Assessing the ability of a large language model to score free text medical student clinical notes: A quantitative study

Harry B Burke, Albert Hoang, Joseph O Lopreiato, Heidi King, Paul Hemmer, Michael Montogmery, Viktoria Gagarin

Submitted to: JMIR Medical Education
on: January 15, 2024

Table of Contents

Assessing the ability of a large language model to score free text medical student clinical notes: A quantitative study

Harry B Burke[1] MD, PhD; Albert Hoang[2] PhD, DSc; Joseph O Lopreiato[1] MD; Heidi King[3] MS; Paul Hemmer[1] MD; Michael Montogmery[1]; Viktoria Gagarin[1] MD

[1]Uniformed Services University of the Health Sciences Bethesda US
[2]Uniformed Services Univrsity of the Health Sciences Bethesda US
[3]Defense Health Agency Falls Church US

Corresponding Author:
Harry B Burke MD, PhD
Uniformed Services University of the Health Sciences
4301 Jones Bridge Road
Bethesda
US

Abstract

Background: Teaching medical students the skills required to acquire, interpret, apply, and communicate clinical information is an integral part of medical education. A crucial aspect of this process involves providing students with feedback regarding the quality of their free-text clinical notes.

Objective: The objective of this project is to assess the ability of ChatGPT 3.5 (ChatGPT) to score medical students' free text history and physical notes.

Methods: This is a single institution, retrospective study. Standardized patients learned a prespecified clinical case and, acting as the patient, interacted with medical students. Each student wrote a free text history and physical note of their interaction. ChatGPT is a large language model (LLM). The students' notes were scored independently by the standardized patients and ChatGPT using a prespecified scoring rubric that consisted of 85 case elements. The measure of accuracy was percent correct.

Results: The study population consisted of 168 first year medical students. There was a total of 14,280 scores. The standardized patient incorrect scoring rate (error) was 7.2% and the ChatGPT incorrect scoring rate was 1.0%. The ChatGPT error rate was 86% lower than the standardized patient error rate. The standardized patient mean incorrect scoring rate of 85 (SD 74) was significantly higher than the ChatGPT mean incorrect scoring rate of 12 (SD 11), p = 0.002.

Conclusions: ChatGPT had a significantly lower error rate than the standardized patients. This suggests that an LLM can be used to score medical students' notes. Furthermore, it is expected that, in the near future, LLM programs will provide real time feedback to practicing physicians regarding their free text notes. Generative pretrained transformer artificial intelligence programs represent an important advance in medical education and in the practice of medicine.

(JMIR Preprints 15/01/2024:56342)
DOI: https://doi.org/10.2196/preprints.56342

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Assessing the ability of a large language model to score free text medical student clinical notes: A quantitative study

Harry B. Burke[1], MD, PhD; Albert Hoang[1], PhD, DSc; Joseph O. Lopreiato[2], MD, MPH; Heidi King[3], MS; Paul Hemmer, MD,[1] Michael Montgomery[1]; Viktoria Gagarin[1], MD, MS

1. Department of Medicine, Uniformed Services University of the Health Sciences, Bethesda, MD

2. Department of Pediatrics, Uniformed Services University of the Health Sciences, Bethesda, MD

3. Patient Safety Program, Defense Health Agency, Falls Church, VA

Key words: medical education, generative artificial intelligence, natural language processing, ChatGPT, generative pretrained transformer, standardized patients, clinical notes, free text notes, history and physical examination

Abstract: 293
Words: 1,979
Table: 1
Figure: 0

Corresponding author:

Harry B. Burke, MD, PhD
Department of Medicine
F. Edward Hébert School of Medicine
Uniformed Services University of the Health Sciences
4301 Jones Bridge Road, Bethesda, MD 20814
301-295-4162
harry.burke@usuhs.edu

Abstract

Background

Teaching medical students the skills required to acquire, interpret, apply, and communicate clinical information is an integral part of medical education. A crucial aspect of this process involves providing students with feedback regarding the quality of their free-text clinical notes.

