BJS, 2023, 1–20
https://doi.org/10.1093/bjs/znad331
Systematic Review

Evaluation of objective tools and artificial intelligence in robotic surgery technical skills assessment: a systematic review

Matthew W. E. Boal[1,2,3], Dimitrios Anastasiou[2,4], Freweini Tesfai[1,2], Walaa Ghamrawi[1], Evangelos Mazomenos[2,4], Nathan Curtis[5], Justin W. Collins[3,6], Ashwin Sridhar[3,6], John Kelly[3,6], Danail Stoyanov[2,7] and Nader K. Francis[1,3,8,*]

[1]The Griffin Institute, Northwick Park & St Marks' Hospital, London, UK
[2]Wellcome/ESPRC Centre for Interventional Surgical Sciences (WEISS), University College London (UCL), London, UK
[3]Division of Surgery and Interventional Science, Research Department of Targeted Intervention, UCL, London, UK
[4]Medical Physics and Biomedical Engineering, UCL, London, UK
[5]Department of General Surgey, Dorset County Hospital NHS Foundation Trust, Dorchester, UK
[6]University College London Hospitals NHS Foundation Trust, London, UK
[7]Computer Science, UCL, London, UK
[8]Yeovil District Hospital, Somerset Foundation NHS Trust, Yeovil, Somerset, UK

*Correspondence to: Nader Francis, The Griffin Institute, Y Block, Northwick Park & St Mark's Hospital, London HA1 3UJ, UK (e-mail: n.francis@griffininstitute.org.uk)

Abstract

Background: There is a need to standardize training in robotic surgery, including objective assessment for accreditation. This systematic review aimed to identify objective tools for technical skills assessment, providing evaluation statuses to guide research and inform implementation into training curricula.

Methods: A systematic literature search was conducted in accordance with the PRISMA guidelines. Ovid Embase/Medline, PubMed and Web of Science were searched. Inclusion criterion: robotic surgery technical skills tools. Exclusion criteria: non-technical, laparoscopy or open skills only. Manual tools and automated performance metrics (APMs) were analysed using Messick's concept of validity and the Oxford Centre of Evidence-Based Medicine (OCEBM) Levels of Evidence and Recommendation (LoR). A bespoke tool analysed artificial intelligence (AI) studies. The Modified Downs–Black checklist was used to assess risk of bias.

Results: Two hundred and forty-seven studies were analysed, identifying: 8 global rating scales, 26 procedure-/task-specific tools, 3 main error-based methods, 10 simulators, 28 studies analysing APMs and 53 AI studies. Global Evaluative Assessment of Robotic Skills and the da Vinci Skills Simulator were the most evaluated tools at LoR 1 (OCEBM). Three procedure-specific tools, 3 error-based methods and 1 non-simulator APMs reached LoR 2. AI models estimated outcomes (skill or clinical), demonstrating superior accuracy rates in the laboratory with 60 per cent of methods reporting accuracies over 90 per cent, compared to real surgery ranging from 67 to 100 per cent.

Conclusions: Manual and automated assessment tools for robotic surgery are not well validated and require further evaluation before use in accreditation processes.

PROSPERO: registration ID CRD42022304901

Introduction

Robotic surgery is increasingly being adopted due to improved vision, dexterity and surgical ergonomics. In selected procedures there is supportive evidence demonstrating non-inferiority and lower morbidity compared to laparoscopy[1–5]. Minimally invasive surgery (MIS) is complex, highly variable and requires technical skill with unfavourable error profiles compared to industrial data[6]. Meanwhile, the addition of new technology into the operating room, with novel technical and non-technical considerations, increases the potential for human error, and therefore patient risk[7]. Of surgical patients, 10–15 per cent in the UK experience adverse events, of which 50 per cent are preventable[8]. Adverse events relating to robotic procedures (10 624) were reported in the USA between 2000 and 2013[9] while a global independent review on health technology hazards identified a lack of robotic surgical training as one of the top 10 risks to patients[10]. This deficit is being addressed through development and standardization of basic and specialty curricula[11–23].

Robotic surgical procedures require high levels of experience. Evaluation of performance in surgery is shifting from time- and operative numbers-based assessment towards proficiency-based training and accreditation[24]. To assist this, objective tools are frequently employed but must be fully evaluated if they are to be used as summative, high-stakes assessment instruments. Traditionally, proficiency in surgery was extrapolated from clinical outcomes such as histopathology, morbidity and mortality, yet these are subject to multifactorial influences. Intraoperative performance analysis has proved to be a fruitful area for

Received: July 11, 2023. **Revised:** September 18, 2023. **Accepted:** September 19, 2023

assessment of performance and intervention delivery[6,25–27]. This facilitates direct formative feedback to guide reduction in proficiency curves, as well as summative assessment[26]. Objective assessments within MIS have demonstrated reliability and clinically relevant validities[6,25–27], leading to the development of tools to aid this[28]. However, previous studies have highlighted the variability in reporting on the validity and reliability of manual tools[28,29], which risks undermining truly objective skills assessment. A full appraisal of the literature on objective assessment tools, therefore, is imperative to inform learning and accreditation processes in robotic surgery.

Prior systematic reviews have focused on one aspect of technical skills assessment[30,31] or combinations of surgical approaches[29]. Other reports provide an overview but lack scope and granularity of type of validity and reliability of assessment or fail to grade the evidence[32]. Finally, given the rapid uptake of robotic techniques and development of artificial intelligence (AI) methods, many reports are now outdated[33–37], requiring up-to-date evaluation[38–40].

The aim of this systematic review is to provide an up-to-date and comprehensive evaluation of objective, technical skill assessment tools in robotic surgery.

Methods

This systematic review followed an *a priori* protocol (PROSPERO registration ID CRD42022304901). The Covidence® platform was used to screen studies, exclude duplications and extract data.

Search strategy

A systematic search of the literature was conducted in line with the PRISMA guidelines[41]. Ovid Embase/Medline, PubMed and Web of Science databases were searched from conception to 22 February 2022. *Table S1* outlines the full search strategy. Searches were performed independently by two authors using medical subject headings (MeSH) terms for ‘Objective’, ‘Assessment’, ‘Tool’, ‘Error’, ‘Skill’, ‘Robot’ and ‘Surgery’, which were combined with Boolean operators ‘AND’ and ‘OR’. Studies from knowledge of the field and references from relevant articles, including one literature[32] and four systematic reviews[28,29,31,42], were additionally screened. Conference proceedings and journal supplement abstracts were considered relevant if meaningful data were available.

Selection of eligible studies

Four reviewers independently screened, reviewed and extracted data, with the primary investigator reviewing all articles. Disagreements were resolved through discussion with the corresponding author.

Included studies followed the PICO question:

- Population—participants being assessed on robotic technical skill.
- Intervention—an objective technical skill assessment tool or method is developed and/or implemented.
- Comparison—to other tools or measurement of assessment.
- Outcome—validity, reliability, accuracy, impact on the participant.

Exclusion criteria were solely laparoscopic and/or open assessment skills, or failures to retrieve the article or an English translation.

Data extraction

All studies

Study details including year, country, participant number, participant expertise level and evaluator type were extracted. Identified studies were grouped based on study and tool types into manual, automated performance metrics (APMs) and evaluation of statistical models or AI algorithms. These domains of technical skill assessment were devised using approaches employed by previous reviews and that reflect different assessment methods. The manual domain is human assessment with subgroups that are global rating scales, procedure-specific and error-based tools. APMs are metrics produced by computer software typically in virtual reality (VR) simulators. Finally, AI algorithms are mathematical models implemented to process input data and estimate skill or clinically related outputs, for example, using kinematic data to predict postoperative urinary incontinence[36] or vision data to predict skill level (*Fig. 1*).

Manual and APM studies

Due to the heterogeneity in methods of technical skill assessment, different approaches were applied to facilitate evaluation. Manual tools and automated performance metrics (non-AI articles) were evaluated using Messick’s validity concept[43] and the Modified Educational Oxford Centre for Evidence-Based Medicine (OCEBM) Levels of Evidence (LoE) and Levels of Recommendation (LoR)[44]. Messick’s concept views validity as a continuous process and a combination of the classical views of face, content, construct and predictive validity, internal consistency, intra- and inter-rater reliability. Instead of viewing these as separate, five aspects that need to be considered for an assessment tool to be valid were assessed (*Table 1*). Strength of correlational analyses and significance was also extracted using standardized definitions.

Artificial intelligence studies

AI specialists contributed to screening, data extraction and evaluation and a bespoke data extraction template was employed to standardize data capture.

Methodological quality assessment

Methodological quality assessment was evaluated using a modifiable Downs–Black checklist (*Table S2*)[45]. Due to study heterogeneity some aspects were not applicable; therefore, taking a pragmatic approach, we modified the score in these circumstances, with a maximum score of 10 available. For AI studies, it was not feasible to apply a relevant methodological quality tool such as the Downs–Black checklist or Medical Education Research Study Quality Instrument (MERSQI)[46], as most study designs are conceptual.

Tables 2–4 summarize the main tools in each domain of technical skill assessment and *Supplementary Tables (Tables S3 to S8)* describe each study’s analysis. Summaries of the remaining tools can be viewed in *Table S9*.

Results

Two thousand, nine hundred and forty-four studies were identified from searches with 85 identified from additional sources. Seven hundred and forty-nine duplicates were removed. Of 2280 studies that were screened and reviewed, 2033 were excluded with 247 studies undergoing data extraction (*Fig. 2*). Two hundred and twenty-seven studies were classified as observational, including Delphi meetings, experimental, cohort and randomized studies

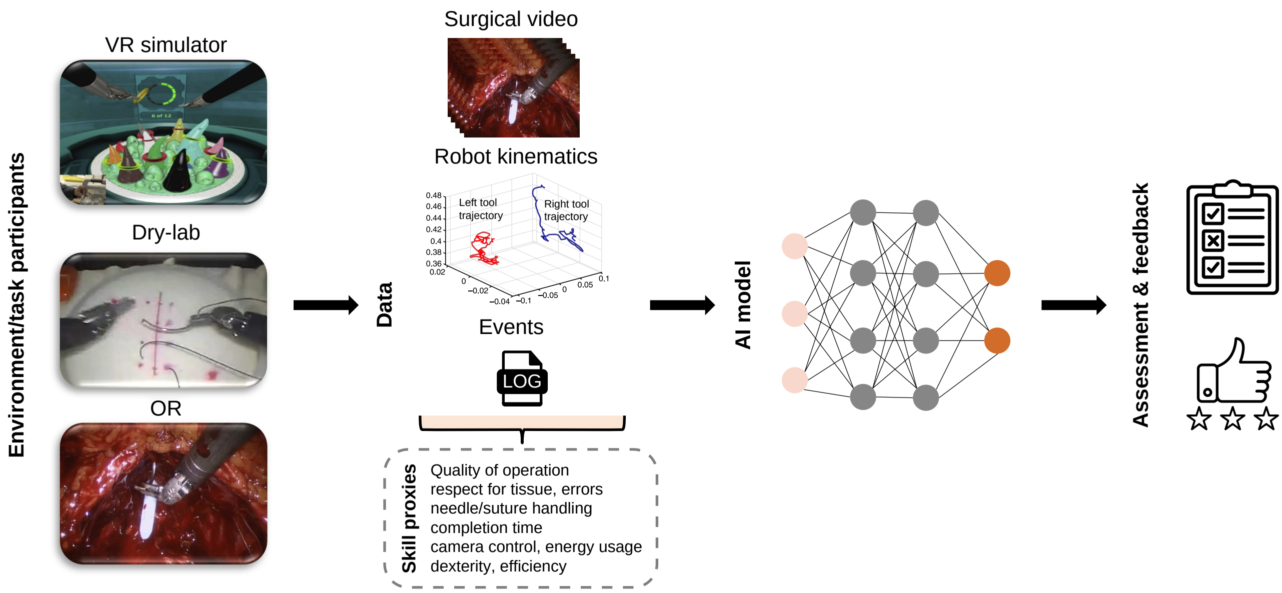

Fig. 1 Artificial intelligence framework for surgical skill assessment

Table 1 Messick's validity framework adapted from[143,144]

Source of validity	Evidence	Examples
Test content, that is face and content validity	The test's content and the construct it is intended to measure	Delphi methodology/expert consensus development Questionnaires for realism and usefulness
Response process	Analysis of raters, that is how well they respond to the test and steps taken to improve the validity	Rater training/orientation/familiarization Randomization Powered study APMs eliminate rater bias
Internal structure, that is reliability	Degree to which domains and aspects of the tool fit the underlying construct	Intra- and inter-rater reliability, internal consistency APMs eliminated rater subjectivity
Relationship to other variables	Evaluating scores' associations/correlations, whether they are positive or negative, strong or weak, and with other variables including discriminative ability	Concurrent, construct and predictive validity Generalizability of the evidence Learning curves
Consequence	Impact of the assessment	Pass/fail/benchmarking of scores Impact or consequences on participants future/ learning

not defined as randomized control trials (RCTs), while a total of 20 RCTs were identified. Of the manual studies, 93 used global rating scales (GRS), 45 procedure- or task-specific tools, 43 error-based, 77 simulator automated performance metrics, 28 non-simulator automated performance metrics and 53 AI studies.

Global rating scale tools

Eight different GRS tools were identified (*Table 2* and *Table S3*). Global Evaluative Assessment of Robotic Skills (GEARS) was the most utilized assessment method, assessed in 58 studies, including 12 RCTs giving a Level 1 recommendation based on 21 studies reporting excellent reliability and 3 low/poor. Interestingly, crowd-sourced GEARS ratings demonstrated excellent inter-rater reliability[47], good to strong/excellent inter-observer group reliability compared to experts[48–51], as well as construct[48,52] and predictive validity[53]. GEARS (all raters) demonstrated supportive evidence of 'relationship to other variables' including concurrent (17 studies), construct (25 studies) and predictive validity (3 studies).

Objective Structured Assessment of Technical Skills (OSATS), Global Operative Assessment of Laparoscopic Skills (GOALS) and Robotic-OSATS (R-OSATS) were used in a total of 34 studies and all received a Level 2 recommendation, despite only one of them being robotic-specific. GEARS has no robust data validating a benchmark for overall and domain scores, whereas GOALS and R-OSATS used the contrasting groups method[54] and the modified Angoff method[55], setting competency at 80 per cent and 70 per cent, respectively. All other tools identified have not been thoroughly evaluated with LoR 3 or 4 (*Table S9*).

