ORIGINAL RESEARCH

Initial Development of an Automated Platform for Assessing Trainee Performance on Case Presentations

Andrew J. King[1], Jeremy M. Kahn[1], Emily B. Brant[1], Gregory F. Cooper[2], and Danielle L. Mowery[3]

[1]Department of Critical Care Medicine, [2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, and [3]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania

ORCID IDs: 0000-0002-9809-0563 (A.J.K.); 0000-0001-9688-5576 (J.M.K.); 0000-0002-0207-8210 (E.B.B.); 0000-0002-9276-773X (G.F.C.); 0000-0003-3802-4457 (D.L.M.)

ABSTRACT

Background: Oral case presentation is a crucial skill of physicians and a key component of team-based care. However, consistent and objective assessment and feedback on presentations during training are infrequent.

Objective: To determine the potential value of applying natural language processing, computer software that extracts meaning from text, to transcripts of oral case presentations as a strategy to assess their quality automatically and objectively.

Methods: We transcribed a collection of simulated oral case presentations. The presentations were from eight critical care fellows and one critical care attending. They were instructed to review the medical charts of 11 real intensive care unit patient cases and to audio record themselves, presenting each case as if they were doing so on morning rounds. We then used natural language processing to convert the transcripts from human-readable text into machine-readable numbers. These numbers represent details of the presentation style and content. The distance between the numeric representation of two different transcripts negatively correlates with the similarity of those two transcripts. We ranked fellows on the basis of how similar their presentations were to the attending's presentations.

Results: The 99 presentations included 260 minutes of audio (mean length: 2.6 ± 1.24 min per case). On average, 23.88 ± 2.65 sentences were spoken, and each sentence had 14.10 ± 0.67 words, 3.62 ± 0.15 medical concepts, and 0.75 ± 0.09 medical adjectives. When ranking fellows on the basis of how similar their presentations were to the attending's presentation, we found a gap between the five fellows with the most similar presentations and the three fellows with the least similar presentations (average group similarity scores of 0.62 ± 0.01 and 0.53 ± 0.01, respectively). Rankings were sensitive to whether presentation style or content information were weighted more heavily when calculating transcript similarity.

Conclusion: Natural language processing enabled the ranking of case presentations on the basis of how similar they were to a reference presentation. Although additional work is needed to convert these rankings, and underlying similarity scores, into actionable feedback for trainees, these methods may support new tools for improving medical education.

Keywords:
natural language processing; clinical rounds; intensive care unit; medical education; deep learning

(Received in original form January 18, 2022; accepted in final form August 8, 2022)

ATS Scholar Vol 3, Iss 4, pp 548–560, 2022

DOI: 10.34197/ats-scholar.2022-0010OC

Oral clinical case presentations are a key component of team-based health care, particularly in academic settings. They allow one team member, typically a trainee, to inform other team members about a patient's history and recent events to engender effective collaboration. Furthermore, the oral case presentation is a useful proxy to assess trainee competency in clinical reasoning and patient care (1). In teaching hospitals, trainees, such as medical students, residents, and fellows, often present cases to senior physicians as part of their professional training. Ideally, senior physicians then provide feedback to the trainee by asking clarifying questions, providing constructive comments, and integrating teachable moments in which the senior physician adds broader clinical context to the current situation (2).

Although this apprenticeship model of teaching oral presentations has many strengths (3), its effectiveness is limited by various factors. Rounding teams are extremely busy, and there is rarely time to provide meaningful feedback (4). Senior physicians are usually not trained in how to give feedback on oral presentations (5). In addition, natural biases and heuristics among teachers mean that feedback lacks objectivity (6, 7). For example, senior physicians experience expert blindness in which their expertise makes it difficult to identify a trainee's knowledge deficiencies (8). As a result of these problems, learners are unlikely to receive objective, actionable, and unbiased feedback on their oral case presentations (9).

There is a crucial need for new approaches to objectively assess and provide feedback to trainees on the quality of their oral case presentations. Novel digital technology might play a role in this regard. For example, trainee presentations could be audio recorded, transcribed using automatic speech recognition software, and analyzed for quality using natural language processing (NLP). Broadly speaking, NLP is software that extracts meaning from text (10). NLP can perform tasks like identifying specific words or phrases, extracting relationships between named entities, or creating a numerical representation of the meaning of a unit of text (10, 11).

Supported by the National Institutes of Health under award numbers R35 HL144804 (National Heart, Lung, and Blood Institute), R01 LM012095 (U.S. National Library of Medicine), and T15 LM007059. The content is solely the responsibility of the authors and does not necessarily represent the official views of the University of Pittsburgh, University of Pennsylvania, or National Institutes of Health.

