Contents lists available at ScienceDirect

The American Journal of Surgery

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

Exploring the roles of artificial intelligence in surgical education: A scoping review

Elif Bilgic[a], Andrew Gorgy[a], Alison Yang[a], Michelle Cwintal[a], Hamed Ranjbar[a], Kalin Kahla[a], Dheeksha Reddy[a], Kexin Li[a], Helin Ozturk[a], Eric Zimmermann[a], Andrea Quaiattini[b,c], Samira Abbasgholizadeh-Rahimi[d,e,f,g], Dan Poenaru[c,h], Jason M. Harley[a,c,i,j,*]

[a] Department of Surgery, McGill University, Montreal, Quebec, Canada
[b] Schulich Library of Physical Sciences, Life Sciences, and Engineering, McGill University, Canada
[c] Institute of Health Sciences Education, McGill University, Montreal, Quebec, Canada
[d] Department of Family Medicine, McGill University, Montreal, Quebec, Canada
[e] Department of Electrical and Computer Engineering, McGill University, Montreal, Canada
[f] Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada
[g] Mila Quebec AI Institute, Montreal, Canada
[h] Department of Pediatric Surgery, McGill University, Canada
[i] Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada
[j] Steinberg Centre for Simulation and Interactive Learning, McGill University, Montreal, Quebec, Canada

ARTICLE INFO

Keywords:
Artificial intelligence
Surgical education
Assessment
Training

ABSTRACT

Background: Technology-enhanced teaching and learning, including Artificial Intelligence (AI) applications, has started to evolve in surgical education. Hence, the purpose of this scoping review is to explore the current and future roles of AI in surgical education.

Methods: Nine bibliographic databases were searched from January 2010 to January 2021. Full-text articles were included if they focused on AI in surgical education.

Results: Out of 14,008 unique sources of evidence, 93 were included. Out of 93, 84 were conducted in the simulation setting, and 89 targeted technical skills. Fifty-six studies focused on skills assessment/classification, and 36 used multiple AI techniques. Also, increasing sample size, having balanced data, and using AI to provide feedback were major future directions mentioned by authors.

Conclusions: AI can help optimize the education of trainees and our results can help educators and researchers identify areas that need further investigation.

1. Introduction

Artificial intelligence (AI) is growing rapidly in healthcare. A *Nature* study found that between 2010 and 2020, publications regarding AI in the fields of healthcare had grown exponentially.[1] A U.S. study evaluated the perceptions of medical students regarding which specialties would be impacted the most by AI.[2] Over 75% of the participants believed that AI would have a major impact on medicine during their lifetimes. The majority of students (over 65%) believed that radiology and surgery would be impacted by AI sooner and to a greater extent than other specialties. The researchers concluded that AI should be integrated into medical education curricula to dispel any misperceptions towards AI, as well as enable students to take advantage of the technology.

Surgical training is a long process that comes with many challenges: trainees are faced with work-hour and operating room (OR) restrictions, cost pressures, and policies intended to reduce patient waiting times. Adding to these challenges, surgical training is transitioning from a traditional time-based framework to a competency-based model. For example, in Canada, the transition is underway and by 2022 all residency programs will have transitioned completely. Competency-based medical education (CBME) is an outcomes-based approach for the design, implementation, assessment and evaluation of a surgical

[*] Corresponding author. Department of Surgery, Faculty of Medicine, McGill University, Montreal General Hospital, 1650 Cedar Ave, R1.112, Montreal, QC, H3G 1A4, Canada.
E-mail address: jason.harley@mcgill.ca (J.M. Harley).

https://doi.org/10.1016/j.amjsurg.2021.11.023
Received 25 August 2021; Received in revised form 19 November 2021; Accepted 22 November 2021
Available online 30 November 2021
0002-9610/© 2021 Elsevier Inc. All rights reserved.

program, using a pre-defined framework of competencies. This approach is composed of iterative, formative and summative assessments. In order to properly assess surgical trainees' competencies, assessments with established validity evidence for specific purposes and contexts within surgical education are needed.

The introduction of AI in surgical training has the potential to ease the current transition of surgical training to a competency-based model.[3] Within surgical education, AI can be defined as: '*An intelligent system/program that acts to fulfill or* support *the fulfillment of educational tasks traditionally performed exclusively by Surgical Educators, through making decisions in a manner similar to educators and providing customized adaptation, including performance assessment and feedback, to surgical trainees*'.[4] Hence, AI platforms can provide automated feedback and assessment, allowing trainees to practice on their own time and without the need for the physical presence of an expert. Although the use of AI technologies has been rapidly increasing in the medical field, it is still relatively new in the context of surgical education.[1] As a result, the extent of information available regarding AI's application in surgical education is not clear. Therefore, the purpose of this scoping review is to explore the current and future roles of AI in surgical education.

2. Objectives

2.1. Determine educational roles of AI

To explore the educational roles of AI in surgical education, focusing on (a) purpose of utilization (development of educational content, augmentation of learning and knowledge development, and assessment of student performance), (b) types of AI applications, (c) skill domains targeted (e.g. knowledge, problem solving, communication etc.), and (d) educational platforms targeted (e.g. multiple-choice questions, video-games, virtual simulations [including virtual patients], virtual or augmented reality simulators etc.).

2.2. Determine educational theories utilized

To identify, categorize, and evaluate the educational theories or frameworks used to guide the implementation of AI in surgical education, taking into account the (a) frequency, (b) type, and (c) relevancy of the used theories/frameworks.

2.3. Determine expectations for AI

To determine students' expectations for AI in regards to (a) contribution to success, (b) AI-directed emotions, (c) perceived usefulness of existing and future AI-enhanced systems, and (d) future-oriented emotions.

2.4. Determine training strategies for usage of AI technologies

To determine how students and teachers are trained in order to use and learn with AI-enhanced educational materials, taking into account the (a) frequency, (b) type of training, and (c) sufficiency of training.

2.5. Determine challenges and future roles of AI

To explore challenges and anticipated future roles of AI in surgical education.

3. Materials and methods

A scoping review was conducted according to Arksey and O'Malley's framework and is reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA). A scoping review was selected as this methodology seeks to identify and map out existing evidence on a particular topic, and specifically, to examine how research is conducted on a certain topic, and to identify and analyze knowledge gaps. They also allow authors to determine the volume of literature and evidence that exists on a given topic, and give an overview of its focus. Scoping reviews are also useful when the topic consists of emerging evidence. AI and its use in surgical education are emerging fields, with new evidence and applications being published at a rapid rate. However, the topic is largely unmapped, and knowledge gaps remain significant. Thus, a scoping review was the best method to follow to achieve the broadly defined objective of this project: to determine what role(s) AI currently plays in surgical education, as well as where it may evolve.

3.1. Identifying relevant studies and study selection

The literature search was conducted by a health sciences librarian (AQ) to identify all relevant literature examining the use of AI for surgical education in undergraduate and postgraduate medical education. The search period covered from 2010 to January 17, 2021, and searches used Medical Subject Headings (MeSH) and keywords as appropriate. The Ovid Medline search strategy was developed in consultation with the research team, including an expert computer scientist/engineer (SR) who focuses on artificial intelligence, and a surgeon (DP) with expertise in surgical education. After the initial Medline strategy was finalized, it was adapted to the remaining databases: Ovid Embase, CINAHL, PsychINFO, ERIC, Scopus, Web of Science, Compendex, and Inspec (see Appendix A for all search strategies). It is important to note that since some of the relevant articles might be published in conference proceedings and IEEE journals, our librarian performed the search to capture these papers as well. The results were compiled, and duplicates removed in EndNote X9.3 (EndNote, Clarivate Analytics). Additional full-text sources were identified through backward citation searching and added to the full-text review.

3.2. Study selection

3.2.1. Inclusion and exclusion criteria

Studies were included if they (1) focused on medical students, surgical interns, and surgical residents, (2) focused on surgery, (3) focused on surgical education, (4) used AI, (5) were empirical (including surveys, interviews, etc.), and (6) were in English. We excluded studies that: (1) only included nurses, dentistry students, paramedics, physicians/surgeons, patients, fellows, or totally unrelated populations, (2) only focused on medicine (outside of surgery), (3) only included medical education (unrelated to surgery), allied health professions education, undergraduate education (other than medicine), high school education, (4) did not include AI, (5) were reviews (scoping, systematic, etc.), conference abstracts, commentaries, editorials, magazine articles, one-off diagrams, posters, supplementary articles, dissertations, letters to the editors, and (6) were in another language than English.

3.2.2. Screening and selection

A total of 10 reviewers were involved in the title/abstract and full-text screening processes. Pilot screenings were performed for both the title/abstract and full text protocol to mitigate any conflicts. With each subsequent pilot screening, the protocol was updated. For the screenings, the actual screening process did not begin until the results of the pilot screening had a 75% agreement. For the title/abstract screening, 6 pairs were formed, and articles were divided amongst the pairs. The screening was conducted using Rayyan software, and each pair completed the screening with the 'blind' setting on.

Once the title/abstract screening process was completed, the results were transferred onto an Excel spreadsheet. Any source that one or both members of the pair decided to include was included for full-text analysis. Reviewers were again divided into pairs, and each pair was assigned a specific number of full-text articles to review independently. Reasons for exclusion were classified as "Conference Abstract" (if the article was published as a conference abstract), “Full-Text Availability” (if no full-text was available), “Language” (if the full-text article was not in English), “Article Type” (if the article was non-empirical), “Population” (if the article did not focus on the population of interest) or “Constructs” (if the article did not focus on surgery, surgical education, or AI).

