

Assessing ChatGPT's educational potential in lung cancer radiotherapy: A readability, clinician, and patient evaluation

Cedric Richlitzki, Sina Mansoorian, Lukas Käsmann, Mircea Gabriel Stoleriu, Julia Kovacs, Wulf Sienel, Diego Kauffmann-Guerrero, Thomas Duell, Nina Sophie Schmidt-Hegemann, Claus Belka, Stefanie Corradini, Chukwuka Eze

Submitted to: JMIR Cancer
on: December 08, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 22

0..... 22

0..... 22

0..... 22

0..... 22

0..... 22

0..... 22

0..... 22

0..... 22

Figures 23

Figure 0..... 24

Figure 0..... 25

Figure 0..... 26

Figure 0..... 27

Assessing ChatGPT's educational potential in lung cancer radiotherapy: A readability, clinician, and patient evaluation

Cedric Richlitzki¹; Sina Mansoorian¹; Lukas Käsmann¹; Mircea Gabriel Stoleriu²; Julia Kovacs³; Wulf Sienel³; Diego Kauffmann-Guerrero^{4,5}; Thomas Duell²; Nina Sophie Schmidt-Hegemann¹; Claus Belka^{1,6,7,8}; Stefanie Corradini¹; Chukwuka Eze¹

¹ Department of Radiation Oncology, University Hospital LMU Munich DE

² Asklepios Lungenfachkliniken GmbH, Asklepios Fachkliniken Munich - Gauting Munich DE

³ Division for Thoracic Surgery Munich, Ludwig-Maximilians-University of Munich (LMU) and Asklepios Medical Center, Munich, Germany. Munich DE

⁴ Department of Medicine V, University Hospital, Ludwig-Maximilians-University Munich, Munich, Germany Munich DE

⁵ Comprehensive Pneumology Center Munich (CPC-M), Member of the German Center for Lung Research (DZL), Munich, Germany Munich DE

⁶ German Cancer Consortium (DKTK), partner site Munich; and German Cancer Research Center (DKFZ) Heidelberg DE

⁷ Comprehensive Pneumology Center Munich (CPC-M), Member of the German Center for Lung Research (DZL) Munich DE

⁸ Bavarian Cancer Research Center (BZKF) Munich DE

Corresponding Author:

Cedric Richlitzki

Department of Radiation Oncology, University Hospital LMU
Marchioninistrasse 15, 81377 Munich, Germany
Munich
DE

Abstract

Background: Artificial intelligence models like ChatGPT have advanced significantly, with GPT-4o offering improved accuracy and contextual understanding. In healthcare, ChatGPT provides accessible explanations of complex medical concepts, aiding patient education and reducing clinician workload. It is particularly effective in simplifying medical jargon, addressing patient questions, and fostering engagement. However, limitations include misinformation risks, outdated data, and potential biases. For lung cancer, the leading cause of cancer-related deaths globally, patients require reliable, comprehensive, and accessible educational tools, particularly for complex treatments like radiotherapy. ChatGPT offers potential as a supplementary resource to meet these needs, though careful oversight is required to address its shortcomings.

Objective: This study aims to evaluate the educational capabilities and limitations of GPT-4 for patients undergoing radiotherapy for lung cancer, focusing on clinician-led assessments of relevance, accuracy, and completeness, patient-led evaluations of educational content, and a readability analysis to assess response accessibility.

Methods: Eight questions related to lung cancer radiotherapy were posed to GPT-4 (July 2024) via OpenAI's web interface. Responses were assessed for readability using the Modified Flesch Reading Ease (FRE) Formula and the 4th Vienna Formula (WSTF). Six clinicians (two radiation oncologists, two medical oncologists, and two thoracic surgeons) experienced in the treatment of lung cancer rated relevance, correctness, and completeness on a five-point Likert scale (1 = strongly disagree, 5 = strongly agree). Patients evaluated comprehensibility, accuracy, relevance, trustworthiness, and willingness to use ChatGPT for future medical questions during post-radiotherapy follow-up. Data were analyzed using descriptive statistics (median, mean, standard deviation) in Microsoft Excel (version 2410). Figures were created in Python (version 3.8) using Matplotlib, with data structured in Pandas DataFrames for analysis and visualization.

Results: ChatGPT's responses were classified as "very difficult" or "difficult to read" in the readability analysis using the Modified Flesch Reading Ease (FRE) and the 4th Vienna Formula (WSTF) (FRE: 23.36 ± 11.16 , WSTF: 13.81 ± 2.01). Clinicians rated relevance (3.7–4.3), correctness (3.5–4.3), and completeness (3.5–4.2), with ChatGPT's response to the question "What follow-up care is required after radiotherapy for lung cancer?" scoring highest across all dimensions. Thirty consecutive patients (48 – 87 years, median: 66 years) who received radiotherapy for lung cancer rated clarity highly ("easy to understand":

4.4 ± 0.61), but trustworthiness and usability scored lower ("confidence in information": 4.0 ± 0.84). These results highlight ChatGPT's strengths in accessibility and relevance, with room for improvement in trustworthiness and usability.

Conclusions: ChatGPT shows promise as a supplementary tool for patient education in radiation oncology, offering clear and relevant information. However, limitations in completeness and trustworthiness necessitate careful review and supplementation by healthcare professionals. Further advancements and standardized evaluation criteria are essential for its effective integration into clinical practice.

(JMIR Preprints 08/12/2024:69783)

DOI: <https://doi.org/10.2196/preprints.69783>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <https://preprints.jmir.org/preprint/69783>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://preprints.jmir.org/preprint/69783>

Original Manuscript

Original Paper

Assessing ChatGPT's educational potential in lung cancer radiotherapy: A readability, clinician, and patient evaluation

Cedric Richlitzki¹, Sina Mansoorian¹, Lukas Käsmann¹, Mircea Gabriel Stoleriu^{3,7}, Julia Kovacs⁷, Wulf Sienel⁷, Diego Kauffmann-Guerrero^{3,5}, Thomas Duell⁶, Nina Sophie Schmidt-Hegemann¹, Claus Belka^{1,2,3,4}, Stefanie Corradini¹, Chukwuka Eze¹

¹ Department of Radiation Oncology, University Hospital LMU, Munich, Germany

² German Cancer Consortium (DKTK), partner site Munich; and German Cancer Research Center (DKFZ), Heidelberg, Germany

³ Comprehensive Pneumology Center Munich (CPC-M), Member of the German Center for Lung Research (DZL), Munich, Germany

⁴ Bavarian Cancer Research Center (BZKF), Munich, Germany

⁵ Department of Medicine V, University Hospital, Ludwig-Maximilians-University Munich, Munich, Germany

⁶ Asklepios Lungenfachkliniken GmbH, Asklepios Fachkliniken Munich - Gauting, Germany

⁷ Division for Thoracic Surgery Munich, Ludwig-Maximilians-University of Munich (LMU) and Asklepios Medical Center, Munich, Germany.

Corresponding Author: Cedric Richlitzki, Department of Radiation Oncology, LMU University Hospital, LMU Munich, Marchioninistrasse 15, 81377 Munich, Germany. E-mail address: Cedric.Richlitzki@med.uni-muenchen.de

Keywords: AI, Artificial Intelligence, Patient Education, ChatGPT, Lung Cancer, NSCLC, Radiotherapy

Background: Large language models (LLMs) such as ChatGPT are increasingly discussed as potential tools for patient education in healthcare. In radiation oncology, where patients are often confronted with complex medical terminology and complex treatment plans, LLMs may support patient understanding and promote more active participation in care. However, the readability, accuracy, completeness and overall acceptance of LLM-generated medical content remain underexplored.

Objective: This study aimed to evaluate the potential of ChatGPT-4o as a supplementary tool for patient education in the context of lung cancer radiotherapy by assessing the readability, content quality, and perceived usefulness of AI-generated responses from both clinician and patient perspectives.

Methods: Eight frequently asked questions about radiotherapy for lung cancer were developed based on clinical experience from a team of clinicians specialized in lung cancer treatment at a university hospital. The questions were submitted individually to ChatGPT-4o (version as of July 2024) using the prompt: "I am a lung cancer patient looking for answers to the following questions." Responses were evaluated using three approaches: (1) a readability analysis applying the Modified Flesch Reading Ease (FRE) formula for German and the 4th Vienna Formula (WSTF); (2) a multicenter expert evaluation by six multidisciplinary clinicians (radiation oncologists, medical oncologists, and thoracic surgeons) specialized in lung cancer treatment using a 5-point Likert scale to assess relevance, correctness and completeness; and (3) a patient evaluation during the first follow-up appointment after radiotherapy, assessing comprehensibility, accuracy, relevance, trustworthiness and willingness to use ChatGPT for future medical questions.

