Assessing the Quality and Reliability of ChatGPT's Responses to Radiotherapy-Related Patient Queries: GPT-3.5 versus GPT-4

Ana Grilo, Catarina Marques, Maria Corte-Real, Elisabete Carolino, Marco Caetano

Submitted to: JMIR Cancer
on: June 27, 2024

Table of Contents

Assessing the Quality and Reliability of ChatGPT's Responses to Radiotherapy-Related Patient Queries: GPT-3.5 versus GPT-4

Ana Grilo[1] PhD, MSc; Catarina Marques[2] Bachelor MIRT; Maria Corte-Real[2] Bachelor MIRT; Elisabete Carolino[1] PhD; Marco Caetano[2] Master Radiotherapy

[1]Health & Technology Research Center, ESTeSL ? Escola Superior de Tecnologia da Saúde, Instituto Politécnico de Lisboa, Portugal Research Center for Psychological Science of the Faculty of Psychology, University of Lisbon Lisboa PT
[2]ESTeSL ? Escola Superior de Tecnologia da Saúde, Instituto Politécnico de Lisboa, Portugal Lisbon PT

Corresponding Author:
Ana Grilo PhD, MSc
Health & Technology Research Center, ESTeSL ? Escola Superior de Tecnologia da Saúde, Instituto Politécnico de Lisboa, Portugal
Research Center for Psychological Science of the Faculty of Psychology, University of Lisbon
Av. D. João II, Lote 4.69.01, Parque das Nações
Portugal
Lisboa
PT

Abstract

Background: Patients frequently resort to the Internet to access cancer information. Nevertheless, these online websites often need more content accuracy and readability. Recently, ChatGPT, an artificial intelligence-powered chatbot, signifies a potential paradigm shift in how cancer patients can access vast medical information. However, given that ChatGPT was not explicitly trained for oncology-related inquiries, the quality of the information it provides still needs to be verified. Evaluating the quality of responses is crucial, as misinformation can foster a false sense of knowledge and security, lead to noncompliance, and delay appropriate treatment.

Objective: This study aims to evaluate the quality and reliability of ChatGPT’s responses to standart patient queries about radiotherapy, comparing the performance of GPT-3.5 and GPT-4.

Methods: Forty commonly asked radiotherapy questions were selected and inserted into both versions. Responses were evaluated by six radiotherapy experts using a General Quality Score (GQS), assessed for consistency and similarity using the cosine similarity score, and analyzed for readability using the Flesch Reading Ease Score (FRES) and Flesch-Kincaid Grade Level (FKGL). Statistical analysis was performed using the Mann-Whitney test.

Results: GPT-4 demonstrated superior performance, with higher GQS and a complete absence of lower scores compared to GPT-3.5. The Mann-Whitney test revealed statistically significant differences in some questions, with GPT-4 generally receiving higher ratings. The cosine similarity score indicated substantial similarity and consistency in responses from both versions. Readability scores for both versions were considered college-level, with GPT-4 scoring slightly better in FRES (35.55) and FKGL (12.71) compared to GPT-3.5 (30.68 and 13.53, respectively). Both versions’ responses were deemed challenging for the public to read.

Conclusions: While GPT-4 generates more accurate and reliable responses than GPT-3.5, both models present readability challenges for the public. ChatGPT reveals potential as a valuable resource for addressing common patient queries related to radiotherapy. However, it’s crucial to acknowledge its limitations, including the risks of misinformation and readability issues.

(JMIR Preprints 27/06/2024:63677)
DOI: https://doi.org/10.2196/preprints.63677

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 v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Assessing the Quality and Reliability of ChatGPT's Responses to Radiotherapy-Related Patient Queries: GPT-3.5 versus GPT-4

Abstract

Background: Patients frequently resort to the Internet to access cancer information. Nevertheless, these online websites often need more content accuracy and readability. Recently, ChatGPT, an artificial intelligence-powered chatbot, signifies a potential paradigm shift in how cancer patients can access vast medical information. However, given that ChatGPT was not explicitly trained for oncology-related inquiries, the quality of the information it provides still needs to be verified. Evaluating the quality of responses is crucial, as misinformation can foster a false sense of knowledge and security, lead to noncompliance, and delay appropriate treatment.

Objective: This study aims to evaluate the quality and reliability of ChatGPT's responses to standard patient queries about radiotherapy, comparing the performance of GPT-3.5 and GPT-4.

Methods: Forty commonly asked radiotherapy questions were selected and inserted into both versions. Responses were evaluated by six radiotherapy experts using a General Quality Score (GQS), assessed for consistency and similarity using the cosine similarity score, and analyzed for readability using the Flesch Reading Ease Score (FRES) and Flesch-Kincaid Grade Level (FKGL). Statistical analysis was performed using the Mann-Whitney test.

Results: GPT-4 demonstrated superior performance, with higher GQS and a complete absence of lower scores compared to GPT-3.5. The Mann-Whitney test revealed statistically significant differences in some questions, with GPT-4 generally receiving higher ratings. The cosine similarity score indicated substantial similarity and consistency in responses from both versions. Readability scores for both versions were considered college-level, with GPT-4 scoring slightly better in FRES (35.55) and FKGL (12.71) compared to GPT-3.5 (30.68 and 13.53, respectively). Both versions' responses were deemed challenging for the public to read.

Conclusions: While GPT-4 generates more accurate and reliable responses than GPT-3.5, both models present readability challenges for the public. ChatGPT reveals potential as a valuable resource for addressing common patient queries related to radiotherapy. However, it's crucial to acknowledge its limitations, including the risks of misinformation and readability issues.

