Natural Language Processing to Estimate Clinical Competency Committee Ratings

Kenneth L. Abbott, MD, MS,[**] Brian C. George, MD, MAEd,[†] Gurjit Sandhu, PhD,[†] Calista M. Harbaugh, MD, MS,[†] Paul G. Gauger, MD,[†] Erkin Ötleş, MEng,[**] Niki Matusko, BS,[†] and Joceline V. Vu, MD[†]

[**]University of Michigan Medical School, Ann Arbor, Michigan; and [†]Department of Surgery, University of Michigan, Ann Arbor, Michigan

OBJECTIVE: Residency program faculty participate in clinical competency committee (CCC) meetings, which are designed to evaluate residents' performance and aid in the development of individualized learning plans. In preparation for the CCC meetings, faculty members synthesize performance information from a variety of sources. Natural language processing (NLP), a form of artificial intelligence, might facilitate these complex holistic reviews. However, there is little research involving the application of this technology to resident performance assessments. With this study, we examine whether NLP can be used to estimate CCC ratings.

DESIGN: We analyzed end-of-rotation assessments and CCC assessments for all surgical residents who trained at one institution between 2014 and 2018. We created models of end-of-rotation assessment ratings and text to predict dichotomized CCC assessment ratings for 16 Accreditation Council for Graduate Medical Education (ACGME) Milestones. We compared the performance of models with and without predictors derived from NLP of end-of-rotation assessment text.

RESULTS: We analyzed 594 end-of-rotation assessments and 97 CCC assessments for 24 general surgery residents. The mean (standard deviation) for area under the receiver operating characteristic curve (AUC) was 0.84 (0.05) for models with only non-NLP predictors, 0.83 (0.06) for models with only NLP predictors, and 0.87 (0.05) for models with both NLP and non-NLP predictors.

CONCLUSIONS: NLP can identify language correlated with specific ACGME Milestone ratings. In preparation for CCC meetings, faculty could use information automatically extracted from text to focus attention on residents who might benefit from additional support and guide the development of educational interventions. (J Surg Ed 78:2046−2051. © 2021 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: Natural language processing, clinical competency committee, resident, assessment, evaluation

COMPETENCIES: Patient Care, Medical Knowledge, Systems-Based Practice, Practice-Based Learning And Improvement, Professionalism, Interpersonal And Communication Skills

INTRODUCTION

Residency programs use a system of assessments to track trainee progress and development. For example, a subset of faculty members participates in clinical competency committee (CCC) meetings, which occur every six months and are designed to evaluate performance and aid in the development of individualized learning plans and interventions.[1] In preparation for the CCC meetings, committee members synthesize performance information from a variety of sources—some formal (e.g., monthly end-of-rotation assessments) and some informal (e.g., conversations).

Artificial intelligence could support the CCC faculty performing these complex holistic reviews by guiding their attention to residents who may benefit from additional support. Natural language processing (NLP) is a form of artificial intelligence that interprets complex human language.[2] In general surgery, Milestones are used to structure CCC meeting discussion and resident assessment.[3,4] It is unknown whether NLP can identify language correlated with specific Accreditation Council

Correspondence: Inquiries to Joceline V. Vu MD, Department of Surgery, University of Michigan, Ann Arbor, MI; 48109. Phone: (703) 200-8623; e-mail: vuj@med.umich.edu

  1931-7204/$30.00 https://doi.org/10.1016/j.jsurg.2021.06.013

for Graduate Medical Education (ACGME) Milestone ratings, but this could help faculty identify residents who may need additional support in a specific performance domain. For example, faculty could review predictions of Milestone ratings, gather additional information about residents who are predicted to have low Milestone ratings, and spend additional CCC meeting time discussing these residents.

With this study, we examine whether NLP can be used to estimate CCC Milestone ratings, using text from end-of-rotation assessments.

