Detection of Residents With Progress Issues Using a Keyword–Specific Algorithm

Gaby Tremblay, MSc
Pierre-Hugues Carmichael, MSc
Jean Maziade, MD, FCMF, CCMF, MSc
Mireille Grégoire, MDCM, FRCPSC

ABSTRACT

Background The literature suggests that specific keywords included in summative rotation assessments might be an early indicator of abnormal progress or failure.

Objective This study aims to determine the possible relationship between specific keywords on in-training evaluation reports (ITERs) and subsequent abnormal progress or failure. The goal is to create a functional algorithm to identify residents at risk of failure.

Methods A database of all ITERs from all residents training in accredited programs at Université Laval between 2001 and 2013 was created. An instructional designer reviewed all ITERs and proposed terms associated with reinforcing and underperformance feedback. An algorithm based on these keywords was constructed by recursive partitioning using classification and regression tree methods. The developed algorithm was tuned to achieve 100% sensitivity while maximizing specificity.

Results There were 41 618 ITERs for 3292 registered residents. Residents with failure to progress were detected for family medicine (6%, 67 of 1129) and 36 other specialties (4%, 78 of 2163), while the positive predictive values were 23.3% and 23.4%, respectively. The low positive predictive value may be a reflection of residents improving their performance after receiving feedback or a reluctance by supervisors to ascribe a "fail" or "in difficulty" score on the ITERs.

Conclusions Classification and regression trees may be helpful to identify pertinent keywords and create an algorithm, which may be implemented in an electronic assessment system to detect future residents at risk of poor performance.

Introduction

Early identification of a resident with progress difficulties who is enrolled in postgraduate medical training is an ongoing challenge. In various studies, between 4.3% and 9.1% of residents show evidence of struggling during training.[1,2] The learning difficulties of residents are frequently identified late in their training,[3] as assessors are often reluctant to mark "in difficulty" or "failure" on in-training evaluation reports (ITERs), or to designate "fail" on other end-of-rotation assessment forms.[4,5] However, the length of narratives and percentage of ambiguous or negative comments on rotation assessments, such as ITERs, indicate a potential need for resident remediation.[3,6] Narrative comments in assessing trainees have been shown to be valuable,[7,8] and must be taken into account when determining learners' progress toward achieving competencies.[9–11]

Although the tools of language analytics have been applied in education, very few articles concerning language analytics in medical education have been published, and those that have tend to focus on undergraduate students.[12] In order to handle all the narrative information becoming available as part of competency-based medical education, we sought to develop a novel computerized semantic analysis, which consists of an algorithm that is able to detect residents with progress issues, based on certain keywords.

Methods

A database containing all ITERs (forms indicating whether preset objectives are met, corresponding narratives, overall score [pass/in difficulty/fail], and general comments) from all residents training in accredited programs at Université Laval between 2001 and 2013 was extracted and anonymized to ensure confidentiality of their track records. The ITERs were split into either family medicine or Royal College of Physicians and Surgeons of Canada (RCPSC) programs (provided as online supplemental material), as the ITER format used in family medicine for the period covered in this study differed significantly from that of other residency programs. The databases included the name of the program, residency level, rotation block number, residency beginning

DOI: http://dx.doi.org/10.4300/JGME-D-19-00386.1

Editor's Note: The online version of this article contains a list of Royal College of Physicians and Surgeons of Canada programs, a list of positive and negative keywords with an English translation, and the technical statistical methodology.

and end dates, residency site, CanMEDS[13] role assessments (ratings and comments), overall rotation evaluation (comments), and number of days of absence during the training period. In general, ITERs were completed by attending faculty within 30 days of rotation completion.

For each database we identified residents with progress issues, defined as having an ITER either rated "in difficulty" (ie, struggling) or "fail." For the purposes of the study, all ITERs from a given resident were kept until a form with the mention "failure" or "in difficulty" appeared. All ITERs following an ITER with identified progress issues were discarded.