3.3. Charting the data

Using Microsoft Excel, a data extraction chart was developed by JH and EB, in collaboration with the aforementioned experts. Data charting included 3 reviewers (EB, AG, AY), and similar to the screening process, pilot chartings were performed prior to the official charting. These reviewers were chosen from the 10 reviewers involved in previous aspects of this review based on their experience in conducting reviews, and their performance in the title/abstract and full-text screenings, as they made the least number of mistakes. Additionally, their performances were reviewed during the pilot data charting, to ensure that they can accurately extract data. Reviewers were assigned specific full-text papers. Two of the reviewers formed a pair and completed the charting independently using the same articles, and the 3rd reviewer performed data charting independently for their unique set of articles. After the pair completed charting, they combined their results to form a single data charting sheet for their assigned papers. The 2 reviewers extracted data in pairs based on a previous determination during the pilot phase that their data complemented each other, so they could extract quality data when in a pair.

3.4. Collating, summarizing, and reporting the results

Data were synthesized in 3 steps: Analysis of the data (numerical and thematic analysis), reporting of findings, and discussion of results and implications.[5]

3.4.1. Analysis of data

Numerical analysis: The distribution of studies, explaining study characteristics and our review objectives 1–4 were highlighted (e.g. number of studies, study design, year of publication, study population, targeted skills and setting, and AI techniques).

Thematic analysis: Texts relevant to challenges and future roles of AI (objective 5) in the discussion sections of the articles were extracted, and an inductive thematic analysis was conducted following Braun and Clarke’s framework. Coding was conducted by one research team member in an iterative fashion, and a codebook was developed with specific themes, sub-themes, and definitions. Frequency counts were tabulated using the codebook to describe the current challenges and future roles of AI that authors of reviewed articles mentioned in their discussion sections.

3.4.2. Reporting the findings

All results are reported in multiple tables, for both the numerical and thematic analysis.

3.5. Quality assessment

The methodological quality of the included studies was assessed by 2 raters (EB, AG) using the Medical Education Research Study Quality instrument (MERSQI).[6] Two pilot quality assessments were completed by the 2 raters to ensure rater consistency (reached an inter-rater agreement of 0.9) prior to the assessment of all studies. After the pilot, each rater was assigned a unique set of studies for assessment. Data is provided as mean (standard deviation).

4. Results

Our scoping review yielded 14,008 unique sources of evidence. Among them, 13,771 were excluded during the title/abstract screening, while 18 hand-searched full-text sources were added, resulting in 262 records for full-text review. After full-text analysis, 93 were included for data charting and synthesis (Fig. 1).[7–99]

4.1. Paper characteristics

Of the 93 papers, 74 were published after 2014, and 44 were published in computer science/engineering journals and 41 in medical/surgical journals. Forty-nine papers mentioned that the study received funding, 24 from government, 18 from private, and 7 from both. Regarding the country of the 1st author, the majority were from the USA (39), followed by Australia (8), Canada (7), and France (7). More information can be found in Table 1.

4.2. Study design

Sixty (65%) studies were cross-sectional, 75 used primary data whereas 15 used secondary data, 86 were single-group studies where all participants completed the same tasks, and 76 focused on assessment without any training intervention. Additionally, 9 studies were multicentre, and the rest were single-centre or not specified. Finally, 80 studies used private datasets, 10 used public datasets, and 3 used both. More information can be found in Table 2.

4.3. Objective on educational roles of AI

4.3.1. Setting, targeted skills, and participants

Among the 93 studies, 84 were conducted in a simulation setting and 5 in an operating room (details in Table 3). Of the 84 simulation studies, the majority of the simulations included benchtop[52] and screen-based Virtual Reality.[20] Among all 93 studies, 89 targeted technical skills, 3 non-technical skills, and 1 medical knowledge. Specifically, 58 studies focused on laparoscopic and robotic skills, while 19 included four or more skills (such as laparoscopic and robotic skills, in addition to other skills). Regarding the participants, 29 included residents, 23 inexperienced/novices/beginners/intermediates/trainees (participant’s exact level is unclear due to author wording), 12 medical students, and 19 a combination of residents and medical students (Table 4).

4.3.2. AI systems

Among the AI modalities, 19 used neural networks (NN), including artificial NN (ANN) and convolutional NN (CNN), 8 used support vector machines (SVM), 4 linear discriminant analysis (LDA), 4 nearest neighbor (NNe), and 36 used multiple techniques (e.g. one study using NNe, LDA, and naïve bayes (NB)) (Table 5). The majority of the studies used AI for assessment, including skill classification.[56] Nine used it to develop and/or provide feedback, and 19 used it for multiple reasons (e.g. task recognition and task duration estimation, and skill classification and feedback) (Table 6).

Regarding performance metrics, which are data captured to measure performance, the majority of studies used a combination of metrics, including rating scales (such as Objective Structured Assessment of Technical Skills (OSATS) or modified-OSATS, and Global Evaluative Assessment of Robotic Skills (GEARS)), as well as automated computer measurements, motion metrics, kinematic metrics, physiological metrics, and ocular metrics (e.g. eye gaze, blink rate, fixation rate, pupil metric, vergence). Additionally, 19 studies used automated computer measurements alone (metrics collected automatically by the simulator, and providing both individual and cumulative score; e.g. time, position and angles of instruments, forces applied on specific structures, volume of any removed tissue etc.), and 16 used motion metrics alone (data collected using external hardware such as devices; e.g. total distance moved, number of movements, total time, and average velocity). Also, among the studies using rating scales,[32] 23 used experts as raters and 3 used both experts and crowd-sourced workers. More information can be

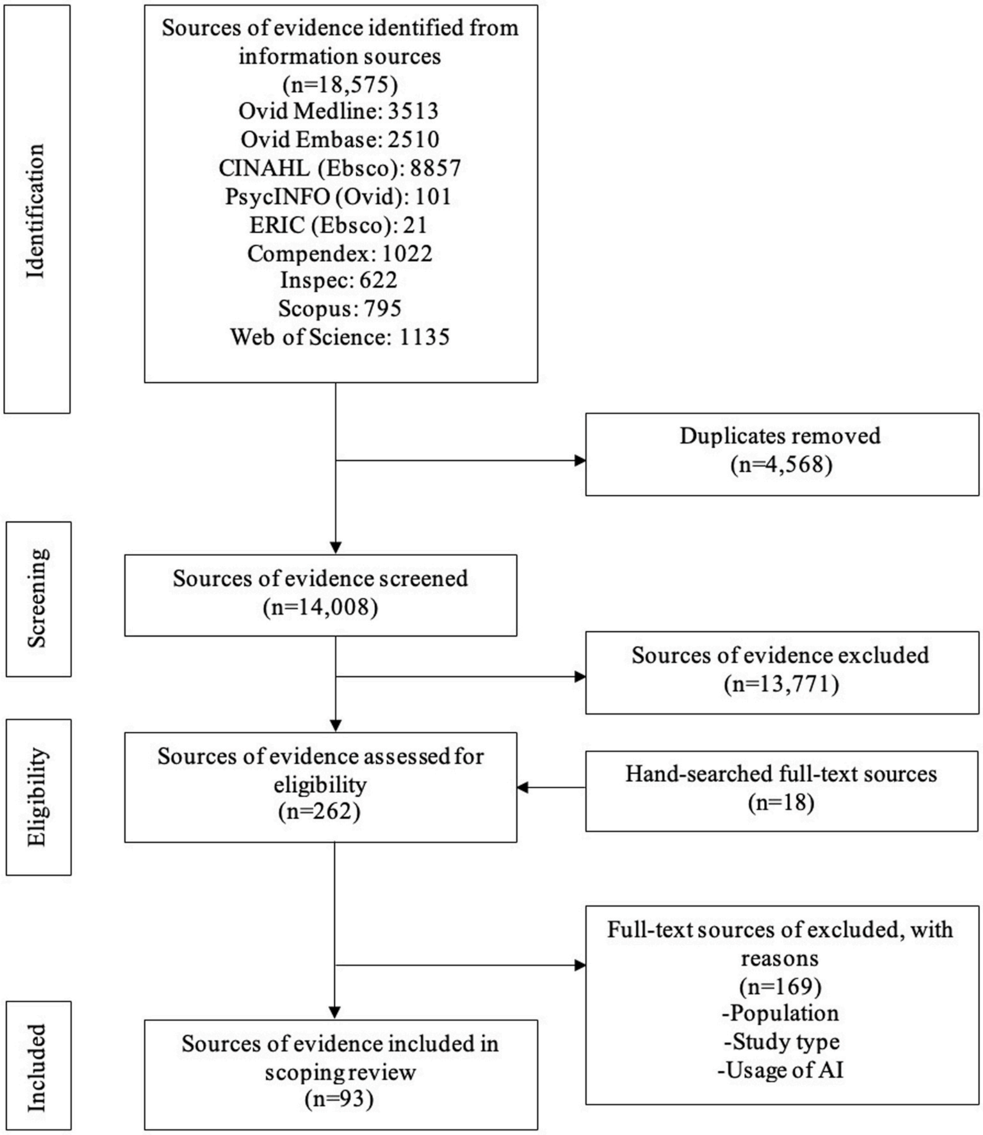
Identification
Sources of evidence identified from information sources (n=18,575)
Ovid Medline: 3513
Ovid Embase: 2510
CINAHL (Ebsco): 8857
PsycINFO (Ovid): 101
ERIC (Ebsco): 21
Compendex: 1022
Inspec: 622
Scopus: 795
Web of Science: 1135
Duplicates removed (n=4,568)
Screening
Sources of evidence screened (n=14,008)
Sources of evidence excluded (n=13,771)
Eligibility
Sources of evidence assessed for eligibility (n=262)
Hand-searched full-text sources (n=18)
Full-text sources of excluded, with reasons (n=169)
-Population
-Study type
-Usage of AI
Included
Sources of evidence included in scoping review (n=93)

Fig. 1. PRISMA-Scr diagram.

found in Table 7.

