HHS Public Access
Author manuscript
*Am J Clin Oncol*. Author manuscript; available in PMC 2025 January 01.

Published in final edited form as:
*Am J Clin Oncol*. 2024 January 01; 47(1): 17–21. doi:10.1097/COC.0000000000001050.

# Physician Assessment of ChatGPT and Bing Answers to American Cancer Society's Questions to Ask About Your Cancer

**James R. Janopaul-Naylor, MD**,
Emory University Department of Radiation Oncology and Memorial Sloan Kettering Cancer Center Department of Radiation Oncology

**Andee Koo**,
Emory University

**David Qian, MD PhD**,
Emory University Department of Radiation Oncology

**Neal S. McCall, MD**,
Emory University Department of Radiation Oncology

**Yuan Liu, PhD MS**,
Rollins School of Public Health at Emory University

**Sagar A. Patel, MD**
Emory University Department of Radiation Oncology

## Abstract

**Objectives:** AI chatbots are a new, publicly available tool for patients to access healthcare related information with unknown reliability related to cancer-related questions. This study assesses quality of responses for common questions for patients with cancer.

**Methods:** From February to March 2023 we queried ChatGPT from OpenAI and Bing AI from Microsoft questions from American Cancer Society's recommended "Questions to Ask About Your Cancer" customized for all stages of Breast, Colon, Lung, and Prostate cancer. Questions were additionally grouped by type (prognosis, treatment, or miscellaneous). Quality of AI chatbot responses were assessed by an expert panel using the validated DISCERN criteria.

**Results:** Of the 117 questions presented to ChatGPT and Bing, the average score for all questions were 3.9 and 3.2 respectively ($p<0.001$) and the overall DISCERN scores were 4.1 and 4.4 respectively. By disease site, the average score for ChatGPT and Bing respectively were 3.9 and 3.6 for prostate cancer ($p=0.02$), 3.7 and 3.3 for lung cancer ($p<0.001$), 4.1 and 2.9 for breast cancer ($p<0.001$), and 3.8 and 3.0 for colorectal cancer ($p<0.001$). By type of question the average score for ChatGPT and Bing respectively were 3.6 and 3.4 for prognostic questions ($p=0.12$), 3.9 and 3.1 for treatment questions ($p<0.001$), and 4.2 and 3.3 for miscellaneous questions ($p=0.001$). For 3 responses (3%) by ChatGPT and 18 responses (15%) by Bing, at least one panelist rated them as having serious or extensive shortcomings.

---

**Corresponding Author** James R. Janopaul-Naylor, 1275 York Ave, New York, NY 10065, janopaj@mskcc.org, 908-542-3429.
**Conflict of Interest**: None

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

**Conclusions:** AI chatbots provide multiple opportunities for innovating healthcare. This analysis suggests a critical need, particularly around cancer prognostication, for continual refinement to limit misleading counseling, confusion, and emotional distress to patients and families.

**Keywords**

Artificial Intelligence; ChatGPT; Patient Information; Health Literacy

## Introduction:

Advances in large language models dominated headlines in 2023 with the public release of Chat Generative Pre-trained Transformer (ChatGPT) version 3.5 on November 30, 2022 and competing chatbots from Microsoft (Bing AI released on March 14, 2023) and Google (Bard released on March 21, 2023.[1-3] These chatbots provide coherent answers to user queries by iteratively predicting the next best word in a response using machine learning algorithms. The responses can alternatingly be so coherent that peer-reviewed journals are concerned about the future integrity of science, but also confidently calculate that 2 x 300 = 500.[4,5] While development of these models relied in part on human oversight, their expertise answering healthcare related questions is relatively unknown.

Patients have increasingly turned to online sources for healthcare questions. Approximately 5% of all online searches ask about medical information.[6] While some sources have rigorous peer review, others present unverified testimonials as sufficient evidence for recommendation. Concerningly, in one study, 96% of patients used unaccredited information when asked health related questions and 25% provided incorrect answers.[7] Furthermore, this misinformation can be widely disseminated on social media with ease.[8] Patients with cancer face challenging, nuanced decisions that can be confusing. Herein, we assess and compare ChatGPT and Bing AI in responding to a frequently asked questions from patients with common cancers.

