Readiness of US General Surgery Residents for Independent Practice

Brian C. George, MD, MAEd,[*] Jordan D. Bohnen, MD, MBA,† Reed G. Williams, PhD,‡ Shari L. Meyerson, MD, MEd,§ Mary C. Schuller, MSEd,§ Michael J. Clark, PhD,¶ Andreas H. Meier, MD, MEd,|| Laura Torbeck, PhD,† Samuel P. Mandell, MD, MPH,[**] John T. Mullen, MD,[*] Douglas S. Smink, MD, MPH,†† Rebecca E. Scully, MD,‡‡ Jeffrey G. Chipman, MD,§§ Edward D. Auyang, MD, MS,¶¶ Kyla P. Terhune, MD, MBA,|||| Paul E. Wise, MD,[***] Jennifer N. Choi, MD,† Eugene F. Foley, MD,††† Justin B. Dimick, MD, MPH,¶ Michael A. Choti, MD,‡‡‡ Nathaniel J. Soper, MD,§ Keith D. Lillemoe, MD,[*] Joseph B. Zwischenberger, MD,§§§ Gary L. Dunnington, MD,† Debra A. DaRosa, PhD,§ and Jonathan P. Fryer, MD, MHPE§, on behalf of the Procedural Learning and Safety Collaborative (PLSC)

Objective: This study evaluates the current state of the General Surgery (GS) residency training model by investigating resident operative performance and autonomy.

Background: The American Board of Surgery has designated 132 procedures as being ''Core'' to the practice of GS. GS residents are expected to be able to safely and independently perform those procedures by the time they graduate. There is growing concern that not all residents achieve that standard. Lack of operative autonomy may play a role.

Methods: Attendings in 14 General Surgery programs were trained to use a) the 5-level System for Improving and Measuring Procedural Learning (SIMPL) Performance scale to assess resident readiness for independent practice and b) the 4-level Zwisch scale to assess the level of guidance (ie, autonomy) they provided to residents during specific procedures. Ratings were collected immediately after cases that involved a categorical GS resident. Data were analyzed using descriptive statistics and supplemented with Bayesian ordinal model-based estimation.

Results: A total of 444 attending surgeons rated 536 categorical residents after 10,130 procedures. Performance: from the first to the last year of training, the proportion of Performance ratings for Core procedures (n = 6931) at ''Practice Ready'' or above increased from 12.3% to 77.1%. The predicted probability that a typical trainee would be rated as Competent after performing an average Core procedure on an average complexity patient during the last week of residency training is 90.5% (95% CI: 85.7%–94%). This falls to 84.6% for more complex patients and to less than 80% for more difficult Core procedures. Autonomy: for all procedures, the proportion of Zwisch ratings indicating meaningful autonomy (''Passive Help'' or ''Supervision Only'') increased from 15.1% to 65.7% from the first to the last year of training. For the Core procedures performed by residents in their final 6 months of training (cholecystectomy, inguinal/femoral hernia repair, appendectomy, ventral hernia repair, and partial colectomy), the proportion of Zwisch ratings (n = 357) indicating near-independence (''Supervision Only'') was 33.3%.

Conclusions: US General Surgery residents are not universally ready to independently perform Core procedures by the time they complete residency training. Progressive resident autonomy is also limited. It is unknown if the amount of autonomy residents do achieve is sufficient to ensure readiness for the entire spectrum of independent practice.

Keywords: assessment, autonomy, independent practice, performance, readiness, resident, simpl, surgical education

(*Ann Surg* 2017;xx:xxx–xxx)

There is increasing concern that some General Surgery (GS) residents are not competent to enter independent practice by they time they complete their residency training.[1–3] National surveys have demonstrated that these concerns are widespread.[4,5] Residents themselves sometimes feel less than fully confident, which may explain the rising rates of graduates seeking additional fellowship training.[6–8] Unfortunately, most of the existing evidence is either indirect or based on opinions.

The lack of high-quality empiric data makes it challenging to make progress—or even agree on the urgency of the task. For example, some have suggested that concerns about the current generation of trainee surgeons are essentially the same concerns voiced by generations of more senior surgeons.[9] Others point to the data we do have and caution against complacency.[10] It has so far been difficult to make a definitive assessment of the true scope of the problem and thereby shift the debate toward how to best identify and address the issues. Furthermore, even if there was consensus on what is needed, measuring success would be difficult without an established baseline against which educational quality improvement efforts can be judged.

Whatever the scope of the current deficits in trainee competence, diminished trainee autonomy almost certainly contributes to the problem.[11,12] The current model of graduate surgical education leans heavily on the principles elucidated by Dr Halsted over

From the [*]Department of Surgery, University of Michigan, Ann Arbor, MI; †Massachusetts General Hospital, Boston, MA; ‡Indiana University, Bloomington, IN; §Northwestern University, Evanston, IL; ¶University of Michigan, Ann Arbor, MI; ||SUNY Upstate Medical University, Syracuse, NY; [**]University of Washington, Seattle, WA; ††Brigham and Williams Hospital, Boston, MA; ‡‡Brigham and Women's Hospital, Boston, MA; §§University of Minnesota, Minneapolis, MN; ¶¶University of New Mexico, Albuquerque, NM; ||||Vanderbilt University, Nashville, TN; [***]Washington University, St. Louis, MO; †††UT Southwestern, Dallas, TX; ‡‡‡University of Wisconsin, Madison, WI; and §§§University of Kentucky, Lexington, KY.

This study was funded by a grant from the American Board of Surgery. The initial development of SIMPL was funded via grants from Massachusetts General Hospital, Northwestern University, and Indiana University. Later development was funded by contributions from the members of the Procedural Learning and Safety Collaborative (PLSC, http://www.procedurallearning.org).

The authors report no conflicts of interest.

Reprints: Brian C. George, MD, MAEd, 1C421 University Hospital, 1500 E. Medical Center Dr., Ann Arbor, MI 48109-5033.
E-mail: bcgeorge@med.umich.edu.

ISSN: 0003-4932/16/XXXX-0001
DOI: 10.1097/SLA.0000000000002414

100 years ago: residents are expected to work closely with supervising surgeons and assume increasing responsibility as they gain proficiency.[13] Since that time, progressive resident autonomy has been the essential feature of surgical training. More recently, patient safety has become an increasingly central concern in clinical training, particularly after the Libby Zion case in New York in 1984.[14] This case and growing public concern led to calls for more supervision of trainees, soon followed by the implementation of new rules that restricted resident autonomy.[15,16] Many of these new rules specifically addressed the supervision of trainees in the operating room (OR). At the same time, the introduction of managed care and resident work hour restrictions posed another set of challenges to the existing system.[17,18] After 3 decades of regulatory, institutional, and cultural change it is unknown how much meaningful operative autonomy surgical residents now achieve.

In this study, we aimed to evaluate the current state of the GS training model by assessing how much autonomy GS residents are provided and how well GS residency programs train and graduate surgeons who are ready for independent practice.

METHODS

Study Population

Faculty and categorical residents from 14 GS programs were enrolled in a prospective multi-institutional observational trial. Programs were recruited to represent a diversity of geography and case mix. For logistical reasons, enrollment was limited to those programs with at least 20 clinically active categorical residents. All enrolled programs were also members of the Procedural Learning and Safety Collaborative (PLSC, http://www.procedurallearning.org, Boston, MA), a multi-institutional surgical education research consortium.

The study was approved by the institutional review board of the lead site and approval or exemption was also obtained from each participating site's local institutional review board. Individual faculty and residents were not required to participate in this study, and participants could opt out at any time without repercussion.

Rating Scales

Attending and residents independently used a smartphone app to answer three questions immediately after completing an operative case. These questions ask raters to use the Zwisch, Performance, and Complexity scales, described below.

Zwisch Scale

The first question asks raters to use the previously validated Zwisch scale to assess the amount of guidance provided by the faculty to the resident (where faculty guidance is defined as the opposite of resident autonomy).[19] This scale consists of 4 levels, each representing progressively less guidance by the attending/more autonomy by the resident: Show and Tell, Active Help, Passive Help, and Supervision Only. In Show and Tell, the attending demonstrates and explains the procedure to the trainee. In Active Help, the attending is directing the flow of the operation, while in Passive Help the resident assumes leadership responsibility. A rating of Supervision Only indicates that the attending provides no guidance and acts only in a supervisory role. We defined the Zwisch level of Passive Help as the threshold for trainees to be categorized as ''Meaningfully Autonomous.'' Similarly, the Zwisch level of Supervision Only is operationally synonymous with ''Nearly Independent'' (Table 1).

Performance Scale

The second question asks raters to assess the trainee's readiness to independently perform procedures similar to the procedure in which they just participated.[20] They perform this assessment using the Performance scale which is adapted from a previously validated scale[21] and has 5 levels: Unprepared/Critical Deficiency, Unfamiliar with Procedure, Intermediate Performance, Practice Ready, and Exceptional Performance. The fourth level in this scale, Practice Ready, is conceptualized as the target goal for training and is defined as: ''Resident is ready to perform this operation safely, effectively, and independently assuming resident consistently performs procedure in this manner.'' For the purposes of summative analysis we defined the System for Improving and Measuring Procedural Learning (SIMPL) Performance level of Practice Ready as the threshold for trainees to be categorized as ''Competent'' (Table 2).