Author contributions: A.J.K. collected the corpus, proposed the experiments, conducted the experiments, analyzed the results, and drafted the manuscript. E.B.B. and G.F.C. contributed to collecting the corpus, designing the experiments, and interpreting the results. J.M.K. and D.L.M. contributed to designing the experiments and interpreting the results. All authors revised the manuscript and approved the final version.

Correspondence and requests for reprints should be addressed to Andrew J. King, Ph.D., 3550 Terrace Street, 603A, Pittsburgh, PA 15261. E-mail: andrew.king@pitt.edu.

This article has a related editorial.

This article has a data supplement, which is accessible from this issue's table of contents at www.atsjournals.org.

In health care, NLP has been applied to progress notes to improve diagnostic accuracy by recognizing words and phrases related to a specific diagnosis like sepsis (12). When applied to presentation transcripts, NLP methods might enable the comparison of multiple transcripts on the basis of the likeness of their meaning (13). In this context, NLP holds significant promise as a technology to provide objective feedback on trainees' presentations, which at their core consist of analyzable text. We investigated a proof-of-concept version of this vision.

METHODS

To better understand the potential of technology to assist with the assessment of trainee performance, we first present a conceptual model of how NLP can be incorporated into practice. Then, we introduce the research setting under which we conducted this work. Next, we describe the collection of a set of simulated case presentations and the application of NLP methods to create a numeric representation of the presentation transcripts. Finally, we define the similarity score we used when ranking presentations on the basis of their likeness to a reference presentation and the parameters we used when performing a sensitivity analysis on the rankings.

Conceptual Model

We developed a conceptual model for workflows in which an automated system for grading trainee presentations in clinical care might be practically used (Figure 1). This model depicts three modes of assessing and providing individualized feedback to trainees. The first mode, traditional assessment, is the existing state; a more senior physician listens to a trainee present a case and provides feedback on the basis of the senior physician's subjective perceptions of the presentation's style, tone, structure, and content.

The second mode, automated comparative assessment, is a proposed system in which a trainee receives assessment and feedback without needing to have a senior physician present for the presentation. It works by having a senior physician and a trainee review a patient's electronic health record (EHR) data, and each presents the case to a microphone as if they were communicating the details of the case to other members of the patient's care team. Both audio recordings are transcribed using automatic speech recognition software, the transcripts are converted from human-readable text into machine-readable numbers using NLP methods, and the distance between the two numerical representations is calculated. A smaller distance corresponds to a more similar presentation. In this mode, the senior physician is considered the comparison subject who provides the reference presentation that the trainee's presentation is judged against.

The validity of automated comparative assessment depends on the expertise of the comparison subject. Oral case presentations are subject to significant variability among physicians owing partly to a widespread lack of teaching and assessment of this skill (1, 14, 15). Toalleviate concerns of inadequate expertise, an alternative formulation replaces the comparison subject with a reference standard created by a group of experts (16). These experts could include senior physicians of various subspecialties and perhaps even experts in other domains such as communication and statistics. Using a reference standard rather than a comparison subject may

