Journal of Robotic Surgery (2023) 17:1401–1409
https://doi.org/10.1007/s11701-023-01521-1

RESEARCH

Objective assessment of intraoperative skills for robot-assisted partial nephrectomy (RAPN)

Rui Farinha[1,2,3] · Alberto Breda[4] · James Porter[5] · Alexandre Mottrie[1,2] · Ben Van Cleynenbreugel[6,7] · Jozef Vander Sloten[8] · Angelo Mottaran[9,10] · Anthony G. Gallagher[1,11,12]

Received: 9 November 2022 / Accepted: 2 January 2023 / Published online: 23 January 2023

Abstract

RAPN training usually takes place in-vivo and methods vary across countries/institutions. No common system exists to objectively assess trainee capacity to perform RAPN at predetermined performance levels prior to in-vivo practice. The identification of objective performance metrics for RAPN training is a crucial starting point to improve training and surgical outcomes. The authors sought to examine the reliability, construct and discriminative validity of objective intraoperative performance metrics which best characterize the optimal and suboptimal performance of a reference approach for training novice RAPN surgeons. Seven Novice and 9 Experienced RAPN surgeons video recorded one or two independently performed RAPN procedures in the human. The videos were anonymized and two experienced urology surgeons were trained to reliably score RAPN performance, using previously developed metrics. The assessors were blinded to the performing surgeon, hospital and surgeon group. They independently scored surgeon RAPN performance. Novice and Experienced group performance scores were compared for procedure steps completed and errors made. Each group was divided at the median for Total Errors score, and subgroup scores (i.e., Novice HiErrs and LoErrs, Experienced HiErrs and LoErrs) were compared. The mean inter-rater reliability (IRR) for scoring was 0.95 (range 0.84–1). Compared with Novices, Experienced RAPN surgeons made 69% fewer procedural Total Errors. This difference was accentuated when the LoErr Expert RAPN surgeon's performance was compared with the HiErrs Novice RAPN surgeon's performance with an observed 170% fewer Total Errors. GEARS showed poor reliability (Mean IRR = 0.44; range 0.0–0.8), for scoring RAPN surgical performance. The RAPN procedure metrics reliably distinguish Novice and Experienced surgeon performances. They further differentiated performance levels within a group with similar experiences. Reliable and valid metrics will underpin quality-assured novice RAPN surgical training.

Keywords Surgical training · Robot-assisted partial nephrectomy · Proficiency-based training · Metrics · Construct validation · Renal cancer

✉ Rui Farinha
ruifarinhaurologia@gmail.com

[1] Orsi Academy, Proefhoevestraat 12, Melle, 9090 Ghent, Belgium

[2] Department of Urology, Onze-Lieve-Vrouw Ziekenhuis, Aalst, Belgium

[3] Department of Urology, São José Hospital, Lisbon, Portugal

[4] Department of Urology, Fundació Puigvert, Universitat Autonoma de Barcelona, Barcelona, Spain

[5] Swedish Urology Group, Swedish Medical Center, Seattle, WA, USA

[6] Department of Urology, University Hospitals Leuven, Louvain, Belgium

[7] Department of Development and Regeneration, KU Leuven, Louvain, Belgium

[8] Department of Mechanical Engineering, Section of Biomechanics, KU Leuven, Louvain, Belgium

[9] Division of Urology, IRCCS Azienda Ospedaliero-Universitaris di Bologna, Bologna, Italy

[10] University of Bologna, Bologna, Italy

[11] Faculty of Medicine, KU Leuven, Louvain, Belgium

[12] Faculty of Life and Health Sciences, Ulster University, Derry, Northern Ireland, UK

Abbreviations

RAPN	Robot-assisted partial nephrectomy
IRR	Inter-rater reliability
PBP	Proficiency-based progression
AGG	Anthony gallagher
GEARS	Global evaluative assessment of robotic skills
Est	Statistical estimate
SE	Standard errors
df	Degree of freedom
t	Test statistic
p	Probability value
CI	Confidence interval
SD	Standard deviation
ROC	Receiver operating characteristic
AUC	Area under the curve

Introduction

Robot-assisted partial nephrectomy (RAPN) is well established as a surgical treatment for T1a renal masses [1–3]. Guidelines on how surgeons should perform RAPN usually come from experienced surgeons and vary across institutions/countries [4, 5]. In addition, surgeon and hospital's RAPN volume directly impact on complication rates [6–10].

The relationship between RAPN approaches and complication rates also informs the way surgeons are trained [9, 11], and increased focus on patient safety compels a paradigm shift in the training methodology. At the start of the twenty-first-century surgeons still acquire the vast majority of their procedure skill on patients. Furthermore, surgical competence is still assessed through process measure (i.e., number of procedures performed or time in training), instead of using meaningful performance metrics [12].

Proficiency-based progression (PBP) training represents a paradigm shift in how surgeons learn new skills, offering objective and validated performance metrics to support and track the progression of their operative skills. It also utilizes a standardized and validated system to objectively evaluate trainee capacity to perform RAPN at a predetermined performance level prior to in-vivo practice on patients [13–16].

Several prospective, randomized and blinded controlled trials have shown that metric-based simulation training to proficiency produces superior surgical skills in comparison with traditional quality-assured training approaches and impact on clinical outcomes [16–22]. This motivated the authors to develop performance metrics for a RAPN procedure which would then underpin a metric-based training program. The identified metrics were subjected to detailed scrutiny and discussion by an international panel of expert RAPN surgeons. The outcome was a clear consensus on the metrics and their operational definitions [23].

This study aims to examine the reliability, construct and discriminative validity of these metrics. Construct validity establishes the extent to which the metrics discriminate between different levels of expertise of RAPN surgeons (i.e., the Experienced group of RAPN surgeons should perform the procedure better than the Novice RAPN surgeons) [24]. Discriminative validity assesses whether the metrics are able to differentiate performance levels within a group of surgeons with similar experience, denoting high sensitivity and specificity [24–26].