Results: Readability analysis classified most responses as "very difficult to read" (university level) or "difficult to read" (upper secondary school), likely due to the use of medical language and long sentence structures. Clinician assessments yielded high scores for relevance (mean: 4.5 ± 0.52) and correctness (mean: 4.3 ± 0.65), but completeness received slightly lower ratings (mean: 3.9 ± 0.59). 30 Patients rated the responses positively for clarity (mean: 4.4 ± 0.61) and relevance (mean: 4.3 ± 0.64), but lower for trustworthiness (mean: 3.8 ± 0.68) and usability (mean: 3.7 ± 0.73). No harmful misinformation was identified in the responses.

Conclusions: ChatGPT-4o shows promise as a supplementary tool for patient education in radiation oncology. While patients and clinicians appreciated the clarity and relevance of the information, limitations in completeness, trust and readability highlight the need for clinician oversight and further optimization of LLM-generated content. Future developments should focus on improving

accessibility, integrating real-time readability adaptation and establishing standardized evaluation frameworks to ensure safe and effective clinical use.

Background

Artificial intelligence (AI) has made remarkable progress in recent years, with models like ChatGPT, launched by OpenAI in November 2022, emerging as critical tools in natural language processing. Built on the GPT architecture, ChatGPT has evolved from GPT-1 (2018) to GPT-4o (May 2024), with each iteration improving accuracy, contextual understanding, and versatility, particularly in generating human-like texts. In addition to ChatGPT, other notable large language models (LLMs) include Google's Bard, which excels in generating creative content and integrating real-time data, Meta's LLAMA, tailored for research and non-commercial applications and Anthropic's Claude, which prioritizes safety and ethical AI interactions.

ChatGPT, in its current form, offers notable advantages in the medical field, especially in patient education and communication [1, 2]. It can provide clear explanations of complex medical concepts, answer patient queries, and assist clinicians in creating educational materials [3, 4]. ChatGPT is a powerful tool for enhancing patient understanding and engagement in treatment plans, leveraging its ability to process and generate text from vast datasets.

In healthcare, ChatGPT has found diverse applications [5]. It is particularly effective for patient education, simplifying complex medical jargon into accessible language, and offering support beyond clinical hours. Patients often have follow-up questions about treatment processes, side effects, safety, and treatment design and delivery [6].

These queries can significantly increase staff workload, potentially exacerbating physician burnout and negatively affecting care quality [7]. LLM chatbots like ChatGPT offer a promising solution to mitigate this burden by answering routine patient inquiries and reducing the workload on healthcare professionals. Furthermore, its ability to simulate conversations enables interactive patient education, improving comprehension and fostering a more informed and empowered patient community [4, 8, 9].

However, ChatGPT has limitations. It lacks critical thinking and contextual judgment, which can lead to misinformation or factually inaccurate responses, commonly referred to as "hallucinations" [10, 11]. Other concerns include the risk of providing incorrect medical information, breaches of patient privacy, generating falsified documents or images, overdependence on AI in medical education, and reinforcing existing biases [12]. Additionally, the current version relies on data up to 2021, limiting its ability to provide up-to-date medical information [13].

Lung cancer, one of the most prevalent cancers and the leading cause of cancer-related deaths worldwide, presents a significant need for patient education [14]. Given the critical implications of this diagnosis with its complex treatment modalities, such as radiotherapy, patients often seek comprehensive, reliable, and easily accessible information online [15].

This study aims to evaluate the educational capabilities and limitations of GPT-4 for patients undergoing radiotherapy for lung cancer. This includes a clinician-led evaluation of response relevance, accuracy, and completeness and a patient-led review of its performance in providing educational content. Additionally, a readability analysis was conducted to assess the accessibility of the responses generated.

Methods

We, a team from the Department of Radiation Oncology at University Hospital LMU, Munich, Germany, designed eight questions about radiotherapy for lung cancer. These questions were developed by collecting input from clinicians primarily involved in lung cancer treatment at our department, asking them which questions they frequently encounter during consultations. To ensure a representative selection, four experienced physicians from our lung cancer core team were

interviewed. The questions covered general information about lung cancer, treatment options, radiotherapy, side effects, quality of life, preparations before treatment, and follow-up care. The questions were presented to GPT-4o (version as of July 2024) via OpenAI's web interface, using the prompt: "I am a lung cancer patient looking for answers to the following questions." Table 1 displays the questions and ChatGPT's responses. The questions and answers were originally in German and translated for this publication. To ensure consistency, each question was posed to ChatGPT five times using the same prompt. Each question was presented individually using a standardized prompt structure, without engaging the model in an ongoing conversation or allowing follow-up questions or AI-generated additional prompts. This approach ensured that each response was independent and comparable across multiple trials. The responses were reviewed for variations in content and length by the primary researcher, and no substantial differences were identified.

Table 1: Questions and responses from ChatGPT-4o

Q1: What is lung cancer? ChatGPT: Lung carcinoma, also known as lung cancer, is a malignant neoplasm in the lung tissue. It develops when cells in the lungs grow uncontrollably and form tumors. There are two main types of lung cancer: 1. non-small cell lung carcinoma (NSCLC): This is the most common type and accounts for about 85% of cases. It includes various subtypes such as adenocarcinoma, squamous cell carcinoma and large cell carcinoma. 2. small cell lung carcinoma (SCLC): This type is more aggressive and grows faster than non-small cell lung carcinoma. It is less common and accounts for about 15% of cases. The main causes of lung cancer are tobacco smoking, inhalation of second-hand smoke, environmental and workplace exposure to certain pollutants and genetic factors. Symptoms can include chronic cough, chest pain, weight loss and shortness of breath. Diagnosis is usually made by imaging (e.g. X-ray, CT scan) and biopsies. Treatment may include surgery, radiotherapy, chemotherapy and targeted therapies, depending on the type and stage of the cancer.	Q2: How can lung cancer be treated? ChatGPT: The treatment of lung cancer depends on the type (non-small cell or small cell) and the stage of the disease: - Surgery: removal of the tumor and surrounding lung tissue, often for early stages of non-small cell lung cancer. - Radiotherapy: High-energy radiation destroys cancer cells, useful for inoperable tumors or for follow-up treatment. - Chemotherapy: Drug treatment to kill or inhibit the growth of cancer cells, used before or after surgery and in advanced stages. - Targeted therapy: Drugs that attack specific molecular targets on cancer cells, particularly effective for certain genetic mutations. - Immunotherapy: Stimulates the immune system to fight cancer cells, especially in advanced non-small cell lung cancer. - Palliative therapy: Alleviates symptoms and improves quality of life in advanced cancer. Treatment is often customized and may include a combination of these methods.
Q3: What are the advantages of radiotherapy over other treatment options? ChatGPT: Radiotherapy offers several advantages in the treatment of lung cancer: - Targeted treatment: precise radiation attacks the tumor directly, sparing the surrounding healthy tissue. - Non-invasive: No surgical intervention required, ideal for patients who cannot undergo surgery. - Complementary to other therapies: Effective in combination with chemotherapy to improve outcomes. - Applicability for inoperability: Useful for tumors that cannot be surgically removed. - Pain relief: Relieves symptoms and improves quality of life in advanced cancer. - Preservation of organ function: Preserves lung function better than complete removal of a lung section. - Outpatient treatment: Often available as an outpatient therapy, minimizing hospitalization. These benefits make radiotherapy a valuable treatment option for various stages of lung cancer.	Q4: How is radiotherapy carried out for lung cancer? ChatGPT: Radiotherapy for lung cancer is carried out in several stages: 1. planning: imaging techniques such as CT or MRI create detailed images of the tumor. A radiotherapist plans the exact dose and position of the radiation. 2. simulation: The patient is placed on a couch in the treatment position, often with special cushions or masks for stabilization. Markings on the skin or special tattoos help with precise alignment. 3. treatment: The actual radiotherapy takes place over several sessions (fractions), typically daily over several weeks. High-energy beams are directed at the tumor from different angles to achieve maximum effect and spare healthy tissue. 4. monitoring: regular reviews and imaging during the treatment phase to adjust therapy and monitor progress. The treatment is painless and is performed on an outpatient basis, with each session lasting only a few minutes.
Q5: What are the possible side effects of radiotherapy for lung cancer? ChatGPT: Radiotherapy for lung cancer can have	Q6: What precautions or preparations are necessary before the radiation treatment? ChatGPT: Several precautions and preparations are

