International Journal of Medical Informatics 223 (2027) 106776

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Review

Artificial intelligence for video-based assessment of surgical performance in robot-assisted minimally invasive surgery in real operative settings: a systematic review

Jiani Tang [a,c,*], Ana Manzano Rodriguez [a,c,d], Erdem Uğur [e], Marlies P. Schijven [a,b,c,*,1]

[a] Amsterdam UMC, Location University of Amsterdam, Department of Surgery, Amsterdam, the Netherlands
[b] Amsterdam Gastroenterology and Metabolism, Amsterdam, the Netherlands
[c] Amsterdam Public Health, Digital Health, Amsterdam, the Netherlands
[d] Data Science Center HAVA-lab, University of Amsterdam, Amsterdam, the Netherlands
[e] Amsterdam UMC, Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Robotic technical skill assessment
Video assessment
Artificial intelligence
Real surgery
Robot-assisted surgery
Minimally invasive surgery

ABSTRACT

Background: Robot-assisted minimally invasive surgery requires specific skills of the surgeon beyond the mastery of conventional laparoscopic techniques. However, efficient and objective methods for evaluating these skills remain limited. Artificial intelligence (AI) offers potential for a more objective, video-based performance assessment. This systematic review aims to evaluate the current evidence of using AI for evaluation of surgical skill in real robotic surgery and outline potential avenues for future research.

Methods: We systematically searched Ovid Medline, Embase, Clarivate Web of Science, ERIC Institute of Education Sciences and Scopus databases from inception to 10 July 2025, for studies evaluating AI-based technical skill assessment using videos of robot-assisted surgery performed on real patients. Study characteristics, AI approaches, assessment targets, reference standards, and model performance were extracted. Risk of bias and applicability were assessed using PROBAST + AI. Owing to differences in assessment constructs and reported performance measures, findings were synthesised using a structured narrative approach.

Results: In total, 2116 articles were identified, of which 5 met our eligibility criteria. AI approaches and assessment targets varied substantially across studies, including prediction of global technical skill scores, analysis of instrument-derived performance metrics, procedure-specific assessment, and surgical error detection. Reference standards included OSATS, GEARS or modified GEARS, PLACE, and procedural experience. Reported performance measures represented different constructs and were not directly comparable across studies. PROBAST + AI identified important methodological limitations, and no study reported external validation using an independent dataset.

Conclusions: AI shows potential for video-based technical skill assessment in real robotic surgery, but current evidence is limited and heterogeneous. Future research should prioritise standardised annotation, multicentre intraoperative datasets, and external validation to support reliable translation into clinical practice.

Abbreviations: AI, artificial intelligence; ASD, automated structure detection; AUC, Area under the curve; CASE, cholecystectomy assessment and surgical evaluation; CI, confidence interval; CV, computer vision; DL, deep learning; GEARS, Global Evaluative Assessment of Robotic Skills; IoU, Intersection over Union; mAP, mean Average Precision; Mask R-CNN, mask region-based convolutional neural network; M-GEARS, Modifiable-Global Evaluative Assessment of Robotic Skills; mtCNN, multitask convolutional neural network; MIS, minimally invasive surgery; ML, machine learning; N–Vs, nerves and vessels; OR, Operating room; OSATS, Objective Structured Assessment of Technical Skill; PLACE, Pelvic Lymphadenectomy Assessment and Completion Evaluation; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-analyses PROBAST+AI, Prediction model Risk of Bias Assessment Tool–Artificial Intelligence RoB, risk of bias; SPaN, score for partial nephrectomy.

* Corresponding authors at: Department of Surgery, Amsterdam UMC, de Boelelaan 1117, 1081 HV Amsterdam, the Netherlands.
E-mail addresses: j.tang2@amsterdamumc.nl (J. Tang), m.p.schijven@amsterdamumc.nl (M.P. Schijven).
[1] Corresponding author during review process.

https://doi.org/10.1016/j.ijmedinf.2026.106776
Received 28 July 2026; Received in revised form 13 October 2026; Accepted 19 October 2026
Available online 20 October 2026
1386-5056/© 2026 Published by Elsevier B.V.

1. Introduction

To date, the use of a robot to help perform minimally invasive surgery (MIS) has become the preferred standard for many surgical procedures in diverse settings [1–3]. Indeed, the number of robotic surgeries (surgeons using a robot to help perform MIS) is on the rise [4,5]. Performing robotic surgery requires a high and specific level of skill of the surgeon [6,7]. Skills, that cannot be merely extrapolated from one's proficiency derived from performing MIS, not even by an expert surgeon. Surgeons using a surgical robot to perform MIS need to develop a specific wrist technique to operate the robot and must get used to the end-effector motion scaling of the system. Also, the need to adjust to the three-dimensional visualization of the operative field, as conventional laparoscopy usually relies on looking at monitors at a relative distance, displaying a two-dimensional image. Furthermore, also in contrast to conventional MIS, they have to get used to a total lack of tactile and force feedback when performing surgery [8]. Until recent, robotic systems were reserved as instruments for expert surgeons, taking the leap from conventional MIS to robotic supported MIS surgery [9]. As more surgical robots became available in more hospitals, and research mounted on the benefits of use, it became more mainstream to use robots for MIS procedures. With this, it became apparent that training on robotic MIS in the core surgical skills curriculum is needed. Hence, it has become of interest to train residents within their training period to gain proficiency using a robotic surgical system, preferrable without the need for surgeons in understaffed departments, usually busy performing surgery, to be present [10,11].

Artificial Intelligence (AI) is experiencing rapid advancements across various domains, including healthcare [12,13]. AI and its subfields, such as machine learning (ML), computer vision (CV), and deep learning (DL) have the potential to supply the user with meaningful information provided that the underlying data is well-collected and reasonably free from bias. In robotic MIS, an important source of data for these intelligent systems may come from endoscopic surgical video analysis. AI has the potential to evaluate the performance of robotic procedures free from observer and bias, such as the associated Hawthorne effect, however, there are other challenges associated [14]. Previous studies have signalled marked variability and concern, reporting on the validity and reliability of metric tools to analyse videos recorded from dry labs in simulative settings [15]. In the clinic, objective assessment of surgical skill is far more complex than in simulative settings. This is due to factors such as (but not limited to) aberrant anatomy, intraoperative bleeding, and surgeon stress. Pedrett et al. (2023) [16] attempted to synthesize the available literature on technical skill assessment for both robotic and laparoscopic surgeries in the operating room. This study was challenged as there was great variability and difference in the generated video image data and modes of operation. For example, in conventional MIS, endoscope manipulation necessitates a dedicated assistant; whereas in robot-assisted surgery, once inserted in the patient, autonomous camera control can be enabled by the primary operator sitting in the robotic console. This introduces interpersonal coordination variability as an important confounding factor when assessing technical skill assessment in conventional MIS when compared to robot-assisted MIS.