Patients and methods

After obtaining Institutional Review Board Approval, and written informed consent from study participants, the authors compared intraoperative RAPN performance scores of 7 Novices and 9 Expert surgeons. A Novice was defined as having performed ≤ 25 RAPNs, and an Expert as having performed ≥ 250 RAPNs.

Procedure videos

RAPN videos to be scored only included left-sided renal tumors, located in the lower pole or middle part of the kidney, with < 4 cm diameter and being $> 50\%$ exophytic.

Assessors

Two consultant RAPN surgeons were trained to be assessors, by a behavioral scientist and education-training expert (AGG). Eight hours of face-to-face meetings and online conference calls were conducted using Zoom (San Jose, California, US) [14, 27]. Both assessors studied the methods of PBP metrics for RAPN and Global Evaluative Assessment of Robotic Skills (GEARS) scoring in detail [23, 28]. Multiple unedited videos of RAPN performed by different surgeons of varying degrees of expertise were used to illustrate what to score.

Each assessor then scored the videos independently until IRR [IRR: agreements/(agreements + disagreements)] ≥ 0.8 [14, 29]. Disagreements, conflicts or uncertainty around scoring entailed further discussion with AGG to improve scoring and assessments.

Once both assessors could independently, consistently and reliably (i.e., IRR ≥ 0.8) score operative performance, 24 complete unedited recordings of RAPN, performed by 7 Novice and 9 Experienced surgeons were scored. Assessors were blinded to the identity or level of expertise of the operating surgeon. Each video was evaluated using both binary metrics and GEARS, and scorings were tabulated. An IRR was considered acceptable if it was ≥ 0.8.

Performance metrics

A procedure Step was defined as a component task, the series aggregate of which constitutes the completion of a specific procedure. An Error was defined as a deviation from optimal performance. A Critical Error was defined as an event or occurrence involving a serious deviation from optimal performance during a procedure that either (1) jeopardizes the success or the desired result of the procedure or (2) creates iatrogenic insult to the patient's tissues [17].

Statistical analysis

Assuming approximately equal variance in each group, data were analyzed with a Mixed Model regression analysis. Fixed effect was Group (i.e., Novice and Experienced Groups) and the repeated measure was the Procedure number used to determine if there was a statistical difference for the primary endpoints (number of completed Steps, Errors, Critical Errors, and Total Errors) between the Novice and the Experienced group.

Results are reported in terms of statistical estimate (Est), standard errors (SE), degree of freedom (*df*), test statistic (*t*), and probability value (*p*). It was assessed the statistical difference between the first and second procedure for the surgeons that have submitted a second RAPN.

The median Total Errors score was calculated for the Novices and Experienced surgeons and a dummy dichotomous variable was created (i.e., based on scores above or below the median Total Errors score). A Novice LoErrs (Total Errors score below the median) and a Novice HiErrs (Total Errors score above the median) subgroups were defined. We used the same approach to create Experienced LoErrs and HiErrs subgroups. The IBM Statistical Package for the Social Sciences, version 28 (SPSS®; IBM, Corp., Armonk, NY, USA) was used.

IRR assessment

RAPN videos were scored for occurrence (i.e., event/ metric unit was observed) by each of the assessors and scores tabulated. The difference and discrepancies between reviewers were compared. Video score IRR was assessed using the formulated Agreements/(Agreements + Disagreements) [14, 15, 30].

Results

Binary metrics scores

The mean IRR was 0.9 (SD = 0.03), and no assessment fell below 0.8. The mean and 95% confidence intervals (CI) for the number of procedure Steps, Errors, Critical Errors, and Total Errors (sum of Errors and Critical Errors) made by the Novice and Experienced surgeons are shown in Figs. 1a–d.

The Novice group on average, completed 10% (i.e., 2) fewer procedure Steps than the Experienced group, but this difference was not statistically significant. The later made 53% fewer procedure Errors than the former, being this difference statistically significant (Est = 5.46, SE = 1.58, $df = 21.968$, $t = 3.454$, $p = 0.002$). Overall, the Critical Errors rate was low. Although the Novice group made more Critical Errors and showed greater score variability, the difference was not statistically significant. On average, the Novice group made 69% more Total Errors than the Experienced group, which was found to be statistically significant (Est = 6.694, SE = 1.92, $df = 21.344$, $t = 3.478$, $p = 0.002$). Procedure numbers (i.e., Procedure 1 and 2) had no significant impact on the statistical results.

Performance data for each group was divided at the median of the Total Errors score, to create two sub-groups. The median Total Errors score for the Experienced surgeons was 10.25. Therefore, Experienced surgeons who made more than 10.25 Total Errors were classified as surgeons above the median who had a high error rate (Experienced HiErrs; $n = 5$), and who made less than 10.25 Total Errors were classified as surgeons below the median who had a low error rate (Experienced LoErrs, $n = 7$). The same approach was used for the Novice surgeons. The median Total Errors score for the Novice surgeons was 17.25. Novice surgeons who made more than 17.25 Total Errors were classified as surgeons above the median who had a high error rate (Novice HiErrs; $n = 6$), and who made less than 17.25 Total Errors were classified as surgeons below the median who had a low error rate (Novice LoErrs, $n = 6$).

A summary of the performances of the resulting four groups is shown in Fig. 2a–d for procedure Steps, Errors, Critical Errors and Total Errors. Differences between groups for procedure Steps were assessed for statistical significance with a Mixed Model regression analysis where the fixed effects were Group (i.e., Novice HiErrs and LoErrs and Experienced HiErrs and LoErrs) and the repeated measure was the Procedure number (i.e., Procedure 1or Procedure 2).

Although differences between the subgroups in the number of procedure Steps completed were small, it was found that the Experienced LoErrs completed significantly more procedure Steps than the Novice HiErrs (Est = − 3.897, SE = 1.702, $df = 18.173$, $t = 2.172$, $p = 0.043$) and the Experienced LoErrs completed more procedure Steps than the Experienced HiErrs although this difference was not statistically significant.