METHODS

Data

We collected deidentified performance assessments for surgical residents who trained at one institution between 2014 and 2018. No residents were excluded. Assessments included monthly end-of-rotation assessments gathered via an online assessment system (MedHub, https://www.medhub.com/) and biannual CCC assessments. End-of-rotation assessments included nine numeric items with anchors that were generally related to the ACGME general surgery Milestones,[3,4] and asked faculty to rate trainees along multiple dimensions, using a 9-point Likert scale, with ratings of 1-3 corresponding with unsatisfactory performance, ratings of 4-6 corresponding with satisfactory performance, and ratings of 7-9 corresponding with superior performance. End-of-rotation clinical assessments also included a tenth numeric item that asked faculty to rate a trainee's overall clinical competence, and one text field for general comments. The CCC assessments included a numeric scale for each of the 16 Milestones grouped within 6 competencies (*patient care*, *medical knowledge*, *systems-based practice, practice-based learning and improvement*, *professionalism,* and *interpersonal and communication skills*) and 8 domains (*care for diseases and conditions*, *coordination of care*, *performance of operations and procedures*, *self-directed learning*, *teaching*, *improvement of care*, *maintenance of physical and emotional health*, and *performance of administrative tasks*). On the CCC assessment scale, which was used for all post-graduate years (PGYs), ratings ranged from 1-8, with a rating of 1 corresponding with a Milestone rating of *critical deficiency*, a rating of 2 corresponding with a Milestone rating of Level 1 (demonstrating Milestones expected of an incoming resident), a rating of 4 corresponding with a Milestone rating of Level 2 (demonstrating additional Milestones, but not yet at mid-residency level), a rating of 6 corresponding with a Milestone rating of Level 3 (demonstrating a majority of Milestones), and a rating of 8 corresponding with a Milestone rating of Level 4 (substantially demonstrates Milestones targeted for residency). CCC assessments also included a text field for comments for each Milestone.

Analysis

Figure 1 summarizes our analytic process. First, we aggregated manually deidentified text from all the end-of-rotation assessments (not CCC assessments) delivered during each CCC assessment period. Since we aimed to detect low performance, we dichotomized CCC ratings into high (≥7, above Milestone Level 3) and low (<7, at or below Milestone Level 3) ratings.

Next, we used the *googleLanguageR* package[5] to connect to Google Cloud Natural Language[6] and complete sentiment analysis of text comments from end-of-rotation assessments. Sentiment analysis is a type of NLP whose demand has been driven by electronic commerce and other industries that wish to interpret large amounts of qualitative

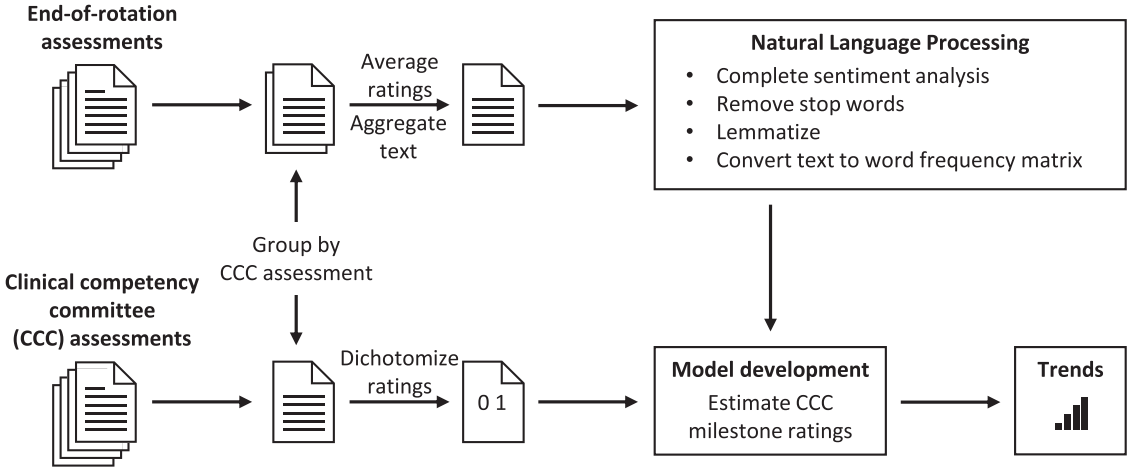

FIGURE 1. Summary of analytic process.

data, such as social media comments, product reviews, or restaurant reviews.[7] Sentiment analysis can extract information related to opinion and translate it into quantitative data, such as positive or negative numeric values for specific words; for example, in the phrase "excellent performance," the noun *performance* has positive sentiment, because *excellent* is positive, and the adjective *excellent* describes the noun *performance*. By contrast, in the phrase "terrible performance," the same noun *performance* has negative sentiment, because *terrible* is negative. Google's NLP software produces numeric scores between -1 and 1, in intervals of 0.1.