An instructional designer reviewed all ITERs written in French and proposed terms associated with positive feedback and underperformance. Terms were determined from 133 216 words entered in the overall performance comments section and 84 365 words entered in the narrative section of each CanMEDS role of the ITERs. The first half of the database was used to make a list of positive and negative keywords that was checked for consistency against the second half of the database. French words that could have a dual meaning (either positive or negative) and conjunctions were discarded. The practical significance of this list was confirmed by the associate dean of postgraduate medical education and by a nonmedical member of the faculty of medicine. The list of these keywords with an English translation is provided as online supplemental material.

A classification rule based on these keywords was constructed by recursive partitioning using classification and regression tree methods.[14] This methodology was preferred due to its flexibility in automatically selecting variables and cutoff values and its ability to produce relatively simple classification rules. Technical methodologic details are presented as online supplemental material. The classification and regression tree algorithm was applied independently to the family medicine and specialized programs data sets and tuned with the aim of obtaining rules with near 100% sensitivity (proportion of actual positives correctly identified as such) and maximal specificity (proportion of actual negatives correctly identified as such). Sensitivity, specificity, and positive and negative predictive values (proportion of true positives and true negatives, respectively) were computed for each derived classification rule. The algorithm was compared to the stringent standard of "fail" or "in difficulty" overall score, either concurrent to the keyword or anytime thereafter. We present only data of the final rules obtained for each data set. Due to the low prevalence of struggling residents, data sets were not split into training and testing sets to avoid increasing performance variability.[15]

What was known and gap
Identification of progress issues in residents often occurs late in training. Semantic analysis of narrative information as part of competency-based medical education can be useful for detecting problems earlier.

What is new
An algorithm that uses keywords from in-training evaluation reports to identify residents at risk of failure.

Limitations
Data were collected from a single site; linguistic patterns could be a result of institutional culture.

Bottom line
Systematic monitoring of resident progress through a prospective computerized semantic analysis using an algorithm may be an effective way to identify residents with progress difficulties.

The Université Laval Ethics Board exempted this project from review.

Statistical analyses were carried out with R 3.2.3 (R Foundation for Statistical Computing, Vienna, Austria).

Results

There was a total of 41 618 ITERs for the 3292 registered residents. The RCPSC database contained 30 073 ITERs from 2163 residents (60% female) training in 36 accredited programs at Université Laval between May 2002 and November 2013. The family medicine database was composed of 11 545 ITERs from 1129 residents (73% female) training between August 2001 and September 2013. There are currently 910 residents registered in the 50 rural and urban, university-based accredited training programs.

Table 1 presents the performance of the chosen rule when classifying residents enrolled in an RCPSC residency program. This classification rule achieves 100% sensitivity while maximizing specificity at 87.7%. In this particular group, progress issues were identified in 78 of the 2163 residents (4%). The classification tree correctly classified these residents as having progress issues. However, the classification tree identified 256 residents as having progress issues, although they did not have an overall score indicating difficulty or failure, resulting in a positive predictive value of 23.4%.

Table 2 presents the performance of the chosen classification rule when classifying residents enrolled in the family medicine program. The classification rule achieves 100% sensitivity while maximizing specificity at 79.2%. In this group, progress issues were identified in 67 of 1129 residents (6%). The classification rule identifies these residents correctly, but it identified 221 residents as having progress

TABLE 1
Statistical Analysis Obtained by Classification Tree for Residents Enrolled in Royal College of Physicians and Surgeons of Canada Specialty Program[a]

		Progress Issues				
		Yes	No	Total		
Decision Tree	Positive	78	256	334	Positive Predictive Value	23.4%
	Negative	0	1829	1829	Negative Predictive Value	100%
		78	2085	2163	Prevalence	3.6%
		Sensitivity	Specificity			
		100%	87.7%			

[a] Progress issues are defined as an "in difficulty" or "fail" overall in-training evaluation report score.

TABLE 2
Statistical Analysis Obtained by Classification Tree for Residents Enrolled in a Family Medicine Program[a]

		Progress Issues				
		Yes	No	Total		
Decision Tree	Positive	67	221	288	Positive Predictive Value	23.3%
	Negative	0	841	841	Negative Predictive Value	100%
		67	1062	1129	Prevalence	5.9%
		Sensitivity	Specificity			
		100%	79.2%			

[a] Progress issues are defined as an "in difficulty" or "fail" overall in-training evaluation report score.

issues, although they did not have an overall score indicating difficulty or failure, resulting in a positive predictive value of 23.3%.

