Journal for ImmunoTherapy of Cancer

Use of artificial intelligence chatbots in clinical management of immune-related adverse events

Hannah Burnette [1], Aliyah Pabani [2], Mitchell S von Itzstein,[3] Benjamin Switzer [4], Run Fan,[5] Fei Ye,[5] Igor Puzanov [4], Jarushka Naidoo [6], Paolo A Ascierto [7], David E Gerber [3], Marc S Ernstoff [8], Douglas B Johnson[1]

To cite: Burnette H, Pabani A, von Itzstein MS, *et al*. Use of artificial intelligence chatbots in clinical management of immune-related adverse events. *Journal for ImmunoTherapy of Cancer* 2024;**12**:e008599. doi:10.1136/jitc-2023-008599

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/jitc-2023-008599).

Accepted 14 May 2024

Check for updates

For numbered affiliations see end of article.

Correspondence to
Dr Douglas B Johnson;
douglas.b.johnson@vumc.org

ABSTRACT

Background Artificial intelligence (AI) chatbots have become a major source of general and medical information, though their accuracy and completeness are still being assessed. Their utility to answer questions surrounding immune-related adverse events (irAEs), common and potentially dangerous toxicities from cancer immunotherapy, are not well defined.

Methods We developed 50 distinct questions with answers in available guidelines surrounding 10 irAE categories and queried two AI chatbots (ChatGPT and Bard), along with an additional 20 patient-specific scenarios. Experts in irAE management scored answers for accuracy and completion using a Likert scale ranging from 1 (least accurate/complete) to 4 (most accurate/complete). Answers across categories and across engines were compared.

Results Overall, both engines scored highly for accuracy (mean scores for ChatGPT and Bard were 3.87 vs 3.5, $p<0.01$) and completeness (3.83 vs 3.46, $p<0.01$). Scores of 1–2 (completely or mostly inaccurate or incomplete) were particularly rare for ChatGPT (6/800 answer-ratings, 0.75%). Of the 50 questions, all eight physician raters gave ChatGPT a rating of 4 (fully accurate or complete) for 22 questions (for accuracy) and 16 questions (for completeness). In the 20 patient scenarios, the average accuracy score was 3.725 (median 4) and the average completeness was 3.61 (median 4).

Conclusions AI chatbots provided largely accurate and complete information regarding irAEs, and wildly inaccurate information (“hallucinations”) was uncommon. However, until accuracy and completeness increases further, appropriate guidelines remain the gold standard to follow

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Large language model chatbots can provide information about a variety of topics, including medical data. However, the utility of LLMs for complex immune-related adverse event (irAE) questions is unclear.

WHAT THIS STUDY ADDS

⇒ We found that ChatGPT provided generally accuate and comprehensive answers to queries about irAEs, though occasional errors were noted.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Clinicians may use ChatGPT as a resource for irAEs, though additional verification is needed.

BACKGROUND

The advent of new artificial intelligence chatbots such as ChatGPT, Google Bard, and many others (hereafter referred to as chatbots) has the potential to change medical diagnostics and treatment drastically. These chatbots, built around large language models, analyze various data sets procured from sources found on the internet and learn from them before producing human-like answers to address inputted queries.[1] The answers generated by the chatbots evolve based on human feedback combined with the availability of new or updated sources of information. This allows the chatbot to provide more complex answers that are better aligned with the end-user’s original intentions.

The ever-increasing extent and availability of medical information presents substantial challenges to physicians. Increasingly, both physicians and patients are turning to chatbots to help make medical information more digestible and accessible. Determining whether chatbot answers are accurate or reliable is important, especially given that patients are increasingly relying on the answers from these chatbots to inform their medical decision-making.[2] Several studies have shown that earlier versions of chatbots provide digestible and fairly accurate information, but may also provide incomplete, inaccurate, or out-of-date answers.[3 4] Many of these studies though, focus on multiple-choice or binary answers, which often do not reflect the open-ended nature of the real-world medical practice. Lastly, chatbot responses may also lack both the emotional aspects of healthcare such as empathy although some studies suggest they perform well in this regard.[5–7]

This study seeks to analyze the accuracy and completeness of chatbot-generated answers surrounding complex, open-ended questions regarding immune-related adverse events (irAEs). These immune-related toxicities impact multiple organs,[8] are treated algorithmically by defined guidelines,[9–11] and are common medical problems for physicians caring for patients with cancer. Further, the diverse range of organs affected, the often non-specific clinical presentations, and the multidisciplinary management required make this a challenging area for clinicians, and thus a potentially attractive area for chatbot-derived assistance.

