Natural Language Processing and Assessment of Resident Feedback Quality

Quintin P. Solano, BS,[*] Laura Hayward, BS,[†] Zoey Chopra, BA,[‡] Kathryn Quanstrom, BA,[§] Daniel Kendrick, MD,[‖] Kenneth L. Abbott, MD, MS,[¶] Marcus Kunzmann, AB,[#] Samantha Ahle, MD, MHS,[**] Mary Schuller, MSEd,[††] Erkin Ötleş, MSE,[‡‡] and Brian C. George, MD, MAEd[§§]

[*]University of Michigan Medical School, Ann Arbor, Michigan; [†]University of Michigan Medical School, Ann Arbor, Michigan; [‡]University of Michigan Medical School, Ann Arbor, Michigan; [§]University of Michigan Medical School, Ann Arbor, Michigan; [‖]Department of Surgery, University of Minnesota Medical School, Minneapolis, Minnesota; [¶]University of Michigan Medical School, Ann Arbor, Michigan; [#]Washington University School of Medicine in St. Louis, St Louis, Missouri; [**]Department of Surgery, Yale School of Medicine, New Haven, Connecticut; [††]Department of Surgery, Michigan Medicine, Ann Arbor, Michigan; [‡‡]Department of Industrial and Operations Engineering , University of Michigan Medical School, University of Michigan, Ann Arbor, Michigan; and [§§]Center for Surgical Training and Research, Michigan Medicine, Ann Arbor, Michigan

OBJECTIVE: To validate the performance of a natural language processing (NLP) model in characterizing the quality of feedback provided to surgical trainees.

DESIGN: Narrative surgical resident feedback transcripts were collected from a large academic institution and classified for quality by trained coders. 75% of classified transcripts were used to train a logistic regression NLP model and 25% were used for testing the model. The NLP model was trained by uploading classified transcripts and tested using unclassified transcripts. The model then classified those transcripts into dichotomized high- and low- quality ratings. Model performance was primarily assessed in terms of accuracy and secondary performance measures including sensitivity, specificity, and area under the receiver operating characteristic curve (AUROC).

SETTING: A surgical residency program based in a large academic medical center.

PARTICIPANTS: All surgical residents who received feedback via the Society for Improving Medical Professional Learning smartphone application (SIMPL, Boston, MA) in August 2019.

RESULTS: The model classified the quality (high vs. low) of 2,416 narrative feedback transcripts with an accuracy of 0.83 (95% confidence interval: 0.80, 0.86), sensitivity of 0.37 (0.33, 0.45), specificity of 0.97 (0.96, 0.98), and an area under the receiver operating characteristic curve of 0.86 (0.83, 0.87).

CONCLUSIONS: The NLP model classified the quality of operative performance feedback with high accuracy and specificity. NLP offers residency programs the opportunity to efficiently measure feedback quality. This information can be used for feedback improvement efforts and ultimately, the education of surgical trainees. (J Surg Ed 78:e72–e77. © 2021 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

ABBREVIATIONS: NLP, Natural language processing SIMPL Society for Improving Medical Professional Learning

KEY WORDS: feedback, medical education, natural language processing, machine learning

COMPETENCIES: Practice-Based Learning and Improvement, Medical Knowledge

INTRODUCTION

Performance feedback is necessary for effective learning. In surgery, feedback supports the development of both

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Correspondence: Inquiries to Quintin P. Solano, B.S., University of Michigan Medical School, 1301 Catherine St. Ann Arbor, MI 48109; e-mail: qsolano@med.umich.edu

  1931-7204/$30.00 https://doi.org/10.1016/j.jsurg.2021.05.012

technical and non-technical skills.[1-6] For this reason, providing residents with performance feedback is an Accreditation Council for Graduate Medical Education (ACGME) core program requirement.[7] To address this need, new workplace-based assessment tools provide a mechanism for faculty to provide trainees with dictated feedback.[8-10] This in turn has led to greater volume of feedback provided to surgical trainees.[8,11] While the *quantity* of feedback is important for learning, it is the *quality* of feedback matters most.[12-14] Within that context, training programs must ensure that faculty use these new tools to provide the high-quality feedback that trainees need. However, current approaches to characterizing the quality of performance feedback are labor and resource intensive, often requiring raters to individually evaluate each piece of feedback in a dataset.[10,15]

Natural language processing (NLP), a set of machine learning methods, may offer an automated solution to this problem. A previous pilot study compared different NLP models applied to narrative data from the Society for Improving Medical Professional Learning (SIMPL) smartphone application (Boston, MA).[8] That study identified which type of NLP model most accurately classified the quality of feedback transcripts of a small sample of surgical trainee feedback.[16] While the initial results were promising, performance assessment was limited by the size of the dataset. Therefore, the expected performance of utilizing NLP tools to automatically assess feedback quality is unknown.

