BJUI
BJU International

Objective assessment of intraoperative skills for robot-assisted radical prostatectomy (RARP): results from the ERUS Scientific and Educational Working Groups Metrics Initiative

Alexandre Mottrie[1,2], Elio Mazzone[1,2,3,4] iD, Peter Wiklund[5,6], Markus Graefen[7], Justin W. Collins[1,8] iD, Ruben De Groote[1,2], Paolo Dell'Oglio[1,9] iD, Stefano Puliatti[1,2,10] and Anthony G. Gallagher[1,11]

[1]*Orsi Academy, Melle*, [2]*Department of Urology, Onze Lieve Vrouw Hospital (OLV), Aalst, Belgium*, [3]*Division of Oncology/Unit of Urology, URI, L'Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Ospedale San Raffaele, Milan*, [4]*Vita-Salute San Raffaele University, Milan, Italy*, [5]*Department of Urology, Karolinska Institutet, Stockholm, Sweden*, [6]*Department of Urology, Icahn School of Medicine at Mount Sinai Health System, New York, NY, USA*, [7]*Martini-Klinik Prostate Cancer Center, University Hospital Hamburg-Eppendorf, Hamburg, Germany*, [8]*Department of Uro-oncology, University College London Hospital (UCLH), London, UK*, [9]*Department of Urology, ASST Grande Ospedale Metropolitano Niguarda, Milan*, [10]*Department of Urology, University of Modena and Reggio Emilia, Modena, Italy, and* [11]*Faculty of Life and Health Sciences, Ulster University, Northern Ireland, UK*

Objective

To develop and seek consensus from procedure experts on the metrics that best characterise a reference robot-assisted radical prostatectomy (RARP) and determine if the metrics distinguished between the objectively assessed RARP performance of experienced and novice urologists, as identifying objective performance metrics for surgical training in robotic surgery is imperative for patient safety.

Materials and methods

In Study 1, the metrics, i.e. 12 phases of the procedure, 81 steps, 245 errors and 110 critical errors for a reference RARP were developed and then presented to an international Delphi panel of 19 experienced urologists. In Study 2, 12 very experienced surgeons (VES) who had performed >500 RARPs and 12 novice urology surgeons performed a RARP, which was video recorded and assessed by two experienced urologists blinded as to subject and group. Percentage agreement between experienced urologists for the Delphi meeting and Mann–Whitney U- and Kruskal–Wallis tests were used for construct validation of the newly identified RARP metrics.

Results

At the Delphi panel, consensus was reached on the appropriateness of the metrics for a reference RARP. In Study 2, the results showed that the VES performed ~4% more procedure steps and made 72% fewer procedure errors than the novices ($P = 0.027$). Phases VIIa and VIIb (i.e. neurovascular bundle dissection) best discriminated between the VES and novices. Limitations: VES whose performance was in the bottom half of their group demonstrated considerable error variability and made five-times as many errors as the other half of the group ($P = 0.006$).

Conclusions

The international Delphi panel reached high-level consensus on the RARP metrics that reliably distinguished between the objectively scored procedure performance of VES and novices. Reliable and valid performance metrics of RARP are imperative for effective and quality assured surgical training.

Keywords

surgical training, robot-assisted radical prostatectomy, proficiency-based training, proficiency-based metrics, construct validation, #Prostate Cancer, #PCSM, #uroonc, #EndoUrology

Introduction

Robot-assisted radical prostatectomy (RARP) has become the most widely used approach for surgical treatment of prostate cancer [1,2]. Increased focus on patients' safety and procedure efficacy has imposed the need to move away from the Halstedian training model where patients may have been exposed to increased risks during the trainee's learning curve. Proficiency based progression (PBP) training offers objective and validated performance metrics to track progression of the trainee and operative skill on a specific task or procedure (which is performed in the training laboratory on virtual simulators and animal models) before the trainees start their clinical practice in the operating theatre on patients [3–6]. In prospective, randomised and blinded studies it has been repeatedly shown that metric-based simulation training to proficiency produces superior surgical skills in comparison to traditional training approaches [3,7–13]. There is also evidence that a PBP simulation training impacts on clinical outcomes [8].

From a clinical standpoint, the technology innovation of robot-assisted surgery should be sustained by improvements in surgical training programmes in order to assure virtually the same clinical outcomes between different centres adopting new robotic platforms. As a first step to achieve this goal, the European Association of Urology (EAU) Robotic Urology Section (ERUS) has designed and developed the first structured curriculum in urology that focusses on RARP [14,15]. The aim of this structured-validated training programme is to propel a surgeon with limited robotic experience towards a complete independent full RARP in order to improve global outcomes of the patients treated during his/her learning curve. However, to date, because of the lack of validated scoring metrics, a full implementation of a PBP-training pathway is still not possible in RARP. There is an imperative to standardise modular training with defined and validated performance metrics in order to enable a PBP-training programme.

Based on this premise, we aimed to develop performance metrics for a RARP procedure and then in a modified Delphi format [16–19] achieve consensus amongst experts on the key steps of RARP and the errors and critical errors related to those steps. We then evaluated whether the performance metrics distinguished between the performances of very experienced surgeons (VES) and less experienced (but trained) novice robotic surgeons performing a standard bilateral nerve-sparing RARP.

