Automated Methods of Technical Skill Assessment in Surgery: A Systematic Review

Marc Levin, BSc,* Tyler McKechnie, BSc,* Shuja Khalid, BASc,†
Teodor P. Grantcharov, MD, PhD, FACS,†,‡ and Mitchell Goldenberg, MBBS†,‡

*Michael G. DeGroote School of Medicine, McMaster University, Hamilton, Ontario, Canada; †Surgical Safety Technologies, Li Ka Shing International Knowledge Institute, Toronto, Ontario, Canada; and ‡Department of Surgery, University of Toronto, Toronto, Ontario, Canada

OBJECTIVE: The goal of the current study is to systematically review the literature addressing the use of automated methods to evaluate technical skills in surgery.

BACKGROUND: The classic apprenticeship model of surgical training includes subjective assessments of technical skill. However, automated methods to evaluate surgical technical skill have been recently studied. These automated methods are a more objective, versatile, and analytical way to evaluate a surgical trainee's technical skill.

STUDY DESIGN: A literature search of the Ovid Medline, Web of Science, and EMBASE Classic databases was performed. Articles evaluating automated methods for surgical technical skill assessment were abstracted. The quality of all included studies was assessed using the Medical Education Research Study Quality Instrument.

RESULTS: A total of 1715 articles were identified, 76 of which were selected for final analysis. An automated methods pathway was defined that included kinetics and computer vision data extraction methods. Automated methods included tool motion tracking, hand motion tracking, eye motion tracking, and muscle contraction analysis. Finally, machine learning, deep learning, and performance classification were used to analyse these methods. These methods of surgical skill assessment were used in the operating room and simulated environments. The average Medical Education Research Study Quality Instrument score across all studies was 10.86 (maximum score of 18).

CONCLUSIONS: Automated methods for technical skill assessment is a growing field in surgical education. We found quality studies evaluating these techniques across many environments and surgeries. More research must be done to ensure these techniques are further verified and implemented in surgical curricula. (J Surg Ed 76:1629−1639. © 2019 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: Surgical training, Technical skills, Automated methods, Surgical technology

COMPETENCIES: Practice-Based Learning and Improvement, Patient Care, Systems-Based Practice, Medical Knowledge, Interpersonal and Communication Skills

INTRODUCTION

The classical model of surgical training involves a master-apprentice relationship between faculty surgeons and surgical trainees. This paradigm has the potential to introduce subjectivity and bias into surgical skill assessment. Recently, new models of medical education, such as the Royal College of Physicians and Surgeons of Canada's (Royal College) Competence-By-Design (CBD), have risen to popularity among surgical training programs. In these models, surgical trainees must demonstrate competency in both in the technical and nontechnical aspects of surgery.[1] Technical skills refer to psychomotor actions or related-mental processes during surgery obtained through practice, such as suturing and knot tying.[2] Nontechnical skills encompass numerous factors such as decision making, communication, and situational awareness.[2] Despite the implementation

Funding: N/A.
Correspondence: Inquiries to Marc Levin, BSc, Michael G. DeGroote School of Medicine, McMaster University, 1280 Main St. W., Hamilton, ON L8S 4K1; e-mail: marc.levin@medportal.ca

of CBD to reduce subjectively in surgical training, variability still exists among assessors in their personal evaluations of trainees' abilities, leading to significant discrepancies in who is deemed competent in this educational model. While the value of subjective evaluation in surgical education is well described, particularly with regards to assessor trainability,[3] human raters introduce biases through factors such as preferences in surgical technique as well as unblinded assessment that a computer is unaffected by.[4]

Currently, surgical training programs aim to objectively evaluate surgical trainees basic surgical skills using tools such as the Objective Structured Assessment of Technical Skills (OSATS) and the Ottawa Surgical Competency Operating Room Evaluation (O-SCORE).[5] Many studies have evaluated the validity of these tools across a variety of environments.[5–7] Despite this, these tools rely on the observations and judgments of an individual,[5] inevitably associated with subjectivity and bias. Furthermore, they require the time and resources of an expert surgeon or trained rater. While the OSATS and O-SCORE, among others, have been translated to different surgical settings in the literature, these tools do not provide any procedure-specific information on how the trainee should strive to improve. The tools are generalizable, without anchors that are related to a specific surgical step.[6] As such, automated and analytically objective ways to evaluate surgical technical skill are needed. In order to objectively evaluate such technical surgical competencies, automated methods for assessing these skills have become increasingly prevalent in the literature. Techniques such as kinetics and computer vision have been studied as a means of accurately assessing technical surgical skills.[8–10] However, the full breadth of these techniques has yet to be defined in the literature.

The purpose of this systematic review is to synthesize the evidence examining automated methods of technical skills assessment in surgery. Additionally, we aim to evaluate the quality of these published methods, and identify techniques requiring greater investigation and development.