Procedure- and task-specific tools

Twenty-six different types of procedure- and task-specific tool were identified in 45 studies (*Table 2*, *Tables S4, S9*). Of these, 22 (48.9 per cent) studies containing 17 (37.8 per cent) different tools used full procedural data[56–77]. With regards to specialty, 21 (46.7 per cent) studies[50,56,57,64,66–74,76,78–84] assessed 15 urology tools, 9 (20 per cent) studies[58–61,63,75,77,85,86] assessed 8 general surgery tools, 3 (6.7 per cent) gynaecology tools[58,65,87], 1 (2.2 per

Table 2 Summary of manual assessment tools

Tool	Study type	Setting	Test content	Response process	Internal structure	Relationship to other variables	Consequences	LoE	LoR
Global rating scale tools									
GEARS	46 Observational 12 RCTs	36 Lab 17 OR 5 Lab & OR	Experts developed based on GOALS[145]	All studies (except one[53]) demonstrated response process	*Inter-rater reliability* Strong/Very strong/ Excellent/High[47–51,71,82,83,145–156] Acceptable/Good/ Moderate[15,48,51,157,158] Low/Poor[50,159] *Intra-rater reliability* Excellent[148] *Internal consistency* Excellent[105,145,160] Low[49]	*Concurrent validity* Other GRS tools[49] Task-/ procedure-specific tools[77,80,81,146,161] Error tools[162] Virtual reality[146,163–166] APMs[52,84,105,154,157] Non-technical skills[167] *Construct validity*[47,52,77,82,84,99,105,145,147,148,150,151,153,154,157,161,163,167–174] *Predictive validity*[53,162,175]	80–100% = good to excellent (arbitrary definition[73]) Benchmarked using expert scores[167,169]	Level 1b[15,83,169] Level 2a[82,149,153,159,160,165,166,168,176,177] Level 2b[47–52,77,80,81,84,99,105,113,145–148,150,151,154–156,161,163,164,167,170,171,174,178] Level 3 [53,71,73,152,158,162,172,179–182,282] Level 4[58,64,175]	Level 1 recommendation
OSATS	20 Observational 3 RCT	19 Lab 3 OR 1 Lab & OR		All but two studies[183,184] demonstrated response process	*Inter-rater reliability* High/Excellent[90,185,186] Good/Moderate[86,187–189] *Intra-rater reliability* Strong[86] *Internal consistency* Excellent[105]NB OSATS/ GEARS combo	*Concurrent validity* Task-/procedure-specific tools[90] Error tools[190] Virtual reality[90,191] APMs[105] Cognitive load[192] *Construct validity*[54,62,105,185,187,188,193,194] *Predictive validity*[86]	Hypothesized mean OSATS category scores 3.5 novice and 4.5 expert was in concordance with results of mean[188] Expert benchmarking[62,184,185]	Level 2a[63,195,196] Level 2b[54,62,90,105,183,185–188,189,192–194] Level 3[86,88,184,190,191,197,198]	Level 2 recommendation (NB if combined with OSATS task-specific, it would still be LoR 2)
GOALS	4 Observational 3 RCTs	5 Lab 1 OR 1 Lab & OR	Expert group added 2 domains to GOALS creating GOALS + [199]	All studies demonstrated response process	*Internal consistency* High[160]	*Concurrent validity* Other GRS tools[160] Virtual reality[160,200] *Construct validity*[54,199,201]	Pass mark defined by contrasting groups method[54] by experts[92]	Level 1b[200] Level 2a[160] Level 2b[54,92,199,201] Level 4[58]	Level 2 recommendation
R-OSATS	4 Observational	4 Lab	Developed from GOALS and OSATS[202]	All studies demonstrated response process	*Inter-rater reliability* Strong/Very high/ Excellent[55,203,204] Acceptable/Moderate/ Good[202,203] *Intra-rater reliability* Very strong[202] Moderate/Good[203]	*Concurrent validity with VR*[204] *Construct validity*[202,204]	Modified Angoff method set threshold competency scores per drill[55]	Level 2b[55,202,204] Level 3[203]	Level 2 recommendation
Procedure- and Task-Specific Assessment Tools									
OSATS *Task-specific* Tools are for separate procedures/ tasks	5 Observational 2 RCTs	7 Lab	No studies	All demonstrated response process	*Inter-rater reliability* High[90]	*Concurrent validity* Non-simulator APMs[89] Simulator APMs[90] *Construct validity*[62,98]	Pass mark based on experts[62]	Level 2a[63] Level 2b[62,90,98] Level 3[88,89,91]	Level 2 recommendation (NB if combined with OSATS GRS, would still be Level 2)
RACE	5 Observational 1 RCT	3 Lab 2 OR 1 Delphi and OR	Delphi consensus[81]	All demonstrated response process	*Inter-rater reliability* Strong/Excellent[50,83] Good/Moderate[50] *Intra-rater reliability* Good[81]	*Concurrent validity* GEARS[81] UVA leak on model[84] EASE suturing tool[205] *Construct validity*[81,84]		Level 1b[83] Level 2b[50,81,84,205] Level 3[78]	Level 2 recommendation
Task-Perform-ance Metrics Tools Tools are for separate procedures/ tasks	1 Observational 2 Delphi and video rating OR videos 2 Delphi and Lab	1 Delphi 1 OR 1 Delphi and OR 2 Delphi and OR	All tools developed through Delphi consensus	All tools demonstrated response process	*Inter-rater reliability* Percentage agreement 0.85–0.96[76,77,93,97]	*Construct validity*[76,77,93,97]	Anastomotic leak test[93]	Level 2b[76,77,93,97] Level 4[75]	Level 2 recommendation

(continued)

Table 2 (continued)

Global rating scale tools

Tool	Study type	Setting	Test content	Response process	Internal structure	Relationship to other variables	Consequences	LoE	LoR
Error Assessment Tools									
FLS Scoring System	8 Observational 1 RCT 2	8 Lab 1 OR and Lab	None	2 demonstrated response process[169,206]	*Inter-rater reliability* Excellent[207] *Intra-rater reliability* Excellent[207] *Internal consistency* Good[207]	*Construct validity*[169,208–210] *Concurrent validity with* GEARS[146]	Expert defined proficiency/pass fail marks[169,207–209]	Level 2a[169] Level 2b[146,206,208,209] Level 3[95,207,209,211]	Level 2 recommendation
Task-Perform-ance Metrics Tools Tools are for separate procedures/ tasks	1 Observational 2 Delphi and video rating OR videos 2 Delphi and Lab	1 Delphi 1 OR 1 Delphi and OR 2 Delphi and OR	All tools developed through Delphi consensus	All tools demonstrated response process	*Inter-rater reliability* Percentage agreement 0.85–0.96[76,77,93,97]	*Construct validity*[76,77,93,97]	Anastomotic leak test[93]	Level 2b[76,77,93,97] Level 4[75]	Level 2 recommendation
Generic Error Rating Tool	2 Observational	2 OR	None	Both demonstrated response process		*Concurrent validity* GEARS[162] Clinical adverse events (presumed intra-op, unclear)[162] Cognitive task load[192]		Level 2b[192] Level 3[162]	Level 3 recommendation

cent) ear, nose and throat[62], 1 (2.2 per cent) microsurgery[88], 8 (17.8 per cent) suturing[89–95], 2 (4.4 per cent) dissection[92,96], 1 (2.2 per cent) vessel dissection and ligation[97] and 1 (2.2 per cent) other dry model tasks[98].

Robotic Anastomosis Competency Evaluation (RACE), Task-Performance Metrics tools and OSATS are the most evaluated tools with a Level 2 recommendation, demonstrating concurrent and construct validity, missing evaluation for predictive and in the consequence domain. One tool, A-OSATS[85], has been evaluated over all five of Messick's domains, but only in one study and lacking predictive validity. In summary, there are no full procedural tools that have been fully evaluated for robotic surgery.

Error tools

Three main tools underwent multiple study evaluations (*Table 2*): the Fundamental Laparoscopic Skills (FLS) scoring system, Generic Error Rating Tool (GERT) and Task-Performance Metrics. The most common error method was the cumulative number of errors, reported in 20 of 42 studies (46.5 per cent; *Table S5*). In 13 studies (69.7 per cent) a composite score was created, while a further study defined arbitrary task-specific time penalties. There is substantial variability in the definition and measurement of errors, often missing robust evaluation on validity and multiple tools were only present in single studies. The FLS scoring system gives a composite score and was used in 9 (20.9 per cent) studies. Task-Performance Metrics tools define errors and were used in 5 (11.6 per cent). The GERT tool assesses a surgical task group, error mode, number, description and mechanism of event and was analysed in 2 (4.6 per cent) studies. All three methods reached Level 2 recommendation. FLS and Task-Performance Metrics tools both had evidence of internal structure and relationship to other variables, with excellent reliability, concurrent and construct validity. There were no reports on predictive validity or benchmarking of these tools.

Simulator automated performance metrics

Ten different simulators were identified (*Table S6*, *Table 3*). Automated performance metrics in simulation environments have been thoroughly evaluated with 39 (50.6 per cent) studies on da Vinci Skills Simulator (dVSS), 17 on (22.1 per cent) dV-Trainer (dV-T) and 9 (11.7 per cent) on RobotiX Mentor. Sixteen (76.2 per cent) of the 20 RCTs in this review involved simulators. These three simulators have been validated in all five Messick domains exhibiting concurrent and construct validity. dVSS and dV-T training also predicted better performance on the console in operative and dry model performances. In addition, more comprehensive evaluation in the consequence domain has been carried out for all three when compared to other assessment tools. Current evidence favours dVSS at Level 1 recommendation. dV-T, RobotiX Mentor, Promis hybrid surgical simulator, Robotic Surgery Simulator (RoSS) and 3D hydrogel models with 'Clinically Relevant Objective or Performance Metrics (CROMS/CRPMS)' all receive Level 2 recommendation. The Versius trainer from CMR Surgical has currently been evaluated at Level 4 recommendation. Simulators[99–101,283] unlikely to be in wide usage were identified and excluded from *Table 3*. Notably the vast majority of studies (70; 90.9 per cent) looked at basic skills, with only 6 (7.8 per cent) reviewing procedure-specific VR[52,84,99,102–104].

Non-simulator automated performance metrics

Of the 28 included studies (*Table 3* and *Table S7*), 16 used da Vinci Application Programming Interface (API) kinematic and system event data, with 6 (21.4 per cent) from the operating room and all within urology. Kinematic and system event data from the da

Table 3 Summary of automated performance metrics

Simulator automated performance metrics

Tool	Study type	Setting	Test content	Response process	Internal structure	Relationship to other variables	Consequences	LoE	LoR
da Vinci (Surgical) Skills Simulator	28 Observational 11 RCTs	30 Lab 9 OR and Lab	14 studies showing face and content validity[83,164,168,201,204,212–220] Note three studies stated dVSS more realistic/helpful than dV-T[83,219,220]	Automated metrics inherently no rater bias 11 RCTs (3 not powered)[159,160,165] 2 Powered non-randomized studies[201,212] Orientation/standardization of training[201,212,215,216,218,221,222]	Automated metrics inherently consistent 1 study measured consistency with consecutive attempts with excellent reliability[217]	*Concurrent validity* GRS tools[160,163–165,200,204] Other VR[219] Error tools[163] Other training method[223] *Construct validity*[163,181,201,204,212–215,221,222,224–226] *Predictive of console performance in the lab*[83,194,200,201] *Predictive of console performance in the OR*[146,166,201,227]	Pass mark defined[92,160,166,168,169,194,201,204,210,216,217,225,228–230]	Level 1b[83,200] Level 2a[159,160,165,166,168,169,177] Level 2b[146,163,164,201,204,212–217,221–226,228,229] Level 3[73,92,181,191,194,210] Level 4[220]	Level 1 recommendation
dV-Trainer	13 Observational 4 RCTs	17 Lab	9 Face/content[83,90,193,218–220,231–233] Note lack of realism for needle driving[231]	Automated metrics inherently no rater bias 4 RCTs (2 not powered)[165,234] Orientation/standardization of training[218,232,235,236]	Automated metrics inherently consistent	*Concurrent validity* GRS[90,165,218] Other VR[219] *Construct validity*[218,219,231–233,236,237] *Predictive of console performance in the lab*[193,234,238]	Novice proficiency criteria developed[149] Defined by expert performance[235] VR index competency score (not benchmarked)[234]	Level 1b[83] Level 2a[149,165,234,235] Level 2b[90,193,218,219,231–233,236,237] Level 3[238,239] Level 4[220]	Level 2 recommendation
RobotiX Mentor	9 Observational 1 RCT	9 Lab 1 Delphi	All studies showed face/content except two[173,240] No difference between RXM and dVSS/dV-T[220] Note limitation of realism of suture[241]	Automated metrics inherently no rater bias Orientation/standardization of training[102,103,173,241–243] Randomized groups[173]	Automated metrics inherently consistent *Test–retest reliability* Excellent[103] *Internal consistency of metrics* Fair[243] Unacceptable to good[143] Unacceptable to poor[103]	*Concurrent with FRS metrics dry lab*[244] *Construct validity*[103,143,240–244] *Predictive of console performance in the lab*[173]	Pass mark defined [241] based on competent surgeons[240,244], with contrasting groups method[103,43,243]	Level 2a[173] Level 2b[102,103,143,241–244] Level 2c[240] Level 4[220]	Level of recommendation 2
Promis™ hybrid surgical simulator	3 Observational	3 Lab	Face and content[245]	Standardized orientation[245,246]	*Internal consistency* Good[247]	*Construct validity*[245–247]		Level 2b[245–247]	Level 2 recommendation
Robotic Surgery Simulator (RoSS)	3 Observational 1 RCT	1 Lab 1 Lab and Delphi	Face[248] and content[12,249,250]	Automated metrics inherently no rater bias Randomized, powered[12]	Automated metrics inherently reliable	*Construct validity*[250] *Predictive of console performance in the lab*[12]		Level 2a[12] Level 2b[250] Level 4[248,249]	Level 2 recommendation
3D hydrogel models—clinically relevant objective/performance metrics (CROMS/CRPMS)	2 Observational	2 Lab	Face and content both studies	Objective metrics Pilot testing[52]		*Concurrent validity with GEARS*[52,84] *Construct validity*[52,84]		Level 2b[52,84]	Level 2 recommendation
da Vinci SimNow[251]	1 RCT	Lab		Automated metrics inherently without rater bias	Automated metrics inherently reliable	*Concurrent validity with da Vinci system event data recorder* *Construct validity*		Level 2a	Level of Recommendation 3
Versius Trainer[172]	1 Observational	Lab		Automated metrics inherently without rater bias	Automated metrics inherently reliable	*Learning curve demonstrated*		Level 3	Level of Recommendation 4
Non-Simulator Automated Performance Metrics									
da Vinci Kinematic and System Event Recorders NB different terms used that is da Vinci API, dVLogger or da Vinci Systems Events data recorder	15 Observational 1 RCT	10 Lab 6 OR		Automated metrics inherently without rater bias 1 study randomized[251]	Automated metrics inherently reliable	*Concurrent validity* GRS[89,154,157] VR[251] Cognitive load[251] Task evoked pupillary response[252] R.E.N.A.L. nephrotomy score and intraop data for example EBL[35]		Level 2a[251] Level 2b[35,37,154,157,198,252–256] Level 3[36,89,189,257,258]	Level of Recommendation 2

(continued)

Table 3 (continued)

Simulator automated performance metrics

Tool	Study type	Setting	Test content	Response process	Internal structure	Relationship to other variables	Consequences	LoE	LoR
						Construct validity[35,37,157,251–256] *Predictive validity*[35–37,256]			
Electromag-netic motion tracker sensor (TrakStar; Ascension Technologies, USA)	5 Observational	All Lab		Automated metrics inherently without rater bias Standardized orientation[259] Experts reviewed metrics[260]	Automated metrics inherently reliable	*Construct validity*[206,259,261,262]		Level 2b[206,259,261,262] Level 3[260]	Level 2 recommendation
SurgTrak™ Motion tracking[235]	1 RCT	Lab		Automated metrics inherently without rater bias Randomized, powered		*Construct validity*		Level 2a	Level 2 recommendation

Vinci systems are the most evaluated APMs on the console achieving Level 2 recommendation. Despite two other APM tools having the same LoR, da Vinci API data evaluation is arguably more useful as concurrent, construct and predictive validity has been demonstrated. The only study[105] looking at non-kinematic data, instrument vibration and forces, showed construct and concurrent validity, with a LoR 3. No study has yet fully validated non-simulator APM data, primarily missing evaluation in the consequence domain.