Then, we used the *tidytext* and *textstem* packages[8,9] to create a frequency matrix of words extracted from text comments. For example, a comment consisting only of "solid performance" would yield a 1 in the column for the word *solid*, a 1 in the column for the word *performance*, and 0 in all columns for other words. In creating this word frequency matrix, we discarded stop words, which are extremely common words of little value in NLP (e.g., *you*, *I*, *the*, *to*, *a*),[2] and used lemmatization, which is a means of identifying variants of the same word;[2] for example, singular *resident* and plural *residents* were both mapped to *resident*.

Next, we used h2o.ai's Driverless AI[10] to estimate the probability of dichotomized CCC assessment ratings. This software automatically engineers composite variables and evaluates thousands of possible predictive models, which may involve a variety of machine learning algorithms, and then creates an ensemble of predictive models that yield the best performance. We created 48 models: 16 models with only non-NLP predictors (i.e., only predictors not derived from NLP of end-of-rotation assessment text, including PGY and mean ratings for each of the 10 domains on end-of-rotation assessments), 16 models with only NLP predictors, and 16 models with all predictors. Outcome variables included each of the 16 numeric ratings on CCC assessments. NLP predictors included Google sentiment score for text comments from aggregated end-of-rotation assessments (not CCC assessments) and the above-described word frequency matrix. We evaluated the performance of each of these models with 3-fold cross validation, which involves splitting datasets of limited size into training and testing subsets (training datasets that incorporate all variables are used to create models, and testing datasets that incorporate all variables except the outcome variable are used to evaluate models). We used predictions from cross validation to calculate area under the receiver operating characteristic curve (AUC), a standard model performance metric given by calculating the definite integral of the curve created by plotting a model's false positive rate against its true positive rate.

We used *R* version 4.0.0[11] to aggregate and analyze all assessment data.

IRB Statement

This study was exempt from review by the University of Michigan Institutional Review Board (IRB).

RESULTS

We analyzed 594 end-of-rotation assessments and 97 clinical competency assessments for 24 general surgery residents (Table 1). NLP of end-of-rotation text yielded 1,930 words, each of which served as a predictor variable. CCC assessment ratings varied by Milestone, with the prevalence of low ratings <7 ranging from 0.23 to 0.57 (Table 2); prevalence of low ratings was greatest for *performance of operations and procedures* under *patient care* and *performance of assignments and administrative tasks* under *professionalism*. Across all models, sensitivity for detection of low ratings ranged from 0.28 to 0.89; accordingly, AUCs ranged from 0.71 to 0.96 (Table 2). AUCs were comparable for models with NLP predictors, non-NLP predictors, and all predictors.

DISCUSSION

We are aware of no previous research applying NLP to the ACGME Milestone rating process. In this study, we used NLP of end-of-rotation assessments to examine whether NLP could identify language correlated with specific Milestone ratings. We found that NLP could be used to estimate dichotomized Milestone ratings on biannual CCC assessments. Information automatically extracted from text could help faculty focus attention on residents who might benefit from additional support.

Many prior studies have applied NLP to analysis of medical records,[12] but little research applies NLP to medical education. A recent review found only a handful of studies of NLP in medical education,[13] and only one of these involved performance assessments. That study classified text into 6 ACGME competencies,[14] but did not relate narrative data to ACGME Milestone ratings.[3,4] We found that NLP can be used to estimate dichotomized Milestone ratings. This extends prior research into NLP in graduate medical education.

Faculty could use NLP to help prepare for CCC meetings. For example, automated analyses of numeric ratings and text comments could be used to predict the probability of a low Milestone rating (likely higher during early PGYs if the same CCC rating scales are used across PGYs) or recommend a numeric Milestone rating. The scope of these analyses might include certain Milestones of interest, Milestones grouped according to competency or domain, or all Milestones. Before a CCC meeting, faculty could gather additional information about residents identified by these

TABLE 1. Comparison of Sample Characteristics and Clinical Competency Committee Assessment Ratings Across Post-Graduate Years.