several side effects, depending on the radiation dose and the area treated: Acute side effects: - Fatigue: frequent feeling of exhaustion. - Skin reactions: Redness, irritation or dryness in the irradiated area. Cough and breathing difficulties: Intensification of existing symptoms. Late side effects: - Scarring: In the lungs can lead to long-term breathing problems. Pneumonia: Radiation pneumonitis, an inflammation of the lung tissue. - Esophagitis: Inflammation and pain in the esophagus, making swallowing difficult. Rare side effects: - Heart problems: With radiation near the heart. - Secondary tumors: Increased risk of other types of cancer in the irradiated area. Side effects vary from person to person and can be mitigated by supportive measures and medication. Close monitoring by the treatment team helps to identify and treat side effects at an early stage.	necessary before radiotherapy for lung cancer: - Consultation and planning: a detailed consultation with the radiotherapist to discuss the treatment plan, possible side effects and the expected course. - Detailed imaging (CT, MRI) for precise localization of the tumor. - Medical preparation: review and adjustment of current medications, as some may affect the radiation effect. - Assessment of general health and lung function tests. - Physical preparation: Making markings or tattoos on the skin for exact alignment during radiation. - Possibly creation of custom-made splints or masks to stabilize the treatment position. - Nutrition and lifestyle advice: Advice on nutrition to strengthen the body. - Advice to stop smoking and avoid alcohol to promote healing. - Psychological support: access to counseling and support to manage anxiety and stress before treatment. These preparations help to make the treatment more effective and safer.
Q7: How does radiation affect the patient's quality of life during and after treatment? ChatGPT: Radiotherapy can affect the patient's quality of life in various ways both during and after treatment: During treatment: - Fatigue: frequent fatigue may limit daily activities. - Skin reactions: Redness and irritation may cause discomfort. - Breathing problems and coughing: Exacerbation of existing symptoms that interfere with everyday life. After treatment: - Recovery phase: fatigue may last a few weeks, return to normality often takes longer. - Long-term effects: Scarring and radiation pneumonitis can cause breathing difficulties. - Emotional distress: Fear of relapse and coping with physical changes. Positive effects: - Symptom relief: reduction of tumor pain and improvement in breathing can increase quality of life. - Cancer control: Effective tumor control or reduction increases well-being and confidence. Close cooperation with the medical team, supportive measures and a healthy lifestyle help to improve quality of life during and after radiotherapy.	Q8: What follow-up care is required after radiotherapy for lung cancer? ChatGPT: After radiotherapy for lung cancer, careful aftercare is crucial. This includes regular check-ups to monitor the response to treatment and detect possible relapses at an early stage. These usually include imaging such as CT or PET scans, blood tests and lung function tests. It is also important to monitor and treat side effects of radiotherapy, such as pneumonia (radiation pneumonitis) or pulmonary fibrosis. Rehabilitation measures can help to improve lung function and increase quality of life. Psychosocial support is also an essential part of aftercare to provide emotional support for patients and their families.

Evaluation of readability

A readability analysis was conducted using the Modified Flesch Reading Ease (FRE) Formula for German. A well-established readability metric for the English language is the Flesch Reading Ease Scale (FRE) [16]. The FRE measures the readability of a text in terms of its average sentence length (ASL) and the average number of syllables per word (ASW). It relies on the fact that short words or sentences are usually easier to understand than longer ones. For this analysis, we have used the modified FRE for the German language by Toni Amstad [17]: $FRE(\text{German}) = 180 - ASL - (58.5 \times ASW)$.

Also, the 4th Vienna Formula (WSTF) was utilized. Unlike the FRE, the Vienna Formula (WSTF) has not been adapted for the German language. Instead, it is based on the work of Bamberger and Vanacek [18], who analyzed German textual material. They derived at least five versions of the Vienna Formula for prose and non-fiction texts. Typically, the fourth WSTF is used for text analysis.

This metric is also based on average sentence length (ASL) and the proportion of words with three or more syllables (MS): $WSTF = 0.2656 \times ASL + 0.2744 \times MS - 1.6939$.

While the FRE and WSTF do not directly map to standard grade levels in the German language, readability categories were approximated to estimated educational levels to provide a practical interpretation of the required comprehension level. This approach allows for a more intuitive understanding of the readability of ChatGPT-generated responses in the context of patient education (details are provided in Table 2).

The readability analysis and score calculation was performed using Python (version 3.8) and its text processing libraries, such as nltk for sentence and word tokenization and a custom syllabification function for the German language. The FRE score was computed directly based on the modified formula for German. The WSTF score was calculated using the 4th Vienna Formula.

Clinician Evaluation

Following the readability analysis, the eight responses were independently evaluated by six clinicians experienced in lung cancer treatment, including two radiation oncologists, two medical oncologists, and two thoracic surgeons, all with 5–12 years of experience working in specialized lung cancer centers with a university teaching function. This multidisciplinary approach ensured a comprehensive evaluation from different medical perspectives while remaining focused on lung cancer treatment. Clinicians received an information sheet outlining the study's procedures and objectives. The question-answer pairs and evaluation sheet were provided in electronic form. Evaluators had no time limit to complete the assessment, scoring each response for relevance, correctness, and completeness using an ordinal five-point Likert scale, with 1 indicating disagreement and 5 indicating complete agreement with the statements that the responses were relevant, correct, and complete, respectively. Respondents were also allowed to add additional comments to their evaluations.

Patient evaluation

Finally, the question-answer pairs were presented to lung cancer patients during their first follow-up appointment after completing radiotherapy. Patients were invited to participate in a study evaluating a large language model (LLMs) for patient education. They received an information sheet outlining the study's procedures and goals and were asked to sign a data security statement and provide informed consent before participation. After consenting, patients were given the question-answer pairs on paper sheets and had as much time as needed to complete the evaluation. The evaluation was based on seven statements to assess ChatGPT's performance in terms of comprehensibility, accuracy, relevance, and trustworthiness using a five-point Likert scale (1 = strongly disagree, 5 = strongly agree) (Figure 1). Additionally, they were asked whether the information made them feel better informed and if they would consider using ChatGPT for future medical questions. Patient responses from the paper forms were manually entered into Microsoft Office Excel (version 2410) by the primary researcher for further processing and analysis.

Please rate your agreement with the following statements regarding information:



	Do not agree at all	Rather disagree	Neutral	Agree	Fully and completely agree
1. The information provided by ChatGPT was easy to understand.	☐	☐	☐	☐	☐
2. The information provided by ChatGPT was consistent with my experience	☐	☐	☐	☐	☐
3. The information provided by ChatGPT was clear and did not contain medical terminology that was difficult to understand.	☐	☐	☐	☐	☐
4. The information provided by ChatGPT was accurate and relevant to the topic of radiotherapy for lung cancer.	☐	☐	☐	☐	☐
5. The information provided by ChatGPT would have helped me to become better informed about radiotherapy for lung cancer in advance.	☐	☐	☐	☐	☐
6. I have confidence in the information I received from ChatGPT.	☐	☐	☐	☐	☐
7. I would also use the ChatGPT search for future medical questions.	☐	☐	☐	☐	☐

Figure 1: Example of a five-point Likert scale presented to patients to rate ChatGPT's responses

Ethical approval

The local Ethics Committee of Ludwig-Maximilians-University Munich (LMU) approved the study protocol in August 2023 (approval number 23-0742). The study was conducted in accordance with the Declaration of Helsinki, and all patients provided signed written consent to participate. To ensure privacy and confidentiality, all collected data were anonymized before analysis and no personal identifiable information was stored or shared. Data were handled in compliance with institutional and national data protection regulations. Participants did not receive any financial or material compensation for their participation in the study.

Statistical analysis

Data are reported using descriptive statistics, including median, mean, and standard deviation. Statistical analyses were performed using Microsoft Office Excel (version 2410). Figures were generated using Python (version 3.8) with the Matplotlib library. Data extracted from tables was structured in Pandas DataFrames for analysis and plotting.

Results

Evaluation of Readability

Table 2 displays the interpretation of readability scores with an estimated educational level. The FRE scores ranged from 6.3 to 42.3, with a mean of 23.4 ± 11.2 , classifying most responses as "very difficult to read" (university level). Similarly, the WSTF scores ranged from 10.6 to 16.8, with a mean of 13.8 ± 2.1 . Most responses were in the "very difficult to read" category, with some being "difficult" (upper secondary, grade 10-12/13) or "fairly difficult" (lower secondary, grade 7-10) (Table 3).

