HealthManagement
Promoting Management and Leadership

Improving Mammography Report Clarity with AI

Large language models (LLMs) such as ChatGPT and Google Gemini are increasingly used to improve patient communication by simplifying complex medical content. In breast imaging, radiology reports are often filled with specialised terminology that can confuse patients, particularly when accessed online before physician consultation. Clearer language can reduce anxiety, support better understanding and encourage more informed decision-making. To be effective, AI-generated translations must remain accurate, easy to follow and emotionally considerate. Comparing different versions reveals how well they convey diagnostic information, follow-up instructions and empathy.

Design and Methodology
Three fictional radiology reports covering BI-RADS categories 3, 4 and 5 were created, each featuring typical clinical terminology and diagnoses. These were processed using three different LLMs—ChatGPT-4, ChatGPT-4o and Google Gemini Advanced 1.5—each instructed to simplify the text into language understandable by non-experts. To reduce model bias, new accounts were used, disabling chat history and training features. Each translation was generated three times per model using a consistent prompt designed to prioritise clarity, factual correctness and avoidance of jargon. Two breast imaging specialists evaluated all 27 translated outputs and selected the most accurate and complete version for each combination of model and BI-RADS category. Reports with hallucinated content were excluded.

Participants were recruited from a regional breast cancer centre, radiology and gynaecology departments. Eligibility required fluency in German. Before distribution, an experienced radiologist provided an overview of the reports and questionnaire. Each participant received one original report and three anonymised AI translations for each BI-RADS category. Reports were rated using a five-point Likert scale across four dimensions: comprehensibility, clarity of recommendations, empathy and additional value. Participants also ranked the reports by overall preference.

Must Read:** **Patient Trust in AI for Mammogram Screening

Demographic data collected included age group, prior mammography, education level and AI experience. Questionnaire data with more than one missing or invalid response were excluded. In total, 40 valid responses were analysed. A cumulative link mixed model was used to evaluate Likert-scale responses, accounting for AI system, question type, BI-RADS category, age, education and prior experience. Random effects controlled for individual rating tendencies. Ranking data were assessed using the Plackett-Luce model.

Key Findings
The participants showed a consistent preference for AI-translated reports across all four evaluation dimensions. Reports from ChatGPT-4 and ChatGPT-4o were rated significantly higher than those from Google Gemini in all categories, including perceived empathy, clarity of follow-up recommendations and added value. In the category of procedural explanation, the highest scores were recorded, while empathy received the lowest. ChatGPT-4o had the highest probability of being preferred (48%), followed by ChatGPT-4 (37%), with Gemini trailing at 15%. Participants aged 50 years or older tended to rate AI reports more positively, particularly those in age groups 3 (50–69) and 4 (70+), compared to participants aged 18–29. Ratings were also higher among individuals without previous mammography experience. In contrast, participants with tertiary education gave lower ratings than those with primary or secondary education. Previous experience with AI did not significantly influence the ratings.

The vast majority of participants supported the further development of AI-generated translations in radiology. Most selected either “yes, definitely” or “rather yes” when asked whether these tools should continue to be explored for patient communication. Only one participant remained undecided, and none were opposed.

Implications and Context
Findings align with prior research showing that LLMs can effectively reduce reading complexity without compromising factual integrity. Unlike previous evaluations conducted solely by medical professionals, this work incorporated direct patient feedback. The emphasis on empathy, clarity of recommendations and added value underscores the importance of aligning AI outputs with patient expectations and emotional needs. Differences in perception by age and education indicate that patient-specific factors should inform the design and implementation of AI tools. Older individuals may benefit more from simplified language, while those with higher education may expect more technical detail or nuance.

Using fictional reports eliminated risks related to data protection and ensured consistency across evaluations. However, evaluating fictional rather than personal reports may have reduced emotional engagement, possibly affecting perceptions of relevance or empathy. The selection of BI-RADS 3, 4 and 5 reports was deliberate, as these categories represent diagnostic uncertainty or malignancy risk, making clarity and emotional sensitivity particularly important. Reports with BI-RADS 1 or 2 may yield different results given their lower diagnostic complexity.

Limitations included a relatively small sample size and language restriction to German speakers. Only 40 questionnaires met quality standards from 76 distributed. Participants were primarily recruited from clinical settings, which may introduce selection bias. The use of general-purpose LLMs rather than domain-specific models may limit the relevance of these findings to more specialised clinical environments. Despite these constraints, consistency across evaluations suggests that AI-translated reports offer tangible benefits for patient communication.

AI-generated translations of mammography and sonography reports improve comprehensibility, perceived empathy and follow-up clarity. ChatGPT-4 and ChatGPT-4o were preferred over Google Gemini across all evaluation criteria. Preferences varied based on age, education and prior experience, highlighting the need for adaptive communication strategies. These findings support the integration of LLMs into clinical practice, with expert oversight to ensure accuracy, emotional appropriateness and patient safety. Broader studies across diverse populations and languages are needed to confirm these results and inform future development of domain-specific tools.

Source: **Academic Radiology**

Image Credit: **iStock**

Published on : Tue, 22 Jul 2025