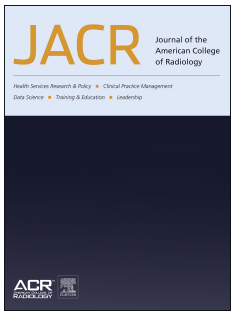
Journal Pre-proof



Enhancing patient communication with Chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions

Emile B. Gordon, M.D., Alexander J. Towbin, M.D., Peter Wingrove, M.D., Umber Shafique, M.D., Brian Haas, M.D., Andrea B. Kitts, M.S., M.P.H., Jill Feldman, M.A., Alessandro Furlan, M.D.

PII: S1546-1440(23)00775-5

DOI: https://doi.org/10.1016/j.jacr.2023.09.011

Reference: JACR 6357

To appear in: *Journal of the American College of Radiology*

Received Date: 24 July 2023

Revised Date: 12 September 2023

Accepted Date: 20 September 2023

Please cite this article as: Gordon EB, Towbin AJ, Wingrove P, Shafique U, Haas B, Kitts AB, Feldman J, Furlan A, Enhancing patient communication with Chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions, *Journal of the American College of Radiology* (2023), doi: https://doi.org/10.1016/j.jacr.2023.09.011.

Enhancing patient communication with Chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions

Emile B. Gordon, M.D.[a], Alexander J. Towbin, M.D.[b], Peter Wingrove, M.D.[a], Umber Shafique, M.D.[c], Brian Haas, M.D.[d], Andrea B. Kitts, M.S., M.P.H.[e], Jill Feldman, M.A.[f], Alessandro Furlan, M.D.[a*]

[a] University of Pittsburgh Medical Center, Department of Radiology
[b] University of Cincinnati, Department of Radiology
[c] Indiana University School of Medicine, Department of Radiology
[d] University of California San Francisco, Department of Radiology and Biomedical Imaging,
[e] Rescue Lung Society
[f] EGFR Resisters

[*]**Corresponding author:** Emile B. Gordon, emile.gordon@duke.edu

Data statement: The author(s) declare(s) that they had full access to all of the data in this study, and the author(s) take(s) complete responsibility for the integrity of the data and the accuracy of the data analysis.

Conflict of interest: The authors declare no financial conflict of interest. Andrea B. Kitts is an Associate Editor at JACR.

Funding: No funding was provided.

Disclosure: During the preparation of this work the author(s) used GPT-4 in order enhance the clarity of select portions of the text. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Author contributions: All authors were responsible for project design, analyzing the data, and manuscript editing. In addition, P.W. conducted most statistical analyses; E.B.G collected the data, analyzed the data, wrote the first draft of the manuscript, and finalized submitted draft.

ChatGPT shows promise in providing accurate, consistent, and relevant responses to patients' imaging-related queries, with prompts further improving responses. However, accuracy was imperfect, and all responses exceeded the recommended reading levels for patient educational materials.

Enhancing patient communication with Chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions

Abstract

Purpose:

To assess ChatGPT's accuracy, relevance, and readability in answering patients' common imaging-related questions and examine the effect of a simple prompt.

Methods:

22 imaging-related questions were developed from categories previously described as important to patients: safety, the radiology report, the procedure, preparation before imaging, meaning of terms, and medical staff. These questions were posed to ChatGPT with and without a short prompt instructing the model to provide an accurate and easy-to-understand response for the average person. Four board-certified radiologists evaluated the answers for accuracy, consistency, and relevance. Two patient advocates also reviewed responses for their utility for patients. Readability was assessed by Flesch Kincaid Grade Level (FKGL). Statistical comparisons were performed using chi-square and paired t-tests.

Results:

264 answers were assessed for both unprompted and prompted questions. Unprompted responses were accurate 83% (218/264) of the time, which did not significantly change for prompted responses (87% [229/264]; P=0.2). The consistency of the responses increased from 72%f (63/88) to 86% (76/88) when prompted (P=0.02). Nearly all responses (99% [261/264]) were at least partially relevant for both question types. Fewer unprompted responses were considered

fully relevant at 67% (176/264), though this increased significantly to 80% when prompted (210/264) (P=0.001). The average FKGL was high at 13.6 [12.9-14.2], unchanged with the prompt (13.0 [12.41-13.60], P=0.2). None of the responses reached the eighth-grade readability recommended for patient-facing materials.

