2025, Volume 57, Issue 10, 714-718, e-ISSN 1938-3800

Family Medicine

ORIGINAL ARTICLE

Using Artificial Intelligence in Clerkship Learner Assessment: A CERA Study

Anthony Dambro, MD[a]; Alyssa Anderson, MD[a]; Karl T. Clebak, MD, MHA[a]; Michael Partin, MD[a]; Juandalyn Burke, PhD, MPH[b]; Misbah Keen, MD, MPH[b]

AUTHOR AFFILIATIONS:

[a] Department of Family and Community Medicine, Penn State College of Medicine, Hershey, PA

[b] Department of Family Medicine, University of Washington, Seattle, WA

CORRESPONDING AUTHOR:

Anthony Dambro, Department of Family and Community Medicine, Penn State College of Medicine, Hershey, PA, adambro@pennstatehealth.psu.edu

HOW TO CITE: Dambro A, Anderson A, Clebak KT, et al. Using Artificial Intelligence in Clerkship Learner Assessment: A CERA Study. *Fam Med.* 2025;57(10):714-718.
doi: 10.22454/FamMed.2025.529827

FIRST PUBLISHED: November 20, 2025

KEYWORDS: artificial intelligence, assessment, clerkship

ABSTRACT

Background and Objectives: The integration of artificial intelligence (AI) in medical education primarily has focused on clinical applications, with limited investigation into its role in learner assessment. This study explores AI usage among family medicine clerkship directors and their perspectives on AI's potential in competency-based medical education assessment.

Methods: Data were collected through the 2024 Council of Academic Family Medicine Educational Research Alliance survey of family medicine clerkship directors. The survey was distributed from June four to July 12, 2024, to 173 directors, achieving a 52.6% (91/173) response rate. We used multivariable linear regression to analyze the relationship between AI usage and a composite favorability score for AI integration in student assessment, adjusting for covariates such as clerkship design and director tenure.

Results: All respondents were physicians leading mandatory clerkships. Female and underrepresented minority directors were more likely to report no prior AI use. Multivariable analysis demonstrated a significant positive association between AI usage and favorability toward AI in learner assessment (coefficient: 2.601; 95% CI: 1.246–3.956; P<.001), even after adjusting for multiple comparisons. Feedback quality, organizational support, and policies did not significantly impact favorability.

Conclusions: AI exposure was significantly associated with favorable attitudes toward AI in learner assessment, while organizational factors had no significant effect. Future studies with larger samples and longitudinal designs may clarify how institutional support and increasing AI exposure influence attitudes over time, informing best practices for AI adoption in medical education.

INTRODUCTION

Artificial intelligence (AI) has evolved dramatically from its early days of symbolic reasoning and rule-based systems to today's advanced machine learning and deep learning models. Historically, AI's initial promise in medicine was limited to data management and simple diagnostic tools. However, recent advancements in AI, including natural language processing and predictive analytics, have expanded its role to encompass more complex tasks such as analyzing images and data or summarizing and generating text.[1]

These advancements have the potential to enhance competency-based medical education (CBME) by offering innovative ways to assess learner performance accurately and efficiently. The intersection with medical education has focused mainly on building knowledge for future providers on how to use AI clinically[2] while other fields already are applying AI to educational tasks.[3] Thus far, most research regarding the integration of AI technologies into medical learner assessment has been completed within surgical residency programs and has focused on mapping narrative formative feedback to entrustable professional activity levels. Other studies have used AI to classify the quality of narrative feedback or to grade learner assignments

using set rubrics.[3-8] Where AI, especially large language models, may be most useful in the medical education space is in the collation and summarization of large amounts of narrative data about a learner's performance. In the context of CBME, an accurate distillation of faculty's descriptions of a learner's skills, and those they lack, will prove incredibly valuable.[8,9] This study aims to investigate the use of AI by family medicine clerkship directors and their perspectives on AI's potential in learner assessment within CBME.

METHODS

Data were gathered as part of the 2024 Council of Academic Family Medicine (CAFM) Educational Research Alliance (CERA) survey of family medicine clerkship directors. The methodology of the CERA clerkship director survey has previously been described in detail.[10,11] CAFM members were invited to propose survey questions for inclusion into the CERA survey. Approved projects were assigned a CERA research mentor to help refine questions. The survey items were pilot tested for content validity, and adjustments were made based on feedback to ensure clarity and relevance to the study objectives. The project was approved by the American Academy of Family Physicians Institutional Review Board in April 2024.

