Communication and Learning Improvement Model for Bedside Skills (CLIMBS): An AI-Based Feedback Model for the Objective Standardized Clinical Examination

Liam Hobson[1], Chandler Lutz[1], David Rodgers[2]
[1]Indiana University School of Medicine; [2]Indiana University School of Medicine, Department of Medicine and Interprofessional Simulation Center

Background: The Objective Standardized Clinical Examination (OSCE) is a tool designed to assess and provide feedback to healthcare students; however, feedback quality is a common point of frustration for students. Common concerns with post-OSCE feedback include that it is generally not timely, comprehensive, and/or individualized. In this retrospective, quasi-experimental pilot study, we created and implemented an artificial intelligence (AI) feedback model and assessed its ability to provide accurate and reliable OSCE feedback.

Methods: Using ChatGPT-4o, we developed the Communication and Learning Improvement Model for Bedside Skills (CLIMBS) to reference a 16-metric OSCE rubric and provide feedback on OSCE transcripts derived from a custom workflow. Using a single, full-marks OSCE recording, we assessed CLIMBS by calculating transcription and feedback accuracy and examining feedback reliability against an artificially-generated 'Poor Performance' transcript and a manually corrected transcript.

Results: The CLIMBS workflow exhibited 92.39% transcript accuracy and good inter-rater reliability (cosine similarity = 0.972 +/- 0.015; ICC2 = 0.744 (95% CI: [0.59, 0.88]). Overall OSCE scores assigned by CLIMBS to the unedited transcript (93.84; 95% CI [91.53, 96.17]) and the manually corrected transcript (87.69, 95% CI [81.37, 94.02]) were statistically different from the 'Poor Performance' transcript (78.08; 95% CI [71.59, 84.57]). The metric-based score pattern distribution of the manual transcript was similar to both the unedited and the 'Poor Performance' transcripts (cosine similarity = 0.890 and 0.819, respectively), while the unedited and 'Poor Performance' transcripts exhibited low overlap with each other (cosine similarity = 0.566).

Conclusion: This pilot study demonstrated that an AI model can analyze and provide accurate summative feedback to OSCE recordings – differentiating good from poor performance. The accuracy and reliability of formative feedback require further study. With future testing to improve inter-rater reliability and examine formative feedback, CLIMBS has potential to improve the timeliness, comprehensiveness, and personalization of OSCE feedback.