

Determining Need for Remediation Through Postrotation Evaluations

JEANNETTE GUERRASIO, MD
ETHAN CUMBLER, MD
ADAM TROSTERMAN, MD
HEIDI WALD, MD
SUZANNE BRANDENBURG, MD
EVA AAGAARD, MD

Abstract

Introduction Postrotation evaluations are frequently used by residency program directors for early detection of residents with academic difficulties; however, the accuracy of these evaluations in assessing resident performance has been questioned.

Methods This retrospective case-control study examines the ability of postrotation evaluation characteristics to predict the need for remediation. We compared the evaluations of 17 residents who were placed on academic warning or probation, from 2000 to 2007, with those for a group of peers matched on sex, postgraduate year (PGY), and entering class.

Results The presence of an outlier evaluation, the number of words written in the comments section, and the percentage of evaluations with negative or ambiguous comments were all associated with the need for remediation ($P = .01$, $P = .001$, $P = .002$, $P < .001$,

respectively). In contrast, United States Medical Licensing Examination step 1 and step 2 scores, total number of evaluations received, and percentage of positive comments on the evaluations were not associated with the need for remediation ($P = .06$, $P = .87$, $P = .55$, respectively).

Discussion Despite ambiguous evaluation comments, the length and percentage of ambiguous or negative comments did indicate future need for remediation.

Conclusions Our study demonstrates that postrotation evaluation characteristics can be used to identify residents as risk. However, larger prospective studies, encompassing multiple institutions, are needed to validate various evaluation methods in measuring resident performance and to accurately predict the need for remediation.

Introduction

Seven percent of residents have difficulties achieving 1 or more of the Accredited Council for Graduate Medical Education (ACGME) competencies, according to a survey of internal medicine program directors,¹ and early intervention is thought to improve the success of remediation efforts.² Although many modes of resident assessment are

available, postrotation evaluations are the most commonly used form because they are quick and inexpensive to complete and review.³

Despite the widespread use of postrotation evaluations for early detection of residents in difficulty, the accuracy of those evaluations in assessing resident performance has been questioned. Multiple studies, across several disciplines and involving residents and medical students, have documented the failure of postrotation evaluations for early identification of at-risk trainees. The failures have been attributed to grade inflation, attending physicians' lack of willingness to document poor performance, and lack of knowledge about how to document performance concerns.⁴⁻⁸ Notably, when deficits are documented in evaluations, the comments section often does not correlate with the numeric ratings given for the corresponding ACGME competencies.⁹

We performed a retrospective case-control study to examine whether postrotation evaluations can identify at-risk residents and to assess whether specific elements of the postrotation evaluation can predict the need for remediation.

All authors are at University of Colorado, **Jeannette Guerrasio, MD**, is Assistant Professor of Medicine and Director of Remediation; **Ethan Cumbler, MD**, is Associate Professor of Medicine; **Adam Trosterman, MD**, is Assistant Professor of Medicine and Clerkship Director; **Heidi Wald, MD**, is Associate Professor of Medicine; **Suzanne Brandenburg, MD**, is Program Director of the Internal Medicine Residency Program; and **Eva Aagaard, MD**, is Vice-Chair of Education.

Funding: The authors report no external funding source for this study.

The authors would like to thank Traci Yamashita, MS, for data analysis and assembling tables and figure.

Corresponding author: Jeannette Guerrasio, MD, University of Colorado, Department of Medicine, 12401 E 17th Ave, Mail Code F782, Aurora, CO 80045, 303.301.4849, Jeannette.Guerrasio@ucdenver.edu

Received June 28, 2011; revision received October 13, 2011; accepted October 18, 2011.

DOI: <http://dx.doi.org/10.4300/JGME-D-11-00145.1>

Methods

Between July 1, 2000, and June 30, 2007, 17 of the 515 internal medicine residents (3%) at the University of Colorado were placed on academic warning or probation and were considered to be in poor academic standing. These residents served as the cases for this study. *Good academic standing* refers to residents who were neither placed on academic warning nor probation throughout the study period. For each case of a resident in poor standing, a list of potential control residents in good standing, matched for sex, postgraduate year (PGY) of training, and entering class, was compiled. Seventeen controls were chosen from this list using a random number generator.

