AEM Education and Training

A GLOBAL JOURNAL OF EMERGENCY CARE

Elevating the human condition during times of emergency

ORIGINAL CONTRIBUTIONS

Application of Frequent, Spaced Multiple-choice Questions as an Educational Tool in the Pediatric Emergency Department
Matthew J. Rustici, Vincent J. Wang, Kate E. Dorney et al. 85

Evaluation of Gender Differences in Ultrasound Milestone Evaluations During Emergency Medicine Residency Training: A Multicenter Study
Josie Acuña, Uwe Stolz, Lori A. Stolz et al. 94

The Implementation of a National Multifaceted Emergency Medicine Resident Wellness Curriculum Is Not Associated With Changes in Burnout
Kelly Williamson, Patrick M. Lank, Nicholas Hartman et al. 103

The Variable Journey in Learning to Interpret Pediatric Point-of-care Ultrasound Images: A Multicenter Prospective Cohort Study
Charisse Kwan, Martin Pusic, Martin Pecaric et al. 111

The Variability of Preferred Infant Lumbar Puncture Insertion Site Between Novice and Experienced Physicians
Jeffrey T. Neal, Jason A. Levy, Rachel G. Rempell et al. 123

Core Content for Pediatric Emergency Medicine Ultrasound Fellowship Training: A Modified Delphi Consensus Study
Erika Constantine, Maria Levine, Alyssa Abo et al. 130

Contents continued inside
Emergency Department Thoracotomy: Development of a Reliable, Validated Checklist for Procedural Training
Hashim Q. Zaidi, Sarah S. Dhake, Danielle T. Miller et al. 139

An Event-based Approach to Measurement: Facilitating Observational Measurement in Highly Variable Clinical Settings
Rosemarie Fernandez, Elizabeth D. Rosenman, Sarah Brolliar et al. 147

BRIEF CONTRIBUTION
The Design and Implementation of a Professional Development Program for Physician Assistants in an Academic Emergency Department
Derek L. Monette, Brian Baccari, Laura Paskind et al. 154

NEW IDEAS IN B-E-D-SIDE TEACHING
Escape the Trauma Room
Mikhail Podlog, Abbas Husain, Josh Greenstein et al. 158

INNOVATIONS REPORT
Engaging Emergency Medicine Influencers in Sex- and Gender-based Medicine: Lessons Learned from the Sex and Gender Interest Group in Emergency Medicine and the SAEM Jeopardy Game
Jeannette Wolfe, Basmah Safdar, Kinjal N. Sethuraman et al. 161

COMMENTARY—INVITED
Gender Influence on Ultrasound Milestone Evaluations Among Emergency Medicine Residents
Resa E. Lewiss 166

COMMENTARY AND PERSPECTIVE
The Irresponsible Use of Social Media Among Medical Students
Mark Curato, Abbas Husain, Kaushal Shah et al. 168

The Responsible Use of Social Media Among Medical Students and Residents
Stephen C. Dorner, Morgan B. Swanson, Shana M. Zucker 171
Application of Frequent, Spaced Multiple-choice Questions as an Educational Tool in the Pediatric Emergency Department

Matthew J. Rustici, MD1, Vincent J. Wang, MD2, Kate E. Dorney, MD3, Joshua Nagler, MD3, P. Jamil Madati, MD4, Patricia Ziegler1, and Genie Roosevelt, MD1

ABSTRACT

Objectives: The objective was to assess the feasibility of using spaced multiple-choice questions (MCQs) to teach residents during their pediatric emergency department (PED) rotation and determine whether this teaching improves knowledge retention about pediatric rashes.

Methods: Residents rotating in the PED from four sites were randomized to four groups: pretest and intervention, pretest and no intervention, no pretest and intervention, and no pretest and no intervention. Residents in intervention groups were automatically e-mailed quizlets with two MCQs every other day over 4 weeks (20 questions total) via an automated e-mail service with answers e-mailed 2 days later. Retention of knowledge was assessed 70 days after enrollment with a posttest of 20 unique, content-matched questions.

Results: Between August 2015 and November 2016, a total 234 residents were enrolled. The completion rate of individual quizlets ranged from 93% on the first and 76% on the 10th quizlet. Sixty-six residents (55%) completed all 10 quizlets. One-hundred seventy-three residents (74%) completed the posttest. There was no difference in posttest scores between residents who received a pretest (61.0% ± 14.5%) and those who did not (64.6% ± 14.0%; mean difference = –3.7, 95% confidence interval [CI] = –8.0 to 0.6) nor between residents who received the intervention (64.5% ± 13.3%) and those who did not receive the intervention (61.2% ± 15.2%; mean difference = 3.2, 95% CI = –1.1 to 7.5). For those who received a pretest, scores improved from the pretest to the posttest (46.4% vs. 60.1%, respectively; 95% CI = 9.7 to 19.5).

Conclusion: Providing spaced MCQs every other day to residents rotating through the PED is a feasible teaching tool with a high participation rate. There was no difference in posttest scores regardless of pretest or intervention. Repeated exposure to the same MCQs and an increase in the number of questions sent to residents may increase the impact of this educational strategy.

Teaching postgraduate resident physicians (residents) in the pediatric emergency department (PED) is challenging due to variability in training program, training level, shift timing, and seasonality of pediatric patient illness. Some learners may not learn key topics due to a lack of exposure.1 Learners completing rotations in pediatric emergency medicine (PEM) work varied shift times and are often not

From the 1Department of Pediatrics, Denver Health Medical Center, Denver, CO; the 2Department of Emergency Medicine, UT Southwestern Medical Center, Dallas, TX; the 3Department of Emergency Medicine, Boston Children’s Hospital, Boston, MA; and the 4Department of Emergency Medicine, Children’s National Medical Center, Washington, DC.

Received March 15, 2019; revision received May 20, 2019; accepted May 29, 2019.
Supported by a University of Colorado Pediatrics Department Educational Grant.
VW is the CEO of PEMQBOOK LLC, which produces a product relevant to the subject material and is the lead editor of both texts. The other authors have no potential conflicts to disclose.
Author contributions: study concept and design—MR, VW, PZ, GR; acquisition of the data—MR, VW, KD, JN, PJM, PZ, GR; analysis and interpretation of the data—MR, VW, GR; drafting of the manuscript—MR, GR; critical revision of the manuscript for important intellectual content—MR, VW, KD, JN, PJM, GR; statistical expertise—GR; and acquisition of funding—MR, GR.
Supervising Editor: Daniel P. Runde, MD.
Address for correspondence and reprints: Matthew J. Rustici, MD; e-mail: matthew.rustici@ucdenver.edu.
AEM EDUCATION AND TRAINING 2020;4:85–93
available to participate in group learning activities. Learner preferences of millennials and digital natives are also shifting toward easy-to-use technology and flexible learning environments.\(^2\)

Educators in the PED have started utilizing asynchronous e-learning strategies using Web modules and more recently podcasts.\(^3\)–\(^7\) These delivery modalities are frequently cited by learners as preferred but often show poor participation rates when integrated into the learner curriculum.\(^5,8\)

Retrieval practice is an effective learning strategy in which learners recall information they have previously seen using concept maps, flashcards, free-recall, or multiple-choice questions (MCQs).\(^9\)–\(^13\) MCQs are commonly used for assessment purposes to determine how much knowledge each learner was able to absorb during a teaching encounter; however, using MCQs as practice items, has been shown to improve retention of knowledge markedly more than passive educational activities (e.g. lectures, reading).\(^14\)–\(^17\) Utilizing MCQs as an effective and low-cost teaching tool for resident physicians completing a rotation in PEM has been incompletely described.\(^18\) Although learning science has informed educators about best learning practices, application of these principles remains challenging and there is potential to use technology to create innovative and low-cost methods to deliver this curriculum.

We hypothesized that sending easily accessible MCQs at frequent, regular intervals during a PEM rotation to residents would be an educational intervention with high participation rates and that completion of these MCQs would improve knowledge on the diagnosis and treatment of pediatric rashes compared to residents who experience standard clinical learning as measured by a posttest 2 months after the completion of the rotation.

**METHODS**

**Study Design**

This was a prospective, randomized controlled, multi-center educational trial using Solomon four-group design to control for pretest sensitization.\(^19\) Institutional review board approvals were obtained at each of the four study sites.

**Study Setting and Population**

This study was performed at four PEM sites (Boston Children’s Hospital, Children’s Hospital Los Angeles, Children’s National Medical Center, Denver Health Medical Center) across the United States. Inclusion criteria for study participants included residents from emergency medicine, pediatrics, combined internal medicine–pediatrics, and family medicine rotating for at least 4 consecutive weeks in a PED. Exclusion criterion was limited to residents who had previously enrolled in the study.

**Study Protocol**

The educational intervention was designed using a six-step process from Kern.\(^20\) A needs assessment was conducted of Denver-based emergency medicine, pediatrics, and family medicine residents, who ranked the usefulness of common teaching topics within PEM. These data suggested that teaching about pediatric rash diagnosis and treatment would be perceived as useful to residents across training years and programs due to low baseline comfort with rashes and the rarity of some rashes in the clinical environment. The goal of this intervention was to determine if residents would consistently engage in the learning intervention and also to determine if this would lead to an increase in resident knowledge of pediatric rash diagnosis and treatment. Principles utilized in developing the intervention were low cost, easy accessibility for participants, and automation of the intervention.

The administration of a pretest can have a significant positive effect on learning.\(^20\) To control for this effect, a Solomon four-group design was used in which participants in both the control group and the intervention group are randomized to receive or not receive a pretest. Stratified block randomization at each site assigned 15 enrollment packets to each of the following four groups: 1) pretest and intervention, 2) pretest and no intervention, 3) no pretest and intervention, and 4) no pretest and no intervention (Figure 1). Study packets with randomization assignments were constructed by the research coordinator at the primary site and mailed to each of the other three sites. Site directors and enrollees were blinded to the randomization, and faculty at each site were not informed about the group of any individual resident.

Potential research participants were approached by a faculty representative at each site within the first 3 days of their rotation. After signing a consent form, the residents opened a sealed research packet that contained an enrollment survey and a pretest if randomized to a pretest group. Residents randomized to the pretest groups completed the 10-question MCQ pretest on paper. The enrollment survey included questions on demographic information, training program and
training year. Participants were asked to rank their comfort level diagnosing and treating pediatric rashes on a 5-point Likert scale (1 = very uncomfortable and 5 = very comfortable) and how many weeks they had previously completed on a pediatric dermatology rotation. Each site director e-mailed the research coordinator the research packet number, resident name, and resident e-mail address within 24 hours of enrollment. Within 24 hours of receiving this information, the research coordinator e-mailed enrollees informing them if they were randomized to the intervention group. The completed consent forms, enrollment surveys, and pretests were batched and mailed to the research coordinator at regular intervals. Residents who chose to enroll were given a study incentive of a printed study version of the 2013 Pediatric Emergency Medicine Question Book (PEMQBook 2013) in which all rash-related questions were removed (78 total questions). These excluded rash questions were offered to residents at the completion of the study. Residents were told prior to enrollment they would need to return the PEMQBook (participation incentive) to their site director if they did not complete the posttest.

Residents who were randomized to the intervention groups were e-mailed a quizlet of two MCQs every 2 days (20 questions total) starting on the day of enrollment and then were sent a 20-question posttest 70 days after enrollment. Residents randomized to the “no intervention” group were sent an e-mail confirming enrollment and then did not receive additional contact until the posttest that was sent 70 days from the time of enrollment.

The 10-question pretest and 20-question posttest were created using questions included in the PEMQBook 2013. These pretest and posttest questions were selected via an iterative process matching content, question quality, difficulty index, and rash category type. Pretest and posttest questions had similar average difficulty indexes (0.796 and 0.818, respectively) based on preexisting data from administration to PEM fellows and attendings who were using PEMQBook 2013 online to study for PEM board certification. The 20 unique MCQs delivered in the teaching intervention consisted of questions taken from a prior edition of this text: 2009 Pediatric Emergency Question Review Book (PEMQBook 2009). All intervention questions and the posttest were administered using the SurveyMonkey survey tool. Residents who had not completed the posttest were sent up to three reminders automatically. All e-mails and e-mail reminders were delivered using MailChimp (Rocket Science Group, LLC) an automated e-mail delivery tool. E-mail campaigns were created at the onset of enrollment and functioned autonomously after participant e-mails were added to the campaign. No resident contacted the study coordinator for assistance accessing the intervention quizlets.

Data Analysis
Pretest and posttest scores were calculated by dividing the number of questions answered correctly by the number of questions on the test (10 for pretest and 20 for posttest). Mean pretest and posttest scores were presented with standard deviations (SD) as their distributions were found to be normal. Posttest scores were compared between those who received a pretest and those who did not to assess pretest sensitization. Mean difference with 95% confidence intervals (CIs) were
given for two group comparisons. Comfort-level diagnosing and treating pediatric rashes was dichotomized for analysis. Chi-square and Fisher’s exact tests were used to analyze categorical variables. The Student’s t-test or ANOVA was used to analyze continuous variables. The paired t-test was used to compare the change from pretest to posttest for participants who received a pretest. ANCOVA was used to analyze the relationship between posttest scores and receipt of the intervention in order to control for potential confounders. Kuder-Richardson 20 (KR20) was calculated for both the pretest and the posttest to measure the internal consistency of resident responses across questions. The number of residents enrolled was based on a prior study that calculated a sample size of 64 participants in two groups to detect a 10-percentage-point difference in posttest scores (i.e., two test questions out of 20) with 80% power and an alpha level of 0.05. This sample size was doubled to 240 to ensure sufficient numbers for cross-group analysis with the Solomon four-group design. Analysis was performed based on randomization group regardless of participation rate in the intervention group. A p-value of ≤0.05 was considered statistically significant. Statistical analyses were performed with SPSS version 22. This project was supported by a University of Colorado Department of Pediatrics Education Grant.

RESULTS

Between August 2015 and November 2016, a total of 239 residents were approached and 234 residents were enrolled in the study (Figure 1). Five residents declined enrollment due to concern that the study would be too time-consuming. There was no difference in group enrollment between the four sites (p = 0.99). There were significant differences among sites in sex distribution, training program, and training year of the residents (Table 1). Prior experience on a pediatric dermatology rotation differed between the sites but not comfort level in diagnosing pediatric rashes (Table 2).

Table 1
Demographics of Study Participants by Study Site

<table>
<thead>
<tr>
<th></th>
<th>Boston Children’s Hospital (n = 60)</th>
<th>Children’s Hospital Los Angeles (n = 60)</th>
<th>Children’s National Medical Center (n = 54)</th>
<th>Denver Health Medical Center (n = 60)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>30.5 (±3.3)</td>
<td>28.9 (±3.1)</td>
<td>30.1 (±3.9)</td>
<td>29.1 (±2.8)</td>
<td>0.04</td>
</tr>
<tr>
<td>Male sex</td>
<td>38 (64)</td>
<td>20 (35)</td>
<td>28 (56)</td>
<td>17 (29)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Training program†</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emergency medicine</td>
<td>53 (90)</td>
<td>5 (8)</td>
<td>27 (54)</td>
<td>8 (14)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Family medicine</td>
<td>3 (5)</td>
<td>6 (10)</td>
<td>10 (20)</td>
<td>16 (28)</td>
<td></td>
</tr>
<tr>
<td>Pediatrics</td>
<td>3 (5)</td>
<td>47 (80)</td>
<td>9 (18)</td>
<td>34 (59)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>1 (2)</td>
<td></td>
<td>4 (8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training year†</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PGY-1</td>
<td>21 (36)</td>
<td>18 (31)</td>
<td>17 (36)</td>
<td>31 (54)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>PGY-2</td>
<td>7 (12)</td>
<td>25 (43)</td>
<td>14 (30)</td>
<td>2 (3)</td>
<td></td>
</tr>
<tr>
<td>PGY-3 or -4</td>
<td>31 (52)</td>
<td>15 (26)</td>
<td>16 (34)</td>
<td>25 (43)</td>
<td></td>
</tr>
</tbody>
</table>

Data are reported as mean (±SD) or n (%).
†Not all participants provided their type of training program or training year.

Table 2
Pediatric Dermatology Comfort and Exposure of Study Participants by Study Site

<table>
<thead>
<tr>
<th></th>
<th>Boston Children’s Hospital (n = 60)</th>
<th>Children’s Hospital Los Angeles (n = 60)</th>
<th>Children’s National Medical Center (n = 54)</th>
<th>Denver Health Medical Center (n = 60)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very comfortable or comfortable diagnosing and treating pediatric rashes†‡</td>
<td>7 (12)</td>
<td>7 (12)</td>
<td>5 (10)</td>
<td>12 (21)</td>
<td>0.35</td>
</tr>
<tr>
<td>Prior pediatric dermatology rotation†</td>
<td>13 (22)</td>
<td>24 (41)</td>
<td>7 (14)</td>
<td>19 (33)</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Data are reported as n (%).
†1 = very uncomfortable, 2 = comfortable, 3 = neutral, 4 = comfortable, 5 = very comfortable.
‡Not all participants responded about comfort level with rashes and prior pediatric dermatology rotation.
In the intervention group, the completion rate of individual quizlets ranged from 93% on the first quizlet to 76% on the 10th quizlet (Figure 2). Resident completion of the intervention was high with 66 (55%) completing all 10 quizlets (Figure 3). Only four participants completed none. One-hundred seventy-three residents (74%) completed the posttest 2 months after their PED rotation.

Mean pretest and posttest scores are presented in Table 3. There was no difference between the four randomized groups on posttest scores. There was no difference in posttest scores between the group that completed a pretest (61.0 ± 14.5) and those who did not (64.6 ± 14.0; mean difference = −3.7, 95% CI = −8.0 to 0.6). Since there was no evidence of pretest sensitization, posttest scores were compared between the residents in the group that received an intervention and those who did not regardless of the pretest. There was no difference in posttest scores between the group that received the intervention (64.5 ± 13.3) and those who did not receive the intervention (61.2 ± 15.2; mean difference = 3.2, 95% CI = −1.1

**Quizlet Completion Percentages**

![Quizlet Completion Percentages](image1)

**Figure 2.** Quizlet completion percentages.

**Intervention Quizlets Completed by Residents**

![Intervention Quizlets Completed by Residents](image2)

**Figure 3.** Number of residents completing quizlets.
to 7.5). The mean difference between posttest (60.1) and pretest scores (46.4) for those who received a pretest was 14.6 (95% CI 9.7 to 19.5). The percentage of individual questions answered correctly ranged from 30% to 94%.

There was no difference in posttest scores by year of training (PGY-1 = 64.2 ± 14.7, PGY-2 = 61.1 ± 16.1, PGY-3 or 4 = 62.6 ± 13.6; p = 0.60) nor by specialty (emergency medicine = 60.3 ± 15.0, family medicine = 66.1 ± 14.5, pediatrics = 63.8 ± 13.7; p = 0.28). Given the differences between sites, site was considered a potential confounder for posttest scores. However, after controlling for site, there was no difference in posttest scores between residents in the intervention group compared to those not exposed to the intervention (p = 0.14).

The 20-question posttest had moderate internal consistency (KR20 = 0.57). The 10-question pretest was less internally consistent (KR20 = 0.38). Item-total correlation of the posttest showed no substantial improvements to internal consistency with the removal of any question.

**DISCUSSION**

We demonstrated that automated, asynchronous learning via regularly spaced “teaching MCQs” is a feasible teaching tool that could be utilized by PED educational directors. Residents enrolled in this study had high rates of completion for both the teaching intervention questions and the posttest, suggesting residents were engaged and willing to complete these electronic tasks. In addition, very few residents refused participation (2%).

Given the average residents’ busy schedules and requirements, finding a teaching modality that is both simple and convenient for the residents to use is paramount to the success of the teaching intervention. Our study’s intervention of regularly scheduled e-mailed MCQs and detailed answers were both easy and convenient for the residents to complete. Time spent communicating with the participants after enrollment was minimal and managed entirely through an automated e-mail campaign. The teaching questions were delivered to each resident’s e-mail, and residents were able to directly click a link in the e-mail to answer the questions via computer or mobile device. No residents contacted the study coordinator for assistance accessing the intervention quizlets suggesting that this teaching modality was both convenient and simple to use by the resident. In prior studies, feedback has been noted as a key component of teaching millennial learners and encouraging intrinsic motivation.2,23 Residents were informed that they would only be able to see explanations to the questions after they completed them, and this may have been motivating for residents to complete the quizlets.

Our rates of posttest completion were markedly higher than prior similar studies (44% dropout rate in Chang et al.5 and 48% completion of the posttest in House et al.18). The posttest in this study was delivered electronically 70 days after enrollment when residents were no longer in the PEM department. Three reminders were sent to participants who had not yet completed the posttest and these multiple reminders likely improved completion rates. Residents may have perceived the posttest as a learning opportunity similar to the teaching questions, although there was no difference in posttest completion rates between those who received the intervention and those who did not. Studies in behavioral economics have noted that people are highly motivated to avoid losing something they have been given (loss aversion).24 We chose to give participants the study incentive at the time of enrollment and the book needed to be returned if participants failed to complete the posttest. This may have motivated residents to complete the posttest rather than lose the book although no site study investigator was able to collect a book from those who did not complete the posttest.

The 20-question posttest had moderate internal consistency (KR20 = 0.57). The 10-question pretest was less internally consistent (KR20 = 0.38). Item-total correlation of the posttest showed no substantial improvements to internal consistency with the removal of any question.

### Table 3
Mean Pretest and Posttest Scores

<table>
<thead>
<tr>
<th>Score</th>
<th>Pretest and Intervention (n = 59)</th>
<th>Pretest and No Intervention (n = 58)</th>
<th>No Pretest and Intervention (n = 40)</th>
<th>No Pretest and No Intervention (n = 45)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest†</td>
<td>44.2 (±17.4)</td>
<td>46.0 (±17.5)</td>
<td>NA</td>
<td>NA</td>
<td>0.58</td>
</tr>
<tr>
<td>Posttest‡</td>
<td>62.0 (±11.9)</td>
<td>60.0 (±16.7)</td>
<td>67.0 (±14.2)</td>
<td>62.6 (±13.6)</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Data are reported as mean (±SD)

†Pretest with 10 questions.
‡Posttest with 20 questions.
Despite the high participation rate in completing the teaching intervention questions, there was no difference in posttest performance between the intervention and the control groups. Other studies using spaced repetition models where MCQs are delivered at regular intervals have seen moderate and persistent improvements in knowledge.\textsuperscript{25–27} Kerfoot et al.\textsuperscript{26} found gains in medical knowledge using a similar model but used the same questions on the pretest, intervention, and posttest and also repeated questions using a spaced repetition algorithm. Our pretest, intervention, and posttest were composed of unique questions to ensure that we were teaching and measuring content knowledge and to avoid improved test performance based on recollection of the questions. It is also possible that repetition of the intervention items is necessary for a larger effect size. In addition, many of the prior efforts using this educational strategy have spaced questions over longer periods of time,\textsuperscript{25,26} with only one prior study using spacing over a 1-month period.\textsuperscript{18} We chose this shorter duration intervention period to match the clinical experience on a given rotation for residents. Although there was no difference in posttests between the intervention and control groups, we did see a significant improvement in performance from the pretest to the posttest suggesting that learning did occur during the study period. This may have been related to clinical exposure or the Hawthorne effect. Participants may have focused their learning efforts on rashes during their PED rotation knowing they would be tested later.

In our study, there was moderate internal consistency of the posttest, which suggests that despite the variability in the types of rashes presented, participants had a moderately consistent performance across the questions. Questions used in the posttest were directly extracted from the PEM boards review book that had been heavily edited and questions psychometrically analyzed prior to being published in the PEMQBook. The mean scores seen with resident administration of the questions were lower than those seen with the difficulty indexes derived from PEM attendings and fellows. This may be due to differences in clinical exposure or baseline knowledge for those earlier in their training or the perceived stakes of completion of the questions. There was one question about the management of Becher’s disease that no resident was able to answer correctly. The difficulty of the questions for participants at the resident level may have reduced the sensitivity of the examination to detect knowledge gains. Despite other studies that have shown a sensitizing effect of a pretest, this was not seen in our cohort. This may be due to the relatively small number of pretest questions leading to an insignificant effect size. There was a 70-day gap between the pre- and posttest, which may have further diminished the effect size. There was no pretest effect in the study by Chang et al.\textsuperscript{5} that also focused on medical trainees, and it is possible that complex medical knowledge is less affected by pretests than more simplistic recall-based learning in other contexts.\textsuperscript{28}

This study was accomplished using simple and inexpensive online tools making it a feasible intervention to introduce into a PED or other clinical rotation. The MailChimp subscription was $120/year and Survey Monkey subscription was $250/year. The educational intervention could have been completed with the free version of Survey Monkey, but this would not allow for exporting of data with unique identifiers needed for research analysis. The intervention had a small administrative time demand with ~3 hours needed during the design phase, mostly related to copying questions into Survey Monkey and designing the MailChimp campaign and another 8 hours required of the study coordinator over the 18-month duration of the study. This time was primarily spent adding the 240 residents to the MailChimp campaign. Using this intervention outside of a research study would be even less time-intensive, and scaling this intervention to more residents would have a minimal impact in administrative time.

Educators working in EDs struggle to employ traditional teaching structures in which learners and teachers are present in the same location at the same time due to the variability in schedules of both faculty and learners. The majority of teaching that occurs on EM or PEM rotations is done in situ in the context of the patients that are being treated. Residents likely receive sufficient education using this model for common patient conditions; however, rare and variable illnesses like many pediatric rashes may need more deliberate and explicit teaching. This model of automatically delivered quizlets initiated at the start of a rotation proved to be an easy and low-cost intervention that had a very high rate of participation by residents. Although there may be limits to the number of quizlets or questions that residents will complete using this model, it could provide an effective asynchronous learning tool to address rare topic areas. This teaching modality can be used with residents of varied specialty and training year as well as residents working in geographically separate sites.
LIMITATIONS

Enrollment by site differed in sex, training program, and training year. The effect of the intervention may have been greater in residents from a particular specialty or training year, which may have limited our ability to see a difference when all site enrollees were combined.

Despite the fact that there was a relatively high level of overall participation, it is possible that there was some selection bias of more motivated residents completing more of the intervention. Given the size of the cohorts, this effect is not likely to be very large.

The generalized learning impact of this intervention was not assessed in this study and so it is unknown whether residents in the intervention group chose to spend more time completing the training questions and less time studying rashes on their own. No study participant was blinded and therefore the educational topic covered in the intervention may also have been known to those in the control group. Also, there may have been contamination between the intervention and no intervention group if residents in the intervention group shared their quizlet questions with residents in the no intervention group as they were simultaneously rotating in the PED. The attendings were unaware of the randomization of individual residents, and as almost all residents approached for study enrollment agreed to participate in the study, there is no reason to believe there was a systematic bias in clinical teaching that would disproportionately affect intervention or control residents.

Another limitation that may have affected the ability to detect knowledge gain differences is the relatively small number of intervention questions and also the relatively small number of questions on posttest. Other studies with similar numbers of enrollees were able to see differences in performance with a similarly sized posttest. However, 20 questions may be insufficient to adequately assess knowledge from a variety of different rash topic areas.

CONCLUSIONS

Teaching foundational knowledge to residents rotating in the pediatric ED is challenging due to variability in resident schedules and rarity of certain illnesses. In-person didactic teaching is also challenging as residents need to work varying shifts and are often unable to take time away from patients while on shift. Delivering short multiple-choice question quizlets every 2 days to residents via an automated e-mail service had high rates of resident participation although there was no difference in posttest scores regardless of pretest or intervention. The time and monetary costs of administration were very small, which makes this educational tool easily adoptable by other residency programs and could be extended to large numbers of residents from different programs with a minimal increase in time or costs. This intervention may be particularly helpful in the teaching of uncommon or seasonal diseases to ensure exposure to core content areas.

REFERENCES

Evaluation of Gender Differences in Ultrasound Milestone Evaluations During Emergency Medicine Residency Training: A Multicenter Study

Josie Acuña, MD\(^1\), Uwe Stolz, PhD\(^2\), Lori A. Stolz, MD\(^2\), Elaine H. Situ-LaCassee, MD\(^1\), Gregory Bell, MD\(^3\), Ross P. Berkeley, MD\(^4\), Jeremy S. Boyd, MD\(^5\), David Castle, DO\(^6\), Kristin Carmody, MD\(^7\), Tiffany Fong, MD\(^8\), Ekjot Grewal, MD\(^9\), Robert Jones, DO\(^10\), SueLin Hilberts, MD\(^11\), Carolyn Kanter, MD\(^12\), Kenneth Kelley, MD\(^13\), Stephen J. Leetch, MD\(^14\), Philip Pazderka, MD\(^15\), Erica Shaver, MD\(^16\), Jeffrey R. Stowell, MD\(^17\), Elaine B. Josephson, MD\(^18\), Daniel Theodoro, MD\(^11\), and Srikar Adhikari, MD\(^1\)

ABSTRACT

**Objectives:** Prior literature has demonstrated incongruities among faculty evaluation of male and female residents’ procedural competency during residency training. There are no known studies investigating gender differences in the assessment of procedural skills among emergency medicine (EM) residents, such as those required by ultrasound. The objective of this study was to determine if there are significant gender differences in ultrasound milestone evaluations during EM residency training.