Artificial intelligence

Fifty-three AI studies were identified (*Table 4* and *Table S8*). The range of participating surgeons across the AI studies varied from 1[106] to 77[107] (median = 8). Most studies employed the publicly available JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS)[109], which features 139 trials from eight surgeons suturing, knot-tying, and needle-passing exercises with kinematic data.

The most common level grouping was between expert and novice surgeons; however, there was significant heterogeneity in how this was defined. In most cases, expertise was defined by a surgeon's caseload with wide variability, for example 50 to over 2500 cases[33,105]. Other studies assigned assessment scores to group surgeons above a predefined threshold as experts and below as novices[108].

Forty-one (77.4 per cent) studies evaluated their AI models using data obtained from simulators or dry lab simulations, while 12 (22.6 per cent) studies used data collected from real surgical procedures. The most frequently used dry lab data set (24/41; 58.5 per cent) was JIGSAWS.

Studies that used real surgical data included procedures such as lymph node dissection[110], laparoscopic cholecystectomy[108,111], robotic assisted radical prostatectomy (RARP) urethrovesical anastomosis[112], and phases[36,113–115], gastrectomy[116] and thyroid surgery[117]. RARP procedures were the most common (7/12), with Chen *et al.*[33] utilizing the largest data set.

The majority approached skill assessment as a classification task (28/53; 52.8 per cent), with the aim of predicting the participant skill level. Twenty (37.7 per cent) studies estimated an assessment score (numerical regression) that corresponds to an assessment tool. Notably, only three[118–120] attempted to estimate the individual domains of the tool, with the remainder predicting the total score.

A few studies adopted a different approach to assess skill; ranking performance[121], estimating the operating field clearness[116], using stylistic behaviour labels[122] and linking skill levels to clinical outcomes in RARP[36,113,115,123].

Of the 53 studies, 20 (37.7 per cent) utilized video data, 29 (54.7 per cent) used kinematics, 7 (13.2 per cent) employed system events and 3 (5.6 per cent) used force data. Furthermore, a few others utilized clinical parameters such as BMI and prostate -specific antigen (PSA)[114], eye-tracking and electroencephalography (EEG) signals, electromyography data (EMG) and galvanic skin response (GSR)[124], surgical gesture sequences[114] and stylistic behaviour components[122,125,126]. Among these studies, 33 (62.3 per cent) used a single input modality (for example, video only), while 20 (37.7 per cent) utilized two or more input modalities.

Twenty-six (49.1 per cent) studies used classic machine learning methods, with support vector machine (SVM) being the most common (13/26 (50 per cent)). Most used APMs as input. Twenty-seven (51 per cent) employed deep learning methods, with 19 (35.8 per cent) using convolutional neural networks (CNN). Video-based deep learning methods used a CNN to extract visual features, which are then either fed to a temporal model[110,111,127–131] or to a simple classifier/regressor[108,112,116,131,132]. Kinematic-based deep learning approaches use either temporal convolutional networks (TCN)[110,133,134] or recurrent neural network (RNN)[129,135] or a combination of the two[119,136,137]. Notably, deep learning approaches have gained popularity in surgical skill assessment (*Fig. S1*).

To evaluate their developed methods, most studies utilized the accuracy metric and Spearman's correlation coefficient (SCC). The accuracy rates and SCC for the models tested on real surgical data ranged from 67 per cent to 100 per cent and 0.41 to 0.64, respectively, but were inferior to simulator/dry-lab data; nearly 60 per cent of classification methods reported accuracy above 90 per cent; while only one study[111] out of 10 reported SCC over 0.90.

Discussion

This systematic review comprehensively analysed the current development and evaluation status of objective technical skills assessment tools in robotic surgery. Despite the plethora of publications, it is evident that full evaluation according to Messick's concept is sparse. This may explain the notable lack of reports showing their implementation within day-to-day practice or curricula. Many manual tools are lacking in scope and are arguably unsuitable to be used as summative tools at their current validation status. Emerging evidence in AI has reached the first in-human studies, but these are predominantly conceptual and require full validation. The current review suggests that research efforts should be focused on validating

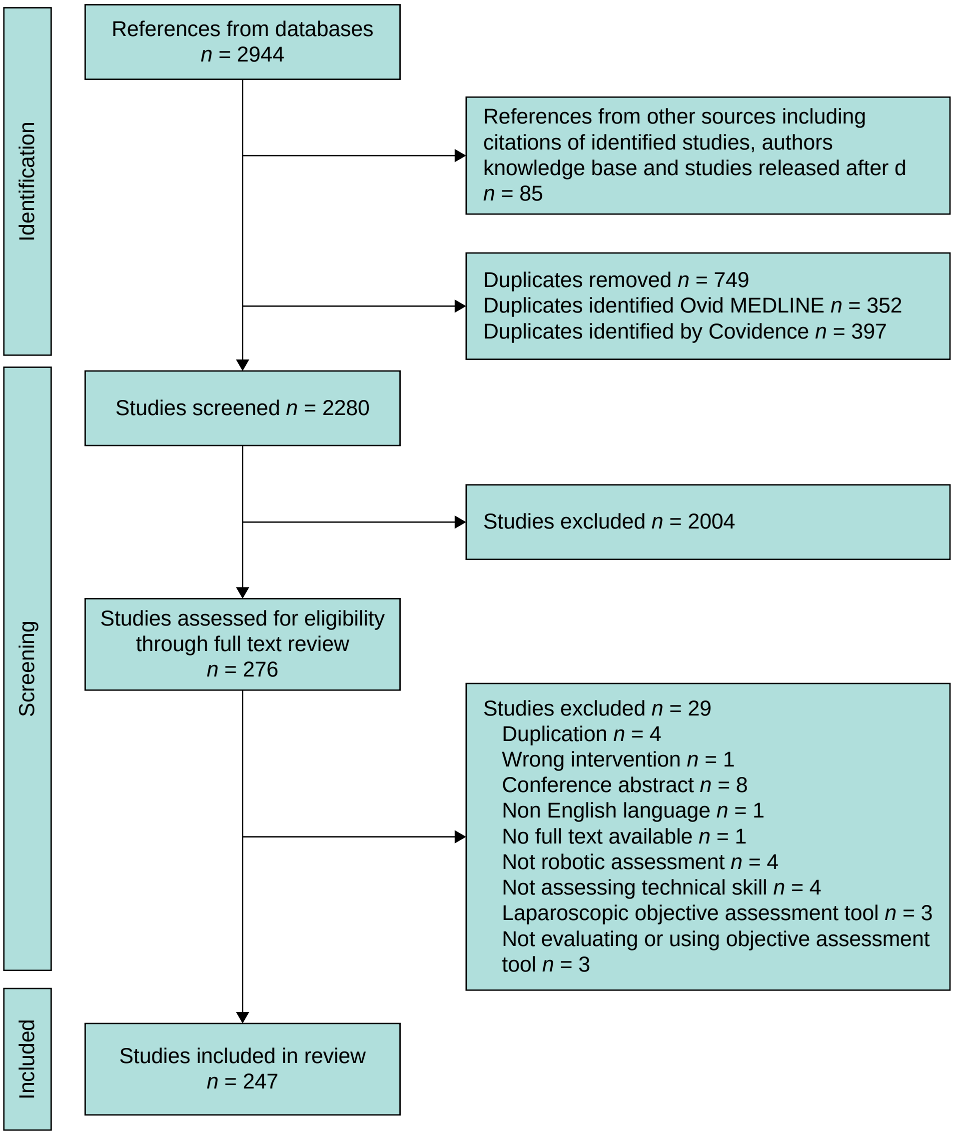

Fig. 2 PRISMA flowchart

and implementing existing instruments rather than seeking any further robotic surgery assessment methods.

GEARS and VR simulators offer clear opportunities for formative and summative assessment within the basic skills curricula. Simulator studies demonstrated VR participants outperforming controls or an improvement in post-VR curriculum assessments in the operative and laboratory setting. GEARS has not been formally benchmarked and given that it is likely the best manual annotation GRS tool to use with AI models warrants further focused evaluation. Meanwhile, given that AI studies often use OSATS or modified GOALS, efforts are necessary to inform the computer science and surgical community to utilize GEARS instead for robotic global technical ratings. Chen *et al.*[28] highlighted gaps in the assessment domains of generic robotic skills assessments for GEARS, which provides an opportunity for modification and re-evaluation. VR simulators allow safe transference of basic skills and have defined competency benchmarks before progression to console training, broadly speaking a score between 80 and 90 per cent.

Procedure-specific VR and 3D-printed hydrogel models provide high-fidelity simulation allowing an opportunity for standardized, safe progression to clinical training. These platforms avoid possible ethical, religious and moral issues that can prevent the use of cadaver or live animals. Only six studies were identified evaluating procedure-specific VR, confirming the need for further development and evaluation of different operative VR and 3D model tasks. However, additional issues including training access and the financial implications of these platforms remain unstudied.

Procedure-specific tools can potentially act as excellent formative and summative assessments often with higher reliability than GRS. Three tools (OSATS task-specific, RACE and Task-Performance Metrics) had the highest LoR; however, importantly there are no reports demonstrating predictive validity or benchmarking full procedural tools. Task-Performance Metrics were all developed through Delphi consensus as proficiency-based progression (PBP) assessment tools and had high reliability through trained expert raters undergoing reliability 'checks'. The tools' structure includes phases and subtasks for each procedure and can be commended for including operation-specific error metrics. Their intended application is within proficiency-based

Table 4 Summary of artificial intelligence studies

<table>
<tr><th>Study setting</th><th>Participant no.</th><th>Tasks/procedure</th><th>Data set</th><th>Data set size (trials)</th><th>Model input</th><th>Model output that is estimates/predictions | performance</th></tr>
<tr><td>Simulators (VR)</td><td>1–9[124]
10–19[122,125,126,263]
≥ 20[107]</td><td>Ring and rail[107,122,124,125,263]
Suture sponge[107,122,125]
Camera targeting, peg board, dots and needles, tubes[124]
Endowrist manipulation, needle control and needle driving skills[126]</td><td>All private</td><td><50[124,263]
50–99[122,125,126]</td><td>Kinematics[107,122,125,126,263]
Skill-related labels[125,126]
EMG signals[125]
Eye-tracking and EEG signals[124]</td><td>Skill level[125,263] | accuracy 65–100%
GEARS[107,126] | accuracy 69–89%
Skill-related labels[122] | accuracy 48–99%</td></tr>
<tr><td>Dry lab</td><td>1–9[118,120,121,127,129–132,134,135,136,198,264–275]
10–19[105,276–278]
≥ 20[152]</td><td>Suturing[120,121,131,132,264–266,276,277]
Needle passing[118–121,127–132,134,136,137,265,267–270,272,274,275,279]
Knot tying[118,120,121,127,129–132,134,136,137,198,265–270,272,274,277,278]
Transection and dissection[264,276]
Peg transfer[105,152]
Ring transfer[280]</td><td>JIGSAWS[110,111,118–121,127–132,134,135,137,265–269,272–275,277,279]</td><td><50[271,274]
50–99[136]
100–149[110,111,118–121,127–132,134,135,137,152,198,265–270,272–275,277–279]
≥ 150[105]</td><td>Kinematics[105,118–120,129,134–137,152,198,264–266,269,271,274–278,281]
Force data[105,264]
System event data[264,271,276,281]
Videos[121,127,128,130–132,267,268,270,272,273,279]</td><td>Skill level[118,127,134–137,198,265,266,268,269,271,274,276,277,280,281] | accuracy 46–100%
Modified OSATS (JIGSAWS)[111,118–120,127,128,131,132,267,270,272] | SCC 0.03–0.93
GEARS[152] | accuracy 52–75%</td></tr>
<tr><td>Operating room</td><td>1–9[106,110,111,115]
10–19[33,112]
≥ 20[36,114,138]</td><td>RARP[33,36,112,114,115,138]
Urethrovesical anastomosis[33,112]
Needle handling/driving[138]
Lymph node dissection[106,110]
Laparoscopic cholecystectomy[108,111]</td><td>HeiChole[111]</td><td><50[106,110–112,117]
50–99[33,114,116]
100–149[115]
≥ 150[36,108,138]</td><td>Videos[106,108,110–112,116,117,138]
Kinematics[33,36,115]
System event data[33,36,115]
Clinical parameters[114,115]</td><td>Skill level[33,108,112,117,138] | accuracy 67–100%
PLACE score[106] | accuracy 83.3%
Modified GOALS[110,111] | SCC 0.46–0.57
3-Month/6-month urinary continence after RARP[36] | AUC 0.67–0.74)
1-Year erectile function recovery[114] | AUC 0.68–0.77</td></tr>
</table>

training, which aims to benchmark phases or modules before moving on to the next stage. However, these among other tools are not yet publicly available, precluding research including external validation efforts.

It is evident that there is a paucity of procedure-specific tools ready for implementation into robotic training curricula. They also lack scope, with the majority in urology, and so as a surgical community it is imperative to both develop tools missing for key operations and fully evaluate existing ones.

Error tools identified in this review typically used cumulative number of errors and have not been fully evaluated within clinical settings. A key aspect in a surgeon's learning curve is to understand the 'what, where, when, how, why' and corrective mechanisms of an error, which no current study has reported. Granular methods of surgeons' technical performance and errors are necessary to train AI, combined with global rating scales and procedure-specific tools, to fully understand the complexities of any operation. Tools should combine each aspect with full comprehensive evaluation before implementation into training curricula. As demonstrated in this review, reliability can most likely be improved through expert, trained raters and quality assurance processes.

APMs and AI are emerging and promising tools to guide training and assessment in robotic surgery. APMs can be considered truly objective, yet need further focused evaluation to understand and benchmark important metrics for construct and predictive validity. While AI models performed well when analysing intraoperative surgical skill data, they generally perform better on simulator/dry lab.

A significant proportion of the AI models tested on simulated data achieved accuracy rates above 90 per cent, while some models tested on real surgical data demonstrated perfect classification performance of surgical skill levels. Despite this, AI-based skill assessment is still in its conceptual stage with four broad areas that need to be addressed: data sets, manual annotation, AI model evaluation and integration into clinical practice.

For the field of automated surgical skill assessment to advance, it is critical to assess models on real surgical data. Additionally, it is crucial to gather data from high-fidelity simulations tasks so AI models can be evaluated for benchmarking and comparison of different methods. Efforts must focus on collecting large, publicly available, diverse data sets, including surgeons with differing levels of expertise and different robotic platforms with matched clinical outcome data. Utilizing diverse data sets will ensure AI models are unbiased and can generalize effectively on unseen surgeons and tasks.