We compared the postrotation evaluations, on the residents in poor academic standing for all clinical rotations before the date on the letter of academic warning or probation, with postrotation evaluations for the matched residents in good academic standing for the corresponding period. Research block evaluations were excluded. All evaluations were completed by attending physicians and contained a 9-point numeric rating scale for an overall score of competence and a score for each of the 6 ACGME competencies. The evaluations used the following descriptors: 1 to 3 unsatisfactory, 4 to 6 satisfactory, and 7 to 9 superior, and the evaluations contained a free-text section for comments.

Additional data collected on all subjects included United States Medical Licensing Examination (USMLE) 1 and 2 scores, total number of evaluations completed on each resident, scores on individual ACGME competencies and on overall performance for each evaluation, and the number of evaluations without comments. The data were collected from each individual's residency program files, which included USMLE scores reported directly from the testing agency. An outlying evaluation was defined as having a single value that was more than 2 points lower than any numeric score on that resident's other evaluations. For the comments section, the principal investigator recorded the total number of words per evaluation and de-identified the text. Each de-identified written comment section was scrutinized by 2 general internal medicine faculty members, blinded to the academic standing of the subject. The number of positive, negative, or ambiguous comments was recorded. A comment was identified as negative if it alerted the reader to an underlying deficiency and positive if it described a resident at or above the level of his or her peers. Comments that could not be clearly identified as positive or negative were classified as ambiguous. Discrepancies between the 2 evaluators' assessments were adjudicated by a third blinded faculty attending, who independently reviewed the de-identified

What was known

Rotation evaluations are used to provide feedback and assess ongoing development, but have not been tested for efficacy in early detection of residents with academic difficulties.

What is new

Attributes of rotation evaluations (presence of "outlier" evaluations, the amount of comments, and the percentage of negative or ambiguous comments) were associated with subsequent identification of a need for remediation.

Limitations

Single institution, single-specialty (internal medicine) sample, retrospective study, with the potential for a history effect due to the implementation of the competencies during the 7-year study period.

Bottom line

"Outlier" evaluations, the amount of text, and negative and ambiguous comments in rotation evaluations identified residents in need for added education or remediation.

comments section and served as a tie-breaker. The negative comments were assigned to the most relevant ACGME competency.

The total number of evaluations for residents in poor and good standing was compared to assess the adequacy of the control group. Descriptive statistics were performed on resident demographics and postrotation evaluation measures. Median scores and interquartile ranges were determined, and significance was calculated for continuous measures using Wilcoxon rank sum test, using the *t* approximation for small samples. Significance was calculated for dichotomous measures using the χ^2 or Fisher exact test when necessary. Receiver operating curves were created for numeric Likert scores for overall scores and each of the ACGME competencies. All analyses were performed using the SAS v. 9.2 (Chicago, IL).

This study was approved by the Colorado Multiple Institutional Review Board.

Results

From 2000 to 2007, 9 residents (53%) were placed on academic warning and 8 (47%) on probation. Of these residents in poor academic standing, 9 were PGY-1 (53%), 3 were PGY-2 (18%), and 5 (29%) were PGY-3; 9 (53%) were women and 8 (47%) were men ($P = .95$). Sixteen of the 17 residents (94%) were noted to have deficiencies in more than 1 ACGME competency on their postrotation evaluations. Of the 16 with multiple deficits, 15 residents (94%) were noted to have medical knowledge deficits, 15 (94%) had patient care deficits, 13 (81%) had interpersonal skills deficits, 16 (100%) had professionalism deficits, 1 (6%) had deficits relating to practice-based learning and improvement, and 2 (13%) had deficits in the

