Development and Validation of Cystectomy Assessment and Surgical Evaluation (CASE) for Robot-Assisted Radical Cystectomy for Men

Ahmed A. Hussein[1,2,*], **Kevin J. Sexton**[1,*], **Paul R. May**[1], **Maxwell V. Meng**[3], **Abolfazl Hosseini**[4], **Daniel D. Eun**[5], **Siamak Daneshmand**[6], **Bernard H. Bochner**[7], **James O. Peabody**[8], **Ronney Abaza**[9], **Eila C. Skinner**[10], **Richard E. Hautmann**[11], **Khurshid A. Guru**[1]

[1]Roswell Park Cancer Institute, Buffalo, NY

[2]Cairo University, Cairo, Egypt

[3]University of California, San Francisco, CA

[4]Karolinska University Hospital, Stockholm, Sweden

[5]Temple University Hospital, Philadelphia, PA

[6]University of Southern California, Los Angeles, CA

[7]Memorial Sloan Kettering Cancer Center, New York, NY

[8]Henry Ford Health Systems, Detroit, MI

[9]OhioHealth, Dublin, OH

[10]Stanford University, Palo Alto, CA

[11]University of Ulm, Ulm, Germany

Abstract

Introduction—We aimed to develop a structured scoring tool: CASE (Cystectomy Assessment and Surgical Evaluation) that objectively measures and quantifies performance during robot-assisted radical cystectomy (RARC) for men

Methods—A multinational 10-surgeon expert panel collaborated towards development and validation of CASE. The critical steps of RARC in men were deconstructed into 9 key domains, each assessed by 5 anchors. Content validation was done utilizing the Delphi methodology. Each anchor was assessed in terms of context, score-concordance, and clarity. The Content Validity Index (CVI) was calculated for each aspect. A CVI $\geq$ 0.75 represented consensus, and this statement was removed from the next round. This process was repeated until consensus was achieved for all statements. CASE was used to assess de-identified videos of RARC to determine reliability and construct validity. Linearly weighted percent agreement was used to assess inter-rater reliability (IRR). A logit model for odds ratio (OR) was used to assess construct validation.

Corresponding Author Khurshid A. Guru, MD, Department of Urology, Roswell Park Cancer Institute, Elm and Carlton Streets, Buffalo, NY 14263, USA. **Tel:** +1 716 845 3389, **Fax:** +1-716-845-3300, Khurshid.Guru@roswellpark.org.
*Ahmed A. Hussein and Kevin J. Sexton had equal contribution.

Results—The expert panel reached consensus on CASE after 4 rounds. The final 8 domains of the CASE included: Pelvic Lymph Node Dissection, Development of the Peri-ureteral Space, Lateral Pelvic Space, Anterior Rectal Space, Control of the Vascular Pedicle, Anterior Vesical Space, Control of the Dorsal Venous Complex, and Apical Dissection. IRR > 0.6 was achieved for all eight domains. Experts outperformed trainees across all domains, but significance was not reached.

Conclusion—We developed and validated a reliable structured, procedure-specific tool for objective evaluation of surgical performance during RARC. CASE may help differentiate novice from expert performances.

Introduction

Radical cystectomy (RC) with pelvic lymph node dissection (LND) represents the gold standard for management of non-metastatic muscle invasive bladder cancer (MIBC) and refractory non-muscle invasive disease. Recent interest has bolstered support for robot-assisted radical cystectomy (RARC) aiming to improve perioperative outcomes, including blood loss, transfusion rates, hospital stay and recovery without compromising oncological efficacy [1, 2]. Consequently, the past decade has witnessed a shift in the utilization of RARC (from <1% in 2004 to 13% in 2010) [2]. Nevertheless, much of the criticism to RARC has been attributed to the steep learning curve associated with this complex procedure. Moreover, there is no standardized way to what defines surgical proficiency for RARC, with some reports supporting at least 20–30 procedures to "flatten" the initial learning curve and then subsequently attempt more complex intracorporeal reconstructions [3].