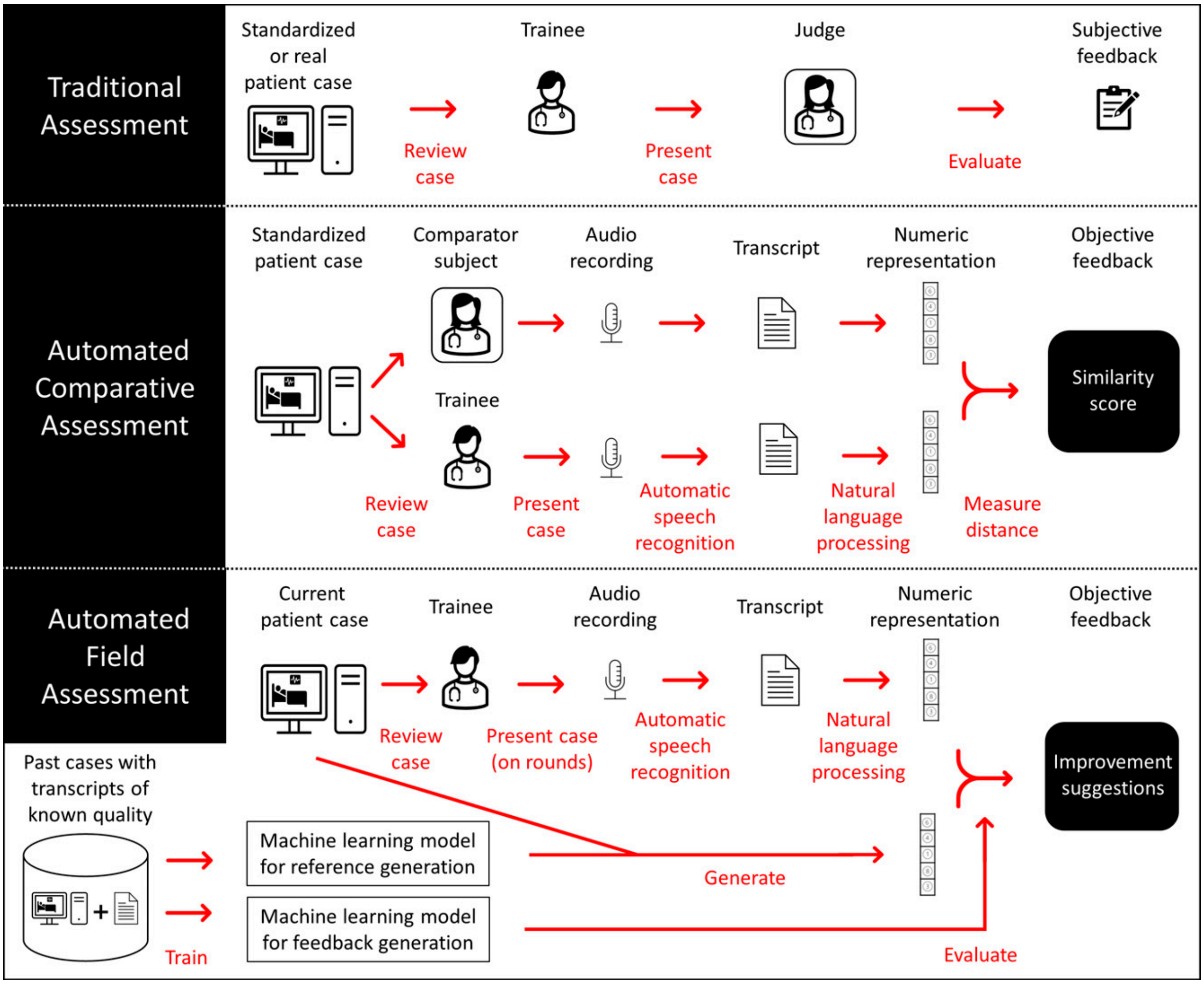

Figure 1. A conceptual model for assessment of oral case presentations. This conceptual model includes three modes: *1*) traditional assessment (the current state); *2*) automated comparative assessment; and *3*) automated field assessment. Automatic comparative assessment requires a reference presentation provided by a comparator subject, such as a senior physician. Both the trainee's and comparator subject's presentations are transcribed, transformed into a numeric representation using natural language processing, and assessed using a similarity score. The automatic field assessment is performed during actual clinical discussions. It requires rounds to be audio recorded and automatically transcribed using automatic speech recognition technology. It also requires a large set of past patient cases with transcripts of known quality. These data are used to train a pair of machine learning models: one for generating the numeric representation of a reference presentation when provided with a patient's electronic health record data and a second for generating feedback when provided with the numeric representations of two presentations of the same patient case.

lead to higher-quality presentation transcripts; however, it also increases the cost of creating them because more people would be involved, and forming a consensus is time-consuming.

The third mode, automated field assessment, is a proposed system in which a microphone is used during clinical rounds to audio record a trainee's actual oral case presentations at the point of care. The audio file is transcribed, converted to numbers, and evaluated against a computer-generated reference standard. This reference standard is generated by a machine learning model trained on a large set of past patient cases with transcribed presentations of known quality. Rather than just a similarity score, this system can also use a second machine learning model to generate trainee feedback if the set of past patient cases also includes examples of suggestions for improvement.

Variations of these automated approaches, and entirely different methods, are possible and expected. Rather than aiming for completeness, this conceptual model sparks a new area of investigation. The experiments included in this manuscript focus on automated comparative assessment.

Research Setting

The research was conducted in the Department of Critical Care Medicine at the University of Pittsburgh and UPMC (formerly the University of Pittsburgh Medical Center), an academic health system in western Pennsylvania. We focused our work on the intensive care unit (ICU) because it is an archetypal example of complex, team-based care in which the quality of case presentations may have an outsized impact on care quality and care team efficiency (17). The UPMC Department of Critical Care Medicine operates an integrated multidisciplinary fellowship subspecialty training program. Individuals with base training in internal medicine, emergency medicine, surgery, anesthesia, neurology, and neurosurgery receive comprehensive clinical training for 1–2 years. Trainees rotate through 10 multidisciplinary ICUs across five hospitals, and clinical training is augmented by additional didactic and simulation-based learning experiences (18). The United States Accreditation Council accredits all six programs for Graduate Medical Education.

