Pretraining and Posttraining Assessment of Residents’ Performance in the Fourth Accreditation Council for Graduate Medical Education Competency

Patient Communication Skills

Rajiv Y. Chandawarkar, MD; Kimberly A. Ruscher, MD, MPH; Aleksandra Krajewski, MD; Manish Garg, PhD; Carol Pfeiffer, PhD; Rekha Singh, MD; Walter E. Longo, MD; Robert A. Kozol, MD; Beth Lesnikoski, MD; Prakash Nadkarni, MD

Hypothesis: Structured communication curricula will improve surgical residents’ ability to communicate effectively with patients.

Design and Setting: A prospective study approved by the institutional review board involved 44 University of Connecticut general surgery residents. Residents initially completed a written baseline survey to assess general communication skills awareness. In step 1 of the study, residents were randomized to 1 of 2 simulations using standardized patient instructors to mimic patients receiving a diagnosis of either breast or rectal cancer. The standardized patient instructors scored residents’ communication skills using a case-specific content checklist and Master Interview Rating Scale. In step 2 of the study, residents attended a 3-part interactive program that comprised (1) principles of patient communication; (2) experiences of a surgeon (role as physician, patient, and patient’s spouse); and (3) role-playing (3-resident groups played patient, physician, and observer roles and rated their own performance). In step 3, residents were retested as in step 1, using a crossover case design. Scores were analyzed using Wilcoxon signed rank test with a Bonferroni correction.

Results: Case-specific performance improved significantly, from a pretest content checklist median score of 8.5 (65%) to a posttest median of 11.0 (84%) \(P = .005\) by Wilcoxon signed rank test for paired ordinal data \((n=44)\). Median Master Interview Rating Scale scores changed from 58.0 before testing \((P = .10)\) to 61.5 after testing \((P = .94)\). Difference between overall rectal cancer scores and breast cancer scores also were not significant.

Conclusions: Patient communication skills need to be taught as part of residency training. With limited training, case-specific skills (herein, involving patients with cancer) are likely to improve more than general communication skills.

Arch Surg. 2011;146(8):916-921

CLEAR AND EMPATHIC COMMUNICATION builds stronger relationships between physicians and patients. In general, surgical residents have no formal curriculum for patient communication and are expected to acquire these skills in practice. Communication skill is recognized as a core competency by the Accreditation Council for Graduate Medical Education, which requires documentation by residency programs of formal processes that assess residents in this competency, provide feedback to residents, and use assessment results to progressively improve residents’ competence.1 One widely accepted method of training and assessment of both clinical skills and interpersonal communication is the use of standardized patients (SPs, also referred to as patients instructors). These nonclinicians (typically actors) are healthy individuals who have been trained to play the role of patients, evaluators, and occasionally instructors in an objective structured clinical examination.2,3 In the surgical context, testing of residents using this examination devised for end-of-life family conferences and disclosure of complications has been described.4-5 The goal of this project was to teach surgical residents to incorporate patient-centered communication skills into their practice, providing emotional support, transition, and continuity of care, as well as information and education, involving family and friends and respecting patient values and preferences.
This article describes the design and evaluation of a pilot project aiming to teach surgical residents patient-centered communication skills.

### METHODS

This prospective study was approved by the institutional review board at the University of Connecticut Health Center; 44 residents from the Department of Surgery were included. A baseline survey determined their demographics and awareness of patient communication. Residents were given the option to participate in a 3-step exercise: pretest, training, and posttest assessment. This exercise, incorporated into the residents’ 80-hour work rule requirements, was held at the same time as the routine didactic educational sessions. There was no compromise in the clinical teaching curriculum or the lecture sessions. Because of factors such as call schedule and availability, only 30 residents had evaluable data for all sessions.

### STUDY DESIGN

#### Weeks 1 and 2: Pretraining Evaluation

Residents were randomized to one of 2 simulated cases in which a surgeon had to inform a patient about a new diagnosis of either breast cancer or rectal cancer. Diagnosis of cancer is very difficult for patients to accept and is possibly the most common situation in surgical practice requiring sympathetic and effective communication. Use of cancer scenarios for both preevaluations and postevaluations also minimizes case-specific variability.6,7

A resident/SP encounter lasted approximately 10 to 15 minutes. The SP spent an additional 5 minutes grading the resident on 2 scales (discussed in the “Assessment” subsection of the “Methods” section): a case-specific content checklist and the Master Interview Rating Scale (MIRS).