METHODS

This cross-sectional study was exempt from institutional review board review given the lack of patient data. Available guidelines for the management of irAEs were reviewed. Based on these guidelines, a total of 50 questions were generated by the senior author (DBJ) and refined/approved by other study authors as representative common questions that arise in clinical settings. Five questions each from nine common irAE categories were generated (gastrointestinal, hepatic, pulmonary, dermatologic, thyroid, pituitary/adrenal, rheumatologic, neuromuscular, cardiac), with an additional five questions about general irAE management. All questions were designed as descriptive and open-ended in nature (online supplemental table 1), but with clearly defined answers present in available guidelines from international committees with expertize in irAEs.[9–11]

Finalized questions were entered into two chatbots (ChatGPT (V.GPT-4) and Google Bard) by the first author (HB) on October 6, 2023. Answers were provided back to the rating physicians. Rating physicians were either members of the Society for Immunotherapy in Cancer immune checkpoint inhibitor and cytokine-related adverse events subcommittee (n=5) or their colleagues with a strong focus on irAE management (n=3). All answers were graded by each rater for accuracy and completeness for both chatbots. Accuracy was graded on a 1–4 point Likert scale, with 1 signifying completely inaccurate, 2 mostly inaccurate, 3 mostly accurate, and 4 accurate. Raters were instructed to grade accuracy results based on guideline content, not personal management style. Similarly, completeness was graded on a 1–4 point Likert scale, with 1 signifying incomplete, 2 missing multiple pieces of key information, 3 missing one piece of key information, and 4 complete. Raters were instructed to grade based on major pieces of key information rather than minor or optional items, specifically giving the example of colitis (major including endoscopic evaluation, minor/optional being fecal calprotectin testing).

Grades were summarized with means, medians, and ranges for each chatbot overall and for each irAE category. Scores for completeness and accuracy were compared between chatbots using Wilcoxon signed-rank tests. Inter-rater agreement was assessed with Kendall's coefficient of concordance since there were >2 raters. The two-sample binomial proportion test was used to compare the proportions of certain ratings between chatbots.

To further judge accuracy and completeness, 20 different clinical scenarios were generated by DBJ and approved by other participating authors, and entered into ChatGPT (not Bard given the amount of time that had elapsed and poorer performance) on March 20, 2024. Two questions were generated from each of the 10 categories, and were judged by four of the rating physicians.

RESULTS

Both chatbots were rated for accuracy and completeness on 50 questions from 10 different categories (see online supplemental file). Both chatbots had relatively high scores overall; ChatGPT scored a median of 3.88 for accuracy (mean 3.87) and 3.88 for completeness (mean 3.83) across all questions and raters. Bard scores were median 3.5 for accuracy (mean 3.5) and 3.5 for completeness (mean 3.46). Inter-rater agreement was fair across all raters (Kendall's correlation coefficients for accuracy and completeness were 0.21 and 0.24 for ChatGPT and 0.27 and 0.24 for Bard).[12] Overall, GPT-4 received significantly higher ratings compared with Bard in both accuracy and completeness (p<0.001).

We then assessed scores stratified by category by pooling scores across five questions per category (maximum of 20 per category). Both mean and median scores for each category for both accuracy and completeness were between 19 and 20 except for general immune checkpoint inhibitor (ICI) questions for ChatGPT (table 1). Median scores for Bard ranged from 15.5 to 19, with similar ranges for mean scores (16–18.5) (table 1). Scores in all categories were rated numerically higher with ChatGPT. This difference reached statistical significance (p<0.05) in one category for accuracy (cardiac) and five categories for completeness (hepatic, dermatologic, thyroid, pituitary/adrenal, and cardiac). An additional six categories for accuracy, and two categories for completeness showed marginal statistical significance (p<0.1) favoring ChatGPT. By category, the "general" category had the lowest scores for ChatGPT with generally high scores across specific irAE categories, whereas Bard seemed to perform highest in dermatologic, rheumatologic, neuromuscular, and cardiac categories.