Simulated Case Presentations

To provide an empirical proof-of-concept of this process, we focused on the more straightforward mode: automated comparative assessment. To do this, we transcribed audio files collected in a laboratory study in which critical care physicians used a novel EHR interface to review the medical records of cases of patients in the ICU before presenting those cases as if they were speaking to an attending physician on rounds (19, 20). The cases consisted of deidentified EHR data of patients admitted to UPMC ICUs who were admitted in 2012 and were experiencing either acute respiratory or renal failure. To simulate an actual ICU discussion, data were restricted to the time between hospital admission and a randomly selected day during the patient's ICU stay (all data after that day were censored) (21). Cases were viewed twice. During the first viewing, the physician was instructed to become familiar with the patient's clinical course since admission. During the second viewing, an additional 24 hours' worth of the patient's data was shown to the physician, who was instructed to prepare and present the case to a microphone as if they were speaking to an attending physician on rounds. The physicians were not instructed on how to present the cases or provided with any presentation templates.

The collection of the corpus was approved by the University of Pittsburgh Institute Review Board (IRB# PRO17050016). Contributions by author D.L.M. were covered under the University of Pennsylvania IRB (#833,938, determined as not human subjects research).

Numeric Representation of the Presentation Transcripts

To automatically assess presentation transcripts, the text must first be transformed into a representation that computers can use, specifically, converting text into numbers. NLP enables this process, but humans must make implementation decisions that affect what information is captured in the numeric representation. Informed by prior work on diagnostic communication (22–25), we implemented NLP methods to convert a transcript into a numeric representation that captures details about presentation style and content.

To represent presentation style, we defined features that relate to how information is conveyed during a case presentation. For example, the number of spoken sentences (i.e., utterances) per transcript and the average number of words, medical concepts, or medical adjectives per utterance. The calculation of these features was supported by two existing NLP tools: MetaMap (26) and pyConTextNLP (27).

To represent presentation content, we used a preexisting neural network model called Bio+ClinicalBERT (28). Neural networks are a class of machine learning models that consists of a series of interconnected nodes analogous to the interconnected neurons of an animal brain. Neural network models are valuable for NLP tasks because the networks can be developed on massive amounts of data independent of the model's specific use case (29). For example, Bio+ClinicalBERT was trained on a combination of PubMed abstracts, PubMed Central manuscripts, and clinical notes from the MIMIC (Medical Information Mart for Intensive Care) dataset (30). During training, the model learns associations between words and stores this knowledge in its connections. To use this model to generate a numeric representation of a presentation's content, we input a transcript into the model and extract numeric values from the model's nodes (29).

Similarity Score

A similarity score is calculated from the distance between the numeric representations of two presentation transcripts: the trainee's and the comparator subject's. Because our numeric representation consists of two components, style and content, we defined a similarity score that first calculates the similarity within each component and then averages the component similarity to calculate the transcript similarity. The full equation for what we call the Style–Content Similarity Score is provided in Equation E1 in the data supplement.

We calculated the similarity score for each fellow–attending pair to rank the fellows across each case. We considered the attending to be the comparison subject because they had the most critical care experience and were, therefore, best positioned to provide the reference presentations. We acknowledge that other approaches for choosing a comparison subject are possible, including having one or more judges review the transcripts to select the best presentation for each case.

Sensitivity Analysis

In the ranking experiment, presentation style and content components were given equal weight (50% style information and 50% content information). It is also possible to calibrate the similarity score by favoring one component over the other. This is achieved using different weight values in the similarity score equation. To show that different weights result in different rankings, we conducted a sensitivity analysis by varying weighting values in 5% increments from 100% to 0% style information and from 0% to 100% content information. We plot the average number of rank changes in the fellow rankings across all cases as the weighting varies.

All analyses were performed in Python (3.7.9) using Matplotlib (3.1.1), NumPy (1.21.1), Pandas (1.3.0), pyConTextNLP (0.7.0.1), PyTorch (1.7.1), SciPy (1.7.0), Tokenizers (0.9.4), and Transformers (4.1.1).

RESULTS

The data supplement provides the characteristics of the nine critical care physicians (eight fellows and one attending) and 11 patient cases (Tables E1 and E2). The corpus contained 260 minutes of presentation (mean length: 2.6 ± 1.2 min per case). A summary of the values from the numeric representation of presentation style is shown as a series of radar plots in Figure 2 (31). The figure demonstrates substantial variation across trainees in all the features shown. Trainees B, C, and D appear notably different from other trainees in terms of the number of utterances, tending to have shorter presentations overall.

Figure 3 shows the rankings of the trainees' presentations across the patient cases on the basis of their Style–Content Similarity Scores. This figure demonstrates a notable separation between the top five performing trainees (A, G, H, E, and F) and the bottom three performing trainees (B, C, and D). Two example transcripts are provided in Table E3. These transcripts are the most similar (trainee H) and least similar (trainee B) presentations to the reference presentation for Case 11. Sensitivity analyses demonstrate that the ranking of trainees varies as the relative importance is shifted between more strongly favoring style versus content information (Figure 4). This expected behavior means that these weighting terms can be used to calibrate the system.

