Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Artificial Intelligence (AI) Assisted Decision Making in Malignant Pleural Mesothelioma: A Comparative Study of AI Responses

Mert DELİKAYA
Ege University Hospital
Fatma SERT
gracilis81@yahoo.com

Ege University Hospital

Research Article

Keywords:

Posted Date: May 30th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6474443/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Objective:

This study evaluates the applicability of artificial intelligence (AI) in clinical decision-making for malignant pleural mesothelioma (MPM) by comparing treatment recommendations from large language models (LLMs) with expert decisions.

Methods:

A retrospective analysis was conducted on 12 MPM cases treated between 2021 and 2023 at a tertiary university hospital. AI-generated recommendations from ChatGPT, Gemini, and Copilot were compared with multidisciplinary tumor board decisions regarding initial treatment strategy, radiotherapy (RT) timing, target volume delineation, dosimetric assessment, and RT plan approval. AI responses were scored using a 5-point Likert scale by three radiation oncologists. Readability and quality assessments were performed using the DISCERN scale and established readability metrics. Statistical analyses included intraclass correlation coefficients, Friedman tests, and Wilcoxon signed-rank tests with Bonferroni correction.

Results:

The study analyzed 12 cases with 60 questions comparing three LLMs in mesothelioma treatment decision-making. ChatGPT demonstrated superior performance with the highest mean score (4.50 ± 0.57) and a median score of 5, significantly outperforming Gemini (mean 3.77 ± 0.43, median 4) and Copilot (mean 3.85 ± 0.52, median 4) ($p < 0.001$).

Category-specific analysis showed that ChatGPT consistently excelled across all decision-making domains, particularly in RT timing and dosimetric data evaluation (median scores of 5). It significantly outperformed the other models in four of five categories: Initial Treatment Recommendation, Radiotherapy Timing, Radiotherapy Planning, and Dosimetric Data Evaluation (all $p < 0.05$). Gemini maintained moderate performance with median scores of 4 across all categories. Copilot showed variable performance with median scores ranging from 3 to 4. In RT Plan Approval, ChatGPT and Gemini performed similarly ($p = 1.000$), while Copilot scored significantly lower ($p = 0.025$).

ChatGPT achieved the highest DISCERN score (70/75, excellent quality), while Copilot (62/75) and Gemini (61/75) were rated as good. Readability analyses classified all AI outputs as "difficult to read," with Copilot being the most readable (Flesch Reading Ease Score = 35.52).

Conclusion:

Among the evaluated AI models, ChatGPT provided the most accurate and clinically relevant recommendations for MPM management. While AI tools show promise in decision support, further

validation is required before integration into clinical workflows. Future research should focus on enhancing readability and reliability for clinical applications.

Introduction

Recent advancements in natural language processing and machine learning, exemplified by tools like ChatGPT, have triggered a paradigm shift in healthcare. Released by OpenAI in November 2022, ChatGPT gained rapid global attention, followed by the emergence of AI-powered programs like Gemini and Co-pilot(1). Trained on massive text datasets, these large language models hold transformative potential for healthcare, including oncology. In oncology, AI is increasingly applied to enhance diagnostic accuracy, personalize treatment plans, and analyze complex datasets, such as genomic profiles and radiological imaging(2, 3). However, the current literature highlights the pressing need for validation studies and real-world implementation to fully harness these technologies' capabilities(4).

Malignant pleural mesothelioma (MPM) is an aggressive malignancy of the pleura strongly associated with asbestos exposure(5). Despite advances in locoregional therapy, MPM continues to have poor oncologic outcomes, with high rates of local recurrence and metastatic spread, leading to dismal survival rates. In Turkey, MPM is a particularly prevalent disease. While its global incidence is approximately 1–2 cases per million people, an average of 500 new cases are diagnosed annually in Turkey(6).

