Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 9: 1–7

DOI: 10.1177/20552076231221620
journals.sagepub.com/home/dhj

Sage

Feasibility and acceptability of ChatGPT generated radiology report summaries for cancer patients

Eric M Chung[1,2], Samuel C Zhang[1,2], Anthony T Nguyen[1,2], Katelyn M Atkins[1,2], Howard M Sandler[1,2] and Mitchell Kamrava[1,2]

Abstract

Objective: Patients now have direct access to their radiology reports, which can include complex terminology and be difficult to understand. We assessed ChatGPT's ability to generate summarized MRI reports for patients with prostate cancer and evaluated physician satisfaction with the artificial intelligence (AI)-summarized report.

Methods: We used ChatGPT to summarize five full MRI reports for patients with prostate cancer performed at a single institution from 2021 to 2022. Three summarized reports were generated for each full MRI report. Full MRI and summarized reports were assessed for readability using Flesch-Kincaid Grade Level (FK) score. Radiation oncologists were asked to evaluate the AI-summarized reports via an anonymous questionnaire. Qualitative responses were given on a 1–5 Likert-type scale. Fifty newly diagnosed prostate cancer patient MRIs performed at a single institution were additionally assessed for physician online portal response rates.

Results: Fifteen summarized reports were generated from five full MRI reports using ChatGPT. The median FK score for the full MRI reports and summarized reports was 9.6 vs. 5.0, ($p < 0.05$), respectively. Twelve radiation oncologists responded to our questionnaire. The mean [SD] ratings for summarized reports were factual correctness (4.0 [0.6], understanding 4.0 [0.7]), completeness (4.1 [0.5]), potential for harm (3.5 [0.9]), overall quality (3.4 [0.9]), and likelihood to send to patient (3.1 [1.1]). Current physician online portal response rates were 14/50 (28%) at our institution.

Conclusions: We demonstrate a novel application of ChatGPT to summarize MRI reports at a reading level appropriate for patients. Physicians were likely to be satisfied with the summarized reports with respect to factual correctness, ease of understanding, and completeness. Physicians were less likely to be satisfied with respect to potential for harm, overall quality, and likelihood to send to patients. Further research is needed to optimize ChatGPT's ability to summarize radiology reports and understand what factors influence physician trust in AI-summarized reports.

Keywords

artificial intelligence, general, digital health, general, cancer, disease, oncology, medicine, radiology, medicine, electronic, general

Submission date: 8 May 2023; Acceptance date: 30 November 2023

Introduction

Large language models (LLMs) are a type of artificial intelligence (AI) that can simplify complex text and generate human-like responses. ChatGPT (OpenAI, San Francisco, CA) is a type of LLM that has gained widespread popularity after its release to the public in November 2022. ChatGPT is trained on vast amounts of text data from various sources

[1]Department of Radiation Oncology, Cedars-Sinai Medical Center, Los Angeles, CA, USA
[2]Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA

Corresponding author:
Eric M Chung, Department of Radiation Oncology, Cedars-Sinai Medical Center, 8700 Beverly Blvd., Los Angeles, CA 90048, USA.
Email: eric.chung@cshs.org

including books, articles, and websites that allow it to communicate in a conversational manner.[1] In response to a text-based prompt, ChatGPT has shown ability to compose essays, write computer code, and even pass difficult exams such as the United States Medical Licensing Exam (USMLE).[2,3] However, one of the most impressive capabilities ChatGPT has demonstrated is the ability to simplify complex text and make it more accessible for a broader audience.

One potential novel application of ChatGPT in the healthcare arena, related to this, is generating simplified radiology report summaries for patients. Following implementation of the information-blocking rule of the twenty-first Century Cures Act in April 2021, patients now have immediate and direct access to their diagnostic imaging reports.[4] Often times, patients access these reports before physicians are able to add interpretive notes, call to disclose results, or schedule appointments.[5] In oncology, these radiology reports can often include serious results such as cancer diagnoses, recurrences, and progression. Additionally, they often include complex terminology can be difficult for patients to understand. For example, one study found that only 4% of radiology reports were readable by the average US adult.[6]

However, composition of high-quality summaries of radiology reports for patients can be time-consuming with the primary responsibility falling on physicians. With increasing physician burnout, that is partially related to the management of electronic health record in-box messages, developing tools to improve in-box message efficiency are needed.[7] Here, we assessed ChatGPT's ability to generate summarized MRI reports for patients with prostate cancer and evaluated the readability and physician satisfaction of an AI-summarized report. Additionally, we assessed current online patient portal response patterns from physicians at our institution.