Differences were observed for the procedure Errors. The Experienced LoErrs made 132% fewer Errors than the Novice HiErrs, being this difference statistically significant (Estimate = 10.971, Standard Error = 1.422,

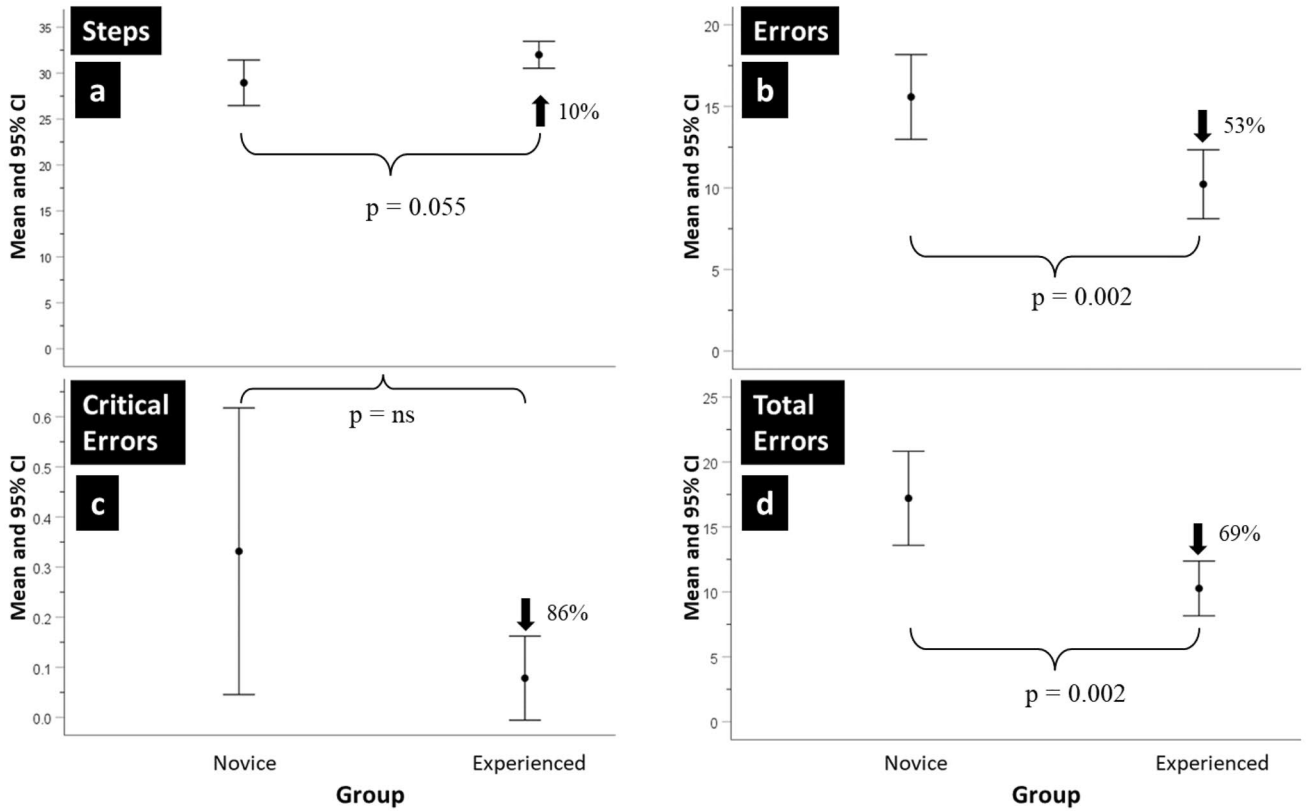


Fig. 1 **a–d** Novice and Experts using binary metrics. The mean and 95% confidence intervals (CI) of the number of Steps completed, Errors, Critical Errors and Total Errors made by Novice and Experienced surgeons when performing the robot-assisted partial nephrectomy (RAPN) procedure

df = 19.975, *t* = 7.718, *p* < 0.001) and 62% fewer Errors than the Experienced HiErrs (Estimate = 5.007, Standard Error = 1.421, *df* = 16.089, *t* = 3.542, *p* = 0.003). In contrast, the Experienced HiErrs made a similar number of procedure Errors as the Novice LoErrs.

Overall, the Critical Error rate was low. Both Novice HiErrs and LoErrs made the most Critical Errors and demonstrated the greatest variability. The Experienced HiErrs made no Critical Errors and the Experienced LoErrs made on average 0.07, although this difference was not statistically significant.

Concerning Total Errors, the Experienced LoErrs made significantly fewer Total Errors than the Novice HiErrs (Est = 13.577, SE = 1.618, *df* = 19.110, *t* = 8.389, *p* < 0.001) and the Experienced HiErrs (Est = 4.984, SE = 1.524, *df* = 13.180, *t* = 3.247, *p* = 0.006). The Experienced HiErrs made a similar number of Total Errors to the Novice LoErrs. In this analysis of procedure Steps, Errors, Critical Errors, and Total Errors, there were also no significant main or interaction effects of procedure number.

GEARS score

Figures 3a and b show the mean and 95% CI score of operative performance using the GEARS assessment instrument. The mean IRR for GEARS was 0.44 (0–0.8).

Figure 3a shows the comparison between Experienced and Novice groups. The former had a 19% higher score than the later, and this difference was not statistically significant.

Both groups were divided into LoErrs and HiErrs subgroups following the approach previously described. Figure 3b shows that the Experienced LoErrs score was 20% higher than the Novice HiErrs and this difference was statistically significant (Est = − 3.344, SE = 1.57, *df* = 18.228, *t* = − 2.13, *p* = 0.047).

None of the other score differences were statistically significant, and although the Experienced HiErrs scored similar to the Experienced LoErrs, they had greater score variability as demonstrated by the larger CI.