Conclusions:

ChatGPT demonstrates the potential to respond accurately, consistently, and relevantly to patients' imaging-related questions. However, imperfect accuracy and high complexity necessitate oversight before implementation. Prompts reduced response variability and yielded more targeted information but did not improve readability.

Relevance and Application:

ChatGPT has the potential to increase accessibility to health information and to streamline the production of patient-facing educational materials, though its current limitations require cautious implementation and further research.

Introduction

The American College of Radiology (ACR) has prioritized effective patient communications in radiology, encouraging its improvement to lower the barrier of healthcare accessibility (1). Patient concerns are often addressed by direct communication with healthcare workers or by curated educational information. They can turn to a range of sources, including websites maintained by radiology departments or by representative bodies such as www.radiologyinfo.org, which hosts an extensive online repository of descriptions of procedures, common radiology language, and the significance of common imaging findings.

While patients can turn to these reliable information sources and knowledgeable

healthcare team members, accessibility is limited by resources such as time, personnel, and funding.

The recent emergence of ChatGPT, a conversational software developed by OpenAI (San Francisco, California), has garnered attention for its ability to generate human-like text responses to natural language inquiries, which shows promise as an innovative tool to augment patient communication in radiology. This generative pre-trained language model (GPLM) employs deep learning algorithms to analyze text and respond to user inquiries or prompts. ChatGPT could be used to create reference materials or serve to answer patients' questions prior to a radiology examination. It has already been tested in several patient-facing medical contexts, including answering questions related to cardiovascular disease prevention (2) and breast cancer screening recommendations (3). Yet, the role of this language model in addressing patients' imaging-related questions remains unexplored.

The aim of this paper is to assess the accuracy, relevance, and readability of answers provided by ChatGPT to common imaging-related questions. A secondary aim of the paper is to evaluate the influence a patient-directed prompt may have on these parameters.

Methods

The study was conducted in March 2023A list of 22 questions was developed to simulate common patient concerns in radiology. These questions were tailored based on the professional expertise of four board-certified radiologists and existing literature (Table 1), using categories identified by Alarifi et al., who conducted a thematic analysis of online questions from platforms like Yahoo Answers, Reddit, Quora, and Wiki Answers. Our questions focused on key areas identified in the study: safety, radiology reports, procedures, preparation before imaging,

meaning of terms, and medical staff interactions (4). Questions about the cost of imaging were excluded.

To assess for consistency and account for variability in responses, every question was posed to ChatGPT (version 3.5) (https://chat.openai.com) independently three times, within unique chat instances. The process was repeated and addended with a structured prompt to assess the modifying effect of a short patient-directed prompt: “Provide an accurate and easy-to-understand response that is suited for an average person.”

The output of both unprompted and prompted questions was independently assessed for accuracy and relevance by four board-certified radiologists who reviewed the responses sequentially. Accuracy was assessed based on the readers' knowledge and available resources, including RadiologyInfo™, the patient-facing public information website operated by the Radiological Society of North America (RSNA), and the ACR (5). Relevance was assessed by addressing the question, "Is the response relevant and helpful to the average patient?" Responses were subsequently graded as 1) not relevant or helpful, 2) partially relevant or helpful, or 3) fully relevant or helpful. Each radiologist also estimated the consistency of the three responses question, labeling the set of answers as either "consistent" or "inconsistent."

The readability of the answers was assessed with the Flesch Kincaid grade level (FKGL), calculated using an online tool (Readability analyzer: https://datayze.com/readability-analyzer).

Results for accuracy, relevance, and consistency were considered independent and pooled. Statistical comparisons between unprompted and prompted responses for these were performed with chi-square tests. Given the study's exploratory focus, which was not intended for immediate clinical implementation, we assessed the risk of Type I errors as relatively modest but Type II errors as more problematic in the context of a rapidly evolving field. As such, we opted

against employing a Bonferroni correction. Unpaired Student *t* tests were used to compare readability and the average response character count. Response character length was calculated with the Excel LEN function. Statistical analyses were conducted in Stata 14.2. Graphs were created in Graphpad Prism 9.0. Heatmaps were created with *R* (*R* Core Team, 2023) using the heatmap package (v1.0.12; Kolde, 2019).