Variable	Post-graduate year (PGY)					p*
	PGY-1	PGY-2	PGY-3	PGY-4	PGY-5	
Assessments (n)	1	3	9	35	49	
Gender = female (%)	0 (0)	0 (0)	3 (33.3)	10 (28.6)	12 (24.5)	0.765
Ethnicity = non-white (%)	0 (0)	2 (66.7)	4 (44.4)	13 (37.1)	19 (38.8)	0.626
Clinical competency committee rating mean (SD)						
Patient care						
1. Care for diseases and conditions	4 (NA)	5.33 (1.15)	6 (0)	7.6 (0.81)	7.92 (0.40)	<0.001
2. Care for diseases and conditions	4 (NA)	5.33 (1.15)	6 (0)	7.54 (0.85)	7.8 (0.61)	<0.001
3. Performance of operations and procedures	4 (NA)	4.67 (1.15)	5.78 (0.67)	6.51 (1.01)	7.27 (1.06)	<0.001
Medical knowledge						
1. Care for diseases and conditions	6 (NA)	5.33 (1.15)	5.33 (1.00)	6.97 (1.12)	7.35 (1.11)	<0.001
2. Performance of operations and procedures	4 (NA)	6 (0)	5.78 (0.67)	6.86 (1.00)	7.55 (0.84)	<0.001
Systems-based practice						
1. Coordination of care	6 (NA)	5.33 (1.15)	6.22 (0.67)	7.66 (0.76)	7.71 (0.71)	<0.001
2. Improvement of care	6 (NA)	6 (2.00)	5.78 (1.56)	7.03 (1.22)	7.43 (0.91)	0.001
Practice-based learning and improvement						
1. Teaching	6 (NA)	6 (0)	5.89 (2.03)	7.6 (0.81)	7.63 (0.78)	<0.001
2. Self-directed learning	6 (NA)	6 (0)	5.56 (1.33)	6.97 (1.40)	7.31 (1.19)	0.002
3. Improvement of care	4 (NA)	6 (0)	5.11 (1.05)	7.2 (0.99)	7.8 (0.61)	<0.001
Professionalism						
1. Care for diseases and conditions	6 (NA)	6 (0)	6.44 (1.33)	7.77 (0.65)	7.67 (0.75)	<0.001
2. Maintenance of physical and emotional health	4 (NA)	5.33 (1.15)	5.56 (1.33)	7.49 (0.89)	7.71 (0.71)	<0.001
3. Performance of assignments and administrative tasks	4 (NA)	3.33 (1.15)	5.11 (1.05)	6.17 (1.64)	7.31 (1.19)	<0.001
Interpersonal and communication skills						
1. Care for diseases and conditions	4 (NA)	5.33 (1.15)	5.56 (0.88)	7.43 (1.04)	7.71 (0.71)	<0.001
2. Coordination of care	6 (NA)	6.67 (1.15)	5.78 (0.67)	7.49 (0.89)	7.71 (0.71)	<0.001
3. Performance of operations and procedures	4 (NA)	5.33 (1.15)	5.78 (0.67)	6.46 (0.98)	7.63 (0.78)	<0.001

PGY: post-graduate year; SD: standard deviation; NA: not applicable

*Analysis of rating change across PGY required exclusion of the lone PGY-1 observation, which had no standard deviation.

TABLE 2. Prevalence of Low Ratings for Each Milestone and Performance of Models Estimating Clinical Competency Committee Assessment Ratings, With and Without Natural-Language Processing Predictors

Competency	Prevalence	Area under receiver operating characteristic curve (AUC)		
	Low ratings, mean (SD) = 0.36 (0.11)	Non-NLP predictors, mean (SD) = 0.84 (0.05)	NLP predictors, mean (SD) = 0.83 (0.06)	All predictors, mean (SD) = 0.87 (0.05)
Patient care				
1. Care for diseases and conditions	0.23	0.86	0.95	0.96
2. Care for diseases and conditions	0.27	0.93	0.88	0.92
3. Performance of operations and procedures	0.57	0.89	0.78	0.95
Medical knowledge				
1. Care for diseases and conditions	0.45	0.81	0.82	0.85
2. Performance of operations and procedures	0.45	0.83	0.82	0.81
Systems-based practice				
1. Coordination of care	0.26	0.79	0.81	0.83
2. Improvement of care	0.40	0.75	0.82	0.81
Practice-based learning and improvement				
1. Teaching	0.28	0.76	0.80	0.81
2. Self-directed learning	0.42	0.78	0.83	0.85
3. Improvement of care	0.33	0.83	0.92	0.92
Professionalism				
1. Care for diseases and conditions	0.23	0.88	0.87	0.94
2. Maintenance of physical and emotional health	0.29	0.86	0.82	0.83
3. Performance of assignments and administrative tasks	0.52	0.83	0.79	0.84
Interpersonal and communication skills				
1. Care for diseases and conditions	0.30	0.83	0.74	0.86
2. Coordination of care	0.29	0.89	0.89	0.90
3. Performance of operations and procedures	0.49	0.88	0.71	0.92