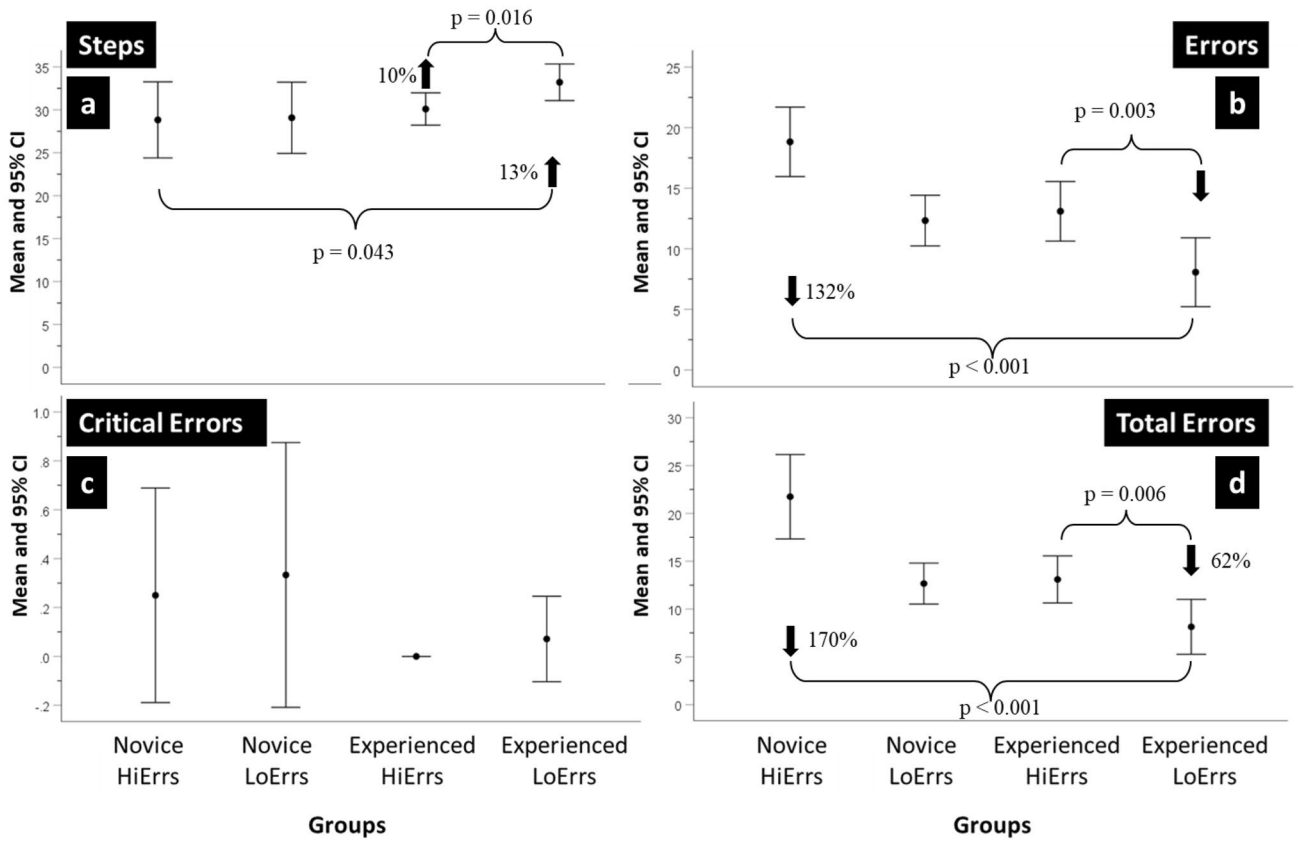


Fig. 2 **a–d** HiErrs and LoErrs using binary metrics. The mean and 95% confidence intervals (CI) of the number of Steps completed, Errors, Critical Errors and Total Errors made by Novice and Experienced surgeons in the 'high' and 'low' errors groups when performing the robot-assisted partial nephrectomy (RAPN) procedure