4.3.3. Simulation training

Seventeen studies conducted simulation training, whereby only 6 of the 17 studies implemented AI-enhanced simulations as a part of the training (other studies gathered participant data for further analysis using AI techniques for purposes such as assessment, skill classification, and learning curve prediction). Out of the 6, 3 were randomized trials, and 3 were group comparisons. The structure of the training sessions varied extensively between the studies. However, all of them had a group that completed the training with an AI-enhanced system, and at least one other group that completed it without AI or with a different AI system. Of the 3 randomized trials, all showed that AI-enhanced simulation training is effective in teaching surgical skills.

4.4. Objectives on educational theories, expectations for AI, and training to use AI technologies

Most of the studies did not provide information regarding the three objectives focusing on educational theories or frameworks, students' expectations from AI, and training to use AI-enhanced technologies. Regarding educational theories, one study briefly mentioned deliberate practice (trainees have repetitive practice based on specific learning objectives and close monitoring with continuous assessment and feedback[100]), transfer of learning (transfer learning from simulation to clinical setting[100]), and constructivism theory (trainees use their prior knowledge and experiences to build on[100]) in their introduction. How these concepts guided the development of their AI-enhanced educational tools was not clear, however. Concerning students' expectations for AI, one simulation training study assessed perceived quality of the AI-enhanced simulation training by asking students about their impressions of the feedback generated from 2 AI-systems using a 5 point Likert scale. Questions included usefulness, clarity, accuracy, timeliness, and extent of feedback. Results showed that students rated feedback generated from both AI techniques as high.

4.5. Objective on challenges and future roles

Appendix B provides a definition of each theme and examples. For current challenges and limitations of developing/implementing AI-enhanced educational technologies specified by the authors, 5 themes emerged: 1) Study design, 2) Data, 3) Feedback, 4) Skills Investigated,

Table 1
Paper characteristics (N = 93).

Year of publication	N
2015–January 2021	74
2010–2014	19
1st author corresponding	
Yes	51
No	22
Not Specified	20
Journal types	
Technical	44
Medical/Surgical	41
Education	8
Country of 1st author	
USA	39
Australia	8
Canada	7
France	7
Greece	4
Spain	3
Germany	3
Italy	3
UK	3
China	2
Mexico	2
Netherlands	2
Singapore	1
Colombia	1
Hungary	1
India	1
Iran	1
Pakistan	1
Sweden	1
Taiwan	1
Turkey	1
NS	1
Funding	
Private (include societies, universities etc)	18
Government	24
Both	7
Quality of published studies	
MERSQI scores (mean (Standard Deviation))*	11.2 (1.36)

*MERSQI Medical Education Research Study Quality instrument.

and 5) Simulator. Table 8 provides the details of each theme, sub-theme, and frequency counts. Overall, the theme 'data' was given the most elaboration by the authors, where 15 studies discussed having a limited sample size, and 7 discussed having unbalanced data (imbalance between skill levels) as limiting factors. Additionally, 13 studies discussed limitations in the metrics/features used, and 6 studies limitations in the AI techniques used.

Regarding future directions of developing/implementing AI-enhanced educational technologies specified by authors, 12 themes emerged: 1) Study design, 2) Data, 3) Feedback, 4) Simulation training, 5) Skills to be targeted, 6) Simulators representing clinical setting, 7) Role of other factors (e.g. emotions, motivations), 8) Legal/ethical considerations, 9) Implementation in other settings (e.g. clinical, simulation), 10) Decision-making for trainee readiness for operating room, 11) Classifying levels based on post-graduate year (PGY) level, case volume, and 11) Testing ergonomics and comfort of new tools. Table 9 provides the details of each theme, sub-theme, and frequency counts. Overall, the themes 'data' and 'feedback' were given the most elaboration by the authors. For data, 40 studies mentioned improving and/or testing other metrics, and 23 studies mentioned the same for AI techniques. Moreover, for feedback, 36 studies discussed the structure of feedback, with 26 discussing improving the feedback content and quality, and 9 discussing how to provide the feedback (e.g. verbal, visual, haptic). Additionally, 16 studies mentioned the need for further experimental studies and 14 mentioned the need for larger sample size in each study, with an increase in the number of skills (and complex skills) targeted.

Table 2
Study design* (N = 93).

Study Design	N
Cross sectional	60
Repeated measure	30
Both (2 databases)	1
Not Specified	2
Primary use of data	75
Secondary use of data	15
Both	3
Single group	86
RCT	3
Group comparison	3
Survey	0
Interview/focus group	0
Not Specified	1
Assessment only (no training)	76
Intervention: Pre/post assessment **	7
Intervention: Post-assessment only**	5
Intervention: Continuous assessment**	4
Not Specified	1

*- Cross sectional (trainees assessed at 1 time point only) OR Repeated measure (e.g., learning/skill is assessed multiple times).
- Secondary use of data (data that was collected previously, outside of the study) OR Primary use of data (data collected as a part of the study).
- Randomized Controlled Trial (RCT, trainees are randomly assigned to 2 or more groups (1 group is usually the control group, whereby those participants do what they would normally do [most common control]) and experimental groups go through their assigned educational interventions such as a particular workshop, simulation training, a lecture etc) OR Group comparison (trainees are assigned to 2 or more groups (not randomized) to complete their assigned educational interventions (can include a control group)) OR Single group (all of the trainees complete the same intervention) OR Survey (survey is distributed to individuals to gain knowledge about their experiences and opinions) OR Interview/focus group (interviews are done with 1 or more people, or focus groups are done to gain a deeper understanding of people's experiences).
**If RCT, Group comparison, or Single group.
- Pre/post assessment (trainees are assessed before and after an educational intervention) OR Post assessment only (trainees are only assessed after the intervention) OR Assessment itself was the only intervention (e.g. participants are assessed in a simulator) OR Assessment and training were done together (e.g. trainees were assessed continuously throughout the training (no pre/post or post only assessment).

4.6. *Study quality*

The mean MERSQI score across all papers was 11.2 (1.36) out of 16.5 (response rate item was not applicable for the published studies). Most studies lost points due to their study design and sampling, as they were often single-centre and single-group, and cross-sectional or pre/post-test studies.

5. Discussion

The results of our review reveal that over the past 5 years, there has been a major growth in the number of papers published within the field of AI in surgical education, suggesting that there is an increase in interest amongst the surgical education community regarding the role of AI in enhancing training and assessment of surgical trainees. During and post-pandemic, AI-enhanced educational technologies could play an even greater role in automated training and assessment of trainees.

The majority of the articles identified utilized AI for assessment or differentiating skill levels, using a variety of metrics such as instrument and hand motion, eye tracking, time/error, and rating scales.

Table 3
Setting and Targeted skills (N = 93).

Setting	N
Simulation	84
Operating Room	5
Both	3
Elsewhere (classroom)	1
Type of Simulation	
Benchtop	52
Screen-based VR	20
Screen-based VS	4
Multiple	4
Cadaver	2
Other (hand washing)	1
Immersive VR	0
Not Specified	4
Type of Skill	
Technical	89
Non-technical	3
Medical knowledge	1
Tasks being assessed*	
Laparoscopy	33
Laparoscopic placing/grasping/transferring of objects (include peg transfer)	15
Laparoscopic needle passing/suturing/knot tying	10
Laparoscopic Pattern cutting	5
Laparoscopic coordinated pulling	1
Laparoscopic hand-eye coordination	1
Laparoscopic matching object orientation/location	1**
Robotics	25
Robotic needle passing/suturing/knot tying	16
Robotic placing/transferring of objects (include peg transfer)	4***
Robotic ring and rail (endowrist manipulation)	3
Robotic Urethrovesical Anastomosis	1
Robotic Lymph Node Dissection	1
Other skills	36
Open needle passing/suturing/knot tying	11
Complete/Cortical mastoidectomy	7
Posterior tympanotomy	3
Subpial brain tumor resection	2
Neurosurgery placing/grasping/transferring of objects (include peg transfer)	1
Cochleostomy	1
Endoscopic *trans*-nasal sinus surgery	1
Closure step of lumbar disk herniation surgery	1
Ligature	1
Minimally invasive prostate cryo- surgery	1
Femoral bone drilling	1
Benign tumor resection from tibia	1
Tracheoesophageal fistula repair (minimally invasive pediatrics)	1
Tissue dissection in nasal septoplasty	1
Removal of L3 lamina (Left L3 hemilaminectomy)	1
Hand washing	1
In-plane and out-of-plane needle insertions	1
Multiple (4 or more tasks) ****	19
Not Specified	2

*Some studies assessed more than 1, so won't add up to 93.
**"localized sphere: the goal is to match the center of the sphere with the center of the image camera and the sphere arrow pointing up".
***1 study used both Peg Board and Pick & Place tasks.
****Open and laparoscopic procedures; Sensorimotor, Subclavian central line placement, Bowel repair, Urinary catheterization, and Laparoscopic ventral hernia tasks; 9 simple endoscopic endonasal tasks (touching 9 endonasal structures); Multiple catheterization tasks; Forward peg transfer, retroflexion peg transfer, snaring, clipping, puncturing, cannulation; Mitral valve repair techniques (annuloplasty, triangular leaflet resection, neo-chordae implantation); laparoscopic tasks of varying degrees of difficulty (simulation and Transperitoneal laparoscopic renal surgery (OR); Robotic suturing, knot tying, needle passing, and peg transfer tasks; Robotic Camera Targeting, Peg Board, Ring and Rail, Sponge Suturing, Dots and Needles, and Tubes; Robotics Manipulation, Suturing, Transection, and Dissection tasks; Laparoscopic Navigation Instrument, Coordination, Grasping, Lifting and Grasping, and Cutting; Laparoscopic Pipe cleaner, Rubber band, Beads, and Circle' Laparoscopic peg transfer, precision cutting, intracorporeal knot tying, ligating loop; Robotic motion training tasks; 6 scenarios of removing a series of virtual brain tumors; Robotics tasks focused on the skills camera control, EndoWrist instrument manipulation, clutching, needle control, and needle driving; Skin pad incision, Tissues dissection, Interrupted stitch, Running suture, Knot Tying exercise.