Table 2: Interpretation of readability scores with estimated educational level: Modified Flesch Reading Ease (FRE) for German, 4th Vienna Formula (WSTF)

Description	FRE	WSTF	Estimated educational level (approx.)
Very difficult to read	0-29	> 14	University level
Difficult to read	30-49	13-14	Upper secondary (Grade 10-12/13)
Fairly difficult to read	50-59	10-13	Lower secondary (Grade 7-10)

Average readability	60-69	8-10	Upper middle school (Grade 6)
Fairly easy to read	70-79	7-8	Lower middle school (Grade 5)
Easy to read	80-89	5-7	Upper elementary (Grade 4)
Very easy to read	90-100	4-5	Lower elementary (Grade 1-3)

Table 3: Readability scores of ChatGPT's Responses to Questions 1-8: Modified Flesch Reading Ease (FRE) and 4th Vienna Formula (WSTF), displaying individual scores of Answers 1-8, Mean $\pm$ Standard Deviation (SD) and Minimum – Maximum

Readability analysis of questions presented to ChatGPT				
Answer	FRE	FRE Interpretation	WSTF	WSTF Interpretation
A1	42.3	Difficult to read	10.6	Fairly difficult to read
A2	12.6	Very difficult to read	16.8	Very difficult to read
A3	23.9	Very difficult to read	14.4	Very difficult to read
A4	35.8	Difficult to read	10.8	Fairly difficult to read
A5	21.1	Very difficult to read	13.6	Difficult to read
A6	28.2	Very difficult to read	14.1	Very difficult to read
A7	16.6	Very difficult to read	14.7	Very difficult to read
A8	6.3	Very difficult to read	15.	Very difficult to read
Min. -	6.3 - 42.3	-	10.6 - 16.8	-
Max. $\pm$	23.4 $\pm$ 11.2	-	13.8 $\pm$ 2.1	-
Mean $\pm$ SD				

Clinician evaluation

Table 4 presents the evaluation of ChatGPT's responses by six clinicians experienced in treating lung cancer: two radiation oncologists, two medical oncologists, and two thoracic surgeons.

The mean scores for relevance ranged from 3.7 ± 0.94 (Responses 2 (treatment) and 3 (advantages of radiotherapy)) to 4.3 ± 0.75 (Response 8 (follow-up)). Correctness scores varied between 3.5 ± 0.50 (Response 7 (quality of life)) and 4.3 ± 0.75 (Response 8 (follow-up)). Completeness ratings ranged from 3.5 ± 0.50 (Responses 2 (treatment), 5 (side effects), and 7 (quality of life)) to 4.2 ± 0.69 (Response 8 (follow up)). Overall, responses showed variability in performance, with relevance and correctness achieving higher mean scores than completeness. Notably, Response 8 (follow up) scored the highest across all three dimensions (Relevance: 4.3 ± 0.75 , Correctness: 4.3 ± 0.75 , Completeness: 4.2 ± 0.69), while Response 7 (quality of life) scored the lowest for correctness (3.5 ± 0.50).

A thoracic surgeon commented that ChatGPT did not discuss chances of treatment success and recurrence rates. A medical oncologist commented that the role of multidisciplinary tumor boards should have been mentioned. A radiation oncologist commented that there was no differentiation between radiotherapy modalities.

Table 4: Clinician ratings of ChatGPT's responses (1–8) for relevance, correctness, and completeness. Scores are based on a 5-point Likert scale, where 1 represents the lowest and 5 the highest score.

Response to questions	Mean (SD)	Ratings on Likert Scale, (%)				
		1	2	3	4	5
Response 1						
Relevance	3.8 (1.07)	0 (0%)	1 (17%)	1 (17%)	2 (33%)	2 (33%)
Correctness	4.2 (0.37)	0 (0%)	0 (0%)	0 (0%)	5 (83%)	1 (17%)
Completeness	3.5 (0.76)	0 (0%)	1 (17%)	1 (17%)	4 (67%)	0 (0%)
Response 2						
Relevance	3.7 (0.94)	0 (0%)	1 (17%)	1 (17%)	3 (50%)	1 (17%)
Correctness	3.7 (0.75)	0 (0%)	0 (0%)	3 (50%)	2 (33%)	1 (17%)
Completeness	3.5 (0.50)	0 (0%)	0 (0%)	3 (50%)	3 (50%)	0 (0%)
Response 3						
Relevance	3.7 (0.94)	0 (0%)	1 (17%)	1 (17%)	3 (50%)	1 (17%)
Correctness	3.7 (0.75)	0 (0%)	0 (0%)	3 (50%)	2 (33%)	1 (17%)
Completeness	3.7 (0.47)	0 (0%)	0 (0%)	2 (33%)	4 (67%)	0 (0%)
Response 4						
Relevance	4.3 (0.47)	0 (0%)	0 (0%)	0 (0%)	4 (67%)	2 (33%)
Correctness	3.8 (0.37)	0 (0%)	0 (0%)	1 (17%)	5 (83%)	0 (0%)
Completeness	3.8 (0.37)	0 (0%)	0 (0%)	1 (17%)	5 (83%)	0 (0%)
Response 5						
Relevance	4.2 (0.90)	0 (0%)	0 (0%)	2 (33%)	1 (17%)	3 (50%)
Correctness	4.0 (0.58)	0 (0%)	0 (0%)	1 (17%)	4 (67%)	1 (17%)
Completeness	3,5 (0.50)	0 (0%)	0 (0%)	3 (50%)	3 (50%)	0 (0%)
Response 6						
Relevance	3.8 (0.90)	0 (0%)	0 (0%)	3 (50%)	1 (17%)	2 (33%)
Correctness	4.0 (0.00)	0 (0%)	0 (0%)	0 (0%)	6 (100%)	0 (0%)
Completeness	4.0 (0.00)	0 (0%)	0 (0%)	0 (0%)	6 (100%)	0 (0%)
Response 7						
Relevance:	4.0 (0.82)	0 (0%)	0 (0%)	2 (33%)	2 (33%)	2 (33%)
Correctness:	3.5 (0.50)	0 (0%)	0 (0%)	3 (50%)	3 (50%)	0 (0%)
Completeness:	3.5 (0.50)	0 (0%)	0 (0%)	3 (50%)	3 (50%)	0 (0%)
Response 8						
Relevance:	4.3 (0.75)	0 (0%)	0 (0%)	1 (17%)	2 (33%)	3 (50%)
Correctness:	4.3 (0.75)	0 (0%)	0 (0%)	1 (17%)	2 (33%)	3 (50%)
Completeness:	4.2 (0.69)	0 (0%)	0 (0%)	1 (17%)	3 (50%)	2 (33%)

Patient evaluation

The responses generated by ChatGPT were evaluated by 30 consecutive patients who underwent radiation therapy for lung cancer between 06/2024 – 10/2024 at the University Hospital LMU Munich during their first follow-up examination six weeks after treatment completion. The median age of the nineteen male and eleven female patients was 66 years (48 – 87 years). Twenty-six of those patients had non-small cell lung cancer (NSCLC), while four patients had small cell lung cancer (SCLC). Twelve patients received concomitant chemotherapy, and ten patients received stereotactic body radiotherapy (SBRT). Five patients were treated using magnetic resonance-guided radiotherapy (MRgRT).

Results of the patient evaluation are summarized in Table 5. The highest-rated statement was "The information provided by ChatGPT was easy to understand," with a mean score of 4.4 ± 0.61 , where 94% of patients rated it as "agree" or "strongly agree." Similarly, the statement "The information provided by ChatGPT was accurate and relevant to radiotherapy for lung cancer" received a high mean score of 4.2 ± 0.83 , with 87% of patients rating it positively. The statement "The information provided by ChatGPT was consistent with my experience" achieved a mean score of 4.1 ± 0.63 , reflecting alignment with patient expectations. Similarly, the statement "The information provided by ChatGPT was clear and did not contain medical terminology that was difficult to understand" received a mean score of 4.1 ± 0.81 , with 80% of patients giving positive feedback. This highlights ChatGPT's strength in delivering accessible and jargon-free information.

In contrast, statements related to usability and trustworthiness received slightly lower ratings. "The information provided by ChatGPT would have helped me to become better informed about radiotherapy for lung cancer in advance" and "I would also use ChatGPT for future medical questions" both scored a mean of 3.9 ± 0.94 . Additionally, the statement "I have confidence in the information I received from ChatGPT" scored 4.0 ± 0.84 .

Table 5: Patient's ratings of statements 1-7. Scores are based on a 5-point Likert scale, where 1 represents the lowest and 5 the highest score (1 = strongly disagree, 5 = strongly agree).

