Evaluating the Performance of ChatGPT at Breast Tumor Board

Y. Xu[1,2], N. Logie, T. Phan[1,2], L. Barbera[1,2], R. A. Nordal[1,2], J. M. Stosky[1,2], S. L. Lee[1,2]
[1]Division of Radiation Oncology, Tom Baker Cancer Centre, Calgary, AB, Canada
[2]Division of Radiation Oncology, University of Calgary, Calgary, AB, Canada

INTRODUCTION

- Chat Generative Pre-trained Transformer (ChatGPT) is a chatbot built on the Generative Pre-trained Transformer (GPT) large language model.
- ChatGPT has performed admirably answering patient questions and writing medical board examinations[1-3], but its ability to provide insight to questions posed by practicing radiation oncologists has not been evaluated.

AIM

- To determine whether ChatGPT can contribute to tumor board discussions by comparing the accuracy and clarity of its responses to physician questions with that of human specialists.

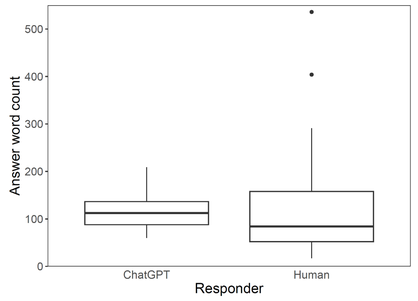
METHOD

- Twenty consecutive breast radiation oncology questions between January and February 2023 that received at least one human answer were curated from *theMedNet*, a physician-only Q&A platform for expert answers to real-world clinical situations.
- These questions were posed to ChatGPT, and its answers were paired with the first chronological human response.
- Board-certified breast radiation oncologists at an academic institution (reviewers) were asked to rate from 1 (strongly disagree) to 5 (strongly agree) the extent to which they agreed with each answer (accuracy score) and whether they felt the response provided clear and specific guidance relevant to the original question (clarity score).
- The DTS test with 10,000,000 bootstrap iterations was used to compare the distribution of answer lengths between ChatGPT and human responders.
- Responses were clustered by question to account for correlation in reviewer ratings for answers to the same question.
- Cluster bootstrapping with a normal approximation over 10,000,000 iterations was performed to compute confidence intervals for the proportion of answers on which reviewers agreed or strongly agreed with ChatGPT and human responders.
- Wilson score intervals with continuity correction were used to estimate the proportion of answers on which ChatGPT receives a higher median accuracy or clarity score than human responders.
- The paired Wilcoxon signed-rank test was used to compare median accuracy and clarity scores across all 20 questions.
- Generalized estimating equations were used to examine the association between the length of ChatGPT's answers and the odds that reviewers will agree with its response (accuracy score 4-5) or find its answer to be clear and specific (clarity score 4-5).

