2025, Volume 57, Issue 6, 424-429, e-ISSN 1938-3800

Family Medicine

ORIGINAL ARTICLE

Evaluating the Agreement Between ChatGPT and the Clinical Competency Committee in Assigning ACGME Milestones for Family Medicine Residents

Michael Partin, MD[a]; Anthony Dambro, MD[a]; Roland Newman, DO[a]; Yimeng Shang, MS[b]; Lan Kong, PhD[b]; Karl T. Clebak, MD, MHA[a]

AUTHOR AFFILIATIONS:

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

[b] Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA

CORRESPONDING AUTHOR:

Michael Partin, Department of Family and Community Medicine, Penn State College of Medicine, Hershey, PA, mpartin@pennstatehealth.psu.edu

HOW TO CITE: Partin M, Dambro A, Newman R, Shang Y, Kong L, Clebak KT. Evaluating the Agreement Between ChatGPT and the Clinical Competency Committee in Assigning ACGME Milestones for Family Medicine Residents. *Fam Med.* 2025;57(6):424-429. doi: 10.22454/FamMed.2025.363712

PUBLISHED: 6 June 2025

KEYWORDS: artificial intelligence, family practice, graduate medical education

ABSTRACT

Background and Objectives: Although artificial intelligence models have existed for decades, the demand for application of these tools within health care and especially medical education are exponentially expanding. Pressure is mounting to increase direct observation and faculty feedback for resident learners, which can create administrative burdens for a Clinical Competency Committee (CCC). This study aimed to assess the feasibility of utilizing a large language model (ChatGPT) in family medicine residency evaluation by comparing the agreement between ChatGPT and the CCC for the Accreditation Council for Graduate Medical Education (ACGME) family medicine milestone levels and examining potential biases in milestone assignment.

Methods: Written faculty feedback for 24 residents from July 2022 to December 2022 at our institution was collated and de-identified. Using standardized prompts for each query, we used ChatGPT to assign milestone levels based on faculty feedback for 11 ACGME subcompetencies. We analyzed these levels for correlation and agreement between actual levels assigned by the CCC.

Results: Using Pearson's correlation coefficient, we found an overall positive and strong correlation between ChatGPT and the CCC for competencies of patient care, medical knowledge, communication, and professionalism. We found no significant difference in correlation or mean difference in milestone level between male and female residents. No significant difference existed between residents with a high faculty feedback word count versus a low word count.

Conclusions : This study demonstrates the feasibility for tools like ChatGPT to assist in the evaluation process of family medicine residents without apparent bias based on gender or word count.

INTRODUCTION

Artificial intelligence (AI) conceptually dates back to the 1950s when an interface between humans and machine intelligence was first described.[1] Around the 1990s and early 2000s, technology advanced to the point where AI tools were being integrated into clinical practice. These advancements laid the foundation for subsequent applications of AI in medical education, creating opportunities to enhance both teaching and evaluation processes.

AI technology has integrated into the world of medical education, spanning learners from undergraduate medical education to independent practicing providers. As of 2020, the doubling time of new medical information available to learners was estimated at just 73 days, compared to 3.5 years in 2010 and 7 years in 1980.[2] Current areas of AI use in medical education include developing curriculum, providing feedback to learners, and delivering content.[3,4] Chatbots are commonly used AI tools that use natural language processing models to interpret queries made by users and produce a response synthesizing large amounts of information from the Internet.[5] Developed by OpenAI (OpenAI, LLC) and launched in 2022, ChatGPT is an example of a large language model chatbot.[6] Applications of chatbots in medical education include summarizing information from evidence-based resources, transforming presentations into question-and-answer flash cards, and generating acronyms or mnemonics to help with retention of information.

