A Step Ahead—Evaluating the Clinical Judgment Skills of Incoming Interns

Andrea Wendling, MD; Philip Baty, MD

Background: Assessing the baseline competency of incoming interns is a challenge to faculty of all residency programs. This article describes and evaluates the Intern Clinical Judgment Evaluation, a tool developed to test the clinical judgment and communication skills of incoming interns. Methods: During orientation week, each intern is evaluated using structured clinical scenarios modeled after “on call” situations. Interns are given feedback on information collection, medical decision making, judgment, understanding of the medical system, and written and verbal communication skills. Data from 4 years is presented, comparing predictions of performance from faculty and from the Intern Clinical Judgment Evaluation to actual year-end intern performance data. Results: There was excellent agreement between the Intern Clinical Judgment Evaluation and the summary of interns’ performance (kappa value=0.8). Faculty predictions were a less accurate predictor of intern performance (kappa value=0.6). The Intern Clinical Judgment Evaluation had a sensitivity of 93% (13/14), specificity of 88% (15/17), positive likelihood ratio (LR+) of 7.9, and negative likelihood ratio (LR-) of 0.08. Comparatively, faculty predictions of resident performance had a sensitivity of 57% (8/14), specificity of 94% (17/18), LR+ of 10.3, and LR- of 0.40. Surveyed interns felt the evaluation process was a positive experience. Conclusions: The Intern Clinical Judgment Evaluation was a valid tool to predict which interns would require more faculty supervision during intern year. Based on solicited feedback from tested interns, it also was a positive testing experience.

(Fam Med 2009;41(2):111-15.)
interpretation of lab data, or communication skills—each as an isolated skill set in a typically station-based format. We designed a similarly structured testing model but focused on clinical performance and judgment rather than on independent skill sets.

The most widely recognized structured testing module for clinical performance has been the Advanced Cardiac Life Support (ACLS) model. This model has been shown effective when measuring retention of skills and performance.\(^{1-13}\) Skills learned during ACLS are used by residents, usually acting as members of a team, during management of acute cardiac arrest. In our experience, interns are rarely responsible for individual decision making during such situations. Interns do, however, often need to function autonomously in less life threatening but still critical situations, mainly at night when taking call. For this reason, we used the crisis-based format of ACLS but designed scenarios to test acute patient care situations that an intern would realistically need to manage alone.

To successfully manage a crisis situation, an intern needs to recognize that a patient is acutely ill, needs to understand that help is needed, and needs the confidence and ability to seek out that help. The intern also needs to be able to communicate the severity of the situation to the senior physician. This scenario is often stressful for incoming interns and could be life-threatening if managed incorrectly.

For these reasons, we developed the Intern Clinical Judgment Evaluation. The evaluation is designed to test each intern's clinical evaluation skills, judgment, and communication skills. Information gathered during the testing process is used to determine which interns should receive extra support early in their intern year, especially during call shifts. Since instituting the evaluation system, our program has identified issues such as gaps in medical knowledge, communication concerns, differing medical system expectations, and discrepancies between confidence and ability. With this information, we have worked to target early interventions with the goal of improving the success of our interns.

**Methods**

**Description of Evaluation Tool**

Over the course of the past 6 years, residents entering the intern class have participated in the Intern Clinical Judgment Evaluation. The evaluation process was developed by a faculty member and consists of two clinical scenarios representing typical acute inpatient on-call situations. A faculty facilitator leads the clinical scenarios for each intern, providing information when requested and cues to guide the intern through the scenarios.

Each scenario begins with the first call from a ward nurse alerting an intern that the nurse is concerned about a patient. The intern is then expected to gather necessary information by asking specific questions and make decisions based on the information at hand. The faculty facilitator uses scripted cards to provide information and cues so each intern receives the same information when it is requested. Complete history and exam skills are not tested; interns only need to ask for the necessary information. Interns are not given information that they do not request.

Prior to beginning the evaluation, the interns are instructed to respond to each scenario as they typically would while on call. They are told to ask for help whenever needed and are reminded of the typical sources of help that are available. If help from a senior team member is requested, the interns need to explain the situation as though they were on call. The facilitator then gives the next correct step in the scenario, redirects the intern if necessary, or gives the requested information. Scenarios are finished when the patient is stabilized; each scenario typically takes 10–15 minutes to complete.

After each clinical scenario, the intern is given brief feedback from the faculty facilitator covering whether help was called for appropriately, any deficiencies in basic skills, and expectations within our system.

We schedule two interns for each hour of testing. The first intern completes an observed clinical scenario then writes a chart note based on the scenario while the second intern is testing. The interns rotate again for the second scenario and then complete anonymous feedback on the experience rather than writing a second note. Chart notes are reviewed after the entire intern class has been tested.