Materials/methods

ChatGPT (GPT-3.5 series, OpenAI, San Francisco, CA) was used to summarize five unique, de-identified prostate MRI reports from patients with prostate cancer performed at our institution from 2021 to 2022 (Figure 1). A separate cohort of 50 prostate MRIs were used to assess physician online portal response patterns. The five full prostate MRIs were chosen at random. No protected health information was entered into ChatGPT to generate these reports at any time. This study qualified as exempt from requiring regulatory approval according to the Cedars-Sinai Medical Center institutional review board. Informed consent requirement was waived by the Cedars-Sinai Center Institutional Review Board. The MRI reports included different prostate cancer scenarios such as newly diagnosed, recurrent, and active surveillance. All five MRIs were summarized by two radiologists experienced in reading MRIs. Reports usually included multiple findings with different complexities. All MRIs and simplified MRI reports are included in Supplemental Figure 1.

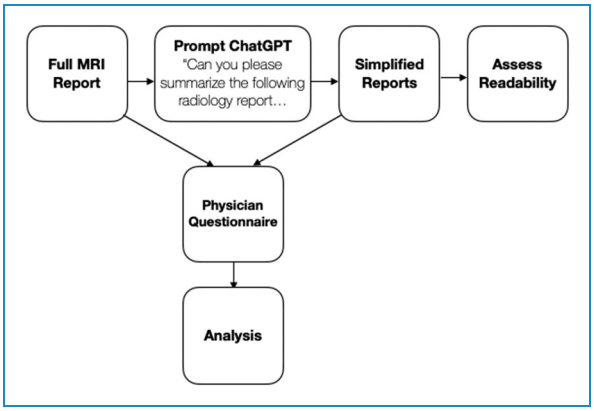

Figure 1. Flowchart of ChatGPT prompt and evaluation by physician questionnaire.

Prompt

Using a standardized prompt, we asked ChatGPT to summarize the full MRI reports into a patient letter at a sixth-grade reading level. The prompt was derived after multiple trials with different prompts including "Please simplify…," "Explain like I'm a 5-year old." The final prompt we used was "Can you please summarize the following radiology report in letter format to a patient at a sixth-grade reading level. Please include size/location of lesion and likelihood of malignancy if applicable."

Readability

To account for variability in text output, we generated three different summarized reports per unique MRI report. We assessed if ChatGPT was able to simplify the full MRI reports at or below a sixth-grade reading level, which is the recommended readability for patient health materials per the American Medical Association (AMA) 2007 Health Literacy Manual.[8] Summarized reports were evaluated for readability using Flesch-Kincaid (FK) Grade Level, which is calculated using the average sentence length and average syllables per word.[9] FK scores of full and summarized reports were compared using Bland-Altman analysis with two-sided tests at a significance level of 0.05.

Physician assessment

Radiation oncologists at a single institution were asked to evaluate the summarized reports with an anonymous questionnaire. In the questionnaire, physicians were shown two full MRI reports that were chosen randomly and three ChatGPT-summarized versions for each full report. For

each summarized report, physicians were asked six questions assessing the following: factual correctness, ease of understanding, completeness, potential for harm, overall quality, and likelihood they would send the report to a patient. A full list of survey questions is shown in Supplemental Figure 2. Qualitative responses were given on a 1 to 5 Likert-type scale ranging from strongly disagree, disagree, neither agree nor disagree, agree, and strongly agree.

Online patient portal response patterns

At our institution, physicians have the option of immediately commenting on patient imaging results through our online patient portal developed by Epic (Epic Systems Corporation, Madison, WI). Fifty consecutive prostate MRIs performed at our institution from newly diagnosed prostate cancer patients referred to our radiation oncology clinic from 2020 to 2022 were assessed to determine current physician online portal response patterns. We evaluated physician response rate to patients, time to response, and length of response via the online patient portal.

Results

Fifteen summarized reports were generated from five full MRI reports using ChatGPT. FK score for each of the five full MRI reports and 15 summarized reports is shown in Table 1. The median FK score for the full reports and summarized reports was 9.6 vs. 5.0. The median difference between FK scores for full and summarized MRI reports was 4.6 (95% CI: 4.14, 5.11, $p < 0.05$). The median word count for the five full MRI reports was 464 words compared to 182 words for the summarized reports ($p < 0.05$).