TABLE 1 BOARD SCORES AND GENERAL EVALUATION CHARACTERISTICS OF RESIDENTS IN GOOD VERSUS POOR ACADEMIC STANDING

Characteristic	Good Standing (N = 17)	Poor Standing (N = 17) ^a	P Value
USMLE 1, median (IQR), n	235 (224, 243)	227 (197, 233), 13	.06
USMLE 2, median (IQR), n	234.5 (223, 248), 10	227 (220, 247), 9	.87
Total evaluations completed, median (IQR)	10 (7, 19)	13 (7, 19)	.55
Any unsatisfactory score on an evaluation (score ≤ 3)	0 (0)	10 (58.8)	< .001
Any outlier score on an evaluation ^b	0 (0)	7 (41.2)	.01

Abbreviations: IQR, interquartile range; USMLE, United States Medical Licensing Examination.

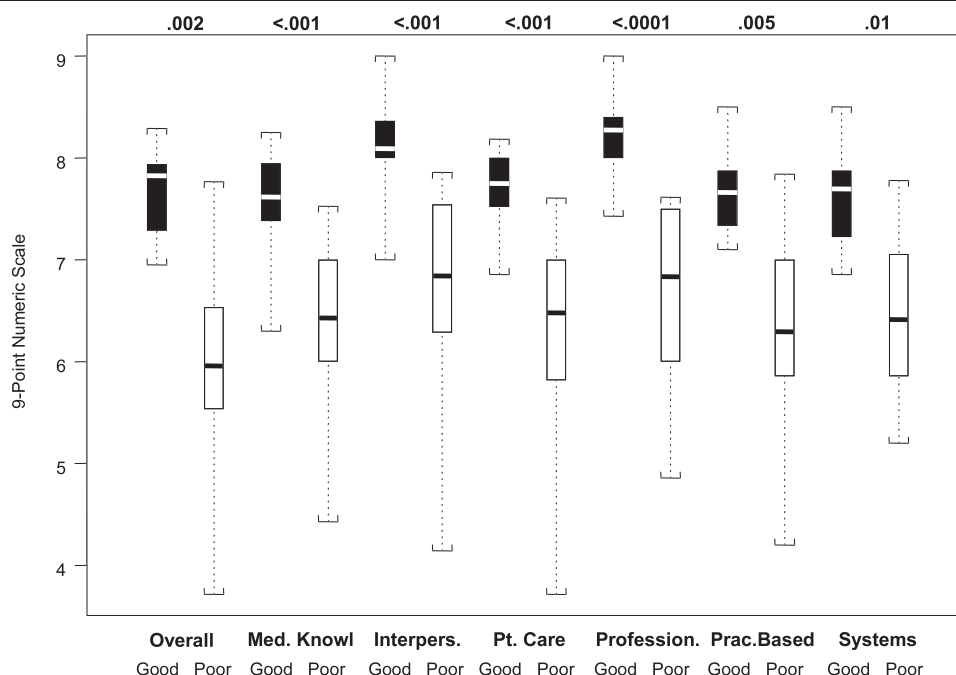
^a Warning (n = 9); probation (n = 8).

^b An outlier score is defined as a single value that was > 2 points lower than any numeric score on that resident's other evaluations.

area of systems-based practice. Bivariate comparisons by gender, PGY level, and USMLE scores revealed no statistically significant differences.

TABLE 1 compares the board scores and postrotation evaluation characteristics of the residents in poor and good standing. There were no significant differences in the USMLE 1 or 2 scores ($P = .06$ and $P = .87$, respectively), nor were there significant differences in the number of

evaluations completed for residents in good versus poor standing ($P = .55$). A median of 13 evaluations were completed before the resident was placed on warning or probation, with a median of 10 evaluations completed for residents in good standing during the respective time frame (TABLE 1). Residents in poor standing were significantly more likely to have received an evaluation with at least one unsatisfactory score (≤ 3) than their peers in good standing

**FIGURE**

OVERALL PERFORMANCE SCORE AND SCORES FOR EACH OF 6 ACCREDITATION COUNCIL FOR GRADUATE MEDICAL EDUCATION COMPETENCIES, WITH BLACK AND WHITE BOXES FOR RESIDENTS IN GOOD AND POOR STANDING, RESPECTIVELY