DISCUSSION

On a corpus of 99 transcribed case presentations, we examined the feasibility

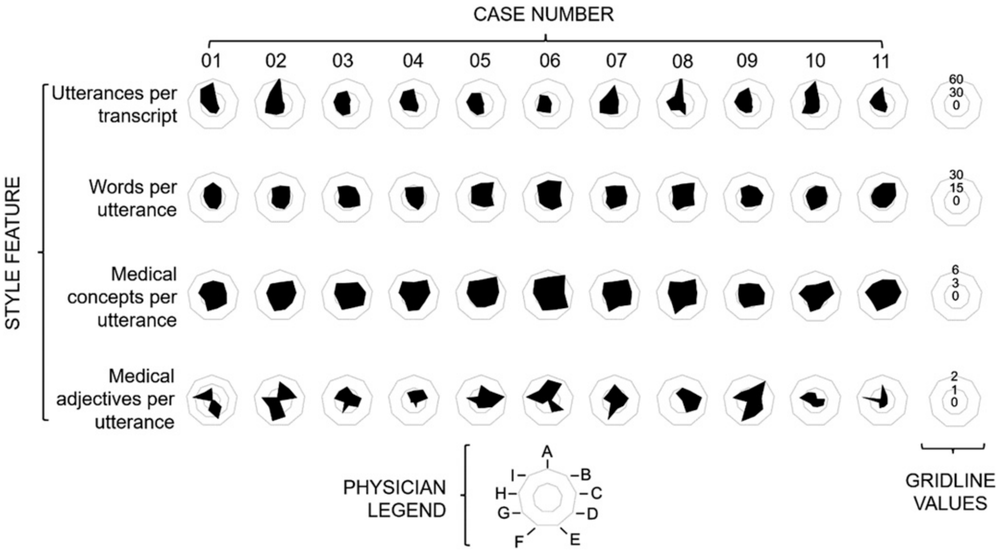

Figure 2. Radar plots showing the values of features from the numeric representation of presentation style. Each radar plot shows the calculated values of one feature for all presentations of a patient case. Each row of plots corresponds to one feature. Each column of plots corresponds to one case (01–11). Within a plot, each corner corresponds to one physician (A–I). The center of each plot is a value of 0; values increase as you move outwards toward the circular gridlines. Gridline values for each row of plots are shown on the right side. An utterance is analogous to a spoken sentence.

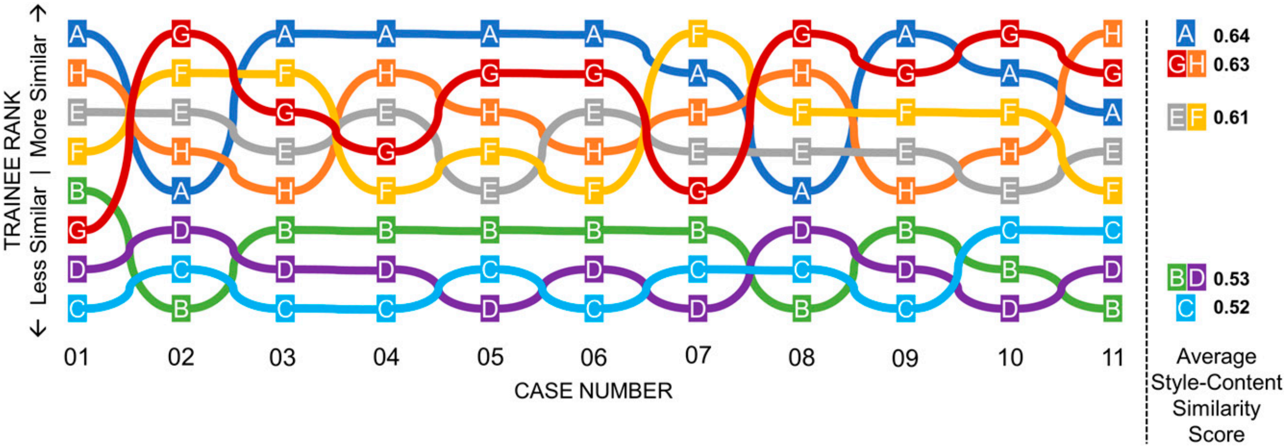

Figure 3. Yarn diagram of automated ranking of eight trainees who presented 11 patient cases. The diagram depicts trainee rank (A–H) for each case (01–11). Trainee rank is on the basis of how similar a trainee's presentation was to the comparator subject's presentation. Similarity was calculated using the Style–Content Similarity Score. The strings connecting a physician's position across columns illustrate changes in rank from one case to the next. Average scores and associated rankings are shown on the right side.

of automatically assessing presentation quality with respect to the presentations of the most experienced physician.