Due to the rapid adoption of robotic techniques and the continuous advancement of AI methods, we need to revisit current observational models for video-based skills assessment in surgery. This, as the use of checklists for skill assessment is tedious, prone to bias, and may even have become outdated as techniques develop. Indeed, current times call for the need for updated evaluations on how to assess robotic MIS using modern computational technology [14,17–19]. Previous systematic reviews have evaluated AI applications in robotic surgery by combining simulation, dry-lab, and kinematic studies. While these reviews provide a broad overview of the field, they do not specifically address the evidence on real operative settings, where clinical complexity may substantially influence AI model development. The aim of this systematic review is focused on evaluating the question: To date, is information enclosed in surgical videos generated from performing robotic-assisted MIS surgery, properly distilled as data supporting reliable AI models assessing surgical performance?

Our objective is to synthesize prior studies, assess them for heterogeneity, and analyse reported results whilst identifying gaps and biases. By focusing exclusively on intraoperative video from real operative settings, this systematic review aims to define the current evidence supporting AI-based assessment of robot-assisted minimally invasive surgery and identify priorities for future research.

Table 1
Metrics for evaluating models of all studies included in this review.

Metric	Formula	Definition
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Proportion of correct predictions among the total number of observations
mAP	$\frac{1}{N}\sum_{i=1}^{N} APi$	calculated by finding Average Precision(AP) for each class and then average over a number of classes
Precision	$\frac{TP}{TP+FP}$	Ratio of true positive predictions to all positive predictions
IoU	$\frac{\lvert GT \cap P\rvert}{\lvert GT \cup P\rvert}$	Performance metric for evaluating the accuracy of fit of a segmentation network and measures the degree of overlap between ground truth and predicted area
AUC	–	Area under a receiver operating characteristic curve
F1 score	$\frac{2TP}{2TP+FP+FN}$	Harmonic mean of precision and recall
Macro-F1	$\frac{F1+F2+\cdots+Fn}{n}$	Arithmetic mean of all the per-class F1 scores

TP: True positives, TN: True negatives, FP: False positives, FN: False negatives, IoU: Intersection over Union, GT: Ground truth area, P: Predicted area, mAP: mean Average Precision, AUC: Area under the curve, Recall: Probability of correct positive predictions in actual positives.

2. Materials and methods

2.1. Search strategy and selection criteria

This systematic review was conducted according to a prespecified protocol registered in PROSPERO (CRD42024534137), and is reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [20] (Appendix p8-9).

In this systematic review, we have searched the OVID/Medline, Embase.com, Clarivate Analytics/Web of Science Core Collection, Ebsco/ERIC, and Elsevier/Scopus databases from inception to May 2024 in collaboration with a medical information specialist (JCFK). The search included controlled and free text terms for synonyms of 'machine learning' or 'artificial intelligence', 'education' or 'learning', 'educational measurement', 'robotic surgical procedures', and 'surgery'. The search was performed without restrictions for methodology, date, or language. The database searches were rerun from inception on July 10, 2025. For OVID/Medline, the search terms were adapted to reflect changes in the controlled vocabulary and hierarchical organisation of artificial intelligence related terms available at the time of the updated search. The review question and eligibility criteria remained unchanged. Full search strategies and exact search dates for both the initial and updated searches are provided in Table S3 (appendix p11-15).

Studies were eligible if they focused on AI assessment of technical surgical performance using video data from robot-assisted surgical procedures performed on real patients in the operating room. Full inclusion criteria using the population, intervention, comparison, and outcomes PICOTS framework are described in Appendix 2 (p3). Studies were excluded if they: (1) used simulation, dry-laboratory, or other non-clinical robotic surgical videos; (2) combined video or image data with non-video sensor or kinematic data as model inputs; (3) focused solely on surgical phase or procedure recognition, anatomical or instrument recognition without subsequent assessment of technical performance, or

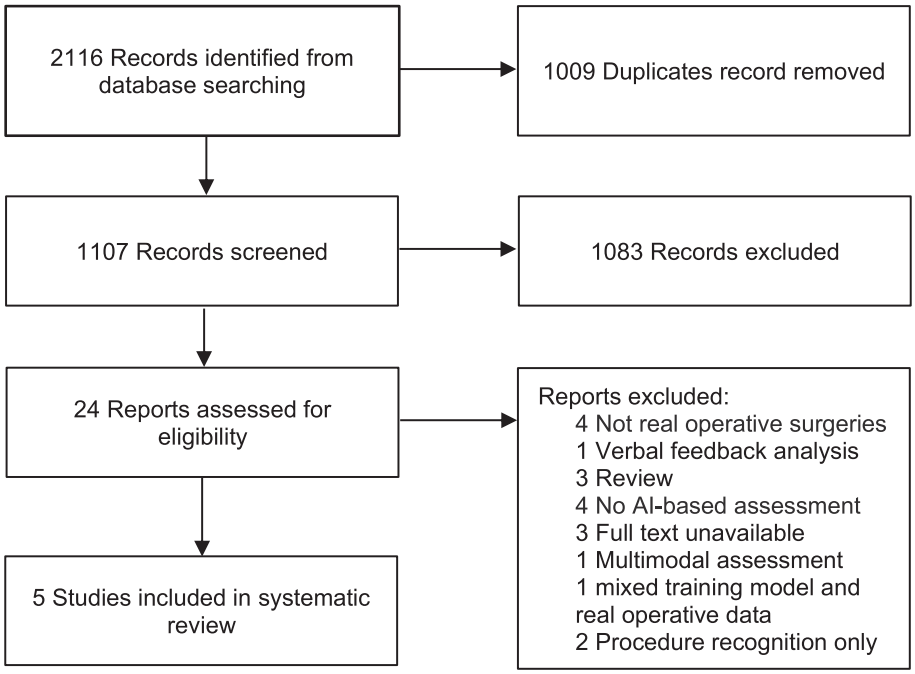


Fig. 1. Study selection.

automated analysis of surgical feedback without assessment of operative technical performance. Additionally, commentaries, expert consensus statements, reviews, editorials, and studies for which the full text could not be retrieved were excluded.

2.2. Study selection and data extraction

JT and EU independently screened the titles and abstracts for eligibility using Rayyan (https://www.rayyan.ai) [21], followed by a second screening of the full texts to exclude inappropriate studies. Disagreements between the reviewers were discussed and resolved for consensus.

Data from eligible studies were extracted into a standardised Microsoft Excel spreadsheet (Microsoft, Redmond, Washington, USA). Extracted study characteristics included publication year, first author's department, country, surgical procedure, assessed surgical task, and data size. AI related characteristics included model architecture, type of AI-based assessment, AI assessment approach, assessment target, reference standard, model output, unit of analysis, evaluation design, and external validation. Performance measures and corresponding results were extracted according to the definitions reported in each original study. Definitions related AI performance metrics encountered in the included studies are provided in Table 1.