If raters assigned a Zwisch level of Show and Tell to a case they were not prompted to assess the trainee's Performance. This was implemented because the Show and Tell stage does not provide any real opportunity for the trainee to demonstrate their readiness for independent practice.[22] As a result, there are more Zwisch ratings than Performance ratings.

Complexity Scale

The final question asks raters to use the Complexity scale to assess whether patient-related and contextual factors made this procedure relatively more or less complex than typical cases of the same type. They could choose from 1 of 3 levels: Easiest 1/3, Average, or Hardest 1/3. To be clear, these relative ratings do not reflect absolute differences in procedure-specific difficulty (ie, Whipple vs appendectomy). Raters were instructed on this point.

Data Collection

Ratings were collected immediately after cases that involved a categorical GS resident with the aid of a smartphone application called SIMPL (Procedural Learning and Safety Collaborative, Boston, MA). This educational and research tool facilitates assessment and feedback of operative performance.

The SIMPL system is described in detail elsewhere,[20] but we include a brief overview here. After completing a procedure, either the attending or resident can use the SIMPL app to create a new evaluation by specifying who participated in the operation, which procedure(s) was performed, and the date of the procedure. Once an evaluation is created by one party (eg, the resident), the system automatically creates a prepopulated evaluation for the other party (eg, the attending) and sends a notification to their phone. At that point, both attending and resident use SIMPL to complete identical 3

TABLE 1. The Zwisch Scale and Its Relationship to Meaningful Autonomy and Near-independence

Zwisch Level	Attending Behaviors	Meaningfully Autonomous?	Nearly Independent?
Show and Tell	Performs >50% of critical portion	–	–
Active Help	Leads the resident (active assist) for >50% of the critical portion	–	–
Passive Help	Follows the lead of the resident (passive assist) for >50% of the critical portion	Yes	–
Supervision Only	Provides no unsolicited advice for >50% of the critical portion	Yes	Yes

TABLE 2. The SIMPL Performance Scale and Its Relationship to Competence

Performance Level	Trainee Behaviors	Competent?
Unprepared/Critical Deficiency	Poorly prepared to perform this procedure and/or included critical performance errors that endangered the safety of the patient or the outcome of the procedure.	–
Unfamiliar with Procedure	Frequent problems regarding technique, execution, smoothness, efficiency, and forward planning.	–
Intermediate Performance	Performance of procedural elements is variable but acceptable for the amount of experience with this procedure. Not yet at the level expected for graduating residents.	–
Practice-Ready	Resident is ready to perform this operation safely, effectively and independently assuming resident consistently performs procedure in this manner.	Yes
Exceptional Performance	One of the best performances I have ever seen. Above the level expected of graduating residents.	Yes

question surveys on their mobile devices (see Rating Instruments, above). They independently complete these ratings without knowledge of the rating assigned by the other participant. Attendings can also dictate formative feedback and save it on the system for later review by the resident. Of note, evaluations expire 72 hours after an operation owing to research suggesting that, beyond this point, assessments no longer reliably contain the necessary details about the performance.[23]

Rater Training

All eligible participants were required to attend a 1-hour rater training session to be enrolled in this trial. Initial rater training sessions were led by one of the study investigators. Local study coordinators were trained to lead additional training sessions for those attendings and trainees who missed the initial training. In both cases, the rater training sessions incorporated a standardized curriculum developed for a previous trial that used only the Zwisch scale.[19,24] That curriculum was augmented to included additional information about the more recently developed SIMPL Performance and Complexity scales.

Study participants were given access to SIMPL after a) they had completed the standardized rater training and b) 70% of eligible attendings and 70% of eligible residents within their program had successfully completed the training. These thresholds were chosen to maximize the network effects of SIMPL implementation within individual programs and thereby improve rater engagement.

Procedural Taxonomy

The SIMPL app provides 1686 distinct procedures from which raters can choose when creating their evaluation. This taxonomy is a modified version of the ACGME case logging taxonomy. For purposes of analysis, the SIMPL procedural taxonomy was mapped, wherever possible, to the American Board of Surgery's 2016 to 2017 SCORE procedural taxonomy.[25] The SCORE taxonomy includes 132 ''Core'' procedures and 85 ''Advanced'' procedures. There were 36 SCORE procedures for which mapping to a SIMPL procedure was not possible (20 Core and 16 Advanced). Most of the unmapped SCORE procedures were gynecologic procedures, genitourinary procedures, or nonoperative skills that do not exist in the SIMPL procedural taxonomy (eg, ''Defibrillation and Cardioversion'' and ''Oxygen Administrative Devices''). In the other direction, a total of 148 SIMPL procedures (8.8%) could not be mapped to a SCORE procedure. These were classified as ''non-SCORE.''

Statistical Analysis

Data were analyzed using descriptive statistics. Subset analyses were performed to specifically examine ratings associated with Core procedures and/or ratings collected for trainees in their final 6 months of residency training. These circumstances were hypothesized to represent the scenario within which a trainee was most likely to be deemed Competent and/or experience Meaningful Autonomy. Correlations between Performance and Zwisch (Autonomy) ratings were also calculated using Pearson r statistics.

Additional model-based analyses were conducted to better understand the role of program, rater, trainee, and other factors contributing to the performance and autonomy scores. Bayesian ordinal mixed models[26,27] were used to predict Performance and Autonomy scores respectively. Using these models we controlled for trainee PGY, linear time trends (each week of an academic year during the study period), the number of trainees present for the case, and the relative patient-related complexity of the procedure (as judged by the rater). In addition, random effects were included to control for rater stringency, individual trainee characteristics, individual program characteristics, and the type of procedure performed. Finally, cluster-level effects were estimated for a procedure being Core versus non-Core. Similar to a standard regression model, the product of that analysis is a set of coefficients for each of the model parameters. For this mixed-model there is 1 coefficient for each level of the fixed effects and 1 coefficient for each possible value of the random effects. Such a model allows for predictions to be made for Core procedures performed at the end of PGY5 while holding other effects constant. In other words, one can see end-of-training expectations for the typical program, average trainee, average rater, etc.

There were only 2 Performance ratings (0.02% of the sample) at the lowest level of ''Unprepared/Critical Deficiency.'' These 2 ratings were collapsed to the subsequent category (''Unfamiliar with Procedure'') for all model-based analyses.

RESULTS

Sample Description

Subjects

Fourteen GS residency programs took part in this study. All enrolled programs were university-based. In the first half of the study (during the 2015–2016 academic year), 872 faculty and 511 categorical residents within these programs were eligible to participate. In the second half of the study (during the 2016–2017 academic year) 892 faculty and 515 residents were eligible to participate. Data were collected for procedures that occurred over the course of 16 months starting in September 2015. During the study period, a total of 444 attendings rated 536 categorical GS residents after 10,130 procedures.

Procedures

Ratings spanned 332 different types of procedures. 73.4% (7437) of ratings were of procedures categorized as ''Core'' to GS while 10.9% (1104) were categorized as ''Advanced.'' The remaining

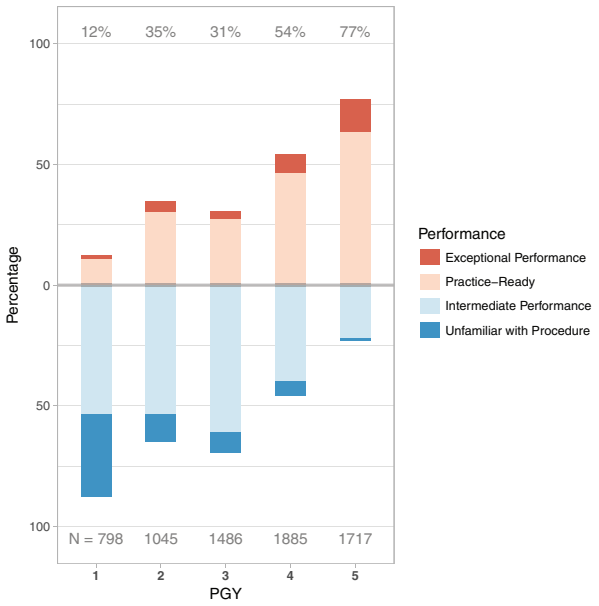

FIGURE 1. The relationship between resident post-graduate year (PGY) and the distribution of unadjusted operative Performance ratings for all residents performing Core procedures. Proportions are aligned such that Performance ratings indicating Competence (Practice-Ready and Exceptional Performance) are above zero. The total number of ratings for each group is shown at the bottom of the plot.