METHODS

Eligibility Criteria

All articles in the medical, surgical, or bioengineering literature describing automated methods for assessing technical skills in surgery were included. For this systematic review, we defined "automated" as a technique which involves reducing human intervention to a minimum or negligible amount. We defined surgical technical skills as any action by the surgeon or trainee in the operating room or simulated operating room environments. Finally, we defined assessment as an objective report detailing a surgeon or trainee's technical proficiency at their task. Studies assessing nontechnical skills in surgery were excluded from this review. Only original research studies published in English in peer reviewed journals were included. Randomized controlled trials, observational, cohort, case-control, case series, and cross-sectional studies were included. Lastly, unpublished abstracts, posters, opinions, case reports, reviews, letters to editors, and editorials were excluded from our search. This review was completed in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines.[11]

Search Strategy

One author conducted a search in Ovid Medline, Web of Science, and EMBASE Classic. The search included all studies covering the period from database inception through August 4th, 2018. The search was designed and conducted by a medical research librarian with input from study investigators. Medical subject headings terms used in the search included "auto-analysis," "pattern recognition," "kinematics," "computer vision," "surgery," "operating room," "surgeon," "clinical competence," "time and motion studies," "educational measurement," "technical skills," "technical skills assessment," "surgical competency," and "technical report" (complete search strategy available in Supplementary Digital Content 1). The references of published studies were searched, as well as gray literature, to ensure that relevant articles were not missed. The titles and abstracts of these articles resulting from these searches were then reviewed. Any article that fit the eligibility criteria was selected for full-text review. Any duplicate articles were removed.

Study Selection

Two authors (ML and TM) assessed each article for inclusion according to the previously described eligibility criteria. Any disagreements in article selection were discussed by both authors until consensus was reached. Discrepancies at the full-text stage were resolved by consensus between 2 reviewers and if disagreement persisted, a third reviewer was consulted.

Data Collection Process

After both authors (ML and TM) agreed upon all articles that were to be included, data were extracted systematically. Data extracted included type of automated method assessment, type of surgery, number of centres in the study, number of participants in the study, and study setting (Table 1). Synthesis of extraction was

TABLE 1. Summary of Automated Methods for Technical Skill Assessment in Surgery

Data Extraction Method	Automated Method	Analysis Plan	Total # of Centers Across all Studies	Total # of Participants Across all Studies	Environment	
					OR	Simulation
Kinetics (sensors)	Tool motion tracking		69	1324	8	53
			59	1115	8	42
		Machine learning	15	328		14
		Deep learning	11	152	3	4
		Performance classification	31	635	5	26
	Hand motion tracking		13	263	0	11
		Machine learning	3	69	0	3
		Deep learning	2	66	0	2
		Performance classification	7	91	0	5
	Eye motion tracking		2	78	0	2
		Machine learning	2	78	0	2
		Deep learning	0	0	0	0
		Performance classification	0	0	0	0
	Muscle contraction analysis		1	11	0	1
		Machine learning	0	0	0	0
		Deep learning	0	0	0	0
		Performance classification	1	11	0	1
Computer vision	Tool motion tracking		56	317	5	11
			51	240	2	9
		Machine learning	5	33	1	3
		Deep learning	1	24	0	1
		Performance classification	45	183	1	5
	Hand motion tracking		5	78	3	2
		Machine learning	2	19	2	0
		Deep learning	1	6	1	0
		Performance classification	2	53	0	2
	Eye motion tracking		1	20	0	1
		Machine learning	0	0	0	0
		Deep learning	0	0	0	0
		Performance classification	1	20	0	1

facilitated by grouping the publications based on common methods of technical assessment. Additionally, data regarding the Medical Education Research Study Quality Instrument (MERSQI) was extracted by 2 independent reviewers.[12] Inconsistencies in MERSQI score were discussed by both authors until consensus was reached and if disagreement persisted, a third reviewer was consulted.

Quality Assessment

The MERSQI was used to evaluate the quality of all studies that were included in the present review. The MERSQI stratifies study quality by using 8 categories: study design, institutions samples, response rate, type of data, validity evidence for evaluation of instrument scores, sophistication of data analysis, appropriateness of data analysis, and assessment outcome (Supplemental Digital Content 2). Studies received a summed MERSQI score based on all 8 domains and a relative study quality was assigned.

RESULTS

Our initial search yielded 1715 articles. After review of titles and abstracts by 2 independent authors, 1621 studies were excluded for failing to satisfy the eligibility criteria. 94 articles were selected for full review to determine inclusion status. Subsequent full-text review lead to the inclusion of 76 studies in the final analysis. A Preferred Reporting Items for Systematic Reviews and Meta-

