Assessment of medical student clinical reasoning by "lay" vs physician raters: Interrater reliability using a scoring guide in a multidisciplinary objective structured clinical ex...

ARTICLE in AMERICAN JOURNAL OF SURGERY · JANUARY 2012
Impact Factor: 2.29 · DOI: 10.1016/j.amjsurg.2011.08.003 · Source: PubMed

Available from: Ming C Tsai
Retrieved on: 12 February 2016
The Association for Surgical Education

Assessment of medical student clinical reasoning by “lay” vs physician raters: inter-rater reliability using a scoring guide in a multidisciplinary objective structured clinical examination

Alexandra J. Berger, B.A., Colleen C. Gillespie, Ph.D., Linda R. Tewksbury, M.D., Ivey M. Overstreet, B.A., Ming C. Tsai, M.D., Adina L. Kalet, M.D., Jennifer B. Ogilvie, M.D.,*

Departments of a Surgery, b Medicine, c Pediatrics, and d Obstetrics and Gynecology, New York University School of Medicine, New York, NY, USA

Abstract

BACKGROUND: To determine whether a “lay” rater could assess clinical reasoning, interrater reliability was measured between physician and lay raters of patient notes written by medical students as part of an 8-station objective structured clinical examination.

METHODS: Seventy-five notes were rated on core elements of clinical reasoning by physician and lay raters independently, using a scoring guide developed by physician consensus. Twenty-five notes were rerated by a 2nd physician rater as an expert control. Kappa statistics and simple percentage agreement were calculated in 3 areas: evidence for and against each diagnosis and diagnostic workup.

RESULTS: Agreement between physician and lay raters for the top diagnosis was as follows: supporting evidence, 89% (κ = .72); evidence against, 89% (κ = .81); and diagnostic workup, 79% (κ = .58). Physician rater agreement was 83% (κ = .59), 92% (κ = .87), and 96% (κ = .87), respectively.

CONCLUSIONS: Using a comprehensive scoring guide, interrater reliability for physician and lay raters was comparable with reliability between 2 expert physician raters.

© 2012 Elsevier Inc. All rights reserved.

KEYWORDS: Clinical reasoning; Assessment; Objective structured clinical examination (OSCE); Interrater reliability; Kappa statistic

The objective structured clinical examination (OSCE) is a useful tool to assess clinical performance and prepare medical students for the United States Medical Licensing Examination (USMLE) Step 2 Clinical Skills section. Since their development by Harden in 1975, the use of OSCEs to measure and standardize professional competencies among medical students has become increasingly widespread and evidence of their validity is mounting. For example, OSCEs administered in the first month of residency have been shown to be accurate predictors of residents’ future performance. Note writing is an integral part of many OSCEs and is intended to measure clinical reasoning and the ability to document clinical reasoning. Accurate patient documentation skill is viewed by medical school deans of education...
throughout the United States as an essential aspect of medical student education and as necessary preparation for residency training.6

During an OSCE, the examinee gathers history and physical examination information from a standardized patient and documents the relevant findings, differential diagnosis, and plan of action in a structured patient note. Assessment of clinical reasoning on the basis of patient notes in an OSCE is often completed by trained physician raters, who evaluate performance on the basis of holistic traits such as interpretation of clinical data, thought process, and logic. This type of holistic or global rating is often thought of as the “gold standard” for clinical skills assessment and appears to have increased validity over task-specific checklists.7

Although a global rating completed by expert physicians may be a suitable method to assess clinical reasoning, this process is extremely resource intense. Our physician raters require, on average, a minimum of 5 minutes to rate each patient note; for our 8-station high-stakes OSCE, this involves rating 8 patient notes for up to 180 medical students: 1,440 notes × 5 minutes = 120 hours, or approximately 15 hours per case. Enlisting faculty members who can or are willing to devote hours to rating patient notes is a daunting task.