To enable the comparison, we applied NLP to convert each transcript into a numeric representation. We calculated the similarity between presentations using a similarity score that accounts for both the style and content components of the numeric representation. This approach successfully separated trainees according to how similar their presentations were to

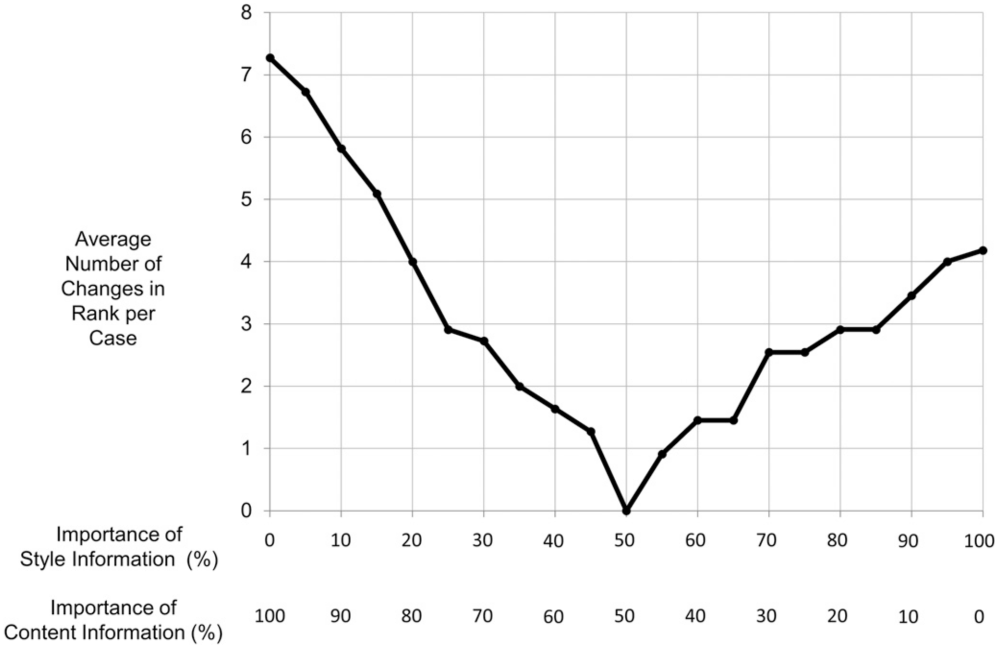

Figure 4. Sensitivity of rank to different weightings of presentation style and content importance. The weighted average between style and content importance was modified in 5% increments from 100% to 0% and 0% to 100%, respectively. The *y-axis* shows the average number of changes in rank per case as the weighting varies. The baseline rank is when importance is split evenly (50% each).

a comparator subject's presentations, providing essential proof of concept.

Strengths and Limitations

Our approach has several advantages over the existing paradigm for giving trainees feedback on their oral presentations. First, it is objective in that it does not rely on the subjective interpretations and availability of senior physicians who are subject to known heuristics and biases. Second, it is quantitative, such that trainees can be compared with each other on the basis of a standard provided by an expert comparator subject. And third, with further development, it will be actionable in that the different components of the final score can be used to give trainees specific and individualized feedback on how to improve. Such feedback could instruct trainees on how to improve their presentation structure, content, repetition of important topics, and degree of detail. Once fully developed, automated comparative assessment could be implemented as an online tool with standardized patient cases in which medical students and residents could visit a website and test their skills against other trainees from across the country or track their progress over time. The adoption of this tool would generate a corpus of rounding presentations with known quality. Such data would be paramount for pretraining the models required for automated field assessment.

This work builds on and extends prior efforts to leverage technology to improve medical education. For example, our conceptual model is at the intersection of the work of Callahan and colleagues, who propose applying technology to address the need for teaching communication skills at a larger scale (32), and Green and colleagues, who are creating guidelines for improving oral case presentation skills (33). Our methods are intended to augment, not replace, other assessments of technical skills (34, 35).

This work has several limitations. First and most importantly, we demonstrate only preliminary proof-of-concept using transcripts obtained during simulated presentations. More conceptual work is needed to determine the ideal features by which presentations should be judged. More empirical work is required to develop and test the system before using it in medical education. Collecting "high-quality" presentations that can serve as training data for the NLP models is a critical next step. In this first step, we tested having a senior physician serve as a comparator subject, with them presenting the same cases as the trainees for our automated comparative assessment. An alternative approach would use a reference standard created by either averaging across the presentations of multiple senior physicians or by enlisting a team of experts to create a census presentation. Each approach has merit and includes tradeoffs of time and resources needed to produce a reference presentation for each case. Future evaluations would evaluate the system's face validity by comparing its presentation assessments against those created by expert human raters. In addition, we did not control for the EHR interface used because our focus was on evaluating methods for measuring differences in case presentations, not on determining why those differences exist.

