Global Surgical Education - Journal of the Association for Surgical Education (2022) 1:58
https://doi.org/10.1007/s44186-022-00051-y

ORIGINAL ARTICLE

Using natural language processing to determine factors associated with high-quality feedback

Kayla M. Marcotte[1,2] · Erkin Ötleş[3] · Angela E. Thelen[1] · Rebecca Gates[1] · Brian C. George[1,4] · Andrew E. Krumm[1,2]

Received: 22 May 2022 / Revised: 1 September 2022 / Accepted: 21 September 2022 / Published online: 2 November 2022

Abstract

Purpose Feedback is a cornerstone of medical education. However, not all feedback that residents receive is high-quality. Natural language processing (NLP) can be used to efficiently examine the quality of large amounts of feedback. We used a validated NLP model to examine factors associated with the quality of feedback that general surgery trainees received on 24,531 workplace-based assessments of operative performance.

Methods We analyzed transcribed, dictated feedback from the Society for Improving Medical Professional Learning's (SIMPL) smartphone-based app. We first applied a validated NLP model to all SIMPL evaluations that had dictated feedback, which resulted in a predicted probability that an instance of feedback was "relevant", "specific", and/or "corrective." Higher predicted probabilities signaled an increased likelihood that feedback was high quality. We then used linear mixed-effects models to examine variation in predictive probabilities across programs, attending surgeons, trainees, procedures, autonomy granted, operative performance level, case complexity, and a trainee's level of clinical training.

Results Linear mixed-effects modeling demonstrated that predicted probabilities, i.e., a proxy for quality, were lower as operative autonomy increased ("Passive Help" $B = -1.29$, $p < .001$; "Supervision Only" $B = -5.53$, $p < 0.001$). Similarly, trainees who demonstrated "Exceptional Performance" received lower quality feedback ($B = -12.50$, $p < 0.001$). The specific procedure or trainee did not have a large effect on quality, nor did the complexity of the case or the PGY level of a trainee. The individual faculty member providing the feedback, however, had a demonstrable impact on quality with approximately 36% of the variation in quality attributable to attending surgeons.

Conclusions We were able to identify actionable items affecting resident feedback quality using an NLP model. Attending surgeons are the most influential factor in whether feedback is high quality. Faculty should be directly engaged in efforts to improve the overall quality of feedback that residents receive.

Keywords Feedback · Feedback quality · Natural language processing · Assessment

Introduction

Residents receive a large volume of feedback during their medical training. However, not all feedback is high quality [1, 2]. Experts agree that high-quality feedback should be specific, action-oriented, and given in a timely manner [3–5]. Residents have indicated that they desire more direct and actionable feedback in their training to increase learning and skill development [6–8]. To improve resident learning, it is important to understand factors affecting the quality of feedback residents receive [9, 10].

While there is a large volume of literature examining resident feedback in medical education, a few studies have analyzed feedback quality across multiple institutions. Assessing feedback quality on a large-scale has proven challenging

✉ Kayla M. Marcotte
kaymar@umich.edu

1 Center for Surgical Training and Research, Department of Surgery, Michigan Medicine, 2800 Plymouth Road, NCRC Building 10 Room A193, Ann Arbor, MI 48109-2800, USA

2 Department of Learning Health Sciences, University of Michigan, Ann Arbor, MI, USA

3 Department of Industrial and Operations Engineering, University of Michigan Medical School, Ann Arbor, MI, USA

4 Department of Surgery, Michigan Medicine, Ann Arbor, MI, USA

due to difficulties with both data collection and analysis. For example, programs often have individual systems for collecting and storing resident feedback. Even if feedback data were successfully aggregated across institutions, efficiently analyzing large volumes of narrative data can be challenging. These factors make it challenging to identify generalizable factors influencing feedback quality. Standardized workplace-based assessments (WBAs) offer a potential solution for assessing resident feedback [11, 12]. WBAs that are administered across multiple programs provide an important source of data for examining what factors are associated with high-quality feedback.

Natural language processing (NLP) can be used to evaluate large amounts of feedback data in a short amount of time [13, 14]. Previous studies have demonstrated that an NLP model could be trained to reliably classify whether feedback provided to general surgery residents was "relevant" and "specific" and/or "corrective." [13, 14] To understand factors affecting the quality of feedback that residents received across multiple general surgery programs, we applied an NLP model to a large corpus of dictated and transcribed feedback from the Society for Improving Medical Professional Learning's (SIMPL) smartphone-based app [15–18]. Analyzing feedback from a large educational registry with a validated NLP model provides one of the first opportunities to identify potentially generalizable factors affecting the feedback trainees receive.

