EDITORIAL

Shifting Approaches for Evaluation of Resident Performance
From Competencies to Milestones

Lia S. Logio, MD

Outcome measures have replaced process measures as the new currency of quality. For example, the quality of care for patients with diabetes is now measured by results of hemoglobin A_{1C} tests, not just the percentage of patients who received the test. The same shift has occurred in graduate medical education (GME). Competency-based education serves to hold residency programs accountable for the outcomes of its graduates.

Related article page 2253

In 2001, the Accreditation Council for Graduate Medical Education (ACGME) implemented its Outcomes Project defining 6 core professional competencies for physicians.[1,2] Residency training programs and advanced subspecialty fellowships were asked to measure the competence of trainees in patient care, medical knowledge, interpersonal and communication skills, professionalism, practice-based learning and improvement, and systems-based practice. For the first time, all ACGME-accredited programs were using a standard language to describe performance of physicians in training.

In July 2013, the ACGME rolled out the Next Accreditation System with one of its key elements to measure and report outcomes through the use of educational milestones.[3] The ACGME invited each specialty to define the stepping-stones within each of the 6 core competencies that represented the usual trajectory of progression in the professional development of that specialty. Through consensus, narrative criteria for each step of these defined milestones were mapped using a developmental scale with the goal rating in each subcompetency labeled "ready for unsupervised practice." These milestones provided trainees with a more explicit set of the expectations for performance at every level with considerably more granular detail about the composite of "good doctoring" within each discipline. Within internal medicine, resident performance shifted from ratings grouped into 6 core competencies to ratings split into 22 reporting subcompetencies. For example, interpersonal and communication skills now include 3 components: communicates effectively with patients and caregivers, communicates effectively in interprofessional teams, and appropriately uses and completes health records.

In this issue of *JAMA*, Hauer and colleagues[4] present a cross-sectional comparative study of grouping vs splitting to measure the performance of 21 284 internal medicine resident trainees—7048 postgraduate year 1 [PGY-1], 7233 PGY-2, 7003 PGY-3 residents—during the 2013-2014 academic year. During the 2013-2014 inaugural year of reporting milestones, programs used both the milestone ratings and the preceding tool, the resident annual evaluation summary (RAES), collected each year by the American Board of Internal Medicine as part of tracking individuals toward board eligibility. With both assessment tools submitted simultaneously, the authors were able to compare data collected by each on the same set of residents. The authors explored the validity of milestones and attempted to determine if use of milestones improved the assessment of competence. Although milestones may lead to better assessment, it is still too early to tell.

The study findings suggest evidence for the validity of milestones through 3 key findings. First, there was a small correlation between the RAES system and the milestones system; corresponding RAES ratings and milestone ratings demonstrated progressively higher Spearman correlations across training years, ranging among competencies from 0.31 to 0.35 for PGY-1 residents to 0.43 to 0.52 for PGY-3 residents.

Second, for graduating residents, poor ratings in medical knowledge using milestones, as previously demonstrated with low medical knowledge scores on RAES,[5] correlated with failure to pass the ABIM certification examination. Among the 6260 PGY-3 residents who attempted the certification examination, higher medical knowledge ratings were correlated with higher examination scores (RAES Spearman r, 0.40; milestone medical knowledge 1 r, 0.37; and milestone medical knowledge 2 r, 0.30). The 618 residents who failed the ABIM examination had lower ratings using both rating systems for medical knowledge than those who passed (RAES rating difference, −0.9; milestone medical knowledge 1 rating difference, −0.3; and milestone medical knowledge 2 rating difference, −0.2).

Third, the 4 milestones that measure professionalism provided a higher level of discrimination through their narrative anchors to identify lapses in professional behavior than did the RAES. For instance, of the 1190 residents across all training years with a professionalism milestone rating of less than 2.5, indicating concern about professional behavior, 1161 (97.6%) were rated as satisfactory (n = 809) or superior (n = 352) in professionalism on the RAES rating system. In addition, a comparison of low ratings in medical knowledge with low rating in professionalism revealed that of the 7003 PGY-3 residents, only 26 had a milestone rating lower than 2.5 on either of the 2 medical knowledge subcompetencies (0.3%), whereas 1190 of the 21 284 in all training

years had a professionalism milestone rating lower than 2.5 (5.6%). With a nearly 10-fold difference in prevalence of poor professionalism scores, the findings suggest that either faculty are particularly harsh when rating professionalism or there is a small but significant problem related to professionalism behaviors in resident trainees. Either way, these findings deserve further study.

The study by Hauer et al also demonstrates that although the correlation between the RAES and the milestone ratings was modest, the steeper slope across years of training seen with the milestone ratings suggests greater discernment and variation across performance. Ratings for PGY-3 residents, however, converge, raising the question of whether the investment of time, talent, and resources to develop and implement the milestone rating system was worthwhile. To answer this, several potential contributing effects of other key elements of the Next Accreditation System should be considered.

One such element, the clinical competency committee, requires programs to operationalize a committee responsible for reviewing the progress of all residents in the program.[6] Members of the committee include active teaching faculty who serve together as an advisory group to the program director. The committee formalizes a scheduled periodic review of resident performance as a mechanism to synthesize data from multiple sources into reportable data to submit to the ACGME every 6 months. The committee has substantially altered monitoring resident performance into a more iterative process with a team of people systematically and routinely looking at the data, a continuous quality improvement model.