Materials and Methods

Study 1: Face and Content Validity (Delphi consensus)

This study received expedited Institutional Review Board approval from Onze Lieve Vrouw Hospital, Aalst, Belgium (OLV, study number: 2019/093). RARP procedure characterisation was performed in five face-to-face meetings. Four urological surgeons (A.M., P.W., J.W.C., M.G.) and a behavioural scientist (A.G.G.) formed the procedure characterisation group. The surgeons had >10 years practise experience of RARP. Procedure characterisation methods are described elsewhere [4–6,17,19,20]. In addition, a glossary of the specific terms used in the present study is reported in Table S1.

Subsequently, a panel of 19 experts from 10 countries (Table 1) then validated the key metrics with a modified Delphi process [21,22]. The Delphi meeting took place in Marseille on 5 September 2018. At the start of the meeting, the concepts of 'PBP' were outlined. The procedure metrics for a reference approach to RARP were also presented. Procedure phases, steps, errors and critical errors were outlined and discussed by the Delphi panel. Following this discussion, the proposed metrics were edited in real time and a vote was taken to establish the level of consensus on the metrics.

Changes in the number of metric units before and after the Delphi meeting were compared for statistical significance with a Wilcoxon signed-rank test. The relationship between the number of metric units before and after was assessed with Pearson's product-moment correlation coefficient.

Study 2: Construct Validity

For the construct validation, we compared the objective scores of intraoperative performance. Two experienced and trained (i.e. Fellow/Consultant level) robotic surgeons were appointed to score the videos of the RARP procedures performed by 12 VES and 12 novice surgeons, using the final version of the RARP performance metrics agreed at the Delphi meeting. For the purpose of video evaluation, only intra- or inter-fascial nerve-sparing RARP cases without lymph node dissection were included. Only full case-videos were evaluated. The surgeons were trained to score the RARP metrics until they consistently achieved a >0.8 inter-rater reliability (IRR). Reviewer training (detailed methodology described elsewhere) [23] was initiated with an 8-h meeting, during which time each metric was studied in detail. Multiple video examples of live cases were shown to illustrate each particular metric. Discussion helped to clarify how each step and error was to be scored, including the nuances and conventions to be used. Full-length practice videos were then independently scored (for occurrence, i.e. event/metrics unit was observed) by each of the reviewers, and the scores tabulated. The differences and discrepancies amongst the reviewers were compared and discussed seeking conformity in scoring. Practice video scoring continued until the reviewers IRR (agreements/[agreements + disagreements]) was consistently ≥0.8. Only then did reviewers progress to scoring study videos.

Table 1 Procedure phases, steps, errors and critical errors and before and after the Delphi meeting.

Phases	No. of steps before DELPHI*	No. of steps after DELPHI	No. of errors before DELPHI*	No. errors after DELPHI	No. of critical errors before DELPHI*	No. critical errors after DELPHI
I. Patient positioning and docking	19 (3)	19	28 (2)	27	8 (2)	10
II. Bladder detachment	5 (1)	5	13 (6)	13	4 (2)	3
III. Endopelvic fascia incision	2 (0)	2	4 (1)	5	3 (2)	3
IV. Bladder neck dissection	11 (4)	10	13 (7)	13	3 (0)	3
V. Dissection of the vas and seminal vesicles	7 (0)	7	9 (0)	9	1 (1)	1
VI. Dissection of posterior space	4 (0)	4	7 (0)	7	1 (1)	1
VIIa: Right lateral dissection of the prostate (intra- or inter-fascial)	8 (1)	8	12 (6)	13	1 (1)	1
VIIb: Left lateral dissection of the prostate (intra- or inter-fascial)	8(1)	8	12 (6)	13	1 (1)	1
IX: Dorsal venous complex	3 (0)	3	9 (0)	9	0 (0)	0
X: Apical dissection	6 (1)	6	13 (2)	12	0 (2)	2
XI: Posterior reconstruction	4 (1)	4	7 (0)	7	0 (0)	0
XII: Vesico–urethral anastomosis ± bladder neck reconstruction	5 (2)	5	16 (4)	13	0 (3)	3
General errors (any phase)	NR	NR	4 (0)	4	1 (0)	1
Assistant errors (any phase)	NR	NR	1 (1)	0	2 (1)	3
Total	82	81	148	145	30	32

*NR, not relevant. *Modifications in parentheses.*

The VES had performed >500 RARPs and the novice surgeons had completed modular training for RARP and performed <10 full RARPs. The video reviewers remained blinded as to the identity of the operator and their status (i.e. VES or novice surgeon). The IRR between the two video reviewers was calculated according to the formula: IRR = number of agreements/(number of agreements + disagreements). Agreement = both reviewers scored an item the same and disagreement = they scored an item differently. This was applied to all performance metrics including error metrics. The IRR was considered to be acceptable if ≥0.8.

Data were used to determine differences in performance when comparing the two groups (VES vs novice) using Mann–Whitney U-tests and for four groups Kruskal–Wallis tests were used. The 95% CIs were derived using bootstrapping. Statistical analysis was performed with the IBM Statistical Package for the Social Sciences (SPSS®; IBM, Corp., Armonk, NY, USA) and with the R software, version 3.5.1 (R Foundation for Statistical Computing, Vienna, Austria; http://www.r-project.org/). Statistical significance for all the analyses was defined as a $P < 0.05$. Comparison between groups was completed for each of the procedural phases and collectively for the entire procedure.