Methods

Study cohort

SIMPL evaluations were collected from September 2015 to September 2021. Evaluations were included from 70 general surgery programs in the United States for categorical trainees across post-graduate year (PGY) 1–5. Incomplete SIMPL evaluations were excluded.

Data source

The SIMPL app provides attending surgeons with the opportunity to rate resident autonomy, resident operative performance, and case complexity, as well as provide narrative feedback data within 72 h of case completion [15, 16]. Autonomy was rated using the Zwisch scale: 1 = "Show and Tell," 2 = "Active Help," 3 = "Passive Help," or 4 = "Supervision Only." [19] Operative performance was rated using the SIMPL Performance Scale: 1 = "Unprepared/Critical Deficiency," 2 = "Inexperienced," 3 = "Intermediate," 4 = "Practice-Ready," or 5 = "Exceptional Performance." Complexity of the case relative so similar procedures could be rated as 1 = "Low Complexity," 2 = "Medium Complexity," or 3 = "High Complexity". Narrative feedback was dictated and then transcribed using Google Cloud Speech-to-Text. Transcribed feedback were then scored using a validated NLP model [13, 14]. This study was deemed exempt by the University of Michigan Institutional Review Board.

Statistical analysis

The NLP model used in the current was trained on text that was coded using the classification scheme and method developed by Ahle et al. [18] This classification scheme categorizes feedback as relevant or irrelevant, specific or general, and corrective and/or encouraging. The process for developing the NLP model and example text aligned to a label of relevant and specific and/or corrective can be found in Solano et al. [13]. The result of applying the trained NLP model to an instance of transcribed feedback from SIMPL was a predicted probability that the feedback was relevant and specific and/or corrective. We used the predicted probabilities as the dependent variable in a linear mixed-effect model where faculty rater, trainee, procedure, and program were modeled as random effects and autonomy, performance, case complexity, and PGY level of a trainee were modeled as fixed effects. Reference levels were set at "Active Help" for autonomy, "Unprepared/Critical Deficiency" for operative performance, "Low" for case complexity, and "PGY 1" for PGY. For autonomy, "Active Help" was used, because trainees who receive an autonomy rating of "Show and Tell" are not given an opportunity to receive a performance rating. To determine how much variation in the quality of narrative feedback was explained by different factors, intraclass correlation coefficients (ICC) were calculated for the random effects included in the model.

Results

We examined 24,531 SIMPL evaluations from 70 general surgery programs using a validated NLP model. Study cohort characteristics are shown in Table 1. Feedback quality scores by PGY, case complexity, autonomy rating, and operative performance are shown in Fig. 1. The box plots for PGY illustrate limited variation between levels. Case complexity, likewise, showed limited variation across rating levels. Both autonomy and performance ratings, however, showed demonstrable differences as autonomy and performance ratings increased, respectively.

Linear mixed-effects results are shown in Table 2. Inferential patterns largely match descriptive patterns for autonomy, performance, and complexity scales whereby predicted probabilities decreased as operative autonomy increased ("Passive Help" $B = -1.29$, $p < 0.001$; "Supervision Only" $B = -5.53$, $p < 0.001$), and trainees who demonstrated

Table 1 Descriptive statistics

	PGY 1	PGY 2	PGY 3	PGY 4	PGY 5
Trainees	740	753	845	803	780
Faculty raters	653	721	805	751	723
Dictated ratings	3411	3868	5586	5398	6268
Mean quality score (SD)	86.6 (20.0)	86.6 (20.0)	86.5 (20.4)	84.1 (22.0)	81.9 (24.0)