Skeptics might ask whether this is just the difference in the scale. The language in the milestone ratings provides highly specific narratives that define deliberate behaviors expected at each level of training. These developmental anchors clearly inform the raters around the stepwise progression of doctoring behaviors with a much richer language for identifying the level at which an individual resident performs. Furthermore, to operationalize the Next Accreditation System, faculty were enlisted and educated on the milestone rating system, providing an opportunity to train faculty evaluators toward a standardized approach in rating resident performance. This inadvertent faculty development combined with participation in clinical competency committee discussions may have calibrated faculty behaviors in their own assessments of resident performance.

Although the study by Hauer et al provides helpful insights, several limitations should be acknowledged. The authors report that a number of milestone ranking data were missing from the set but do not report for which of the 22 subcompetencies nor to what degree data were missing. For any rating system to be accurate and reliable, it must be based on directly observable behaviors and not surrogate markers. Competency-based education relies on direct observation, but the ability to routinely watch learners perform their various tasks creates a logistical challenge.[7,8] Both the RAES ratings and the milestones ratings rely heavily on the subjective assessments of teaching faculty. As with the ACGME Outcomes Project, the years that follow major educational model shifts require a steep learning curve for programs to interpret new concepts and to translate the language and ideas for teaching faculty who supervise the clinical work of the trainees.[9] Because this analysis used data from the inaugural year of implementation, milestones performance is likely to change with additional experience.

Two potentially important features of the RAES system are not represented in the milestone system. The dichotomous scoring (satisfactory or unsatisfactory) for moral and ethical behavior may be more judicious than rating this construct on a developmental scale. The RAES system also included a final "overall clinical competence" rating for the whole performance of doctoring. Sometimes the sum is greater than the parts. Overall clinical competency potentially measures how an individual physician balances the complex interdependence between the knowledge, skills, and attitudes that are core to the practice of medicine.

With the level of government support provided to graduate medical education through the Centers for Medicare & Medicaid Services, the trained physicians who are the product of GME programs are a public good. As such, scrutiny to ensure these programs are preparing physicians to practice in the health care environment of today and tomorrow is a worthy investment, especially in light of evidence demonstrating that a strong influence on how physicians practice in the long term may come from how they learned to practice during residency.[10] Both RAES competencies and milestone subcompetencies aim to measure the outcome of training. The study by Hauer and colleagues provides a good first attempt to discern whether milestones are better than the evaluation system they replaced. It is not entirely clear yet. Further investigation in prospective longitudinal studies is warranted, especially studies that focus on the value of milestones to the trainees and the measurable influence of focused feedback to which residents can respond with a goal of improving over the course of training.

Ultimately, something is lost and something is gained in the shift from grouping to splitting, from judging to coaching, from process to outcome in this model of continuous professional development. Because attainment of "ready for unsupervised practice" is not required for program completion, it must be assumed that these data will represent part of a continuum on the performance of doctoring that may begin as early as medical school and last through maintenance of certification. Hauer and colleagues have provided important evidence on the validity of educational milestones in internal medicine. But to achieve the ultimate goal, milestones must demonstrate their power to provide meaningful and actionable feedback to individuals on how to improve their role as practicing physicians. Accomplishing this will require an explicit set of expectations in an environment replete with mentoring relationships, honest discourse, and humility to work on deliberate practice toward improvement, an aspiration for everyone in health care.

ARTICLE INFORMATION

Author Affiliation: Medicine, Weill Cornell Medical College, New York, New York.

Corresponding Author: Lia S. Logio, MD, Department of Internal Medicine, Weill Cornell Medical College, 525 E 68th St, M507, PO Box 130, New York, NY 10065 (lil9051@med.cornell.edu).

Conflict of Interest Disclosures: The author has completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and reports receiving royalties from McGraw Hill Publishing.

REFERENCES

1. ACGME. Outcome project. http://www.ucdenver.edu/academics/colleges/medicalschool/departments/pediatrics/meded/fellowships/Documents/ACGME%20Outcome%20Project.pdf. 1999. Accessed October 31, 2016.

2. Swing SR. The ACGME outcome project: retrospective and prospective. *Med Teach*. 2007;29(7):648-654.

3. Nasca TJ, Philibert I, Brigham T, Flynn TC. The next GME accreditation system—rationale and benefits. *N Engl J Med*. 2012;366(11):1051-1056.

4. Hauer KE, Vandergrift J, Hess B, et al. Correlations between ratings on the resident annual evaluation summary and the internal medicine milestones and association with ABIM certification examination scores among US internal medicine residents, 2013-2014. *JAMA*. doi:10.1001/jama.2016.17357.

5. Shea JA, Norcini JJ, Kimball HR. Relationships of ratings of clinical competence and ABIM scores to certification status. *Acad Med*. 1993;68(10)(suppl):S22-S24.

6. Andolsek K, Padmore J, Hauer KE, Holmboe E. Clinical competency committees: a guidebook for programs. https://www.acgme.org/Portals/0/ACGMEClinicalCompetencyCommitteeGuidebook.pdf. Accessed October 31, 2016.

7. Williams RG, Dunnington GL, Mellinger JD, Klamen DL. Placing constraints on the use of the ACGME milestones: a commentary on the limitations of global performance ratings. *Acad Med*. 2015;90(4):404-407.

8. Holmboe ES. Realizing the promise of competency-based medical education. *Acad Med*. 2015;90(4):411-413.

9. Malik MU, Diaz Voss Varela DA, Stewart CM, et al. Barriers to implementing the ACGME Outcome Project: a systematic review of program director surveys. *J Grad Med Educ*. 2012;4(4):425-433.

10. Asch DA, Nicholson S, Srinivas S, Herrin J, Epstein AJ. Evaluating obstetrical residency programs using patient outcomes. *JAMA*. 2009;302(12):1277-1283.