Within graduate medical education, chatbots have been ethically analyzed regarding proper uses for applicants applying to residency programs as well as for programs evaluating applicants.[7,8] However, a gap exists in the literature regarding

the use of AI technologies such as ChatGPT to assist in the evaluation process of resident performance, progression, and readiness for independent practice. The Accreditation Council for Graduate Medical Education (ACGME) has developed milestones for each medical specialty, serving as a road map of progression through residency.[9] ACGME milestones are divided into six core competencies: patient care (PC), medical knowledge (MK), professionalism (Prof), interpersonal and communication skills (ICS), practice-based learning and improvement (PBLI), and systems-based practices (SBP). The core competencies are then divided into subcompetencies, each with a scale from 1 to 5. Generally speaking, level 1 represents a novice or starting resident, while level 4 is a graduation target for programs. Level 5 describes aspirational behaviors, including role modeling or mentoring others in the core competency. The Clinical Competency Committee (CCC) represents a group of residency program faculty and leadership who meet regularly to evaluate resident performance across competencies and assign levels for specialty-specific milestones.

Family medicine residency programs are undergoing numerous changes as an emphasis is placed on competency-based medical education, which prioritizes real-time formative feedback using direct observation.[10] Learners ultimately benefit from an increased volume of evaluations, but that can perceivably create an increased administrative burden for residency faculty, leadership, and the CCC, especially when faced with decreasing protected administrative time.[11] AI chatbots such as ChatGPT represent a potential tool for residency faculty to aid in the evaluation process and assist CCCs in assigning milestone levels based on faculty written feedback. This study aimed to assess the agreement between ChatGPT and CCC-assigned ACGME milestone levels for family medicine residents at our institution based on faculty written feedback.

METHODS

This study was approved by the Penn State College of Medicine Institutional Review Board. All active residents during the period of July 2022 to December 2022 at our institution (an academic medical center) were selected spanning postgraduate years (PGY) 1 to 3. Written faculty comments and feedback were accessed in the residency management program New Innovations for the Fall 2022 semester. Written feedback sources included end-of-rotation evaluations and formative one-time shift cards. Evaluators included both family medicine faculty and nonfamily medicine faculty, accounting for when residents rotated in other specialties. Clinical settings included outpatient continuity clinics, outpatient specialty clinics, inpatient adult and pediatric medicine, labor and delivery, emergency medicine, and outpatient procedure clinics. Faculty comments from these different sources were compiled for each resident in this study. Each compilation of comments underwent a manual de-identification process to ensure that the identity of residents and faculty alike was undiscoverable. In doing so, names and pronouns were replaced with an X in the faculty comments. Further, comments were reviewed to ensure that no discoverable clinical scenarios or patient protected health information was included. Prior to de-identification, the gender and postgraduate year of each resident was recorded. After de-identification, the total word count of faculty feedback was recorded for each dataset. Each resident was assigned a random identification number for data tracking while maintaining anonymity.

OpenAI's software ChatGPT 4o-mini was selected for this study because it does not require a subscription, and thus the tool is accessible to all users. This was the current OpenAI software available at the time of data analysis and writing of the manuscript. A standardized and sequential process of interfacing with ChatGPT took place for each resident's de-identified and compiled faculty feedback. This process integrated the CARE (context, action, result, example) prompt engineering framework developed by Juuzt AI to standardize and enhance the desired outcomes from ChatGPT.[12] First, a new chat in ChatGPT was opened and the following prompt entered:

> Find below the Family Medicine Residency ACGME milestones and associated level for "patient care 1," "patient care 2," "patient care 3," "patient care 4," "patient care 5," "medical knowledge 1," "medical knowledge 2," "professionalism 1," "professionalism 2," "interpersonal and communication skills 1," and "interpersonal and communication skills 2."

Before submitting to ChatGPT, the ACGME milestones for the listed subcompetencies were copied and pasted verbatim following the preceding prompt to allow the language model to become familiar with the behaviors expected for achieving a certain level in the milestone.[13] These 11 family medicine subcompetencies, of the total 19, were selected as the most common areas of faculty provided written feedback to residents. Further, these tend to be areas in which residents often require remediation or more intentionally directed education. The subcompetencies within SBP and PBLI were excluded for this study. After the initial prompt was submitted, a second standardized prompt was inserted, stating:

> Find below comments provided by faculty members for family medicine resident "X." Based on these comments please assign a level for the following ACGME milestones based on the rubric above: "patient care 1," "patient care 2," "patient care 3," "patient care 4," "patient care 5," "medical knowledge 1," "medical knowledge 2," "professionalism 1," "professionalism 2," "interpersonal and communication skills 1," and "interpersonal and communication skills 2." Scores can be fractionated (eg, 1.5, 2.5, 3.5, 4.5). All behaviors in a given level need to be met in order to achieve that level.