Table 4
Study participants (N = 93).

Specialty	N
MIS (Laparoscopy)	28
MIS (Robotics)	25
Open Surgery (General skills)	10
Ear nose throat	10
Neurosurgery	4
Orthopaedic/Neurosurgery	3
MIS/Open Surgery (General skills)	3
Orthopaedic Surgery	2
MIS (Flexible Endoscopy)	1
MIS/Open Surgery (Pediatric)	1
Endovascular Surgery	1
General Surgery	1
Trauma Surgery	1
Others (hand washing)	1
Not Specified	2
Population*	
Residents	29**
Inexperienced/Novices/Beginner/Intermediate/Trainees	23
Combination (medical student, resident)	19
Medical students	12
Trainees	6
Non-experts	1
Medical interns	1
Non-surgeons	1
Surgeons with broad range of skill	1
Not Specified	3

MIS Minimally Invasive Surgery.
*Based on classifications by the authors; each category could include experts and other health professionals outside of the inclusion criteria of the review.
**In study by Germain Forestier, 2018, they use multiple databases, and 1 database includes resident, other includes novices/intermediates (counted separately).

Additionally, the majority of articles used multiple AI techniques to determine the one(s) that perform the best in addressing their specific purposes. This suggests that there is no single way to develop and implement AI-enhanced educational technologies, and that there are multiple factors authors should consider before designing their study and determining metrics and AI techniques to focus on. Some factors that could be taken into account include whether an AI system is used in a clinical versus a simulation setting, being used for simpler or more complex skills, or for assessment or feedback purposes. Additionally, the type of simulation used (e.g. benchtop, virtual simulation, cadaver) and resources available for collecting and labeling performance data could be important to consider as well. Also, each AI technique has its own benefits and limitations, and this was discussed by authors as a consideration when deciding which techniques to investigate for the defined problem/task.

The majority of the studies was conducted in a simulation setting, likely because researchers can control various factors that might affect the performance of the AI systems (e.g. lighting conditions, standardization of the tasks being performed, and availability of fixed cameras and equipment for recording data).[75] On the other hand, in the clinical setting, each case is unique, the operative field is not standardized, and it is more difficult to capture data (e.g. the camera is not fixed and there are sterilization and distancing requirements limiting data capture). However, as a future direction, some authors mentioned that research should investigate the role of AI for training and assessment in the clinical setting. Currently, assessments in the clinical setting are done using forms, where attending surgeons are usually the raters. However, this approach has some limitations when it comes to the need of surgeons to take time off from their busy schedules to complete assessments for each resident multiple times. Therefore, trainee progress can be

Table 5
Specifics of the AI algorithms.

AI Algorithm	Number of studies, with specifics of the algorithms if available
Neural Networks (NN)	Total: 19 6, Artificial NN 1, Deep convolutional NN (CNN) with dense optical flow 1, Deep CNN 1, CNN 1, CNN (YOLOv3 and Faster R- CNN) 1, Clip- Based CNN and long-term dynamic model (LTDM) using Markov random fields (MRFs) 1, Generative adversarial network (GAN) (cross-domain conditional GAN and baseline GAN) 1, SATR-DL (CNN-Gated Recurrent Unit (GRU) network) 1, Mask Region-based CNN and GAN 1, 3D ConvNet 1, Recurrent NN (simple RNNs, long short-term memory, gated recurrent units, and mixed history RNNs) 1, Fully CNN 1, ConvNet, Recursive NN 1, radial basis function neural networks (RBF) and multilayer percepron neural networks (MPN)
Support Vector Machine (SVM)	8
Discriminant Analysis (DA)	Total: 4 1, linear DA (LDA), reduced and simple 3, LDA
Nearest Neighbor (NNe)	4
Regression models	Total: 3 1, 2 models 1, 6 models 1
Decision Trees (DT) (include random forest (RF) and boosted trees)	2
k-means, fuzzy c-means	2
Hidden Markov Models (HMM)	Total: 2 1 1, Relative and 2-class HMM
Dynamic Time Warping (DTW)	1
Naïve Bayes (NB)	1
Fuzzy logic	1
Intelligent Tutoring System (details not specified)	1
String-matching algorithm	1
Optical flow algorithms	1
Template-matching approach	1
Hidden-state Conditional Random Fields (HCRF)	1
Multiple	Total: 36 1, Emerging patterns (EP) classifier, HMM, NB, DT 1, RF, EP, feedback using rule-based methods (details not specified) 1, NN, rule-based methods (NS) 1, RF, NB, DT, RF-NNe 1, NNe, LDA, NB 1, MM, K-means clustering 1, SVM, DTW 1, Logistic regression and NN 1, SVM, NNe 1, Decision forest, NN, Boosted DT, DTW (using both regression and classification ML) 1, SVM, Gaussian Mixture Multivariate Autoregressive Models (GMMAR) 1, kNNe, Ascendant Hierarchical Clustering (AHC) using DTW 1, LDA, nonlinear NN 2, LDA, SVM, adaptive network-based fuzzy inference system (ANFIS) 1, ANN, RF, K-Star 2, Linear Dynamical Systems (LDS), HMM, SAX-VSM algorithm based classification (Symbolic Aggregate approXimation (SAX), Vector Space Model (VSM)) 1, support vector regression, elastic net regression, regression trees, K nearest neighbors, RF (using both regression and classification ML) 1, k-means clustering, SVM, markov chains 1, HMM, RF, Bayesian approach 1, SVM, LDS, GMMAR 1, K-Nearest Neighbors, Parzen Window, SVM, Fuzzy K-NNe 1, RF, nearest neighbor 1, K-NNe, NB, DA, SVM 1, HMM, fuzzy logic 1, ZeroR and J48 ML algorithms, DT 1, SVM, deep Convolutional neural network–Long Short-term Memory (CNN-LSTM) neural network 1, NNe, SVR 1, SVM, fuzzy C-means clustering 1, HMM, DTW 1, logistic regression, NB, SVM 1, NB, SVM 1, SVM, HMM 1, SVM, K-NNe, LDA, NB, DT 1, DT, fuzzy rule-based assessment, zero-rule regression, linear regression, SVM regression, NNe regression, random regression forest
Not Specified/Unclear	5

Table 6
Purpose of using Artificial Intelligence (N = 93).

AI used for	N
Assessment (including skill classification)	56
Feedback	9
Multiple*	19
Others**	9

*Skill classification, Outline weights of metrics; Task recognition, Task duration estimation; Skill classification, Task recognition; Maneuver and Gesture recognition; Skill classification, Feedback; Gesture recognition, Skill classification; Metric selection, Skill classification; Skill classification, Development of parameters; Stage prediction, Feedback; Trajectory segmentation, Skill classification; Skill classification, Phase recognition; Pattern recognition, Skill classification; Gesture and Skill classification; Skill classification, Performance measure identification; Simulation development, Student perception; Skill classification, Development of new measure.

**Improvement classification, Generation of Stereopairs, Instructive video retrieval, Learning curve prediction, Needle recognition, Score prediction, Trust classification, Video alignment, Workload classification.

tracked in real-time and automatically, and with AI, students can be provided with performance feedback and guidance in a more timely and sustainable way.

Furthermore, most authors focused on laparoscopic and robotic surgeries. Since these types of procedures are minimally invasive, and surgeons perform the surgeries using a monitor, it is easy to record and store videos, making it feasible for the application of AI. Additionally, videos from these operations directly show the area of the operation, whereas in open procedures, the view of the performance might be partially or fully blocked with the surgeon's head or body. Nonetheless, focusing on the OR, with minimally invasive procedures, the camera is still frequently moving, so even though it is easier to record, the camera is not fixed. Also, for open procedures, even though the camera is

Table 7
Metrics and raters (N = 93).

Metrics	N
Automated computer measurements*	19
Motion	16
Surgical gestures	4
Kinematic	3
Rating scales**	2
End-product assessment	2
Force	1
Multiple***	38
Others****	4
Not Specified	4
Raters***	**N = 32**
Experts (fellows and surgeons)	23
Experts and crowd-workers	3
Crowd-workers	1
Rater experienced in using the tool	1
Experts and non-experts	1
Not specified	3

*These are metrics collected automatically by the simulator, and providing both individual and cumulative score; e.g. time, position and angles of instruments, forces applied on specific structures, volume of any removed tissue etc.
**Video commentary assessment tool, Welling Scale.
***Automated computer measurements, kinematics, motion, rating scales (Objective Structured Assessment of Technical Skills (OSATS) or modified-OSATS, Global Evaluative Assessment of Robotic Skills (GEARS) and others), time/error, heart rate, electromyography, galvanic skin response, electroencephalogram (EEG), cortical hemodynamic data, stylistic behaviours, eye metrics, written feedback, number of trials.
****Task phases; Surgical process; Surgical steps; Written exam scores.
*****For studies using rating scales (alone or in combination).

somewhat fixed, the exact location is often adjusted multiple times by the surgical team to optimize lighting, and the lighting constantly changes. Therefore, overall, within clinical setting (minimally invasive or open), there are limitations that the simulation setting does not have, making simulation settings easier to capture data and apply AI techniques.

