RDME Research and Development in Medical Education
Res Dev Med Educ, 2021, 10, 2
doi: 10.34172/rdme.2021.002
https://rdme.tbzmed.ac.ir
TUOMS PRESS

Original Research

Discriminatory ability of milestones: An analysis of milestone variability by obstetrics and gynecology subspecialty

Emily Hinchcliff[1*], Kaitlyn James[2,3], Kristina Dzara[2,4,5], Lori R. Berkowitz[2,4]

[1]Department of Gynecologic Oncology and Reproductive Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX
[2]Department of Obstetrics, Gynecology, and Reproductive Biology, Massachusetts General Hospital, Boston, MA
[3]Deborah Kelly Center for Outcomes Research, Department of Obstetrics, Gynecology, and Reproductive Biology, Massachusetts General Hospital, Boston MA
[4]Harvard Medical School, Boston MA
[5]Brigham Education Institute, Brigham and Women's Hospital, Boston, MA

Article info

Article History:
Received: 9 Feb. 2021
Accepted: 21 Feb. 2021
epublished: 15 Mar. 2021

Keywords:
Professionalism
Clinical Competency Committees
Competency-based medical Education
Milestones
Residency
Obstetrics and gynecology

Abstract

Background: Little evidence exists regarding Accreditation Council for Graduate Medical Education (ACGME) milestone discriminatory ability. This short report describes variability in milestone scores by category to determine their utility in discerning high and low performers in a single Obstetrics and Gynecology residency.

Methods: A Clinical Competency Committee (CCC) design was implemented with four subcommittees, each responsible for a predetermined milestones subset: Obstetrics, Gynecology, Ambulatory Practice, and Professional Activities. Milestone scores for 44 residents per year over four biannual evaluation cycles (2014-2016) were evaluated, for a total of 176 independent evaluations.

Results: Findings indicate that discriminatory ability, assessed by variability between resident scores, differed by subcommittee. Subcommittees that were primarily tasked with evaluating clinical- and procedural-based milestones demonstrated lower discriminatory ability among trainees.

Conclusion: Greater Professional Activity milestone variability indicates better differentiation; future research should determine correlation of these findings with other professionalism performance metrics and novel intervention strategies.

Introduction

In residency training, evaluation milestones are defined by the Accreditation Council for Graduate Medical Education (ACGME) with the goal of assessing and tracking trainee performance.[1,2] Within each subspecialty, core competencies create subspecialty-specific objective milestones within a defined developmental framework from novice to proficient; trainees must demonstrate increasing levels of autonomy as they progress.[3] However, little data exists to determine the milestones' ability to differentiate within residents. This discriminatory ability, while not necessarily essential for tracking competence, may provide programs with indicators of high and low performers and allow for subsequent intervention. Recent literature indicates that ACGME milestones may fall short in identifying struggling trainees, with only 22% having language to describe critical deficiencies.[4]

The Integrated Residency Program in Obstetrics and Gynecology at the Brigham and Women's Hospital/Massachusetts General Hospital (BWH/MGH) has a Clinical Competency Committee (CCC) structure that includes four independent subcommittees evaluating different milestones subgroupings. Thus BWH/MGH is uniquely positioned to provide essential information regarding milestones and their discriminatory ability. As training programs continue within the milestones' implementation discovery phase, reports of successes and challenges in the early adoption period are crucial.

Materials and Methods

CCC subcommittees

The CCC design within the BWH/MGH residency was based on the premise that specialized clinical faculty would have greater interaction with trainees in their area

***Corresponding author:** Emily Hinchcliff, Email: emhinchcliff@mdanderson.org

of expertise, and ideally should limit their evaluations to be within those areas to maximize knowledge of trainee performance and allow for more accurate milestone assessment. The CCC structure included four independent subcommittees evaluating different milestone subgroupings in an attempt to minimize evaluator bias about global performance of a trainee by assigning evaluators solely in their area of expertise and interactions with each trainee.

The four CCC subcommittees included Obstetrics, Gynecology, Ambulatory Practice, and Professional Activities. For the three clinical subcommittees, the main scope of practice fell within that subcommittee; for example, subspecialists in surgical fields such as Gynecologic Oncology were assigned to the Gynecology CCC, Maternal Fetal Medicine (MFM) subspecialists to the Obstetrics CCC, and Family Planning subspecialists to the Ambulatory Practice Committee. Each subcommittee was tasked with evaluating all trainees on a subset of milestones, pre-determined by residency leadership and relating directly to the subcommittee members' scope of practice. The tools used to assess these competencies incorporated data from multiple sources, including global assessment of performance (rotation evaluations from multiple raters and over multiple time points) as well as completion of administrative tasks.