Then, the de-identified summary for the individual resident was inserted verbatim following the prompt and submitted to ChatGPT. After ChatGPT generated a milestone level for the requested subcompetency, these data were transferred to a password-protected spreadsheet for later analysis. A new chat was then opened to prevent confounding influence from previous faculty comments or chats. The preceding process was then repeated for each resident's faculty feedback.

New Innovations data from July 2022 to December 2022 were extracted and de-identified to include milestone levels already assigned by the CCC and submitted to ACGME in 2022 for the subcompetencies included in the study. Statistical analysis was performed using the software programs R (R Foundation), SPSS (IBM), and Microsoft Excel. A Pearson's correlation coefficient (r_p) was determined to assess the linear strength and direction between milestone levels assigned by ChatGPT and the CCC. Pearson's correlation coefficients between 0.6 and 1.0 indicate a strong positive correlation, while coefficients between 0.4 and 0.59 indicate a moderate positive correlation. Spearman's rank correlation coefficient (r_s) was calculated to evaluate the monotonic relationship between the two groups. A mean difference was calculated between levels assigned by ChatGPT and the CCC for each subcompetency. A paired t test was calculated to determine whether a significant difference existed in the means of milestone levels between ChatGPT and the CCC using a *P* value threshold of <.05. A concordance correlation coefficient (r_c) was calculated to assess agreement from both a precision and accuracy standpoint between ChatGPT and the CCC. Each subcompetency was analyzed for agreement individually and aggregated into competencies of PC, MK, Prof, and ICS using averages. A subgroup analysis was performed looking at aggregate competencies between male and female residents, postgraduate year, and high versus low word count on faculty feedback. High word count was defined as above the median word count (416) of all residents and low word count below this threshold.

RESULTS

Study Population

Within the studied timeframe of July 2022 to December 2022, 24 total residents were included in the analysis. That included nine PGY-1 residents, seven PGY-2 residents, and eight PGY-3 residents. Of the 24 residents, 11 were female and 13 were male. The average word count of faculty feedback for all residents was 473, with a PGY-1 average of 557, PGY-2 of 518, and PGY-3 of 340. The average word count for female residents was 432 and for male residents was 508.

Correlation by Competency

In this analysis, 11 family medicine ACGME subcompetencies were studied, with the subcompetencies then aggregated into four distinct competencies (PC, MK, Prof, and ICS). Of the 16 areas analyzed, 15 showed a strong positive correlation (r_p= 0.6–1.0) between ChatGPT and the CCC using Pearson's correlation coefficient (Table 1). The only domain to show moderate correlation was ICS1 (patient- and family-centered communication).[13] Three domains showed very strong correlation, including PC2 (care of patients with chronic illness), aggregate PC, and all competencies aggregated together. The concordance correlation coefficient, which assesses agreement in milestone assignment, showed moderate to strong agreement between 13 of 16 areas between ChatGPT and the CCC (r_c=0.4–1.0). The three areas showing weak agreement were ICS1, ICS2 (interprofessional and team communication), and aggregate ICS.

The average mean difference between milestone levels (1–5) assigned was 0.58 for all subcompetencies and competencies. Of the 16 domains, six showed a mean difference of ≤0.5. All but two domains (PC3 and PC5) had a *P* value of <.05.

Correlation by Gender

In general, we found no major differences in Pearson's correlation coefficient between female and male residents in aggregate PC, MK, ICS, and all competencies analyzed together (Table 2). Within the professionalism aggregate, we found a noticeable difference in Pearson's correlation coefficient between female (0.46) and male (0.80). Spearman's rank correlation coefficient for this domain was 0.585 (95% CI; 0.103, 0.865) for females and 0.793 (95% CI; 0.476, 0.938) for males. These differences are not significant because the confidence intervals overlap. We also found no major differences in the mean difference for male and female residents when comparing milestone level assignment between ChatGPT and the CCC for all studied domains.