Moreover, it is important to note that in many papers, the participants were not clearly defined. Especially in papers published in technical journals, the authors used words such as “novice”, “intermediate”, and “trainee” without specifying the level of the trainees (medical students, residents, and year of medical school or residency), which made the full-text screening process difficult during our review, as the reviewers were unsure whether to include or exclude an article, based on the participants.

Our review had 5 objectives. However, the 3 objectives related to theory or framework used to guide the development and usage of AI, trainee emotions and perceived usefulness of AI, and training trainees to use AI-enhanced educational technologies were only briefly addressed by 2 out of 93 studies. Theory is important in designing a research question, selecting and interpreting data, and understanding what we observe as relationships between concepts.[101] However, there was only one study that briefly mentioned constructivism theory and other concepts, without specifying how these guided the development of their AI-enhanced educational technology. This could be due to the fact that even though authors might have based their technologies on various educational theories and frameworks (e.g. deliberate practice and expertise, social learning theory, constructivism etc), due to the word limit and nature of journals within the domain, there might not have been space to provide this type of information. Nonetheless, even with these limitations, researchers could still either briefly discuss the usage of specific theories, or at least cite the theories in their introduction or methods. Moreover, only one simulation training study assessed trainees’ perceived quality of the AI-enhanced simulation training. Many studies collected performance data, and applied AI techniques later, so trainees were not exposed to AI-enhanced training and assessment technologies, which could explain this finding. Nonetheless, when new technologies are being developed and implemented, there is a need for buy-in from all stakeholders, including trainees, educators, hospital/faculty leadership etc. Therefore, evidence is required to demonstrate that these technologies are effective and efficient at assessing and improving skill level, which requires a rigorous research phase where high quality evidence is collected. Hence, how theory is used to guide the development of these technologies as educational tools, and constructs related to how trainees perceive these educational tools are essential to help determine the best way to implement these technologies in the training programs.

Additionally, perceptions of trainees regarding AI enhanced education could also be impacted by their knowledge of AI. When searching the literature, we identified a study by researchers at the University of

Table 8
Frequency counts of themes and sub-themes related to current challenges of developing/implementing AI-enhanced educational technologies, specified by authors.

			Sufficient/ Appropriate	Limited/ Inappropriate
1 Study design	1.0 Number of institutions		1	1
	1.1 Interinstitutional similarities		1	0
	1.2 Timing of studies		0	1
	1.3 Participants	1.3.1 Skill level diversity	0	4
2 Data	2.0 Sample size		0	15
	2.1 Types of data	2.1.1 Metrics/features used	11	13
		2.1.2 ‘Ground truth’ measures for sample labeling/video annotation	3	7
	2.2 Balance of data		0	7
	2.3 Capturing data	2.4.1 Equipment requirements	3	5
		2.4.2 Feasibility of receiving data from raters (e.g. crowd-workers, expert, peer etc)	1	2
		2.4.3 Number of raters	0	1
	2.4 Analysis of data	2.5.1 AI algorithms used	21	7
		2.5.2 Equipment	0	1
		2.5.3 Transparency of analysis	2	1
		2.5.4 Time efficiency and accuracy	2	4
		2.5.5 Using public datasets	0	2
	2.5 Access to data		0	1
3 Feedback	3.0 Providing real-time/instant feedback	3.0.1 Skill gain	1	1
	3.1 Content of feedback		3	3
4 Skills investigated	4.0 Complexity of skills		0	2
	4.1 Number of skills/tasks		0	5
5 Simulator	5.0 Representing clinical setting		0	8

Table 9
Frequency count of themes and sub-themes regarding AI-related future directions specified by authors.

			Mentioned by authors
1 Study design	1.0 Multiple institutions		2
	1.1 Conducting further studies	1.1.1 Group comparisons (including RCT)	4
		1.1.2 Details not specified	10
	1.2 Participants	1.2.1 Similar skill levels	2
		1.2.2 Different levels of experience	7
	1.3 Comparison to clinical outcomes and other measures		10
2 Data	2.0 Large sample size		14
	2.1 Type of data	2.1.1 Improving/Testing other or combination of metrics	40
		2.1.2 Getting better ground-truth measures/annotations	5
	2.2 Balanced numbers of participants per groups		3
	2.3 Techniques to generate synthetic data		1
	2.4 Capturing data	2.5.1 Using new technologies	2
		2.5.2 Using crowd-workers	2
	2.5 Analysis of data	2.6.1 Improving/Testing AI algorithms	23
		2.6.2 Adaptability of analysis	2
		2.6.3 Time efficiency/accuracy of analysis	1
3 Feedback	3.0 Real-time feedback		6
	3.1 Structure of feedback	3.1.1 Developing/Improving feedback content and quality (e.g. metrics, benchmarks)	26
		3.1.2 Mechanisms of feedback (e.g. verbal, visual, haptic)	9
		Timing of feedback	1
	3.2 Outcome of feedback	3.2.1 Improving skill gain	5
		3.2.2 Improving skill retention	3
		3.2.3 Detecting learning curve	3
		3.2.4 Helping overcome skill decay	2
4 Simulation training	4.0 Longer training time		1
	4.1 Implement in parallel to clinical exposure		1
5 Skills to be targeted			13
	5.0 Higher number of skills/tasks		
	5.1 Higher complexity/whole procedures		8
	5.2 Anatomical variations		4
6 Simulators representing clinical setting			4
7 Role of other factors (e.g. emotions, motivations)			4
8 Legal/ethical considerations			1
9 Implementation in other settings (e.g. clinical, simulation)			18
10 Certification and recredentialing			1
11 Classifying levels based on PGY level, case volume			1
12 Testing ergonomics and comfort of new tools			1

Toronto who surveyed medical students about their knowledge of AI, perceptions on the role of AI, and preferences surrounding how to integrate AI competencies in their curriculum.[102] The researchers found that most students had a general understanding of AI, but would need their medical schools to adequately prepare them to use these tools as a part of their clinical practice. Their findings wasn't specific to implementing AI to be used for educational purposes; rather it was about usage of AI in the clinical setting, and teaching students fundamental knowledge and skills to use these technologies in the clinical setting. While this type of usage is not the focus of this review, we believe that if trainees have a fundamental knowledge of AI, and a practical understanding of how to use these technologies, we could assume that students would know the limitations and benefits of AI-enhanced educational tools, and use/interpret their performance results accordingly.

Also, based on our thematic analysis, for both challenges and future roles, a major theme identified was 'data', where authors discussed the limitations of having small sample sizes and unbalanced data, all potentially effecting the performance of the AI-enhanced technologies. Also, authors called for studies with larger sample size, balanced (e.g. similar numbers of expert and novice data) and high-quality data, and more experimental studies. This is especially important since the performance of the AI-enhanced technologies are dependent on the quality of the data used for development: if the datasets used lack enough data, do not include diverse data, and do not have properly defined ground truth measures to compare the data to, the performance of the technology will suffer. Additionally, authors suggested improving or testing other metrics, including combinations of metrics, and other AI techniques to address their purposes for developing or using AI-enhanced educational technologies. Specifically for future directions, many authors mentioned the potential benefit of AI technologies providing feedback, with important considerations regarding the content and mechanisms of feedback for skill gain and retention. Ultimately, these results show that even though current data are promising, there is a need to go beyond small studies and explore metrics/techniques that are better able to capture expertise levels and develop high quality AI-enhanced systems to provide automated feedback and assessment.

Even though quality assessments are not mandatory for scoping reviews, we used MERSQI to report the quality of the included articles. In this tool, how much each item is weighed is based on a comprehensive process undertaken by the developers of the tool during the development phase. For example, based on hierarchies of research designs, "single group cross-sectional or single group posttest only" provide a lower level of evidence of effectiveness and receive lower points compared to "nonrandomized, 2 group" or "randomized controlled trial". Due to a similar reasoning, multicentre studies receive higher points compared to single centre studies. In terms of what it means to receive one score versus the other, we were unable to identify any established thresholds regarding high versus moderate versus low scores, but it means that if studies lost points (e.g. for the included studies here, mean was 11.2 out of 16.5), there are some areas within their research quality that could be improved, in order to provide higher levels of evidence. As specified in the results, most studies in our review lost points due to their study design and sampling. Therefore, as identified through the thematic analysis and reflected in the MERSQI scores, future studies within this field could focus on developing experiments considering the current limitations of published studies.

Even though this scoping review allowed us to reach a broad range of empirical studies to address our five objectives, we found that there is a lack of consensus around a definition of AI, including in surgical education, and in our opinion, the ones that were available did not have enough depth to be used in a surgical education context. Hence, it could be possible that not all of the AI in surgical education related articles were retrieved. Retrieving relevant articles is inherently tied to the terminology used by authors of published articles, and in biomedical databases, controlled vocabulary such as MeSH. To help address this gap, our recently developed brief and accessible definition of AI for surgical education was used by our multidisciplinary research team, including our librarian and experts, to develop the search strategy, maximizing the retrieval of relevant articles.[4] A part of our definition is about AI systems making decisions in a manner similar to educators. While researchers continue to study the mechanisms of educator reasoning and how AI systems mimic, or not, those exact processes, it is important to note that our definition reflects similarities between educators and AI systems, in general, including reaching similar decisions, and going through decision-making processes that share common features (e.g., using valid criteria for assessment—whether identified through supervised or discovered through unsupervised learning).