With wide adoption of robot-assisted surgery in urology, the American Urologic Association (AUA) highlighted the need for validated, procedure-specific tools to assess proficiency and assist in credentialing for robot-assisted procedures [4]. These tools should aim at objective quantification of surgical performance in a standardized manner using evidence-based quality indicators, and to provide structured feedback to trainees. In this context, we aimed to develop a structured scoring tool, the Cystectomy Assessment and Surgical Evaluation (CASE) that objectively measures and quantifies performance during RARC for men.

Methods

A multi-institutional study was conducted between April 2016 and May 2017 using de-identified videos of RARCs performed by expert surgeons, fellows and chief residents. The study comprised 3 phases:

Phase 1: Content Development and Validation

A panel of 10 experienced open and/or robotic surgeons developed the structure of CASE by deconstructing the critical steps of RARC into 9 key domains; each was assessed on a 1-to-5

Likert scale. Specific description for anchors 1, 3 and 5 were provided, meanwhile each domain was assessed utilizing 5 anchors evaluating surgical principles, technical proficiency and safety, where Anchor 1—represented worst, and Anchor 5—defined ideal performance.

Delphi methodology was utilized for validation of CASE structure and content. Anchor descriptions for each of the 9 domains were assessed by the panel in terms of appropriateness of operative skill being assessed, concordance between the statement and the score assigned, and clarity and unambiguous of wording. Experts were contacted separately to provide their feedback on the content of CASE to minimize the "bandwagon" or "halo" effect.

Responses from the panel experts were then collected and analyzed by two independent coordinators (medical student and a Urology fellow). The Content Validity Index (CVI) was used to validate the scoring system [5]. CVI is the proportion of expert surgeons who rated each anchor description as 4 or 5 on the Likert scale. Consensus for an anchor description was considered when the CVI was   0.75. If for a given anchor the CVI was <0.75, comments of the panel were incorporated by the coordinator and redistributed to the panel. The process was repeated until consensus was achieved (Figure 1).

Phase 2: Inter-Rater Reliability

Reliability is the ability of CASE to yield consistent results when utilized by different raters. Ten videos of different proficiency levels (for each domain) were assessed by the panel members. The videos were assigned randomly to reviewers so that each video was reviewed by at least 4 different raters. Raters were blinded to operator level of experience.

Phase 3: Construct Validation

Construct validation refers to the ability of CASE to differentiate between expert and novice performance. Scores given by the panel members for attendings and trainees (fellows and residents) were compared. All domains of the procedure were assessed except the "Control of Vascular Pedicle" domain, for which there were no available videos of trainees performing the step, and the Lymph Node Dissection domain, which has already been assessed using a previously validated tool dedicated for assessment of lymph node dissection after RARC [6].

Statistical Analysis

Descriptive statistics were calculated for scores in each domain. A linearly weighted percent agreement was used to assess the inter-rater reliability. A logit model for odds ratio (OR), which refers to the odds of an expert surgeon to score higher than 3 for a particular domain when compared to a trainee, was used to assess construct validation. Statistical significance was set at alpha level 0.05. All statistical analyses were performed using SAS 9.4 (SAS Institute, Cary, NC).

Results

Content Development and Validation

The expert surgeon panel included 4 open and 6 robotic surgeons (Table 1). All of the panel members perform lymph node dissection routinely with RC, 4 of them perform it prior to bladder extirpation. Based on the panel consensus, domain 9 "Tissue Disposition and Specimen Removal" was removed from the scoring system. Consensus (CVI 0.75) was achieved for all statements after 4 rounds: 12 statements (11%) in the first round, 64 (61%) in the second, 25 (24%) in the third and 4 (4%) after the fourth (Figure 2).The final eight-domain CASE included: Pelvic Lymph Node Dissection, Development of the Peri-ureteral Space, Lateral Pelvic Space, Anterior Rectal Space, Control of the Vascular Pedicle, Anterior Vesical Space, Control of the Dorsal Venous Complex (DVC), and Apical Dissection (Figure 3).

