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

Matthew W. E. Boal1,2,3*, Dimitrios Anastasiou2,4, Freweini Tesfai1,2, Walaa Ghamrawi1, Evangelos Mazomenos2,4, Nathan Curtis5, Justin W. Collins3,6, Ashwin Srithar3,6, John Kelly3,6, Danail Stoyanov2,7* and Nader K. Francis1,3,8,*

1The Griffin Institute, Northwick Park & St Mark’s Hospital, London, UK
2Wellcome/ESPRC Centre for Interventional Surgical Sciences (WEISS), University College London (UCL), London, UK
3Division of Surgery and Interventional Science, Research Department of Targeted Intervention, UCL, London, UK
4Medical Physics and Biomedical Engineering, UCL, London, UK
5Department of General Surgery, Dorset County Hospital NHS Foundation Trust, Dorchester, UK
6University College London Hospitals NHS Foundation Trust, London, UK
7Computer Science, UCL, London, UK
8Yeovil 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 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. Minimally invasive surgery (MIS) is complex, highly variable and requires technical skill with unfavourable error profiles compared to industrial data. 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. Of surgical patients, 10–15 per cent in the UK experience adverse events, of which 50 per cent are preventable. Adverse events relating to robotic procedures (10 624) were reported in the USA between 2000 and 2013 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. This deficit is being addressed through development and standardization of basic and specialty curricula.

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. 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
assessment of performance and intervention delivery\textsuperscript{6,25–27}. This facilitates direct formative feedback to guide reduction in proficiency curves, as well as summative assessment\textsuperscript{28}. Objective assessments within MIS have demonstrated reliability and clinically relevant validities\textsuperscript{6,25–27}, leading to the development of tools to aid this\textsuperscript{28}. However, previous studies have highlighted the variability in reporting on the validity and reliability of manual tools\textsuperscript{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\textsuperscript{30,31} or combinations of surgical approaches\textsuperscript{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\textsuperscript{32}. Finally, given the rapid uptake of robotic techniques and development of artificial intelligence (AI) methods, many reports are now outdated\textsuperscript{33–37}, requiring up-to-date evaluation\textsuperscript{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 \textit{a priori} protocol (PROSPERO registration ID CRD42022304901). The Covidence\textregistered{} 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\textsuperscript{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\textsuperscript{36} and four systematic reviews\textsuperscript{26,27,31,42}, 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\textsuperscript{16} 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\textsuperscript{43} and the Modified Educational Oxford Centre for Evidence-Based Medicine (OCEBM) Levels of Evidence (LoE) and Levels of Recommendation (LoR)\textsuperscript{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)\textsuperscript{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)\textsuperscript{46}, as most study designs are conceptual.

Tables 2–4 summarize the main tools in each domain of technical skill assessment and \textit{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.
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, good to strong/excellent inter-observer group reliability compared to experts, as well as construct and predictive validity. 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 and the modified Angoff method, 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. With regards to specialty, 21 (46.7 per cent) studies assessed 15 urology tools, 9 (20 per cent) surgery tools, 3 (6.7 per cent) gynaecology tools, 5 (11 per cent) orthopaedic tools and 1 (2.2 per cent) ophthalmology tools.
# Table 2 Summary of manual assessment tools