2.3. Risk of bias and applicability assessment

The methodological quality of the included studies was assessed using PROBAST + AI, a risk of bias and applicability assessment tool developed for studies evaluating prediction models using artificial intelligence methods [22]. Assessments were performed separately for model development and model evaluation when applicable. For model development, concerns regarding development quality were assessed across the relevant domains. For model evaluation, risk of bias was assessed separately. Applicability concerns were evaluated in relation to the review question, specifically the use of video-based AI for assessment of technical surgical performance in real robot-assisted operative settings.

2.4. Data synthesis

A structured narrative synthesis was conducted to summarise the findings of the included studies. Studies were compared according to the type of AI-based assessment, assessment target, unit of analysis, evaluation design, and reported performance. AI performance measures were interpreted according to their definitions and intended purposed in the original studies.

Given differences in the constructs being assessed and the methods used to quantify model performance, results were synthesised with their respective assessment contexts. Reported measures included direct skill score prediction accuracy, associations between AI-derived performance metrics and expert ratings, performance of computer vision components used to derive skill-related measures, and surgical error detection metrics. These measures were not considered interchangeable estimates of a common outcome and were therefore reported using the definitions and scales provided by the original studies. Where a study reported multiple models, outcomes, or performance measures derived from the same or overlapping samples, these findings were synthesised at the study level rather than treated as independent observations.

3. Results

3.1. Study selection

The initial search conducted to May 2024, identified 2297 records and resulted in 5 eligible studies (Supplementary Fig. S1, page 16). The final search rerun from inception to July 10, 2025, identified 2116 records. After removing duplicates, titles and abstracts of the remaining 1107 studies were screened. Twenty-four studies remained, 19 of which were excluded after full-text screening. In summary, five studies met the eligibility criteria and were thus included in this systematic review (Fig. 1) [23–27]. Detailed reasons for full-text exclusion are provided in Supplementary Appendix 1 (page 2).

3.2. Study characteristics and AI-based assessment approaches

The characteristics and methodological approaches of the 5 studies are summarised in Table 2, with complete extracted information

Table 2
Characteristics of the studies in this systematic review.

Characteristic	No. of studies	References
First author's disciplinary affiliation		
Technical discipline	5	Baghdadi et al. (2018) [23], Yang et al. (2022) [24], Wang et al. (2023) [25], Strong et al. (2024) [26], Sirajudeen et al. (2024) [27]
Country of origin		
United States	3	Baghdadi et al. (2018) [23], Yang et al. (2022) [24], Wang et al. (2023) [25]
United Kingdom	1	Sirajudeen et al. (2024) [27]
Japan	1	Strong et al. (2024) [26]
Specialty		
Urology	3	Baghdadi et al. (2018) [23], Wang et al. (2023) [25], Sirajudeen et al. (2024) [27]
General Surgery	2	Yang et al. (2022) [24], Strong et al. (2024) [26]
Type of operative assessment		
Procedure-specific technical performance	1	Baghdadi et al. (2018) [23]
AI-derived instrument characteristics associated with technical skill	1	Yang et al. (2022) [24]
Prediction of expert-evaluated technical skill scores	1	Wang et al. (2023) [25]
Instrument use patterns according to procedural experience	1	Strong et al. (2024) [26]
Technical skill estimation and surgical error detection	1	Sirajudeen et al. (2024) [27]
Reference standard		
PLACE	1	Baghdadi et al. (2018) [23]
OSATS	2	Wang et al. (2023) [25], Sirajudeen et al. (2024) [27]
GEARS	2	Yang et al. (2022) [24], Wang et al. (2023) [25]
Modified GEARS	1	Sirajudeen et al. (2024) [27]
OCHRA-based error. annotations	1	Sirajudeen et al. (2024) [27]
Procedural experience	1	Strong et al. (2024) [26]

Data are the number of studies unless otherwise indicated. Categories are not mutually exclusive where individual studies used more than one technical skill assessment or reference standard. PLACE, Pelvic Lymphadenectomy Assessment and Completion Evaluation; GEARS, Global Evaluative Assessment of Robotic Skills; OSATS, Objective Structured Assessment of Technical Skills; M−GEARS, modified Global Evaluative Assessment of Robotic Skills; OCHRA, Observational Clinical Human Reliability Analysis.

provided in Table S4 (supplement page 17–18). The studies evaluated video-based AI across different robot-assisted surgical procedures and operative tasks, including pelvic lymph node dissection during robot-assisted radical cystectomy [23], peritoneal closure during robotic rectopexy [24], resection, renorrhaphy, and bolster placement during robot-assisted partial nephrectomy [25], infrapyloric lymph node dissection during robotic distal gastrectomy [26], and dorsal venous complex suturing during robot-assisted radical prostatectomy [27]. Dataset sizes and analytical units varied across studies, including procedure-level data, operative clips or task videos, images, and frame-level or temporal sequences.

Although all included studies analysed operative video, technical performance was operationalised using substantially different AI-based assessment approaches. Baghdadi et al. [23] used automated structure detection to derive operative-field features for assessment against performance based on Pelvic Lymphadenectomy Assessment and Completion Evaluation (PLACE) (All the details of these standard scores are in Appendix 3 and Table S2 (page 4–7 and 10). Yang et al. [24] used instrument segmentation and tracking to derive objective motion metrics and examined their associations with expert evaluated Global Evaluative Assessment of Robotic Skills (GEARS) domains. Wang et al. [25] used video-derived features and temporal modelling to predict GEARS and OSATS scores. Strong et al. [26] used automated instrument recognition to quantify instrument use patterns and subsequently compared these measures between experience groups. Sirajudeen et al. [27] combined video-based technical skill estimation with surgical error detection.

The underlying AI approaches included automated structure detection, Mask R-CNN-based instrument segmentation and tracking, convolutional and temporal neural networks, semantic segmentation, and transformer-based feature extraction (Fig. 2a; Table S4).

3.3. Reference standards and model performance

The assessment targets and reference standards differed substantially across studies. Baghdadi et al. [23] assessed performance using the PLACE. Yang et al. [24] examined associations between AI-derived instrument-motion metrics and GEARS domains rated by experts, whereas Wang et al. [25] directly predicted GEARS and OSATS scores. Sirajudeen et al. [27] used OSATS and modified GEARS for technical-skill estimation and OCHRA-based annotations for surgical error detection. Strong et al. [26] did not use a validated technical skill rating scale, instead, AI-derived instrument use measures were compared between groups defined by procedural experience.