Correlation by Postgraduate Year

All postgraduate years showed a positive correlation between ChatGPT- and CCC-assigned milestone levels using Pearson's correlation coefficient (Table 3). The majority of competency domains for PGY-1 residents showed a weak to moderate correlation, with the lowest being 0.02 for MK. PGY-3 residents consistently showed moderate to high correlation. PGY-2 residents showed the strongest correlation, with strong to very strong levels according to Pearson's correlation coefficient. The mean difference was greatest among all domains for the PGY-1 residents, with scores generally being level 1 or higher assigned from ChatGPT compared to the CCC. The PGY-2 analysis showed mean differences more consistent with overall mean differences, as seen in Table 1, with the majority <0.5. The mean difference for the PGY-3 cohort for PC, MK, and overall aggregate was negative, suggesting that the CCC assigned a higher score on average than ChatGPT. The remaining domains of ICS and Prof were 0.19, suggesting a small mean difference between ChatGPT and the CCC.

Correlation by Word Count

Based on the median word count for all resident feedback being 416, we identified 12 residents in the high word count group (≥416 words) and 12 in the low word count group (<416 words). Both high and low word counts have a positive and moderate to strong correlation between all areas using

TABLE 1. Comparison of ChatGPT- and CCC-Assigned Milestone Levels (1–5) by ACGME Subcompetencies and Aggregate Competencies (N=24)

Subcompetency and competency	Pearson	Concordance correlation coefficient (95% CI)	Mean difference (ChatGPT-CCC)	*P* value
PC1	0.73	0.465 (0.238, 0.644)	0.71	<.01
PC2	0.84	0.670 (0.460, 0.809)	0.50	<.01
PC3	0.78	0.694 (0.478, 0.831)	0.19	.15
PC4	0.72	0.549 (0.302, 0.726)	0.46	<.01
PC5	0.73	0.640 (0.395, 0.800)	0.19	.16
MK1	0.78	0.55 (0.321, 0.718)	0.65	<.01
MK2	0.66	0.546 (0.265, 0.742)	0.46	<.01
Prof1	0.68	0.433 (0.189, 0.626)	0.75	<.01
Prof2	0.62	0.453 (0.175, 0.665)	0.60	<.01
ICS1	0.57	0.362 (0.109, 0.572)	0.79	<.01
ICS2	0.68	0.380 (0.160, 0.563)	0.88	<.01
PC aggregate	0.87	0.656 (0.472, 0.785)	0.41	<.01
MK aggregate	0.74	0.559 (0.312, 0.735)	0.55	<.01
Prof aggregate	0.68	0.456 (0.206, 0.651)	0.68	<.01
ICS aggregate	0.65	0.379 (0.151, 0.569)	0.83	<.01
All aggregate	0.82	0.562 (0.352, 0.718)	0.56	<.01

Abbreviations: CCC, Clinical Competency Committee; ACGME, Accreditation Council for Graduate Medical Education; PC, patient care; MK, medical knowledge; Prof, professionalism; ICS, interpersonal and communication skills; CI, confidence interval

TABLE 2. Comparison of ChatGPT- and CCC-Assigned Milestone Levels (1–5) by ACGME Competency and Gender (Female n=11, Male n=13)

Competency	Pearson		Mean difference (ChatGPT-CCC)	
	Female	Male	Female	Male
PC aggregate	0.88	0.85	0.35	0.45
MK aggregate	0.78	0.68	0.57	0.54
Prof aggregate	0.46	0.80	0.75	0.62
ICS aggregate	0.61	0.69	0.77	0.88
All aggregate	0.82	0.83	0.54	0.58

Abbreviations: CCC, Clinical Competency Committee; ACGME, Accreditation Council for Graduate Medical Education; PC, patient care; MK, medical knowledge; Prof, professionalism; ICS, interpersonal and communication skills

TABLE 3. Comparison of ChatGPT-and CCC-Assigned Milestone Levels (1–5) by ACGME Competency and Postgraduate Year (PGY-1, n=9; PGY-2, n=7; PGY-3, n=8)