For an exploratory analysis, we engaged two patient advocates to provide a lay perspective on the utility of both prompted and unprompted responses and to identify if there was a preference between the two. The utility was measured on a 1-3 scale where 1 is not relevant or helpful, and 3 is fully relevant and helpful.

Results

22 questions were posed to ChatGPT three times each, both with and without an accompanying prompt, resulting in a total of 132 independent responses. These responses were independently evaluated by four-board certified radiologists leading to a total of 528 evaluations (264 each for unprompted and prompted questions). The variability in the grading of accuracy and consistency among these radiologists is illustrated in Figure S1. The radiologists also performed 176 additional evaluations for consistency across the three outputs for each question, resulting in 88 evaluations each for unprompted and prompted questions.

The consistency and accuracy of responses are summarized in Table 2. 83% (218/264) of the unprompted responses were assessed as accurate. There was no statistically significant difference in the percentage of accurate answers between unprompted and prompted questions (83% vs. 87%, P=0.2). ChatGPT answered questions for most topics accurately (range of 79% to 96%), except for the "safety" topic, which had a comparably low accuracy rating at 71%

(34/48). ChatGPT answered 72% (63/88) of the three repeated questions consistently, which significantly increased to 86% (76/88) after adding the patient-directed prompt (P=0.02).

The radiologist reviewers rated very few responses as irrelevant to the question asked. 98.5 % (260/264) and 98.8% (261/264) of answers were determined to be at least partially relevant for unprompted and prompted questions, respectively (Table 2). Only 67% (176/264) of unprompted responses were considered fully relevant. When prompted, the number of fully relevant responses increased significantly to 80% (210/264) (P=0.001).

The readability of the responses was assessed by Flesch Kincaid grade level (FKGL), summarized in Figure 1. The average FKGL for unprompted and prompted responses were not significantly different, despite the request to tailor the response to an average person, measuring 13.6 [12.9-14.2], and 13.0 [12.41-13.60], respectively (P=0.2). None of the responses to either unprompted or prompted questions were at or below the eighth-grade reading level generally recommended for patient-facing health educational materials (Fig 1B). A minority of all responses reached below a twelfth-grade level for both unprompted (30% [20/66]) and prompted (41% [27/66]) questions.

The average question length was 1145 [CI 1006-1284] total characters, which was unchanged when questions were prompted (1147 [CI 1035-1258] total characters; P=0.98).

Our initial findings of the lay perspective on the utility of responses are divergent from that of our experts (Table 4). Although the two patient advocates deemed most responses to be at least partially relevant and helpful (92-97% [122-128/132]), only a minority of unprompted (42%) and a slight majority of prompted responses (57%) were considered fully relevant. When given a choice between unprompted or prompted responses, the patient advocates showed a preference for the latter 71% [31/44] (versus 30% [13/44]).

Discussion

Our results demonstrate that the popular online tool, ChatGPT has the potential to provide accurate and relevant answers to patient-directed imaging-related questions. These findings contribute to the growing body of literature that highlights the potential of ChatGPT to automate time-consuming tasks within the healthcare setting. Automating the development of patient health educational materials and providing on-demand access to medical questions holds great promise to improve patient access to health information.

While the accuracy, consistency, and relevance of the ChatGPT responses to imaging-related questions are impressive for a GPLM, they are imperfect; by clinical standards, the frequency of inaccurate statements that we observed precludes its use without careful human supervision or review.

Our results also show that the readability of the ChatGPT responses far exceeded the recommended eighth-grade level for patient health educational materials (6, 7). The ability to understand health information presented to patients is crucial for their capacity to make informed medical decisions. As it currently stands, the high complexity of the responses clouds the promise of true patient access to health information. However, the higher levels of complexity exhibited by ChatGPT's output are also observed for most patient-facing resources that are currently available across medical fields, including within radiology, which are the likely sources for the model's pre-training (8, 9).

The readability of the responses did not change significantly when accompanied by a prompt to provide a response to an average person. The lack of improvement may indicate a resistance to the model's ability to tailor its readability to recommended levels, which would be a deviation from its pre-trained knowledge, or could represent a deficiency of our prompt.

