Educational Impact of Automated Feedback Systems in Surgical Training: A Systematic Review With Quantitative Synthesis

Gauri Harshawardhan Godbole, MSc,[†,‡] Daniel Hawkins, MSc,[†,‡] Kingsley Ewool, MBcHB, MSc, MRCS,[‡] Mauro Henrique Batista Camacho, MD, MSc,[§] Rezaul Karim, MBBS, MRCS, MSc,[‖], and Bijendra Patel MBBS, MS, FRCS[‡]

[†]Barts and The London School of Medicine and Dentistry, London, United Kingdom; [‡]Queen Mary University of London, London, United Kingdom; [§]City St Georges Universtiy London, London, United Kingdom; and [‖]Barts Health NHS Trust, London, United Kingdom

OBJECTIVE: To evaluate the impact of automated feedback systems (AFS) on technical surgical skill acquisition in individuals undergoing surgical skills training.

DESIGN: PRISMA-guided systematic review of studies published between May 2013 and December 2024. Four databases were searched. Eligible studies compared AFS with no feedback or assessed the impact of AFS on technical skill. Risk of bias was assessed using ROB-2 and ROBINS-I, and certainty of evidence with GRADE. The review was prospectively registered with PROSPERO (CRD420251058650).

SETTING: Simulation-based training environments, including bench models and virtual reality simulators.

PARTICIPANTS: Fourteen studies involving 814 trainees were included.

RESULTS: All studies reported improvement in technical skills with AFS; 9 demonstrated significant within-group gains, with a mean improvement of 38.1% ($p = 0.0046$). Six studies contributed to pooled analysis, showing a moderate-to-large benefit (Hedges' $g = 0.81$, 95% CI: 0.45-1.00, $p < 0.0001$). Secondary outcomes consistently favored AFS: learner satisfaction increased by 60% (MD = 1.16, 95% CI: 0.65-1.67, $p < 0.01$), path length decreased by 41% (95% CI: 10.3%-71.7%, $p = 0.02$), speed improved by 9.4% (MD = 3.1 mm/s, 95% CI: 0.4-5.8, $p = 0.04$), and applied force was reduced by 11.8% (95% CI: 4.5%-19.2%, $p = 0.01$).

CONCLUSIONS: AFS are associated with moderate-to-large improvements in technical performance, particularly for foundational repetitive surgical tasks. While gains are often task-specific and largely confined to simulation settings, evidence supports AFS as a valuable adjunct to early surgical training. Integration into structured programmes, alongside expert oversight and contextual teaching, is essential to maximize benefit and ensure safe transferability to clinical practice. Small study numbers, task-specific designs, and heterogeneity limit interpretation. (J Surg Ed 83:103879. © 2026 The Author(s). Published by Elsevier Inc. on behalf of Association of Program Directors in Surgery. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/))

KEYWORDS: automated feedback systems, surgical education, technical skills, simulation training, artificial intelligence, performance metrics

INTRODUCTION

Surgical education has undergone major reform over the past 3 decades, shifting from traditional apprenticeship-based training[1] to structured, competency-driven frameworks.[2] The European Working Time Directive (EWTD), introduced in 1993 and fully implemented by 2009, capped trainee hours at 48 per week to safeguard well-being.[3] While well-intentioned, this reform significantly

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Correspondence: Inquiries to Gauri Harshawardhan Godbole, MSc, Barts and The London School of Medicine and Dentistry, Queen Mary University London, Charterhouse Square, London, EC1M 6BQ, United Kingdom; e-mail: gaurihgodbole@gmail.com

  1931-7204/$30.00 https://doi.org/10.1016/j.jsurg.2026.103879

reduced hands-on operative exposure and disrupted continuity of care-challenges that were later compounded by the COVID-19 pandemic and increasing workforce pressures in the NHS.[4] These constraints rendered traditional, time- and volume-based training models increasingly inadequate for ensuring surgical competence. These constraints have created demand for scalable, reproducible methods to support technical skill acquisition outside the operating room.

It is well established that proficient surgeons are associated with better patient outcomes, with disparities in technical skill contributing to over 25% of the variation in postoperative outcomes.[5] However, traditional training methods often fall short, with challenges including inconsistent feedback, limited access to expert mentors, and reduced hands-on exposure.[6] Simulation training has become a cornerstone of surgical training, offering a safe and controlled environment in which trainees can repetitively practise technical tasks without consequences. The definition of a simulator can be broadly described as "an imitation of some real thing, state of affairs, or process for the practice of skills, problem solving, and judgment."[7] However, its educational value is highly dependent on expert supervision to deliver timely, constructive feedback. Given pressures on faculty time and resources, there is increasing interest in technologies that can supplement direct supervision.

Artificial intelligence (AI), as defined by the European Commission, refers to "systems that display intelligent behavour by analyzing their environment and taking actions-with some degree of autonomy-to achieve specific goals."[8] In medical education, AI is gaining traction for its potential to personalize learning, automate assessment, and provide consistent, real-time feedback. A 2019 integrative review found that most applications focus on undergraduate training, with formative feedback emerging as the most common use-particularly to address persistent gaps in timely performance evaluation.[9] These findings have translated into practical applications: platforms like Geeky Medics now incorporate AI-powered chatbots that simulate OSCE-style patient encounters, helping users refine empathy, structure their communication, and build clinical reasoning. With over 45,000 users engaging in more than 700 interactive scenarios.[10] Automated Feedback Systems (AFS) represent a novel application of artificial intelligence within surgical education. By analysing metrics such as motion efficiency, instrument force, and task completion time, AFS provide trainees with real-time, objective feedback that is immediate and standardised.[11] Early studies suggest that AFS can reliably differentiate between novice and expert performance and improve efficiency in simulation-based tasks.[12-15] Importantly, they also offer the potential for personalised, data-driven feedback at scale, which could address current limitations in faculty availability.