The primary aim of this study was to assess the performance of an NLP model to characterize the quality of feedback provided to surgical trainees. To do this we use a much larger dataset. The NLP model was trained using a set of coded transcripts, subsequently tested, and then analyzed for performance.

MATERIALS AND METHODS

Study Population

We collected SIMPL transcripts of dictated operative performance feedback from a single academic surgical residency program, all recorded in August 2019. The University of Michigan institutional review board deemed this study exempt from review.

Data Collection

Dictated feedback was collected using the SIMPL smartphone app, which was developed to facilitate post-operative evaluation of surgical residents' intra-operative performance. Feedback was transcribed via Google Cloud Speech-to-Text (Mountain View, CA), de-identified by a study coordinator, and then coded for quality.

Quality Assessment

All transcripts were evaluated by two separate teams with 2 coders each. The coders were medical students who were trained on a set of "warm-up transcripts" coded by surgeons in a previous study.[15] Their codes were then compared to expert ratings and discrepancies were discussed to improve rater accuracy. Transcripts were assessed in phases of 500. After every phase and prior to initiation of any subsequent phase, coding discrepancies were identified, and each coder team met to discuss their coding decisions and refine coding schema for subsequent phases. During coding of the training and study data sets, authors DK and BG were consulted for clarification when questions arose about the meaning of text in the transcripts. For each discordant code, each team reached consensus on a single quality code for each transcript. These final codes were then used to train the NLP model.

Coders classified the feedback following the methods described by Ahle et al[15], with an initial classification of the transcripts as "relevant" or not. If a transcript was coded as "relevant", subsequent coding would assign the transcript as being "specific", "corrective", both, or neither. These binary attributes were assessed by each trained rater. Coding for each transcript occurred at the sentence level; if any sentence within a transcript qualified as "relevant", "specific", and/or "corrective", the entire transcript would be coded as such.

Transcripts rated as both "specific" and "corrective" were classified as effective (*E*); transcripts rated as specific or corrective but not both were classified as mediocre (*M*); transcripts rated as relevant but neither specific nor corrective were classified as ineffective (*I*). Transcripts not rated as relevant were classified as other (*O*). These codes were further dichotomized for specific analyses, with transcripts rated as *E* or *M* classified as high quality, and transcripts rated as *I* or *O* classified as low quality (Table 1).

TABLE 1. Transcript Coding System

Initial Coding System	Final Classification	Dichotomized Classification
Relevant AND specific AND corrective	Effective (*E*)	High-quality
Relevant AND (specific OR corrective)	Mediocre (*M*)	High-quality
Relevant NOT (specific AND corrective)	Ineffective (*I*)	Low-quality
NOT relevant	Other (*O*)	Low-quality

Statistical Analysis

NLP models were constructed using the Python[17] programming language with the aid of the SKLearn[18], Pandas[19], and Numpy[20] frameworks. Transcripts were pre-processed into bag-of-word vectors[21] with varying n-gram sizes, ranging from length 1 to 5. The data was randomly split (75%/25%) into a training set and a testing set.

Logistic regression models were chosen based on results from a pilot study.[17] Model hyperparameters and pipeline parameters (e.g. n-gram size) were assessed using a 5-fold cross-validation grid search on the training set. Once the best parameters were found, the models were trained on the full training set and evaluated on the testing set.

The primary outcomes were the predictive accuracies of both the individual, and dichotomized coding systems. The individual class (*E, M, I, O*) performance ratings were calculated using class weighted metrics (i.e., micro weighting). Secondary outcomes were sensitivity, specificity, and negative and positive predictive values, and area under the receiver operating characteristic curve (AUROC) of the NLP model. Confidence intervals were estimated using bootstrap sampling. The test dataset was resampled with replacement 1,000 times to generate bootstrap samples of the performance metrics. This bootstrap analysis enabled estimation of the predictive model's variation in accuracy, and the other performance measures, in relation to the distribution of transcript quality labels.

