Comparing Machine Learning Models and Human Raters When Ranking Medical Student Performance Evaluations

Jonathan Kibble, PhD
Jeffrey Plochocki, PhD

ABSTRACT

Background The Medical Student Performance Evaluation (MSPE), a narrative summary of each student's academic and professional performance in US medical school is long, making it challenging for residency programs evaluating large numbers of applicants.

Objective To create a rubric to assess MSPE narratives and to compare the ability of 3 commercially available machine learning models (MLMs) to rank MSPEs in order of positivity.

Methods Thirty out of a possible 120 MSPEs from the University of Central Florida class of 2020 were de-identified and subjected to manual scoring and ranking by a pair of faculty members using a new rubric based on the Accreditation Council for Graduate Medical Education competencies, and to global sentiment analysis by the MLMs. Correlation analysis was used to assess reliability and agreement between student rank orders produced by faculty and MLMs.

Results The intraclass correlation coefficient used to assess faculty interrater reliability was 0.864 ($P<.001$; 95% CI 0.715-0.935) for total rubric scores and ranged from 0.402 to 0.768 for isolated subscales; faculty rank orders were also highly correlated ($r_s=0.758$; $P<.001$; 95% CI 0.539-0.881). The authors report good feasibility as the rubric was easy to use and added minimal time to reading MSPEs. The MLMs correctly reported a positive sentiment for all 30 MSPE narratives, but their rank orders produced no significant correlations between different MLMs, or when compared with faculty rankings.

Conclusions The rubric for manual grading provided reliable overall scoring and ranking of MSPEs. The MLMs accurately detected positive sentiment in the MSPEs but were unable to provide reliable rank ordering.

Introduction

The Medical Student Performance Evaluation (MSPE) is a narrative summary of the professional and academic performance of medical students in the United States at the end of their core clinical clerkships.[1] The MSPE is generated by medical schools that are being evaluated based partly on the graduation outcomes of their students,[2] and as such the process has faced criticisms in trying to balance transparency about a trainee's weaknesses with student advocacy.[3] Another challenge is making the letter usable when processing hundreds of residency applications. The need for efficiency argues for standardization, but this risks homogeneity in the letters. Despite these caveats, the MSPE is cited by most program directors as very important when determining whom to select for residency.[4]

Reliable assessment of textual data such as the MSPE is difficult and time-consuming, a problem well known by educators evaluating narrative data.[5] Machine learning models (MLMs) are a resource that may be used to improve the reliability and efficiency of MSPE assessment. MLMs work by using computational models to analyze unstructured textual data and identify sentiment (attitude, opinion, emotion).[6-8] In such models, strings of words are represented as sets of numbers that serve as input for trained algorithms that parse their syntactic and semantic meaning to assign a sentiment to the text. The output of the model is a score that reflects the confidence of the sentiment classification derived from how well it comports with the data used to train the algorithm.

The aims of this project were: (1) to develop a rubric for faculty to assess MSPE narrative comments more objectively and obtain preliminary validity evidence for its use; and (2) to compare faculty rubric scores with MLM outputs for rapidly screening and ranking MSPEs.

Methods

Setting and Participants

The University of Central Florida (UCF) matriculates 120 medical students per year. The medical curriculum consists of 2 years of basic science education followed by 2 years of clinical training. This study was undertaken in 2022 using MSPEs from the graduated class of 2020, selected because they were based on performance in core clerkships completed prior to any disruptions from the COVID-19 pandemic.

DOI: http://dx.doi.org/10.4300/JGME-D-22-00678.1

Interventions

Thirty out of 120 MSPEs were selected, based on a sample size calculation that assumed a correlation coefficient of 0.5 between 2 raters, with the probabilities of type I and II errors assumed to be 0.05 and 0.2, respectively. We further applied a purposeful sampling approach that ensured we sampled proportionally across each of our class rank categories (good, very good, excellent, outstanding), which are determined by a faculty committee. Our MSPE included verbatim comments written by attending and resident physicians from every core clerkship, plus clerkship director summaries. For each student, comments were extracted, concatenated into a single narrative, and saved as a de-identified plain text file for analysis.

KEY POINTS

What Is Known
Medical Student Performance Evaluations (MSPEs) can be lengthy, and program directors would likely appreciate novel ways to review these more efficiently.