6. Conclusions

In conclusion, AI is relatively new in the surgical field and the benefits, limitations, and applications of AI in surgical education are still not clear, as evident from our results. Given the novelty of using AI to facilitate surgical education, this scoping review provides a first step in providing the scientific and educational community with an overview of what has been done so far and what is needed. Our findings show that currently, the major focus of AI is on performance assessment or skill classification of trainees for technical skills within the simulation setting. Hence, future studies should focus on conducting experimental studies, exploring multiple metrics and AI techniques, using AI for feedback generation, and investigating applications in the operating room, and for non-technical skills.

Funding

This work was supported by The Social Sciences and Humanities Research Council of Canada [ID 611-2020-0157].

Declaration of competing interest

All authors have no relevant conflicts of interest or financial ties to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.amjsurg.2021.11.023.

References

1. Bertalan M, Grg Marton. A short guide for medical professionals in the era of artificial intelligence. *NPJ Digit Med*. 2020;3(1):1–8. https://doi.org/10.1038/s41746-020-00333-z.
2. Park CJ, Yi PH, Siegel E. Medical student perspectives on the impact of artificial intelligence on the practice of medicine. *Curr Probl Diagn Radiol*. 2020. https://doi.org/10.1067/j.cpradiol.2020.06.011.
3. Winkler-Schwartz A, Bissonnette V, Mirchi N, et al. Artificial intelligence in medical education: best practices using machine learning to assess surgical expertise in virtual reality simulation. *J Surg Educ*. 2019;76(6):1681–1690. https://doi.org/10.1016/j.jsurg.2019.05.015.
4. Bilgic E, Gorgy A, Young M, Abbasgholizadeh-Rahimi S, Harley JM. Artificial intelligence in surgical education: considerations for interdisciplinary collaborations. *Surg Innov*. 2021. https://doi.org/10.1177/15533506211059269. In press.
5. Bussieres AE, Al Zoubi F, Stuber K, et al. Evidence-based practice, research utilization, and knowledge translation in chiropractic: a scoping review. *BMC Compl Alternative Med*. 2016;16:216. https://doi.org/10.1186/s12906-016-1175-0.
6. Gorbanev I, Agudelo-Londono S, Gonzalez RA, et al. A systematic review of serious games in medical education: quality of evidence and pedagogical strategy. *Med Educ Online*. 2018;23(1), 1438718. https://doi.org/10.1080/10872981.2018.1438718.
7. Ahmad MA, Mansoor SB, Khan ZA, Aqeel W, Kabir S. *Benchmarking Expert Surgeons' Path for Evaluating a Trainee Surgeon's Performance*. 2013:57–62. https://doi.org/10.1145/2534329.2534345.
8. Ahmidi N, Ishii M, Fichtinger G, Gallia GL, Hager GD. An objective and automated method for assessing surgical skill in endoscopic sinus surgery using eye-tracking and tool-motion data. *Int Forum Allergy Rhinol*. 2012;2(6):507–515. https://doi.org/10.1002/alr.21053.
9. Ahmidi N, Poddar P, Jones JD, et al. Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty. *Int J Comput Assist Radiol Surg*. 2015;10(6):981–991. https://doi.org/10.1007/s11548-015-1194-1.
10. Alonso-Silverio GAP, Pérez-Escamirosa FP, Bruno-Sanchez RlM, et al. Development of a laparoscopic box trainer based on open source hardware and artificial intelligence for objective assessment of surgical psychomotor skills. *Surg Innovat*. 2018;25(4):380–388. https://doi.org/10.1177/1553350618777045.
11. Andreu-Perez J, Leff DR, Shetty K, Darzi A, Yang GZ. Disparity in frontal lobe connectivity on a complex bimanual motor task aids in classification of operator skill level. *Brain Connect*. 2016;6(5):375–388. https://doi.org/10.1089/brain.2015.0350.
12. Anh NX, Nataraja RM, Chauhan S. Towards near real-time assessment of surgical skills: a comparison of feature extraction techniques. *Comput Methods Progr Biomed*. 2020;187:105234. https://doi.org/10.1016/j.cmpb.2019.105234.
13. Azari DP, Frasier LL, Quamme SRP, et al. Modeling surgical technical skill using expert assessment for automated computer rating. *Ann Surg*. 2019;269(3):574–581. https://doi.org/10.1097/SLA.0000000000002478.
14. Baby B, Srivastav VK, Singh R, Suri A, Banerjee S. Neuro-endo-activity-tracker: an automatic activity detection application for Neuro-Endo-Trainer: neuro-Endo-activity-tracker. In: *Presented at: 2016 International Conference on Advances in Computing, Communications and Informatics*. ICACCI); 2016.
15. Baloul MS, Yeh VJ, Mukhtar F, et al. Video commentary & machine learning: tell me what you see, I tell you who you are. *J Surg Educ*. 2020;S1931–7204(20). https://doi.org/10.1016/j.jsurg.2020.09.022, 30372-X.
16. Bencteux V, Saibro G, Shlomovitz E, et al. Automatic task recognition in a flexible endoscopy benchtop trainer with semi-supervised learning. *Int J Comput Assist Radiol Surg*. 2020;15(9):1585–1595. https://doi.org/10.1007/s11548-020-02208-w.
17. Bissonnette V, Mirchi N, Ledwos N, et al. Artificial intelligence distinguishes surgical training levels in a virtual reality spinal task. *J Bone Joint Surg Am*. 2019; 101(23), e127. https://doi.org/10.2106/JBJS.18.01197.
18. Brown JD, CE OB, Leung SC, Dumon KR, Lee DI, Kuchenbecker KJ. Using contact forces and robot arm accelerations to automatically rate surgeon skill at peg transfer. *IEEE Trans Biomed Eng*. 2017;64(9):2263–2275. https://doi.org/10.1109/TBME.2016.2634861.
19. Cavallo F, Sinigaglia S, Megali G, et al. Biomechanics-machine learning system for surgical gesture analysis and development of technologies for minimal access surgery. *Surg Innovat*. 2014;21(5):504–512. https://doi.org/10.1177/1553350613510612.