Reliability

Linearly weighted percent agreement was > 0.6 for all domains. The highest agreement was for "Anterior Vesical Space" (0.81), followed by "Apical Dissection" (0.76) and "Anterior Rectal Space" (0.75) (Table 2).

Construct Validation

The expert surgeon group outperformed the trainees in all domains assessed but did not reach statistical significance. The odds for experts to score higher than trainees were highest for "Apical dissection" (Mean difference 0.6, Odds ratio [OR] 2.35, 95% Confidence interval [CI] 0.52–10.74) and "Peri-ureteral Space" (mean difference 0.6, OR 2.16, 95% CI 0.63–7.47). It was least for "DVC Control" (mean difference 0.2, OR 1.09, 95% CI 0.27–4.47) (Table 3) (Figure 4).

Discussion

Continuous monitoring of surgical quality is a duty of modern day practice. Although outcomes are mainly driven by disease-related factors, the quality of operative management is mandatory for satisfactory outcomes [1, 7–9]. Variability between surgeons and institutions has been shown to significantly affect patient outcomes [10, 11]. It remains crucial to ensure adequate, comprehensive training of trainees to assure graduation of competent and safe surgeons. Although surgical simulators have been developed to allow surgical training without affecting patient outcomes, they have only shown limited predictive validity in transferring skills gained to the real operative environment [12]. Operating with robot assistance has become an integral aspect of modern urological practice, however, standards for competency are yet to be determined [13]. Overall number of procedures performed or console hours have been used suggested as a measure for technical proficiency. However, there is no sufficient evidence to support the use of either, with wide variation in the number of procedures required to achieve technical competency [12, 14]. This highlights the need for an objective and well defined tool to assess surgical competency. The Global Evaluative Assessment of Robotic Skills (GEARS) has been developed and validated to assess basic robot-assisted surgical skills (as depth perception, bimanual dexterity, etc) but

lacked procedure-specific goals [15]. Prostatectomy Assessment and Competency Evaluation (PACE) is a validated procedure-specific tool for radical prostatectomy that has been developed using a similar expert-based methodology by our group [16]. Prior studies have shown that what constitutes 'technical proficiency' varies widely among surgeons [14, 17]. To our knowledge, CASE is the first objective assessment tool for RARC developed using real surgical performances and with involvement of both open and robotic expert surgeons. Reliability of all steps has been demonstrated among the ten experts. Utilizing the same principles, our group has developed and validated several objectives, structured and procedure-specific scoring tools for different robot-assisted procedures [6, 16, 18].

Deconstruction a procedure into smaller modules (modular training) has been shown effective in transferring challenging surgical skills to trainees in a step-wise manner while simultaneously shorten the learning curve [19, 20]. The structure of CASE combines the principles of modular training with specific descriptions of the scores assigned to reduce any bias among raters. The scores can be further utilized to track the progress of trainees, and identify weaknesses and thereby guide and tailor training. Although skill-assessment tools have been developed to assess performance on simulators, the majority were not validated in a real operative setting [21]. CASE was able to distinguish between expert surgeons and trainees for all steps examined despite the lack of statistical significance. This may be attributed to the smaller number of video-recordings, or the lack of wide variation among the videos, as most of the trainees were experienced (either chief residents or fellows) with significant console hours under their belt. CASE has been developed and validated expert surgeons in the field and based on real operative performances to ensure skill acquisition based on procedure-specific goals.

The training of new surgeons in robotic surgery faces obstacles such as limited time for training (especially with the new duty-hour restrictions), high cost, and lack of well-defined standards and the need for ensuring oncologic safety [16, 22]. Ethical concerns in terms of disclosing the degree of trainees' participation and the potential higher risk of complications represent another key concern [23]. The high cost of the training equipment, such as simulators or setting up a fully equipped robot in an animal or dry lab is big limitation [16]. In the OR, the isolation of the operating console surgeon away from the patient and any form of visual cueing or body language from the remotely placed mentor to the trainee adds an additional layer of complexity to the teaching process versus a hands-on approach (traditional open—or even laparoscopic environment), where the mentor remotely placed and cannot as readily redirect the trainee or intervene if needed [24].