An alternative to global rating is analytic scoring, in which experts design a scoring rubric that can then be used to train nonexpert raters to assess individual components of patient notes (eg, by keyword or phrase matching). Mertler8 described the difference between holistic and analytic scoring, stating that whereas a holistic approach requires the rater to score the overall product as a whole, without judging the component parts, an analytic approach has the rater score separate parts of the product individually and sum them to reach a total score. Analytic scoring has been used to explore ratings completed by a variety of nonphysician medical raters, including nurses, medical students, and billing clerks,9 and has been found to be an effective method of giving feedback.10 Analytic scoring has been shown to have increased reliability over global ratings; however, even using regression-based techniques to control for confounding, analytic scoring has not yet been shown to accurately predict global rating scores.11,12 Such results suggest that global ratings by experts may capture aspects of clinical reasoning that are missed by lay raters, even when using a comprehensive scoring rubric.

If nonmedical raters could be trained to reliably assess OSCE patient notes, the burden on physician resources would be greatly minimized. In this study, we determined whether a nonmedical “lay” rater could be trained to accurately assess clinical reasoning in patient notes using an analytic scoring model. We developed a comprehensive scoring guide for a high-stakes, multidisciplinary, 4th-year medical student abdominal pain OSCE. Interrater reliability was measured between 2 physician raters and a nonmedical lay rater.

Methods

A comprehensive scoring guide was developed to assess medical student patient notes for a multidisciplinary abdominal pain case. The case, which is briefly presented in Figure 1, is a 42-year-old woman with acute left lower quadrant abdominal pain. This case was specifically designed to have several possible differential diagnoses, ranging across several different clinical specialties, and was used as 1 of 8 cases in a high-stakes (passing grade required for graduation) OSCE administered to rising 4th-year medical students. In the patient note, the students were expected to list their top 3 differential diagnoses in order of likelihood and include supporting evidence for the diagnosis, evidence against the diagnosis, and diagnostic tests, with expected results, needed to confirm each diagnosis (Fig. 2). The scoring guide was written in a collaborative effort by physicians across different clinical specialties, including sur-

---

**Figure 1** The clinical case: a multidisciplinary, high-stakes OSCE.

---

**Figure 2** Sample patient note.

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Supporting Evidence</th>
<th>Evidence Against</th>
<th>Diagnostic Tests and Expected Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
surgery, obstetrics and gynecology, pediatrics, and internal medicine. A condensed version of the scoring guide is shown in Figure 3.

A 3-point scale was used to score 3 core areas (supporting evidence for and against each diagnosis and diagnostic workup) in each of the top 3 diagnoses, for a possible total score of 27. Overall clinical reasoning (on a 4-point scale) was based on total points calculated from the 3 core areas across the top 3 diagnoses. To emphasize the multidisciplinary nature of the case, the overall clinical reasoning score was downgraded if the top 3 diagnoses did not include at least 1 obstetric or gynecologic diagnosis (Fig. 3).

An initial random sample of 25 4th-year medical student–written patient notes was rated by a physician and lay rater independently, after reviewing the scoring guide together for 30 minutes. Kappa statistics and simple percentage agreement were calculated to compare ratings in 3 core areas: supporting evidence for and against each possible diagnosis and diagnostic workup. The 2 raters met for 30 minutes to discuss discrepancies and then rated another random sample of 50 patient notes, for a total of 75. Kappa and percentage agreement statistics were again calculated. A 2nd physician rater independently rereated the initial 25 notes, as an “expert” control.

To assess internal consistency, Cronbach’s α coefficient was calculated for the physician and lay raters using scores from the 3 core areas for each of the top 3 diagnoses.