$N = 70$ programs

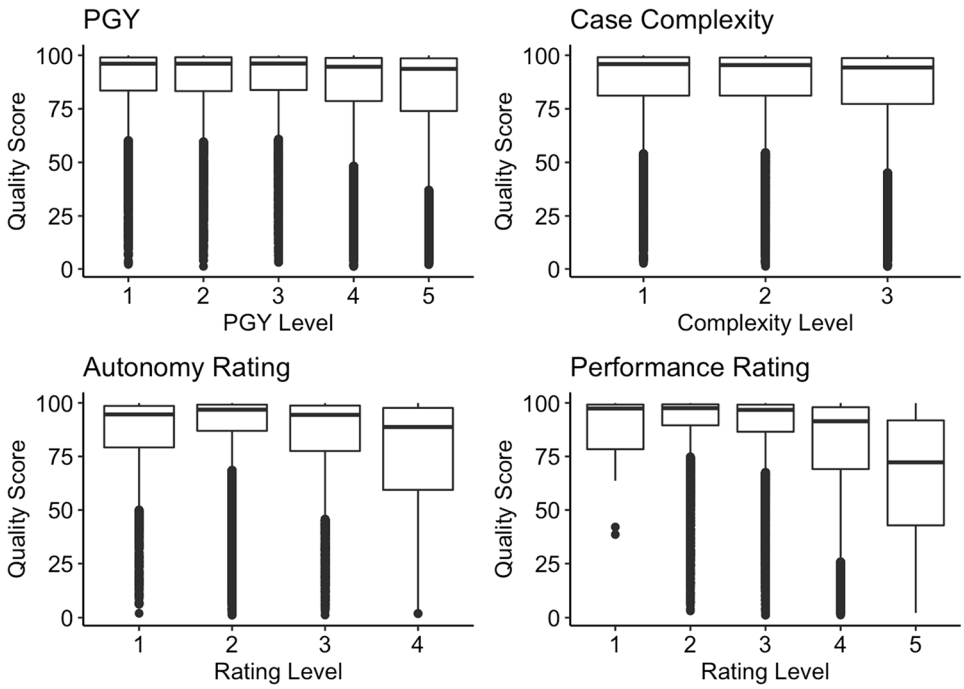

Fig. 1 Breakdown of quality scores by PGY, Case Complexity, Autonomy, and Performance. Autonomy: 1 = "Show and Tell," 2 = "Active Help," 3 = "Passive Help," or 4 = "Supervision Only." [19]. Operative performance: 1 = "Unprepared/Critical Deficiency," 2 = "Inexperienced," 3 = "Intermediate," 4 = "Practice-Ready," or 5 = "Exceptional Performance." Case Complexity: 1 = "Low Complexity," 2 = "Medium Complexity," or 3 = "High Complexity"

"Exceptional Performance" received lower quality feedback (B = -12.50, $p < 0.001$). For random effects, we found small effects for the role that individual trainees, different procedures, or programs played. We observed that approximately 36% of the variation in feedback quality was attributable to the individual attending surgeon providing narrative feedback.

Discussion

This study suggests that despite the many factors that could influence the quality of feedback faculty provide to general surgery residents, the most important factor is the attending surgeon—regardless of the specific resident or procedure. This means that despite the trainee and his or her performance, individual attending surgeons provide consistently high- or low-quality feedback. Deviations from this high-level takeaway can be observed when a trainee, on average, is granted the highest level of autonomy and/or performs to an exceptional standard. At these highest levels, trainees are likely to receive the lowest quality of feedback, meaning that the feedback is not likely to be relevant, specific, and/or corrective. This could be due to faculty triaging the more detailed feedback they give toward lower-performing residents, while providing lower quality feedback (for example, "Good job!" or "Nice work!") to higher performers. The robustness and potential generalizability of these findings is enhanced by the scale at which we were able to examine feedback in this paper.

Since faculty accounted for a large proportion of the variation in the quality of feedback that trainees received, understanding how faculty chooses to construct and gives feedback could uncover opportunities for improving feedback quality more generally [20]. Previous studies have demonstrated that teaching faculty how to give effective feedback can improve the feedback provided [21, 22]. Investing in faculty develops programs that address the components of effective feedback and strategies for giving high-quality feedback could increase the overall quality of feedback given to residents [3, 23]. Faculty can also raise quality when they receive feedback about their feedback