RESULTS

A total of 2,416 transcripts were coded for quality and are described in Table 2. Overall, 1,014 (42%) were coded as Effective (*E*) and 1,811 (75%) of the transcripts were high quality (*E, M*). Examples of high quality and low-quality feedback are shown in Table 3.

The accuracy of the model when rating individual pieces of feedback as *E, M, I,* or *O* was 0.65 (95% confidence interval: 0.61, 0.65), with sensitivity of 0.46 (0.43, 0.49), specificity of 0.87 (0.86, 0.89), positive predictive value of 0.50 (0.48, 0.53), and negative predictive value of 0.87 (0.86-0.89). Individual class performance metrics are presented in Appendix Table 1.

When ratings were dichotomized (high vs. low quality), the model accuracy for classifying low quality feedback was 0.83 (0.80, 0.86), with sensitivity of 0.37 (0.33, 0.45), specificity of 0.97 (0.96, 0.98), positive predictive value of 0.80 (0.74, 0.85), negative predictive value of 0.83 (0.80, 0.85), and area under the receiver operating characteristic curve of 0.86 (0.83, 0.87; Fig. 1).

TABLE 2. Individual Transcript Codes

Classification	N (%)
Individual Class	
Effective (*E*)	1,014 (42%)
Mediocre (*M*)	797 (33%)
Ineffective (*I*)	604 (25%)
Other (*O*)	2 (<1%)
Dichotomized	
High-Quality (*E, M*)	1,811 (75%)
Low-Quality (*I, O*)	605 (25%)

DISCUSSION

We investigated the performance characteristics of NLP models tasked with characterizing the quality of feedback provided to surgical trainees. NLP models can classify feedback quality with high accuracy and specificity. However, sensitivity was much lower, indicating that the algorithm can most reliably identify low quality feedback. The NLP model described in this report may be useful for measuring the effects of feedback interventions in surgical training programs.

This study validates the results of a previous pilot study examining the capabilities of NLP, however, in this study we utilized a larger sample size to further improve classification performance in the hope that this technology might be used on a larger scale.[16] Model metrics from this study are comparable to those from studies of NLP outside medical education settings.[22-24] Ramachandran et al. utilized NLP to automatically assess the quality of research reviews and reported accuracies of 0.32-0.67.[22] Our model achieved relatively high accuracy, and this highlights the potential of NLP for future feedback quality improvement in a medical education context.

NLP models may be a novel tool to both measure and help improve feedback. The importance of providing effective, high quality feedback is clear, yet the measurement of feedback quality is resource intensive.[15,25] NLP models can reduce this burden by automatically characterizing feedback quality in near real time. Surgical residency programs could use automated characterizations of feedback quality to improve the feedback their residents receive. For example, faculty who consistently provide low quality feedback might be provided with additional faculty development resources. Furthermore, automated feedback classification might be used to develop and test new ideas for improving feedback, and to assess the impact of implementing existing methods for improving feedback quality.[9,26,27]

Our approach, while piloted within a general surgery residency program, is likely generalizable to other procedural specialties. Although, the feedback provided to trainees in non-procedural specialties likely features

TABLE 3. Examples of Transcripts and Quality Classification

Final Classification	Categorization	Examples
Effective	Relevant, specific, and corrective	Dr. *[RES_NAME] is very familiar with the basic parts of a surgical procedure such as creating the laparotomy incision and dividing tissue being shown to him between clamps period for recommended next steps I suggest he work on specific skill sets such as dissection around the rectum, division of mesenteric vessels, creation of an ostomy period his skills are on
Mediocre	Relevant and specific	Very good job just getting to know the catheters in the others we have available will assist excellent tissue handling and judgment with the wires and others
	Relevant and corrective	Outstanding performance overall *[RES_NAME] as usual work on communicating with the staff a little bit more I tend to take over a little bit so that it's harder for you to complete some that so but otherwise it's great to give you a key
Ineffective	Relevant	Provided invaluable help demonstrate a good judgment glad you're there
Other	Irrelevant	He's okay I just wanted to give you some feedback on your M&M presentation this morning

*[RES_NAME]: De-identified resident name.

different terminology and verbiage, our methods could be used to develop NLP models for use in such settings.