Competency	Pearson			Mean difference (ChatGPT-CCC)		
	PGY-1	PGY-2	PGY-3	PGY-1	PGY-2	PGY-3
PC aggregate	0.59	0.82	0.43	0.98	0.33	−0.16
MK aggregate	0.02	0.77	0.46	1.17	0.43	−0.03
Prof aggregate	0.20	0.85	0.69	1.22	0.54	0.19
ICS aggregate	0.37	0.69	0.63	1.50	0.71	0.19
All aggregate	0.36	0.93	0.56	1.15	0.45	−0.01

Abbreviations:CCC, Clinical Competency Committee; ACGME, Accreditation Council for Graduate MedicalEducation; PGY, postgraduate year; PC, care of patients; MK, medical knowledge;Prof, professionalism; ICS, interpersonal and communication skills

TABLE 4. Comparison of ChatGPT- and CCC-Assigned Milestone Levels (1–5) by ACGME Competency and Word Count

Competency	Pearson		Mean difference (ChatGPT-CCC)	
	High	Low	High	Low
PC aggregate	0.89	0.85	0.48	0.34
MK aggregate	0.77	0.73	0.52	0.58
Prof aggregate	0.75	0.59	0.69	0.67
ICS aggregate	0.64	0.70	0.98	0.69
All aggregate	0.85	0.8	0.61	0.51

Note: Median word count for all residents was 416 with high (n=12) being ≥416 and low (n=12) being <416.
Abbreviations: CCC, Clinical Competency Committee; ACGME, Accreditation Council for Graduate Medical Education; PC, patient care; MK, medical knowledge; Prof, professionalism; ICS, interpersonal and communication skills

Pearson's correlation coefficient (Table 4). Spearman's rank correlation coefficients were all moderate to strong between high and low word counts with all corresponding confidence intervals overlapping, suggesting no significant difference between the two groups. The majority of mean differences between high and low word count were similar for each domain. The greatest difference was in ICS for the high word count group (0.98) and low word count group (0.69).

DISCUSSION

This feasibility study analyzed the correlation and agreement between ACGME milestone levels assigned to family medicine residents between use of ChatGPT and our institution's CCC based on written faculty evaluations. For all domains studied in the main analysis and subgroup analysis, a positive Pearson's correlation coefficient was found. This suggests that ChatGPT is consistent in aligning with our institution's CCC in a linear relationship. Almost all subcompetencies and competencies demonstrated a strong to very strong correlation. Further, the concordance correlation efficient showed moderate concordance between the two groups in all domains except ICS1, ICS2, and ICS aggregate. These findings may represent limited comments about a resident's communication style in faculty evaluations. Further, these findings raise the question of limitations in using ChatGPT to assess progression in this competency, whereas historically this competency has been evaluated using direct observation in both simulated and real-time clinical scenarios. ChatGPT's performance is dependent on the quality and specificity of the input data, which may inherently bias its assessments if faculty comments lack unbiased, objective, and detailed descriptions of specific resident behaviors.

Concerns have been raised in the literature about potential biases inherent to using AI tools.[14] Based on this analysis, we found no major differences between male and female residents in terms of correlation or mean difference between ChatGPT and the CCC. This finding highlights the importance of de-identifying information when using this tool, not only to protect a learner's educational record but also to prevent inherent biases of the tool itself. We found a modest difference in strength of correlation noted between female (0.46) and male (0.8) within the professionalism aggregate competency. The mean difference was not significantly different between the two (0.75 and 0.62, respectively) but raises the question of whether more comments are made for males about professionalism strengths or weaknesses than for females.

ChatGPT and the CCC were positively correlated for all domains regardless of postgraduate year. However, the strongest correlation existed among the PGY-2 cohort and the weakest among the PGY-1 cohort. The mean difference was greatest in the PGY-1 cohort, suggesting the possibility of inflated scores from ChatGPT for more junior residents. Conversely, most PGY-3 mean differences showed a higher score from the CCC than from ChatGPT, suggesting possible deflation of scores for more senior residents.