FIGURES 1 and 2, respectively, present the chosen classification rules for RCPSC and family medicine residents in classification tree form. Each node of the tree represents a simple binary criterion, which includes both a keyword and its frequency, with movement down the tree going to the left when the criterion is met.

Discussion

In this retrospective study, we were able to demonstrate that an algorithm based on keywords associated with a suboptimal performance would help a program director identify a struggling resident. The algorithm correctly ranked all residents who had difficulty progressing, as evidenced by the 100% sensitivity and 100% negative predictive value.

This ability of the algorithm to detect all residents with progress issues is embedded in its design. Specificity was maximized, knowing that this compromise would give a lower positive predictive value. In our opinion, the consequences of delaying the detection of a resident in difficulty are much more important than reviewing the file of an otherwise well-performing resident. A total of 334 residents were identified as having progress issues in the RCPSC data set (FIGURE 1). However, this database includes assessment forms from 36 programs and covers a period of 12 years. Therefore, each year, on average, a program director would have to review less than 1 resident file that was falsely identified by the algorithm as being in difficulty. As for family medicine, the algorithm presented in FIGURE 2 proved to be the most effective in detecting residents in difficulty (100% sensitivity and negative predictive value). Considering the large size of the family medicine program, the 221 false positives over the 12-year period represent only 1 or 2 cases per year per teaching site, among a group of residents already well known to the teaching site director. For the purpose of this study, the standard against which the algorithm was tested is the overall global score. It is likely that some of the false positives represent struggling residents who improved their performance following feedback from their supervisors, and joined the well-performing cohort thereafter. Alternatively, supervisors could describe underperformance in the narratives without assigning the corresponding overall score "in difficulty" or "fail."[4,5] Therefore, some false positives may include true strugglers.

The algorithms in FIGURES 1 and 2 highlight a series and frequency of keywords needed to detect struggling residents. They also provide some insight into the evaluation practices of assessors. For example, the high frequency of "good" suggests that this word is generally overused by assessors, as it is commonly used in most ITERs, even to describe struggling residents. The keyword "lag" has coincided with several occurrences of the keyword "excellent" (FIGURE 1). Likewise, some positive keywords, such as "interested," were associated with negative performance, as indicated in FIGURE 2. This might suggest that specific encouragement wording may be preferentially used in ITERs of struggling residents.

The results of this study parallel the findings found in a previous study of 34 internal medicine residents' ITERs, reviewed individually by blinded faculty members, in which the number of words in the comment section and the percentage of ITERs with negative or ambiguous comments were associated with serious progress issues.[6] Using a similar design, a retrospective study of general surgical residents demonstrated that 84% of struggling residents could be identified in their first year of training.[16]

Considering that automated essay scoring of the first part of the Canadian Medical Council examinations has been shown to be reliable,[17] the use of an automated computer semantic analysis could facilitate the work of program directors and the office of postgraduate medical education. Given the low prevalence of residents in difficulty, a keyword approach would be a valuable flagging tool for program directors with large resident cohorts and those with little experience, as well as for the postgraduate associate dean. A subanalysis of ITERs for each program would have been an interesting