The results showed that Cronbach’s α coefficient was low for both the physician rater (.54) and the lay rater (.58), suggesting that the individual domains of clinical reasoning may be relatively independent. Agreement between the physician and lay rater in the initial 25-note sample was as follows: supporting evidence, 84% (κ = .69); evidence against, 71% (κ = .62); and diagnostic workup, 73% (κ = .69). After additional training and consensus development, agreement improved substantially for evidence against (87%; κ = .81) and diagnostic workup (84%; κ = .74) and remained similar for supporting evidence. Similar patterns and magnitudes of agreement were found within each of the highest frequency diagnoses, namely, ectopic pregnancy, appendicitis, and pelvic inflammatory disease (Table 1). Agreement in the initial 25-note sample was highest for the diagnosis listed with highest frequency (ectopic pregnancy): supporting evidence, 92% (κ = .81); evidence against, 92% (κ = .86);
and diagnostic workup, 71% (κ = .42). After additional training, agreement improved significantly for diagnostic workup, 84% (κ = .69), and remained similar in the other assessment domains. The mean agreement across the top 5 differential diagnoses is shown in Table 2.

The overall clinical reasoning score was calculated based on total points across 3 core areas in the scoring guide for each student’s top 3 differential diagnoses. The score was then categorized on the basis of point totals in combination with presence of at least 1 gynecologic diagnosis into the following 4 categories: poor, fair, good, and excellent (Fig. 3). Percentage agreement for overall clinical reasoning across all differential diagnoses was 73% (κ = .56) for physician versus lay raters and 84% (κ = .68) for physician versus physician raters. Figure 4 shows minimum differences by rater in terms of the number of students falling into each of the overall score categories, ranging from 1 to 6 students, depending on the category. The average point difference between lay and physician raters’ total points (absolute value) was 1.2 ± 1.5, out of 27 possible points.

### Comments

Clinical reasoning is a complex entity that is not easily operationalized or assessed. Performance on the USMLE Step 2 Clinical Knowledge section has been shown to have minimal redundancy with performance on the Step 2 Clinical Skills section; therefore, it is essential that both components be adequately assessed. The OSCE-based USMLE Step 2 Clinical Skills section evaluates competency in the following categories: integrated clinical encounter (including data gathering from history and physical examination and documentation), communication and interpersonal skills, and spoken English proficiency. Within the Clinical Skills module, data gathering and documentation scores have been shown to have a moderate to high correlation, but history gathering and physical examination have been shown to have only a modest to moderate relationship. All other components have low correlations. Because documentation of a patient note involves the synthesis of data gathering and other OSCE components into a differential diagnosis and management plan, it has been used as a surrogate for clinical reasoning. The patient note is the only instrument that allows for evaluation of all components of OSCE performance simultaneously, enabling the assessment of competency in individual areas, as well as overall clinical reasoning skills.

It has been shown that inconsistent rating of patient notes in the USMLE Step 2 Clinical Skills module contributes more substantially to measurement error than case specificity. Providing raters with comprehensive training and a detailed scoring rubric has been associated with an increase in the generalizability of resulting scores. Researchers have developed a number of different methods for assessing note-writing skills, including Web-based peer review or checklist-based strategies. Methods of evaluating patient notes to assess clinical reasoning must be consistent, reliable, and feasible.