Identified AI studies used different ways to evaluate their methods making direct comparisons challenging and reducing external validity. Testing models on the JIGSAWS data set has highlighted the performance gap between cross-validation schemes such as Leave-One-User-Out (LOUO) and more relaxed schemes such as Leave-One-Super-Trial-Out (LOSO). However, before automated skill assessment can be used in clinical practice it must first be ensured the models can generalize to unseen surgeons. To achieve this, evaluation should be performed with cross-validation schemes (for small data sets), or with large external test sets containing trials from unseen surgeons from different hospitals to ensure generalizability[138–140]. LOSO still remains useful in situations where the performance of a specific surgeon is tracked for proficiency curve analyses.

To achieve integration into clinical practice, it is essential that models can provide not only accurate predictions but also clear, understandable justifications for their decisions that clinicians

Table 5 Summary of all assessment domains

Assessment domain	Advantages	Disadvantages	Further research
Global Rating Scales	Quick to fill in. GEARS is capable of strong validity and reliability. Likely good formative tools for generic technical skill.	Subjective, risk of low reliability. Miss granularity of operative steps; therefore, not to be used solely for summative assessment in procedural training. No evidence used in day-to-day formative/summative assessment or incorporated into curricula.	Benchmarking alongside procedure-specific tools. Incorporation into curricula. Expert consensus on tool to use for training AI, and a standardized method for example experts who are trained with reliability tests.
Procedure-specific	Valid and reliable often as it is easy to agree on what steps have been done that is binary. Useful formative and summative assessments.	Potential to miss how well the surgeon performs an operation. Currently only a few tools and predominantly within urology. No fully evaluated or benchmarked tools exist for robotics.	Benchmarking with GRS tools. Development and full evaluation of more tools within different specialties. Incorporation into training curricula.
Error methods	Evidence of validity and reliability. Important part of formative assessment for the surgeon to improve and summative to indicate competency.	Difficult to define certain aspects for example how and why it occurs. Detailed error analysis takes time consuming retrospective video analysis.	Combine with GRS and procedure-specific operative analysis including benchmarking. Further evaluation of existing tools within MIS. AI recognition of near-misses or pre-errors, errors, and critical errors, through manual annotation.
Automated performance metrics	Promising objective tools which have some validity evidence. No concerns re: reliability or bias. Simulators have been thoroughly evaluated and are a good tool for basic skills transfer to the console.	Simulator APMs for procedure-specific simulation are lacking development and evaluation. Non-simulator APMs—current understanding is limited as to their meaning for formative/summative assessment and patient outcome. Data are protected by industry. Requires expertise in computer science collaborating with industry and clinicians; therefore, used only in research currently.	Simulator evaluation on procedure-specific tasks. Evaluation of emerging robotic platforms VR and APMs. Evaluation of APM data in the clinical setting. Creation of more open data sets or perhaps registries should be considered. Availability of APM data sets depends on collaborating with industry to ensure it can be publicly available.
Artificial intelligence	Promising initial results, with the potential to transform formative and summative assessment, particularly if evaluated with APMs.	In its infancy. Ensuring AI is correctly assessing performance requires highly reliable, manually annotated videos, which is time-consuming, particularly given the numbers needed to train then test. Current results are from small data sets. More challenging to evaluate in the clinical setting with more variability, for example camera and patient movements. External validity of methods to surgeons outside of the research data set. Risk of blocking innovation in the future?	Open data sets/registries. Evaluation with manually annotated/rated videos to help train AI, alongside APMs. Development of more complex laboratory data sets to initially evaluate models, then transfer into multispecialty clinical setting.

can trust. As the role of AI in healthcare continues to expand there is increasing awareness of the potential pitfalls and the need for guidance to avoid them[141], including a recent statement from the World Health Organization[142].

Increasingly detailed and informative feedback beyond simple scores or skill level labels can help to personalize surgical training. Although there have been some efforts to develop explainable AI models and feedback mechanisms[111,127,128,133], more research is needed to fully address these issues, focusing on developing methods that are more transparent and interpretable, for example written reports and error-detection capabilities to provide more informative context-specific feedback. Indeed, research is needed investigating human factors with educational specialists to elucidate the best way for skill assessment to be presented and when.

To credential surgeons as competent for independent practice, blinded expert video rating is considered an essential part of accreditation[21]. This requires fully evaluated objective summative assessment tools. Often, surgeons undergoing robotic training are already credentialed, adding additional challenges to standardizing pathways and ensuring patient safety. Undoubtedly, there are many routes to competency and now also emergent robotic systems to consider.

This review has highlighted many assessment domains, with their advantages, disadvantages and future research needs (Table 5). To achieve implementation of validated and reliable tools into curricula, collaboration between surgical societies is required. Through expert consensus and large, multicentre, international studies, single tools for each procedure should be developed and fully evaluated. Only then, should they be implemented within curricula as formative and summative tools, or in the evaluation of APMs and AI.

Limitations

This comprehensive review standardized data extraction with Messick's concept and modified OCEBM guidelines. Nevertheless, due to marked study heterogeneity this was difficult at times, and was particularly evident when utilizing the OCEBM guidance, with previous systematic reviews disagreeing on studies' LoE. Not only this, but some studies have a higher LoE, despite demonstrating less validity evidence than others. It is likely that guidelines require updating as surgical data science evolves. The application of methodological quality tools was found to be impractical for assessing AI studies, primarily as most are in their conceptual stage of development. Future research should focus on developing and piloting a new AI-specific study quality assessment tool.

Conclusion

A large number of manual, automated and artificial intelligence tools in robotic surgery exist. There is huge variability in approach to assessment and the level of evaluation among all domains of robotic technical skill assessment, with few having been well validated. In addition, there is a lack of scope and most tools are presently only used within the research setting, despite the unmet need for both objective formative and summative tools to inform learning and accreditation, respectively. Collaboration between surgical societies, AI scientists and industry, with large high-quality studies and open data sets, appears the most efficient way forward to aid diffusion and implementation of objective assessment tools in clinical practice to enhance training and patient safety.

Author contributions

Matthew Boal (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing—original draft, Writing—review & editing), Dimitrios Anastasiou (Data curation, Formal analysis, Investigation, Methodology, Writing—original draft), Freweini Tesfai (Formal analysis, Investigation, Writing—review & editing), Walaa Ghamrawi (Investigation, Writing—review & editing), Evangelos Mazomenos (Formal analysis, Investigation, Methodology, Writing—review & editing), Nathan Curtis (Conceptualization, Writing—review & editing), Justin Collins (Conceptualization, Formal analysis, Supervision, Writing—review & editing), Ashwin Sridhar (Conceptualization, Supervision, Writing—review & editing), John Kelly (Conceptualization, Supervision, Writing—review & editing), Danail Stoyanov (Conceptualization, Formal analysis, Supervision, Writing—review & editing), and Nader Francis (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Supervision, Writing—review & editing)

Funding

The authors have no funding to declare.

Acknowledgements

This systematic review was registered with PROSPERO: registration ID CRD42022304901.

Disclosure

Justin W. Collins is an associate medical director at CMR Surgical Ltd, UK. The other authors declare no conflict of interest.

Supplementary material

Supplementary material is available at *BJS* online.

Data availability

Data for this review can be reproduced on reasonable request to the corresponding author.

References

1. Khajeh E, Aminizadeh E, Dooghaie Moghadam A, Nikbakhsh R, Goncalves G, Carvalho C *et al.* Outcomes of robot-assisted surgery in rectal cancer compared with open and laparoscopic surgery. *Cancers (Basel)* 2023;**15**:839
2. Hopkins MB, Geiger TM, Bethurum AJ, Ford MM, Muldoon RL, Beck DE *et al.* Comparing pathologic outcomes for robotic versus laparoscopic surgery in rectal cancer resection: a propensity adjusted analysis of 7616 patients. *Surg Endosc* 2020;**34**:2613–2622
3. Markar SR, Karthikesalingam AP, Venkat-Ramen V, Kinross J, Ziprin P. Robotic vs. laparoscopic Roux-en-Y gastric bypass in morbidly obese patients: systematic review and pooled analysis. *Int J Med Robot* 2011;**7**:393–400
4. Safiejko K, Tarkowski R, Koselak M, Juchimiuk M, Tarasik A, Pruc M *et al.* Robotic-assisted vs. standard laparoscopic surgery for rectal cancer resection: a systematic review and meta-analysis of 19,731 patients. *Cancers (Basel)* 2022;**14**:1036
5. Kamarajah SK, Bundred J, Saint MO, Jiao LR, Manas D, Abu Hilal M *et al.* Robotic versus conventional laparoscopic pancreaticoduodenectomy: a systematic review and meta-analysis. *Eur J Surg Oncol* 2020;**46**:6–14
6. Curtis NJ, Dennison G, Brown CSB, Hewett PJ, Hanna GB, Stevenson ARL *et al.* Clinical evaluation of intraoperative near misses in laparoscopic rectal cancer surgery. *Ann Surg* 2021;**273**:778–784
7. Collins JW, Dell'Oglio P, Hung AJ, Brook NR. The importance of technical and non-technical skills in robotic surgery training. *Eur Urol Focus* 2018;**4**:674–676
8. Vincent C, Neale G, Woloshynowych M. Adverse events in British hospitals: preliminary retrospective record review. *BMJ* 2001;**322**:517–519
9. Collins JW, Levy J, Stefanidis D, Gallagher A, Coleman M, CecilT *et al.* Utilising the Delphi process to develop a proficiency-based progression train-the-trainer course for robotic surgery training. *Eur Urol* 2019;**75**:775–785
10. ECRI. *Top 10 Health Technology Hazards for 2015*. 2014. https://www.ecri.org/Resources/Whitepapers_and_reports/Top_Ten_Technology_Hazards_2015.pdf (accessed 21 August 2023)
11. Raza SJ, Froghi S, Chowriappa A, Ahmed K, Field E, Stegemann AP *et al.* Construct validation of the key components of fundamental skills of robotic surgery (FSRS) curriculum—a multi-institution prospective study. *J Surg Educ* 2014;**71**:316–324
12. Stegemann AP, Ahmed K, Syed JR, Rehman S, Ghani K, Autorino R *et al.* Fundamental skills of robotic surgery: a multi-institutional randomized controlled trial for validation of a simulation-based curriculum. *Urology* 2013;**81**:767–774
13. Satava R, Smith R, Patel V, Advincula A, Aggarwal R, al Ansari A *et al.* Fundamentals of robotic surgery: outcomes measures

and curriculum development principle investigators. *Soc Laproendosc Surg* 2012
14. Goh AC, Aghazadeh MA, Mercado MA, Hung AJ, Pan MM, Desai MM *et al.* Multi-institutional validation of fundamental inanimate robotic skills tasks. *J Urol* 2015;**194**:1751–1756
15. Satava RM, Stefanidis D, Levy JS, Smith R, Martin JR, Monfared S *et al.* Proving the effectiveness of the fundamentals of robotic surgery (FRS) skills curriculum: a single-blinded, multispecialty, multi-institutional randomized control trial. *Ann Surg* 2020;**272**: 384–392
16. Schmiederer I S, Torices-Dardon A, Ferrari-Light D M, Charbel Abboud E, Villani V, Lau J N *et al.* Developing a robotic general surgery training curriculum: identifying key elements through a Delphi process. *J Surg Educ* 2021;**78**:e129–e136
17. Smith R, Patel V, Satava R. Fundamentals of robotic surgery: a course of basic robotic surgery skills based upon a 14-society consensus template of outcomes measures and curriculum development. *Int J Med Robot* 2014;**10**:379–384
18. Challacombe BM, Ahmed K, Soomro N, Dasgupta P, Shamim Khan M, Cross W *et al. British Association of Urological Surgeons (BAUS) Robotic Surgery Curriculum- Guidelines for Training. [cited 2021 Sep 29].* https://www.baus.org.uk/professionals/baus_business/publications/83/robotic_surgery_curriculum/ (accessed 29 September 2022)
19. Veronesi G, Dorn P, Dunning J, Cardillo G, Schmid RA, Collins J *et al.* Outcomes from the Delphi process of the Thoracic Robotic Curriculum Development Committee. *Eur J Cardiothorac Surg* 2018;**53**:1173–1179
20. Szold A, Bergamaschi R, Broeders I, Dankelman J, Forgione A, Langø T *et al.* European Association of Endoscopic Surgeons (EAES) consensus statement on the use of robotics in general surgery. *Surg Endosc* 2015;**29**:253–288
21. Vanlander AE, Mazzone E, Collins JW, Mottrie AM, Rogiers XM, van der Poel HG *et al.* Orsi Consensus Meeting on European Robotic Training (OCERT): results from the first multispecialty consensus meeting on training in robot-assisted surgery. *Eur Urol* 2020;**78**:713–716
22. Ruiz M G, Alfieri S, Becker T, Bergmann M, Boggi U, Collins J *et al.* Expert consensus on a train-the-trainer curriculum for robotic colorectal surgery. *Colorectal Dis* 2019;**21**:903–908
23. Palagonia E, Mazzone E, De Naeyer G, D'Hondt F, Collins J, Wisz P *et al.* The safety of urologic robotic surgery depends on the skills of the surgeon. *World J Urol* 2020;**38**:1373–1383
24. Stefanidis D, Huffman EM, Collins JW, Martino MA, Satava RM, Levy JS. Expert consensus recommendations for robotic surgery credentialing. *Ann Surg* 2022;**276**:88–93
25. Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR *et al.* Surgical skill and complication rates after bariatric surgery. *N Engl J Med* 2013;**369**:1434–1442
26. Hanna GB, Mackenzie H, Miskovic D, Ni M, Wyles S, Aylin P *et al.* Laparoscopic colorectal surgery outcomes improved after national training program (LAPCO) for specialists in England. *Ann Surg* 2020;**275**:1149–1155
27. Curtis NJ, Foster JD, Miskovic D, Brown CSB, Hewett PJ, Abbott S *et al.* Association of surgical skill assessment with clinical outcomes in cancer surgery. *JAMA Surg* 2020;**155**: 590–598
28. Chen J, Cheng N, Cacciamani G, Oh P, Lin-Brande M, Remulla D *et al.* Objective assessment of robotic surgical technical skill: a systematic review. *J Urol* 2019;**201**:461–469
29. Vaidya A, Aydin A, Ridgley J, Raison N, Dasgupta P, Ahmed K. Current status of technical skills assessment tools in surgery: a systematic review. *J Surg Res* 2020;**246**:342–378
30. Levin M, McKechnie T, Khalid S, Grantcharov TP, Goldenberg M. Automated methods of technical skill assessment in surgery: a systematic review. *J Surg Educ* 2019;**76**:1629–1639
31. Lam K, Chen J, Wang Z, Iqbal FM, Darzi A, Lo B *et al.* Machine learning for technical skill assessment in surgery: a systematic review. *NPJ Digit Med* 2022;**5**:24
32. Kutana S, Bitner DP, Addison P, Chung PJ, Talamini MA, Filicori F. Objective assessment of robotic surgical skills: review of literature and future directions. *Surg Endosc* 2021;**36**:3698–3707
33. Chen AB, Liang S, Nguyen JH, Liu Y, Hung AJ. Machine learning analyses of automated performance metrics during granular sub-stitch phases predict surgeon experience HHS public access. *Surgery* 2021;**169**:1245–1249
34. Hung AJ, Chen J, Gill IS. Automated performance metrics and machine learning algorithms to measure surgeon performance and anticipate clinical outcomes in robotic surgery. *JAMA Surg* 2018;**153**:770–771
35. Ghodoussipour S, Reddy SS, Ma R, Huang D, Nguyen J, Hung AJ. An objective assessment of performance during robotic partial nephrectomy: validation and correlation of automated performance metrics with intraoperative outcomes. *J Urol* 2021;**205**:1294–1302
36. Hung AJ, Ma R, Cen S, Nguyen JH, Lei X, Wagner C. Surgeon automated performance metrics as predictors of early urinary continence recovery after robotic radical prostatectomy—a prospective bi-institutional study. *Eur Urol Open Sci* 2021;**27**: 65–72
37. Chen J, Chu T, Ghodoussipour S, Bowman S, Patel H, King K *et al.* Effect of surgeon experience and bony pelvic dimensions on surgical performance and patient outcomes in robot-assisted radical prostatectomy. *BJU Int* 2019;**124**:828–835
38. Kumar A, Smith R, Patel VR. Current status of robotic simulators in acquisition of robotic surgical skills. *Curr Opin Urol* 2015;**25**:168–174
39. Moglia A, Ferrari V, Morelli L, Ferrari M, Mosca F, Cuschieri A. A systematic review of virtual reality simulators for robot-assisted surgery. *Eur Urol* 2016;**69**:1065–1080
40. Julian D, Tanaka A, Mattingly P, Truong M, Perez M, Smith R. A comparative analysis and guide to virtual reality robotic surgical simulators. *Int J Med Robotics Comput Assist Surg* 2018; **14**:e1874
41. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;**339**:b2535
42. Goldenberg MG, Lee JY, Kwong JCC, Grantcharov TP, Costello A. Implementing assessments of robot-assisted technical skill in urological education: a systematic review and synthesis of the validity evidence. *BJU Int* 2018;**122**:501–519
43. Messick S. Foundations of validity: meaning and consequences in psychological assessment. *ETS Res Rep Ser* 1993;**1993**:i-18
44. Carter FJ, Schijven MP, Aggarwal R, Grantcharov T, Francis NK, Hanna GB *et al.* Consensus guidelines for validation of virtual reality surgical simulators. *Surg Endosc* 2005;**19**: 1523–1532
45. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health* 1998;**52**:377–384
46. Cook DA, Reed DA. Appraising the quality of medical education research methods: the medical education research study quality instrument and the Newcastle–Ottawa scale-education. *Acad Med* 2015;**90**:1067–1076

