Comparing Students' Clinical Grades to Scores on a Standardized Patient Note-Writing Task

Benjamin D. Gallagher, MD[1], *Saman Nematollahi, MD*[2], *Henry Park, PhD*[3], *and Salila Kurra, MD*[4]

[1]Department of Internal Medicine, Yale School of Medicine New Haven, CT, USA; [2]Department of Medicine, Johns Hopkins University School of MedicineBaltimore, MD, USA; [3]Center for Education Research and Evaluation, Columbia University Vagelos College of Physicians & SurgeonsNew York, NY, USA; [4]Department of Medicine, Columbia University Vagelos College of Physicians & SurgeonsNew York, NY, USA.

BACKGROUND: Few assessments capture the diagnostic impressions medical students form immediately following patient encounters. However, notes written for objective structured clinical examinations (OSCEs) allow learners to document their clinical reasoning in real time. The University of Illinois at Chicago College of Medicine (UIC-COM) has developed a rubric for scoring patient notes (PNs) in their OSCE for senior students.

OBJECTIVE: To validate the UIC-COM PN Scoring Rubric as a measure of clinical reasoning by comparing PN scores from a similar exam at the Columbia University Vagelos College of Physicians and Surgeons (VP&S) to clinical rotation performance.

DESIGN: Cross-sectional analysis.

PARTICIPANTS: From a total of 146 third-year medical students who completed the OSCE at VP&S in spring 2017, we selected 60 at random, 20 from each tertile of clinical rotation performance.

MAIN MEASURES: We scored these students' PNs using the rubric's four sections—Documentation, Differential Diagnosis, Justification, and Workup, each scored from 1 to 4—and calculated a composite score (maximum 100). We used one-way ANOVA to examine differences in scores between clinical rotation performance tertiles.

KEY RESULTS: Students in the bottom, middle, and top clinical rotation performance tertiles had mean Documentation scores of 2.54, 2.63, and 2.88, respectively ($p = 0.02$, bottom vs. top tertile). Mean composite scores were 61.98, 64.05, and 67.86, respectively ($p = 0.02$, bottom vs. top tertile).

CONCLUSIONS: We showed an association between PN scores and clinical rotation performance. Since clinical rotation grades incorporate multiple types of assessments of students' clinical reasoning skills, we believe that this correlation lends validity evidence to using the note-writing task as a measure of clinical reasoning. Future directions include expanding the task to different stages of learners, to real life patient encounters, and to formative rather than summative assessments of note-writing skills.

KEY WORDS: undergraduate medical education; clinical reasoning; clinical assessment; OSCE; simulation.

J Gen Intern Med 35(11):3243–7
DOI: 10.1007/s11606-020-06019-2
© Society of General Internal Medicine 2020

INTRODUCTION

Learning clinical reasoning is a key objective in the training of early physicians.[1] At the undergraduate level, teachers of medical students employ a variety of methods to assess clinical reasoning skills. Many of these entail reviewing students' patient notes (PNs), either presented orally or submitted in written form to a preceptor.[2] PNs of this type often involve significant input from supervisors, medical textbooks, online resources, and the primary literature. As such, they are several steps removed from the clinical impression a student forms immediately following the patient encounter.

PNs written for an objective structured clinical examination (OSCE), on the other hand, provide the opportunity to evaluate clinical reasoning at the point of care. In the USA, many medical schools hold OSCEs for their senior students to assess readiness for residency and prepare for the United States Medical Licensing Examination (USMLE) Step 2 Clinical Skills test.[3] These OSCEs consist of timed standardized patient (SP) encounters, each ending with a templated PN. In the PN, students document pertinent elements of the history and physical, propose a ranked differential diagnosis with supporting evidence, and suggest an initial workup. Researchers at the University of Illinois at Chicago College of Medicine (UIC-COM) have developed a validated rubric for scoring PNs in their OSCE for senior students.[4–7] Most recently, this group has published data from five cases shared across seven medical schools, with a total of 990 students participating.[7]

Because clinical rotation grades comprise different types of assessments of clinical reasoning, with the input of multiple evaluators, comparing PN scores to clinical rotation performance would lend further validity evidence to the use of OSCE note-writing as a measure of clinical

These results were presented as an oral presentation at the Society of General Internal Medicine (SGIM) Annual Meeting, Denver, CO, April 14, 2018.