Table 2 Linear mixed-effects model results

Fixed effects	B	SE	*t* Value
Intercept	78.28***	4.26	18.36
Autonomy[1]			
Passive help	− 1.29***	0.33	− 3.93
Supervision only	− 5.53***	0.49	− 11.20
Performance[2]			
Inexperienced	3.87	4.20	0.92
Intermediate	3.12	4.20	0.75
Practice-ready	− 3.01	4.20	− 0.72
Exceptional	− 12.50***	4.26	− 2.94
Complexity[3]			
Medium	0.98**	0.36	2.77
High	− 0.05	0.42	− 0.11
PGY[4]			
2	1.79*	0.51	2.33
3	1.05*	0.51	2.09
4	1.55***	0.54	2.90
5	1.15*	0.56	2.03
Random effects	Variance	ICC	
Trainee	6.03	0.01	
Faculty Rater	183.32	0.36	
Procedure	8.09	0.02	
Program	18.06	0.04	
Residual	194.56		

$p < 0.05$ *; $p < 0.01$ **; $p < 0.001$. *** based on *t* value. ICC = intraclass correlation

Reference categories: "Active help" [1]; "Unprepared/critic deficiency" [2]; "Low complexity" [3]; "PGY 1" [4]

[24]. Future efforts to advance quality could build upon evidence-based strategies combined with effective audit and feedback. [25]

This study is not without limitations. First, only 39% of the SIMPL evaluations we examined included narrative feedback. This percentage of evaluations that include narrative feedback suggests that we are missing feedback instances that are not captured with SIMPL, such as verbal feedback given to residents during an operation. While this may limit the generalizability of our results to contexts outside of the operating room, the robust amount of feedback instances that were captured in our analysis strengthen our confidence identifying factors associated with high-quality feedback. Furthermore, programs that utilize the SIMPL application may be skewed toward giving more feedback, and therefore could overrepresent the amount of high-quality feedback that residents receive. Finally, while the NLP model provided the opportunity examine feedback quality at an unprecedented scale, the underlying model itself requires continual analysis and maintenance to understand the text-based features it attends to in scoring an instance of feedback. We plan to build on this study by examining how the NLP model could identify feedback quality while accounting for additional variables, such as PGY.

A key strength of our study was applying a validated NLP to assess feedback quality of nearly 25,000 evaluations with narrative feedback. The innovative approach used in this paper, while limited in important respects, allowed us to identify potentially generalizable factors associated with high-quality feedback in general surgery training using WBAs.

Conclusions

A highly influential factor in whether residents receive high-quality feedback is the attending surgeon. Efforts to enhance feedback quality should directly engage with faculty to understand their decisions about how to give feedback. By partnering with faculty, the overall quality of feedback that residents receive could be improved.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no financial disclosures.