#### Week 3: Training

The residents attended a 90-minute workshop, in which they participated in an interactive teaching session on the ability to communicate with patients. This included a lecture by a surgeon (professor of surgery) about his experiences as a patient and as a relative of a patient (in this case, his spouse) whom he accompanied to all her physician visits. Furthermore, the director of the Clinical Skills Program at the University of Connecticut Health Center (C.P.) instructed residents formally on communication with patients.

The next session contained approximately 30 minutes of role-playing (15 minutes) followed by a postsession analysis (15 minutes). Residents were divided into groups of 3, playing the roles of patient, surgeon, and observer; the observer scored the encounter using the pretest scales. After discussion/analysis of performance within the group, the observer summarized the analysis for the entire group, providing feedback to the “surgeon” role-player regarding strengths and areas for improvement.

#### Weeks 4 and 5: Posttraining Evaluation

The residents had a follow-up evaluation with an SP, using a scenario identical to that conducted at week 1, except that residents who had taken part in a breast cancer simulation now participated in a rectal cancer simulation and vice versa.

### Experiment Design

A total of 5 SPs—3 (female) for breast cancer and 2 (male) for rectal cancer—participated in the study as independent contractors (unassociated with the residency program) after 6 hours’ training. They were not blinded as to whether a particular evaluation was before or after training. One female SP and both male SPs were involved in both weeks 1 and 3. The residents knew that they were dealing with simulations.

### Institutional Review Board Concerns

The baseline survey and demographic data were anonymous; the residents were not required to enter quasi-identifying information such as age, sex, year of residency, and surgical specialty. In the evaluation of clinical and communications skills before and after training, residents identified themselves with a random 4- or 5-digit numeral that was used for both the pretest and posttest. This number was selected by the resident and was not able to be correlated with the resident’s identity, because the baseline survey did not use this number.

There was no pressure by faculty for residents’ participation. In addition, residents were free to drop out of the study at any time, without repercussions.

Because of these design constraints, analyses such as correlation of communications skills with baseline beliefs or year of residency could not be performed.

### Assessment

#### Content Checklist

Table 1. Items on Case-Specific Content Checklist

<table>
<thead>
<tr>
<th>Item</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Performs an interval history of biopsy recovery</td>
</tr>
<tr>
<td>2.</td>
<td>Asks whether there was pain or discomfort from surgery</td>
</tr>
<tr>
<td>3.</td>
<td>Gives a warning shot about the bad news</td>
</tr>
<tr>
<td>4.</td>
<td>States that the diagnosis is cancer</td>
</tr>
<tr>
<td>5.</td>
<td>Explains the type of cancer</td>
</tr>
<tr>
<td>6.</td>
<td>Asks for patient’s emotions</td>
</tr>
<tr>
<td>7.</td>
<td>Asks about patient’s knowledge of the disease</td>
</tr>
<tr>
<td>8.</td>
<td>Gives prognosis</td>
</tr>
<tr>
<td>9.</td>
<td>Discusses treatment</td>
</tr>
<tr>
<td>10.</td>
<td>Describes and provides referral to oncologist</td>
</tr>
<tr>
<td>11.</td>
<td>Provides assurance of continuing care</td>
</tr>
<tr>
<td>12.</td>
<td>Asks about support and offers to meet with others to explain</td>
</tr>
<tr>
<td>13.</td>
<td>Offers to meet again and provides other contacts for information</td>
</tr>
</tbody>
</table>

*All responses are recorded as yes/no.

### General Communication Skills Assessment

The MIRS is widely used for evaluation of training programs for imparting communication skills to physicians.6,9 It was originally developed as a 27-item questionnaire at the University of Arizona College of Medicine by Stillman et al.10 The MIRS, which is rated by a patient instructor, addresses general communication aspects (eg, pacing of interview, questioning skills, summarizing, lack of jargon, and empathy). Each item is rated on a scale of 1 (poorest) to 5 (best).

Several subset MIRS versions exist. Woods6 and O’Sullivan et al6 used slightly different 6-question subsets for nutrition counseling and geriatric training, respectively. We used a 16-item subset of a version developed by Theresa A. Thomas of Eastern Virginia Medical School (copyright 2000). In this version, the items are partially anchored; that is, to partially reduce the
Table 2. Demographic Data on Baseline Survey Respondents

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residents, No.</td>
<td>44</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>25</td>
</tr>
<tr>
<td>Female</td>
<td>17</td>
</tr>
<tr>
<td>Not reported</td>
<td>2</td>
</tr>
<tr>
<td>Age, median (range), y</td>
<td>29 (25-49)</td>
</tr>
<tr>
<td>Postgraduate year level</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Not reported</td>
<td>3</td>
</tr>
<tr>
<td>Specialty&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Surgery</td>
<td>25</td>
</tr>
<tr>
<td>Ear, nose, and throat</td>
<td>2</td>
</tr>
<tr>
<td>Urology</td>
<td>2</td>
</tr>
<tr>
<td>Oromaxillary and facial surgery</td>
<td>3</td>
</tr>
</tbody>
</table>

<sup>a</sup>Two were not reported.

effect of interobserver variability, there are clear and detailed textual definitions for each item of what grades 1, 3, and 5 imply. (Grades 2 and 4 are not defined explicitly.) The supplementary Appendix shows our subset with the anchoring information (http://www.archsurg.com).