What Is New
The authors created 3 machine learning models (MLMs)—a type of artificial intelligence (AI)—to process and rank MSPEs, and a rubric for human raters to use in ranking. MLMs did not correlate with human rankings, but the rubric for manual rating showed promise.

Bottom Line
As we begin to contemplate the use of AI in residency application review, this article provides early data in how it performs compared to a manual process. Readers may consider the manual rubric provided for their own program review.

Human Rater (Faculty) Rubric

The rubric (TABLE 1) was created by the authors and based on the 6 Accreditation Council for Graduate Medical Education (ACGME) competencies: Medical Knowledge, Patient Care and Procedural Skills, Interpersonal and Communication Skills, Professionalism, Practice-Based Learning and Improvement (PBLI), and Systems-Based Practice (SBP).[9] The ACGME framework was selected for its direct relevance to assessment of residents and because our MD program learning objectives are organized around these competencies. Subcompetencies were used as descriptive anchors within each competency to guide raters as to the characteristics of student performance in each category.[9] For the first 4 competencies, we arbitrarily selected a 10-point score scale, but we used a 5-point scale for PBLI and SBP to reflect the fact that we have fewer learning objectives in these areas for third-year medical students. The authors formed a rater pair to pilot test the rubric by independently scoring 30 MSPE narratives. A subscore was assigned to each competency, and these values were summed to produce a total score out of 50 points. We then assigned a global rating score, also out of 50 points, to determine if a simple overall impression would have value.

Machine Learning Models

Three MLMs were selected: MonkeyLearn (MonkeyLearn Inc, Montevideo, Uruguay), Lexalytics (Lexalytics, Amherst, MA), and MeaningCloud (MeaningCloud LLC, New York, NY). These programs used standard English lexicons and provided not only a sentiment polarity (positive, neutral, negative), but also a numeric value for degree of positivity or confidence. Lexalytics reports a sentiment score on a scale of -1 to +1; MeaningCloud reports a percent of confidence in the sentiment determination; and MonkeyLearn reports an overall percent of positive sentiment. The selected MLMs were also accessible on a low budget, since they offered free trial use for small projects such as this one. Although narrative data can be manually entered into the MLM interfaces, the Information Technology Department at UCF worked for 2 months to develop an in-house application programming interface costing $4,300 to allow batch upload of the narratives.

MLMs, including those used in this study, work by modeling text as numerical data. The text is converted to numerical data by representing each word as a set of numbers, called a vector, that can be plotted in multidimensional space. The semantic meaning of each word in the model is determined by its relationship to the other words in this space. The smaller the distance between 2 vectors, the greater the semantic similarity of the words they represent. The context of each word can by modeled by making the values of the vectors dependent on the words that precede and follow it in a sentence. This allows the model to determine that, for example, the word "bad" has a negative sentiment, but when used in the phrase "not too bad" it has a more positive sentiment.

The main limitation of MLMs is their dependency on their training data, which is typically obtained from online product reviews. The training data is used to set the values of the vectors in the MLM that represent each word. When the data from the MSPEs is entered into the MLM, it assigns the vectors derived from the training data to the words in the text and uses them to determine the sentiment with an associated confidence score. The confidence score is proportional to the distances of the vectors to a particular sentiment. If word usage in the training data does not align well with word usage in the MSPE text, the sentiment and confidence score will be inaccurate. Similarly, if word usage in the training