Keywords: Artificial intelligence; ChatGPT; Large language model; Radiotherapy; Patient information.

Introduction

In an increasingly digitized society, patients frequently resort to the Internet to access information about cancer [1–3]. However, despite being one of the patients' most favored informational modalities, online websites often need more content accuracy and better readability [1].

Recently, artificial intelligence (AI)-powered chatbots, such as Chat Generative Pre-trained Transformer (ChatGPT), signify a potential paradigm shift in how cancer patients can access a vast amount of medical information [1,3,4]. The rise of these AI platforms, accessible to the general public, has escalated notably over the past year since OpenAI released version 3.5 of ChatGPT (GPT-3.5) on November 30th, 2022 [4–13], which amassed over 1 billion users in March 2023 [4].

ChatGPT, a large language model (LLM) [6,14–17], utilizes natural language processing to offer varied responses to the same query, considering the context of the conversation and individual

user preferences [18]. Through text-to-text communication, ChatGPT can engage with humans [12] and aims to deliver responses resembling human interactions [6,14,18]. This model has undergone extensive training on a diverse corpus of text data, enabling it to comprehend and respond to natural language queries across various topics [14]. Additionally, ChatGPT can compose emails, essays, and medical reports, as well as solve problems and provide clarifications [10,13,19,20].

On March 14th, 2023, OpenAI announced the release of ChatGPT-4 (GPT-4), which became available through a subscription-based model [9,12,16]. This new version has demonstrated outstanding performance across numerous academic and professional benchmarks, providing more refined and varied responses compared to GPT-3.5 [21].

In this context, ChatGPT has emerged as a contender to traditional search engines like Google due to its capacity to filter through vast quantities of data and provide easily comprehensible responses [4,6]. Consequently, ChatGPT presents itself as a potentially reliable source of medical information to both the public and cancer patients, capable of offering insights regarding radiotherapy [4,22]. This is particularly significant given the general public's limited knowledge about this treatment [15,23] and the fears about its possible side effects [24]. Providing patients with comprehensive radiotherapy information at the appropriate stages may enhance adherence to the treatment plan, as inadequate information can lead to increased uncertainty, needless anxiety, and distress among patients and their families [24–26]. Additionally, poorly informed patients are likely to be dissatisfied with their care, have difficulty coping [25] and often have many follow-up questions regarding the treatment process. Moreover, cancer patients often feel uncomfortable discussing body image and sexual health with clinicians. Consequently, patient communication with ChatGPT may lower these barriers [26].

However, given that ChatGPT was not explicitly trained for oncology-related inquiries, the quality of the information it provides remains unverified [7,14,26]. Evaluating the quality of responses is crucial, as misinformation can foster a false sense of knowledge and security, lead to noncompliance, and result in delays in receiving appropriate treatment [4,14,15]. Nevertheless, various limitations of ChatGPT have been identified, such as its tendency to provide unreliable or inaccurate information, potentially generating incorrect or misleading responses [14,16]. Furthermore, it has been observed to fall below the expected educational level [4], as health-related materials intended for patient consumption are typically recommended to have a reading level equivalent to fifth and sixth grade [4,27]. Additionally, the training data for GPT-3.5 is outdated, limited to information available up to September 2021, lacking access to newer knowledge beyond that date [5,28,29]. To address this constraint, GPT-4 introduces a novel feature wherein external plug-ins can be utilized [22].

To date, limited research has been conducted on the application of language models in the medical domain, and the effectiveness of ChatGPT in patient education remains indeterminate [14]. While literature addressing ChatGPT's capabilities has proliferated in recent months, there remains a lack of data regarding the quality and reliability of the responses it provides [11,18].

This study aimed to evaluate the quality and reliability of ChatGPT's responses to common patient queries regarding radiotherapy to ascertain its potential as a reliable source of information for patients. Additionally, it aims to compare the performance of GPT-3.5 with GPT-4 in generating responses to the same radiotherapy queries.

Methods

To determine the most common patient queries regarding radiotherapy, an assessment was

conducted on articles by Halkett et al. [24,25,30], Zeguers et al. [31], and the National Cancer Institute [32]. One hundred twenty-eight questions were formulated based on topics delineated in the examined articles. These were designed to address patients' informational needs at various stages of their radiotherapy treatment, and the key messages typically communicated by healthcare professionals during this period [24]. From this set of questions, forty queries were selected for insertion on ChatGPT, excluding duplicates and queries specific to pathologies or specialized treatments. This exclusion aimed to ensure that the responses could apply to all patients receiving radiotherapy, thereby reflecting their primary concerns and doubts. Therefore, the final set of questions was subcategorized into general information (n = 14), planning and treatment (n = 16), and side effects (n = 10) (Table 1). These dimensions were chosen to assess potential strengths and weaknesses of responses relating to different topics of radiotherapy. Question phrasing was intentionally written in the first person to approximate how an average patient may enter their question into ChatGPT [33].