Understanding and addressing these limitations is essential to effectively increase the accessibility of ChatGPT and could potentially involve the use of more detailed prompting.

Prompts play a critical role in affecting the output of GPLMs by providing context and serving as a source of external knowledge that the model can effectively leverage (10). They can profoundly affect the output, possibly even dissuading the model from reaching an accurate response that it would have otherwise concluded (11). While our prompt did not significantly change the percentage of accurate answers, it elicited a significant increase in fully relevant responses, potentially highlighting the value of a carefully crafted prompt to provide more targeted information. Additionally, our prompt may have introduced unmeasured differences that were noticed by our patient advocates, who preferred the responses with additional prompting. The efficacy of proper prompt engineering should be evaluated in the medical context, which includes patient involvement for patient-facing materials, to optimize the output of accurate and relevant medical information.

The layperson's assessment of response utility was unexpectedly critical of responses both with and without a prompt, assessing fewer than half as fully relevant/helpful. This contrasts with physicians' evaluations, although we did not tailor this secondary aim for statistical comparison. This discrepancy underscores the need for more inclusive evaluations of patient-facing materials, challenging the approach of using only expert reviewers.

There are several limitations to our study. The rapidly evolving nature of technology introduces unpredictability in evaluating the platform's performance, which may change in the future. Additionally, radiologists —not patients— wrote the questions and evaluated the answers, which does not fully capture the unpredictable variety of ways patients seek, consume, or understand medical information. Variability in question context and phrase is expected to affect

the output just as prompts do; a straightforward example is that the model responds in the same language that the question was asked. Additionally, it is unknown how ChatGPT would perform when questions are posed and responded to in a language other than English. Furthermore, some areas of knowledge may have been less accessible or represented during the model's pre-training, which may lead to less accurate responses; for example, while overall accurate, our study showed that certain topics such as "safety" could potentially be less accurate or relevant. Additionally, as an exploratory study, extrapolation of the findings is limited; the inherent stochastic output of GPLMs and the tendency to "confabulate", or confidently produce factually incorrect responses, may become more clinically apparent in larger cohorts.

In conclusion, our exploratory analysis underscores the potential of ChatGPT to streamline time-consuming tasks in patient health education. However, the immediate application of this system necessitates a cautious approach and further investigation. It is vital to tackle challenges related to readability and mitigate the risks of presenting potentially misleading information to patients. Exploring strategies such as effective, prompt engineering will contribute to optimizing ChatGPT's output, ensuring its safety and effectiveness for patient use.

Summary statement:

ChatGPT shows promise in providing accurate, consistent, and relevant responses to patients' imaging-related queries, with prompts further improving responses. However, accuracy was imperfect, and all responses exceeded the recommended reading levels for patient educational materials.

Take Home Points

- ChatGPT provided an accurate response to patients' common imaging-related questions 83% of the time and 87% when asked with a simple prompt, although this difference was not statistically significant (P=0.2).
- Although almost always partially relevant (99%), the proportion of ChatGPT responses considered fully relevant significantly rose from 67% to 80% (P=0.001) when a simple prompt accompanied the questions. Prompting also improved the response consistency from 72% to 86% (P=0.02).
- ChatGPT responses were uniformly complex, with no response reaching the recommended eighth-grade level for patient-facing materials (average Flesch Kincaid grade level of 13.6 unprompted and 13.0 prompted [P=0.2]).

References

1. Siewert B, Bruno MA, Fleishon HB, et al. Summary of the 2022 ACR Intersociety Meeting. Journal of the American College of Radiology 2023;20(5):479–486.

2. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. JAMA [Internet] 2023 [cited 2023 Feb 24]; Available from: https://jamanetwork.com/journals/jama/fullarticle/2801244

3. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT. Radiology

[Internet] Radiological Society of North America; 2023 [cited 2023 Apr 5]; Available from: https://pubs.rsna.org/doi/10.1148/radiol.230424

4. Alarifi M, Patrick T, Jabour A, Wu M, Luo J. Understanding patient needs and gaps in radiology reports through online discussion forum analysis. Insights Imaging 2021;12(1):50.

5. RadiologyInfo [Internet]. Radiologyinfo.org 2023 [cited 2023 Mar 20]. Available from: https://www.radiologyinfo.org/en

6. Doak CC, Doak LG, Root JH. Teaching Patients with Low Literacy Skills. AJN The American Journal of Nursing 1996;96(12):16M.