TABLE 1
Scoring Rubric for Faculty Assessment of MSPE Narratives

Competency (Score Scale); [Observed Frequency of Comments in the Sample, %]	Rater 1 Median Score (IQR)	Rater 2 Median Score (IQR)	ICC Single Rater (95% CI) *P* value	ICC Average Score (95% CI) *P* value
Medical Knowledge (score out of 10); [12%] ▪ Good fund of knowledge; investigative and analytical approach ▪ Applies knowledge to clinical situations and solves problems ▪ Able to teach others	9 (2)	7 (2)	0.624[a] (0.345-0.801) *P*<.001	0.768[a] (0.513-0.890) *P*<.001
Patient Care and Procedural Skills (score out of 10); [19%] ▪ Gathers essential and accurate information about the patient ▪ Counsels patients and family members ▪ Proposes good differential diagnoses and contributes to assessment and plan ▪ Performs essential medical procedures at third-year medical student level	9 (1)	8 (2)	0.585[a] (0.290-0.778) *P*<.001	0.738[a] (0.449-0.875) *P*<.001
Interpersonal and Communication Skills (score out of 10); [21%] ▪ Creates and sustains a therapeutic relationship with patients and families ▪ Works effectively as a member of a health care team	9 (1)	8 (2)	0.535[a] (0.222-0.748) *P*<.001	0.697[a] (0.363-0.856) *P*<.001
Professionalism (score out of 10); [35%] ▪ Demonstrates professional conduct and accountability; reliable, dependable, strong work ethic ▪ Demonstrates humanism and cultural proficiency ▪ Maintains personal emotional and physical well-being, seeks personal improvement	9 (0)	9 (1)	0.286 (-0.077-0.582) *P*=.06	0.444 (-0.167-0.736) *P*=.06
Practice-Based Learning and Improvement (score out of 5); [10%] ▪ Uses scientific evidence ▪ Seeks to improve the practice of medicine ▪ Seeks and responds to feedback well	4 (2)	4 (1)	0.252 (-0.114-0.557) *P*=.09	0.402 (-0.256-0.715) *P*=.09
Systems-Based Practice (score out of 5); [3%] ▪ Works effectively in various health care delivery settings ▪ Helps to coordinate patient care, advocates for patient ▪ Works well in interprofessional teams	3 (2)	3 (0)	0.400[a] (0.052-0.661) *P*=.013	0.571[a] (0.099-0.796) *P*=.013
Faculty summed total score (out of 50)	41 (9)	39 (7)	0.761[a] (0.556-0.879) *P*<.001	0.864[a] (0.715-0.935) *P*<.001
Faculty global rating score (out of 50)	45 (9)	42 (7)	0.726[a] (0.500-0.859) *P*<.001	0.841[a] (0.666-0.924) *P*<.001

[a] Denotes significant ICC value (*F* test, n=30).
Abbreviations: MSPE, Medical Student Performance Evaluation; IQR, interquartile range; ICC, intraclass correlation coefficient.

TABLE 2
Correlation Between Machine Learning Models and Human Raters

	Faculty Rater 1	Faculty Rater 2	Lexalytics	MonkeyLearn	MeaningCloud
Faculty Rater 1	1				
Faculty Rater 2	r_s=0.758[a] (95% CI 0.539-0.881) *P*<.001	1			
Lexalytics	r_s=0.264 (95% CI -0.117-0.578) *P*=.16	r_s=0.155 (95% CI -0.228-0.497) *P*=.41	1		
MonkeyLearn	r_s=-0.005 (95% CI -0.374-0.366) *P*=.98	r_s=-0.022 (95% CI -0.388-0.351) *P*=.91	r_s=0.010 (95% CI -0.362-0.378) *P*=.96	1	
MeaningCloud	r_s=-0.332 (95% CI -0.625-0.043) *P*=.07	r_s=-0.275 (95% CI -0.586-0.105) *P*=.14	r_s=-0.137 (95% CI -0.482-0.246) *P*=.47	r_s=0.141 (95% CI -0.242-0.486) *P*=.46	1

[a] Denotes significant correlation between rank orders generated by each scorer/program.
Abbreviation: r_s, Spearman's rank correlation coefficient.

data differs greatly among the MLMs, their computed sentiment scores may vary greatly.

Outcomes Measured

The performance of the new hand-scored rubric was assessed using correlation of the faculty scores for each competency, for total scores and for a global rating score. The ability of MLMs to assess MSPEs was determined by using raw sentiment/confidence scores to rank the students in the sample from 1 to 30 then by assessing correlations between MLMs, and by comparison with faculty ratings.