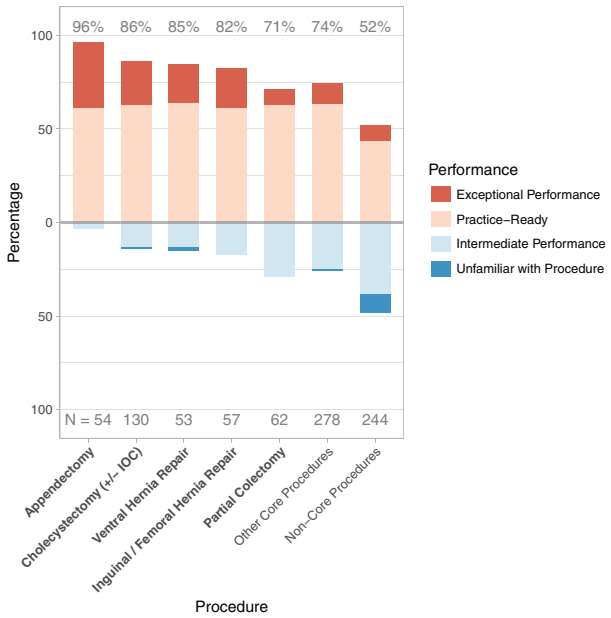

FIGURE 2. The distribution of trainee Performance ratings for trainees in the last 6 months of their residency training while performing the 5 most frequently rated American Board of Surgery-defined Core procedures (in bold). Performance ratings for the remaining Core procedures as well as ratings for non-Core procedures are also included. Proportions are aligned such that Performance ratings indicating readiness for independent practice (Practice-Ready and Exceptional Performance) are above zero. The total number of ratings for each group is shown at the bottom of the plot.

15.7% (1589) of ratings were for procedures not included in the SCORE curriculum. The 5 most frequently rated Core procedures were cholecystectomy, inguinal/femoral hernia repair, appendectomy, ventral hernia repair, and partial colectomy. A combined rating was used for the 13% of cases that included multiple procedures. On average, trainees received 18.9 ratings each (median 10, maximum 202).

Trainee Performance

Descriptive Results

Unadjusted trainee Performance ratings increased in each year except from PGY 2 to PGY 3. Tabulated across the entire academic year, PGY 5 trainees were rated as Competent (Practice Ready or above) for 77.1% of Core cases (Fig. 1). When restricted to the final 6 months of training and still limited to Core procedures, that proportion rises to 80%. For the 5 most frequently rated Core procedures performed in the last 6 months, PGY 5 trainees were rated as Competent for 84.3%. For less frequently rated Core procedures performed during that same time period, PGY 5 trainees were rated as Competent for 74.5% (Fig. 2).

Model-based Results

Figure 3 shows the estimated coefficients for the fixed and random effect covariates in a mixed model used to predict resident Performance under various conditions. Changes in PGY have the largest positive impact on estimated performance. Positive effects were also found for week within year, which is an additional and independent effect on top of being promoted from 1 year to the next and represents changes in performance throughout the year. Having multiple procedures in a single case tends to decrease rated performance, as does increased patient complexity. Having multiple trainees in a case does not significantly alter rated performance, in contrast to its estimated effect on trainee autonomy (see the model-based results for Autonomy, below). Analysis of the random effects demonstrates that the attending raters account for the most rating-to-rating variation. The effect of the specific program is negligible.

By the end of training, the predicted probability that a trainee would be rated as Competent after performing a Core procedure is 90.5% (95% CI: 85.7%–94%). This estimate must be interpreted with caution, as it assumes a patient of average relative complexity, a Core procedure of average absolute difficulty, and further controls for other factors such as trainee ability and rater stringency. When these conditions do not hold ratings vary substantially (see Fig. 4 for some examples). For non-Core procedures the predicted probability of being rated as Competent is 86.1% (95% CI: 80.1%–90.5%). Note that after using this model to control for confounding factors (including the difficulty of the rated procedures), there is no longer a plateau in Performance between PGY 2 and 3.

Trainee Autonomy

Descriptive Results

Comparison of the descriptive and model-based analyses of the overall Performance and Autonomy ratings (Fig. 4) suggests that Performance and Autonomy are highly correlated. This qualitative assessment was confirmed by calculating the correlation of all observed Performance and Autonomy measures (Pearson r = 0.66). The correlation of these ratings for PGY 5 residents

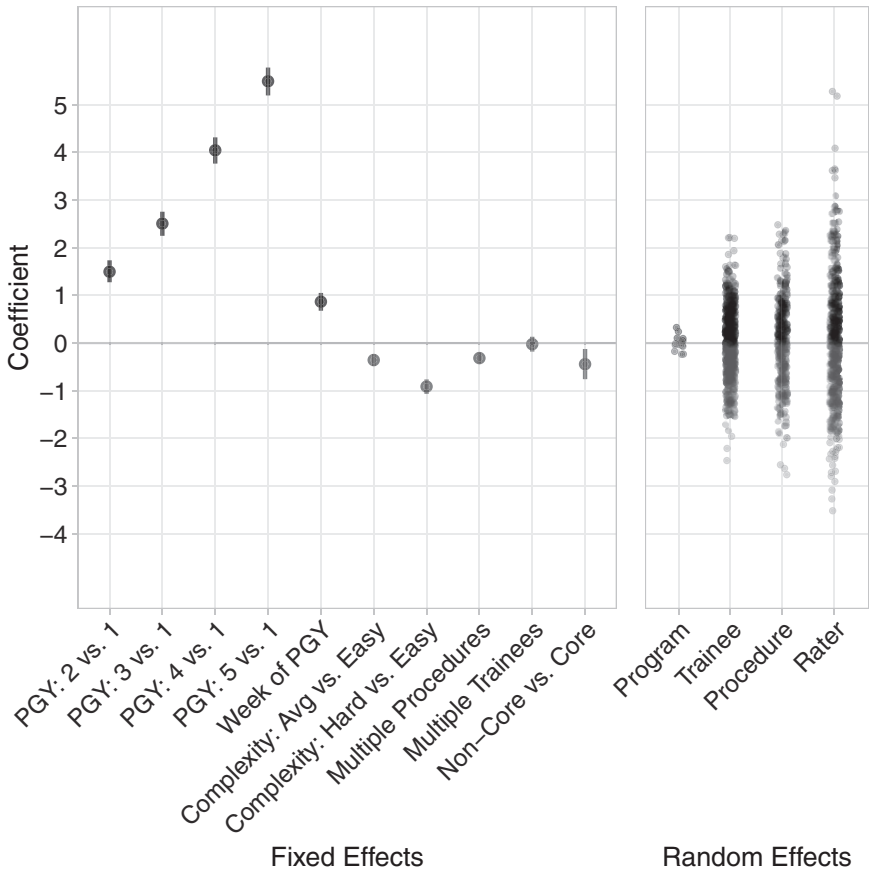

FIGURE 3. The fixed effect coefficients (on the left) and random effect coefficients (on the right) for the covariates used in a Bayesian model used to predict trainee Performance. The coefficients for the fixed effects are similar to the coefficients for a standard regression model in that there is one coefficient for each level of the fixed effects. For the random effects, however, there are many more coefficients because these variables have many more levels. As a result the random effect coefficients are difficult to include on a single plot except as a group (as we have done here). One can roughly compare the influence of fixed effects to random effect by comparing fixed effect estimates with the standard deviation of the distribution of random effect estimates. ''Week of PGY'' is scaled to represent the effect over an entire year. Note that the random effects have been jittered to improve the visualization of the distribution.

specifically was nearly identical (Pearson r = 0.64). As such, the results for Autonomy closely parallel those for Performance.

Over the course of their final 6 months in training, categorical GS residents experienced Meaningful Autonomy (corresponding to Zwisch ratings of Passive Help or Supervision Only) for 69.3% of all cases. When only analyzing Core procedures, this rises to 77.4% (Fig. 5). When the analysis is restricted only to include the 5 most frequently rated Core procedures performed during the last 6 months of training, PGY 5 residents experienced Meaningful Autonomy for 84% of procedures. For less frequently rated Core procedures during that same time period, PGY 5 trainees were rated as Meaningfully Autonomous for 69.1% (Fig. 6).

We also examined how frequently residents achieved Near Independence (ie, the highest Zwisch level, Supervision Only) during the last 6 months of training. During this time period, PGY 5 residents experience Near Independence for 27.1% of all the procedures, 33.3% of Core procedures, and 40.3% of the 5 most frequently rated Core procedures (also in Fig. 6).

Model-based Results

Similar to the analysis for Performance, Figure 7 shows the estimated regression coefficients for the fixed and random effect covariates in a mixed model used to estimate Autonomy. Changes in PGY have the largest positive effect. Positive effects were also found for week of the academic year in which the procedure took place. This effect of timing within the academic year is an additional and independent effect on top of trainee promotion from 1 year to the next. Having multiple procedures in a single case tends to decrease rated Autonomy, as does increasing complexity. Attending raters and type of procedure account for more rating-to-rating variation than the individual trainee and the effect of the program is negligible. In contrast to estimations for Performance, having multiple trainees in a case increases the expected amount of trainee Autonomy.