CASE can serve as a standardized framework for the benchmarks for RARC. It can be used to evaluate trainee surgeons in the execution of the various steps of RARC and monitor their progress during their residency of fellowship training. The modular approach also ensures acquisition of the desired skills and the training program objectives. It can be used by institutions as a quality measure and for auditing performances, and credentialing and surgeon remediation. CASE has been shown to be highly reliable and yield similar results when the applied by different raters, demonstrating its objectivity. The scoring system can be utilized consistently across various institutions, with various using a tool with consistent and

standardized language. This is despite, and in contrast to, previously identified variation in what actually is meant by 'technical proficiency' among different surgeons [14, 17].

Despite the uniqueness of this study, we encountered some significant limitations. First, there were no videos for trainees performing the "Control of Vascular Pedicle" step. A larger pool of videos for the validation phase would have been more likely to reach statistical significance. Although the quality of surgical performance has been previously linked with outcomes, we did not correlate CASE scores with patient outcomes in the current study. Despite the importance of non-technical skills, such as communication, leadership and decision making, CASE was specifically designed to only assess procedure-specific surgical skills.

Conclusion

We developed and validated a reliable structured, procedure-specific tool for objective evaluation of surgical performance during RARC. CASE may differentiate novice from expert performances. It can be used to provide structured feedback for surgical quality assessment and facilitate more effective training programs.

Acknowledgments

References

1. Raza SJ, et al., Long-term oncologic outcomes following robot-assisted radical cystectomy: results from the International Robotic Cystectomy Consortium. Eur Urol, 2015 68(4): p. 721–8. [PubMed: 25985883]

2. Leow JJ, et al., Propensity-matched comparison of morbidity and costs of open and robot-assisted radical cystectomies: a contemporary population-based analysis in the United States. Eur Urol, 2014 66(3): p. 569–76. [PubMed: 24491306]

3. Wilson TG, et al., Best practices in robot-assisted radical cystectomy and urinary reconstruction: recommendations of the Pasadena Consensus Panel. Eur Urol, 2015 67(3): p. 363–75. [PubMed: 25582930]

4. Lee JY, et al., Best practices for robotic surgery training and credentialing. J Urol, 2011 185(4): p. 1191–7. [PubMed: 21334030]

5. Polit DF, Beck CT, and Owen SV, Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Res Nurs Health, 2007 30(4): p. 459–67. [PubMed: 17654487]

6. Hussein AA, et al., Development, validation and clinical application of Pelvic Lymphadenectomy Assessment and Completion Evaluation: intraoperative assessment of lymph node dissection after robot-assisted radical cystectomy for bladder cancer. BJU Int, 2017 119(6): p. 879–884. [PubMed: 27987527]

7. Schiffmann J, et al., Contemporary 90-day mortality rates after radical cystectomy in the elderly. Eur J Surg Oncol, 2014 40(12): p. 1738–45. [PubMed: 25454826]

8. Eisenberg MS, et al., The SPARC score: a multifactorial outcome prediction model for patients undergoing radical cystectomy for bladder cancer. J Urol, 2013 190(6): p. 2005–10. [PubMed: 23770147]

9. Hussein AA, et al., Development and Validation of a Quality Assurance Score for Robot-Assisted Radical Cystectomy: a 10-Year Analysis. Urology, 2016.

10. Herr HW, et al., Surgical factors influence bladder cancer outcomes: a cooperative group report. J Clin Oncol, 2004 22(14): p. 2781–9. [PubMed: 15199091]

11. Cooperberg MR, Odisho AY, and Carroll PR, Outcomes for radical prostatectomy: is it the singer, the song, or both? J Clin Oncol, 2012 30(5): p. 476–8. [PubMed: 22215744]

12. Rashid HH, et al., Robotic surgical education: a systematic approach to training urology residents to perform robotic-assisted laparoscopic radical prostatectomy. Urology, 2006 68(1): p. 75–9. [PubMed: 16844450]

