

Performance of ChatGPT in Radiology: A Systematic Review

Pedram Keshavarz, Sara Bagherieh, Seyed Ali Nabipoorashrafi, Hamid Chalian, Amir Ali Rahsepar, Grace Hyun J Kim, Cameron Hassani, Steven S. Raman, Arash Bedayat

Department of Radiological Sciences, David Geffen School of Medicine at UCLA, USA

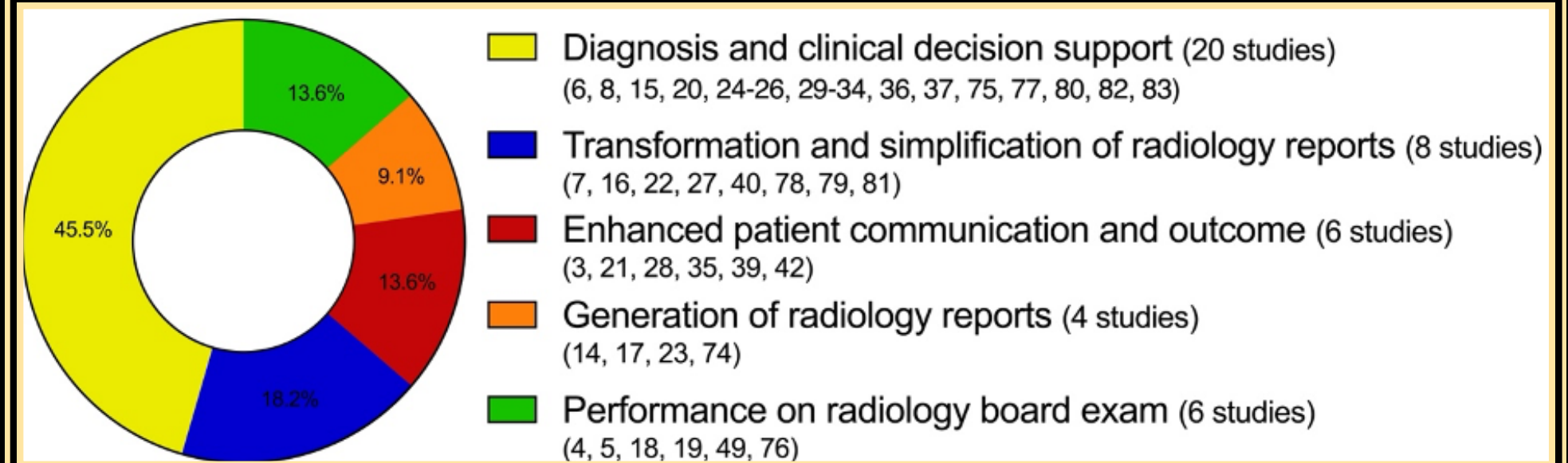
Background

Several studies have been published to determine ChatGPT's applicability in various radiology-related domains. The use of ChatGPT has elicited a debate about the advantages and disadvantages of AI technologies in routine clinical practice including concerns about potential biases in the ChatGPT training datasets constraining its use.

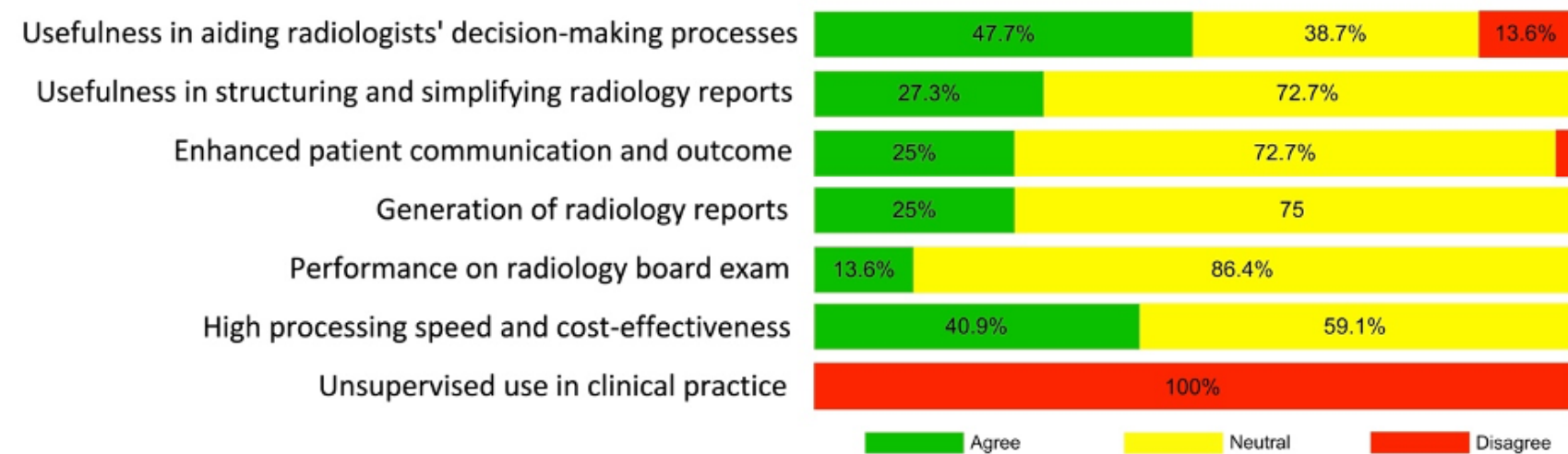
Materials and Methods

After a comprehensive review of PubMed, Web of Science, Embase, and Google Scholar, we derived a cohort of published studies up to January 1, 2024, utilizing ChatGPT for clinical radiology applications. **44 studies** covered the potential utilization of ChatGPT in **five** clinical areas of radiology.

ChatGPT's Topics in Five Different Clinical Areas of Radiology



Review Criteria of Included Studies



Results

Among 44 studies evaluating ChatGPT, 37 (84.1%) demonstrated high performance. Of the 24 studies reporting performance metrics, 19 (79.2%) measured a mean accuracy of 70.5%, and 5 (20.8%) reported an 83.6% agreement with reference standards. Low performance was identified in 7 studies (15.9%) in areas of diagnosis and clinical decision support (13.6%) and patient communication (2.3%). ChatGPTv4 outperformed v3.5 in 10 of 11 studies (90.9%), showing advancements in complex reasoning and medical terminology accuracy.

Conclusion

Although ChatGPT's effectiveness has been shown in 84.1% of radiology studies, there are still multiple pitfalls and limitations to address. It's too soon to confirm its complete proficiency and accuracy, and more extensive and comprehensive studies utilizing diverse datasets and pre-training techniques are required to verify ChatGPT's role in radiology.