By the end of training, the predicted probability that a trainee would experience Meaningful Autonomy for a typical Core procedure (ie, average difficulty Core procedure on an average complexity patient) is 91.4% (95% CI: 87.7%–94.2%) (Fig. 4, above). For relatively more complex patients (ie, Hardest 1/3), end-of-year estimated Meaningful Autonomy falls to 79.5% (95% CI: 72.2%–85.4%) while for less complex patients (ie, Easiest 1/3) it rises to 94.9% (95% CI: 92.3%–96.6%). End-of-year estimated Meaningful Autonomy for the 5 most frequently rated Core procedures is shown in Figure 8. For comparison this figure also includes data for those procedures rated at least 10 times and with the least amount of Meaningful Autonomy.

DISCUSSION

The results of this study demonstrate that, while both resident performance and autonomy increased over the course of training, gaps remain. For example, when examining the ratings for all Core procedures performed in the final 6 months of residency residents were deemed Competent and reached Meaningful Autonomy for 80% and 77.4% of cases, respectively. These proportions are lower for Core procedures performed less frequently. After using a model-based approach to adjust for potential confounding variables, our results predict somewhat higher proportions although those estimates assume a hypothetical ''average'' procedure, patient, and trainee.

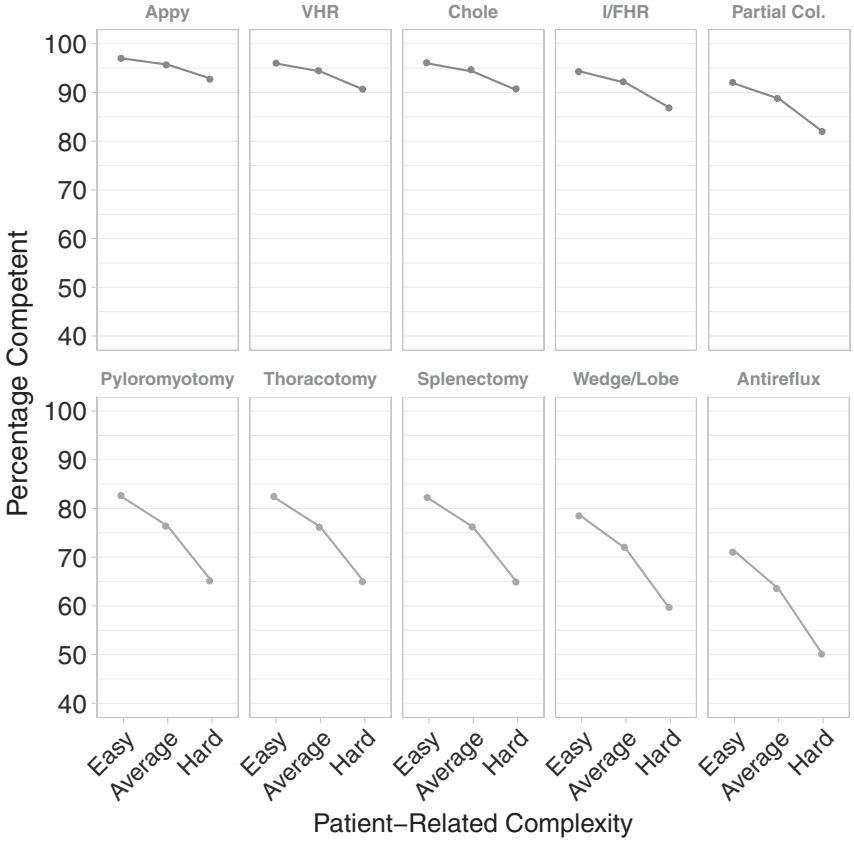

FIGURE 4. Estimated end-of-year Competence for specific procedures, stratified by relative patient-related complexity and controlling for covariates such as trainee variation and rater stringency. The top row displays the results for the 5 most frequently rated Core procedures while the bottom row includes results for the 5 lowest rated Core procedures with at least 10 ratings. Note that Competence includes Performance ratings at practice ready or above. Appy indicates appendectomy; VHR, ventral hernia repair; Chole, Cholecystectomy (+/− intraoperative cholangiogram); I/FHR, Inguinal/Femoral Hernia Repair; Partial Col, Partial Colectomy; Thoractomy, Exploratory Thoracotomy – Open and Thoracoscopic; Splenectomy, Splenectomy/Splenorrhaphy (Trauma); Wedge/Lobe, Partial Pulmonary Resection—Open and Thoracoscopic; Antireflux, Antireflux Procedures.

The expected proportion falls substantially when these assumptions do not hold true.

The other major finding from this study is that maximum autonomy (Near-Independence, ie, Zwisch level Supervision Only) is achieved for only a fraction of Core procedures, even for trainees in the last 6 months of training. It is now rare for residents to operate without the faculty directly participating in the case.

Implications

The ACGME has outlined 132 procedures as being Core to the practice of GS and expects that trainees reach competence in these procedures prior to graduation from residency. To this end, the ACGME has defined a series of GS ''Milestones'' that programs must use to assess the competence of residents as they progress through training.[28] For the operative performance Milestone, the graduation target is level 4 which indicates that ''the resident can perform *most* of the Core operations.'' Our findings suggest that residents do not always meet that standard. Residents appear to more typically achieve Milestones level 2 or 3 (*some* or *many* Core procedures) in the final year of training. Depending on how one defines ''most,'' this is true even for PGY 5 residents performing any of the top 5 most commonly rated Core procedures. This suggests that either the format of training must be modified, standards for what defines competence must be adjusted, or expectations of what constitutes GS must change.

Significance and Relationship to Other Research

This study overcomes some of the limitations of previous work. It is the largest study to date describing the progression of trainee operative performance and autonomy during GS residency. Data were derived from a prospectively collected, multi-institutional database that includes ratings from a large group of residents for numerous procedures across multiple residency training programs. Furthermore, assessment was performed in line with recently published practice guidelines for the assessment of trainee operative performance.[29] Those guidelines recommend that rating instruments are short, include global items for performance and autonomy, and be aligned with the clinician's way of thinking about performance. These recommendations naturally dovetail with recent suggestions from Ten Cate et al[30] that assessing autonomy may be a better method of assessing performance.

This study builds on other research also suggesting that GS residents are not universally ready for independent practice. The best evidence in this regard comes from a survey of programs directors performed by Mattar et al in 2012.[4] In that research, 26.1% of the respondents did not believe that incoming fellows could take GS call with rare need for assistance with cases. In that same study, 17.8% of program directors did not believe that incoming fellows could independently perform a laparoscopic cholecystectomy. A national survey of US surgical residents in 2009 revealed that 27.5% were concerned that they would not be prepared to practice independently upon graduation.[31] Finally, in a national survey of practicing surgeons administered in 2011, 21% of younger surgeons (age less than 45) reported pursuing fellowship training in part because they did not feel ready to practice.[5] The magnitude of these results is consistent with the findings in the current study and supports our conclusions about proficiency.

There is also a growing belief that the current generation of trainees experience less autonomy than prior generations.

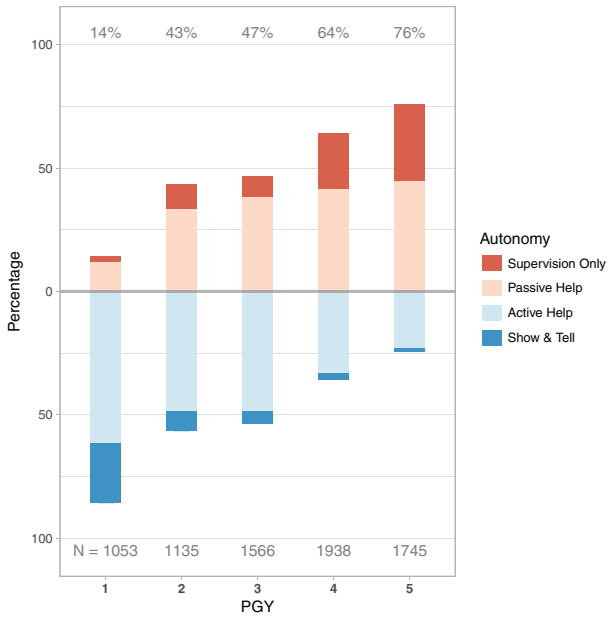

FIGURE 5. The relationship between resident postgraduate year (PGY) and the distribution of unadjusted operative Autonomy (Zwisch) ratings for all residents performing Core procedures. Proportions are aligned such that Autonomy ratings indicating Meaningful Autonomy (Active Help and Supervision Only) is above zero. The total number of ratings for each group is shown at the bottom of the plot.

