

## MODELING SURGICAL RESIDENT PERFORMANCE

Sara Waxberg & Caroline G.L. Cao  
Department of Mechanical Engineering  
Tufts University, Medford, MA

### ABSTRACT

The goals of this work were to evaluate the methods used for assessing surgical residents and to model surgical resident performance. Currently, residents are evaluated by the attending surgeons using a one-page paper evaluation form after each rotation in a particular department. These subjective questionnaires require the evaluators to rate residents in competency areas that are thought to define a successful surgeon. An electronic database was created from resident performance records collected over the past 33 years from the Department of Surgery at the Tufts-NEMC. A usability study examined the effectiveness of the design of the evaluation form and the competency measures. Analysis showed nine changes in format between 1972 and 2005, varying in the competencies rated and rating scales used. Regression analysis was used to model the performance of surgical residents. Results showed that judgment ( $p < 0.0001$ ), initiative ( $p < 0.0001$ ), and reaction to stress ( $p = 0.0206$ ) were significant predictors of a successful outcome. This model may be used to predict the success of new residents and possibly target weaknesses in the surgical education curriculum.

### INTRODUCTION

To become a general surgeon, a medical school graduate must complete an additional five years of residency training within the General Surgery (GS) Department at a teaching hospital. During their surgical education, residents must complete a number of rotations, or training within various departments and/or hospitals, to obtain a range of experience. After each rotation, the attending surgeons evaluate their residents by questionnaire, which use rating scales for a number of qualities commonly associated with an effective surgeon. In addition to evaluations by the attendings, self evaluations, peer evaluations, and the American Board of Surgery In-Training Examination (ABSITE) are normally considered before a resident is allowed to move on to the next year of residency training. Despite the long history of surgical education and training, none of these formats for evaluating residents has been standardized. Research has shown that scores from the ABSITE did not show a significant relationship with either skill testing on a laparoscopic video trainer or intraoperative assessment of skills during a laparoscopic cholecystectomy (Scott et al, 2000).

Clearly, the ABSITE alone is not an adequate method for evaluating residents. This means that the questionnaire at the end of each rotation becomes the primary means of resident assessment. In 1999, the Accreditation Council for Graduate Medical Education (ACGME), in an attempt to standardize the assessment of resident performance, instituted a required coverage of six competencies (patient care, medical knowledge, professionalism, system-based practice, practice-based learning and improvement, and interpersonal and communication skills) on the evaluation forms administered by attendings. "Patient care" rates the residents on their ability to be "compassionate, appropriate, and effective in the treatment of health problems and the promotion of health" for patients. "Medical knowledge" is demonstrated by an understanding of biomedical, clinical, and cognate sciences and their application to patient care. "Practice-based learning and improvement" refers to a resident's ability to investigate, evaluate, and improve upon patient care practices. The "interpersonal and communication skills" of the residents should result in effective exchange of information with patients, their families, and other health care professionals. Having "professionalism"

suggests that the residents can carry out their professional responsibilities, keep to ethical principles, and be sensitive to a diversity of patients (ACGME, 2003). However, the evaluation tool used for assessing these competencies is not standardized across teaching hospitals.

A retrospective study was conducted to examine the evaluation forms used by the attending surgeons to judge the residents at the Tufts-New England Medical Center (NEMC) in Boston, MA. We have modeled the performance of the residents in an attempt to identify weaknesses in the current residents' competency repertoire and to suggest guidelines for standardized evaluation and training. This paper presents preliminary results of our examination of the evaluation forms and modeling of the performance data.

## METHODS

### Data source

Records for 365 surgical residents over the past 33 years at the Tufts-NEMC were collected. The residents had an average of 22 evaluations over the course of the training program, with a range of one to 76 evaluations per record. A total of 7863 evaluations were available for analysis. The evaluation tool is a one-page paper-based questionnaire (see sample in Figure 1). Each questionnaire contains standard information such as the dates of the rotation, evaluators' names, rotation name, post-graduate year, overall performance rating, and comments. Other topics of interest such as chief resident potential or interest in teaching changed over time. An electronic database was created from these questionnaires to allow for sorting and analysis. Since the questionnaire had evolved over the years, there were variations in the questions asked and the rating scales used.