Objective

The goal of this project is to assess the ability of ChatGPT 3.5 (ChatGPT) to score medical students' free text history and physical notes.

Methods

This is a single institution, retrospective study. Standardized patients learned a prespecified clinical case and, acting as the patient, interacted with medical students. Each student wrote a free text history and physical note of their interaction. ChatGPT is a large language model (LLM). The students' notes were scored independently by the standardized patients and ChatGPT using a prespecified scoring rubric that consisted of 85 case elements. The measure of accuracy was percent correct.

Results

The study population consisted of 168 first year medical students. There was a total of 14,280 scores. The ChatGPT incorrect scoring rate was 1.0% and the standardized patient incorrect scoring rate was 7.2%. The ChatGPT error rate was 86% lower than the standardized patient error rate. The ChatGPT mean incorrect scoring rate of 12 (SD 11) was significantly lower than the standardized patient mean incorrect scoring rate of 85 (SD 74), $P = .002$.

Conclusions

ChatGPT demonstrated a significantly lower error rate than that of standardized patients. This is the first study to assess the ability of a GPT program to score medical students' standardized patient-based free text clinical notes. It is expected that, in the near future, LLM programs will provide real time feedback to practicing physicians regarding their free text notes. Generative pretrained transformer artificial intelligence programs represent an important advance in medical education and in the practice of medicine.

Introduction

Teaching medical students the skills required to acquire, interpret, apply, and communicate medical information is an integral part of medical education. A crucial aspect of this process involves providing students with feedback regarding the quality of their free-text clinical notes. Various methods have been employed to systematically assess clinical notes, notably, QNOTE [1,2] but they depend on human raters. This reliance presents numerous challenges, including rater recruitment and training, and raters having the availability and inclination to perform the reviews. Furthermore, humans are susceptible to biases, fatigue, and misinterpretation.

An attractive innovative alternative to human raters is to use a natural language processing (NLP) program to score student notes. An NLP program is a computer-based algorithm that automatically detects specific meanings in free text. The potential advantages of using an NLP program to grade student notes include that: it is systematic; it is objective; it avoids human bias, fatigue, and misinterpretation; it is essentially free to run; it can assess any number of notes in seconds; and it can grade notes in real time in order to provide immediate student feedback.

A new type of NLP program was introduced in November 2022, namely, ChatGPT 3.5 (ChatGPT) (OpenAI, San Francisco, CA) [3] a large language model (LLM) based on the generative pretrained transformer (GPT) artificial intelligence algorithm. It has achieved a 91.7% score on United States Medical Licensing Examination (USMLE) style questions [4]. Furthermore, its score on a clinical knowledge test was 87.3%, on medical genetics its score was 91.7%, on anatomy its score was 89.2%, and on professional medicine its score was 92.4% [4]. Its medical-related capabilities include that it can improve clinician empathy [5], respond to patient questions [6], perform differential diagnoses [7], classify radiology reports [8], write discharge summaries [9], provide accurate prevention advice to patients [10], and predict suicide risk [11]. ChatGPT has been compared to human raters in terms of grading short-answer pre-clerkship medical questions. The

ChatGPT-human Spearman correlations for a single assessor ranged from 0.6 to 0.7 [12].

We assess the ability of ChatGPT to accurately score medical student history of present illness, physical examination, and assessment and plan free text notes and we compare its scores to standardized patients' scoring of the clinical notes. We hypothesized that ChatGPT would be more accurate than standardized patients. To our knowledge, this is the first study to assess the ability of a GPT program to score medical student standardized patient-based clinical free text notes.

Methods

This is a single institution, retrospective study. Standardized patients are people who volunteer to interact with medical students to assist in their clinical training. They are trained on a prespecified medical case and, as the patient, they interact with a first-year medical student as if they had the medical condition they were trained on, including responding to clinical questions and undergoing an examination by the medical student. The students document their interaction with standardized patients in free text clinical notes. They write a chief complaint, history of the present illness, review of systems, physical examination, and differential diagnosis featuring three rank ordered diagnoses. In addition, they provide their pertinent positives and negatives and they suggest follow-up tests. At our medical school, standardized patients provide verbal feedback to students regarding their interaction and they score their students' notes. They have 7 – 10 days to score the student notes and send the results to the course instructor. They do not provide any grading feedback to the students. The advantage of standardized patients over actual patients for training medical students is that the medical students experience, and therefore their clinical notes, is based on a consistent clinical presentation.