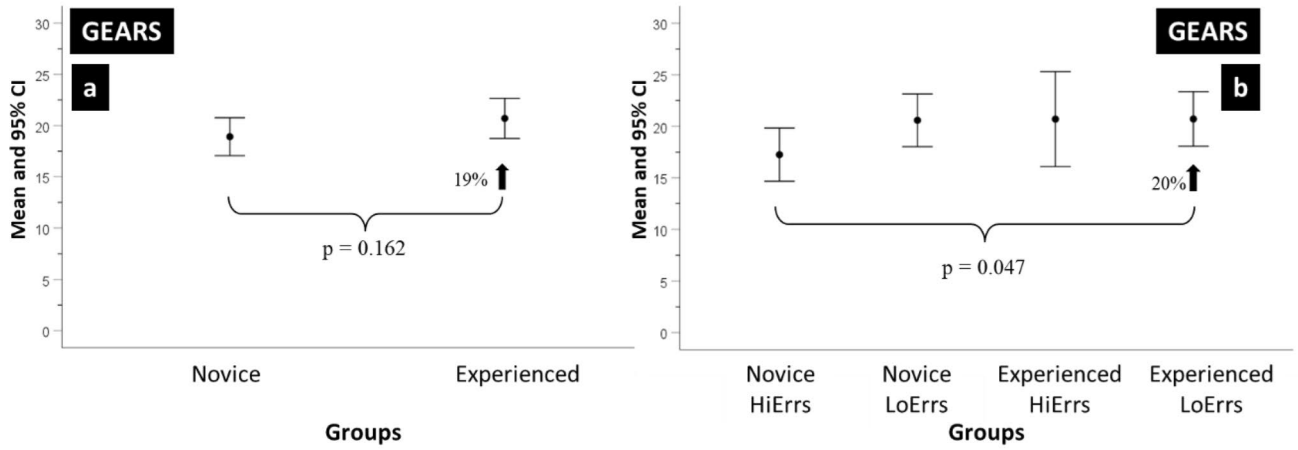
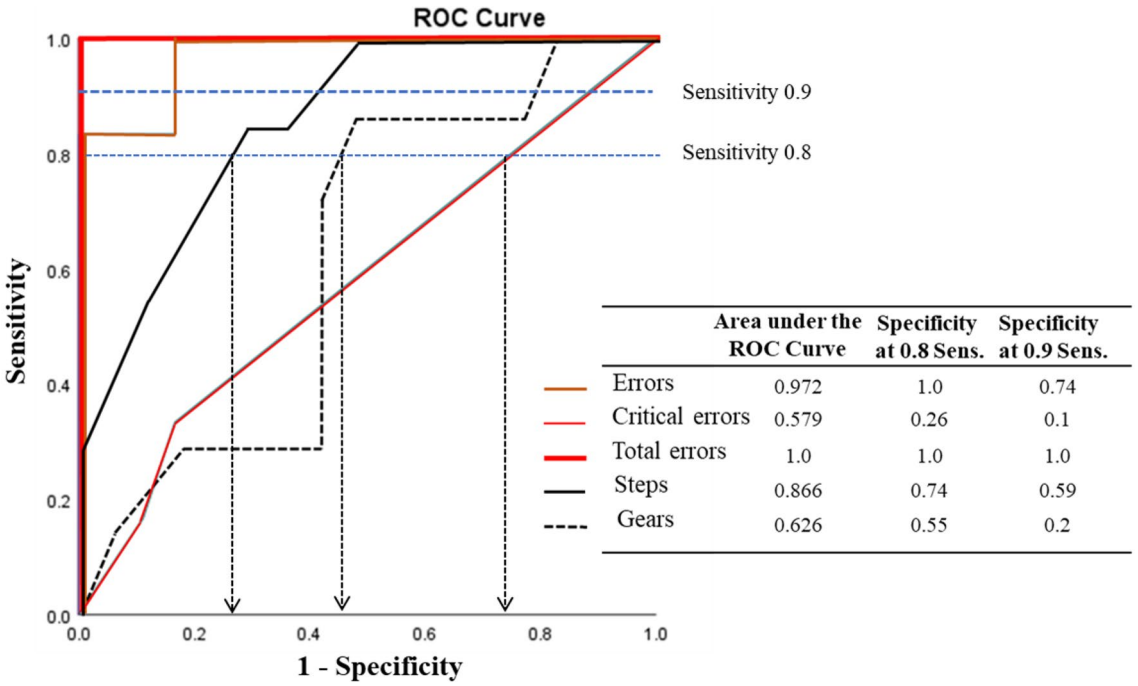


Fig. 3 **a**, **b** HiErrs and LoErrs using GEARS. The mean and 95% confidence intervals (CI) of the number GEARS scores **a** Novice and Experienced surgeons and **b** Novice and Experienced surgeons in the 'high' and 'low' errors groups when performing the robot-assisted partial nephrectomy (RAPN) procedure



	Area under the ROC Curve	Specificity at 0.8 Sens.	Specificity at 0.9 Sens.
Errors	0.972	1.0	0.74
Critical errors	0.579	0.26	0.1
Total errors	1.0	1.0	1.0
Steps	0.866	0.74	0.59
Gears	0.626	0.55	0.2

Fig. 4 ROC curve. The Receiver operator characteristic (ROC) Area Under the Curve showing the Checklist (i.e., Procedure Steps, Errors, Critical Errors and Total Errors) and GEARS Assessment discrimination (i.e., Specificity) levels for surgeon groups as Sensitivity thresholds were varied (i.e., 0.8 Sensitivity and 0.9 Sensitivity levels)

Receiver operating characteristic (ROC) analysis

We evaluated the capacity of the five assessments (i.e., Steps, procedure Errors, Critical Errors, Total Errors, and GEARS score) to discriminate the Experienced LoErrs performance in comparison to the other subgroups at a sensitivity threshold of 0.8 and 0.9.

The Area Under the Curve (AUC) for the Critical Errors was the lowest (i.e., AUC = 0.579) and the Total Errors AUC was the highest (i.e., AUC = 1.0) (Fig. 4).

The interpolation on the Specificity of a Sensitivity level of 0.8 and 0.9 was also analyzed. At a Sensitivity level of 0.8, the Specificity was 1.0 for Total Errors, 0.74 for procedure Errors, 0.55 for GEARS and 0.26 for Critical Errors. Only Total Errors demonstrated an excellent level of Specificity at a Sensitivity of 0.9 since all other measures had a Specificity level < 0.8 (Fig. 4).

Discussion

The increasing use of RAPN necessitates the requirement for better training and probably the establishment of a standardized PBP training program with the goal to improve surgical outcomes [31, 32]. Implementation of better training will allow the objective, transparent and fair assessment of surgical skills. A Delphi meeting with experienced surgeons, established the face and content validity of the RAPN metrics used in this study [23]. We then sought to establish evidence supporting the construct validity of the corresponding metrics.

We demonstrated that RAPN metrics could be scored reliably with an IRR > 0.8 and that the metrics reliably discriminated between Experienced and Novice RAPN surgeons. The metrics which discriminate best were the number of Errors and Total Errors. The Experienced group made 53% fewer objectively assessed procedure Errors than the Novice group, and the later made 69% more Total Errors than the former. We then implemented a methodology previously described [33–38], which partitioned each of the two groups at their median Total Errors score (i.e., LoErrs and HiErrs).

After delineating the four surgical performance subgroups (Experienced LowErrs, Experience HiErrs, Novice LowErrs and Novice HiErrs) key findings emerged. This novel approach to the analysis of operative performance showed a greater capacity to differentiate objectively assessed surgical performance. The Experienced LowErrs made between 132 and 170% fewer Errors and Total Errors, respectively, than the Novice HiErrs. As reported in previous studies it also became clear that surgical experience and seniority did not always translate into better surgical performance [39]. We found that the Experienced LoErrs group made 62% fewer Total Errors than Experienced HiErrs and that the

Experienced HiErrs performed at the same level as the Novice LowErrs.

It might be argued that objective assessment of intraoperative performance on one occasion is a poor indicator of surgical skill, but published evidence has already shown that peer-assessed surgical skills strongly predict clinical outcomes [33, 36].