References

1. Jackson JL, Kay C, Jackson WC, Frank M. The quality of written feedback by attendings of internal medicine residents. J Gen Intern Med. 2015;30(7):973–8. https://doi.org/10.1007/s11606-015-3237-2.
2. Bing-You RG, Trowbridge RL. Why medical educators may be failing at feedback. JAMA. 2009;302(12):1330–1.
3. Jug R, Jiang X "Sara", Bean SM. Giving and receiving effective feedback: a review article and how-to guide. Arch Pathol Lab Med 2019;143(2):244–250. https://doi.org/10.5858/arpa.2018-0058-RA
4. Hewson MG, Little ML. Giving feedback in medical education. J Gen Intern Med. 1998;13(2):111–6. https://doi.org/10.1046/j.1525-1497.1998.00027.x.
5. Lefroy J, Watling C, Teunissen PW, Brand P. Guidelines: the do's, don'ts and don't knows of feedback for clinical education. Perspect Med Educ. 2015;4(6):284–99. https://doi.org/10.1007/s40037-015-0231-7.
6. Jensen AR, Wright AS, Kim S, Horvath KD, Calhoun KE. Educational feedback in the operating room: a gap between resident and faculty perceptions. Am J Surg. 2012;204(2):248–55. https://doi.org/10.1016/J.AMJSURG.2011.08.019.
7. Bello RJ, Sarmiento S, Meyer ML, et al. Understanding surgical resident and fellow perspectives on their operative performance feedback needs: a qualitative study. J Surg Educ. 2018;75(6):1498–503. https://doi.org/10.1016/J.JSURG.2018.04.002.
8. Rose JS, Waibel BH, Schenarts PJ. Disparity between resident and faculty surgeons' perceptions of preoperative preparation, intraoperative teaching, and postoperative feedback. J Surg Educ. 2011;68(6):459–64. https://doi.org/10.1016/J.JSURG.2011.04.003.
9. Kornegay JG, Kraut A, Manthey D, et al. Feedback in medical education: a critical appraisal. AEM Educ Train. 2017;1(2):98–109. https://doi.org/10.1002/aet2.10024.
10. McKendy KM, Watanabe Y, Lee L, et al. Perioperative feedback in surgical training: a systematic review. Am J Surg. 2017;214(1):117–26. https://doi.org/10.1016/J.AMJSURG.2016.12.014.
11. Norcini JJ. Workplace assessment. In: Understanding medical education: evidence, theory and practice. Wiley, 2013; 2013. https://books.google.com/books?hl=en&lr=&id=EWsKAgAAQBAJ&oi=fnd&pg=PA279&dq=Norcini+Workplace+assessment&ots=5SNX3lR8VE&sig=GZgxsYdM7MDc_xsrIiKvbnj1fv8. Accessed 15 May 2022.
12. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide No. 31. Med Teach. 2007;29(9-10):855-871. https://doi.org/10.1080/01421590701775453
13. Solano QP, Hayward L, Chopra Z, et al. Natural language processing and assessment of resident feedback quality. J Surg Educ. 2021;78(6):e72–7. https://doi.org/10.1016/J.JSURG.2021.05.012.
14. Ötles E, Kendrick DE, Solano QP, et al. Using natural language processing to automatically assess feedback quality: findings from 3 surgical residencies. Acad Med. 2021. https://doi.org/10.1097/ACM.0000000000004153.
15. George BC, Bohnen JD, Schuller MC, Fryer JP. Using smartphones for trainee performance assessment: a SIMPL case study. Surgery. 2020;167(6):903–6. https://doi.org/10.1016/j.surg.2019.09.011.
16. Bohnen JD, George BC, Williams RG, et al. The Feasibility of real-time intraoperative performance assessment with SIMPL (system for improving and measuring procedural learning): early experience from a multi-institutional trial. J Surg Educ. 2016;73(6):e118–30. https://doi.org/10.1016/j.jsurg.2016.08.010.
17. Williams RG, George BC, Bohnen JD, et al. A proposed blueprint for operative performance training, assessment, and certification. Ann Surg. 2021. https://doi.org/10.1097/SLA.0000000000004467.
18. Ahle SL, Eskender M, Schuller M, et al. The quality of operative performance narrative feedback: a retrospective data comparison between end of rotation evaluations and workplace-based assessments. Ann Surg. 2020. https://doi.org/10.1097/SLA.0000000000003907.
19. DaRosa DA, Zwischenberger JB, Meyerson SL, et al. A theory-based model for teaching and assessing residents in the operating room. J Surg Educ. 2013;70(1):24–30. https://doi.org/10.1016/j.jsurg.2012.07.007.
20. Kogan JR, Conforti LN, Bernabeo EC, Durning SJ, Hauer KE, Holmboe ES. Faculty staff perceptions of feedback to residents after direct observation of clinical skills. Med Educ. 2012;46(2):201–15. https://doi.org/10.1111/j.1365-2923.2011.04137.x.
21. Junod Perron N, Nendaz M, Louis-Simonet M, et al. Effectiveness of a training program in supervisors' ability to provide feedback on residents' communication skills. Adv Heal Sci Educ. 2013;18(5):901–15. https://doi.org/10.1007/s10459-012-9429-1.
22. Minehart RD, Rudolph J, Pian-Smith MCM, Raemer DB. Improving faculty feedback to resident trainees during a simulated case: a randomized, controlled trial of an educational intervention. Anesthesiology. 2014;120(1):160–71. https://doi.org/10.1097/ALN.0000000000000058.
23. Zendejas B, Toprak A, Harrington AW, Lillehei CW, Modi BP. Quality of dictated feedback associated with SIMPL operative assessments of pediatric surgical trainees. Am J Surg. 2021;221(2):303–8. https://doi.org/10.1016/j.amjsurg.2020.10.014.
24. Baker K. Clinical teaching improves with resident evaluation and feedback. Anesthesiology. 2010;113(3):693–703. https://doi.org/10.1097/ALN.0b013e3181eaacf4.
25. Springer MV, Sales AE, Islam N, et al. A step toward understanding the mechanism of action of audit and feedback: a qualitative study of implementation strategies. Implement Sci. 2021;16(1):35. https://doi.org/10.1186/s13012-021-01102-6.