The study case and scoring rubric, "Suzy Whitworth," were developed by the Association for Standardized Patient Educators and adapted by the Mid-Atlantic Consortium of Clinical Skills

Centers in June 2018, with additional formatting edits in January 2019. The standardized patients were trained on this case and its scoring rubric. The case contained 85 scorable elements that were expected to be present in the students' notes. Three scoring rubric examples are, "Notes chief complaint of shortness of breath (SOB; dyspnea; difficulty breathing; can't catch my breath)," "Notes sudden onset (acute; all of the sudden; all at once," and "Notes timing a few hours ago this morning; upon awakening; today)." The rubric combined the 85 scorable elements into 12 classes. ChatGPT and the standardized patients scored as either correct or incorrect each of the 85 elements in the deidentified students' notes. An error was either an incorrect answer or the absence of an answer. A reviewer checked the standardized patient scoring and the ChatGPT scoring and a second reviewer checked the first reviewer's scores.

ChatGPT is an LLM based on the generative pretrained transformer (GPT) artificial intelligence algorithm. It was pretrained on 45 terabytes of data and it consists of attention, which connects and weights natural language meanings, and an artificial neural network, which organizes and stores the meanings [13]. It accepts natural language input and provides natural language output. For each medical student and for each rubric, the researcher created a new prompt that asked ChatGPT if the rubric's meaning was contained in the student's free text note.

For ChatGPT and the standardized patients the measure of accuracy was the percent correct for each of the 12 categories and across the 12 categories. The Student's t-test compared the mean error rate across the 12 classes for ChatGPT with the mean error rate across the 12 classes for the standardized patients using the "R" language (R Project). (https://www.r-project.org/).

Results

The study population consisted of 168 first-year medical students, the case scoring rubric consisted of 85 elements, resulting in a total of 14,280 scores. There were four standardized patients, each working with one fourth of the students. The incorrect scoring (error) rates for the standardized

patients and ChatGPT are shown in Table 1.

Table 1. Incorrect scoring rates for ChatGPT and the standardized patients across free text note categories and across all categories.

	Scoring opportunities for the 168 students	Standardized patient error rate (%)	ChatGPT error rate (%)
Chief complaint	840	135 (16.1)	17 (2.0)
History of present illness	1,512	226 (14.9)	35 (2.3)
Review of systems	1,008	67 (6.6)	7 (0.7)
Past medical history	1,512	43 (2.8)	21 (1.4)
Physical exam	2,352	181 (7.7)	25 (1.1)
Diagnosis – pulmonary emboli	168	3 (1.8)	0 (0)
PE evidence	2,352	182 (7.7)	8 (0.3)
Diagnosis - pneumonia	168	0 (0)	0 (0)
Pneumonia evidence	1,848	66 (3.6)	4 (0.2)
Diagnosis - pneumothorax	168	0 (0)	7 (4.2)
Pneumothorax evidence	1,176	54 (4.6)	5 (0.4)
Diagnostic studies	1,008	66 (6.5)	16 (1.6)
Total[a]	14,280	1,023 (7.2)	145 (1.0)

[a]ChatGPT vs. standardized patient, $P = .002$

The category error rates for the standardized patients and the ChatGPTs were (respectively): chief complaint, 135, 17; history of present illness, 226, 35; review of systems, 67, 7; past medical history, 43, 21; physical examination, 181, 25; diagnosis #1, 3, 0; evidence for diagnosis #1, 182, 8; diagnosis #2, 0, 0; evidence for diagnosis #2, 66, 4; diagnosis #3, 0, 7; evidence for diagnosis #3, 54, 5; and diagnostic studies, 66, 16. The ChatGPT incorrect scoring rate was 1.0% and the standardized patient incorrect scoring rate was 7.2%. The ChatGPT error rate was 86% lower than the standardized patient error rate. The ChatGPT mean incorrect scoring rate of 12 (SD 11) was significantly lower than the standardized patient mean incorrect scoring rate of 85 (SD 74), $P = .002$.