Despite these promising developments, key questions remain. It is unclear whether AFS primarily function as assessment tools or whether they actively enhance technical skill acquisition in trainees.[16] Furthermore, variability in feedback modalities, study designs, and outcome measures has limited comparability across studies.[17]

This systematic review aims to evaluate the educational impact of AFS on technical surgical skill development. Specifically, it investigates whether real-time, automated feedback improves performance outcomes, and explores its influence on learner perceptions, cognitive workload, and training efficiency.

METHODS

Study Design

This systematic review was conducted in accordance with the PRISMA 2020 guidelines (Fig. 1) to ensure transparency and reproducibility. The protocol was prospectively registered with PROSPERO (ID: CRD420251058650). Ethical approval was not required, as only previously published data were included.

- Eligibility Criteria (Refer to supplementary materials Table 4):

  Studies were eligible if they:

- Included medical students, residents, or other surgical trainees without independent proficiency in the target task.
- Evaluated automated or AI-driven feedback systems (AFS) providing real-time or postperformance feedback during technical surgical training.
- Reported technical performance outcomes, using validated global rating scales (e.g., OSATS, GRS) or objective metrics (e.g., motion analysis, force, efficiency measures).
- Were randomized controlled trials or cohort studies, published between May 2013-December 2024, in English (or with English translation).

Exclusion criteria included:

- Studies involving senior trainees/experts already competent in the assessed task.
- AFS used only for skill classification.
- Interventions focused solely on nontechnical or cognitive skills (e.g., decision-making, communication).
- Single-arm studies or studies with unavailable outcome data after author contact.

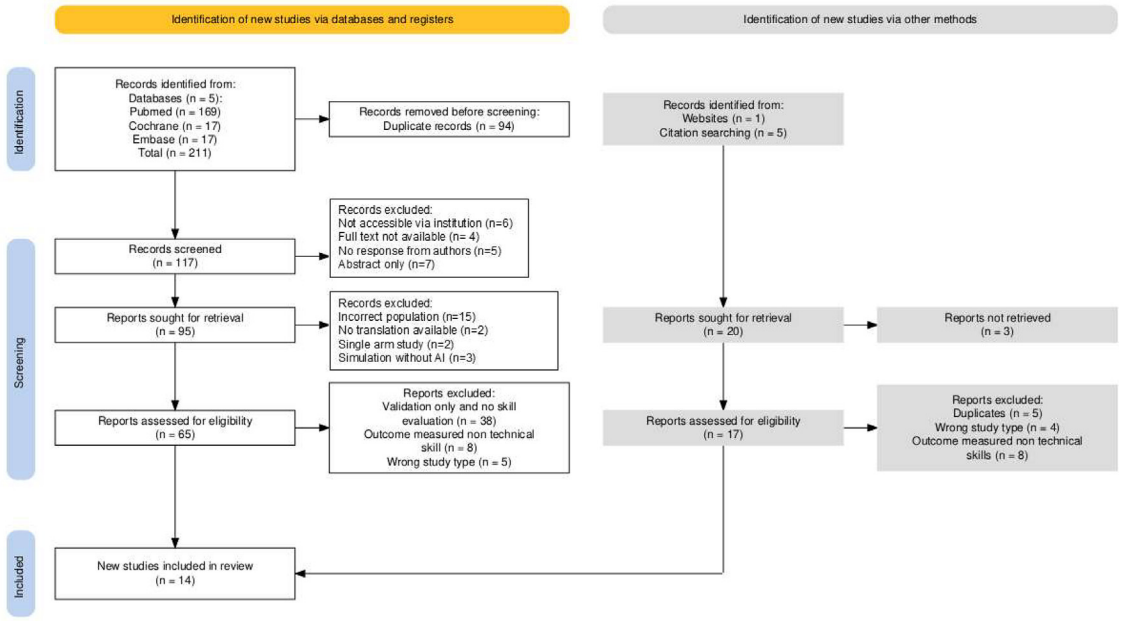

FIGURE 1. PRISMA 2020 flow diagram of study selection 53- Outlines the process of study identification, screening, eligibility assessment, and inclusion. From 211 records identified, 14 studies met inclusion criteria. Reasons for exclusion are detailed at each stage.

Information Sources and Search Strategy

A systematic search was performed in PubMed, Embase, ScienceDirect, and the Cochrane Library (December 2024-March 2025). The search covered studies published from May 2013 to December 2024.

Search strategies combined MeSH terms and free-text keywords relating to *trainee, novice, surgical skills, automated feedback, artificial intelligence, and surgical education*. Example PubMed and Cochrane search strategies are provided in Appendix A Tables 2-4. Boolean operators (AND/OR) were applied to optimise sensitivity.

Study Selection

Search results were imported into EndNote, with duplicates automatically removed. Two reviewers independently screened titles and abstracts against eligibility criteria. Full texts of potentially relevant studies were retrieved and assessed for inclusion. Disagreements were resolved through discussion and consensus.

Data Collection Process

Data were extracted into a structured spreadsheet by 1 reviewer and independently verified by a second reviewer. Extracted variables included:

- Study details (author, year, country, study design).
- Population characteristics (sample size, sex, level of training).
- Task performed and type of AFS used.
- Comparator group (e.g., no feedback/traditional teaching).
- Primary and secondary outcome measures.

When outcome data were incomplete, study authors were contacted. Studies with irretrievable data were excluded.

Outcomes:

- *Primary outcome*: improvement in technical skill, measured via validated composite rating instruments (e.g., OSATS, GRS, OPRS) or study-specific global rating tools.
- *Secondary outcomes*: Learner attitude toward the training system using Likeart scale (1-5), Cognitive demand (%), Change in Instrument path length/ distance travelled (%), Instrument speed / velocity (mm/s), Change in Instrument force.

Risk of Bias Assessment

Risk of bias was assessed at the study level using:

- *RoB 2* for randomized controlled trials.[18]
- *ROBINS-I* for nonrandomized studies.[19]

Certainty of Evidence

The certainty of evidence for each outcome was assessed using the *GRADE* framework (considering risk of bias, inconsistency, indirectness, imprecision, and publication bias.