The treatment of MPM is complex and typically involves a multidisciplinary approach. Standard therapies include surgery (Extrapleural pneumonectomy (EPP) and pleurectomy decortication (P/D)), chemotherapy, and RT, often used in combination for eligible patients (7). Chemoterapy with pemetrexed and platinum-based agents remains the cornerstone of systemic treatment. Recent advancements in immunotherapy, such as immune checkpoint inhibitors targeting PD-1/PD-L1, have shown promise in improving outcomes(8). However, MPM continues to pose significant challenges due to its aggressive nature and resistance to conventional treatments, underscoring the need for further research and novel therapeutic strategies.

Although EPP is still performed in highly selected cases, radical P/D is increasingly favored whenever feasible for patients with MPM. In MPM, determining and implementing RT after surgery RT is particularly challenging due to the disease’s anatomical complexity and the need to protect critical structures like the lungs and heart. The intricacies of tailoring RT plans make it a demanding decision that often requires expert consensus.

Recent studies have investigated the use of AI tools, such as ChatGPT, in comparing treatment recommendations with those made by multidisciplinary tumor boards in oncology(9). Building on this concept, our study focuses on MPM treatment, evaluating the alignment between AI-generated suggestions from ChatGPT, Gemini, and Co-pilot with multidisciplinary council decisions. By analyzing complex cases of MPM, we aim to explore how these AI-supported tools can assist in tailoring treatment strategies, including surgery, chemotherapy, and radiotherapy, while maintaining the expertise and

precision required for managing this challenging disease. Our aim is to assess the applicability of AI-supported decision-making algorithms in the treatment of difficult disease groups like MPM, where treatment planning is particularly complex, considering established clinical guidelines.

Material and Method

Patient selection

In our study, we investigated the data of the patients who were treated by the same thoracic surgeon, the same medical oncologist and the same radiation oncologist in a tertiary university hospital between January 2021 and December 2023 with the decision made by the thoracic oncology council. To evaluate AI responses regarding the treatment of MPM, a structured approach was developed. Data from 12 patients who underwent multidisciplinary council-based treatment and follow-up at our clinic were utilized. Patient-specific data included age, performance status, smoking and alcohol use, comorbidities, pathology, disease stage, imaging reports, and pulmonary function tests. All patients provided informed consent prior to their inclusion in the study.

Steps of Artificial Intelligence Preparation Algorithms

Before providing patient-specific details, each AI system was prompted with the instruction: “Consider yourself an expert radiation oncologist specializing in mesothelioma.” The goal was to elicit decision-making responses rather than general advice.

The evaluation encompassed five key aspects of clinical decision-making. First, each AI system was asked to provide an initial treatment recommendation tailored to the clinical scenario. Next, the timing of RT within the treatment plan was assessed. This was followed by queries regarding the definition of RT target volumes to ensure precise and consistent treatment planning. The AI systems were then tasked with evaluating the dosimetric parameters of treatment plans, focusing on their adherence to clinical guidelines and safety standards. Finally, each system was asked to approve or reject the treatment plans based on the provided clinical details and dosimetric data. The questionnaire used in this study is provided as Supplementary File.

To ensure the scope of the discussion was clear, these aspects were reiterated at the beginning of each section during the evaluation process. This setup was implemented on October 10, 2024, using web-based chat interfaces of AI models. Responses were generated using a standard browser on a single computer without activating any additional features.

To minimize bias in the AI responses, all questions were reviewed for grammatical and syntactic accuracy before submission. Given the probabilistic algorithms of large language models (LLMs) that utilize random sampling for various outputs, there was a possibility of obtaining different responses to the same question. To maintain consistency in analysis, only the first response from each LLM to each

question was recorded and considered. No exclusion criteria or restrictions were applied to the responses.

AI response evaluation and statistical analyses

The accuracy and informativeness of the responses provided by the AI systems were assessed using a 5-point Likert scale. On this scale, a score of 1 indicated very poor quality or unacceptable inaccuracies, corresponding to “strongly disagree”; 2 indicated poor accuracy with potentially harmful errors, corresponding to “disagree”; 3 reflected moderate inaccuracies that could be misinterpreted, corresponding to “neither agree nor disagree”; 4 indicated good quality with only minor, harmless inaccuracies, corresponding to “agree”; and 5 represented excellent accuracy with no errors, corresponding to “strongly agree.”