Received January 8, 2020
Accepted June 30, 2020
Published online July 13, 2020

reasoning skills. Therefore, we applied the UIC-COM rubric to PNs written by senior medical students in an OSCE at the Columbia University Vagelos College of Physicians and Surgeons (VP&S) and examined the relationship between PN scores and core clinical rotation grades.

METHODS

VP&S Senior OSCE

Each year, students in their sixth semester at VP&S (i.e., the second half of the third year) take part in an OSCE. The OSCE consists of 10 SP encounters developed by VP&S faculty, each limited to 15 min. The case scenarios are designed to highlight common chief complaints in the core specialties of medical practice. Participants then have 10 min to write a templated PN on a computer (Fig. 1). SPs use checklists unique to individual cases to rate students on the completeness of their histories and physicals as well as on their communication skills. These SP scores are reported back to the participants, who also review selected video recordings of their SP encounters in small groups with core teaching faculty from the physical diagnosis course. The PNs are not routinely scored except for those students receiving remediation or those who request feedback on their note-writing skills. Performance on the OSCE has no bearing on medical school grades or residency placement.

Participants

This study was approved by the Institutional Review Board of Columbia University Irving Medical Center. In spring 2017, 146 third-year students completed the OSCE. In the third-year class that year, 48% were women, 53% self-described as White, 17% Asian, 13% Hispanic, and 8% Black. An email explaining the purpose of the study was sent to the students who completed the OSCE. When they matriculated to VP&S, they were made aware that their coursework might be used for educational research and were allowed to opt out at that time. Consent for participation in the study was therefore implied. In the email, they were given the opportunity to opt out of this specific study, and two students chose to do so. Those students' data were excluded from the analysis.

We stratified students' academic performance by the grades earned in eight core clinical rotations. Due to an accelerated pre-clinical curriculum at VP&S, these rotations take place in the fourth and fifth semesters of medical school (i.e., the second half of the second year and first half of the third year). Final grades are based on subjective evaluations by faculty and residents, the score on the National Board of Medical Examiners (NBME) subject exam, and (for some rotations) performance on rotation-specific OSCEs. These components are weighted differently in each rotation. In addition, each rotation has a different evaluation form, but most have an item for clinical reasoning skills. In the third-year class in 2017, the median (IQR) number of evaluations per student per rotation was 2.04 (1.06–4.20). However, this figure underestimates the number of evaluators per student per rotation because some evaluations comprise the input of multiple evaluators.

The available grades are "honors," "high pass," "pass," "low pass," and "fail," but the majority of grades awarded are "honors" or "high pass." We therefore stratified the class into bottom, middle, and top tertiles based on the number of "honors" grades per student, and selected 20 students at random from each tertile, yielding a study sample of 60 students. Students in the bottom, middle, and top tertiles had median 1, 4, and 7 "honors" grades, respectively.