The 2024 list of survey recipients included 179 clerkship directors (164 in the United States, 15 in Canada). This list was generated by starting with the 169 respondents to the 2023 clerkship directors survey. Results from a separate department chairs survey identified five additional clerkship directors. Responses from 2023 clerkship directors survey invitations identified five additional clerkship directors.[11]

During the survey, 12 survey recipients indicated they were no longer the clerkship director and gave a replacement name and email. The new clerkship director was then sent an invitation to participate in the survey. One survey was sent to a residency program email address. The residency program was deleted from the pool. Five undeliverable email addresses were removed from the pool, yielding a final pool size 173 survey recipients (158 in the United States and 15 in Canada), which we believe represents all family medicine clerkship directors in North America. The survey was open from June 4, 2024, through July 12, 2024. Six reminders to nonresponders and partial responders were sent: five weekly and one on July 12, 2024.

The final survey opened with a brief introduction presenting a scenario in which generative AI and natural language models are used as an alternative to traditional Likert-scale assessments to process narrative data on student performance. With this context in mind, participants rated their sentiments on a 5-point Likert scale. They were asked about their comfort with using AI in student assessment, how they believed students might respond, and how AI could influence current practices—specifically regarding assessment quality, bias, standardization, and time savings. To better understand participants' institutional contexts, the survey also asked about the quality of feedback they typically receive from faculty about students, their perceptions of organizational support for AI in education, and the existence of any institutional policies guiding AI use. Finally, participants were asked about their own experience with AI, categorized as no use, use of a single model, or use of multiple models.

We calculated descriptive statistics using Microsoft Excel. We collapsed responses on the 5-point Likert scale into three categories for analysis: unfavorable (responses 1 and 2), neutral (response 3), and favorable (responses 4 and 5). We did this recoding to facilitate group comparisons and interpret trends in overall sentiment.

During the multivariable linear regression, the primary outcome, a composite favorability score, was derived from the sum of 5-point Likert scale responses for the six questions that assessed respondents' views on using AI in the educational space. The composite favorability score exhibited a symmetric distribution in a histogram plot, confirming its suitability for analysis with multivariable linear regression models. During our analysis, we recoded the primary exposure (AI usage) and secondary exposures (organizational support for AI integration, AI policies on AI integration, and perceived feedback quality) as binary responses. Covariates such as gender, clerkship length, clerkship design, and clerkship director tenure were considered in the model. Missing data were handled using multiple imputation. Two authors (A.D. and M.K.) used ChatGPT 4o (OpenAI)[12] to run the analysis separately to ensure that the results were congruent.[13,14]

RESULTS

We considered 93 total responses to the CERA survey. We counted two respondents who answered only the initial question "Are you the clerkship director?" as nonresponders. We also eliminated another responder who answered only two additional demographic questions not relevant to this study. Another eight respondents did not complete the AI survey questions, so the final 82 responses yielded an overall response rate of 47.4% (82/173).

All respondents were physicians, and all reported that the family medicine clerkship was mandatory. A total of 45.7% (37/81) of respondents reported having used AI in some personal or professional context. Relevant demographic information on the respondent sample is described in Table 1.

Table 2 compares the demographic data with prior AI use. Of the respondents that identified as female, 28/48 (58.3%) reported having no prior use of AI platforms. Seven participants self-identified as underrepresented in medicine (URiM), six of whom reported no prior use of AI (85.7%). Those with fewer years since graduation from residency (defined as graduation within 17 years, given the average time since graduation of the sample) were more likely to have experimented with AI (24/41, 58.5%) than those that graduated earlier (14/40, 35%).