Study Protocol

Biannual evaluation milestone scores were obtained for all residents and deidentified for the first two evaluation cycles following milestone implementation in Fall 2014, Spring 2015, Fall 2015, and Spring 2016. The first two cycles were analyzed to capture any early implementation validity concerns. All analyses were performed in Stata/IC, Version 14.2 (StataCorp LP, College Station, TX), with a P value of <0.05 considered statistically significant.

Milestone assessment: comparison across CCC subgroups

To determine the milestones' ability to discern between high- and low-performing residents, milestone subgroup standard deviations were analyzed. This analysis was based on the assumption that, while the majority of residents will cluster around the expected milestone performance score for the respective year of training, the standard deviation across milestones represents the separation between highest and lowest performance. Tightly clustered milestone scores are less able to discern differences in performance, while broad separation represents greater difference between residents. To analyze milestone score variation across PGY classes overall, Fall 2014 and Fall 2015 scores were combined. The standard deviations between the cumulative PGY class Fall scores were compared by milestone using Levene's test for homogeneity of variances.

Results

Overview

There were four milestones assessment cycles from Fall 2014 to Spring 2016 with 44 residents per year, resulting in a total of 176 independent resident evaluations. The CCC subcommittee structure was feasible and well-received by faculty committee members.

Analysis of milestones across CCC subgroups with advancing training

While absolute numerical scores were relatively similar across subcommittees, variability of scores differed significantly (Figure 1). All CCC subcommittees demonstrated statistically significant differences between PGY1 and at least one other training year. In the OB subcommittee, the variability in PGY1 and PGY2 was significantly smaller than that of PGY3 and PGY4 (P <0.02 and P<0.01, respectively), while in the GYN subgroup PGY1 was different than PGY2 (P <0.02). No other year comparisons were statistically significantly different. PGY1 in the Ambulatory subgroup was an outlier, with a significantly smaller variation of scores (SD 0.61) when compared to all other PGY years within the same subgroup (P <0.001 for all), while no other PGY years differed. The outlier nature of these milestones may be due to program rotation design. In our program, trainees start their GYN continuity clinic experience, which contributes significantly to Ambulatory milestones, in PGY2; thus, these milestones are likely less relevant in PGY1 and scores are likely more similar.

Within the Professional Activities CCC subcommittee, there was significantly more variation in the distribution of milestone scores and a larger overall difference. The standard deviation increased consistently as the training level increased, from 1.63 to 3.80 to 4.62 to 6.19 from PGY1 to PGY4 respectively. PGY1 was significantly

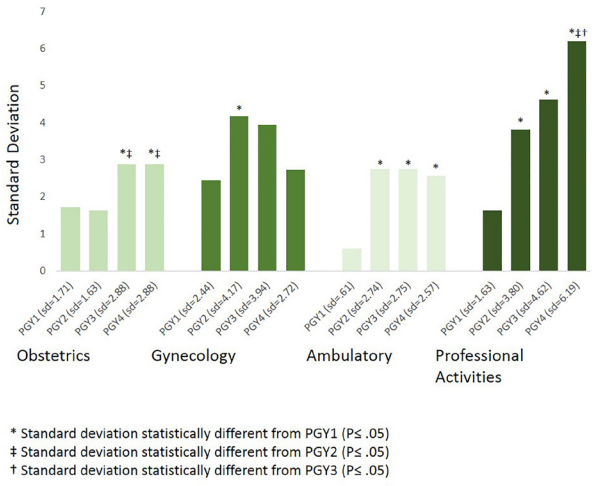

Figure 1. Milestone Score Variation Across Subcommittees. Standard deviation of milestone scores listed by PGY year for each of the four subcommittee designations: Obstetrics, Gynecology, Ambulatory, Professional Activities.

different than all other years (P <0.34), while PGY2 was not significantly different than PGY3 but differed from PGY4 (P <0.004). PGY3 and PGY4 demonstrated a trend toward significant differences (P <0.053). To summarize, there was no clear trend in the variability across training years for OB, GYN, and Ambulatory milestones. While there were some differences, the overall difference in SD was small (OB: 1.17, GYN: 1.73, Ambulatory: 2.14). However, Professional Activities milestones demonstrated significantly more variability between years, and showed a clear trend toward wider standard deviations as residents progressed in their training years.