47. Holst D, Kowalewski TM, White LW, Brand TC, Harper JD, Sorenson MD *et al.* Crowd-sourced assessment of technical skills: an adjunct to urology resident surgical simulation training. *J Endourol* 2015;**29**:604–609
48. Holst D, Kowalewski TM, White LW, Brand TC, Harper JD, Sorensen MD *et al.* Crowd-sourced assessment of technical skills: differentiating animate surgical skill through the wisdom of crowds. *J Endourol* 2015;**29**:1183–1188
49. Powers MK, Boonjindasup A, Pinsky M, Dorsey P, Maddox M, Su LM *et al.* Crowdsourcing assessment of surgeon dissection of renal artery and vein during robotic partial nephrectomy: a novel approach for quantitative assessment of surgical performance. *J Endourol* 2016;**30**:447–452
50. Ghani KR, Miller DC, Linsell S, Brachulis A, Lane B, Sarle R *et al.* Measuring to improve: peer and crowd-sourced assessments of technical skill with robot-assisted radical prostatectomy. *Eur Urol* 2016;**69**:547–550
51. Vernez SL, Huynh V, Osann K, Okhunov Z, Landman J, Clayman RV. C-SATS: assessing surgical skills among urology residency applicants. *J Endourol* 2017;**31**:S95–100
52. Ghazi A, Melnyk R, Hung AJ, Collins J, Ertefaie A, Saba P *et al.* Multi-institutional validation of a perfused robot-assisted partial nephrectomy procedural simulation platform utilizing clinically relevant objective metrics of simulators (CROMS). *BJU Int* 2021;**127**:645–653
53. Ghani KR, Comstock B, Miller DC, Dunn RL, Kim T, Linsell S *et al.* PNFBA-02 technical skill assessment of surgeons performing robotic-assisted radical prostatectomy: relationship between crowdsourced review and patient outcomes. *J Urol* 2017;**197**:e609
54. Tunitsky E, Murphy A, Barber MD, Simmons M, Jelovsek JE. Development and validation of a ureteral anastomosis simulation model for surgical training. *Female Pelvic Med Reconstr Surg* 2013;**19**:346–351
55. Siddiqui NY, Tarr ME, Geller EJ, Advincula AP, Galloway ML, Green IC *et al.* Establishing benchmarks for minimum competence with dry lab robotic surgery drills. *J Minim Invasive Gynecol* 2016;**23**:633–638
56. Hussein AA, Sexton KJ, May PR, Meng M V, Hosseini A, Eun DD *et al.* Development and validation of surgical training tool: cystectomy assessment and surgical evaluation (CASE) for robot-assisted radical cystectomy for men. *Surg Endosc* 2018; **32**:4458–4464
57. Hussein AA, Abaza R, Rogers C, Boris R, Porter J, Allaf M *et al.*PD07-09 Development and validation of an objective scoring tool for minimally invasive partial nephrectomy: scoring for partial nephrectomy (SPaN). *J Urol* 2018;**199**: e159–e160
58. Stefanidis D, Anderson-Montoya B, Higgins RV, Pimentel ME, Rowland P, Scarborough MO *et al.* Developing a coaching mechanism for practicing surgeons. *Surgery* 2016;**160**:536–545
59. Petz W, Spinoglio G, Choi GS, Parvaiz A, Santiago C, Marecik S *et al.* Structured training and competence assessment in colorectal robotic surgery. Results of a consensus experts round table. *Int J Med Robot* 2016;**12**:634–641
60. Panteleimonitis S, Popeskou S, Aradaib M, Harper M, Ahmed J, Ahmad M *et al.* Implementation of robotic rectal surgery training programme: importance of standardisation and structured training. *Langenbecks Arch Surg* 2018;**403**:749–760
61. Eddahchouri Y, van Workum F, van den Wildenberg FJH, van Berge Henegouwen MI, Polat F, van Goor H *et al.* European consensus on essential steps of minimally invasive Ivor Lewis and McKeown esophagectomy through Delphi methodology. *Surg Endosc* 2022;**36**:446–460
62. Sobel RH, Blanco R, Ha PK, Califano JA, Kumar R, Richmon JD. Implementation of a comprehensive competency-based transoral robotic surgery training curriculum with *ex vivo* dissection models. *Head Neck* 2016;**38**:1553–1563
63. Willuth E, Hardon SF, Lang F, Haney CM, Felinska EA, Kowalewski KF *et al.* Robotic-assisted cholecystectomy is superior to laparoscopic cholecystectomy in the initial training for surgical novices in an *ex vivo* porcine model: a randomized crossover study. *Surg Endosc* 2022;**36**:1064–1079
64. Ghani K, Guru K, Aly A, Lane B, Sarle R, Linsell S *et al.* MP20-14 Variation in technical skill of surgeons performing robot-assisted prostatectomy. *J Urol* 2016;**195**:e218
65. Frederick PJ, Szender JB, Hussein AA, Kesterson JP, Shelton JA, Anderson TL *et al.* Surgical competency for robot-assisted hysterectomy: development and validation of a robotic hysterectomy assessment score (RHAS). *J Minim Invasive Gynecol* 2017;**24**:55–61
66. Hussein AA, Ghani KR, Peabody J, Sarle R, Abaza R, Eun D *et al.* Development and validation of an objective scoring tool for robot-assisted radical prostatectomy: prostatectomy assessment and competency evaluation. *J Urol* 2017;**197**: 1237–1244
67. Beulens AJW, Brinkman WM, Van Der Poel HG, Vis AN, Van Basten JP, Meijer RP *et al.* Linking surgical skills to postoperative outcomes: a Delphi study on the robot-assisted radical prostatectomy. *J Robot Surg* 2019;**13**:675–687
68. Lovegrove C, Bruce E, Raison N, Challacombe B, Novara G, Mottrie A *et al.* MP51-16 Development and content validation of a training and assessment tool for RAPN. *J Urol* 2017;**197**:e700
69. Lovegrove C, Ahmed K, Novara G, Guru K, Mottrie A, Challacombe B *et al.* Modular training for robot-assisted radical prostatectomy: where to begin? *J Surg Educ* 2017;**74**:486–494
70. Lovegrove C, Novara G, Mottrie A, Guru KA, Brown M, Challacombe B *et al.* Structured and modular training pathway for robot-assisted radical prostatectomy (RARP): validation of the RARP assessment score and learning curve assessment. *Eur Urol* 2016;**69**:526–535
71. Chow AK, Wong R, Monda S, Bhatt R, Sands KG, Vetter J *et al. Ex vivo* porcine model for robot-assisted partial nephrectomy simulation at a high-volume tertiary center: resident perception and validation assessment using the global evaluative assessment of robotic skills tool. *J Endourol* 2021;**35**:878–884
72. Davis JW, Kamat A, Munsell M, Pettaway C, Pisters L, Matin S. Initial experience of teaching robot-assisted radical prostatectomy to surgeons-in-training: can training be evaluated and standardized? *BJU Int* 2010;**105**:1148–1154
73. Volpe A, Ahmed K, Dasgupta P, Ficarra V, Novara G, Van Der Poel H *et al.* Pilot validation study of the European Association of Urology Robotic Training Curriculum. *Eur Urol* 2015;**68**:292–299
74. Iqbal U, Jing Z, Ahmed Y, Elsayed AS, Rogers C, Boris R *et al.* Development and validation of an objective scoring tool for robot-assisted partial nephrectomy: scoring for partial nephrectomy. *J Endourol* 2022;**36**:647–653
75. Tou S, Gómez Ruiz M, Gallagher AG, Matzel KE, Amin S, Bianchi P *et al.* European expert consensus on a structured approach to training robotic-assisted low anterior resection using performance metrics. *Colorect Dis* 2020;**22**:2232–2242
76. Mottrie A, Mazzone E, Wiklund P, Graefen M, Collins JW, De Groote R *et al.* Objective assessment of intraoperative skills for robot-assisted radical prostatectomy (RARP): results from the ERUS Scientific and Educational Working Groups Metrics Initiative. *BJU Int* 2021;**128**:103–111

77. Gómez Ruiz M, Tou S, Gallagher AG, Cagigas Fernández C, Cristobal Poch L, Matzel KE. Intraoperative robotic-assisted low anterior rectal resection performance assessment using procedure-specific binary metrics and a global rating scale. *BJS Open* 2022;**6**:zrac041
78. Khan H, Kozlowski JD, Hussein AA, Sharif M, Ahmed Y, May P *et al.* Use of robotic anastomosis competency evaluation (RACE) tool for assessment of surgical competency during urethrovesical anastomosis. *Canad Urol Assoc J* 2019;**13**:E10–E16
79. Hussein AA, Hinata N, Dibaj S, May PR, Kozlowski JD, Abol-Enein H *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**:879–884
80. Hung AJ, Bottyan T, Clifford TG, Serang S, Nakhoda ZK, Shah SH *et al.* Structured learning for robotic surgery utilizing a proficiency score: a pilot study. *World J Urol* 2017;**35**:27–34
81. Raza SJ, Field E, Jay C, Eun D, Fumo M, Hu JC *et al.* Surgical competency for urethrovesical anastomosis during robot-assisted radical prostatectomy: development and validation of the robotic anastomosis competency evaluation. *Urology* 2015;**85**:27–32
82. Chowriappa A, Raza SJ, Fazili A, Field E, Malito C, Samarasekera D *et al.* Augmented-reality-based skills training for robot-assisted urethrovesical anastomosis: a multi-institutional randomised controlled trial. *BJU Int* 2015;**115**:336–345
83. Hoogenes J, Wong N, Al-Harbi B, Kim KS, Vij S, Bolognone E *et al.* A randomized comparison of 2 robotic virtual reality simulators and evaluation of trainees' skills transfer to a simulated robotic urethrovesical anastomosis task. *Urology* 2018;**111**:110–115
84. Witthaus MW, Farooq S, Melnyk R, Campbell T, Saba P, Mathews E *et al.* Incorporation and validation of clinically relevant performance metrics of simulation (CRPMS) into a novel full-immersion simulation platform for nerve-sparing robot-assisted radical prostatectomy (NS-RARP) utilizing three-dimensional printing and hydrogel casting technology professional innovation introduction. *BJU Int* 2020;**125**:322–332
85. Schmidt MW, Haney CM, Kowalewski KF, Bintintan V V, Abu Hilal M, Arezzo A *et al.* Development and validity evidence of an objective structured assessment of technical skills score for minimally invasive linear-stapled, hand-sewn intestinal anastomoses: the A-OSATS score. *Surg Endosc* 2022;**36**:4529–4541
86. Hogg ME, Zenati M, Novak S, Chen Y, Jun Y, Steve J *et al.* Grading of surgeon technical performance predicts postoperative pancreatic fistula for pancreaticoduodenectomy independent of patient-related variables. *Ann Surg* 2016;**264**:482–489
87. Moloney K, Janda M, Frumovitz M, Leitao M, Abu-Rustum NR, Rossi E *et al.* Development of a surgical competency assessment tool for sentinel lymph node dissection by minimally invasive surgery for endometrial cancer. *Int J Gynecol Cancer* 2021;**31**:647–655
88. Willems JIP, Shin AM, Shin DM, Bishop AT, Shin AY. A comparison of robotically assisted microsurgery versus manual microsurgery in challenging situations. *Plast Reconstr Surg* 2016;**137**:1317–1324
89. Suh I, Mukherjee M, Oleynikov D, Siu KC. Training program for fundamental surgical skill in robotic laparoscopic surgery. *Int J Med Robot* 2011;**7**:327–333
90. Egi H, Hattori M, Tokunaga M, Suzuki T, Kawaguchi K, Sawada H *et al.* Face, content and concurrent validity of the Mimic® dV-trainer for robot-assisted endoscopic surgery: a prospective study. *Eur Surg Res* 2013;**50**:292–300
91. Møller SG, Dohrn N, Brisling SK, Larsen JCR, Klein M. Laparoscopic *versus* robotic-assisted suturing performance among novice surgeons: a blinded, cross-over study. *Surg Laparosc Endosc Percutan Tech* 2020;**30**:117–122
92. Vaccaro CM, Crisp CC, Fellner AN, Jackson C, Kleeman SD, Pavelka J. Robotic virtual reality simulation plus standard robotic orientation *versus* standard robotic orientation alone: a randomized controlled trial. *Female Pelvic Med Reconstr Surg* 2013;**19**:266–270
93. 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* 2021;**35**:4285–4294
94. Chang L, Satava RM, Pellegrini CA, Sinanan MN. Robotic surgery: identifying the learning curve through objective measurement of skill. *Surg Endosc* 2003;**17**:1744–1748
95. Singh H, Modi HN, Ranjan S, Dilley JWR, Airantzis D, Yang GZ *et al.* Robotic surgery improves technical performance and enhances prefrontal activation during high temporal demand. *Ann Biomed Eng* 2018;**46**:1621–1636
96. Vanstrum EB, Ma R, Maya-Silva J, Sanford D, Nguyen JH, Lei X *et al.* Development and validation of an objective scoring tool to evaluate surgical dissection: dissection assessment for robotic technique (DART). *Urol Pract* 2021;**8**:596–604
97. Puliatti S, Amato M, Mazzone E, Rosiello G, De Groote R, Piazza P *et al.* Development and validation of the metric-based assessment of a robotic vessel dissection, vessel loop positioning, clip applying and bipolar coagulation task on an avian model. *J Robot Surg* 2022;**16**:677–685
98. Menhadji A, Abdelshehid C, Osann K, Alipanah R, Lusch A, Graversen J *et al.* Tracking and assessment of technical skills acquisition among urology residents for open, laparoscopic, and robotic skills over 4 years: is there a trend? *J Endourol* 2013;**27**:783–788
99. Hung AJ, Shah SH, Dalag L, Shin D, Gill IS. Development and validation of a novel robotic procedure specific simulation platform: partial nephrectomy. *J Urol* 2015;**194**:520–526
100. Balasundaram I, Aggarwal R, Darzi A. Short-phase training on a virtual reality simulator improves technical performance in tele-robotic surgery. *Int J Med Robot* 2008;**4**:139–145
101. Van Der Meijden O, Schijven M, Broeders I, Der Meijden V. The SEP 'Robot'™: a valid virtual reality robotic simulator for the da Vinci surgical system? *Surg Tech Int* 2010;**19**:51–58
102. Ebbing J, Wiklund PN, Akre O, Carlsson S, Olsson MJ, Höijer J *et al.* Development and validation of non-guided bladder-neck and neurovascular-bundle dissection modules of the RobotiX-Mentor® full-procedure robotic-assisted radical prostatectomy virtual reality simulation. *Int J Med Robot* 2021;**17**:e2195
103. Olsen RG, Bjerrum F, Konge L, Jepsen JV, Azawi NH, Bube SH. Validation of a novel simulation-based test in robot-assisted radical prostatectomy. *J Endourol* 2021;**35**:1265–1272
104. Turner TB, Kim KH. Mapping the robotic hysterectomy learning curve and re-establishing surgical training metrics. *J Gynecol Oncol* 2021;**32**:e58
105. Gomez ED, Aggarwal R, McMahan W, Bark K, Kuchenbecker KJ. Objective assessment of robotic surgical skill using instrument contact vibrations. *Surg Endosc* 2016;**30**:1419–1431
106. Baghdadi A, Hussein AA, Ahmed Y, Cavuoto LA, Guru KA. A computer vision technique for automated assessment of surgical performance using surgeons' console-feed videos. *Int J Comput Assist Radiol Surg* 2019;**14**:697–707