Reported performance measures were not directly comparable across studies because they represented different assessment targets and statistical quantities. Results are presented within the context of each study's assessment approach and reference standard (Supplementary Table S4).

Overall, the included studies differed in the construct being assessed, reference standard, unit of analysis, model evaluation design, and definition of model performance. No common performance measure was identified across the 5 studies. In addition, no study reported external validation using an independent dataset.

3.4. Risk of bias and applicability

PROBAST + AI identified important methodological limitations across the included studies (Table 3). All 5 studies were judged to have a high overall risk of bias for model evaluation. Key concerns varied across studies and included small numbers of independent procedures or surgeons, limited representation of technical skill spectrum, non-independence or clustering of observations, uncertainty regarding separation of development and evaluation data, and absence of external validation.

For model development, high overall quality concerns were identified in Baghdadi et al. [23], Wang et al. [25], Strong et al. [26], Sirajudeen et al. [27]. Development assessment did not apply to Yang et al. [24], because the study evaluated associations between AI-derived performance metrics and expert skill ratings rather than developing a supervised technical skill prediction model.

Applicability concerns were low for Baghdadi et al. [23], Yang et al. [24], Wang et al. [25], and Sirajudeen et al. [27]. Strong et al. [26] had high applicability concern for the outcome domain because procedural experience, rather than a validated technical skill reference standard, was used as the comparator for the skill related analysis. Detailed signalling question assessments and study-specific rationales are provided in Table S5 and S6. respectively.

4. Discussion

This systematic review provides a focused synthesis of the current evidence for AI-based technical skill assessment using intraoperative video from real robot-assisted minimally invasive surgery. Although previous studies have evaluated broader applications of AI in robotic

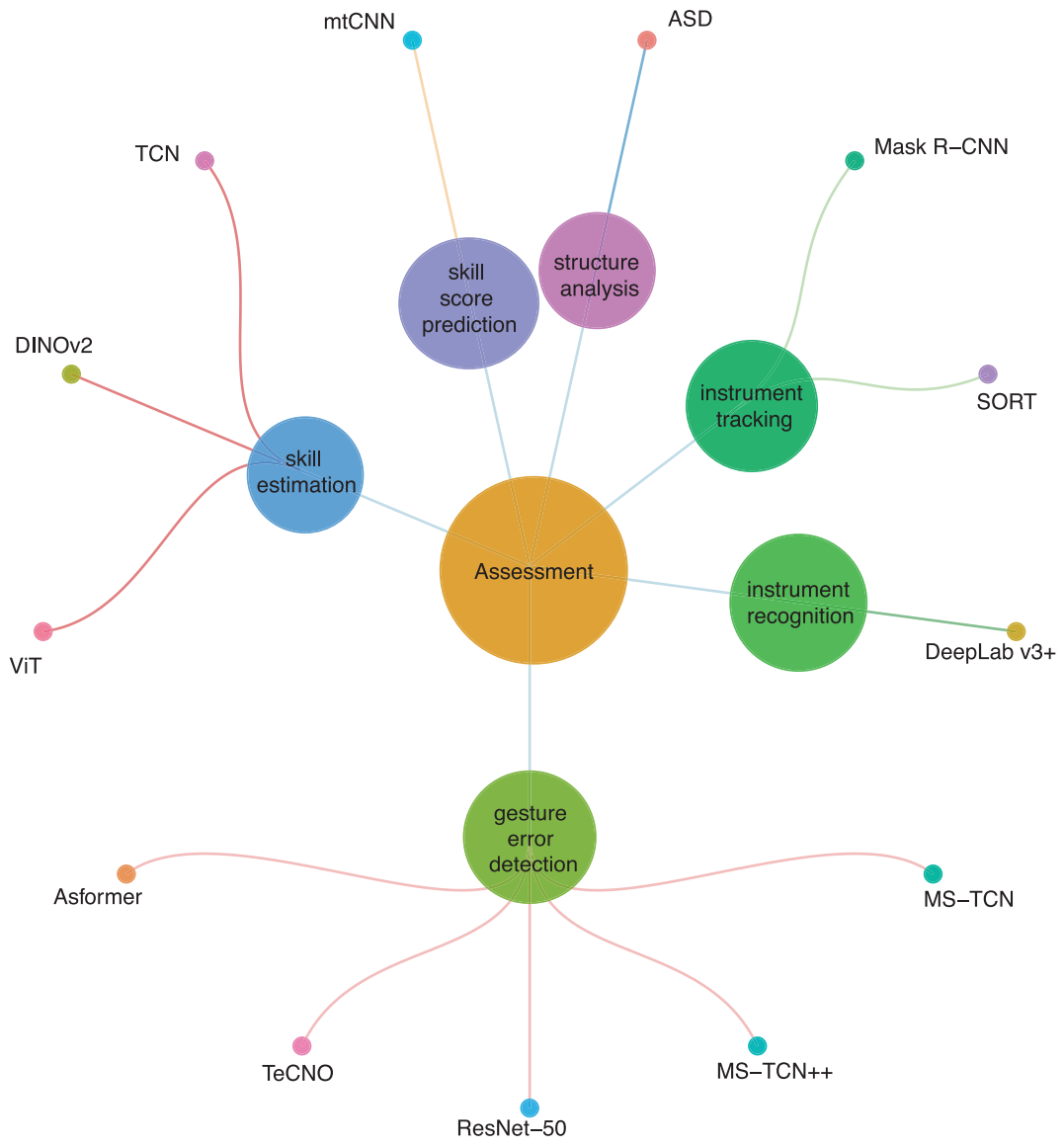


Fig. 2a. Distribution of AI techniques applied for skills evaluation across the included studies in this systematic review.

surgery, our findings demonstrate that evidence derived from real operative settings remains limited and methodologically heterogeneous.

The included studies demonstrated substantial clinical and methodological heterogeneity, particularly in the constructs being assessed, reference standards, units of analysis, AI approaches, and definitions of model performance. Importantly, labelled performance measures did not necessarily represent the same underlying outcome. For example, reported"accuracy" could refer to correct classification of technical performance, exact matching of predicted and skill scores rated by experts, performance of an upstream computer vision component, or a value derived from prediction error. These measures therefore cannot be interpreted as interchangeable estimates of a common AI performance outcome. The current evidence should consequently be interpreted at the individual study level and within the context of each assessment approach. In addition, the absence of external validation and the methodological limitations identified by PROBAST + AI restrict conclusions regarding the generalisability of the reported model performance. Despite these limitations, the available studies demonstrate the potential of AI-based approaches for objective technical skill assessment and support further methodological development in this emerging field.