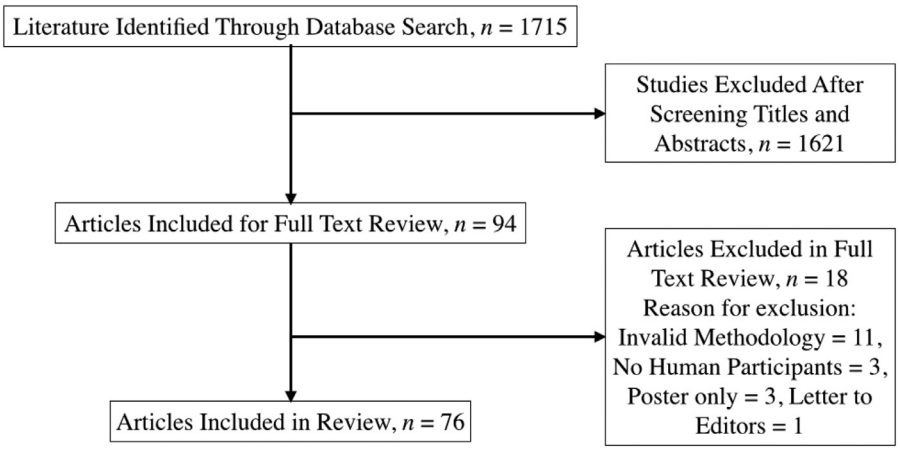

FIGURE 1. Preferred reporting items for systematic reviews and meta-analyses (PRISMA) flowchart.

FIGURE 2. Automated methods pathway from data extraction to analysis item.

Analyses flow diagram describes our search (Fig. 1). The included articles are displayed in Supplemental Digital Content 3.

Methods of Technical Skill Assessment

Figure 2 summarizes the different kinds of automated technical skill assessments extracted and synthesized from the reviewed articles. Current methods for automated technical skills assessment in surgery include tool motion tracking, hand motion tracking, eye motion tracking, and muscle contraction analysis. Data extraction in studies was completed using either kinetics (sensors) and/or computer vision. Then, machine learning, deep learning, or performance classification were employed for further analysis (Fig. 2). An expert engineer in computer vision and machine learning technologies was consulted to define these categories and exemplar papers in each category are reported.

MERSQI Score

Table 2 depicts the median and interquartile range MERSQI score assigned to studies. The average MERSQI score across all studies was 10.86 (maximum score of 18). Table 3 describes the specific validity evidence for evaluation instrument scores. The majority of studies (76.3%) scored a response of validity evidence for relationships to other variables. A total of 14.5% of studies reported only descriptive analysis while 85.5% of studies reported beyond descriptive analysis. A total of 100% of studies had appropriate data analysis.

Data Extraction Method 1: Kinetics

Kinetics is commonly defined in physics as the study of the way forces act on motion. Eight of these studies were completed in the operating room and 53 in a simulated environment. One study evaluated surgical kinetics both in the operating room and in a simulated environment. In this review, any study evaluating forces, torques, angular velocity, or acceleration were included in this category. For example, Chami et al.[13] evaluated the force and torques of hand and surgical tool interfaces using sensors during knee arthroscopy. They found that force metrics correlated with the surgical experience of participants. Cundy et al. [14] used a force transducer

TABLE 2. MERSQI Appraisal of Studies Describing Automated Methods for Technical Skill Assessment in Surgery

Data Extraction Method	Automated Method	Median	IQR
Kinetics		11	0.5 (10.5-11)
	Tool motion tracking	11	0.5 (10.5-11)
	Hand motion tracking	11	0.5 (10.5-11)
	Eye motion tracking	10.5	N/A
	Muscle contraction analysis	11	N/A
Computer vision		11	0.5 (10.5-11)
	Tool motion tracking	11	1 (10-11)
	Hand motion tracking	11	0.5 (10.5-11)
	Eye motion tracking	11	N/A

N/A = not applicable.

retrofitted to a laparoscopic box trainer simulator in their study, and concluded that force metrics calculated in this way could be used to differentiate different levels of surgical skill. They found that more experienced surgeons tended to exert the least amount of force on the surgical tools. They conclude that there is potential for force-related metrics to be a valuable differentiator of surgical skill, and that the force a surgical trainee applies to the surgical instrument, and subsequently the patient's tissues, has implications for surgical safety and quality.[14]

The median MERSQI score for all tool kinetics studies was 11 points out of a possible 18. A total of 76.7% of the validity evidence for evaluation instrument scores for kinetics studies was relationships to other variables. A total of 8.3% of the kinetics studies used internal structure validity and the remaining 15% of studies used combinations of content, internal structure, and relationships to other variables validity. One kinetics study utilized all of content, internal structure, and relationships to other variables validity.[35]

Data Extraction Method 2: Computer Vision

Computer vision techniques aim to allow computers to understand images or videos in real time, using programmed algorithms. These algorithms allow computers to automate human visual tasks by extracting features directly from the provided image data. Five of these studies were completed in the operating room and 11 in a simulated environment. Jain et al.[15] used computer vision to evaluate surgeons performing tracheoesophageal fistula repair on a simulated model. They found that their algorithm could provide accurate and timely quality-control feedback to surgeons during tracheoesophageal fistula repair. Such findings have implications for surgical trainee assessment and education, wherein trainees could be immediately alerted to potential surgical adverse events and provided with live learning opportunities. Similarly, Law et al.[16] also used computer vision analysis to evaluate surgical skill during robotic surgery techniques by analysing 146,309 images from 12 different surgeons. They were able to train machine learning models using features extracted from computer vision techniques to accurately stratify surgeons based on technical skill.