At our program, a second faculty member observes each scenario via live video. We have found through trial and error that having two faculty members observe and score each scenario gives us a more complete assessment. Both this faculty observer and the faculty facilitator complete evaluation worksheets at the completion of each scenario. Worksheets include checklists for specific skills such as requesting and interpreting vital signs and clinical tests, developing basic treatment plans, understanding and communicating the severity of situations, and clinical judgment skills such as recognizing personal limitations and level of support needed. Using the worksheets, a consensus summary sheet is completed and submitted to the program director within 2–3 days. This summary focuses on the intern's ability to collect information, ability to interpret basic tests, medical decision-making, judgment, understanding of the medical system, confidence level, and written and verbal communication skills. The summary includes positive feedback as well as formal recommendations regarding areas on which each intern should focus during the first few months of intern year. Interns review the completed summary sheet when completed with their faculty advisor.
We dedicate two faculty members for 1 day of testing (8–10 interns). The faculty facilitator spends another 4–5 hours assimilating information and completing summary sheets.

Assessment of Tool

For 4 of the past 5 years, we have compared faculty members’ predictions of resident performance to the Intern Clinical Judgment Evaluation predictions. To obtain faculty prediction data, all faculty were surveyed during intern orientation week about their predictions of each intern’s clinical performance. Survey questions were graded on a 5-point Likert scale ranging from 1 (not likely) to 5 (very likely). Questions included “Do you predict resident will be able to recognize when she/he needs help? Will resident ask for help when needed? Will resident be able to assess the severity of situations? Will resident be able to communicate the severity of situations when necessary? Will resident’s chart notes be organized and provide an accurate representation of events?” A sixth survey question was worded “Please rate the resident’s medical knowledge base” and was graded on a 5-point Likert scale ranging from 1 (very poor) to 5 (exceptional). Predictions were obtained anonymously and were used only for research purposes. By obtaining faculty predictions of performance during orientation week, we felt these predictions would be based on the information known about each intern from his or her application and residency interview, inter-faculty discussions during the ranking period, and initial impressions during orientation week. We felt this was the baseline for how performance was predicted in most programs prior to interns beginning ward rotations. Numerical Likert values for faculty responses were pooled and averaged for each intern to obtain a prediction score. Prediction scores among residents in the same class year were then compared. We arbitrarily defined any resident scoring at least one standard deviation below the mean to have a prediction of needing more supervision or assistance; all other residents were considered to have a prediction of needing normal supervision.

Based on information obtained during the Intern Clinical Judgment Evaluation, we recorded whether we had predicted that our various interns were ready to function with normal responsibilities or whether they needed more supervision or assistance. Each year, faculty surveys were completed and returned prior to releasing orientation evaluation data, and orientation evaluation summaries were finalized prior to reviewing faculty predictions.

At the completion of intern year, the program director was interviewed regarding each intern’s performance. This structured interview included information gathered from the intern’s performance reviews, inservice testing, and written summary of performance. This summary interview was considered the gold standard regarding each intern’s clinical performance. Interns were divided into two groups based on this year-end summary. If an intern did not require formal remediation plans, did not need to repeat clinical rotations, and had positive performance reviews, then they were considered to have had “normal levels of supervision.” If an intern needed a formal remediation plan, needed to repeat clinical rotations due to performance or professional issues, or had negative performance reviews that required faculty intervention then they were considered to have had “more faculty supervision.”

Intern Clinical Judgment predictions, faculty predictions, and program director’s summary of performance were compared using kappa values. Sensitivities and specificities as well as likelihood ratios were obtained and compared for Intern Clinical Judgment predictions and faculty predictions using the program director’s year-end summary of performance as the gold standard. All data were analyzed using SPSS for MS Windows version 6.1.

Resident Survey

For the past 3 years we have also requested feedback from tested interns at the completion of testing. This has been in the form of anonymous typed, open-ended feedback recorded and saved in individual Word documents. Interns were asked to give both positive and negative feedback to improve the testing system. Pooled Word documents were reviewed yearly after faculty members had completed and submitted testing summary sheets.

Institutional Review

This project received a waiver from formal review by the St. Mary’s Health Center Institutional Review Board.

Results

Thirty-one residents were evaluated during 4 years between 2002 and 2006. One class of residents during that time period was not evaluated due to the retirement of a faculty facilitator; in subsequent years we resumed testing. Only one resident during the tested years was not evaluated due to difficulty obtaining a visa in time to attend orientation.

Of the 31 tested interns, 15 were determined to need extra supervision or assistance based on the Intern Clinical Judgment Evaluation summary (Table 1). Nine were predicted to need more supervision based on the faculty predictions. Of the 31 interns tested, 14 required more faculty time than the average based on the program director’s summary of the year.