TABLE 2 CHARACTERISTICS OF THE COMMENTS ON RESIDENT EVALUATIONS BASED ON ACADEMIC STANDING

Characteristic	Good Standing (N = 17)	Poor Standing (N = 17) ^a	P Value
Percentage of evaluations without comments, median (IQR)	12.0 (5.3, 16.7)	8.3 (0, 14.3)	.11
Average No. of words written over all evaluations, median (range)	30.3 (13.8–66.6)	60.9 (19.0–329.5)	.002
Any negative comments, No. (% of residents)	12 (70.6)	17 (100.0)	.04
Percentage of evaluations with negative comments, median (IQR)	11.1 (0, 20.0)	38.5 (25.0, 53.3)	.001
Percentage of evaluations with positive comments, median (IQR)	87.5 (83.3, 92.0)	81.5 (69.2, 90.9)	.17
Percentage of evaluations with ambiguous comments, median (IQR)	18.5 (14.3, 33.3)	54.5 (45.8, 70.0)	< .001

Abbreviation: IQR, interquartile range.

^a Warning (n = 9); probation (n = 8).

(59%, [10/17] versus 0%, $P < .001$). Additionally, residents in poor standing were more likely to have received an outlier evaluation compared with residents in good standing (41%, [7/17] versus 0%, $P = 0.01$).

The FIGURE shows the overall performance score and the scores for each of 6 ACGME competencies, with the black boxes representing residents in good standing, and the white boxes representing the residents in poor standing.

The FIGURE shows that the median scores for the poor standing group were lower than those for the good standing group. In addition, the range of scores received by the residents in poor standing was much greater than that of their matched peers in good academic standing. The median overall score for a resident in good standing was 7.8 with a 5% to 95% range of 7 to 8.3, whereas the median overall score for the resident in poor standing was 5.9 with a 5% to 95% range of 3.7 to 7.8 ($P = .002$). In addition, although the median score in each competency was lower for those later placed on academic warning or probation, residents in poor standing still received many superior ratings of 7, 8, or 9. Notably, our review revealed that 61%, (294/480) of these superior ratings were assigned in the same competencies in which the resident's standing was ultimately cited as sufficiently unsatisfactory to warrant letter of warning or academic probation.

Important for the assessment of the ability of post-rotation evaluations to predict academic difficulty, receiver operating curves demonstrated that neither the overall score nor individual competency scores were able to predict the need for remediation. For interpersonal skills and professionalism having a lowest score greater than 8 was predictive of good standing, but the receiver operating curve did not demonstrate which lowest numeric rating scores were predictive of poor standing.

TABLE 2 contains the results of the analysis of the comments section of the postrotation evaluations. Residents in poor standing had almost double the average

number of words written over all postrotation evaluations compared with residents in good standing ($P = .002$). Notably, the percentage of evaluations with negative comments and the percentage of evaluations with ambiguous comments were both greater for the residents in poor standing ($P = .001$ and $P < .001$, respectively). Although all 17 (100%) of the residents in poor standing received at least one evaluation with negative comments, 12 (71%) of residents in good standing also received at least one evaluation with a negative comment.

As expected, PGY-1 residents received significantly fewer evaluations before being placed on academic warning or probation ($P = .01$) compared with PGY-2 and PGY-3 residents. These PGY-1 residents received a greater average number of words in the comments section (81 versus 35; $P = .04$) and a greater percentage of total evaluations with ambiguous comments (63.6 versus 48.1; $P = .03$) compared with their more senior colleagues in poor standing.

Discussion

This study of resident postrotation evaluations at the University of Colorado identified several evaluation characteristics that predated a resident's change in academic status from good standing to either warning or probation. In particular, we found that residents who were ultimately placed on poor academic status were more likely to have low ratings on the ACGME competencies, increased numeric score variability for each category assessed, ambiguous written comments, negative written comments, and longer written comments when compared with colleagues in good academic standing. Based on the data, however, no single criterion emerged that could be used in isolation to identify those at high risk. Very low numeric rating scores of 3 or less were the one characteristic present for residents in poor standing that were not present for residents in good standing.