There were multiple questions that received ratings of 4 from all eight reviewers, including 22/50, 44% (ChatGPT accuracy) and 16/50, 32% (ChatGPT completeness) compared with 2/50, 4% (Bard accuracy) and 1/50, 2% (Bard completeness) (p<0.001). Ratings of 1 (fully inaccurate or incomplete) were uncommon, given for 2/800 ChatGPT rater-responses (0.3%) and 9/800 Bard

Table 1 Scores for accuracy and completeness for each engine in each category

Accuracy					Completeness				
	ChatGPT (N=8)	Bard (N=8)	Overall (N=16)	P value		ChatGPT (N=8)	Bard (N=8)	Overall (N=16)	P value
ICI – general					ICI – general				
Mean (SD)	17.9 (1.81)	17.3 (2.05)	17.6 (1.90)	0.52	Mean (SD)	17.3 (2.25)	17.0 (1.60)	17.1 (1.89)	1
Median (Min, Max)	18.0 (16, 20)	17.0 (15, 20)	17.5 (15, 20)		Median (Min, Max)	17.0 (14, 20)	17.0 (15, 20)	17.0 (14, 20)	
Gastrointestinal					Gastrointestinal				
Mean (SD)	19.0 (1.07)	18.3 (1.67)	18.6 (1.41)	0.202	Mean (SD)	18.8 (1.28)	18.0 (1.77)	18.4 (1.54)	0.188
Median (Min, Max)	19.0 (17, 20)	18.0 (16, 20)	19.0 (16, 20)		Median (Min, Max)	19.0 (16, 20)	18.0 (15, 20)	19.0 (15, 20)	
Hepatic					Hepatic				
Mean (SD)	19.4 (0.518)	16.9 (2.53)	18.1 (2.19)	0.057	Mean (SD)	19.1 (0.835)	16.0 (3.07)	17.6 (2.71)	0.041
Median (Min, Max)	19.0 (19, 20)	16.5 (14, 20)	19.0 (14, 20)		Median (Min, Max)	19.0 (18, 20)	15.5 (12, 20)	18.5 (12, 20)	
Pulmonary					Pulmonary				
Mean (SD)	19.0 (0.756)	16.0 (4.11)	17.5 (3.25)	0.093	Mean (SD)	19.1 (0.835)	17.3 (2.87)	18.2 (2.26)	0.106
Median (Min, Max)	19.0 (18, 20)	16.0 (9, 20)	19.0 (9, 20)		Median (Min, Max)	19.0 (18, 20)	18.0 (12, 20)	19.0 (12, 20)	
Dermatologic					Dermatologic				
Mean (SD)	19.5 (0.756)	18.5 (1.69)	19.0 (1.37)	0.134	Mean (SD)	19.6 (0.518)	18.0 (1.60)	18.8 (1.42)	0.035
Median (Min, Max)	20.0 (18, 20)	18.5 (15, 20)	19.5 (15, 20)		Median (Min, Max)	20.0 (19, 20)	18.0 (16, 20)	19.0 (16, 20)	
Thyroid					Thyroid				
Mean (SD)	19.8 (0.463)	16.6 (3.38)	18.2 (2.83)	0.058	Mean (SD)	19.4 (0.916)	17.3 (2.92)	18.3 (2.36)	0.035
Median (Min, Max)	20.0 (19, 20)	16.5 (11, 20)	20.0 (11, 20)		Median (Min, Max)	20.0 (18, 20)	17.5 (11, 20)	19.0 (11, 20)	
Pituitary and adrenal					Pituitary and adrenal				
Mean (SD)	19.8 (0.707)	16.9 (2.80)	18.3 (2.47)	0.057	Mean (SD)	19.8 (0.463)	16.9 (2.85)	18.3 (2.47)	0.036
Median (Min, Max)	20.0 (18, 20)	16.5 (14, 20)	20.0 (14, 20)		Median (Min, Max)	20.0 (19, 20)	17.0 (12, 20)	19.5 (12, 20)	
Rheumatologic					Rheumatologic				
Mean (SD)	19.9 (0.354)	18.3 (1.98)	19.1 (1.61)	0.098	Mean (SD)	19.4 (0.744)	17.8 (2.25)	18.6 (1.82)	0.058
Median (Min, Max)	20.0 (19, 20)	19.0 (16, 20)	20.0 (16, 20)		Median (Min, Max)	19.5 (18, 20)	18.5 (14, 20)	19.0 (14, 20)	
Neuromuscular					Neuromuscular				
Mean (SD)	19.6 (0.744)	18.4 (2.00)	19.0 (1.59)	0.099	Mean (SD)	19.5 (0.535)	17.8 (2.43)	18.6 (1.93)	0.056
Median (Min, Max)	20.0 (18, 20)	19.0 (15, 20)	20.0 (15, 20)		Median (Min, Max)	19.5 (19, 20)	18.0 (13, 20)	19.0 (13, 20)	
Cardiac					Cardiac				
Mean (SD)	19.6 (0.518)	18.0 (1.93)	18.8 (1.60)	0.034	Mean (SD)	19.4 (0.744)	17.0 (2.56)	18.2 (2.20)	0.036
Median (Min, Max)	20.0 (19, 20)	18.0 (14, 20)	19.0 (14, 20)		Median (Min, Max)	19.5 (18, 20)	17.0 (12, 20)	19.0 (12, 20)	