Statement		Mean (SD)	Ratings on Likert Scale, n (%)				
			1	2	3	4	5
1.	The information provided by ChatGPT was easy to understand.	4.4 (0.61)	0 (0%)	0 (0%)	2 (7%)	14 (47%)	14 (47%)
2.	The information provided by ChatGPT was consistent with my experience.	4.1 (0.63)	0 (0%)	0 (0%)	5 (17%)	18 (60%)	7 (23%)
3.	The information provided by ChatGPT was clear and did not contain medical terminology that was difficult to understand.	4.1 (0.81)	0 (0%)	1 (3%)	5 (17%)	13 (43%)	11 (37%)
4.	The information provided by ChatGPT was accurate and relevant to the topic of radiotherapy for lung cancer.	4.2 (0.83)	0 (0%)	2 (7%)	2 (7%)	14 (47%)	12 (40%)
5.	The information provided by ChatGPT would have helped me to become better informed about radiotherapy for lung cancer in advance.	3.9 (0.94)	0 (0%)	3 (10%)	6 (20%)	12 (40%)	9 (30%)
6.	I have confidence in the information I received from ChatGPT.	4.0 (0.84)	0 (0%)	1 (3%)	8 (27%)	12 (40%)	9 (30%)
7.	I would also use the ChatGPT search for future medical questions.	3.9 (0.94)	0 (0%)	3 (10%)	6 (20%)	12 (40%)	9 (30%)

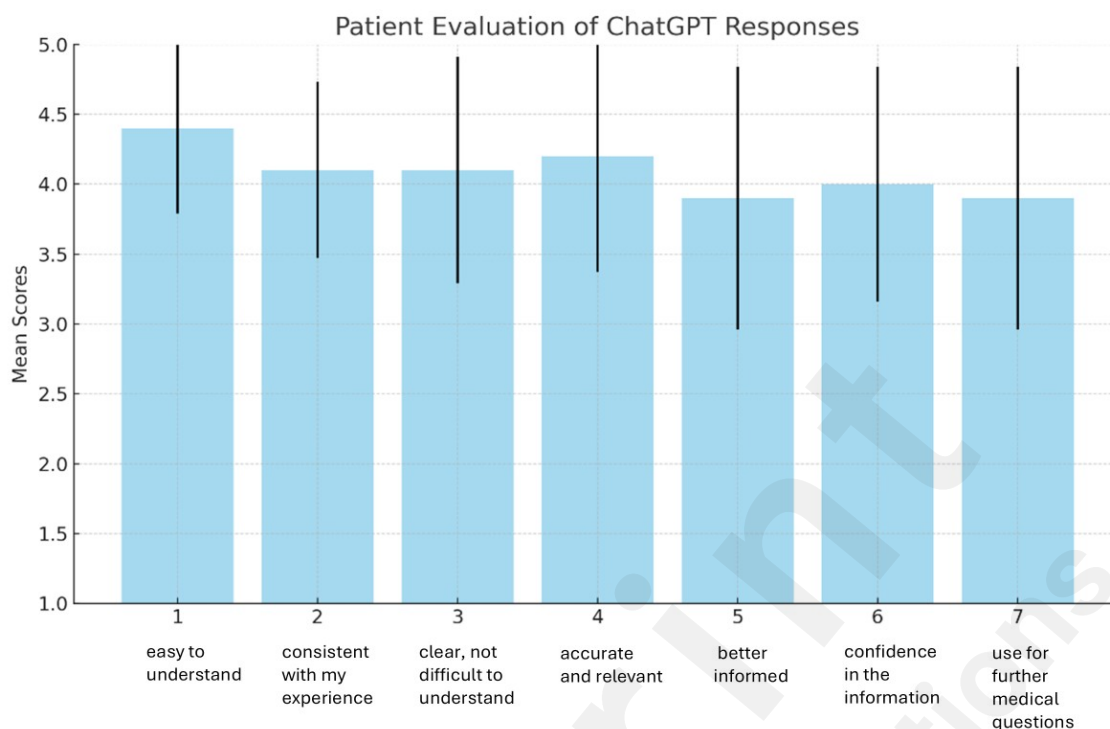


Figure 2: Mean patient ratings with standard deviations (SD) of ChatGPT's responses to seven statements, evaluated using a 5-point Likert scale. Statements assessed the answers' clarity, relevance, accuracy, and usefulness. Error bars indicate standard deviations, reflecting variability in patient responses.

Discussion

Providing accessible and understandable information is a key component of patient-centered care, particularly in oncology. Research has shown that cancer patients often seek information from sources other than their healthcare providers, with the Internet serving as a primary resource [19]. However, existing online resources frequently fail to address patient's specific questions, especially in radiation oncology and often exceed recommended complexity levels [20, 21]. Against this background, our study explores the potential of the most widely used and broadly adopted LLM, ChatGPT [22], making it a relevant and practical model for evaluating real-world applications in patient communication and education in lung cancer radiotherapy.

Principal findings

This study evaluated the benefits and risks of using ChatGPT to educate patients undergoing radiotherapy for lung cancer. The analysis included a multifaceted evaluation of ChatGPT-generated content, including readability assessment, clinician evaluation and patient feedback. The main findings indicate that while ChatGPT's responses are often technically complex and rated as "difficult to read" based on objective readability measures (FRE and WSTF), patients still perceived the information as clear and understandable. Clinicians rated the responses positively for relevance and correctness but noted some limitations in completeness.

Comparison to prior work

The readability analysis of ChatGPT's responses revealed that the FRE and WSTF scores classified most responses as "very difficult to read" or "difficult to read", which may limit accessibility, particularly for individuals with lower health literacy. The low readability scores are primarily due to the extensive use of complex medical terminology and long sentence structures, which increase the calculated ASL and ASW values, thereby reducing readability. While the FRE and WSTF scores do not have direct grade-level equivalents in German, texts classified as "very difficult to read" or

"difficult to read", typically require upper secondary education or higher for full comprehension. Additionally, the complexity of the prompt can influence the readability of responses, as more detailed inquiries tend to generate longer, more technical answers, which may further reduce readability. These findings are consistent with previous studies in other medical domains, which have similarly reported low readability scores for AI-generated content, suggesting that readability challenges are a common limitation across various medical specialties and not specific to lung cancer [23-26]. Moreover, lung cancer education inherently involves complex terminology, multidisciplinary treatment approaches and a broad spectrum of disease presentations, all of which may contribute to lower readability scores compared to simpler medical topics. These findings align with studies indicating that cancer-related information on the Internet is generally not well-tailored to patient's needs [27]. Despite this, the patient evaluation showed that ChatGPT's responses were perceived as easy to understand (mean score: 4.4 ± 0.61), possibly because the survey was conducted post-therapy when patients were already familiar with relevant topics and terminology. One possible strategy to improve readability in patient education materials is the fine-tuning of LLM's with curated, patient-friendly datasets or the integration of real-time readability adjustments that simplify sentence structure while maintaining medical accuracy. Additionally, a hybrid approach involving AI-generated content reviewed by clinicians may enhance accessibility without compromising correctness.

The clinician evaluation of ChatGPT highlighted its strengths in relevance and correctness but noted limitations in completeness. Response 8, for example, performed best across all dimensions (Relevance: 4.3 ± 0.75 , Correctness: 4.3 ± 0.75 , Completeness: 4.2 ± 0.69), while Response 7 demonstrated inconsistencies, scoring the lowest for correctness (3.5 ± 0.50). These findings align with other studies assessing ChatGPT's accuracy in answering questions about lung cancer [28, 29] and other queries in radiotherapy [30, 31]. Interestingly, another study found that ChatGPT achieved high qualitative ratings for factual accuracy, conciseness and completeness, closely mirroring expert responses [32]. The lower completeness scores suggest that ChatGPT responses, while relevant and mostly accurate, may omit critical clinical details. This limitation could be mitigated by refining prompting strategies to ensure more detailed outputs or integrating clinician oversight in AI-assisted patient education.

Patients rated ChatGPT highly for clarity and relevance, but usability and trust received comparatively lower scores. Statements like "I would also use ChatGPT for future medical questions" (3.9 ± 0.94) and "I have confidence in the information I received from ChatGPT" (4.0 ± 0.84) highlight areas where trust and reliability could be improved. Lower trustworthiness and usability scores suggest that while patients find ChatGPT-generated responses clear and relevant, concerns remain regarding the credibility of AI-generated medical information. Future implementations could improve trust through clinician oversight, AI transparency measures and integration with evidence-based sources.

Considerations for clinical integration

LLM's like ChatGPT are often approached cautiously in healthcare due to concerns about trust, security, privacy and ethics [33, 34]. While ChatGPT is sometimes criticized for lacking a human touch and empathy [35], studies have found its responses to be more empathetic than those of clinicians in specific scenarios [36], especially for sensitive health topics where patients may feel uncomfortable consulting clinicians, non-sentient chatbot tools may offer valuable support [32].