Data Synthesis and Analysis

Both qualitative and quantitative syntheses were undertaken. For quantitative synthesis, pre- and postintervention scores were extracted where available to calculate change scores and effect sizes (Hedges' g).[20] Standard deviations for change scores were reconstructed using established formulae, assuming a correlation coefficient of $r = 0.50$ (with sensitivity analyses at $r = 0.30$ and $r = 0.70$).[21] Forest plots were generated using the metafor package in R (version 4.4.2).[20]

Secondary outcomes (satisfaction, cognitive demand, efficiency metrics) were narratively synthesised and presented descriptively. Direction of effect was indicated as *AFS* > *NF* or *AFS* < *NF*.

RESULTS

The search yielded 211 records, of which 14 studies met the eligibility criteria (Appendix A). These comprised 12 randomized controlled trials (8 full RCTs, 2 pilot RCTs, 2 proof of concept RCTs), 1 parallel cohort, 1 prospective pilot, enrolling a total of 814 trainees, including 628 participants in AFS groups.

Most participants were novices - medical students and early postgraduate residents. Interventions ranged from basic psychomotor tasks such as suturing, knot-tying, and probing, to more complex procedures including laparoscopic cholecystectomy, robotic suturing, and CT-guided puncture. Feedback modalities included VR simulators, robotic console metrics, sensor-augmented box trainers, and AI-driven platforms. Outcomes were measured using validated global rating scales such as OSATS, GRS, and mOSATS, alongside objective motion and force metrics (Table 1).

Risk of bias was generally low among RCTs, although 2 were judged to have "some concerns." Among the 4 nonrandomized studies, 1 was at low risk, 2 at moderate risk, and 1 at high risk due to self-selection and confounding (Fig. 2).

Across all 14 studies, technical skill improved in AFS groups. Nine of the no-feedback (NF) control groups also improved, while 2 declined, suggesting a potential risk of unguided error reinforcement (Table 2) Quantitative synthesis of 6 studies demonstrated a moderate-to-large benefit of AFS (Mean Hedges' $g = 0.81$, 95% CI: 0.35-1.40), with mean skill improvement of 31% (median 13%) (Fig. 3). Narrative synthesis of the remaining 8 studies also consistently favoured AFS, though methodological and outcome heterogeneity precluded pooling.

Secondary outcomes supported these findings. Four studies (n = 300) assessing learner satisfaction, each using a 1−5 Likert scale, demonstrated significantly higher scores in AFS groups (mean difference +1.16 points, $p < 0.001$). Two studies (n = 121) assessed cognitive load using different instruments—a 1−5 Likert-type Cognitive Load Index and a reaction-time workload ratio. Because these measures were not comparable, results were summarised narratively. Both studies reported higher cognitive load with AFS: Frithioff et al.[35] found a 10% increase in workload, while Yilmaz et al.[36] reported a small increase in extraneous load on a Likert scale compared with expert teaching. Motion efficiency, assessed in 2 studies (n = 68), showed reduced path length (−15%) and increased instrument speed (+3.1 mm/s). Three studies (n = 98) evaluated force metrics, demonstrating a reduction in applied force (−28.8%), although certainty was limited by variation in measurement methods (Table 3). A consolidated summary of all primary and secondary outcomes, including absolute effects, effect sizes, and certainty of evidence ratings, is provided in the Summary of Findings table (Refer to Supplementary Materials: Table 5).

DISCUSSION

This review examined the impact of automated feedback systems (AFS) on technical skill acquisition in surgical training. Fourteen studies were included, covering tasks from basic suturing to advanced procedures such as tumor resection, minimally invasive surgery, and drilling. Across studies, AFS was associated with moderate-to-large improvements in technical performance (pooled Hedges' $g = 0.81$), particularly in repetitive, precision-based tasks such as suturing and drilling.[30] Task efficiency measures (instrument path length, speed and force modulation) and learner satisfaction consistently favoured AFS, suggesting that these systems enhance both technical outcomes and trainee confidence.

Task Specific Improvements

Performance improvements varied by task type. Suturing and anastomosis tasks demonstrated the most consistent and substantial benefits, reflecting their dependence on motion-based metrics and trajectory optimization. Tumor resection studies also showed improvements in efficiency and safety metrics, though findings were limited by single-center designs and heterogeneous methodologies. Minimally invasive and image-guided procedures demonstrated smaller but positive effects, likely due to