This evaluation was conducted independently by 3 radiation oncologists (S.O., F.S. and D.Y), each possessing extensive medical knowledge and clinical experience in MPM. The evaluators were blinded to the source of the responses. Inter-rater reliability was assessed by calculating intraclass correlation coefficients. Any discrepancies in the Likert scale ratings for individual responses were resolved through discussion until a consensus was reached. The final agreed-upon scores were recorded for further analysis **(Table 1)**.

The reliability and clarity of responses provided by LLM’s were evaluated using the widely accepted DISCERN scale(10). The DISCERN scale is specifically designed to provide a comprehensive evaluation of various aspects of a publication (in this study, responses from an LLM), including reliability, the quality of treatment-related information, and overall content quality. It is a three-part scoring system consisting of 16 questions, each rated on a scale from 1 to 5.

The first section assesses the reliability of a publication with eight questions, the second section evaluates treatment-related information with seven questions, and the third section appraises overall content quality with a single question. However, the final question in the third section is excluded from scoring, making the evaluation based on 15 questions. The DISCERN scoring system ranges from 16 to 75 points, classifying content quality as follows: excellent (63–75 points), good (51–62 points), moderate (39–50 points), poor (27–38 points), or very poor (16–26 points). LLMs were assessed based on criteria such as avoiding the presentation of uncertain information as definitive, discussing the reliability of information even in the absence of scientific references, and accurately describing the methods, advantages, and disadvantages of treatment options. DISCERN scores were determined by evaluating the collective responses of each LLM to all study questions rather than scoring each question individually. The final scores for each LLM were established through a consensus approach among three investigators.

The comprehensibility and complexity of responses were evaluated using three widely accepted readability measures: the Flesch Reading Ease Score, the Flesch-Kincaid Grade Level, and the Coleman-Liau Index. The Flesch Reading Ease Score assesses readability based on average sentence length (in

words) and average word length (in syllables), with scores ranging from 0 to 100. Higher scores indicate greater readability. This score has been shown to correlate strongly with other readability formulas.

The Flesch-Kincaid Grade Level is a widely used readability metric developed by the United States military and validated in prior studies. It is significantly influenced by word count and syllable count. The Coleman-Liau Index, another readability measure, assumes that word length is a more accurate predictor of readability than syllables. It calculates the average number of letters per 100 words and the average sentence length. Both the Flesch-Kincaid Grade Level and the Coleman-Liau Index assess the educational level required to comprehend a given text, with higher values indicating more complex content(11).

All statistical analyses were performed using SPSS v27 (IBM Corp., Armonk, NY, USA). Descriptive statistics, including mean, standard deviation, median, minimum, and maximum values, were calculated for each AI model (ChatGPT, Gemini, and Copilot). The Kolmogorov-Smirnov and Shapiro-Wilk tests were conducted to assess normality. As all distributions significantly deviated from normality ($p < 0.001$ for all models), non-parametric tests were applied for further comparisons.

To evaluate overall differences between the three LLM's, the Friedman test was used. Post-hoc pairwise comparisons were performed using the Wilcoxon signed-rank test, with Bonferroni correction applied to adjust for multiple comparisons. Additionally, the intraclass correlation coefficient (ICC) was calculated to assess inter-rater reliability among the LLM's. The two-way mixed-effects model was applied using consistency definition, and Cronbach's alpha was reported to determine internal consistency reliability.

All statistical tests were two-tailed, and a p-value <0.05 was considered statistically significant unless otherwise stated.