## Materials and Methods:

In this cross-sectional study conducted between February and March 2023, ChatGPT[9] and Bing AI[10] were queried. We used these two large language models as they were the first available at the time of this study. Both are large language models that produce highly readable responses for almost all ages.[11] ChatGPT version 3.5 was used as the most updated version at time of data collection. ChatGPT provided answers to queries without references or links. In contrast Bing AI would provide similarly formatted text answers, but would also include links at the bottom of answers for source material that informed the response.

The two chatbots were asked a subset of questions from the American Cancer Society's handout of "Questions to Ask About Your Cancer"[12] We aimed to include common questions that would be applicable to a broad range of patients. We excluded questions about specific logistics (e.g. When and where will [diagnostics tests] be done), finances (e.g.Who can help me figure out what my insurance covers), or specific clinical scenarios (e.g. Why do you think this treatment isn't working). The selected questions were personalized

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

to different stages (I-IV), cancer types (Breast, Colon, Lung, and Prostate), and types of treatment (surgery, radiation, systemic therapy). For example, “What are my chances of surviving Stage 1 breast cancer?”. The full list of responses and questions is detailed in Supplemental Table 1.

Overall chatbot quality was independently assessed by a panel of four oncologists with 5-10 years of clinical experience treating a broad spectrum of cancer types. The panelists used the DISCERN scale, a validated questionnaire to assess the quality of consumer health resources.[13] Additionally, individual chatbot responses were scored on an ordinal scale of 1 to 5 as follows: 5-minimal shortcomings, 3-potentially important but not serious shortcomings, and 1-serious or extensive shortcomings. Questions were stratified by cancer type and question type. Question type was defined by the investigators, using three categories: prognosis (e.g. what are chances of surviving cancer or chances it will come back), treatment (e.g. what are the side effects of treatment or what can we do if it comes back), and miscellaneous (e.g. should I get genetic testing or how do we know if the treatment worked). Using the average evaluator score per question, the chatbots were compared using Wilcoxon signed rank test with a 0.05 significance level. This study received IRB exemption.

## Results:

Of all 117 questions presented, the mean scores for ChatGPT were higher than Bing (3.87+/−0.47 vs 3.19+/−0.65, $p<0.001$). There were three responses (3%) by ChatGPT in which at least one panelist rated as having serious or extensive shortcomings. There were 18 responses (15%) by Bing in which at least one panelist rated as having serious or extensive shortcomings. A sample of the lowest rated responses is in Table 1. The overall mean DISCERN score for ChatGPT and Bing respectively were 4.13+/−1.31 and 4.44+/−0.63 (Table 2).

The mean scores were higher for ChatGPT than Bing in each disease site, including breast (4.07+/−0.40 vs 2.94+/−0.64 $p<0.001$), colorectal (3.84+/−0.43 vs 2.97+/−0.51 $p<0.001$), lung (3.66+/−0.53 vs 3.30+/−0.55 $p<0.001$), and prostate (3.94+/−0.45 vs 3.59+/−0.71, $p=0.02$). By question type, the mean score for ChatGPT and Bing were similar for prognosis (3.62+/−0.67 vs 3.41+/−0.68 $p=0.12$), but ChatGPT had significantly higher scores for treatment (3.88+/−0.38 vs 3.11+/−0.54 $p<0.001$) and miscellaneous questions (4.16+/−0.41 vs 3.27+/−0.98 $p=0.001$). Figure 1 presents histograms of rating by cancer type or type of question and chatbot.

## Discussion:

ChatGPT and Bing AI provided numerous cogent responses for common cancer patient questions. However, some answers were misleading, inaccurate, or incomplete with varied reliability based on disease site and type of question. Further work is ongoing and needed to refine these publicly available resources to limit potential confusion and emotional distress.

