Journal Pre-proof

Exploring the capabilities of Large Language Models such as ChatGPT in radiation oncology

Fabio Dennstädt , Janna Hastings , Paul Martin Putora , Erwin Vu , Galina Fischer , Krisztian Süveg , Markus Glatzer , Elena Riggenbach , Hông-Linh Hà , Nikola Cihoric

PII: S2452-1094(23)00228-2
DOI: https://doi.org/10.1016/j.adro.2023.101400
Reference: ADRO 101400

To appear in: *Advances in Radiation Oncology*

Received date: 24 August 2023
Accepted date: 16 October 2023

Please cite this article as: Fabio Dennstädt , Janna Hastings , Paul Martin Putora , Erwin Vu , Galina Fischer , Krisztian Süveg , Markus Glatzer , Elena Riggenbach , Hông-Linh Hà , Nikola Cihoric , Exploring the capabilities of Large Language Models such as ChatGPT in radiation oncology, *Advances in Radiation Oncology* (2023), doi: https://doi.org/10.1016/j.adro.2023.101400

Exploring the capabilities of Large Language Models such as ChatGPT in radiation oncology

Fabio Dennstädt[1], Janna Hastings[2,3], Paul Martin Putora[1,4], Erwin Vu[1], Galina Fischer[1], Krisztian Süveg[1], Markus Glatzer[1], Elena Riggenbach[4], Hông-Linh Hà[4], Nikola Cihoric[4]

Affiliations

1 – Department of Radiation Oncology, Kantonsspital St. Gallen, St. Gallen, Switzerland.

2 – School of Medicine, University of St. Gallen, St. Gallen, Switzerland.

3 – Institute for Implementation Science in Health Care, University of Zurich, Zurich, Switzerland.

4 – Department of Radiation Oncology, University of Bern, Bern, Switzerland.

Author responsible for statistical analyses:

Fabio Dennstädt

Telephone number:

+49 1601723236

mailing address:

Klinik für Radioonkologie, Kantonsspital St. Gallen, Rorschacherstrasse 95, 9000 St. Gallen, SWITZERLAND

e-mail address:

fabio.dennstaedt@kssg.ch

Conflict of interest statement

Dr. Cihoric is a technical lead for the SmartOncology© project and medical advisor for Wemedoo AG, Steinhausen AG, Switzerland.

The authors declare no other conflicts of interest.

Funding statement

No funding was received for this study.

Data sharing statement

All data generated and analyzed during this study are included in this published article (and its supplementary information files).

Acknowledgements

Not applicable.

[Information about authors anonymized for Review]

ABSTRACT

Purpose: Technological progress of machine learning and natural language processing (NLP) led to the development of large language models (LLMs), capable of producing well-formed text responses and providing natural language access to knowledge. Modern conversational LLMs such as ChatGPT have shown remarkable capabilities across a variety of fields, including medicine. These models may assess even highly specialized medical knowledge within specific disciplines, such as radiation therapy. We conducted an exploratory study to examine the capabilities of ChatGPT to answer questions in radiation therapy.

Methods and Materials: A set of multiple-choice questions about clinical, physics and biology general knowledge in radiation oncology as well as a set of open-ended questions were created. These were given as prompts to the LLM ChatGPT, and the answers were collected and analyzed. For the multiple-choice questions, it was checked how many of the answers could be clearly assigned to one of the answers and the portion of correct answers was determined. For the open-ended questions, independent blinded radiation oncologists evaluated the quality of the answers regarding correctness and usefulness on a 5-point Likert scale. Furthermore, the evaluators were asked to provide suggestions for improving the quality of the answers.

Results: For 70 multiple-choice questions, ChatGPT gave valid answers in 66 cases (94.3%). In 60.61% of the valid answers, the selected answer was correct (50.0% of clinical questions, 78.6% of physics questions and 58.3% of biology questions). For 25 open-ended questions, 12 answers of ChatGPT were considered as “acceptable”, “good” or “very good” regarding both correctness and helpfulness by all six participating radiation oncologists. Overall, the answers were considered “very good” in 29.3%/28%, “good” in 28%/29.3%, “acceptable” in 19.3%/19.3%, “bad” in 9.3%/9.3% and “very bad” in 14%/14% regarding correctness/helpfulness.

Conclusions: Modern conversational LLMs such as ChatGPT can provide satisfying answers to many relevant questions in radiation therapy. As they still fall short of consistently providing correct information, it is problematic to use them for obtaining medical information. As LLMs will further improve in the future, they are expected to have an increasing impact not only on general society, but also on clinical practice, including radiation oncology.

Keywords: Natural language processing, artificial intelligence, large language models, radiation oncology, text generation, Chat-GPT

Word count: 4839

Journal Pre-proof

INTRODUCTION

Advancements in natural language processing (NLP) led to the development of language models that are able to process large amounts of textual data. These recently developed Large Language models (LLMs) have shown remarkable capabilities in the interpretation of text and in the generation of well-formed text. Some of the most powerful models used today are based on the generative pretrained transformer 3/3.5/4 (GPT-3/GPT-3.5/GPT-4) models, developed by OpenAI. ChatGPT, an LLM with 175 billion parameters based on GPT-3.5 and further training through extensive human feedback (1), has achieved impressive results in different subjects and tasks that usually require profound knowledge and extensive understanding and reasoning for humans to perform (2), (3). Since its release in November 2022, it has gained a lot of attention both publicly and scientifically, due to its good performance and wide knowledge in a variety of fields.

With the rapid technological advancements of LLMs and the newly arisen capabilities they show, it is very likely that these will have major implications for healthcare (4) – in particular as the technology continues to evolve. However, at the current time, still relatively little is known about what the new generation of LLMs may be used for in the clinical environment, not only in the context of general medical question-answering but also in clinical routine and in specialized medical fields. Artificial intelligence (AI) and NLP may be of particular interest in radiation oncology, being a very specialized technical and data-driven medical discipline that requires very domain-specific expertise beyond general medical knowledge.

To explore the current capabilities of modern conversational LLMs in radiation therapy, the International Society for Radiation Oncology Informatics (ISROI) performed a study to descriptively evaluate the capabilities of ChatGPT in answering domain-specific questions related to radiation oncology.