7. Agency for Healthcare Research and Quality. Assess, Select, and Create Easy-to-Understand Materials: Tool #11 [Internet]. 2020. Available from: https://www.ahrq.gov/health-literacy/improve/precautions/tool11.html

8. Hansberry DR, Agarwal N, Baker SR. Health Literacy and Online Educational Resources: An Opportunity to Educate Patients. American Journal of Roentgenology American Roentgen Ray Society; 2015;204(1):111–116.

9. Rooney MK, Santiago G, Perni S, et al. Readability of Patient Education Materials From High-Impact Medical Journals: A 20-Year Analysis. Journal of Patient Experience SAGE Publications Inc; 2021;8:2374373521998847.

10. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. Drazen JM, Kohane IS, Leong T-Y, editors. N Engl J Med 2023;388(13):1233–1239.

11. Zuccon G, Koopman B. Dr ChatGPT, tell me what I want to hear: How prompt knowledge impacts health answer correctness [Internet]. arXiv; 2023 [cited 2023 Apr 18]. Available from: http://arxiv.org/abs/2302.13793

Table and Figure legends:

Table 1: Common imaging-related topics and patient questions

Table 2: Accuracy and consistency of responses

Table 3: Responses assessed as at least partially or fully relevant.

Table 4: Utility of responses, as assessed by patient advocates.

Figure 1: Readability of ChatGPT responses to questions accompanied without (–) or with (+) a prompt. A) Violin plots of Flesch Kincaid grade level (FKGL) show the distribution of responses with median and IQR; striped lines at FKGL=12 and FKGL=8 for reference, the latter of which is the upper limit of the recommended readability of patient-facing educational materials B) Heatmap of the FKGL for ChatGPT responses by topic and prompt status. Each row represents a question, and each cell represents an individual response to questions posed in triplicate.

Figure S1: Individual reviewer scores of ChatGPT responses to questions accompanied without or with a prompt. A) Heatmap of accuracy scores assigned by reviewers for ChatGPT responses by topic and prompt status. Each row represents a question, and each cell represents an individual response to questions posed in triplicate. B) Heatmap of the relevance scores assigned by reviewers for ChatGPT responses by topic and prompt status. Each row represents a question, and each cell represents an individual response to questions posed in triplicate.

Table 1: Common imaging-related topics and patient questions

Topic	Question
Report	*1. What does "clinical correlation" mean in a radiology report?* *2. What does it mean if a mass is "stable' in the radiology report?* *3. In the radiology report, my liver is described as "steatotic". What does it mean?* *4. What does it mean if my tumor is "hypermetabolic" on my PET-CT scan?*
Safety	*5. Do X-rays or CT scans cause cancer?* *6. What are the risk of MRI during pregnancy?* *7. What are the risks of Gadolinium deposition?* *8. I have an allergy to shellfish. Can I get iodine CT contrast?*
Procedure	*9. How long does it take to obtain a brain MRI?* *10. What does it feel like to get IV contrast?* *11. What is it like to be inside an MR machine?* *12. What are the risks of an imaging-guided renal biopsy?*
Preparation	*13. Why do I need to drink contrast before some CTs but not others?* *14. Why do I need intravenous contrast for my CT scan?* *15. What can I do to reduce my anxiety during an MRI scan?* *16. Why do I need to isolate myself from others after some but not other nuclear medicine exams?*
Meaning	*17. What is the difference between an open MRI and a closed MRI?* *18. What is the difference between a CT and an MRI?* *19. What is the difference between a CT and an ultrasound?* *20. How is a blocked vessel opened in interventional radiology?*
Medical Staff	*21. Who is a radiologist?* *22. Who can answer questions about my radiology report?*