TABLE 1. Study Characteristics

Study	Authors	Study Design	Country	Task Performed	AFS Used	Sample Size (M/F)	Control (n)	Traditional Training (n)*	AFS (n)	Level of Training (n)	Assessment Tool Used
1	Yang, Y. Y. and B. Shulruf[22]	Prospective pilot study	China	Suturing/ligature	WasedaKyoto-Kagaka suture no. 2 refined II) WKS-2RII system	72 (38/34)	25	24	23	5th year (72)	GRS TP (technical performance)
2	Yilmaz, R., et al.[23]	Randomized control trial	Canada	Tumor resection (brain tumor)	NeuroVR	97 (39/58)	32	32	33	1st year(56) 2nd year (10) 3rd year (4) 4th year (1)	ICEMS OSATS
3	Fazlollahi, A. M., et al.[24]	Randomized control trial	Canada	Tumor resection (brain tumor)	Da Vinci (kinematics and system events data)	70 (29/41)	23	24	23	Preparatory (26) 1st years (25) 2nd years (19)	ICEMS OSATS
4	Ma, R., et al.[25]	Pilot study	United States	Suturing (VUA)	NeuroVR	42 (17/ 25)	20	N/A	22	Undergraduates (8) Medical students (34)	EASE (End-To-End Assessment of Suturing Expertise)
5	Yilmaz, R., et al.[26]	Randomized control trial	Canada	Resection (brain tumor)	AI Algorithm (prev developed)	120 (71/49)	30	28	Visual (29) Visuo-spatial (33)	1st year(75) 2nd year(31) 3rd year (8) 4th year(6)	Benchmarks
6	Fazlollahi, A. M., et al.[27]	Cohort study	Canada	Resection (brain tumor)	Virtual Operative Assistant (VOA)	46 (19/ 27)	23	N/A	23	Preparatory (19) First year (16) Second year (11)	90 Benchmarks
7	Khan, D. Z., et al.[28]	RCT stage 1, proof of concept study	United Kingdom	Endoscopic pit surgery	AI-assisted indexing and performance analysis	104	41	N/A	63		mOSATS
8	Wu, S., et al.[29]	RCT stage 1, proof of concept study		Laparoscopic cholecystectomy	SmartCoach	18 (17/1)	9	N/A	9	Resident (14) Fellow (4)	LCRF
9	Ji, Z., et al.[30]	Parallel cohort Study	China	CT guided puncture	ChatGPT	90 (49/11)	30	N/A	Template + Digital image (30) ChatGPT + Template+digital image (30)	Juniors (29) Intermediate (46) Seniors (15)	Benchmarks
10	Wijewickrema, S., et al.[31]	Randomized control trial	Australia	Drilling	Trained a classifier	24	12	N/A	12	MBBS (13) MD (10) PhD (1)	"stroke" (their own performance metric)
11	Vallabhajosula, S., et al.[32]	Randomized control trial	United states	Robotic bimanual carrying, needle passing, suture tying	Augmented visual feedback	22	N/A	N/A	Relative phase feedback (5) Speed feedback (5) Grip Force feedback (6) Video feedback (6)	"novice" users	Expert benchmarks
12	Singapogu, R. B., et al.[33]	Randomized control trial	United states	Laparoscopic grasping, probing, or sweeping	fundamentals of laparoscopic skills (FLS) trainer	30 (15/5) - 5 ppl didnt want to respond	N/A	N/A	Grasping (n = 9) Probing (n = 10) Sweeping (n = 10)	undergraduate students	Forces
13	Ma, R., et al.[34]	Randomized control trial	United States	Robotic suturing	Da Vinci (kinematics and system events data)	23 (15/8)	12	N/A	11	PGY 0 (12) PGY 1 (4) PGY 2 (4) PGY 3 (3)	cumulative of specific outcomes
14	Frithioff, A., et al.[35]	Randomized control trial	United Kingdom	VUA	Visible Ear Simulator (ver 2.1)	24 (10/14)	12	N/A	12	Medical students	Final product score mean metrics-based score

Table showing study characteristics- Summary of study characteristics for all included trials.

*Some studies featured a third arm involving traditional or standard instruction; while these are reported in the table for completeness, they were excluded from the primary synthesis, which focused specifically on comparisons between automated feedback systems (AFS) and no feedback.

their greater cognitive demands. Drilling tasks showed improvements in both efficiency and fine-control behaviors, with some evidence of short-term skill retention. These findings collectively suggest that AFS may be particularly effective in tasks where performance can be decomposed into quantifiable motor behaviors. A summary of task-specific quantitative outcomes is provided in Table 2.

Across studies, the tools used to assess performance varied widely -from validated instruments like ICEMS[24,36] or mOSATS,[28] to proprietary AI-derived metrics such as "bleeding risk" or "coordination index."[37] Even when domains like "efficiency" or "safety" were consistently named, the actual measurement techniques varied-sometimes based on expert scoring, while others used objective kinematic data such as instrument path length or speed.[27,33] Consequently, improvements across studies may appear comparable in name while representing fundamentally different constructs - a recurring issue seen in the wider literature.[38] This highlights a central issue in surgical education research: surgical skill is inherently task-specific, and no single metric can universally capture competence across procedures. The educational value of AFS therefore depends on the extent to which system-generated metrics align with the real-world demands of each task. Developing core outcome sets and validating AI-derived metrics against clinical outcomes will be essential to improving interpretability and comparability across future studies.

Accuracy/Reliability of AFS

Although most included studies benchmarked performance against expert standards, few formally evaluated AFS validity. One study demonstrated high agreement between AFS and expert feedback,[31] but automated classification of performance remains vulnerable to misinterpreting nontraditional but clinically acceptable techniques. Many systems provide little transparency regarding how thresholds are defined, how errors are weighted, or how confident the algorithm is in its assessment.[39]

Recent work outside the included studies suggests that AI-driven feedback systems may generate inconsistent explanations across learner groups, raising concerns about fairness. Emerging frameworks such as the Task-Weighted Integrative Index (TWIX) aim to align AI-generated feedback with human-labelled behavior, improving both accuracy and equity.[40] As AFS evolve from passive scoring tools to active feedback agents, ensuring interpretability and minimizing bias will be increasingly important.

Feedback modality also appears to influence learning. One RCT in this review found that visual or visuospatial feedback was more effective than numerical metrics alone, likely because it supports quicker error recognition and reduces the cognitive effort required to interpret performance data.[37] However, system reliability continues to vary across tasks, learner backgrounds, and procedural complexity.[41] Until feedback-generation frameworks become more standardized and transparent, AFS should function as an adjunct to expert mentorship

Improvement in Confidence and Learner Engagement

AFS also influenced learner experience. Four studies reported higher confidence and satisfaction among learners using automated feedback, and these perceptions generally corresponded with objective performance gains. Two studies observed increased cognitive workload with AFS, which may represent productive mental effort associated with interpreting and integrating multiple streams of feedback gains.[26,35]

These findings align with Self-Determination Theory and principles of deliberate practice, which highlight the importance of competence, autonomy, and structured feedback in motivating sustained learning.[42,43,44] However, reliance on self-reported outcomes and short follow-up periods limits conclusions about long-term impact on engagement, skill retention, and clinical performance.

Implications for Surgical Education

Integrating AFS in Simulation Based Training

The findings of this review suggest that Automated Feedback Systems (AFS) can meaningfully enhance early-stage surgical training by improving both core technical metrics (e.g., force modulation/ instrument speed), and learner-related outcomes like confidence and motivation. The scalability and efficiency of AFS directly address many of the limitations of traditional simulation-based training.[45]

Automated feedback can reduce dependence on continuous expert supervision, offering a scalable way to support early skills training. Although formal cost analyses of AFS are lacking, simulation-based training more broadly has demonstrated strong cost-effectiveness, with several studies showing substantial reductions in training costs as programmes scale.[46,47] As an adjunct to simulation, AFS has the potential to amplify these economic and educational benefits by providing consistent, automated guidance during routine skill acquisition. However, dedicated cost-effectiveness studies are needed to determine whether these advantages translate into measurable savings at programme level.