TABLE 1. Demographics of Responders

Gender, n (%)	
Female	52 (57.8)
Male	37 (41.1)
Did not identify	1 (1.1)
Self-identify as underrepresented, n (%)	
No	80 (88.9)
Yes	8 (8.9)
Did not identify	2 (2.2)
Median years as a clerkship director	6
Average number of years since residency graduation	17.75
Average protected time for clerkship	29.80%
Average number of students per class	145.25
Clerkship design, n (%)	
Block only	66 (74.2)
Longitudinal only	10 (11.2)
Hybrid	13 (14.6)
Median length of clerkship	5 weeks

Table 3 shows aggregated participant responses to the individual survey questions. Participants were uncomfortable (42.7%) or neutral (41.5%) about using AI in student assessment. A majority of respondents (57.3%) were optimistic about the time savings AI may offer. Participants were largely neutral (63.4%) about the impact of AI on bias in assessment.

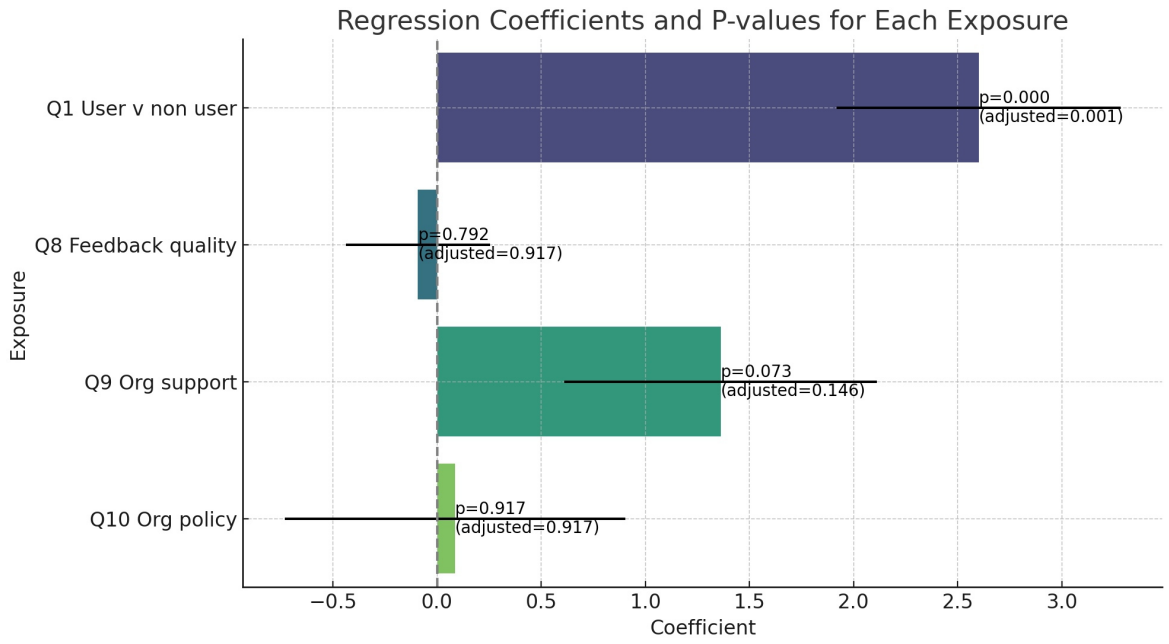
In the multivariable linear regression analysis, after adjusting for covariates (total clerkship time, clerkship design, current years as a clerkship director, and gender), the coefficient for AI user status was 2.601 (95% CI: 1.246–

TABLE 2. Experience With Creating Documents With AI by Demographics

Variable	Experience with AI, n (%)	No experience with AI, n (%)
Gender		
Male	18 (22.2)	15 (18.5)
Female	20 (24.7)	28 (34.6)
Self-identify as underrepresented		
Yes	1 (1.3)	6 (7.5)
No	36 (45.0)	37 (46.3)
Years as CD		
<6	19 (23.5)	16 (19.8)
6 +	19 (23.5)	27 (33.3)
Years since residency graduation		
<17	24 (29.6)	17 (21.0)
17 +	14 (17.3)	26 (32.1)
Protected time		
<30%	16 (19.8)	16 (19.8)
30+%	22 (27.2)	27 (33.3)
Number of students per class		
<145	16 (29.6)	24 (29.6)
145+	22 (27.2)	19 (23.5)
Clerkship design		
Block only	27 (33.3)	33 (40.7)
Long only	4 (4.9)	6 (7.4)
Hybrid	7 (8.6)	4 (4.9)
Length of clerkship (weeks)		
<5	19 (23.5)	21 (25.9)
5+	19 (23.5)	22 (27.2)

Abbreviations: AI, artificial intelligence; CD, clerkship director; long, longitudinal.