METHODS

Study design

The objective of the study was to explore how well ChatGPT (gpt-3.5-turbo) was able to answer relevant clinical questions as well as more basic general knowledge questions about radiation oncology. For this purpose, a two-part evaluation approach was used.

Part 1: Evaluation of multiple-choice questions about general knowledge in radiation oncology

To investigate the broader "general knowledge" of ChatGPT in radiation oncology, a test consisting of multiple-choice questions (with four answers A-D per question) was created. The questions were considered to be easily understandable, with unambiguously only one defined answer being clearly correct. The questions also contained the instruction “Provide only the correct letter (A, B, C, or D) as answer” for the LLM to select one of the answers. Questions were grouped into the three thematic groups clinical, physics, and biology. A total of 70 questions, each with four possible answers, were created upon agreement by the three study coordinators [anonymized for Review]. The questions were considered suitable (meaning clearly understandable and unambiguously answerable) by all three physicians. Thematically, 44 questions were clinical questions, 14 questions were physics questions, and 12 questions were biology questions.

The questions were posed to ChatGPT via the Web interface provided by OpenAI (5). To reduce bias due to the retention of previous questions and answers, a new chat session was started for each question. The text was entered in English language and no adaptations were made to the answers provided by ChatGPT.

Part 2: Physician-based evaluation of answers to relevant clinical questions

Radiation oncology is a complex medical field with many factors to consider and many uncertainties in clinical decision-making. As a result, many relevant clinical questions often do not just have “one clear answer”, that can definitely be identified as either “correct” or “incorrect”. Therefore, to do an evaluation for some of these more complex relevant questions, we used a physician-based evaluation approach in the second part of the study.

With the objective of evaluating the answers of ChatGPT to open-ended questions relevant for radiation therapy, a list of text-based questions/tasks (without multiple-choice answers) was created. Open-ended questions/tasks (intended to cover different aspects of radiation therapy) were created by the three study coordinators. After several adaptations and revisions regarding content and formulation, the three clinicians agreed upon a list of 25 questions/tasks, which were considered relevant for radiation therapy and adequate for the study. The questions were grouped into the topics “Patient evaluation / indication”, “Treatment planning”, “Plan evaluation”, “Treatment and side effects”, and “Others”.

Physicians from the radiation oncology departments of the [anonymized for Review] were asked to evaluate the quality of the answer to each question using an evaluation form. To reduce bias, the evaluators were not informed about the whole study design and were not told that the answers were given by an AI. The evaluating physicians just received the study documents with the request to evaluate the quality of answers to medical questions.

The two main relevant factors defining the overall quality of an answer were correctness and usefulness. While the correctness and usefulness of an answer clearly correlate, they are not necessarily identical (e.g., the question “Why should a patient with breast cancer receive adjuvant radiotherapy?” could be answered with “To treat the disease”. In such a case, the answer would be correct but not very useful). To address this, the physicians doing the evaluation were asked to separately evaluate the quality of an answer regarding correctness and regarding usefulness. A 5-point Likert scale (1–5; very bad – bad – acceptable – good – very good) was used for the evaluation.

Since there may be disagreements due to limited medical knowledge about individual circumstances, there is not always one clear answer to a given open-ended question. Therefore, the radiation oncologists were asked to do the evaluation based on generally accepted medical knowledge and not to insist on personal beliefs and opinions.

Furthermore, the physicians were asked whether adaptations to the answers should be made to improve their quality and to provide comments about how to do so. The physicians were allowed to search medical literature to check on the scientific background of a specific question.

Seven radiation oncologists at the radiation therapy departments of [anonymized for Review] were contacted for participation in the study without being told that the answers were given by an AI. Six physicians agreed to participate and returned the filled-out evaluation form. The participating radiation oncologists had a median of 6.5 years of clinical experience in radiation oncology (range 1.5–10 years).

The overall study design (Parts 1 and 2) is illustrated in Fig. 1.

Ethical considerations

No approval from an ethics committee was required for this study. A declaration of non-responsibility was issued by the local ethics committee of [anonymized for Review].

Data and statistical analysis

After the collection of the answers to the multiple-choice questions for part 1 of the study, each answer was evaluated to determine whether a clear assignment to one of the provided answers (A–D) was possible. Answers of ChatGPT that failed to select one of the four provided answers were defined as invalid. The portion of valid, correct, and incorrect answers was determined for all questions, as well as for each of the clinical, physics, and biology questions.

As for part 2 of the study, the performance of ChatGPT as rated by the physicians with the evaluation forms was examined. The values on the 5-point-Likert scale ranging from 1 ("very bad") to 5 ("very good") were used to obtain a score value for the quality of an answer regarding correctness and usefulness. The overall score value of ChatGPT for individual questions was calculated as the mean of the values given by the individual raters.

The evaluation of the individual radiation oncologists as well as interrater agreement (IRA) were determined. IRA on individual questions/tasks was determined by calculating r_{WG}, and overall agreement on all items was determined by calculating r_{WG} (J) (6). Furthermore, the intraclass correlation coefficient (ICC) was calculated with a two-way mixed model with absolute agreement (7). R_{WG}, r_{WG} (J), and ICC can have values between 0 and 1, with low values indicating a low level of agreement and values close to 1 indicating a high level of agreement. Statistical analyses were performed using SPSS 29.0.0.0 and Microsoft Excel.

The comments given by the physicians to improve the quality of the answers were examined by content analysis. Each comment was assigned to one or several of the following three categories: “comment mentioning errors or inaccuracies in the answer”, “comment recommending further details or clarification to the answer” and “comment not directly related to the quality of the answer”. The frequencies of these categories were determined.

RESULTS

Performance of ChatGPT in answering multiple-choice questions

For 66 of the 70 answers (94.3%) given by ChatGPT to the multiple-choice questions, a clear assignment to one of the four provided answers was possible. For the other four questions, the LLM did not select one of the answers but provided the information that it is an AI language model with knowledge cutoff in September 2021 and that it is unable to answer the question. As all the questions could in fact be answered with knowledge available prior to September 2021, these four answers were deemed invalid.

For 40 questions, ChatGPT selected the correct answer (57.14% of all questions, 60.61% of validly answered questions). Regarding the thematic subgroups, 22 of the 44 clinical questions (50%), 11 of the 14 physics questions (78.57%), and 7 of the 12 biology questions (58.33%) were answered correctly (Fig. 2). All the questions, together with the answers from ChatGPT, are provided in Appendix 1.