General Information	Planning and Treatment
1. Why is radiotherapy recommended?	1. Can I maintain my daily routine and activities during radiotherapy?
2. What does radiotherapy involve?	2. Can I keep working while undergoing radiotherapy treatments?
3. When should radiotherapy and chemotherapy be combined?	3. Are complementary medicines recommended while undergoing radiotherapy treatments?
4. What's the cost of radiotherapy treatment?	4. What's the planning appointment in radiotherapy and what does it involve?
5. Who will be providing my radiotherapy treatment?	5. Why is computed tomography (CT) planning necessary in radiotherapy?
6. How does the radiotherapy treatment machine work?	6. Why are tattoos useful in radiotherapy CT planning?
7. What impact will radiotherapy treatment have on my life?	7. What happens on the first day of radiotherapy treatment?
8. What impact will radiotherapy treatment have on my health in the future?	8. Will the radiotherapy treatment schedule be adjusted to my availability?
9. During the period of radiotherapy, will I have to follow a particular diet?	9. What am I expected to do during the radiotherapy treatment?
10. Will radiotherapy make me radioactive?	10. Does the radiotherapy machine make noise?
11. What does radiotherapy do to healthy cells?	11. How close is the radiotherapy treatment machine going to get?
12. How long does radiotherapy take to work?	12. What happens during radiotherapy treatment?
13. Can I be cured of my disease through radiotherapy treatments?	13. Is there a possibility of experiencing pain due to the radiotherapy treatment?
14. What will happen after the radiotherapy treatment is finished?	14. How long does a radiotherapy session last?
	15. What should I wear for radiotherapy treatment?
	16. Will there be follow-up after the end of radiotherapy treatments?
Side Effects	

1. What are the side effects of radiotherapy?
2. What skin care should I have during and after radiotherapy?
3. Am I going to feel tired after the radiotherapy treatments?
4. What hygiene care should be taken after radiotherapy treatments?
5. Which steps should be taken to reduce radiotherapy side effects?
6. Will the radiotherapy treatment be interrupted if I experience adverse side effects?
7. Who can I go to if the radiotherapy side effects become too burdensome?
8. Will radiotherapy affect my fertility?
9. Will radiotherapy cause hair loss?
10. Will radiotherapy cause permanent damage?

Table 1. Common patient queries regarding radiotherapy by dimension inserted in ChatGPT.

Responses were collected from ChatGPT from 06.04.24 to 09.04.24. Each question was queried to both versions of ChatGPT in the English language. Each query was entered separately using the "New Chat" function, acknowledging that ChatGPT considers the context of the conversation, which could influence the responses. The queries were then regenerated in each version of ChatGPT, and both responses were documented to analyze the consistency. It is important to note that no patient records were used in this research, therefore, ethics committee approval was not required.

Various methods were employed, as described below, to assess the quality and reliability of the response content, response consistency, response readability, and similarity between responses from GPT-3.5 and GPT-4.

Quality and Reliability

To evaluate the quality and reliability of the information provided by ChatGPT, we employed a 5-point Likert scale, known as the General Quality Score (GQS), which has been used in previous studies [14,34]. This assessment criteria included accuracy, use of lay language, information flow, usefulness, and empathy. The 5-point Likert scale was defined as follows: (1) inaccurate information, poorly organized text, missing important details, and not helpful for patients, (2) limited accuracy, some relevant information is present, but still not easily understandable for patients, (3) adequately accurate information and some important details are explained in plain language, (4) accurate information, well-organized text, and most relevant details are presented in a patient-friendly manner, (5) extremely accurate information, well-structured text, and all relevant details are presented in a compassionate and patient-friendly way [14].

The mean GQS was calculated by averaging the ratings given by six independent radiotherapy experts. Among these six experts, three evaluated the responses from GPT-3.5, while the remaining three evaluated the responses from GPT-4. Each expert evaluated only the responses from one of ChatGPT's versions to reduce potential bias during the evaluation process, thereby decreasing the likelihood of altering assessments and enhancing their credibility [35]. Furthermore, the authors that analysed the obtained results were unaware of the identity of the radiotherapy experts.

Consistency and Similarity

The consistency and similarity of the responses were evaluated using the cosine similarity

score. This method involves transforming the text information provided by ChatGPT into vectors, and then calculating the cosine of the angle between two vectors, which indicates how similar the responses are to each other. This score was calculated using an online tool. The cosine similarity score ranges from 0 to 1, where a score of 0 indicates complete dissimilarity between the texts, and a score of 1 indicates complete similarity [14,26].

To assess the similarity between the responses generated by GPT-3.5 and GPT-4, the initial responses to the same question provided by both versions were inserted into the online tool to determine the cosine similarity score between them.

The consistency of the responses generated by ChatGPT was assessed by regenerating the same question into both versions and calculating the cosine similarity score between the two responses to the same question. By regenerating the same question, we aim to assess whether ChatGPT can provide consistent information or if its responses vary widely.

Readability

To evaluate readability, responses from both versions were assessed using an online Flesch Reading Ease test score calculator. This calculator determined the responses' readability using two different indices: the Flesch Reading Ease Score (FRES) and the Flesch-Kincaid Grade Level (FKGL). These readability tests use mathematical formulas that consider factors such as sentence length and word count. The FRES is a numerical score ranging from 0 to 100, with higher numbers indicating better readability, meaning the content is easier to read and understand [8,36,37], and corresponding to a lower grade level [4,36,37] (Table 2). The FKGL score indicates the average number of years of education needed to comprehend a text, with lower scores suggesting better readability [8,36,37] and correlating to the equivalent school level [4,36] (Table 3).

Table 2. Flesch Reading Ease Score.

Score	Grade Level	Summary
90 - 100	5th grade	Very easy to read
80 - 90	6th grade	Easy to read
70 - 80	7th grade	Fairly easy to read
60 - 70	8th & 9th grade	Plain English
50 - 60	10th to 12th grade	Fairly difficult to read
30 - 50	College	Difficult to read
10 - 30	College graduate	Very difficult to read
0 - 10	Professional	Extremely difficult to read

Table 3. Flesch-Kincaid Grade Level Score.