Despite out-performing Bing on response-level assessments for all cancer types and treatment related questions, ChatGPT scored lower on the overall DISCERN scale. The

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

version of ChatGPT used in this study rarely provided sources, while Bing consistently provided citations—albeit of mixed reliability. Newer versions of ChatGPT, although subscription based, as well as emerging competitors such as Bard offer more clarity as to the sources of information or the ability to search more classic tools such as Google to verify the claims.[14]

Both ChatGPT and Bing, however, need increased consistency of accurate responses before more widespread use by patients. Both resources provided multiple responses that were wholly inaccurate or misleading. Refinement of publicly available large language models is already underway. However, providing accurate, complete responses in the absence of patient-level data is a formidable task. The challenge is compounded because these models, flawed but improving, are already available and being used by patients.

There are multiple limitations to this study. While attempting to investigate generalizable oncologic questions that patients may ask across stages, cancer types, and treatments, this study has limited specificity. However, the medical community is only scratching the surface of the implications of these resources and further work is needed to provide more specific guidance about accuracy and utility of these tools in different scenarios. Bias is possible in the evaluation of individual questions as well as in the overall DISCERN scores. We attempted to mitigate this by using multiple reviewers with varying experience and a validated evaluation instrument. This study did not evaluate the newest version of ChatGPT nor other large language models that have been released since the data collection portion of this study. Our results remain important to highlight potential pitfalls that may or may not have been addressed in the ongoing, rapid refinement of these AI chatbots. Similarly, with the rise of prompt engineering, it is possible there are better ways to phrase or ask these questions for higher fidelity responses.[15] We attempted to iteratively query three different questions with different vernacular or phrasings, but subjectively we found the responses to be similar enough to merit an identical score. As such we only report the results for a single query using language as similar as possible to the American Cancer Society's question sheet. Finally, the responses with the lowest scores ranged from misinformation to inaccuracies of omission. While both can be harmful, further work could better elucidate the differential impact and prevalence of such errors.

We assessed ChatGPT and Bing responses to common patient questions about Breast, Colorectal, Lung, and Prostate cancer. While many responses were helpful, thorough, and reliable, some were inaccurate or incomplete, particularly for answers pertaining to cancer prognostication. As patients increasingly turn to new web-based health information, this analysis suggests a critical need for continual refinement to limit misleading counseling, confusion, and emotional distress to patients and families.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

## Funding:

Research reported in this publication was supported in part by the Biostatistics Shared Resource of Winship Cancer Institute of Emory University and NIH/NCI under award number P30CA138292. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health

## Data Access Statement:

Materials and data not available in manuscript and supplemental material will be made available upon reasonable request to the corresponding author.

## References

1. Mehdi Yusuf. Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web. Off Microsoft Blog. Published online 2023. Accessed August 17, 2023. https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/
2. What's new with Bard. Accessed August 17, 2023. https://bard.google.com/updates
3. OpenAI. ChatGPT — Release Notes | OpenAI Help Center. Published online 2023. Accessed August 17, 2023. https://help.openai.com/en/articles/6825453-chatgpt-release-notes
4. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. Nature. 2023;614(7947):224–226. doi:10.1038/d41586-023-00288-7 [PubMed: 36737653]
5. Zumbrun J. ChatGPT Needs Some Help With Math Assignments. Wall Street Journal. Published 2023. Accessed February 21, 2023. https://www.wsj.com/articles/ai-bot-chatgpt-needs-some-help-with-math-assignments-11675390552
6. Swire-Thompson B, Lazer D. Public health and online misinformation: Challenges and recommendations. Annu Rev Public Health. 2019;41:433–451. doi:10.1146/annurev-publhealth-040119-094127 [PubMed: 31874069]
7. Quinn S, Bond R, Nugent C. Quantifying health literacy and eHealth literacy using existing instruments and browser-based software for tracking online health information seeking behavior. Comput Human Behav. 2017;69:256–267. doi:10.1016/j.chb.2016.12.032
8. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science (80- ). 2018;359(6380):1146–1151. doi:10.1126/science.aap9559
9. ChatGPT [Large Language Model] March 2023 Version. https://chat.openai.com/chat
10. Bing AI [Large Language Model] March 2023 Version.
11. Murgia E, Pera MS, Landoni M, Huibers T. Children on ChatGPT Readability in an Educational Context: Myth or Opportunity? In: Association for Computing Machinery (ACM); 2023:311–316. doi:10.1145/3563359.3596996
12. Portion PO, Bmi Y. Questions to ask your doctor. :17. Accessed May 26, 2023. https://www.cancer.org/cancer/managing-cancer/making-treatment-decisions/questions-to-ask-your-doctor.html
13. Charnock D, Shepperd S. Learning to DISCERN online: Applying an appraisal tool to health websites in a workshop setting. Health Educ Res. 2004;19(4):440–446. doi:10.1093/her/cyg046 [PubMed: 15155597]
14. Bard. Accessed July 19, 2023. https://bard.google.com/
15. Giray L. Prompt Engineering with ChatGPT: A Guide for Academic Writers. Ann Biomed Eng. Published online 2023. doi:10.1007/s10439-023-03272-4