When approaching use of AI tools, one might assume that a submission with a higher word count will assist the language model in generating a more closely aligned response to a human counterpart. However, this study did not demonstrate a significant difference between higher word count faculty feedback versus lower word count. This finding suggests that although the word count may be higher, the quality of the feedback may be lower, less specific, or redundant. This study did not assess the correlation between ChatGPT and the CCC based on number of total faculty comments per resident submitted, but represents an area of future study. Overall, faculty comments may be concise but highlight the importance of using specific situations or milestone-aligned language if using an AI tool in evaluations.

Limitations of this study included analyzing data from only one institution and one semester of faculty feedback. Additionally, milestone levels may vary from query to query with ChatGPT despite the same input data, raising questions about reproducibility and precision. This discrepancy requires further research on the integrity of using these tools in this process. These tools are likely to improve as large language models are continuously refined. This study also was limited by assessing capabilities of AI tools in residency evaluation from strictly a numerical standpoint of milestone levels. Although tools like ChatGPT have potential for providing summative and longitudinal feedback, these features were not assessed in this study. Finally, this study did not quantify the amount of time saved using AI technology to assist in the evaluation process. Future studies are needed to assess the reduction in adminis-

trative burden, especially for larger residency programs, when using tools like ChatGPT.

CONCLUSIONS

This retrospective study is one of the first to assess the feasibility of using a large language model to assist in evaluating family medicine residents against ACGME milestones. As the guideline for more direct observation and increased faculty feedback grows, financial pressures may lead to reduced faculty protected time; tools like ChatGPT can perceivably increase efficiency and reduce administrative burden. This study supports the feasibility of using de-identified faculty written feedback to predict ACGME milestones for subcompetencies within family medicine. Future directions include integrating tools like ChatGPT into the CCC workflow in real time to study its efficacy. Further, these tools have the potential to predict the longitudinal trajectory of a resident's performance based on past comments and to suggest areas of improvement in a concise manner. Finally, research is needed on the utility of these tools in residency remediation and individualized learning plans.

PRESENTATIONS

Society of Teachers of Family Medicine Annual Spring Conference, May 7, 2024, Los Angeles, California.

REFERENCES

1. Turing AM. Computing machinery and intelligence. *Mind.* 1950;59(236):433-460.
2. Densen P. Challenges and opportunities facing medical education. *Trans Am Clin Climatol Assoc.* 2011;122:48-58.
3. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. *JMIR Med Educ.* 2019;5(1):13930.
4. Wu D, Xiang Y, Wu X. Artificial intelligence-tutoring problem-based learning in ophthalmology clerkship. *Ann Transl Med.* 2020;8(11):700.
5. Ghorashi N, Ismail A, Ghosh P, Sidawy A, Javan R. AI-powered chatbots in medical education: potential applications and implications. *Cureus.* 2023;15(8):43271.
6. OpenAI. Introducing ChatGPT. 2022. https://openai.com/blog/chatgpt.
7. Chen JX, Bowe S, Deng F. Residency applications in the era of generative artificial intelligence. *J Grad Med Educ.* 2024;16(3):254-256.
8. Mangold S, Ream M. Artificial intelligence in graduate medical education applications. *J Grad Med Educ.* 2024;16(2):115-118.
9. ACGME Common Program Requirements (Residency). *Accreditation Council for Graduate Medical Education.* . 2024. https://www.acgme.org/globalassets/pfassets/programrequirements/cprresidency_2023.pdf.
10. Newton W, Cagno CK, Hoekzema GS, Edje L. Core outcomes of residency training 2022 (provisional). *Ann Fam Med.* 2023;21(2):191-194.
11. Ringwald BA, Auciello S, Ginty J, Jefferis M, Stacey S. Administrative time expectations for residency core faculty: a CERA study. *Fam Med.* 2024;56(7):428-434.
12. Juuzt AI. Advancing AI prompt engineering with the CARE Framework. 2024. https://juuzt.ai/knowledge-base/prompt-frameworks/the-care-framework.
13. Family Medicine Milestones. *Accreditation Council for Graduate Medical Education.* 2019. https://www.acgme.org/globalassets/pdfs/milestones/familymedicinemilestones.pdf.
14. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med.* 2022;28(1):31-38.