This study has limitations. First, all the feedback we analyzed was collected via SIMPL at a single academic institution and may not be representative of feedback delivered in other settings or with other tools. Second, the raters of feedback quality were medical students who, due to limited experience, may have miscoded some transcripts, though we attempted to mitigate this via training with previously coded transcripts and frequent consultations with practicing surgeons. Although some exemplar and difficult transcripts were discussed with the practicing surgeons, the majority of the coded transcripts were not audited by them. Third, transcripts of audio feedback sometimes contained transcription errors requiring reasoned guesses concerning meaning, and some of these guesses may have been incorrect. Finally, factors like age, race, gender, and accents may impact transcription quality and content, when moving to implement these models we must be mindful of the risk of related biases in NLP model output. Notwithstanding these limitations, this report highlights the potential utility and provides a benchmark for the study of NLP in medical education.

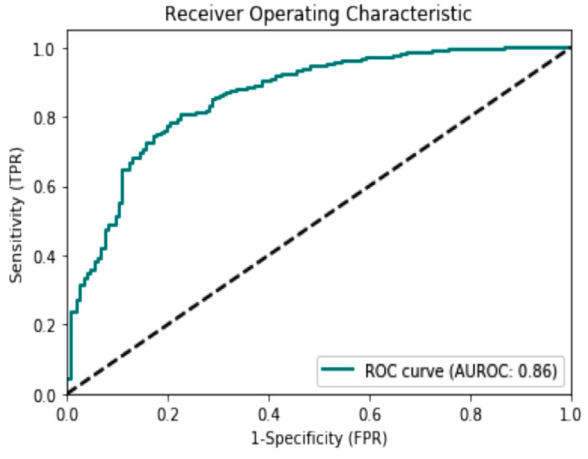

FIGURE 1. Receiver operating characteristic curve for natural language processing dichotomized classification of feedback transcripts.

CONCLUSIONS

An NLP model is able to classify operative performance feedback quality with high accuracy and specificity and modest sensitivity. NLP could serve as effective approach for automated classification of feedback quality. That information can ultimately be used to improve feedback and in turn accelerate learning for surgical trainees.

ACKNOWLEDGEMENT

The authors would like to acknowledge the program directors, coordinators, faculty, and residents from the programs who are members of the Society for Improving Medical Professional Learning (SIMPL). Without their educational efforts and ongoing support, this study would not have been possible.

REFERENCES

1. Bjerrum F, Maagaard M, Led Sorensen J. Effect of Instructor feedback on skills retention after laparoscopic simulator training: Follow-up of a randomized trial. *Journal of Surgical Education*. 2015;72:53–60. https://doi.org/10.1016/j.jsurg.2014.06.013.
2. Boyle E, Al-Akash M, Gallagher AG, Traynor O, Hill ADK, Neary PC. Optimising surgical training: Use of feedback to reduce errors during a

simulated surgical procedure. *Postgraduate Medical Journal*. 2011;87:524–528. https://doi.org/10.1136/pgmj.2010.109363.

3. Boyle E, O'Keeffe DA, Naughton PA, Hill ADK, McDonnell CO, Moneley D. The importance of expert feedback during endovascular simulator training. *Journal of Vascular Surgery*. 2011;54. https://doi.org/10.1016/j.jvs.2011.01.058.

4. Trehan A, Barnett-Vanes A, Carty MJ, McCulloch P, Maruthappu M. The impact of feedback of intraoperative technical performance in surgery: a systematic review. *BMJ Open*. 2015;5. https://doi.org/10.1136/bmjopen-2014-006759. e006759.

5. Porte MC, Xeroulis G, Reznick RK, Dubrowski A. Verbal feedback from an expert is more effective than self-accessed feedback about motion efficiency in learning new surgical skills. *American Journal of Surgery*. 2007;193:105–110. https://doi.org/10.1016/j.amjsurg.2006.03.016.

6. Wigton R, Patil K, Hoellerich V. The effect of feedback in learning clinical diagnosis. *Journal of Medical Education*. 1986;61:816–822.

7. Acgme. *ACGME Common Program Requirements (Residency)*. 2018.

8. Bohnen JD, George BC, Williams RG. The Feasibility of Real-Time Intraoperative Performance Assessment With SIMPL (System for Improving and Measuring Procedural Learning): Early Experience From a Multi-institutional Trial. *Journal of Surgical Education*. 2016;73. https://doi.org/10.1016/j.jsurg.2016.08.010. e118-e130.

