Kylan Dwayne Steele
L2
LSU Health Sciences Center, New Orleans, LA

Rachel Fiore PhD CCC-SLP[1]
[1]Standardized Patient Project, Louisiana State University Health Sciences Center, New Orleans, LA 70112

"Checking AI against Human Experts in Evaluating the Communication of Simulated Doctor-Patient Interactions"

Background: Effective doctor-patient communication plays a critical role in promoting treatment adherence, patient understanding, and improved clinical outcomes. As a result, the evaluation of communication skills has become a prominent topic of discussion in medical education. Traditional assessments, such as those used in Objective Structured Clinical Examinations (OSCEs), rely on human raters using itemized checklists but are resource-intensive and subject to variability. Recent advances in artificial intelligence (AI) offer potential for standardized, scalable communication assessment; however, the ability of AI to accurately evaluate more nuanced aspects of patient interaction, particularly those involving non-verbal cues or affective tone remains unclear.

Objective: To examine the alignment, discrepancies, and consistency between the evaluation methods of artificial intelligence and human experts when assessing clinical communication performance.

Methods: This study employed a cross-sectional, mixed-methods design, comparing AI-derived ratings of clinical communication performance to expert human evaluations, across a set of four simulated doctor-patient encounters filmed at LSUHSC's patient simulation labs. Videos were organized into two scenarios, each consisting of two videos—one portraying low to moderate communication skills and the other portraying an improved version of the same encounter with more effective communication skills by the physcian. Videra Health's multimodal AI assessment platform analyzed each video using a 5-point Likert scale across 12 communication domains. Human raters from LSUHSC's standardized patient working group completed the same evaluations, along with open-ended questions assessing trust-related communication.

Results: AI and human ratings showed an 85.4% agreement within one Likert point across all four videos. The resulting mean absolute deviation (0.68) and average bias (+0.34 points) indicated a less than one-point difference in AI and human consensus ratings, with AI scores showing a consistent mild overestimation. Notably, while AI demonstrated strong directional accuracy evaluating communication quality, it showed weaker agreement in more interpretive or affective domains, particularly those involving emotional tone, trust-building, and multi-step communication behaviors.

Conclusions: AI-based communication assessment aligns closely with human ratings for structured, observable behaviors and demonstrates strong promise as a scalable tool for formative feedback in medical education. However, given current models decreased sensitivity to affective and contextual nuances involved in doctor-patient interactions, findings best support AI's use as a complementary adjunct to human expert judgement rather than a replacement.