Score	School Level
0 - 3	Kindergarten/Elementary
3 - 6	Elementary
6 - 9	Middle School
9 - 12	High School
12 - 15	College
15 - 18	Post-Graduate

Statistical Analysis

The data were analyzed using the SPSS statistical software, version 29.0 for Windows®. The results were considered significant at a 5% significance level (P = .05). An exploratory data analysis was carried out using frequency analysis (n, %) for qualitative data and mean and standard deviation (SD) for quantitative data. The Mann-Whitney test was applied to compare evaluations between the two versions of ChatGPT.

Results

Quality and Reliability

GPT-3.5 received primarily mid-range scores, with most evaluations at levels 3 (n = *14*) and 4 (n = *18*), indicating generally accurate and comprehensible responses. Notably, some responses had the highest rating of 5 (n = *5*), providing extremely accurate and well-structured information. However, it also received low scores of 1 (n = *1*) and 2 (n = *2*) suggesting limited or inaccurate information.

Conversely, GPT-4 received predominantly the highest score of 5 (n = *25*), indicating a superior ability to provide accurate and well-structured information. Some responses were assigned a score of 4 (n = *14*), and one response received a score of 3 (n = *1*), illustrating that it frequently provided responses that were accurate, well-organized, and accessible to patients. Remarkably, GPT-4 exhibited a complete absence of lower scores (1 and 2). The score breakdown by question dimension is shown in Figure 1.

Figure 1. Scores assigned by radiotherapy experts to the total number of responses in each dimension from (a) ChatGPT-3.5 and (b) ChatGPT-4.

Considering the general information dimension, statistically significant differences were detected between the two versions of ChatGPT regarding question 3 (P = .043). Regarding planning and treatment, statistically significant differences were detected in questions 5 (P = .043), 7 (P = .043), 9 (P = .037) and 12 (P = .043). Regarding side effects, statistically significant differences were detected in question 4 (P = .025). In either situation, GPT-4 showed higher ratings (Table 4).

Table 4. Comparison of general information, planning and treatment and side effects between the two versions of ChatGPT. Mann-Whitney test results.

Questions		Test Statistic		
		ChatGPT-3.5	ChatGPT-4	*P*
		Mean Rank		
General Information	Q.1.	2,33	4,67	.099
	Q.2.	2,33	4,67	.105
	Q.3.	2,00	5,00	.043
	Q.4.	3,33	3,67	.814
	Q.5.	3,17	3,83	.637
	Q.6.	3,17	3,83	.653
	Q.7.	2,83	4,17	.346
	Q.8.	2,17	4,83	.072
	Q.9.	2,50	4,50	.121
	Q.10.	2,83	4,17	.346
	Q.11.	2,33	4,67	.099
	Q.12.	2,33	4,67	.099
	Q.13.	3,50	3,50	1
	Q.14.	2,50	4,50	.114
Planning and Treatment	Q.1.	3,33	3,67	.796
	Q.2.	2,50	4,50	.121
	Q.3.	4,00	3,00	.317
	Q.4.	2,33	4,67	.105
	Q.5.	2,00	5,00	.043
	Q.6.	2,17	4,83	.068
	Q.7.	2,00	5,00	.043
	Q.8.	3,00	4,00	.317
	Q.9.	2,00	5,00	.037
	Q.10.	3,00	4,00	.317
	Q.11.	2,67	4,33	.246
	Q.12.	2,00	5,00	.043
	Q.13.	2,83	4,17	.346
	Q.14.	3,17	3,83	.637
	Q.15.	2,50	4,50	.121
	Q.16.	3,00	4,00	.317
Side Effects	Q.1.	2,50	4,50	.121
	Q.2.	3,33	3,67	.817
	Q.3.	3,00	4,00	.317
	Q.4.	2,00	5,00	.025
	Q.5.	2,67	4,33	.261
	Q.6.	3,50	3,50	1
	Q.7.	3,00	4,00	.487
	Q.8.	2,50	4,50	.121
	Q.9.	2,50	4,50	.121
	Q.10.	3,00	4,00	.456

Consistency and Similarity

Regarding similarity and consistency, a cosine similarity score ranging from 0 to 1 was calculated, as previously described. Concerning similarity, the mean (SD) cosine similarity between GPT-3.5 and GPT-4 responses was 0.80 (0.04), indicating a reasonably good similarity between the two versions of the ChatGPT. Notably, question 11 in the planning and treatment dimension exhibited the lowest similarity, with a value of 0.68. With respect to consistency, the cosine similarity mean (SD) for GPT-3.5 and GPT-4 responses was 0.85 (0.04) and 0.84 (0.03),

respectively. In both versions, the consistency was demonstrated to be good or very good, with values ranging between 0.74 and 0.92.

Readability

Word count, sentence count, FRES and FKGL for both versions are summarized in Table 5. A significant disparity was observed in the mean (SD) word count between GPT-3.5 and GPT-4 [265.18 (102.49) versus 385.83 (73.20)]. Additionally, the sentence count was higher in GPT-4 compared to GPT-3.5 [19.88 (6.90) versus 16.15 (8.61)].

The FRES mean (SD) for GPT-3.5 and GPT-4 responses were 30.68 (12.89) and 35.55 (12.24), respectively. This indicates that the responses generated by the two versions were considered college-level and difficult to read. The FKGL mean (SD) for GPT-3.5 and GPT-4 responses were 13.53 (2.63) and 12.71 (2.65), respectively. This suggests that an average of 14 years of education (college-level) is required to understand the responses generated by GPT-3.5, whereas the responses from GPT-4 necessitate an average of 13 years of education (college-level) for comprehension.