### Table 1

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Percentage agreement (κ)</th>
<th>Supporting evidence</th>
<th>Evidence against</th>
<th>Diagnostic tests and expected results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ectopic pregnancy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physician rater vs lay rater</td>
<td>89% (.72)</td>
<td>89% (.81)</td>
<td>79% (.58)</td>
<td></td>
</tr>
<tr>
<td>Physician rater vs physician rater</td>
<td>83% (.59)</td>
<td>92% (.87)</td>
<td>96% (.87)</td>
<td></td>
</tr>
<tr>
<td>Appendicitis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physician rater vs lay rater</td>
<td>86% (.64)</td>
<td>77% (.64)</td>
<td>91% (.85)</td>
<td></td>
</tr>
<tr>
<td>Physician rater vs physician rater</td>
<td>73% (.47)</td>
<td>100% (1.00)</td>
<td>100% (1.00)</td>
<td></td>
</tr>
<tr>
<td>Pelvic inflammatory disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physician rater vs lay rater</td>
<td>88% (.78)</td>
<td>68% (.48)</td>
<td>60% (.41)</td>
<td></td>
</tr>
<tr>
<td>Physician rater vs physician rater</td>
<td>88% (.75)</td>
<td>63% (.37)</td>
<td>75% (.62)</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th></th>
<th>Supporting evidence</th>
<th>Evidence against</th>
<th>Diagnostic tests and expected results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physician rater vs lay rater</td>
<td>79%</td>
<td>80%</td>
<td>77%</td>
</tr>
<tr>
<td>Physician rater vs physician rater</td>
<td>77%</td>
<td>91%</td>
<td>90%</td>
</tr>
</tbody>
</table>
the traditional “subjective, objective, assessment, plan” (SOAP) note, which is similar to that used in the USMLE Step 2 Clinical Skills module and includes sections for history, physical examination, differential diagnoses and diagnostic workup.\(^\text{14,15}\) SOAP is a useful format because it is straightforward, reproducible, problem based, and widely used among health care practitioners. As an assessment tool, the quality of written SOAP format patient notes has been shown to correlate moderately with other measures of clinical ability, such as doctor-patient communication and data-gathering ability.\(^\text{16}\)

However, because of the SOAP note’s complexity, analytic scoring of this format can neglect to consider factors such as the presence of erroneous, potentially dangerous findings, or the absence of pertinent negative findings. The SOAP note format can also be challenging to assess when there is a discrepancy between the quality and/or organization of the history and physical examination and the accuracy of the differential diagnosis and workup. Therefore, SOAP notes have traditionally been scored by trained physician raters who can assess each note carefully, both holistically and in terms of its constituent elements.

For this study, a restructured patient note format was developed to simplify analytic scoring and account for potentially dangerous omissions. Our case of acute abdominal pain in a reproductive-age woman necessitated consideration of both general surgical and gynecologic causes. Points were deducted if students did not include at least 1 gynecologic diagnosis. We also developed a comprehensive scoring guide to connect supporting evidence for and against each differential diagnosis to the subsequent diagnostic workup. Each patient note was then assessed by 2 physician raters and 1 lay rater. Interrater reliability between both the physician and lay raters and the 2 physician raters was determined after 2 sessions of lay rater training. After two 30-minute training sessions using the detailed scoring guide, interrater reliability for physician and lay raters improved and was similar to that shown between 2 physician raters.

The use of trained physician raters to assess each patient note requires faculty resources that are often scarce. Our high-stakes OSCE assesses up to 180 students yearly, with 8 cases per student, which generates 1,440 individual patient notes. To reduce the burden on physician raters, educational researchers have examined the use of nonphysician raters, such as medical students and other medical professionals. In one study, medical student raters gave slightly better than average scores compared with faculty physician raters.\(^\text{17}\) In another study, assessments by medical student raters correlated closely with assessments by faculty raters.\(^\text{18}\) However, both of these studies used observational checklists and global ratings alone. They also failed to examine note writing as a means to measure clinical reasoning and did not provide the nonphysician raters with a comprehensive scoring guide.

The high interrater reliability between lay and faculty raters in this study demonstrates that nonmedical professionals, when trained appropriately and provided with a very specific, detailed rating rubric, can score patient notes in close alignment with physician raters. Lay rating of patient notes on core components of clinical reasoning is reliable and practical. Our next steps will focus on exploring the degree to which lay rating of the patient note is a valid method for assessing clinical reasoning by linking clinical reasoning scores with practice-related and patient-related outcomes. We will continue to incorporate analysis of interrater reliability to provide quality improvement and targeted rater retraining.

Conclusions

The findings of this study suggest that with adequate training, lay raters may act as examiners in the assessment of the patient note clinical reasoning score in a multidisciplinary, high-stakes OSCE.

References