## Global rating scale tools

<table>
<thead>
<tr>
<th>Tool</th>
<th>Study type</th>
<th>Setting</th>
<th>Test content</th>
<th>Response process</th>
<th>Internal structure</th>
<th>Relationship to other variables</th>
<th>Consequences</th>
<th>LoE</th>
<th>LoR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEARS</td>
<td>46 Observational</td>
<td>36 Lab &amp; OR</td>
<td>Experts developed based on GOALS</td>
<td>All studies (except one) demonstrated response process</td>
<td>Inter-rater reliability</td>
<td>Concurrent validity</td>
<td>80-100% = good to excellent (arbitrary definition) Benchmarking used expert scores</td>
<td>Level 2</td>
<td>Level 1 recommendation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Inter-rater reliability Strong/Very strong/Excellent/High</td>
<td></td>
<td></td>
<td></td>
<td>LoR</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Acceptable/Good/Moderate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low/Poor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Intra-rater reliability Strong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Construct validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Excellent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Internal consistency Strong/Excellent/High/Strong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Intra-rater reliability Strong/Excellent/High/Strong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Predictive validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Level 1b, 2b, 3b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Level 2b, 3b, 4b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OSATS</td>
<td>20 Observational</td>
<td>19 Lab &amp; OR</td>
<td>Expert group added 2 domains to GOALS creating OSATS to create GOALS+</td>
<td>All but two studies demonstrated response process</td>
<td>Inter-rater reliability</td>
<td>Concurrent validity</td>
<td>Hypothesized mean OSATS category scores 3.5 novice and 4.5 expert in concordance with results of mean Expert benchmarking</td>
<td>Level 2</td>
<td>Level 2 recommendation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Inter-rater reliability High/Excellent/High/Excellent/High</td>
<td></td>
<td></td>
<td></td>
<td>LoR</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Acceptable/Good/Moderate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low/Poor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Intra-rater reliability Strong/Excellent/High/Strong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Construct validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Excellent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Internal consistency Strong/Excellent/High/Strong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Intra-rater reliability Strong/Excellent/High/Strong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Predictive validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Level 1b, 2b, 3b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GOALS</td>
<td>4 Observational</td>
<td>5 Lab &amp; OR</td>
<td>Expert group added 2 domains to GOALS creating OSATS to create GOALS+</td>
<td>All studies demonstrated response process</td>
<td>Inter-rater reliability</td>
<td>Concurrent validity</td>
<td>Pass mark defined by contrasting groups method by experts</td>
<td>Level 2</td>
<td>Level 2 recommendation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Inter-rater reliability High/Excellent/High/Excellent/High</td>
<td></td>
<td></td>
<td></td>
<td>LoR</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Acceptable/Good/Moderate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low/Poor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Intra-rater reliability Strong/Excellent/High/Strong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Construct validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Excellent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Internal consistency Strong/Excellent/High/Strong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Intra-rater reliability Strong/Excellent/High/Strong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Predictive validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Level 1b, 2b, 3b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-OSATS</td>
<td>4 Observational</td>
<td>4 Lab</td>
<td>Expert group added 2 domains to GOALS creating OSATS to create GOALS+</td>
<td>All studies demonstrated response process</td>
<td>Inter-rater reliability</td>
<td>Concurrent validity</td>
<td>Modified Angoff method set threshold competency scores per drill</td>
<td>Level 2</td>
<td>Level 2 recommendation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Inter-rater reliability Strong/Excellent/High/Excellent/High</td>
<td></td>
<td></td>
<td></td>
<td>LoR</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Acceptable/Good/Moderate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low/Poor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Intra-rater reliability Strong/Excellent/High/Strong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Construct validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Excellent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Internal consistency Strong/Excellent/High/Strong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Intra-rater reliability Strong/Excellent/High/Strong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Predictive validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Level 1b, 2b, 3b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Procedure- and Task-Specific Assessment Tools

<table>
<thead>
<tr>
<th>Tool</th>
<th>Study type</th>
<th>Setting</th>
<th>Test content</th>
<th>Response process</th>
<th>Internal structure</th>
<th>Relationship to other variables</th>
<th>Consequences</th>
<th>LoE</th>
<th>LoR</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSATS</td>
<td>5 Observational</td>
<td>7 Lab</td>
<td>No studies</td>
<td>All demonstrated response process</td>
<td>Inter-rater reliability</td>
<td>Concurrent validity</td>
<td>Pass mark based on experts</td>
<td>Level 2</td>
<td>Level 2 recommendation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Inter-rater reliability High</td>
<td></td>
<td></td>
<td></td>
<td>LoR</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Non-simulator APMs, Simulator APMs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Construct validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Construct validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RACE</td>
<td>5 Observational</td>
<td>3 Lab</td>
<td>Delphi consensus</td>
<td>All demonstrated response process</td>
<td>Inter-rater reliability</td>
<td>Concurrent validity</td>
<td>Pass mark based on experts</td>
<td>Level 2</td>
<td>Level 2 recommendation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Inter-rater reliability Strong/Excellent/High/Excellent/High</td>
<td></td>
<td></td>
<td></td>
<td>LoR</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Acceptable/Good/Moderate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low/Poor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Intra-rater reliability Strong/Excellent/High/Strong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task-Perform-ance Metrics Tools</td>
<td>1 Observational</td>
<td>2 Delphi and OR videos</td>
<td>All tools developed through Delphi consensus</td>
<td>All tools demonstrated response process</td>
<td>Percentage agreement 0.85–0.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(continued)
cent) ear, nose and throat, 1 (2.2 per cent) microsurgery, 8 (17.8 per cent) suturing, 2 (4.4 per cent) dissection, 1 (2.2 per cent) vessel dissection and ligation, and 1 (2.2 per cent) other dry model tasks.