Results

A total of 12 cases and 60 questions were included in the analysis. The descriptive statistics for each LLM (ChatGPT, Gemini, and Copilot) are presented in **Table 2**. The highest mean score was observed for ChatGPT (Mean = 4.50, SD = 0.567), followed by Copilot (Mean = 3.85, SD = 0.515) and Gemini (Mean = 3.77, SD = 0.427). The median score for ChatGPT was 5, while for Gemini and Copilot, it was 4. Upon evaluating the Likert scores from three LLMs across all 60 questions, a statistically significant difference was found (Friedman test, $p < 0.001$). Post-hoc pairwise comparisons were conducted using the Wilcoxon signed-rank test, with Bonferroni correction applied to account for multiple comparisons. The results indicated that ChatGPT demonstrated significantly higher scores than Gemini ($p < 0.001$), suggesting a substantial difference in performance. Similarly, ChatGPT also outperformed Copilot, with a statistically significant difference ($p < 0.001$). However, no statistically significant difference was observed between Gemini and Copilot ($p = 0.197$), indicating that these two models performed comparably. To evaluate the performance of LLM models in different decision-making steps, separate Friedman tests were conducted for the five predefined categories (Initial Treatment Recommendation, Radiotherapy Timing, Radiotherapy Planning, Dosimetric Data Evaluation, Radiotherapy Plan Approval).

On a categorical basis, there were statistically significant differences between the responses provided by the three LLMs. ChatGPT consistently demonstrated the highest performance across all categories, significantly outperforming Gemini and Copilot in Initial Treatment Recommendation ($p = 0.001$), Radiotherapy Timing ($p < 0.001$), Radiotherapy Planning ($p < 0.001$), and Dosimetric Data Evaluation ($p < 0.001$). In these categories, ChatGPT had significantly higher scores than Gemini and Copilot ($p < 0.05$ for all comparisons), while Gemini and Copilot showed no significant difference in most cases. In the Radiotherapy Plan Approval category, ChatGPT and Gemini performed similarly ($p = 1.000$), whereas Copilot had significantly lower scores ($p = 0.025$). These findings indicate that ChatGPT consistently provided superior recommendations across various clinical decision-making stages, while Copilot and Gemini exhibited comparable but lower performance.

A comparison of the median scores of three LLMs in different question categories is given in **Table 2**. ChatGPT consistently demonstrated superior performance, achieving a median score of 5 in critical areas such as radiotherapy timing and dosimetric data evaluation, with 4-5 scores across most tasks. Gemini maintained a more moderate performance, consistently scoring a median of 4 across all categories, showing reliable but less exceptional results. Copilot exhibited variable performance, with median scores of 3-4, occasionally matching ChatGPT's performance in specific domains like radiotherapy planning. Initial treatment recommendations showed the most variability, with ChatGPT scoring 4 in 10 out of 12 cases, Gemini consistently scoring 3, and Copilot alternating between 3 and 4. Notably, in radiotherapy plan approval, all three models converged to a more uniform performance, with median scores of 4, suggesting a degree of consistency in this particular assessment area.

The reliability of the ratings across the three AI systems was assessed using intraclass correlation coefficient (ICC). The analysis revealed a moderate level of agreement with a single measures ICC of 0.560 (95% CI: 0.416-0.689) and an average measures ICC of 0.792 (95% CI: 0.681-0.869), both statistically significant ($F(59,118) = 4.813$, $p < 0.001$). These findings suggest acceptable consistency among the AI systems' performance across the evaluation categories, indicating that the observed differences between systems, particularly the superior performance of ChatGPT, reflect genuine variations in capability rather than measurement inconsistency.

ChatGPT-4 exhibited superior reliability compared to other LLMs, scoring 70 points on the DISCERN scale, indicating ‘excellent’ quality; while Copilot and Gemini received 62 and 61 (good quality) points, respectively.

All three LLMs provided responses categorized as ‘difficult to read’ according to the Flesch Reading Ease Score. Among them, Copilot showed the highest readability with a score of 35.52, followed by ChatGPT (33.78) and Gemini (33.56). In terms of complexity, as measured by the Flesch–Kincaid Grade Level, Copilot scored 13.51, ChatGPT 13.28, and Gemini 13.01. These results indicate that while all models generated text with similar readability challenges, Copilot's responses were slightly easier to read, whereas Gemini's had the lowest grade level requirement.