Out of 18 physicians invited, a total of 12 radiation oncologists completed the questionnaire with varying levels of experience: resident (25%), attending <5 years (33%), attending 5–10 years (17%), and attending >10 years (25%). The mean [SD] ratings for all summarized reports across the six aspects evaluated in the questionnaire were as follows: factual correctness (4.0 [0.6]), ease of understanding (4.0 [0.7]), completeness (4.1 [0.5]), potential for harm (3.5 [0.9]), overall quality (3.4 [0.9]), and likelihood to send to a patient (3.1 [1.1]) (Table 2).

Table 1. Readability scores for full and summarized reports.

	Report #1	Report #2	Report #3	Report #4	Report #5
Full MRI reports (n = 5)					
FK score	9.9	10.3	9.5	8.8	9.2
Word count	336	464	374	566	587
Summarized reports (n = 15)					
Median FK score (range)	3.8 (3.6–6.8)	5.8 (4–6.8)	5 (3.8–6)	5 (2.8–6.2)	4.8 (4.4–5.1)
Median word count (range)	187 (181–230)	125 (119–166)	182 (125–230)	219 (138–221)	190 (182–237)

Table 2. Results of physician questionnaire.

Question	Factual correctness	Ease of understanding	Completeness	Potential for harm	Overall quality	Likelihood to send to patient
All reports (n = 6)						
Mean score (SD)	4.0 (0.6)	4.0 (0.7)	4.1 (0.5)	3.5. (0.9)	3.4 (0.9)	3.1 (1.1)
Summarized reports for MRI #1 (n = 3)						
Mean score (SD)	4.0 (0.7)	4.3 (0.6)	4.1 (0.4)	3.5 (1.0)	3.6 (0.9)	3.1 (1.2)
Summarized reports for MRI #2 (n = 3)						
Mean score (SD)	4.0 (0.7)	3.7 (0.7)	4.1 (0.6)	3.4 (0.7)	3.3 (0.9)	3.1 (1.0)

The majority of respondents agreed or strongly agreed that the summarized reports were factually correct (89%), easy to understand (78%), and complete (93%) (Figure 2). However, fewer respondents agreed or strongly agreed that the reports had low potential for harm (51%), were of high overall quality (53%), and would be sent to a patient (46%). Although less than 3% of respondents disagreed/strongly disagreed with regard to correctness, ease of understanding, and completeness; 14% and 26% of respondents disagreed/strongly disagreed with regard to overall quality and willingness to send to patient, respectively.

Physician responses via the online patient portal are outlined in Table 3. Overall, physicians responded to patients via the online portal for 14/50 (28%) prostate MRIs. Telephone encounters were documented in an additional 4/50 (8%) patient MRIs. The median response time was 1 day (range 0–5). Online patient portal responses by physicians were usually short, with a median word count of 28 words (range 10–76). All physician online portal responses are shown in Supplemental Figure 3.

Discussion

With the recent policy changes related to the twenty-first Century Cures Act, patients now have direct access to their imaging results with even high-stake results such as cancer diagnoses being released immediately.[4] There is now an emphasis on the need for effective communication for oncology physicians when dealing with patient portal-delivered results.[10] Waiting for oncology test results often increases patient anxiety and direct release now cuts down on wait times for receiving results by 7–14 days.[11,12] However, the ambiguity of portal-delivered results can increase anxiety, with patients now seeking information from the internet, family, and friends, which can be sources for potential misinterpretation of results.[13] LLMs such as ChatGPT represent a potential way to compose high-quality summaries of radiology reports for patients. While this application is promising, LLMs are not specifically trained for understanding or accuracy of statements and have the potential to generate false or misleading statements, including "hallucinations."[14–16] Furthermore, the utility of ChatGPT in clinical-decision making capacities has yet to be demonstrated. A recent study assessing ChatGPT's performance in offering cancer treatment recommendations revealed a significant portion of recommendations partially deviated from NCCN guidelines and "hallucinations" were present in 12.5% of outputs.[17]

It is necessary for physician review to validate the responses generated by ChatGPT to ensure accuracy and appropriateness.

Our study demonstrates the feasibility of using ChatGPT to generate summarized radiology reports for prostate MRIs. The AI-generated summaries significantly reduced the readability score to a sixth-grade level, making the information more accessible to the average patient.[8] In addition, our results indicate that current physician online portal response patterns are inconsistent, with only 28% of patients receiving a physician response to their prostate MRI at our institution. Furthermore, physician responses are often short, with a median word count of 28 words, and often do not provide patients with limited information or follow-up instructions (Supplemental Figure 3).