Table 5. Word count, sentence count, Flesch Reading Ease Score and Flesch-Kincaid Grade Level score of responses from ChatGPT-3.5 and ChatGPT-4.

	Questions	ChatGPT-3.5				ChatGPT-4			
		Word count	Sentence count	FRES	FKGL	Word count	Sentence count	FRES	FKGL
General Information	Q.1.	332	22	35,31	12,08	378	25	32,36	12,5
	Q.2.	414	27	35,97	12,05	453	28	35,97	12,26
	Q.3.	304	18	24,11	14,09	340	17	25,8	14,63
	Q.4.	188	7	13,53	18	268	15	25,81	14,1
	Q.5.	246	18	28,58	12,67	305	21	21,78	13,83
	Q.6.	378	27	35,96	11,72	431	27	36,55	12,13
	Q.7.	422	27	35	12,26	358	27	41,9	10,71
	Q.8.	389	22	32,74	13,09	351	16	30,79	14,41
	Q.9.	332	25	41,23	10,81	311	25	57,92	8,27
	Q.10.	84	5	17,56	14,98	223	16	21,59	13,71
	Q.11.	304	26	36,9	11,02	352	21	37,45	12,2
	Q.12.	178	7	33,21	14,94	298	15	47,28	11,6
	Q.13.	177	8	27,61	14,91	231	9	23,3	16,39
	Q.14.	348	22	35,92	12,18	410	13	19,24	18
Planning and Treatment	Q.1.	374	28	49,19	9,72	402	30	52,44	9,27
	Q.2.	229	8	21,88	17,32	369	22	52,94	10,04
	Q.3.	165	8	15,25	16,99	359	20	16,19	10,93
	Q.4.	361	21	26,51	13,83	433	25	39,01	12,12
	Q.5.	316	16	16,79	15,82	378	24	26,8	13,43
	Q.6.	214	12	15,19	15,57	332	21	30,51	12,93
	Q.7.	358	22	34,35	12,51	472	27	48,93	10,78
	Q.8.	140	5	20,7	17,33	232	15	33,24	12,47
	Q.9.	361	27	40,94	10,87	416	31	43,54	10,52
	Q.10.	76	4	30,59	13,71	106	5	31,28	14,16
	Q.11.	164	6	0	18	314	16	33,07	13,52
	Q.12.	335	24	37,1	11,55	388	28	39,71	11,16
	Q.13.	247	16	37,38	11,88	315	19	37,73	12,12
	Q.14.	144	5	0	18	264	13	28,24	14,37
	Q.15.	327	24	52,01	9,39	337	22	62,5	8,35
	Q.16.	183	7	19,42	17,05	339	20	37,9	12,18
Side Effects	Q.1.	340	26	48	9,81	324	11	26,28	16,91
	Q.2.	300	26	57,23	8,14	354	33	52,8	8,56
	Q.3.	150	7	36,19	13,54	270	9	32,57	16,17
	Q.4.	411	23	43,17	11,68	361	28	48,22	9,74
	Q.5.	397	21	17,81	15,47	418	17	13,49	17,49
	Q.6.	137	5	32,05	15,6	349	20	37,38	12,38
	Q.7.	298	21	45,09	10,5	371	27	38,28	11,33
	Q.8.	164	8	23,53	15,07	277	10	17,15	17,75
	Q.9.	108	5	43,91	12,5	214	15	57,94	8,72
	Q.10.	212	10	29,29	14,44	330	12	26,13	16,45

Discussion

Principal Results

The power and utility of AI platforms in healthcare, such as ChatGPT, is rapidly evolving and improving, carrying the potential to significantly improve patient education [5,38]. Our results demonstrate that although both versions of ChatGPT can accurately answer queries about radiotherapy, they reveal reduced accuracy when responding to queries related to planning and treatment, as Valentini et al. has similarly demonstrated [39]. However, the absence of lower scores (1 and 2) the radiotherapy experts gave to GPT-4's responses signify a noteworthy improvement in its response quality and reliability. Therefore, GPT-4 showcased superior performance compared to GPT-3.5, as supported by several studies [9,21,22,33,36].

Although most responses were correct or close to correct, upon comparing the accuracy of responses between GPT-4 and GPT-3.5 in the three dimensions, it became evident that GPT-4 consistently offered improved elucidation of specific concepts relevant to radiotherapy treatment. In question 10 of the general information dimension, GPT-4 specifically delineated that patients are non-radioactive and may safely interact with others post-treatment (*"You can safely be around others, including children and pregnant women, without any risk of exposing them to radiation"*). Additionally, within the side effects dimension, in questions 2 and 3, GPT-4 emphasized that the intended creams to use throughout radiotherapy treatment should be only those recommended by the healthcare provider (*"Apply a fragrance-free moisturizer recommended by your healthcare provider"*) and specified strategies to mitigate fatigue, a treatment-related side effect. Moreover, for questions 7, 11, and 12, within the planning and treatment dimension, GPT-3.5 demonstrated a propensity to diverge from directly addressing the queried issue, in contrast to GPT-4. However, in GPT-4´s response to the 13th planning and treatment question, specific information was inaccurately presented as it erroneously stated that radiotherapy induces direct pain (*"Direct Pain from Treatment Site: Radiotherapy can cause localized pain at the site of treatment"*).

Moreover, there were a few occasions in both versions where a lack of information was demonstrated. For instance, in question 7 of the side effects dimension, neither version mentioned that radiation therapists could serve as advisors for patients experiencing severe side effects.