Discussion

Although artificial intelligence algorithms in the field of radiation oncology are generally focused on contouring and treatment planning, publications on their use in decision-making processes have increased in recent years. To our knowledge, this study is the first in the field of radiation oncology to compare the decision-making performance of LLMs. In the literature, LLMs have been compared in various specialties and topics, with ChatGPT often demonstrating superior performance(12–14). Similarly, in our study, it exhibited consistent superiority across all five decision-making categories.

The most aggressive therapy offered to patients with malignant pleural mesothelioma (MPM) is the trimodality approach, which includes chemotherapy, surgery, and radiotherapy. Despite the promising potential of immunotherapy, this rare disease presents complex decision-making challenges for many oncology specialists(15).

Since MPM is most often confined to the ipsilateral pleura, local control is a key concern. However, treating the entire pleura requires a large radiation field, which increases the risk of toxicity. The choice between extrapleural pneumonectomy (EPP) and pleurectomy/decortication (P/D), as well as the clinical setting, determines how radiotherapy (RT) is delivered. For patients with good performance status who undergo EPP, we suggest hemithoracic adjuvant RT to improve local control, while acknowledging that retrospective data do not show a clear survival benefit. For patients with good performance status who undergo P/D, hemithoracic pleural intensity-modulated radiotherapy (IMRT) is an option if appropriate expertise is available, given the technical challenges. IMRT has been evaluated in patients undergoing EPP, with early reports showing significantly increased toxicity, including fatal radiation pneumonitis. However, subsequent studies have demonstrated improvements in toxicity management. One potential disadvantage of IMRT is the radiation dose delivered to the contralateral lung, increasing the risk of pneumonitis. A higher mean lung dose and increased lung volume receiving 5, 10, or 20 Gy have been associated with a greater risk of lung toxicity(16). Current guidelines do not recommend RT after P/D if conventional techniques are used or if sufficient clinical expertise is lacking. Due to these complexities, MPM RT management varies among clinicians. In our study, the same cases were evaluated by three radiation oncologists specializing in MPM, and the agreement on LLM responses was found to be moderate (ICC = 0.560). This suggests that even in experienced centers, MPM management can involve differing approaches.

The use of LLMs in medicine is rapidly expanding, with these models being utilized not only by clinicians but also by patients(12, 17, 18). Although LLMs are widely accessible, the readability and complexity of their outputs increase in proportion to the complexity of the input data. In our study, all LLM responses were classified as "difficult to read." Nevertheless, individuals without any medical education may still gain insight into complex multidisciplinary treatment decisions for conditions like MPM. Therefore, oncologists should be aware of this potential and recognize that they may increasingly encounter patients who are well-informed about their treatment options.

ChatGPT, released on November 30, 2022, is one of the most widely used LLMs and serves as the foundational model for many other AI-based applications, including the LLM evaluated in our study. In studies on similar topics, ChatGPT has consistently outperformed other models, and in our study, it demonstrated a similarly superior performance. LLMs improve as they receive more input data, and ChatGPT's frequent use may have contributed to this result. It should be noted that the evaluations in our study were conducted as of October 10, 2024, and repeating the same analysis in the future may yield different and potentially better results.

The fact that all LLMs, especially ChatGPT, achieved high median scores in the categories of Radiotherapy Planning and Dosimetric Data Evaluation may be due to the larger and more well-defined body of literature available on these topics compared to other categories.