Discussion

We found that milestone subspecialty category grouping – GYN, OB, Ambulatory, and Professional Activities – resulted in resident evaluation variation. While overall scores were similar across CCC groups, the range of scores was broadest in the Professional Activities subcommittee. The first conclusion that can be reached from these results is that the discriminatory ability of milestones, at least during early implementation, appears limited. There has been significant literature addressing the validity and utility of milestone metrics across trainee specialties. Literature suggests that since the adoption of ACGME milestones, programs have maintained consistency in ratings over time and validity assessments have demonstrated discriminatory ability between trainee years and increasing scores with advancing training.[5-8] However, more recent evaluations have begun to highlight some of the challenges and potential inaccuracies within the milestone scoring system. Interprogram variability has been reported, as has variability between specialties.[9,10] In terms of milestone accuracy within a particular evaluation, Beeson et al. evaluated the rate of "straight line scoring" (SLS, defined as a resident being assigned the same score across milestone subcompetencies) and showed that a small but meaningful number of programs submitted SLS ratings. Because of the statistical improbability of SLS, any SLS ratings reduce the validity assertions of the milestone assessments. SLS rates have also been found to vary by year of training and between procedural and medical subspecialties.[11]

In terms of the discriminatory ability of milestones, our study mirrors what has been found in other specialties. In family medicine residency, a study found that individual residents differed only based on their year of training and there were no identifiable differences between residents at similar levels.[12] Similarly, when program directors were surveyed regarding the discriminatory ability of milestones, 44% of urology program directors felt that they never or almost never accurately distinguished between residents.[13] Therefore, this report adds to the growing body of literature that milestone scores may not capture key differences between trainees.

However, the second conclusion that can be reached from this data is that there was greater variability in the Professional Activities subcommittee, indicating that these milestones' discriminatory ability was greatest. We cannot conclusively determine the cause, and it is beyond the scope of this research to determine if low performance in this domain was associated with additional poor performance metrics or subsequent individualized remediation programs. It is possible that other CCCs are driven by the inclusion of primarily skills-based milestones, which may be perceived as a dichotomous measure of whether a resident can or cannot perform a procedure independently. Non-skills-based Professional Activities milestones may be evaluated on a continuous spectrum and may be sensitive to increased granularity to assess aptitude and detail the rate of progress. Alternatively, an argument could be made that Professional Activities milestones are more subjective, resulting in greater variability; however, many of these milestones depend on administrative tasks, which are not subjective. Additionally, milestone subjectivity does not explain the stable variability within the procedural skills CCC subcommittees as residents advanced in their training while there was broadening variability in the Professional Activities subcommittee.

Interestingly, that Professional Activities milestones may be better able to capture performance disparities is consistent with prior literature from other fields, such as general medicine, which also noted greater variability in professionalism competencies.[14,15] A 10-year retrospective review of Canadian residents also found professionalism to be a core competency in which problem residents had difficulty compared to their residency counterparts.[16] Additionally, a longitudinal analysis of both qualitative and quantitative evaluation data within general surgery, including milestone levels, found that the highest number of ACGME-related subthemes within qualitative comments were related to professionalism, indicating that this may be a competency where evaluators have more suggestions for resident performance improvement.[17]

Conclusion

Dividing the CCC into subspecialty committees provided a unique approach allowing for analysis of milestone scoring by subspecialty category. In the first two years of implementation, the Professional Activities subcommittee exhibited greater variability than three other clinical skills-based committees. This indicates that these milestone subcompetencies may be better able to discern between residents throughout their training. Given that greater variability is present even among residents in the first year of training, further research is needed to determine if the detected variability correlates with other metrics of performance. If so, it could serve as an early warning sign of poor performance, and the delineation of professionalism-based milestones versus clinically/surgically-based milestones attainment may be a more

relevant way to analyze milestone ratings. Importantly, the impact of these results and subsequent interventions during training, as well as the prediction of individual trainee success along the entirety of a medical career, remains to be seen.

Ethical approval

The research was approved by the Partners Human Research Committee.

Competing interests

The authors report no conflicts of interest.

Authors' contributions

EH and LRB developed the study, obtained ethics approval, and collected the data. KJ analyzed the data. EH, KJ, KD, and LRB interpreted results. EH drafted the manuscript. EH, KJ, KD, and LRB reviewed and finalized the manuscript.

Acknowledgements

The authors thank Amy Stagg, MD and Ruth Tuomala, MD for their support.