rater-responses (1.1%). Ratings of 2 (mostly incorrect or missing multiple key pieces of information) were of similar incidence for ChatGPT (4/800 rater-responses, 0.5%), though more common for Bard (83/800 rater-responses, 10.4%) (p<0.001).

To assess utility in specific clinical scenarios, we provided 20 different patient-specific scenarios (see online supplemental file) into ChatGPT. These answers were also rated highly; mean accuracy was 3.73 (median 4) and mean completeness was 3.61 (median 4). Of the 80 physician-answers, scores were 4 (n=53), 3 (n=23), 2 (n=4), and 1 (n=0).

DISCUSSION

In this study, we found that chatbots, particularly ChatGPT (V.GPT-4), provided generally accurate and complete information surrounding irAEs. Questions were open-ended (not multiple choice), mirroring real-life situations rather than board examinations. The median rating for many questions was 4 (fully accurate and complete), and egregiously wrong answers were uncommon. Thus, these engines appear promising for use in receiving guidance for irAEs.

Although both engines had a reasonably high degree of accuracy and completeness, it appeared that ChatGPT was further advanced in providing accurate and comprehensive information compared with Bard. Ratings of 3 or 4 predominated for ChatGPT (794 of 800 rater-responses), thus showing consistently high grades across physician raters. As a new technology, it is likely that chatbots will change and upgrade rapidly though, thus comparisons between engines may be rapidly outdated. It is also likely that different engines will ultimately be optimized for distinct tasks and prioritize different capabilities (eg, accuracy vs comprehensiveness). In addition, chatbots may be designed to maximize other goals, such as conciseness (eg, avoiding extraneous information) or delivering information at a specific educational attainment level. These goals are also important to maximize high-yield information delivery to busy clinicians. Of note, ChatGPT and other engines have shown promise in providing high-quality medical information across a range of medical conditions.[13–15] This includes general immune-oncology questions,[16] urological cancers,[17] and preoperative counseling for head and neck cancer surgery.[18]

Interestingly, ratings of 1 (fully incorrect or incomplete) were very uncommon, suggesting that outright "hallucinations" were very rare. At the outset of these technologies, this phenomenon appeared to occur with troubling frequency.[19] The rarity of egregiously wrong answers in this data set suggests that such hallucinations may be a surmountable problem, at least in this type of focused question set with concrete answers available in publicly available guidelines. However, it could be argued that less frequent wrong answers may increase the impact of residual incorrect information, since increasing trust in the outcomes may decrease reliance on other more validated sources.