Despite ongoing concerns about "hallucinations," where LLMs generate plausible but incorrect answers [10, 11], no potential harm was identified in ChatGPT's responses in this study. OpenAI, the developer of ChatGPT, acknowledges the possibility of inaccurate outputs [37], likely contributing to healthcare provider's reluctance to adopt LLM chatbots for patient communication and education [37]. However, other studies have shown that ChatGPT can provide highly accurate and complete

responses comparable to virtual patient-clinician communication in radiation oncology [32]. To minimize the risk of misinformation, future AI-driven patient education tools should incorporate source attribution, real-time fact-checking and clinician oversight. Additionally, models specifically trained on verified medical datasets may help reduce the occurrence of incorrect or misleading information.

Strengths and limitations

First, a key strength of this study is its comprehensive evaluation approach, combining readability metrics with both clinician and patient assessments.

This study has several limitations. First, the questions were formulated by the study team based on input from clinicians experienced in lung cancer treatment, rather than being directly collected from patients. While this approach ensured clinical relevance and reflected frequently encountered consultation topics, it may have limited the diversity of clinical scenarios and the representativeness of the findings from a broader patient perspective. Second, the study was conducted in German, which could affect the generalizability of results, as ChatGPT's performance may vary across languages [38, 39]. Third, this study used GPT-4o, released by OpenAI in May 2024, a paid subscription model with superior accuracy and coherence compared to the free GPT-3.5 version, which may limit accessibility by the general population. Fourth, another key limitation of this study is the lack of standardized criteria for assessing AI-generated medical responses. Future research should focus on developing structured evaluation frameworks, integrating expert consensus and establishing domain-specific benchmarks to ensure consistent assessment of AI-generated content.

Fifth and finally, while our study focused on ChatGPT, the most popular and earliest publicly released conversational LLM [22], Other models, such as Bard (Google), LLAMA (Meta), and Claude (Anthropic), show promise in addressing oncology queries. However, no comparative analysis of these alternative models was conducted in this study, as the primary objective was to assess the feasibility and quality of ChatGPT's responses as a widely used reference model in patient education. ChatGPT was selected due to its widespread adoption and superior response quality demonstrated in prior studies compared to other LLMs.[24, 40, 41]. Previous research has highlighted differences among LLMs in terms of response accuracy, completeness and readability in healthcare applications, including radiotherapy [40-45]. Future studies should explore how various LLMs perform specifically in the context of patient education in lung cancer, to identify the most suitable tools for clinical integration.

Conclusion and future directions

In conclusion, ChatGPT demonstrates significant potential as a supplementary tool for patient education in radiation oncology, particularly for patients undergoing radiotherapy for lung cancer. Its ability to provide clear and relevant information highlights its value in enhancing patient understanding and engagement in their treatment journey. However, limitations in completeness, accuracy, and trust underscore the importance of careful review and supplementation by healthcare professionals.

Further development of AI tools should focus on improving readability, through fine-tuning on patient-friendly datasets or integrating real-time readability adaptation, while maintaining medical accuracy. Incorporating clinician oversight into AI-generated content could enhance both reliability and trust. Additionally, the development of standardized evaluation frameworks for AI-generated health information will be essential to ensure consistent quality assessment. With continued research and refinement, ChatGPT and similar technologies have the potential to revolutionize patient education and support healthcare providers in delivering accurate, accessible and personalized care.

Acknowledgments

CRedit authorship contribution statement

Cedric Richlitzki: Data curation, Conceptualization, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft. **Sina Mansoorian:** Methodology, Validation, Data curation, Writing – review & editing. **Lukas Käsmann:** Methodology, Validation, Data curation, Writing – review & editing. **Mircea Gabriel Stoleriu:** Methodology, Validation, Data curation, Writing – review & editing. **Julia Kovacs:** Methodology, Validation, Data curation, Writing – review & editing. **Wulf Sienel:** Methodology, Validation, Data curation, Writing – review & editing. **Diego Kauffmann-Guerrero:** Methodology, Validation, Data curation, Writing – review & editing. **Thomas Duell:** Methodology, Validation, Data curation, Writing – review & editing. **Nina Sophie Schmidt-Hegemann:** Methodology, Validation, Data curation, Writing – review & editing. **Claus Belka:** Conceptualization, Resources, Investigation, Formal analysis, Methodology, Project administration, Supervision, Validation, Writing – review & editing. **Stefanie Corradini:** Conceptualization, Resources, Investigation, Methodology, Formal analysis, Project administration, Supervision, Validation, Writing – review & editing. **Chukwuka Eze:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – review & editing. Generative AI (ChatGPT-4o) was used to generate the eight answers evaluated in this study, as described in the Methods section.

Conflicts of Interest

No funding was received to conduct this study. The authors have no competing interests to declare relevant to this article's content.

Abbreviations

AI	artificial intelligence
ASL	average Sentence Length
ChatGPT	Chat Generative Pre-Trained Transformer
FRE	Modified Flesch Reading Ease
GPT	Generative Pre-trained Transformer
LLM	large language model
LMU	Ludwig-Maximilians-University Munich
MRgRT	magnetic resonance-guided radiotherapy
NLP	natural language processing
NSCLC	non-small cell lung cancer
RI	readability Index
SBRT	stereotactic body radiotherapy
SCLC	small cell lung cancer
SD	standard deviation
WSTF	4th Vienna Formula

Data availability Statement

Clinical study reports, detailed data tables, and programming code are available on request: Cedric Richlitzki, Department of Radiation Oncology, LMU University Hospital, LMU Munich, Marchioninistrasse 15, 81377 Munich, Germany. E-mail address: Cedric.Richlitzki@med.uni-muenchen.de.

Figures and Tables

Table 1: Questions and responses from ChatGPT-4o

Figure 1: Example of a five-point Likert scale presented to patients to rate ChatGPT's responses

Table 2: Interpretation of readability scores with estimated educational level: Modified Flesch Reading Ease (FRE) for German, 4th Vienna Formula (WSTF)

Table 3: Readability scores of ChatGPT's Responses to Questions 1-8: Modified Flesch Reading Ease (FRE) and 4th Vienna Formula (WSTF), displaying individual scores of Answers 1-8, Mean $\pm$ Standard Deviation (SD) and Minimum – Maximum

Table 4: Clinician ratings of ChatGPT's responses (1–8) for relevance, correctness, and completeness. Scores are based on a 5-point Likert scale, where 1 represents the lowest and 5 the highest score

Table 5: Patient's ratings of statements 1-7. Scores are based on a 5-point Likert scale, where 1 represents the lowest and 5 the highest score (1 = strongly disagree, 5 = strongly agree)

Figure 2: Mean patient ratings with standard deviations (SD) of ChatGPT's responses to seven statements, evaluated using a 5-point Likert scale. Statements assessed the answers' clarity, relevance, accuracy, and usefulness. Error bars indicate standard deviations, reflecting variability in patient responses.

References

- [1] Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N Engl J Med*. 2023;388:1233-9.
- [2] Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer. *JAMA Oncol*. 2023;9:1437-40.
- [3] Monje S, Ulene S, Gimovsky AC. Identifying ChatGPT-written Patient Education Materials Using Text Analysis and Readability. *Am J Perinatol*. 2024.
- [4] Shao CY, Li H, Liu XL, Li C, Yang LQ, Zhang YJ, et al. Appropriateness and Comprehensiveness of Using ChatGPT for Perioperative Patient Education in Thoracic Surgery in Different Language Contexts: Survey Study. *Interact J Med Res*. 2023;12:e46900.
- [5] Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29:1930-40.
- [6] Atwood TF, Brown DW, Juang T, Moore KL, McConnell KA, Steers JM, et al. A review of patient questions from physicist-patient consults. *J Appl Clin Med Phys*. 2020;21:305-8.
- [7] Akbar F, Mark G, Warton EM, Reed ME, Prausnitz S, East JA, et al. Physicians' electronic inbox work patterns and factors associated with high inbox work duration. *J Am Med Inform Assoc*. 2021;28:923-30.
- [8] Nielsen JPS, von Buchwald C, Grønhøj C. Validity of the large language model ChatGPT (GPT4) as a patient information source in otolaryngology by a variety of doctors in a tertiary otorhinolaryngology department. *Acta Otolaryngol*. 2023;143:779-82.
- [9] Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*. 2023;29:721-32.
- [10] Chen S, Kann BH, Foote MB, Aerts H, Savova GK, Mak RH, et al. Use of Artificial Intelligence Chatbots for Cancer Treatment Information. *JAMA Oncol*. 2023;9:1459-62.
- [11] van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature*. 2023;614:224-6.
- [12] Liu Z, Zhang L, Wu Z, Yu X, Cao C, Dai H, et al. Surviving ChatGPT in healthcare. *Front Radiol*. 2023;3:1224682.
- [13] Whiles BB, Bird VG, Canales BK, DiBianco JM, Terry RS. Caution! AI Bot Has Entered the Patient Chat: ChatGPT Has Limitations in Providing Accurate Urologic Healthcare Advice. *Urology*. 2023;180:278-84.
- [14] Leiter A, Veluswamy RR, Wisnivesky JP. The global burden of lung cancer: current status and

future trends. *Nat Rev Clin Oncol*. 2023;20:624-39.

