Journal of University College of Medicine & Dentistry

Editorial

The Essential Role of Human Oversight in the Era of AI-Enabled Medical Assessment

Ahsan Sethi[1*], Mariyah Hidayat[2]

[1]College of Health Sciences, QU Health, Qatar University, Doha, Qatar

[2]University College of Medicine and Dentistry, The University of Lahore, Pakistan

*Corresponding Author

Ahsan Sethi

asethi@qu.edu.qa

Submission: 1st November 2025
Revision: 10th December, 2025
Acceptance: 25th December, 2025

DOI: https://doi.org/10.51846/jucmd.v5i1.4842

Cite this article as:

Sethi A, Hidayat M. The Essential Role of Human Oversight in the Era of AI-Enabled Medical Assessment. *J Univ Coll Med Dent*.2026;5(1):4-5.

Every medical educator recognizes the importance of assessment in health professions education.[1] A decision made in an examination room or an OSCE station can determine whether a learner progresses, repeats a year, or leaves the profession altogether. Such judgments are rarely based on scores alone. These decisions are shaped by the context, training of assessors, students' competence, response to uncertainty, professionalism, and performance in ambiguous situations. In medicine, assessment is therefore not merely a technical exercise; it is a moral responsibility with direct implications for patient safety and public trust.[1,2]

Across health professions education, AI systems are increasingly being advocated and used to generate assessment items, score written responses, detect plagiarism, and predict learner performance.[3] While their appeal is understandable, particularly in an era of expanding student numbers, faculty workload, and demands for standardization, the central question confronting medical education is not whether AI can assist assessment, but whether it should be allowed to influence judgments that carry profound professional and ethical consequences.[4]

AI systems excel at identifying patterns and reproducing statistical regularities from large datasets. However, apparent competence should not be mistaken for understanding. The classic example of *Clever Hans*, the horse that appeared to perform arithmetic while merely responding to subtle human cues, illustrates how convincing performance can mask shallow cognition.[5] Similarly, the concept of the *stochastic parrot* highlights how large language models generate fluent responses by predicting sequences rather than by reasoning or comprehension.[6] In assessment, this distinction matters deeply. Algorithms may score correctness, but they cannot interpret intent; they may analyse observable behaviour, but they cannot authentically discern empathy, integrity, or moral judgment, which are the core attributes that define a competent doctor.[2]

These limitations become particularly evident when assessment is viewed through Miller's Pyramid of Clinical Competence. AI may reasonably support assessments at the *Knows* and *Knows How* levels, such as factual recall or structured problem-solving. At the *Shows How* and *Does* levels, however, competence is demonstrated through performance, interaction, and professional conduct in real or simulated contexts. Here, assessment is inherently relational and value-laden. Reducing such judgments to algorithmic outputs risks oversimplifying complex human behaviour and misclassifying competence in ways that may be educationally and ethically harmful.[7]

AI undoubtedly has the potential to enhance assessment by improving efficiency and consistency.[8] AI automated assessment can reduce personal bias, make marking more consistent, and save time, especially when large numbers of students are assessed. Research also shows that AI-based scoring can be useful for structured assessments and can help spot differences in scoring.[8] However, growing evidence also highlights important shortcomings. Studies comparing AI and human evaluators in OSCEs have reported disagreement in scoring, limited sensitivity to nuanced communication skills, and difficulty interpreting contextual or culturally embedded expressions of professionalism.[8]

Rather than eliminating bias, inadequately contextualized AI systems may simply reproduce it in less visible ways. These concerns are especially salient in Global South contexts. Most AI systems are trained predominantly on datasets derived from the Global North, reflecting specific linguistic norms, cultural expectations, and professional behaviours. Learners from different cultural, linguistic, or religious backgrounds, including those in many Muslim-majority societies, may show

empathy, respect, or reasoning in ways that these systems do not easily recognize as standard. Without careful attention to how training data are generated and whose behaviours are represented, AI-based assessment risks disadvantaging those who do not conform to dominant cultural templates.[9]

Some early efforts have tested the use of AI to analyse video-recorded patient interactions in order to assess communication skills and empathy.[10] Although these methods are interesting, they raise important concerns. Can empathy truly be judged using visible behaviour alone? Who decides the rules used by the algorithm? What should be done when the judgment of a human assessor differs from that of a machine? Until clear evidence and transparent standards are available, such tools should be used with caution, only as support and not as the basis for final decisions.[4]

The responsible integration of AI into medical assessment, therefore, requires clear governance and layered accountability. Regulatory bodies and universities should lead by establishing guidelines that define acceptable and unacceptable uses of AI, particularly for high-stakes decisions. Universities must translate these principles into institutional policies that ensure transparency, auditability, and data protection. Academic leaders bear responsibility for resourcing faculty development and ensuring the integrity of assessment. Educators and assessors, ultimately, must retain authority, adopting a human-in-the-loop approach in which AI supports but never replaces professional judgment.[11]

In medicine, judgment requires experience, compassion informed by context, and accountability grounded in professional responsibility, all the features that algorithms may not possess.[12] The challenge before us is not whether AI should be used, but where its authority must end. As medical educators, we are now confronted with the question: While technology offers certainty without understanding, are we prepared to surrender assessment to AI without being accountable for the decisions that shape our learners, our patients, and our profession?

References

1. Sethi A, Khan NA, Zaib AJ. Exploring "failure to fail" behaviour among examiners of undergraduate medical programs. *Medical Teacher*. 2025;1–13. https://doi.org/10.1080/0142159X.2025.2593493

2. ten Cate O, Carraccio C. Medical competence: the interplay between individual ability and the learning environment. *Medical Teacher*. 2023;45(2):115–118. https://doi.org/10.1080/0142159X.2022.2145683

3. Tolsgaard MG, Pusic M, Sebok-Syer S, et al. The fundamentals of artificial intelligence in medical education research: AMEE Guide No. 156. *Medical Teacher*. 2023;45(6):565–573. https://doi.org/10.1080/0142159X.2023.2191995

4. Masters K, MacNeil H, Benjamin J, Carver T, Nemethy K, Valanci-Aroesty S, et al. Artificial intelligence in health professions education assessment: AMEE Guide No. 178. *Medical Teacher*. 2025;1–15.

5. Ethical use of artificial intelligence in health professions education. *Medical Teacher*. 2023;45(9):889–896. https://doi.org/10.1080/0142159X.2023.2243457

6. Pfungst O. *Clever Hans (The Horse of Mr. von Osten)*. New York: Holt; 1911. https://doi.org/10.1037/10852-000

7. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*; 2021. https://doi.org/10.1145/3442188.3445922

8. Miller GE. The assessment of clinical skills/competence/performance. *Academic Medicine*. 1990. https://doi.org/10.1097/00001888-199009000-00045

9. Tekin M, et al. Is artificial intelligence the future of evaluation in medical education? *BMC Medical Education*. 2025;25:54. https://doi.org/10.1186/s12909-025-07241-4

10. Birhane A. Algorithmic injustice: a relational ethics approach. *Patterns*. 2021;2(2):100205. https://doi.org/10.1016/j.patter.2021.100205

11. Stamer T, Steinhäuser J, Flägel K. Artificial intelligence supporting the training of communication skills in health care professions: a scoping review. *Journal of Medical Internet Research*. 2023;25:e43311. https://doi.org/10.2196/43311

12. Sethi A. Artificial intelligence in health professions education. *Journal of Shalamar Medical and Dental College*. 2024;5(1):1–3. https://doi.org/10.53685/jshmdc.v5i1.208