Tempering this enthusiasm is the fact that most questions did not universally receive a rating of 4 (fully accurate and/or complete) on all questions. This could reflect subjective disagreement by highly experienced physicians, but could also suggest that these chatbots may not be reliable as stand-alone sources of medical information. A potentially important future direction could include training chatbots specifically on irAE and other cancer-specific guidelines, as has been done with other corpus of texts. Until those types of advances, available guidelines remain a golden standard when making medical decisions. It is also important to note that ratings were subjective, and could differ with different clinicians (and could be impacted based on the particular Likert scale used). It is also possible that new features or upgrades worsen the model performance; this will be difficult to assess.

In conclusion, current iterations of chatbots provide fairly accurate and complete information to many questions surrounding irAEs, though important differences are present between different chatbots. Additional research and validation are needed prior to using these engines as "stand-alone" resources.

Author affiliations

[1]Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA
[2]Department of Oncology, Johns Hopkins University, Baltimore, Maryland, USA
[3]Harold C Simmons Comprehensive Cancer Center, The University of Texas Southwestern Medical Center, Dallas, Texas, USA
[4]Department of Medicine, Roswell Park Comprehensive Cancer Center, Buffalo, New York, USA
[5]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, USA
[6]RCSI Cancer Centre, Beaumont Hospital, Dublin, Ireland
[7]Department of Melanoma, Cancer Immunotherapy and Development Therapeutics, Istituto Nazionale Tumori IRCCS Fondazione Pascale, Napoli, Campania, Italy
[8]ImmunoOncology Branch (IOB), Developmental Therapeutics Program, Cancer Therapy and Diagnosis Division, National Cancer Institute (NCI), National Institutes of Health, Bethesda, Maryland, USA

X Benjamin Switzer @BenSwitzerDO, Jarushka Naidoo @DrJNaidoo and Paolo A Ascierto @PAscierto

Contributors DBJ and HB contributed to conception, design, and writing the manuscript. AP, MSvl, BS, RF, FY, IP, JN, PAA, DEG, and MSE reviewed the manuscript and participated in the research. DBJ is responsible for the overall content as guarantor. As this study is based on assessing AI's potential impact on the healthcare field, AI was used to generate answers to 50 questions and 20 scenarios so they could be assessed for accuracy and completeness by physicians. AI was not used to write the manuscript in any way nor was it used to analyze the data.

Funding DBJ receives funding from the NCI R01CA227481. Susan and Luke Simons Directorship for Melanoma, the James C. Bradford Melanoma Fund, the Van Stephenson Melanoma Fund.

Competing interests AP reports grants and personal fees from Bristol-Myers Squibb; personal fees from AstraZeneca, Pfizer, Merck, Roche, and Canadian Agency for Drugs and Technologies in Health; and grants from Alberta Cancer Foundation outside the submitted work. IP has served on advisory boards for Nektar, Iovance, Nouscom, I-O Bio and has stock ownership in Ideaya. DBJ has served on advisory boards or as a consultant for BMS, Catalyst Biopharma, Iovance, Mallinckrodt, Merck, Mosaic ImmunoEngineering, Novartis, Pfizer, Targovax, and Teiko, has received research funding from BMS and Incyte, and has patents pending for use of MHC-II as a biomarker for immune checkpoint inhibitor response, and abatacept as treatment for immune-related adverse events. DEG reports research funding from AstraZeneca, BerGenBio, Karyopharm, and Novocure; stock ownership in Gilead; service on advisory boards or consulting for AstraZeneca, Catalyst Pharmaceuticals, Daiichi-Sankyo, Elevation Oncology, Janssen Scientific Affairs, LLC, Jazz Pharmaceuticals, Regeneron Pharmaceuticals, Sanofi; U.S. patent 11,747,345, patent applications 17/045,482, 63/386,387, 63/382,972, 63/382,257; and is Co-founder and Chief Scientific Officer of OncoSeer Diagnostics, LLC.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information. All data associated with this manuscript has been provided in the form of supplemental materials and can be found in online supplemental table 1.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs

