Radiography 32 (2026) 103469

Contents lists available at ScienceDirect

Radiography

journal homepage: www.elsevier.com/locate/radi

Breast cancer patients' questions about radiotherapy-induced skin toxicity: A comparative analysis of ChatGPT-4.0 and Gemini 2.5 flash

M. Caetano [a,*], D. Vicente [a], L. Caetano [a], E. Carolino [a,b], A. Grilo [a,c]

[a] Lisbon Higher School of Health, Polytechnic University of Lisbon, Lisbon, Portugal
[b] Health & Technology Research Center of Lisbon Higher School of Health (H&TRC), Polytechnic University of Lisbon, Lisbon, Portugal
[c] Research Center for Psychological Science of the Faculty of Psychology, University of Lisbon to CICPSI, Faculty of Psychology – Lisbon University, Portugal

ARTICLE INFO

Article history:
Received 14 January 2026
Received in revised form
9 April 2026
Accepted 28 May 2026
Available online xxx

Keywords:
Radiotherapy
Breast cancer
Skin toxicity
Patient education
Artificial intelligence
Large language models

ABSTRACT

Introduction: Artificial Intelligence (AI) is increasingly being deployed in health communication, a trend particularly visible in specialised domains, such as breast cancer radiotherapy. The subsequent clinical and educational value is fundamentally determined by the reliability and clarity of the responses generated. However, the comparative performance of different AI models in addressing patient concerns about radiotherapy side effects remains unclear, creating uncertainty regarding the optimal tool selection for patient education and support. This study aimed to compare the performance of ChatGPT 4.0 and Gemini 2.5 FLASH using a mixed-methods analytical-descriptive approach.
Methods: Twenty-three unique questions, derived from a literature review and rephrased to simulate patient enquiries about the skin effects of radiotherapy, were submitted to both models. Sixteen expert Radiation Therapists (RTTs) independently assessed the responses using a seven-point Likert scale. The analyses included semantic cosine similarity and linguistic readability (Flesch Reading Ease/Flesch-Kincaid Grade Level), with statistical comparisons performed using Mann–Whitney tests.
Results: Gemini 2.5 FLASH achieved higher median scores (6/7) than ChatGPT 4.0 (5/7), demonstrating particular strengths in clinical detail and empathy. Conversely, ChatGPT 4.0 produced more direct and structured answers, although it occasionally simplified complex concepts. Models showed low semantic similarity (median 0.78). Readability analysis revealed that ChatGPT aligned with an 8th-grade level, whereas Gemini operated at an 11th-grade level. Expert agreement was robust, with Gemini achieving greater consistency ($\alpha = 0.78$; $\kappa = 0.70$) than ChatGPT ($\alpha = 0.72$; $\kappa = 0.65$).
Conclusion: Gemini was more effective for complex psychosocial issues, whereas ChatGPT provided better accessible guidance, necessitating ongoing professional validation for reliable integration into patient education workflows.
Implications for practice: The use of AI in counselling patients undergoing breast radiotherapy enhances accessibility but requires ongoing professional validation to ensure clinical reliability.

Introduction

Artificial Intelligence (AI) has emerged as a transformative tool in healthcare, leveraging computational systems with learning and reasoning capabilities to support patient needs for reliable, accurate, and timely information.[1,2] Current literature suggests that the utilization of specific Artificial Intelligence (AI) tools for health-related inquiries is notable. Specifically, in a cohort of 2406 respondents, 21.5% (n = 517) reported using ChatGPT to acquire online health information.[3] Although their use in Europe remains low,[4] concerns regarding the reliability and accuracy of these technologies are particularly relevant in specialised contexts, such as radiotherapy, where clear and precise communication may directly influence patient outcomes.[3,4]

In breast cancer treatment, radiotherapy targets malignant cells using ionising radiation while preserving the surrounding healthy tissue.[5,6] One of the most common side effects is skin toxicity, which ranges from mild erythema to moist desquamation, and in some cases, chronic skin changes.[7,8] These potential complications frequently raise concerns among patients. Indeed, up to

* Corresponding author. Lisbon Higher School of Health, Polytechnic University of Lisbon, Lisbon, Portugal,
E-mail addresses: marco.caetano@essl.ipl.pt (M. Caetano), dinsi613@gmail.com (D. Vicente), leonor.ds.caetano@gmail.com (L. Caetano), etcarolino@essl.ipl.pt (E. Carolino), ana.grilo@essl.ipl.pt (A. Grilo).

https://doi.org/10.1016/j.radi.2026.103469

50% of patients with breast cancer report anxiety about radiotherapy-induced skin toxicity.[9] The focus of these concerns typically shifts during the treatment course. Before therapy, patients often enquire about potential skin changes and preparations.[10,11] During therapy, questions revolve around managing acute symptoms, such as dermatitis,[10–12] and toward the conclusion of therapy, attention turns to healing strategies and long-term care.[12]

AI chatbots, including ChatGPT 4.0, (OpenAI, November 2022), and Gemini 2.5 FLASH (Google DeepMind, December 2023), are designed to provide context-sensitive responses by processing user prompts through transformer-based architectures, thereby enabling context-aware and coherent outputs.[1,13] Because general-purpose large language models are not developed as clinical decision support systems, their outputs may occasionally contain inaccurate, incomplete, or misleading information.[14–16] Emerging evidence indicates that these systems can provide relevant and structured answers to patient queries and are increasingly being explored as tools for health information provision and patient education.[16–18] Therefore, it is imperative to evaluate AI responses in this context.