The median MERSQI score for all computer vision studies was 11 points out of a possible 18. A total of 81.3% of the validity evidence for evaluation instrument scores for computer vision studies was relationships to other variables. The remaining 18.7% of computer vision studies used combinations of content, internal structure, and relationships to other variables validity.

Automated Method 1: Tool Motion Tracking

Common surgical tools include laparoscopic instruments, robotic accessories, electrocautery, catheters,

TABLE 3. MERSQI Validity Evidence for Evaluation Instrument Scores

Validity Evidence for Evaluation Instrument Scores Response Options	Number of Studies	Percentage of Total Studies (%)
Content	1	1.3
Internal structure	7	9.2
Relationships to other variables	58	76.3
Internal structure and content	1	1.3
Content and relationships to other variables	4	5.3
Internal structure and relationships to other variables	4	5.3
Internal structure, content, and relationships to other variables	1	1.3

needles, needle drivers, and scissors. In this review, any study quantifying the kinematics of these various surgical tools in terms of motion, velocity, or path of travel was included in this category. Eight studies found using tool motion tracking were completed in the operating room and 42 in a simulated environment. For example, Estrada et al.[17] utilized electromagnetic sensors to track the kinematic movement of a catheter tip during simulated endovascular surgery. Their findings demonstrated that tool motion tracking effectively differentiated surgeons in terms of skill level and correlated strongly to structured grading assessment of surgical skill. Similarly, Hofstad et al.[18] analyzed tool motion in a simulated environment, evaluating a set of 9 motion-related metrics for the tips of laparoscopic instruments. Once again, these factors effectively distinguished expert from novice performers.

The median MERSQI score for all tool motion tracking studies was 11 points out of a possible 18. A total of 74.1% of the validity evidence for evaluation instrument scores for tool motion tracking studies was relationships to other variables. A total of 10.3% of the tool motion tracking studies used internal structure validity and the remaining 15.6% of tool motion tracking studies used combinations of content, internal structure, and relationships to other variables validity. One tool motion tracking study utilized all of content, internal structure, and relationships to other variables validity.[35]

Automated Method 2: Hand Motion Tracking

Hand motion information can be gathered by tracing the movement of the hands in time and space during the performance of a surgical task. All studies using hand motion tracking were completed in a simulated environment. Tasks analyzed included suturing, knot tying, cutting, and drilling. In a study conducted by Zirkle et al.[19], hand motion tracking using computer vision as a data extraction method was utilized to quantify distance covered, time, speed, and total number of movements during a simulated cortical mastoidectomy. They found that analyses of hand motions were an accurate reflection of surgeon experience in the drilling task required for this operation. More recently, a study by Watson[20] utilized a 6-degree of freedom analysis of hand motion with kinetics as a data extraction method during a simulated venous anastomosis task. The results demonstrated that increased amounts of surgical experience correlated with more complex hand motions. This may be indicative of a higher order surgical flow that novice surgeons could analyze and learn from.

The median MERSQI score for all hand motion tracking studies was 11 points out of a possible 18. A total of 83.3% of the validity evidence for evaluation instrument scores for hand motion tracking studies was relationships to other variables. The remaining 16.7% of hand motion tracking studies used combinations of content, internal structure, and relationships to other variables validity.

Automated Method 3: Eye Motion Tracking

Eye motion tracking systems use sensors to monitor and track the movements of surgeons' eyes. Two studies were found in our review using eye motion tracking to assess technical skills of surgeons in simulated environments.[8,21] Ahmidi et al.[21] measured the eye-gaze position of surgeons during simulated endoscopic sinus surgery tasks. Snaineh et al.[8] recorded eye features during laparoscopic skills simulations. Both Snaineh et al.[8] and Ahmidid et al.[21] found eye motion tracking effective and accurate at assessing surgeon skill levels with up to 94.6% accuracy.

The median MERSQI score for all eye motion tracking studies was 10.5 points out of a possible 18. Both eye motion tracking studies used relationship to other variables to for validity evidence for evaluation instrument scores.

Automated Method 4: Muscle Contraction Analysis

Muscle contraction analysis can be used to assess ergonomics during completion of a surgical task. It is a novel idea that evaluates instrument handling and posture in an attempt to reduce fatigue, and subsequent error during surgery.[22] This systematic review identified a single study that used infrared markers for sensing body movement to analyze muscle contraction.[22] They calculated how muscles are used and their mean work during the performance of surgical tasks using a virtual laparoscopic simulator. They found that their muscle contraction analysis results clearly differed between surgeons and medical students.

The MERSQI score for this study was 11 points out of a possible 18. The single muscle contraction analysis study used relationship to other variables to for validity evidence for evaluation instrument scores.