Results

Study 1: Delphi Consensus Meeting

The RARP phases, the number of steps, errors and critical errors before and after the Delphi panel meeting are shown in Table 1. The median number of RARP cases performed by these panel experts was 1500 and median age of panel experts

was 49 years. Additional demographic data about the Delphi panel are reported in Table S2. Changes and edits to the metrics made in real time by the Delphi panel mainly focussed on the precision of the language and operational definitions of procedure steps and errors. Table 3 summarises the changes proposed, voted on, and accepted by the Delphi panel. The number of steps and errors decreased during the Delphi consensus, but the number of critical errors increased. None of these changes were statistically significant (steps: $Z = -1.0$, $P = 0.3$; errors: $Z = -1.5$, $P = 0.1$; critical errors: $Z = -1.6$, $P = 0.1$). Metric units before and after the Delphi were strongly positively correlated (steps, $r = 0.9$, $P < 0.001$; errors, $r = 0.9$, $P < 0.001$; critical errors, $r = 0.9$, $P < 0.001$). A summary of the RARP characterisation after Delphi panel, including a brief description of steps and errors, is reported in Table 2 and 3.

Study 2: Construct Validity

Overall, 12 VES and 12 novice (but trained) surgeons were evaluated. The median age of the VES vs novice surgeons was, respectively 59 and 36 years. Additional demographic data about the evaluated surgeons is reported in Table S3. One VES was removed from the analysis because he used a different approach for RARP (posterior Retzius-sparing) where the identified metrics were not applicable. The mean IRR between the two raters for Study 2 assessments was 0.85.

The individual subject, as well as median and quartile, summary scores for the number of steps made during the RARP procedure by the novice surgeons and VES are shown in Fig. 1A. The median number of procedure steps completed was 48 in the VES group and 46 in the novice group, with an

Table 2 Summary of different RARP procedure metric phases and steps*.

I. Patient positioning and docking

1. WHO checklist completed
2. Patient is anaesthetised on the table
3. Secure placement of patient for Trendelenburg.
4. Check for pressure between patient and padding
5. Positioning of the patient for side docking or between the legs docking
6. Observation of the patient's vital signs when put into Trendelenburg
7. Draping of the patient and arranging suction, cables and other tools in the surgical field
8. Placement of vesical catheter and emptying of bladder
9. Pneumoperitoneum induction
10. Establish internal view
11. Lysis of abdominal adhesions
12. Port placement
13. Patient placed in Trendelenburg position
14. Docking of the robot
15. Adjust depth of the trocars
16. Connection of diathermy cables to the instruments
17. Check suction
18. Instrument insertion
19. Check for free access of instruments

II. Bladder detachment

20. Instrument positioning
21. Incision of the peritoneum
22. Dissection in the Retzius space
23. Coagulation of the median umbilical ligaments and cutting of ligaments to drop the bladder
24. Remove fat over pubo-prostatic ligaments and anterior prostate

III. Endopelvic fascia incision

25. Instrument positioning
26. Incision and development of the endopelvic fascia to allow visibility of the lateral prostate

IV. Bladder neck dissection

27. Define the border between the bladder and the prostate
28. Provide and maintain bladder stretch
29. Start dissection of bladder neck
30. Extend midline incision
31. Visualisation and opening of the urethra
32. Traction on catheter tip with deflated balloon
33. Cut posterior aspect of the urethra and continue with posterior dissection of the bladder neck
34. If non-bladder neck preserving technique, identify the ureteric orifices
35. Lift the prostate and cut through the longitudinal posterior vesico-prostatic fibres
36. Bilateral clips on the remaining lateral anterior aspect of the bladder pedicles attached to the prostate

V. Dissection of vas deferens and seminal vesicles (SV)

37. Instrument positioning
38. Identify the vas deferens, lift with additional arm and using traction dissect it down to the tip of the SV
39. Repeat step 38 on the opposite side
40. Identification and control of the SV arteries by pin-point diathermy or clips
41. Lift up the SV with the additional arm and blunt and sharp dissection to define the plane between the SV and Denonvilliers' fascia
42. Repeat step 40 on the opposite side
43. Repeat step 41 on the opposite side

VI. Dissection of posterior space between the prostate and the rectum

44. Instrument positioning
45. Lift the SV
46. Incision with cold scissors of Denonvilliers' fascia
47. Access and progressive blunt dissection down to apex of the posterior prostatic space

Table 2 (continued)

VII. Right lateral dissection of the prostate

48. Instrument positioning. Use the additional arm during this phase to mobilise the prostate to visualise the dissection area
49. Lift the right SV anteriorly
50. Identification and clipping on the remaining prostatic pedicle
51. Identification, clipping with 'small' clips and cutting on vessels entering the base of the prostate
52. Antegrade dissection of the NVB
53. Complete high anterior release
54. Progression on the anatomical plane
55. Complete the dissection to the level of the apex

VIII. Left lateral dissection of the prostate

56. Instrument positioning. Use the additional arm during this phase to mobilise the prostate to visualise the dissection area
57. Lift the left SV anteriorly
58. Identification and clipping on the remaining prostatic pedicle
59. Identification, clipping with 'small' clips and cutting on vessels entering the base of the prostate
60. Antegrade dissection of the NVB
61. Complete high anterior release
62. Progression on the anatomical plane
63. Complete the dissection to the level of the apex