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, 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 (46.5 per cent) studies. 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, unlike the dVSS and dV-T, were not included in the review of robotic surgery.

### 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 Vinci Systems 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, unlike the dVSS and dV-T, were not included in the review of robotic surgery.
### Table 3 Summary of automated performance metrics

#### Simulator automated performance metrics

<table>
<thead>
<tr>
<th>Tool</th>
<th>Study type</th>
<th>Setting</th>
<th>Test content</th>
<th>Response process</th>
<th>Internal structure</th>
<th>Relationship to other variables</th>
<th>Consequences</th>
<th>LoE</th>
<th>LoR</th>
</tr>
</thead>
<tbody>
<tr>
<td>da Vinci Surgical Skills Simulator</td>
<td>11 RCTs</td>
<td>30 Lab</td>
<td>14 studies showing face and content validity. Note three studies stated dVSS more realistic/helpful than dV.</td>
<td>Automated metrics inherently no rater bias</td>
<td>11 RCTs (3 not powered)</td>
<td>Concurrent validity</td>
<td>Pass mark defined</td>
<td>Level 1 recommendation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2 Powered non-randomized studies</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3 Randomization of training</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3 RCTs, 1 study measured with consecutive attempts with excellent reliability.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dV-Trainer</td>
<td>4 RCTs</td>
<td>17 Lab</td>
<td>9 Face/content</td>
<td>Automated metrics inherently no rater bias</td>
<td>4 RCTs (2 not powered)</td>
<td>Concurrent validity</td>
<td>Pass mark defined</td>
<td>Level 1 recommendation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3 Randomization of training</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RobotIX Mentor</td>
<td>1 RCT</td>
<td>1 Lab</td>
<td>All studies showed face/content except two.</td>
<td>Automated metrics inherently no rater bias</td>
<td>Orientation/standardization of training</td>
<td>Concurrent validity with FRS of dry lab</td>
<td>Pass mark defined</td>
<td>Level 2 recommendation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Randomized groups.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Promi™ hybrid surgical simulator</td>
<td>3 Lab</td>
<td>Face and content</td>
<td>Standardized orientation</td>
<td>Internal consistency</td>
<td>Construct validity</td>
<td>Level 2 recommendation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Robotic Surgery Simulator (RosSi)</td>
<td>1 RCT</td>
<td>1 Lab and Delphi</td>
<td>Face and content</td>
<td>Automated metrics inherently no rater bias</td>
<td>Construct validity. Predictive of console performance in the lab.</td>
<td>Level 2 recommendation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3 Randomization of training</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3D hydrogel models —clinically relevant objective/ performance metrics (CROMS/ CEPMS)</td>
<td>2 Lab</td>
<td>1 Lab and Delphi</td>
<td>Face and content both studies</td>
<td>Automated metrics inherently no rater bias</td>
<td>Construct validity</td>
<td>Level 2 recommendation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pilot testing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Randomized, powered</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>da Vinci SimNow™</td>
<td>1 RCT</td>
<td>Lab</td>
<td>Automated metrics inherently without rater bias</td>
<td>Automated metrics inherently reliable</td>
<td>Concurrent validity with da Vinci system event data recorder</td>
<td>Level 3 recommendation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Automated metrics inherently reliable</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Automated metrics inherently reliable</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Versius Trainer™</td>
<td>1 Lab</td>
<td>Automated metrics inherently without rater bias</td>
<td>Automated metrics inherently reliable</td>
<td>Construct validity</td>
<td>Level 2 recommendation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Automated metrics inherently reliable</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Automated metrics inherently reliable</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Non-Simulator Automated Performance Metrics

<table>
<thead>
<tr>
<th>Tool</th>
<th>Study type</th>
<th>Setting</th>
<th>Test content</th>
<th>Response process</th>
<th>Internal structure</th>
<th>Relationship to other variables</th>
<th>Consequences</th>
<th>LoE</th>
<th>LoR</th>
</tr>
</thead>
<tbody>
<tr>
<td>da Vinci Kinematic and System Event Recorders</td>
<td>1 RCT</td>
<td>Lab</td>
<td>Automated metrics inherently without rater bias</td>
<td>Automated metrics inherently reliable</td>
<td>Concurrent validity</td>
<td>Pass mark defined</td>
<td>Level 1 recommendation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1 study randomized</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>da Vinci API, dVLogger or da Vinci Systems Events data recorder</td>
<td>6 OR</td>
<td>Automated metrics inherently without rater bias</td>
<td>Automated metrics inherently reliable</td>
<td>Concurrent validity</td>
<td>Pass mark defined</td>
<td>Level 2 recommendation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(continued)
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 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 to 77 (median = 8). Most studies employed the publicly available JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS), 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. Other studies assigned assessment scores to group surgeons above a predefined threshold as experts and below as novices.