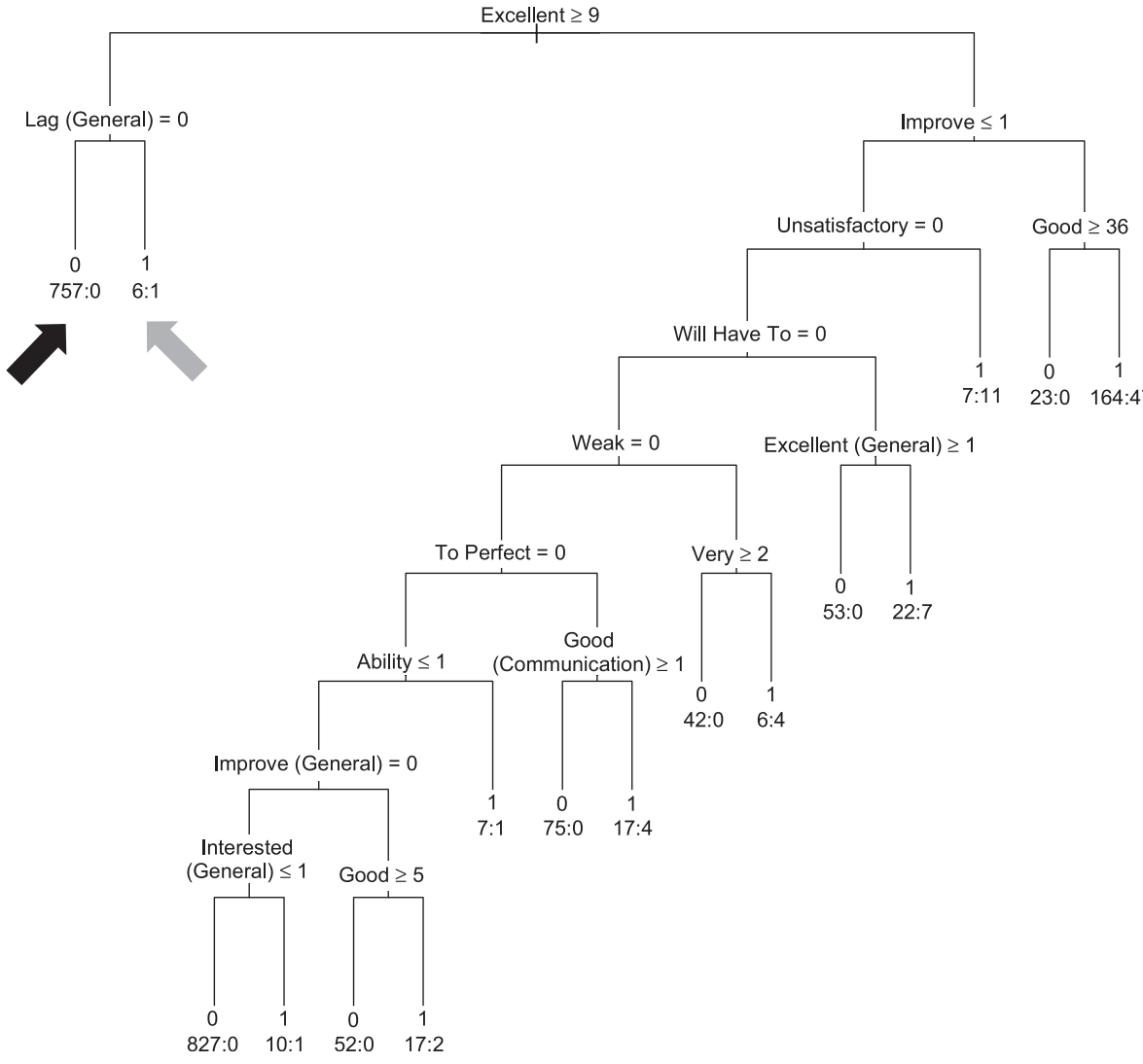

FIGURE 1

Classification Decision Tree for Residents of Royal College of Physicians and Surgeons of Canada Specialties

Note: Each node of the tree represents a simple binary criterion, which includes both a definite keyword and its frequency, with movement down the tree going to the *left* when the criterion is met. When a resident has 9 or more occurrences of the keyword "excellent" in all of his or her in-training evaluation reports (ITERs), the system then looks for the keyword "lag" in the general comment section. If it is absent (lag = 0), then the resident is classified as having no progress issues (in the "0" branch to the left as shown by the black arrow). 757:0 indicates that 757 residents were not in difficulty and none had issues following this algorithm. If "lag" occurs at least once in all ITERs ("lag" ≠ 0), the resident is classified as having progress issues (in the "1" branch as shown by the gray arrow). 6:1 indicates that 6 residents did not have issues and 1 did. Words between parentheses indicate the specific section of the ITER where the keyword is searched. "General" indicates the overall performance section of the ITER. When no words between parentheses appear under the frequency of occurrences, the system searches all sections of the ITER.

addition, but the low rate of residents who experienced progress difficulties and data confidentiality concerns made this impossible.