Currently, there is a dearth of published studies that have systematically evaluated the accuracy and feasibility of using AI models to provide cancer-related information, especially in the context of breast cancer radiodermatitis. This study aimed to evaluate the accuracy, clarity, and quality of the outputs generated by ChatGPT 4.0 and Gemini 2.5 FLASH and assess their suitability as AI-based educational resources for patients across the critical stages of radiotherapy treatment—before, during, and after—with a specific focus on managing breast cancer-related skin toxicity. To address this gap, the present study adopts a mixed-methods analytical framework that integrates expert clinical assessment with computational language analysis. Responses generated by ChatGPT 4.0 and Gemini 2.5 FLASH were evaluated by experienced RTTs using a structured seven-point Likert scale across multiple quality dimensions. Complementary analyses, including semantic similarity metrics, readability indices, and inter-rater reliability statistics, were conducted to enhance methodological robustness and enable triangulation of findings.

Methods

Study design and methodological framework

This study used a descriptive-analytical mixed-methods design. The responses of ChatGPT 4.0 and Gemini 2.5 FLASH were compared through expert evaluations. This methodology involves prompt development, data collection, and output analysis.

Prompt Generation

To identify common patient inquiries regarding radiotherapy-induced skin toxicity, a review of English-language literature published since 2020 was conducted, yielding 12 relevant studies via keywords as "radiotherapy," "breast cancer," "skin reactions," and "information needs." An initial list of 64 questions was compiled from the articles identified. Seventeen items unrelated to cutaneous toxicity were excluded,[13,14,19,20] leaving 47 questions. These were subsequently categorised by treatment phase, and a further 24 duplicates were removed,[13,14,19,21,22] resulting in a pool of 23 unique questions. The final selection, comprising 13 questions derived from Halkett et al.[20,22–24] and 10 from other studies,[24–26] focused on precision and clinical relevance (Fig. 1). All questions were reviewed by two of the authors to ensure validity.

All prompts and AI-generated responses were presented in English. The participants were European RTTs accustomed to using English as a standard professional communication language in clinical and research settings, particularly in international collaborations. Although formal language proficiency testing was not performed, participation required sufficient proficiency to read and critically evaluate English-language clinical information.

To emulate patient–AI interactions, all questions were reformulated using first-person phrasing, with each prompt consistently initiated by the statement, “I am a patient with breast cancer …”. These questions were reformulated by the study authors based on items derived from published literature and medical information sources addressing patient information needs. Although this approach aimed to preserve the original semantic meaning and informational intent, the reformulation was not

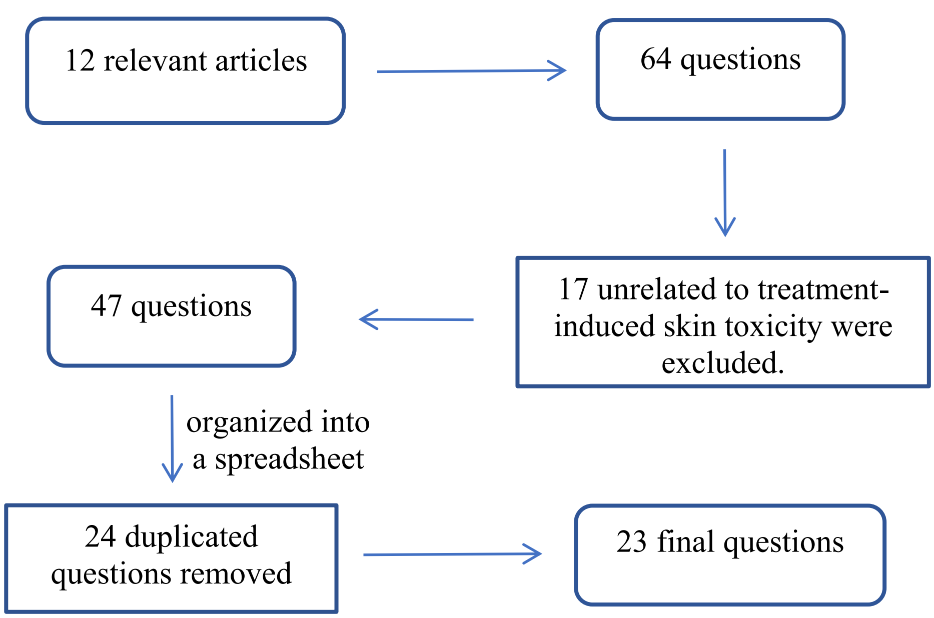

Figure 1. Flowchart illustrating the selection and refinement process of patient questions.

validated by patients or lay individuals and may have introduced subtle framing bias. This limitation reflects a broader methodological challenge in AI evaluation studies, in which clinician-generated prompts may not fully capture the variability and linguistic characteristics of real patient queries.