ChatGPT-generated summaries on the other hand have the advantage of not being limited by time, which allows

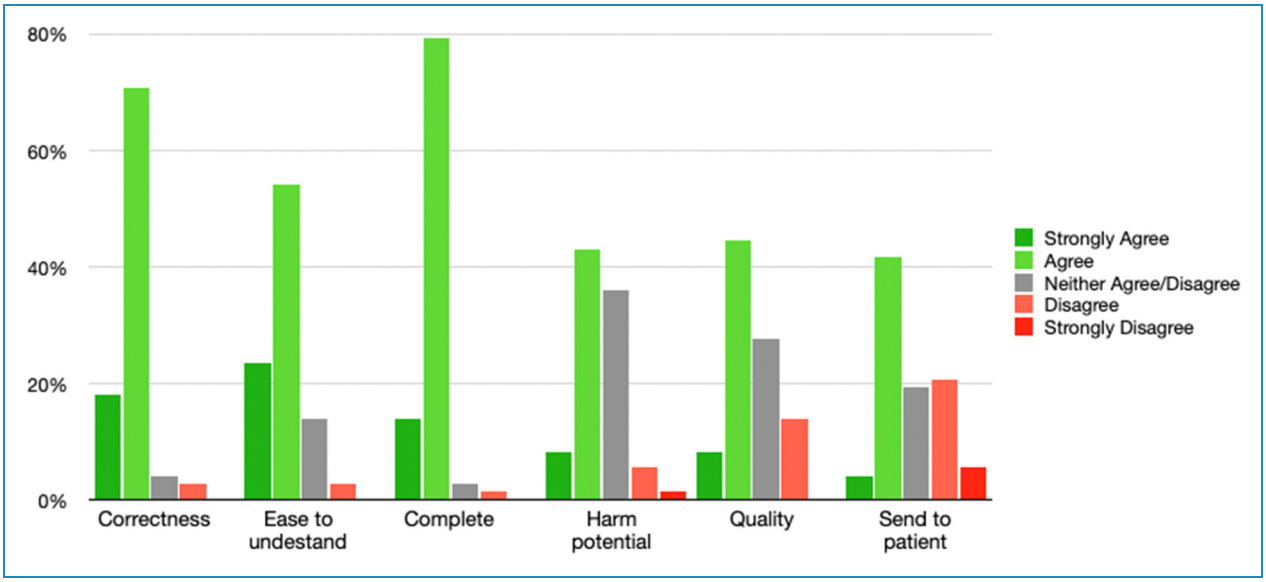

Figure 2. Overall physician rating by question.

Table 3. Online patient portal response patterns (n = 50).

Physician responses	
Online patient portal	14 (28%)
Telephone	4 (8%)
Response time	
Median days (range)	1 (0–5)
Response length	
Median word count (range)	28 (10–76)

it to generate more robust and detailed explanations for patients. Additionally, ChatGPT has demonstrated the ability to generate human-like and empathetic responses. A recent study demonstrated that ChatGPT chatbots exhibited almost a 10 times higher proportion of empathetic vs very empathetic responses compared to physicians in addressing patient questions.[14] Other recent studies have also demonstrated the potential utility of LLMs to generate after visit instructions, follow-ups, and summarizing medical records.[18–20]

In a recent multi-site survey of current patient online portal preferences, of 8139 patients surveyed, the majority (96%) expressed a preference for receiving test results through the online patient portal immediately upon reporting and before being contacted by a health care practitioner.[21] Most respondents reported that reviewing the results had either a positive effect or no effect on their level of worry. However, in a subset of respondents with abnormal results, they experienced additional worry. Despite this, the overwhelming majority (95%) would still rather continue to receive immediately released results through the portal. Other studies have also demonstrated that patients receiving abnormal results are at increased risk for negative emotions, often due to difficulty in interpreting the results in the context of their own health.[22,23] ChatGPT-generated summaries can potentially alleviate some of this worry by providing clearer and more informative explanations of the results. Additionally, by integrating AI-generated summaries into the patient portal, clinicians can better focus on discussing the implications of the results and addressing patient concerns.