In most of the responses, ChatGPT used a typical structure characterized by a succinct introductory paragraph, followed by five or six bullet points delineating the responses, and culminating in a short concluding paragraph. Additionally, in a fair number of responses generated by GPT-3.5 ($n = 25$) and GPT-4 ($n = 28$), a statement was included advising that the information provided should always be discussed with the healthcare providers, consistent with prior studies [33,40,41]. The cosine similarity score indicated a reasonably substantial similarity and consistency, and while subtle changes in sentence structure were noted, the answers remained consistent, implying accuracy [3]. However, the bibliography used to obtain the information were not disclosed in either version, indicating their incapacity to inform users of the contentious nature of certain information [7,19].

Concerning readability, all responses were considered more difficult to read than the sixth grade reading level recommended for patient consumption, a concern highlighted in prior studies [4,27]. This finding suggests that although the content was predominantly accurate, it was presented at a level too advanced for the public, particularly for individuals with lower health literacy [27,42]. Due to the heightened challenges these patients face in understanding their disease, radiation treatment, and potential side effects, this bears particular significance [26].

A similar study in the radiation oncology field was conducted by Yalamanchili et al. to

determine the quality of GPT-3.5's responses to patient care questions. The authors verified that the LLM generated accurate, comprehensive, and concise responses with minimal risk of harm, using language identical to human experts, however at a higher reading level. To overcome this issue, it was suggested that direct prompts such as "Explain this to me like I am in fifth grade" could assist in generating simplified responses. These findings indicate the potential of the LLM as a valuable resource for addressing patient queries in radiation oncology and other medical fields [26].

This investigation reveals that this platform, particularly GPT-4, can generate high quality responses to radiotherapy queries. Therefore, ChatGPT is a convenient and powerful tool, empowering patients by granting them access to accurate medical information, thereby facilitating a more effective shared decision-making process [6,36,43]. Additionally, it holds promise for offering clinical guidance, suggesting treatment options, and serving as a valuable resource for medical education [42,43]. Hence, it could potentially serve as an alternative to current online resources [26].

Limitations

This study encountered several limitations. Firstly, the formulation and phrasing of queries in both versions may have influenced ChatGPT's performance. Additionally, the queries were exclusively written in English, which restricted responses to a single language. Although the total number of questions was comparable to other studies, the optimal quantity of queries needed to evaluate the model effectively remains undetermined. Furthermore, the scoring process inherently involves subjectivity, particularly with GQS, as different raters may interpret and prioritize quality aspects differently. Moreover, this study was conducted within a specific time frame (April 2024), and ChatGPT is expected to improve continuously over time. Repeating this study later could improve response quality.

The other limitations of our study are a result of the limitations of ChatGPT itself. Firstly, it should be noted that the information provided by GPT-3.5 is only current up to September 2021. Meanwhile, GPT-4 has a limited number of questions that can be posed within a specific timeframe and is exclusively accessible through paid subscription, potentially constraining the public's access to more accurate information. Finally, ChatGPT is one of many AI models available, making it uncertain whether the responses obtained represent the general characteristics of all LLMs. Consequently, further research is essential to fully comprehend ChatGPT's role in patient education.

Conclusions

Our study revealed that GPT-3.5 and GPT-4 generate reliable responses, with GPT-4 demonstrating superior performance. Although the readability scores for GPT-4 were slightly better, both versions were considered difficult for the public to read. Therefore, caution is advised regarding potential misinformation and readability issues. Additionally, GPT-4's paid subscription may exacerbate existing healthcare disparities. Consequently, while ChatGPT reveals potential as a valuable resource for patients, accurately addressing common patient queries about radiotherapy, its limitations must be acknowledged.

Acknowledgments

AMG and MC contributed to the conceptualization, methodology, writing/review and editing, supervision, and project administration. CM and MCR were responsible for the conceptualization, methodology, investigation, resources, data curation and writing the original

draft. EC conducted the statistical analysis.

Conflicts of Interest

None declared.

Abbreviations

AI: artificial intelligence
ChatGPT: Chat Generative Pre-trained Transformer
GPT-3.5: version 3.5 of ChatGPT
LLM: large language model
GPT-4: ChatGPT-4
GQS: General Quality Score
FRES: Flesch Reading Ease Score
FKGL: Flesch-Kincaid Grade Level
SD: standard deviation

References

1. Moazzam Z, Cloyd J, Lima HA, Pawlik TM. Quality of ChatGPT Responses to Questions Related to Pancreatic Cancer and its Surgical Care. Ann Surg Oncol. Springer Science and Business Media Deutschland GmbH; 2023. p. 6284–6286. PMID:37349615
2. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT. Radiology Radiological Society of North America Inc.; 2023 May 1;307(4). PMID:37014239
3. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. JNCI Cancer Spectr Oxford University Press; 2023 Apr 1;7(2). doi: 10.1093/jncics/pkad015
4. Davis RJ, Ayo-Ajibola O, Lin ME, Swanson MS, Chambers TN, Kwon DI, Kokot NC. Evaluation of Oropharyngeal Cancer Information from Revolutionary Artificial Intelligence Chatbot. Laryngoscope John Wiley and Sons Inc; 2024 May 1;134(5):2252–2257. PMID:37983846
5. Gabriel J, Shafik L, Alanbuki A, Larner T. The utility of the ChatGPT artificial intelligence tool for patient education and enquiry in robotic radical prostatectomy. Int Urol Nephrol Springer Science and Business Media B.V.; 2023 Nov 1;55(11):2717–2732. PMID:37528247
6. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, Ayoub W, Yang JD, Liran O, Spiegel B, Kuo A. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol Korean Association for the Study of the Liver; 2023 Jul 1;29(3):721–732. PMID:36946005
7. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, Staubli SM. Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. J Med Internet Res JMIR Publications Inc.; 2023;25. PMID:37389908
8. Wei K, Fritz C, Rajasekaran K. Answering head and neck cancer questions: An assessment of ChatGPT responses. American Journal of Otolaryngology - Head and Neck Medicine and Surgery W.B. Saunders; 2024 Jan 1;45(1). PMID:37844413
9. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, Cielo D, Oyelese AA, Doberstein CE, Telfeian AE, Gokaslan ZL, Asaad WF. Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. Neurosurgery Wolters Kluwer Medknow Publications; 2023 Nov 1;93(5):1090–1098. PMID:37306460