References

1. Hauer KE, Cate OT, Boscardin CK, Iobst W, Holmboe ES, Chesluk B, et al. Ensuring resident competence: a narrative review of the literature on group decision making to inform the work of clinical competency committees. J Grad Med Educ. 2016;8(2):156-64. doi: 10.4300/jgme-d-15-00144.1.
2. Schwind CJ, Williams RG, Boehler ML, Dunnington GL. Do individual attendings' post-rotation performance ratings detect residents' clinical performance deficiencies? Acad Med. 2004;79(5):453-7. doi: 10.1097/00001888-200405000-00016.
3. Withiam-Leitch M, Olawaiye A. Resident performance on the in-training and board examinations in obstetrics and gynecology: implications for the ACGME Outcome Project. Teach Learn Med. 2008;20(2):136-42. doi: 10.1080/10401330801991642.
4. Kinnear B, Bensman R, Held J, O'Toole J, Schauer D, Warm E. Critical deficiency ratings in milestone assessment: a review and case study. Acad Med. 2017;92(6):820-6. doi: 10.1097/acm.0000000000001383.
5. Aagaard E, Kane GC, Conforti L, Hood S, Caverzagie KJ, Smith C, et al. Early feedback on the use of the internal medicine reporting milestones in assessment of resident performance. J Grad Med Educ. 2013;5(3):433-8. doi: 10.4300/jgme-d-13-00001.1.
6. Beeson MS, Holmboe ES, Korte RC, Nasca TJ, Brigham T, Russ CM, et al. Initial validity analysis of the emergency medicine milestones. Acad Emerg Med. 2015;22(7):838-44. doi: 10.1111/acem.12697.
7. Hamstra SJ, Yamazaki K, Barton MA, Santen SA, Beeson MS, Holmboe ES. A national study of longitudinal consistency in ACGME milestone ratings by clinical competency committees: exploring an aspect of validity in the assessment of residents' competence. Acad Med. 2019;94(10):1522-31. doi: 10.1097/acm.0000000000002820.
8. Turner TL, Bhavaraju VL, Luciw-Dubas UA, Hicks PJ, Multerer S, Osta A, et al. Validity evidence from ratings of pediatric interns and subinterns on a subset of pediatric milestones. Acad Med. 2017;92(6):809-19. doi: 10.1097/acm.0000000000001622.
9. Heath JK, Dine CJ. ACGME milestones within subspecialty training programs: one institution's experience. J Grad Med Educ. 2019;11(1):53-9. doi: 10.4300/jgme-d-18-00308.1.
10. Hu K, Hicks PJ, Margolis M, Carraccio C, Osta A, Winward ML, et al. Reported pediatrics milestones (mostly) measure program, not learner performance. Acad Med. 2020;95(11S):S89-S94. doi: 10.1097/acm.0000000000003644.
11. Hinchcliff E, Gunther J, Ponnie AE, Bednarski B, Onstad M, Shafer A, et al. A not so perfect score: factors associated with the rate of straight line scoring in oncology training programs. J Cancer Educ. 2020. doi: 10.1007/s13187-020-01855-6.
12. Peabody MR, O'Neill TR, Peterson LE. Examining the functioning and reliability of the family medicine milestones. J Grad Med Educ. 2017;9(1):46-53. doi: 10.4300/jgme-d-16-00172.1.
13. Sebesta EM, Cooper KL, Badalato GM. Program director perceptions of usefulness of the accreditation Council for Graduate Medical Education Milestones System for urology resident evaluation. Urology. 2019;124:28-32. doi: 10.1016/j.urology.2018.10.042.
14. Park YS, Zar FA, Norcini JJ, Tekian A. Competency evaluations in the next accreditation system: contributing to guidelines and implications. Teach Learn Med. 2016;28(2):135-45. doi: 10.1080/10401334.2016.1146607.
15. Tekian A, Borhani M, Tilton S, Abasolo E, Park YS. What do quantitative ratings and qualitative comments tell us about general surgery residents' progress toward independent practice? Evidence from a 5-year longitudinal cohort. Am J Surg. 2019;217(2):288-95. doi: 10.1016/j.amjsurg.2018.09.031.
16. Sullivan C, Murano T, Comes J, Smith JL, Katz ED. Emergency medicine directors' perceptions on professionalism: a Council of Emergency Medicine Residency Directors survey. Acad Emerg Med. 2011;18 Suppl 2:S97-103. doi: 10.1111/j.1553-2712.2011.01186.x.
17. Park YS, Riddle J, Tekian A. Validity evidence of resident competency ratings and the identification of problem residents. Med Educ. 2014;48(6):614-22. doi: 10.1111/medu.12408.