**Methods:** We used a stratified, random cluster sample of Accreditation Council for Graduate Medical Education (ACGME) EM residency programs to conduct a longitudinal, retrospective cohort analysis of resident ultrasound milestone evaluation data. Milestone evaluation data were collected from a total of 16 ACGME-accredited EM residency programs representing a 4-year period. We stratified milestone data by resident gender, date of evaluation, resident postgraduate year, and cohort (residents with the same starting date).

From the \(^1\)Department of Emergency Medicine, The University of Arizona, Tucson, AZ; the \(^2\)Department of Emergency Medicine, The University of Cincinnati, Cincinnati, OH; the \(^3\)Department of Emergency Medicine, University of Iowa Hospitals and Clinics, Iowa City, IA; the \(^4\)Department of Emergency Medicine, University of Nevada Las Vegas School of Medicine, Las Vegas, NV; the \(^5\)Department of Emergency Medicine, Vanderbilt University Medical Center, Nashville, TN; the \(^6\)Department of Emergency Medicine, Sparrow Hospital, Lansing, MI; the \(^7\)Department of Emergency Medicine, New York University School of Medicine, New York, NY; \(^8\)Department of Emergency Medicine, Johns Hopkins University School of Medicine, Baltimore, MD; the \(^9\)Department of Emergency Medicine, Mount Sinai St. Luke’s Mount Sinai West Hospitals, New York, NY; the \(^10\)Department of Emergency Medicine, MetroHealth Medical Center, Cleveland, OH; the \(^11\)Department of Emergency Medicine, Washington University School of Medicine, St. Louis, MO; the \(^12\)Department of Emergency Medicine, Temple University Hospital, Philadelphia, PA; the \(^13\)Department of Emergency Medicine, University of California Davis, Sacramento, CA; the \(^14\)Department of Emergency Medicine, Orlando Health, Orlando, FL; the \(^15\)Department of Emergency Medicine, Western Michigan University Homer Stryker M.D. School of Medicine, Kalamazoo, MI; the \(^16\)Department of Emergency Medicine, West Virginia State University, Morgantown, WV; the \(^17\)Department of Emergency Medicine, Maricopa Integrated Health System, Phoenix, AZ; and the \(^18\)Department of Emergency Medicine, Lincoln Medical and Mental Health Center, Bronx, NY.

Received November 12, 2018; revision received June 9, 2019; accepted June 17, 2019.

Supported by the Department of Emergency Medicine, University of Arizona.

The authors have no potential conflicts of interest to disclose.

Author contributions: JA—study concept and design, acquisition of the data, analysis and interpretation of the data, drafting of the manuscript, and critical revision of the manuscript for important intellectual content; US—analysis and interpretation of the data, drafting of the manuscript, critical revision of the manuscript for important intellectual content, and statistical expertise; LAS—study concept and design, acquisition of the data, analysis and interpretation of the data, drafting of the manuscript, critical revision of the manuscript for important intellectual content, and statistical expertise; EHS—drafting of the manuscript and critical revision of the manuscript for important intellectual content; GB, RB, JB, DC, KC, TF, EG, RJ, SH, CR, KJ, SL, PP, ES, JS, EJ, and DT—acquisition of the data; SA—study concept and design, acquisition of the data, analysis and interpretation of the data, drafting of the manuscript, critical revision of the manuscript for important intellectual content, and statistical expertise.

A related article appears on page 166.

Supervising Editor: Sorabh Khandelwal, MD.

Address for correspondence and reprints: Josie Acuña, MD; e-mails: jacuna@aemrc.arizona.edu or jacuna86@gmail.com.

AEM EDUCATION AND TRAINING 2020;4:94–102
Results: A total of 2,554 ultrasound milestone evaluations were collected from 1,187 EM residents (750 men [62.8%] and 444 women [37.1%]) by 104 faculty members during the study period. There was no significant overall difference in mean milestone score between female and male residents [mean difference = 0.01 (95% confidence interval {CI} = −0.04 to 0.05)]. There were no significant differences between female and male residents’ mean milestone scores at the first (baseline) PGY1 evaluation [mean difference = −0.04 (95% CI = −0.09 to 0.003)] or at the final evaluation during PGY3 (mean difference = −0.02 [95% CI = −0.03 to 0.06]).

Conclusions: Despite prior studies suggesting gender bias in the evaluation of procedural competency during residency training, our study indicates that there were no significant gender-related differences in the ultrasound milestone evaluations among EM residents within training programs throughout the United States.

There are gender gaps throughout medicine, and the field of emergency medicine (EM) is no exception.1,2 Despite significant advancements over past decades, there are fewer women than men in EM residency programs and in positions of leadership in academic emergency departments.3 It is postulated that this difference in representation in EM is likely related to a broader imbalance in medicine, with the disparity starting early in medical education and residency training. Several studies have explored gender differences in resident training evaluations among various specialties, with findings that support bias against female residents.4,5

The Accreditation Council for Graduate Medical Education (ACGME)-accredited EM programs utilize educational milestones to evaluate residents and track progression toward unsupervised practice. The concept of milestones for residency training was introduced in 2008, developed to assist faculty in standardizing evaluations of resident physicians by identifying and tracking resident attainment of core knowledge and skills throughout their training.6 Residents are assessed throughout the year by faculty and milestones are recorded for each resident biannually for review by the program leadership. This directs their academic plan and methods for further education. Prior literature on gender bias has demonstrated that despite having similar skills and fund of knowledge at the beginning of residency, female EM residents attained lower-level milestones, most evident by the time of graduation when compared to their male counterparts.5

Bedside ultrasound is one of the required milestone achievements for EM residents and is recognized as one of the essential procedural skills for an emergency physician. Prior studies among other specialties have demonstrated significant incongruities among male and female trainees regarding the assessment of procedural skill competency during residency training. Studies have demonstrated disparities not only in the level of autonomy female residents were given while performing procedures when compared to male trainees but also in the overall number and type of procedures performed.7–9 Unfortunately, these previous studies are poorly generalizable to EM training as many were single-center studies among non-EM residents, often utilizing institution-specific evaluation scales. Therefore, it remains unclear after a review of the prior literature whether such a difference is present in EM training as well. We sought to study potential gender differences in evaluations of residents’ procedural skills by analyzing the ultrasound milestones data for EM residents throughout the United States. A pilot study performed at our institution suggested a gender bias in the evaluation of ultrasound milestone levels among EM residents, which is consistent with other gender bias studies.10 However, a more robust study is needed to assess whether this disparity exists among EM residency programs throughout the country. To our knowledge, there are no previous multicenter studies specifically investigating gender differences in the evaluation of psychomotor or procedural skills among EM residents, such as those required by ultrasound. The objective of this study was to investigate the possible existence of gender differences in ultrasound milestone evaluations during EM residency training across the country.

METHODS

Study Design
This was a longitudinal, retrospective analysis of ACGME resident milestone evaluation data from a representative sample of EM residency programs in the United States. The research data were collected from a total of 16 ACGME-accredited EM residency programs from Fall 2014 to Spring 2018. Both 3-year and 4-year residency programs were included in the analysis. An institutional review board (IRB) approval was obtained at the institution initiating the study. Additional IRB approval was obtained from the...
collaborating programs as deemed necessary by each institution.

**Study Population/Selection**

A list of all eligible programs (the sampling frame) was compiled from the ACGME list of accredited EM residency programs. Programs not accredited through the ACGME and osteopathic programs were excluded. Survey sampling was similar that in a previous study by Stolz et al. Programs were stratified by geographic region and the designation of geographic region was based on U.S. Census Bureau Geographic areas, with the four geographic regions being West, Midwest, Northeast, and South. Four programs were randomly selected from each stratum, with alternate programs identified to account for nonresponse of originally chosen programs. The study population includes EM residents trained at the randomly chosen residency programs over a 4-year period between Fall 2014 to Spring 2018. The milestones evaluation data on these residents were collected biannually as a part of their residency training and competency assessment. Evaluations were performed by faculty members at each program. The ultrasound subcompetency outlines five advancing proficiency levels ranging from level 1 through 5. EM residents are expected to meet at least a level 4 subcompetency requirement prior to graduation. Ultrasound competency is typically assessed through the number of ultrasound examinations performed that have passed quality assurance and direct observation of skills. All former or current residents who had milestone evaluations recorded over a 4-year period were included in the study. After randomization, programs were excluded if they elected to not be part of the study after an invitation to participate was extended. Programs were also excluded after randomization if they had not been in existence prior to Fall 2014 as these programs would not have the full 4 years of data required for participation in the study.

**Study Protocol**

All programs were assigned a random number using a random number generator and then sorted in descending order. The first four programs per stratum were designated as primary programs, and the next four, as backup programs in the event that any primary programs were not able to participate. Thus, for each stratum, four primary programs along with four backup programs were randomly chosen for participation. Each of these programs had an on-site faculty member who was chosen to be the site coordinator. This faculty member was asked to coordinate data collection and facilitate any IRB process deemed necessary by their institution. The site coordinator compiled ultrasound milestone evaluations recorded on the residents from their program from Fall 2014 to Spring 2018. The data came from the Accreditation Data System, which is entered semiannually. These data had been culminated by clinical competency committees at each institution per ACGME program requirements. All data were deidentified. The resident milestones were recorded along with the gender of each resident. Once the deidentified data were collected on the standardized data collection sheet, it was sent securely to the primary investigator. The site coordinator also completed and returned a short survey that reported the gender of the faculty evaluators, a description of the primary practice setting (academic, community, or county), the average number of ultrasound examinations completed by residents prior to graduation and the presence/absence of a required ultrasound rotation for residents. The presence or absence of a fellowship training program was determined from the Emergency Ultrasound Fellowships website. Data from all programs were deidentified and pooled to create a composite data set.

**Data Analysis**

Based on our previous survey of US ACGME EM residency programs, we determined that stratifying by geographic region and then sampling EM residencies (clusters) would result in a variance inflation factor (the increase in the variance relative to simple random sampling) of 2.9 based on a previous survey study. We a priori considered a 10% difference in means to be meaningful (e.g., for a mean score of 1.5 we wanted to be able to detect a significant increase or decrease of 0.15; for a mean of 2.00, a 0.20 change; etc.). We used a mean ultrasound milestone evaluation score from a convenience sample (single program, single evaluation period: mean ± SD = 2.48 ± 1.06) to estimate that in order to detect a 10% change/difference between two groups (mean group 1 = 2.25, mean group 2 = 2.5) with a 1:3 allocation ratio (women represent approximately 35% of EM residents nationwide), power = 80%, and p = 0.05, we would need 590 evaluation scores. Given the variance inflation factor of 2.9, we estimated that we would need 1,711 evaluation scores to have sufficient power to detect our a priori proposed difference. We estimated...
that each sampled program would provide 120 scores (representing several classes of residents across eight evaluation periods) and that to get at least 1,711 scores and sample an equal number of programs for each stratum, we would need to sample four programs per stratum for a total of 16 programs and 1,920 total evaluations to have sufficient power. Ultrasound milestone evaluation scores ranged from 1 to 5, and evaluators commonly used 0.5 increments.

We accounted for the survey design and sampling weights for all analyses using the survey functions (“svy”) in Stata (version 14.2; StataCorp). Sampling weights were the inverse probability of sampling each cluster within each stratum and we applied a finite population correction. We used generalized linear multilevel mixed-effects regression to estimate mean milestone evaluation data, with individual residents considered a random effect. Residents were nested within programs and programs were nested within strata. We constructed an overall regression model that included gender, evaluation date, the postgraduate year (PGY) for each milestone evaluation, and an interaction term between PGY at each evaluation and evaluation date. We also included several other potential confounding factors as judged relevant by the authors to examine confounding of the relationship between milestone scores and gender (presence of an ultrasound fellowship, size of the program, type of program [community vs. university-based], percentage of female EM faculty, self-reported [by program] number of ultrasound examinations by each resident prior to graduation), and 3-year versus 4-year program. The distribution of scores was approximately normal, and we found no transformation of the data that improved normality. Preliminary analyses indicated that the distribution of milestone scores and regression residuals met the assumption of normality; consequently, we used the continuous, untransformed milestone evaluation scores for analyses. We used the final regression models to estimate differences (i.e., marginal mean differences) between mean evaluation scores for male and female residents after stratifying by PGY at each evaluation and evaluation date. This generated mean differences for 32 male–female resident comparisons representing seven distinct resident cohorts. Regression models were also used to analyze differences in the increase of mean evaluation scores from PGY-1 to PGY-3 for all residents that had both PGY-1 and PGY-3 evaluations. We also analyzed differences in the increase from PGY-1 to PGY-2 and from PGY-2 to PGY-3 between male and female evaluation scores. Finally, we also examined mean differences between men and women’s scores for their first PGY-1 evaluation and their last PGY-3 evaluation.

RESULTS

Demographic Characteristics

A total of 2,554 milestone evaluations were collected from 1,186 EM residents (750 men [63.2%] and 436 women [36.8%]) by 104 faculty members during the study period. Our study included evaluations from 16 EM training programs (10 academic programs, four community programs, and two county programs). The training programs represent all four U.S. Census–designated regions of the United States (Northeast, Midwest, South, West) in a mix of rural, suburban, and urban settings. Training programs ranged in size from 29 to 62 residents and 18 to 133 physicians on faculty. Fifteen of the 16 programs required that residents rotate through an ultrasound rotation as part of their residency training. Nine of the 16 programs had an ultrasound fellowship program.

Table 1 shows the demographic and study characteristics for our sample as well as the national estimates for the number of residents and proportions and means. While the overall proportion of female EM residents in our sample was 36.8%, we estimated the proportion of female EM residents nationally to be 37.2% (95% confidence interval [CI] = 32.3 to 42.5) after applying survey sampling weights and taking the survey design into account.

Figure 1 shows the mean ultrasound milestone evaluation scores for the seven cohorts sampled in our survey, stratified by gender, date of evaluation, and PGY in residency. Table 2 shows the mean differences and associated 95% CIs for all male–female comparisons graphically shown in Figure 1. Only paired differences within a cohort between male and female scores for the same evaluation date were analyzed. Only one paired difference (Fall 2014/PGY-1 scores for the 2014/2015 cohort) was statistically significant with a mean difference (female–male residents) of −0.28 (95% CI = −0.51 to −0.06).

Our initial combined multilevel mixed effects model included resident gender as the primary independent variable along with evaluation date, PGY at evaluation date, and a PGY by evaluation date interaction term and showed no statistically significant difference between male and female residents’ scores. After
various factors were controlled for and after it was determined whether any of them were confounders for the relationship between gender and milestone scores, there were no differences between mean ultrasound evaluation scores between men and women. The mean difference for evaluation scores for female–male residents was 0.01 (95% CI = 0.04 to 0.05), evaluation date was controlled for, PGY at each evaluation, presence of ultrasound fellowship, size of program, university versus community program, the number of self-reported (by programs) ultrasound examinations by each resident upon graduation, 3-year versus 4-year programs, and the percentage of female EM faculty for each program (Table 3).

A comparison of the initial evaluation scores for PGY-1 residents showed that women’s ultrasound evaluation scores were on average 2% lower (mean difference female–male = −0.04 [95% CI = −0.09 to 0.003]) compared to men’s scores; however, this was not statistically significant based on our a priori threshold of a 10% difference. For all PGY-3 residents, women had a slightly higher final mean milestone evaluation score (mean difference female–male scores = 0.02 [95% CI = −0.03 to 0.06]), but again this difference (<1%) was not statistically significant.

We also compared the change in scores between male and female residents from initial PGY-1 score to the final PGY-3 score for the two cohorts (2014/2015 and 2015/2016) that had data for PGY-1 and -3 residents. The overall difference for increase of ultrasound milestone scores (mean increase females–males) from first PGY-1 score to final PGY-3 score was not statistically different. There was also no statistically significant difference in the mean increase in ultrasound milestone evaluation scores (from first PGY-1 score to last PGY-2 score) between men and women (mean increase female–male) for the 2016/2017 cohort (−0.14 [95% CI = −0.36 to 0.07]) or the change from first PGY-2 score to final PGY-3 score for cohort 2013/2014 (+0.19 [95% CI = −0.07 to 0.46]).

### DISCUSSION

This is the one of the largest nationally representative studies to date examining whether gender differences exist in ultrasound milestone evaluations. We found that, overall, there is no evidence suggesting significant gender disparity between male and female EM resident ultrasound milestone evaluations. While there are consistent small differences between male and female mean evaluation scores, the magnitude of these differences did not reach statistical significance, nor did they meet our a priori threshold of a 10% change. These findings may be unexpected given the recent

**Table 1**

Survey sample characteristics and national population estimates for emergency medicine residents in ACGME-accredited residency programs

<table>
<thead>
<tr>
<th>EM Resident/residency characteristic</th>
<th>Survey sample, n (%)</th>
<th>National estimate (95% CI)</th>
<th>National estimate, percentage (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>1,186 (100)</td>
<td>61,964 (52,448 to 71,479)</td>
<td>100</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>436 (36.8)</td>
<td>23,052 (18,357 to 27,746)</td>
<td>37.2 (32.3 to 42.5)</td>
</tr>
<tr>
<td>Male</td>
<td>750 (63.2)</td>
<td>38,861 (32,051 to 45,670)</td>
<td>62.8 (57.5 to 67.7)</td>
</tr>
<tr>
<td>Program affiliation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University</td>
<td>773 (65.1)</td>
<td>42,605 (22,983 to 62,226)</td>
<td>68.8 (38.5 to 88.6)</td>
</tr>
<tr>
<td>Community</td>
<td>418 (34.9)</td>
<td>19,359 (3,129 to 35,588)</td>
<td>31.2 (11.4 to 61.5)</td>
</tr>
<tr>
<td>Program size</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤35 residents</td>
<td>210 (17.7)</td>
<td>10,860 (956, to 22,676)</td>
<td>17.5 (4.8 to 47.5)</td>
</tr>
<tr>
<td>36+ residents</td>
<td>917 (82.3)</td>
<td>51,104 (31,646 to 70,561)</td>
<td>82.5 (52.5 to 95.2)</td>
</tr>
<tr>
<td>Self-reported (by program) number of completed ultrasound exams by residents upon graduation – mean (95% CI)</td>
<td>308 (301 to 315)</td>
<td>–</td>
<td>304 (225 to 383)</td>
</tr>
<tr>
<td>Presence of ultrasound fellowship</td>
<td>785 (66.1)</td>
<td>40,145 (19,969 to 60,320)</td>
<td>64.8 (34.3 to 86.7)</td>
</tr>
<tr>
<td>Percent female faculty – mean (95% CI)</td>
<td>32.2 (31.6 to 32.8)</td>
<td>–</td>
<td>35.6 (28.7 to 38.5)</td>
</tr>
<tr>
<td>Residency type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-year</td>
<td>326 (27.5)</td>
<td>19,083 (56, to 38,734)</td>
<td>30.8 (9.6 to 65.1)</td>
</tr>
<tr>
<td>3-year</td>
<td>861 (72.5)</td>
<td>42,881 (23,911 to 61,850)</td>
<td>69.2 (34.9 to 90.4)</td>
</tr>
</tbody>
</table>

Abbreviations: CI, confidence interval; EM, emergency medicine.
evidence of gender bias in medical training across multiple specialties. A recent comprehensive study by Dayal et al.\(^5\) revealed a wide gender gap in the milestone evaluations of male and female EM residents. The study found that despite having similar skills and knowledge at the beginning of residency training, female EM residents were evaluated lower on the vast majority of the 23 EM competencies/subcompetencies by the time of graduation when compared to their male counterparts.

A limitation of this study was its use of a convenience sample of residency programs. We hoped to overcome this limitation with a robust and representative study design that produces generalizable findings.

A number of factors could have contributed to the lack of gender bias in this study, contrary to that seen in recent literature. Perhaps the recent awareness and advancements toward gender equality are encouraging faculty members to deliberately address and reveal their own overt and covert biases. Traditionally, ultrasound evaluations are performed by faculty with ultrasound expertise who rely on objective factors for evaluating residents, such as direct supervision of psychomotor skills, image quality/acquisition, number of ultrasound examinations performed, quality assurance reviews, and performance on knowledge-based examinations. This is supported Amini et al.,\(^15\) who demonstrated that nearly all ACGME EM residency programs currently use image quality assurance and direct observation during clinical shifts during ultrasound skills assessments. These assessments are more objective and therefore might be less prone to bias compared to more subjective assessments. By choosing to evaluate a more objective milestone, greater levels of bias are likely required to demonstrate a difference between male and female evaluations.

The goal of this study was to identify a hypothesized gender bias for resident milestone achievement from a representative sample of programs; therefore, the decision was made to systematically stratify EM residency programs by geographic region. Since gender bias could be influenced by regional cultural practices and beliefs, it was presumed that stratification by geography might result in the greatest variance reduction (compared to simple random sampling). We used a population-based stratified cluster sample of EM residency programs that allowed us to generate national estimates for United States EM residency ultrasound milestone characteristics for men and women, and our findings are generalizable to all EM residents in the United States. Previous studies on this topic utilized convenience samples and therefore had a significant chance for selection bias. Our systematic method of sampling was chosen to minimize the risk of introducing any selection bias. Our findings may therefore better reflect the national status of gender differences among EM resident evaluations than prior studies, although our findings are limited in scope as they are focused on a single subcompetency.

**LIMITATIONS**

Our study has several limitations. Our focus on a single subcompetency limits generalizability in this regard, but it informs the point-of-care ultrasound community on how we are evaluating EM residents. There are also concerns regarding the current practice of using

---

**Figure 1.** Mean ultrasound milestone evaluation score estimates for emergency medicine (EM) residents in the United States. All estimates account for the complex survey sampling design of the study and represent national estimates for all EM residents in the United States from Fall (Fa) 2014 to Spring (Sp) 2017. Estimates for female residents are represented by circles and estimates for males by squares. I-bars represent 95% CIs. Integers 1 to 4 represent the postgraduate year for the fall and spring scores in the same academic year. Estimates represent seven distinct cohorts of EM residents (shaded alternating gray and black), each labeled parenthetically with the cohorts’ starting academic year in italics.
the ultrasound subcompetency to assess resident progression. The PC12 subcompetency has been largely criticized by experts in the point-of-care ultrasound community. A publication by Nelson et al., published in 2016, discussed several measurement issues with the current ultrasound subcompetency. As written, the ultrasound milestones do not address procedural guidance, the ability to utilize emergency ultrasound protocols and clinical algorithms, documentation of ultrasound examinations, and understanding of the limitations of emergency ultrasound. These omissions were addressed by Nelson et al. and incorporated into their proposed revision of the PC12 subcompetency. Some programs use the PC12 subcompetency as written currently, and some programs use the suggested and supported version as proposed by Nelson et al. A limitation to our study is that we are unaware if any of the participating programs are using the modified version of the subcompetency. These data could have been obtained using the survey we sent out to the programs and incorporated into our analysis.

Table 2
Mean differences for ultrasound milestone evaluation scores between male and female emergency medicine residents

<table>
<thead>
<tr>
<th>Cohort (starting academic year)</th>
<th>Post-graduate year</th>
<th>Evaluation date</th>
<th>Mean ultrasound milestone evaluation score difference: females-males (95% confidence interval)†</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011/2012</td>
<td>4</td>
<td>Fall 2014</td>
<td>0.30 (−0.12, 0.73)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spring 2015</td>
<td>0.14 (−0.07, 0.35)</td>
</tr>
<tr>
<td>2012/2013</td>
<td>3</td>
<td>Fall 2014</td>
<td>0.16 (−0.10, 0.42)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spring 2015</td>
<td>0.00 (−0.20, 0.19)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Fall 2015</td>
<td>0.11 (−0.11, 0.34)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spring 2016</td>
<td>0.11 (−0.12, 0.33)</td>
</tr>
<tr>
<td>2013/2014</td>
<td>2</td>
<td>Fall 2014</td>
<td>−0.19 (−0.49, 0.11)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spring 2015</td>
<td>−0.15 (−0.41, 0.11)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Fall 2015</td>
<td>−0.12 (−0.29, 0.06)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spring 2016</td>
<td>−0.02 (−0.11, 0.07)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Fall 2016</td>
<td>−0.18 (−0.46, 0.10)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spring 2017</td>
<td>−0.18 (−0.47, 0.11)</td>
</tr>
<tr>
<td>2014/2015</td>
<td>1</td>
<td>Fall 2014</td>
<td>−0.28 (−0.51, −0.06)‡</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spring 2015</td>
<td>−0.18 (−0.48, 0.13)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Fall 2015</td>
<td>−0.10 (−0.35, 0.15)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spring 2016</td>
<td>−0.03 (−0.23, 0.18)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Fall 2016</td>
<td>−0.01 (−0.14, 0.12)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spring 2017</td>
<td>−0.06 (−0.14, 0.01)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Fall 2017</td>
<td>−0.10 (−0.30, 0.11)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spring 2018</td>
<td>−0.22 (−0.49, 0.05)</td>
</tr>
<tr>
<td>2015/2016</td>
<td>1</td>
<td>Fall 2015</td>
<td>−0.05 (−0.27, 0.16)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spring 2016</td>
<td>−0.19 (−0.49, 0.11)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Fall 2016</td>
<td>−0.02 (−0.34, 0.29)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spring 2017</td>
<td>−0.07 (−0.30, 0.16)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Fall 2017</td>
<td>−0.01 (−0.19, 0.17)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spring 2018</td>
<td>0.01 (−0.08, 0.10)</td>
</tr>
<tr>
<td>2016/2017</td>
<td>1</td>
<td>Fall 2016</td>
<td>0.11 (−0.20, 0.41)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spring 2017</td>
<td>0.07 (−0.27, 0.40)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Fall 2017</td>
<td>−0.04 (−0.29, 0.21)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spring 2018</td>
<td>−0.03 (−0.19, 0.13)</td>
</tr>
<tr>
<td>2017/2018</td>
<td>1</td>
<td>Fall 2016</td>
<td>−0.03 (−0.33, 0.27)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spring 2017</td>
<td>0.07 (−0.17, 0.32)</td>
</tr>
</tbody>
</table>

Bold differences: mean female scores > male scores.

†Differences calculated from final regression model (generalized linear mixed model) as marginal mean difference between female and male mean scores for each post-graduate year and milestone evaluation date, controlling for all other variables in the model (see text for all covariates).

‡Statistically significant.
An additional limitation is that faculty evaluators are not consistent among all the programs. At programs with a robust ultrasound program, an ultrasound faculty member would be the evaluator, but other programs may have their residency program director make these determinations. Resident ultrasound rotations are also not standardized, so the length, structure, curricula, and materials can vary greatly from one program to the next. Timing of the ultrasound rotations can also affect the scores, since milestone evaluations are performed twice per academic year. If the resident had not completed their ultrasound rotation by the time the evaluations are completed, then there is one fewer data point.

We attempted to determine if evaluators were male or female, but most programs had both male and female faculty assess their residents, which made it difficult to form any conclusions on gender biases based on the gender of the evaluators. We did not collect data on the race of the evaluators as well, so we cannot comment on how that would potentially impact our results. In addition, we did not collect data related to nonultrasound milestones, which would have been ideal for comparison. We suggested that given the objective nature of ultrasound milestone evaluations, the scores may have been less vulnerable to the effects of implicit bias. This argument could have been supported had we obtained and compared data from some of the less objective milestones.

**CONCLUSIONS**

Our study results indicate there was no significant difference in the faculty evaluation of ultrasound
milestones among emergency medicine residents within training programs throughout the United States.

REFERENCES

The Implementation of a National Multifaceted Emergency Medicine Resident Wellness Curriculum Is Not Associated With Changes in Burnout

Kelly Williamson, MD\textsuperscript{1,2}, Patrick M. Lank, MD, MS\textsuperscript{3}, Nicholas Hartman, MD\textsuperscript{4}, Dave W. Lu, MD, MS, MBE\textsuperscript{5,6}, Natasha Wheaton, MD\textsuperscript{7,8}, Jennifer Cash, MD\textsuperscript{9}, Jeremy Branzetti, MD\textsuperscript{10,11}, Elise O. Lovell, MD\textsuperscript{1,2}, and for the Emergency Medicine Education Research Alliance (EMERA)

ABSTRACT

Background: The Accreditation Council for Graduate Medical Education Common Program Requirements effective 2017 state that programs and sponsoring institutions have the same responsibility to address well-being as they do other aspects of resident competence.

Objectives: The authors sought to determine if the implementation of a multifaceted wellness curriculum improved resident burnout as measured by the Maslach Burnout Inventory (MBI).

Methods: We performed a multicenter educational interventional trial at 10 emergency medicine (EM) residencies. In February 2017, we administered the MBI at all sites. A year-long wellness curriculum was then introduced at five intervention sites while five control sites agreed not to introduce new wellness initiatives during the study period. The MBI was readministered in August 2017 and February 2018.