Despite their promise, AFS remains underutilized. Most implementations are confined to small pilot studies, and few training programmes have established frameworks

FIGURE 2. Risk of bias assessment. Left: Cochrane Risk of Bias 2 (RoB-2) traffic plot and summary for randomized controlled trials. Right: ROBINS-I traffic plot and summary for non randomized studies. Colorblind friendly versions can be found in supplementary materials (Fig 3 and 4).

for interpreting automated feedback or integrating it into progression decisions. Wider adoption will require clearer mapping of task complexity to AFS suitability, standardized and objective metrics for proficiency, benchmarking against validated tools such as OSATS and GEARS, and evaluation of skill transfer and retention in clinical environments. Developing validated, clinically relevant criteria for interpreting AFS-generated metrics-particularly for defining proficient performance-would enable training pathways to shift toward dynamic, performance-based progression rather than fixed-time models.

AFS as a Complement to Mentorship

AFS provide consistent, real-time feedback that supports deliberate practice and helps learners identify technical errors independently. However, feedback without interpretation can be misapplied, overprioritized, or taken out of context, particularly when systems optimize narrow metrics that do not fully reflect procedural competence. Notably, 1 study reported that although AFS improved safety metrics, it also introduced trade-offs-such as reduced dominant-hand efficiency and slower tumor resection.[27] Similarly, a classifier-based system labelled trainees as "proficient" based on force thresholds despite persistent technical errors, illustrating the risk of false positives when system-defined metrics are used in isolation.[31] Across the studies included in this review, hybrid models that combine AFS with expert instruction tended to achieve the strongest improvements, particularly for complex or judgement-heavy procedures. These findings underscore the need for human oversight: mentors contextualise automated cues, teach adaptability, and judge whether variations in force, trajectory, or speed are appropriate for a given scenario. Accordingly, AFS should function as an adjunct within structured, educator-led curricula-standardising routine feedback and extending faculty reach-while expert mentorship provides the contextual, adaptive guidance required for safe and holistic surgical performance.

Limitations

This review is limited by the small sample sizes of included studies (often fewer than 30 participants per arm[29,34,35] and heterogeneity in both tasks and outcome measures. Only 6 of the 14 studies contributed to pooled analysis due to inconsistent reporting of baseline data, and outcome definitions (e.g., "efficiency" or "safety") varied widely across trials. Although most randomized controlled trials were at low risk of bias, nonrandomized designs introduced moderate-to-high risk, particularly through confounding and assessor bias (Yang[22]). GRADE certainty was therefore moderate for most outcomes, with high certainty achieved only for learner satisfaction. Finally, all studies were conducted in simulated environments-using artificial models, VR platforms, or benchtop trainers so the extent to which such gains transfer to real-world surgical environments remains uncertain. Simulation cannot fully replicate clinical complexity, including

TABLE 2. Primary Outcomes

Task Type	Study	Specific Task Performed	Sample Size (NF/AFS)	Improvement Post Task Completion (+/=/-)		Automated feedback Group Skill Gain (%)	Δ Improvement (Automated Feedback vs No Feedback)
				No Feedback	AFS		
Suturing/ Anastomosis	Yang et al.[22]	Suturing/ligature	NF: 25 AFS: 23	↑	↑	+ 22.5%	+ 7% vs NF
	Ma et al.[25]	Robotic suturing (VUA)	NF: 20 AFS: 22	↓	↑	+ 9.75%	+10.91% vs NF
	Vallabhajosula et al.[32]	Robotic bimanual carrying, needle passing, suture tying	AFS: 22	No Data	↑	+ 30.28%	No data
	Ma et al.[34]	Robotic Suturing	NF: 12 AFS: 11	↑	↑	+ 39.85%	+ 6.8% vs NF
Tumor resection	Yilmaz et al.[36]	Brain tumor resection	NF: 32 AFS: 33	↑	↑	No data	+0.266 points vs NF
	Fazlollahi et al.[24]	Brain tumor resection	NF: 23 AFS: 33	↓	↑	+ 124.56%	+130% vs NF
	Yilmaz et al.[37]	Brain tumor resection	NF: 30 AFS: 62	↑	↑	Reported improvement; no data	No data
	Fazlollahi et al.[27]	Brain tumor resection	NF: 23 AFS: 23	↑	↑	+31. 06%	+ 28.7% vs NF
Minimally invasive procedures	Khan et al.[28]	Endoscopic pituitary surgery	NF: 41 AFS: 63	↑	↑	Reported improvement; no data	No data
	Wu et al.[29]	Laparoscopic cholecystectomy	NF: 9 AFS: 9	↑	↑	+ 29%	+17.2% vs NF
	Singapogu et al.[33]	Laparoscopic grasping, probing, or sweeping	AFS: 29	No Data	↑	+ 20.62%	No Data
Image-guided procedure	Ji et al.[30]	CT-guided puncture	NF: 30 AFS: 60	↑	↑	+ 35.6%	+ 15.6% vs NF
Drilling	Wijewickrema et al.[31]	Temporal bone drilling	AFS: 22	No Data	↑	Reported improvement; no data	Reported improvement; no data
	Frithioff et al.[35]	Anatomical mastoidectomy with posterior tympanotomy	NF: 12 AFS: 12	↑	↑	No Data	+12.7% vs NF
Studies Summary	14 studies		NF: 204 AFS: 424	11/14 Performance declined in 2 studies	14/14 Improvement observed in all Studies	9/14 Mean: 38.14% Range: 114.81% Median: 30.28%	9/14 All studies favored AFS: Range:123.2% Median: 15.6%

Primary outcomes- across included studies. The symbol (↑) indicates improvement in technical skill, and the symbol (↓) indicates a decline in technical skill. Outcomes reflect within-group improvements in technical skill following training with or without automated feedback systems (AFS). Comparative data between AFS and no-feedback (NF) groups is presented where available.