Each prompt was submitted individually to ChatGPT 4.0 and Gemini 2.5 FLASH under identical conditions to enable a direct comparison of the generated responses. Appendix A, Supplementary Material 1 outlines the set of common patient queries regarding radiotherapy before, during, and finishing treatment that were included in this analysis.

Data collection

The 23 prompts were submitted to ChatGPT 4.0 and Gemini 2.5 FLASH on 11 April 2025. To ensure independent responses, a new chat session was initiated for each question. The outputs were subsequently copied into a Microsoft Word document and labelled by model and prompt, with no modifications.

Outcomes

Quality and reliability

The quality of the AI-generated responses was assessed using a seven-point Likert scale[27] applied across eight dimensions: format, accuracy, consistency, organisation, clarity, precision, relevance, and audience suitability. The specific definitions of the scales are provided in Appendix A, Supplementary Material 2. The evaluation survey, developed and administered via Google Forms, was distributed using a purposive sampling approach through professional radiotherapy networks (e.g., ESTRO RTT Alliance and national RTT societies) and direct email invitations were sent to radiotherapy departments across several European countries (Belgium, Finland, Malta, Norway, Portugal, and Switzerland).

The invitation email outlined the study objectives, inclusion criteria (minimum of two years of clinical experience in a radiotherapy setting), and the voluntary and anonymous nature of participation. Interested Radiation Therapists (RTTs) accessed the evaluation survey via a secure Google Forms link. No personal or identifiable data were collected. The survey was initially deployed on May 12, 2025, with a follow-up reminder sent on May 19, 2025, to encourage participation.

By the cut-off date of 23 May 2025 16 Radiation Therapists (RTTs) had completed the assessment. All participants were qualified and professionally registered Radiation Therapists practising in accredited European radiotherapy centres, ensuring a consistent baseline level of clinical expertise. In addition, the inclusion criteria required a minimum of two years of clinical experience in a radiotherapy setting, guaranteeing that the evaluators possessed the practical knowledge necessary to critically assess the accuracy and quality of AI responses.[28] To minimise potential bias, each RTT independently evaluated the responses generated by either ChatGPT 4.0 or Gemini 2.5 FLASH, with eight RTTs assigned to each AI model. Importantly, the evaluators were blinded to the identity of the AI model being assessed.

Consistency and similarity

Cosine similarity analysis was employed to evaluate inter- and intra-model consistency and similarity. This method converts textual responses into vector representations and calculates the cosine of the angle between these vectors to estimate the degree of semantic overlap between them. The similarity scores ranged from 0 (no similarity) to 1 (complete similarity).[29] Calculations were performed using a web-based tool specifically designed for natural language similarity assessments. The analysis was conducted exclusively on English-language responses generated by the AI models and did not involve evaluator input, as it was based on vector representations of the AI-generated text. Consequently, the similarity calculations were not influenced by participants' linguistic backgrounds.

As described earlier, cosine similarity was used to assess inter-model similarity by comparing the initial responses generated by ChatGPT 4.0 and Gemini 2.5 FLASH for each question to determine the degree of alignment between their contents. Intra-model consistency was assessed by submitting the same question three times to ChatGPT 4.0 and Gemini 2.5 FLASH in different sessions. The similarity between these independently generated responses was used to assess whether the model provided a stable and consistent output over time.

Readability

The linguistic accessibility of AI-generated responses was assessed using two established readability indices: Flesch Reading Ease Score (FRES)[30] and Flesch-Kincaid Grade Level (FKGL).[31] The scores were calculated using an online readability calculator[26]. The FRES produces values between 0 and 100, with higher scores indicating easier-to-read content. In contrast, the FKGL reflects the U.S. school grade level required to understand a text, with lower values corresponding to more straightforward and accessible language. Appendix A, Supplementary Material 4, provides a summary table detailing the word count, sentence count, FRES, and FKGL scores for responses generated by ChatGPT 4.0 and Gemini 2.5 FLASH. These readability assessments provided both subjective and objective measures of suitability for patients with varying health literacy levels.

Statistical analysis

Statistical analyses were performed using IBM SPSS Statistics software (version 29). A significance level of 5% ($p < 0.05$) was used for all analyses. Descriptive statistics, including frequency analysis (n, %) for categorical variables and medians with interquartile ranges (IQR) for ordinal variables, are presented.

Data normality was assessed using the Shapiro–Wilk test. Owing to the non-normal distributions, the Mann–Whitney U test was used to compare the performance scores between the two AI models. Effect sizes were calculated to assess the magnitude of these differences.

To examine the inter-rater reliability across evaluators, Krippendorff's alpha and Fleiss' kappa were calculated to ensure robust agreement metrics for expert assessments across the eight quality dimensions.

Ethical Considerations

This study did not involve patients or identifiable patient data and was therefore exempt from institutional ethical approval. All participating RTTs provided voluntary and anonymous responses. Prior to starting the assessment, the RTTs were fully informed of the study objectives and explicitly accepted participation. Furthermore, as all data originated from publicly accessible platforms (ChatGPT 4.0 and Gemini 2.5 FLASH), no additional permissions were required for the utilisation or publication of the AI content.