Analysis Plan 1: Machine Learning

Machine learning is the use of mathematical representations for the modelling of a process. A typical machine learning process can be broken down as follows: problem definition, choosing a model type that is best suited to solve that problem, collecting data for training the model, converting the data into a formal tool that can be used to train the model, training the model, and finally tweaking the model. For example, in the case of automated methods for surgical skills assessment, machine learning algorithms would be trained to

effectively analyze these methods. In our review, machine learning analysis was found in both kinetic and computer vision studies. In total, 3 machine learning studies were completed in the operating room and 22 studies in simulated environments. For example, Forestier
et al.[23] implemented machine learning as an analysis plan to discover and rank discriminative and interpretable patterns of surgical sill via the recording of surgical motions. Specifically, these used a publicly available dataset to classify gestures of 10 surgeons in the Operating room (OR). In another study, Oquendo et al.[24] used a validated and novel machine learning algorithm to evaluate the performance of 32 surgical residents during a pediatric laparoscopic suturing task.

Analysis Plan 2: Deep Learning

Deep learning is defined as a specialized subset of machine learning whereby there are more restrictions places on the aforementioned process of machine learning. Specifically, the model type that is chosen to best solve the problem defined is usually deep convolutional networks. This type of learning results in more sophisticated evaluation of surgical technical skill assessment than machine learning. Hence, it may add to the learning experience associated with the automated method of surgical skill assessment. Our review found that 4 deep learning studies were completed in the operating room and 7 studies in simulated environments. For example, Dosis et al.[25] used data from 5 surgeons performing laparoscopic cholecystectomies to further train their dexterity based motion device.

Analysis Plan 3: Performance Classification

Performance classification is completed using either a machine learning classifier or a deep learning classifier. For example, the trained models accept input in a certain format, run it through the model and produce an integer value that corresponds to one of the performance outcomes from subjective evaluation. Most performance studies use N to represent the number of performance outcome categories. For example, $N = 5$ if classification is done to associate an OSATS score to the performance of a surgeon, or $N = 3$ if surgeon performance is simply broken down into unacceptable, acceptable, or excellent, or $N = 2$ if surgeon performance is classified as novice or expert. Chellali et al.[26] used performance classification to analyze surgical skill on a Virtual Basic Laparasopic Skill Trainer to differentiate between expert (Postgraduate Year 5, fellow and attending surgeons) and novice (Postgraduate Year 1-4). This is an example of a study using performance classification categories of $N = 2$. Alternatively, Shafiei et al.[27] used kinetic data extraction techniques during a simulated Da Vinci robot task to differentiate motor skill performance between expert, intermediate and novice surgeons. This is an example of a study using performance classification categories of $N = 3$.

DISCUSSION

Technical surgical skills among surgical trainees are commonly evaluated by senior or more experienced surgeons. Such evaluation techniques are often subjective and often not detailed enough for an all-encompassing assessment. The use of CBD models is increasing in surgical residency programs internationally. As such, the creation and validation of automated and objective methods of surgical technical skill assessment is a growing and exciting area of research. In this systematic review, existing automated methods that exist in evaluating technical skills in surgery are outlined.

This review highlights the heterogeneity that exists in automated methods of technical skill evaluation in surgery. Despite the diversity in the literature, we identified an automated methods pathway including kinetics and computer vision data extraction methods. Automated methods included tool motion tracking, hand motion tracking, eye motion tracking, and muscle contraction analysis. Finally, machine learning, deep learning, and performance classification were used to analyze these methods. Tool motion tracking was the most frequently employed methods, while hand motion tracking and muscle contraction analysis were only utilized in a single study each. The most common procedural approaches studied were laparoscopic and robotic minimally invasive surgery. Furthermore, most studies were completed in a simulated environment rather than in the OR. We posit these trends exist due to ease of methodology as well as concerns for patient safety. For example, it also may be less distracting for a surgeon if there is a motion/force probe on their laparoscopic instrument in comparison to the same probe on their hand.

The MERSQI score was used to assess the overall quality of each study in this systematic review. The average MERSQI score across all 76 studies included in this systematic review is 10.86. Using a benchmark of 14/18 to determine "high quality" literature, this implies that from an educational perspective, these articles do not meet this standard.[28] It is not coincidental that the average MERSQI scores across studies involving tool kinetics, computer vision, motion tracking, and muscle contraction was 11/18. Most studies included in this review were nonrandomized, 2 group studies at a single institution. The majority of studies used objective data and analyzed in a descriptive manner. Hence, despite variation