The category of RT plan approval, however, represents one of the greatest challenges encountered in real clinical scenarios. Although guidelines are followed for MPM RT, achieving optimal dose-volume curves remains difficult even with the best techniques and equipment. As a result, clinicians almost always must weigh the risks and benefits when approving a treatment plan. This challenge is particularly evident in patients receiving adjuvant RT after surgery, as there are no further local control options available afterward. In such cases, clinicians may accept a suboptimal RT plan while considering potential toxicities. One of the most critical aspects of our study was testing the ability of LLMs to take initiative by evaluating the risks and benefits in a manner similar to real clinicians. In this category, both ChatGPT and Gemini demonstrated accurate dosimetric evaluations and aligned closely with real clinicians' perspectives, whereas Copilot was somewhat more cautious ($p = 0.025$).

The limitations of this study include the potential influence of uncontrollable variables, such as device location, which may have introduced regional biases during response generation. Another limitation stems from the method used to assess readability and complexity indices. Although moderete intraclass correlation coefficients were observed among the three observers, the final score was determined through a consensus approach, which may have affected the objectivity of the ratings. Nevertheless, this method is widely accepted in studies involving evaluations by multiple researchers.

Possible limitations regarding the statistical analyses include the small sample size, which may have reduced the ability to detect significant differences within each category, and the use of a nonparametric test, which may have lower statistical power compared to parametric tests. Additionally, a unique limitation of this study is that MPM was chosen as the test subject for LLMs, a cancer type with limited literature and available data. This may have led to less reliable results compared to cancer types with more extensive research and clinical experience. However, by selecting MPM and its treatment approach, our aim was to assess not only the literature review capabilities of LLMs but also their initiative in approving complex radiotherapy plans. This is particularly relevant as it may assist radiation oncologists who lack experience in MPM radiotherapy in evaluating treatment decisions.

The strengths of this study include being one of the first to evaluate LLMs in radiation oncology beyond contouring and planning, specifically in treatment decision-making. Furthermore, the LLMs were

instructed to behave as radiation oncologists, and their approaches were assessed using real patient data and clinical scenarios. The differing perspectives of the three observers and the LLMs were demonstrated in our study, highlighting the heterogeneity of treatment decisions in MPM management from a novel perspective.

Conclusion

In conclusion, this evaluation of three leading large language models (ChatGPT, Gemini, and Copilot) across radiotherapy-related tasks demonstrates significant performance variations. ChatGPT consistently outperformed the other models in most categories, particularly in radiotherapy timing, planning, and dosimetric data evaluation. Gemini displayed steady but moderate performance across all categories, while Copilot showed variable results with occasional peaks matching ChatGPT's performance. These findings suggest that while all three LLMs demonstrate potential for supporting radiotherapy-related decision-making, ChatGPT currently offers the most reliable expertise across the assessed domains. Future research should explore how these models perform in real-world clinical settings and evaluate potential benefits of incorporating LLM technologies into radiotherapy workflows, with appropriate clinical oversight.

Abbreviations

AI: Artificial Intelligence; MPM: Malignant Pleural Mesothelioma; LLM: Large Language Model; RT: Radiotherapy; EPP: Extrapleural Pneumonectomy; P/D: Pleurectomy Decortication; IMRT: Intensity-Modulated Radiotherapy**.**

Declarations

Ethics approval and consent to participate: This retrospective study was conducted in accordance with the principles of the Declaration of Helsinki. Ethical approval was obtained from the Ethics Committee of Ege University Faculty of Medicine. Written informed consent was obtained from all participants prior to their inclusion in the study.

Consent for publication: Not applicable.

Availability of data and materials: The datasets generated and/or analysed during the current study are not publicly available due to patient confidentiality but are available from the corresponding author on reasonable request and with permission from the Ege University Ethics Committee**.**

Competing Interests: Not applicable.

Funding: Not applicable.

Authors' contributions:

MD and FS jointly contributed to the development of the research question and the formulation of the study hypothesis. Both authors reviewed the current literature and clinical background to define the scientific rationale of the study. Both MD and FS actively participated in designing the study methodology. They collaborated on determining inclusion and exclusion criteria, outlining data collection methods, defining clinical and radiological endpoints, and planning the statistical analysis strategy. MD and FS were both responsible for the logical interpretation of the study findings. They contributed to the analysis of results in relation to the hypothesis and participated in drafting and organizing the results and discussion sections of the manuscript. MD and FS were equally involved in executing the study, including patient follow-up, collection and verification of clinical and radiological data, and maintaining the accuracy of the study database. Both authors contributed to the preparation, writing, and revision of the final manuscript.