Of the 14 interns who required more assistance, 13 were predicted using the Intern Clinical Judgment Evaluation but only eight were predicted by faculty sur-
Two interns identified as needing more assistance during the Intern Clinical Judgment Evaluation actually required little assistance from faculty; both were identified early in the year and were released to normal supervision levels. Only one intern who required more than the normal amount of supervision from faculty was not identified by the evaluation tool. Based on the program director’s summary, this intern’s problem area was one of professionalism and honesty.

Correlation between the two sets of predictions and the program director’s summary were then compared. The results showed excellent correlation between the Intern Clinical Judgment Evaluation predictions and the summary of interns’ performance (kappa value=0.8) while the faculty predictions showed only fair to good correlation (kappa value=0.6). The Intern Clinical Judgment Evaluation when compared to actual performance had a sensitivity of 93% (13/14), specificity of 88% (15/17), LR+ of 7.9, and LR- of 0.08. The faculty prediction when compared to actual performance had a sensitivity of 57% (8/14), specificity of 94% (17/18), LR+ 10.3, and LR- 0.40.

The results of the resident feedback survey were overwhelmingly positive. Twenty-four residents were asked for feedback over 3 years. Every resident surveyed was positive about the experience. Most residents responded that they were less nervous about taking call after completing the scenarios. Several residents asked for more scenarios throughout the year. Interestingly, when commenting, several residents called the process a “teaching module” rather than an “evaluation module,” even though the process was clearly explained as an evaluation tool. One resident felt it was difficult to imagine the patient, and two residents commented that being videotaped was stressful; however, even these three residents were positive about the experience.

**Discussion**

The Intern Clinical Judgment Evaluation showed excellent correlation with actual intern performance. Comparatively, faculty predictions of intern performance based on admissions data and early impressions were more likely to overestimate interns’ potential—although faculty were usually correct if they predicted an intern would need more supervision, they tended to miss more of these interns than were missed when the Intern Clinical Judgment Evaluation was used. Based on feedback obtained from residents, it was also a positive evaluation experience.

Our program uses the information gathered from the Intern Clinical Judgment Evaluation to help predict which interns may have difficulty during clinical rotations and to guide discussions regarding expectations and resources. The evaluation is used to guide faculty in developing early interventions and teaching goals to maximize each intern’s clinical performance, especially during the early months of the intern year.

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Comparison of Predictions From Intern Clinical Judgment Evaluation and Faculty Surveys With Interns’ Actual Performance</strong></td>
</tr>
<tr>
<td><strong>Actual Performance</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Orientation evaluation</strong></td>
</tr>
<tr>
<td>More supervision</td>
</tr>
<tr>
<td>Normal supervision</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Faculty prediction</strong></td>
</tr>
<tr>
<td>More supervision</td>
</tr>
<tr>
<td>Normal supervision</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

PPV—positive predictive value  
NPV—negative predictive value  
LR+ = positive likelihood ratio  
LR- = negative likelihood ratio
Limitations

Our study has several limitations. One limitation of our resident survey includes the fact that the clinical scenarios were completed at a single program with a small number of faculty facilitators. In a different situation, residents may not be as comfortable during the testing.

Second, although residents were assured of anonymity, being new to a program, they may not have felt comfortable sharing negative feedback. The overwhelmingly positive feedback regarding an observed structured evaluation experience, however, is in agreement with several other published studies.14-16

A limitation of the faculty prediction survey is that faculty members, depending on their involvement with interviewing, had varying amounts of contact with interns prior to orientation week. All faculty would have been at ranking meetings when potential interns were discussed, however, and this limited amount of information is what predictions were based on prior to our formal evaluation process.

Another limitation is that the Intern Clinical Judgment Evaluation may have influenced the program director evaluation that was used as the gold standard; this would have similarly been influenced by the faculty predictions of performance. Faculty time was measured subjectively, which is another limitation of the study.

Conclusions

When designing our Intern Clinical Judgment Evaluation, our primary goal was to help our interns to be more successful early in their residency career. This is in line with Epstein’s previously published recommendations that an evaluation system should be a form of learning that inspires confidence.17 Based on the results of our study, we found that the Intern Clinical Judgment Evaluation was a valid tool to predict which interns would require more faculty supervision during intern year. Based on solicited feedback from tested interns, it also appeared to be a well-received educational testing experience.

Acknowledgment: This study was presented at the Society of Teachers of Family Medicine 2008 Annual Spring Conference in Baltimore.

References


Corresponding Author: Address correspondence to Dr Wendling, 11133 Summerhill Way, Charlevoix, MI 49720. 231-547-2533. Fax: 231-582-5338. aharmeling@msn.com.