Analysis of Outcomes

Faculty rating scores were presented as medians and interquartile ranges. Interrater reliability for the pair of raters was assessed using intraclass correlation coefficients (ICCs). To analyze 2 specific raters, a 2-way mixed-effects ICC model was selected, prioritizing consistency.[10] ICC estimates and 95% confidence intervals are reported for both single rater reliability and for the average of both raters; significance difference of ICC estimates from zero was assessed using *F* tests. The scores generated by MLMs were continuous variables and are presented as means and standard deviations (SDs). To compare human and machine raters, the summed total scores for human raters and the positivity/confidence scores for MLMs were converted into rank orders, and correlations were computed using Spearman rank correlation. The SPSS statistical package version 28 (SPSS Inc, Chicago, IL) was used for all statistical tests.

This project was exempted by the UCF Institutional Review Board as secondary research that used deidentified data and did not require participant consent.

Results

The tenor of MSPE narratives was highly positive, with median total scores out of 41 and 39 out of 50 for the 2 faculty raters. TABLE 1 shows that ICCs for faculty interrater reliability were significant for the overall rubric scores, global rating scores, and for the subscales of Medical Knowledge, Patient Care and Procedural Skills, Interpersonal and Communication Skills, and SBP, but did not reach significance for the Professionalism or PBLI subscales.

All 3 MLMs classified every MSPE as having a "positive" overall sentiment. Lexalytics scores ranged from +0.12 to +0.57 arbitrary units, with a mean (SD) of 0.49 (0.09); the positivity reported by MeaningCloud ranged from 93% to 100%, with a mean (SD) of 99% (2); MeaningCloud reported a range of confidence in the positivity determination between 76% and 92% with a mean (SD) of 87% (3). TABLE 2 shows the results of a correlation matrix for the MSPE rank orders (first=highest score, 30th=lowest scores). The rank orders from the faculty were significantly correlated, whereas there were no significant correlations between faculty ratings and MLM scores, and no significant correlations between any of the MLM rank orders.

Discussion

The important findings of the study were that (1) a novel rubric performed well for objectively assessing the overall strength of an MSPE narrative, and (2) commercially available MLMs were unable to reliably screen and rank MSPE narratives.

To our knowledge there are no other published rubrics for the purpose of assessing an MSPE narrative. Our pilot data was promising in that our faculty pair produced acceptable ICCs for total scores and high correlation between their rank orders. We found the rubric straightforward to use alongside the narrative assessments. There are tensions in using a scoring rubric to summarize and rank MSPEs, which should be acknowledged. Most fundamentally, the purpose of developing detailed narrative assessments of a student's clinical performance is in part to get away from reducing complex human skills, attitudes, and behaviors to simple numbers.[11] Similarly, these narratives are often developed for the purpose of providing feedback to the learner,[12] rather than for competitive selection for residency. Finally, the importance of holistic review, which assesses applicants in light of their unique backgrounds and is a key part of addressing inequity in the physician workforce,[13] should be acknowledged. However, we contend that a rubric to numerically summarize the long MSPE narrative could lend more objectivity to the selection process without detracting from holistic review. In the end, it is a reality that residency programs must reduce the applicant list by more than 90% when selecting interviewees.[4] Therefore, tools to help in this selection process may have utility.

The second goal of the study was to determine if MLMs could reliably rank students along a gradient of sentiment polarity. While the tools we selected did report an accurate outcome of 100% positivity in the MSPE sample, they were unable to work with the range restriction of narratives that were, for the most part, describing various shades of excellence. The absence of frankly negative statements supports concerns that MSPEs may exaggerate the quality of graduates.[14] For MLMs untrained on MSPE data, it will be necessary to develop a bespoke reference set to parse out the outstanding from the merely excellent. In a pilot study using student perception of instruction, where one might expect more common language phrases to be used, we were able to show significant correlations between the output of the MLMs and faculty raters,[15] establishing the idea that MLMs can be developed to support applications in medical education.

Limitations of this study include the use of a single pair of nonclinician raters and the use of a limited dataset from a single institution. Other institutions may use different methods for constructing MSPEs, so the approach described here may not be generalizable.[16] Field testing during actual residency selection will be needed to provide additional validity evidence for the instrument. We used only a subset of commercially available MLMs that provided a numerical output, which should not rule out the likelihood that such tools will have utility in the future.

Conclusions

The new rubric for hand grading MSPEs was found to be a reliable tool as indicated by a high interobserver agreement for total scores or student rank ordering. However, machine learning algorithms as utilized in this study could not undertake automated sentiment analysis of MSPEs.