While this algorithm made it possible to accurately identify residents in the database who have shown progress issues, it remains unable to determine the lead time between the first use of the negative keywords and the global ITER rating of "in difficulty" or "fail." Moreover, a computerized algorithm does not understand the subtleties in the diplomatic language sometimes used in assessing trainees.[11] The negative predictive value and sensitivity could also vary with a different data set. Another limitation was the separate data set for family medicine residents, since at the time of the study this program used an ITER form that was significantly different from the RCPSC specialties. Moreover, all data were collected at a single university. Thus, some of the linguistic patterns could be a result of a broader institutional culture, potentially limiting its generalizability. Language could be used differently according to the gender[18] or ethnic background of trainees, inducing a

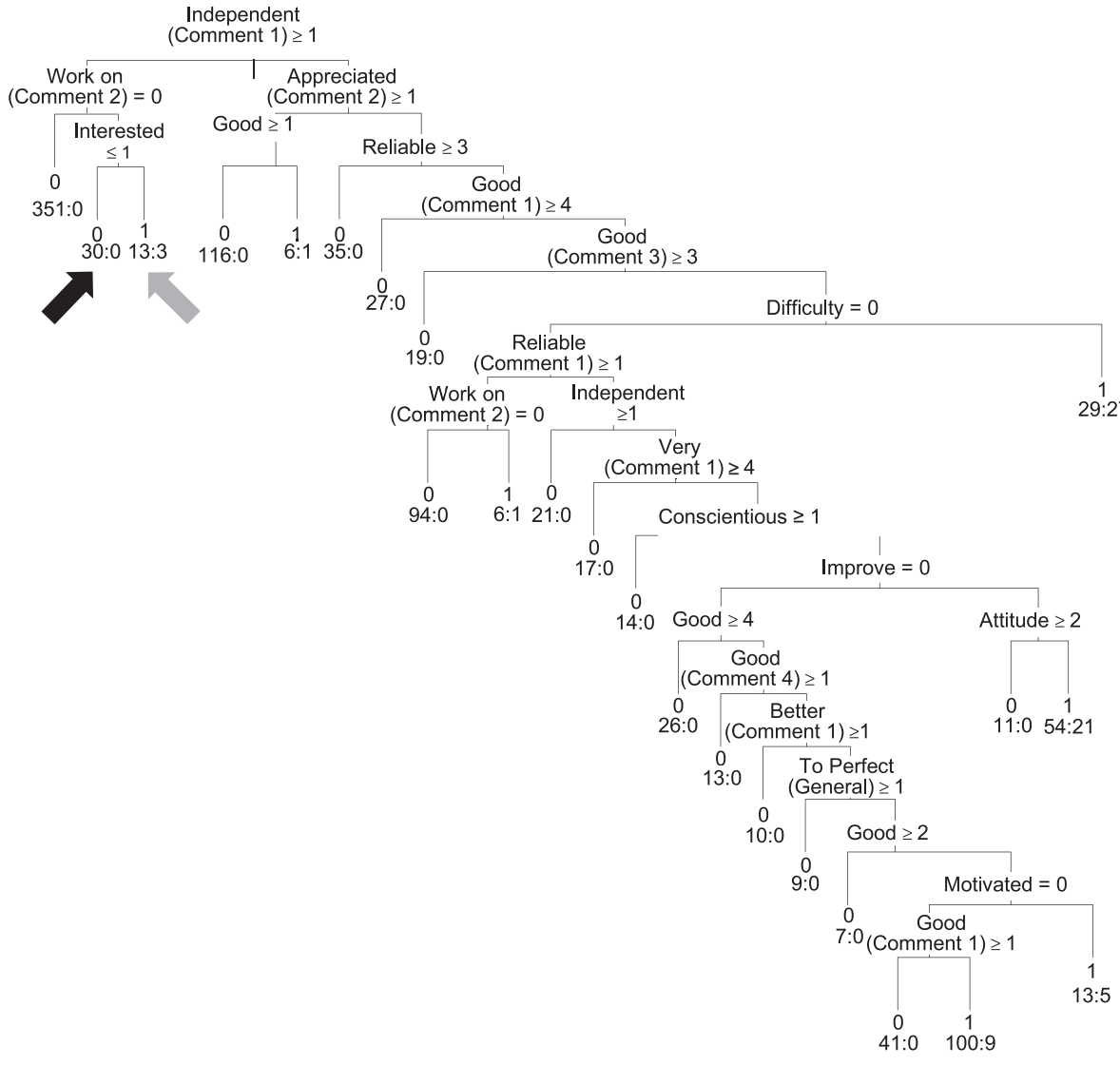

FIGURE 2

Classification Decision Tree for Family Medicine

Note: Comment 1 refers to the medical expert section of the family medicine in-training evaluation report (ITER), comment 2 to communication skills and professionalism, comment 3 to collaboration and management, and general to the overall performance section. Words between parentheses indicate the specific section of the ITER where the keyword is searched. When no words between parentheses appear under the frequency of occurrences, the system searches all sections of the ITER.

Each node of the tree represents a simple binary criterion, which includes both a definite keyword and its frequency, with movement down the tree going to the *left* when the criterion is met. When a resident has more than 1 occurrence of the keyword "independent" in the medical expert section