IX. Dorsal venous complex (DVC) dissection

64. Instrument positioning
65. Cutting of DVC at the level of the prostatic apex
66. Closure of the DVC with suture

X. Apical dissection

67. Instrument positioning
68. Preservation of the urethra
69. Transection of the urethra
70. Transection of any remnants of tissue attached to the prostate
71. Bagging of the prostate
72. Reduce pneumoperitoneum to check for bleeding

XI. Posterior reconstruction

73. Instrument positioning.
74. Closure of the DVC with suture
75. Posterior reconstruction
76. Second layer of suture incorporating posterior aspect of the bladder, remnants of prostate-vesical muscle and bladder mucosa with posterior urethral stump

XII. Vesico–urethral anastomosis (VUA)

77. Use suture with two needles. Closing with running suture from 6 to 12 o'clock anticlockwise on the right side and running suture from 6 to 12 o'clock clockwise on the left side
78. Before closing the anterior aspect of the VUA, push catheter into bladder under direct vision
79. Tie the suture at the completion of the VUA
80. Remove the needles by assistant
81. Leak test for the VUA

DVC, dorsal venous complex; NVB, neurovascular bundle; SV, seminal vesicles; VUA, vesico–urethral anastomosis. *The description of the steps reported in the current table is a summary of the full description of the metrics used for the actual procedure evaluation.

absolute difference in median number of steps of 2 (95% CI −3 to 10, $P = 0.09$). The relative difference in median number of steps performed was 3.7% (95% CI −4 to 24) in favour of the VES group. Results for the number of errors

Table 3 Summary of different RARP procedure metric errors and critical errors (CE)*.

1. Non completion of the step
2. Non-sterile technique
3. No communication with the anaesthetist (CE)
4. Using Veress needle
5. Failure to check for port access and instrument access for planned port placement
6. Trauma to mesenteric vessels or omental vessels
7. Damage to inferior epigastric artery
8. Failure to mark port placement appropriately
9. Port site incision too large
10. Ports not placed appropriately
11. Ports not placed perpendicular to skin
12. Inappropriate distance from patient according to the robotic system
13. Failure to check conflict of robotic arms with patient body
14. Incorrect depth of port placement
15. Incorrect energy setting
16. Failure to check assistant access
17. Operating with poor vision
18. Collisions between instruments
19. Trauma to bowel or major vessels (CE)
20. Port damage to bowel or major vessels (CE)
21. Blind insertion of the instruments (CE)
22. Incorrect use of instruments
23. Uncontrolled tearing of tissue
24. Failure to remove loose clips
25. Collisions of the 4th arm
26. Failure to go lateral to the median umbilical ligament
27. Damage to anterior abdominal wall muscle
28. Inadequate tension on the tissues that prevents progress of the dissection
29. Damage to the bladder
30. Diathermy damage to NVB
31. Excessive tension that results in bleeding or trauma to dissection planes
32. Damage the accessory pudendal artery.
33. Failure to control superficial dorsal venous complex
34. Entry to DVC
35. Failure to identify and repair damage to the bladder (CE)
36. Damage to the obturator nerve (CE)
37. Damage to major iliac vessels (CE)
38. Failure to maintain tissue traction
39. Cut into the prostate
40. Buttonhole in the bladder or trigonal damage
41. Excessive bleeding that obscures anatomy
42. Failure to apply traction to the catheter or the prostate
43. Undermined bladder neck
44. Entering adenomectomy plane
45. Excessive traction on bladder neck
46. Plane too cranial that endangers ureteral orifices
47. Failure of traction between prostate and bladder
48. Deep clips that goes into the NVB
49. Damage to ureteral orifices (CE)
50. Damage to ureters (CE)
51. Tearing of the vas deferens or SV
52. Failure to control bleeding from the vas deferens artery
53. Charring of the tissues
54. Neurovascular tissue attached to the SV.
55. Denonvilliers' fascia attached the SV
56. Damage to the rectum (CE)
57. Damage to NVB
58. NVB is bluntly dissected off prostate inappropriately
59. Inappropriate traction on NVB
60. Inappropriate clips placement
61. Failure to preserve periurethral tissue
62. Failure to stop arterial bleeding
63. Failure to control venous bleeding that compromises visualisation
64. Incorporating urethra in the suture
65. Rupture of the suture
66. Failure to rotate the prostate
67. Progression of the apical dissection with poor visualisation of the anatomy

Table 3 (continued)

68. Cut into apical prostatic tissue
69. Sutures placed into rhabdosphincter
70. Cut into rhabdosphincter (CE)
71. Clips placed inappropriately
72. Suture cuts through sphincteric structure
73. Failure to approximate tissues
74. Trauma to urethral stump
75. Trauma to bladder neck
76. Failure to include mucosa in the suture
77. Suture placed into rhabdosphincter inappropriately
78. Suture placed through NVB
79. Suture cutting through the urethra (CE)
80. Suture through ureteral orifices or through ureter (CE)
81. Failure to introduce catheter under direct vision
82. Suturing catheter into the VUA
83. Failure to complete leak test
84. Leakage from the VUA
85. Failure to recognise leakage
86. Failure to correct leakage
87. Rupture of VUA by overfilling the bladder

