HealthManagement
Promoting Management and Leadership

LLMs’ BI-RADS Categorisations Can Negatively Affect Patients

The widespread availability of large language models (LLMs) has sparked considerable interest in their potential applications within healthcare, prompting investigations into their efficacy across various clinical tasks. In radiology, LLMs have been tested for tasks ranging from processing radiologic request forms to providing imaging recommendations and differential diagnoses. While publicly available LLMs have shown promise in simpler tasks like transforming free-text reports to structured ones, challenges arise in more complex tasks requiring nuanced reasoning and deeper clinical knowledge, particularly when operating in languages other than English. Concerns over interreader agreement in assigning BI-RADS categories in breast imaging reports have led to intensive evaluation of natural language processing tools' performance in this area. Studies have demonstrated that with extensive training on large datasets, these tools can accurately extract BI-RADS features and predict pathologic outcomes. However, research on the agreement between human readers and generically trained LLMs in assigning BI-RADS categories across different languages and the impact of discordant assignments on clinical management is lacking. A recent study published in Radiology aims to assess such agreement and its implications for clinical practice.

Comprehensive Evaluation of Breast Imaging Reports

The study analysed breast imaging reports from women who underwent MRI, mammography, and/or ultrasound across three hospitals. These reports were written in English, Italian, or Dutch, reflecting the linguistic diversity of the patient population. To ensure balanced representation, reports were categorised across the BI-RADS spectrum and different imaging modalities. Each centre curated four sets of reports based on imaging modality: MRI, mammography alone, ultrasound alone, and mammography combined with ultrasound. Reports categorised as BI-RADS 0 or BI-RADS 6 were excluded due to potential biases in interpretation. The inclusion criteria required complete descriptions of imaging findings according to BI-RADS descriptors, impressions, and BI-RADS assignment by the original radiologists.

Assessing Agreement in BI-RADS Category Assignments

The study included reports from two centres retrieved between May 2020 and October 2023, while reports from the third centre were randomly sampled from a dataset collected between January 2000 and December 2020. In total, 2400 reports were analysed, with each language group comprising 800 reports. The agreement between the original and reviewing radiologists in assigning BI-RADS categories was found to be nearly perfect across all reports, regardless of language, imaging modality, or clinical management category. However, discrepancies were observed when comparing the performance of LLMs (specifically GPT-4, GPT-3.5, and Bard) with human review. LLMs exhibited higher rates of assigning different BI-RADS categories compared to human review, which could potentially lead to detrimental changes in clinical management. Conversely, LLMs had lower rates of positive changes in clinical management compared to human review. These findings underscore the importance of careful consideration when integrating LLMs into clinical practice, particularly in tasks where accurate categorization is critical for patient management.

Limitations and Regulatory Challenges

A total of 2400 breast imaging reports were analysed, with each language group comprising 800 reports. The study found nearly perfect agreement among human readers, as reflected by a Gwet AC1 coefficient of 0.91. However, the agreement between human readers and LLMs, including GPT-4, GPT-3.5, and Bard, was moderate, with Gwet AC1 coefficients ranging from 0.42 to 0.52. The frequency of discordant BI-RADS category assignments leading to negative changes in clinical management varied significantly between human-human and human-LLM comparisons. For instance, disagreements were observed to be more prevalent between human readers and LLMs compared to human-human agreements.

Bridging the Gap Between Large Language Models and Clinical Practice

The study emphasised the current limitations of LLMs in handling complex medical reasoning tasks, such as those encountered in breast imaging interpretation. This aligns with previous research findings in liver and lung imaging tasks. Despite the availability of LLM chatbots like ChatGPT and Bard, their lack of transparency and bias control raises concerns regarding their suitability for healthcare-related tasks. Moreover, the study underscored the need for regulatory oversight of publicly available LLMs, especially considering their potential use by both patients and healthcare professionals for diagnostic purposes. Calls for specific approval of LLMs as medical devices have been made, but regulatory frameworks are yet to be established. Limitations of the study included the absence of access to images and clinical data for LLMs, which could have enhanced their performance. Additionally, the evaluation of LLM assignment repeatability was conducted at a single time point, and the study focused on only three languages, potentially limiting generalizability to other linguistic contexts.

In conclusion, while LLMs demonstrated moderate agreement with human-assigned BI-RADS categories, the study highlighted significant discrepancies that could impact clinical management decisions. This underscores the imperative for regulatory oversight and the development of context-trained LLM extensions to enhance their reliability and utility in medical settings.

Source: Radiology

Image Credit: iStock

Published on : Thu, 9 May 2024