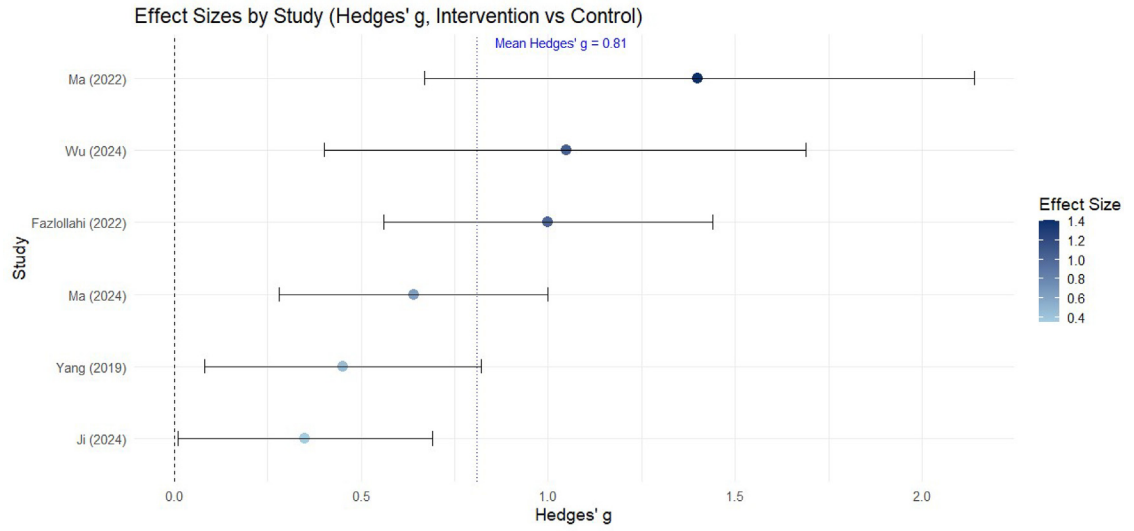

FIGURE 3. Forest plot showing effect sizes for technical skill improvement across studies comparing AFS to no feedback: Each dot represents the standardized mean difference for a study, and horizontal lines indicate 95% confidence intervals. The color gradient reflects effect size magnitude. The vertical dotted line marks the pooled mean effect (Hedges' $g = 0.81$), indicating a moderate-to-large overall benefit of AFS.

TABLE 3. Secondary Outcomes

Outcome	Studies (n)	Participants (n)	Mean NF (Post)	Mean AFS (Post)	Mean % Change (NF)	Mean % Change (AFS)	Direction of Favor
Learner attitude (satisfaction/confidence)	4	300	3.36	4.52	12.5	60	AFS > NF
Cognitive demand (%)	2	121	N/A	N/A	No data	No data	AFS < NF
Instrument path length (mm)	2	68	15.1	514.6	−26	−41	AFS >NF
Instrument speed (mm/s)	2	68	13.5	16.6	12.5	−9.44	AFS >NF
Instrument force (N)	3	98	0.98	51.6	17	−11.83	AFS >NF

Summary of secondary outcomes across included studies comparing automated feedback systems (AFS) to no feedback (NF)-Outcomes include learner-reported satisfaction/confidence, cognitive demand, and objective performance metrics such as instrument path length, speed, and force. Mean postintervention scores and percentage changes are shown where available. Direction of favor indicates which group demonstrated superior performance; reductions in force and path length were considered improvements.

patient variability, time pressure, and interprofessional dynamics. Moreover, AFS systems cannot evaluate intraoperative decision-making, adaptability, or teamwork. Trainees may also behave differently in independent versus supervised conditions. As such, while the findings support AFS as an adjunct for early technical training, they do not justify its standalone use for assessing clinical readiness or licensing decisions.

CONCLUSIONS

This systematic review evaluated the impact of automated feedback systems (AFS) on technical surgical skill acquisition. Across 14 studies, AFS was associated with moderate-to-large improvements in overall technical performance, particularly in foundational, repetitive tasks such as suturing. Improvements were consistently observed in finer performance domains-including task speed and instrument path length. Furthermore, several studies reported increased learner confidence and engagement, supporting the utility of AFS in early-stage simulation-based training. However, the benefits of AFS were often task- and domain-specific, with some studies highlighting trade-offs between safety and efficiency. Substantial heterogeneity in task types, outcome measures, and feedback modalities, along with small sample sizes and limited baseline data, restricted the scope of

quantitative synthesis. Furthermore, as all included studies were conducted in simulated environments, the extent to which these performance gains translate to real-world surgical competence remains uncertain.[37]

Overall, the evidence supports the use of AFS as a promising adjunct for enhancing early procedural skill development. However, these systems should be integrated within structured training pathways that include expert mentorship and contextualized teaching. AFS are not a substitute for clinical judgment or experience but can help optimize deliberate practice, standardize technical benchmarks, and extend faculty reach-particularly in resource-constrained settings.

Future Directions

Future research should prioritise standardisation of outcome metrics and task selection, enabling comparability across studies. Validation of AFS-derived metrics against clinical outcomes is essential to ensure their educational and practical relevance. Longitudinal studies assessing skill retention and transfer to the operative environment are needed, alongside economic evaluations to determine scalability and cost-effectiveness. Finally, as AI-driven feedback becomes more sophisticated, issues of interpretability, fairness, and integration into competency-based curricula will be central to ensuring its safe and effective adoption.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the author(s) used ChatGPT to improve language and readability. The author (s) reviewed and edited the content as needed and take full responsibility for the content of the publication.

ACKNOWLEDGMENTS

The author gratefully acknowledges Professor Bijendra Patel and Dr. Mauro Henrique Batista Camacho for their guidance and mentorship throughout this work. Thanks are also due to Md Rezaul Karim for his encouragement, and to Mr. Kingsley Ewool and Mr. Daniel Hawkins for their collaboration and feedback. Appreciation is extended to Queen Mary University of London and the Barts Cancer Institute for their resources and support.