In our study, we found that ChatGPT demonstrated an excellent capability to summarize multiple findings in a report and generate a concise, clear, and comprehensive report. Indeed, the majority of physicians found the AI-summarized reports in our study were factually correct, easy to understand, and complete. However, there were concerns regarding potential for harm, overall quality, and likelihood of sending the reports to patients. These concerns may be attributed to several factors that influence physician trust in AI-summarized reports. Firstly, physicians may be hesitant to rely on AI-generated summaries due to concerns about potential inaccuracies, misinterpretations, or omissions of important information.[24,25] Additionally, a known limitation of LLMs like ChatGPT is the potential for "hallucinations," where the model includes findings and information that were not included in the original report.[26,27] In our study, we noticed differences between AI-generated summaries regarding length, level of detail, and content even when given the same prompt. Random results such as these are inherent to LLMs, but can decrease physician trust in AI-generated reports. In a recent study evaluating feasibility of ChatGPT to summarize radiology reports, the authors found that surveyed radiologists were satisfied with the correctness and completeness of ChatGPT-generated reports.[28] However, 51% of radiologists found incorrect passages in the summaries and instances of missed key medical findings, imprecise language, and potential for misinterpretation. Currently, ChatGPT does not have a built-in template for its summarized report generation. In the future, providing ChatGPT with a consistent template with clear instructions for formatting may help ChatGPT generate more reliable and accurate radiology report summaries.

ChatGPT has already demonstrated the potential of LLMs in healthcare with multiple applications being explored ranging from documentation, clinical practice support, education, remote patient monitoring, and medication management to name a few.[29] Although ChatGPT offers significant promise, extensive research will be needed to understand how to best deploy LLMs and address concerns before its widespread implementation. This study emphasizes the need for more efficient communication through online patient portals and potential of AI-generated summaries in addressing challenges faced by both patients and physicians. However, the acceptability of this technology from a patient perspective must be considered, as concerns of automation and depersonalization of care could lead to resistance in adopting this technology.[20] Future research should focus on eliciting patient perspectives and exploring factors that influence physician trust in AI-generated radiology report summaries.

Our study has several limitations including relatively small sample size and a single institution study, which may limit the generalizability of our findings. Currently, we do not envision this technology being used without human interaction and further large sample research will be needed to validate the use of LLMs to summarize radiology reports. Second, we focused exclusively on prostate cancer MRI reports, and the results may not be directly applicable to other types of radiology reports or cancer diagnoses. Third, the anonymous questionnaire used in

our study may be subject to response bias, as physicians with strong opinions on AI-generated summaries may be more likely to participate. Additionally, our study found that despite high ratings regarding summaries being "factually correct, easy to understand, and complete" there were concerns regarding "potential for harm, overall quality, and likelihood of sending the report to patients." Our study is unable to identify potential causes for this discrepancy, and it is beyond the current scope of our study. Finally, it's possible the physicians communicated results with patients and did not document this in the electronic medical record.

Conclusion

In conclusion, this study demonstrates the novel feasibility of using ChatGPT to generate patient-friendly summaries of prostate cancer MRI reports, significantly reducing the readability score to a sixth-grade level. This has the potential to make radiology reports more accessible and understandable for patients, particularly in the context of online patient portals, where patients often receive results directly without physician guidance. The majority of physicians agreed that AI-generated summaries were factually correct, easy to understand, and complete. However, there were concerns regarding potential for harm, overall quality, and willingness to send the reports to patients. Further studies will be needed to understand patient perspectives as well as factors that influence physician trust in AI-generated summaries.

Contributorship: EC and MK researched literature and conceived the study. EC and MK was involved in study development, data collection and data analysis. EC wrote the first draft of the manuscript. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

Data sharing statement: All data generated and analyzed during this study are included in this published article (and its supplementary information files).

Declaration of conflicting interests: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article: KA—honoraria Onclive 2021. HS—consulting/advisory role at Janssen, other relationship at Caribou Publishing. MK—serves on the board of directors for American Brachytherapy Society and Association for Directors of Radiation Oncology Programs, reports advisory board fees for Theragenics, serves on the Data and Safety Monitoring Board for Alessa Therapeutics and GammaTile, and receives book royalties from Springer Publishing.

Ethics approval: This study qualified as exempt from requiring regulatory approval according to the Cedars-Sinai Medical Center Institutional Review Board. Informed consent requirement was waived by the Cedars-Sinai Center Institutional Review Board.

Funding: The author(s) received no financial support for the research, authorship, and/or publication of this article.

Guarantor: M.K.

ORCID iD: Eric M Chung https://orcid.org/0000-0001-9651-3801

Supplemental material: Supplemental material for this article is available online.