CE, critical error; DVC, dorsal venous complex; NVB, neurovascular bundle; SV, seminal vesicles; VUA, vesicourethral anastomosis. *Some of the errors (e.g., excessive bleeding that obscures anatomy or damage to NVB) are repeated in more than one phase or steps. Consequently, the number of errors reported in the present table is lower than the overall number of errors (n = 145) that can be scored in the evaluation of a full RARP procedure after considering repetitions for each phase or step. The 87 errors (or CE) were unique performance units. These may be repeated for different procedure phases. For example, in phase II, damage to the bladder can occur in steps 18 and 21 and it is explicitly identified as a potential CE for each step.

made by both groups and individual members of both groups are shown in Fig. 1B. The median number of procedure errors performed was seven in the VES group and 27 in the novice group, with an absolute difference in median number of steps of −20 (95% CI −23 to −1, P = 0.027) in favour of the VES group. The relative difference in median number of steps performed was −72% (95% CI −85 to −1) in favour of the VES group. However, the VES group also demonstrated considerably greater performance variability than the novice group and two of the VES group performed worse than the weakest performing novice subject.

To further investigate these findings the performance error scores, for each phase of the procedure for each group were divided at the median score to create two sub-groups, i.e. VES performing in the upper half (UH, i.e. fewest errors) and lower half (LH, i.e. most errors) and the same for the novice group. The results of this analysis are shown in Fig. 2 and Table 4. The results show that the VES-UH group consistently made few or no errors across all the phases. In contrast the VES-LH and both novice groups in general made more errors. This difference was greatest for phases VIIa and VIIb.

Overall differences in error scores between the performance of the VES-UH group and the other three groups were compared for significance with Kruskal–Wallis H-tests. The VES-UH group made significantly fewer errors than the novice-LH group (absolute difference in medians −25 errors, 95% CI −28 to −21; P < 0.001), the novice-UH group

Fig. 1 The median (IQR) and individual surgeon scores of **A.** procedure steps completed, and **B.** the number of errors made by the novice surgeons and VES.
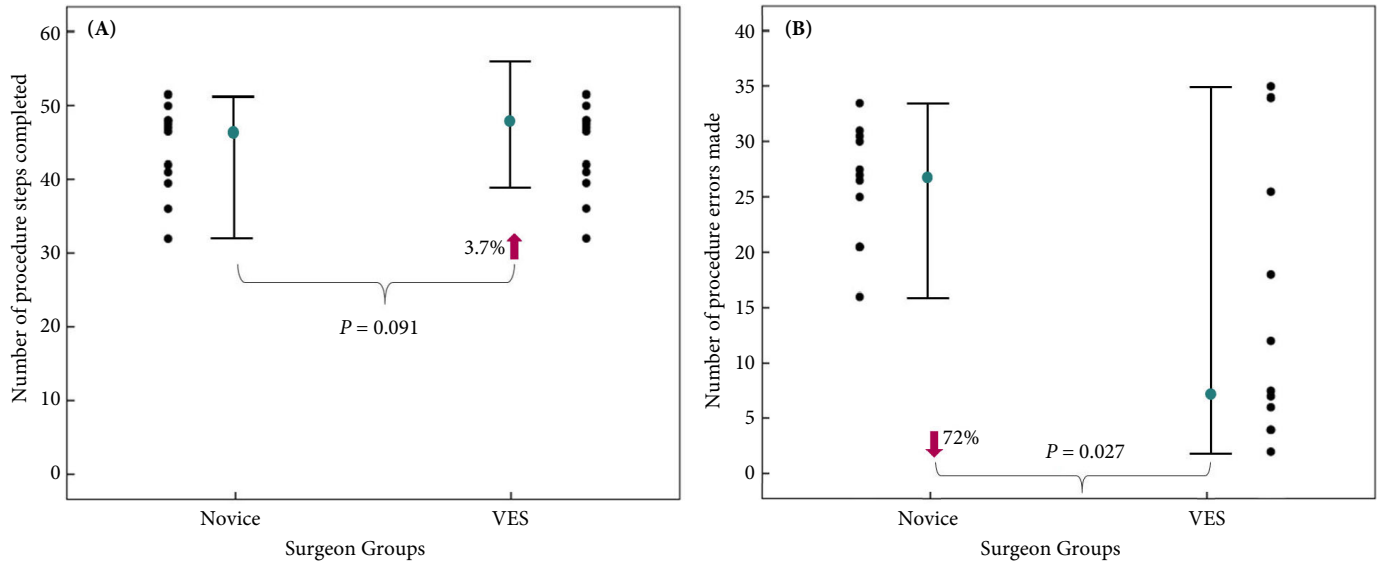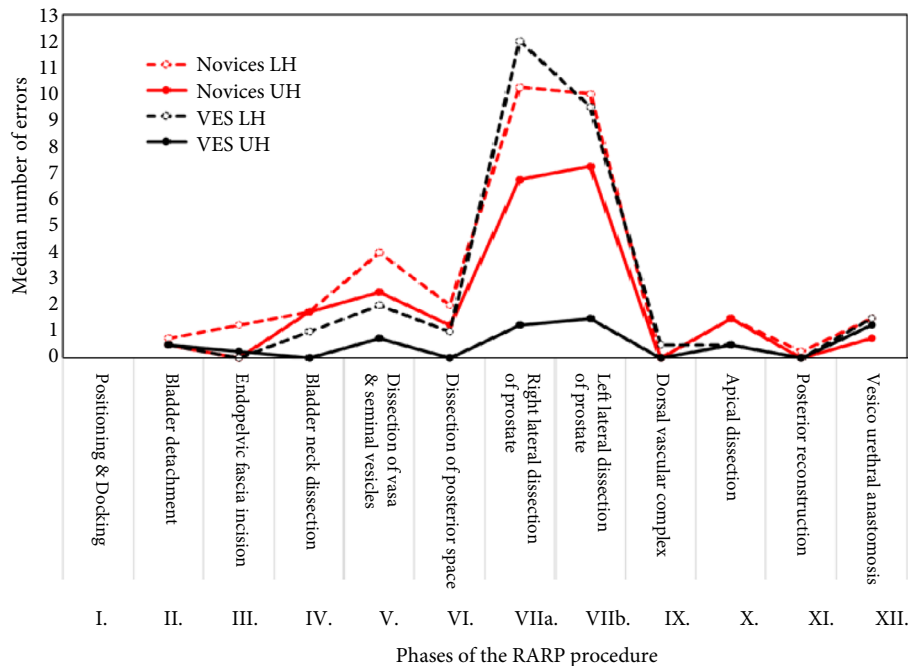