Results: Of 523 potential respondents, 437 (83.5%) completed at least one MBI assessment. When burnout was assessed as a continuous variable, there was a statistically significant difference in the depersonalization component favoring the control sites at the baseline and final survey administrations. There was also a higher mean personal accomplishment score at the control sites at the second survey administration. However, when assessed as a dichotomous variable, there were no differences in global burnout between the groups at any survey administration and burnout scores did not change over time for either control or intervention sites.

Conclusions: In this national study of EM residents, MBI scores remained stable over time and the introduction of a multifaceted wellness curriculum was not associated with changes in global burnout scores.

Burnout, the triad of emotional exhaustion, depersonalization, and low personal accomplishment, arises from mismatch between an individual and their work environment in six key areas: workload, control, reward, community, fairness, and values.\textsuperscript{1} Physician burnout is widespread (45\%–55\%),\textsuperscript{2,3} with emergency...
medicine (EM) physicians reporting some of the highest levels among all specialties (55%–70%).2–5 Resident physicians report significantly higher levels of burnout (60%) than their age-adjusted nonphysician peers.6 For both attending and resident physicians, burnout negatively impacts patient care due to low professionalism as well as being associated with adverse patient safety events and low patient satisfaction.7–9 Burnout is associated with lower mental health,10 substance use disorders,11,12 and suicidal ideation.13,14 In residents, burnout is associated with career choice regret15 and in practicing physicians is associated with a reduction in clinical hours and intent to leave a medical practice.16,17

The impact of resident burnout and the importance of wellness on resident education and training were recently reinforced by the Accreditation Council for Graduate Medical Education’s (ACGME) release of new Common Program Requirements that require residency programs to regularly monitor burnout and have in place wellness promotion initiatives.18 However, methods through which programs can best monitor resident burnout and promote wellness remain unclear. In an attempt to address these issues, a consortium of content experts and graduate medical educators developed a novel and multifaceted wellness curriculum, including didactic presentations, corresponding nondidactic elements, individualized interactive instruction assignments, and additional Internet-based resources, for use among EM resident training programs.19 We aimed to determine if implementation of this wellness curriculum across multiple training programs improved resident burnout.

**METHODS**

**Study Design**

This study was a multicenter prospective educational trial performed at 10 ACGME-accredited EM residencies in the United States. Members of the Emergency Medicine Education Research Alliance (EMERA) were core faculty at the time of study initiation at all sites. The study was reviewed by each institution’s institutional review board and received approval at each site prior to study initiation.

**Subjects**

Eligible subjects for this study were PGY-1 to -4 EM residents at the participating programs during the study period of February 2017 to 2018. There were no further exclusion criteria. Participation in the survey study was voluntary.

**Study Protocol**

**Survey Instrument.** The survey instrument was sent to eligible participants at all study sites at three different time points in the study: February 2017, August 2017, and February 2018. The instrument was designed for completion in 15 minutes and consisted of 34 total questions, including demographic information, the Maslach Burnout Inventory,20 and four additional published wellness instruments: a quality-of-life assessment, the Primary Care Evaluation of Mental Disorders Patient Health Questionnaire 2 question depression screen (Prime-MD PHQ-2), an appraisal of career satisfaction, and a work–life balance rating.21–24 Informed consent was obtained from all subjects. The survey was administered either as a paper survey or via online proprietary software (SurveyMonkey) at the preference of the site study leader. Follow-up for nonresponders was programspecific, either in-person or via e-mail.

**Curriculum Intervention.** Prior to study initiation, each site self-selected as either a control site or an intervention site. A year-long multifaceted wellness curriculum was then introduced at five intervention sites while the other five control sites agreed not to introduce new wellness initiatives during the study period. The wellness curriculum included standardized bimonthly structured didactic elements presented by the study investigator at each site, individualized interactive instruction assignments, and additional Internet-based resources.19 Intervention sites delivered the didactic lectures and additional resources within a predetermined time frame so that each site completed their intervention prior to administration of the February 2018 survey.

**Data Analysis.** In addition to the questions that make up the Maslach Burnout Inventory (MBI) and wellness information in the survey instrument described above, basic demographic information was also obtained and included respondent age, sex, ethnicity, and PGY classification. Results of the components of the MBI are presented as both continuous and dichotomous data. “Global burnout” was defined as having both an emotional exhaustion score > 26 and a depersonalization score > 12 at any single survey administration.20,25
Descriptive statistics are presented as total number (n) and percentages with 95% confidence intervals (CIs) for categorical variables. Continuous variables are displayed as either means with standard deviation for normally distributed variables or as medians with interquartile ranges (IQR) for nonnormally distributed variables. Univariable analyses were performed using chi-square or Student’s t-test as appropriate for continuous or categorical variables. Logistic regression was performed to obtain adjusted odds ratios (ORs) for the intervention sites. Analysis was performed using a statistical package program (R version 3.3.2 [2016-10-31]).

RESULTS

The response rate for the February 2017 data collection was 285 of 382 (75%), for August 2017 was 247 of 386 (64%), and for February 2018 was 228 of 386 (59%). Of a total 523 potential respondents at the 10 different study sites, there were a total of 437 individuals who completed at least one survey (83.5%, 95% CI = 80.32% to 86.68%). A total of 769 completed surveys were collected across the three different survey administrations; 85 residents (16.3%, 95% CI = 13.1% to 19.4%) completed all three. There were no significant differences in age, sex, ethnicity, or PGY training year distribution between the control and intervention sites (Table 1).

Mean component scores and proportions of residents meeting criteria for global burnout were compared between control and intervention groups at each of the survey administration times (Table 2). There was a significant difference in MBI scores between intervention and control groups at baseline with a higher mean depersonalization score at the intervention sites (13.68 vs. 11.87, p = 0.02). At the second data collection, the only significant difference between sites was a higher mean personal accomplishment score at the control sites (40.26 vs. 38.50, p = 0.02). At the conclusion of the study, the only significant difference was a higher mean depersonalization score in the intervention sites (13.37 vs. 11.69, p = 0.04) (Table 3). When assessing burnout as a dichotomous variable, there was no difference in

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Demographics of Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>Control</td>
</tr>
<tr>
<td>Age (years), median (IQR)</td>
<td>29 (28–32)</td>
</tr>
<tr>
<td>Sex, % female (95% CI)</td>
<td>35.3% (28.1%–42.5%)</td>
</tr>
<tr>
<td>Ethnicity, % underrepresented in medicine (95% CI)</td>
<td>10.3% (5.4%–15.3%)</td>
</tr>
<tr>
<td>PGY, n</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td>IQR = interquartile range.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2</th>
<th>MBI Components of Burnout by Study Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>Intervention</td>
</tr>
<tr>
<td>Emotional exhaustion</td>
<td></td>
</tr>
<tr>
<td>Survey 1 (continuous)</td>
<td>21.1 (±9.1)</td>
</tr>
<tr>
<td>Survey 2 (continuous)</td>
<td>21.0 (±9.3)</td>
</tr>
<tr>
<td>Survey 3 (continuous)</td>
<td>21.3 (±9.5)</td>
</tr>
<tr>
<td>Depersonalization</td>
<td></td>
</tr>
<tr>
<td>Survey 1 (continuous)</td>
<td>11.9 (±6.2)†</td>
</tr>
<tr>
<td>Survey 2 (continuous)</td>
<td>11.2 (±6.5)</td>
</tr>
<tr>
<td>Survey 3 (continuous)</td>
<td>11.7 (±5.9)†</td>
</tr>
<tr>
<td>Personal accomplishment</td>
<td></td>
</tr>
<tr>
<td>Survey 1 (continuous)</td>
<td>38.8 (±5.8)</td>
</tr>
<tr>
<td>Survey 2 (continuous)</td>
<td>40.3 (±5.2)†</td>
</tr>
<tr>
<td>Survey 3 (continuous)</td>
<td>39.1 (±5.9)</td>
</tr>
</tbody>
</table>

Data are reported as mean (±SD). MBI = Maslach Burnout Inventory. †p < 0.05.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Proportion of Respondents Screening Positive for MBI Components of Burnout by Study Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>Intervention</td>
</tr>
<tr>
<td>Emotional Exhaustion</td>
<td></td>
</tr>
<tr>
<td>Survey 1 (dichotomous)</td>
<td>28.1 (±7.3)</td>
</tr>
<tr>
<td>Survey 2 (dichotomous)</td>
<td>25.5 (±7.0)</td>
</tr>
<tr>
<td>Survey 3 (dichotomous)</td>
<td>28.1 (±7.5)</td>
</tr>
<tr>
<td>Depersonalization</td>
<td></td>
</tr>
<tr>
<td>Survey 1 (dichotomous)</td>
<td>42.5 (±8.0)</td>
</tr>
<tr>
<td>Survey 2 (dichotomous)</td>
<td>38.3 (±7.8)</td>
</tr>
<tr>
<td>Survey 3 (dichotomous)</td>
<td>41.0 (±8.2)</td>
</tr>
<tr>
<td>Personal accomplishment</td>
<td></td>
</tr>
<tr>
<td>Survey 1 (dichotomous)</td>
<td>15.1 (±5.8)†</td>
</tr>
<tr>
<td>Survey 2 (dichotomous)</td>
<td>6.0 (±3.8)</td>
</tr>
<tr>
<td>Survey 3 (dichotomous)</td>
<td>12.2 (±5.4)</td>
</tr>
</tbody>
</table>

Data are reported as % (±SD). MBI = Maslach Burnout Inventory. †p < 0.05.
burnout between the intervention and control groups at any of the survey administrations (Table 4). In addition, average burnout scores did not change significantly over time for either the intervention or the control sites. When controlling for age, sex, and ethnicity using logistic regression, there remained no significant difference between intervention and control sites for odds of global burnout (Table 5).

**DISCUSSION**

In this year-long national study of EM residents, the introduction of a multifaceted wellness curriculum was not associated with changes in burnout scores. MBI scores remained stable over time. This study represents the first EM multicenter educational intervention trial to assess the effects of implementation of a formalized wellness curriculum on EM resident burnout.

The authors conducted a systematic literature review and drew upon previously published experience when creating the first published multifaceted EM wellness curriculum. From a systems standpoint, work hour limits have been associated with positive effects on burnout scores. Other attempts at mitigating resident burnout at an individual level have been restricted by small sample size and single residency program design. Positive effects from self-care workshops and meditation have been observed.

More comprehensive resident wellness curricula have been published. In general, conclusions about effectiveness are limited by lack of assessment and/or by the small number of participants and single-site design. A wellness curriculum developed over 6 years at the William Beaumont Family Medicine Residency Program emphasized how a curriculum including both residents and faculty members, and a “wellness champion” led to durable culture change. At the Oregon Health and Science University, a Resident and Faculty Wellness Program providing educational outreach and psychological counseling/psychiatric evaluation has demonstrated 10-year growth in utilization of services as well as high satisfaction from participants.

In the years since our curriculum was developed, academic leaders in EM, internal medicine, and pediatrics have recognized the need for a more comprehensive approach to creating and distributing well-being resources. The 2017 Emergency Medicine Resident Wellness Consensus Summit used a learning network of residents and attending physicians to create a 17 module resident wellness curriculum, educator toolkit resources, resident needs assessment, and program-level planning tool as well as a wellness-targeted technology database. The Collaborative for Healing and Renewal in Medicine (CHARM), supported by the Alliance for Academic Internal Medicine (AAIM) is a clearinghouse for learner wellness resources and scholarly activity. The American Academy of Pediatrics has developed a 14 module curriculum concentrating on the disclosure of life-altering diagnoses, provider’s response to challenging patient care experiences, and provider resilience. The Pediatric Resident Burnout and Resilience Study Consortium is a collaboration of over 40 pediatric and medicine–pediatric training programs, with the aim of developing and studying best practices to prevent and mitigate burnout in pediatric residents. The University of Arizona Center for Integrative Medicine Pediatric Integrative Medicine in Residency includes a 100 hour self-care curriculum that has been piloted at five pediatric residency programs and emphasizes approaches to a healthy lifestyle as well as concepts of pediatric integrative medicine.

Both the ACGME and the American Medical Association have placed increased emphasis on trainee wellness and resilience. In addition, the Accreditation Council for Graduate Medical Education (ACGME) Back to Bedside Initiative supports resident-driven projects that augment meaning in work through humanism and connection with patients. Organizational strategies to reduce burnout and promote work

<table>
<thead>
<tr>
<th>Table 4</th>
<th>Proportion (%) of Respondents Screening Positive for Global Burnout by Study Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Global Burnout</td>
</tr>
<tr>
<td>Survey 1</td>
<td>17.1 (±6.1)</td>
</tr>
<tr>
<td>Survey 2</td>
<td>18.8 (±6.3)</td>
</tr>
<tr>
<td>Survey 3</td>
<td>21.6 (±6.4)</td>
</tr>
</tbody>
</table>

Data are reported as % (±SD).

<table>
<thead>
<tr>
<th>Table 5</th>
<th>Adjusted ORs† (95% CI) for Global Burnout Among Respondents at Intervention Sites Compared to Control Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adjusted OR</td>
</tr>
<tr>
<td>Survey 1</td>
<td>0.62 (0.32–1.15)</td>
</tr>
<tr>
<td>Survey 2</td>
<td>0.89 (0.39–2.03)</td>
</tr>
<tr>
<td>Survey 3</td>
<td>1.05 (0.46–2.44)</td>
</tr>
</tbody>
</table>

†When controlling for age, sex, and ethnicity with control sites being the reference.
engagement are less common but likely more impactful than initiatives targeting individual care providers.  

In this study, there are a number of factors that may have contributed to the absence of an effect of the wellness curriculum on burnout and burnout scores in both intervention and control groups remaining stable over time. Following Maslach’s definition, burnout was defined as both emotional exhaustion > 26 and depersonalization > 12. The burnout prevalence in our study population was considerably lower than the approximately 50% previously reported.  

Prior studies have defined burnout somewhat inconsistently, often categorizing burnout as either high emotional exhaustion or high depersonalization, which may have led to a prevalence overestimate. The overall lower prevalence in our study population may have contributed to difficulty detecting smaller changes in burnout. Additionally, over the past several years there has been increased promotion and awareness of resident wellness nationally, which may have decreased the impact of our wellness curriculum. There has been a particular recognition since our curriculum was introduced that burnout is driven by systems issues and, for residents, impacted by the learning environment. While our curriculum included discussion on occupational wellness, the focus remained largely on the individual. Also, the control sites, though not introducing new initiatives in their programs during the study period, may have had preexisting programs or exposure to institutional initiatives that diluted differences between intervention and control sites. Three different site principal investigators, including two from intervention sites, changed institutions during the study period, which likely affected the consistency of engagement at those sites. Finally, resident compliance with the curriculum was variable. While the lectures were delivered during each program’s protected weekly conference time, attendance was subject to the constraints of resident vacation and schedule. While we promoted participation in all aspects of the curriculum, individual participation in the curriculum was not rigidly enforced nor monitored, so not all aspects of the curriculum received equal involvement. Residents particularly enjoyed dedicated time for resident bonding, residency wellness activities, and a wellness retreat. Conversely, individualized interactive instruction, designating wellness champions, and assigned readings about wellness were considered less useful and not as routinely accessed.  

The personal experience of the study investigators correlates with themes published in the literature over the past several years. Well-being interventions will be better received if they are personalized and encouraged but voluntary, respect the already high time burden on residents and opportunity costs of introducing new curricular elements, consider the culture of each program and resources of the institution, draw from a menu of possibilities, engage residents in the development and implementation of initiatives, and importantly target both the learning and the work environments as well as the individual. National platforms of resources, specialty specific or through organizations such as the ACGME, include standardized needs assessment surveys, well-being interventions, and implementation guidelines.  

LIMITATIONS  

There are important limitations to this study. While our sample is relatively large for a study of an educational intervention, it is still a convenience sample and was not subject to power analysis. No studies have defined a cutoff for a clinically significant change in MBI scores; however, studies have demonstrated associations between one-point increases in depersonalization or emotional exhaustion and the odds of self-reported medical error. One-point increases in emotional exhaustion have been associated with increased likelihood of decreasing professional work hours, and each one-point increase in burnout scores on each of the MBI subscales has been correlated with increased likelihood of reporting suicidal ideation. It is possible that the study was underpowered; however, the lack of consistent trends in the data argues against this interpretation. Of note, the initial and postintervention measurements were close to the date of the annual EM in-training examination in February. Both wellness and burnout are dynamic processes that fluctuate throughout the year and this time of year is well known to be stressful for residents. In addition, the control and intervention sites self-selected based on availability of resources and ability to introduce the year-long curriculum into their residencies. However, despite this self-selection, the intervention and control groups were well-matched. When burnout was assessed as a dichotomous variable, there was no difference in global burnout between groups at the baseline survey administration. When assessed as a
continuous variable, the only difference between groups was in the depersonalization component of burnout.

The objective of the study was a comparison of the intervention and control groups, rather than tracking individuals’ changes in burnout scores over time. Only 16.3% of residents completed all three MBIs, at least in part due to the February 2017 to February 2018 research design, as a portion of the study population graduated after the February 2017 survey administration and others entered residency in the summer of 2017. This design precluded a portion of the eligible participants from completing all three survey administrations. Eligible residents who completed a portion of the curriculum were included in the corresponding data collections to maximize the power of comparisons between the intervention and control sites at each data collection point.

Optimal solutions to burnout consider both the person and the organization and address excessive workload, inefficiencies, and lack of support; improve autonomy and work–home integration; and reinforce purpose, meaning, and accomplishment. Future research on the impact of burnout interventions for resident physicians will benefit from longitudinal study and investigations across specialties and practice environments and evaluate the combined effects of individual and organizational strategies.

CONCLUSIONS

In this national study of emergency medicine residents, the introduction of a multifaceted wellness curriculum was not associated with a change in burnout scores. Maslach Burnout Inventory scores remained stable over time. Further study is needed to determine best practices to lessen resident burnout.

The authors acknowledge Erin Quattromani, MD, Brooks Orb, MD, Doug Franzen, MD, and Jessica Klein, MD.

References

8. Lu DW, Weygandt PL, Pinchbeck C, Strout TD. Emergency medicine trainee burnout is associated with lower patients’ satisfaction with their emergency department care. AEM Educ Train 2018;2:86–90.
17. Shanafelt TD, Mungo M, Schmitgen J, et al. Longitudinal study evaluating the association between physician burnout...
The Variable Journey in Learning to Interpret Pediatric Point-of-care Ultrasound Images: A Multicenter Prospective Cohort Study

Charisse Kwan, MD¹, Martin Pusic, MD, PhD², Martin Pecaric, PhD³, Kirstin Weerdenburg, MD⁴, Mark Tessaro, MD¹, and Kathy Boutis, MD, MSc¹

ABSTRACT

Objectives: To complement bedside learning of point-of-care ultrasound (POCUS), we developed an online learning assessment platform for the visual interpretation component of this skill. This study examined the amount and rate of skill acquisition in POCUS image interpretation in a cohort of pediatric emergency medicine (PEM) physician learners.

Methods: This was a multicenter prospective cohort study. PEM physicians learned POCUS using a computer-based image repository and learning assessment system that allowed participants to deliberately practice image interpretation of 400 images from four pediatric POCUS applications (soft tissue, lung, cardiac, and focused assessment sonography for trauma [FAST]). Participants completed at least one application (100 cases) over a 4-week period.

Results: We enrolled 172 PEM physicians (114 attendings, 65 fellows). The increase in accuracy from the initial to final 25 cases was 11.6%, 9.8%, 7.4%, and 8.6% for soft tissue, lung, cardiac, and FAST, respectively. For all applications, the average learners (50th percentile) required 0 to 45, 25 to 97, 66 to 175, and 141 to 290 cases to reach 80, 85, 90, and 95% accuracy, respectively. The least efficient (95th percentile) learners required 60 to 288, 109 to 456, 160 to 666, and 243 to 1040 cases to reach these same accuracy benchmarks. Generally, the soft tissue application required participants to complete the least number of cases to reach a given proficiency level, while the cardiac application required the most.

Conclusions: Deliberate practice of pediatric POCUS image cases using an online learning and assessment platform may lead to skill improvement in POCUS image interpretation. Importantly, there was a highly variable rate of achievement across learners and applications. These data inform our understanding of POCUS image interpretation skill development and could complement bedside learning and performance assessments.
The use of emergency point-of-care ultrasound (POCUS) in pediatric emergency medicine (PEM) can improve patient outcomes and expedite patient disposition. As such, learning POCUS has become an increasing priority in PEM, with the learning experience typically including introductory courses and skill acquisition that relies on case-by-case clinical exposure. However, this approach may lead to challenges in achieving POCUS proficiency across a wide range of learners. Opportunities for bedside feedback may be limited by the number of POCUS-trained attendings available at a given site. This deprives the learner of immediate feedback, one of the most powerful methods of skill acquisition. Further, since the baseline pathology rate in pediatrics is relatively low, relying on case-by-case exposure to achieve sufficient skill may take years with clinical practice alone.

Gaining POCUS expertise is multifaceted and complex since the technique includes image acquisition, image interpretation, and integration of interpretation into clinical decision making. To facilitate learning of complex tasks, instructional design models recommend intentionally alternating part-task with whole-task training. In this light, e-learning provides an opportunity to expose learners to an image interpretation learning experience (i.e., part-task) that could complement the resource intensive face-to-face teaching and learning at the bedside that addresses all facets of POCUS (i.e., whole-task). However, to date, most POCUS simulation and online education focuses on image acquisition, specific conditions (e.g., cardiac tamponade), or adult applications. As such, existing teaching tools are limited in being able to significantly improve learner exposure to the part-task of pediatric POCUS image interpretation. However, Web-based learning and assessment image banks that provide intentional sequencing and targeted analytic feedback on hundreds of cases have demonstrated success for increasing electrocardiogram and musculoskeletal radiograph image interpretation skill. Using these types of learning platforms, the presentation of images is simulated to mirror how clinicians interpret them at the bedside. Specifically, cases are presented with a brief clinical stem, standard images and views, and juxtaposition of normal and abnormal cases. The cases are also presented in large numbers so that learners can learn similarities and differences between diagnoses, identify weaknesses, and build up a global representation of possible diagnoses. After each case the system provides visual and text feedback, which allows for deliberate practice and an ongoing measure of performance as part of the instructional strategy. This type of learning assessment platform could be applied to the image interpretation component of POCUS and potentially improve our understanding of POCUS image interpretation skill development.

We developed a POCUS image interpretation learning and assessment system that included four common pediatric POCUS applications (100 cases/application): soft tissue, lung, cardiac, and focused assessment ultrasound for trauma (FAST). We sought to determine PEM physician performance metrics and the number of cases and time within which most participants could achieve specific performance benchmarks.

### METHODS

#### Education Intervention

We used previously established methods to develop the education intervention. Deidentified images in the four POCUS applications acquired using the departmental ultrasound machine (Zonare Medical Systems) were exported from the local POCUS archive (Q-Path, Telexy Healthcare) in JPEG (still images) or MPEG-4 (cine-clips) formats. From this pool, two PEM POCUS fellowship-trained physicians selected a consecutive sample of 100 cases per application (400 total) that demonstrated acceptable image quality and a spectrum of pathology and normal anatomy. Any images with embedded technical clues that pointed to a diagnosis were excluded. Further, each application contained 50 cases with and 50 cases without pathology. For lung and cardiac applications, both video and still images were used because recognition of movement is key to interpretation. For soft tissue and FAST (except cardiac view) still images alone were used since these were more likely to adequately capture the subtleties distinguishing abnormal and normal cases and were more cognitively efficient for learners.

As the skill of assigning clinical significance and a specific diagnosis to POCUS findings (e.g., pneumonia) is best assessed at the bedside, the main goal of this education intervention was to develop the skill of distinguishing normal from any abnormal POCUS findings on video/still images, alongside forcing the learner to visually locate any abnormality for cases allocated as “abnormal.” The key initial educational outcome of accurately distinguishing normal from abnormal cases (vs. confirming a specific diagnosis) is also in keeping with
the user goals pediatric POCUS at the bedside since identifying an abnormal POCUS image should then prompt the physician to consider additional tests or consultations as needed to confirm or refine the diagnosis. Nevertheless, learning the possible specific diagnoses from POCUS imaging findings is also important. As a result, this information is presented with every case for the participant to consider in the text feedback. This approach satisfies the essential learning goals for pediatric POCUS, while providing additional information to participants to also learn specific diagnostic interpretations for each case.

All images were reviewed jointly by two POCUS experts and classified as POCUS images as normal versus abnormal, where abnormal images had changes that could be considered pathological (Table 1). A third POCUS expert then independently provided normal/abnormal classifications, and discrepancies between the classifications were resolved by consensus (κ = 0.86, 95% confidence interval [CI] 0.81 to 0.91). Two POCUS experts then subsequently marked up the abnormal images using graphics to highlight pathology, creating clickable areas over the pathology with 2 to 3 mm of allowance just outside the abnormal area. These images were then embedded into a template generated using a Flash-integrated development environment.

A website was developed using HTML, PHP, and a mysql database. In brief, once a participant was provided unique access, they were taken to the online system and presented with 100 cases per application. For each case, they considered a brief clinical description stem and an unmarked POCUS image (still ± video). Cases were presented in random order unique to each participant. After review of the case, the user declared the case as definitely normal, probably normal, probably abnormal, or definitely abnormal. The definitely/probably assignments permitted the user to express the certainty of their response. If their answer was in the “abnormal” category, the user was then required to designate one area of abnormality using an interactive system. When ready, the participant submitted their response, after which they received immediate visual and written feedback on the correctness of their response, diagnosis of the case, and normal anatomy (demonstration at https://imagesim.com/demo/ or Figure 1). Prior to launch, the system was pilot tested on five participants (one POCUS nonexpert, four POCUS experts) for technical and content issues and revised in consequence.

### Table 1
Pathology Presented by the Learning and Assessment System for Each POCUS Application

<table>
<thead>
<tr>
<th>Pathology (n = 50)†</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Soft tissue</strong></td>
<td></td>
</tr>
<tr>
<td>Cellulitis and abscess</td>
<td>29 (58)</td>
</tr>
<tr>
<td>Cellulitis</td>
<td>19 (38)</td>
</tr>
<tr>
<td>Foreign body</td>
<td>2 (4)</td>
</tr>
<tr>
<td><strong>Lung</strong></td>
<td></td>
</tr>
<tr>
<td>Pneumonia, with and without pleural effusion</td>
<td>25 (50)</td>
</tr>
<tr>
<td>Inflammation/bronchiolitis</td>
<td>20 (40)</td>
</tr>
<tr>
<td>Pneumothorax</td>
<td>3 (6)</td>
</tr>
<tr>
<td>Pulmonary edema</td>
<td>2 (4)</td>
</tr>
<tr>
<td><strong>Cardiac</strong></td>
<td></td>
</tr>
<tr>
<td>Pericardial effusion</td>
<td>32 (64)</td>
</tr>
<tr>
<td>Poor function</td>
<td>18 (36)</td>
</tr>
<tr>
<td><strong>FAST</strong></td>
<td></td>
</tr>
<tr>
<td>Multiple areas of free fluid</td>
<td>25 (50)</td>
</tr>
<tr>
<td>Pelvic free fluid</td>
<td>8 (16)</td>
</tr>
<tr>
<td>Right or left upper quadrant free fluid</td>
<td>12 (24)</td>
</tr>
<tr>
<td>Pericardial effusion</td>
<td>5 (10)</td>
</tr>
</tbody>
</table>

FAST = focused assessment sonography in trauma.
†50 cases with pathology per application; 50 cases without pathology to make up 100 total.

### Study Design and Setting
The education tool was developed in collaboration with two tertiary care pediatric centers and an industry partner. This research was undertaken using a multicenter prospective cohort design. We recruited a convenience sample of PEM physicians in the United States and Canada from September to November 2018.

### Selection of Participants
An e-mail was sent to PEM division heads, PEM fellowship program directors, and P2 Network (an international collaborative of pediatric POCUS physicians, https://p2network.com/) site leads asking them to forward the e-mail to their respective fellows and attending physicians. Interested participants contacted the study coordinator. We also recruited five “expert” PEM POCUS physicians (separate from study team) who had each completed a PEM POCUS fellowship and performed at least 1,500 bedside scans. This study was approved by the institutional review boards at the study institutions.

### Measurements
Secure entry was ensured via unique participant login credentials and access to the system was available 24 hours per day, 7 days per week. We collected information on type of learner (fellow vs. attending), geographic...
location, and self-reported POCUS scans completed (none, <50, 51–100, 101–500, 501–1,000). Participants were asked to complete an introductory tutorial and at least one of the four 100-case POCUS applications. The system automatically time stamped the time a case was started to the time a case response was submitted. Participants were given a time limit of 4 weeks to minimize decay of learning that may confound results. Participants were given a time limit of 4 weeks to minimize decay of learning that may confound results. At 2 weeks, participants who had not completed at least one application were sent an e-mail reminder.