Results

Quality and reliability

The RTT evaluations of ChatGPT 4.0 and Gemini 2.5 FLASH showed distinct patterns (Fig. 2). ChatGPT 4.0 received the highest ratings of 5 ($n = 73$) and 6 ($n = 74$), with some ratings of 2 ($n = 1$), 3 ($n = 6$), and 7 ($n = 20$). These level 5–6 evaluations indicate accurate, clear, and well-organized responses. Gemini 2.5 FLASH performed more consistently, with predominant ratings at level 6 ($n = 104$) and 7 ($n = 37$), showing accurate, precise, well-structured answers with patient-centred language. It received ratings of 4 ($n = 5$) and 5 ($n = 38$), but none below 4.

Significant differences were detected between the platforms for several questions (Appendix A, Supplementary material 3). During treatment, questions 3 ($P = 0.048$; effect size = 0.493), 6 ($P = 0.084$; effect size = 0.432), and 7 ($P = 0.027$; effect size = 0.552) showed significant differences in scores. Regarding the questions about finishing treatment, differences were observed for questions 1 ($P = 0.071$; effect size = 0.452), 4 ($P = 0.026$; effect size = 0.557), and 5 ($P = 0.034$; effect size = 0.530), with Gemini 2.5 FLASH receiving higher ratings. Effect sizes ranging from moderate to high indicated differing rating distributions between groups. For the remaining questions, the p-values exceeded 0.05 with low effect sizes, indicating overlapping distributions. Although not always statistically significant, Gemini 2.5 FLASH consistently received higher ratings. Several questions showed moderate or high effect sizes (≥ 0.3), suggesting practical differences between ChatGPT 4.0 and Gemini 2.5 FLASH.

The analysis of Krippendorff's alpha (α) and Fleiss' Kappa (κ) showed moderate to substantial evaluator agreement. Both models exceeded the minimum acceptability ($\alpha \geq 0.67$) for reliability for all questions. Gemini 2.5 FLASH showed higher agreement ($\alpha = 0.78$; $\kappa = 0.70$) than ChatGPT ($\alpha = 0.72$; $\kappa = 0.65$), suggesting more consistent content.

Consistency and similarity

Cosine similarity (0–1 scale) measured the similarity between ChatGPT 4.0 and Gemini 2.5 FLASH with a median of 0.78 (IQR = 0.05), indicating low similarity, particularly for Q6 at 0.74. Consistency measurements showed median similarities of 0.81 (IQR = 0.02) for ChatGPT and 0.85 (IQR = 0.03) for Gemini, with values ranging from 0.76 to 0.89.

Readability

Appendix A, Supplementary Material 4 shows the differences in the readability of the AI responses across the treatment phases. *Before* treatment, ChatGPT 4.0 produced responses with FRES values of 45–68 and FKGL scores of 6.51–10.41, indicating moderate accessibility. Gemini 2.5 FLASH had lower FRES (34–60) and higher FKGL scores (8.64–12.67), indicating more complex content, particularly for sensitive topics such as sexuality.

During treatment, ChatGPT maintained a clearer response (FRES 40–66, FKGL 6.85–11.95) and used direct formats for self-care instructions. Gemini maintained its detailed style (FRES 39–56, FKGL up to 12.87), which may pose a challenge to patients with lower health literacy.

Finishing Treatment, ChatGPT responses remained readable (FRES 45–64, FKGL 7.56–12.42), although topics such as late effects increased complexity. Gemini maintained comprehensive content (FKGL 9.39–11.99), emphasising detail over accessibility.

ChatGPT 4.0 is more suitable for patients with limited health literacy, whereas Gemini excels in providing detailed explanations for more proficient readers. Tool selection should consider both the treatment phase and the audience literacy level.

Discussion

This study evaluated the quality of responses provided by ChatGPT 4.0 and Gemini 2.5 FLASH to patient enquiries concerning radiotherapy-induced skin toxicity. The mixed-methods approach adopted in this study, integrating expert clinical evaluation with computational analyses (semantic similarity and readability metrics), enabled a multidimensional assessment of AI-generated health information. Such triangulation is particularly valuable in this context, as it captures both clinical quality and linguistic accessibility, which are essential for patient-facing applications. The objective was to assess the accuracy, clarity, and appropriateness of these responses as educational resources for use before, during, and finishing treatment. Our findings corroborate the increasing recommendation[32] that generative AI models should undergo rigorous and independent evaluation in clinical environments, with particular attention to accuracy, readability, empathy, and consistency of outcomes.