Schwind et al⁹ reviewed the evaluations of 30 surgical residents with deficiencies requiring remediation and noted that only 2% to 4% were identified via the numeric rating scale and 8% to 20% in the written comments. Our data confirm these prior concerns that postrotation evaluation scores and comments alone were insufficient to identify at-risk residents. Within internal medicine, 60% of program directors had difficulty convincing problem residents of their deficiencies because of a lack of honest and accurate written evaluations from attending physicians.¹ A lack of consistency in ratings and comments between evaluators, as seen in our study, contributes to these challenges. In addition, the frequency of ambiguous comments in at-risk resident evaluations further speaks to broad-based lack of direct feedback and constructive criticism. Moreover, the subtle differences in numeric evaluations and written comments may not provide program directors with the documentation needed to act on identified deficiencies.

The USMLE step scores were not significant in determining the need for remediation, which further supports the study research synthesis that USMLE step 1 and 2 scores are not correlated with reliable measures of medical students', residents', and fellows' clinical skill acquisition.¹⁰

Although the evaluations detected deficiencies in all 6 of the ACGME competencies, a "gold standard" for determining which residents need remediation does not exist. As described by Holmboe and Hawkins,¹¹ many tools are needed to adequately and thoroughly assess residents' knowledge, patient-care skills, and attitudes, including the in-training exam, monthly attending evaluations, directly observed mini and/or full clinical evaluation exercises, self-evaluations, student-teaching evaluations, and simulations as well as 360° evaluations, medical-record audits, quality-improvement projects, and clinical question logs.

Conclusion

Our study provides evidence for the continuing use of end-of-rotation evaluations by faculty to identify residents in need of remediation. More research is needed on the information extracted from these assessment tools to determine whether they can identify the early warnings of at-risk residents. Future research will hopefully lead to a validated, predictive risk index for determining learners at high risk for needing remediation.

Until then, further emphasis can be placed on following residents with high variability of numeric rating scores between evaluations, those who receive more negative or ambiguous comments than their peers, and those who have lengthy written comments.

References

- 1 Yao DC, Wright SM. A national survey of internal medicine residency program directors regarding problem residents. *JAMA*. 2000;284(9):1099–1104.
- 2 Wiese J. COUGAR: *Proceedings of Resident Remediation Program at Tulane*. Aurora, CO: Fall 2008. <http://www.med.umkc.edu/cme/online/wiese050710b.shtml>
- 3 Holmboe ES. A systems approach to evaluation. In: *Proceedings from American Board of Internal Medicine Workshop*. Aurora, CO: November 19, 2008.
- 4 Dudek NL, Marks MB, Regehr G. Failure to fail: the perspectives of clinical supervisors. *Acad Med*. 2005;80(10, suppl):S84–S87.
- 5 Shapiro J, Prislis MD, Larsen KM, Lenahan PM. Working with the resident in difficulty. *Fam Med*. 1987;19(5):368–375.
- 6 Speer AJ, Solomon DJ, Ainsworth MA. An innovative evaluation method in an internal medicine clerkship. *Acad Med*. 1996;71(1, suppl):S76–S78.
- 7 Magarian GJ, Mazur DJ. Evaluation of students in medicine clerkships. *Acad Med*. 1990;65(5):341–345.
- 8 Boisselle PM. A remedy for resident evaluation and remediation. *Acad Radiol*. 2005;12(7):894–900.
- 9 Schwind CJ, Williams RG, Boehler ML, Dunnington GL. Do individual attending' post rotation performance ratings detect residents' clinical performance deficiencies? *Acad Med*. 2004;79(5):453–457.
- 10 McGaghie WC, Cohen ER, Wayne DB. Are United States Medical Licensing Exam step 1 and 2 scores valid measures for postgraduate medical residency selection decisions? *Acad Med*. 2011;86(1):48–52.
- 11 Holmboe ES, Hamkins, RE. Methods for evaluating the clinical competence of residents in internal medicine: a review. *Ann Intern Med*. 1998;129(1):42–48.