Many studies included in previous systematic reviews used surgical videos mainly from the public dataset 'JIGSAWS' as input data [16]. This dataset is captured using the da Vinci Surgical System and results from eight surgeons with different levels of skill performing five repetitions of three elementary surgical tasks on a bench-top model. Hence, it is highly likely that this dataset is biased in itself, as experience level varies considerably a lot between surgeons and there were only eight surgeons involved. The five studies included here used videos of surgery on patients to construct training and testing datasets. In general, assessments conducted on public datasets such as JIGSAW or private datasets in dry lab environments can provide quantitative scores that reflect the performance of fundamental surgical skill tasks. These controlled settings allow for the evaluation of specific technical skills under standardized conditions. However, real-world surgical operations include a much higher degree of intricacy that is not captured in dry lab environments. One imperative challenge is the variation in real patient tissue textures because standardized synthetic tissue textures in dry labs are not the same, especially for patients who have undergone treatments such as chemotherapy or radiotherapy [28]. These characteristics make surgical manipulation more challenging but are absent in dry lab simulations. Additionally, haemorrhage is common in real patient operations; however, it is not a realistic event that can be vividly replicated in dry labs [29]. For example, when suturing, bleeding control depends on the surgeon's ability, which calls for a degree of competence and

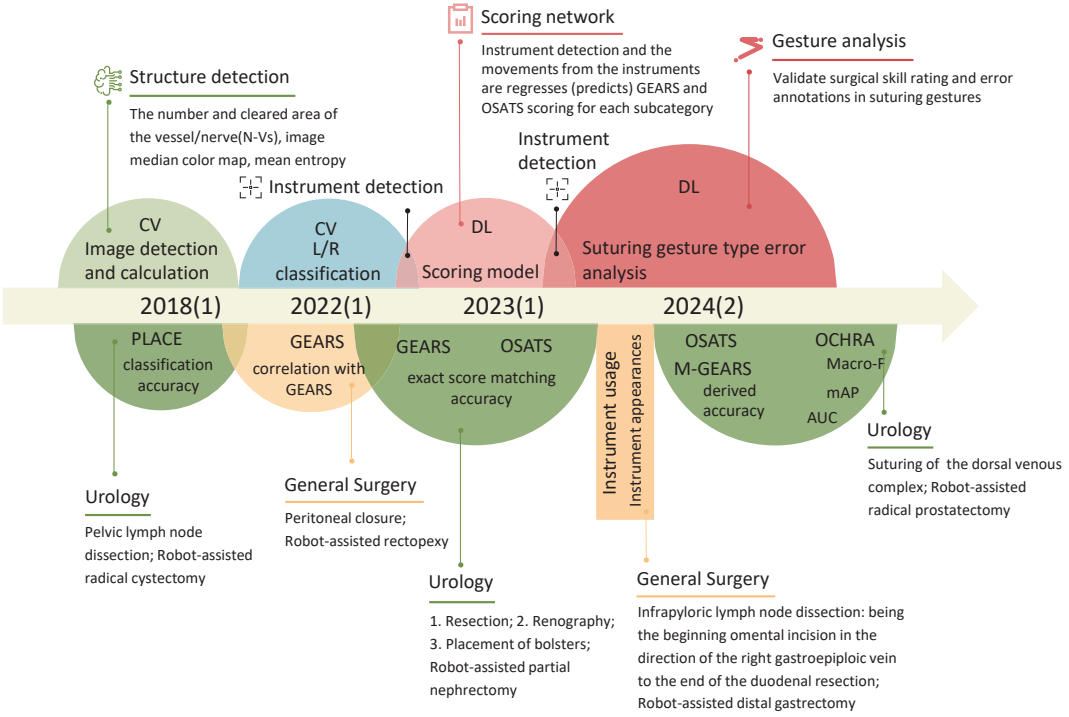


Fig. 2b. Key attributes of studies included in the review. The figure summarizes: (1) spectrum of AI techniques employed and corresponding surgical specialties; (2) criteria and metrics used for performance evaluation; and (3) the chronological distribution of publications.

Table 3
Summary of PROBAST + AI assessment of the included studies.

	study	Part A: Development					Part B: Evaluation					Applicability	
study	type	D1	D2	D3	D4	O	D1	D2	D3	D4	O	Development	Evaluation
Baghdadi 2018	D and E	●	▲	▲	●	●	●	▲	▲	●	●	▲	▲
Yang 2022	E	NA	NA	NA	NA	NA	▲	▲	▲	●	●	NA	▲
Wang 2023	D and E	●	▲	▲	●	●	●	▲	▲	●	●	▲	▲
Strong 2024	D and E	●	▲	●	●	●	●	▲	●	●	●	●	●
Sirajudeen 2024	D and E	●	▲	▲	●	●	●	▲	▲	●	●	▲	▲

● Low risk/Low concern; ▲ High risk/High concern.
D1, Participants and data sources; D2, Predictors; D3, Outcome; D4, Analysis; NA, not applicable. For model development, domain and overall ratings indicate concern regarding model development quality (low or high concern). For model evaluation, domain and overall ratings indicate risk of bias (low or high risk). Applicability ratings indicate concern regarding applicability (low or high concern).

decision-making that is challenging to evaluate in artificial environments. Anatomical variations among patients also complicate real-world surgical procedures. Furthermore, some patients present a unique set of anatomical features, which can deviate from the standard models used in dry labs [30,31]. Therefore, studies concentrating on AI-guided evaluations conducted on real patients provide a more comprehensive and accurate reflection of a surgeon's true capabilities than those studies involving surgeon evaluations in simulators or dry labs. This is because the former group of studies captures the nuances and challenges intrinsic to real-world clinical practice. This discrepancy has important implications for surgical training and credentialing, where reliable and context-aware evaluation of technical performance is essential. Rating scales such as OSATS are often used in assessing performance, and GEARS or M-GEARS (modified vision) are also used in assessing robotic surgical performance. In our review, the PLACE rating was also used. All of these scales are ordinal scales. This level of measurement of a variable can significantly limit the analysis of data, and AI modelling typically requires an interval for outcomes to be properly categorized, ranked, and evenly spaced so that variables can be ordered, and the distance between them can be measured and computational models developed. It is typically known that ordinal scales may set up models for uncertainty in AI algorithm development, especially if data is multisource in nature or prone to bias [32]. Also, methods like OSATS are subjective, and often even professionals do not fully agree with each other [33]. These concepts are hard for AI to measure. It is challenging to predict something from an entire video because there is a lot of complex information, both over time and space, which is difficult for AI to handle to date.