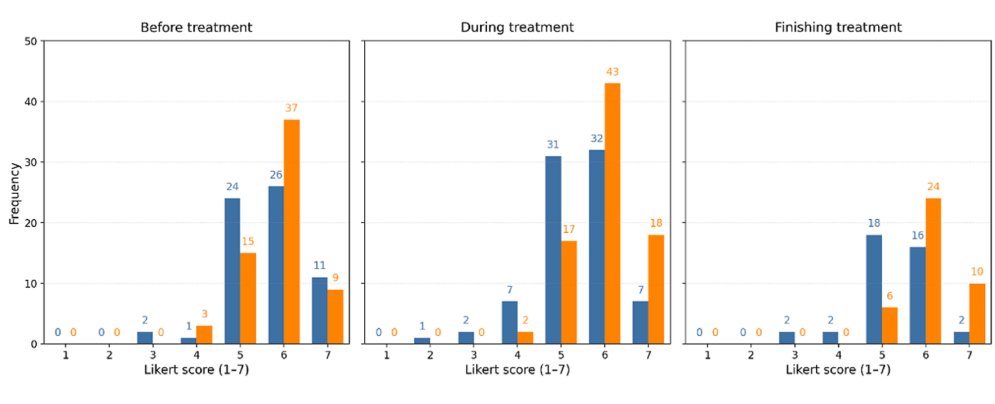

Figure 2. Number of scores assigned by radiotherapy experts to the total number of responses in each dimension from (Blue) ChatGPT 4.0 and (Orange) Gemini 2.5 FLASH. Each radiotherapy expert evaluated responses to questions collected before (8 questions), during (10 questions), and finishing (5 questions) the radiotherapy treatment. Scores were assigned using a seven-point Likert scale (1–7), where 1 = Extremely Poor, 2 = Very Poor, 3 = Poor, 4 = Fair, 5 = Good, 6 = Very Good, and 7 = Excellent. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Main findings and implications

Phased treatment analysis revealed the complementary strengths of the models. Gemini 2.5 FLASH consistently achieved higher median scores for quality and reliability (6/7) than ChatGPT 4.0 (5/7), a superiority stemming from its ability to respond with greater clinical detail, accuracy, and an empathetic, patient-centred tone. This aligns with the growing interest in AI and compassionate technologies in healthcare[33] and the findings of Sharma et al. (2024),[34] who reported particularly positive affective tones in Gemini's outputs. An empathetic tone is especially valuable in radio-oncology, where it can support treatment adherence and facilitate psychosocial adjustments.[35] Additionally, Gemini's greater consistency, reflected in higher inter-rater agreement ($\alpha = 0.78$; $\kappa = 0.70$), suggests that it is a more reliable source of complex and sensitive topics, such as those related to sexuality (Q8) or the daily burden of illness (Q5), consistent with recent analyses showing that Gemini achieved higher reliability in expert scoring.[36]

Simultaneously, the elevated empathy ratings ascribed to Gemini necessitate rigorous scrutiny. Within the field of oncology, such responses may create a semblance of empathy, prompting patients to anthropomorphise the system and ascribe authentic emotional understanding to what is, in fact, the replication of acquired linguistic patterns.[37] This dynamic was apparent in this study. Gemini often initiated responses with expressions such as “It's understandable that you're concerned …” and concluded by reinforcing support “Your sexuality is an important aspect of your overall well-being. Do not hesitate to seek help …”. While these formulations are designed to convey warmth and reassurance, they may inadvertently lead patients to attribute human-like intentions to the system itself. This misperception can foster unwarranted trust, potentially discouraging patients from seeking professional guidance and positioning AI tools as substitutes for clinician communication with patients. Such dependence poses a risk of undermining the therapeutic alliance, which is fundamental to effective oncology care,[38] and may increase emotional vulnerability by fostering a sense of support that lacks genuine human presence.[39] Collectively, these concerns underscore the necessity for research on the implications of perceived empathy in AI-mediated health communication and the formulation of safeguards to guide the application of these systems in sensitive clinical settings.

In contrast, ChatGPT-4.0 adopts a more direct and fact-focused communication style. In response to questions regarding the impact of radiotherapy on sexuality, it opens with unambiguous statements such as, "Yes, radiation therapy for breast cancer can affect your sexuality — both physically and emotionally," followed by brief reassurance and factual points. This straightforward structure, combined with its high readability (8th-grade level), makes ChatGPT particularly suitable for conveying practical guidance, including hygiene care and bra use (Q6). Nevertheless, this simplicity poses the risk of oversimplifying clinically significant nuances,[40] potentially constraining the depth necessary for addressing complex psychosocial issues. Although its readability is consistent with previous findings,[41,42] the responses generated by ChatGPT surpass the 6th-grade reading level recommended for patient education materials.[43] To address these limitations, a viable strategy involves equipping patients with fundamental prompting techniques that enable chatbots to generate text at a lower reading level (e.g., "translate to a 6th-grade reading level"). Previous research has demonstrated that this method can significantly enhance patient comprehension.[14,44,45]

These findings collectively highlight the critical need to balance perceived empathy with linguistic accessibility when incorporating large language models into oncologic communication.

Strengths and limitations

This study has several strengths, including the implementation of a treatment phase-specific framework and the evaluation conducted by a panel of 16 RTTs from six European countries, providing a geographically diverse and professionally qualified evaluator group. Our analysis was distinguished by its specific focus on skin toxicity in breast cancer radiotherapy, an area of significant clinical relevance.

Although the expert panel size was relatively small, which may limit generalisability, the study design generated a substantial number of independent evaluations across multiple response dimensions and treatment phases. In addition, inter-rater reliability was formally assessed using Krippendorff's alpha and Fleiss' kappa, supporting the consistency and robustness of the expert scoring.