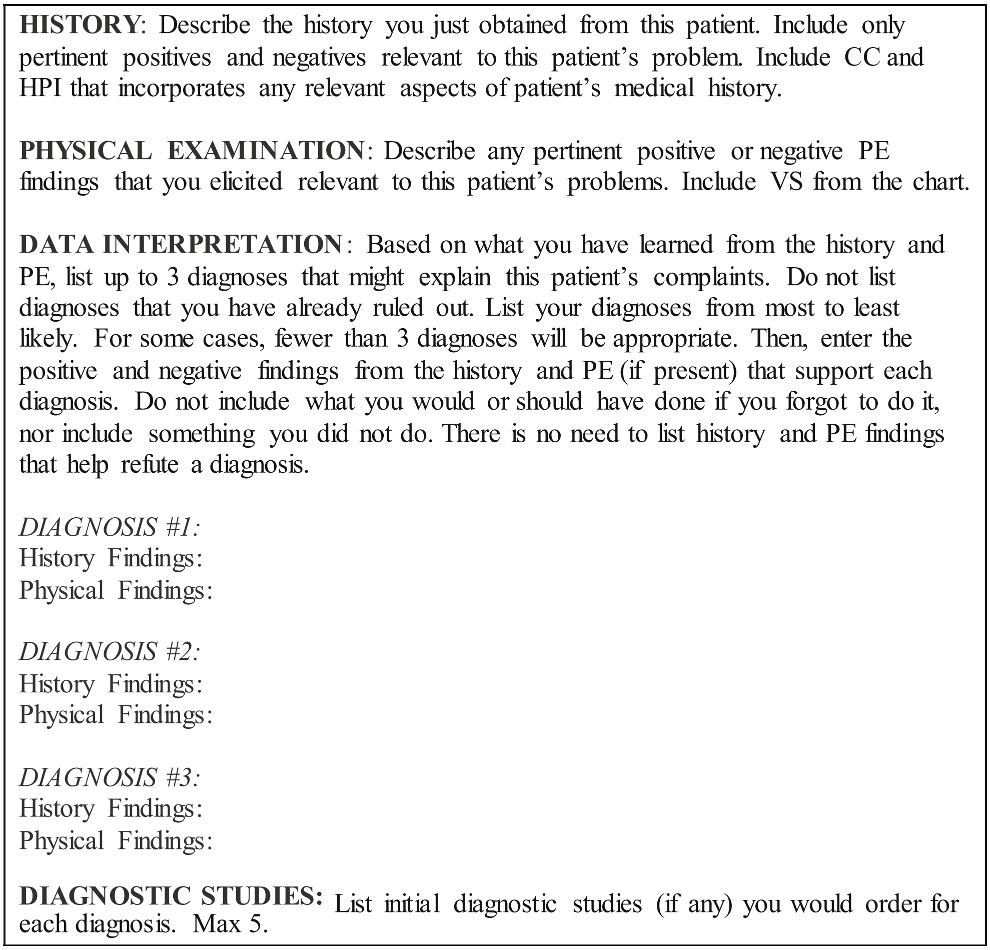
Scoring Rubric

We used the UIC-COM PN Scoring Rubric to score the PNs (see Table 1 in Park *et al.*).[7] The rubric contains four sections, each scored on a scale of 1 to 4: Documentation, Differential Diagnosis, Justification, and Workup. To generate a composite score with a maximum of 100 points, the Documentation, Differential, and Justification sections are each worth 30 points, and the Workup section is worth 10 points. Each section score level is worth 25% of the maximum number of points for that section (e.g., 7 points for a Documentation score of 1, 15 points for a score of 2, 23 points for a score of 3, and 30 points for a score of 4).

Two third-year internal medicine residents reviewed the case materials for the SP scenarios and together wrote one exemplar note for each case. Using the exemplar notes, both residents scored 10 PNs selected at random and compared scores to ensure agreement about use of the rubric. Then one of the residents scored all participants' PNs (600 total), and the other resident scored 10 participants' PNs (100 total) to determine inter-rater reliability. Percent exact agreement for all unweighted PN section scores was 70%. Kappa (SE) was 0.55 (0.03) and quadratically weighted kappa (SE) was 0.75 (0.05). Results for the four PN sections are shown in Table 1.

Statistical Analyses

We used descriptive statistics to report PN section and composite scores. We used one-way ANOVA to examine differences in scores between clinical rotation performance tertiles. When the effect of clinical rotation tertile was found to be significant, we performed pairwise comparisons with Sidak's test for multiple comparisons. Using Cohen's method, we estimated that including 3 performance groups with 20 students in each group would provide 80% power to detect a large difference in PN scores with an alpha of 0.05. This method operationally defines an effect size index for one-way ANOVA as the ratio of the SD of the group means to the SD for the overall population; a large effect size index is 0.40.[8] Analyses were performed using SPSS (version 21.0, IBM Corp., Armonk, NY).