Discussion

ChatGPT, a large language model, had a significantly lower error rate than that of standardized patients. This suggests that a LLM can be used to score medical students' notes.

Natural language processing programs have been used in several medical education settings. Medical education NLPs have been based on keywords, expert systems, statistical algorithms, and combinations of these approaches. Here we review these studies. DaSilva and Dennick [14] transcribed medical student problem-based verbal learning sessions and used an NLP program to count the frequency of technical words. Zhang et al. [15] implemented both a naïve Bayes approach and a supervised support vector machine method to assess resident performance evaluations. Their sentiment accuracies were 0.845 for naïve Bayes and 0.937 for the support vector machine. Spickard et al. [16] used an electronic scoring system to detect 25 core clinical problems in medical students' clinical notes. They achieved a 75% PPV on 16 of the 25 problems. Denny et al. [17] examined whether students mentioned advance directives and/or altered mental status in their clinical notes. For advance directives, their sensitivity was 69% and their PPV was 100% and, for mental status, their sensitivity was 100% and their PPV was 93%. Sarker et al. [18] used a semi-supervised NLP method to assess students' free text notes. Their accuracy over 21 cases and 105 notes was a sensitivity of 0.91 and a positive predictive value of 0.87. Two recent papers from the University of Michigan's Department of Surgery [19,20] assessed resident feedback and competency. Solano et al. [19] dichotomized the narrative surgical feedback given to residents into high and low quality and trained a logistic regression model to distinguish between them. Their model achieved a sensitivity of 0.37, a specificity of 0.97, and a receiver operating characteristic (ROC) of 0.86. Otles et al. [20] assessed narrative surgical resident feedback using a variety of statistical methods. The support vector machine algorithm achieved the best result with a maximum mean accuracy of 0.64. Abbott [21] studied whether an NLP program could assess the clinical competency committee ratings of residents in terms of the language that correlated in the 16 Accreditation Council for Graduate

Medical Education (ACGME) Milestones. The ROCs for the 16 milestones ranged from 0.71 to 0.95 and the mean ROC was 0.83. Neves et al. [22] examined the ability of RapidMiner Studio, a machine learning program, to the quality of attending feedback on resident performance. Their accuracies ranged from 74.4% to 82.2%.

If NLP programs are to be used to automate the grading of students' notes, they must achieve an acceptable accuracy. Sarker et al. [18] suggested that any method of scoring medical notes should achieve an accuracy close to 100%. Regrettably, none of the reported medical education NLPs achieved an acceptable accuracy. In our study, standardized patients also failed to achieve an acceptable accuracy. ChatGPT did attain an accuracy close to 100% and is, therefore, suitable for scoring students' free text notes.

A potential limitation of this study is that it has been suggested that GPT-based methods have the potential to generate unreliable answers under certain circumstances. We did not find that to be true in our study. Another potential limitation is that, although ChatGPT is free to the public, it does have resource requirements. It used 45TB of data, it has 175 billion parameters, and it runs on supercomputers residing in the cloud. This is a great deal of computing power for student notes. Fortunately, there are open source GPTs, for example, Meta's Llama, that can be run on a workstation. We would like to have examined the standardized patient validity literature but, to our knowledge, there has never been such a study. Finally, assessing note error does not directly address clinical reasoning.

An important advantage of an LLM is that it can be used to score student notes in real time and to provide students with immediate feedback regarding their clinical free text note performance. This affords students with an important learning opportunity because their interaction will be fresh in their mind. Another advantage is that the scoring is accurate and objective so students will no longer have to worry about human error and bias. A disadvantage of ChatGPT was that it was time consuming. Fortunately, there are now compound GPTs that can perform the entire assessment of all

the elements and all the students at one time. In terms of clinical reasoning, in the future we will be asking medical students, as part of their write-up of their clinical note, to provide their clinical reasoning and we can have a GPT assess the quality of their reasoning.