Performance of ChatGPT in answering open-ended questions as evaluated by physicians

Out of the total of 6x25=150 evaluations for the open-ended questions, the correctness of the answers given by ChatGPT was 44 times “very good” (29.3%), 42 times “good” (28%), 29 times “acceptable” (19.3%), 14 times “bad” (9.3%), and 21 times “very bad” (14%). The mean score values ranged from 1.50 to 5.00 (mean of all score values of 3.49; median of all score values of 3.67).

The correctness of 13 answers was considered “very bad” or “bad” by at least one of the evaluators, leaving 12 answers that were considered “acceptable”, “good”, or “very good” by all physicians. Four answers were concordantly considered “good” or “very good”, with one answer reaching a perfect result being concordantly rated as “very good”. Results for the correctness are presented in Fig. 3.

Slightly different but similar results were obtained for the usefulness of the answers with 42 times “very good” (28%), 44 times “good” (29.3%), 29 times “acceptable” (19.3%), 14 times (9.3%) “bad” and 21 times (14%) “very bad”. 15 answers were deemed “bad” or “very bad” by at least one of the physicians. The same four answers that were considered “good” or “very good” regarding correctness were also concordantly deemed “good” or “very good” regarding usefulness. The mean score values ranged from 1.50 to 5.00 (mean of all score values of 3.48; median of all score values of 3.5). Individual results for usefulness are presented in the appendix (Fig. A1).

Comments for improving the quality of answers

For 24 questions (96%), at least one of the radiation oncologists provided a comment to improve the quality of the answer. Overall, comments were provided in 75 of 150 cases. 40 comments addressed some kind of error or inaccuracy of an answer. 41 comments recommended adding further details or clarification to an answer to improve its quality. Four comments did not directly relate to the quality of the answer. A comment addressing an error/inaccuracy was made in 16 answers (64%), while a recommendation to add further details was given in 20 answers (80%) by at least one of the radiation oncologists. The results of the classification of the comments on individual questions or tasks are presented in the appendix (Table A1).

Evaluations of individual physicians and interrater agreement

The evaluations of the six physicians varied (Fig. 4). An answer was considered "very bad" or "bad" regarding correctness and/or usefulness in 0, 4, 7, 8, 8, and 11 cases (median 7.5) by the individual physicians.

IRA on the evaluation of individual questions/tasks was determined by the calculation of r_{WG}, which yielded results ranging from 0.00 to 1.00. r_{WG} values had a median of 0.6 (correctness) and 0.45 (usefulness). IRA was low, with r_{WG} values of 0.00 for 3 of the 25 questions regarding both correctness and usefulness (Questions 11, 16, and 23).

Overall agreement was determined by calculation of $r_{WG}(J)$, which had a value of 0.959 for correctness and 0.956 for usefulness. Results for the individual questions/tasks are presented in the appendix (Table A2). The additionally calculated ICC had a value of 0.892 (95%CI 0.810–0.946) for correctness and 0.886 (95%CI 0.799–0.943) for usefulness.

Answers of ChatGPT to individual questions

In the following, some individual answers given by ChatGPT to the open-ended questions are presented. The total list of open-ended questions and answers is provided in Appendix 2.

Examples of "acceptable/good" answers

One exemplary question from the group "Patient evaluation / indication" is Question 3: "What are the indications for adjuvant axillary radiotherapy in a patient with breast cancer?". The answer given by ChatGPT was:

"Adjuvant axillary radiotherapy is typically indicated in breast cancer patients with positive axillary lymph nodes, as a means to reduce the risk of local regional recurrence. This treatment is usually offered in conjunction with other adjuvant therapies such as systemic therapy and surgery. The decision to offer adjuvant axillary radiotherapy is based on several factors including the number of positive axillary lymph nodes, the size of the primary tumor, and other patient-specific factors."

Regarding correctness, the answer was evaluated as "very good" by one physician, as "good" by two physicians, and as "acceptable" by three physicians. It yielded a score of 3.67 regarding correctness and of 3.50 regarding usefulness (Fig. 3 and Fig. A1).

As mentioned, for one question, ChatGPT yielded a perfect result, with all physicians considering it "very good" (score of 5.0 for both correctness and usefulness). The corresponding question is Question 21: "A patient with limited disease small-cell lung cancer had chemotherapy with concurrent radiotherapy for 6 weeks. After 4 weeks of the treatment, he is beginning to develop symptoms of pain during swallowing (VAS 6/10), which he describes as burning. He also suffers from impaired swallowing, which is why he reduced eating. What is the most likely reason for the symptoms and how could you help the patient?".

Answer: "*The most likely reason for the patient's symptoms of pain during swallowing and impaired swallowing is radiation-induced esophagitis, which is a common side effect of concurrent chemoradiotherapy for lung cancer. The symptoms can be due to inflammation and injury of the esophageal mucosa caused by the radiation.*

To help the patient, the following measures can be taken:

1. *Pain management: Over-the-counter pain relievers or prescription pain medication can be used to manage the pain.*

2. *Diet modification: Soft, bland foods and frequent small meals can help alleviate symptoms and improve swallowing. Avoiding spicy and acidic foods can also help.*
3. *Mouth care: Good oral hygiene, such as regular mouth rinses and gargles, can help to soothe the throat and prevent infections.*
4. *Lubricating agents: Using agents such as honey, glycerin, or saliva substitutes can help to lubricate the throat and improve swallowing.*

It is important to consult a medical professional for proper evaluation and management of the patient's symptoms, as the above measures may not be sufficient to alleviate symptoms in all cases. Additionally, if the symptoms persist or worsen, further investigations such as endoscopy may be necessary to rule out other causes."

Example of a "bad" answer

While these two examples show that some of the answers of ChatGPT are of acceptable or good quality, the answers to some other questions were of poor quality.

The two answers with the worst results were about fractionation calculations using the linear-quadratic (LQ) model. One example is Question 17: "What is 20x3Gy in EQD2 based on the linear-quadratic model using an α/β=2?".