Nevertheless, several limitations should be acknowledged. The time required to complete the evaluation form (approximately 20 min) may have diminished adherence, potentially affecting statistical power. The absence of direct comparisons with other AI platforms or established clinical educational resources constrains the interpretability of the findings. Furthermore, the results reflect the performance of the AI models at a single time point (April 2025), and given the rapid evolution of these technologies, future iterations may yield different outcomes.

Importantly, AI-generated responses do not account for individual patient context, including comorbidities, concurrent treatments, and other clinical factors, which may limit their applicability in complex clinical scenarios. Therefore, these outputs should be interpreted as general informational support and not as a substitute for personalised clinical advice. Additionally, the prompts were reformulated by the study authors and not validated by patients, which may limit ecological validity.

Future studies involving larger and more diverse multidisciplinary panels as well as direct patient participation would further strengthen the external validity of these findings.

Future Directions

While these findings underscore the potential of AI-mediated communication, they also highlight the ongoing challenges concerning the psychosocial impact on patients. Rigorous validation and continuous refinement of models using authoritative medical sources are imperative to maintain confidence among patients and professionals in the context of radiotherapy for breast cancer. Future research should involve larger and more diverse evaluator panels and incorporate a broader multidisciplinary group, including advanced practice RTTs, oncology nurses, and clinical oncologists, to enable a more comprehensive assessment of AI-generated responses from multiple clinical perspectives. Additionally, the development of patient profiles and inclusion of patient perspectives are essential to better align chatbot performance with varying informational needs and levels of health literacy, as well as to enhance ecological validity. Comparative analyses involving other AI systems with distinct cultural sensitivities (e.g. DeepSeek) and established clinical resources may further elucidate the specific strengths of each model. Finally, the systematic collection of patient feedback is crucial for evaluating the long-term real-world effectiveness of these tools, understanding the implications of perceived empathy in AI-mediated communication, complementing expert assessments, and guiding the safe and ethically sound integration of AI into radiotherapy practice.

Conclusion

A comparative evaluation of ChatGPT-4.0 and Gemini 2.5 FLASH revealed functional complementarity in their use as patient education tools for breast cancer radiotherapy. Gemini 2.5 FLASH demonstrated greater clinical depth, a more empathetic tone, and higher qualitative acceptance by RTTs, suggesting a particular value for addressing complex psychosocial concerns. Conversely, ChatGPT-4.0 offers superior clarity and accessibility, making it especially suitable for straightforward practical guidance.

Taken together, these findings indicate that the optimal use of AI chatbots in clinical practice lies not in selecting a single "best" model, but in deploying each system selectively according to patients' health literacy and informational needs. However, their role in supporting education on skin toxicity management must remain grounded in ongoing clinical oversight and systematic evaluation. Ultimately, their clinical utility will depend on the development of robust, regularly updated validation frameworks that safeguard reliability and safety and align with evolving oncology standards of care.

Ethics approval and consent to participate

Not applicable.

This manuscript is a comprehensive literature review that does not involve the collection of primary data from human participants or animals. Therefore, ethical approval or consent was not required.

Availability of data

Not applicable.

Author contributions

MC: Conceptualisation, Methodology; Validation, Writing - Review & Editing, Visualisation, Supervision.

LC: Conceptualisation, Methodology, Investigation, Writing - Original Draft, Visualisation.

DV: Conceptualisation, Methodology, Investigation, Writing - Original Draft, Visualisation.

AG: Conceptualisation, Methodology; Validation, Writing - Review & Editing, Visualisation, Supervision.

EC: Validation, Writing - Review & Editing, Supervision.

Generative AI use

During the preparation of this work, the author(s) used Paperpal for writing. After using this tool/service, the author(s) reviewed and edited the content as needed and took (s) full responsibility for the content of the publication.

Funding

The authors have no funding sources to declare.

Conflict of interest statement

None.

Acknowledgements

Not applicable.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radi.2026.103469.