Outcomes

By surveying 150 members of the P2 network, we determined that the performance benchmarks of 80, 85, 90, and 95% accuracy could be considered educationally meaningful for a variety of contexts. Thus, we provided data that demonstrated learning curves of the participants and the median number of cases that participants need to complete to achieve the latter performance benchmarks (primary outcome). Since we anticipated that there would be a variation between participants to achieve these performance benchmarks, we provided this data separately for the average (50th percentile) and least efficient learner (95th percentile) of nonexpert participants.

We also examined the change in accuracy per application between the first and last 25 cases. Further, we measured the effect size of learning gains, changes in sensitivity and specificity per application, and differences in learning gains between fellows versus attendings and made comparisons between applications. We examined if there was any association between demographic variables and the odds of achieving expert-level performance. Expert-level data were also used to examine relations-to-other variables validity (construct), where we would expect that expert performance would be significantly higher than those that are relative nonexperts. Finally, to examine feasibility, we calculated average amount of time spent by participants per case and per 100-case module.

Data Analyses

Sample Size. From our previous work, the educationally important difference between initial and final scores was approximately 10% and the proportion of discordant pairs is about 12%. We also assumed that $\alpha = 0.05$ and $\beta = 0.80$. Using a McNemar test power analysis (PASS 11), we calculated a minimal sample size of 95 of PEM POCUS nonexpert participants per application.
**Scoring.** Participants were scored only on the broad category selections of “normal” or “abnormal” and not the subassignments of “probably or definitely” normal/abnormal. Specifically, normal items were scored dichotomously, while abnormal items were scored correct if the participant classified the case as abnormal and indicated one correct region of abnormality.

**Number of Cases to a Performance Benchmark.** We modeled the learning curves of each individual participant with a random coefficients hierarchical logistic regression model. Using these learning curves, we predicted the median number of cases required for a participant to attain a performance benchmark by solving the individual regression equation for the number of cases required to achieve the log odds (logit) that would correspond to the selected performance benchmark. From these data, we also plotted the proportion of participants that would reach 95% accuracy for a given number of cases to determine if these histograms were normally distributed by testing for both skewness and kurtosis of the distributions.

**Learning Outcomes.** Using the initial (pre) and final (post) 25 cases, we calculated change in accuracy, sensitivity, specificity, and Cohen’s d effect sizes for each application, with respective 95% CI. We analyzed for the association of achieving POCUS expert accuracy performance (dependent variable) and a priori selected independent variables using a logistic regression model. The independent variables included years in practice (≤5 years vs. >5 years), academic setting (university affiliated children’s hospital vs. other), POCUS training during fellowship (yes vs. no), number of any type of scan completed prior to participation (≥100 vs. <100), and accuracy score on initial 20 cases (≥80% vs. <80%).

**Comparisons Between Applications.** Independent and dependent normally distributed parametric data were compared with a Student’s t-test and paired Student’s t-test, respectively. Analysis of variance (ANOVA) testing was used to compare three or more means from independent proportions, and the Bonferroni test was used to perform post hoc analyses.

**Time Commitment.** We determined the median time it took to complete each 100-case application with respective interquartile range. We compared the time commitment between applications using the Kruskal-Wallis test.

Significance was set at p < 0.01 to account for multiple testing. All analyses except the regression analyses were carried out using SPSS (Version 23, IBM 2015). The regression models were performed using Stata (Version 14, StataCorp LLC).

**RESULTS**

**Study Participants**

We enrolled 177 PEM physicians. Of these, 172 were PEM POCUS learner physicians (fellows [n = 65] and attendings [n = 107]) and five were PEM POCUS experts (Figure 2). Participants represented 28 (56%) of the 50 US states and five (50%) of the 10 Canadian provinces. There were significantly more fellows than attendings that received POCUS training as part of their fellowship.

---

**Figure 2.** POCUS PEM physician participation. Sum total of all specific application participants was greater than 128 since most participants completed more than one application. FAST = focused assessment sonography for trauma; PEM = pediatric emergency medicine; POCUS = point-of-care ultrasound.
of their fellowship experience (82.0% vs. 30.8%; difference = 51.2% [95% CI = 36.2–61.7]; Table 2).

At least one application was completed by 128 (74.4%) participants (Figure 2), 88 (68.8%) of whom completed all four applications, while 11 (8.6%) completed three, seven (5.5%) completed two, and 22 (17.2%) completed one application.

**Number of Cases to Performance Benchmarks**

A qualitative review of the modeled learning curves demonstrates significant variation between participants in rates of achieving higher performance (Figure 3). For the average (50th percentile) learners, the predicted median number of cases needed across our four applications was 0 to 45 for 80% accuracy, 25 to 97 for 85% accuracy, 66 to 175 for 90% accuracy, and 141 to 290 for 95% accuracy (Figure 4A). The least efficient (95th percentile) of learners would have to complete 60 to 288 cases to achieve 80% accuracy, 109 to 456 cases for 85% accuracy, 160 to 666 cases for 90% accuracy, and 243 to 1,040 cases for 95% accuracy. Participants required the highest number of cases for the cardiac application and the lowest number for the soft tissue application to reach a specified benchmark (Figure 4B).

The distribution of the number of participants that would achieve the expert benchmark for a given number of cases was skewed for all applications (p < 0.0001). The lung and cardiac distributions also demonstrated kurtosis (p < 0.0001; Figure 5).

**Performance Outcomes**

The pre/post change in accuracy for each application is detailed in Table 3. The Cohen’s d-effect sizes for each application were moderate to large and ranged

---

**Table 2**  
Participant Demographics

<table>
<thead>
<tr>
<th>Demographic Variable</th>
<th>PEM Fellow N=65</th>
<th>PEM Attending N=107</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant from the United States, no. (%)</td>
<td>52 (80.0)</td>
<td>67 (62.6)</td>
<td>0.02</td>
</tr>
<tr>
<td>Years since completing postgraduate training, no. (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 6 years</td>
<td>NA†</td>
<td>44 (41.1)</td>
<td>NA†</td>
</tr>
<tr>
<td>6-10 years</td>
<td></td>
<td>20 (18.7)</td>
<td></td>
</tr>
<tr>
<td>11-15 years</td>
<td></td>
<td>13 (12.1)</td>
<td></td>
</tr>
<tr>
<td>16-20 years</td>
<td></td>
<td>10 (9.3)</td>
<td></td>
</tr>
<tr>
<td>&gt; 20 years</td>
<td></td>
<td>20 (18.7)</td>
<td></td>
</tr>
<tr>
<td>Female sex, no. (%)</td>
<td>44 (72.1)</td>
<td>72 (63.2)</td>
<td>0.23</td>
</tr>
<tr>
<td>Employment Setting, no. (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University affiliated general/community hospital</td>
<td>3 (4.6)</td>
<td>6 (5.6)</td>
<td>0.53§</td>
</tr>
<tr>
<td>University affiliated children’s hospital</td>
<td>61 (93.8)</td>
<td>99 (92.5)</td>
<td></td>
</tr>
<tr>
<td>Non-university affiliated general/community hospital</td>
<td>2 (3.1)</td>
<td>1 (0.9)</td>
<td></td>
</tr>
<tr>
<td>Non-university affiliated children’s hospital</td>
<td>0</td>
<td>1 (0.9)</td>
<td></td>
</tr>
<tr>
<td>Type of Point-of-Care Ultrasound Training, no. (%)‡</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>2 (3.1)</td>
<td>13 (12.1)</td>
<td>&lt;0.0001§</td>
</tr>
<tr>
<td>Integrated into emergency medicine residency training</td>
<td>3 (4.6)</td>
<td>9 (8.4)</td>
<td></td>
</tr>
<tr>
<td>Integrated into pediatric residency training</td>
<td>3 (4.6)</td>
<td>1 (0.9)</td>
<td></td>
</tr>
<tr>
<td>Integrated into PEM fellowship training</td>
<td>53 (82.0)</td>
<td>33 (30.8)</td>
<td></td>
</tr>
<tr>
<td>Workshops/ Courses</td>
<td>7 (10.8)</td>
<td>75 (70.70)</td>
<td></td>
</tr>
<tr>
<td>Institutional faculty training</td>
<td>NA</td>
<td>61 (57.0)</td>
<td></td>
</tr>
<tr>
<td>Self-directed learning</td>
<td>8 (12.3)</td>
<td>27 (25.2)</td>
<td></td>
</tr>
<tr>
<td>Number of educational/clinical point-of-care ultrasound examinations completed, no. (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>2 (3.1)</td>
<td>5 (4.4)</td>
<td>0.12</td>
</tr>
<tr>
<td>&lt;50</td>
<td>30 (45.2)</td>
<td>43 (40.4)</td>
<td></td>
</tr>
<tr>
<td>51-100</td>
<td>17 (26.2)</td>
<td>20 (18.4)</td>
<td></td>
</tr>
<tr>
<td>101-500</td>
<td>15 (23.1)</td>
<td>28 (26.3)</td>
<td></td>
</tr>
<tr>
<td>501-1000</td>
<td>1 (0.2)</td>
<td>15 (14.0)</td>
<td></td>
</tr>
</tbody>
</table>

†NA - Not applicable  
‡Participants could choose all that apply  
§p-value comparing distributions in each learner group
Figure 3. Predicted learning curves for soft tissue, lung, cardiac, and FAST. FAST = focused assessment sonography for trauma.

Figure 4. (A) The x-axis represents the benchmark level of proficiency as chosen by the educator. The y-axis represents the predicted number of cases required to achieve that benchmark based on the logistic regression model described in the text. The two panels represent two different educational contexts. (A) The median learner is represented—50% of learners required fewer cases to achieve the benchmark; 50% required more. (B) A higher ambition is presented: the number of cases required of the marginal learner so that 95% of the learners will have achieved the given benchmark (i.e., 95% of learners would need to do fewer; 5% would need to do even more). FAST = focused assessment sonography for trauma.
from 0.6 to 1.0. There were no differences in fellow versus attending final accuracy, sensitivity, or specificity scores for soft tissue, lung, or cardiac applications. For FAST, however, attendings had higher final accuracy, (6.0% difference; 95% CI = 1.8, to 10.2) and sensitivity (7.5% difference; 95% CI = 0.6 to 14.4). There was no association of PEM POCUS physician learners achieving expert performance with any of the baseline variables: >100 POCUS scans experience versus <100 (odds ratio [OR] = 1.2, 95% CI = 0.5 to 2.6), initial accuracy >80% versus <80% (OR = 2.1, 95% CI = 0.8 to 5.6), <5 versus >5 years in practice (OR = 1.4, 95% CI = 0.5 to 3.5), children’s hospital versus other setting (OR = 1.6, 95% CI = 0.3 to 8.8), POCUS training in fellowship versus none (OR = 2.1, 95% CI = 0.9 to 4.5). Per application, PEM POCUS expert mean (95% CI) accuracy scores were soft tissue 96.0% (92.3% to 99.7%), lung 96.0% (93.9% to 98.1%), cardiac 90.0% (81.8% to 98.2%), and FAST 93.0% (88.0% to 98.0%). Expert final scores were significantly higher than those of nonexpert participant scores, with an accuracy difference of 7.3% (4.4% to 10.4%).

Comparisons Between Applications
The final accuracy sensitivity and specificity (Table 3) differed between applications (ANOVA p < 0.001), with post hoc Bonferroni analyses showing that the cardiac application had lower final accuracy (p < 0.01), sensitivity (p < 0.01), and specificity (p < 0.01) relative to each of the other applications. For the soft tissue application, there were greater learning gains for normal cases (specificity = +18.9%) versus abnormal cases (sensitivity = +4.9%), with a difference of 14.0% (95% CI = 9.8% to 18.2%). There were no differences in learning gains for abnormal versus normal cases in the other applications.

Time Commitment
The initial (first 25 case) mean time per case was 31.7 seconds, which decreased to 21.2 seconds on the final 25 cases (difference = −10.5 seconds, 95% CI = −8.6 to −12.4). The median time (interquartile range) in
minutes it took to complete each 100-case application was as follows: soft tissue (still image) 29.5 (20.9 to 40.7), lung (video + still image) 52.3 (39.7 to 72.7), cardiac (video + still image) 52.0 (39.7 to 72.7), and FAST (video + four still images) 63.6 (49.1 to 76.6) minutes (p < 0.0001).

**Missing Data**
There were no differences in the demographics of the 44 of 172 (25.6%) who dropped out versus the 128 who completed at least one 100-case application. Further, there were no differences in initial accuracy of those participants who completed the 100-case application versus those who did not (mean initial accuracy difference = 0.14%, 95% CI = –4.2 to 3.9).

**DISCUSSION**
We demonstrated that the case numbers required to reach the performance benchmarks ranged considerably for both the average and least efficient learners. Further, we noted that the deliberate practice of POCUS image interpretation led to skill improvement within a 100-case online learning experience, and most participants needed only about 2 to 3 hours to achieve the highest performance benchmark accuracy of 95% for all applications except cardiac, which required 3 to 10 hours. These data could inform education strategies and potentially add to the skill development of POCUS image interpretation.

Most organizations are faced with decisions about ideal credentialing POCUS standards to allow their physicians to safely practice at the bedside. This complex discussion often includes consensus building, stakeholder engagement, and the use of standardized competency-setting methods. Our results can inform these organizational discussions on the variable journey learners take to a given performance benchmark, while the broader POCUS community considers which performance benchmarks are most appropriate. Our data also demonstrate that the numbers of cases required to achieve a performance benchmark were not normally distributed, with skewed distribution tails, indicating that a minority of learners required considerably more cases to attain each benchmark. This has policy implications for education guidelines based on an “average” learner. Specifically, our data suggest that a shift away from the current standard of recommending specific numbers of cases for POCUS proficiency (e.g., 25–50 cases) in favor of learners achieving a performance-based competency benchmark. This approach is also in keeping with evolving performance-based competency frameworks in residency training, which promote greater accountability and documentation of actual capability.

Separating out the POCUS image interpretation task and then integrating knowledge gained into whole task activities can reduce cognitive load during face-to-face bedside POCUS teaching. This may result in more efficient and effective learning and greater learning satisfaction as demonstrated in similar forms of blended learning. Alternatively, one could argue that POCUS image interpretation may be inherently easier to learn at the bedside, where one can pursue visual hints, make comparisons to unaffected areas, or better integrate the clinical context. However, these bedside advantages are challenged by the learner also needing to simultaneously acquire the images and conditions that lead to suboptimal image acquisition may limit the learning of image interpretation. Further, given the relatively low frequency of pathology in pediatrics, bedside learning may not efficiently offset the fact that many cases may be needed for most learners to achieve practice-ready standards, whereby a 50% or higher abnormal rate has previously been shown to be optimal for achieving an acceptable user sensitivity. Given the pace of case exposure, bedside education may also not be very effective at identifying specific areas of individual participant weaknesses or
applications that are more difficult to learn. For example, our results provided evidence that cardiac and normal soft tissue cases were more difficult than other POCUS applications. Future research should explore the interaction between learning the image interpretation skill via online deliberate practice and the real-time application of POCUS at the bedside.36,37

The education intervention demonstrated a large effect size for the soft tissue and lung applications and a moderate effect size for the cardiac and FAST applications.30 One possible explanation for these disparities is that cardiac and FAST applications were more difficult to interpret, due to increased complexity of anatomy (cardiac, FAST), differences in number of views (FAST), or low a priori skill due to low rates of pathology of these applications at the bedside. Strategies that may enhance diagnostic performance outcomes using the education intervention in this study include scaffolding (e.g., embedding electronic hints or coaching),38,39 more deliberate review of incorrect cases (including referring to supplemental resources), and/or repeating cases as many times as needed to reach a desired performance outcome.

None of the participant baseline variables that we tested predicted for achieving expert-level performance. While it is not possible to be certain of the reasons for this from our data, we can consider some possible explanations. With respect to the variable of number of scans (100 vs. >100), the number alone may not be sufficient to predict for achieving expert level if a participant did not routinely receive feedback on images acquired at the bedside, which may limit a participant’s ability to learn from each scan performed.7 POCUS training during PEM fellowship versus no training during fellowship may not have alone impacted participant ability to achieve expert scores since many of our study attending-level participants also engaged in institutional and other POCUS workshops/courses. Further, since almost all our participants worked at children’s hospitals, we were likely underpowered to examine the impact of practice setting. Finally, since scores are weighted for the 25 most recent cases and case interpretation difficulty varied over the 100-case experience, initial accuracy scores may not have predicted for subsequent and final scores.

**LIMITATIONS**

Image interpretations were based on the expert opinion of three POCUS experts and the interobserver agreement between these opinions was high; however, expert opinions may be subject to error.40 Participants were able to select the applications and about one-quarter of our enrolled participants did not complete the minimum study intervention; therefore, our results may be biased by increased engagement for the selected applications and/or toward more POCUS-motivated PEM physicians. Since this is a part-task education intervention removed from the bedside and utilizes a higher proportion of pathological cases than is typically experienced at the bedside, it is uncertain how knowledge gained via this tool will translate to patient-level skill and outcomes. Some participants required considerably more than the 100 cases prescribed within the study design. Like other models of educational outcome distributions, we have the most information about persons at the mean and the extreme predictions run the risk of spectrum bias.

Over the 4-week study period, other factors may have contributed to the study’s reported learning outcomes. However, our data demonstrates that most participants completed the cases in one sitting so it is unlikely additional POCUS exposure influenced our study results significantly. The changes in participant performance were reported using a pre-post designs, which may be subject to confounders that impact the validity of the results. Finally, this study included PEM physician participants and thus may not be generalizable to other categories of physician learners.

**CONCLUSIONS**

Deliberate practice of pediatric point-of-care ultrasound image cases using an online learning and assessment platform may lead to skill improvement in point-of-care ultrasound image interpretation. Further, this method allows an efficient review of a larger number of cases than would be typically available with bedside practice alone. Our results also demonstrated that the rate of learning the image interpretation task of point-of-care ultrasound can be highly variable across learners. These data could inform education strategies and potentially add to our understanding of how the skill of point-of-care ultrasound image interpretation is acquired among pediatric emergency medicine physicians.

We acknowledge the efforts of Ms. Kelly Sobie who facilitated participant recruitment and providing access information to the study participants. We also thank the pediatric emergency medicine physicians who participated in this research.
References

The Variability of Preferred Infant Lumbar Puncture Insertion Site Between Novice and Experienced Physicians

Jeffrey T. Neal, MD1, Jason A. Levy, MD1, Rachel G. Rempell, MD2, and Rebecca L. Vieira, MD1

ABSTRACT

Background: We sought to determine if vertebral interspace selection for performance of infant lumbar puncture (LP) varies between less experienced trainees and more experienced pediatric emergency medicine (PEM) attending physicians.

Methods: We performed an observational prospective study using a convenience sample of infants aged 0 to 12 months presenting to a single emergency department. Trainees with limited LP experience (defined as less than 10 infant LPs performed) marked their preferred LP insertion site with an invisible ultraviolet pen. PEM attending physicians subsequently marked their preferred LP insertion site with a visible pen. A trained sonographer then performed a bedside ultrasound to confirm interspace concordance or discordance. Our primary outcome was the proportion of concordant marked insertion sites.

Results: Of the 110 patients enrolled, 102 (92.8%) completed study procedures. Trainee and PEM attending LP interspace markings were concordant in 27% of cases. Trainees marked a preferred interspace below the level of the attending in 55% of patients: 29 (28.4%) marked one spot inferior, 20 (19.6%) marked two spots inferior, and seven (6.9%) marked three spots inferior in relation to the attending.

Conclusions: There is variability of preferred LP insertion site based on provider experience. Trainees with limited LP experience tended to mark insertion spaces more caudal than those marked by the attending physicians in an area where the subarachnoid space is slightly smaller.

Lumbar puncture (LP) is a frequently performed procedure in infants to evaluate for meningitis. Traditional technique involves palpation of anatomic landmarks followed by “blind” needle insertion into the selected interspinous space. A survey of pediatric interns entering residency demonstrated limited experience, poor knowledge, and low confidence with LP.1 Upon graduating, pediatric residents are expected to perform this procedure independently and successfully; however, on average, they perform a median of 12 infant LPs during residency and commonly supervise other trainees in this procedure despite a low success rate.2

Lumbar punctures in children are often unsuccessful and, not surprisingly, less physician experience has...
been identified as a risk factor. According to several pediatric studies, the reported rate of unsuccessful LP (traumatic puncture or inability to obtain cerebrospinal fluid [CSF]) can be as high as 50%. Research on the effect of simulation on LP success has been equivocal, likely due to additional factors that are not able to be duplicated, such as patient movement or challenges with patient positioning and stabilization. Point-of-care ultrasonography (POCUS) allows providers to visualize the anatomic landmarks of LP, which may help improve success rates. In children, POCUS has been utilized to increase provider confidence in choosing the LP insertion site and to select alternative suitable interspaces. Most importantly, it has been utilized to mark the LP insertion site, leading to increased success.

Selection of the proper insertion site is a major contributor to LP success. Based on palpation landmarks, the recommendation is to insert the needle in the midline of the interspinous space above or below Tuffier’s line (i.e., L3 to L4 or L4 to L5). Anecdotally, those with less experience tend to choose an insertion site that is significantly lower than Tuffier’s line (caudally), where the subarachnoid space tapers, or too lateral (off the midline) potentially explaining the higher failure rates observed in trainees.

More education is needed for trainees surrounding the procedural technique of LP given their limited exposure during training and high failure rates. Understanding why trainees fail is the first step in determining where to focus educational efforts. Increasing the proportion of successful LPs could lead to decreased patient pain, decreased hospitalizations, and lower hospital costs by reducing the number of LPs resulting in uninterpretable or unobtainable CSF.

Our goal was to determine if preferred LP insertion site varied based on procedural experience. We hypothesized that trainees would mark more caudal insertion sites than pediatric emergency medicine (PEM) attending physicians.

METHODS

Study Design and Setting
We performed an observational prospective study of patients in a single urban academic pediatric emergency department (ED) that has approximately 60,000 patient visits per year and is staffed with medical students, physician assistants, trainees (pediatrics, medicine-pediatrics and emergency medicine residents, and PEM fellows), and PEM attending physicians. In the 0- to 12-month age group, a mean of 270 LPs are performed annually at the study institution. No specific LP training program existed aside from standard hospital policy for procedures during the entirety of this study. The institutional review board approved this study, and the study was registered at ClinicalTrials.gov (NCT02949869).

Study Population
During the study period, any trainee with limited infant LP experience (defined as less than 10 infant LPs) and any PEM-trained attending physician were invited to participate. All trainees and PEM attending physicians provided written informed consent prior to participating in the study.

Over an 18-month period from January 2017 to June 2018, we recruited a convenience sample of infants aged 0 to 12 months who presented for care to the study site ED. Research assistants screened patients for eligibility. Exclusion criteria included prior history of LP in the preceding 72 hours; known spinal cord abnormality, such as tethered cord; or clinically unstable patients (up to the discretion of the treating physician). Sampling of patients was limited to study sonographer availability, which typically occurred during daytime hours. Written informed consent was obtained by a research assistant or study investigator prior to enrollment.

Study Protocol
Once the subject was consented and enrolled, a clinical assistant held the patient in the lateral decubitus position, while the trainee marked the preferred LP interspace insertion site with an invisible ultraviolet skin marker using the landmark technique. The PEM attending physician subsequently entered the room and marked the preferred LP interspace insertion site with a visible skin marker. The sonographer then unveiled the trainee marking with an ultraviolet light and marked the trainee spot with a visible skin marker. A bedside ultrasound (US) was then performed utilizing a Mindray TE7 linear high-frequency transducer to identify trainee and attending interspace concordance or discordance. Attempts were made to keep the patient in the same position at all steps of the protocol, but due to patient comfort and provider availability, this was not always possible.
The patient was placed in a standardized lateral decubitus position with the spine flexed as if planning for LP. The termination of the conus medullaris was identified. The markings were defined as “below the level of the conus” if the conus did not cross the interspace. The markings were then correlated to the closest intervertebral space; if the marking was over a spinous process, the closest intervertebral space was selected. If the trainee and attending physician markings were discordant (not over the same interspace), the number of spaces cephalad or caudal in relation to the attending physician marking was sonographically determined and recorded. The subarachnoid space width (with the probe oriented in the longitudinal plane) and area (with the probe oriented in the transverse plane) were then measured at each marking (see Figure 1). If the spaces marked by the attending and trainee were concordant, only one measurement was performed. On a select number of random patients, a second sonographer repeated the protocol to obtain images and measurements to evaluate sonographer agreement.

Three study sonographers performed the USs (two PEM board-certified physicians with RDMS certification [RLV, JAL] and one pediatric emergency medicine fellow [JTN]). Two additional sonographers (both PEM board-certified physicians with additional PEM POCUS fellowship training) independently performed the USs to determine interspace concordance and consistency of subarachnoid space width and area measurements. Each sonographer reviewed sonographic technique for visualization of landmarks and marking procedure with the principal investigator prior to enrolling patients to standardize practice. All US images and clips were reviewed by the PEM fellow for completion utilizing a Web-based computer platform (Telexy Health).

**Methods and Measurements**

Baseline demographic and clinical characteristics of the patient population were collected by a research assistant on a standardized data sheet. Trainee information including residency program, level of training, self-reported prior infant LP experience (stratified into groups of zero, one to five, and six to 10), and self-reported prior total LP experience (stratified into groups of zero, one to five, six to 10, 11 to 20, and >20) were recorded. Stratifications were determined a priori. A follow-up chart review was performed by research coordinators and the principal investigator on each patient following hospital discharge and missing information, such as final diagnosis, was recorded.

Sample size calculation was based on the main outcome of the proportion of correlated skin markings between trainee and attending physician. There are limited to no data on this question, so we estimated there would be an 80% correlation. We calculated a sample size of 110 to detect this 80% correlation with 7.5% precision and an alpha of 0.05.

**Outcomes**

Our primary outcome was the proportion of concordant trainee- and attending physician–marked LP insertion sites as determined by ultrasonography. A second trained sonographer performed an examination for approximately 10% of patients (n = 14) to determine the inter-rater reliability of interspace agreement.

Our secondary outcomes included numeric interspace difference between trainee- and attending
physician–marked sites as well as mean subarachnoid space width (cm) and subarachnoid space area (cm²) for trainees versus attending physicians.

**Data Analysis**

Descriptive statistics were used to report baseline and demographic characteristics of the study population. To assess our primary outcome, we calculated simple proportions of concordance rates between trainee- and physician-marked insertion sites. To assess our secondary outcomes of selected subarachnoid space and width, we calculated means with 95% confidence intervals (CIs). To assess inter-rater agreement between sonographers for qualitative factors, a kappa statistic was calculated. To estimate correlations between individual measurements made by sonographers on the same target, an intraclass correlation coefficient was calculated. All analyses were performed utilizing STATA 13.0 (StataCorp).

**RESULTS**

**Characteristics of Study Subjects**

We present the characteristics of trainee providers (Table 1) and study patients (Table 2). Sixty-eight trainees and 52 attendings were involved in marking. Eighty-six (84%) of the trainees who participated in the study were interns. Of the trainees, 51 (50%) had no prior infant LP experience, whereas overall, 12 (12%) had no prior clinical LP experience on a patient of any age. No trainee or attending marked more than three different subjects, except for two attendings who each marked five subjects.

A total of 110 patients were enrolled. Eight patients withdrew after initial written consent due to parental preference or discharge prior to study completion. The median age of the patients was 140 days (range = 4–354 days). Three study sonographers performed all of the bedside ultrasonography on the patients. Investigator 1 (JTN) performed 99 examinations, investigator 2 (RLV) performed two examinations, and investigator 3 (JAL) performed one examination. Two additional sonographers assisted with kappa measurements.

**Main Results**

Trainee and attending physician preferred LP insertion sites were concordant in 27% (95% CI = 19% to 37%) of cases. Trainees marked a preferred interspace below the level of the attending in 55% (95% CI = 45% to 65%) of patients, and above the level of the attending in 18% (95% CI = 10% to 25%) of patients. In the cases where trainees marked below the level of the attending, 29 (28.4%) marked one spot inferior, 20 (19.6%) marked two spots inferior, and seven (6.9%) marked three spots inferior in relation to the attending (see Figure 2).