Fig. 2 The median number of errors made during the 12 different phases of the procedure by the VES and novice surgeon groups which were both divided at their median point into LH and UH scores.

(absolute difference in medians −15 errors, 95% CI −21 to −13; P = 0.004) and the VES-LH group (absolute difference in medians −20 errors, 95% CI −31 to −7; P = 0.006). These results are summarised in Fig. 2, which shows that the VES-LH group error performance profile across the phases of the procedure appeared more similar to the novice group than the VES-UH group.

Discussion

The increasing use of robot-assisted technology for surgery imposes the need to set standardised training pathways to optimise patient care and safety [24–26]. As such, to improve patient outcomes after RARP, robotic training and education needs to be modernised and augmented. To achieve this, the

Table 4 Summary descriptive data demonstrating the relative performance profiles of the novice surgeons and VES objectively assessed error performance for the RARP procedure.

	Sum of errors	Number of errors, median (IQR)
Novices-LH	180	30 (27–31)
Novices-UH	129	21 (19–25)
VES-LH	124	29 (16–34)
VES-UH	30	5 (3–7)

skills and performance levels of trainees need to be objectively and quantitatively assessed and verified before operating on real patients. However, a full implementation of standardised PBP-based training for RARP will only be possible when objective and valid metrics are available. In pursuit of this goal, we developed the performance metrics for a reference approach to a RARP procedure. Our present analysis revealed several noteworthy findings.

First, the RARP metrics developed by the procedure characterisation group and their operational definitions were presented to a Delphi panel and were very well received. Consensus among 19 international experts was high. Specifically, they concurred that the 12 phases and 81 steps did characterise a reference approach to RARP procedures by trainees at the start of their learning curve. It is important to remark that the panel did not advocate that the identified approach has to be considered right and other approaches wrong. Conversely, the Delphi panel agreed that this is the more frequently and commonly used approach for RARP and, therefore, it will be easier to learn and more comfortable to use for the trainees at the beginning of their learning curve.

Second, the results showed that the performance metrics scored by two independent reviewers distinguished between the performance of VES and procedure novices who underwent a pre-defined modular training proposed by the ERUS [14,15]. The VES group completed significantly more phases and steps of the procedure compared to the novice group. More importantly, they also made fewer objectively assessed intraoperative procedure errors. The largest performance differences in the number of errors were observed for phases VIIa and VIIb. These data indicate that neurovascular bundle dissection is the step where the highest rate of discrepancy was observed, probably due to its higher complexity compared to other phases of the procedure. In summary, these metrics were able to discriminate with high reliability performance differences between the VES and novice groups and support construct validity evidence.

Third, we observed considerable variability in the performance of the VES group, particularly for error scores. Two of the VES performed worse than the weakest trainees (i.e. number of errors made) in the novice group. When we divided the scores of the two groups at the median score

point performance variability particularly in the VES group was further elucidated. The VES performing in the lower half of their group demonstrated the largest performance variability. These findings are of concern, but not new [10,27,28].

Our present results have demonstrated that surgical experience and seniority do not always translate into optimal objectively assessed surgical performance, an observation noted by Begg *et al.* [29] when investigating morbidity after RP. Whilst the goal of the studies reported here was to develop valid performance metrics to improve robotic surgical training for novices, objective metrics may also be used to define surgical performance quality in surgeons with extensive previous robotic experience. It might be argued that objective assessment of intraoperative performance on one occasion is a poor indicator of surgical skill. However, published evidence challenges this view. Birkmeyer *et al.* [30] found that objective peer-assessed surgical skills strongly predicted clinical outcomes of patients undergoing minimally invasive bariatric surgery. Surgeons' assessed as performing in the lowest quartile had significantly higher complication rates, re-admissions, and mortality. Of note, it has been shown that simulated and real-world performances are highly correlated [31]. Furthermore, these findings do not appear to be an anomaly and replicate previous observations [10,28,32,33].

