The American Journal of Surgery 221 (2021) 505–508

Contents lists available at ScienceDirect

The American Journal of Surgery

journal homepage: www.americanjournalofsurgery.com

Original Research Article

Is there a gender bias in milestones evaluations in general surgery residency training?

Lauren Kwasny [a], Saad Shebrain [a, *], Gitonga Munene [a, b], Robert Sawyer [a]

[a] Western Michigan University Homer Stryker M.D. School of Medicine, 1000 Oakland Dr, Kalamazoo, MI, 49008, USA
[b] West Michigan Cancer Center, 200 N Park St, Kalamazoo, MI, 49007, USA

ARTICLE INFO

Article history:
Received 16 July 2020
Received in revised form
7 December 2020
Accepted 7 December 2020

Keywords:
General surgery education
Evaluation
Gender disparity
Self-assessment
ACGME core Competency
Milestones

ABSTRACT

Background: Studies of gender disparity in surgical training have yielded conflicting results. We hypothesize that there is no influence of gender on resident self-evaluation Milestone (SEM) scores and those assigned by the Clinical Competency Committee (CCC).
Methods: 42 residents (25 male & 17 female) and faculty completed 300 Accreditation Council for Graduate Medical Education (ACGME) Milestone evaluations over a 4-year period. Two-way ANOVA, intraclass correlations coefficients, and general linear mixed models were used for analysis.
Results: CCC Milestone scores from 150 evaluations, 51 (34%) for female residents and 99 (66%) for male residents, were compared to corresponding SEM scores. There is a high interrater reliability (self vs. CCC). There was a significant increase in scores with advancing PGY levels ($p < 0.001$). No effect of gender on Milestones scores ($p > 0.05$) was noted.
Conclusions: We found no significant differences in Milestones scores between male and female residents as determined by the CCC. Both scores improved significantly as residents progressed in training.

Introduction

Gender imbalance in medical-surgical fields varies among specialties. 2018–2019 data from the AAMC shows that female residents represent 41.3% in general surgery and 83.5% in obstetrics and gynecology. Female residents are the least represented in orthopedic surgery, at only 15.4%.[1] Given the inequality of gender representation, institutions must increase awareness and implement strategies to mitigate risk of gender disparity when evaluating resident performance. Studies have presented conflicting conclusions regarding the effect of gender bias when assessing surgical residents. Watson et al. in a recent multi-institutional study found a slight, but statistically significant, discordance in how female and male residents scored themselves on the ACGME Milestones. They found that females scored themselves lower than male residents on corresponding evaluations from the CCC.[2] Conversely, a study by Meier found no significant score differences per PGY level between the raters (resident and CCC).[3]

Providing comprehensive assessment of general surgery residents spanning the entirety of their training is critical. In 2014, the ACGME implemented competency-based developmental outcomes (*i.e.* Milestones).[4] Residency programs determine semi-annual Milestones scores via their Clinical Competency Committees (CCC). A strong correlation between resident SEM and CCC evaluations of performance has been reported.[5] Our study aims to evaluate for any gender disparity in Milestones evaluations of trainees at a single surgical residency program.

Methods

In this IRB-approved retrospective study, we analyzed the self-assessment Milestones of 42 general surgery residents, in addition to the corresponding CCC Milestone evaluations completed by the faculty, over a period of four years (January 2015–December 2019). The six ACGME clinical competencies (*i.e.* patient care, medical knowledge, practice-based learning and improvement, systems-based practice, interpersonal and communication skills, and professionalism) are stratified over 16 Milestone practice domains. Numerical values progressing in 0.5-point increments from 1 to 4 are assigned to each level of developmental progression within the Milestones, with a value of 0 assigned to any critical deficiencies. Summative scores for all domains were calculated, with total scores ranging from 0 to 64. Additionally, individual

* Corresponding author.
E-mail address: saad.shebrain@med.wmich.edu (S. Shebrain).

https://doi.org/10.1016/j.amjsurg.2020.12.020
0002-9610/

scores for each competency were calculated. In an effort to attain less-biased, comprehensive evaluations of our residents, the CCC in our program is comprised of an average of eight faculty members (range 7–12), diverse in both gender and type and location of practice. In preparation for the CCC meeting, the program coordinator generates a master CCC report for each resident based on evaluations completed via New Innovations® over the previous 6 months. Each resident receives, on average, seven evaluations from male and female faculty (range 5–13 evaluations). This report has two components, 1) Milestones data for each resident, and 2) additional performance data and text comments (*e.g.* ABSITE and mock oral scores, intraoperative and oral presentation evaluations, and feedback from their peers and nursing staff). One week before the CCC meeting, the master report is sent via secured email to all members of the committee for review. During the CCC meeting, the residents' performance data is presented along with any reported concerns. Committee members then comment on each resident based on their individual review of the information. Following discussion, the committee arrives at a final consensus regarding Milestones scores.