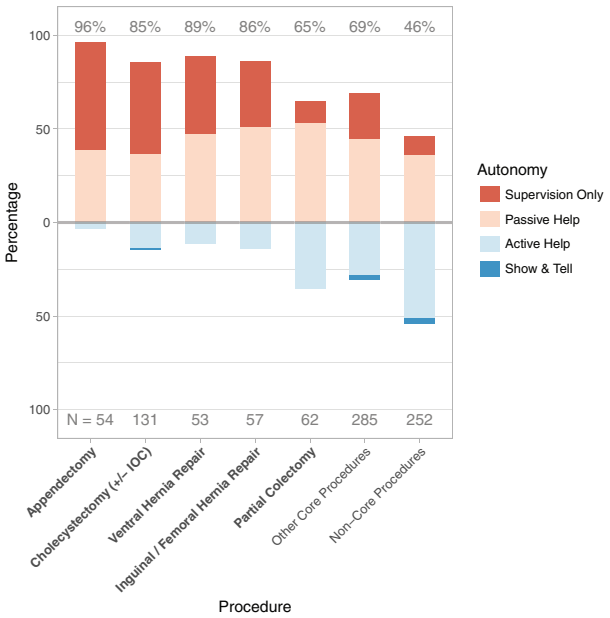

FIGURE 6. The distribution of unadjusted trainee Autonomy (Zwisch) ratings for trainees in the last 6 months of their residency training while performing the 5 most frequently rated American Board of Surgery-defined Core procedures (in bold). Zwisch ratings for the remaining Core procedures as well as ratings for non-Core procedures are also included. Proportions are aligned such that Zwisch ratings indicating Meaningful Autonomy (Passive Help and Supervision Only) are above zero. The total number of ratings for each group is shown at the bottom of the plot.

Unfortunately, there are few empiric data to confirm this belief. A preliminary examination using NSQIP data does suggest that less time is being dedicated to teaching, which may also indicate less autonomy.[32] In a qualitative analysis performed as part of Mattar et al's study described above, many program directors believed that there was a "generalized lack of autonomy and independence during residency that delayed progress, or at least required a "catching-up" phase at the beginning of fellowship."[4] It is unknown if residents directly entering independent practice face similar challenges in their early career.

It remains unclear how much autonomy is necessary for trainees to achieve competence prior to graduation. We can only speculate that the "correct" amount of autonomy is that which enables residents to learn the skills necessary to operate independently without exposing the patient to undue risk. As any teaching surgeon can attest, achieving this balance is difficult.

Alternative Interpretations

Some might consider the levels of competence and autonomy achieved by residents, especially for the most common Core procedures, to be reasonable. One might also be reassured that approximately 80% of all graduating residents pursue additional clinical training in the form of a fellowship.[33] Still, that leaves 20% of graduating trainees to directly enter independent practice. Our data suggest that some of these individuals may not have achieved Competence for a substantial fraction of even the most frequently performed Core procedures. We can only hope that trainees who choose not to pursue additional training are those who are more prepared for independent practice. Self-reported trainee data does provide some support for that belief[33] although these opinions have not been confirmed empirically.

Addressing the Challenges

The descriptive and model-based analyses demonstrate that the pace of progression of both autonomy and performance is too slow to reliably achieve training goals within the 5 years of GS residency. The performance deficits may be due in part to the limited autonomy residents receive during training. Alternatively, the limited autonomy may be the result of inadequate trainee performance. Most likely there is a causal interplay in both directions since autonomy is both an input to learning and the consequence of improving skills.

While inadequate trainee performance may contribute to reduced autonomy, there may be modifiable factors that could be addressed to break this cycle. First, there are numerous systemic demands on residents and attendings that make it hard to provide intermediate residents with increased autonomy. For example, most health care systems encourage attendings to finish cases quickly and many surgical departments base faculty salary and/or bonuses on clinical productivity. Cases performed with residents are well known to take longer than those without residents, and attendings may attempt to mitigate that effect by maintaining operative leadership and control.[34,35] Similarly, expanding service demands on residents constrained by the 80-hour work week may adversely affect trainee time spent in the OR, especially as a second assistant. This may be especially true in the intermediate years, since senior and chief residents' time in the OR is typically more protected. Another possible explanation is that faculty are not trained (and some may not be comfortable) providing trainees with meaningful autonomy even when those residents are otherwise ready for increased responsibility. This may be especially true for junior faculty. Again, these

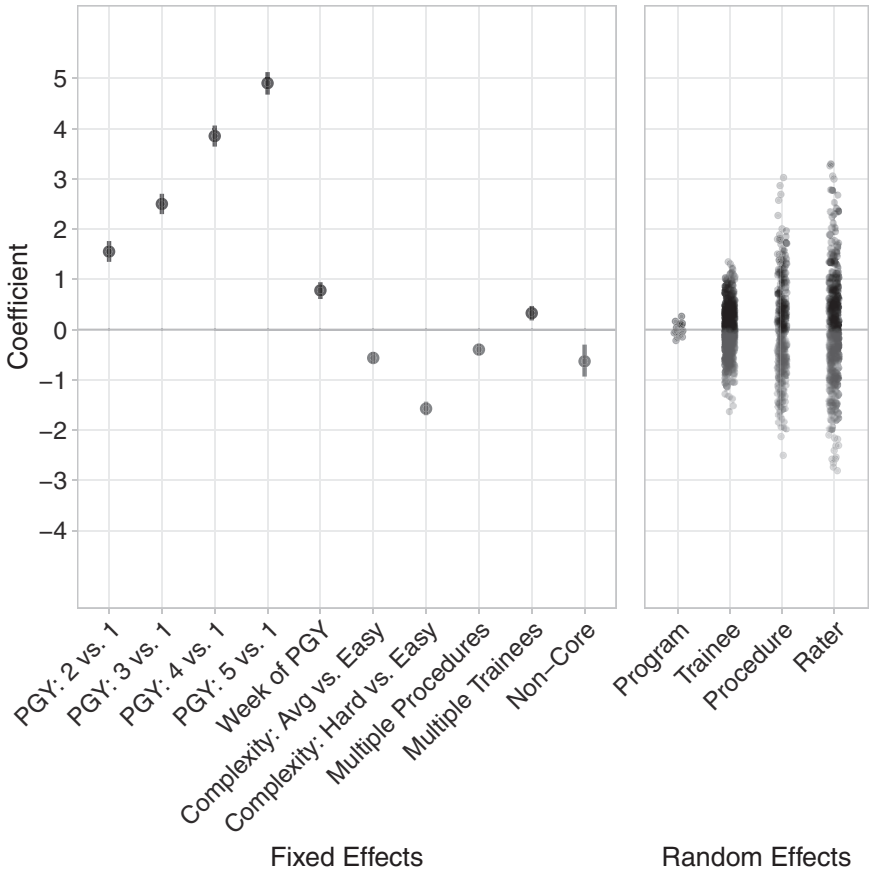

FIGURE 7. Coefficients for the covariates used in a Bayesian model used to predict trainee Autonomy. The fixed effect coefficients are shown on the left and the random effect coefficients are shown on the right. ''Week of PGY'' is scaled to represent the effect over an entire year. Note that the random effects have been jittered to improve the visualization of the distribution. See Figure 3 for a more detailed explanation about how to interpret this plot.

challenges may disproportionately affect more junior residents as senior residents often have the administrative power to select the most experienced teachers for themselves. These and other factors might contribute to trainees, particularly in the intermediate years, not being granted the autonomy that they might otherwise reasonably achieve if such pressures did not exist. This in turn may slow the progression of their learning, ultimately threatening their ability to achieve meaningful autonomy and competence prior to graduation.

There are no easy solutions, but wider implementation of previously described interventions may help. For example, operative performance might be improved by providing residents the opportunity to use simulation to rehearse component skills before applying their knowledge in the OR. This could be expected to support a virtuous cycle whereby trainees achieve more autonomy earlier and thereby further accelerate their learning. Operative learning can likely also be improved with increased preoperative communication between resident and attending surgeon. Specifically, improvements in the resident educational experience can be realized when trainees discuss their learning and performance goals with attendings prior to cases.[36] These strategies could be taught to both trainees and attendings.

Limitations

We recognize several limitations. This study only includes data from residents training at large academic medical centers and as such our results may not be generalizable to other training settings. Resident and attending participation at each enrolled site was voluntary and so there may be a bias introduced into our results due to the nonrandom subset of attendings and residents who chose not to participate. Furthermore, most residents received ratings of their performance for only a minority of the procedures in which they participated. The procedures chosen for assessment may be different than those that were not. If true, we would hypothesize that these biases would lead to falsely elevated average ratings. In addition, there was wide variation in the number of assessments performed by both attendings and residents. Our results may therefore be more heavily influenced by data from more frequently rated trainees. Furthermore, current analysis disregards demographic information about residents and raters, something we hope to incorporate in future studies.

Our results from the model-based analysis indicate that residents' skills have less impact on ratings than the rating habits of the attending. To overcome this, greater efforts will be needed to set benchmarks of performance at different stages in training and to disseminate a common vocabulary and shared expectations around which residents' technical performance can be more reliably assessed.