Answer: "*The equivalent dose in 2 Gy fractions (EQD2) can be calculated using the linear-quadratic model and an α/β of 2 as follows:*

*EQD2 = 20 x 3Gy * (2 + (2/α/β)) = 20 x 3Gy * (2 + (2/2)) = 20 x 3Gy * 4 = 240 Gy*

So, the EQD2 for 20 fractions of 3 Gy each with an α/β of 2 is 240 Gy."

In this case, the answer of ChatGPT can clearly be defined as incorrect (the used formula by ChatGPT is incorrect and the correct answer would be 75 Gy (see also (8)). The answer was considered as “very bad” for both correctness and usefulness by 5 of the 6 radiation oncologists.

DISCUSSION

We aimed to explore the capabilities of ChatGPT in answering questions in the context of radiation therapy. We showed that some of the answers given by the model may be of good or very good quality.

ChatGPT has been tested across a wide range of subjects and has been shown to achieve success in e.g., business management (9) and law school exams (10). When applied within medicine, ChatGPT showed remarkable results in medical question answering and performed comparable to the level of a third-year medical student (11). Furthermore, ChatGPT performed near the level of the passing threshold on the United States Medical Licensing Exam in a study by Kung et al. (12). In another study by Ayers et al., evaluators preferred the responses of ChatGPT over physician responses to patient questions from a social media forum in 78.6% of cases (13).

Our findings show that ChatGPT may also provide some helpful and correct answers in radiation therapy, with an anticipated success rate of about 50–70% correct answers in a multiple-choice test like the one used in our study. For the more complex treatment-related questions, under the evaluation criterion that the correctness of given answers should be deemed “acceptable” or better by all clinicians in our physician-based evaluation, the model would have fulfilled this requirement in 12 out of 25 questions (48%).

With the continuing progress in the field of LLMs and the fine-tuning of models or application of other optimization techniques, the performance of future LLMs is likely to be considerably improved.

To avoid false and possibly harmful answers, models may also be adapted to behave in a more cautious way, like giving medical answers only if well-established medical knowledge exists.

To overcome such problems, current research focuses on combining models with explicit knowledge bases (14).

LLMs in medicine

While models such as ChatGPT can provide some correct and useful answers in radiation therapy, they are in principle rather general models without special optimization for the medical domain (15). Other models have been developed specifically for application in medicine. One of the most powerful models is Med-PaLM, developed by Google (16). As reported by the researchers involved in the development of Med-PaLM, it can provide helpful answers, often near the level of clinicians. The newer generation, Med-PaLM 2, was shown to answer medical exam questions at an “expert doctor level”. It reached an accuracy of 85% on US medical licensing style questions, outperforming its predecessor by 18% (17). An important thing to keep in mind is the rapid pace of progress (the results of Med-PaLM were published in December 2022; the announcement of the results of Med-PALM 2 was just 4 months later, in March 2023). Since Med-PaLM is not available to the public, we were not able to use it in our study.

The recent advancements of LLMs offer immense possibilities for application in medicine. With the advancements in AI continuing, it is likely that new AI technologies will profoundly change healthcare (18), (19). LLMs offer language capabilities, a key feature for processing data based on domain-specific knowledge, which will be essential for future applications in medicine. The multimodal capabilities of the newest models, e.g., those combining images and text, unlock an even wider set of possibilities for processing medically relevant data in sophisticated ways.

LLMs also have the potential to assist physicians in their daily clinical lives. For example, LLMs could be very helpful in administrative work. In a pilot study by Ali et al., ChatGPT wrote patient clinic letters with high scores regarding factual correctness and humanness as evaluated by physicians (20).

Further use of LLMs may include medical education (12), research (21) or application in clinical-decision support systems (22).

For now, it remains unclear how the recent advancements will impact general society and medicine. LLMs have begun to be used for medical advice with unknown consequences (4). With the fast progress in the field, models such as Chat-GPT, MedPaLM, or GPT-4 are just the predecessors of models that may be much more powerful and may considerably impact clinical practice (23).

LLMs in radiation therapy

NLP and LLMs may be of particular interest in radiation therapy (24). In general oncology as well as radiation oncology, physicians are faced with complex medical situations with many individual factors. Profound medical knowledge, which frequently changes due to new therapeutic options and new findings from clinical trials, is essential to make adequate decisions. With a lot of uncertainty and limited knowledge in individual oncological situations, AI-based support of clinical decision-making is of high interest (25). Furthermore, radiation oncology is in part a very technical and data-driven discipline, characterized by a high level of data processing (26). Radiation oncology information systems (ROCIS) are broadly used to manage data about patient treatment schedules, treatment plans, treatment delivery, and documentation (27), (28). Many steps involved in radiation therapy can be assessed and supported using IT and AI systems. The application of ROCIS facilitates direct workflow integration of such systems in clinical care. This allows implementation of AI solutions not only for circumscribed tasks within radiation therapy, but more general in multidisciplinary oncological situations (like e.g., application of AI systems in the multidisciplinary treatment of prostate cancer (29)). LLMs may play a key role in the future design of comprehensive oncological data systems. If the models can be coupled appropriately with medical evidence, LLMs might indeed be highly valuable for radiation therapy (30). Our study shows that modern LLMs have the potential to provide useful answers not only regarding general subjects, but also in highly specialized topics of radiation therapy.

Problems and drawbacks

Despite the impressive capabilities of the new LLMs, it has been repeatedly shown that they have considerable limitations, so their output needs to be interpreted with great caution (31).

One of the major issues with using advanced conversational models as sources of medical advice is that they may “hallucinate”, meaning that an LLM may generate text with illusory statements not based on correct data (32). An answer given by an LLM consists of a sequence of words which is the result of statistical calculations. Which sequence of words is created depends on its probability, as determined during the training of the model. Sequences of text that occur more commonly in training data are assigned higher probabilities during the foundational training phase of the model, and sequences of text that are formulated in a suitable way within a dialogue context are assigned higher probabilities during the instruction training phase of the model. How well a model can answer domain-specific questions therefore depends on the design, training data and size of the model. However, the model is not directly coupled to evidence but rather represents a synthesis of its training data with generative capability. Therefore, there is no constraint preventing it from generating incorrect statements that appear as if they were evidence-based. In oncology and medicine in general, this obviously presents a considerable problem regarding the safety and application of such a model. As mentioned, the combination of language models with explicit knowledge bases is a promising future direction to enable overcoming a part of this problem (14).