AI has increasingly been used to extract and analyse data from surgical videos to assess technical surgical skills [34]. The relationships between some subfields of AI, along with their associated techniques and

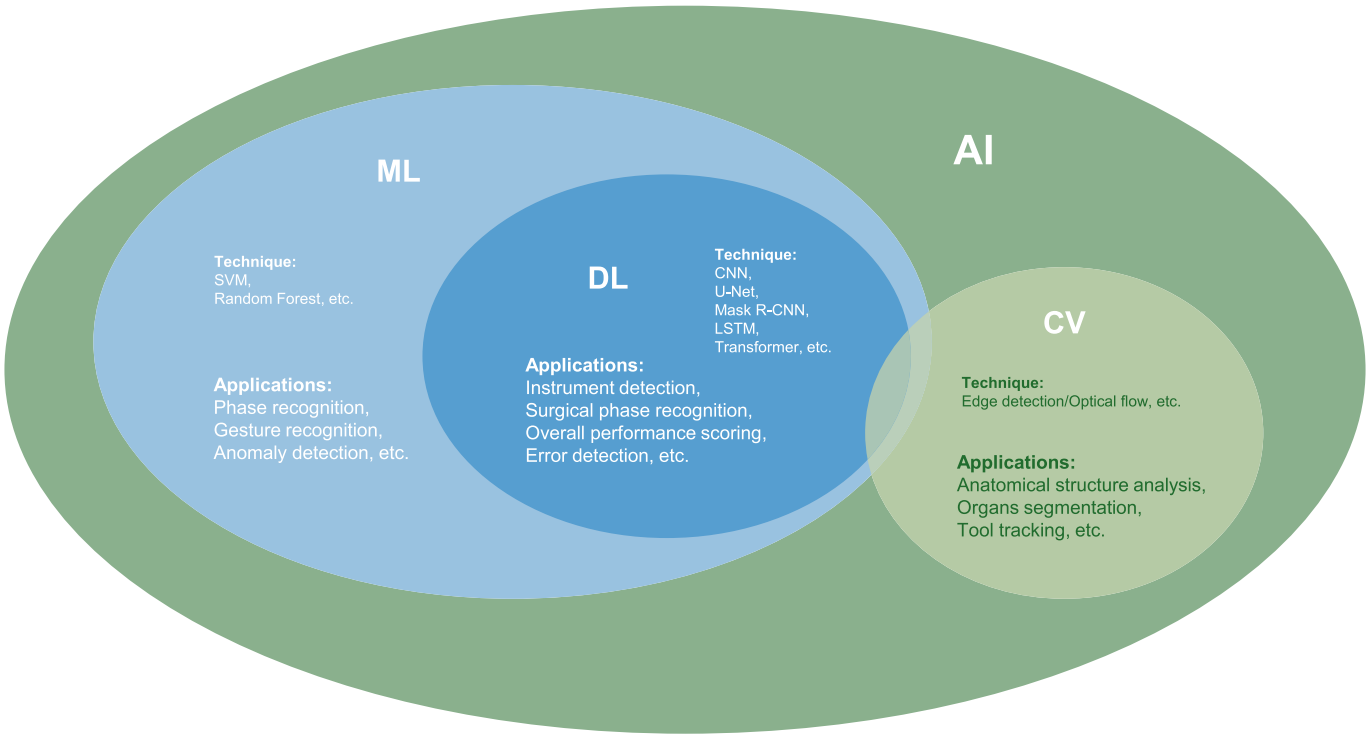


Fig. 3. Conceptual overview of CV, ML, and DL in surgical video analysis. ML is a key approach that enables systems to learn patterns from data and improve performance without explicit programming. A specialized subset of ML, called DL, uses neural networks to automatically detect detailed features. In surgical videos, this could include aspects such as the precision and smoothness of instrument movements. CV is a subfield of AI dedicated to understanding and processing visual data, using ML and DL techniques to perform a variety of tasks.

applications, are illustrated in Fig. 3. Of note, with the development of AI, some applications, such as phase recognition, have gradually transitioned from ML to DL, leveraging techniques like CNN and Transformers. Together, these AI methods support automated and detailed evaluation of surgical performance. While these approaches show promise for comprehensive skill assessment, preliminary evidence suggests that, at present, DL-based instrument recognition or CV techniques are limited to evaluating only certain aspects of robotic technical skills. This limitation may partially account for the inability of these models to achieve optimal accuracy relative to conventional assessment methods. Different AI techniques can work together to create a more complete system for assessing surgical skills. Foundation models, which are large AI models trained on diverse data types and tasks, can support this by combining information such as instrument use, movement patterns, and the flow of the procedure to provide objective evaluations of technical performance. Indeed, data relevant for performing surgery—such as assessment of pre- and post-operative imaging, instrument trajectories, time, intraoperative imaged include detection of landmarks, and times of instrument collisions—are very different in nature and all of importance. A single ML algorithm is unlikely to succeed in making the best use of these pluriform data and interpret correctly. Pluriform large datasets need to be processed by combined AI models, a foundational model, to produce a detailed and reliable analysis of the surgeon's performance and support the assessment of a range of surgical tasks. Such a model acts as a versatile base that can be adapted and fine-tuned for specific applications, rather than training separate models from scratch for each task on insufficient datasets.

Three of the five included studies incorporated expert evaluated technical skill instruments, including OSATS, GEARS, or M-GEARS, as assessment targets or reference standards [24,25,27]. However, their use differed substantially across studies. Wang et al. directly predicted GEARS and OSATS scores, Yang et al. examined associations between AI-derived instrument motion metrics and individual GEARS domains, and Sirajudeen et al. evaluated both technical skill estimation and surgical error detection. Baghdadi et al. used PLACE, a procedure-specific assessment instrument, whereas Strong et al. quantified instrument use patterns and compared these measures between groups defined by procedural experience rather than against a validated technical skill rating scale. These approaches represent different facets of technical performance rather than interchangeable measures of a single construct. In addition to the use of these two fundamental evaluations (OSATS and GEARS), it could be better to incorporate more specific evaluation criteria tailored to different surgical procedures, such as cholecystectomy assessment and surgical evaluation (CASE), and the score for partial nephrectomy (SPaN) [35,36]. Furthermore, the difficulty of surgery may vary depending on the unique condition of each patient. Therefore, assessing the surgical technique of a surgeon solely on the basis of the number of instruments used and their trajectories may be inaccurate without considering other pertinent criteria. Future research should prioritise the development of standardised evaluation frameworks.

Some limitations must be accounted for in this systematic review. The main limitation is that, as a burgeoning technology, only a few related research articles have concentrated on the use of AI models to assess the technical skills on real patients to be discussed. Other limitations include the lack of generalisability and methodological limitations of the five included studies. These studies are not multicentre trials, as the datasets were collected from only one hospital. Additionally, some procedures were performed by skilled surgeons, resulting in high scores and limiting the ability of the study to evaluate trainees. Moreover, in AI models detected on two-dimensional views, errors may have occurred due to the absence of depth information and instrument detection, potentially affecting the accuracy of the results. Datasets could incorporate surgical videos from various levels of surgeons across multiple medical centres, along with additional accumulated errors, to enhance the comprehensiveness of AI models. Robust external validation and the inclusion of clinically meaningful performance metrics will be essential to enable safe and effective integration of AI into surgical

practice. AI models could also integrate patient complications into the scoring system and combine algorithms for various types of subject detection for a more holistic assessment.