in automated method across the different studies, the methods used to in the studies were quite homogenous. All studies used appropriate data analysis and the majority used sophisticated data analysis reported more than just descriptive results. After in-depth analysis of the specific validity of the studies included, the majority reported relationships to other variables by creating expert-novice technical skill comparisons using their automated method. Further breakdown of the validity scores by data extraction method revealed similar trends across kinetics and computer vision, with the majority of studies from both methods using relationships to other variables validity. When validity scores were analysed by automated method, similar trends ensued across eye motion tracking, tool motion tracking, hand motion tracking, and muscle contraction, with relationships to other variables being the primary source of validity. One exemplar study incorporated all of content, internal structure, and relationships to other variables validity.[35] They used kinetics, tool motion tracking, and machine learning algorithms to automatically evaluate robotic surgical trainees and surgeons on their performance of a peg transfer skill on a da Vinci Standard robot. The tool that they used to extract kinetic data from the surgeon's tools' motions was created. Content structure was achieved in this study as all the authors are mechanical engineering experts and developed their own hardware and software specifically for their study. Internal structure was implemented by inter-rater reliability by having 3 different expert rater surgeons evaluating all participants using GEARS. The study mentions the implementation of a calibration process and intraclass correlation coefficients to ensure strong inter-rater reliability. The authors demonstrated that contact forces and robot arm accelerations can automatically rate surgeon skill at a level that is good to excellent in comparison with human raters using GEARS. The minority of studies included in this review implemented content validity or internal structure, limiting the overall validity of these studies. The majority of studies regardless of extraction method or automated method used relationship to other variables validity, stressing the dearth of validity among studies in this review. Furthermore, this emphasizes the lack of superiority of one extraction method or automated method, from a validity perspective. It is important to identify this limitation among the majority of the automated method of technical skill assessment literature and encourage future studies to strive to incorporate all of content, internal structure, and relationships to other variables validity.

Bias exists in human rating of surgical technical skills.[29,30] The current models used in surgical technical skills assessment include subjective rating tools such as the OSATS and GEARS.[6,7,31] Despite studies validating these tools across a variety of surgical training environments, they rely on input from individuals, inevitably skewing the results according to the rater's personal bias.[6,7,31] The implementation of automated technical skills assessments in surgical scenarios has the potential to mitigate such biases and has implications for more equitable assessment of surgical trainees across institutions, provinces, and countries. Areas in surgical training programs, where these fair assessment measures would be beneficial, include CBD observations and Royal College credentialing examinations. Currently the Royal College and American Board of Surgery's surgical licensing exams do not include intraoperative technical skills assessments. The Royal Australasian College of Surgeons lists technical expertise as a core competency of surgical training despite not assessing intraoperative technical skill on licensing exams. Whereas, the Royal College of Surgeons' Intercollegiate Surgical Curriculum Program in the United Kingdom is designed for assessing the incremental development of technical and operative skills. As these assessments are a significant component of their training curricula, automating such assessments would benefit their program through increased objectivity, efficiency, and uniformity. We envision that sections of surgical specialty licensing exams be added wherein surgical trainees across countries such as Canada, the United States of America, Australia, and the United Kingdom uniformly complete specialty-specific surgical technical skills tasks intraoperatively. These tasks would be assessed automatically via one of the aforementioned technical skill assessment techniques. Resultantly, graduating surgical trainees would not only be able to demonstrate clinical and surgical problem-solving abilities, but also technical skills when obtaining licensure. In order for such large scale changes to be implemented, future collaboration between experts in the field of automated methods of surgical skill assessment and medical education is required.

There are certain limitations associated with the present review. Firstly, most of the studies were completed in simulated environments. While parts of surgical training do occur in simulation labs, traditionally the majority of training occurs in the OR. As such, we cannot be certain that the methods used to automatically assess technical surgical skill in the simulated environments would translate directly to the OR environment. For example, issues such as physical space in the OR for the required equipment for automated assessment, this equipment's sterility, and the comfort of the surgeon using this equipment could all prevent the seamless transition of these techniques from simulation to the OR. Secondly, the included studies lack comparison between two or more kinds of automated technical skill assessment using the same outcome metrics. Although some studies

included 2 different automated methods of assessment, they were not directly comparing the two. For example, kinetic data extraction methods were used to evaluate both tool motion tracking and hand motion tracking as insight items.[32,33] As some studies combined rather than compared automated methods, we were similarly unable to do comparative analyses of the different types of methods used in the studies. Hence, we cannot make inferences or recommendations regarding whether one automated method for surgical skill assessment is superior to others. Thirdly, while we have mentioned that nonautomated methods of technical skill assessment can be resource intensive, it is also important to recognize the unique resources required for automating methods of technical skill assessment such as computing power and analytic expertise. It is important to recognize that currently, automated methods and machine learning algorithms inherently require aspects of human input for creation. Such resources were not commented on in this review. Finally, a limitation of this whole body of literature is a lack of consensus on terminology between studies. For example, studies used terminology of kinetics and kinematics interchangeably, while other studies used the terms computer vision and tool motion analysis to describe similar techniques.

There are a multitude of directions in which this field can continue expanding. Addressing the limitations outlined above, future work should undertake the evaluation of these automated techniques in the clinical environment to ensure their generalizability, and to assess their relationship with patient outcomes and surgical safety. Future studies must also compare the efficacy of different automated methods within the same study. For example, whether certain variables such as tool motion tracking, hand motion tracking, or muscle contraction analysis are more effective than others at information behavioral change. Moreover, uniform approaches to evaluating the efficacy of these methods should be identified, as this would allow for educators to more easily understand the relative utility of automated assessments in their respective learning environments. Additionally, consensus must be reached among experts in the field regarding the terminology for the techniques that are being used in these studies. This would further allow comparisons to be made among the various automated assessment techniques. A reason for this lack of language agreement may be that some of these studies were not completed under the guidance of engineers or experts in the automated technology space. While Satava et al.[34] defined terms for objective surgical assessment, we urge that collaboration between surgical and technical experts ensue in future automated methods studies, to ensure updated and accurate terminology usage. Lastly, studies in this systematic review rarely provided information on how to apply the data from automated methods of technical skill assessment to trainee learning objectives. According to our MERSQI calculations, the majority of the included studies failed to report trainee behavior as an outcome of their study. While we have demonstrated the utility of automated methods of technical skills assessment in surgery, more research to translate these such data into these learner-related contexts must be completed.