Before meeting with the program director for their semiannual meetings following the CCC meeting, residents are required to complete a Milestones self-evaluation via New Innovations©. To avoid any potential bias when determining final Milestones scores, resident self-assessment scores are not distributed to the CCC. Instead, the program coordinator adds the resident self-assessment data before the PD meeting. Any significant discrepancy between the scores is discussed with the residents. After meeting with residents, the final scores determined by the CCC are submitted to the ACGME.

Statistical analysis

All evaluation forms were de-identified, and the Milestones scores were entered into an Excel spreadsheet database on a password-protected server. We created two groups of scores: 1) a summative score for all Milestone practice domains, calculated by adding all 16 Likert item scores and 2) scores for each of the six core competencies.

Frequency (percent) of evaluations was calculated based on resident gender at each PGY level. To assess disparity related to resident gender, the mean (standard deviation) differences were calculated between resident self-assessment milestone scores and corresponding CCC assessment milestone scores at different PGY levels. To assess both degree of correlation and agreement between raters (self and CCC) on milestones scores, an Intra-class correlation coefficient (ICC) was used. Two-way ANOVA was conducted to explore the effect of PGY level and gender on self-assessment and CCC assessment milestone scores. Additionally, due to the nature of our data, the correlation between subjects' scores, the presence of unequal numbers of repeated measurements, and unequal variances, we used a general linear mixed model to analyze the change of summative Milestones scores over time for those residents who had complete datasets for more than two time points (*e.g.* a resident at end of PGY3 will have 6 time points), and to assess the effect of gender on these scores. A significance level of 0.05 was considered. Statistical analyses were performed using SPSS 25 software (IBM, Armonk, NY).

Results

A total of 300 evaluations for 42 residents, 25 male (60%) and 17 female (40%), across all PGY levels, were analyzed. Of 150 SEM evaluations, 51 (34%) and 99 (66%) were completed by female and male residents, respectively, at different PGY levels. Total evaluations completed per PGY level (male/female) over 4-year period include: 43 PGY-1 (28/15), 34 PGY-2 (19/15), 26 PGY-3 (16/10), 24 PGY-4 (18/6), and 23 PGY-5 (18/5). A Chi-squared goodness-of-fit test was performed on frequencies of evaluation for male/female residents in each PGY level. There was no statistically significant difference between genders ($p = 0.379$). Of the 42 residents, 25 (18 male, and 7 female) had evaluations at least three time points during the study period, and 16 (11 mal, and 5 female) had evaluations at least 4 time points. Inter-rater reliability between raters (self and CCC) was calculated for summative milestone scores and domain scores. A high degree of reliability was found between SEM and CCC ratings. Overall, the average measure Intraclass Correlation Coefficient (ICC) was 0.946 with a 95% confidence interval from 0.925 to 0.961 ($F(149,149) = 18.371$, $p < 0.001$), Table 1. We compared the mean score differences between male and female residents by PGY level and ACGME Core competency as rated by residents and CCC (Table 2). The only statistically significant different score was observed in self-reported scores at PGY1.

A two-way fixed-effects ANOVA was performed to explore the effect of PGY level and gender on the means of CCC Milestones scores and whether interaction between these two factors exists. There was a statistically significant main effect for PGY level ($p < 0.001$) with a large effect size, and partial eta-squared value of 0.827. Neither the effect of gender nor interaction effect of PGY level and gender were statistically significant ($p = 0.294$, $p = 0.480$, respectively). We performed pairwise comparisons of gender at each PGY level, with a Bonferroni adjustment to control the inflation of the type I error rate. At all PGY levels, no statistically significant difference between male and female residents as CCC Milestones means are similar and $p > 0.05$. However, on SEM, only at PGY1 are self-milestones means different for female (20.57) and male (25.11). When considering assessment differences in the six ACGME core competencies by the CCC, the only statistically significant difference observed was in medical knowledge, MK ($p = 0.027$). However, on SEM, female residents scored themselves lower in medical knowledge, MK ($p = 0.032$), and patient care, PC ($p = 0.041$), compared to their male counterparts. These differences were statistically significant. Table 2. General linear mixed model analysis showed no gender effect on the means of summative Milestones scores over time for those residents who had complete datasets for more than two time points (Fig. 2).