Furthermore, one should be aware that LLMs are not equally powerful in different tasks and partly still have limited capacities. As an example, in our study, ChatGPT failed to consistently answer questions requiring fractionation calculation. It has been shown that LLMs have limited performance when solving arithmetic reasoning and calculation tasks (33). Unlike natural language understanding, calculations typically have a single correct answer, making the task of generating accurate solutions more challenging. Moreover, they require specific abstraction and reasoning skills that are not well supported by the architecture and training of language models.

Another problem arises from the fact that every model is dependent on the data it was trained on. This can lead to wrong and biased results, as LLMs may adopt unwanted features such as gender or ethnic biases (34) (35).

Despite some good results, ChatGPT failed to consistently provide correct and good answers for many of the questions in our study. Since the consequences of wrong advice can be severe in medicine, the quality bar for clinical application of such technologies is very high, which is why LLMs in their current form should not be used directly for clinical decision-making, although they may provide supplementary language-related functionality in larger decision-making applications. While LLMs will improve and will likely play an important role in future healthcare, they will likely always have limitations that users should be aware of. In any case, LLMs cannot and should not be used to replace human doctors but to assist them in their work (4).

Usage of LLMs by patients

In the current state, it is not advisable to use LLMs for seeking medical advice. However, models such as ChatGPT have gained a lot of attention in the last months and are easily accessible. Furthermore, GPT-4 has been introduced into the Bing Web Search of Microsoft (36) and both Microsoft and Google have announced plans to further implement the new models into their software products. It is thus very likely that radiation oncologists and other clinicians will soon have consultations with patients who have previously consulted an LLM such as ChatGPT before attending the appointment with their treating physician. Clinicians should therefore be aware of the capabilities and limitations of these new technologies. While the support of radiation oncologists in daily clinical life by LLMs may not yet be a reality, LLMs will already have an impact on patients seeking information about their oncological situation.

Limitations of the study

Our study has several limitations. In general, the evaluation of LLMs in medicine is challenging and currently a subject of open discussion (16). While the performance of LLMs such as Med-PaLM is assessed using benchmarks like medical question-answering datasets, this approach fails to encompass all relevant factors needed in daily clinical life. Furthermore, despite the vast amount of medical literature available, the best advice for an individual patient's situation is not always known. Many relevant questions in radiation oncology do not have one defined correct answer, but an answer may be of higher or lower quality. In our study, we used a set of multiple-choice questions about basic knowledge as well as a physician-based evaluation to assess the quality of answers given by ChatGPT. However, the physician-based evaluation is prone to the personal beliefs and subjective factors of the clinicians and may fail to obtain an objective assessment. As we have also seen by comparing the evaluations of the different participating physicians, the interrater agreement for some questions/tasks was quite poor. Even though we saw an overall high level of agreement, a consensus on individual answers may not always be reached. This limits the possibility of assessing the quality of a given answer in some situations.

Furthermore, our study used a limited number of 70 multiple-choice questions and 25 open-ended questions. While the questions were created with the intention of covering different facets of radiation therapy, our study does not provide a comprehensive or systematic evaluation of LLMs in radiation therapy. Overall, the study can only be of descriptive nature, and the results do not allow further generalization.

In future work, testing the performance of LLMs in a more systematic way would ideally encompass a larger set of questions/tasks, evaluated by many physicians comparing different models and prompting techniques, as well as comparing it to the performance of clinicians and medical trainees. Furthermore, it should be noted that the development of benchmarks to evaluate the performance of LLMs is additionally challenging due to the lack of transparency about the training data used in model development. Ideally, models should be evaluated on their performance on questions they have not seen in their training. Very complex and effortful systematic studies will be necessary to evaluate the role of LLMs in the clinical practice of future healthcare. This is beyond the scope of the current study, which was initiated by the ISROI to initially assess the capabilities of these new technologies in radiation oncology.

CONCLUSIONS

We have shown that ChatGPT can provide correct and useful answers to some questions that are relevant in radiation therapy. Since such models are currently not reliable and may lead to inaccurate or wrong answers, their output should be taken with caution. Nevertheless, clinicians should be aware of the capabilities and problems of LLMs, as patients may use them to seek medical advice. As the technology continues to evolve rapidly, LLMs are anticipated to have a major impact on the practice and future of medicine and radiation oncology.

List of abbreviations

AI – artificial intelligence

GPT – generative pre-trained transformer

ICC – intraclass correlation coefficient

IRA – interrater agreement

ISROI – International Society for Radiation Oncology Informatics

NLP – Natural Language Processing

LLM – Large Language Model

LQ model – linear-quadratic model

ROCIS – radiation oncology information system

1. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback [Internet]. arXiv; 2022 [cited 2023 Apr 15]. Available from: http://arxiv.org/abs/2203.02155

2. Bang Y, Cahyawijaya S, Lee N, Dai W, Su D, Wilie B, et al. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. 2023 [cited 2023 Apr 15]; Available from: https://arxiv.org/abs/2302.04023

3. ChatGPT [Internet]. Available from: https://openai.com/blog/chatgpt

4. Will ChatGPT transform healthcare? Nat Med. 2023 Mar;29(3):505–6.

5. ChatGPT Web Interface [Internet]. Available from: chat.openai.com

6. James LR, Demaree RG, Wolf G. Estimating within-group interrater reliability with and without response bias. J Appl Psychol. 1984 Feb;69(1):85–98.

7. Hallgren KA. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutor Quant Methods Psychol. 2012 Feb 1;8(1):23–34.

8. McMahon SJ. The linear quadratic model: usage, interpretation and challenges. Phys Med Biol. 2018 Dec 19;64(1):01TR01.

9. Terwiesch C. Would Chat GPT Get a Wharton MBA? New White Paper By Christian Terwiesch.

10. Choi JH, Hickman KE, Monahan A, Schwarcz DB. ChatGPT Goes to Law School. SSRN Electron J [Internet]. 2023 [cited 2023 Apr 15]; Available from: https://www.ssrn.com/abstract=4335905

11. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023 Feb 8;9:e45312.

12. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. Dagan A, editor. PLOS Digit Health. 2023 Feb 9;2(2):e0000198.

13. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med. 2023 Jun 1;183(6):589.