10. Hopkins AM, Logan JM, Kichenadasse G, Sorich MJ. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. JNCI Cancer Spectr. Oxford University Press; 2023. PMID:36808255
11. Mukherjee S, Durkin C, PeBenito AM, Ferrante ND, Umana IC, Kochman ML. Assessing ChatGPT's Ability to Reply to Queries Regarding Colon Cancer Screening Based on Multisociety Guidelines. Gastro Hep Advances. American Gastroenterological Association; 2023. p. 1040–1043. doi: 10.1016/j.gastha.2023.07.008
12. Uprety D, Zhu D, West H. ChatGPT—A promising generative AI tool and its implications for cancer care. Cancer. John Wiley and Sons Inc; 2023. p. 2284–2289. PMID:37183438
13. Guckenberger M, Andratschke N, Ahmadsei M, Christ SM, Heusel AE, Kamal S, Kroese TE, Looman EL, Reichl S, Vlaskou Badra E, von der Grün J, Willmann J, Tanadini-Lang S, Mayinger M. Potential of ChatGPT in facilitating research in radiation oncology? Radiotherapy and Oncology Elsevier Ireland Ltd; 2023 Nov 1;188. PMID:37659658
14. Coskun B, Ocakoglu G, Yetemen M, Kaygisiz O. Can ChatGPT, an Artificial Intelligence Language Model, Provide Accurate and High-quality Patient Information on Prostate Cancer? Urology Elsevier Inc.; 2023 Oct 1;180:35–58. PMID:37406864
15. Chow JCL, Wong V, Sanders L, Li K. Developing an AI-Assisted Educational Chatbot for Radiotherapy Using the IBM Watson Assistant Platform. Healthcare (Switzerland) Multidisciplinary Digital Publishing Institute (MDPI); 2023 Sep 1;11(17). doi: 10.3390/healthcare11172417
16. Li J, Zhong J, Li Z, Xiao Y, Wang S. Ectopic Pituitary Neuroendocrine Tumor: A Case Report Written With the Help of ChatGPT. Cureus Springer Science and Business Media LLC; 2023 Oct 14; doi: 10.7759/cureus.46999
17. Lai UH, Wu KS, Hsu TY, Kan JKC. Evaluating the performance of ChatGPT-4 on the United Kingdom Medical Licensing Assessment. Front Med (Lausanne) Frontiers Media SA; 2023;10. doi: 10.3389/fmed.2023.1240915
18. Trager MH, Queen D, Bordone LA, Geskin LJ, Samie FH. Assessing ChatGPT responses to common patient queries regarding basal cell carcinoma. Arch Dermatol Res. Institute for Ionics; 2023. p. 2979–2981. PMID:37668714
19. Szczesniewski JJ, Tellez Fouz C, Ramos Alba A, Diaz Goizueta FJ, García Tello A, Llanes González L. ChatGPT and most frequent urological diseases: analysing the quality of information and potential risks for patients. World J Urol Springer Science and Business Media Deutschland GmbH; 2023 Nov 1;41(11):3149–3153. PMID:37632558
20. Emile SH, Horesh N, Freund M, Pellino G, Oliveira L, Wignakumar A, Wexner SD. How appropriate are answers of online chat-based artificial intelligence (ChatGPT) to common questions on colon cancer? Surgery (United States). Elsevier Inc.; 2023. p. 1273–1275. PMID:37482439
21. Holmes J, Liu Z, Zhang L, Ding Y, Sio TT, McGee LA, Ashman JB, Li X, Liu T, Shen J, Liu W. Evaluating large language models on a highly-specialized topic, radiation oncology physics. Front Oncol Frontiers Media SA; 2023;13. doi: 10.3389/fonc.2023.1219326
22. Huang Y, Gomaa A, Semrau S, Haderlein M, Lettmaier S, Weissmann T, Grigo J, Tkhayat H Ben, Frey B, Gaipl U, Distel L, Maier A, Fietkau R, Bert C, Putz F. Benchmarking ChatGPT-4 on a radiation oncology in-training exam and Red Journal Gray Zone cases: potentials and challenges for ai-assisted medical education and decision making in radiation oncology. Front Oncol Frontiers Media SA; 2023;13. doi: 10.3389/fonc.2023.1265024
23. Smith SK, Zhu Y, Dhillon HM, Milross CG, Taylor J, Halkett G, Zilliacus E. Supporting patients with low health literacy: What role do radiation therapists play? Supportive Care in Cancer 2013 Nov;21(11):3051–3061. PMID:23812495
24. Halkett GKB, Kristjanson LJ, Lobb E, O'Driscoll C, Taylor M, Spry N. Meeting breast cancer