[15] Jia X, Pang Y, Liu LS. Online Health Information Seeking Behavior: A Systematic Review. *Healthcare (Basel)*. 2021;9.

[16] Flesch R. A new readability yardstick. *J Appl Psychol*. 1948;32:221-33.

[17] Amstad T. Wie verständlich sind unsere Zeitungen?: Studenten-Schreib-Service; 1978.

[18] Bamberger R, Vanecek E. Lesen-Verstehen-Lernen-Schreiben: die Schwierigkeitsstufen von Texten in deutscher Sprache: Jugend und Volk; 1984.

[19] Reifegerste D, Rosset M, Czerwinski F, Baumann E, Gaisser A, Kludt E, et al. Understanding the Pathway of Cancer Information Seeking: Cancer Information Services as a Supplement to Information from Other Sources. *J Cancer Educ*. 2023;38:175-84.

[20] Prabhu AV, Hansberry DR, Agarwal N, Clump DA, Heron DE. Radiation Oncology and Online Patient Education Materials: Deviating From NIH and AMA Recommendations. *Int J Radiat Oncol Biol Phys*. 2016;96:521-8.

[21] Rosenberg SA, Francis DM, Hullet CR, Morris ZS, Brower JV, Anderson BM, et al. Online patient information from radiation oncology departments is too complex for the general population. *Pract Radiat Oncol*. 2017;7:57-62.

[22] Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton EW, et al. A Systematic Review of ChatGPT and Other Conversational Large Language Models in Healthcare. *medRxiv*. 2024.

[23] Aydın FO, Aksoy BK, Ceylan A, Akbaş YB, Ermiş S, Kepez Yıldız B, et al. Readability and Appropriateness of Responses Generated by ChatGPT 3.5, ChatGPT 4.0, Gemini, and Microsoft Copilot for FAQs in Refractive Surgery. *Turk J Ophthalmol*. 2024;54:313-7.

[24] Shi R, Liu S, Xu X, Ye Z, Yang J, Le Q, et al. Benchmarking four large language models' performance of addressing Chinese patients' inquiries about dry eye disease: A two-phase study. *Heliyon*. 2024;10:e34391.

[25] Swisher AR, Wu AW, Liu GC, Lee MK, Carle TR, Tang DM. Enhancing Health Literacy: Evaluating the Readability of Patient Handouts Revised by ChatGPT's Large Language Model. *Otolaryngol Head Neck Surg*. 2024;171:1751-7.

[26] Srinivasan N, Samaan JS, Rajeev ND, Kanu MU, Yeo YH, Samakar K. Large language models and bariatric surgery patient education: a comparative readability analysis of GPT-3.5, GPT-4, Bard, and online institutional resources. *Surg Endosc*. 2024;38:2522-32.

[27] Abreu AA, Murimwa GZ, Farah E, Stewart JW, Zhang L, Rodriguez J, et al. Enhancing Readability of Online Patient-Facing Content: The Role of AI Chatbots in Improving Cancer Information Accessibility. *J Natl Compr Canc Netw*. 2024;22.

[28] Haver HL, Lin CT, Sirajuddin A, Yi PH, Jeudy J. Evaluating ChatGPT's Accuracy in Lung Cancer Prevention and Screening Recommendations. *Radiol Cardiothorac Imaging*. 2023;5:e230115.

[29] Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard. *Radiology*. 2023;307:e230922.

[30] Dennstädt F, Hastings J, Putora PM, Vu E, Fischer GF, Süveg K, et al. Exploring Capabilities of Large Language Models such as ChatGPT in Radiation Oncology. *Adv Radiat Oncol*. 2024;9:101400.

[31] Pandey VK, Munshi A, Mohanti BK, Bansal K, Rastogi K. Evaluating ChatGPT to test its robustness as an interactive information database of radiation oncology and to assess its responses to common queries from radiotherapy patients: A single institution investigation. *Cancer Radiother*. 2024;28:258-64.

[32] Yalamanchili A, Sengupta B, Song J, Lim S, Thomas TO, Mittal BB, et al. Quality of Large Language Model Responses to Radiation Oncology Patient Care Questions. *JAMA Netw Open*. 2024;7:e244630.

[33] Jeyaraman M, K SP, Jeyaraman N, Nallakumarasamy A, Yadav S, Bondili SK. ChatGPT in Medical Education and Research: A Boon or a Bane? *Cureus*. 2023;15:e44316.

[34] Naik N, Hameed BMZ, Shetty DK, Swain D, Shah M, Paul R, et al. Legal and Ethical

Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility? *Front Surg.* 2022;9:862322.

[35] Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical Considerations of Using ChatGPT in Health Care. *J Med Internet Res.* 2023;25:e48009.

[36] Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, 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.

[37] Sinsky CA, Shanafelt TD, Ripp JA. The Electronic Health Record Inbox: Recommendations for Relief. *J Gen Intern Med.* 2022;37:4002-3.

[38] Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis. *J Med Internet Res.* 2024;26:e60807.

[39] Petrić Howe N. ChatGPT has a language problem - but science can fix it. *Nature.* 2024.

[40] Abbas A, Rehman MS, Rehman SS. Comparing the Performance of Popular Large Language Models on the National Board of Medical Examiners Sample Questions. *Cureus.* 2024;16:e55991.

[41] D'Anna G, Van Cauter S, Thurnher M, Van Goethem J, Haller S. Can large language models pass official high-grade exams of the European Society of Neuroradiology courses? A direct comparison between OpenAI chatGPT 3.5, OpenAI GPT4 and Google Bard. *Neuroradiology.* 2024;66:1245-50.

[42] Trapp C, Schmidt-Hegemann N, Keilholz M, Brose SF, Marschner SN, Schönecker S, et al. Patient- and clinician-based evaluation of large language models for patient education in prostate cancer radiotherapy. *Strahlenther Onkol.* 2025;201:333-42.

[43] Iannantuono GM, Bracken-Clarke D, Karzai F, Choo-Wosoba H, Gulley JL, Floudas CS. Comparison of Large Language Models in Answering Immuno-Oncology Questions: A Cross-Sectional Study. *Oncologist.* 2024;29:407-14.

[44] Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. *Front Med (Lausanne).* 2024;11:1477898.

[45] Rydzewski NR, Dinakaran D, Zhao SG, Ruppin E, Turkbey B, Citrin DE, et al. Comparative Evaluation of LLMs in Clinical Oncology. *Nejm ai.* 2024;1.

Supplementary Files

Clinician ratings of ChatGPT's responses (1–8) for relevance, correctness, and completeness. Scores are based on a 5-point Likert scale, where 1 represents the lowest and 5 the highest score.

URL: <http://asset.jmir.pub/assets/724489038f6002074c609ce70915a543.docx>

Interpretation of readability scores: Modified Flesch Reading Ease (FRE) for German, 4th Vienna Formula (WSTF).

URL: <http://asset.jmir.pub/assets/d79bb540654e801e689f08a018372a15.docx>

Patients' ratings of statements 1-7. Scores are based on a 5-point Likert scale, where 1 represents the lowest and 5 the highest score (1 = strongly disagree, 5 = strongly agree).

URL: <http://asset.jmir.pub/assets/69ef3f3b51e2bd187eaf1dfe02b88096.docx>

Questions and responses from ChatGPT-4.

URL: <http://asset.jmir.pub/assets/40f105da6bfb6672d9742710a9163f52.docx>

Readability scores of ChatGPT's Responses to Questions 1-8: Modified Flesch Reading Ease (FRE) and 4th Vienna Formula (WSTF), displaying individual scores of Answers 1-8, Mean $\pm$ Standard Deviation (SD) and Minimum – Maximum.

URL: <http://asset.jmir.pub/assets/c81672dce4898613cc7882a4206289ba.docx>

Revised_Ms_with_tracked_changes_280425.