The mean trainee- and attending physician–marked subarachnoid space widths were 0.83 (95% CI = 0.78

---

**Table 1**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Specialty</strong></td>
<td></td>
</tr>
<tr>
<td>Pediatrics</td>
<td>45 (44.1)</td>
</tr>
<tr>
<td>Emergency medicine</td>
<td>57 (55.9)</td>
</tr>
<tr>
<td><strong>Trainee level</strong></td>
<td></td>
</tr>
<tr>
<td>PGY-1</td>
<td>86 (84.3)</td>
</tr>
<tr>
<td>PGY-2</td>
<td>16 (15.7)</td>
</tr>
<tr>
<td>PGY-3/4</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td><strong>Prior infant LP experience</strong></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>51 (50.0)</td>
</tr>
<tr>
<td>1–5</td>
<td>47 (46.1)</td>
</tr>
<tr>
<td>6–10</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td><strong>Prior total LP experience</strong></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>12 (11.8)</td>
</tr>
<tr>
<td>1–5</td>
<td>41 (40.2)</td>
</tr>
<tr>
<td>6–10</td>
<td>26 (25.5)</td>
</tr>
<tr>
<td>11–20</td>
<td>18 (17.6)</td>
</tr>
<tr>
<td>&gt;20</td>
<td>1 (1.0)</td>
</tr>
</tbody>
</table>

†Missing data from four trainees regarding prior LP experience. LP = lumbar puncture.
to 0.89) and 0.91 (95% CI = 0.87 to 0.96) cm, respectively. The mean difference in width was 0.08 (95% CI = 0.04 to 0.12, \( p = 0.00006 \)). The mean trainee- and attending physician–marked subarachnoid space areas were 0.63 (95% CI = 0.56 to 0.70) and 0.72 (95% CI = 0.65 to 0.79) cm\(^2\), respectively. The mean difference in area was 0.09 (95% CI = 0.04 to 0.14, \( p = 0.00036 \)) cm\(^2\). In two cases (2%), an attending physician marked an interspace at the level of the conus. In four cases (7%), the trainee marked a space with little to no measurable CSF (see Figure 3).

Fourteen subjects had a second sonographer obtain images and measurements. The kappa for sonographer identification of interspace correlation was 0.65 (95% CI = 0.21 to 1.0). The intraclass correlation for quantitative measurements between sonographers was 0.98 (95% CI = 0.97 to 0.99).

**DISCUSSION**

Trainees have lower LP success rates than attending physicians.\(^2,3,9,19\) LP success rates reflect multiple factors including adequate preparation, patient cooperation or stabilization, and proper procedural technique. We demonstrated wide variability between trainee- and attending physician–preferred LP insertion sites. Although trainees may have selected interspaces that would yield CSF, most trainees selected lower inter-spinous spaces with smaller subarachnoid spaces, which may explain decreased success rates reported in the literature.

Our study is the first to use POCUS to determine the relation between trainee- and attending physician–preferred insertion sites and to measure the variability in size of the subarachnoid space. US has been previously used to determine optimal LP positioning,\(^20\) to increase the confidence of the proceduralist,\(^13\) and to mark the insertion site.\(^4,15\) In one US study, 41% of selected insertion sites had two additional suitable spaces superior to the selected site.\(^14\) As trainees tended to mark a preferred insertion site more caudal than attending physicians, our study supports the anecdotal suggestion to attempt “one space higher” after an initial unsuccessful trainee attempt. Alternatively, a screening US examination could be performed prior to infant LP to identify and mark the conus, allowing clinicians to select a safe interspace for needle insertion with the largest subarachnoid area, thus improving the chance of success.
Additional research is needed to determine whether or not outcomes (e.g., LP success) vary with selected insertion site. Given that we found slightly larger subarachnoid spaces at more superior insertion sites, one could theorize that these sites are more likely to be associated with LP success. Previous studies have examined the effect of positioning or fluid boluses on the size of the subarachnoid space, but no studies to date have studied the relationship between subarachnoid space size and LP success.

Research has demonstrated that unsuccessful LPs lead to higher costs and higher rates of hospitalization for the youngest infants. This study suggests that the selection of a suboptimal insertion site may contribute to increased trainee LP failure rates. Given the decreasing exposure during residency where the average graduating resident now performs a median of 12 LPs, it is important to supplement clinical experience with other learning opportunities. It is possible that the lack of a more standardized LP training procedure at this institution contributed to these results. Effectively teaching appropriate LP insertion site may be an important step toward improving trainee success in this procedure.

**LIMITATIONS**

Our study has several limitations. First, this study is subject to observation bias (specifically, the Hawthorne effect). This might have caused attending physicians or trainees to mark a spot higher or lower (e.g., a “safer” space) than normal given that they were being studied. It is possible that some providers may have changed behavior to fit the “expected” results of the study.

Second, the study is not tied to outcomes (e.g., LP success). Although the trainees most frequently marked one space more caudal than attending physicians, we could not determine if this would have made a clinical difference. Even though the subarachnoid space was slightly smaller at the trainee-selected insertion site, it is possible the trainee-selected spaces would result in obtaining CSF. Given that the diameter of a 22-gauge LP spinal needle (the size typically utilized for infant LPs) is 0.07 cm, it is unclear if the statistically significant difference in width of 0.08 cm would be clinically significant.

Third, bedside US examinations were limited to three sonographers, potentially limiting the generalizability of this study to other institutions with limited POCUS experience. The kappa statistic of 0.65 demonstrates a moderate level of inter-rater agreement for obtaining qualitative measurements, but the CI ranges from 0.21 (minimal agreement) to 1.00 (perfect agreement), so no definitive conclusions can be made. Nevertheless, the high intraclass correlation of 0.96 (95% CI = 0.88 to 0.99) demonstrates high reliability for quantitative measurements, and other studies have suggested that physicians can quickly and reliably identify the sonographic landmarks needed for LP.

Fourth, the study was conducted at a single academic institution where the majority of procedures are performed by trainees with no specific LP training program beyond what they receive in intern orientation. Therefore, our findings may not be applicable to other settings where rigorous LP simulation exists or trainees have more experience with this procedure.

Finally, although the selected insertion sites were blinded between the trainee and attending physician as well as to the second sonographer in select cases, the selection of sites was not blinded to the initial sonographer. In addition, the sonographers were all investigators on this project. These factors could have introduced bias in the sonographic measurements of the subarachnoid space.

**CONCLUSIONS**

In conclusion, there is variability of preferred infant lumbar puncture insertion sites based on provider experience. Trainees with limited lumbar puncture experience tend to mark insertion sites one interspace caudal to the sites selected by the attending pediatric emergency medicine physicians.

The authors acknowledge Michael C. Monuteaux, ScD, of Boston Children’s Hospital for his guidance on the statistical analysis of the data. We thank all of the research assistants for their help in the recruitment of the patients for this study.

**References**


Core Content for Pediatric Emergency Medicine Ultrasound Fellowship Training: A Modified Delphi Consensus Study

Erika Constantine, MD¹, Marla Levine, MD², Alyssa Abo, MD³, Alex Arroyo, MD⁴, Lorraine Ng, MD⁵, Charisse Kwan, MD⁶, Janette Baird, PhD⁷, and Allan E. Shefrin, MD⁸, on behalf of the P2 Network Point-of-care Ultrasound Fellowship Delphi Group*

ABSTRACT

Background: Pediatric emergency medicine (PEM) point-of-care ultrasound (POCUS) fellowships exist to provide learners with expertise in ultrasound (US) education, administration, and research oversight. Currently, there are no standardized goals or objectives for these programs, resulting in considerable variability in PEM POCUS fellowship training.

Methods: A modified Delphi survey of PEM and general emergency medicine (EM) POCUS experts in Canada and the United States was conducted to obtain consensus regarding the most important curricular components of a PEM POCUS fellowship training program. Participants were solicited from the P2 Network mailing list and from PEM and EM POCUS fellowship directors listed on the Society of Clinical Ultrasound Fellowships and the Canadian Society of POCUS-EM Fellowships websites. Curricular components considered as part of the survey included US skills, educational skills, administrative skills, and research requirements. Consensus was considered to have been reached when ≥80% of respondents agreed to either include or exclude the component in fellowship training.

Results: Round 1 of the survey was sent to 311 participants. A total of 118 (37.9%) completed eligibility for the survey, and 92 (78.0%) met eligibility criteria. Of those, 80 (67.8% of eligible participants) completed the first round of the survey. Round 2 of the survey was sent to those who completed part 1, and 64 (80.0%) completed that round. During Round 1, consensus was achieved for 15 of 75 US applications, seven of seven educational skills, nine of 11 administrative skills, and four of six research requirements. In Round 2 of the survey, consensus was reached on two additional US skills, but no additional administrative skills or research requirements.
Conclusions: With a consensus-building process, the core content for PEM POCUS fellowship training was defined. This can help POCUS educators formulate standardized curricula to create consistent training in POCUS fellowship graduates.

Over the past two decades, the use of point-of-care ultrasound (POCUS) has increased exponentially in the emergency care setting.1–4 Training programs have recognized the importance of ultrasound (US) in the care of both adult and pediatric patients.5–11 The Accreditation Council for Graduate Medical Education (ACGME) has designated POCUS as a core competency for general emergency medicine (EM) residency graduates.12,13 POCUS is also listed in the core objectives of the pediatric emergency medicine (PEM) fellowship training guidelines by the American Board of Pediatrics and the Royal College of Physicians and Surgeons of Canada.6 Additionally, a consensus curriculum for core POCUS skills to be taught as part of PEM fellowship were recently published as a guide for PEM fellowship directors and US directors.14 Point-of-care US fellowship programs were created within the field of EM with goals to refine POCUS skills and develop proficiency in US education, program development, and oversight.7 There are over 100 EM POCUS fellowship training programs across North America and core content guidelines have been published to guide general EM POCUS fellowship directors.7 Because general EM POCUS fellowships are currently not ACGME certified, these core content guidelines are important for providing a framework for US educators. Prior to 2011, there were few training opportunities in POCUS specific to PEM. Many of the early adopters of PEM POCUS obtained POCUS training through EM POCUS fellowships or through self-directed learning, using general EM POCUS courses, third-party courses, or online educational modalities. There are now 13 PEM POCUS fellowship programs listed on the Society of Clinical Ultrasound Fellowships (SCUF) website and one PEM POCUS fellowship listed on the Canadian Society of POCUS-EM Fellowship website. Several other advanced POCUS training programs are offered, either independent of SCUF or in conjunction with PEM fellowship training. Current PEM US fellowships are loosely based on general EM US fellowships in that they usually require completion of 1000 scans during fellowship as well as an US-based research project. However, there are no standardized educational objectives within these PEM POCUS advanced training programs, leading to significant variability in PEM POCUS fellowship curricula.15,16 To create a standardized framework for PEM POCUS fellowship training, we conducted a modified Delphi consensus process to delineate the core content for PEM POCUS fellowships and advanced PEM US training programs.

METHODS

The Modified Delphi Process

The Delphi process is an iterative process that aims to achieve consensus opinion among a group of experts based on their responses to a series of questionnaires.17 The process was chosen for this particular study because it allowed for the solicitation of anonymous opinions from a broad group of experts in the field of POCUS. Additionally, the process is more cost-effective than in-person meetings, allowing for greater participation for a geographically diverse cohort.

Participants were presented with the various proposed curricular components and were asked to rate their importance for acceptability for inclusion in the curriculum on a scale of 1 to 5, 1 being “not at all important” and 5 being “very important.” Consensus was achieved when ≥80% of respondents agreed on a particular item (either for inclusion or for exclusion from the curriculum). This level of agreement has been used in previous Delphi studies, including studies designed to better define POCUS curriculum for PEM fellows.14 Between rounds, the data were analyzed and the survey was modified to include the data and comments. Modified surveys were then presented to the participants for consideration in the subsequent round. Consensus was achieved when ≥80% of respondents agreed on a particular item (either for inclusion or for exclusion from the curriculum). This level of agreement has been used in previous Delphi studies, including studies designed to better define POCUS curriculum for PEM fellows.14 Between rounds, the data were analyzed and the survey was modified to include the data and comments. Modified surveys were then presented to the participants for consideration in the subsequent round. Participants were informed of which curricular components had reached consensus, and these components were removed from consideration for subsequent rounds. Precise data regarding overall group response for nonconsensus items were not provided to survey respondents between rounds to ensure the independence of responses. Additionally, because this was an anonymous survey, there was no mechanism to provide participants with their own Round 1 response to allow them to make an informed decision in Round 2. The decision to pursue further rounds
was based on whether there were any significant changes in nonconsensus items between rounds.

**Survey Tool Development**

The content of the survey tool used for this study was designed by a group of 16 experts in PEM POCUS and general EM POCUS education. All are members of the P2 Network. The P2 Network is an international group of Pediatric POCUS experts and enthusiasts, open to all who are interested. It was formed in 2014 with the intention of sharing expertise, building research collaborations, and offering mentorship in the field of pediatric POCUS.18

The survey was developed using the same framework as a previously published curriculum for general EM POCUS fellowships.7 The survey was divided into four main sections: 1) US applications, 2) educational skills, 3) administration skills, and 4) research skills. Twenty-one basic US skills, previously determined to be core skills required as part of a PEM fellowship, were not included in the survey, and participants were informed of this at the beginning of the survey.14 Additional basic and advanced POCUS skills deemed relevant to emergency care of the pediatric patient were included in the survey. A broad definition of what was considered relevant to pediatric practice was used. The survey included questions relevant to the structure of POCUS training and skills required for completion of PEM POCUS fellowship training. The survey also solicited suggestions for additional components that should be considered as part of a formal PEM US fellowship curriculum. The survey was reviewed and approved by all members of the research team prior to its implementation. The study was approved by the institutional review board (IRB) of the primary study site, Rhode Island Hospital, in Providence, Rhode Island.

**Survey Participants**

Participation in the survey was solicited from individuals who practiced in Canada or the United States, had personally completed more than 1500 US scans, and had PEM POCUS leadership positions or training. PEM POCUS leadership or training was defined as meeting one of the following criteria: 1) has completed a PEM POCUS fellowship, 2) serves as a PEM POCUS lead or director, 3) serves as a PEM POCUS fellowship director, or 4) serves as a general EM POCUS fellowship director and teaches PEM POCUS skills as a part of this role. This definition of “expert” was agreed upon by the study authors because it acknowledges the variety of training backgrounds in the current field of PEM POCUS and requires more scans than is typically required of a newly graduating PEM POCUS fellow. Study authors and contributing authors who met inclusion criteria for survey participation were included among those invited to participate. All had an expertise similar to that required for participation in the survey. Potential survey participants were solicited from the P2 Network mailing list and from PEM and EM POCUS fellowship directors listed on the Society of Clinical Ultrasound Fellowships and the Canadian Society of POCUS-EM Fellowships websites,15,16 which to our knowledge, offers the most comprehensive list of general and pediatric POCUS fellowships offered in Canada and the United States.

**Survey Administration and Data Analysis**

REDCap (v. 7.1.2, Vanderbilt University), a secure Web application for building and managing online surveys and databases, was used to distribute the survey and collect the results. Initial participation was solicited via e-mail, and up to three reminders, 1 week apart, were sent to nonresponders. All questions within the survey required a response, although participants could choose not to finish the survey. Data were analyzed using SAS (v.9.4, SAS Institute).

**RESULTS**

We conducted Round 1 of the survey in September 2017. A total of 311 individuals were invited to participate. Of the 280 people from this list with known affiliations, 73.9% indicated PEM as their specialty. A total of 118 responded (37.9%) and 92 (78.0% of respondents) met eligibility criteria. Eighty participants (87% of eligible) completed Round 1. We conducted Round 2 of the survey in March 2018 and 64 experts (80% of eligible) completed this portion of the survey (Figure 1).

The majority of study participants reported working in the United States, and 59.7% of respondents saw only children as part of their practice. Experts were largely involved in POCUS curriculum development for EM residency, PEM fellowship, and EM POCUS fellowships. Many also reported involvement in medical school and pediatrics residency POCUS training. Most of the survey participants reported having a PEM fellowship, an EM POCUS fellowship, or a PEM POCUS fellowship at their institution. A total of
2.6% of experts reported having no PEM or US fellowship at all at their institutions (Table 1).

After the first round, there was consensus to include 15 PEM POCUS applications and 20 PEM POCUS fellowship skills in the domains of research, education, and administration. In Round 2, consensus was reached for two additional PEM POCUS applications (POCUS for long-bone fracture identification and US guided arterial line placement), but no additional consensus was achieved for skills in any of the other domains (Table 2). The applications and skills for inclusion in a PEM POCUS fellowship curriculum are summarized in Figure 2. There was no expert consensus reached on the remaining 58 PEM POCUS applications, two administrative skills, and two research requirements. Percent agreement for these items can be found in Table 3.

Our survey also included questions regarding general fellowship structure that could serve as a guideline for those developing a PEM POCUS fellowship. In Round 1, experts reached consensus that the completion of an IRB-approved research project that can lead to publication would meet the requirements for scholarly activity. In Round 2, this was broadened to include the completion of a non-IRB-reviewed project, such as a podcast or blog, non-peer-reviewed publication, or other quality improvement or educational initiative. Submission of a case report or case series to fulfill the scholarly activity requirement in fellowship did not reach consensus.

Experts reached consensus (90% agreement) in the first round that a PEM POCUS fellowship should be a 1-year program completed after PEM fellowship is successfully completed. Combined PEM/PEM US fellowships completed within a 2 year for EM residents or 3- or 3.5-year time frame for pediatrics residents did not reach consensus.

Experts were asked in Round 1 whether third-party certification should be a requirement for completion of PEM POCUS fellowship, with a request for a “yes,” “no,” or “I don’t know” response. This item did not reach consensus (13.8% “yes,” 77.5% “no,” and 8.8% “I don’t know”) in Round 1. In Round 2, this question was revisited because study authors believed that the “I don’t know” option allowed survey respondents to remain undecided, potentially preventing consensus on this topic. As such, the question was changed to remove the “I don’t know” response option, and an additional question, “Should eligibility for third-party certification be a requirement for completion of PEM POCUS fellowship,” was added to further characterize participants’ thoughts on third party certification. In Round 2, 88.9% of respondents answered “no” to the question of whether third party certification should be required as part of fellowship completion. This was the sole item of the survey that was rejected by experts. However, for the question regarding eligibility for certification, consensus was still not reached (46.0% “yes,” 54.0% “no”).

Finally, our survey solicited comments from participants regarding additional elements to be considered as part of a fellowship curriculum. There was one suggestion for leadership training and another for interdisciplinary collaboration and scanning opportunities with various experts such as cardiology, radiology, or anesthesia. These were not felt to be a strong enough response to consider including in Round 2 of the survey. No other suggestions for curriculum content were offered.

The Delphi process was terminated after two rounds as only two additional items reached consensus out of 64 remaining items surveyed that had not achieved consensus on the first round.
To our knowledge, this is the first modified Delphi consensus study outlining core content for PEM POCUS fellowships. Through a consensus process, we were able to outline 17 PEM POCUS applications and 20 PEM POCUS fellowship skills in the domains of research, education, and administration, recommended as part of PEM POCUS fellowship training. Currently, there is a lack of a standardized curriculum or structure for PEM POCUS fellowships, with each program crafting its own goals and objectives. In October 2018, the American Board of Medical Specialties approved a Designation of Focused Practice in Advanced Emergency Medicine Ultrasonography under the American Board of Emergency Medicine, but no similar designation exists for the American Board of Pediatrics. The resulting flexibility in the structure of PEM POCUS fellowships has resulted in significant variability in the outcome of training. By outlining core elements of PEM POCUS training that should be included in all US fellowship programs, new PEM US fellowship graduates should demonstrate uniformity in knowledge and expertise, regardless of where they trained.

In designing our survey tool, we were deliberately broad in our inclusion of US applications to be considered for fellowship training, primarily because uses

---

**Table 1**
Round 1 Expert Participant Demographics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country of work (n = 77)</td>
<td></td>
</tr>
<tr>
<td>United States</td>
<td>63 (81.8)</td>
</tr>
<tr>
<td>Canada</td>
<td>14 (18.2)</td>
</tr>
<tr>
<td>Years of practice since residency (n = 77)</td>
<td>Median = 6 (IQR = 3.8)</td>
</tr>
<tr>
<td>Primary patients seen in ED (n = 77)</td>
<td></td>
</tr>
<tr>
<td>Children only</td>
<td>46 (59.7)</td>
</tr>
<tr>
<td>Adults only</td>
<td>24 (31.1)</td>
</tr>
<tr>
<td>Children and adults</td>
<td>7 (9.1)</td>
</tr>
<tr>
<td>US curriculum development involvement (n = 80)</td>
<td></td>
</tr>
<tr>
<td>Medical school</td>
<td>30 (37.5)</td>
</tr>
<tr>
<td>Pediatric residency</td>
<td>22 (27.5)</td>
</tr>
<tr>
<td>EM residency</td>
<td>41 (51.2)</td>
</tr>
<tr>
<td>PEM fellowship</td>
<td>58 (72.5)</td>
</tr>
<tr>
<td>PEM POCUS fellowship</td>
<td>32 (4.0)</td>
</tr>
<tr>
<td>EM POCUS fellowship</td>
<td>38 (47.5)</td>
</tr>
<tr>
<td>Other (APPs/global health/critical care/surgical subspecialties)</td>
<td>3 (3.8)</td>
</tr>
<tr>
<td>Fellowships available at expert institution (n = 80)</td>
<td></td>
</tr>
<tr>
<td>PEM</td>
<td>53 (66.3)</td>
</tr>
<tr>
<td>PEM POCUS</td>
<td>27 (33.8)</td>
</tr>
<tr>
<td>EM POCUS</td>
<td>47 (58.8)</td>
</tr>
<tr>
<td>None</td>
<td>2 (2.5)</td>
</tr>
</tbody>
</table>

1n = 77 due to three participants choosing not to complete the demographic portion of the survey.

APP = advanced practice provider; PEM = pediatric emergency medicine; POCUS = point-of-care ultrasound; US = ultrasound.

**Table 2**
Number of Items Receiving ≥80% Consensus in Each Round

<table>
<thead>
<tr>
<th>Domain (Number of Items Rated)</th>
<th>Round 1</th>
<th>Round 2</th>
<th>Total Items Reaching Consensus, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEM POCUS applications (75)</td>
<td>15</td>
<td>2</td>
<td>17 (22.7)</td>
</tr>
<tr>
<td>Educational skills (7)</td>
<td>7</td>
<td>0 (none asked)</td>
<td>7 (100)</td>
</tr>
<tr>
<td>Administration skills (11)</td>
<td>9</td>
<td>0</td>
<td>9 (81.8)</td>
</tr>
<tr>
<td>Research requirements (6)</td>
<td>4</td>
<td>0</td>
<td>4 (66.7)</td>
</tr>
</tbody>
</table>

PEM = pediatric emergency medicine; POCUS = point-of-care ultrasound.

---

**DISCUSSION**

To our knowledge, this is the first modified Delphi consensus study outlining core content for PEM POCUS fellowships. Through a consensus process, we were able to outline 17 PEM POCUS applications and 20 PEM POCUS fellowship skills in the domains of research, education, and administration, recommended as part of PEM POCUS fellowship training. Currently, there is a lack of a standardized curriculum or structure for PEM POCUS fellowships, with each program crafting its own goals and objectives. In October 2018, the American Board of Medical Specialties approved a Designation of Focused Practice in Advanced Emergency Medicine Ultrasonography under the American Board of Emergency Medicine, but no similar designation exists for the American Board of Pediatrics. The resulting flexibility in the structure of PEM POCUS fellowships has resulted in significant variability in the outcome of training. By outlining core elements of PEM POCUS training that should be included in all US fellowship programs, new PEM US fellowship graduates should demonstrate uniformity in knowledge and expertise, regardless of where they trained.

In designing our survey tool, we were deliberately broad in our inclusion of US applications to be considered for fellowship training, primarily because uses
<table>
<thead>
<tr>
<th>System</th>
<th>POCUS Applications Assumed Mastered in PEM Fellowship*</th>
<th>PEM POCUS Fellowship Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiovascular</td>
<td>Identify non-traumatic pericardial effusion</td>
<td>Assess IVC for volume status</td>
</tr>
<tr>
<td></td>
<td>Identify traumatic pericardial effusion</td>
<td>Identify tamponade physiology</td>
</tr>
<tr>
<td></td>
<td>Identify cardiac standstill</td>
<td>Chamber size and comparison</td>
</tr>
<tr>
<td></td>
<td>Evaluate cardiac function</td>
<td></td>
</tr>
<tr>
<td>Lung</td>
<td>Identify hemothorax</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Identify pleural fluid/effusion</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Identify pneumothorax</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Identify lung consolidation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Identify pulmonary edema</td>
<td></td>
</tr>
<tr>
<td>Soft Tissue/</td>
<td>Identify abscess</td>
<td></td>
</tr>
<tr>
<td>Musculoskeletal</td>
<td>Identify cellulitis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Identify soft tissue foreign body</td>
<td></td>
</tr>
<tr>
<td>Abdomen</td>
<td>Identify free peritoneal fluid in trauma</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Identify intussusception</td>
<td></td>
</tr>
<tr>
<td>Renal/Genitourinary</td>
<td>Assess bladder volume</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Identify intrauterine pregnancy</td>
<td></td>
</tr>
<tr>
<td>Ocular</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Procedural Skills</td>
<td>Abscess incision and drainage</td>
<td>Perform US guided nerve blocks</td>
</tr>
<tr>
<td></td>
<td>Central line placement</td>
<td>US guided arthrocentesis</td>
</tr>
<tr>
<td></td>
<td>Peripheral intravenous access</td>
<td>US guided arterial line placement</td>
</tr>
<tr>
<td></td>
<td>Soft tissue foreign body localization/removal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pericardiocentesis</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Educational Skills</th>
<th>Administrative Skills</th>
<th>Research Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Development and Dissemination of Educational Content</strong></td>
<td><strong>Quality Assurance and Improvement</strong></td>
<td>Critical analysis of medical literature</td>
</tr>
<tr>
<td>• Assess and develop curriculum content</td>
<td>• US image assessment and feedback</td>
<td>Research project development</td>
</tr>
<tr>
<td>• Develop lecture presentation content and organization</td>
<td>• Reporting and management of incidental US findings</td>
<td>Research project abstract and manuscript preparation</td>
</tr>
<tr>
<td>• Engage learners during lecture presentations</td>
<td>• US coding and billing</td>
<td>Completion of a scholarly project during fellowship</td>
</tr>
<tr>
<td>• Utilize multiple hands-on education methods</td>
<td><strong>Leadership</strong></td>
<td></td>
</tr>
<tr>
<td>• Organize an ultrasound instructional course</td>
<td>• Education oversight</td>
<td></td>
</tr>
<tr>
<td><strong>Competency Assessment</strong></td>
<td>• Equipment and software acquisition and maintenance oversight</td>
<td></td>
</tr>
<tr>
<td>• Understand competency pathways</td>
<td>• Research oversight</td>
<td></td>
</tr>
<tr>
<td>• Assess competency</td>
<td>• Workflow solution oversight</td>
<td></td>
</tr>
<tr>
<td><strong>Relationships and Networks</strong></td>
<td><strong>Research Requirements</strong></td>
<td></td>
</tr>
<tr>
<td>• Hospital credentialing and privileging</td>
<td>Critical analysis of medical literature</td>
<td></td>
</tr>
<tr>
<td>• Demonstrate knowledge of local, national and international US organizations</td>
<td>Research project development</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. Skills and requirements recommended upon completion of a PEM POCUS fellowship. *Consensus on POCUS curriculum as part of PEM fellowship are outlined by Shefrin et al.14 PEM = pediatric emergency medicine; POCUS = point-of-care ultrasound.
POCUS Curriculum Items With no Agreement to Include or Exclude by Modified Delphi: Percentage of Experts Who Ranked Item “Important/Very Important” After Round 2

<table>
<thead>
<tr>
<th>Table 3</th>
<th>PEM POCUS FELLOWSHIP CORE CONTENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>US applications</td>
<td></td>
</tr>
<tr>
<td>Cardiac applications</td>
<td></td>
</tr>
<tr>
<td>Cardiac output assessment</td>
<td>64.1</td>
</tr>
<tr>
<td>Assess IVC/aorta ratio</td>
<td>37.5</td>
</tr>
<tr>
<td>Cardiac valve assessment</td>
<td>29.7</td>
</tr>
<tr>
<td>Identify regional wall motion abnormality</td>
<td>28.1</td>
</tr>
<tr>
<td>Aortic root assessment</td>
<td>23.4</td>
</tr>
<tr>
<td>Gastrointestinal applications</td>
<td></td>
</tr>
<tr>
<td>Identify small bowel obstruction</td>
<td>65.6</td>
</tr>
<tr>
<td>Identify pneumoperitoneum</td>
<td>46.9</td>
</tr>
<tr>
<td>Identify ileus</td>
<td>45.3</td>
</tr>
<tr>
<td>Identify gallbladder polyps</td>
<td>25.0</td>
</tr>
<tr>
<td>Identify gallbladder masses</td>
<td>15.6</td>
</tr>
<tr>
<td>Genitourinary applications</td>
<td></td>
</tr>
<tr>
<td>Assessment for renal stones</td>
<td>78.1</td>
</tr>
<tr>
<td>Identify testicular torsion</td>
<td>71.9</td>
</tr>
<tr>
<td>Identify renal cysts</td>
<td>64.1</td>
</tr>
<tr>
<td>Identify congenital renal abnormalities</td>
<td>59.4</td>
</tr>
<tr>
<td>Identify hydrocele</td>
<td>59.4</td>
</tr>
<tr>
<td>Fetal dating</td>
<td>54.7</td>
</tr>
<tr>
<td>Identify epididymoorchitis</td>
<td>54.7</td>
</tr>
<tr>
<td>Identify adnexal torsion</td>
<td>51.6</td>
</tr>
<tr>
<td>Assessment for scrotal abscess and cellulitis</td>
<td>51.6</td>
</tr>
<tr>
<td>Assessment for adnexal pathology (abscess, cysts)</td>
<td>48.4</td>
</tr>
<tr>
<td>Identify renal masses</td>
<td>45.3</td>
</tr>
<tr>
<td>Assessment for renal jets</td>
<td>45.3</td>
</tr>
<tr>
<td>Identify varicocele</td>
<td>45.3</td>
</tr>
<tr>
<td>Identify testicular masses</td>
<td>43.8</td>
</tr>
<tr>
<td>Identify retained products of conception</td>
<td>35.9</td>
</tr>
<tr>
<td>Assessment of renal parenchyma</td>
<td>28.1</td>
</tr>
<tr>
<td>Identify uterine masses or cysts</td>
<td>26.6</td>
</tr>
<tr>
<td>Identify testicular cysts</td>
<td>14.1</td>
</tr>
<tr>
<td>Renal Doppler assessment</td>
<td>7.8</td>
</tr>
<tr>
<td>Ocular applications</td>
<td></td>
</tr>
<tr>
<td>Identify retinal detachment</td>
<td>78.1</td>
</tr>
<tr>
<td>Identify ocular foreign body</td>
<td>70.3</td>
</tr>
<tr>
<td>Identify lens dislocation</td>
<td>62.5</td>
</tr>
<tr>
<td>Identify vitreous hemorrhage</td>
<td>60.9</td>
</tr>
<tr>
<td>Identify vitreous detachment</td>
<td>57.8</td>
</tr>
<tr>
<td>Pupillary assessment</td>
<td>45.3</td>
</tr>
<tr>
<td>Extraocular muscle assessment</td>
<td>31.3</td>
</tr>
<tr>
<td>Identify retrobulbar hematoma</td>
<td>31.2</td>
</tr>
<tr>
<td>Musculoskeletal applications</td>
<td></td>
</tr>
<tr>
<td>Identify necrotizing fasciitis</td>
<td>76.6</td>
</tr>
<tr>
<td>Identify skull fracture</td>
<td>75.0</td>
</tr>
<tr>
<td>Identify rib/sternal fracture</td>
<td>53.1</td>
</tr>
</tbody>
</table>

(Continued)

for POCUS are continually expanding and clinical applicability varies by hospital setting. Although many items included in the survey did not meet criteria for inclusion in a formal PEM US fellowship curriculum, this does not preclude individual programs from expanding their curriculum based on local disease patterns, clinical needs, and local expertise. Because POCUS is a rapidly expanding field, and new modalities and applications for POCUS frequently arise, the proposed curriculum will need ongoing reassessments to maintain consistency among training programs to accommodate changes in the uses of POCUS.