9. Ahmed M, Arora S, Russ S, Darzi A, Vincent C, Sevdalis N. Operation debrief: a SHARP improvement in performance feedback in the operating room. *Annals of surgery*. 2013;258:958–963.

10. Shaughness G, Georgoff PE, Sandhu G. Assessment of clinical feedback given to medical students via an electronic feedback system. *Journal of Surgical Research*. 2017;218:174–179. https://doi.org/10.1016/j.jss.2017.05.055.

11. George BC, Bohnen JD, Williams RG. Readiness of US General Surgery Residents for Independent Practice. *Annals of Surgery*. 2017;266:582–594. https://doi.org/10.1097/SLA.0000000000002414.

12. Ende J. Feedback in Clinical Medical Education. *JAMA*. 1983;250:777–781. https://doi.org/10.1001/jama.1983.03340060055026.

13. Grantcharov T.P., Schulze S., Kristiansen V.B.. The impact of objective assessment and constructive feedback on improvement of laparoscopic performance in the operating room. *Surgical Endoscopy and Other Interventional Techniques*. 2007;21:2240-2243. http://doi:10.1007/s00464-007-9356-z.

14. Hattie J, Timperley H. The Power of Feedback. *Review of Educational Research*. 2007;77:81–112. https://doi.org/10.3102/003465430298487.

15. Ahle SL, Eskender M, Schuller M. The Quality of Operative Performance Narrative Feedback. *Annals of Surgery*. 2020. https://doi.org/10.1097/sla.0000000000003907. Publish Ah(Xx):1-4.

16. Otles E., Kendrick D., Solano Q., Using Natural Language Processing to Automatically Assess Feedback Quality. Academic Medicine.

17. van Rossum, Guido D Jr, Fred L. PythonTutorial. Centrum voor Wiskunde en Informatica. The Netherlands: Amsterdam; 1995.

18. Pedregosa F, Varoquaux G, Gramfort A. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12:2825–2830.

19. McKinney W, Data Structures for Statistical Computing in Python. In: van der Walt S, Millman J, eds.; 2010. doi:10.25080/Majora-92bf1922-00a

20. Harris CR, Millman KJ, van der Walt SJ. Array programming with NumPy. *Nature*. 2020;585:357–362. https://doi.org/10.1038/s41586-020-2649-2.

21. Brownlee J. A Gentle Introduction to the Bag-of-Words Model. Deep Learning for Natural Language Processing, 2020.

22. Ramachandran L, Gehringer EF, Yadav RK. Automated Assessment of the Quality of Peer Reviews using Natural Language Processing Techniques. *International Journal of Artificial Intelligence in Education*. 2017;27:534–581. https://doi.org/10.1007/s40593-016-0132-x.

23. Xiong W, Litman D, Schunn C. Natural Language Processing techniques for researching and improving peer feedback. *Journal of Writing Research*. 2012;4:155–176.

24. Cho K. Machine Classification of Peer Comments in Physics. *Educational Data Mining*. 2008: 192–196. Published online.

25. Ali AS, Bussey M, O'flynn KJ, Eardley I. Quality of feedback using workplace based assessments in urological

training. *Journal of Clinical Urology*. 2012;5:39–43. https://doi.org/10.1016/j.bjmsu.2011.10.001.

26. Domson GF, Appelbaum N, Kates S. Instituting a Postoperative Feedback Process for Orthopedic Surgery Residents. *Journal of Surgical Education*. 2019;76:1200–1204. https://doi.org/10.1016/j.jsurg.2019.03.007.

27. Shaughnessy MP, Ahle SL, Oliveira K, Longo WE, Yoo PS. Improving Satisfaction With Operating Room Feedback: An Effective, Low-Profile, No-Cost Intervention. *Journal of Surgical Education*. 2019;76. https://doi.org/10.1016/j.jsurg.2019.10.002. e138-e145.

APPENDIX 1. Individual Class Performance Metrics

	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value
Effective (*E*)	0.88	0.89	0.67	0.89
Mediocre (*M*)	0.53	0.76	0.57	0.76
Ineffective (*I*)	0.43	0.85	0.77	0.85
Other (*O*)	0.00	0.99	0.00	0.99

Performance metrics for the individual classification of transcripts.