Discussion

In this study, we evaluated gender-related differences in self-reported and CCC Milestones scores. Similar to previous studies, our data demonstrated a strong correlation between resident self-assessment and CCC assessment across PGY levels, and is in accordance with published self-assessment studies that demonstrate a tendency for female residents to rate themselves lower

Table 1
Intraclass correlation coefficients between raters (residents and CCC).

ACGME Core Competency	Resident-CCC Rater		
	ICC	95% CI[a]	p-value
-Patient Care (PC)	0.944	0.920, 0.960	<0.001
-Medical Knowledge (MK)	0.915	0.883, 0.939	<0.001
-System-Based Practice (SBP)	0.912	0.879, 0.936	<0.001
-Practice-Based Learning (PBL)	0.926	0.897, 0.946	<0.001
-Professionalism (PROF)	0.912	0.879, 0.936	<0.001
-Interpersonal & Communication Skills (ICS)	0.918	0.887, 0.941	<0.001
-Summative Milestone Score	0.946	0.925, 0.961	<0.001

[a] CI, Confidence Interval.

Table 2
SEM and Clinical Competency Committee Assessment (CCC) of Male and Female Surgical Residents per Postgraduate year (PGY), and ACGME Core Competency.

Milestones by:	SEM Milestones (SEM)			CCC Milestones (CCC)		
A. PGY	Male	Female	p-value	Male	Female	p-value
-PGY-1	25.10 (6.56)	20.57 (4.30)	**0.020**	26.14 (6.22)	23.17 (6.02)	0.138
-PGY-2	36.42 (5.85)	35.17 (5.30)	0.522	35.71 (5.60)	33.23 (6.48)	0.240
-PGY-3	46.19 (5.16)	45.45 (2.97)	0.686	44.19 (3.53)	45.05 (3.12)	0.533
-PGY-4	51.50 (4.50)	50.58 (2.80)	0.647	50.03 (4.07)	48.50 (2.35)	0.396
-PGY-5	58.72 (3.90)	58.30 (3.66)	0.829	58.05 (4.14)	59.20 (3.55)	0.581
B. ACGME-CC[a]						
-PC	7.80 (2.65)	6.86 (2.64)	**0.041**	7.46 (2.61)	6.67 (2.60)	0.081
-MK	5.10 (1.80)	4.44 (1.67)	**0.032**	5.00 (1.62)	4.36 (1.70)	**0.027**
-SBP	5.33 (1.74)	5.00 (1.71)	0.175	5.24 (1.65)	4.72 (1.62)	0.067
-PBL	7.56 (2.62)	6.71 (2.66)	0.062	7.60 (2.51)	6.90 (2.51)	0.101
-PROF	7.85 (2.61)	7.00 (2.55)	0.060	7.87 (2.42)	7.20 (2.40)	0.101
-ICS	8.00 (2.50)	7.20 (2.62)	0.82	7.90 (2.34)	7.12 (2.40)	0.061
-Summative Scores	41.6 (13.66)	37.2 (13.43)	0.053	41.00 (12.76)	37.00 (12.80)	0.066

PC, patient care.
MK, medical knowledge.
SBP, system-based practice.
PBL, practice-based learning.
PROF, professionalism.
ICS, Interpersonal & Communication Skills.
[a] CC, Core Competency.