FIGURE 1. Regression Coefficients and *P* Values for Each Exposure



https://doi.org/10.22454/FamMed.2025.529827

TABLE 3. Survey Prompts and Participant Responses

Survey prompt	Unfavorable, n (%)	Neutral, n (%)	Favorable, n (%)
How comfortable do you feel about using AI during the student assessment process?	35 (42.7)	34 (41.5)	13 (15.9)
How do you anticipate your students would perceive the use of AI during their assessment process?	16 (19.5)	35 (42.7)	31 (37.8)
How do you think the integration of AI will affect the assessment methods currently used in clerkships?	14 (17.1)	39 (47.6)	29 (35.4)
How will AI impact bias in summary assessments (MSPE/Dean's letter)?	10 (12.2)	52 (63.4)	20 (24.4)
AI would save me time in completing summary assessments.	12 (14.6)	23 (28.0)	47 (57.3)
How will the use of AI to synthesize written feedback affect the standardization of reports on student performance?	14 (17.1)	35 (42.7)	33 (40.2)
Narrative comments received from faculty members at my institution provide an excellent summary of student performance.	27 (32.9)	21 (25.6)	34 (41.5)
My organization is supporting AI integration in education.	20 (24.7)	29 (35.8)	32 (39.5)
My organization has policies around AI use in education.	37 (45.7)	21 (25.9)	23 (28.4)

Abbreviations: AI, artificial intelligence; MSPE, medical student performance evaluation.

3.956; $P<.001$). This positive association remained significant after adjusting for multiple comparisons using the Benjamini-Hochberg procedure, with an adjusted P value of .001. Secondary exposures of feedback quality (coefficient: -0.0919; 95% CI: 0.782–0.598; $P=.792$), organization support (coefficient: 1.3621; 95% CI: -0.129–2.854; $P=.073$), and organizational policies (coefficient: 0.0860; 95% CI: -1.540–1.712; $P=.917$) were not significantly associated (Figure 1).

DISCUSSION

This study provides insights into the perceptions and experiences of family medicine clerkship directors regarding the integration of AI into learner assessment. Less than half of respondents had reported using AI for professional or personal reasons. The majority of respondents were uncomfortable or neutral about using AI in the student assessment process. They were more optimistic about students' perceptions of its use in assessment and AI having a positive impact on current assessment methods. Directors were clearly most favorable toward the prospect of time savings with these tools but were decidedly neutral about AI's impact on bias in assessment.

As hypothesized, those that had prior exposure to AI were overall more favorable in their views. The significant positive association between prior exposure to AI and a favorable attitude toward its integration suggests that familiarity fosters acceptance. Directors who have personally interacted with AI platforms are more inclined to appreciate the potential benefits in enhancing student assessments. The lack of significant associations with secondary exposures such as the quality of feedback from preceptors, organizational policies, and support indicates that individual experience with AI may outweigh institutional factors in shaping attitudes. However, the observed trend regarding organizational support hints at its potential influence. Plausibly, institutions endorsing AI integration may provide resources or training that enhance(s) directors' comfort levels, thereby positively affecting their perceptions. Future studies with larger sample sizes could elucidate this relationship further. The demographic disparities observed, particularly the lower likelihood of AI experience among female directors and those self-identifying as URiM, raise important considerations. Others have found similar disparities in technology uptake.[15–18] While the lack of exposure might be due to a lack of interest, other studies suggest that it more likely reflects broader systemic issues such as access to technological resources or varying levels of encouragement to engage with emerging technologies. This hypothesis would need to be further evaluated in future studies.

This study had several limitations. The response rate, while acceptable for a national cohort, still left room for nonresponse bias. Those with strong opinions about AI, either positive or negative, may have been more inclined to participate. Given that all data were self-reported opened opportunities for recall, extreme responses, and demand biases. Additionally, the cross-sectional design captured perceptions at a single point in time, which may not account for the rapidly evolving nature of AI technologies and their acceptance in the educational setting. Furthermore, the nature of survey data limited our ability to interpret nuanced responses. For example, decreased comfort with using AI in assessment could stem from unfamiliarity with the technology or could reflect ethical or pedagogical concerns.