Finally, one must be careful when interpreting the model-based results. These predictions control for various factors that, in practice, may vary substantially from day to day. For example, an early-career surgeon would expect to see diverse types of patients and perform many different procedures each with varying difficulty. Most of our model-based estimates, in contrast, are for a typical trainee performing a hypothetical typical procedure for a typical patient. The prediction results included in this study can only be interpreted within that narrow context.

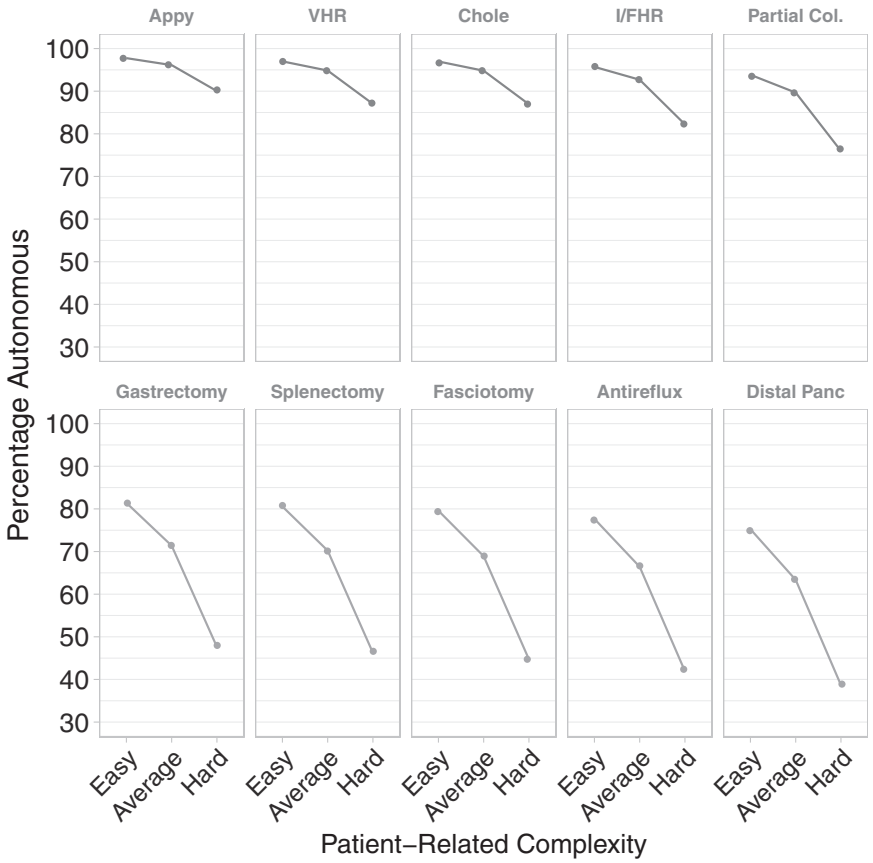

FIGURE 8. Estimated end-of-year meaningful autonomy for specific procedures, stratified by relative patient-related complexity and controlling for covariates such as trainee variation and rater stringency. The top row displays the results for the 5 most frequently rated Core procedures while the bottom row includes results for the five lowest rated Core procedures with at least 10 ratings. Note that meaningful autonomy includes Zwisch ratings at passive help or above. Appy indicates Appendectomy; VHR, ventral hernia repair; Chole, Cholecystectomy (+/− intraoperative cholangiogram); I/FHR, inguinal/femoral hernia repair; Partial Col, partial colectomy; Gastrectomy, Partial or Total Gastrectomy; Splenectomy, Splenectomy/Splenorrhaphy (Trauma); Fasciotomy, Fasciotomy for Injury (Trauma); Antireflux, Antireflux Procedures; Distal Panc, Distal Pancreatectomy.

Future Directions

Improving Education Through Feedback and Assessment

We believe that the ubiquitous presence of smartphones and the ease of use of mobile software will continue to support large-scale research of this type. Other, equally important, benefits of this technology are purely educational. We do not explore these issues here, but smartphone assessment applications may enable more immediate and meaningful feedback between resident and attending, whether as ratings or as dictated comments. Furthermore, changing the training culture to include more frequent assessment of proficiency and autonomy may also affect the preoperative conversation and enable everyone to more clearly define trainees' learning objectives. Reflecting these views, other specialties have begun to use similar methods to assess and give feedback to their own trainees.[37,38]

Developing Standards for Competency-based Education

Quality improvement (QI) methodologies were originally developed in the manufacturing industry but are increasingly being applied to health care and, more specifically, surgery.[39] QI methodologies have, however, rarely been applied to medical education. This is in part due to the lack of objective trainee performance data. This barrier might be overcome by leveraging mobile ''micro-assessment'' tools similar to those described herein. The ultimate outcome of applying QI methodologies to surgical education would be rigorously defined standards of operative performance.

New standards are needed because we can no longer presume that all surgical residents are competent by the end of training. Fortunately, the tools to assess trainee competence are maturing. Given this context, it may be feasible to more broadly perform operative competency assessment rather than simply measuring trainee case volume. High-frequency operative performance rating data could ultimately be used to set standards of operative performance and, similarly, ensure that they are being met. It could also provide early-career surgeons better awareness of their limitations and provide them with information about when they should ask for help. This represents a clear path toward a competency-based educational system in GS that situates residency training within the continuum of surgical learning.

CONCLUSIONS

US GS residents are not universally ready to independently perform the most common Core procedures by the time they complete residency training. Significant gaps remains for less common Core and non-Core procedures. Resident autonomy is also limited and may make it more difficult to ensure resident readiness for independent practice. Future efforts should focus on establishing measurable standards of operative performance that should be met by graduating surgical residents.

ACKNOWLEDGMENTS

The authors thank the program directors, coordinators, faculty, and residents from the participating programs for making this study possible. They also wish to thank all the institutional members of PLSC for their help in designing, developing, and supporting SIMPL.