Interviews and literature review were conducted to gather information on the evaluation methods at a national, state, and hospital-specific level. The past director of the Office of Surgical Education was interviewed to gather information on the organization at the Tufts-NEMC, evaluation methods, and ACGME requirements for the surgical program. The current chief resident also provided

insight on the surgical education curriculum, scheduling of residents, and ACGME requirements for the residents, through an interview.

**NEMC General Surgery - Resident Evaluation Form**

We appreciate your help in the continuous process of evaluating our residents. Please indicate below how well this resident performed on his/her most recent rotation on your service:

Name: \_\_\_\_\_ Dates: \_\_\_\_\_ Service: \_\_\_\_\_

3 = Excellent 2 = Satisfactory 1 = Improvement necessary X = Can't evaluate

<b>Clinical Skills:</b>	(Circle one each)
• Data gathering and interpretation	3 2 1 X
• Technical ability	3 2 1 X
• Judgment	3 2 1 X
• Organizational skills/Administrative skills	3 2 1 X
<b>Medical Knowledge:</b>	
• Evidence of self education	3 2 1 X
• Conference participation and preparation	3 2 1 X
• Overall knowledge base	3 2 1 X
• Teaching skills and commitment	3 2 1 X
<b>Professionalism:</b>	
• Reliability and responsiveness	3 2 1 X
• Honesty	3 2 1 X
• Initiative	3 2 1 X
• Timeliness and punctuality	3 2 1 X
• Accurate and timely completion of medical records	3 2 1 X
• Protection of patient confidentiality	3 2 1 X
• Understanding of and adherence to Ethical Principles	3 2 1 X
<b>System Based Practice</b>	
• Practices evidence based and cost effective surgery	3 2 1 X
• Effective use of resources in tests and consult ordering	3 2 1 X
• Works with available resources to achieve timely and appropriate disposition and discharge	3 2 1 X
• Awareness and Management of Practice Environment	3 2 1 X
<b>Interpersonal and Communication Skills</b>	
• Verbal Communication Skills	3 2 1 X
• Accuracy and completion of written notes	3 2 1 X
• Reaction to stress	3 2 1 X
• Ability to work collaboratively with Nurses and other care providers	3 2 1 X

**Narrative Evaluation (required) please indicate any specific areas in need of attention:**

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

**Overall Performance (circle one):**

Exceeded Expectations for NEMC resident	At "par" for NEMC resident	Better performance was expected	Unsatisfactory
---	----------------------------	---------------------------------	----------------

Signature: \_\_\_\_\_ Date: \_\_\_\_\_

Figure 1. Sample evaluation form from 2005

### Data analysis

Descriptive statistics were obtained from the quantitative data, such as averages and frequency of counts. A logistic regression analysis was also performed to determine the relative importance of each aspect in evaluating a resident.

In order to compensate for variations in scale between the versions of the evaluations, scores were normalized. Scores were normalized to a number between zero and one by dividing the score by the total number of levels on the rating scale (i.e. three on a scale of five would be 0.6).

The data set was further limited by excluding the variables that were not consistent over the years. Other than demographics, ten variables were included in the model: length of rotation, number of evaluators, fund of knowledge, technical ability, judgment, conference participation, reliability, initiative, reaction to stress, and interpersonal skills.

A logistic regression model was employed to fit a sample set of data of 1405 evaluations for 70 residents. Our model defined a successful outcome as receiving an overall normalized score of 0.8 or greater.

### Data categories

Several data categories were common in all questionnaires:

- 1) Subject – The subject number randomly assigned.
- 2) Order – The chronological order of rotations for each resident.
- 3) Start and End Date – The start and end of date range for the evaluation.
- 4) Length of Rotation – The number of days included in the date range.
- 5) Date Evaluated – The date of evaluation or date reviewed by the attending physician.
- 6) Time to Evaluation – The number of days between the end date and date evaluated.
- 7) Evaluators – The attending physicians or staff that completed the questionnaire.
- 8) Number of Evaluators – The number of people who completed the evaluation.
- 9) PGY – Post-graduate year, or level of experience, of the resident.
- 10) Service/Rotation – The department, hospital, or title of the rotation completed.
- 11) Topic Scores – Numbers were assigned to each of the intervals for the varying rating scales starting at zero (“can’t evaluate”) and increasing with performance for variables such as reliability, technical ability, etc.
- 12) Comments/Additional Comments/Narrative Evaluation – Keywords were coded for open-ended questions.