107. Dubin AK, Julian D, Tanaka A, Mattingly P, Smith R. A model for predicting the GEARS score from virtual reality surgical simulator metrics. *Surg Endosc* 2018;**32**:3576–3581
108. Lavanchy JL, Zindel J, Kirtac K, Twick I, Hosgor E, Candinas D *et al.* Automation of surgical skill assessment using a three-stage machine learning algorithm. *Sci Rep* 2021;**11**:5197
109. Ahmidi N, Tao L, Sefati S, Gao Y, Lea C, Haro BB *et al.* A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Trans Biomed Eng* 2017; **64**:2025–2041
110. Liu D, Li Q, Jiang T, Wang Y, Miao R, Shan F *et al.* Towards unified surgical skill assessment. In: *Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation, 2021, 9522–9531
111. Li Z, Gu L, Wang W, Nakamura R, Sato Y. Surgical skill assessment via video semantic aggregation. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2022. doi:10. 48550/arXiv.2208.02611
112. Law H, Ghani K, Deng J. Surgeon technical skill assessment using computer vision based analysis. *Proc Mach Learn Res* 2017;**68**:88–99
113. Hung AJ, Chen J, Che Z, Nilanon T, Jarc A, Titus M *et al.* Utilizing machine learning and automated performance metrics to evaluate robot-assisted radical prostatectomy performance and predict outcomes. *J Endourol* 2018;**32**:438–444
114. Ma R, Ramaswamy A, Xu J, Trinh L, Kiyasseh D, Chu TN *et al.* Surgical gestures as a method to quantify surgical performance and predict patient outcomes. *NPJ Digit Med* 2022;**5**:187
115. Hung AJ, Chen J, Ghodoussipour S, Oh PJ, Liu Z, Nguyen J *et al.* A deep-learning model using automated performance metrics and clinical features to predict urinary continence recovery after robot-assisted radical prostatectomy. *BJU Int* 2019;**124**: 487–495
116. Liu D, Jiang T, Wang Y, Miao R, Shan F, Li Z. Surgical skill assessment on in-vivo clinical data via the clearness of operating field. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2019
117. Lee D, Yu HW, Kwon H, Kong HJ, Lee KE, Kim HC. Evaluation of surgical skills during robotic surgery by deep learning-based multiple surgical instrument tracking in training and actual operations. *J Clin Med* 2020;**9**:1964
118. Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller PA. Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. *Int J Comput Assist Radiol Surg* 2019;**14**:1611–1617
119. Benmansour M, Malti A, Jannin P. Deep neural network architecture for automated soft surgical skills evaluation using objective structured assessment of technical skills criteria. *Int J Comput Assist Radiol Surg* 2023;**18**:929–937
120. Zia A, Essa I. Automated surgical skill assessment in RMIS training. *Int J Comput Assist Radiol Surg* 2018;**13**:731–739
121. Doughty H, Damen D, Mayol-Cuevas W. Who's better? Who's best? Pairwise deep ranking for skill determination. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018
122. Ershad M, Rege R, Majewicz Fey A. Automatic and near real-time stylistic behavior assessment in robotic surgery. *Int J Comput Assist Radiol Surg* 2019;**14**:635–643
123. Ma R, Vanstrum EB, Lee R, Chen J, Hung AJ. Machine learning in the optimization of robotics in the operative field. *Curr Opin Urol* 2020;**30**:808–816
124. Wu C, Cha J, Sulek J, Sundaram CP, Wachs J, Proctor RW *et al.* Sensor-based indicators of performance changes between sessions during robotic surgery training. *Appl Ergon* 2021;**90**: 103251
125. Ershad M, Rege R, Fey AM. Automatic surgical skill rating using stylistic behavior components. IEEE Xplore. 2018
126. Ershad M, Rege R, Majewicz Fey A. Meaningful assessment of robotic surgical style using the wisdom of crowds. *Int J Comput Assist Radiol Surg* 2018;**13**:1037–1048
127. Wang T, Wang Y, Li M. Towards accurate and interpretable surgical skill assessment: a video-based method incorporating recognized surgical gestures and skill levels. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2020
128. Anastasiou D, Jin Y, Stoyanov D, Mazomenos E. Keep your eye on the best: contrastive regression transformer for skill assessment in robotic surgery. *IEEE Robot Autom Lett* 2023;**8**: 1755–1762
129. Oğul BB, Gilgien MF, Şahin PD. Ranking robot-assisted surgery skills using kinematic sensors. In: Chatzigiannakis I, De Ruyter B, Mavrommati I (eds.), *Ambient Intelligence*. 2019, 11912. Available from: http://link.springer.com/10.1007/978-3-030-34255-5
130. Soleymani A, Sadat Asl AA, Yeganejou M, Dick S, Tavakoli M, Li X. Surgical skill evaluation from robot-assisted surgery recordings. In: *2021 International Symposium on Medical Robotics, ISMR 2021*. Institute of Electrical and Electronics Engineers Inc., 2021
131. Parmar P, Morris BT. Learning to score Olympic events. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2017
132. Xiang X, Tian Y, Reiter A, Hager GD, Tran TD. S3D: Stacking segmental P3D for action quality assessment. In: *Proceedings —International Conference on Image Processing, ICIP*, 2018
133. Fawaz H, Forestier G, Weber J, Idoumghar L, Muller PA. Evaluating surgical skills from kinematic data using convolutional neural networks. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018
134. Wang Z, Majewicz Fey A. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *Int J Comput Assist Radiol Surg* 2018;**13**:1959–1970
135. Anh NX, Nataraja RM, Chauhan S. Towards near real-time assessment of surgical skills: a comparison of feature extraction techniques. *Comput Methods Programs Biomed* 2020; **187**:105234
136. Nguyen XA, Ljuhar D, Pacilli M, Nataraja RM, Chauhan S. Surgical skill levels: classification and analysis using deep neural network model and motion signals. *Comput Methods Programs Biomed* 2019;**177**:1–8
137. Wang Z, Fey AM. Improving surgical skill assessment and task recognition in robot-assisted surgery with deep neural networks. IEEE Xplore. 2018
138. Kiyasseh D, Ma R, Haque TF, Miles BJ, Wagner C, Donoho DA *et al.* A vision transformer for decoding surgeon activity from surgical videos. *Nat Biomed Eng* 2023;**7**:780–796
139. Kiyasseh D, Laca J, Haque TF, Miles BJ, Wagner C, Donoho DA *et al.* A multi-institutional study using artificial intelligence to provide reliable and fair feedback to surgeons. *Commun Med* 2023;**3**:42
140. Kiyasseh D, Laca J, Haque TF, Otiato M, Miles BJ, Wagner C *et al.* Human visual explanations mitigate bias in AI-based assessment of surgeon skills. *NPJ Digit Med* 2023;**6**:54
141. Collins JW, Marcus HJ, Ghazi A, Sridhar A, Hashimoto D, Hager G *et al.* Ethical implications of AI in robotic surgical

training: a Delphi consensus statement. *Eur Urol Focus* 2022;**8**: 613–622
142. World Health Organization. *World Health Organization. 2023 [cited 2023 May 26]. WHO Calls for Safe and Ethical AI for Health*. https://www.who.int/news/item/16-05-2023-who-calls-for-safe-and-ethical-ai-for-health (accessed 26 May 2023)
143. Scott SI, Dalsgaard T, Jepsen JV, von Buchwald C, Andersen SAW. Design and validation of a cross-specialty simulation-based training course in basic robotic surgical skills. *Int J Med Robot* 2020;**16**:1–10
144. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. *Adv Simul (Lond)* 2016;**1**:31
145. Goh AC, Goldfarb DW, Sander JC, Miles BJ, Dunkin BJ. Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. *J Urol* 2012;**187**:247–252
146. Aghazadeh MA, Mercado MA, Pan MM, Miles BJ, Goh AC. Performance of robotic simulated skills tasks is positively associated with clinical robotic surgical performance. *BJU Int* 2016;**118**:475–481
147. von Rundstedt FC, Aghazadeh MA, Scovell J, Slawin J, Armstrong J, Silay S *et al.* Validation of a simulation-training model for robotic intracorporeal bowel anastomosis using a step-by-step technique. *Urology* 2018;**120**:125–130
148. Aghazadeh MA, Jayaratna IS, Hung AJ, Pan MM, Desai MM, GillIS *et al.* External validation of global evaluative assessment of robotic skills (GEARS). *Surg Endosc* 2015;**29**: 3261–3266
149. Whitehurst SV, Lockrow EG, Lendvay TS, Propst AM, DunlowSG, Rosemeyer CJ *et al.* Comparison of two simulation systems to support robotic-assisted surgical training: a pilot study (swine model). *J Minim Invasive Gynecol* 2015;**22**:483–488
150. Sánchez R, Rodríguez O, Rosciano J, Vegas L, Bond V, Rojas A *et al.* Robotic surgery training: construct validity of global evaluative assessment of robotic skills (GEARS). *J Robot Surg* 2016;**10**:227–231
151. Bur AM, Gomez ED, Newman JG, Weinstein GS, O'Malley BW, Rassekh CH *et al.* Evaluation of high-fidelity simulation as a training tool in transoral robotic surgery. *Laryngoscope* 2017; **127**:2790–2795
152. Brown JD, Brien CE O, Leung SC, Dumon KR, Lee DI, Kuchenbecker KJ. Using contact forces and robot arm accelerations to automatically rate surgeon skill at peg transfer. *IEEE Trans Biomed Eng* 2017;**64**:2263–2275
153. Vargas M V, Moawad G, Denny K, Happ L, Misa NY, Margulies S *et al.* Transferability of virtual reality, simulation-based, robotic suturing skills to a live porcine model in novice surgeons: a single-blind randomized controlled trial. *J Minim Invasive Gynecol* 2017;**24**:420–425
154. Oh PJ, Chen J, Hatcher D, Djaladat H, Hung AJ. Crowdsourced versus expert evaluations of the vesico-urethral anastomosis in the robotic radical prostatectomy: is one superior at discriminating differences in automated performance metrics? *J Robot Surg* 2018;**12**:705–711
155. Raison N, Gavazzi A, Abe T, Ahmed K, Dasgupta P. Virtually competent: a comparative analysis of virtual reality and dry-lab robotic simulation training. *J Endourol* 2020;**34**:379–384
156. Fukuoka K, Teishima J, Inoue S, Hayashi T, Matsubara A. The influence of reviewer's occupation on the skill assessment of urethrovesical anastomosis in robot-assisted radical prostatectomy. *Asian J Endosc Surg* 2021;**14**:451–457
157. Hung AJ, Chen J, Jarc A, Hatcher D, Djaladat H, Gill IS. Development and validation of objective performance metrics for robot-assisted radical prostatectomy: a pilot study. *Journal of Urology* 2018;**199**:296–304
158. Yu N, Saadat H, Finelli A, Lee JY, Singal RK, Grantcharov TP *et al.* Quantifying the "assistant effect" in robotic-assisted radical prostatectomy (RARP): measures of technical performance. *J Surg Res* 2021;**260**:307–314
159. Kelly JD, Kowalewski TM, Brand T, French A, Nash M, Meryman L *et al.* Virtual reality warm-up before robot-assisted surgery: a randomized controlled trial. *J Surg Res* 2021;**264**:107–116
160. Kiely DJ, Gotlieb WH, Lau S, Zeng X, Samouelian V, Ramanakumar AV *et al.* Virtual reality robotic surgery simulation curriculum to teach robotic suturing: a randomized controlled trial. *J Robot Surg* 2015;**9**:179–186
161. Guni A, Raison N, Challacombe B, Khan S, Prokar D, Ahmed K. Development of a technical checklist for the assessment of suturing in robotic surgery. *Surg Endosc* 2018;**32**:44024407
162. Goldenberg MG, Goldenberg L, Grantcharov TP. Surgeon performance predicts early continence after robot-assisted radical prostatectomy. *J Endourol* 2017;**31**:858–863
163. Hung AJ, Jayaratna IS, Teruya K, Desai MM, Gill IS, Goh AC. Comparative assessment of three standardized robotic surgery training methods. *BJU Int* 2013;**112**:864–871
164. Ramos P, Montez J, Tripp A, Ng CK, Gill IS, Hung AJ. Face, content, construct and concurrent validity of dry laboratory exercises for robotic training using a global assessment tool. *BJU Int* 2014;**113**:836–842
165. Dubin AK, Smith R, Julian D, Tanaka A, Mattingly P. A comparison of robotic simulation performance on basic virtual reality skills: simulator subjective versus objective assessment tools. *J Minim Invasive Gynecol* 2017;**24**:1184–1189
166. Almarzouq A, Hu J, Noureldin YA, Yin A, Anidjar M, Bladou F *et al.* Are basic robotic surgical skills transferable from the simulator to the operating room? A randomized, prospective, educational study. *Canad Urol Assoc J* 2020;**14**:416–422
167. Ross T, Raison N, Wallace L, Wood T, Lovegrove C, Van der Poel H *et al.* PD41-06 robot-assisted training—expert performance in full immersion simulation, setting the benchmark (concurrent validity). *J Urol* 2017;**197**:e809
168. Valdis M, Chu MWA, Schlachta CM, Kiaii B. Validation of a novel virtual reality training curriculum for robotic cardiac surgery a randomized trial. *Innovations (Phila)* 2015;10: 383–388
169. Valdis M, Chu MWA, Schlachta C, Kiaii B. Evaluation of robotic cardiac surgery simulation training: a randomized controlled trial. *J Thorac Cardiovasc Surg* 2016;**151**:1498–1505.e2
170. Monda SM, Weese JR, Anderson BG, Vetter JM, Venkatesh R, Du K *et al.* Development and validity of a silicone renal tumor model for robotic partial nephrectomy training. *Urology* 2018; **114**:114–120
171. Timberlake MD, Garbens A, Schlomer BJ, Kavoussi NL, Kern AJM, Peters CA *et al.* Design and validation of a low-cost, high-fidelity model for robotic pyeloplasty simulation training. *J Pediatr Urol* 2020;**16**:332–339
172. Butterworth J, Sadry M, Julian D, Haig F. Assessment of the training program for Versius, a new innovative robotic system for use in minimal access surgery. *BMJ Surg Interv Health Tech* 2021;**3**:e000057
173. Raison N, Patrick H, Abe T, Aydin A, Ahmed K, Prokar D. Procedural virtual reality simulation training for robotic surgery: a randomised controlled trial. *Surg Endosc* 2021;**35**:6897–6902
174. Tarr ME, Anderson-Montoya BL, Vilasagar S, Myers EM. Validation of a simulation model for robotic sacrocolpopexy. *Female Pelvic Med Reconstr Surg* 2022;**28**:14–19