References

1. Association of American Medical Colleges. Recommendations for revising the Medical Student Performance Evaluation (MSPE). Published May 2017. Accessed September 2, 2022. https://www.aamc.org/download/470400/data/mspe-recommendations.pdf
2. Liaison Committee on Medical Education. Data collection instrument for full accreditation surveys. Accessed September 2, 2022. https://lcme.org/publications/
3. Hauer K, Giang D, Kapp M, Sterling R. Standardization in the MSPE: key tensions for learners, schools, and residency programs. *Acad Med*. 2021;96(1):44-49. doi:10.1097/ACM.0000000000003290
4. National Resident Matching Program. Results of the 2021 NRMP Program Director Survey. Published August 2021. Accessed September 2, 2022. https://www.nrmp.org/wp-content/uploads/2021/11/2021-PD-Survey-Report-for-WWW.pdf
5. Bird JB, Friedman KA, Arayssi T, Olvet DM, Conigliaro RL, Brenner JM. Review of the Medical Student Performance Evaluation: analysis of the end-users' perspective across the specialties. *Med Educ Online*. 2021;26(1):1876315. doi:10.1080/10872981.2021.1876315
6. Lin Q, Zhu Y, Zhang S, Shi P, Guo Q, Niu Z. Lexical based automated teaching evaluation via 339 students' short reviews. *Comp Appl Eng Educ*. 2019;27(1):194-205. doi:10.1002/cae.22068
7. Do HH, Prasad PW, Maag A, Alsadoon A. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert Sys Appl*. 2019;118(6):272-299. doi:10.1016/j.eswa.2018.10.003
8. D'Andrea A, Ferri F, Grifoni P, Guzzo T. Approaches, tools and applications for sentiment analysis implementation. *Int J Comp Appl*. 2015;125(3):26-33. doi:10.5120/ijca2015905866
9. Exploring the ACGME Core Competencies. *New England Journal of Medicine Knowledge+*. Published June 2, 2016. Accessed June 27, 2022. https://knowledgeplus.nejm.org/blog/exploring-acgme-core-competencies/

10. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012
11. Schuwirth LWT, van der Vleuten CPM. A history of assessment in medical education. *Adv in Health Sci Educ Theory Pract*. 2020;25(5):1045-1056. doi:10.1007/s10459-020-10003-0
12. Schuwirth LW, van der Vleuten CP. Programmatic assessment: from assessment of learning to assessment for learning. *Med Teach*. 2011;33(6):478-485. doi:10.3109/0142159X.2011.565828
13. Association of American Medical Colleges. Holistic review. Accessed June 27, 2022. https://www.aamc.org/initiatives/holisticreview
14. Puscas L. Viewpoint from a program director they can't all walk on water. *J Grad Med Educ*. 2016;8(3):314-316. doi:10.4300/JGME-D-16-00237.1
15. Plochocki J, Kibble J. Sentiment analysis machine learning model congruence: a case study using neuroscience module evaluations. *FASEB J*. 2022;36(suppl 1). doi:10.1096/fasebj.2022.36.S1.R3256
16. Buchanan AO, Strano-Paul L, Saudek K, et al. Preparing effective narrative evaluations for the Medical School Performance Evaluation (MSPE). *MedEdPORTAL*. 2022;18:11277. doi:10.15766/mep_2374-8265.11277

Both authors are with University of Central Florida College of Medicine. **Jonathan Kibble, PhD,** is Professor of Medical Education; and **Jeffrey Plochocki, PhD,** is Associate Professor of Medical Education.

Funding: This study was funded by a grant from the Department of Medical Education, University of Central Florida, College of Medicine.

Conflict of interest: The authors declare they have no competing interests.

Preliminary data were presented at the Association of American Medical Colleges Annual Meeting of the Group on Student Affairs, April 7-9, 2022, Denver, CO.

The authors would like to thank Andres Calvete and Micah Marshall from the Department of Health Information Technology, the University of Central Florida, College of Medicine, for development of an interface to batch upload and download data to and from Machine Learning Models.

Corresponding author: Jonathan Kibble, PhD, University of Central Florida College of Medicine, jonathan.kibble@ucf.edu

Received September 12, 2022; revision received May 10, 2023; accepted June 1, 2023.