13. Lerner MA, et al., Does training on a virtual reality robotic simulator improve performance on the da Vinci surgical system? J Endourol, 2010 24(3): p. 467–72. [PubMed: 20334558]

14. Abboudi H, et al., Learning curves for urological procedures: a systematic review. BJU Int, 2014 114(4): p. 617–29. [PubMed: 24053179]

15. Goh AC, et al., Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. J Urol, 2012 187(1): p. 247–52. [PubMed: 22099993]

16. Hussein AA, et al., Development and Validation of an Objective Scoring Tool for Robot-Assisted Radical Prostatectomy: Prostatectomy Assessment and Competency Evaluation. J Urol, 2017 197(5): p. 1237–1244. [PubMed: 27913152]

17. Apramian T, et al., How Do Thresholds of Principle and Preference Influence Surgeon Assessments of Learner Performance? Ann Surg, 2017.

18. Frederick PJ, et al., Surgical Competency for Robot-Assisted Hysterectomy: Development and Validation of a Robotic Hysterectomy Assessment Score (RHAS). J Minim Invasive Gynecol, 2017 24(1): p. 55–61. [PubMed: 27780777]

19. Stolzenburg JU, et al., Modular training for residents with no prior experience with open pelvic surgery in endoscopic extraperitoneal radical prostatectomy. Eur Urol, 2006 49(3): p. 491–8; discussion 499–500. [PubMed: 16359780]

20. Volpe A, et al., Pilot Validation Study of the European Association of Urology Robotic Training Curriculum. Eur Urol, 2015 68(2): p. 292–9. [PubMed: 25454612]

21. Vassiliou MC, et al., A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg, 2005 190(1): p. 107–13. [PubMed: 15972181]

22. Ahmed N, et al., A systematic review of the effects of resident duty hour restrictions in surgery: impact on resident wellness, training, and patient outcomes. Ann Surg, 2014 259(6): p. 1041–53. [PubMed: 24662409]

23. McAlister C, Breaking the Silence of the Switch--Increasing Transparency about Trainee Participation in Surgery. N Engl J Med, 2015 372(26): p. 2477–9. [PubMed: 26107052]