Table 2: Accuracy and consistency of responses

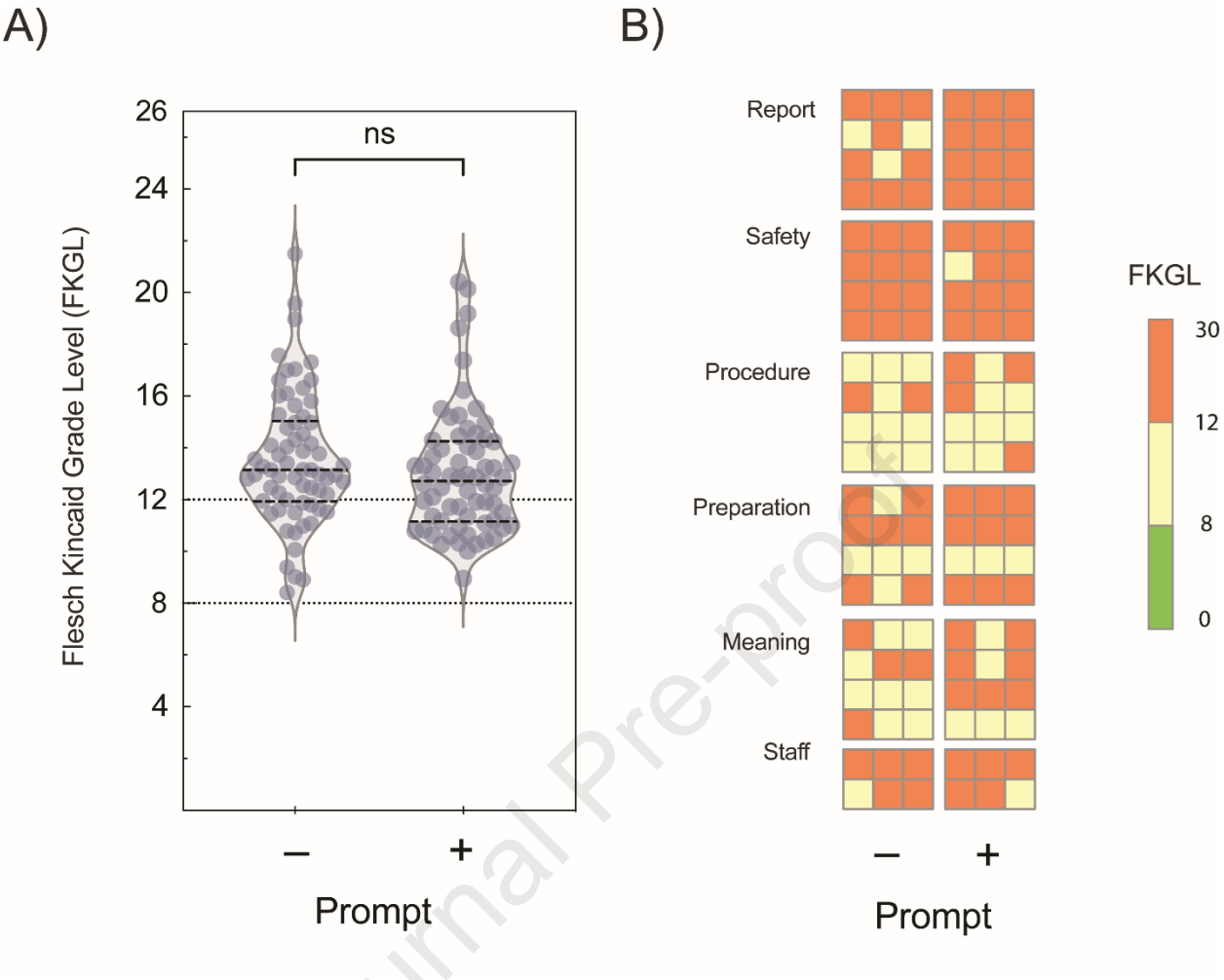
	Consistency (%, [n/N])			**Accuracy** (%, [n/N])		
	No prompt	**Prompted**	*P*	**No prompt**	**Prompted**	*P*
All questions	71.6 (63/88)	86.4 (76/88)	0.016*	82.6 (218/264)	86.7 (229/264)	0.2
Report	87.5 (14/16)	100 (16/16)	0.1	89.6 (43/48)	85.4 (41/48)	0.5
Safety	62.5 (10/16)	75 (12/16)	0.4	70.8 (34/48)	81.3 (39/48)	0.2
Procedure	62.5 (10/16)	81.3 (13/16)	0.2	87.5 (42/48)	91.7 (44/48)	0.5
Preparation	75 (12/16)	87.5 (14/16)	0.4	79.2 (38/48)	77.1 (37/48)	0.8
Meaning	62.5 (10/16)	87.5 (14/16)	0.1	79.2 (38/48)	91.7 (44/48)	0.08
Medical Staff	87.5 (7/8)	87.5 (7/8)	1.0	95.8 (23/24)	100.0 (24/24)	0.3