The results of GEARS assessments demonstrated weak levels of IRR and were lower than the IRR levels observed when using RAPN binary metrics (0.44 vs 0.9). GEARS assessment also struggled to differentiate the objectively assessed performance of Experienced and Novice surgeons.

This is the first report using metrics that objectively characterize intraoperative RAPN performance, reporting quantitative evidence to support construct and discriminative validity.

These results provide a stepping stone for the construction of a standardized RAPN training program following a PBP methodology. Specifically, they provide the tool to establish performance benchmarks (i.e., proficiency levels), that trainees must demonstrate before training progression. Trainees would only be allowed to operate on the patient after demonstrating the ability to perform to the quantitatively defined performance level. They can (i) have as many training trials as they wish, (ii) be supported by faculty who know and can score the metrics reliably and iii) know how to use the metrics for deliberate rather than repeated practice [40]. Therefore, the trainee is offered metric-based training that is objective, transparent and fair and produces performances that have been shown to be ~ 60% better than traditional training courses [41]. That said, he does not progress to performing the procedure on patients until demonstrating the requisite benchmarks [13, 15, 42].

Our study found that the results were not supportive of the reliability and validity of GEARS scoring on RAPN performance, adding to the accumulating evidence that Lickert scales are not robust enough to score surgical performance.

Study limitations

- The number of RAPN videos evaluated might limit the generalizability of our analysis and any firm conclusions about performance variability by very experienced operators. There is, however, a steady stream of reports on very experienced surgeons that perform poorly on straightforward and familiar tasks when their performance is objectively assessed by reviewers blinded to experience level [35, 37, 38, 43, 44];
- The RAPN metrics used in this study are only applicable to left-sided transperitoneal RAPN. Right-sided and retroperitoneal approaches, cannot be scored using these metrics, although the large majority of evidence on RAPN techniques and outcomes refer to the transperitoneal approach;
- These metrics were developed for straightforward cases and patient characteristics (i.e., age, body mass index, previous abdominal surgery or other comorbidity indexes) were not taken into account which may have influenced the differences in performance. However, if a trainee cannot do a straightforward case, they probably should not be performing a more complex one;
- Although *Novice* surgeons were required to complete the RAPN independently, we cannot exclude a marginal bias derived from the impact of clinical supervision from an *Experienced* surgeon. This, however, would mean that the differences observed in this study are an underestimation of the real differences.

Conclusions

We report RAPN metrics that reliably and consistently discriminate the intraoperative performance of Expert and Novice Surgeons. The number of Total Errors demonstrated the highest level of specificity and sensitivity in this discrimination. GEARS demonstrated poor reliability in scoring RAPN surgical performance. These metrics lay the foundation to implement a simulation-based PBP training program.

Patient summary

We developed metrics for scoring RAPN performance. Two experienced surgeons, trained to use RAPN metrics, scored anonymized video recorded RAPN performance of novice and expert surgeons. Our study showed that the RAPN metrics consistently and reliably identify "true" Experts and "true" Novices.

Acknowledgements Alessandro Antonelli (Department of Urology, ASST Spedali Civili Hospital, University of Brescia, Italy), Antonio Grosso (Department of Urology, University of Florence, Careggi Hospital, Italy), Andrea Minervini (Department of Urology, University of Florence, Careggi Hospital, Italy), Alessandro Princiotta (Department of Urology, ASST Spedali Civili Hospital, University of Brescia, Italy), Ben Challacombe (Department of Urology, Guy's and St Thomas Hospitals NHS Foundation Trust, UK), Christophe Vaessen (Department of Urology, Pitié-Salpêtrière University Hospital Group, France), Cosimo De Carne (Department of Urology, University of Modena and Reggio Emilia, Italy), Erdem Canda (Department of Urology, Koç University School of Medicine, Istanbul, Turkey), Evangelos Malovrouvas, (Department of Urology, University Hospital of Galway, County Galway, Ireland), Farleigh Reeves (Department of Urology, Guy's and St Thomas Hospitals NHS Foundation Trust, UK), Geert De Naeyer (Department of Urology, Onze-Lieve-Vrouw Ziekenhuis, Belgium), Josep Gaya (Department of Urology, Fundació Puigvert, Universitat Autonoma de Barcelona, Spain), Nicoló Buffi (Department of Urology, Humanitas Clinical and Research Centre, Italy), Paolo Verri (Department of Urology, Fundació Puigvert, Universitat Autonoma de Barcelona, Spain), Ruben De Groote (Department of Urology,

Onze-Lieve-Vrouw Ziekenhuis, Belgium), Stephan Buse (Department of Urology and Urologic Oncology, Alfried Krupp Krankenhaus, Essen, Germany), Stefano Puliatti (Department of Urology, University of Modena and Reggio Emilia, Italy). Alessandro Antonelli, Antonio Grosso, Andrea Minervini, Alessandro Princiotta, Ben Challacombe, Christophe Vaessen, Cosimo De Carne, Erdem Canda, Evangelos Malovrouvas, Farleigh Reeves, Geert De Naeyer, Josep Gaya, Nicoló Buffi, Paolo Verri, Ruben De Groote, Stephan Buse, Stefano Puliatti.

Author contributions Rui Farinha had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Study concept and design: Farinha, Gallagher. Acquisition of data: Farinha, Breda, Porter, Mottrie, Van Cleynenbreugel, Vander Sloten, Mottaran, Gallagher. Analysis and interpretation of data: Farinha, Gallagher. Drafting of the manuscript: Farinha, Gallagher. Critical revision of the manuscript for important intellectual content: Farinha, Breda, Porter, Mottrie, Van Cleynenbreugel, Vander Sloten, Gallagher. Statistical analysis: Farinha, Gallagher. Obtaining funding: Mottrie, Gallagher. Administrative, technical or material support: Farinha, Gallagher. Supervision: Gallagher. Other: None.

Funding The present research project has been conducted by Rui Farinha as part of his PhD studies in KU Leuven, Belgium, and of the ongoing project for the ERUS and ORSI Academy. Medtronic (Minneapolis, USA) provided the unrestricted educational grant for this study but did not influence in selecting the experts, design and conduct of the research, data collection, analysis and preparation of the manuscript.