Taken together, this is the first report to objectively characterise intraoperative performance for RARP by proposing the scoring of operative procedure steps, errors, and critical errors. The metrics were reviewed in detail, edited, and agreed by an international group of experienced urological robotic surgeons. Construct validity for the metrics was also demonstrated by comparing video recorded performance of VES and novice surgeons performing a straightforward RARP procedure. Ideally, these results should be used to standardise robot-assisted surgical training by introducing a PBP methodology. Specifically, these metrics will also be used to establish performance benchmarks (i.e. proficiency levels), which trainees must unambiguously demonstrate before training progression. Additionally, trainees should not progress to performing the procedure on real patients until they have demonstrated that they 'know' how to do the procedure and can 'do it' to a quantitatively defined performance level.

Our present study does have limitations. First, the limited number of the video recorded procedures evaluated may limit the generalisability of our analysis and any firm conclusions about performance variability by very experienced operators. Second, the reported metrics are only applicable to RARP with the classical anterior transperitoneal approach. Different techniques, such as extraperitoneal RARP, lateral or posterior Retzius-sparing approaches, cannot be scored using the performance metrics in their present configuration. However,

it is important to note that the large majority of evidence available on RARP outcomes and techniques refer to the standard anterior approach. Third, variation in patients' characteristics (i.e. age, body mass index, previous abdominal surgery or other comorbidity index) that were not taken into account in the present study may have influenced the reported differences in performance. Fourth, despite that the novice surgeons were required to complete the RARP independently as part of their course, we cannot exclude that the results reported for the novice-UH group may have been marginally biased by the impact of clinical supervision from an experienced surgeon.

Conclusions

Using a modified Delphi process, we achieved consensus among a group of very experienced international experts for a PBP approach to RARP training. We have also found that the metrics demonstrated construct validity and discriminative validity. Overall, these newly developed metrics reliably distinguished between the objectively assessed intraoperative RARP performance of VES and novice robotic surgeons. Errors metrics showed the greatest capacity to distinguish performances. These metrics lay the foundation to implement a simulation-based PBP training programme for modular RARP training.

Conflict of Interest

Alexandre Mottrie reports grants from Intuitive, during the conduct of the study; personal fees from ORSI Academy, outside the submitted work; Justin W Collins reports grants from Medtronic, personal fees from Medtronic, personal fees from Intuitive Surgical, personal fees from CMR Surgical, outside the submitted work; Elio Mazzone, Peter Wiklund, Markus Graefen, Paolo Dell'Oglio, Ruben De Groote, Stefano Puliatti and Anthony G Gallagher certify that all their conflicts of interest, including specific financial interests and relationships and affiliations relevant to the subject matter or materials discussed in the manuscript (e.g. employment/ affiliation, grants or funding, consultancies, honoraria, stock ownership or options, expert testimony, royalties, or patents filed, received, or pending), are the following: None.

References

1 **Mazzone E, Mistretta FA, Knipper S et al.** Contemporary North-American assessment of robot-assisted surgery rates and total hospital charges for major surgical uro-oncological procedures. *J Endourol* 2019; 33: 438–47

2 **Leow JJ, Chang SL, Meyer CP et al.** Robot-assisted versus open radical prostatectomy: a contemporary analysis of an all-payer discharge database. *Eur Urol* 2016; 70: 837–45

3 **Angelo RL, Ryu RKN, Pedowitz RA et al.** A Proficiency-based progression training curriculum coupled with a model simulator results in the acquisition of a superior arthroscopic Bankart skill set. *Arthroscopy* 2015; 31: 1854–71

4 **Gallagher A.** Metric-based simulation training to proficiency in medical education: what it is and how to do it. *Ulster Med J* 2012; 81: 107–13

5 **Gallagher AG, O'Sullivan GC.** *Fundamentals of Surgical Simulation: Principles and Practice.* New York, NY: Springer Publishing Company, Inc., 2011

6 **Gallagher AG, Ritter EM, Champion H et al.** Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. *Ann Surg* 2005; 241: 364–72

7 **Breen D, O'Brien S, McCarthy N, Gallagher A, Walshe N.** Effect of a proficiency-based progression simulation programme on clinical communication for the deteriorating patient: a randomised controlled trial. *BMJ Open* 2019; 9: e025992

8 **Srinivasan KK, Gallagher A, O'Brien N et al.** Proficiency-based progression training: an "end to end" model for decreasing error applied to achievement of effective epidural analgesia during labour: a randomised control study. *BMJ Open* 2018; 8: e020099

9 **Cates CU, Lönn L, Gallagher AG.** Prospective, randomised and blinded comparison of proficiency-based progression full-physics virtual reality simulator training versus invasive vascular experience for learning carotid artery angiography by very experienced operators. *BMJ Simul Technol Enhanc Learn* 2016; 2: 1–5

10 **Pedowitz RA, Nicandri GT, Angelo RL, Ryu RKN, Gallagher AG.** Objective assessment of knot-tying proficiency with the fundamentals of arthroscopic surgery training program workstation and knot tester. *Arthroscopy* 2015; 31: 1872–9

11 **Van Sickle KR, Ritter EM, Baghai M et al.** Prospective, randomized, double-blind trial of curriculum-based training for intracorporeal suturing and knot tying. *J Am Coll Surg* 2008; 207: 560–8

12 **Ahlberg G, Enochsson L, Gallagher AG et al.** Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. *Am J Surg* 2007; 193: 797–804