**STATISTICAL ANALYSIS**

All results are reported using only medians and ranges because the individual variables are not parametric. For the MIRS, although individual items are based on (ordinal) Likert scales, some studies<sup>11,12</sup> have incorrectly treated total score as parametric. However, applying a Shapiro-Wilks test for normality to our MIRS total score data showed the parametric assumption to be violated ($P < .001$).

**RESULTS**

**DEMOGRAPHICS AND BASELINE SURVEY QUESTIONNAIRE**

The residents’ demographic data and responses to a baseline survey questionnaire are summarized in Table 2 and Table 3, respectively. Most residents strongly agreed that patient communication is important.

**CONTENT CHECKLIST SCORES: BEFORE AND AFTER TRAINING**

Scores could range from 0 to 13 (13, perfect score). The posttest median score (11.0 [84%]) was significantly higher than the pretest score of 8.5 (65%) ($P = .005$, Wilcoxon signed rank test for paired ordinal data) ($n = 44$), indicating that residents’ case-specific performance was clearly improved by training.

The differences between overall rectal cancer scores (median, 9.5) and breast cancer scores (median, 10.0) were not significant. This indicated that these clinical scenarios were interchangeable and that there was no inherent bias in comparing performance in either of these cases.

**GENERAL COMMUNICATIONS SKILLS: MIRS SCORE BEFORE AND AFTER TRAINING**

**MIRS Total Scores**

The MIRS scores can range from 16 to 80 (80, perfect score), with a mid-point of 48. The medians and ranges for the pretest and posttest scores were 58.0 (range, 26-72) and 61.5 (range, 31-74). Differences between the pretest and posttest were not significant, by the Wilcoxon signed rank test.

**Differences on Individual MIRS Items**

As stated in the “Methods” section, the score on an individual item ranged from 1 (poorest) to 5 (best). After using Wilcoxon signed rank test with a Bonferroni correction for multiple hypothesis testing, the only significant difference between the pretest and posttest scores was seen for 1 item, support systems ($P = .03$), the median of which improved from 2.5 to 4.

Most important, the difference between overall rectal cancer scores and breast cancer scores was not significant. Median scores for both were 59.0, indicating that general communications skills was not influenced by the type of cancer.

**COMMENT**

Our results show a discrepancy between measured improvement in condition-specific communication skills and measured improvement in general communication skills. The former was highly significant, whereas the latter was marginal and not statistically significant. We now discuss the possible reasons for this discrepancy.

**DURATION AND TYPE OF TRAINING**

Although case-specific skills might be taught and reinforced in a single day (eg, through the use of checklists, as required by the Joint Commission in contexts such as the Universal Protocol<sup>13</sup>), this may not suffice to ensure improved performance on most MIRS items. In this study, the support systems item was the only one that improved significantly. Inspection of this item in the supplementary Appendix indicates that optimal performance only requires the clinician to ask about emotional support, financial support, health care access, and other resources and to suggest appropriate community resources. Performance on this item is amenable to improvement using an easily memorized checklist.

Factors such as lack of jargon, verbal facilitation skills, and effective summarizing can also be taught. However, it is likely that sustained coaching with repeated practice, rather than a one-time session, will be required to see improvement in performance.

Finally, factors such as empathy, nonverbal communication, and concern for another’s viewpoint have a sig-
significant innate component, often involving attitudes, and are more difficult to change. Ishikawa et al,14 while trying to teach medical students nonverbal skills in a single training session, found that although training increased students’ awareness of nonverbal communication, it did not improve actual performance. Skeptics such as Davis15 believe that empathy cannot be taught. Finally, short-term, 1-time interventions or interventions of a primarily pedagogical nature may have little permanent effect in changing attitudes: Jewe’s16 before-and-after data on more than 500 undergraduates showed that completing a business ethics course did not significantly affect respondents’ ethical attitudes.

