Evaluating Agreement Between AI and Human Graders in Medical School Simulation

An Initial Look at a Multi-Institutional Study

Judah Gruen, MD[1]; Ameer H. Shakur, PhD[2]; Lindsey Stockton, MD[3]; John Cunningham, MD[4]; Luyang Jin, MD[4]; Michael J. Holcomb, MS[2]; Andrew R. Jamieson, PhD[2]

[1] Department of Internal Medicine, UT Southwestern Medical Center, Dallas, TX
[2] Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center, Dallas, TX
[3] Department of Internal Medicine, UT Tyler School of Medicine, Tyler, TX
[4] Department of Internal Medicine, Long School of Medicine, UT Health San Antonio, San Antonio, TX

THE UNIVERSITY OF TEXAS SYSTEM
REAL Health AI
Harnessing AI to Transform Health Care

Learn More

Try it out yourself!

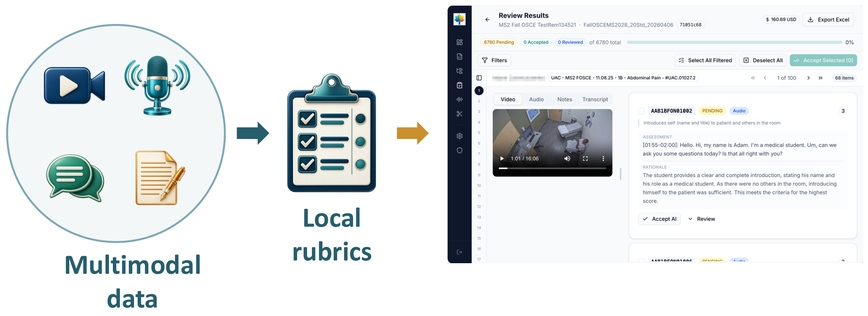
BACKGROUND

- Objective Structured Clinical Examinations (OSCEs) assess clinical reasoning, communication, and documentation through simulations with standardized patients.
- Faculty grading requires substantial time, and differences in rubric interpretation can affect consistency.
- AI may support more scalable and consistent assessment, but its agreement with faculty judgment requires rigorous evaluation before and during adoption.
- AI deployment at UT Southwestern demonstrated multimodal AI grading with human review [1]. This study extends that work to evaluation across institutions using local rubrics.

STUDY OVERVIEW

- Through the **UT-REAL Health AI Pilot Program**, UT Southwestern and five partner medical schools are developing a reproducible approach to AI-assisted OSCE assessment.
- This preliminary retrospective review compares AI and human rubric-item ratings across **six clinical cases at UT Tyler and UT Health San Antonio**.
- **Faculty reviewed disagreements** to identify unclear scoring criteria, guide rubric revisions, and reassess agreement.
- Retrospective reviews at **UT Houston, UT Rio Grande Valley, and UT Medical Branch** are planned next.

PROJECT MAPLES



RESULTS

Case	Learners	Items	Agreement
UT Tyler			
Pulmonary embolism	10	69	**80.8%**
Lichen planus	10	50	**75.1%**
Vitamin B12 deficiency	10	68	**81.6%**
Total	**10**	**187**	**79.0%**
UTHSA			
Cough & DOE	102	26	**86.3%**
Insomnia & confusion	105	8	**83.1%**
Low back pain	108	7	**63.1%**
Total	**219**	**41**	**81.3%**

Agreement: Percentage of rubric-items where AI and human scores matched exactly.

DISCUSSION

- **Faculty review of AI grades and rationales highlighted rubric quality considerations.**

"The management plan selects no more than five clinically coherent first-line conservative options."
Concern: Counting options could reward unsafe combinations.
Revision: Six response levels → four, with explicit safety penalties.

"The assessment prioritizes acute lumbar disc herniation with right-sided lumbosacral radiculopathy as the leading diagnosis."
Concern: Exact wording may penalize equivalent diagnoses.
Revision: Accept equivalent terms; read diagnosis and rationale together.

"The imaging plan makes a coherent decision on whether to image the low back."
Concern: Scoring expectations exceeded the learner prompt.
Revision: Credit justified MRI or justified deferral with follow-up and return precautions.

- UTSA agreement increased from 63.1% to 68.3% after rubric revision; further refinement is needed.
- UT Tyler achieved 75.1–81.6% agreement after a year of collaborative rubric refinement.
- **Iterative rubric refinement** is necessary to create the clarity needed for consistent grading.
- **Next steps:** Evaluate whether formative feedback explaining AI score deductions on individual items helps students learn and supports teaching.

REFERENCES
[1] Ngo HT, Shakur AH, Holcomb M, et al. Prospective Deployment of Multimodal AI Grading for Medical Student OSCEs. npj Digital Medicine. Accepted for publication.

UT Tyler — THE UNIVERSITY OF TEXAS AT TYLER
UT Health San Antonio
UT Southwestern Medical Center