References

1. Floridi L and Chiriatti M. GPT-3: its nature, scope, limits, and consequences. *Minds Mach* 2020; 30: 681–694.
2. Kitamura FC. ChatGPT is shaping the future of medical writing but still requires human judgment. *Radiology* 2023; 307: e230171.
3. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023; 2: e0000198.
4. 21st Century Cures Act: Interoperability, Information Blocking, and the ONC Health IT Certification Program, US Department of Health and Human Services, 2020
5. Giardina TD, Baldwin J, Nystrom DT, et al. Patient perceptions of receiving test results via online portals: a mixed-methods study. *J Am Med Inform Assoc* 2018; 25: 440–446.
6. Martin-Carreras T and Kahn CE Jr. Coverage and readability of information resources to help patients understand radiology reports. *J Am Coll Radiol* 2018; 15: 1681–1686
7. Murphy DR, Meyer AN, Russo E, et al. The burden of inbox notifications in commercial electronic health records. *JAMA Intern Med* 2016; 176: 559–560.
8. Osborn CY, Weiss BD, Davis TC, et al. Measuring adult literacy in health care: performance of the newest vital sign. *Am J Health Behav* 2007; 31: S36–S46.
9. Kincaid JP. *United States. National technical Information S: derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy enlisted personnel*. Springfield, VA, U.S.: Department of Commerce, National Technical Information Service, 1975.
10. Hahne J, Carpenter BD, Epstein AS, et al. Communication skills training for oncology clinicians after the 21st century cures act: the need to contextualize patient portal-delivered test results. *JCO Oncol Pract* 2023; 19: 99–102.
11. Wiljer D, Leonard KJ, Urowitz S, et al. The anxious wait: assessing the impact of patient accessible EHRs for breast cancer patients. *BMC Med Inform Decis Mak* 2010; 10: 46.
12. Woolen SA, Kazerooni EA, Steenburg SD, et al. Optimizing electronic release of imaging results through an online patient portal. *Radiology* 2019; 290: 136–143.
13. Avdagovska M, Menon D and Stafinski T. Capturing the impact of patient portals based on the quadruple aim and

benefits evaluation frameworks: scoping review. *J Med Internet Res* 2020; 22: e24568.
14. 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: 589–596.
15. Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. *Radiology* 2023; 307: e230163.
16. Singh R, Reardon T, Srinivasan VM, et al. Implications and future directions of ChatGPT utilization in neurosurgery. *J Neurosurg* 2023.
17. Chen S, Kann BH, Foote MB, et al. Use of artificial intelligence chatbots for cancer treatment information. *JAMA Oncol* 2023; 9: 1459–1462.
18. Ayoub NF, Lee YJ, Grimm D, et al. Comparison between ChatGPT and Google search as sources of postoperative patient instructions. *JAMA Otolaryngol Head Neck Surg* 2023; 149: 556–558.
19. Ali SR, Dobbs TD, Hutchings HA, et al. Using ChatGPT to write patient clinic letters. *Lancet Digit Health* 2023; 5: e179–e181.
20. Patel SB and Lam K. ChatGPT: the future of discharge summaries? *Lancet Digit Health* 2023; 5: e107–e108.
21. Steitz BD, Turer RW, Lin CT, et al. Perspectives of patients about immediate access to test results through an online patient portal. *JAMA Netw Open* 2023; 6: e233572.
22. Giardina TD, Modi V, Parrish DE, et al. The patient portal and abnormal test results: an exploratory study of patient experiences. *Patient Exp J* 2015; 2: 148–154.
23. Steitz BD, Wong JIS, Cobb JG, et al. Policies and procedures governing patient portal use at an academic medical center. *JAMIA Open* 2019; 2: 479–488.
24. Geis JR, Brady AP, Wu CC, et al. Ethics of artificial intelligence in radiology: summary of the joint European and north American multisociety statement. *Can Assoc Radiol J* 2019; 70: 329–334.
25. Li Y and James L. Mckibben J: trust between physicians and patients in the E-health era. *Technol Soc* 2016; 46: 28–34.
26. Zhou C, Neubig G, Gu J, et al.: Detecting hallucinated content in conditional neural sequence generation. *Findings of the association for computational linguistics: ACL-IJCNLP*. ACL Anthology, 2021, pp. 1393–1404.
27. Ebrahimi B, Howard A, Carlson DJ, et al. ChatGPT: can a natural language processing tool be trusted for radiation oncology use? *Int J Radiat Oncol Biol Phys* 2023; 116: 977–983.
28. Jeblick K, Schachtner B, Dexl J, et al.: ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. ArXiv arXiv:2212.14882, 2023
29. Cascella M, Montomoli J, Bellini V, et al.: Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst* 2023; 47: 33.