175. Sarcona J, Mikhail D, Tabibzadeh A, Nassau D, Kozel Z, Vira M *et al.* MP34-07 Correlating crowd-sourced assessment of technical skills (CSATS) with post-operative complication rates in urological surgery. *J Urol* 2020;**203**:e505
176. Liang MI, McCann GA, Rath KS, Backes FJ, Cansino C, Salani R. Training the next generation of robotic surgeons using guided mentorship: a randomized controlled trial. *J Minim Invasive Gynecol* 2014;**21**:1075–1079
177. Carter SC, Chiang A, Shah G, Kwan L, Montgomery JS, Karam A *et al.* Video-based peer feedback through social networking for robotic surgery simulation: a multicenter randomized controlled trial. *Ann Surg* 2015;**261**:870–875
178. Bendre HH, Rajender A, Philip, Barbosa V, Wason SEL. Robotic dismembered pyeloplasty surgical simulation using a 3D-printed silicone-based model: development, face validation and crowdsourced learning outcomes assessment. *J Robot Surg* 2020; **14**:897–902
179. Chen C, White L, Kowalewski T, Aggarwal R, Lintott C, Comstock B *et al.* Crowd-sourced assessment of technical skills: a novel method to evaluate surgical performance. *Journal of Surgical Research* 2014;**187**:65–71
180. Goldenberg MG, Nabhani J, Wallis CJD, Chopra S, Hung AJ, Schuckman A *et al.* Feasibility of expert and crowd-sourced review of intraoperative video for quality improvement of intracorporeal urinary diversion during robotic radical cystectomy. *Canad Urol Assoc J* 2017;**11**:331–336
181. Mills JT, Hougen HY, Bitner D, Krupski TL, Schenkman NS. Does robotic surgical simulator performance correlate with surgical skill? *J Surg Educ* 2017;**74**:1052–1056
182. Addison P, Yoo A, Duarte-Ramos J, Addy J, Dechario S, Husk G *et al.* Correlation between operative time and crowd-sourced skills assessment for robotic bariatric surgery. *Surg Endosc* 2021;**35**:5303–5309
183. Tarr ME, Rivard C, Petzel AE, Summers S, Mueller ER, Rickey LM *et al.* Robotic objective structured assessment of technical skills: a randomized multicenter dry laboratory training pilot study. *Female Pelvic Med Reconstr Surg* 2014;**20**: 228–236
184. Knab LM, Zureikat AH, Zeh HJ, Hogg ME. Towards standardized robotic surgery in gastrointestinal oncology. *Langenbecks Arch Surg* 2017;**402**:1003–1014
185. Tam V, Zenati M, Novak S, Chen Y, Zureikat AH, Zeh HJ *et al.* Robotic pancreatoduodenectomy biotissue curriculum has validity and improves technical performance for surgical oncology fellows. *J Surg Educ* 2017;**74**:1057–1065
186. Rice MJK, Zenati MS, Novak SM, Al Abbas AI, Zureikat AH, Zeh HJ *et al.* Crowdsourced assessment of inanimate biotissue drills: a valid and cost-effective way to evaluate surgical trainees. *J Surg Educ* 2019;**76**:814–823
187. Curry M, Malpani A, Li R, Tantillo T, Jog A, Blanco R *et al.* Objective assessment in residency-based training for transoral robotic surgery. *Laryngoscope* 2012;**122**:2184–2192
188. Alemozaffar M, Narayanan R, Percy AA, Minnillo BB, Steinberg P, Haleblian G *et al.* Validation of a novel, tissue-based simulator for robot-assisted radical prostatectomy. *J Endourol* 2014;**28**:995–1000
189. Hernandez JD, Bann SD, Munz Y, Moorthy K, Datta V, Martin S *et al.* Qualitative and quantitative analysis of the learning curve of a simulated surgical task on the da Vinci system. *Surg Endosc* 2004;**18**:372–378
190. Hutchinson K, Li Z, Cantrell LA, Schenkman NS, Alemzadeh H. Analysis of executional and procedural errors in dry-lab robotic surgery experiments. *Int J Med Robot* 2022;**18**:e2375
191. Vogell A, Wright V, Wright K. An evaluation of the utility of robotic virtual reality simulation in gynecologic resident surgical education. *J Minim Invasive Gynaecol*. 2014;**21**:S84–S85
192. Chen QY, Zhong Q, Liu ZY, Li P, Wang JB, Lin JX *et al.* Surgical outcomes, technical performance and surgery burden of robotic total gastrectomy for locally advanced gastric cancer: a prospective study. *Ann Surg* 2021;**276**:e434–e443
193. Korets R, Mues AC, Graversen JA, Gupta M, Benson MC, Cooper KL *et al.* Validating the use of the Mimic dV-trainer for robotic surgery skill acquisition among urology residents. *Urology* 2011;**78**:1326–1330
194. Ahmad SB, Rice M, Chang C, Hamad A, Kingham TP, He J *et al.* Will it play in Peoria? A pilot study of a robotic skills curriculum for surgical oncology fellows. *Ann Surg Oncol* 2021;**28**:6273–6282
195. Zwart MJW, Jones LR, Fuente I, Balduzzi A, Takagi K, Novak S *et al.* Performance with robotic surgery *versus* 3D- and 2D-laparoscopy during pancreatic and biliary anastomoses in a biotissue model: pooled analysis of two randomized trials. *Surg Endosc* 2022;**36**:4518–4528
196. Moncayo S, Compagnon R, Caire F, Grosos C, Bahans C, Ilhero P *et al.* Transition effects from laparoscopic to robotic surgery skills in small cavities. *J Robot Surg* 2020;**14**:525–530
197. Lee JY, Mattar T, Parisi TJ, Carlsen BT, Bishop AT, Shin AY. Learning curve of robotic-assisted microvascular anastomosis in the rat. *J Reconstr Microsurg* 2012;**28**:451–456
198. Vedula SS, Malpani A, Ahmidi N, Khudanpur S, Hager G, Chen CCG. Task-level *vs.* Segment-level quantitative metrics for surgical skill assessment. *J Surg Educ* 2016;**73**:482–489
199. Hung AJ, Ng CK, Patil MB, Zehnder P, Huang E, Aron M *et al.* Validation of a novel robotic-assisted partial nephrectomy surgical training model. *BJU Int* 2012;**110**:870–874
200. Hung AJ, Patil MB, Zehnder P, Cai J, Ng CK, Aron M *et al.* Concurrent and predictive validation of a novel robotic surgery simulator: a prospective, randomized study. *J Urol* 2012;**187**:630–637
201. Culligan P, Gurshumov E, Lewis C, Priestley J, Komar J, Salamon C. Predictive validity of a training protocol using a robotic surgery simulator. *Female Pelvic Med Reconstr Surg* 2014;**20**:48–51
202. Siddiqui NY, Galloway ML, Geller EJ, Green IC, Hur HC, Langston K *et al.* Validity and reliability of the robotic objective structured assessment of technical skills. *Obstet Gynecol* 2014;**123**:1193–1199
203. Polin MR, Siddiqui NY, Comstock BA, Hesham H, Brown C, Lendvay TS *et al.* Crowdsourcing: a valid alternative to expert evaluation of robotic surgery skills. *Am J Obstet Gynecol* 2016; **215**:644.e1–644.e7
204. Newcomb LK, Bradley MS, Truong T, Tang M, Comstock B, Li YJ *et al.* Correlation of virtual reality simulation and dry lab robotic technical skills. *J Minim Invasive Gynecol* 2018;**25**: 689–696
205. Haque TF, Hui A, You J, Ma R, Nguyen JH, Lei X *et al.* An assessment tool to provide targeted feedback to robotic surgical trainees: development and validation of the end-to-end assessment of suturing expertise (EASE). *Urol Pract* 2022;**9**:532–539
206. Hutchins AR, Manson RJ, Lerebours R, Farjat AE, Cox ML, Mann BP *et al.* Objective assessment of the early stages of the learning curve for the Senhance surgical robotic system. *J Surg Educ* 2018;**76**:201–214
207. Arain NA, Dulan G, Hogg DC, Rege R V, Powers CE, Tesfay ST *et al.* Comprehensive proficiency-based inanimate training for robotic surgery: reliability, feasibility, and educational benefit. *Surg Endosc* 2012;**26**:2740–2745

208. Dulan G, Rege RV, Hogg DC, Gilberg-Fisher KM, Arain NA, Tesfay ST *et al.* Proficiency-based training for robotic surgery: construct validity, workload, and expert levels for nine inanimate exercises. *Surg Endosc* 2012;**26**:1516–1521
209. Dulan G, Rege RV, Hogg DC, Gilberg-Fisher KM, Arain NA, Tesfay ST *et al.* Developing a comprehensive, proficiency-based training program for robotic surgery. *Surgery* 2012;**152**:477–488
210. Bric J, Connolly M, Kastenmeier A, Goldblatt M, Gould JC. Proficiency training on a virtual reality robotic surgical skills curriculum. *Surg Endosc* 2014;**28**:3343–3348
211. Suh IH, Lagrange CA, Oleynikov D, Siu KC. Evaluating robotic surgical skills performance under distractive environment using objective and subjective measures. *Surg Innov* 2016;**23**:78–89
212. Hung AJ, Zehnder P, Patil MB, Cai J, Ng CK, Aron M *et al.* Face, content and construct validity of a novel robotic surgery simulator. *Journal of Urology* 2011;**186**:1019–1025
213. Kelly DC, Margules AC, Kundavaram CR, Narins H, Gomella LG, Trabulsi EJ *et al.* Face, content, and construct validation of the da Vinci skills simulator. *Urology* 2012;**79**:1068–1072
214. Alzahrani T, Haddad R, Alkhayal A, Delisle J, Drudi L, Gotlieb W *et al.* Validation of the da Vinci surgical skill simulator across three surgical disciplines: a pilot study. *J Canad Urol Assoc* 2013;**7**:520
215. Lyons C, Goldfarb D, Jones SL, Badhiwala N, Miles B, Link R *et al.* Which skills really matter? Proving face, content, and construct validity for a commercial robotic simulator. *Surg Endosc* 2013;**27**:2020–2030
216. Foell K, Finello A, Yasufuku K, Benardini M, Waddell T, Pace K *et al.* Robotic surgery basic skills training: evaluation of a pilot multidisciplinary simulation-based curriculum. *Canad Urol Assoc J* 2013;**7**:430
217. Cecilie Havemann M, Dalsgaard T, Led Sørensen J, Røssaak K, Brisling S, Jul Mosgaard B *et al.* Examining validity evidence for a simulation-based assessment tool for basic robotic surgical skills. *J Robot Surg* 2019;**13**:99–106
218. Perrenot C, Perez M, Tran N, Jehl JP, Felblinger J, Bresler L *et al.*The virtual reality simulator dV-trainer is a valid assessment tool for robotic surgical skills. *Surg Endosc* 2012;**26**:2587–2593
219. Liss MA, Abdelshehid C, Quach S, Lusch A, Graversen J, Landman J *et al.* Validation, correlation, and comparison of the da Vinci Trainer™ and the da Vinci Surgical Skills Simulator™ using the Mimic™ software for urologic robotic surgical education. *J Endourol* 2012;**26**:1629–1634
220. Hertz AM, George EI, Vaccaro CM, Brand TC. Head-to-head comparison of three virtual-reality robotic surgery simulators. *J Soc Laparoendosc Surg* 2018;**22**:e2017.00081
221. Sheth SS, Fader AN, Tergas AI, Kushnir CL, Green IC. Virtual reality robotic surgical simulation: an analysis of gynecology trainees. *J Surg Educ* 2014;**71**:125–132
222. Connolly M, Seligman J, Kastenmeier A, Goldblatt M, Gould JC. Validation of a virtual reality-based robotic surgical skills curriculum. *Surg Endosc* 2014;**28**:1691–1694
223. Brown K, Mosley N, Tierney J. Battle of the bots: a comparison of the standard da Vinci and the da Vinci Surgical Skills simulator in surgical skills acquisition. *J Robot Surg* 2017;**11**:159–162
224. Finnegan KT, Meraney AM, Staff I, Shichman SJ. Da Vinci Skills simulator construct validation study: correlation of prior robotic experience with overall score and time score simulator performance. *Urology* 2012;**80**:330–336
225. Liss MA, Kane CJ, Chen T, Baumgartner J, Derweesh IH. Virtual reality suturing task as an objective test for robotic experience assessment. *BMC Urol* 2015;**15**:63
226. Yamany T, Woldu SL, Korets R, Badani KK. Effect of postcall fatigue on surgical skills measured by a robotic simulator. *J Endourol* 2015;**29**:479–484
227. Vogell A, Gujral H, Wright K, Ruthazer R. Impact of a robotic simulation program on resident surgical performance. *Am J Obstet Gynaecol* 2015;**213**:874–875
228. Brinkman WM, Luursema JM, Kengen B, Schout BMA, Witjes JA, Bekkers RL. Da Vinci Skills Simulator for assessing learning curve and criterion-based training of robotic basic skills. *Urology* 2013;**81**:562–566
229. Robison W, Patel SK, Mehta A, Senkowski T, Allen J, Shaw E *et al.* Can fatigue affect acquisition of new surgical skills? A prospective trial of pre- and post-call general surgery residents using the da Vinci Surgical Skills simulator. *Surg Endosc* 2018;**32**:1389–1396
230. Gleason A, Servais E, Quadri S, Manganiello M, Cheah YL, Simon CJ *et al.* Developing basic robotic skills using virtual reality simulation and automated assessment tools: a multidisciplinary robotic virtual reality-based curriculum using the da Vinci skills simulator and tracking progress with the intuitive learning platform. *J Robot Surg* 2022;**16**: 1313–1319
231. Kenney PA, Wszolek MF, Gould JJ, Libertino JA, Moinzadeh A. Face, content, and construct validity of dV-Trainer, a novel virtual reality simulator for robotic surgery. *Urology* 2009;**73**: 1288–1292
232. Kang SG, Cho S, Kang SH, Haidar AM, Samavedi S, Palmer KJ *et al.* The tube 3 module designed for practicing vesicourethral anastomosis in a virtual reality robotic simulator: determination of face, content, and construct validity. *Urology* 2014;**84**:345–350
233. Schreuder HWR, Persson JEU, Wolswijk RGH, Ihse I, Schijven MP, Verheijen RHM. Validation of a novel virtual reality simulator for robotic surgery. *ScientificWorldJournal* 2014;**2014**:507076
234. Cho JS, Hahn KY, Kwak JM, Kim J, Baek SJ, Shin JW *et al.* Virtual reality training improves da Vinci performance: a prospective trial. *J Laparoendosc Adv Surg Tech* 2013;**23**:992–998
235. Lendvay TS, Brand TC, White L, Kowalewski T, Jonnadula S, Mercer LD *et al.* Virtual reality robotic surgery warm-up improves task performance in a dry laboratory environment: a prospective randomized controlled study. *J Am Coll Surg* 2013;**216**:1181–1192
236. Ruparel RK, Taylor AS, Patel J, Patel VR, Heckman MG, Rawal B *et al.* Assessment of virtual reality robotic simulation performance by urology resident trainees. *J Surg Educ* 2014;**71**:302–308
237. Sethi AS, Peine WJ, Mohammadi Y, Sundaram CP. Validation of a novel virtual reality robotic simulator. *J Endourol* 2009;**23**: 503–508
238. Kim JY, Bin KS, Pyun JH, Kim HK, Cho S, Lee JG *et al.* Concurrent and predictive validation of robotic simulator tube 3 module. *Korean J Urol* 2015;**56**:756–761
239. Schommer E, Patel VR, Mouraviev V, Thomas C, Thiel DD. Diffusion of robotic technology into urologic practice has led to improved resident physician robotic skills. *J Surg Educ* 2017; **74**:55–60
240. Watkinson W, Raison N, Abe T, Harrison P, Khan S, Van der Poel H *et al.* Establishing objective benchmarks in robotic virtual reality simulation at the level of a competent surgeon using the RobotiX mentor simulator. *Postgrad Med J* 2018;**94**: 270–277
241. Leijte E, De Blaauw I, Rosman C, Botden SMBI. Assessment of validity evidence for the RobotiX robot assisted surgery simulator on advanced suturing tasks. *BMC Surg* 2020;**20**:183