There were a number of items included as part of our survey that offer some guidance regarding duration of fellowship, scholarly work requirements, and need
for third-party certification eligibility. Although we achieved consensus that fellowship duration should be 1 year after completion of a PEM fellowship, no consensus was achieved regarding the possibility of a combined PEM/PEM POCUS fellowship. Although a 3-year PEM fellowship can be tailored to include many components of a PEM POCUS fellowship, whether an individual can complete all components of both PEM and PEM POCUS fellowships in a 3-year time frame, while still satisfying ACGME requirements for a PEM fellowship needs further evaluation.

LIMITATIONS

Our study has several limitations. First, it is possible that there were POCUS experts that were not included in our electronic mailing because they did not belong to the organizations from which we solicited our mailing list. Second, the survey tool we generated contained items that are current to the practice of POCUS in the ED setting. Other items specific to local practice variation may have been appropriate to consider as part of the modified Delphi process. Third, expert participants had a median of 6 years in practice since residency. While this makes for a relatively young group of experts, this is the current reality for PEM POCUS because training in its use is relatively recent. Finally, although the initial mailing list included a large number of experts, the response rate was 37.9%, which may have affected survey results.

Our decision to design this study as a modified Delphi, without providing more precise data on where the group was leaning after Round 1, may have affected the results of our study. In particular, for those skills that almost reached consensus (for example, identifying renal stones or identifying retinal detachment), knowing that the group was leaning toward agreement to include these items may have swayed respondents toward an “important/very important” response. However, because we were seeking more independent responses from participants, we did not feel that providing this data was warranted.

CONCLUSIONS

Through a modified Delphi process, we created a guideline for core content to be included in a pediatric emergency medicine point-of-care ultrasound fellowship curriculum that excludes previously published competencies for an ultrasound curriculum for pediatric emergency medicine fellows. National adoption and implementation of the core components in pediatric emergency medicine ultrasound fellowships will allow for standardization of education of our future pediatric emergency medicine point-of-care ultrasound leaders.

References


Appendix

P2 Network Point-of-care Ultrasound Fellowship Delphi Group

Alyssa Abo, MD, Division of Emergency Medicine, George Washington University School of Medicine and Health Sciences, Washington, DC
Alex Arroyo, MD, Division of Pediatric Emergency Medicine, Department of Emergency Medicine, Montefiore Medical Center, Brooklyn, NY
Kiyettha H. Alade, MD, Department of Pediatrics, Section of Emergency Medicine, Baylor College of Medicine, Houston, TX
Erika Constantine, MD, Division of Pediatric Emergency Medicine, Department of Emergency Medicine, Alpert Medical School of Brown University, Providence, RI
David O. Kessler, MD, MSc, Department of Emergency Medicine, Columbia University Vagelos College of Physicians and Surgeons, New York, NY
Charisse Kwan, MD, Division of Emergency Medicine, Department of Pediatrics, University of Toronto, Toronto, Ontario, Canada
Samuel Lam, MD, Department of Emergency Medicine, Sutter Medical Center, Sacramento, CA
Stephanie K. Leung, MD, Department of Pediatrics, Section of Emergency Medicine, Baylor College of Medicine, Houston, TX
Marla Levine, MD, Department of Pediatrics, University of Texas at Austin, Dell Medical School, Dell Children’s Medical Center of Central Texas, Austin, TX
Resa E. Lewiss, MD, Department of Emergency Medicine, Thomas Jefferson University Hospital, Philadelphia, PA
Lorraine Ng, MD, Department of Emergency Medicine, Columbia University Vagelos College of Physicians and Surgeons, New York, NY
Allan Shefrin, MD, Department of Pediatrics and Emergency Medicine, University of Ottawa, Ottawa, Ontario, Canada
Adam B. Sivitz, MD, Department of Emergency Medicine, Newark Beth Israel Medical Center, Children’s Hospital of New Jersey, Newark, NJ
Amanda Toney, MD, Department of Emergency Medicine, Denver Health Medical Center, University of Colorado School of Medicine, Aurora, CO
Atim Uya, MD, Department of Pediatrics, University of California at San Diego, Rady Children’s Hospital, San Diego, CA
Fred Warkentine, MD, Department of Pediatrics, Division of Pediatric Emergency Medicine, University of Louisville/Norton Children’s Hospital, Louisville, KY
Emergency Department Thoracotomy: Development of a Reliable, Validated Checklist for Procedural Training

Hashim Q. Zaidi, MD, Sarah S. Dhake, MD, Danielle T. Miller, MD, Priyanka Sista, MD, Matthew J. Pirotte, MD, Abra L. Fant, MD, MS, and David H. Salzman, MD, MEd

ABSTRACT

Objectives: Emergency department thoracotomy (EDT) is a rare and challenging procedure. Emergency medicine (EM) residents have limited opportunities to perform the procedure in clinical or educational settings. Standardized, reliable, validated checklists do not exist to evaluate procedural competency. The objectives of this project were twofold: 1) to develop a checklist containing the critical actions for performing an EDT that can be used for future procedural skills training and 2) to evaluate the reliability and validity of the checklist for performing EDT.

Methods: After a literature review, a preliminary 22-item checklist was developed and disseminated to experts in EM and trauma surgery. A modified Delphi method was used to revise the checklist. To assess usability of the checklist, EM and trauma surgery faculty and residents were evaluated performing an EDT while inter-rater reliability was calculated with Cohen’s kappa. A Student’s t-test was used to compare the performance of participants who had or had not performed a thoracotomy in clinical practice. Item-total correlation was calculated for each checklist item to determine discriminatory ability.

Results: A final 22-item checklist was developed for EDT. The overall inter-rater reliability was strong (κ = 0.84) with individual item agreement ranging from moderate to strong (κ = 0.61 to 1.00). Experts (attending physicians and senior residents) performed well on the checklist, achieving an average score of 80% on the checklist. Participants who had performed EDT in clinical practice performed significantly better than those that had not, achieving an average of 80.7% items completed versus 52.3% (p < 0.05). Seventeen of 22 items had an item-total correlation greater than 0.2.

Conclusions: A final 22-item consensus-based checklist was developed for the EDT. Overall inter-rater reliability was strong. This checklist can be used in future studies to serve as a foundation for curriculum development around this important procedure.
Trauma is a leading cause of death for persons age 1 to 44.\textsuperscript{1} Penetrating trauma continues to represent a particularly lethal problem with a higher prehospital and emergency department (ED) mortality compared to blunt trauma.\textsuperscript{2,3} Carefully selected patients presenting to the ED with penetrating thoracic or abdominal trauma may benefit from an emergency department thoracotomy (EDT).\textsuperscript{4,5} When performing an EDT, the physician emergently enters the thoracic cavity with the goals of identifying and temporizing direct damage from a penetrating injury. Specifically, the physician may relieve cardiac tamponade, directly control hemorrhage, or cross-clamp the thoracic aorta to provide hemorrhage control and prioritize cardiac and cerebral perfusion. While the indications for this procedure have become more selective over the past several decades, it remains a critical and potentially lifesaving procedure within the scope of practice for trauma and emergency medicine (EM) physicians.\textsuperscript{5} EDT is an invasive, technically challenging, and resource-intensive process.\textsuperscript{5,7} It is often performed with little preparatory time and almost always performed by an EM physician or trauma surgeon as opposed to a thoracic surgeon. If performed incorrectly, the lifesaving potential of an EDT may be attenuated and there may be an increased risk of bloodborne pathogen exposure to providers. EDTs are rarely performed, and studies of EM residents show a lack of opportunities to develop competency in this procedure.\textsuperscript{8,9} In addition to a paucity of real-world experience, opportunities for deliberate practice are by most accounts nonexistent. Cadaveric models may be considered for procedural training; however, the expense of models and the need for repetitive deliberate practice to ensure competency make cadaveric models cost-prohibitive for education on a widespread basis.\textsuperscript{10} The other major obstacle with cadaveric models lies with the difficulty in repeatedly simulating several of the key components for which an emergent thoracotomy is performed, such as identifying and relieving cardiac tamponade, controlling hemorrhage of cardiac injuries, and repairing cardiac wounds. Even when practice opportunities for EDT arise in clinical or educational contexts, standardized, reliable, and valid checklists do not exist to ensure procedural competency. In contrast, other rare yet lifesaving procedures, such as cricothyrotomies, have valid and reliable tools for assessing performance.\textsuperscript{11} Standardized checklists have been shown to improve trainee performance for many common procedures.\textsuperscript{12,13} To our knowledge there is no validated checklist for performing an EDT. The objectives of this project were twofold: first, to develop a checklist containing the critical actions for performing an EDT that can be used for future procedural skills training, and second, to evaluate the reliability and validity of the checklist for EDT.

**METHODS**

**Study Design**

This is a prospective checklist creation and validation study. Multiple experts in EM and trauma surgery were recruited and a modified Delphi method was used for checklist creation.\textsuperscript{14,15} The checklist was then validated by an additional group of physicians who were observed performing the procedure by two raters. This study was reviewed by the institutional review board at Northwestern University Feinberg School of Medicine and deemed to be exempt.

**Study Setting and Population**

The checklist was developed and validated at an urban academic medical center in 2017.

**Study Protocol**

To identify items to include in the checklist, the authors conducted a review of literature and textbooks for relevant content to performing the procedure. The literature was searched using PubMed for a resuscitative or EDT. Search terms included “thoracotomy AND simulation OR simulator,” “thoracotomy AND curriculum,” “thoracotomy AND teaching OR instruction OR practice OR education,” and “thoracotomy AND residents/residency.” A total of 318 articles were identified in the initial literature search. Of these, 20 articles were selected for relevance as they contained information on the steps necessary to successfully perform an EDT. An additional resource included review of the “Resuscitative Thoracotomy” chapter of Roberts & Hedges’ *Clinical Procedures in Emergency Medicine & Acute Care, 7th Edition*.\textsuperscript{16} Based on this review, a preliminary 22-item binary-response checklist was developed by the authors HQZ, SNS, and DTM. The checklist was developed to contain steps starting with preparation of equipment through performing the procedure and stopping at open cardiac massage based on accepted intervenable injuries that are likely to yield the most favorable outcomes as described by guidelines published by major U.S.
A group of experts including six emergency physicians and two trauma surgeons were recruited. All experts were board certified in their respective fields. The group contained two academic trauma surgeons, five academic emergency physicians, and one emergency physician in community practice, all of whom had clinical experience performing EDT and training physicians in the procedure. Geographically, the experts practiced clinically across the United States, six in the Midwest, one in the South, and one in the West. The experts were initially asked to review the preliminary checklist developed from the literature review described above. Specifically, they were asked whether or not the checklist reflected the critical steps for performing EDT. Subsequently, they were asked to comment on whether or not the checklist needed additions or removal of steps to be a complete description of how to perform an EDT. Responses were compiled, and a revised checklist was again distributed. This process was repeated and a total of three rounds of revision resulted in consensus among the experts regarding the critical steps of the EDT.

Once final consensus from the expert panel was achieved, the checklist was reviewed for usability. One of the authors (HQZ) performed an EDT on the simulator while two additional authors (DHS and SNS) observed the procedure, followed along on the checklist, and made comments to further clarify the checklist and to confirm the steps were sufficiently described and able to be scored appropriately by the evaluator. This final step resulted in changes neither to the items on the checklist nor to the order of steps, but did result in clarification of descriptions of a correct performance of an item.

After the checklist was finalized, nine physicians who were not involved in the development of the checklist were recruited to perform an EDT on a simulated model. They included six surgical residents (PGY-4 to -7), two EM attending physicians, and one attending trauma surgeon. None of these physicians were involved in the creation of the checklist or had prior access to the checklist. The simulated model was designed and built at Northwestern Simulation and allowed for all of the critical steps of the procedure to be performed and observed (Figure 1).

Upon arrival to the simulation center, the purpose of the procedural simulation was explained to the participants. They were provided a brief clinical scenario describing a patient with a single thoracic penetrating gunshot wound with loss of vital signs just prior to arrival to the ED. They were instructed that an EDT was indicated and were asked to perform the procedure. The participants were informed that the patient would be arriving in 3 minutes. During that time, each individual prepared and gathered any equipment necessary to perform an EDT. Typical personal protective equipment (PPE) and thoracotomy equipment was made available. The group was instructed to verbalize what items they needed and correctly identify them in the thoracotomy tray. After either 3 minutes had passed or the individual had completed all correct preparatory items, the evaluator indicated that the patient had arrived and the participant could proceed to perform the procedure.

Measures

Performance of the EDT was observed in real time by two raters (DHS, ALF) who independently scored each participant. These two raters were not a part of the expert panel who developed the initial checklist content. Each step was scored by the raters as being performed “correct” or “incorrect.” The raters were positioned adjacent to the simulator and participant for the duration of the procedure. Additionally, cameras were positioned above and behind the participants so raters were also able to observe from additional angles, after their real-time evaluation, if they deemed they needed additional views to complete their evaluation.

Data Analysis

Reliability, as measured by inter-rater reliability of the two observers, was evaluated with Cohen’s kappa coefficient for each checklist item and the checklist as a whole. The confidence interval was set at 95%. All statistical analyses were performed using Microsoft Excel v14.7. Descriptive statistics were completed on participants’ performance on the checklist. Participants were then grouped according to experience by years of training and by self-reported completion of an EDT in clinical practice. An independent Student’s t-test was used to compare the performance between these groups to determine concurrent validity. Finally, an item-total correlation was calculated to determine the ability of each checklist item to discriminate between...
RESULTS

A final 22-item checklist was developed for the EDT (see Data Supplement S1, available as supporting information in the online version of this paper, which is available at http://onlinelibrary.wiley.com/doi/10.1002/aet2.10387/full). Overall inter-rater reliability was strong, with a Cohen’s kappa coefficient of 0.84. Individual item results are detailed in Table 1, with individual results ranging from moderate to strong ($\kappa = 0.61$ to 1.00), with the exception of items 8 (extend incision) and 10 (insert spreader), which had minimal agreement ($\kappa = 0.00$), and item 9 (manually spread ribs), which had minimal disagreement ($\kappa = -0.29$).

Attending physicians ($n = 3$) completed an average of 73.9% of checklist items correctly when their performance on the EDT was scored via the checklist. Senior surgery residents (PGY-6 and above; $n = 1$) and attendings ($n = 3$) completed an average of 79.5% of items correctly. The cohort of participants who had completed at least one EDT in clinical practice ($n = 7$) completed an average of 80.7% of items correctly. Those participants who had not completed a real EDT scored an average of 52.3% of items correctly ($n = 2$). A Student’s $t$-test was used to compare these groups (Table 2). There was no significant difference between attendings and residents ($n = 3$, $n = 6$; $T = 0.28$, $p = 0.78$) or between senior residents/attendings and junior residents ($n = 4$, $n = 5$; $T = 0.057$, $p = 0.54$). However, when comparing those who had not done a thoracotomy in clinical practice to those who had, a significant difference in checklist item completion was found ($n = 7$, $n = 2$; $T = 3.47$, $p < 0.05$). Item discrimination statistics are shown in Table 3. Item-total correlation was less than 0.2 for only five of the 22 items (Table 3).

DISCUSSION

Through the process of literature review, expert consensus using a modified Delphi method, pilot testing, and refinement, we have developed a checklist for the experienced and inexperienced participants as a measure of convergent validity.

Figure 1. Thoracotomy simulation trainer with EDT steps being performed. (A) Incision of the chest wall. (B) Incision of the pericardium. (C) Cross-clamping of the aorta. (D) Open cardiac massage being performed. EDT = ED thoracotomy.
performance of an EDT. When the checklist was presented to the experts, agreement was high on a majority of the items in the initial rounds. However, there were a few items which required additional discussion regarding three domains of EDT performance: preferred thoracotomy equipment, methods to temporize cardiac wounds, and approaches to identify and cross-clamp the aorta. These differences were hypothesized to be related to practice variation. Experts had variable responses on their preferred thoracotomy equipment. For example, multiple tools were recommended to cut through bone. To achieve consensus, we included a variety of potential equipment with the emphasis on making sure the provider had some method to cut through bone if requested (item 2). There was also variation regarding the approach for the management of cardiac wounds (item 18). Some responses were consistent with established literature and expert agreement to use sutures with pledgets.\textsuperscript{16,19} However, our experts also suggested other methods of repair, such as staples or Foley catheter insertion, with the latter being amenable to transfusion directly into the cardiac chamber (item 18). The final item requiring additional discussion involved cross-clamping the aorta, specifically with regards to identifying the difference between the aorta and esophagus. Although experts agreed that the aorta and esophagus must be distinguished to cross-clamp (item 19), their preferred methods varied. For these items without consensus evidence and significant clinical equipoise, we opted to include all suggested possible approaches in the checklist. As a result, multiple approaches for temporizing cardiac wounds were included as acceptable options in the checklist. Both of the suggested approaches for identification of the aorta including placing a nasogastric tube or utilizing anatomic and tactile differences were included as acceptable for the checklist item.

When using the checklist in a simulated procedural scenario, raters found the checklist easy to use, and overall inter-rater reliability was strong (Table 1). The majority of individual items on the checklist also had a strong kappa coefficient. The items that had lower inter-rater agreement included extending the incision (item 8), manually spreading the ribs (item 9), and inserting the rib spreader correctly (item 10). There are several factors that likely resulted in this low inter-rater agreement. First, several of the items may have been more subjective than initially anticipated. This may have been most apparent in the scoring of item 9–manually spreading the ribs and item 8–extending the incision. Second, while both raters were adjacent to the EDT trainer and the participant, it is possible that subtle differences in viewpoints while watching the procedure may have impacted the scoring.

While these steps did show lower inter-rater reliability, during the process of checklist development, they were identified as critical steps according to expert consensus and review of literature on performing an emergent thoracotomy and were therefore not

<table>
<thead>
<tr>
<th>Item</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Don PPE</td>
<td>1.00</td>
</tr>
<tr>
<td>2. Gather equipment</td>
<td>1.00</td>
</tr>
<tr>
<td>3. Check equipment</td>
<td>0.77</td>
</tr>
<tr>
<td>4. Assemble spreader</td>
<td>1.00</td>
</tr>
<tr>
<td>5. Position patient</td>
<td>1.00</td>
</tr>
<tr>
<td>6. Prepare chest</td>
<td>1.00</td>
</tr>
<tr>
<td>7. Incision</td>
<td>1.00</td>
</tr>
<tr>
<td>8. Extend incision</td>
<td>0.00</td>
</tr>
<tr>
<td>9. Manually spread ribs</td>
<td>-0.29</td>
</tr>
<tr>
<td>10. Insert spreader</td>
<td>0.00</td>
</tr>
<tr>
<td>11. Open spreader</td>
<td>1.00</td>
</tr>
<tr>
<td>12. Identify heart</td>
<td>1.00</td>
</tr>
<tr>
<td>13. Identify phrenic nerve</td>
<td>1.00</td>
</tr>
<tr>
<td>14. Lift pericardium</td>
<td>1.00</td>
</tr>
<tr>
<td>15. Incise pericardium</td>
<td>0.61</td>
</tr>
<tr>
<td>16. Deliver heart</td>
<td>1.00</td>
</tr>
<tr>
<td>17. Identify cardiac injury</td>
<td>1.00</td>
</tr>
<tr>
<td>18. Control cardiac hemorrhage</td>
<td>1.00</td>
</tr>
<tr>
<td>19. Identify aorta</td>
<td>1.00</td>
</tr>
<tr>
<td>20. Cross-clamp aorta</td>
<td>1.00</td>
</tr>
<tr>
<td>21. Cardiac massage</td>
<td>0.61</td>
</tr>
<tr>
<td>22. Maintain sterility</td>
<td>0.77</td>
</tr>
<tr>
<td>Total</td>
<td>0.84</td>
</tr>
</tbody>
</table>

PPE = personal protective equipment.

<table>
<thead>
<tr>
<th>Cohort (A vs. B)</th>
<th>A</th>
<th>B</th>
<th>t-test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attending physician (n = 3) vs. all residents (n = 6)</td>
<td>73.9%</td>
<td>71.4%</td>
<td>0.28</td>
<td>0.78</td>
</tr>
<tr>
<td>EDT experience (n = 7) vs. no EDT experience (n = 2)</td>
<td>80.7%</td>
<td>52.2%</td>
<td>3.47</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

EDT = ED thoracotomy; PPE = personal protective equipment.
removed. We hypothesize the lower correlation to be due to visual limitations of the model when having two raters watching the same technical procedure. Further clarification and delineation of how to score these items correctly could improve agreement. Prompting the participant to more routinely verbalize the actions during the procedure could minimize this discrepancy. These items will be highlighted in subsequent rater training for procedure evaluations to reduce discrepancy and maximize visibility of performance. We are actively reevaluating with a larger sample size to determine if inter-rater reliability can be improved for these few items.

Despite the few items mentioned, overall strong inter-rater agreement of the two observers over multiple tests demonstrates reliability of the checklist.\textsuperscript{18} While demonstrating validity with clinical thoracotomies in real time is impractical due to the rarity and nature of the procedure, we are able to demonstrate concurrent validity by the significant performance difference of physicians who had completed an EDT in real life versus those who had not. In other words, the checklist can differentiate between people with experience in performing an EDT versus those who have not. This may translate to assessing whether a person has acquired sufficient skills to be considered proficient at performing an EDT after training. The checklist’s convergent validity is demonstrated by the item-total correlation. A value greater than 0.2 for an item suggests that there is good discriminatory ability for that item, which was the case for 17 of the 22 items in this checklist.\textsuperscript{20,21} Good discriminatory ability means that someone who performs well overall on the checklist would likely score “correct” on a given item.

Although there were five items with poor discriminatory ability, they were kept in the checklist due to their consensus through the modified Delphi as necessary steps of the procedure. Several of these steps, such as donning PPE or identifying the heart, are quite fundamental and remain a critical part of the procedure. Interestingly, identifying and cross-clamping the aorta was negatively discriminatory, meaning that the novices were more likely to perform this step compared to experts. Several possible explanations could account for this finding. These include the small sample size as well as an impact of previous clinical experience. It is also possible that the sample of novices may have approached the procedure in an algorithmic fashion, whereas experts did not perform this step as they may have integrated previous clinical experience and believe cross-clamping the aorta would be ineffective based on the visualized cardiac injury.

With the emergence of endovascular procedures such as resuscitative endovascular balloon occlusion of the aorta (REBOA), one might argue that the role of EDT may become less important. REBOA is a novel tool to help obtain hemorrhagic control of a noncompressible subdiaphragmatic injury in a trauma patient with profound shock\textsuperscript{22} and overlaps with the EDT step of cross-clamping the aorta. However, for cardiac or thoracic injuries particularly with the suspicion for cardiac tamponade, myocardial injury, or other thoracic vascular injury resulting in cardiac arrest, the resuscitative EDT remains the standard of care. Any suspected thoracic injury is a contraindication for the use of endovascular balloon therapy.\textsuperscript{23} In addition, recent joint statements American College of Surgeons Committee on Trauma and the American College of Emergency Physicians have recommended that EM

\begin{table}[h]
\centering
\caption{Discriminant Ability and Item Discrimination of a 22-item EDT Checklist}
\begin{tabular}{|c|c|c|c|c|c|}
\hline
\textbf{Item} & \textbf{Has Performed EDT (Mean Percent Correct\textsuperscript{1})} & \textbf{Item-total Correlation} \\
\hline
1. Don PPE & 100.0 & 100.0 & 0.000 & \\
2. Gather equipment & 50.0 & 50.0 & 0.160 & \\
3. Check equipment & 0.0 & 58.3 & 0.297 & \\
4. Assemble spreader & 0.0 & 33.3 & 0.377 & \\
5. Position patient & 50.0 & 100.0 & 0.745 & \\
6. Prepare chest & 50.0 & 66.7 & 0.240 & \\
7. Incision & 0.0 & 83.3 & 0.883 & \\
8. Extend incision & 75.0 & 100.0 & 0.355 & \\
9. Manually spread ribs & 50.0 & 58.3 & 0.231 & \\
10. Insert spreader & 75.0 & 100.0 & 0.355 & \\
11. Open spreader & 50.0 & 100.0 & 0.404 & \\
12. Identify heart & 100.0 & 100.0 & 0.000 & \\
13. Identify phrenic nerve & 50.0 & 100.0 & 0.745 & \\
14. Lift pericardium & 50.0 & 83.3 & 0.224 & \\
15. Incise pericardium & 50.0 & 91.7 & 0.607 & \\
16. Deliver heart & 50.0 & 66.7 & 0.543 & \\
17. Identify cardiac injury & 50.0 & 100.0 & 0.745 & \\
18. Control cardiac hemorrhage & 50.0 & 100.0 & 0.745 & \\
19. Identify aorta & 100.0 & 66.7 & -0.420 & \\
20. Cross-clamp aorta & 100.0 & 83.3 & -0.278 & \\
21. Cardiac massage & 50.0 & 91.7 & 0.702 & \\
22. Maintain sterility & 0.0 & 41.7 & 0.590 & \\
\hline
\textbf{Total} & 52.7 & 80.7 & & \\
\hline
\end{tabular}
\textsuperscript{EDT = ED thoracotomy; PPE = personal protective equipment.}
\textsuperscript{1}Mean percentage of items performed correctly for both raters combined.
\end{table}
physicians without critical care training should not perform REBOA. At this time there does not appear to be an alternative for the patient suffering from traumatic arrest with suspected thoracic injuries amenable to an EDT. Education and training for emergency physicians in this rare but critical procedure will need to continue until a suitable alternative is demonstrated.

LIMITATIONS

There are several limitations to our study. While we recruited experts from multiple geographic areas and types of clinical environments to participate in the checklist development process, the overall sample size was small and the majority of our experts were from urban academic institutions, which may limit the generalizability of this checklist. However, we believe the impact of this to be minimal as performing an EDT is generally agreed upon to be a procedure occurring in trauma centers. Differences in resources between community, rural, and urban environments may affect the application of the checklist in other settings. The development of the checklist may have also been limited by the content expertise of our panel. All of our experts were physicians and having members trained in human factors design may have added additional rigor to the checklist.