Acknowledgments: The authors express their sincere gratitude to Prof. Dr. Serdar Özkök and Prof. Dr. Deniz Yalman for their invaluable contributions in evaluating the AI-generated responses.

Questionnaire: The questionnaire used in our study was developed specifically for this research. It was designed by our research team based on the clinical characteristics and treatment pathways of the 12 malignant pleural mesothelioma (MPM) patients included in the study. Therefore, it is not based on any previously published questionnaire, scale, or validated tool. We have uploaded the English version of the questionnaire as a supplementary file and cited it accordingly in the main manuscript.

Corresponding author:

Fatma Sert (corresponding author)
Mail: gracilis81@yahoo.com

References

1. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259-65.
2. Huynh E, Hosny A, Guthier C, Bitterman DS, Petit SF, Haas-Kogan DA, et al. Artificial intelligence in radiation oncology. Nature Reviews Clinical Oncology. 2020;17(12):771-81.
3. Kawamura M, Kamomae T, Yanagawa M, Kamagata K, Fujita S, Ueda D, et al. Revolutionizing radiation therapy: the role of AI in clinical practice. Journal of Radiation Research. 2023;65(1):1-9.
4. Parkinson C, Matthams C, Foley K, Spezi E. Artificial intelligence in radiation oncology: A review of its current status and potential application for the radiotherapy workforce. Radiography. 2021;27:S63-S8.
5. Bibby AC, Tsim S, Kanellakis N, Ball H, Talbot DC, Blyth KG, et al. Malignant pleural mesothelioma: an update on investigation, diagnosis and treatment. European Respiratory Review.25(142):472-86.
6. Ergin DDM. 3 Soruda Mezotelyoma Turkish Society Of Thoracic Surgery [Available from: https://www.tgcd.org.tr/3-soruda-mezotelyoma-akciger-zari-kanseri/.

7. Sayan M, Eren MF, Gupta A, Ohri N, Kotek A, Babalioglu I, et al. Current treatment strategies in malignant pleural mesothelioma with a treatment algorithm. Adv Respir Med. 2019;87(5):289-97.
8. López-Castro R, Fuentes-Martín Á, Medina del Valle A, García Peña T, Soro García J, López González L, et al. Advances in Immunotherapy for Malignant Pleural Mesothelioma: From Emerging Strategies to Translational Insights. Open Respiratory Archives. 2024;6(3):100323.
9. Nardone V, Marmorino F, Germani MM, Cichowska-Cwalińska N, Menditti VS, Gallo P, et al. The Role of Artificial Intelligence on Tumor Boards: Perspectives from Surgeons, Medical Oncologists and Radiation Oncologists. Current Oncology. 2024;31(9):4984-5007.
10. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. Journal of Epidemiology and Community Health. 1999;53(2):105-11.
11. Friedman DB, Hoffman-Goetz L. A Systematic Review of Readability and Comprehension Instruments Used for Print and Web-Based Cancer Information. Health Education & Behavior. 2006;33(3):352-73.
12. Kayabaşı M, Köksaldı S, Durmaz Engin C. Evaluating the reliability of the responses of large language models to keratoconus-related questions. Clin Exp Optom. 2024:1-8.
13. Holmes J, Liu Z, Zhang L, Ding Y, Sio TT, McGee LA, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. Front Oncol. 2023;13:1219326.
14. 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. The Oncologist. 2024;29(5):407-14.
15. Davis A, Ke H, Kao S, Pavlakis N. An Update on Emerging Therapeutic Options for Malignant Pleural Mesothelioma. Lung Cancer (Auckl). 2022;13:1-12.
16. Ashton M, O'Rourke N, Currie S, Rimner A, Chalmers A. The role of radical radiotherapy in the management of malignant pleural mesothelioma: A systematic review. Radiother Oncol. 2017;125(1):1-12.
17. 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.
18. Rahimli Ocakoglu S, Coskun B. The Emerging Role of AI in Patient Education: A Comparative Analysis of the Accuracy of Large Language Models for Pelvic Organ Prolapse. Medical Principles and Practice. 2024;33(4):330-7.