HISTORY: Describe the history you just obtained from this patient. Include only pertinent positives and negatives relevant to this patient's problem. Include CC and HPI that incorporates any relevant aspects of patient's medical history.

PHYSICAL EXAMINATION: Describe any pertinent positive or negative PE findings that you elicited relevant to this patient's problems. Include VS from the chart.

DATA INTERPRETATION: Based on what you have learned from the history and PE, list up to 3 diagnoses that might explain this patient's complaints. Do not list diagnoses that you have already ruled out. List your diagnoses from most to least likely. For some cases, fewer than 3 diagnoses will be appropriate. Then, enter the positive and negative findings from the history and PE (if present) that support each diagnosis. Do not include what you would or should have done if you forgot to do it, nor include something you did not do. There is no need to list history and PE findings that help refute a diagnosis.

DIAGNOSIS #1:
History Findings:
Physical Findings:

DIAGNOSIS #2:
History Findings:
Physical Findings:

DIAGNOSIS #3:
History Findings:
Physical Findings:

DIAGNOSTIC STUDIES: List initial diagnostic studies (if any) you would order for each diagnosis. Max 5.

Figure 1 Patient note template used in an OSCE for senior medical students at the Columbia University Vagelos College of Physicians & Surgeons (New York, NY), 2017. OSCE objective structured clinical examination

RESULTS

Documentation, Differential Diagnosis, Justification, Workup, and composite scores are shown in Table 2. Section and composite scores varied across the case scenarios. For example, the mean Differential Diagnosis score for each case ranged from 1.55 to 2.97.

Students in the bottom, middle, and top clinical rotation performance tertiles had mean (SD) Documentation scores of 2.54 (0.35), 2.63 (0.44), and 2.88 (0.28), respectively (F statistic = 4.54, $p = 0.02$; $p = 0.02$ for bottom vs. top tertile) (Table 3). While there was a trend toward increased scores with better clinical rotation performance in the other three sections, these differences were not statistically significant. Mean (SD) composite scores were 61.98 (6.34), 64.05 (7.32), and 67.86 (6.14), respectively (F statistic = 4.05, $p = 0.02$; $p = 0.02$ for bottom vs. top tertile).

DISCUSSION

In this study, we aimed to further validate a rubric developed by Park *et al.* for scoring PNs written for OSCEs similar to the USMLE Step 2 Clinical Skills exam.[4–7] To this end, we compared PN scores with clinical rotation grades. We found small but statistically significant differences in Documentation and composite scores between students in the bottom and top tertiles of clinical rotation performance. On the other sections, there was a trend toward better PN scores in higher clinical rotation performance tertiles, but this did not meet statistical

Table 1 Inter-rater reliability for patient notes from an OSCE for senior medical students at the Columbia University Vagelos College of Physicians & Surgeons (New York, NY), 2017, scored by two raters ($n = 10$)

	Documentation	DDx	Justification	Workup	All Sections
Exact agreement (%)	59	93	58	69	70
Kappa (SE)	0.31 (0.08)	0.89 (0.04)	0.31 (0.08)	0.50 (0.07)	0.55 (0.03)
Weighted Kappa (SE)	0.56 (0.07)	0.94 (0.03)	0.51 (0.08)	0.69 (0.05)	0.75 (0.05)

OSCE objective structured clinical examination, DDx differential diagnosis, SE standard error

Table 2 Patient note scores from an OSCE for senior medical students at the Columbia University Vagelos College of Physicians & Surgeons (New York, NY), 2017 (n = 60)

	Documentation	DDx	Justification	Workup	Composite*
Mean	2.68	2.15	2.90	2.65	64.63
SD	0.38	0.28	0.40	0.41	6.90
Range	1.79–3.43	1.57–2.61	1.68–3.72	1.76–3.44	48.20–78.00
Lowest case mean†	2.30	1.55	2.40	2.13	58.70
Highest case mean	3.00	2.97	3.17	3.00	72.50