Hannah Burnette http://orcid.org/0009-0009-7784-634X
Aliyah Pabani http://orcid.org/0009-0006-7605-8497
Benjamin Switzer http://orcid.org/0000-0001-8150-1963
Igor Puzanov http://orcid.org/0000-0002-9803-3497
Jarushka Naidoo http://orcid.org/0000-0002-3470-8686
Paolo A Ascierto http://orcid.org/0000-0002-8322-475X
David E Gerber http://orcid.org/0000-0002-7812-6741
Marc S Ernstoff http://orcid.org/0000-0002-8132-7069

REFERENCES

1 Shen Y, Heacock L, Elias J, *et al*. Chatgpt and other large language models are double-edged swords. *Radiology* 2023;307:e230163.
2 Pan A, Musheyev D, Bockelman D, *et al*. Assessment of artificial intelligence Chatbot responses to top searched queries about cancer. *JAMA Oncol* 2023;9:1437–40.
3 Chen S, Kann BH, Foote MB, *et al*. Use of artificial intelligence Chatbots for cancer treatment information. *JAMA Oncol* 2023;9:1459–62.
4 Goodman RS, Patrinely JR, Stone CA Jr, *et al*. Accuracy and reliability of Chatbot responses to physician questions. *JAMA Netw Open* 2023;6:e2336483.
5 El-Metwally A, Toivola P, AlAhmary K, *et al*. The epidemiology of migraine headache in Arab countries: A systematic review. *ScientificWorldJournal* 2020;2020:4790254.
6 Ayers JW, Poliak A, Dredze M, *et al*. Comparing physician and artificial intelligence Chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med* 2023;183:589–96.
7 Maida E, Moccia M, Palladino R, *et al*. Chatgpt vs. neurologists: a cross-sectional study investigating preference, satisfaction ratings and perceived empathy in responses among people living with multiple sclerosis. *J Neurol* 2024.
8 Johnson DB, Chandra S, Sosman JA. Immune checkpoint inhibitor toxicity in 2018. *JAMA* 2018;320:1702–3.
9 Brahmer JR, Abu-Sbeih H, Ascierto PA, *et al*. Society for Immunotherapy of cancer (SITC) clinical practice guideline on immune checkpoint inhibitor-related adverse events. *J Immunother Cancer* 2021;9:e002435.
10 Brahmer JR, Lacchetti C, Schneider BJ, *et al*. Management of immune-related adverse events in patients treated with immune checkpoint inhibitor therapy: American Society of Clinical Oncology clinical practice guideline. *JCO* 2018;36:1714–68.
11 Thompson JA, Schneider BJ, Brahmer J, *et al*. NCCN guidelines insights: management of Immunotherapy-related toxicities, version 1.2020. *J Natl Compr Canc Netw* 2020;18:230–41.
12 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159.
13 Zhang Y, Dong Y, Mei Z, *et al*. Performance of large language models on benign prostatic hyperplasia frequently asked questions. *Prostate* 2024;84:807–13.
14 El Haj M, Boutoleau-Bretonnière C, Gallouj K, *et al*. Chatgpt as a diagnostic aid in Alzheimer's disease: an exploratory study. *J Alzheimers Dis Rep* 2024;8:495–500.
15 Sciberras M, Farrugia Y, Gordon H, *et al*. Accuracy of information given by ChatGPT for patients with inflammatory bowel disease in relation to ECCO guidelines. *J Crohns Colitis* 2024;2024:jjae040.
16 Iannantuono GM, Bracken-Clarke D, Karzai F, *et al*. Comparison of large language models in answering Immuno-oncology questions: a cross-sectional study. *Oncologist* 2024;29:407–14.
17 Ozgor F, Caglar U, Halis A, *et al*. Urological cancers and ChatGPT: assessing the quality of information and possible risks for patients. *Clin Genitourin Cancer* 2024;22:454–7.
18 Lee JC, Hamill CS, Shnayder Y, *et al*. Exploring the role of artificial intelligence Chatbots in preoperative counseling for head and neck cancer surgery. *Laryngoscope* 2024;134:2757–61.
19 McGowan A, Gui Y, Dobbs M, *et al*. Chatgpt and bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Res* 2023;326:115334.