Tables

Table 1. Scores of the responses provided by large language models to each question, measured on a likert scale (1, strongly disagreed; 2, disagreed; 3, neither agreed nor disagreed; 4, agreed answers; 5, strongly agreed).

	CHATGPT	GEMINI	COPILOT
INITIAL TREATMENT RECOMMANDATIONS			
Patient- No1	4	3	3
Patient- No2	4	3	3
Patient- No3	4	3	4
Patient- No4	4	3	4
Patient- No5	4	3	3
Patient- No6	4	3	3
Patient- No7	4	3	3
Patient- No8	4	3	3
Patient- No9	4	3	4
Patient- No10	4	3	4
Patient- No11	3	3	3
Patient- No12	3	3	3
TIMING OF RADIOTHERAPY			
Patient- No1	5	4	4
Patient- No2	5	4	4
Patient- No3	5	4	4
Patient- No4	5	4	4
Patient- No5	5	4	4
Patient- No6	5	4	4
Patient- No7	5	4	4
Patient- No8	5	4	4
Patient- No9	5	4	4
Patient- No10	4	3	4
Patient- No11	4	4	4
Patient- No12	4	3	4

RADIOTHERAPY PLANNING			
Patient- No1	5	4	5
Patient- No2	5	4	5
Patient- No3	5	4	4
Patient- No4	5	4	4
Patient- No5	5	4	5
Patient- No6	5	4	5
Patient- No7	5	4	4
Patient- No8	4	4	4
Patient- No9	5	4	4
Patient- No10	5	4	4
Patient- No11	5	4	4
Patient- No12	5	4	4

DOSIMETRIC EVALUATION			
Patient- No1	5	4	4
Patient- No2	5	4	4
Patient- No3	5	4	4
Patient- No4	5	4	4
Patient- No5	5	4	4
Patient- No6	5	4	4
Patient- No7	5	4	4
Patient- No8	5	4	4
Patient- No9	5	4	4
Patient- No10	5	4	4
Patient- No11	5	4	4
Patient- No12	5	4	4

APPROVAL OF RADIOTHERAPY PLAN			
Patient- No1	4	4	4
Patient- No2	4	4	4
Patient- No3	4	4	4
Patient- No4	4	4	3
Patient- No5	4	4	3
Patient- No6	4	4	4
Patient- No7	4	4	3
Patient- No8	4	4	3
Patient- No9	4	4	3
Patient- No10	4	4	4
Patient- No11	4	4	4
Patient- No12	4	4	4

Table 2. Median scores of three large language models in different question categories.

1, strongly disagreed; 2, disagreed; 3, neither agreed nor disagreed; 4, agreed answers; 5, strongly agreed. *Friedman test.

Category	ChatGPT-4 Median (Min – Max)	Gemini Median (Min – Max)	Copilot Median (Min – Max)	*p**
Initial treatment recommandations	4 (3 - 4)	3 (3 - 3)	3 (3 - 4)	<.001
Timing of radiotherapy	5 (4- 5)	4 (3 - 4)	4 (4 - 4)	<.001
Radiotherapy planning	5 (4 - 5)	4 (4 - 4)	4 (4 - 5)	<.001
Dosimetric evaluation	5 (5 - 5)	4 (4 - 4)	4 (4 - 4)	<.001
Approval of radiotherapy plan	4 (3 - 4)	4 (3 - 4)	4 (3 - 4)	<.001
Total	4 (4 - 4)	4 (4 - 4)	4 (3 - 4)	<.001

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SUPPLEMENTARY.docx