## RESULTS AND DISCUSSION

Typically, a surgical residency program accepts medical school graduates who indicate a desire to specialize in GS. Upon entering the program, the residents’ initial impressions made on the attendings often suffer from a “halo/devil” effect (Thorndike, 1920). There is a chance that because the program has already accepted the residents, it may be very hard to get rid of them

and/or the attending surgeons want them to do well because they have picked them and are therefore more lenient on the scores they give. Also, perhaps if a student has a bad first interaction with an attending the first week of their residency, the attending surgeon may feel animosity and carry that along for the resident’s entire time in the program. This can potentially skew the scores given on the questionnaire. This is something that needs to be taken into consideration when designing a resident evaluation.

Our analysis showed that rotation assignments were determined by scheduling needs, not by an education curriculum. First year residents did not always complete the same rotations or the same order of rotations. Even though the rotations not covered during their first year would be completed in the later years, the residents who missed out earlier in their careers may be at a disadvantage in acquiring necessary skills. The order of exposure may be critical to effectiveness in their practice. Our on-going analysis examines this by comparing the rotation sequences to scores and overall outcome for the resident (if they finish the program or not) (results not available).

The program at the Tufts-NEMC did not have a prescribed method of objective evaluation for technical skill. There were no standard means for determining whether a resident had the motor skills necessary to complete surgical procedures effectively. In a survey conducted by Baldwin, Paisley, and Brown (1999), technical skills were found to be “absolutely necessary” by the majority of physicians surveyed. Therefore, a “technical skills” component is needed within the program to determine competence and to supplement the subjective questionnaire.

An assessment of the tool used over the last 33 years revealed that the questionnaire has changed its general form nine times. The differences were in the competencies being rated and the rating scales used. Scales have changed in the number of intervals, from three to five, and their meanings (e.g. adequate vs. satisfactory vs. average). The aspects rated by the attending surgeons have changed in number and in which of the six competencies were covered. The earliest evaluation only had eight topics while the most recent covers 25 topics.

A usability evaluation of the currently implemented form revealed that only a scale of one to three or not applicable is used (see Figure 1). This scale is quite small. A three-point scale confines the attending to an above average, average, or below average rating. Usually a scale should include four or five levels in order to allow for a distinction between the superb and the neutral (Frery, 1996). There was not any accompanying instruction or training for filling out the questionnaire. Many of the words used within the categories were ambiguous. Another point that should be considered is the overall performance ranking, a measure that compares residents to others within the same program. This could potentially cause a resident to receive lower or higher marks biased by the most recent group within an attending surgeon's rotation. The attending would most likely have a lingering memory from the most recent residents and compare the residents within that group as opposed to the overall Tufts-NEMC resident ideal. An increase in the number of levels and rewording of those levels is necessary.

We observed that over time the average normalized overall grade for residents has increased. It must be noted that a score of 1.0 does not mean that the resident had achieved perfection; rather that he/she was above par for a Tufts-NEMC resident. This suggests that either resident quality has improved or that the evaluators have become more lenient over time.