CONCLUSIONS

In conclusion, in this systematic review we identified an automated methods pathway with kinetics and computer vision data extraction methods, automated methods included tool motion tracking, hand motion tracking, eye motion tracking, and muscle contraction analysis, which were analyzed using machine learning, deep learning and performance classification. Characteristics of the studies were described and categorized, and their educational quality was calculated using the MERSQI. While tool motion tracking was most commonly investigated in the minimally invasive surgery and simulation environments, there is certainly room for expansion of this exciting and innovative area of research. Implications of this work include the potential to drastically alter the way surgical trainees learn and are assessed. In the era of CBD medical education, we believe that automated methods represent the next step toward objective analysis of surgical trainees. Applications of these tools range from low-stakes assessments for trainee learning, to summative assessment at the surgical licensing examination level and beyond.

REFERENCES

1. Sargeant J, Bhanji F, Holmboe E, et al. Assessment and Feedback for Continuing Competence and Enhanced Expertise in Practice. Royal College of Physicians and Surgeons of Canada.
2. Agha RA, Fowler AJ, Sevdalis N. The role of non-technical skills in surgery. *Ann Med Surg*. 2015. https://doi.org/10.1016/j.amsu.2015.10.006.
3. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the "black box" differently: assessor cognition from three research perspectives. *Med Educ*. 2014;48:1055–1068. https://doi.org/10.1111/medu.12546.
4. Williams RG, Verhulst S, Colliver JA, Dunnington GL. Assuring the reliability of resident performance appraisals: more items or more observations?

Surgery. 2005;137:141–147. https://doi.org/10.1016/j.surg.2004.06.011.

5. Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa surgical competency operating room evaluation (O-SCORE): a tool to assess surgical competence. *Acad Med*. 2012;87:1401–1407. https://doi.org/10.1097/ACM.0b013e3182677805.

6. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997. https://doi.org/10.1002/bjs.1800840237.

7. Hatala R, Cook DA, Brydges R, Hawkins R. Constructing a validity argument for the objective structured assessment of technical skills (OSATS): a systematic review of validity evidence. *Adv Heal Sci Educ*. 2015. https://doi.org/10.1007/s10459-015-9593-1.

8. Snaineh S, Seales B. Minimally invasive surgery skills assessment using multiple synchronized sensors.. In: International Symposium on Signal Processing and Information Technology; 2016. p. 314–319.

9. Aggarwal R, Grantcharov T, Moorthy K, et al. An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. *Ann Surg*. 2007;245:992–999. https://doi.org/10.1097/01.sla.0000262780.17950.e5.

10. Azari DP, Frasier LL, Quamme SRP, et al. Modeling surgical technical skill using expert assessment for automated computer rating. *Ann Surg*. 2019;269:574–581. https://doi.org/10.1097/SLA.0000000000002478.

11. Moher D, Liberati A, Tetzlaff J, Altman DG, Prisma Group. Preferred reporting items for systematic reviews and Meta-Analyses: The PRISMA statement. *Ann Intern Med*. 2009;151:264–269. https://doi.org/10.1371/journal.pmed.1000097.

12. Cook DA, Reed DA. Appraising the quality of medical education research Methods: The medical education research study quality instrument and the newcastle-ottawa scale-education. *Acad Med*. 2015. https://doi.org/10.1097/ACM.0000000000000786.

13. Chami G, Ward JW, Phillips R, Sherman KP. Haptic feedback can provide an objective assessment of arthroscopic skills. *Clin Orthop Relat Res*. 2008;466:963–968. https://doi.org/10.1007/s11999-008-0115-9.

14. Cundy TP, Thangaraj E, Rafii-Tari H, et al. Force-Sensing enhanced simulation environment (For-Sense) for laparoscopic surgery training and assessment. *Surgery*. 2015;157:723–731. https://doi.org/10.1016/j.surg.2014.10.015.

15. Jain S, Barsness KA, Argall B. Automated and objective assessment of surgical training: detection of procedural steps on videotaped performances. In: 2015 Int Conf Digit Image Comput Tech Appl DICTA; 2015. https://doi.org/10.1109/DICTA.2015.7371233.

16. Law H, Ghani K, Deng J. Surgeon technical skill assessment using computer vision based analysis. *Proceedings of Machine Learning for Healthcare*. 2017;68:88–99.