20. Chen L, Zhang Q, Zhang P, Li B. Multimedia IICo, Expo I. Instructive video retrieval for surgical skill coaching using attribute learning. *Proc - IEEE Int Conf Multimed Expo*. 2015. https://doi.org/10.1109/ICME.2015.7177389.
21. Chmarra MK, Klein S, de Winter JC, Jansen FW, Dankelman J. Objective classification of residents based on their psychomotor laparoscopic skills. *Surg Endosc*. 2010;24(5):1031–1039. https://doi.org/10.1007/s00464-009-0721-y.
22. Civelek T, Vidinli IB. Resection of benign tumor in Tibia with a high speed burr by haptic devices in virtual reality environments. In: *Proc. - IEEE Int. Conf. Multimed. Expo 2018:92-97. Presented at: 22nd World Multi-Conference on Systemics, Cybernetics and Informatics*. WMSCI); 2018.
23. Costantini G, Saggio G, Sbernini L, Lorenzo ND, Paolo FD, Casali D. Surgical skill evaluation by means of a sensory glove and a neural network. In: *Presented at: Proceedings of the International Joint Conference on Computational Intelligence*. vol. 3. 2014. https://doi.org/10.5220/0005030301050110. Rome, Italy.
24. DiPietro R, Ahmidi N, Malpani A, et al. Segmenting and classifying activities in robot-assisted surgery with recurrent neural networks. *Int J Comput Assist Radiol Surg*. 2019;14(11):2005–2020. https://doi.org/10.1007/s11548-019-01953-x.
25. Engelhardt S, Sharan L, Wolf I, et al. Cross-domain conditional generative adversarial networks for stereoscopic hyperrealism in surgical training. *Lect Notes Comput Sci*. 2019;11768 LNCS:155–163. https://doi.org/10.1007/978-3-030-32254-0_18.
26. Ershad M, Rege R, Fey AM. Meaningful assessment of robotic surgical style using the Wisdom of crowds. *Int J Comput Assist Radiol Surg*. 2018;13(7):1037–1048. https://doi.org/10.1007/s11548-018-1738-2.
27. Ershad M, Rege R, Fey AM. Automatic surgical skill rating using stylistic behavior components. *Annu Int Conf IEEE Eng Med Biol Soc*. 2018:1829–1832. https://doi.org/10.1109/EMBC.2018.8512593.
28. Ershad M, Rege R, Majewicz Fey A. Automatic and near real-time stylistic behavior assessment in robotic surgery. *Int J Comput Assist Radiol Surg*. 2019;14(4):635–643. https://doi.org/10.1007/s11548-019-01920-6.
29. Forestier G, Lalys F, Riffaud L, Trelhu B, Jannin P. Classification of surgical processes using dynamic time warping. *J Biomed Inf*. 2012;45(2):255–264. https://doi.org/10.1016/j.jbi.2011.11.002.
30. Forestier G, Petitjean F, Senin P, et al. Surgical motion analysis using discriminative interpretable patterns. *Artif Intell Med*. 2018;91:3–11. https://doi.org/10.1016/j.artmed.2018.08.002.
31. Forestier G, Petitjean F, Senin P, Despinoy F, Jannin P. Th Conference on Artificial Intelligence in Medicine At. Discovering discriminative and interpretable patterns for surgical motion analysis. *Lect Notes Comput Sci*. 2017;10259 LNAI:136–145. https://doi.org/10.1007/978-3-319-59758-4_15.
32. Frischknecht AC, Kasten SJ, Hamstra SJ, et al. The objective assessment of experts' and novices' suturing skills using an image analysis program. *Acad Med*. 2013;88 (2):260–264. https://doi.org/10.1097/ACM.0b013e31827c3411.
33. Funke I, Mees ST, Weitz J, Speidel S. Video-based surgical skill assessment using 3D convolutional neural networks. *Int J Comput Assist Radiol Surg*. 2019;14(7): 1217–1225. https://doi.org/10.1007/s11548-019-01995-1.
34. Gao Y, Kruger U, Intes X, Schwaitzberg S, De S. A machine learning approach to predict surgical learning curves. *Surgery*. 2020;167(2):321–327. https://doi.org/10.1016/j.surg.2019.10.008.
35. Gray RJ, Kahol K, Islam G, Smith M, Chapital A, Ferrara J. High-fidelity, low-cost, automated method to assess laparoscopic skills objectively. *J Surg Educ*. 2012;69 (3):335–339. https://doi.org/10.1016/j.jsurg.2011.10.014.
36. Haidegger T, Nagy M, Lehotsky A, Szilagyi L. Digital imaging for the education of proper surgical hand disinfection. *Med Image Comput Comput Assist Interv*. 2011;(6893):619–626. https://doi.org/10.1007/978-3-642-23626-6_76.
37. Holden MS, Xia S, Lia H, et al. Machine learning methods for automated technical skills assessment with instructional feedback in ultrasound-guided interventions. *Int J Comput Assist Radiol Surg*. 2019;14(11):1993–2003. https://doi.org/10.1007/s11548-019-01977-3.
38. Horeman T, Rodrigues SP, Jansen FW, Dankelman J, van den Dobbelsteen JJ. Force parameters for skills assessment in laparoscopy. *IEEE Trans Haptics*. 2012;5 (4):312–322. https://doi.org/10.1109/TOH.2011.60.
39. Huang FC, Mohamadipanah H, Mussa-Ivaldi FA, Pugh CM. Combining metrics from clinical simulators and sensorimotor tasks can reveal the training background of surgeons. *IEEE Trans Biomed Eng*. 2019;66(9):2576–2584. https://doi.org/10.1109/TBME.2019.2892342.
40. Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller P-A. Evaluating surgical skills from kinematic data using convolutional neural networks. *Med Image Comput Comput Assist Interv – MICCAI*. 2018;2018:214–221 (Chapter 25).
41. Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller PA. Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. *Int J Comput Assist Radiol Surg*. 2019;14(9): 1611–1617. https://doi.org/10.1007/s11548-019-02039-4.
42. Ismail Fawaz H, Forestier G, Weber J, et al. Automatic alignment of surgical videos using kinematic data. *Lect Notes Comput Sci*. 2019;11526 LNAI:104–113. https://doi.org/10.1007/978-3-030-21642-9_14.
43. Jain S, Barsness KA, Argall B. Automated and objective assessment of surgical training: detection of procedural steps on videotaped performances. In: *Presented at: 2015 International Conference on Digital Image Computing: Techniques and Applications*. DICTA); 2015:1–6. https://doi.org/10.1109/DICTA.2015.7371233.
44. Jun SK, Narayanan MS, Agarwal P, et al. Robotic Minimally Invasive Surgical skill assessment based on automated video-analysis motion studies. In: *Presented at: Proceedings of the IEEE RAS and EMBS International Conference on Biomedical Robotics and Biomechatronics*. 2012:25–31. https://doi.org/10.1109/BioRob.2012.6290869.
45. Kerwin T, Wiet G, Stredney D, Shen HW. Automatic scoring of virtual mastoidectomies using expert examples. *Int J Comput Assist Radiol Surg*. 2012;7(1): 1–11. https://doi.org/10.1007/s11548-011-0566-4.
46. Kirby GSJ, Kwasnicki RM, Hargrove C, et al. Wireless body sensor for objective assessment of surgical performance on a standardised FLS task. In: *Presented at: Proceedings of the 9th International Conference on Body Area Networks*. 2014. https://doi.org/10.4108/icst.bodynets.2014.257019. London, United Kingdom.
47. Kowalewski KF, Garrow CR, Schmidt MW, Benner L, Muller-Stich BP, Nickel F. Sensor-based machine learning for workflow detection and as key to detect expert level in laparoscopic suturing and knot-tying. *Surg Endosc*. 2019;33(11): 3732–3740. https://doi.org/10.1007/s00464-019-06667-4.
48. Kumar R, Jog A, Malpani A, et al. Assessing system operation skills in robotic surgery trainees. *Int J Med Robot*. 2012;8(1):118–124. https://doi.org/10.1002/rcs.449.
49. Kumar R, Jog A, Vagvolgyi B, et al. Objective measures for longitudinal assessment of robotic surgery training. *J Thorac Cardiovasc Surg*. 2012;143(3):528–534. https://doi.org/10.1016/j.jtcvs.2011.11.002.
50. Laverde R, Rueda C, Amado L, et al. Artificial neural network for laparoscopic skills classification using motion signals from Apple Watch. *Annu Int Conf IEEE Eng Med Biol Soc*. 2018:5434–5437. https://doi.org/10.1109/EMBC.2018.8513561.
51. Liang H, Shi MY. Surgical skill evaluation model for virtual surgical training. *Appl Mech Mater*. 2010;40–41:812–819. https://doi.org/10.4028/www.scientific.net/AMM.40-41.812.
52. Lin S, Qin F, Bly RA, Moe KS, Hannaford B. Automatic sinus surgery skill assessment based on instrument segmentation and tracking in endoscopic video. *Multiscale Multimodal Med Imag*. 2020:93–100 (Chapter 12).
53. Loukas C, Gazis A, Kanakis MA. Surgical performance analysis and classification based on video annotation of laparoscopic tasks. *J Soc Laparoendosc Surg*. 2020;24 (4). https://doi.org/10.4293/JSLS.2020.00057.
54. Loukas C, Georgiou E. Multivariate autoregressive modeling of hand kinematics for laparoscopic skills assessment of surgical trainees. Article. *IEEE Trans Biomed Eng*. 2011;58(11):3289–3297. https://doi.org/10.1109/TBME.2011.2167324, 6015537.
55. Loukas C, Georgiou E. Performance comparison of various feature detector-descriptors and temporal models for video-based assessment of laparoscopic skills. *Int J Med Robot*. 2016;12(3):387–398. https://doi.org/10.1002/rcs.1702.
56. Malpani A, Vedula SS, Chen CC, Hager GD. A study of crowdsourced segment-level surgical skill assessment using pairwise rankings. *Int J Comput Assist Radiol Surg*. 2015;10(9):1435–1447. https://doi.org/10.1007/s11548-015-1238-6.
57. Mei Q, Chainey J, Asgar-Deen D, Aalto D. Detection of suture needle using deep learning. *J Med Robot Res*. 2020;4. https://doi.org/10.1142/s2424905x19420054, 03n04.
58. Mirchi N, Bissonnette V, Ledwos N, et al. Artificial neural networks to assess virtual reality anterior cervical discectomy performance. *Oper Neurosurg*. 2020;19(1): 65–75. https://doi.org/10.1093/ons/opz359/5674993.
59. Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Maestro RFD. The Virtual Operative Assistant- an explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLoS One*. 2020;15(2). https://doi.org/10.1371/journal.pone.0229596.
60. Monserrat C, Lucas A, Hernandez-Orallo J, Ruperez MJ. Automatic supervision of gestures to guide novice surgeons during training. *Surg Endosc*. 2014;28(4): 1360–1370. https://doi.org/10.1007/s00464-013-3285-9.
61. Nguyen XA, Ljuhar D, Pacilli M, Nataraja RM, Chauhan S. Surgical skill levels: classification and analysis using deep neural network model and motion signals. *Comput Methods Progr Biomed*. 2019;177:1–8. https://doi.org/10.1016/j.cmpb.2019.05.008.
62. Oquendo YA, Riddle EW, Hiller D, Blinman TA, Kuchenbecker KJ. Automatically rating trainee skill at a pediatric laparoscopic suturing task. *Surg Endosc*. 2018;32 (4):1840–1857. https://doi.org/10.1007/s00464-017-5873-6.
63. Oropesa I, Sanchez-Gonzaez P, Chmarra MK, et al. Supervised classification of psychomotor competence in minimally invasive surgery based on instruments motion analysis. *Surg Endosc*. 2014;28(2):657–670. https://doi.org/10.1007/s00464-013-3226-7.
64. Oropesa I, Sánchez-González P, Sánchez-Margallo JA, García-Novoa J, Sánchez-Margallo FM, Gómez EJEVA. Endoscopic video analysis of the surgical scene for the assessment of MIS psychomotor skills. In: *Presented at: XIII Mediterranean Conference on Medical and Biological Engineering and Computing*. vol. 2014. 2013: 52–56 (Chapter 13).
65. Oussi N, Loukas C, Kjellin A, et al. Video analysis in basic skills training: a way to expand the value and use of BlackBox training? *Surg Endosc*. 2018;32(1):87–95. https://doi.org/10.1007/s00464-017-5641-7.
66. Peng W, Xing Y, Liu R, Li J, Zhang Z. An automatic skill evaluation framework for robotic surgery training. *Int J Med Robot*. 2019;15(1), e1964. https://doi.org/10.1002/rcs.1964.
67. Perez-Escamirosa F, Alarcon-Paredes A, Alonso-Silverio GA, et al. Objective classification of psychomotor laparoscopic skills of surgeons based on three different approaches. *Int J Comput Assist Radiol Surg*. 2020;15(1):27–40. https://doi.org/10.1007/s11548-019-02073-2.
68. Rafii-Tari H, Payne CJ, Bicknell C, et al. Objective assessment of endovascular navigation skills with force sensing. *Ann Biomed Eng*. 2017;45(5):1315–1327. https://doi.org/10.1007/s10439-017-1791-y.
69. Richstone L, Schwartz MJ, Seideman C, Cadeddu J, Marshall S, Kavoussi LR. Eye metrics as an objective assessment of surgical skill. *Ann Surg*. 2010;252(1): 177–182. https://doi.org/10.1097/SLA.0b013e3181e464fb.