CONCLUSIONS

In conclusion, as of 2024, family medicine clerkship directors were largely uncomfortable with including AI in student assessment, though prior exposure to AI significantly influenced favorability toward the idea. As AI technologies continue to evolve, providing opportunities for educators to engage with these tools could facilitate their integration into medical education. Institutions might consider implementing

workshops or pilot programs to familiarize faculty with AI applications in assessment, especially to help address any possible gender or racial barriers to entry.

Future research should explore, more deeply, why directors are uncomfortable and consider ways that would mitigate faculty concerns. As AI becomes more embedded in medicine, assessing the longitudinal changes in attitudes and the impact of institutional initiatives aimed at promoting AI literacy among educators also will be of interest moving forward.

PRESENTATIONS

Clerkship Director Sentiments on Using Artificial Intelligence in Clerkship Learner Assessment: A CERA Study. Presentation at 2025 Annual Society of Teachers of Family Medicine Spring Conference, May 3-7, 2025. Salt Lake City.

REFERENCES

1. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. 2017;2(4):230–243.
2. Masters K. Artificial intelligence in medical education. *Med Teach*. 2019;41(9):976–980.
3. González-Calatayud V, Prendes-Espinosa P, Roig-Vila R. Artificial intelligence for student assessment: a systematic review. *Appl Sci*. 2021;11(12):5467.
4. Gin BC, Ten Cate O, O'Sullivan PS, Hauer KE, Boscardin C. Exploring how feedback reflects entrustment decisions using artificial intelligence. *Med Educ*. 2022;56(3):303–311.
5. Gudgel BM, Melson AT, Dvorak J, Ding K, Siatkowski RM. Correlation of ophthalmology residency application characteristics with subsequent performance in residency. *J Acad Ophthalmol*. 2021;13(2):e151–e157.
6. Stahl CC, Jung SA, Rosser AA, et al. Natural language processing and entrustable professional activity text feedback in surgery: a machine learning model of resident autonomy. *Am J Surg*. 2021;221(2):369–375.
7. Ötleş E, Kendrick DE, Solano QP, et al. Using natural language processing to automatically assess feedback quality: findings from 3 surgical residencies. *Acad Med*. 2021;96(10):1457–1460.
8. Abbott KL, George BC, Sandhu G, et al. Natural language processing to estimate clinical competency committee ratings. *J Surg Educ*. 2021;78(6):2046–2051.
9. Booth GJ, Ross B, Cronin WA, et al. Competency-based assessments: leveraging artificial intelligence to predict subcompetency content. *Acad Med*. 2023;98(4):497–504.
10. Seehusen DA, Mainous AG III, Chessman AW. Creating a centralized infrastructure to facilitate medical education research. *Ann Fam Med*. 2018;16(3):257–260.
11. Kost A, Ellenbogen R, Biggs R, Paladine HL. Methodology, respondents, and past topics for 2024 CERA clerkship director survey. *PRiMER*. 2025;9:7.
12. OpenAI. Website. OpenAI, LLC. https://openai.com
13. Ordak M. Using ChatGPT in statistical analysis: recommendations for *JACC* Journals. *JACC Adv*. 2024;3(2):100776.
14. Huang Y, Wu R, He J, Xiang Y. Evaluating ChatGPT-4.0's data analytic proficiency in epidemiological studies: a comparative analysis with SAS, SPSS, and R. *J Glob Health*. 2024;14:04070.
15. Haluza D, Wernhart A. Does gender matter? Exploring perceptions regarding health technologies among employees and students at a medical university. *Int J Med Inform*. 2019;130:103948.
16. Møgelvang A, Bjelland C, Grassini S, Ludvigsen K. Gender differences in the use of generative artificial intelligence chatbots in higher education: characteristics and consequences. *Edu Sci*. 2024;14(12):1363.
17. Lee D, Rutsohn P. Racial differences in the usage of information technology: evidence from a national physician survey. *Educ Sci*. 2012;9:1.
18. Joseph L. The adoption and diffusion of computing and internet technologies in historically Black colleges and universities. *Int J Appl Manag Technol*. 2008;6(2):86–112.