The inter-rater reliability of our checklist and the concurrent validity was assessed with a small sample size. A larger sample size would have provided greater power to more precisely understand the instrument’s characteristics. For the few items that had lower inter-rater agreement, this may reflect a limitation based on the vantage of the rater, the checklist itself, or the model utilized. Further study into these items in particular with a larger sample size and using nonresearcher evaluators may help to assess and delineate the source of the lower inter-rater agreement.

The checklist development occurred in the context of asking the expert panel to identify key steps in performing this critical procedure, the validation of the checklist, and the assessment of inter-rater reliability was performed using a simulated model. It is possible that the model used possessed the potential limitation of overlooking key critical actions or steps in the procedure that would only be uncovered or revealed in testing the checklist on model with better fidelity characteristics such as a cadaver or human patient. However, given that cadaveric models may be cost-prohibitive for training on a widespread basis, the infrequent yet critically important nature of being able to correctly perform the procedure, and the potential application to combat training, we believe that this checklist fills the longempty niche of a standardized tool for the assessment of competency in EDT. Future studies can assess the transferability and performance of this expert consensus derived checklist to the completion of an EDT on a cadaver or other simulated models.

Finally, this checklist is based on a hypothetical scenario of a gunshot wound leading to a single anterior ventricular injury that was simulated using a thoracotomy model. While this is a valid indication for an EDT, we did not specifically seek to create a checklist for a particular injury but rather for general performance of EDT. A variety of injuries could be discovered upon gaining surgical access to the left hemithorax. Given this phase of the project focused on developing a checklist for the performance of an EDT by an emergency physician, we focused the scenario on one that combines the most common indication with a repairable injury that would be most likely to yield a favorable outcome. There may be items on the checklist that are more critical or less critical depending on the specific injury leading to the decision to perform EDT. To further validate the checklist, it must be applied to a broader range of simulated scenarios and clinical practice.

CONCLUSION

We describe the development of a valid, reliable checklist for performing an ED thoracotomy. Given the critical nature of this procedure with limited opportunities for practice, we believe that the creation of this validated checklist can serve as the foundation for curricular development. Future studies should be conducted to further validate this checklist across a more diverse range of practice environments, with additional types of scenarios and, ultimately, in the clinical environment. It is our goal to improve education for an ED thoracotomy to further procedural training and competence with the ultimate aim of improved patient outcomes.

References


Supporting Information

The following supporting information is available in the online version of this paper available at http://onlinelibrary.wiley.com/doi/10.1002/aet2.10387/full

Data Supplement S1. EDT Checklist.
An Event-based Approach to Measurement: Facilitating Observational Measurement in Highly Variable Clinical Settings

Rosemarie Fernandez, MD1, Elizabeth D. Rosenman, MD2, Sarah Brolliar2, Anne K. Chipman, MD, MS2, Colleen Kalynych, MSH, EdD3, Marie C. Vrablik, MD, MCR2, Joseph R. Keebler, PhD4, and Elizabeth H. Lazzara, PhD4

ABSTRACT

Background: Translational research in medical education requires the ability to rigorously measure learner performance in actual clinical settings; however, current measurement systems cannot accommodate the variability inherent in many patient care environments. This is especially problematic in emergency medicine, where patients represent a wide spectrum of severity for a single clinical presentation. Our objective is to describe and implement EBAM, an event-based approach to measurement that can be applied to actual emergency medicine clinical events.

Methods: We used a four-step event-based approach to create an emergency department trauma resuscitation patient care measure. We applied the measure to a database of 360 actual trauma resuscitations recorded in a Level I trauma center using trained raters. A subset (n = 50) of videos was independently rated in duplicate to determine inter-rater reliability. Descriptive analyses were performed to describe characteristics of resuscitation events and Cohen’s kappa was used to calculate reliability.

Results: The methodology created a metric containing both universal items that are applied to all trauma resuscitation events and conditional items that only apply in certain situations. For clinical trauma events, injury severity scores ranged from 1 to 75 with a mean (±SD) of 21 (±15) and included both blunt (254/360; 74%) and penetrating (86/360; 25%) traumatic injuries, demonstrating the diverse nature of the clinical encounters. The mean (±SD) Cohen’s kappa for patient care items was 0.7 (±0.3).

From the 1Department of Emergency Medicine and the Center for Experiential Learning and Simulation, College of Medicine, University of Florida, Gainesville, FL; the 2Department of Emergency Medicine, University of Washington, Seattle, WA; the 3Department of Emergency Medicine, Office of Educational Affairs, University of Florida College of Medicine–Jacksonville, Jacksonville, FL; and the 4Department of Human Factors and Behavioral Neurobiology, College of Arts and Sciences, Embry-Riddle Aeronautical University, Daytona Beach, FL.

Received June 2, 2019; revision received September 9, 2019; accepted September 11, 2019.

Funding and support for this project was provided by the Agency for Healthcare Research and Quality (1R18HS022458-01A1 [RF]) and the Department of Defense Congressionally Directed Medical Research Program (W81XWH1810089 [RF]). The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; or preparation, approval, or decision to submit the manuscript.

Conflict of interest: RF’s institution has received grant money from the Agency for Healthcare Research and Quality, the Department of Defense, and the Washington State Department of Labor and Industries to conduct research conceived and written by RF; RF reports personal payment from Physio-Control, Inc. for speaker fees. EDR’s institution has received grant money from the Agency for Healthcare Research and Quality and the Department of Defense to conduct research conceived and written by RF. EDR reports personal payment from Physio-Control, Inc. for speaker fees. SB’s institution has received grant money from the Agency for Healthcare Research and Quality and the Department of Defense to conduct research conceived and written by RF. AKC reports no conflict of interest. CK’s institution has received grant money from the Agency for Healthcare Research and Quality and the Department of Defense to conduct research conceived and written by RF. JRK reports no conflict of interest. EHL reports no conflict of interest. MCV reports no conflict of interest. Author contributions: concept and design—RF, EDR, JRK, and EHL; acquisition, analysis, or interpretation of data—RF, EDR, SB, CK, AKC, and MCV; drafting of manuscript—RF and EDR; critical revision of the manuscript for important intellectual content—RF, EDR, CK, AKC, MCV, SB, JRK, and EHL; statistical expertise—JRK and EHL; and obtained funding—RF.

Supervising Editor: Sally Santen, MD, PhD.

Address for correspondence and reprints: Rosemarie Fernandez MD; e-mail: fernandez.r@ufl.edu.


© 2019 by the Society for Academic Emergency Medicine doi: 10.1002/aet2.10395 ISSN 2472-5390
Conclusion: We present an event-based approach to performance assessment that may address a major gap in translational education research. Our work centered on assessment of patient care behaviors during trauma resuscitation. More work is needed to evaluate this approach across a diverse array of clinical events.

Medical education research aims to improve physician training, thus making the delivery of health care safer and more effective. As noted by McGaghie, the “downstream goals of medical education research are to demonstrate that educational interventions contribute to physician competence measured in the classroom, educational laboratory, and patient care setting.” This suggests an important role for translational medical education research, where testing of interventions often begins with knowledge assessments and performance in simulated environments (T1) and advances to assessments of actual clinical performance (T2) before demonstrating impact on patient outcomes (T3).

There is a gap in evaluating the impact of educational interventions on clinical performance. Most commonly, assessments target knowledge and skill acquisition, without evaluating whether these skills translate into behavioral change and improved clinical care. Measuring individual skills, trained behaviors, and team performance in the clinical setting are key components of effective translational research in medical education; however, the development of metrics remains a limitation. It is challenging to create a behavioral checklist that is reliable, can discriminate different levels of performance, and can accommodate the variability encountered in the actual clinical environment. Measures that perform well in the simulated setting do not necessarily translate to clinical use. Measures intended for clinical care may be too general to discriminate performance quality or may rely too heavily on subject matter expertise for accurate scoring.

Highly variable tasks, such as those commonly present in emergency medicine, pose further challenges to measuring learner performance during actual clinical events. In the emergency department (ED), patients represent a wide range of complexity and potential clinical instability within a single diagnosis. In such situations, variability in the patient’s condition and the clinical environment significantly impacts which behaviors are indicated and in what order. We present an event-based approach to measurement (EBAM) that allows for the measurement of complex behavioral processes across highly variable clinical events using trauma team resuscitations as an example.

METHODS

EBAM

Event-based approach to measurement adapts the concepts used in event-based training for clinical observation-based measurement design. Event-based training follows a conceptual design process that supports the

Figure 1. The event-based approach to measurement (EBAM) process.
systematic introduction of work-related tasks to purposefully elicit the behaviors or competencies that are the focus of training interventions.\(^9\) We translated this concept to develop EBAM, a four-step process (Figure 1) that allows for the development of measures that are 1) modifiable for different patient conditions and tasks, 2) reliable and supported by validity evidence,\(^{10}\) and 3) directly linked to evidence-based practice. As with event-based training, the core content of EBAM is based on identification of clinical events and triggers. In event-based training the triggers are predefined and controlled, which ensures that the learner has an opportunity to reach all learning objectives. In EBAM, the triggers are dependent on clinical factors (e.g., patient condition, team behaviors), which ensures the assessment only includes pertinent items. This can be applied to individual or team-level performance depending on the research target.

We applied EBAM for the purpose of evaluating team-level clinical care during trauma resuscitations. While Advanced Trauma Life Support (ATLS) provides general guidelines for trauma care, the execution of trauma resuscitation is much more complex and dependent on highly variable patient factors. There are universal events that contain behaviors that should be performed for every trauma regardless of the patient, the etiology of the trauma, or environmental factors. However, there are other conditional behaviors that depend on the patient’s state, for example, the presence of shock physiology. A “shock” event would contain behavioral responses expected for a patient in clinical shock, but would not be expected when caring for a hemodynamically stable patient. Together, the universal and shock behavioral responses then become measurement items that can flex to accommodate both hemodynamically stable and unstable patients. In Figure 2, we provide an example that illustrates how EBAM can handle the variable need for procedural tasks during trauma resuscitations.

**Figure 2.** Example of event-based measurement system (event + trigger + measurement item) In this example, there are several triggers (T1 – T5), either alone or in combination, that should prompt the behavioral responses and thus measurement items, listed in the event Procedure: Thoracostomy. Initiating the Procedure: Thoracostomy also activates a second, related event: Procedure: Universal. This event contains items universally used during major resuscitation-related procedures. It would be expected that other procedures, such as central venous catheter placement would also trigger the Procedure: Universal event. Abbreviations: CXR = chest x-ray; CVC = central venous catheter; PTX = pneumothorax; E-FAST = extended focused assessment with sonography in trauma; SBP = systolic blood pressure.
Study Design and Setting

Trauma resuscitation events (n = 360) were recorded at Harborview Medical Center, an urban, Level I trauma center in the University of Washington health care system. The University of Washington Human Subjects Division approved this study.

Trauma Resuscitation Events

Eligible resuscitations included all adult patients presenting to the ED meeting American College of Surgeons recommendations and Harborview Medical Center institutional criteria for trauma evaluation.11,12 Prisoners and pregnant women were excluded. Patient characteristics were obtained from the Harborview Medical Center trauma registry.

Trauma Team Patient Care Measure Development

Step 1: Incorporate Evidence-based Practice. Step 1 focused on facilitating the inclusion of evidence-based practices for trauma management during measure development. We identified published adult trauma patient care checklists that have been applied to both simulated13 and live14–17 patient care events, as well as standards of trauma care, including ATLS.18 Subject matter experts (SMEs), four board-certified emergency medicine physicians and one clinical nurse, reviewed the guidelines and assessment measures to identify items that 1) were appropriate for trauma resuscitations, 2) apply across all types and severity of trauma resuscitations (universal items), and 3) were indicated in certain clinical presentations but were not universally relevant (conditional items). For conditional items the SMEs determined under what conditions they should be performed. Finally, SMEs noted which items were time-sensitive. All behavioral and time-based items, universal and conditional, were included in a single patient care behavioral measure.

Step 2: Define Events. In Step 2, we defined and described trauma resuscitation behavioral events. Universal items from Step 1 were grouped into events expected to occur during every trauma resuscitation, such as “performs a primary survey.” The conditional items were also placed into events as appropriate. For example, the event “shock, presumed hemorrhagic” included items like “initiates blood transfusion.” SMEs reviewed all universal and conditional events and items to ensure that the items were appropriately assigned.

Step 3: Define Event Triggers. Step 3 focused on defining the clinical triggers that prompt the conditional events. If an event is “triggered,” then behaviors in that event should be included in the final performance assessment. The investigators worked with SMEs to identify observable clinical triggers. We prioritized triggers that were as specific as possible. For example, the item “initiates blood transfusion,” as mentioned above, is within the event “shock, presumed hemorrhagic” and is triggered by two or more low blood pressures (systolic blood pressure < 90 mm Hg). Other possible triggers, such as elevated heart rate, may occur for multiple reasons (e.g., pain) and thus were not considered specific enough to serve as a trigger for initiating a blood transfusion. We also had to limit triggers to items that could be reliability observed on video. A low hematocrit may prompt a blood transfusion; however, this laboratory value was not easily observable. All event triggers were reviewed by SMEs after they had observed 10 video-recorded resuscitations to ensure that triggers met the criteria of being 1) observable, 2) clinically appropriate, and 3) specific.

Step 4: Test and Refine Measure. In Step 4 the measure was tested to ensure that proposed behaviors, events, and triggers could be reliably observed. Preliminary determination of reliability during measure development informed ongoing rater training and prompted modifications to the EBAM items that were confusing, poorly defined, or assigned to an inappropriate event. Initial ratings were performed as a group with a “think-aloud” approach to ensure that important behavioral items and events were captured to the best of our ability.19

Steps 1 through 4 resulted in a list of trauma resuscitation patient care items (Data Supplement S1, Table S1, available as supporting information in the online version of this paper, which is available at http://onlinelibrary.wiley.com/doi/10.1002/aet2.10395/full) that included universal and conditional behaviors, behavioral triggers, and the time to key behaviors. The involvement of SMEs and clinical guidelines (ATLS) helped to establish evidence of content validity. These items were used to code recorded trauma resuscitations. Trigger variables were then used during data analysis to identify which items were appropriate for each individual trauma resuscitation. The total possible points (denominator) ranged from 20 to 38 based on the number of conditional items deemed appropriate for that particular patient resuscitation.
**Data Coding**

Coding was performed using Noldus Observer® XT (Leesburg, VA) software. Primary coders were part-time research assistants who also worked as hospital volunteers in high-acuity patient care settings (ED and pediatric intensive care unit) but who were not health care providers. A subset of resuscitations was coded in duplicate by a board-certified emergency physician. Additionally, board-certified emergency medicine physicians coded in duplicate all items requiring clinical judgment. The overall coding approach maximized the use of nonclinical experts and limited the coding burden for the physician investigators by using nonclinical coders for all scored items and for the triggers that did not require clinical judgment. An example of an item requiring clinical judgment was “vasopressors indicated,” which triggered the conditional item “team orders vasopressor infusion.” An example of a trigger not requiring clinical judgment was “two consecutive low blood pressures,” which triggered the conditional item “team initiates blood transfusion.” Additional examples are provided in Data Supplement S1, Table S1. All coders trained until they reached a Cohen’s $\kappa > 0.75$ across a range of performance episodes that varied with regard to complexity, illness severity, and patient care requirements. Rater training represents evidence of response process validity. To determine inter-rater reliability, 14% of resuscitations (50/360) were coded in duplicate by one of the investigators (LR, RF). Disagreements were reviewed by the coders and reconciled. Inter-rater reliability represents evidence of internal structure validity.

**Data Analysis**

Statistical analyses were performed using SPSS Statistics version 19 (IBM Corp.) and the open-source statistical program R version 3.5. We calculated Cohen’s kappa to determine the degree of inter-rater reliability on trauma resuscitation patient care measure items. We computed descriptive statistics (mean and standard deviation [SD]) for all resuscitation patient characteristics.

**RESULTS**

From March 2016 through February 2018, we recorded 360 trauma resuscitation events and were able to match 342 with descriptive characteristics in the Harborview Medical Center trauma registry. All 360 resuscitations were coded for trauma resuscitation patient care items. Overall inter-rater reliability across all items was Cohen’s $\kappa = 0.7$ (0.3). Characteristics of the trauma resuscitations are described in Data Supplement S1, Table S2, and demonstrate a wide range of clinical events, with injury severity scores ranging from 1 to 75 with a mean ($\pm$SD) of 21 ($\pm$5).

**DISCUSSION**

In this article we present an approach to observational measurement for research that captures granular metrics while accommodating high levels of clinical variability. Such granularity can be important when assessing behaviors and actions in medical education–based clinical trials. It is important that specific aspects of training can be directly linked with specific behaviors. Current assessment tools meant for general resident assessment (e.g., direct observation of procedural skills [DOPS] and mini-Clinical Evaluation Exercise [mini-CEX]) capture global performance but cannot fully explicate changes in performance at a highly detailed level.

The example used in this article focuses on assessment of trauma resuscitative care. However, EBAM may be applicable to other topics, including other areas of team-based clinical care (e.g., medical resuscitations), procedures, and interpersonal skills (e.g., managing conflict or delivering bad news). The strength of EBAM is its ability to handle patient variability in a systematic way. Thus, undifferentiated or unstable patients (e.g., rapid response clinical events), rapidly changing clinical environments (e.g., disaster settings or low-resourced settings), and highly complex patient care events are well suited to EBAM. While we evaluated team performance, we believe that this same approach could be used for individual assessments. The ability to use event-based training for simulation-based research and then transform this work into EBAM for clinical observation provides a mechanism for translational research in graduate medical education and interdisciplinary training.

**Limitations**

While our approach addresses some of the challenges in observational measurement, there are important limitations to recognize. We presented reliability for the entire measure and not at the item level. This was done because there was considerable variability regarding the number of times certain items were indicated and kappa is sensitive to prevalence. Reliability for rarely occurring items may not be accurately reflected. We were able to use video recordings of trauma resuscitations, thus enabling coders to watch the same behaviors...
multiple times to help ensure accurate coding. EBAM could be applied to live observations as well; however, one might have to sacrifice the detail necessary for high-level research, especially in chaotic environments such as resuscitations. Further research would be needed to determine how the methodology could be best structured to assess live events. Additionally, both event-based training and EBAM focus on observable behaviors. Thus, to accurately measure cognitive skills, there must be acceptable behavioral proxies that accurately represent the cognitive construct of interest. Finally, EBAM is resource-intensive. To adapt to patient variability, the measure must consider a number of conditional events and triggers and how these connect. As a result the test/refine process can be more time-consuming than what is needed with simulation-based measure development. While we acknowledge this resource burden, we feel EBAM provides a mechanism for the highly detailed measurement necessary for research, thus justifying the extra effort.

**CONCLUSIONS**

In conclusion, we describe an approach to behavioral assessment that addresses a major barrier to medical education–based translational research. Additional psychometric testing is needed to evaluate the application of this technique to other clinical events.

**References**


Supporting Information

The following supporting information is available in the online version of this paper available at http://onlinelibrary.wiley.com/doi/10.1002/aet2.10395/full

Data Supplement S1. Supplemental material.
The Design and Implementation of a Professional Development Program for Physician Assistants in an Academic Emergency Department

Derek L. Monette, MD, Brian Baccari, PA-C, Laura Paskind, PA-C, David Reisman, MHA, and Elizabeth S. Temin, MD, MPH

ABSTRACT
Physician assistants (PAs) are expanding their role in academic emergency departments (EDs). There are no published models for how to integrate PAs into departmental educational activities, scholarship, and operations outside of a PA residency approach. We created a professional development program for PAs that would provide them with opportunities to integrate into all aspects of our department mission and provide them with a forum for personal growth and ongoing education. The program provides PAs with resources including protected time and mentorship to become a content expert in an academic area of interest. We review our 5-year experience creating and implementing this program, which has grown from six PAs in 2013 to 24 PAs in 2018. These PAs now have formal roles in five of our eight divisions, participating in education, administrative, and research activities. The retention rate for PAs in this program is 90.2% versus 85.7% for PAs at our department who are not in the program. Our experience and results demonstrate the value of investing in the professional development and continued education of PAs at an academic ED versus the traditional model of service and the potential for integration into all aspects of an academic ED’s mission.

BACKGROUND
Physician assistants (PAs) have expanded their presence in emergency departments (EDs) at academic medical centers (AMCs) in parallel with increasing ED visits nationally. Most AMCs now employ PAs and almost two-thirds utilize PAs in the ED. Evolving payment models, a decline in government funding for research and education, and a national physician shortage are some of the leading factors cited as contributing to this change.

Physician assistants as a group currently care for more than 50% of ED patients at our hospital and have become natural stakeholders in departmental scholarship, committees, and operations. Other AMCs report that PAs could play an important role in these activities. We suspected that integrating them into the departmental mission may improve ED efficiency, as well as job satisfaction and intention to leave practice, but to our knowledge, there are no published models for how to integrate PAs into these endeavors. To address our department need and gap in the literature, we created a professional development program named the PA-II Program and describe its objectives, development, implementation, and outcomes.

OBJECTIVES OF INNOVATION
The primary objectives of this innovation were threefold: 1) to develop a program for PAs to integrate into department operations, 2) to provide a forum for PAs to pursue professional development and personal
growth, and 3) to enhance ongoing PA education. A secondary aim was to provide an incentive to help with PA retention.

DEVELOPMENT PROCESS

Job enrichment theory framed our design of a professional development program for PAs. This theory states that increasing the variety, autonomy, and complexity of one’s work may improve job satisfaction and reduce burnout. Further, investing in professional growth on an individual level may improve the organization as a whole and its environment for PAs. PA leadership defined the logistics of the program. Funding comes from our department in recognition of its commitment to PA professional development and investment in retention.

In brief, PAs who have worked for 3 years in the ED are eligible to become a PA II, which is considered a promotion and comes with a salary increase and 10 hours (0.0625 FTE) of protected and paid time each month to pursue an academic or scholarly activity. PAs submit a formal proposal (Data Supplement S1, available as supporting information in the online version of this paper, which is available at http://onlinelibrary.wiley.com/doi/10.1002/aet2.10382/full) outlining a potential role; its responsibilities; and how it may impact the group, department, or hospital. PA leadership supports them in this process by helping PAs explore their outside interests, identify a mentor, and identify existing or new opportunities within the department.

IMPLEMENTATION

A designated PA II director meets with individual PA IIs every 6 months to review progress and identify opportunities for expanding their project. The PA II director is a senior PA who has been with the department for more than 9 years and has approximately 0.1 FTE per month of protected and paid time to perform this role. PA leadership meets monthly for 1 hour with the PA II director to provide support, review PA II progress, and identify any barriers to a PA II’s professional growth within the program.

In addition to PA leadership, designated faculty within our department are natural mentors for PA IIs and provide support through existing structures. For example, a PA II with interest in disaster medicine participates in monthly meetings with faculty from the MGH Center for Disaster Medicine and collaborates with them on biothreats and hazardous materials trainings. These faculty provide feedback to the PA II and PA leadership on the PA II’s contributions to their group and progress in developing expertise.

Adoption of an innovation is most likely to succeed when there is a process for participants to create and share information. Therefore, we have imbedded forums for sharing this information so that the entire group may benefit from the individual’s added expertise. For example, PA IIs with focus on ultrasound and simulation share their knowledge with rotating PA students, new hires, and PAs orienting to our critical care pod. Other PA II’s have positions on department-based interdisciplinary teams that develop strategies to improve the health care for our homeless population and disseminate information about these initiatives to providers in the department.

We document the number and proportion of PA II’s in the PA group as part of surveillance of the program. We share PA and PA II retention rates before and after implementing the program as a potential measure of success of the PA II program and a surrogate for PA job satisfaction.

OUTCOMES

The program has expanded from six PA II’s (25% of the group) in 2013, the first year of implementation, to 24 PA II’s (68%) by 2018. There are now 14 unique PA II roles within six of eight departmental divisions. Example roles include simulation and ultrasound education, departmental quality and patient safety, and wilderness medicine. More recently PA IIs have explored interdepartmental roles in palliative care and pharmacology (see Data Supplement S2, available as supporting information in the online version of this paper, which is available at http://onlinelibrary.wiley.com/doi/10.1002/aet2.10382/full, for a full list.)

We have measured the annual retention rate for the PAs in our department since the PAs started in our department in 2010. Annual retention rate is presented as a percent and equal to one minus the ratio of PAs whose employment with our department ended during a given year relative to the total number of PAs on staff in the department at the end of that same year. The average retention rate for all PAs before the PA II program (2010–2012) was 87.3%, approximately the same as the average retention rate since starting the PA II program (2013–2018, 87.5%).
However, the retention rate for PA IIs was higher than PAs not in the program (PA Is) in every year except 2017 (Figure 1). The average retention rate for PA IIs is 90.2% compared with 85.7% for PA Is.

**DISCUSSION**

We have developed a formal program for PAs at an academic ED to pursue professional development in line with the AMC’s mission of academic growth for the betterment of the individual and the organization. As PA groups expand and assume a larger presence within the health care system, it will be important to provide opportunities for professional growth and promotion. Although the participants in our program are PAs in an academic ED, we believe that this model could easily be applied to advanced practice providers (APPs) in other specialties.

Academic EDs are uniquely positioned to integrate PAs into the nonclinical activities of the department. We have found that existing resources, such as division faculty, ongoing research projects, and department committees, are eager to integrate PAs into their teams. This meets our first two objectives to integrate PAs into department operations and create forums for personal growth. Our third objective, to enhance PA education, is met by supporting PAs in developing roles dedicated to teaching, quality improvement, and clinical care, as well as dissemination of information. PAs in these roles not only teach their PA colleagues, but also collaborate as part of interprofessional teams in education initiatives within the department. These collaborations align with our hospital mission with regard to continued learning and have the potential to secure publication and grant funding.

As the PA II program expands, one of our primary challenges is maintaining surveillance of the program. Some projects and content areas are more amenable to a PA II position as they are already contained within an existing division or committee. However, some PAs have had interests in topics without a preexisting organized structure within our department. We view this as an opportunity to create interprofessional partnerships, and in these instances, we try to pair the PAs with faculty who have shared professional interests.

Finally, the annual retention rate for the MGH ED PA program is approximately the same since beginning the PA II program. However, the retention rate for PA IIs is greater than PA Is in every year except for one (2017). This suggests that the opportunities for professional development provided by our program may increase our ability to keep experienced PAs working in our academic ED for a longer period. Further, retention rate does not account for new hires (PA Is), which limits our ability to discern whether the PA II program motivates PA Is to continue employment. However, our turnover rate for PA IIs (9.8%) is less than the most recently reported national benchmark (14.2% in 2018). Together these data suggest efficacy of the PA II program.

In summary, the PA II program is a novel professional development program and promotional pathway for PAs at an AMC. It harnesses their individual academic and clinical interests in ways that improve personal accomplishment and we provide channels for PAs to share their growing expertise with their peers and the rest of the department. Our data suggest that this may improve job satisfaction and help reduce expensive workforce turnover. We believe that this ultimately contributes to group cohesion and a shared

![Figure 1. PA retention rate after initiating PA II program. PA = physician assistant.](image-url)
experience for the team outside of the traditional model of service for APPs.

References


Supporting Information

The following supporting information is available in the online version of this paper available at http://onlinelibrary.wiley.com/doi/10.1002/aet2.10382/full

Data Supplement S1. MGH Department of Emergency Medicine PA II Application.
Data Supplement S2. PA II Roles & Responsibilities.
NEW IDEAS IN B-E-D-SIDE TEACHING

Escape the Trauma Room

Mikhail Podlog, DO\textsuperscript{1}, Abbas Husain, MD\textsuperscript{1}, Josh Greenstein, MD\textsuperscript{1}, and Snaha Sanghvi, DO\textsuperscript{2}

ABSTRACT

Traditional conference didactics may not be effectively meeting the learning needs of today’s emergency medicine (EM) residents, and educators are employing various interactive approaches to engage learners. Escape Room is a popular adventure game in which participants must work together to solve a series of puzzles to escape a locked room. Our aim was to adapt this game design to teach core EM content and procedural aptitude. Upon entering the “locked” room residents were faced with a series of puzzles involving concepts such as toxicology antidotes, ventilator management, echocardiogram interpretations, airway foreign body removal, arterial line transducer setup, and using a cast cutter. Using teamwork and limited clues learners had to work together to “escape” the room. Afterward, a didactic summary was given to enhance knowledge retention. The Escape Room construct was successfully adopted as an engaging model to teach EM core content and procedure skills while simultaneously fostering team building. Feedback received was overwhelmingly positive. This unique alternative educational activity can be easily implemented at any EM residency program as an effective alternative educational tool.