SUBJECTIVITY OF EVALUATION

In the MIRS, the use of anchoring of individual items with textual definitions is supposed to reduce the effect of interevaluator variability due to subjective factors, but if one inspects the definitions of the grades, 1 means complete failure, 3 is partial, and 5 implies perfect performance. However, the choices of grade 2 or grade 4 (which also imply less-than-perfect performance but which are below and above what the observer might consider a midlevel performance) are subjective and may be influenced by interpersonal factors rather than the factor that is supposedly being evaluated in isolation.

Unless the session is videotaped so that it can be reviewed by the SP, it may well be that having only 5 minutes to accurately score 16 items on the MIRS (in addition to 13 items on the checklist) may overwhelm the SP. In this situation, the SP is likely to be influenced by factors such as the overall bearing and pleasantness of the resident.

In the study of O’Sullivan et al,9 which used 3 separate SPs to evaluate geriatric residents in 3 separate clinical scenarios (without a pre-post design), the interrater agreement was reported as poor; however, statistics were not reported. In that study, the SP-resident encounter was observed through closed-circuit television by a preceptor (a faculty member) who rated the resident on problem-specific clinical skills, and the overall agreement between the preceptor’s rating and the SP’s rating for a given encounter was also poor. These authors cited the work of Reznick et al,17 whose data indicate that a minimum of 10 scenarios is required to achieve a reliability of 0.85 to 0.9.

NATURE OF PARTICIPANTS

The MIRS was originally developed to evaluate medical students. Data for MIRS scores across diverse populations of residents and medical students are not publicly available, so the average total and per-item score is unknown. One may assume that residents, simply through greater sustained contact with patients, attain a somewhat greater communication skill level compared with students. The median per-item score of the residents evaluated before training was 3.625 (58.0/16), which changed to 3.844 (61.5/16). In other words, the residents as a whole were above average in the ratings at baseline, so they did not have as much need to improve.

VALIDITY CONCERNS

Interdependence Between Individual MIRS Items

Individual items on the MIRS, as with most rating scales, are intended to be orthogonal, ie, they are supposed to measure separate and independent aspects of communication. It is not clear, however, that this is the case. For example, although the training booklet distributed to US Peace Corps workers18 states, “facilitation requires skills in asking questions, paraphrasing, and summarizing,” the MIRS “verbal facilitation” item is scored separately from summarizing ability and effective question strategy.

Nonorthogonal rating scales can lead to double-counting of particular skills, biasing the score based on whether the person being rated scores highly or poorly on those skills. Orthogonality, or its lack, can be measured experimentally using the statistical technique of factor analysis.19 The widespread use of factor analysis in psychometric/behavioral contexts was promulgated by Cattell.20 In factor analysis, one evaluates a suitably large set of respondents on the scale and computes an inter-item correlation matrix that is the starting point for computations. In principle, if respondents’ scores on a particular item show a high positive correlation with scores on another item (eg, effective summarizers also question effectively and poor summarizers are poor questioners), then both of these items are likely to be measuring the same underlying performance factor, ie, they are not orthogonal.

Factor analysis is routinely applied as part of the validation phase of multi-item rating scales because identification of orthogonality can result in simplification of

<table>
<thead>
<tr>
<th>Table 3. Baseline Survey Questionnaire Data: Summary of Responsesa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statement</td>
</tr>
<tr>
<td>Communicating effectively with patients is critical to surgical practice</td>
</tr>
<tr>
<td>I have excellent communications skills</td>
</tr>
<tr>
<td>I am better at communicating with patients and their families than my peers</td>
</tr>
<tr>
<td>I have been given the proper training in communication with patients</td>
</tr>
<tr>
<td>I can communicate effectively even with very difficult patients</td>
</tr>
<tr>
<td>I can effectively discuss end-of-life decisions with my patients</td>
</tr>
<tr>
<td>I can effectively break bad news to patients</td>
</tr>
<tr>
<td>There is too little time to establish a meaningful relationship with my patients and families</td>
</tr>
</tbody>
</table>

aTotal number of respondents, 44.
bResponses to each question were recorded on a Likert scale in which 1 indicates strongly disagree; 2, slightly disagree; 3, slightly agree; and 4, strongly agree.
the scale (so that, for example, it need only include 15 items instead of 30, making it easier to administer). We could not, however, locate any published work validating MIRS from this aspect.

**Discrimination Concerns**

Interestingly, the MIRS scores of 2 of the highest-scoring residents dropped from 72 and 69 to 62 and 60 after training. The score of another resident decreased from 52 before training to 31 after training. If the assumption is that the MIRS is measuring only general communication skills (which was the intention behind the test), then such skills are unlikely to deteriorate dramatically within 2 weeks; one must conclude either that the MIRS has poor discrimination or that other subjective (eg, interpersonal) factors are influencing the overall score (or both).