DDx differential diagnosis, SD standard deviation

**To generate a composite score with a maximum of 100 points, the Documentation, Differential, and Justification sections are each worth 30 points, and the Workup section is worth 10 points. Each section score level is worth 25% of the maximum number of points for that section (e.g., 7 points for a Documentation score of 1, 15 points for a score of 2, 23 points for a score of 3, and 30 points for a score of 4)*

†Case mean refers to the mean score for a given section in a single case scenario

significance. We hypothesize that the differences between tertiles was most substantial in the Documentation score because the response in this section is an open-ended narrative that requires the student to translate the history and physical she has obtained into an argument supporting her differential diagnosis. This is closer than the other sections to the note-writing tasks students have previously experienced in their core clinical rotations.

Clinical rotation grades incorporate several kinds of assessments that touch on clinical reasoning: subjective evaluations of case write-ups and oral presentations and contributions to rounds, the NBME subject exam, and the rotation-specific OSCE. The grades also represent the impressions of multiple evaluators who provide feedback to students on many types of clinical activities. Therefore, we believe that the correlation between clinical rotation performance and PN scores lends validity evidence to the use of the note-writing task as a measure of clinical reasoning skills. While clinical grades are influenced by factors unrelated to clinical reasoning (e.g., communication skills and professionalism), one would expect shared variance between clinical reasoning measured by PN scores and by the different grade components. If anything this would dilute the association between PN scores and clinical grades.

However, it is important to discuss why there was not a larger difference in PN scores between students at the top and bottom of the class. At VP&S, the majority of grades awarded for clinical rotations are “honors” or “high pass.” In our sample, the median number of “honors” grades (out of a maximum of eight) was 1 in the bottom tertile and 7 in the top tertile. So the small differences in PN scores cannot be attributed to negligible distinctions in clinical rotation performance. Nor were all VP&S students scoring at an unusually high level on the PNs, such that any variance due to clinical rotation performance would be inconsequential; to the contrary, VP&S students scored similarly to a sample of 990 students from seven other medical schools.[7]

Instead, we propose several other explanations for our findings. First, because the OSCE is a formative assessment with no impact on medical school grades or residency placement, students’ effort on the note-writing task may have been submaximal, weakening any association between the low-stakes PN scores and high-stakes clinical rotation grades. Second, students’ lack of familiarity with the OSCE’s format and the time pressures of the note-writing task may have depressed performance overall, leveling out any differences due to diagnostic acumen. Third, the PN data were anonymized and therefore scoring was less subject to bias than clinical rotation grades, which comprise mostly subjective evaluations. Lastly, the PNs in the OSCE are a “point of care” assessment of clinical reasoning skills, unlike the oral case presentations and written histories and physicals that contribute to clinical rotation grades, and that often involve significant input from supervisors, medical textbooks, online resources, and the primary literature. The discrepancy in achievement on these tasks suggests that students at this stage of their training may rely heavily on outside resources to formulate clinical impressions, and that utilization of these resources is what distinguishes bottom- from top-performing students.

Table 3 Patient note scores from an OSCE for senior medical students at the Columbia University Vagelos College of Physicians & Surgeons (New York, NY), 2017, compared with clinical rotation performance tertile (n = 60)

	Documentation, mean (SD)	DDx, mean (SD)	Justification, mean (SD)	Workup, mean (SD)	Composite, mean (SD)
All students (n = 60)	2.68 (0.38)	2.15 (0.28)	2.90 (0.40)	2.65 (0.41)	64.63 (6.90)
Bottom tertile (n = 20)	2.54 (0.35)	2.10 (0.25)	2.77 (0.41)	2.55 (0.40)	61.98 (6.34)
Middle tertile (n = 20)	2.64 (0.44)	2.13 (0.30)	2.88 (0.39)	2.71 (0.44)	64.05 (7.32)
Top tertile (n = 20)	2.88 (0.28)*	2.22 (0.30)	3.05 (0.38)	2.71 (0.39)	67.86 (6.14)†