Notably, the first authors of all five included studies were affiliated with engineering, computer science, data science, medical physics, or other technical disciplines (Table S4). While the quality of the articles and AI model architectures has demonstrated measurable improvements from 2018 to 2025, the evaluation criteria for robotic technical skill assessment remain underdeveloped without increased multidisciplinary collaboration, particularly from practicing surgeons, to establish clinically validated benchmarks (Fig. 2b). Therefore, our review is also a call-to-action encouraging surgical specialists and AI engineers to cross their disciplines and have a more formalised interdisciplinary collaboration during the developmental phase of AI-integrated skill evaluation systems. This initiative should prioritise: (1) surgeon-led development of evidence-based assessment frameworks through Delphi consensus processes; (2) engineer-driven implementation of validated algorithmic architectures with objective performance analytics. Sustained cross-disciplinary engagement will accelerate translational applications of emerging AI technologies in the surgical environment.

In summary, AI holds significant potential for automating the technical skill assessment of surgeons during robot-assisted surgery. However, few studies have focused on automated AI models for assessing technical skills during robot-assisted surgery with real patients, reflecting the early stage of this research field. Consequently, the development of robust AI models with standardised and clinically relevant performance metrics, supported by rigorous validation, is required before clinical application. Such advances are essential to facilitate the safe and reliable translation of AI assessment from experimental development to real operative settings.

Role of the funding source

There was no funding source for this study.

Category

Review.

Previous communication

The paper is not based on a previous communication to a society or meeting.

Data sharing

Extracted data from the current study will be made available on request to the corresponding author.

CRediT authorship contribution statement

Jiani Tang: Writing – review & editing, Writing – original draft, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ana Manzano Rodriguez:** Writing – review & editing, Writing – original draft. **Erdem Uğur:** Validation, Investigation, Data curation. **Marlies P. Schijven:** Writing – review & editing, Supervision, Resources, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We would like to acknowledge the help of Hans Ket, Medical Information Specialist, University Library, Vrije Universiteit Amsterdam with the literature search.

During the preparation of this work, the author(s) used ChatGPT (OpenAI) to assist with improving grammar, sentence structure, and language clarity, as well as with refining code used for data visualization. The author(s) reviewed and edited all AI-assisted output as needed and take full responsibility for the content of the published article.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2026.106776.

References

[1] S. Zamboni, B. Foerster, M. Abufaraj, T. Seisen, M. Roupret, P. Colin, A. De la Taille, C. Di Bona, B. Peyronnet, K. Bensalah, R. Herout, M.P. Wirth, V. Novotny, F. Soria, P. Chlosta, A. Antonelli, C. Simeone, P. Baumeister, A. Mattei, F. Montorsi, G. Simone, M. Gallucci, K. Matsumoto, P.I. Karakiewicz, A. Briganti, E. Xylinas, S. F. Shariat, M. Marco Moschini, Upper Tract Urothelial Carcinoma Collaboration Group, European Association of Urology - Young Academic Urologists (EAU-YAU), Urothelial carcinoma working group, Incidence and survival outcomes in patients with upper urinary tract urothelial carcinoma diagnosed with variant histology and treated with nephroureterectomy, BJU Int. 124 (2019) 738–745.

[2] P. Dell'Oglio, F. Turri, A. Larcher, F. D'Hondt, R. Sanchez-Salas, B. Bochner, J. Palou, R. Weston, A. Hosseini, A.E. Canda, J. Bjerggaard, G. Cacciamani, K. Ø. Olsen, I. Gill, T. Piechaud, W. Artibani, P.J. van Leeuwen, A. Stenzl, J. Kelly, P. Dasgupta, C. Wijburg, J.W. Collins, M. Desai, H.G. van der Poel, F. Montorsi, P. Wiklund, A. Mottrie, ERUS Educational Working Group and the YAU Working Group on Robot-assisted Surgery, Definition of a structured training curriculum for robot-assisted radical cystectomy with intracorporeal ileal conduit in male patients: a delphi consensus study led by the ERUS educational board, Eur. Urol. Focus 8 (2022) 160–164.

[3] S. Kalata, J.R. Thumma, E.C. Norton, J.B. Dimick, K.H. Sheetz, Comparative safety of robotic-assisted vs laparoscopic cholecystectomy, JAMA Surg. 158 (2023) 1303–1310.

[4] T. Leal Ghezzi, O. Campos Corleta, 30 years of robotic surgery, World J. Surg. 40 (2016) 2550–2557.

[5] I.G. Jeong, Y.S. Khandwala, J.H. Kim, D.H. Han, S. Li, Y. Wang, S.L. Chang, B. Chung, Association of robotic-assisted vs laparoscopic radical nephrectomy with perioperative outcomes and health care costs, 2003 to 2015, JAMA 318 (2017) 1561–1568.

[6] M.G. Goldenberg, Evidence that surgical performance predicts clinical outcomes, World J. Urol. 38 (2020) 1595–1597.

[7] Y. Wang, S. Wilder, M. Hijazi, M.D. Myles, M. Mirza, M. Van Til, T. Maatman, K. R. Ghani, B.R. Lane, C.G. Rogers, M.U.S.I. Collaborative, Surgeon skill and perioperative outcomes in robot-assisted partial nephrectomy, JAMA Netw. Open 2024 (2024), https://doi.org/10.1001/jamanetworkopen.2024.21696, Jul 1.

[8] E.B. Wilson, H. Bagshahi, V.D. Woodruff, Overview of general advantages, limitations, and strategies, in: K. Kim (Ed.), Robotics in General Surgery, Springer, New York, 2014, pp. 17–22.

[9] T. Williamson, S.E. Song, Robotic surgery techniques to improve traditional laparoscopy, JSLS (2022), https://doi.org/10.4293/JSLS.2022.00002. Apr-Jun, 2022.

[10] B. Van Cleynenbreugel, B.K. Somani, Future of robotic surgical education: the (R) evolution from training the barber-surgeon to subspecialist robot-surgeon, Eur. Urol. 78 (2020) 717–718.

[11] W.S. Loh, R.A. Howard, B.T. Fry, J.R. Thumma, E.C. Norton, J.B. Dimick, K. H. Sheetz, Learning curves for robotic-assisted ventral hernia repair, JAMA Netw. Open 2024 (2024), https://doi.org/10.1001/jamanetworkopen.2024.48521, Dec 2.