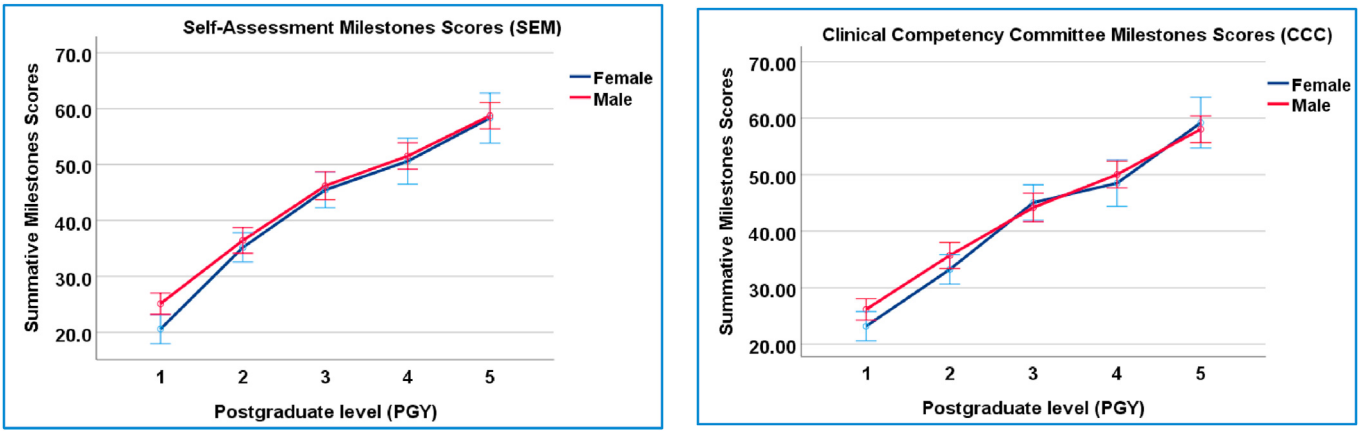

Fig. 1. A: Plot illustrating two simple main effects (gender and PGY level) on Milestones scores as completed by Residents (SEM).
Fig. 1B: Plot illustrating two simple main effects (gender and PGY level) on Milestones scores as assigned by Clinical Competency Committee (CCC).

than male residents at the same level of training or academic achievement.[6] However, in our study, the only statistically significant differences were observed at PGY1, where female residents scored themselves lower than male residents in medical knowledge (MK) and patient care (PC), and in PGY-1 MK when evaluated by CCC. Lyle et al.[5] reported female residents scored themselves lower than male residents in MK. In a large national survey of US general surgery residents, Yeo et al. found that 68.3% of PGY1 residents reported worrying about hurting patients.[7] Bucholz et al. reported that several factors were found to influence the residents' confidence level including gender, social factors, upbringing, and culture. They also found that female residents were almost twice as likely as male residents to worry about their ability to perform procedures by the end of training.[8] In a recent systematic review of fifteen survey studies about general surgery resident confidence, Elfenbien et al.[9] found conflicting data about the definition of confidence. He also pointed out that, in social cognitive theory, the term "self-efficacy" most closely describes the idea surgeons seem to be attempting to capture when discussing confidence. "Self-efficacy" is situation-specific and measurable, whereas confidence is individually understood and interpreted. Self-efficacy is a social phenomenon shaped not only by the objective acquisition of skills and technical expertise, but also by the absorption of the attitudes and opinions of others.[9] Self-inefficacy felt by a resident and observed by a faculty surgeon can be used as motivation for self-improvement and self-reflection, which are necessary traits in continuing professional development. Such reflections can be achieved by providing residents with appropriate mentorship and guidance as soon as they start in training in addition to a providing an environment that encourages close, collegial, and respectful relationships between residents and teaching surgical faculty. Such an environment facilitates resident/attending interactions and improves communication surrounding patient management.[8] Resident participation in the informed-decision-making process of patient care, especially early in training, can increase the residents' sense of involvement and importance, leading to more independence of thought as they progress in training.[8] This participation will require residents to actively invest in and enhance their medical knowledge along with skills of patients care, including procedural skills, as outlined in the ACGME milestones. A

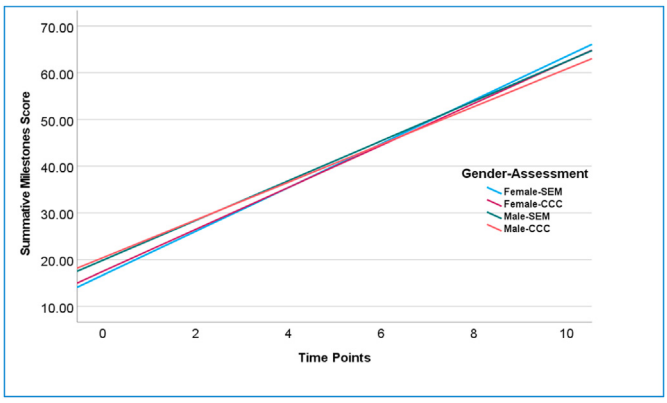

Fig. 2. Plot demonstrating changes in Milestones (SELF vs CCC) scores over multiple time points.

survey study by Binenbaum et al.[10] identified certain highly rated factors by residents that foster the development of confidence, and competency including general learning of medical or surgical knowledge, refining history and physical exam skills, and developing competency in advanced procedural skills and reading imaging studies, as well as exposure to many patients and clinical situations. By implementing preemptive and prophylactic pre-residency measures, such as an integrated cognitive and proficiency-based skills curriculum, an improvement in confidence with enhanced baseline abilities for internship can be achieved in abbreviated time.[11]

We found no differences in the summative scores among resident gender when evaluated by CCC. A possible explanation for this is that in small-to-medium sized programs, such as ours, residents work closely with attendings in a more direct-observational mode. This close relationship increases residents' opportunity to receive frequent feedback from faculty and facilitates mentorship. The diversity of the CCC helps the committee to assign unbiased final Milestones scores. Engaging faculty in discussion about gender bias in resident evaluation, and ensuring each resident receives adequate numbers of evaluations from a diverse group of faculty, helps provide residents with relatively accurate milestones scores. Additionally, in small programs, struggling and low-performing residents can be identified early in training so that there may be early intervention by the program director.