17. Estrada S, O'Malley MK, Duran C, Schulz D, Bismuth J. On the development of objective metrics for surgical skills evaluation based on tool motion. In: Conf Proc - IEEE Int Conf Syst Man Cybern; 2014. p. 3144–3149. https://doi.org/10.1109/smc.2014.6974411.

18. Hofstad EF, Våpenstad C, Chmarra MK, Langø T, Kuhry E, Mårvik R. A study of psychomotor skills in minimally invasive surgery: what differentiates expert and nonexpert performance. *Surg Endosc Other Interv Tech*. 2013;27:854–863. https://doi.org/10.1007/s00464-012-2524-9.

19. Zirkle M, Roberson DW, Leuwer R, Dubrowski A. Using a virtual reality temporal bone simulator to assess otolaryngology trainees. *Laryngoscope*. 2007. https://doi.org/10.1097/01.mlg.0000248246.09498.b4.

20. Watson RA. Quantification of surgical technique using an inertial measurement unit. *Simul Healthc*. 2013. https://doi.org/10.1097/SIH.0b013e318277803a.

21. Ahmidi N, Hager GD, Ishii L, Gallia GL, Ishii M. Robotic path planning for surgeon skill evaluation in minimally-invasive sinus surgery. *Med Image Comput Comput Assist Interv*. 2012;15:471–478. https://doi.org/10.1007/978-3-642-33415-3_58.

22. Cavallo F, Pietrabissa A, Megali G, et al. Proficiency assessment of gesture analysis in laparoscopy by means of the surgeons musculo-skeleton model. *Ann Surg*. 2012;255:394–398. https://doi.org/10.1097/SLA.0b013e318238350e.

23. Forestier G, Petitjean F, Senin P. Artificial Intelligence in Medicine. 2017;10259. https://doi.org/10.1007/978-3-319-59758-4.

24. Oquendo YA, Riddle EW, Hiller D, Blinman TA, Kuchenbecker KJ. Automatically rating trainee skill at a pediatric laparoscopic suturing task. *Surg Endosc Other Interv Tech*. 2018. https://doi.org/10.1007/s00464-017-5873-6.

25. Dosis A, Aggarwal R, Bello F, et al. Synchronized video and motion analysis for the assessment of procedures

in the operating theater. *Arch Surg*. 2005;140:293–299. https://doi.org/10.1001/archsurg.140.3.293.

26. Chellali A, Ahn W, Sankaranarayanan G, et al. Preliminary evaluation of the pattern cutting and the ligating loop virtual laparoscopic trainers. *Surg Endosc*. 2015;91:165–171. https://doi.org/10.1016/j.chemosphere.2012.12.037.Reactivity.

27. Shafiei SB, Cavuoto L, Guru KA. Motor skill evaluation during robot-assisted Surgery. In: Volume 5A: 41st Mechanisms and Robotics Conference; 2017. https://doi.org/10.1115/DETC2017-67607.

28. Lin H, Lin E, Auditore S, Fanning J. A narrative review of high-quality literature on the effects of resident duty hours reforms. *Acad Med*. 2016;91:140–150. https://doi.org/10.1097/ACM.0000000000000937.

29. Gerull KM, Loe M, Seiler K, McAllister J, Salles A. Assessing gender bias in qualitative evaluations of surgical residents. *Am J Surg*. 2019;217:306–313 https://doi.org/10.1016/j.amjsurg.2018.09.029.

30. Vogt VY, Givens VM, Keathley CA, Lipscomb GH, Summitt RL Jr. Is a resident's score on videotaped objective structured assessment of technical skills affected by revealing the resident's identity? *Am J Obstet Gynecol*. 2003;189:688–691. https://doi.org/10.1067/S0002-9378(03)00887-1.

SUPPLEMENTARY INFORMATION

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.jsurg.2019.06.011.

31. Aghazadeh MA, Jayaratna IS, Hung AJ, et al. External validation of global evaluative assessment of robotic skills (GEARS). *Surg Endosc Other Interv Tech*. 2015;29:3261–3266. https://doi.org/10.1007/s00464-015-4070-8.

32. Kumar R, Jog A, Vagvolgyi B, et al. Objective measures for longitudinal assessment of robotic surgery training. *J Thorac Cardiovasc Surg*. 2008;45:788–802. https://doi.org/10.1038/jid.2014.371.

33. Islam G, Kahol K, Li B, Smith M, Patel VL. Affordable, web-based surgical skill training and evaluation tool. *J Biomed Inform*. 2016;59:102–114. https://doi.org/10.1016/j.jbi.2015.11.002.

34. Satava RM, Cuschieri A, Hamdorf J. Metrics for objective assessment: preliminary summary of the surgical skills workshop. *Surg Endosc Other Interv Tech*. 2003;17:220–226. https://doi.org/10.1007/s00464-002-8869-8.

35. Brown JD, O Brien CE, Leung SC, Dumon KR, Lee DI, Kuchenbecker KJ. Using contact forces and robot arm accelerations to automatically rate surgeon skill at peg transfer. *IEEE Trans Biomed Eng*. 2017;64:2263–2275. https://doi.org/10.1109/TBME.2016.2634861.