of all ITERs, the system then looks for the keyword "work on" in the communication and professionalism section. If it is present ("work on" $\neq$ 0), the system looks for "interested" in all sections of the ITER. If there are fewer than 2 occurrences, the resident is classified as having no progress issues (in the "0" branch to the left as shown by the black arrow). 30:0 indicates that 30 residents were not in difficulty and none had issues following this algorithm. If "interested" occurs at least twice in all ITERs (interested not $<$ 2), the resident is classified as having progress issues (in the "1" branch as shown by the gray arrow). 13:3 indicates that 13 residents in fact did not have issues and 3 did.

bias. If used inappropriately, such an algorithm could lead to false labeling of residents as strugglers. Finally, for publishing purposes, the keywords of the algorithms presented in this article were translated into English, but the statistical analysis was done using ITERs written in French. Using this algorithm in a language other than French would require transcultural validation of the keywords.

Additional prospective analyses are required to provide validity evidence for the use of the keywords of the algorithm in current cohorts. Further study to assess the algorithm's efficacy for earlier detection of underperforming trainees and to determine whether one set of keywords could be used for all programs are key next steps now that all ITERs share the same structure.

Conclusions

Systematic monitoring of resident progress through a prospective computerized semantic analysis using an algorithm derived from a classification with regression trees may be an effective way to identify residents in difficulty, especially given the need to analyze increasing numbers of narrative evaluations as part of competency-based medical education.