24. Tiferes J, et al., The Loud Surgeon Behind the Console: Understanding Team Activities During Robot-Assisted Surgery. J Surg Educ, 2016 73(3): p. 504–12. [PubMed: 27068189]
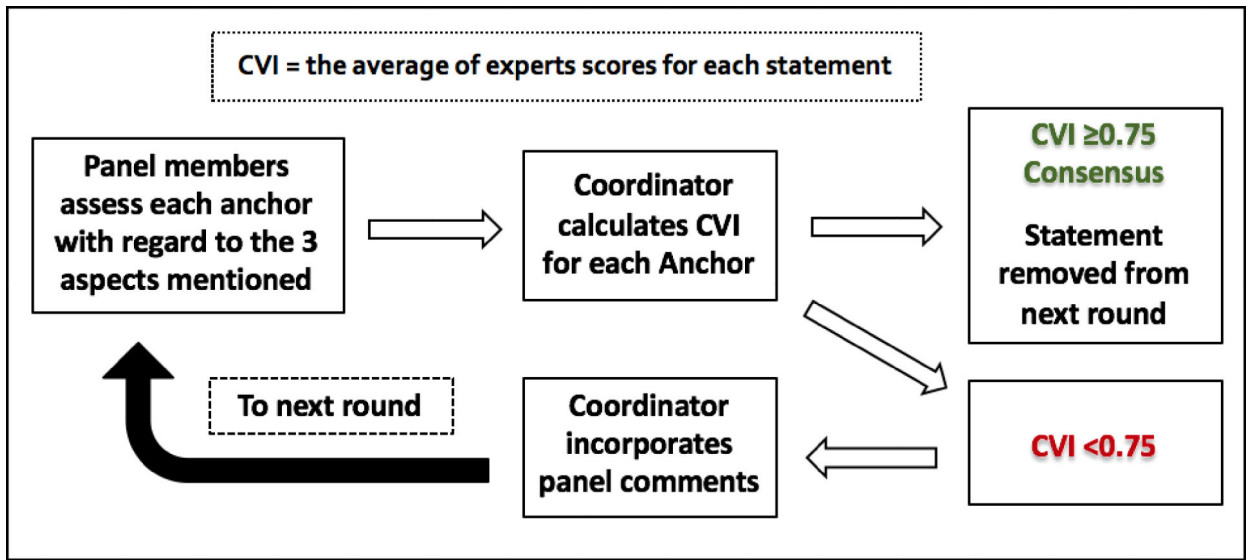
Figure 1.
Delphi Methodology
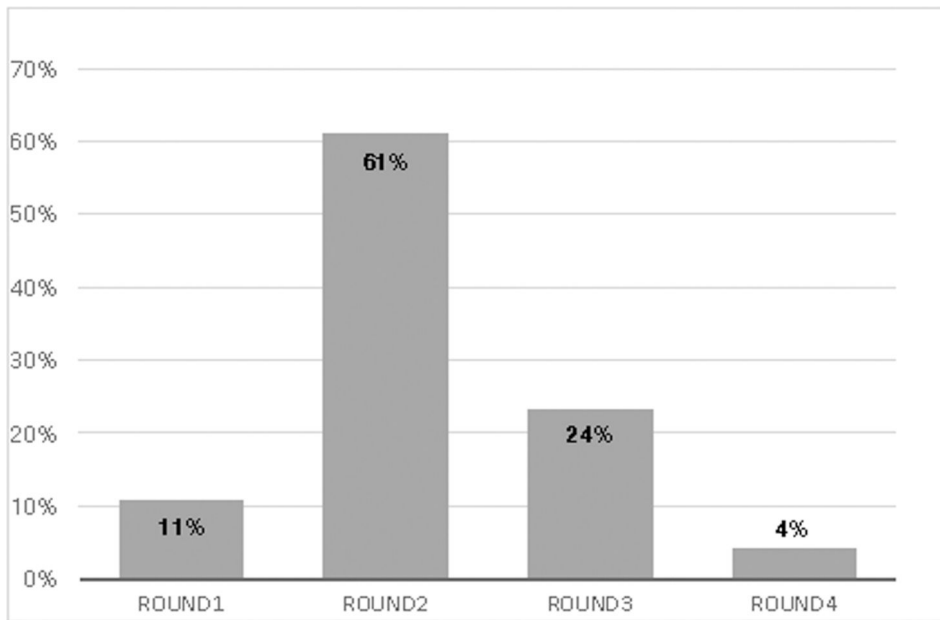
Figure 2.
Content Development: Consensus/Round
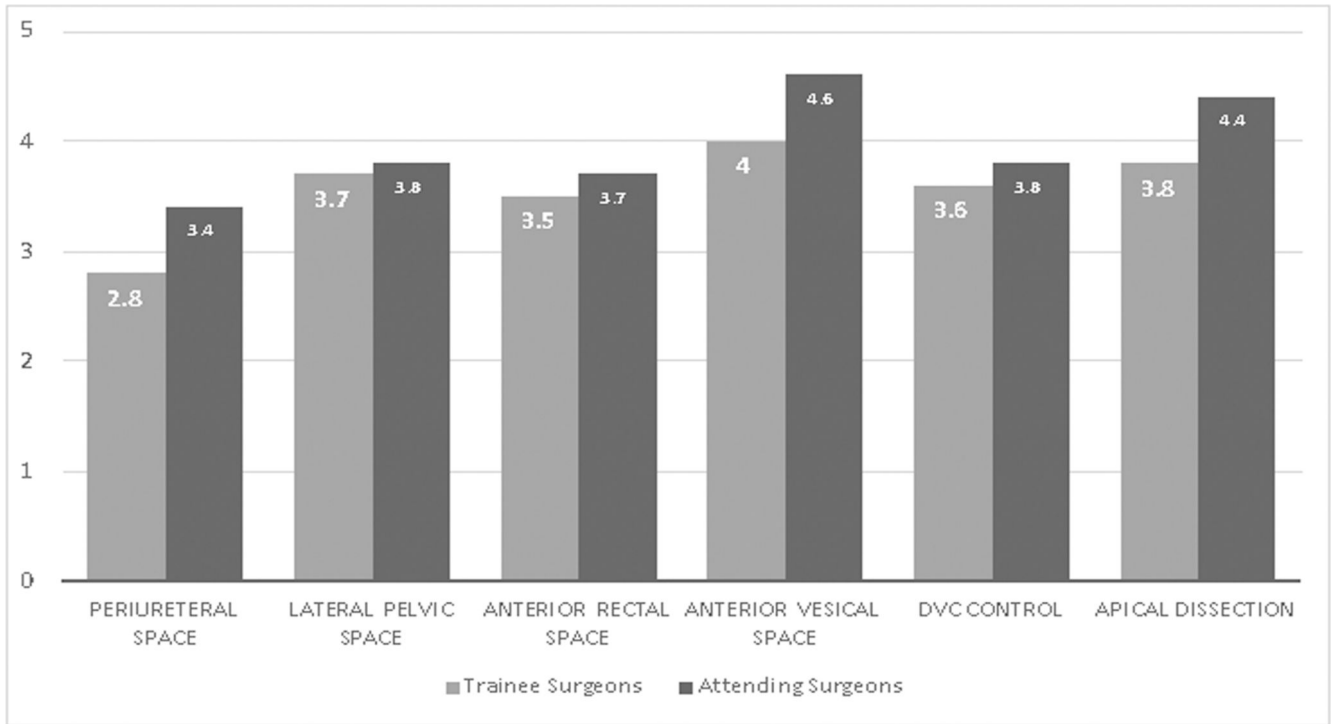
Figure 3.
CASE