REFERENCES

1. Bolwell JS. Surgical practice, then and now: the 5th to the 21st century. *Bullet Royal Coll Surg Engl*. 2020;102(3):94–101. https://doi.org/10.1308/rcsbull.2020.94.
2. Skjold-Ødegaard B, Søreide K. Competency-based surgical training and entrusted professional activities: perfect match or a procrustean bed? *Ann Surg*. 2021;273(5):e174.
3. Fitzgerald JEF, Caesar BC. The European working time directive: a practical review for surgical trainees. *Int J Surg*. 2012;10(8):399–403. https://doi.org/10.1016/j.ijsu.2012.08.007.
4. Advancing The Surgical Workforce: 2023 UK surgical workforce census report. 2024. https://www.rcseng.ac.uk/-/media/Files/RCS/Standards-and-research/RCS-Advancing-The-Surgical-Workforce-Digital.pdf. Access 25th July 2025.
5. Stulberg JJ, Huang R, Kreutzer L, et al. Association between surgeon technical skills and patient outcomes. *JAMA Surg*. 2020;155(10):960–968. https://doi.org/10.1001/jamasurg.2020.3007.
6. Poljo A, Sortino R, Daume D, et al. Educational challenges and opportunities for the future generation of surgeons: a scoping review. *Langenbecks Arch Surg*. 2024;409(1):82. https://doi.org/10.1007/s00423-024-03270-7.
7. Rosen KR. The history of medical simulation. *J Crit Care*. 2008;23(2):157–166. https://doi.org/10.1016/j.jcrc.2007.12.004.
8. European C. A definition of AI: main capabilities and disciplines. 2018. https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_-december_1.pdf. Access 21st July 2025.
9. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. *JMIR Med Educ*. 2019;5(1):e13930. https://doi.org/10.2196/13930.
10. Gupta N, Khatri K, Malik Y, et al. Exploring prospects, hurdles, and road ahead for generative artificial intelligence in orthopedic education and training. *BMC Med Educ*. 2024;24(1):1544. https://doi.org/10.1186/s12909-024-06592-8.
11. Park JJ, Tiefenbach J, Demetriades AK. The role of artificial intelligence in surgical simulation. *Front Med Technol*. 2022;4:1076755. https://doi.org/10.3389/fmedt.2022.1076755.
12. Hla DA, Hindin DI. Generative AI & machine learning in surgical education. *Curr Probl Surg*. 2024:101701. https://doi.org/10.1016/j.cpsurg.2024.101701.

13. Nakajima K, Kitaguchi D, Takenaka S, et al. Automated surgical skill assessment in colorectal surgery using a deep learning-based surgical phase recognition model. *Surg Endosc*. 2024;38(11):6347–6355. https://doi.org/10.1007/s00464-024-11208-9.

14. Prevezanou K, Seimenis I, Karaiskos P, Pikoulis E, Lykoudis PM, Loukas C. Machine learning approaches for evaluating the progress of surgical training on a virtual reality simulator. Article. *Appl Sci (Switz)*. 2024;14(21):9677. https://doi.org/10.3390/app14219677.

15. Lam K, Chen J, Wang Z, et al. Machine learning for technical skill assessment in surgery: a systematic review. *NPJ Digit Med*. 2022;5(1):24. https://doi.org/10.1038/s41746-022-00566-0.

16. Isaac S, Phillips MR, Chen KA, Carlson R, Greenberg CC, Usability Khairat S. Acceptability, and implementation of artificial intelligence (AI) and machine learning (ML) techniques in surgical coaching and training: a scoping review. *J Surg Educ*. 2024;81(7):994–1003. https://doi.org/10.1016/j.jsurg.2024.03.018.

17. Kankanamge D, Wijeweera C, Ong Z, et al. Artificial intelligence based assessment of minimally invasive surgical skills using standardised objective metrics - A narrative review. *Am J Surg*. 2024;241:116074. https://doi.org/10.1016/j.amjsurg.2024.116074.

18. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019;366:l4898. https://doi.org/10.1136/bmj.l4898.

19. Sterne JAC, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2016;355:i4919. https://doi.org/10.1136/bmj.i4919.

20. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3):1–48. https://doi.org/10.18637/jss.v036.i03.

21. Furukawa TA, Barbui C, Cipriani A, Brambilla P, Watanabe N, Churchill R. Imputing standard deviations in meta-analyses when only summary information is available. *Psychol Med*. 2006;36(6):849–855. https://doi.org/10.1017/S0033291706007742.

22. Yang YY, Shulruf B. Expert-led and artificial intelligence (AI) system-assisted tutoring course increase confidence of Chinese medical interns on suturing and ligature skills: prospective pilot study. *J Educ Eval Health Prof*. 2019;16:7. https://doi.org/10.3352/jeehp.2019.16.7.

23. Yilmaz R, Bakhaidar M, Alsayegh A, et al. Real-time multifaceted artificial intelligence vs In-person instruction in teaching surgical technical skills: a randomized controlled trial. *Sci Rep*. 2024;14(1):15130. https://doi.org/10.1038/s41598-024-65716-8.

24. Fazlollahi AM, Bakhaidar M, Alsayegh A, et al. Effect of artificial intelligence tutoring vs expert instruction on learning simulated surgical skills among medical students: a randomized clinical trial. *J Art* JAMA Netw Open. 2022;5(2):e2149008. https://doi.org/10.1001/jamanetworkopen.2021.49008.

25. Ma R, Kiyasseh D, Laca JA, et al. Artificial intelligence-based video feedback to improve novice performance on robotic suturing skills: a pilot study. *J Art* J Endourol/Endourol Soc. 2024;38(8):884–891. https://doi.org/10.1089/end.2023.0328.

26. Yilmaz R, Fazlollahi AM, Winkler-Schwartz A, et al. Effect of feedback modality on simulated surgical skills learning using automated educational systems: a four-arm randomized control trial. *J Surg Educ*. 2024;81(2):275–287. https://doi.org/10.1016/j.jsurg.2023.11.001.