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

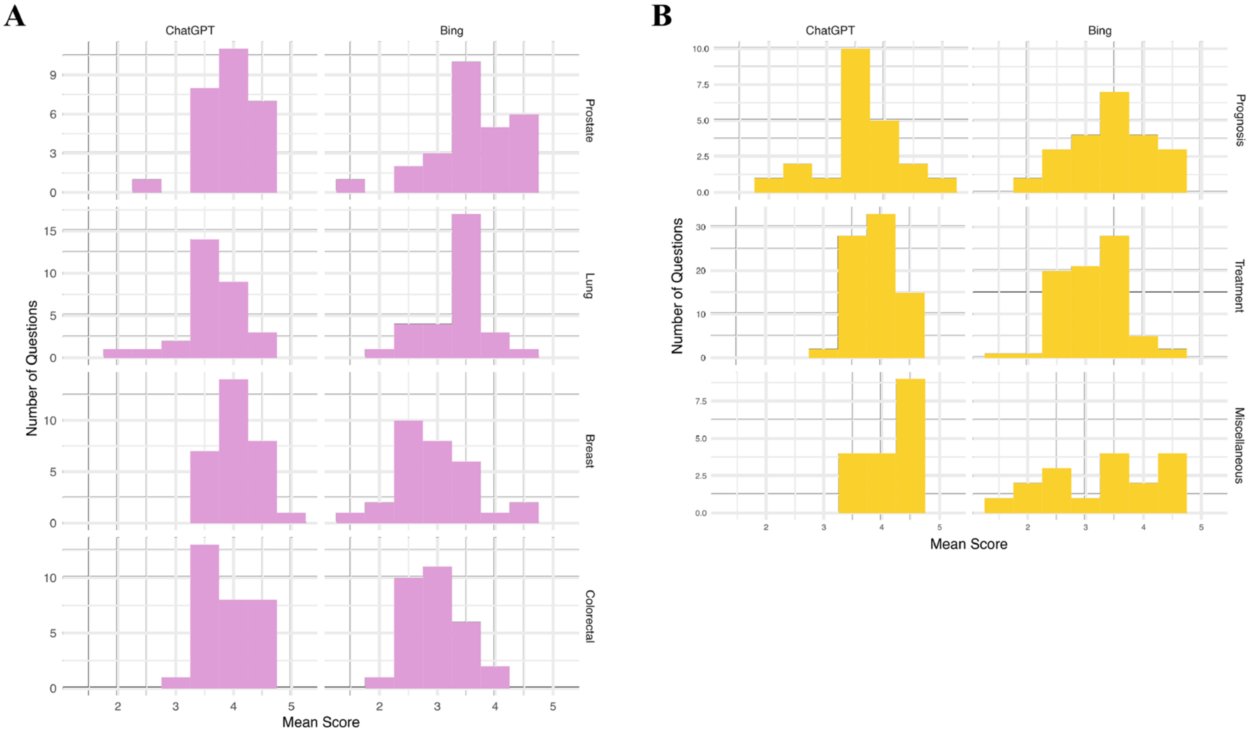

**Figure 1.**

Histograms of mean scores for questions stratified by Chat Bot (ChatGPT or Bing AI) and (A) cancer type (Prostate, Lung, Breast, or Colorectal) or (B) question type (Prognosis, Treatment, or Miscellaneous). Ordinal scoring scale: 5-minimal shortcomings, 3-potentially important but not serious shortcomings, and 1-serious or extensive shortcomings.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