Table 3: Responses assessed as at least partially or fully relevant

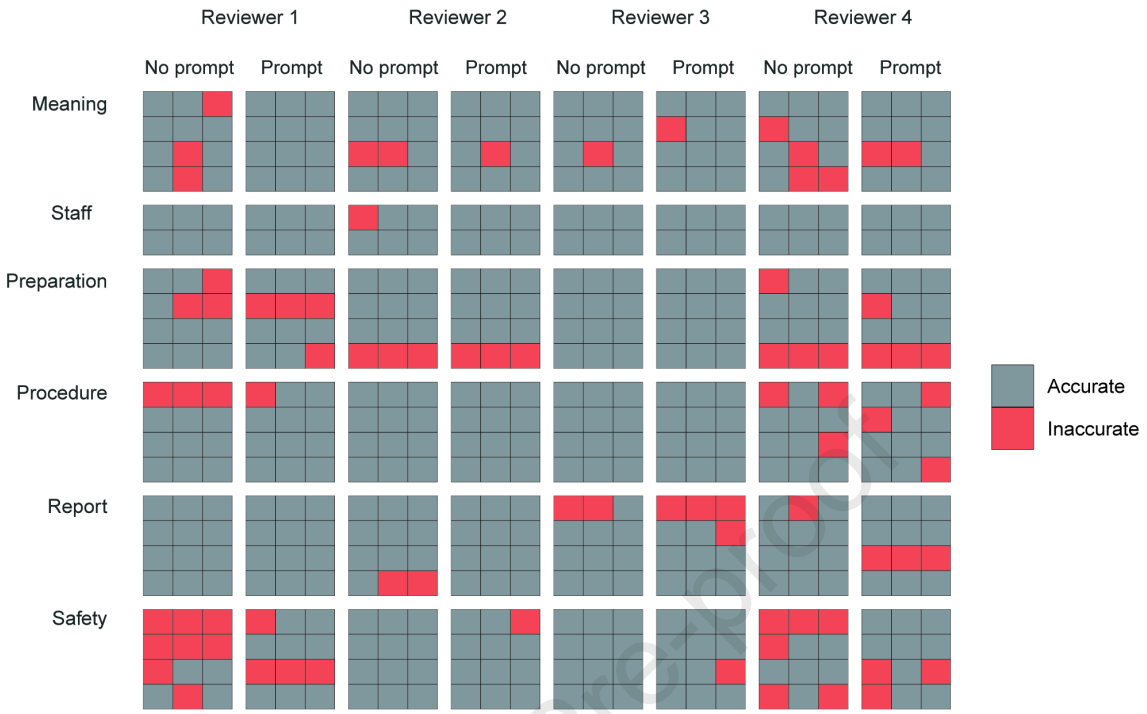
	At least partially relevant (%, [n/N])			**Fully relevant** (%, [n/N])		
	No prompt	**Prompt**	*P*	**No prompt**	**Prompt**	*P*
All questions	98.5 [260/264]	98.9 [261/264]	0.70	66.7 [176/264]	79.6 [210/264]	0.001**
Report	100.0 [48/48]	100.0 [48/48]		77.1 [37/48]	91.7 [44/48]	0.049*
Safety	100.0 [48/48]	100.0 [48/48]		50.0 [24/48]	64.6 [31/48]	0.15
Procedure	97.9 [47/48]	100.0 [48/48]	0.32	70.8 [34/48]	79.2 [38/48]	0.35
Preparation	95.8 [46/48]	93.8 [45/48]	0.65	64.6 [31/48]	68.8 [33/48]	0.67
Meaning	97.9[47/48]	100 [48/48]	0.32	54.2 [26/48]	91.7[44/48]	0.001***
Medical Staff	100 [24/24]	100 [24/24]		100 [24/24]	83.3 [20/24]	0.037*

Table 4: Utility of responses, as assessed by patient advocates.