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, laparoscopic cholecystectomy, robotic assisted radical prostatectomy (RARP) urethrovesical anastomosis, and phases of gastrectomy and thyroid surgery. RARP procedures were the most common (7/12), with Chen et al. 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 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, estimating the operating field clearness, using stylistic behaviour labels and linking skill levels to clinical outcomes in RARP.

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), eye-tracking and electroencephalography (EEG) signals, electromyography data (EMG) and galvanic skin response (GSR).

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 or to a simple classifier/regressor. Kinematic-based deep learning approaches use either temporal convolutional networks (TCN) or recurrent neural network (RNN) or a combination of the two.

Notably, deep learning approaches have gained popularity in surgical skill assessment.

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 utilized 10 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...
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

---

**Fig. 2 PRISMA flowchart**
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.

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
Artificial intelligence: Promising initial results, with the potential to transform formative and summative assessment, particularly if evaluated with APMs.

- To credential surgeons as competent for independent practice, blinded expert video rating is considered an essential part of accreditation. 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 incorporated into curricula. Expert consensus on tool to use for training AI, and a standardized method for example experts who are trained with reliability tests.

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, including a recent statement from the World Health Organization.

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, 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.
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), Evangeli 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


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
training: a Delphi consensus statement. Eur Urol Focus 2022; 8:613–622


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


216. Foell K, Finello A, Yasufuku K, Benardini M, Waddell T, Pace K
213. Kelly DC, Margules AC, Kundavaram CR, Narins H, Gomella LG,
220. Hertz AM, George EI, Vaccaro CM, Brand TC. Head-to-head
223. Brown K, Mosley N, Tierney J. Battle of the bots: a comparison of
224. Finnegan KT, Meraney AM, Staff I, Shichman SJ. Da Vinci Skills
Brisling S, Jul Mosgaard B
Assoc J
virtual reality simulator dV-trainer is a valid assessment tool for
surgical skills.
for a simulation-based assessment tool for basic robotic
surgical skills.
construct validity for a commercial robotic simulator.
Journal of Urology 2011;186:1019–1025
Kelly DC, Margules AC, Kundavaram CR, Narins H, Gomella LG,
Trabulsi EJ et al. Face, content, and construct validation of the
Alzahrani T, Haddad R, Alkhayal A, Delisle J, Drudi L, Gotlieb W
et al. Validation of the da Vinci surgical skill simulator across
Lyons C, Goldfarb D, Jones SL, Badhivala N, Miles B, Link R et al.
Cecilie Havemann M, Dalsgaard T, Led Sørensen J, Røssaak K.
209. Dulan G, Rege RV, Hogg DC, Gilbert-Fisher KM, Arain NA,
208. Dulan G, Rege RV, Hogg DC, Gilbert-Fisher KM, Arain NA,
Eric J, Connolly M, Kastenmeier A, Goldblatt M, Gould JC.
233. Schreuder HWR, Persson JEU, Wolswijk RGH, Ihse I, Schijven MP,
223. Brown K, Mosley N, Tierney J. Battle of the bots: a comparison of
Brisling S, Jul Mosgaard B
Assoc J
virtual reality simulator dV-trainer is a valid assessment tool for
surgical skills.
for a simulation-based assessment tool for basic robotic
surgical skills.
construct validity for a commercial robotic simulator.
Journal of Urology 2011;186:1019–1025
Kelly DC, Margules AC, Kundavaram CR, Narins H, Gomella LG,
Trabulsi EJ et al. Face, content, and construct validation of the
Alzahrani T, Haddad R, Alkhayal A, Delisle J, Drudi L, Gotlieb W
et al. Validation of the da Vinci surgical skill simulator across
Lyons C, Goldfarb D, Jones SL, Badhivala N, Miles B, Link R et al.
Connolly M, Seigman J, Kastenmeier A, Goldblatt M, Gould JC.
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


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