[12] B.K. Beaulieu-Jones, W. Yuan, G.A. Brat, A.L. Beam, G. Weber, M. Ruffin, I. S. Kohane, Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? NPJ Digit Med 4 (2021) 62.

[13] J.S. Brownstein, B. Rader, C.M. Astley, H. Tian, Advances in artificial intelligence for infectious-disease surveillance, N. Engl. J. Med. 388 (2023) 1597–1607.

[14] A.J. Hung, J. Chen, I.S. Gill, Automated performance metrics and machine learning algorithms to measure surgeon performance and anticipate clinical outcomes in robotic surgery, JAMA Surg. 153 (2018) 770–771.

[15] I. Funke, S.T. Mees, J. Weitz, S. Speidel, Video-based surgical skill assessment using 3D convolutional neural networks, Int. J. Comput. Assist. Radiol. Surg. 14 (2019) 1217–1225.

[16] R. Pedrett, P. Mascagni, G. Beldi, N. Padoy, J.L. Lavanchy, Technical skill assessment in minimally invasive surgery using artificial intelligence: a systematic review, Surg. Endosc. 37 (2023) 7412–7424.

[17] A.B. Chen, S. Liang, J.H. Nguyen, Y. Liu, A.J. Hung, Machine learning analyses of automated performance metrics during granular sub-stitch phases predict surgeon experience, Surgery 169 (2021) 1245–1249.

[18] J.E. Knudsen, U. Ghaffar, R. Ma, A.J. Hung, Clinical applications of artificial intelligence in robotic surgery, J. Robot. Surg. 18 (2024) 102.
[19] A. Hamilton, The future of artificial intelligence in surgery, Cureus 2024 (2024), https://doi.org/10.7759/cureus.63699, Jul 2.
[20] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, L.A. McGuinness, L.A. Stewart, J. Thomas, A.C. Tricco, V.A. Welch, P. Whiting, D. Moher, The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, BMJ 372 (2021) n71.
[21] M. Ouzzani, H. Hammady, Z. Fedorowicz, A. Elmagarmid, Rayyan-a web and mobile app for systematic reviews, Syst. Rev. 5 (2016) 210.
[22] K.G.M. Moons, J.A.A. Damen, T. Kaul, L. Hooft, et al., PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods, BMJ 24 (388) (2025 Mar) e082505.
[23] A. Baghdadi, A.A. Hussein, Y. Ahmed, L.A. Cavuoto, K.A. Guru, A computer vision technique for automated assessment of surgical performance using surgeons' console-feed videos, Int. J. Comput. Assist. Radiol. Surg. 2018 (2018), https://doi.org/10.1007/s11548-018-1881-9, Nov 20.
[24] J.H. Yang, E.D. Goodman, A.J. Dawes, J.V. Gahagan, M.M. Esquivel, C.A. Liebert, C. Kin, S. Yeung, B.H. Gurland, Using AI and computer vision to analyze technical proficiency in robotic surgery, Surg. Endosc. 2022 (2022), https://doi.org/10.1007/s00464-022-09781-y, Dec 19.
[25] Y. Wang, Z. Wu, J. Dai, T.N. Morgan, A. Garbens, H. Kominsky, J. Gahan, E. C. Larson, Evaluating robotic-assisted partial nephrectomy surgeons with fully convolutional segmentation and multi-task attention networks, J. Robot. Surg. 17 (2023) 2323–2330.
[26] J.S. Strong, T. Furube, M. Takeuchi, H. Kawakubo, Y. Maeda, S. Matsuda, K. Fukuda, R. Nakamura, Y. Kitagawa, Evaluating surgical expertise with AI-based automated instrument recognition for robotic distal gastrectomy, Ann Gastroenterol Surg 8 (2024) 611–619.
[27] N. Sirajudeen, M. Boal, D. Anastasiou, J. Xu, D. Stoyanov, J. Kelly, J.W. Collins, A. Sridhar, E. Mazomenos, N.K. Francis, Deep learning prediction of error and skill in robotic prostatectomy suturing, Surg. Endosc. 38 (2024) 7663–7671.
[28] M. Deptula, J. Zielinski, A. Wardowska, M. Pikula, Wound healing complications in oncological patients: perspectives for cellular therapy, Postepy Dermatol Alergol 36 (2019) 139–146.
[29] L. Say, D. Chou, A. Gemmill, Ö. Tunçalp, A.B. Moller, J. Daniels, A.M. Gülmezoglu, M. Temmerman, L. Alkema, Global causes of maternal death: a WHO systematic analysis, Lancet Glob. Health 2014 (2014), https://doi.org/10.1016/S2214-109X(14)70227-X, May 5.
[30] E.V. Prado Neto, A. Petroianu, Anatomical variations of portal venous system: importance in surgical clinic, Arq. Bras. Cir. Dig. 2022 (2022), https://doi.org/10.1590/0102-6720202210002e1666, Jun 24.
[31] K. Manolakos, K. Zygogiannis, C. Mousa, T. Demesticha, V. Protogerou, T. Troupis, Anatomical variations of the iliohypogastric nerve: a systematic review of the literature, Cureus 2022 (2022), https://doi.org/10.7759/cureus.24910, May 11.
[32] R.R. Yager, Ordinal scale based uncertainty models for AI, Inf. Fusion 64 (2020) 92–98.
[33] R.C. Caskey, L. Owei, R. Rao, E.W. Riddle, A.D. Brooks, D.T. Dempsey, J.B. Morris, C.J. Neylan, N.N. Williams, K.R. Dumon, Integration of hands-on team training into existing curriculum improves both technical and nontechnical skills in laparoscopic cholecystectomy, J. Surg. Educ. 74 (2017) 915–920.
[34] D.A. Hashimoto, G. Rosman, D. Rus, O.R. Meireles, Artificial Intelligence in Surgery: promises and Perils, Ann. Surg. 268 (2018) 70–76.
[35] A.A. Hussein, K.J. Sexton, P.R. May, M.V. Meng, A. Hosseini, D.D. Eun, S. Daneshmand, B.H. Bochner, J.O. Peabody, R. Abaza, E.C. Skinner, R. E. Hautmann, K.A. Guru, Development and validation of surgical training tool: cystectomy assessment and surgical evaluation (CASE) for robot-assisted radical cystectomy for men, Surg. Endosc. 32 (2018) 4458–4464.
[36] T. Klatte, V. Ficarra, C. Gratzke, J. Kaouk, A. Kutikov, V. Macchi, A. Mottrie, F. Porpiglia, J. Porter, C.G. Rogers, P. Russo, R.H. Thompson, R.G. Uzzo, C. G. Wood, I.S. Gill, A literature review of renal surgical anatomy and surgical strategies for partial nephrectomy, Eur. Urol. 68 (2015) 980–992.