A second limitation is that our methods require the introduction of technology to record and automatically transcribe case presentations. Although this technology currently exists, it is not routinely used in medical education. In live clinical settings, recording case presentations introduces additional concerns related to privacy and data security. Yet there is precedence for video recording trainees as part of medical training, and we suspect that with time there will be growing acceptance of using voice technology as a training tool. Furthermore, we did not apply or evaluate automatic speech recognition (also known as speech-to-text) technology on our transcripts, opting instead to rely on human transcriptionists. A separate formative evaluation of automatic speech recognition technologies and summative evaluation of the system's complete workflow are warranted.

Third, it is also important to consider the trainee's stage of professional development when developing these systems. All our trainees were critical care fellows and were, therefore, relatively advanced in their training. The same comparators and assessment algorithms might not be appropriate for assessing the performance of more junior trainees like residents or medical students because the reference presentations should correspond to the expectations placed on the trainee.

Finally, as with other automated solutions, there is a risk that automated algorithms could paradoxically reinforce ingrained biases rather than mitigate them (36, 37). Specifically, care will need to be taken to ensure that this approach does not harm students from disadvantaged backgrounds or those with English as a second language who may use speech patterns that differ from historical norms. Another point to consider when assessing rounding presentations is that they do not occur in a vacuum. They are part of the multidisciplinary discussion in which team members dynamically interact. So, team dialog must be accounted for to not penalize a trainee for holding discussions with the team.

Conclusions

We presented a conceptual model for providing trainees with consistent and objective feedback on oral presentations and applied this model to simulated presentations. This model contains two novel modes of automated assessment. Automated comparative assessment is intended to be used in educational settings and requires a reference presentation provided by a comparator subject or subjects. Automated field assessment is intended to be used during real conversations in actual practice, such as during multidisciplinary rounds. It requires a training corpus of presentations with known quality. In either mode, case presentations need to be transcribed and processed with NLP. There is a need for further research into how oral case presentation style and content are best represented and assessed.

We envision a future in which technology is used to augment the training of medical students and junior physicians by providing automated assessment and individualized feedback on oral case presentations. The methods presented here join a growing toolbox for medical assessment and training (6). Technologies like these may eventually be commonplace, regularly providing professional feedback to all members of multidisciplinary care teams (38).

Acknowledgment

The corpus was collected with guidance from Drs. Gilles Clermont, Shyam Visweswaran, and Harry Hochheiser. The authors also thank Drs. Lori Shutter and Derek Angus for helpful discussions that lead to these experiments.

Author disclosures are available with the text of this article at www.atsjournals.org.

REFERENCES

1. Williams DE, Surakanti S. Developing oral case presentation skills: peer and self-evaluations as instructional tools. *Ochsner J* 2016;16:65–69.

2. Holmboe ES. Faculty and the observation of trainees' clinical skills: problems and opportunities. *Acad Med* 2004;79:16–22.

3. Newble D, Dawson B, Dauphinee D, Page G, Macdonald M, Swanson D, *et al.* Guidelines for assessing clinical competence. *Teach Learn Med* 1994;6:213–220.

4. Kerlin MP, Harhay MO, Vranas KC, Cooney E, Ratcliffe SJ, Halpern SD. Objective factors associated with physicians' and nurses' perceptions of intensive care unit capacity strain. *Ann Am Thorac Soc* 2014;11:167–172.

5. Haber RJ, Lingard LA. Learning oral presentation skills: a rhetorical analysis with pedagogical and professional implications. *J Gen Intern Med* 2001;16:308–314.

6. Epstein RM, Hundert EM. Defining and assessing professional competence. *JAMA* 2002;287: 226–235.

7. Kassebaum DG, Eaglen RH. Shortcomings in the evaluation of students' clinical skills and behaviors in medical school. *Acad Med* 1999;74:842–849.

8. Bransford J, Brown A, Cocking R, editors. *How people learn: brain, mind, experience, and school: expanded edition.* Washington, DC; National Academies Press; 2000.

9. Melvin L, Cavalcanti RB. The oral case presentation: a key tool for assessment and teaching in competency-based medical education. *JAMA* 2016;316:2187–2188.

10. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc* 2011;18:544–551.

11. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, *et al.* Deep learning in clinical natural language processing: a methodical review. *J Am Med Inform Assoc* 2020;27:457–470.

12. Yan MY, Gustad LT, Nytrø Ø. Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review. *J Am Med Inform Assoc* 2022;29:559–575.