Figure 4.
Construct Validation: Mean scores (trainees vs attendings) for each domain

Table 1.

Characteristics of the 10-surgeon panel which participated in development and validation of CASE

Panel Characteristics	n (%)
Urology Practice (years)	
5–10	2 (20)
10–20	6 (60)
>20	2 (20)
Experience	
Formal RAS training	5 (50)
RAS performed >500	4 (40)
Open RC performed >500	5 (50)
RARC performed >50	4 (40)
Procedural Preferences	
Routinely performs LND with cystectomy	10 (100)
LND before cystectomy	4 (40)
LND after cystectomy	6 (60)
Standard LND template	1 (10)
Extended LND template	6 (60)
Super-Extended LND template	3 (3)
Institutional Characteristics	
Annual urologic procedure volume >500	6 (60)
Annual RC volume >50	6 (60)
Proportion of RC performed robot-assisted 25–50%	3 (30)
All RC performed robot-assisted	3 (60)

RAS, robot-assisted surgery, LND, lymph node dissection, RC, radical cystectomy

Table 2.

Inter-rater Reliability for 8 domains of CASE, using Linearly Weighted Percent Agreement

Domain	Agreement
Periureteral Space	0.68
Lateral Pelvic Space	0.73
Anterior Rectal Space	0.75
Vascular Pedicle	0.72
Anterior Vesical Space	0.81
DVC Control	0.71
Apical Dissection	0.76

DVC, dorsal venous complex

Table 3.

Construct Validation for 8 domains of CASE: average scores for trainees and attending surgeons

Domain	Mean Trainee Score	Mean Expert Score	Odds ratio	95% Confidence Interval
Periureteral Space	2.8	3.4	2.16	0.63–7.47
Lateral Pelvic Space	3.7	3.8	1.36	0.37–5.00
Anterior Rectal Space	3.5	3.7	1.70	0.28–10.51
Vascular Pedicle	N/A*			
Anterior Vesical Space	4.0	4.6	N/A**	
DVC Control	3.6	3.8	1.09	0.27–4.47
Apical Dissection	3.8	4.4	2.35	0.52–10.74

DVC, dorsal venous complex

N/A*- No videos were available for trainees to assess this step, N/A** data was not enough for calculations