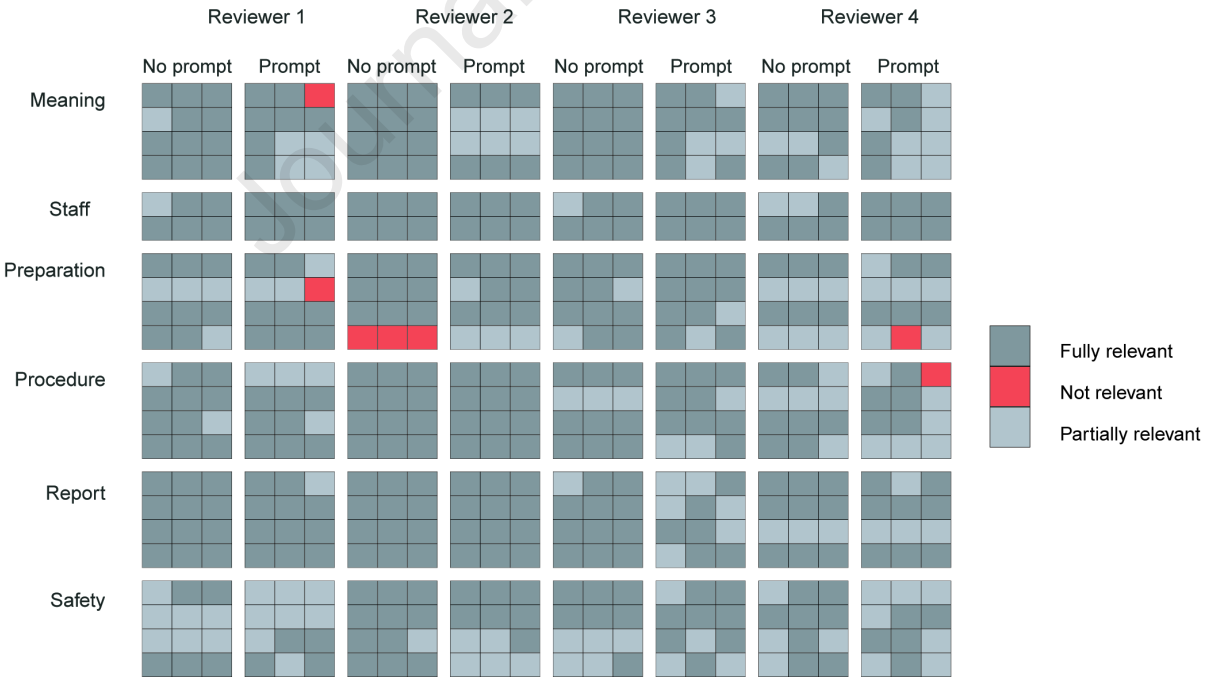
	At least partially relevant/helpful (%, [n/N])		**Fully relevant/helpful** (%, [n/N])	
	No prompt	**Prompt**	**No prompt**	**Prompt**
All questions	92 [122/132]	97 [128/132]	42 [55/132]	57 [75/132]
Report	96 [23/24]	100 [24/24]	54 [13/24]	75 [18/24]
Safety	92 [22/24]	96 [23/24]	38 [9/24]	58 [14]
Procedure	96 [22/24]	100 24/24	50 [12/24]	63 [15/24]
Preparation	83 [20/24]	96 [23/24]	29 [7/24]	38 [9/24]
Meaning	92 [22/24]	100 [24/24]	42 [10/24]	63 [15/24]
Medical Staff	100 [12/12]	83 [10/12]	33 [4/12]	33 [4/12]



A) Accuracy scores



B) Relevance scores



Declaration of interests

☐ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☒ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

One of our co-authors, Andrea Borondy Kitts, is an Associate Editor at JACR.

- ChatGPT provided an accurate response to patients' common imaging-related questions 83% of the time and 87% when asked with a simple prompt, although this difference was not statistically significant (P=0.2).
- Although almost always partially relevant (99%), the proportion of ChatGPT responses considered fully relevant significantly rose from 67% to 80% (P=0.001) when a simple prompt accompanied the questions. Prompting also improved the response consistency from 72% to 86% (P=0.02).
- ChatGPT responses were uniformly complex, with no response reaching the recommended eighth-grade level for patient-facing materials (average Flesch Kincaid grade level of 13.6 unprompted and 13.0 prompted [P=0.2]).