REFERENCES

1. Bell RH. Why Johnny cannot operate. *Surgery*. 2009;146:533–542.
2. Lewis FR, Klingensmith ME. Issues in general surgery residency training—2012. *Ann Surg*. 2012;256:553–559.
3. Elfenbein DM. Confidence crisis among general surgery residents: a systematic review and qualitative discourse analysis. *JAMA Surg*. 2016;151:1166–1175.
4. Mattar SG, Alseidi AA, Jones DB, et al. General surgery residency inadequately prepares trainees for fellowship: results of a survey of fellowship program directors. *Ann Surg*. 2013;258:440–449.
5. Napolitano LM, Savarise M, Paramo JC, et al. Are general surgery residents ready to practice? A survey of the American College of Surgeons board of governors and young fellows association. *J Am Coll Surg*. 2014;218:1063–1072.
6. Borman KR, Vick LR, Biester TW, et al. Changing demographics of residents choosing fellowships: Longterm data from the American Board of Surgery. *J Am Coll Surg*. 2008;206:782–788.
7. Coleman JJ, Esposito TJ, Rozycki GS, et al. Early subspecialization and perceived competence in surgical training: are residents ready? *J Am Coll Surg*. 2013;216:764–771.
8. Fronza JS, Prystowsky JP, DaRosa D, et al. Surgical residents' perception of competence and relevance of the clinical curriculum to future practice. *J Surg Educ*. 2012;69:792–797.
9. Friedell ML, VanderMeer TJ, Cheatham ML, et al. Perceptions of graduating general surgery chief residents: are they confident in their training? *J Am Coll Surg*. 2014;218:695–703.
10. Soper NJ, DaRosa DA. Presidential address: engendering operative autonomy in surgical training. *Surgery*. 2014;156:745–751.
11. Halpern SD, Detsky AS. Graded autonomy in medical education—managing things that go bump in the night. *N Engl J Med*. 2014;370:1086–1089.
12. Malangoni MA. Protecting patients while advancing education. *J Am Coll Surg*. 2014;219:787.
13. Rankin JS. William Stewart Halsted. *Ann Surg*. 2006;243:418–425.
14. Kennedy TJ, Regehr G, Baker GR, et al. Progressive independence in clinical training: a tradition worth defending? *Acad Med*. 2005;80:S106.
15. Bell BM. Supervision, not regulation of hours, is the key to improving the quality of patient care. *JAMA*. 1993;269:403–404.
16. Kachalia A, Studdert DM. Professional liability issues in graduate medical education. *JAMA*. 2004;292:1051–1056.
17. Teman NR, Gauger PG, Mullan PB, et al. Entrustment of general surgery residents in the operating room: factors contributing to provision of resident autonomy. *J Am Coll Surg*. 2014;219:778–787.
18. Ludmerer KM. *Time to Heal: American Medical Education From the Turn of the Century to the Era of Managed Care EDN Revised edition.* Oxford: Oxford University Press; 2005.
19. George BC, Teitelbaum EN, Meyerson SL, et al. Reliability, validity, and feasibility of the Zwisch scale for the assessment of intraoperative performance. *J Surg Educ*. 2014;71:90–96.
20. Bohnen JD, George BC, Williams RG, et al. The feasibility of real-time intraoperative performance assessment with SIMPL (System for Improving and Measuring Procedural Learning): early experience from a multi-institutional trial. *J Surg Educ*. 2016;73:e118–e130.
21. Williams RG, Sanfey H, Dunnington GL. A controlled study to determine measurement conditions necessary for a reliable and valid operative performance assessment: a controlled prospective observational study. *Ann Surg*. 2012;256:177–187.
22. DaRosa DA, Zwischenberger JB, Meyerson SL, et al. A theory-based model for teaching and assessing residents in the operating room. *J Surg Educ*. 2013;70:24–30.
23. Williams RG, Chen X, Sanfey H, et al. The measured effect of delay in completing operative performance ratings on clarity and detail of ratings assigned. *J Surg Educ*. 2014;71:e132–e138.
24. George BC, Teitelbaum EN, DaRosa DA, et al. Duration of faculty training needed to ensure reliable or performance ratings. *J Surg Educ*. 2013;70:703–708.
25. Resident Education SC on. SCORE: Curriculum Outline for General Surgery 2016–2017. 2016. Available at: http://www.absurgery.org/xfer/curriculumoutline2016–17_book.pdf. Accessed January 1, 2017.
26. Buerkner PC. Brms: an R package for Bayesian multilevel models using Stan [computer program]. *J Stat Softw*. 2016.
27. Carpenter B, Gelman A, Hoffman M, et al. Stan: a probabilistic programming language [computer program]. *J Stat Softw*. 2016;20.
28. Cogbill TH, Malangoni MA, Potts JR, et al. The general surgery milestones project. *J Am Coll Surg*. 2014;218:1056–1062.
29. Williams RG, Kim MJ, Dunnington GL. Practice guidelines for operative performance assessments. *Ann Surg*. 2016;264:934–948.
30. Ten Cate O, Hart D, Ankel F, et al. Entrustment decision making in clinical training. *Acad Med*. 2016;91:191–198.
31. Yeo H, Viola K, Berg D, et al. Attitudes, training experiences, and professional expectations of us general surgery residents: a national survey. *JAMA*. 2009;302:1301–1308.
32. Bohnen JD, Chang DC, George BC. Changing trends in operating room times between teaching and non-teaching cases: less time for learning? *J Am Coll Surg*. 2015;221:S49.
33. Klingensmith ME, Cogbill TH, Luchette F, et al. Factors influencing the decision of surgery residency graduates to pursue general surgery practice versus fellowship. *Ann Surg*. 2015;262:449–455.
34. Advani V, Ahad S, Gonczy C, et al. Does resident involvement effect surgical times and complication rates during laparoscopic appendectomy for uncomplicated appendicitis? An analysis of 16,849 cases from the ACS-NSQIP. *Am J Surg*. 2012;203:347–352.
35. Papandria D, Rhee D, Ortega G, et al. Assessing trainee impact on operative time for common general surgical procedures in ACS-NSQIP. *J Surg Educ*. 2012;69:149–155.
36. Roberts NK, Williams RG, Kim MJ, et al. The briefing, intraoperative teaching, debriefing model for teaching in the operating room. *J Am Coll Surg*. 2009;208:299–303.
37. Kobraei EM, Bohnen JD, George BC, et al. Uniting evidence-based evaluation with the ACGME plastic surgery milestones: a simple and reliable assessment of resident operative performance. *Plast Reconstr Surg*. 2016;138:349e–357e.
38. Kozin ED, Bohnen JD, George BC, et al. Novel mobile app allows for fast and validated intraoperative assessment of otolaryngology residents. *OTO Open*. 2017;1:2473974X16685705.
39. Nicolay CR, Purkayastha S, Greenhalgh A, et al. Systematic review of the application of quality improvement methodologies from the manufacturing industry to surgical healthcare. *Br J Surg*. 2012;99:324–335.

DISCUSSANTS

Dr Ara Darzi (London, UK):

Thank you, Mr President. Firstly, may I congratulate Brian and his colleagues from the Procedural Learning and Safety Collaborative for the significant and informative piece of work that is both substantial and also unique in its findings.

In essence, it does reaffirm the concerns that some general surgery residents may not be ready for independent practice following the completion of surgical training, something we also share on the other side of the pond in recent times. I think the methodology combining residents' performance, autonomy and having regard to case complexity is unique and underpinned with measurement tools that are innovative, simple, and validated such as SIMPL and the Zwisch scale are commendable.

I was taken by the discrepancy between the competence, in other words, the performance, but also the autonomy and the independence of practice. And the point about autonomy, it's a 2-way thing. Firstly, it could be given, not taken. And I think the other issue here is that although 85% of PG-5 are deemed practice ready, why do they not have the confidence for autonomous practice? So I'd like a little bit more about that.

I remember from my own training days, I certainly had a significant amount of autonomy, but was not necessarily practice ready. It seems that we've gone completely to the other side of this argument.

I think it's also worthy to ask about 2 or 3 questions. The first is more of an ethical issue. If this study is published today and the Boston Globe gets it and the front headline is 1 in 5 of our surgical residents are not capable of independent practice, how do you think people or how do you think patients and the public will take that?

The second is the issue which you very eloquently highlighted, the strong correlation between improvement in performance and autonomy based on complexity with the years of training. And that's probably the most positive correlation that goes. The more the years, the more competency. I mean, that's quite obvious.

The question I had is if you happen to be the regulator, considering that 80% of your residents are doing fellowships, would you make the case saying in actual fact we should increase of number of residency trainings from PG-5 to PG-6 or at least make your fellowships as compulsory?

The final question I had is—and I will touch this later—is the issue of measuring competency in terms of operating in the OR and certainly of independent. Do we have any idea what their competencies were in many of the nontechnical skills that is outside the operating theater—managing patients on the ward, doing the ward round, managing patients between ward rounds, and so on and so forth?

But all in all, I really wanted to congratulate you. It's a fantastic study. A huge amount of effort has gone into it. The methodology is very sound and well done and plenty to learn in terms of trying to structure or commission the training years for the future.

Dr Brian George (Ann Arbor, MI):

Thank you. I agree there are real challenges to autonomy. Given my research interest, I try all the time to give my residents autonomy, but it's terrifying. And there are structural impediments, too, including productivity pressures, which are probably one of the biggest challenges. As the attending you have to keep things moving, especially when doing multiple elective cases in a single day.

I also think it is important to acknowledge the ethics of all of this. How do you appropriately grant autonomy—it is granted, as you say—to trainees without really knowing whether they are ready? This is especially challenging in the modern era with such high faculty-to-resident ratios because it is more difficult to engender the trust that might improve the autonomy situation. I see Dr Minter nodding because that is her research. I think that's a very important factor.

While there are a lot of challenges to improving autonomy, it's important to try. Several members in our collaborative feel very strongly that it is difficult to know someone is ready for independent practice until they demonstrate independent practice. Not for every case, but for at least some subset of them at the end of training.

Your second question was about the ethics of making this data public, or maybe it was about the politics of making this data public. I, too, am concerned. I'm a surgeon. I love our profession, and I don't want to give any of us a bad name, but I don't think that's what this data says. I think this data says that for some procedures, residents are competent, but that we still have gaps. I don't think these findings are surprising, and I hope that this data can be used to inform the ongoing debate that we're having within our profession about how to make things better.

Your third question was about fellowships, and if I understood it correctly, you're asking whether they might be protective in some sense and whether they should be required. 80% of graduating general surgery residents do go into fellowship, and I expect that they probably get a lot more autonomy. That was certainly my experience. But it's important not to forget that there are 20% of people that go directly into independent practice. In fact, one of the studies I'm very interested in doing is looking at the early career patient outcomes of those people who do go directly into practice, and eventually we'd like to link those to some of this data and see if we might be able to predict those people that need more training.

Whether fellowship should be required or not is above my pay grade, but I hope not. Five years is already a big ask for a lot of medical students when considering this profession. And I think that if we increase the training requirements, we're going to have a supply problem that we'll then need to address. We will in effect be trading one problem for another.

And finally, your final question is whether it's adequate to just assess operative competency, and I would answer no, it is not. Operating is the core of our profession, so I think it's one of the most important things to assess but it's not everything. I hope that over time we can use tools like SIMPL, or even SIMPL, to assess some of those other dimensions of performance. Thank you.

Dr Richard H. Bell, Jr. (Philadelphia, PA):

Thank you, Dr George. I enjoyed that very much. I think it's really great to see that we are focusing the spotlight on this area, We didn't pay enough attention to it in the past.

I wanted to make a background comment about the 133 cases. That list was not created in a smoke filled room at the American Board of Surgery. It arose from a study that was done by the Board, the RRC and the APDS in 2006 in which we polled all of the program directors in the United States. We sent them a list of 300 and some odd procedures, which is what the Board was collecting in those days, and asked them which of those 300+ procedures they thought were really essential to practice. That's where the original list of 122 cases came from, and they became the essential or the core operations for the SCORE curriculum.