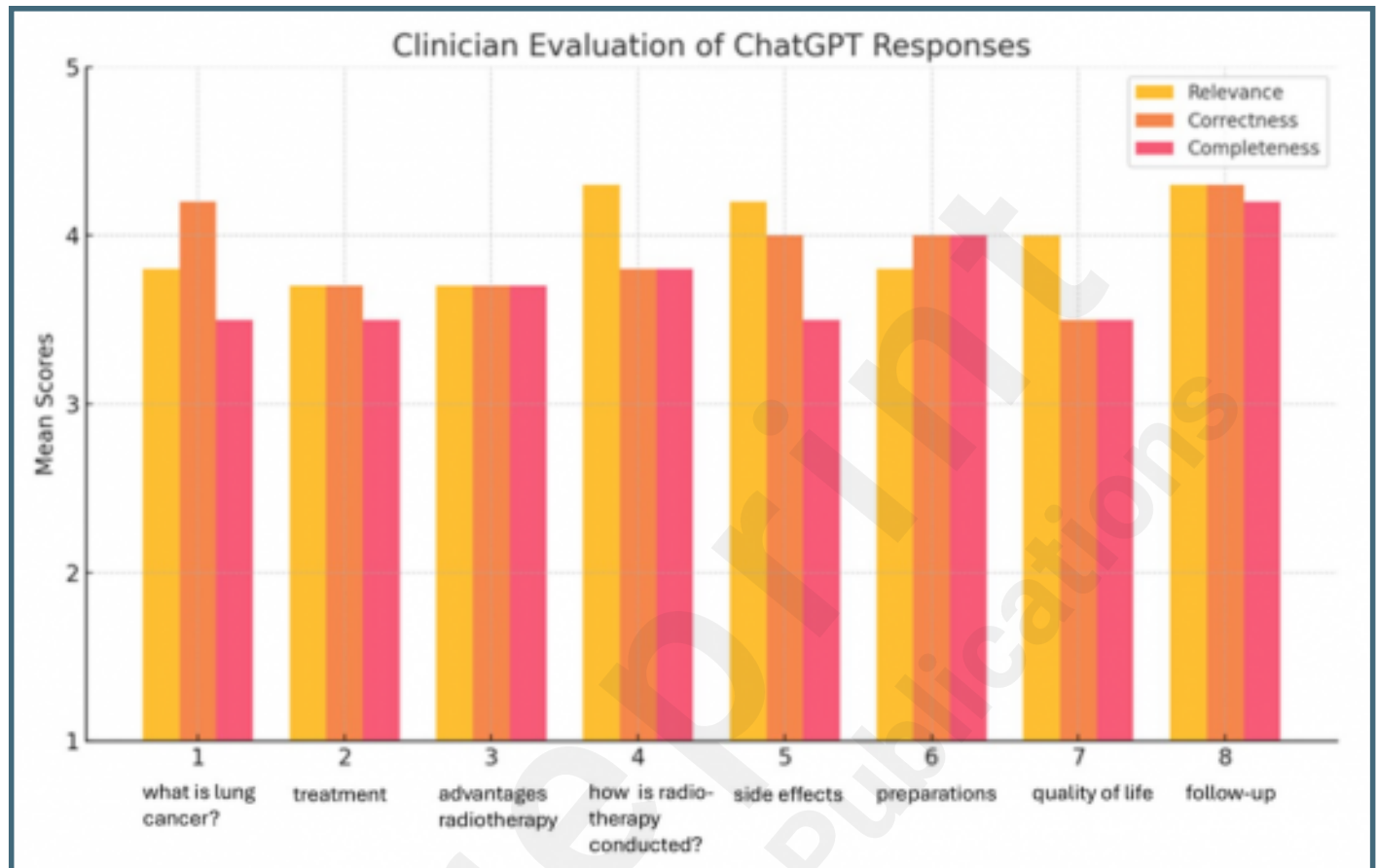
URL: <http://asset.jmir.pub/assets/ef72cc5d91d6400ff6f0d12c69b576c9.docx>

Response to Editorial and Reviewer Comments - 2.

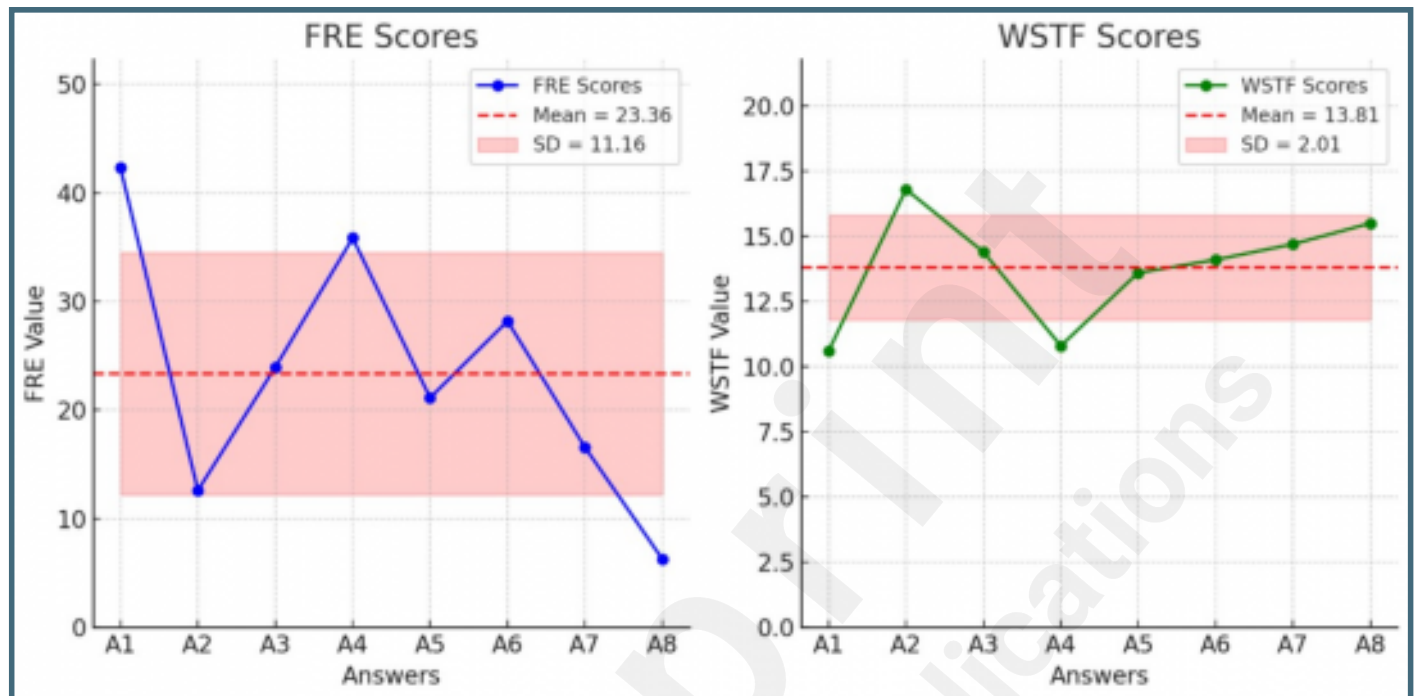
URL: <http://asset.jmir.pub/assets/b26613b20225286e70f308ef9a650d1f.pdf>

Figures

Mean clinician ratings of ChatGPT's responses (1–8) for relevance, correctness, and completeness. Scores are based on a 5-point Likert scale, where 1 represents the lowest and 5 the highest score. Relevance, correctness, and completeness were evaluated for each response, providing an overview of ChatGPT's performance across these dimensions.



Individual values of the readability analysis of the ChatGPT Answers (A1-A8) to the presented questions (Q1-8) along with the Mean $\pm$ Standard Deviation (SD). Modified Flesch Reading Ease (FRE) and 4th Vienna Formula (WSTF) Scores: The blue (FRE) and green (WSTF) lines represent the FRE / WSTF values for each answer (A1-A8). The red dashed line indicates the mean score, and the shaded area represents the SD.



Example of a five-point Likert scale presented to patients to rate ChatGPT's responses.

Please rate your agreement with the following statements regarding information:







	Do not agree at all	Rather disagree	Neutral	Agree	Fully and completely agree
1. The information provided by ChatGPT was easy to understand.	☐	☐	☐	☐	☐
2. The information provided by ChatGPT was consistent with my experience	☐	☐	☐	☐	☐
3. The information provided by ChatGPT was clear and did not contain medical terminology that was difficult to understand.	☐	☐	☐	☐	☐
4. The information provided by ChatGPT was accurate and relevant to the topic of radiotherapy for lung cancer.	☐	☐	☐	☐	☐
5. The information provided by ChatGPT would have helped me to become better informed about radiotherapy for lung cancer in advance.	☐	☐	☐	☐	☐
6. I have confidence in the information I received from ChatGPT.	☐	☐	☐	☐	☐
7. I would also use the ChatGPT search for future medical questions.	☐	☐	☐	☐	☐

Mean patient ratings with standard deviations (SD) of ChatGPT's responses to seven statements, evaluated using a 5-point Likert scale. Statements assessed the answers' clarity, relevance, accuracy, and usefulness. Error bars indicate standard deviations, reflecting variability in patient responses.

