 2023.
3. Questions and answers about breast implant-associated anaplastic large cell lymphoma (BIA-ALCL). Available at https://www.fda.gov/medical-devices/breast-implants/questions-and-answers-about-breast-implant-associated-anaplastic-large-cell-lymphoma-bia-alcl. Accessed April 25, 2023.
4. Breast implant safety. Available at https://www.plasticsurgery.org/patient-safety/breast-implant-safety. Accessed April 25, 2023.
5. Bing. Available at https://www.bing.com/new. Accessed April 19, 2023.
6. ChatGPT. Available at https://chat.openai.com/. Accessed April 19, 2023.
7. American Council of Academic Plastic Surgeons. 2022 Inservice Exam. Available at https://acaplasticsurgeons.org/InService-Exams/. Accessed May 8, 2023.
8. Breast implant-associated anaplastic large cell lymphoma (BIA-ALCL). Available at www.plasticsurgery.org/patient-safety/bia-alcl-summary. Accessed May 7, 2023.
9. American Society of Plastic Surgones. ASPS statement on breast implant associated-squamous cell carcinoma (BIA-SCC). Available at https://www.plasticsurgery.org/for-medical-professionals/publications/psn-extra/news/asps-statement-on-breast-implant-associated-squamous-cell-carcinoma. Accessed May 7, 2023.
10. Hopkins A, Logan J, Kichenadasse G, et al. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *JNCI Cancer Spectr.* 2023;7:pkad010.

11. Johnson S, King A, Warner E, et al. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. *JNCI Cancer Spectr.* 2023;7:pkad015.
12. Huculak M. What is Bing Chat? An introduction to Microsoft's AI chatbot. Available at https://www.windowscentral.com/software-apps/bing/what-is-bing-chat-an-introduction-to-microsofts-ai-chatbot#:~:text=Bing%20Chat%20is%20an%20AI,how%20humans%20will%20answer%20questions. Accessed April 24, 2023.
13. Fowler G, Merrill J. The AI bot has picked an answer for you. Here's how often it's bad. *The Washington Post.* Available at https://www.washingtonpost.com/technology/2023/04/13/microsoft-bing-ai-chatbot-error/. Published April 13, 2023. Accessed April 24, 2023.
14. Yang S. The abilities and limitations of ChatGPT. Austin, TX: Anaconda. Available at https://www.anaconda.com/blog/the-abilities-and-limitations-of-chatgpt. 2022. Accessed April 24, 2023.
15. Open AI. What is ChatGPT. Available at https://help.openai.com/en/articles/6783457-what-is-chatgpt. Accessed April 24, 2023.