27. Fazlollahi AM, Yilmaz R, Winkler-Schwartz A, et al. AI in surgical curriculum design and unintended outcomes for technical competencies in simulation training. *JAMA Netw Open*. 2023;6(9):e2334658. https://doi.org/10.1001/jamanetworkopen.2023.34658.

28. Khan DZ, Newall N, Koh CH, et al. Video-based performance analysis in pituitary surgery - part 2: artificial intelligence assisted surgical coaching. *World Neurosurg*. 2024;190:e797–e808. https://doi.org/10.1016/j.wneu.2024.07.219.

29. Wu S, Tang M, Liu J, et al. Impact of an AI-based laparoscopic cholecystectomy coaching program on the surgical performance: a randomized controlled trial. Article. *Int J Surg*. 2024;110(12):7816–7823. https://doi.org/10.1097/JS9.0000000000001798.

30. Ji Z, Jiang Y, Sun H, et al. Enhancing puncture skills training with generative AI and digital technologies: a parallel cohort study. *BMC Med Educ*. 2024;24(1):1328. https://doi.org/10.1186/s12909-024-06217-0.

31. Wijewickrema S, Piromchai P, Zhou Y, et al. Developing effective automated feedback in temporal bone surgery simulation. *Otolaryngol Head Neck Surg*. 2015;152(6):1082–1088. https://doi.org/10.1177/0194599815570880.

32. Vallabhajosula S, Judkins TN, Mukherjee M, Suh IH, Oleynikov D, Siu KC. Skills learning in robot-assisted surgery is benefited by task-specific augmented

feedback. *Surg Innov*. 2013;20(6):639–647. https://doi.org/10.1177/1553350613484590.

33. Singapogu RB, DuBose S, Long LO, et al. Salient haptic skills trainer: initial validation of a novel simulator for training force-based laparoscopic surgical skills. *Surg Endosc*. 2013;27(5):1653–1661. https://doi.org/10.1007/s00464-012-2648-y.

34. Ma R, Lee RS, Nguyen JH, et al. Tailored feedback based on clinically relevant performance metrics expedites the acquisition of robotic suturing skills-an unblinded pilot randomized controlled trial. *J Urol*. 2022;208(2):414–424. https://doi.org/10.1097/ju.0000000000002691.

35. Frithioff A, Frendø M, von Buchwald JH, Trier Mikkelsen P, Sølvsten Sørensen M, Arild Wuyts Andersen S. Automated summative feedback improves performance and retention in simulation training of mastoidectomy: a randomised controlled trial. *J Laryngol Otol*. 2022;136(1):29–36. https://doi.org/10.1017/s0022215121003352.

36. Yilmaz R, Bakhaidar M, Alsayegh A, et al. Real-time multifaceted artificial intelligence vs In-person instruction in teaching surgical technical skills: a randomized controlled trial. *Sci Rep*. 2024;14(1):15130.

37. Yilmaz R, Fazlollahi AM, Winkler-Schwartz A, et al. Effect of feedback modality on simulated surgical skills learning using automated educational systems: a four-arm randomized control trial. *J Art.* J Surg Educ. 2024;81(2):275–287. https://doi.org/10.1016/j.jsurg.2023.11.001.

38. Dick L, Boyle CP, Skipworth RJE, Smink DS, Tallentire VR, Yule S. Automated analysis of operative video in surgical training: scoping review. *BJS Open*. 2024;8(5). https://doi.org/10.1093/bjsopen/zrae124.

39. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. *Ann Surg*. 2018;268(1):70–76. https://doi.org/10.1097/sla.0000000000002693.

40. Kiyasseh D, Laca J, Haque TF, et al. A multi-institutional study using artificial intelligence to provide reliable and fair feedback to surgeons. *Commun Med (Lond)*. 2023;3(1):42. https://doi.org/10.1038/s43856-023-00263-3.

41. Matsuda T, McDougall EM, Ono Y, et al. Positive correlation between motion analysis data on the LapMentor virtual reality laparoscopic surgical simulator and the results from videotape assessment of real laparoscopic surgeries. *J Endourol*. 2012;26(11):1506–1511. https://doi.org/10.1089/end.2012.0183.

42. Manninen M, Dishman R, Hwang Y, Magrum E, Deng Y, Yli-Piipari S. Self-determination theory based instructional interventions and motivational regulations in organized physical activity: a systematic review and multivariate meta-analysis. *Psychol Sport Exerc*. 2022;62:102248. https://doi.org/10.1016/j.psychsport.2022.102248.

43. Mitchell SA, Boyer TJ. Deliberate Practice in Medical Simulation. Treasure Island (FL): StatPearls Publishing; 2025. *StatPearls*Copyright © 2025, StatPearls Publishing LLC.

44. Higgins M, Madan CR, Patel R. Deliberate practice in simulation-based surgical skills training: a scoping review. *J Surg Educ*. 2021;78(4):1328–1339. https://doi.org/10.1016/j.jsurg.2020.11.008.

45. Shahrezaei A, Sohani M, Taherkhani S, Zarghami SY. The impact of surgical simulation and training technologies on general surgery education. *BMC Med Educ*. 2024;24(1):1297. https://doi.org/10.1186/s12909-024-06299-w.

46. Isaranuwatchai W, Brydges R, Carnahan H, Backstein D, Dubrowski A. Comparing the cost-effectiveness of simulation modalities: a case study of peripheral intravenous catheterization training. *Adv Health Sci Educ Theory Pract*. 2014;19(2):219–232. https://doi.org/10.1007/s10459-013-9464-6.

47. Pezel T, Clémence T, Bohbot Y, et al. Cost-effectiveness analysis of simulation-based training in transesophageal echocardiography: insights from the SIMULATOR trial. *Arch Cardiovasc Dis*. 2024;117(Supplement 1):S158. https://doi.org/10.1016/j.acvd.2023.10.290.

SUPPLEMENTARY INFORMATION

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.jsurg.2026.103879.