Data availability All data were obtained from the experiments conducted by the authors.

Declarations

Competing interests The authors declare no competing interests.

Conflict of interest Rui Farinha certifies that all 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. Ljungberg B, Bensalah K, Canfield S, Dabestani S, Hofmann F, Hora M et al (2015) EAU guidelines on renal cell carcinoma: 2014 update. Eur Urol 67:913–924. https://doi.org/10.1016/j.eururo.2015.01.005
2. Thompson RH, Boorjian SA, Lohse CM, Leibovich BC, Kwon ED, Cheville JC et al (2008) Radical nephrectomy for pT1a renal masses may be associated with decreased overall survival compared with partial nephrectomy. J Urol 179:463–468. https://doi.org/10.1016/j.juro.2007.09.077
3. Roos FC, Steffens S, Junker K, Janssen M, Becker F, Wegener G et al (2014) Survival advantage of partial over radical nephrectomy in patients presenting with localized renal cell carcinoma. BMC Cancer 14:372. https://doi.org/10.1186/1471-2407-14-372
4. Alan WP, McConnell DH, Craig AP (2020) Campbell Walsh Wein Urology
5. Joseph ASJr., Stuart SH (2019) Hinman's Atlas of Urologic Surgery. 4th edn. Saunders
6. Tang AB, Lamaina M, Childers CP, Mak SS, Ruan Q, Begashaw MM et al (2021) Perioperative and long-term outcomes of robot-assisted partial nephrectomy: a systematic review. Am Surg 87:21–29. https://doi.org/10.1177/0003134820948912
7. Buffi NM, Saita A, Lughezzani G, Porter J, Dell'Oglio P, Amparore D et al (2020) Robot-assisted partial nephrectomy for complex (PADUA score ≥10) tumors: techniques and results from a multicenter experience at four high-volume centers. Eur Urol 77:95–100. https://doi.org/10.1016/j.eururo.2019.03.006
8. Casale P, Lughezzani G, Buffi N, Larcher A, Porter J, Mottrie A (2019) Evolution of robot-assisted partial nephrectomy: techniques and outcomes from the transatlantic robotic nephron-sparing surgery study group. Eur Urol 76:222–227. https://doi.org/10.1016/j.eururo.2018.11.038
9. Peyronnet B, Tondut L, Bernhard J-C, Vaessen C, Doumerc N, Sebe P et al (2018) Impact of hospital volume and surgeon volume on robot-assisted partial nephrectomy outcomes: a multicentre study. BJU Int 121:916–922. https://doi.org/10.1111/bju.14175
10. Larcher A, Muttin F, Peyronnet B, De Naeyer G, Khene Z-E, Dell'Oglio P et al (2019) The learning curve for robot-assisted partial nephrectomy: impact of surgical experience on perioperative outcomes. Eur Urol 75:253–256. https://doi.org/10.1016/j.eururo.2018.08.042
11. Birkmeyer JD, Stukel TA, Siewers AE, Goodney PP, Wennberg DE, Lucas FL (2003) Surgeon volume and operative mortality in the United States. N Engl J Med 349:2117–2127. https://doi.org/10.1056/NEJMsa035205
12. Asch DA, Weinstein DF (2014) Innovation in medical education. N Engl J Med 371:794–795. https://doi.org/10.1056/NEJMp1407463
13. Gallagher AG (2012) Metric-based simulation training to proficiency in medical education:- what it is and how to do it. Ulster Med J 81:107–113
14. Anthony GG, O'Sullivan GC (2012) Fundamentals of surgical simulation; principles and practices. Springer-Verlag, London
15. Gallagher AG, Ritter EM, Champion H, Higgins G, Fried MP, Moses G et al (2005) Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. Ann Surg 241:364–372. https://doi.org/10.1097/01.sla.0000151982.85062.80
16. Breen D, O'Brien S, McCarthy N, Gallagher A, Walshe N (2019) Effect of a proficiency-based progression simulation programme on clinical communication for the deteriorating patient: a randomised controlled trial. BMJ Open 9:e025992. https://doi.org/10.1136/bmjopen-2018-025992
17. Angelo RL, Ryu RKN, Pedowitz RA, Beach W, Burns J, Dodds J et al (2015) A Proficiency-based progression training curriculum coupled with a model simulator results in the acquisition of a superior arthroscopic bankart skill set. Arthrosc J Arthrosc Relat Surg Off Publ Arthrosc Assoc North Am Int Arthrosc Assoc 31:1854–1871. https://doi.org/10.1016/j.arthro.2015.07.001
18. Kallidaikurichi Srinivasan K, Gallagher A, O'Brien N, Sudir V, Barrett N, O'Connor R et al (2018) 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 8:e020099. https://doi.org/10.1136/bmjopen-2017-020099
19. Pedowitz RA, Nicandri GT, Angelo RL, Ryu RKN, Gallagher AG (2015) Objective Assessment of knot-tying proficiency with the fundamentals of arthroscopic surgery training program workstation and knot tester. Arthrosc J Arthrosc Relat Surg Off Publ Arthrosc Assoc North Am Int Arthrosc Assoc 31:1872–1879. https://doi.org/10.1016/j.arthro.2015.06.021
20. Van Sickle KR, Ritter EM, Baghai M, Goldenberg AE, Huang I-P, Gallagher AG et al (2008) Prospective, randomized, double-blind trial of curriculum-based training for intracorporeal suturing and knot tying. J Am Coll Surg 207:560–568. https://doi.org/10.1016/j.jamcollsurg.2008.05.007
21. Ahlberg G, Enochsson L, Gallagher AG, Hedman L, Hogman C, McClusky DA et al (2007) Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. Am J Surg 193:797–804. https://doi.org/10.1016/j.amjsurg.2006.06.050
22. Seymour NE, Gallagher AG, Roman SA, O'Brien MK, Bansal VK, Andersen DK et al (2002) Virtual reality training improves operating room performance results of a randomized, double-blinded study. Ann Surg 236:458–464. https://doi.org/10.1097/00000658-200210000-00008