14. Colon-Hernandez P, Havasi C, Alonso J, Huggins M, Breazeal C. Combining pre-trained language models and structured knowledge. 2021 [cited 2023 Aug 20]; Available from: https://arxiv.org/abs/2101.12294

15. Kocoń J, Cichecki I, Kaszyca O, Kochanek M, Szydło D, Baran J, et al. ChatGPT: Jack of all trades, master of none. 2023 [cited 2023 Apr 15]; Available from: https://arxiv.org/abs/2302.10724

16. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large Language Models Encode Clinical Knowledge. 2022 [cited 2023 Apr 15]; Available from: https://arxiv.org/abs/2212.13138

17. Google: Our latest health AI research updates [Internet]. Available from: https://blog.google/technology/health/ai-llm-medpalm-research-thecheckup/

18. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022 Jan;28(1):31–8.

19. Aung YYM, Wong DCS, Ting DSW. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. Br Med Bull. 2021 Sep 10;139(1):4–15.

20. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health. 2023 Apr;5(4):e179–81.

21. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare. 2023 Mar 19;11(6):887.

22. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Assessing the Value of ChatGPT for Clinical Decision Support Optimization [Internet]. Health Informatics; 2023 Feb [cited 2023 Apr 15]. Available from: http://medrxiv.org/lookup/doi/10.1101/2023.02.21.23286254

23. Lee P, Goldberg C, Kohane I. The AI revolution in medicine: GPT-4 and beyond. 1st ed. Hoboken: Pearson; 2023.

24. Bitterman DS, Miller TA, Mak RH, Savova GK. Clinical Natural Language Processing for Radiation Oncology: A Review and Practical Primer. Int J Radiat Oncol. 2021 Jul;110(3):641–55.

25. Lambin P, Zindler J, Vanneste BGL, De Voorde LV, Eekers D, Compter I, et al. Decision support systems for personalized and participative radiation oncology. Adv Drug Deliv Rev. 2017 Jan;109:131–53.

26. ASTRO: The Future of Oncology Data is Happening Now [Internet]. Available from: https://www.astro.org/Blog/June-2021/The-Future-of-Oncology-Data-is-Happening-Now

27. Zhang B, Chen S, D'Souza WD, Yi B. A systematic quality assurance framework for the upgrade of radiation oncology information systems. Phys Med. 2020 Jan;69:28–35.

28. Fong de los Santos LE, Herman MG. Radiation oncology information systems and clinical practice compatibility: Workflow evaluation and comprehensive assessment. Pract Radiat Oncol. 2012 Oct;2(4):e155–64.

29. Baydoun A, Jia AY, Zaorsky NG, Kashani R, Rao S, Shoag JE, et al. Artificial intelligence applications in prostate cancer. Prostate Cancer Prostatic Dis [Internet]. 2023 Jun 9 [cited 2023 Sep 30]; Available from: https://www.nature.com/articles/s41391-023-00684-0

30. Ebrahimi B, Howard A, Carlson DJ, Al-Hallaq H. ChatGPT: Can a Natural Language Processing Tool be Trusted for Radiation Oncology Use? Int J Radiat Oncol. 2023 Apr;S0360301623003541.

31. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus [Internet]. 2023 Feb 19 [cited 2023 Apr 15]; Available from: https://www.cureus.com/articles/138667-artificial-hallucinations-in-chatgpt-implications-in-scientific-writing

32. Guerreiro NM, Alves D, Waldendorf J, Haddow B, Birch A, Colombo P, et al. Hallucinations in Large Multilingual Translation Models. 2023 [cited 2023 Apr 15]; Available from: https://arxiv.org/abs/2303.16104

33. Imani S, Du L, Shrivastava H. MathPrompter: Mathematical Reasoning using Large Language Models. 2023 [cited 2023 Apr 15]; Available from: https://arxiv.org/abs/2303.05398

34. Digutsch J, Kosinski M. Overlap in meaning is a stronger predictor of semantic activation in GPT-3 than in humans. Sci Rep. 2023 Mar 28;13(1):5035.

35. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners. 2020 [cited 2023 Apr 15]; Available from: https://arxiv.org/abs/2005.14165

36. Microsoft Bing Blogs: Confirmed: the new Bing runs on OpenAI’s GPT-4 [Internet]. Available from: https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAI%E2%80%99s-GPT-4