13. Alam F, Afzal M, Malik KM. Comparative analysis of semantic similarity techniques for medical text. *2020 International Conference on Information Networking (ICOIN)* 2020;106–109.

14. Kroboth FJ, Hanusa BH, Parker S, Coulehan JL, Kapoor WN, Brown FH, *et al.* The inter-rater reliability and internal consistency of a clinical evaluation exercise. *J Gen Intern Med* 1992;7: 174–179.

15. Taylor TAH, Swanberg SM. A comparison of peer and faculty narrative feedback on medical student oral research presentations. *Int J Med Educ* 2020;11:222–229.

16. Hripcsak G, Wilcox A. Reference standards, judges, and comparison subjects: roles for experts in evaluating system performance. *J Am Med Inform Assoc* 2002;9:1–15.

17. Diabes MA, Ervin JN, Davis BS, Rak KJ, Cohen TR, Weingart LR, *et al.* Psychological safety in intensive care unit rounding teams. *Ann Am Thorac Soc* 2021;18:1027–1033.

18. Tour our intensive care units. University of Pittsburgh Department of Critical Care Medicine; 2022 [accessed 2022 Sept 1]. Available from: https://ccm.pitt.edu/node/1191.

19. King AJ, Cooper GF, Clermont G, Hochheiser H, Hauskrecht M, Sittig DF, *et al.* Using machine learning to selectively highlight patient information. *J Biomed Inform* 2019;100:103327.

20. King AJ, Calzoni L, Tajgardoon M, Cooper GF, Clermont G, Hochheiser H, *et al.* A simple electronic medical record system designed for research. *JAMIA Open* 2021;4:ooab040.

21. King AJ, Cooper GF, Clermont G, Hochheiser H, Hauskrecht M, Sittig DF, *et al.* Leveraging eye tracking to prioritize relevant medical record data: comparative machine learning study. *J Med Internet Res* 2020;22:e15876.

22. Bordage G, Lemieux M. Semantic structures and diagnostic thinking of experts and novices. *Acad Med* 1991; 66(9, Suppl)S70–S72.

23. Bordage G, Zacks R. The structure of medical knowledge in the memories of medical students and general practitioners: categories and prototypes. *Med Educ* 1984;18:406–416.

24. Greimas AJ. *Structural semantics: an attempt at a method.* Lincoln, NE: University of Nebraska Press; 1983.

25. Gibson E, Futrell R, Piantadosi SP, Dautriche I, Mahowald K, Bergen L, *et al.* How efficiency shapes human language. *Trends Cogn Sci* 2019;23:389–407.

26. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010;17:229–236.

27. Chapman BE, Lee S, Kang HP, Chapman WW. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. *J Biomed Inform* 2011;44: 728–737.

28. Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, *et al.* Publicly available clinical BERT embeddings [preprint]. arXiv preprint arXiv:190403323; 2019 [accessed 2022 Sept 1]. Available from: https://arxiv.org/abs/1904.03323.

29. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding [preprint]. arXiv preprint arXiv:181004805; 2018 [accessed 2022 Sept 1]. Available from: https://arxiv.org/abs/1810.04805.

30. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3:160035.

31. Saary MJ. Radar plots: a useful way for presenting multivariate health care data. *J Clin Epidemiol* 2008;61:311–317.

32. Callahan ME, Brant EB, Mohan D, Norman MK, Arnold RM, White DB. Leveraging technology to overcome the “scalability problem” in communication skills training courses. *ATS Scholar* 2021;2: 327–340.

33. Green EH, Hershman W, DeCherrie L, Greenwald J, Torres-Finnerty N, Wahi-Gururaj S. Developing and implementing universal guidelines for oral patient presentation skills. *Teach Learn Med* 2005;17:263–267.

34. Adamson R, Morris AE, Sun Woan J, Ma IWY, Schnobrich D, Soni NJ. Development of a focused cardiac ultrasound image acquisition assessment tool. *ATS Scholar* 2020;1:260–277.

35. Lam K, Chen J, Wang Z, Iqbal FM, Darzi A, Lo B, *et al.* Machine learning for technical skill assessment in surgery: a systematic review. *NPJ Digital Medicine* 2022;5:1–16.

36. Spring J, Abrahams C, Ginsburg S, Piquette D, Guasch FM, Kiss A, *et al.* Impact of gender on clinical evaluation of trainees in the intensive care unit. *ATS Scholar* 2021;2:442–451.

37. Capers Q IV. How clinicians and educators can mitigate implicit bias in patient care and candidate selection in medical education. *ATS Scholar* 2020;1:211–217.

38. Rak KJ, Kahn JM, Linstrum K, Caplan EA, Argote L, Barnes B, *et al.* Enhancing implementation of complex critical care interventions through interprofessional education. *ATS Scholar* 2021;2:370–385.