References

1. Michalsen A. The role of complementary and alternative medicine (CAM) in rheumatology — it's time for integrative medicine. *J Rheumatol*. 2013;40: 547–549. https://doi.org/10.3899/jrheum.130107.
2. Busch F, Hoffmann L, Rueger C, Van Dijk EH, Kader R, Ortiz-Prado E, et al. Current applications and challenges in large language models for patient care: a systematic review. *Commun Med*. 2025;5. https://doi.org/10.1038/s43856-024-00717-2.
3. Ayo-Ajibola O, Davis RJ, Lin ME, Riddell J, Kravitz RL. Characterizing the adoption and experiences of users of artificial intelligence-generated health information in the United States: cross-sectional questionnaire study. *J Med Internet Res*. 2024;26:e55138. https://doi.org/10.2196/55138.
4. Yun HS, Bickmore T. Online health information-seeking in the era of large language models: cross-sectional web-based survey study. *J Med Internet Res*. 2025;27:e68560. https://doi.org/10.2196/68560.
5. Manem VS, Taghizadeh-Hesary F. Advances in personalized radiotherapy. *BMC Cancer*. 2024;24. https://doi.org/10.1186/s12885-024-12317-3.
6. Corish S, Fulton BA, Galbraith L, Coltart K, Duffton A. Impact of patient information format on the experience of cancer patients treated with radiotherapy. *Technical Innovations & Patient Support in Radiation Oncology*. 2024;30:100252. https://doi.org/10.1016/j.tipsro.2024.100252.
7. Ramseier JY, Ferreira MN, Leventhal JS. Dermatologic toxicities associated with radiation therapy in women with breast cancer. *International Journal of Women's Dermatology*. 2020;6:349–356. https://doi.org/10.1016/j.ijwd.2020.07.015.
8. Hymes SR, Strom EA, Fife C. Radiation dermatitis: clinical presentation, pathophysiology, and treatment 2006. *J Am Acad Dermatol*. 2006;54:28–46. https://doi.org/10.1016/j.jaad.2005.08.054.
9. Shaverdian N, Wang X, Hegde JV, Aledia C, Weidhaas JB, Steinberg ML, et al. The patient's perspective on breast radiotherapy: initial fears and expectations versus reality. *Cancer*. 2018;124:1673–1681. https://doi.org/10.1002/cncr.31159.
10. Andersen ER, Eilertsen G, Myklebust AM, Eriksen S. Women's experience of acute skin toxicity following radiation therapy in breast cancer. *JMDH*. 2018;11:139–148. https://doi.org/10.2147/jmdh.s155538.
11. Salvo N, Barnes E, Van Draanen J, Stacey E, Mitera G, Breen D, et al. Prophylaxis and management of acute radiation-induced skin reactions: a systematic review of the literature. *Curr Oncol*. 2010;17:94–112. https://doi.org/10.3747/co.v17i4.493.
12. Schnur JB, Ouellette SC, Dilorenzo TA, Green S, Montgomery GH. A qualitative analysis of acute skin toxicity among breast cancer radiotherapy patients. *Psychooncology*. 2011;20:260–268. https://doi.org/10.1002/pon.1734.
13. Trager MH, Queen D, Bordone LA, Geskin LJ, Samie FH. Assessing ChatGPT responses to common patient queries regarding basal cell carcinoma. *Arch Dermatol Res*. 2023;315:2979–2981. https://doi.org/10.1007/s00403-023-02705-3.
14. Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur Radiol*. 2023;34:2817–2825. https://doi.org/10.1007/s00330-023-10213-1.
15. Patel R, Dembek C, Won Y, Kadakia A, Huang X, Zeni C, et al. A real-world data analysis of electronic health records to investigate the associations of predominant negative symptoms with healthcare resource utilisation, costs and treatment patterns among patients with schizophrenia. *BMJ Open*. 2024;14: e084613. https://doi.org/10.1136/bmjopen-2024-084613.
16. Tan TF, Thirunavukarasu AJ, Campbell JP, Keane PA, Pasquale LR, Abramoff MD, et al. Generative artificial intelligence through ChatGPT and other large language models in ophthalmology: clinical applications and challenges. *Ophthalmology Science*. 2023;3:100394. https://doi.org/10.1016/j.xops.2023.100394.
17. Baglivo F, De Angelis L, Casigliani V, Arzilli G, Privitera GP, Rizzo C. Exploring the possible use of AI chatbots in public health education: feasibility study. *JMIR Med Educ*. 2023;9:e51421. https://doi.org/10.2196/51421.
18. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare*. 2023;11:887. https://doi.org/10.3390/healthcare11060887.
19. Bin Akhtar Z. From bard to gemini: an investigative exploration journey through Google's evolution in conversational AI and generative AI. *Comput Artif Intell*. 2024;2:1378. https://doi.org/10.59400/cai.v2i1.1378.
20. Halkett GKB, Kristjanson LJ, Lobb E, Little J, Shaw T, Taylor M, et al. Information needs and preferences of women as they proceed through radiotherapy