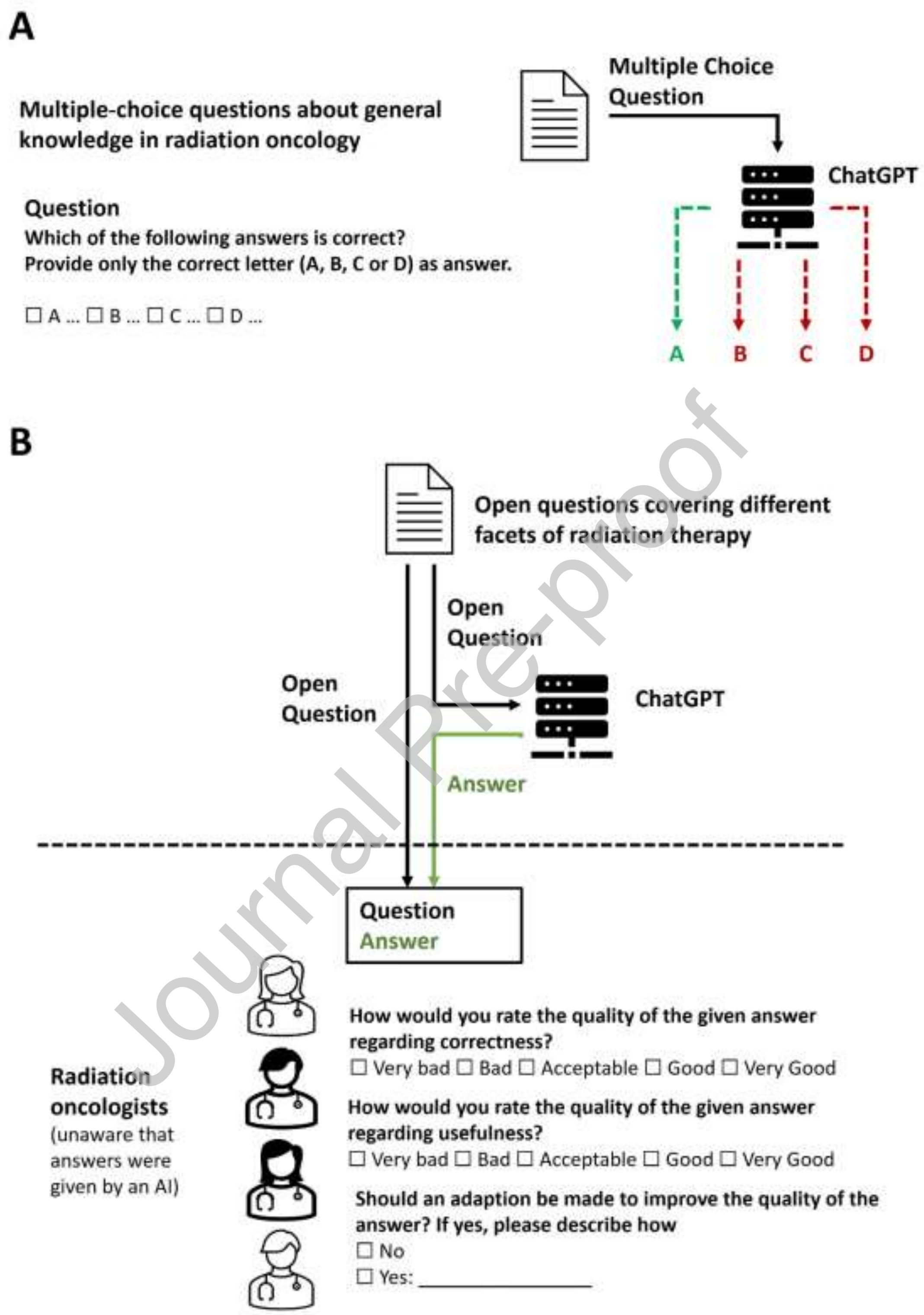

Fig. 1: Schematic illustration of the study design. A – Multiple-choice questions with one correct answer choice were answered by ChatGPT. The portion of valid and of correct answers was determined. B – Open-ended questions/tasks of radiation therapy were answered by ChatGPT. The answers were then evaluated by

independent radiation oncologists. To avoid a possible bias, the physicians were not informed that the answers were given by an AI.

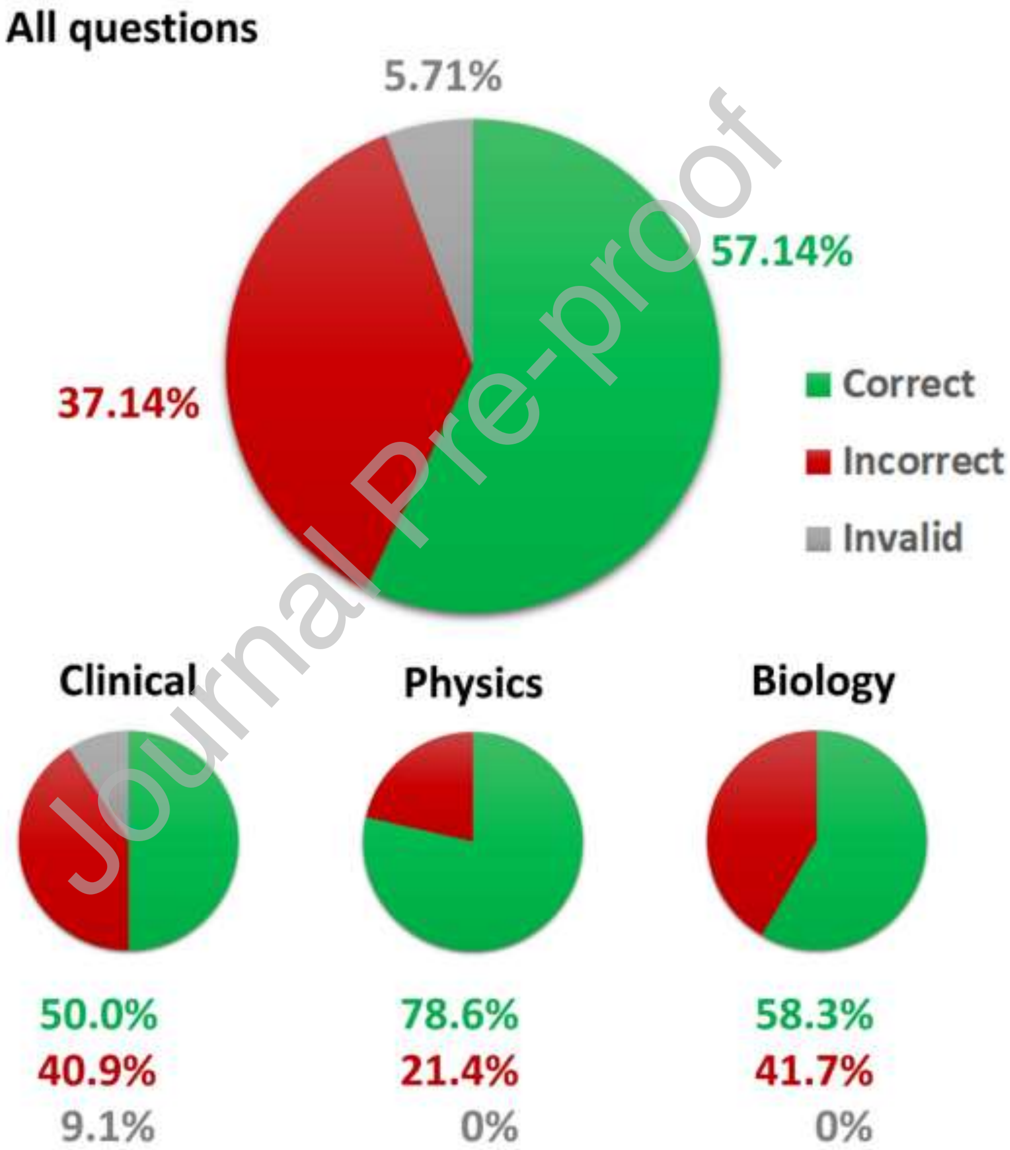

Fig. 2: Portion of correct, incorrect and invalid answers of ChatGPT to the multiple-choice questions.