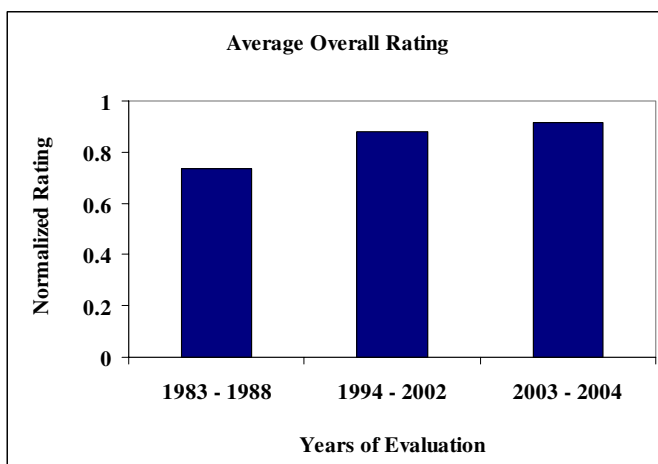


Figure 2. Overall score for residents over time

The average of the overall scores was 0.895, but both the median and mode were 1.0. This means that many of the evaluators rated their residents as excellent or that they exceeded the expectations for a Tufts-NEMC resident. Perhaps an alteration of the levels of scoring scales is necessary. If everyone is above the expectation of a Tufts-NEMC resident, then perhaps the standards also need to be raised to better differentiate the residents from one another.

The lowest normalized score for all aspects rated over all templates was 0.737 for “fund of knowledge”. This suggests that attending surgeons felt that the residents did not do enough reading and studying on the issues for that department. The highest normalized score was for “reliability” at 0.862. This is a personal characteristic rather than a talent or ability perhaps making it hard to give residents poor ratings. In addition, it is very hard to say that someone is not reliable. It can be an insult to someone's personal character. This item may need to be rephrased for intention or excluded.

The findings from the logistic regression analysis allowed us to recognize which competency area(s) could best predict a successful outcome. Results showed that judgment ( $p < 0.0001$ ), initiative ( $p < 0.0001$ ), and reaction to stress ( $p = 0.0206$ ) were significant predictors of a successful outcome within residency training. These characteristics above all others are personal characteristics. There is a possibility that raters associated overall score with a resident's personality rather than technical ability or fund of knowledge. These results agree with a previous study done by Risucci (1989) that demonstrated that ratings given to residents by attending surgeons were primarily influenced by interpersonal skills and secondly by ability. This may suggest that there is a weakness in the current method of evaluation and further analysis is warranted.

There were some limitations in collecting and analyzing the data for this study. Due to the sensitive nature of the content, ABSITE scores, entrance interview scores, and personal attributes of the residents were not accessible. We could not link the residents' information to their present status. In addition, since the study was retrospective, there were inherent limitations that we could not control, such as the change in versions of the questionnaire or missing data points.

## CONCLUSION

Identifying the weaknesses of individual residents can help customize training in the early stages of residency. Recognizing a successful sequence of rotations and the components most predictive of success within a surgical residency program can help to focus graduate medical programs and surgeons in the future. The outcomes of this research will include a performance model of surgical residents and recommendations for a standardized computer-based evaluation tool for surgical performance.

## ACKNOWLEDGEMENTS

This work was supported by a grant from the NSF (IIS-0238284). We also thank Durwood Marshall for his assistance with our data analysis.

## REFERENCES

- Accreditation Council of Graduate Medical Education. (2003). Common program requirements. [On-line] Retrieved February 27, 2005 from <http://www.acgme.org>.
- Baldwin, P.J., Paisley, A.M., & Paterson Brown, S. (1999). Consultant surgeons' opinion of the skills required of basic surgical trainees. *British Journal of Surgery*, 86(8), 1078-1082.
- Frary, Robert B. (1996). Hints for designing effective questionnaires. *Practical Assessment, Research, & Evaluation*, 5(3). [On-line] Retrieved February 24, 2006 from <http://PAREonline.net/getvn.asp?v=5&n=3>.
- Risucci, D.A., Tortolani, A.J., & Ward, R.J. (1989). Ratings of surgical residents by self, supervisors, and peers. *Surgical, Gynecology & Obstetrics*, 169(6), 519-526.
- Scott, D.J., Valentine, R.J., Bergen, P.C., Rege, R.V., Laycock, R., Tesfay, S.T., & Jones, D.B. (2000). Evaluating surgical competence with the American Board of Surgery in-training examination, skill testing, and intraoperative assessment. *Surgery*, 128(4), 613-622.
- Thorndike, E.L. (1920). A constant error on psychological rating. *Journal of Applied Psychology*, 4, 25-29.