for breast cancer. *Patient Educ Counsel*. 2011;86:396–404. https://doi.org/10.1016/j.pec.2011.05.010.
21. Ahamad A, Wallner P, Salenius S, Ross R, Fernandez E. Information needs expressed during patient-oriented oncology consultations: quantity, variation, and barriers. *J Cancer Educ*. 2018;34:488–497. https://doi.org/10.1007/s13187-018-1329-5.
22. Halkett GK, Kristjanson LJ. Validity and reliability testing of two instruments to measure breast cancer patients' concerns and information needs relating to radiation therapy. *Radiat Oncol*. 2007;2. https://doi.org/10.1186/1748-717x-2-43.
23. 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*. 2009;19:538–547. https://doi.org/10.1111/j.1365-2354.2009.01090.x.
24. Smith SK, Nathan D, Taylor J, Van Gelder E, Dixon A, Halkett GKB, et al. Patients' experience of decision-making and receiving information during radiation therapy: a qualitative study. *Eur J Oncol Nurs*. 2017;30:97–106. https://doi.org/10.1016/j.ejon.2017.08.007.
25. Flockton A, Leong A, Gilfillan D, Larsen P. A multi-centre survey of New Zealand cancer patients' preferences for radiation treatment information. *J Med Radiat Sci*. 2023;71:91–99. https://doi.org/10.1002/jmrs.745.
26. Skalla KA, Bakitas M, Furstenberg CT, Ahles T, Henderson JV. Patients' need for information about cancer therapy. *Oncol Nurs Forum*. 2004;31:313–319. https://doi.org/10.1188/04.onf.313-319.
27. Kusmaryono I, Wijayanti D, Maharani HR. Number of response options, reliability, validity, and potential bias in the use of the likert scale education and social science research: a literature review. *INT J EDUC METHODOL*. 2022;8: 625–637. https://doi.org/10.12973/ijem.8.4.625.
28. Coffey M, Leech M. Introduction to the ESTRO european qualifications framework (EQF) 7 and 8: benchmarking radiation therapist (RTT) advanced education. *Technical Innovations & Patient Support in Radiation Oncology*. 2018;8:19–21. https://doi.org/10.1016/j.tipsro.2018.09.008.
29. Guo K. Testing and validating the cosine similarity measure for textual analysis. *SSRN Journal*. 2022. https://doi.org/10.2139/ssrn.4258463.
30. Flesch R. A new readability yardstick. *J Appl Psychol*. 1948;32:221–233. https://doi.org/10.1037/h0057532.
31. Kincaid JP, Fishburne J, Robert PR, Richard LC, S B. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Defense Technical Information Center; 1975. https://doi.org/10.21236/ada006655.
32. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Med Inf Decis Making*. 2025;25. https://doi.org/10.1186/s12911-025-02954-4.
33. Morrow E, Zidaru T, Ross F, Mason C, Patel KD, Ream M, et al. Artificial intelligence technologies and compassion in healthcare: a systematic scoping review. *Front Psychol*. 2023;13. https://doi.org/10.3389/fpsyg.2022.971044.
34. Sharma P, Garg N, Khanna P, Duggal S, Gondode PG. End-of-life care patient information Leaflets-A comparative evaluation of artificial intelligence-generated content for readability, sentiment, accuracy, completeness, and suitability: ChatGPT vs google gemini. *Indian J Crit Care Med : Peer-Reviewed, Official Publication of Indian Society of Critical Care Medicine*. 2024;28: 561–568. https://doi.org/10.5005/jp-journals-10071-24725.
35. Fogarty LA, Somerfield MR, Mcdonnell K, Curbow BA, Wingard JR. Can 40 seconds of compassion reduce patient anxiety? *J Clin Orthod*. 1999;17:371. https://doi.org/10.1200/jco.1999.17.1.371.
36. Leutz-Schmidt P, Palm V, Mathy RM, Grözinger M, Kauczor H-U, Jang H, et al. Performance of large language models ChatGPT and gemini on workplace management questions in radiology. *Diagnostics*. 2025;15:497. https://doi.org/10.3390/diagnostics15040497.
37. Ciechanowski L, Przegalinska A, Magnuski M, Gloor P. In the shades of the uncanny valley: an experimental study of human–chatbot interaction. *Future Gener Comput Syst*. 2018;92:539–548. https://doi.org/10.1016/j.future.2018.01.055.
38. Back A, Tulsky J, Arnold R. *Mastering communication with seriously ill patients*. Cambridge University; 2009. https://doi.org/10.1017/cbo9780511576454.
39. Taylor SE. *Social support: a review*. Oxford University; 2011:190–214. https://doi.org/10.1093/oxfordhb/9780195342819.013.0009.
40. Najdi J, Alameddine B, Armache A, Zein M, Azar WS, El-Hajj A. Assessment of the validity of ChatGPT-3.5 responses to patient-generated queries following BPH surgery. *Sci Rep*. 2025;15. https://doi.org/10.1038/s41598-025-13077-1.
41. Grilo A, Marques C, Corte-Real M, Carolino E, Caetano M. Assessing the quality and reliability of chatgpt's responses to radiotherapy-related patient queries: comparative Study with GPT-3.5 and GPT-4. *JMIR Cancer*. 2025;11:e63677. https://doi.org/10.2196/63677.
42. 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. https://doi.org/10.1001/jamanetworkopen.2024.4630.
43. Akinleye SD, Edelstein DM, Culbertson MD, Hashem J, Montuori M, Garofolo-Gonzalez G. *Readability of the Most commonly accessed online patient education materials pertaining to pathology of the hand. Hand (New York, N,Y)*. vol. 13. 2017:705–714. https://doi.org/10.1177/1558944717726138.
44. Rouhi AD, Ghanem YK, Yolchieva L, Saleh Z, Joshi H, Moccia MC, et al. Can artificial intelligence improve the readability of patient education materials on aortic stenosis? A pilot study. *Cardiol Ther*. 2024;13:137–147. https://doi.org/10.1007/s40119-023-00347-0.
45. Musheyev D, Pan A, Gross P, Kamyab D, Kaplinsky P, Spivak M, et al. Readability and information quality in cancer information from a free vs paid chatbot. *JAMA Netw Open*. 2024;7:e2422275. https://doi.org/10.1001/jamanetworkopen.2024.22275.