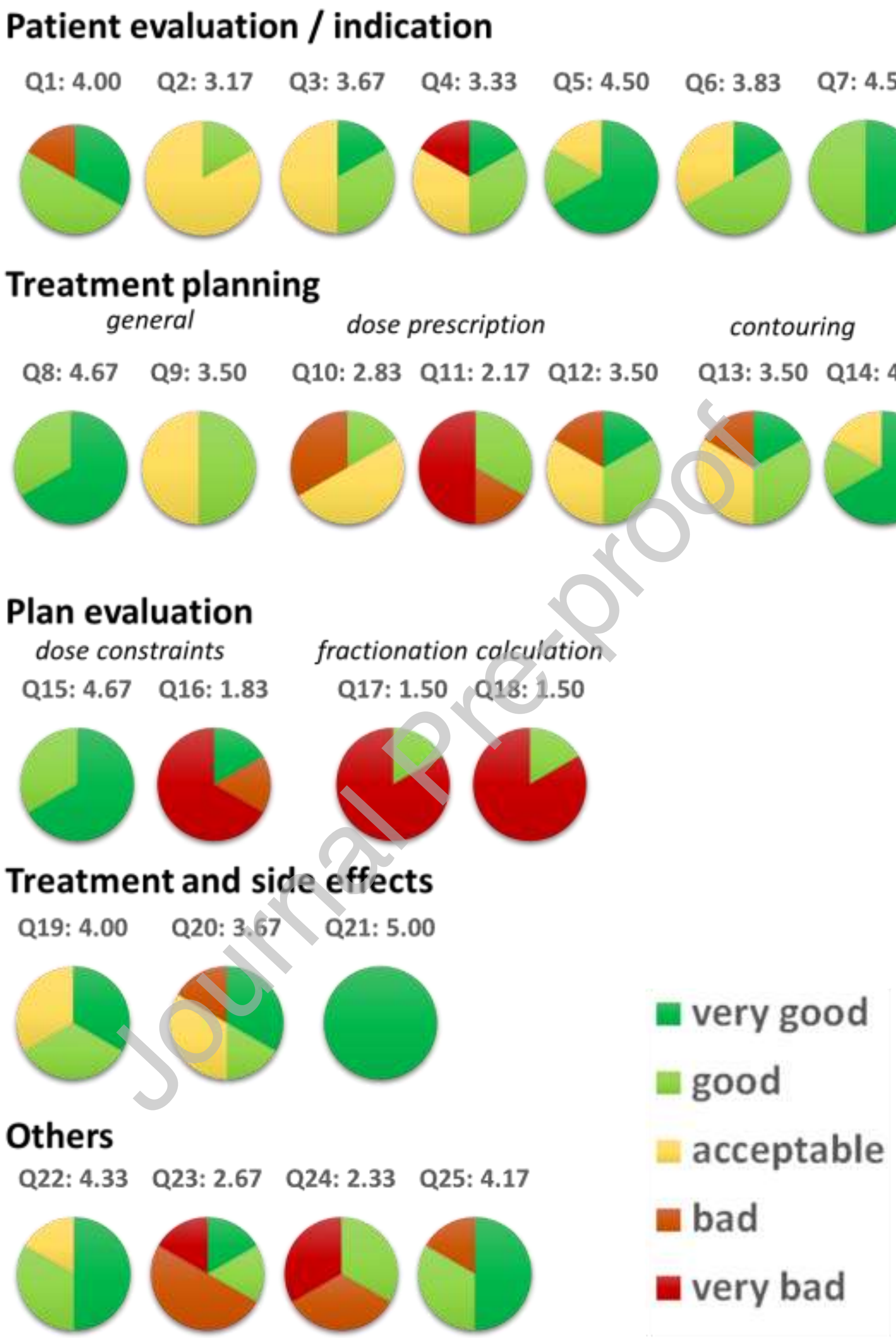

Fig. 3: Evaluation results regarding the correctness of the answers given by ChatGPT. Score values of each answer are calculated as mean of individual score values given by the radiation oncologists.

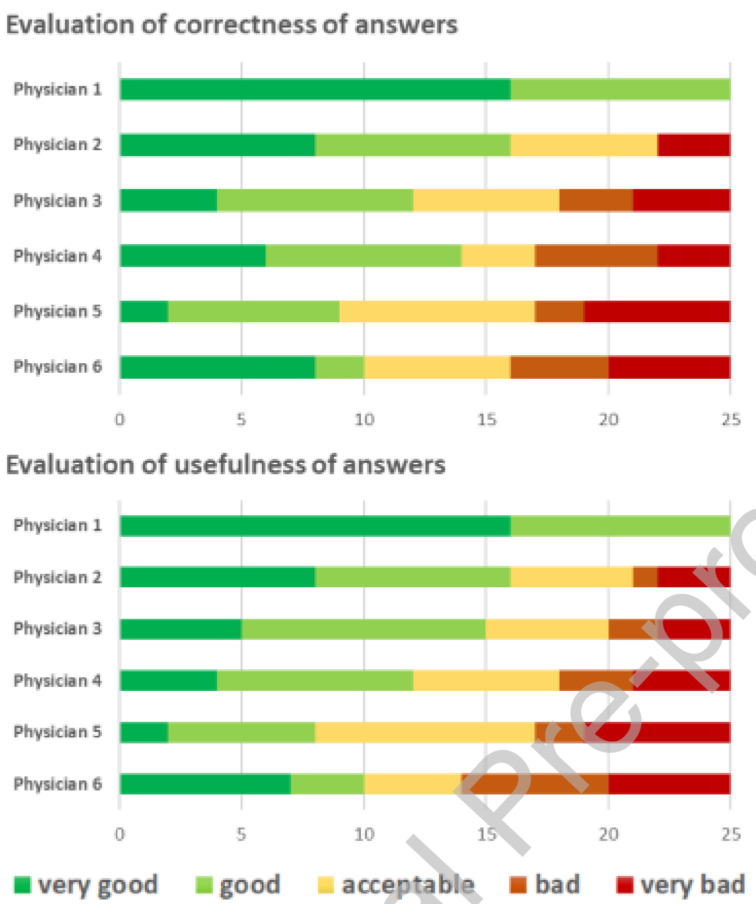

Fig. 4 Evaluation of correctness and usefulness by individual radiation oncologists.

Declaration of interests

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

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

Dr. Nikola Cihoric reports a relationship with Wemedoo AG that includes: consulting or advisory.