It should be noted that the use of an LLM to score clinical notes need not be limited to medical students. Indeed, it is expected that in the near future GPT-based artificial intelligence natural language programs will be applied to providing real time feedback to practicing physicians regarding their free text clinical notes.

In conclusion, ChatGPT demonstrated a significantly lower error rate than that of standardized patients. This is the first study to assess the ability of a GPT program to score medical students' standardized patient-based free text clinical notes. Generative pretrained transformer artificial intelligence programs represent an important advance in medical education and in the practice of medicine.

Declarations

Availability of data and materials
The dataset used in this research is not publicly available because it is of student scores, but it is available from the corresponding author on reasonable request.

Ethics approval and consent to participate
Ethical approval was waived as per section 46.104(d). (https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-A/section-46.104)

Competing interests
The author(s) declare no competing interests.

Funding
Support for this project was provided by the Patient Safety and Quality Academic Collaborative, a joint Defense Health Agency - Uniformed Services University program. Funder did not participate in the design, execution, or analysis of this project.

Author's contributions
HB, AH, JL, PH made substantial contributions to the conception and design of the work; HB, AH, JL, HK, MM, VK made substantial contributions to the acquisition, analysis, and interpretation of data; HB drafted the paper.

Acknowledgements
None

Authors' information
The opinions and assertions expressed in this paper are those of the authors and do not reflect the official policy or position of the Department of Defense, the Defense Health Agency, or the Uniformed Services University of the Health Sciences.

References

1. Burke HB, Hoang A, Becher D, Fontelo P, Liu F, Stephens M, Pangaro LN, Sessums LL, O'Malley P, Baxi NS, Bunt CW, Capaldi VF, Chen JM, Cooper BA, Djuric DA, Hodge JA, Kane S, Magee C, Makary ZR, Mallory RM, Miller T, Saperstein A, Servey J, Gimbel RW. QNOTE: An instrument for measuring the quality of EHR clinical notes. J Am Med Inform Assoc. 2014;21(5):910-916. [doi:10.1136/amiajnl-2013-002321] [PMID: 24384231]
2. Burke HB, Sessums LL, Hoang A, Becher DA, Fontelo P, Liu F, Stephens M, Pangaro LN, O'Malley PG, Baxi NS, Bunt CW, Capaldi VF, Chen JM, Cooper BA, Djuric DA, Hodge JA, Kane S, Magee C, Makary ZR, Mallory RM, Miller T, Saperstein A, Servey J, Gimbel RW. Electronic health records improve note quality. J Am Med Inform Assoc. 2015 Jan;22(1):199-205. [doi:10.1136/amiajnl-2014-002726] [PMID: 25342178]
3. ChatGPT. OpenAI. URL: https://openai.com/blog/chatGPT [accessed 2023-8-11]
4. Singhal K, et al. Towards expert-level medical question answering with large language models. 2023. URL: https://arxiv.org/abs/2305.09617 [doi:10.48550/arXiv.2305.09617]
5. Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nature Machine Intelligence. 2023 Jan;5(1):46-57. [doi:10.1038/s42256-022-00593-2]
6. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, Faix DJ, Goodman AM, Longhurst CA, Hogarth M, Smith DM. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med. 2023;183(6):589-596. [doi:10.1001/jamainternmed.2023.1838] [PMID: 37115527]
7. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. International Journal of Environmental Research and Public Health. 2023; 20(4):3378. [10.3390/ijerph20043378] [PMID: 36834073]
8. Olthof AW, Shouche P, Fennema EM, IJpma FFA, Koolstra RHC, Stirler VMA, van Ooijen PMA, Cornelissen LJ. Machine learning based natural language processing of radiology reports in orthopaedic trauma. Comput Methods Programs Biomed. 2021; Sep:208:106304. [doi:10.1016/j.cmpb.2021.106304] [PMID: 34333208]
9. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health. 2023; Mar;5(3):e107-e108. [doi:10.1016/S2589-7500(23)00021-3] [PMID: 36754724]
10. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. JAMA. 2023; Mar 14;329(10):842-844. [doi:10.1001/jama.2023.1044] [PMID: 36735264]
11. Burkhardt HA, Ding X, Kerbrat A, Comtois KA, Cohen T. From benchmark to bedside: transfer learning from social media to patient-provider text messages for suicide risk prediction. J Am Med Inform Assoc. 2023 May 19;30(6):1068-1078. [doi:10.1093/jamia/ocad062] [PMID: 37043748]
12. Morjaria L, Burns L, Bracken K, Ngo QN, Lee M, Levinson AJ, Smith J, Thompson P, Sibbald M. Examining the Threat of ChatGPT to the Validity of Short Answer Assessments in an