NLP: natural language processing; SD: standard deviation

analyses, and during a CCC meeting, faculty could spend additional time discussing these residents. Faculty could also track estimates of CCC ratings over time. Since AUCs for models using NLP predictors are comparable to AUCs for models using all predictors, priority might be given to incorporating data sources that do not already include numeric information (e.g., messages existing outside of the MedHub performance assessment system). Priority might also be given to analysis of text that addresses gaps in numeric data (e.g., *improvement of care* under *systems-based practice*). Alternately, faculty rater training could be used to enhance the quality of text feedback for specific Milestones.

This study has limitations. First, the development of predictive models can entail tradeoffs between performance and interpretability (e.g., the ability to see how specific predictors account for variance in each Milestone rating). This increases the risk of an NLP model obscuring bias related to gender, ethnicity, or other variables that should have no bearing on performance ratings. Therefore, implementation of these methods should be preceded by attempts at detection and mitigation of biases that NLP might propagate from written assessments. Second, our study incorporated assessments from only 24 residents at a single institution and these findings might not generalize to other groups of residents. However, the pattern of high AUC means and small AUC standard deviations across models, despite such a small sample, is reassuring. Despite these limitations, our findings should provide medical educators with useful information on how NLP might support complex holistic review processes.

CONCLUSION

NLP can identify language correlated with specific ACGME Milestone ratings. In preparation for CCC meetings, faculty could use information automatically extracted from text to focus attention on residents who might benefit from additional support and guide the development of educational interventions.

FUNDING AND CONFLICTS OF INTEREST

This project was funded by a University of Michigan Medical School Capstone for Impact grant. The authors have no conflicts of interest to disclose.

REFERENCES

1. Accreditation Council for Graduate Medical Education. Clinical competency committees. Accessed May 6, 2020. Available at: https://www.acgme.org/Portals/0/ACGMEClinicalCompetencyCommitteeGuidebook.pdf
2. Manning CD, Schütze H. Foundations of statistical natural language processing. MIT Press; 2000. p. 680.
3. Kuo LE, Hoffman RL, Morris JB, et al. A milestone-based evaluation system-the cure for grade inflation? *J Surg Educ*. 2015;72(6):e218–e225. https://doi.org/10.1016/j.jsurg.2015.09.012.
4. Accreditation Council for Graduate Medical Education, American Board of Surgery. The general surgery milestone project. Accessed July 22, 2018. Available at: http://www.acgme.org/Portals/0/PDFs/Milestones/SurgeryMilestones.pdf?ver=2015-11-06-120519-653
5. Edmondson M. googleLanguageR: Call Google's 'Natural Language' API, 'Cloud Translation' API and 'Cloud Speech' API. Accessed February 7, 2020. Available at: https://cran.r-project.org/web/packages/googleLanguageR/index.html
6. Google LLC. Cloud Natural Language. Accessed February 7, 2020. Available at: https://cloud.google.com/natural-language/
7. Pang B, Lee L. Opinion mining and sentiment analysis. *Found Trend Inf Retr*. 2008;2(1−2):1–135.
8. Silge J, Robinson D. tidytext: text mining and analysis using tidy data principles in R. Accessed February 7, 2020. http://dx.doi.org/10.21105/joss.00037
9. Rinker TW. textstem: Tools for stemming and lemmatizing text. Accessed February 7, 2020. Available at: http://github.com/trinker/textstem
10. h2o.ai Inc. Driverless AI. Accessed February 7, 2020. Available at: https://www.h2o.ai/products/h2o-driverless-ai/
11. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Accessed May 11, 2020. Available at: https://www.R-project.org/
12. Dias RD, Gupta A, SJ Yule. Using machine learning to assess physician competence: a systematic review. *Acad Med*. 2018. https://doi.org/10.1097/acm.0000000000002414.
13. Chary M, Parikh S, Manini AF, Boyer EW, Radeos M. A review of natural language processing in medical education. *West J Emerg Med*. 2019;20(1):78–86. https://doi.org/10.5811/westjem.2018.11.39725.
14. 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;2012:1459–1468.