Assessing the Use of ChatGPT for Appropriate Imaging Decision-Making in Breast Imaging Using ACR Appropriateness Criteria

Samuel Escobar[1]; Justin Rhee[1]; Hossam A. Zaki[1]; Elizabeth H. Dibble, MD[2]; Robert C. Ward, MD[2]; Ana P. Lourenco MD[2]; Grayson Baird, PhD[2,3]

[1]The Warren Alpert Medical School, Brown University, Providence, RI; [2]Department of Diagnostic Imaging, The Warren Alpert Medical School, Brown University, and Brown University Health, Providence, RI; [3]Brown Radiology Human Factors Lab, Providence RI

INTRODUCTION

The rapid rise in diagnostic imaging has highlighted a growing need for tools that assist in selecting appropriate studies. Despite the development of evidence-based guidelines such as the American College of Radiology (ACR) Appropriateness Criteria (AC), the integration of these tools into clinical practice remains limited. Large language models (LLMs), like ChatGPT, have shown early promise in supporting clinical decision-making across various specialties. In this study, we investigated the ability of ChatGPT to assess imaging appropriateness in breast imaging scenarios using the ACR AC, and we compared its performance to that of experienced breast imaging radiologists.

OBJECTIVES

Aim 1: Assess the performance of ChatGPT in selecting appropriate imaging modalities in breast imaging scenarios using the American college of Radiology (ACR) Appropriateness Criteria (AC).

Aim 2: Compare the agreement of ChatGPT with the ACR AC to that of breast imaging radiologists at a single institution in selecting appropriate imaging modalities.

METHODS

We selected ten clinical variants from the breast imaging section of the ACR Appropriateness Criteria, which yielded a total of 81 imaging decisions. Each scenario was evaluated by ChatGPT-3.5, ChatGPT-4, and ChatGPT-4o using standardized prompts to rate imaging modalities on a 1–9 appropriateness scale. Each model was queried four times per variant. Four attending breast radiologists independently completed the same ratings. Ratings from both LLMs and radiologists were compared to the ACR standard using generalized estimating equations (GEEs) and Bland-Altman plots to assess accuracy, bias, and consistency.

RESULTS

Radiologists demonstrated the highest agreement with the ACR Appropriateness Criteria, showing the lowest mean bias (0.24, p = 0.489) and the smallest slope bias. All ChatGPT models significantly overestimated appropriateness, with larger mean biases (ChatGPT-4o: 2.46, ChatGPT-4: 1.76, ChatGPT-3.5: 2.47; all p < 0.001). Additionally, ChatGPT responses varied notably across repeated queries, with intra-model variability comparable to or greater than that of the radiologists.

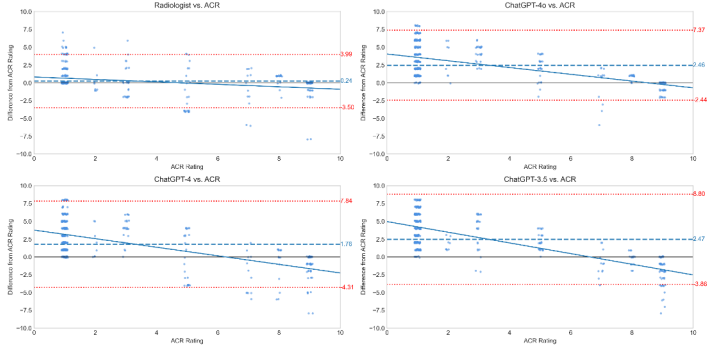


Figure 1. Deviations of radiologist (upper left), ChatGPT-4o (upper right), ChatGPT-4 (bottom left), ChatGPT-3.5 (bottom right) from ACR appropriateness ratings at each ACR appropriateness rating value. Mean deviation is shown in blue while the 95% confidence interval on the deviations is visualized in red.

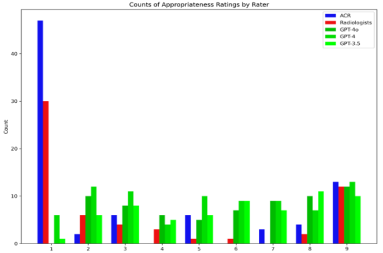


Figure 2. Total counts of appropriateness by each model of ChatGPT (green), radiologists (red), and the ACR (blue) on a scale of 1-9.

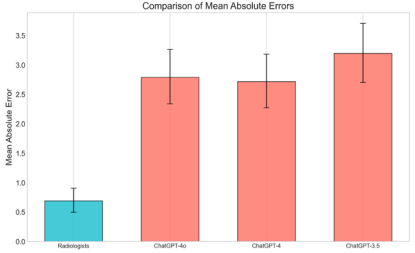


Figure 3. Mean absolute errors between radiologists and each model of ChatGPT with standard error bars.

DISCUSSION

While ChatGPT demonstrated some capacity to rate imaging appropriateness, its performance was inconsistent and often misaligned with established guidelines. The LLMs showed a tendency to overrate appropriateness and lacked reproducibility, which raises concerns about their use in real-world clinical decision-making. Notably, the variability in ChatGPT responses, despite identical prompts, underscores challenges in reliability and trust. Human oversight remains essential, especially considering the potential medicolegal implications of relying on AI-generated recommendations.

CONCLUSION

Although ChatGPT shows potential as a clinical decision support tool, current models do not yet match the accuracy, consistency, or guideline adherence of expert radiologists. Until reproducibility and alignment with clinical standards improve, LLMs should be used with caution and with expert supervision.

REFERENCES

1. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89-94. doi:10.1038/s41586-019-1799-6
2. Ritchie AJ, Sanghera C, Jacobs C, et al. Computer Vision Tool and Technician as First Reader of Lung Cancer Screening CT Scans. J Thorac Oncol. 2016;11(5):709-717. doi:10.1016/j.jtho.2016.01.021
3. Hassan AE, Ringheanu VM, Rabah RR, Preston L, Tekle WG, Qureshi AI. Early experience utilizing artificial intelligence shows significant reduction in transfer times and length of stay in a hub and spoke model. Interv Neuroradiol. 2020;26(5):615-622. doi:10.1177/1591019920953055
4. Briganti G. How ChatGPT works: a mini review. Eur Arch Otorhinolaryngol. 2024;281(3):1565-1569. doi:10.1007/s00405-023-08337-7
5. Kurth DA, Karmazyn BK, Waldrip CA, Chatfield M, Lockhart ME. ACR Appropriateness Criteria® Methodology. J Am Coll Radiol. 2021;18(11):S240-S250. doi:10.1016/j.jacr.2021.03.021