Undergraduate Medical Program. J Med Educ Curric Dev. 2023 Sep 28;10:23821205231204178. [doi:10.1177/23821205231204178] [PMID: 37780034]

13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention Is All You Need. June 12, 2017. [doi:10.48550/arXiv.1706.03762] URL:https://arxiv.org/abs/1706.03762
14. Da Silva AL, Dennick R. Corpus analysis of problem-based learning transcripts: an exploratory study. Med Educ. 2010 Mar;44(3):280-8. [doi:10.1111/j.1365-2923.2009.03575.x] [PMID: 20444059]
15. Zhang R, Pakhomov S, Gladding S, Aylward M, Borman-Shoap E, Melton GB. Automated assessment of medical training evaluation text. AMIA Annu Symp Proc. 2012 Nov 3, 2012;2012:1459-68. [PMID: 23304426]
16. Spickard A 3rd, Ridinger H, Wrenn J, O'brien N, Shpigel A, Wolf M, Stein G, Denny J. Automatic scoring of medical students' clinical notes to monitor learning in the workplace. Med Teach. 2014 Jan;36(1):68-72. [doi:10.3109/0142159X.2013.849801] [PMID: 23304426]
17. Denny JC, Spickard 3rd A, Speltz PJ, Porier R, Rosenstiel DE, Powers JS. Using natural language processing to provide personalized learning opportunities from trainee clinical notes. J Biomed Inform. 2015 Aug; 56:292-9. [doi:10.1016/j.jbi.2015.06.004] [PMID: 26070431]
18. Sarker A, Klein AZ, Mee J, Harik P, Gonzalez-Hernandez G. An interpretable natural language processing system for written medical examination assessment. J Biomed Inform. 2019 Oct; 98:103268. [doi:10.1016/j.jbi.2019.103268] [PMID: 31421211]
19. Solano QP, Hayward L, Chopra Z, Quanstrom K, Kendrick D, Abbott KL, Kunzmann M, Ahle S, Schuller M, Ötleş E, George BC. Natural Language Processing and Assessment of Resident Feedback Quality. J Surg Educ. 2021 Nov-Dec;78(6):e72-e77. [doi:10.1016/j.jsurg.2021.05.012] [PMID: 34167908]
20. Ötleş E, Kendrick DE, Solano QP, Schuller M, Ahle SL, Eskender MH, Carnes E, George BC. Using Natural Language Processing to Automatically Assess Feedback Quality: Findings From 3 Surgical Residencies. Acad Med. 2021 Oct 1;96(10):1457-1460. [doi:10.1097/ACM.0000000000004153] [PMID: 33951682]
21. Abbott KL, George BC, Sandhu G, Harbaugh CM, Gauger PG, Ötleş E, Matusko N, Vu JV. Natural Language Processing to Estimate Clinical Competency Committee Ratings. J Surg Educ. 2021 Nov-Dec;78(6):2046-2051. [doi:10.1016/j.jsurg.2021.06.013] [PMID: 34266789]
22. Neves SE, Chen MJ, Ku CM, Karan S, DiLorenzo AN, Schell RM, Lee DE, Diachun CAB, Jones SB, Mitchell JD. Using Machine Learning to Evaluate Attending Feedback on Resident Performance. Anesth Analg. 2021 Feb 1;132(2):545-555. [doi:10.1213/ANE.0000000000005265] [PMID: 33323789]