13 **Seymour NE, Gallagher AG, Roman SA et al.** Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Ann Surg* 2002; 236: 454–8

14 **Mottrie A, Novara G, van der Poel H, Dasgupta P, Montorsi F, Gandaglia G.** The European association of urology robotic training curriculum: an update. *Eur Urol Focus* 2016; 2: 105–8

15 **Volpe A, Ahmed K, Dasgupta P et al.** Pilot validation study of the European association of urology robotic training curriculum. *Eur Urol* 2015; 68: 292–9

16 **Hegarty J, Howson V, Wills T et al.** Acute surgical wound-dressing procedure: description of the steps involved in the development and validation of an observational metric. *Int Wound J* 2019; 16: 641–8

17 **Crossley R, Liebig T, Holtmannspoetter M et al.** Validation studies of virtual reality simulation performance metrics for mechanical thrombectomy in ischemic stroke. *J Neurointerv Surg* 2019; 11: 775–80

18 **Mascheroni J, Mont L, Stockburger M, Patwala A, Retzlaff H, Gallagher AG.** International expert consensus on a scientific approach to training novice cardiac resynchronization therapy implanters using performance quality metrics. *Int J Cardiol* 2019; 289: 63–9

19 **Angelo RL, Ryu RKN, Pedowitz RA, Gallagher AG.** Metric development for an arthroscopic Bankart procedure: assessment of face and content validity. *Arthroscopy* 2015; 31: 1430–40

20 **Kojima K, Graves M, Taha W, Cunningham M, Joeris A, Gallagher AG.** AO international consensus panel for metrics on a closed reduction and fixation of a 31A2 pertrochanteric fracture. *Injury* 2018; 49: 2227–33

21 **Collins JW, Levy J, Stefanidis D et al.** Utilising the Delphi process to develop a proficiency-based progression train-the-trainer course for robotic surgery training. *Eur Urol* 2019; 75: 775–85

22 **Puliatti S, Mazzone E, Amato M, De Groote R, Mottrie A, Gallagher AG.** Development and validation of the objective assessment of robotic

suturing and knot tying skills for chicken anastomotic model. *Surg Endosc* 2020 [Online ahead of print]. https://doi.org/10.1007/s00464-020-07918-5

23 **Gallagher AG, Ryu RK, Pedowitz RA, Henn P, Angelo RL.** Inter-rater reliability for metrics scored in a binary fashion-performance assessment for an arthroscopic Bankart repair. *Arthroscopy* 2018; 34: 2191–8

24 **Palagonia E, Mazzone E, De Naeyer G et al.** The safety of urologic robotic surgery depends on the skills of the surgeon. *World J Urol* 2020; 38: 1373–83

25 **Mazzone E, Dell'Oglio P, Mottrie A.** Outcome report of the first ERUS robotic urology curriculum-trained surgeon in Turkey: the importance of structured and validated training programs for global outcome improvement. *Turk. J Urol* 2019; 45: 189–90

26 **Puliatti S, Mazzone E, Dell'Oglio P.** Training in robot-assisted surgery. *Curr Opin Urol* 2020; 30: 65–72

27 **Gallagher AG, Richie K, McClure N, McGuigan J.** Objective psychomotor skills assessment of experienced, junior, and novice laparoscopists with virtual reality. *World J Surg* 2001; 25: 1478–83

28 **Gallagher AG, Henn PJ, Neary PC et al.** Outlier experienced surgeon's performances impact on benchmark for technical surgical skills training. *ANZ J Surg* 2018; 88: E412–7

29 **Begg CB, Riedel ER, Bach PB et al.** Variations in morbidity after radical prostatectomy. *N Engl J Med* 2002; 346: 1138–44

30 **Birkmeyer JD, Finks JF, O'Reilly A et al.** Surgical skill and complication rates after bariatric surgery. *N Engl J Med* 2013; 369: 1434–42

31 **Gallagher AG, Seymour NE, Jordan-Black J-A, Bunting BP, McGlade K, Satava RM.** Prospective, randomized assessment of transfer of training (ToT) and transfer effectiveness ratio (TER) of virtual reality simulation training for laparoscopic skill acquisition. *Ann Surg* 2013; 257: 1025–31

32 **Gallagher AG, Smith CD, Bowers SP et al.** Psychomotor skills assessment in practicing surgeons experienced in performing advanced laparoscopic procedures. *J Am Coll Surg* 2003; 197: 479–88

33 **Angelo RL, Pedowitz RA, Ryu RK, Gallagher AG.** The Bankart performance metrics combined with a shoulder model simulator create a precise and accurate training tool for measuring surgeon skill. *Arthroscopy* 2015; 31: 1639–54

Correspondence: Elio Mazzone, Division of Oncology/Unit of Urology URI, IRCCS Ospedale San Raffaele, Via Olgettina 60, Milan 20132, MI, Italy.

e-mail: eliomazzone@gmail.com

Abbreviations: IRR, inter-rater reliability; LH, lower half; (RA)RP, (robot-assisted) radical prostatectomy; PBP, proficiency based progression; UH, upper half; VES, very experienced surgeons.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Glossary of terms.
Table S2. Demographic characteristics of the Delphi panel experts.
Table S3. Demographic characteristics of the 12 VES and 12 novice surgeons evaluated for the construct validation.