## Table 1:

ChatGPT and Bing responses with the lowest scores based on expert panel review.

| Question | Response | ChatBot | Concern |
|---|---|---|---|
| Will I need surgery for my prostate cancer | As an AI language model, I cannot provide medical advice or diagnose medical conditions. It is best to consult with a qualified medical professional who can evaluate your individual case and recommend the appropriate course of treatment for you. They will consider factors such as the stage and aggressiveness of the cancer, your age and overall health, and your personal preferences before making any treatment recommendations. | ChatGPT | Does not provide any information. Inadequate response. |
| How long will it take for my side effects to go away from my prostate cancer radiation | I'm sorry to hear that you are experiencing side effects from your prostate cancer radiation. According to WebMD[1], some of the long-term side effects of radiation therapy for prostate cancer may last up to a year or longer after treatment[1]. These may include skin changes, enlarged pores, sensitivity issues, and tissue thickening[1]. However, these side effects may vary depending on the type and dose of radiation you received, your overall health, and other factors[4]. Therefore, it is best to consult your doctor about your specific situation and how to manage your symptoms.<br>Some questions you may want to ask your doctor are:<br>- What type and dose of radiation did I receive?<br>- How long can I expect my side effects to last?<br>- What can I do to relieve my discomfort or pain?<br>- Are there any medications or supplements that can help me?<br>- Do I need any follow-up tests or treatments?<br>I hope this helps you find some relief and answers. Please take care of yourself.🙏<br>Source: Conversation with Bing, 2/26/2023(1) Prostate Cancer Radiation Therapy: Treatment and Side Effects - WebMD. https://www.webmd.com/prostate-cancer/guide/prostate-cancer-radiation-therapy Accessed 2/26/2023.<br>(2) Side Effects of Radiation for Prostate Cancer Vs. Benefits. https://www.healthline.com/health/prostate-cancer/side-effects-of-radiation-for-prostate-cancer Accessed 2/26/2023.<br>(3) What Are The After Effects Of Radiation For Prostate Cancer. https://www.healthyprostateclub.com/what-are-the-after-effects-of-radiation-for-prostate-cancer/ Accessed 2/26/2023.<br>(4) Side effects of prostate cancer radiotherapy. https://www.cancerresearchuk.org/about-cancer/prostate-cancer/treatment/radiotherapy/external-radiotherapy/side-effects-external-radiotherapy Accessed 2/26/2023.<br>(5) External beam radiation for prostate cancer - Mayo Clinic. https://www.mayoclinic.org/tests-procedures/external-beam-radiation-for-prostate-cancer/about/pac-20384743 Accessed 2/26/2023. | Bing AI | Does not discuss any of most common, dangerous, or bothersome side effects. |
| What are my chances of surviving Stage 2 lung cancer | The survival rates for stage 2 lung cancer depend on several factors, including the size of the tumor, its location in the lung, whether it has spread to nearby lymph nodes or other organs, and the patient's overall health and age. Survival rates are generally better for people with smaller tumors that have not spread to lymph nodes or other organs.<br>According to the American Cancer Society, the overall 5-year survival rate for stage 2 non-small cell lung cancer (NSCLC) is approximately 36%. This means that out of 100 people diagnosed with stage 2 NSCLC, about 36 will still be alive 5 years after diagnosis. For stage 2 small cell lung cancer (SCLC), the 5-year survival rate is lower, with an estimated 13% of people surviving 5 years or more.<br>It is important to note that survival rates are only estimates and cannot predict individual outcomes. | ChatGPT | ChatGPT provides inaccurately low survival rates for both small cell and non-small cell lung cancer |
| Is there anything I can do to make my breast cancer treatment work better | There are several options available to supplement traditional medicine for breast cancer treatment. These include acupuncture, detoxification diets, traditional Chinese medicine, and antioxidants among other alternatives known as complementary and alternative medicine (CAM)[1]. Some targeted drugs can also help other types of treatment work better[2].<br>Your doctor will determine your breast cancer treatment options based on factors such as your type of breast cancer, its stage and grade, size, and whether the cancer cells are sensitive to hormones[3].<br>Source: Conversation with Bing, 3/3/2023(1) Alternative Treatments for Breast Cancer: What Works? - Healthline. https:// | Bing AI | Bing recommends several unproven strategies that could interact negatively with standard of care treatments |