23. Rui F, Alberto B, James P, Alexandre M, Ben Van C, Jozef VS, RAPN-DS gorup AG. International expert consensus on a metric-based characterization of Robot-Assisted Partial Nephrectomy (RAPN). Manuscr Submitt Publ n.d.
24. Gallagher AG, Ritter EM, Satava RM (2003) Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. Surg Endosc. https://doi.org/10.1007/s00464-003-0035-4
25. Gallagher AG, Lederman AB, McGlade K, Satava RM, Smith CD (2004) Discriminative validity of the minimally invasive surgical trainer in virtual reality (MIST-VR) using criteria levels based on expert performance. Surg Endosc 18:660–665. https://doi.org/10.1007/s00464-003-8176-z
26. Mascheroni J, Mont L, Stockburger M, Patwala A, Retzlaff H, Gallagher AG (2019) International expert consensus on a scientific approach to training novice cardiac resynchronization therapy implanters using performance quality metrics. Int J Cardiol 289:63–69. https://doi.org/10.1016/j.ijcard.2019.04.036
27. Gallagher AG, Ryu RKN, Pedowitz RA, Henn P, Angelo RL (2018) Inter-rater reliability for metrics scored in a binary fashion-performance assessment for an arthroscopic bankart repair. Arthrosc J Arthrosc Relat Surg Off Publ Arthrosc Assoc North Am Int Arthrosc Assoc 34:2191–2198. https://doi.org/10.1016/j.arthro.2018.02.007
28. Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondré K, Stanbridge D et al (2005) A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg 190:107–113. https://doi.org/10.1016/j.amjsurg.2005.04.004
29. Kazdin AE (1977) Artifact, bias, and complexity of assessment: the ABCs of reliability. J Appl Behav Anal 10:1311161. https://doi.org/10.1901/jaba.1977.10-141
30. Alan EK (2013) Behavior modification in applied settings. 7th edn. Waveland Press, Inc.
31. Palagonia E, Mazzone E, De Naeyer G, D'Hondt F, Collins J, Wisz P et al (2020) The safety of urologic robotic surgery depends on the skills of the surgeon. World J Urol 38:1373–1383. https://doi.org/10.1007/s00345-019-02901-9
32. Mazzone E, Dell'Oglio P, Mottrie A (2019) Outcomes report of the first ERUS robotic urology curriculum-trained surgeon in Turkey: the importance of structured and validated training programs for global outcomes improvement. Turkish J Urol 45:189–190. https://doi.org/10.5152/tud.2019.19019
33. Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR et al (2013) Surgical skill and complication rates after bariatric surgery. N Engl J Med 369:1434–1442. https://doi.org/10.1056/NEJMsa1300625
34. Mascheroni J, Mont L, Stockburger M, Patwala A, Retzlaff H, Gallagher AG (2020) A validation study of intraoperative performance metrics for training novice cardiac resynchronization therapy implanters. Int J Cardiol 307:48–54. https://doi.org/10.1016/j.ijcard.2020.02.003
35. Mottrie A, Mazzone E, Wiklund P, Graefen M, Collins JW, De Groote R et al (2020) Objective assessment of intraoperative skills for robot-assisted radical prostatectomy (RARP): results from the ERUS Scientific and Educational Working Groups Metrics Initiative. BJU Int. https://doi.org/10.1111/bju.15311
36. Curtis NJ, Foster JD, Miskovic D, Brown CSB, Hewett PJ, Abbott S et al (2020) Association of surgical skill assessment with clinical outcomes in cancer surgery. JAMA Surg 155:590. https://doi.org/10.1001/jamasurg.2020.1004
37. Gómez Ruiz M, Tou S, Gallagher AG, Cagigas Fernández C, Cristobal Poch L, Matzel KE (2022) Intraoperative robotic-assisted low anterior rectal resection performance assessment using procedure-specific binary metrics and a global rating scale. BJS Open. https://doi.org/10.1093/bjsopen/zrac041
38. Kojima KE, Graves M, Taha W, Ghidinelli M, Struelens B, Aliaga JAA et al (2022) Discrimination, reliability, sensitivity, and specificity of metric-based assessment of an unstable pertrochanteric 31A2 intramedullary nailing procedure performed by experienced and novice surgeons. Injury 53:2832–2838. https://doi.org/10.1016/j.injury.2022.05.056
39. Begg CB, Riedel ER, Bach PB, Kattan MW, Schrag D, Warren JL et al (2002) Variations in morbidity after radical prostatectomy. N Engl J Med 346:1138–1144. https://doi.org/10.1056/NEJMsa011788
40. Ericsson KA, Krampe RT, Tesch-Römer C (1993) The role of deliberate practice in the acquisition of expert performance. Psychol Rev 100:363–406. https://doi.org/10.1037/0033-295X.100.3.363
41. Mazzone E, Puliatti S, Amato M, Bunting B, Rocco B, Montorsi F et al (2020) A systematic review and meta-analysis on the impact of proficiency-based progression simulation training on performance outcomes. Ann Surg. https://doi.org/10.1097/SLA.0000000000004650
42. Gallagher AG, O'Sullivan GC (2011) Fundamentals of surgical simulation: principles and practice. Springer Publishing Company, Incorporated https://doi.org/10.1007/978-0-85729-763-1
43. Gallagher AG, Smith CD, Bowers SP, Seymour NE, Pearson A, McNatt S et al (2003) Psychomotor skills assessment in practicing surgeons experienced in performing advanced laparoscopic procedures. J Am Coll Surg 197:479–488. https://doi.org/10.1016/S1072-7515(03)00535-0
44. Crothers IR, Gallagher AG, McClure N, James DTD, McGuigan J (1999) Experienced laparoscopic surgeons are automated to the "fulcrum effect": an ergonomic demonstration. Endoscopy 31:365–369. https://doi.org/10.1055/s-1999-26