70. Riojas M, Feng C, Hamilton A, Rozenblit J. Knowledge elicitation for performance assessment in a computerized surgical training system. *Appl Soft Comput J*. 2011;11 (4):3697–3708. https://doi.org/10.1016/j.asoc.2011.01.041.

71. Rojas-Muñoz E, Couperus K, Wachs JP. The AI-Medic: an artificial intelligent mentor for trauma surgery. *Comput Methods Biomech Biomed Eng Imag Vis*. 2020: 1–9. https://doi.org/10.1080/21681163.2020.1835548.

72. Saffarzadeh M, Cho S, Ohu I, Zihni A, Awad M. Recurrence quantification analysis for surgical motions in minimally invasive surgery. *Int J Biomed Eng Technol*. 2016; 21:159. https://doi.org/10.1504/IJBET.2016.077181.

73. Saggio G, Santosuosso GL, Cavallo P, et al. Gesture recognition and classification for surgical skill assessment. In: *Presented at: 2011 IEEE International Symposium on Medical Measurements and Applicationsl*. 2011:662–666. https://doi.org/10.1109/MeMeA.2011.5966681.

74. Sehrawat A, Keelan R, Shimada K, Wilfong DM, McCormick JT, Rabin Y. Simulation-based cryosurgery intelligent tutoring system prototype. *Technol Cancer Res Treat*. 2016;15(2):396–407. https://doi.org/10.1177/1533034615583187.

75. Sgouros NP, Loukas C, Koufi V, Troupis TG, Georgiou E. An automated skills assessment framework for laparoscopic training tasks. *Int J Med Robot*. 2018;14(1). https://doi.org/10.1002/rcs.1853.

76. Shafiei SB, Hussein AA, Muldoon SF, Guru KA. Functional brain states measure mentor-trainee Trust during robot-assisted surgery. *Sci Rep*. 2018;8(1):3667. https://doi.org/10.1038/s41598-018-22025-1.

77. Sharma Y, Plötz T, Hammerld N, et al. Automated surgical OSATS prediction from videos. In: *Presented at: 2014 IEEE 11th International Symposium on Biomedical Imaging*. ISBI); 2014:461–464.

78. Siyar S, Azarnoush H, Rashidi S, et al. Machine learning distinguishes neurosurgical skill levels in a virtual reality tumor resection task. *Med Biol Eng Comput*. 2020;58(6):1357–1367. https://doi.org/10.1007/s11517-020-02155-3.

79. Tan X, Chng C-B, Su Y, Lim K-B, Chui C-K. Robot-Assisted training in laparoscopy using deep reinforcement learning. *IEEE Robot Autom Lett*. 2019;4(2):485–492. https://doi.org/10.1109/lra.2019.2891311.

80. Wang Z, Fey AM. SATR-DL: improving surgical skill assessment and task recognition in robot-assisted surgery with deep neural networks. In: *Presented at: 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. EMBC); 2018:1793–1796. https://doi.org/10.1109/EMBC.2018.8512575.

81. Wang Z, Majewicz Fey A. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *Int J Comput Assist Radiol Surg*. 2018;13(12):1959–1970. https://doi.org/10.1007/s11548-018-1860-1.

82. Watson RA. Use of a machine learning algorithm to classify expertise: analysis of hand motion patterns during a simulated surgical task. *Acad Med*. 2014;89(8): 1163–1167. https://doi.org/10.1097/ACM.0000000000000316.

83. Wijewickrema S, Ioannou I, Zhou Y, et al. Region-specific automated feedback in temporal bone surgery simulation. In: *Presented at: 2015 IEEE 28th International Symposium on Computer-Based Medical Systems*. 2015:310–315. https://doi.org/10.1109/CBMS.2015.13.

84. Wijewickrema S, Ma X, Piromchai P, et al. Providing automated real-time technical feedback for virtual reality based surgical training: is the simpler the better? *Lect Notes Comput Sci*. 2018;10947 LNAI:584–598. https://doi.org/10.1007/978-3-319-93843-1_43.

85. Wijewickrema S, Zhou Y, Bailey J, Kennedy G, O'Leary S. Provision of automated step-by-step procedural guidance in virtual reality surgery simulation. In: *Presented at: Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*. 2016:69–72. https://doi.org/10.1145/2993369.2993397.

86. Winkler-Schwartz A, Yilmaz R, Mirchi N, et al. Machine learning identification of surgical and operative factors associated with surgical expertise in virtual reality simulation. *JAMA Netw Open*. 2019;2(8). https://doi.org/10.1001/jamanetworkopen.2019.8363. e198363-e198363.

87. Wu C, Cha J, Sulek J, et al. Sensor-based indicators of performance changes between sessions during robotic surgery training. *Appl Ergon*. 2021;90, 103251. https://doi.org/10.1016/j.apergo.2020.103251.

88. Wu C, Cha J, Sulek J, et al. Eye-tracking metrics predict perceived Workload in robotic surgical skills training. *Hum Factors*. 2020;62(8):1365–1386. https://doi.org/10.1177/0018720819874544.

89. Ying-Ying Y, Boaz S. An expert-led and artificial intelligence system-assisted tutoring course to improve the confidence of Chinese medical interns in suturing and ligature skills: a prospective pilot study. *J Educ Eval Health Prof*. 2019. https://doi.org/10.3352/jeehp.2019.16.7.

90. Yost MJ, Gardner J, Bell RM, et al. Predicting academic performance in surgical training. *J Surg Educ*. 2015;72(3):491–499. https://doi.org/10.1016/j.jsurg.2014.11.013.

91. Zahedi E, Khosravian F, Wang W, Armand M, Dargahi J, Zadeh M. Towards skill transfer via learning-based guidance in human-robot interaction: an application to orthopaedic surgical drilling skill. *J Intell Rob Syst*. 2020;98(3-4):667–678. https://doi.org/10.1007/s10846-019-01082-2.

92. Zhang Q, Li B, th ICoCV, Pattern Recognition CPORUSA. Relative hidden markov models for evaluating motion skill. In: *Presented at: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2013:548–555. https://doi.org/10.1109/CVPR.2013.77.

93. Zhou Y, Bailey J, Ioannou I, Wijewickrema S, Kennedy G, O Leary S. Constructive real time feedback for a temporal bone simulator. *Med Image Comput Comput Assist Interv*. 2013;16(Pt 3):315–322. https://doi.org/10.1007/978-3-642-40760-4_40.

94. Zhou Y, Bailey J, Ioannou I, Wijewickrema S, O'Leary S, Kennedy G. Pattern-based real-time feedback for a temporal bone simulator. In: *Presented at: Proceedings of the 19th ACM Symposium on Virtual Reality Software and Technology*. 2013. https://doi.org/10.1145/2503713.2503728. Singapore.

95. Zhou Y, Ioannou I, Wijewickrema S, Bailey J, Kennedy G, O'Leary S. Automated segmentation of surgical motion for performance analysis and feedback. In: *Presented at: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. 2015:379–386 (Chapter 47).

96. Zia A, Essa I. Automated surgical skill assessment in RMIS training. *Int J Comput Assist Radiol Surg*. 2018;13(5):731–739. https://doi.org/10.1007/s11548-018-1735-5.

97. Zia A, Sharma Y, Bettadapura V, Clements MA, Essa I, Sarin EL. Automated assessment of surgical skills using frequency analysis. *Lecture Notes Comput Sci (including subseries Lect Notes in Comput Sci*. 2015;9349:430–438. https://doi.org/10.1007/978-3-319-24553-9_53.

98. Zia A, Sharma Y, Bettadapura V, Sarin EL, Essa I. Video and accelerometer-based motion analysis for automated surgical skills assessment. *Int J Comput Assist Radiol Surg*. 2018;13(3):443–455. https://doi.org/10.1007/s11548-018-1704-z.

99. Zia A, Sharma Y, Bettadapura V, et al. Automated video-based assessment of surgical skills for training and evaluation in medical schools. *Int J Comput Assist Radiol Surg*. 2016;11(9):1623–1636. https://doi.org/10.1007/s11548-016-1468-2.

100. Chauvin SW. Applying educational theory to simulation-based training and assessment in surgery. *Surg Clin*. 2015;95(4):695–715. https://doi.org/10.1016/j.suc.2015.04.006.

101. Duffy MC, Lajoie SP, Pekrun R, Lachapelle K. Emotions in medical education: examining the validity of the Medical Emotion Scale (MES) across authentic medical learning environments. *Learn InStruct*. 2020:70. https://doi.org/10.1016/j.learninstruc.2018.07.001.

102. Mehta N, Harish V, Bilimoria K, et al. Knowledge and attitudes on artificial intelligence in healthcare: a provincial survey study of medical students. *Mededpublish*. 2021;10(1). https://doi.org/10.15694/mep.2021.000075.1.