Our study has several limitations. First, it is retrospective study, with all inherent limitations of this kind of research. Second, it was conducted at a single, small-to-moderate sized program, which restricts sample size, and therefore may not be accurately extrapolated to larger institutions where there are fewer female faculty and contact between residents and attending surgeons is less frequent. Third, the database currently encompasses limited recurrences of Milestones evaluations making it difficult to elicit more granular data. We must also keep in mind that Milestones are not the only way that residents are evaluated, and although gender bias may not be detectable in our milestones evaluations, it may be present in other evaluation tools outside of the scope of this study. The results should therefore be interpreted with caution and with understanding that additional, multi-institutional research of gender impact on milestone evaluation and self-assessment in resident education is needed to confirm the validity of these findings. In spite of these limitations, we believe that providing residents with appropriate coaching, mentorship, and informal and formal feedback early in training improves their management skills and knowledge, thereby helping to close the gender gap in self-assessment evaluation.

Conclusion

We found no significant differences in Milestones scores between male and female residents as determined by the CCC at all PGY levels. However, on self-assessed Milestones scores, there was a difference at PGY1. Both scores improved significantly as residents progressed to the next level. The effect of gender perception early in training is an opportunity for programs to create a common ground in understanding the concept of consistent Milestones evaluation, and the need for early mentoring and coaching of new trainees. We believe support of this kind is a critical component of addressing gender disparity in surgical residency training.

References

1. https://www.aamc.org/data-reports/students-residents/interactive-data/report-residents/2019/table-b3-number-active-residents-type-medical-school-gme-specialty-and-sex.
2. Watson RS, Borgert AJ, O Heron CT, et al. A multicenter prospective comparison of the accreditation Council for graduate medical education milestones: clinical competency committee vs. Resident SEM. *J Surg Educ*. 2017;74(6):e8–e14. https://doi.org/10.1016/j.jsurg.2017.06.009.
3. Meier AH, Gruessner A, Cooney RN. Using the ACGME milestones for resident self-evaluation and faculty engagement. *J Surg Educ*. 2016;73(6):e150–e157. https://doi.org/10.1016/j.jsurg.2016.09.001.
4. Milestones: Milestones. Program and institutional accreditation: next accreditation system 2015. http://www.acgme.org/acgmeweb/tabid/430/ProgramandInstitutionalAccreditation/NextAccreditationSystem/Milestones.aspx.
5. Lyle B, Borgert AJ, Kallies KJ, Jarman BT. Do attending surgeons and residents see eye to eye? An evaluation of the accreditation Council for graduate medical education milestones in general surgery residency. *J Surg Educ*. 2016;73(6):e54–e58. https://doi.org/10.1016/j.jsurg.2016.07.004.
6. Minter RM, Gruppen LD, Napolitano KS, Gauger PG. Gender differences in the SEMSE-MSof surgical residents. *Am J Surg*. 2005;189(6):647–650. https://doi.org/10.1016/j.amjsurg.2004.11.035.
7. Yeo H, Viola K, Berg D, et al. Attitudes, training experiences, and professional expectations of US general surgery residents: a national survey [published correction appears in JAMA. *J Am Med Assoc*. 2009;302(12):1301–1308.
8. Bucholz EM, Sue GR, Yeo H, Roman SA, Bell Jr RH, Sosa JA. Our trainees' confidence: results from a national survey of 4136 US general surgery residents. *Arch Surg*. 2011;146(8):907–914. https://doi.org/10.1001/archsurg.2011.178.
9. Elfenbein DM. Confidence crisis among general surgery residents: a systematic review and qualitative discourse analysis. *JAMA Surg*. 2016;151(12):1166–1175. https://doi.org/10.1001/jamasurg.2016.2792.
10. Binenbaum G, Musick DW, Ross HM. The development of physician confidence during surgical and medical internship. *Am J Surg*. 2007;193(1):79–85. https://doi.org/10.1016/j.amjsurg.2006.07.009.
11. Naylor RA, Hollett LA, Castellvi A, Valentine RJ, Scott DJ. Preparing medical students to enter surgery residencies. *Am J Surg*. 2010;199(1):105–109. https://doi.org/10.1016/j.amjsurg.2009.09.003.