patients' information needs during radiotherapy: what can we do to improve the information and support that is currently provided? Eur J Cancer Care (Engl) 2010 Jul;19(4):538–547. PMID:19708930
25. Halkett GKB, Kristjanson LJ, Lobb E, Little J, Shaw T, Taylor M, Spry N. Information needs and preferences of women as they proceed through radiotherapy for breast cancer. Patient Educ Couns 2012 Mar;86(3):396–404. PMID:21664788
26. Yalamanchili A, Sengupta B, Song J, Lim S, Thomas TO, Mittal BB, Abazeed ME, Teo PT. Quality of Large Language Model Responses to Radiation Oncology Patient Care Questions. JAMA Netw Open American Medical Association; 2024;E244630. PMID:38564215
27. Young JN, Ross O'Hagan, Poplausky D, Levoska MA, Gulati N, Ungar B, Ungar J. The utility of ChatGPT in generating patient-facing and clinical responses for melanoma. J Am Acad Dermatol Elsevier Inc.; 2023 Sep 1;89(3):602–604. PMID:37207955
28. Talyshinskii A, Naik N, Hameed BMZ, Zhanbyrbekuly U, Khairli G, Guliev B, Juilebø-Jones P, Tzelves L, Somani BK. Expanding horizons and navigating challenges for enhanced clinical workflows: ChatGPT in urology. Front Surg. Frontiers Media SA; 2023. doi: 10.3389/fsurg.2023.1257191
29. Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. J Med Internet Res JMIR Publications Inc.; 2023;25. PMID:37379067
30. Halkett GKB, Kristjanson LJ. Validity and reliability testing of two instruments to measure breast cancer patients' concerns and information needs relating to radiation therapy. Radiation Oncology 2007 Nov 25;2(1). PMID:18036247
31. Zeguers M, De Haes HCJM, Zandbelt LC, Ter Hoeven CL, Franssen SJ, Geijsen DD, Koning CCE, Smets EMA. The information needs of new radiotherapy patients: How to measure? Do they want to know everything? and if not, why? Int J Radiat Oncol Biol Phys 2012 Jan 1;82(1):418–424. PMID:21075556
32. Cancer Institute N. Radiation Therapy and You: Support for People with Cancer. Available from: www.cancer.gov
33. Hermann CE, Patel JM, Boyd L, Growdon WB, Aviki E, Stasenko M. Let's chat about cervical cancer: Assessing the accuracy of ChatGPT responses to cervical cancer questions. Gynecol Oncol Academic Press Inc.; 2023 Dec 1;179:164–168. PMID:37988948
34. Onder CE, Koc G, Gokbulut P, Taskaldiran I, Kuskonmaz SM. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. Sci Rep Nature Research; 2024 Dec 1;14(1). PMID:38167988
35. Parker RA, Berman NG. Blinding in Observational Studies. Planning Clinical Research Cambridge University Press; 2016. p. 334–340. doi: 10.1017/cbo9781139024716.029
36. Chiarelli G, Stephens A, Finati M, Cirulli GO, Beatrici E, Filipas DK, Arora S, Tinsley S, Bhandari M, Carrieri G, Trinh QD, Briganti A, Montorsi F, Lughezzani G, Buffi N, Rogers C, Abdollah F. Adequacy of prostate cancer prevention and screening recommendations provided by an artificial intelligence-powered large language model. Int Urol Nephrol Springer Science and Business Media B.V.; 2024; doi: 10.1007/s11255-024-04009-5
37. Thia I, Saluja M. ChatGPT: Is This Patient Education Tool for Urological Malignancies Readable for the General Population? Res Rep Urol Dove Medical Press Ltd; 2024;16:31–37. doi: 10.2147/RRU.S440633
38. Kuşcu O, Pamuk AE, Sütay Süslü N, Hosal S. Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer? Front Oncol Frontiers Media SA; 2023;13. doi: 10.3389/fonc.2023.1256459
39. Valentini M, Szkandera J, Smolle M, Scheipl S, Leithner A, Andreou D. Artificial intelligence large language model ChatGPT: is it a trustworthy and reliable source of information for sarcoma patients? Front Public Health Frontiers Media SA; 2024;12. PMID:38584922

40. Rogasch JMM, Metzger G, Preisler M, Galler M, Thiele F, Brenner W, Feldhaus F, Wetz C, Amthauer H, Furth C, Schatka I. ChatGPT: Can You Prepare My Patients for [18F] FDG PET/CT and Explain My Reports? . Journal of Nuclear Medicine Society of Nuclear Medicine; 2023 Dec 14;jnumed.123.266114. doi: 10.2967/jnumed.123.266114
41. Braun EM, Juhasz-Böss I, Solomayer EF, Truhn D, Keller C, Heinrich V, Braun BJ. Will I soon be out of my job? Quality and guideline conformity of ChatGPT therapy suggestions to patient inquiries with gynecologic symptoms in a palliative setting. Arch Gynecol Obstet Springer Science and Business Media Deutschland GmbH; 2024 Apr 1;309(4):1543–1549. PMID:37975899
42. Braithwaite D, Karanth SD, Divaker J, Schoenborn N, Lin K, Lancaster PM, Health G, Richman I, Hochhegger B, Schonberg M, Israel B. Evaluating ChatGPT's Accuracy in Providing Screening Mammography Recommendations among Older Women: Articial Intelligence and Cancer Communication. 2024; doi: 10.21203/rs.3.rs-3911155/v1
43. Choi J, Kim JW, Lee YS, Tae JH, Choi SY, Chang IH, Kim JH. Availability of ChatGPT to provide medical information for patients with kidney cancer. Sci Rep Nature Research; 2024 Dec 1;14(1). PMID:38233511

Supplementary Files

Figures

Untitled.