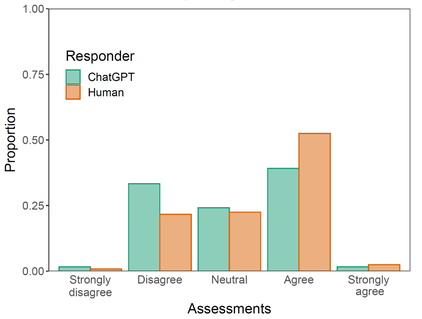
RESULTS

- Six board-certified breast radiation oncologists evaluated answers to the 20 questions, resulting in 120 distinct assessments of each of ChatGPT and human responders.
- Topics encompassed by questions:
  - Decision to offer radiotherapy and dose fractionation for early breast cancer (40%)
  - Recurrent breast cancer and/or breast reirradiation (25%)
  - Target volume selection for locally advanced breast cancer (10%)
  - Radiotherapy for oligometastatic breast cancer (5%)
  - Timing of radiotherapy with respect to implant reconstruction (5%)
  - Treatment of medically inoperable breast cancer (5%)
  - Impact of pembrolizumab-induced pneumonitis on radiotherapy decisions (5%)
  - Treatment toxicity and decision-making for patients with lichen sclerosus (5%)
- ChatGPT had a median answer length of 112.5 words (IQR, 87.75-136.5), compared to 84.0 words (IQR, 52.0-157.75) for human answers. The distribution of answer word counts differed between ChatGPT and human responders (P=0.047).
- The reviewers agreed or strongly agreed with ChatGPT responses on 49 (41%; 95% CI, 28-54) of assessments and human responders on 66 (55%; 95% CI, 43-67) of assessments.
- ChatGPT achieved a higher median accuracy score than human responders on 7 (35%; 95% CI, 16-59) questions, whereas humans outperformed ChatGPT on 8 (40%; 95% CI, 20-64) questions; there was no significant difference in median scores (P=0.29).
- There was agreement or strong agreement that ChatGPT provided clear and specific guidance on 38 (32%; 95% CI, 22-42) of assessments compared to 45 (38%; 95% CI, 27-48) assessments of human answers.
- ChatGPT had a higher median clarity score on 7 (35%; 95% CI, 16-59) questions, whereas human responders had a higher median clarity score on 9 questions (45%; 95% CI, 24-68). No differences were detected in median clarity score across all questions (P=0.75).
- On 3 questions (15%; 95% CI, 4-39), ChatGPT surpassed human responders on both median accuracy score and median clarity score.
- Human responders similarly outperformed ChatGPT in both metrics on 3 (15%; 95% CI, 4-39) questions.
- The word count of ChatGPT's answers was not associated with reviewers agreeing with its response (OR per word, 0.99; 95% CI, 0.98-1.00; P=0.18) or findings its answer to be clear and specific (OR per word, 1.01; 95% CI, 1.00-1.02; P=0.06).

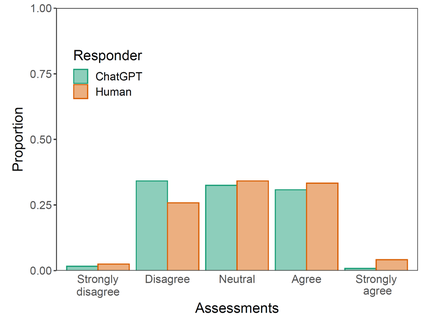
A. Answer length by word count



B. To what extent do you agree with the answer?



C. Does the answer provide clear and specific guidance?



D. Accuracy and clarity of answers

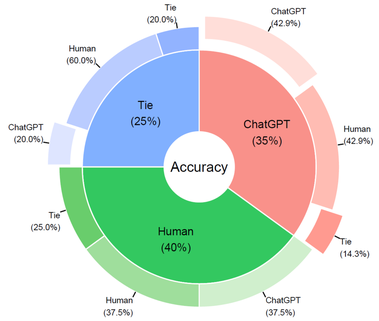


Figure D compares the accuracy and clarity of ChatGPT and human answers to the 20 questions. The inner annulus indicates the proportion of answers by each respondent receiving a higher median accuracy score, and the outer annulus indicates the proportion of answers receiving a higher median clarity score.

CONCLUSIONS

- There was no detectable difference in the accuracy or clarity of answers provided by ChatGPT and human responders in this sample of 20 challenging breast radiation oncology questions.
- ChatGPT outperformed human responders in the accuracy and clarity of its answers to some questions, suggesting that it has the potential to contribute meaningfully to discussions about real-world clinical problems.

REFERENCES

[1]Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine*. https://doi.org/10.1001/jamainternmed.2023.1838
[2]Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS digital health*, *2*(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198
[3]Bhayana, R., Krishna, S., & Bleakney, R. R. (2023). Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. *Radiology*, *307*(5), e230582. https://doi.org/10.1148/radiol.230582

CONTACT INFORMATION

Please reach out to yang.xu@albertahealthservices.ca with inquiries.