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

| Question | Response | ChatBot | Concern |
|---|---|---|---|
| | www.healthline.com/health/breast-cancer/alternative-treatments-for-breast-cancer Accessed 3/3/2023.<br>(2) Targeted Drug Therapy ∣ Breast Cancer Treatment. https://www.cancer.org/cancer/breast-cancer/treatment/targeted-therapy-for-breast-cancer.html Accessed 3/3/2023.<br>(3) Breast cancer - Diagnosis and treatment - Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475 Accessed 3/3/2023.<br>(4) Treatment of Breast Cancer Stages I-III - American Cancer Society. https://www.cancer.org/cancer/breast-cancer/treatment/treatment-of-breast-cancer-by-stage/treatment-of-breast-cancer-stages-i-iii.html Accessed 3/3/2023.<br>(5) Hormone therapy for breast cancer - Mayo Clinic - Mayo Clinic - Mayo Clinic. https://www.mayoclinic.org/tests-procedures/hormone-therapy-for-breast-cancer/about/pac-20384943 Accessed 3/3/2023. | | |

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

## Table 2:

DISCERN questionnaire and panelist scores for ChatGPT and Bing

| DISCERN Question | Scale | ChatGPT Score | Bing Score |
|---|---|---|---|
| Are the aims clear | 1=no, 3=partially, 5=yes | 5 | 5 |
| Does it achieve its aims | 1=no, 3=partially, 5=yes | 4 | 4 |
| Is it relevant | 1=no, 3=partially, 5=yes | 5 | 5 |
| Is it clear what sources of information were used to compile the publication? | 1=no, 3=partially, 5=yes | 2 | 4 |
| Is it clear when the information used or reported in the publication was produced? | 1=no, 3=partially, 5=yes | 1 | 4 |
| Is it balanced and unbiased? | 1=no, 3=partially, 5=yes | 4 | 4 |
| Does it provide details of additional sources of support and information? | 1=no, 3=partially, 5=yes | 2 | 4 |
| Does it refer to areas of uncertainty? | 1=no, 3=partially, 5=yes | 4 | 4 |
| Does it describe how each treatment works? | 1=no, 3=partially, 5=yes | 5 | 5 |
| Does it describe the benefits of each treatment? | 1=no, 3=partially, 5=yes | 5 | 5 |
| Does it describe the risks of each treatment? | 1=no, 3=partially, 5=yes | 5 | 5 |
| Does it describe what would happen if no treatment is used? | 1=no, 3=partially, 5=yes | 5 | 5 |
| Does it describe how the treatment choices affect overall quality of life? | 1=no, 3=partially, 5=yes | 5 | 5 |
| Is it clear that there may be more than one possible treatment choice? | 1=no, 3=partially, 5=yes | 5 | 4 |
| Does it provide support for shared decision-making? | 1=no, 3=partially, 5=yes | 5 | 5 |
| Based on the answers to all of the above questions, rate the overall quality of the publication as a source of information about treatment choices | 1=serious or extensive shortcomings; 3=potentially important but not serious shortcomings; 5=minimal shortcomings | 4 | 3 |
|  | **Mean** | 4.13 | 4.44 |
|  | **Standard Deviation** | 1.31 | 0.63 |