DDx differential diagnosis, SD standard deviation

**p = 0.02 for bottom vs. top tertile*

†p = 0.02 for bottom vs. top tertile

In the future, we plan to use the PN scoring rubric for students who ask for feedback on the written portion of the OSCE. (Currently, we lack the resources to provide such feedback to all students.) It would be informative to use the rubric for note-writing tasks experienced by learners of other levels (e.g., pre-clinical medical students or medical residents), or to track performance on the task in the same learners over time. There is also the possibility of broadening the use of the rubric to real-life patient encounters, though the benefits of standardization would be lost. Moreover, medical students' PN scores have yet to be compared with future clinical performance in residency. Such an analysis could lend predictive validity to the PN scoring rubric. Perhaps, most excitingly, the note-writing task may be a useful tool in formative assessments of students and trainees as they learn the basics of history taking, the physical exam, and clinical reasoning.

Our study has several strengths. We applied a validated PN scoring rubric to a new medical student population and a new set of ten SP scenarios. We achieved good inter-rater reliability without requiring intensive rater training. In addition, we compared PN scores with performance in clinical rotations, which has not been reported previously. There were also some limitations. There was heterogeneity across clinical rotations in the components of the final grade and the weighting of these components, and we were unable to compare PN scores with individual grade components. Our sample size was small and thus we lacked statistical power to show small differences in PN scores by clinical rotation performance tertile. Only one clinician scored all the PNs, and we had only two raters for our analysis of inter-rater reliability. Finally, the period of clinical performance to which we compared the PN scores came before, not after, the OSCE, so the use of PN scores as a means of predicting future clinical reasoning skills is necessarily circumscribed.

CONCLUSIONS

We showed an association between the quality of notes written for an OSCE and clinical rotation performance in senior medical students. As a result, educators should consider using this brief note-writing task as a "point of care" test of key clinical reasoning skills learned in medical school.

Contributors: *None.*

Corresponding Author: *Benjamin D. Gallagher, MD; Department of Internal Medicine, Yale School of Medicine New Haven, CT, USA (e-mail: benjamin.gallagher@yale.edu).*

Compliance with Ethical Standards:

Conflict of Interest: *The authors declare that they do not have a conflict of interest.*

REFERENCES

1. **Bowen JL.** Educational Strategies to Promote Clinical Diagnostic Reasoning. *N Engl J Med* 2006;355(21):2217-2225.
2. **Epstein RM**. Assessment in Medical Education. *N Engl J Med* 2007;356(4):387-396.
3. **Haist SA, Katsufrakis PJ, Dillon GF.** The Evolution of the United States Medical Licensing Examination (USMLE): Enhancing Assessment of Practice-Related Competencies. *JAMA.* 2013;310(21):2245-2246.
4. **Park YS, Lineberry M, Hyderi A, Bordage G, Riddle J, Yudkowsky R.** Validity evidence for a patient note scoring rubric based on the new patient note format of the United States Medical Licensing Examination. *Acad Med* 2013;88(10):1552-1557.
5. **Yudkowsky R, Park YS, Hyderi A, Bordage G.** Characteristics and Implications of Diagnostic Justification Scores Based on the New Patient Note Format of the USMLE Step 2 CS Exam. *Acad Med* 2015;90(11 Suppl):S56-62.
6. **Park YS, Hyderi A, Bordage G, Xing K, Yudkowsky R.** Inter-rater reliability and generalizability of patient note scores using a scoring rubric based on the USMLE Step-2 CS format. *Adv Health Sci Educ* 2016:1-13.
7. **Park YS, Hyderi A, Heine N, et al.** Validity Evidence and Scoring Guidelines for Standardized Patient Encounters and Patient Notes From a Multisite Study of Clinical Performance Examinations in Seven Medical Schools. *Acad Med* 2017;92:S12-S20.
8. **Cohen J.** A power primer. *Psychol Bull* 1992;112(1):155-159.

Publisher's Note: *Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.*