BACKGROUND

Traditional conference didactics may not effectively meet the learning needs of today’s emergency medicine (EM) residents.\textsuperscript{1} Educators are moving away from hour-long lectures and are employing various interactive approaches to engage learners. Existing strategies include small-group discussion, simulation, flipped-classroom teaching, and interactive activities.\textsuperscript{2} Escape Room is a popular adventure game, used for entertainment and team building, in which participants must work together to solve a series of puzzles to escape a locked room. The concept of an Escape Room educational activity offers the potential to expand an educator’s repertoire of active learning methods. This novel activity has been recently described in the literature for nursing and medical student education with some success.\textsuperscript{3,4} Our aim was to adapt this game design to teach core EM content and procedural aptitude, while creating an engaging team-building activity. We included a variety of core EM topics of different difficulties to benefit a diverse group of learners. Core content included toxicology antidotes, airway and ventilator management, and echocardiogram (ECG) interpretation. Procedural skills included arterial line transducer setup, airway foreign body (FB) retrieval, and cast removal. We hypothesize that this innovative session would engage learners more than traditional conference didactics.

EXPLANATION

The learners were divided into three equal groups of about 10 learners per group (one or two medical students and an equal number of junior and senior residents). The three group attempted the “escape” consecutively since the same room was used, and each group had 1 hour to escape. Each group was initially briefed on the activity outside the “locked” trauma room. Upon entering, learners were faced with a series of puzzles. The first three puzzles, once solved, yielded a series of numbers that would open a locked box (Figure 1). These puzzles included a toxicology antidote...
matching puzzle, a maze with airway and ventilator management questions, and an ECG interpretation station. Once the locked box was opened, the learners retrieved additional clues that pointed them to the other remaining puzzles scattered around the room. These included airway FB removal, setting up an arterial line transducer, using a cast cutter, and solving a jigsaw puzzle for a dermatology visual diagnosis prompt. The answers to these final four puzzles helped decipher a phone number that, once called, would “unlock the room.” A full description of the activity is provided in Data Supplement S1 (available as supporting information in the online version of this paper, which is available at http://onlinelibrary.wiley.com/doi/10.1002/aet2.10410/full). The instructors were available to provide subtle clues if the learners were stuck on a topic for an extended period of time. This activity was piloted prior to implementation to ensure that all groups had sufficient time to “escape.” Learners were debriefed after completion of the activity with answers and a brief explanation to each puzzle. One week later the entire group was given a didactic summary of the topics covered to provide spaced repetition and enhance knowledge retention.

DESCRIPTION

The Escape Room construct was successfully adopted as an engaging model to teach EM core content and procedure skills while simultaneously fostering team building. To gauge effectiveness, residents completed an anonymous survey based on a 1 to 5 Likert scale after the educational activity with an 87% completion rate. Eighty-two percent rated this activity as highly educational. Ninety-four percent stated that the topics covered were very relevant to EM. Everyone answered that they would want to do this activity again. Written comments received included “Definitely more engaging than traditional activities,” “Gives you an incentive to learn and develop common skills such as handling the cast cutter, reading ECGs, learning antidotes to common medications, and learning to troubleshoot the vent,” and “Probably the best activity we ever did during any conference.” For reference, our standard conference day is 4 hours of didactics with occasional intermixed small groups and simulation sessions. This novel activity can be easily implemented at any EM residency program as an effective alternative educational tool.

References


Supporting Information

The following supporting information is available in the online version of this paper available at http://onlinelibrary.wiley.com/doi/10.1002/aet2.10410/full

Data Supplement S1. Detailed activity walkthrough with explanations and pictures of individual stations.
Engaging Emergency Medicine Influencers in Sex- and Gender-based Medicine: Lessons Learned from the Sex and Gender Interest Group in Emergency Medicine and the SAEM Jeopardy Game

Jeannette Wolfe, MD1, Basmah Safdar, MD2, Kinjal N. Sethuraman, MD3, Marna R. Greenberg, DO4, Tracy E. Madsen, MD5, Angela F. Jarman, MD6, and Alyson J. McGregor, MD5

ABSTRACT

The Sex and Gender in Emergency Medicine (SGEM) interest group of the Society of Academic Emergency Medicine (SAEM) was established to increase research and to disseminate knowledge about the influence of sex and/or gender in acute care medicine and on patient outcomes. To help facilitate these goals, over the past 4 years, SGEM has created, delivered, and honed a Jeopardy-like scientific quiz game for the annual SAEM national meeting. Here we describe the SAEM Jeopardy Game’s development, implementation, evolution, and outcomes as well as our targeted approach to access and engage emergency medicine stakeholders in its participation.

NEED FOR INNOVATION

More than 25 years ago, the National Institutes of Health (NIH) implemented a policy requiring grant applicants to include both males and females in human clinical trials. In 2015, they expanded this mandate to include preclinical biomedical research.1-3 Even with the NIH prompts and continually mounting evidence identifying clinically important sex and gender differences in disease, pharmacology, therapeutic interventions, and medical outcomes, the earnest inclusion of the variables of sex and gender into mainstream medical research and clinical care has been slow to occur. As many sex- and gender-based medicine (SGBM) principles are directly related to acute care, we, as leaders of Sex and Gender in Emergency Medicine (SGEM) Interest Group at SAEM, identified a need to accelerate education surrounding SGBM within academic emergency medicine (EM).4-7 This innovation paper describes how a Jeopardy-like game may facilitate knowledge translation of SGBM principles to academic emergency physicians (EPs) to enhance their use in clinical care, teaching, and research.

BACKGROUND

The first didactic featuring sex as a biologic variable in EM was presented at the SAEM Annual Meeting in...
2007—it was attended by two registrants. The SGBM in EM awareness gap was further highlighted by a 2011 systematic review showing that although 18% of EM-driven original research publications reported sex and gender as independent variables, only 2% analyzed them as a primary or secondary outcome.\(^8\) This gap was the impetus for the 2014 Gender-Specific Research in Emergency Care Consensus Conference.\(^9\) The conference was attended by EM researchers, clinicians, and policy leaders, and it established a research template to identify and correct knowledge gaps in EM initiated projects.\(^9\) Importantly, it also was a catalyst for the formation of SGEM. Since its inception, SGEM has contributed didactic sessions to every national SAEM conference and now has over 230 members. A keystone project of SGEM has been its creation of the SAEM Jeopardy game (SJG).

**OBJECTIVE OF INNOVATION**

The SJG was created as a novel platform to educate and engage the EM community about the importance of sex and gender in emergency care.

**DEVELOPMENT PROCESS**

Gaming methods have been described for over 20 years as augmentation tools for teaching biomedical sciences, and EM has embraced their use in ultrasound and simulated educational sessions such as Sonogames and SimWars.\(^{10,11}\) SJG was developed as a game using a PowerPoint template. In the inaugural 2016 game, SGBM categories were chosen by SGEM expert consensus. Since then, questions have been updated primarily by the lead author using multiple sources including SGBM reference texts, SGBM flagged research from Google Scholar Alerts, practice changing studies collated by SGEM, and research highlighted in popular media.\(^{12,13}\) The questions are then peer-reviewed and revised by other SJG committee members. The format of questions varies and has included questions with binary or short answers as well as those requiring interpretations of echocardiograms, graphs, or ultrasounds. Embedded videos featuring prominent EM SGBM researchers have also been used. These videos included both a question and a commentary by each researcher underscoring the added value that SGBM brought to their work. In 2019, an action plan category was added to give EPs more concrete tools to increase SGBM’s awareness and inclusion at their own institutions.\(^{14,15}\) A sample of SAEM Jeopardy categories, questions, and answers from previous games can be found in Table 1.

**IMPLEMENTATION**

The SJG was conceived to bring the routine consideration of sex and gender into mainstream clinical care for EPs. It aligned well with SGEM’s mission to increase knowledge translation of SGBM within EM.

### Table 1
A Sample of SAEM Jeopardy Categories, Questions, and Answers From Previous Games

<table>
<thead>
<tr>
<th>Category</th>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic definitions</td>
<td>This refers to one’s chromosomal and hormonal make-up</td>
<td>Biological sex (&quot;Every cell has a sex and every individual is gendered&quot;)*</td>
</tr>
<tr>
<td>Heart to Heart (cardiology)</td>
<td>Higher stroke risk from this rhythm (with EKG showing atrial fibrillation)</td>
<td>Female (follow-up slide shows CHADS VASC score showing that female sex is an independent variable)</td>
</tr>
<tr>
<td>Procedures</td>
<td>In female ED patients &gt; 80, half who got this common procedure did not have any clear indication for it</td>
<td>Foley catheter (^38)</td>
</tr>
<tr>
<td>Crash and Boom (trauma)</td>
<td>In 2017, males made up this percentage of people fatally injured in motor vehicle crashes (within 10%)</td>
<td>71% (^39)</td>
</tr>
<tr>
<td>Sex and Drugs (pharmacology)</td>
<td>Ethanol and likely zolpidem get chewed up by this stomach enzyme that is revved up by testosterone</td>
<td>Alcohol dehydrogenase</td>
</tr>
<tr>
<td>Research</td>
<td>In a recent sampling of 2008-2013 randomized controlled trials from major journals, this percentage had overrepresentation of males (within 10%)</td>
<td>43% (^6)</td>
</tr>
<tr>
<td>Diversity</td>
<td>In 2019, The Lancet published a study evaluating millions of papers and correlated the likelihood of reporting biological sex in write-up and published papers to this</td>
<td>Female gender of first or last author (^38)</td>
</tr>
</tbody>
</table>

\*This phrase is adapted from the Institute of Gender and Health in the Canadian Institutes of Health Research.
We started this initiative by targeting prominent EM stakeholders as game participants. By engaging them, the objective was to have these stakeholders champion the further dissemination of SGBM through their own influential networks. In addition, attracting prominent EM individuals to participate in SJG was a strategy to increase audience attendance and participation. Past SJG participants include SAEM presidents and members of the board of directors, department chairs, journal editors, prominent EM researchers, and high-profile social media influencers. Most recently we added resident and medical students board of directors (RAMs) team. Each year, stakeholders have been invited to captain a team of three players to compete against two other teams.

The game itself is played with three teams, a host, a score keeper, and at least one judge. Audience participation is encouraged both by answering questions missed by the team and by live marketing through social media. The host rewards anyone who can correctly answer the question with a small prize or a SJG branded T-shirt. At the end of the SJG, prizes are awarded for three categories: overall winner based on most points, team spirit, and social media influencer. The team receiving the loudest and most enthusiastic response from the audience gets the award for team spirit. Finally, the social media award is given to the team and its followers that post the most tweets using their team-specific hashtag that is registered with Symplur’s Healthcare hashtag project.¹⁶

Since its inception, the sponsorship of SJG has also expanded. SJG initially started out as an didactic solely created by SGEM. In 2018, SGEM partnered with the Academy of Women in Academy Emergency Medicine (AWAEM) and with the Academy of Diversity and Inclusion in Emergency Medicine (ADIEM). This was a natural collaboration as many gender-related health outcome disparities are exacerbated by race and sexual minority status, and several recent articles have suggested that diverse research teams are more likely to include the variables of sex and gender in their study designs.⁵,¹⁷ These relationships increased our ability to market the program and expand our audience.

Funding of this SJG has also expanded over the years. The first 2 years were funded by the Division of Sex and Gender in Emergency Medicine at Alpert Medical School at Brown University. Our initial success formed the basis for obtaining SAEM grants in successive years to cover the SJG promotional costs and awards. We have successfully partnered with SAEM’s administration resulting in key benefits such as high-profile time slots (e.g., the opening reception) and recruitment of prominent stakeholders for the game.

In 2019, we shifted from institutional-based teams to organizational ones. Specifically, the last SJG had SAEM Board of Directors competing against journal editors and RAMS. The RAMS team won and appeared to be particularly knowledgeable about the gender-based questions, possibly highlighting a generational based knowledge gap in this area.

Besides sharing the content of the SJG with an increasing number of EPs at the SAEM national conference, versions of the game have been used to disseminate SGBM knowledge more broadly throughout the medical community. To date the game has been replayed at institutional grand rounds and digitally.

**OUTCOMES**

Since initiation, SJG’s attendance has steadily increased and in 2019 included over 100 participants. Informal feedback from past team members suggests that SJG has changed their perception concerning the relevance of this material to EM.

I think “gamifying” the topic of sex and gender facilitates the discussion of a challenging topic. The number of people in the audience who often gasped at the answers, because it was new knowledge was telling. For me, there were so many areas that the questions touched upon, where I was completely unaware.—SAEM Board Member
shared with other medical educators at the 2018 Sex and Gender Health Education Summit. Finally, we are working to permanently house the game on an open-access platform “Sex and Gender Specific Health” sponsored by Laura W. Bush Institute for Women’s Health and Texas Tech University Health Sciences https://www.sexandgenderhealth.org/. In addition, hundreds of educational SGBM-related tweets, many of them sparked by competition for the social media award, have circulated throughout the EM community due to the game. Finally, the membership of SGEM interest group has increased a likely byproduct of increased awareness of SGBM principles.

REFLECTIVE SUMMARY

Sex and Gender in Emergency Medicine Interest Group has used several novel methods including SJG to access and engage a diverse EM audience including its key stakeholders about the importance of SGBM in acute care. Besides continuing to optimize SJG’s presence and impact at the SAEM national meeting, we intend to expand its use to other high-profile meetings to increase knowledge dissemination. A stronger format for program evaluation would be encouraged for those considering scholarly pursuits with the gamification/serious games type of interventions. Several manuscripts used in recent systematic reviews18,19 provide insights on how we could have evaluated our program. In this literature the most common outcomes measured were improvements in knowledge, problem solving, skills, and attitudes; rarely do they offer a measure of patient outcomes.20 One of the most adaptable for our program would be modeled as a “before-and-after” assessment questionnaire similar to that used by Diehl when they improved scores on content, knowledge, and attitudes of clinicians on insulin therapy using a game.21 These will be important considerations as we begin collecting longitudinal data about SJG’s influence and effectiveness in SGBM knowledge translation to EM.

REFERENCES


The multicenter retrospective study by Acuña et al.\textsuperscript{1–3} is both interesting and timely given recent acknowledgments of gender bias in medicine.\textsuperscript{1–3} I commend the authors for a well-done research paper. In particular the methodology, discussion of results, and acknowledgement of limitations are academically thoughtful.

Bias impacts women physicians in training, and bias affects their performance over time. Certain unconscious forms of bias prevent women from achieving what would be anticipated based on their education and training. Implicit gender bias, for example, is a set of commonly held beliefs about a certain group. An example of implicit gender bias: faculty determining that women are not as procedurally adept as men and therefore women are evaluated with lower scores and reach graduation level competency at a slower rate than men. Relatedly, stereotype threat is a phenomenon whereby members of a certain group characterized by negative stereotypes perform below their actual abilities.\textsuperscript{4} Stereotype threat is realized when the emotional impact of such stereotypes causes, for example, women stress, anxiety, or disappointment thus leading to underperformance and underachievement. For example, if women trainees had higher miss rates for intubation and perform more poorly due to a threatening environment not felt to be gender equitable, this would be stereotype threat at play. The consequences of implicit gender bias and stereotype threat accumulate over time and may be seen by studying emergency medicine resident ultrasound subcompetency evaluations.

The authors seek to describe whether significant gender differences exist in ultrasound milestone evaluations for emergency medicine trainees. Interestingly in their single-institution pilot study, the authors found gender bias. When they broadened their query to multicenter (i.e., 16 ACGME-accredited programs), this bias was not found.

In 2012 the American Board of Emergency Medicine and the Accreditation Council for Graduate Medical Education published the emergency medicine milestones.\textsuperscript{5} The 23 subcompetencies provide a framework for resident training and evaluation. The Patient Care 12 (PC12) subcompetency delineates ultrasound proficiency levels 1 through 5. Specific milestones are listed in a column for each level.\textsuperscript{6} In 2016, Nelson, myself, and co-authors expressed our concern in the lack of applicability of PC12. We offered a consensus-based subcompetency revision.\textsuperscript{7} This delineates specific milestones that clarify resident performance and map a progression in learning and ultrasound competency. This revision also deemphasizes the single number focus, which serves to measure technical competency. Numbers alone are not an indicator of integration and clinical competency. The authors for the present research study on gender bias thoughtfully account for this in their limitations.

Truthfully, implicit gender bias is difficult to study and yet simultaneously it is well described in the literature. For example, emergency medicine faculty direct observations and written evaluations for emergency medicine women trainees are qualitatively different and assess them as less capable than men.\textsuperscript{8} Dayal et al.\textsuperscript{9} described the implicit gender bias they found in their comparison of male and female resident milestone evaluations for emergency medicine trainees.
The quantitative rate of attainment of the milestones was higher for men than for women for all subcompetencies.

Every emergency medicine program director is motivated for residents to graduate having met the assessment requirements outlined by the 23 subcompetencies. Thus, simply reporting a level achieved at the end of training would not capture implicit gender bias. What the authors for this study utilize and what is important is an analysis of the scores at points in time and the trend of those scores over time for women and for men.

Intuitively, we know that women are as capable and as competent as men. Thus, if the measurements are truly objective, then generally we would not expect gender bias in the evaluation of emergency medicine trainees. However, gender bias is unlikely captured in the recording and reporting of objective numbers. Teasing out bias may require determining the gender distribution of the faculty evaluators. Teasing this out may require the nuance of analyzing the subjective comments and written evaluations. Arguably, these two elements can be the next scholarly inquiry for these authors.

**Resa E. Lewiss, MD**
(resa.lewiss@jefferson.edu)
*Thomas Jefferson University, Philadelphia, PA*

*Supervising Editor: John H. Burton, MD.*

## References

The Irresponsible Use of Social Media Among Medical Students

Mark Curato, DO1, Abbas Husain, MD2, Kaushal Shah, MD3, Marc Kanter, MD4, Daniel Egan, MD5, Holly Thompson, MD6, Mark Silverberg, MD7, Michael Jones, MD8, Laura Melville, MD9, Sally Bogoch, MD10, Anand Swaminathan, MD11, Jennifer Beck-Esmay, MD12, Thomas Nguyen, MD13, Tina Dulani, MD14, Elizabeth Fernandez, MD15, Frosso Adamakos, MD16, Geoff Jara-Almonte, MD17, and Pinaki Mukherji, MD18

Emergency medicine (EM) remains a competitive specialty,1 and the residency application process is arduous. Medical students are applying to and interviewing at a record number of programs,2 and an additional stressor has been added with the introduction of the standardized video interview.3 Students have, for years, perceived the residency match process as opaque and high-stakes, and the competitiveness of our specialty in tandem with the novel facets of the process may be increasing that stress. The informal sharing of rotation and interview experiences and unpublished program information has always been a source of comfort and clarity to the applicant during this challenging time. The current generation of students is unique relative to their predecessors in that they expect program information to be transparent, detailed, and easily available. Indeed, a gap may exist between what is expected by these applicants and what is provided by the AAMC, medical schools, hospitals, and residency programs. We suspect that such a disconnect between expectations and reality in the midst of a stress-laden process has resulted in unprofessional coping mechanisms by some students, and these have been immortalized in online digital media.

In recent years, as digital media have permeated many segments of our daily lives, students have turned to online forums such as Student Doctor Network (SDN)4 and Reddit5 to seek crowdsourced data and a supportive community. During the 2018 to 2019 application season, a shared open-access and freely editable online spreadsheet was utilized by medical students applying for postgraduate training in the United States. The document contained candid reflections on rotations, interviews, rank lists, and anecdotal data from presumed applicants. There are several examples in which students’ opinions, frustrations, and dissatisfaction are expressed with sexually explicit, hateful, misogynistic, violent, homophobic, racist, crude, and threatening language.5–7 In addition, there is language that is less vulgar yet still overtly unprofessional and disparaging to schools, residency programs, named individuals, and anonymous coposters. While a majority of these comments were made with “throwaway” accounts, the context in which they appear makes clear...
that authors are indeed medical students and not imposters purporting to be so.

After reading this strong and hurtful language, our community of GME and UME educators in NYC felt shock, disappointment, anger, and embarrassment. The AMA has a clear statement on “professionalism in the use of social media.” In addition, the Federation of State Medical Boards has set guidelines for social media use. The students in these forums are clearly in breach of these guidelines. Collectively, we felt a sense of betrayal by individuals whom we will soon welcome into our profession and specialty.

As emergency physicians, we develop a sacred doctor–patient relationship with vulnerable individuals whose circumstances have denied them the benefit of vetting and selecting us. We ask as a foundation of our profession to be regarded by the public as honest, altruistic, kind, caring, and competent, and by the nature of our doctor–patient relationships we are usually afforded those assumptions prima facie. Sentiments and language such as those cited in the above document threaten irreparable damage to the foundation of trust upon which our house of medicine rests. While our initial instinct was to defend our venerated profession against the threat posed by the authors of these reprehensible posts (including punitive measures in the event that any authors were actually identified), this reaction gave way to deeper thought and an evolved assessment of the way forward.

The first step is to admit that we have a problem, all of us. The medical students at issue are not aberrations. Their aspirations and ideals as well as their anxieties and flaws are no different than ours were. As educators, we are an amalgam of teacher, mentor, parent, and boss. Perhaps, more importantly, we are their vision of us. The medical students at issue are not aberrations.

We accept that lashing out is a function of the anxiety brought about by being involved in a match process that seems inherently flawed, unfair, and arbitrary. While unthinkable upon our initial discovery of the document, our dominant reaction now is one of reflection and a path forward. We recommend the inclusion of social media professionalism in medical student curricula with a specific focus on such behavior during the match process. We urge all those in medical student and residency leadership not only to actively discourage such conversation but also to directly address issues of professionalism with trainees. This type of language should not be tolerated at any level. We call upon medical schools, residency programs, and other stakeholders to actively seek out evidence of betrayal of professional standards and basic virtue within their spheres of influence. Once discovered, we further call on that righteous majority to aggressively condemn such content and also to identify, understand, educate, and reform those who create it, for there is no place for hate, xenophobia, or vitriol in a just society, let alone among our venerated profession.

The ALL NYC EM Board is composed of Mark Curato, Abbas Husain, Kaushal Shah, Marc Kanter, Daniel Egan, Holly Thompson, Mark Silverberg, Michael Jones, Laura Melville, Sally Bogoch, Anand Swaminathan, Jennifer Beck-Esmay, Thomas Nguyen, Tina Dulani, Elizabeth Fernandez, Frosso Adamakos, Geoff Jara-Almonte, and Pinaki Mukherji. All NYC Emergency Medicine Conference, Inc., is a 501© nonprofit organization dedicated to the education of emergency medicine residents and is managed by our board of dedicated EM educators from the greater NYC area.

References

1. NRMP Program Results 2015-2019 Main Residency Match. 2019. Available at: https://mk0nrpm3oyqui6wqfm.kinstacd
The Responsible Use of Social Media Among Medical Students and Residents

Stephen C. Dorner, MD, MPH, MSc1,2, Morgan B. Swanson3, and Shana M. Zucker4

We applaud the ALL NYC EM Board in its thoughtful call outlining a path forward for increasing professionalism in our daily lives and particularly with respect to social media amid the modern challenges of the 21st century. We similarly condemn sexually explicit, hateful, misogynistic, homophobic, racist, crude, and threatening language that finds its way into the public forums that medical students and residents frequent, such as SDN1 and Reddit.2 However, we propose that the purpose, content, and impact of these forums fundamentally constitute responsible use of social media and highlight issues permeating the medical profession in need of public, open scrutiny.

We need more frank discussion about barriers to entering the medical profession that are inherently discriminatory. Emergency medicine is indeed a progressively competitive specialty and the residency application process across all specialties is increasingly arduous for all the reasons that have been suggested, not the least of which is the cost of applying to, traveling to, and interviewing at an ever-growing number of programs.3 In turn, the residency application process is increasingly weighted with competitive advantages afforded to only the wealthiest applicants at medical schools with abundant resources. The informal sharing of rotation and interview experiences via social media is not only a source of comfort and clarity to applicants, but also often an instrumental tool in narrowing down programs and is often a tool to help applicants overcome barriers and selective advantages afforded to others.

That is to say nothing of the costs to qualify to apply to residency. The USMLE Step Examinations—a total of four examinations designed to afford health care consumers “a high degree of confidence that doctors who have passed all three Steps of the USMLE have met a common standard”4 are inappropriately manipulated by some programs in an attempt to “weed out” applicants who pass the examinations but with “lower” passing scores. Such applicants who pass a Step examination are both ineligible to retake the test and also siloed into a class of applicants deemed to be too “low-performing” for acceptance into some programs, despite the USMLE’s stated intent. The result is an ever-increasing system of parallel, unofficial costs wherein applicants purchase preparation courses, supplementary textbooks, and online reviews to augment their already extensive medical education simply to excel on the USMLE examinations.

Online forums such as SDN and Reddit serve as a window into the inner workings of this system, casting light on program directors’ and faculty members’ honest, unfiltered commentary unlikely to be polished and published on official websites. How should applicants learn about unpublished cutoffs and tiers at programs who will gladly accept an application fee but never afford them an interview—much less a position? These open forums are often one of the only sources for unfiltered information in an obscure system that seems intent to exploit a vulnerable population. Transparency in reporting program-specific application criteria should be a priority.

From the 1Harvard Affiliated Emergency Medicine Residency Program, Harvard Medical School; the 2Department of Emergency Medicine, Massachusetts General Hospital, Boston, MA; the 3Carver College of Medicine, University of Iowa, Iowa City, IA; and the 4Tulane University School of Medicine, New Orleans, LA.

Received August 28, 2019; accepted August 28, 2019.

The authors have no relevant financial information or potential conflicts of interest to disclose.

All authors contributed to the study concept, design, drafting of the manuscript, and critical revisions. A related article appears on page 168.

Supervising Editor: Wendy C. Coates, MD.

Address for correspondence and reprints: Stephen C. Dorner, MD, MPH, MSc; e-mail: sdomer@mgh.harvard.edu.

AEM EDUCATION AND TRAINING 2020;4:171–173

© 2019 by the Society for Academic Emergency Medicine
doi: 10.1002/aet2.10393
ISSN 2472-5390

171
“filters” could help applicants efficiently apply to programs most likely to review their application and, in turn, reduce the burden of the continuously increasing number of applications received by residency program directors.

This leads to the need for more open discussion about the blurring line between the business of medicine and our training environments. Contract management groups were originally developed as a business model for staffing community EDs, but now are increasingly entering the education space and managing residency programs. There has been significant debate on the subject, including public position statements from professional societies painting cautionary tales about the role of CMGs in academic medicine. For those applicants who have never had contact with someone in a CMG, much less the executive board of a professional society, there is little information available to inform decision making. These forums serve as an opportunity for applicants to network with individuals who have more information and learn what may be otherwise privileged information.

A collaborative, collegial spirit is the true underlying theme of these forums. When Hahnemann Hospital recently closed, some 580 residents’ positions were eliminated amid reports that the funds previously supporting their positions would be auctioned to the highest bidder. As hospital leaders, CMS, and the rest of the corporate financial stakeholders engaged in negotiations, 580 newly unemployed residents and families were struggling to find practical information to help them land on their feet. Online forums created a medium for open and honest communication about what was happening in their hospital, to their positions, and about their future job prospects in a way that friends or colleagues would converse and commiserate with one another in any similar high-stakes situation.

The tone of these forums reflects the raw, visceral, and admittedly unprofessional sentiments often driven by anxiety among individuals who feel powerless in a flawed and seemingly arbitrary system. We completely agree with the ALL NYC EM Board that these forums expose the ugly underbelly of medicine and the system of applying to residency. We must point out that this includes the violations of NRMP guidelines during the resident application process, ranging from the relatively benign to the egregious. Programs asking applicants about significant others should know better. Institutions with rampant misogyny and racism should be vilified by name to alert and protect future applicants. Many of us have experienced that the system is riddled with unreported violations. However, in the words of one Reddit use, “snitches get stitches and no one who has matched is concerned with going back and trying to report places.” The system is not structured to incentivize reporting; it stifles it. When faced with an inappropriate question by a program director, does one answer the question or call them out on their violation? How then does embarrassing a program director influence an applicant’s rank? How often do programs informally communicate with one another—in privileged forums that aren’t public and openly accessible? Hopefully we can all agree that addressing injustice should not carry unjust repercussions.

We, as a profession, need to decide if this system reflects the best of us and inspires the best in us. We could not agree more with the ALL NYC EM Board that as emergency physicians, we do develop a sacred doctor–patient relationship with vulnerable individuals and agree that as a foundation of our profession we must protect the public perception as an honest, altruistic, kind, caring, and competent body of professionals. Let us therefore hold ourselves to the highest standards across the entirety of our profession.

We agree that we must hold our institutions accountable to the same professionalism standards to which we hold our applicants. We join the ALL NYC EM Board in calling on that righteous majority to aggressively denounce the betrayal of professional standards and basic virtue within their spheres of influence and also to identify, understand, educate, and reform those who create it. There is no place for hate, xenophobia, or vitriol in a just society—nor disenfranchisement, discrimination, or exploitation.

In our view, the principal issue is not the vulgarity of the language at use in social media. Rather, the principal issue is the ever-higher, complex, discriminatory, and opaque barriers to joining our ranks. We need more open conversations reflecting the collaborative spirit underlying our vocations, embracing the modern networking mediums available to facilitate information exchange.

References