**Problem Dependence**

The MIRS is supposed to measure general skills that would apply to any patient communication scenario. Individual MIRS items, such as empathy, verbal and nonverbal facilitation skills, and lack of jargon, may be independent of the specific clinical problem. However, other items, such as organization, types of questions, pacing of interview, and achievement of closure, may have a significant problem-sensitive component: knowledge of and experience with the clinical condition being discussed may lead to greater fluency, with more focused questioning and superior plan formulation.

Schuwirth and van der Vleuten’ pointed out that even with standardized patient simulations that use yes/no–based checklists to minimize interrater variability, intercase variability continues to be the main factor that influences scores. They stated that this can be addressed only through use of a sufficiently large number of scenarios.

**Reliability Concerns**

In our pilot study, the effect of interrater variability may have been amplified because of the non-uniform distribution of SPs with respect to pretraining and posttraining evaluation. We analyzed the data of each SP by performing a 2-way nonparametric analysis of variance (the Friedman test) computing the within-SP differences and prepost differences. Although the overall differences were not significant, we observed that the SP who had the lowest overall MIRS scores (median, 50) was one who was involved only in posttraining evaluation; the SP involved only in the pretest had a median MIRS score of 57.5.

These differences may be because the 2 SPs evaluated different sets of residents; however, interrater variability cannot be ruled out. The study of Chipman et al,’ reported in the *Journal of Surgical Education*, used 2 patient instructors (acting as patient and relative) per resident per objective structured clinical examination; these SPs had undergone a previous training session of 3 hours. In their study, which evaluated the abilities of 8 residents on a multi–Likert item scale devised by the authors, inter-SP agreement measured using Cronbach α varied from 0.70 to as low as 0.42; the last value was considered poor. These data add support to the concern that interrater variability is a significant factor in all objective structured clinical examinations involving evaluation by nonprofessional raters.

Despite the concerns discussed herein, improvement in case-specific skills, measured by the same raters, was highly significant. There are several possible reasons for this; one interpretation may be that yes/no scales are less influenced by subjective factors than are Likert-type scales when used by nonprofessional raters.

**Residents’ assessment of their patient communication skills** indicates that there is an immediate need for a formal educational curriculum. Our results show that case-specific improvements seem more amenable to measurable improvement than general communications skills, at least with the limited short-term training that we used. Such skills can be assessed over a longer period, perhaps by incorporating this model and assessments from year to year.

Surgical and nonsurgical residency programs will benefit by helping residents incorporate patient needs and opinions into the care team’s decision-making process. Principles such as emotional support, transition and continuity of care, provision of information and education, involvement of family and friends, and respect for patient values and preferences will form the basis of our educational series.

The educational content described in this study is simple and can be easily accomplished in any teaching hospital or community health center setting. We recommend that these educational measures be included in the residency training curriculum with modifications that may be specific to each program. Without communication skills, even the best surgical training would be rendered ineffective.

**Accepted for Publication:** April 8, 2011.

**Correspondence:** Rajiv Y. Chandawarkar, MD, Division of Plastic Surgery, Department of Surgery, University of Connecticut Health Center, Mail Code 1601, 263 Farmington Ave, Farmington, CT 06030 (Chandawarkar@uchc.edu).

**Author Contributions:** Study concept and design: Chandawarkar, Krajewski, Pfeiffer, Singh, Longo, and Kozol.

**Acquisition of data:** Chandawarkar, Ruscher, Garg, Pfeiffer, and Singh.

**Analysis and interpretation of data:** Garg, Pfeiffer, Kozol, Lesnikoski, and Nadkarni.

**Drafting of the manuscript:** Chandawarkar, Garg, Lesnikoski, and Nadkarni.

**Critical revision of the manuscript for important intellectual content:** Chandawarkar, Ruscher, Krajewski, Garg, Pfeiffer, Singh, Longo, Kozol, Lesnikoski, and Nadkarni.

**Statistical analysis:** Pfeiffer and Nadkarni.

**Obtained funding:** Chandawarkar.

**Administrative, technical, and material support:** Chandawarkar, Ruscher, Krajewski, Garg, Pfeiffer, and Singh.

**Study supervision:** Chandawarkar, Longo, and Kozol.

**Financial Disclosure:** None reported.

**Funding/Support:** This study was supported by the 2010 Picker Foundation Educational Challenge (Dr Chandawarkar).
REFERENCES

15. Davis CM. What is empathy, and can empathy be taught? Phys Ther. 1990;70(11):707-715.