References

1. Reamy BV, Harman JH. Residents in trouble: an in-depth assessment of the 25-year experience of a single family medicine residency. *Fam Med*. 2006;38(4):252–257.
2. Park YS, Riddle J, Tekian A. Validity evidence of resident competency ratings and the identification of problem residents. *Med Educ*. 2014;48(6):614–622. doi:10.1111/medu.12408.
3. Schwind CJ, Williams RG, Boehler ML, Dunnington GL. Do individual attendings' post-rotation performance ratings detect residents' clinical performance deficiencies? *Acad Med*. 2004;79(5):453–457. doi:10.1097/00001888-200405000-00016.
4. Yepes-Rios M, Dudek N, Duboyce R, Curtis J, Allard RJ, Varpio L. The failure to fail underperforming trainees in health professions education: a BEME systematic review: BEME Guide No. 42. *Med Teach*. 2016;38(11):1092–1099. doi:10.1080/0142159X.2016.1215414.
5. Dudek NL, Marks MB, Regehr G. Failure to fail: the perspectives of clinical supervisors. *Acad Med*. 2005;80(10 suppl):84–87. doi:10.1097/00001888-200510001-00023.
6. Guerrasio J, Cumbler E, Trosterman A, Wald H, Brandenburg S, Aagaard E. Determining need for remediation through postrotation evaluations. *J Grad Med Educ*. 2012;4(1):47–51. doi:10.4300/JGME-D-11-00145.1.
7. Hatala R, Sawatsky AP, Dudek N, Ginsburg S, Cook DA. Using in-training evaluation report (ITER) qualitative comments to assess medical students and residents: a systematic review. *Acad Med*. 2017;92(6):868–879. doi:10.1097/ACM.0000000000001506.
8. Ginsburg S, van der Vleuten CPM, Eva KW. The hidden value of narrative comments for assessment: a quantitative reliability analysis of qualitative data. *Acad Med*. 2017;92(11):1617–1621. doi:10.1097/ACM.0000000000001669.
9. Cohen GS, Henry NL, Dodd PE. A self-study of clinical evaluation in the McMaster clerkship. *Med Teach*. 1990;12(3–4):265–272.
10. Frohna A, Stern D. The nature of qualitative comments in evaluating professionalism. *Med Educ*. 2005;39(8):763–768. doi:10.1111/j.1365-2929.2005.02234.x.
11. Ginsburg S, Regehr G, Lingard L, Eva KW. Reading between the lines: faculty interpretations of narrative evaluation comments. *Med Educ*. 2015;49(3):296–306. doi:10.1111/medu.12637.
12. Saqr M. A literature review of empirical research on learning analytics in medical education. *Int J Health Sci (Qassim)*. 2018;12(2):80–85.
13. Royal College of Physicians and Surgeons of Canada. CanMEDS 2015 Physician Competency Framework. http://www.royalcollege.ca/rcsite/documents/canmeds/canmeds-full-framework-e.pdf. Accessed October 2, 2019.
14. Therneau T, Atkinson B, Ripley B. rpart: Recursive Partitioning and Regression Trees. R package version 4.1-10. 2015. https://cran.r-project.org/package=rpart. Accessed November 4, 2019.
15. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York, NY: Springer; 2009.
16. Williams RG, Roberts NK, Schwind CJ, Dunnington GL. The nature of general surgery resident performance problems. *Surgery*. 2009;145(6):651–658. doi:10.1016/j.surg.2009.01.019.
17. Gierl MJ, Latifi S, Lai H, Boulais AP, De Champlain A. Automated essay scoring and the future of educational assessment in medical education. *Med Educ*. 2014;48(10):950–962. doi:10.1111/medu.12517.
18. Isaac C, Chertoff J, Lee B, Carnes M. Do students' and authors' genders affect evaluations?: a linguistic analysis of medical student performance evaluations. *Acad Med*. 2011;86(1):59–66. doi:10.1097/ACM.0b013e318200561d.

Gaby Tremblay, MSc, is Accreditation Officer, Postgraduate Medical Education Office, Faculty of Medicine, Université Laval, Québec, Canada; **Pierre-Hugues Carmichael, MSc,** is Biostatistician, Centre d'excellence sur le vieillissement de Québec, Centre intégré universitaire de santé et de services sociaux de la Capitale-Nationale, Québec, Canada; **Jean Maziade, MD, FCMF, CCMF, MSc,** is Professor, Department of Family Medicine, Faculty of Medicine, Université Laval, Québec, Canada; and **Mireille Grégoire, MDCM, FRCPSC,** is Associate Dean, Postgraduate Medical Education and Professor, Department of Surgery, Faculty of Medicine, Université Laval, Québec, Canada.

Funding: This work was funded by the Fondation Université Laval, fonds d'innovation en pédagogie des sciences de la Santé Gilles-Cormier.

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

This work was previously presented and won the *JGME* Top Research in Residency Education award at the International Conference on Residency Education, Québec, Canada, October 17–22, 2017.

The authors would like to thank Louise Arsenault, instructional designer; Dr Miriam Lacasse, QMA-CMA-MDM, Chair of Educational Leadership in Health Science Education, Faculty of Medicine, Université Laval; and Guy Lavoie, IT director (retired), Faculty of Medicine, Université Laval.

Corresponding author: Mireille Grégoire, MDCM, FRCPSC, Université Laval, 1050 Avenue de la Médecine, Bureau 4623, Québec G1V 0A6 Canada, 418.656.5955, mireille.gregoire@fmed.ulaval.ca

Received May 30, 2019; revision received September 16, 2019; accepted September 17, 2019.