I think it's interesting that the list gets re-evaluated every year and yet it hasn't changed much. It's expanded a little bit up to 133 in the last 10 years. So that makes me think that it is a pretty accurate estimate of the procedures that general surgeons need to be able to do in practice. However, if you look at that list of 133 procedures, what we found when we reported a study in the Annals of Surgery in 2012, was that there are about 20 procedures that are fairly commonly performed during residency. And the five procedures that you chose to examine are all in that top 20. Lap chole is number 1 and the others are numbers 3, 4, 9, and 18. So you're looking at some of the most commonly performed procedures.

After about number 20, the number of repetitions done for these procedures really starts to fall off fast and actually for the bottom 60 cases, the median resident experience was zero in 2004 to 2005 when we first collected the data.

So I would argue that there may be procedures that are done frequently enough that we're going to be able to use tools like this to assess the level of competency, but we're left with a large number of procedures that general surgeons are going to have to do once in a while, which they rarely or never experienced during residency.

So the question I wanted to ask you is: I gather that you looked at all procedures initially. How many of the 133 were represented in your data set? In other words, were there a large number of procedures on that list of 133 that nobody reported ever doing? Thank you.

Dr Brian George (Ann Arbor, MI):

Thank you. I don't know the precise number off the top of my head, but if my memory serves, I believe it was 20 to 30 of the 132 core procedures we didn't observe in our data set. Part of that was actually structural for our study. For example, we didn't collect gynecologic procedures or genitourinologic procedures for general surgery residents which are included in that Core list. So we had no chance to actually collect data on those. But I think in addition to those, there were another 20 procedures that we don't have data on.

I will say that, similar to your results and Dr Malangoni's results, there is a steep drop-off in numbers of performances for these less-frequently performed procedures. One of the last slides I showed demonstrated that there is also a steep drop-off in related competence

for those less-frequently performed procedures. Still, I suspect, there's probably some skill transfer. If I do one splenectomy during training but I'm a master surgeon otherwise, I can probably get through it without difficulty because I know how to operate. But our data doesn't capture that, so I can't tell you the degree of skill transfer.

Dr Selwyn Rogers (Chicago, IL):

The question is really this: For a simple appendectomy, there were still 3% of people not competent. Is the problem the training or is the problem the person?

Dr Brian George (Ann Arbor, MI):

I don't know, but my short answer would be both, I think. There are likely to be people that will require more than 5 years to be trained to competence. At least our data would suggest that. But I also think there's tremendous room for improvement on our end. There are a lot of things we can't change. We can't change productivity pressures, but I think we can train teachers to be better—we all have room to improve. And I also think that we can start to set standards. I said this earlier, but if we have standards that we can measure people against, we can use those as benchmarks throughout training and then as part of the credentialing process as trainees graduate and enter practice. This will allow us to avoid having people enter practice before they're ready.

Dr Steven Stain (Albany, NY):

There is currently a measure by the Residency Review Committee for Surgery regarding independence for practice. The RRC frequently cites programs who do not have a statement for their finishing residents that they are independent for practice. So I think this paper substitutes the faculty's opinion of that versus the program director.

My question is about the volume of surgery that the residents do to get to competence. Do you have any data on whether the residents who had higher rates of competence did more operations? Recognizing this, both the board and the RSC has increased the number of cases required for accreditation of programs starting in 2018 from 750 to 850.

Dr Brian George (Ann Arbor, MI):

The RRC defines the operative performance Milestone level 4 as the graduation target. Level 4 states that trainees can perform ''most'' Core procedures. Our data suggests that most of our trainees don't achieve that level.

In terms of case numbers, we do have a preliminary analysis—we haven't published it yet—looking at learning curves, and people that have logged more cases in the case log are predicted to have more autonomy and competence as they go forward. This is what you'd expect and I don't think that's surprising. We don't yet have a real sense of what number is required to achieve specific levels of competence, but our near-term goal is to establish case number requirements based on observed performance data. The Holy Grail is to follow residents out into practice, look at their early career outcomes, link that back to their learning outcomes during residency, and finally use that to calculate a standard of performance that would result in optimal patient outcomes in the future. I think given the quantity of data that we're able to collect that that is a feasible approach. I hope as a profession we can move in that direction over the next decade.

Dr Mary Klingensmith (St. Louis, MO):

Very important work. Thank you. A question about this as a tool for faculty development. Do you know over time did some faculty change their behavior as a result of using the app? Similarly, could you disclose this information to programs and let the programs sort of reveal how many faculty are allowing more of this passive help and supervision only?

Dr Brian George (Ann Arbor, MI):

I have definitely seen changes in behavior, both personally as a trainee and also now as a faculty member. Once you start doing frequent assessment of autonomy and performance, it starts a conversation—especially if you're doing it after every case. What that does is it starts to change the culture. And we have seen this in our participating institutions. I have overheard people in the hall who are not part of this project talking about what Zwisch level they were aiming for during the case. I think that's progress. I also think that changing the culture around assessment and setting targeted learning goals is a great way to make us even better teachers.

Dr Rebecca Minter (Dallas, TX):

Beautifully presented, and congratulations on your work.

To build on the prior comment by Dr Klingensmith, is the 22% of residents that are rated as being ready for Supervision Only related to resident ability, or is it more likely related to the need for faculty and resident development? I think that's really a critical question that we need to consider. Can you reflect on that please?

The second question was just methodological. As I look at the core procedures that you selected, and I think back in recent years of how frequently I've done an inguinal hernia or a laparoscopic cholecystectomy with a chief resident — particularly in their last 6 months, it's probably single digits. So I wonder, do you know in your data set what are the total numbers of these core procedures that are being done and how often are they being done by the chief residents? Is that something we need to think about going forward? Do we have the ability to assess Chief Residents performing those core procedures?

Dr Brian George (Ann Arbor, MI):

The top 5 procedures constitute 70% of our data set. In fact, there's enough data that I could look at the last 6 months only and restrict the analysis to that.

Dr Rebecca Minter (Dallas, TX):

But what is the frequency for the chief resident performing that operation versus the junior residents?

Dr Brian George (Ann Arbor, MI):

There are, yes. Of those top 5 procedures, I think in total there were 500 observed performances in the last 6 months.

You also asked if this is the rater or if it is the trainee, and I will also add that we performed an adjusted analysis of autonomy as well as performance. If you adjust for rater stringency the predicted rates of autonomy varied widely. The one that shocked me was for partial colectomy. Residents are predicted to have a 20% chance of having Supervision Only conditions in their last day of residency training when supervised by a ''typical'' attending. So I don't think it's just the raters or just the trainees. I think it's probably the entire system, and we need to address it.

Dr Carla Pugh (Madison, WI):

Thank you. My questions relate to the potential use of this data from a modeling perspective. How many of the persons that were shown to be incompetent for appendectomy were also incompetent for colorectal and other procedures?

And the second question is, what potential is there to use this database from a predictive standpoint for those in their second and

third year so that we can intervene well before they get to their chief year?

Dr Brian George (Ann Arbor, MI):

In our model we included both trainee and procedure and it is possible to look at an individual's predicted performance across different procedures. We could also do that early in training. The model fit is quite good across all PGY levels, so we can easily plug in data for a PGY-2 trainee and use the model to make a prediction about where they might end up by the end of training. If they are an outlier then we could remediate them earlier, before they have invested four or four-and-a-half years of their life in the process.

Dr John Mellinger (Springfield, IL):

Great presentation. It sounded like 90% of your folks did some kind of frame of reference training ahead of time. You know from Reed William's prior work with our OPRS tool from SIU that about two-thirds of the variation in these numeric are Likert type or, in this case, Zwisch Scale assessments comes from the faculty member. His work suggested that ten separate faculty had to evaluate a resident over the course of a year to have high reliability. It didn't look to me from your numbers like there were very many residents who got ten different faculty to assess them. Can you just comment on that methodologic feature?

Lastly, you have this rich database of comments. I'd be very interested if you're considering doing a qualitative analysis of those which might add a lot of color and texture to the sketch you've given by the reported scales. Thank you very much.

Dr Brian George (Ann Arbor, MI):

Thank you. I'll answer the last one first. We do have a very rich data set with comments, and, unfortunately, that is outside of the scope of this study. There is, however, someone in our collaborative working on analyzing those at the single-institution level and I hope beyond.

In regards to rater reliability, that is a real issue. In fact, Dr Reed Williams that you just mentioned is going to be publishing another paper that suggests you need 60 evaluations over the course of a year to produce reliable estimates of trainee operative performance. That's a lot. And we don't have that number in our study. But that's one reason we used a model-based approach for part of this analysis, to adjust for some of those rater idiosyncrasies and increase the precision of our estimates. But overall we really need more data.

I hope that as a profession we can move more towards a continuous assessment model using a very simple assessment tool. It doesn't have to be fancy to provide substantial benefits. This type of assessment model increases the amount of data available for not only assessment purposes but also educational purposes. Thank you very much.