242. Whittaker G, Aydin A, Raison N, Kum F, Challacombe B, Khan MS *et al.* Validation of the RobotiX mentor robotic surgery simulator. *J Endourol* 2016;**30**:338–346
243. Hovgaard LH, Andersen SAW, Konge L, Dalsgaard T, Larsen CR. Validity evidence for procedural competency in virtual reality robotic simulation, establishing a credible pass/fail standard for the vaginal cuff closure procedure. *Surg Endosc* 2018;**32**: 4200–4208
244. Alshuaibi M, Perrenot C, Hubert J, Perez M. Concurrent, face, content, and construct validity of the RobotiX mentor simulator for robotic basic skills. *Int J Med Robot* 2020;**16**:e2100
245. McDonough PS, Tausch TJ, Peterson AC, Brand TC. Initial validation of the ProMIS surgical simulator as an objective measure of robotic task performance. *J Robot Surg* 2011;**5**: 195–199
246. Jonsson M, Mahmood M, Askerud T, Hellborg H, Ramel S, Wiklund P *et al.* Promis^{TM} can serve as a da Vinci simulator—a construct validity study. *J Endourol* 2011;**25**:345–350
247. Chandra V, Nehra D, Parent R, Woo R, Reyes R, Hernandez-Boussard T *et al.* A comparison of laparoscopic and robotic assisted suturing performance by experts and novices. *Surgery* 2010;**147**:830–839
248. Seixas-Mikelus SA, Kesavadas T, Srimathveeravalli G, Chandrasekhar R, Wilding GE, Guru KA. Face validation of a novel robotic surgical simulator. *Urology* 2010;**76**:357–360
249. Seixas-Mikelus SA, Stegemann AP, Kesavadas T, Srimathveeravalli G, Sathyaseelan G, Chandrasekhar R *et al.* Content validation of a novel robotic surgical simulator. *BJU Int* 2011;**107**:1130–1135
250. Chowriappa AJ, Shi Y, Raza SJ, Ahmed K, Stegemann A, Wilding G *et al.* Development and validation of a composite scoring system for robot-assisted surgical training—the robotic skills assessment score. *J Surg Res* 2013;**185**:561–569
251. Cowan A, Chen J, Mingo S, Reddy SS, Ma R, Marshall S *et al.* Virtual reality *vs* dry laboratory models: comparing automated performance metrics and cognitive workload during robotic simulation training. *J Endourol* 2021;**35**:1571–1576
252. Nguyen JH, Chen J, Marshall SP, Ghodoussipour S, Chen A, Gill IS *et al.* Using objective robotic automated performance metrics and task-evoked pupillary response to distinguish surgeon expertise. *World J Urol* 2020;**38**:1599–1605
253. Verner L, Oleynikov D, Holtmann S, Haider H, Zhukov L. Measurements of the level of surgical expertise using flight path analysis from da Vinci robotic surgical system. *Stud Health Technol Inform* 2003;**94**:373–378
254. Narazaki K, Oleynikov D, Stergiou N. Objective assessment of proficiency with bimanual inanimate tasks in robotic laparoscopy. *J Laparoendosc Adv Surg Tech* 2007;**17**:47–52
255. Judkins TN, Oleynikov D, Stergiou N. Objective evaluation of expert and novice performance during robotic surgical training tasks. *Surg Endosc* 2009;**23**:590–597
256. Hung AJ, Oh PJ, Chen J, Ghodoussipour S, Lane C, Jarc A *et al.* Experts *vs* super-experts: differences in automated performance metrics and clinical outcomes for robot-assisted radical prostatectomy. *BJU Int* 2019;**123**:861–868
257. Narazaki K, Oleynikov D, Stergiou N. Robotic surgery training and performance identifying objective variables for quantifying the extent of proficiency. *Surg Endosc* 2006;**20**:96–103
258. Lefor AK, Harada K, Dosis A, Mitsuishi M. Motion analysis of the JHU–ISI gesture and skill assessment working set II: learning curve analysis. *Int J Comput Assist Radiol Surg* 2021;**16**:589–595
259. Tausch TJ, Kowalewski TM, White LW, McDonough PS, Brand TC, Lendvay TS. Content and construct validation of a robotic surgery curriculum using an electromagnetic instrument tracker. *J Urol* 2012;**188**:919–923
260. Walker JL, Nathwani JN, Mohamadipanah H, Laufer S, Jocewicz FF, Gwillim E *et al.* Residents' response to bleeding during a simulated robotic surgery task. *J Surg Res* 2017;**220**:385–390
261. Nisky I, Okamura AM, Hsieh MH. Effects of robotic manipulators on movements of novices and surgeons. *Surg Endosc* 2014;**28**:2145–2158
262. Nisky I, Hsieh MH, Okamura AM. The effect of a robot-assisted surgical system on the kinematics of user movements. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, EMBS, 2013, 6257–6260
263. Jog A, Itkowitz B, Liu M, DiMaio S, Hager G, Curet M *et al.* Towards integrating task information in skills assessment for dexterous tasks in surgery and simulation. In: *Proceedings—IEEE International Conference on Robotics and Automation*, 2011, 5273–5278
264. Kumar R, Jog A, Malpani A, Vagvolgyi B, Yuh D, Nguyen H *et al.* Assessing system operation skills in robotic surgery trainees. *Int J Med Robot* 2011;**8**:118–124
265. Tao L, Elhamifar E, Khudanpur S, Hager GD, Vidal R. Sparse hidden Markov models for surgical gesture classification and skill evaluation. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012
266. Fard MJ, Ameri S, Darin Ellis R, Chinnam RB, Pandya AK, Klein MD. Automated robot-assisted surgical skill evaluation: predictive analytics approach. *Int J Med Robotics Comput Assist Surg* 2018;**14**:e1850
267. Pan JH, Gao J, Zheng WS. Action assessment by joint relation graphs. In: *Proceedings of the IEEE International Conference on Computer Vision*, 2019
268. Funke I, Mees ST, Weitz J, Speidel S. Video-based surgical skill assessment using 3D convolutional neural networks. *Int J Comput Assist Radiol Surg* 2019;**14**:1217–1225
269. Ming Y, Cheng Y, Chunchen W, Meng L, Guang Z, Feng C. Automated Objective Basic Surgical Skills Assessment: Overall Kinematic Performance Assessment Method. IEEE Xplore. 2020:74–78
270. Tang Y, Ni Z, Zhou J, Zhang D, Lu J, Wu Y *et al.* Uncertainty-aware score distribution learning for action quality assessment. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020
271. Lyman WB, Passeri MJ, Murphy K, Siddiqui IA, Khan AS, Iannitti DA *et al.* An objective approach to evaluate novice robotic surgeons using a combination of kinematics and stepwise cumulative sum (CUSUM) analyses. *Surg Endosc* 2021;**35**:2765–2772
272. Yu X, Rao Y, Zhao W, Lu J, Zhou J. Group-aware contrastive regression for action quality assessment. In: *Proceedings of the IEEE International Conference on Computer Vision*, 2021
273. Zhang J, Nie Y, Lyu Y, Yang X, Chang J, Zhang JJ. SD-Net: joint surgical gesture recognition and skill assessment. *Int J Comput Assist Radiol Surg* 2021;**16**:1675–1682
274. Soleymani A, Li X, Tavakoli M. A domain-adapted machine learning approach for visual evaluation and interpretation of robot-assisted surgery skills. *IEEE Robot Autom Lett* 2022;**7**: 8202–8208
275. Juarez-Villalobos L, Hevia-Montiel N, Perez-Gonzalez J. Machine learning based classification of local robotic surgical skills in a training tasks set. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Institute of Electrical and Electronics Engineers Inc., 2021, 4596–4599

276. Kumar R, Jog A, Vagvolgyi B, Nguyen H, Hager G, Chen CCG *et al.* Objective measures for longitudinal assessment of robotic surgery training. *J Thorac Cardiovasc Surg* 2012;**143**:528–534
277. Ahmidi N, Gao Y, Béjar B, Vedula SS, Khudanpur S, Vidal R *et al.* String motif-based description of tool motion for detecting skill and gestures in robotic surgery. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013
278. Malpani A, Swaroop Vedula S, Chiung C, Chen G, Hager GD. A study of crowdsourced segment-level surgical skill assessment using pairwise rankings. *Int J CARS* 2015;**10**:1435–1447
279. Lajkó G, Elek RN, Haidegger T. Endoscopic image-based skill assessment in robot-assisted minimally invasive surgery. *Sensors* 2021;**21**:5412
280. Takács K, Haidegger T. Adaptive neuro-fuzzy inference system for automated skill assessment in robot-assisted minimally invasive surgery. In: *INES 2021—IEEE 25th International Conference on Intelligent Engineering Systems, Proceedings*, 2021
281. Brown KC, Bhattacharyya KD, Kulason S, Zia A, Jarc A. How to bring surgery to the next level: interpretable skills assessment in robotic-assisted surgery. *Visc Med* 2020;**36**:463–470
282. Goldenberg MG, Grantcharov TP. A novel method of setting performance standards in surgery using patient outcomes. *Ann Surg* 2019;**269**:79–82
283. Gavazzi A, Bahsoun A, Van Haute W, Ahmed K, Elhage O, Jaye P *et al.* Face, content and construct validity of a virtual reality simulator for robotic surgery (SEP robot). *Ann R Coll Surg Engl* 2011;**93**:152–156

European Colorectal Congress

3 – 6 December 2023, St.Gallen, Switzerland

OVERVIEW
Sun, 3 Dec 2023

MASTERCLASS
PROCTOLOGY DAY
ROBOTIC COURSE
DAVOSCOURSE@ECC

SCIENTIFIC PROGRAMME
Mon, 4 Dec – Wed, 6 Dec 2023

DIVERTICULAR DISEASE

Gut microbiome and surgery
Phil Quirke, Leeds, UK

Diet in diverticular disease
Pamela Buchwald, Lund, SE

Decision making in the management of acute complicated Diverticulitis beyond the guidelines
Seraina Faes, Zurich, CH

Diverticular Abscess – Always drainage or who benefits from Surgery?
Johannes Schultz, Oslo, NO

Perforated Diverticulitis: Damage Control, Hartmann's Procedure, Primary Anastomosis, Diverting Loop
Reinhold Kafka-Ritsch, Innsbruck, AT

When to avoid protective stoma in colorectal surgery
Antonino Spinelli, Milano, IT

ENDOMETRIOSIS

Endometriosis – what is the role of the abdominal surgeon
Tuynman Juriaan, Amsterdam, NL

Challenges in Surgery of Endometriosis – always interdisciplinary?
Peter Oppelt, Linz, AT; Andreas Shamiyeh, Linz, AT

A gaze in the crystal ball: Where is the role of virtual reality and artificial Intelligence in colorectal surgery
Müller Beat, Basel, CH

MALIGNANT COLORECTAL DISEASE

Cytoreductive Surgery and Intraperitoneal Chemotherapy – facts and hopes
Michel Adamina, Winterthur, CH

Metastatic Colorectal Cancer – surgical approaches and limits
Jürgen Weitz, Dresden, DE

Extended lymph node dissection for rectal cancer, is it still under debate?
Miranda Kusters, Amsterdam, NL

Organ preservation functional outcome in rectal cancer treatment – in line with patient's needs? (Robot – laparoscopic – open surgery?)
Hans de Wilt, Nijmegen, NL

ROBOTICS

Advances in Robotic Surgery and what we learnt so far
Parvaiz Amjad, Portsmouth, UK

Challenging the market: Robotic (assistant) Devices and how to choose wisely (Da Vinci – Hugo Ras – Distalmotion ua)
Khan Jim, London, UK

TAMIS - Robotic Transanal Surgery, does it make it easier?
Knol Joep, Genk, BE

Live Surgery – Contonal Hospital of St.Gallen
Walter Brunner, St.Gallen, CH;
Salvadore Conde Morals, Sevilla, ES;
Friedrich Herbst, Vienna, AUT;
Amjad Parvaiz, Portsmouth, UK

Video Session

Lars Pahlmann Lecture
Markus Büchler, Lisboa, PRT

Honorary Lecture
Bill Heald, Lisboa, PRT

Information & Registration **www.colorectalsurgery.eu**