https://doi.org/10.1038/s44387-025-00032-8

Combining real-time AI and in-person expert instruction in simulated surgical skills training - Randomized crossover trial

Recai Yilmaz[1,2] ✉, Ahmad Alsayegh[1,3,4], Mohamad Bakhaidar[1,3,4], Ali M. Fazlollahi[1,5], Nour Abou Hamdan[1], Trisha Tee[1], Albert Shalmiev[5], Denis Laroche[6], Carlo Santaguida[3,5], Daniel A. Donoho[2,7] & Rolando F. Del Maestro[1,3,5]

Traditional surgical training has significant limitations, lacking objectivity and standardization. Deploying AI tools with conventional expert-mediated teaching may uncover areas where AI could complement experts and enhance surgical training through real-time performance assessment and feedback alongside risk mitigation. This randomized crossover trial assessed learning outcomes in two training sessions involving in-person expert instruction and real-time AI feedback using previously validated tumor resection simulations. Receiving expert feedback before real-time AI instruction led to greater performance improvement in trainee performance scores compared to the opposite order, with a mean difference of 0.67 95%CI [0.43–0.91], $p < 0.001$. Diminishing returns were observed with human expert feedback, which were not seen with AI feedback, such as increased injury and bleeding risk. In surgical procedural training, AI feedback may efficiently maintain peak performance after an initial learning phase led by human experts. AI-integrated surgical curricula should consider the relative benefits of both AI and expert feedback.

In the modern paradigm of surgical apprenticeship, trainees practice with multiple mentors and develop their competencies through ongoing self-reflection and feedback. This learner-focused model offers trainees the opportunity to learn the multifaceted skills of being a surgeon from a diverse set of instructors each with their unique style of practice[1]. Technical skills constitute one domain of expertise that requires attaining a level of proficiency[2,3]. However, traditional methodologies fall short in objectively measuring trainees' technical capacity, real-time quantifying key parameters that may influence patient outcome, and efficient delivery of the feedback instruction[4,5]. Meanwhile, the field of surgery is experiencing a wave of growing technologies capable of performance data acquisition and analysis. In this scene, artificial intelligence (AI) has become the new tool to not only discern trainees' competency but also offer real-time personalized instructions from data-driven insights[5,6]. Virtual reality simulators provide unique advantages through real-time data collection from surgical performance during realistically simulated tasks, enabling real-time AI assistance while trainees practice skills.

Learners could now gain a further active role in their skills acquisition by practicing with intelligent tutoring systems[5–7]. Previous studies have compared the efficacy of learning from an intelligent tutor with that of remote and in-person instructors in simulation training[8]. These findings demonstrate clear advantages of using AI to teach basic skills to novice learners, such as significantly reduced instructor time for achieving a significantly higher performance score[8,9]. However, relying on these limited AI settings led to unintended learning outcomes that negatively affected performance, notably in operative efficiency and movement[10]. Such limitations highlight the multifaceted nature of surgical technical skills and suggest shifting AI's role from replacing to augmenting surgical instructors within the training paradigm. Moreover, the concept of hybrid intelligence, which aims to optimize the complementary strengths of humans and AI, is particularly relevant in surgical training[11]. Because of the difference in contextual information available to each mode of instruction[12], human-AI collaboration can enhance surgical skills and decision-making, ultimately resulting in safer and more efficient surgical interventions.

This study investigates the combined effect of in-person human instruction with real-time AI tutoring by comparing the effect of two sequences of hybrid training on medical students' skill acquisition in a

[1]Neurosurgical Simulation and Artificial Intelligence Learning Centre, McGill University, Montreal, QC, Canada. [2]Division of Neurosurgery and Pediatrics, Children's National Medical Center, Washington, DC, USA. [3]Department of Neurology and Neurosurgery, Montreal Neurological Institute and Hospital, McGill University, Montreal, QC, Canada. [4]Division of Neurosurgery, Department of Surgery, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia. [5]Faculty of Medicine and Health Sciences, McGill University, Montreal, Canada. [6]National Research Council Canada, Boucherville, Qc, Canada. [7]School of Medicine and Health Sciences, George Washington University, Washington, DC, USA. ✉e-mail: recai.yilmaz@mail.com

crossover randomized trial. We hypothesized that early exposure to AI tutoring will result in a higher performance and receiving human instruction later will provide an additional significant boost. Our discussion examines the relative strengths and limitations of both AI and in-person expert feedback and explores the developments for integrating real-time AI into realistic settings such as the operating room.

Results

Participants

Twenty-five students (mean age [SD]: 21.6 [3.2], 12 [%48] women), currently enrolled in four medical (MD) programs across Canada, participated in this study (Table 1). The exclusion criteria included participation in previous trials involving the NeuroVR (CAE Healthcare) neurosurgery simulator.

Performance assessment using the ICEMS score (Fig. 1a)

Performance was assessed using the ICEMS's composite score, which was also evaluated for alignment with the OSATS rating (Objective Structured Assessment of Technical Skills). An average score was calculated for each task, and the improvement throughout the practice tasks and the transfer of skills to the test task was assessed.

There was a statistically significant interaction between the groups and trials on the ICEMS composite score for the second training session, $F(2.72, 62.47) = 15.34$, $p < 0.001$, partial $\eta 2 = 0.40$. However, there was no significant interaction for the first training session, $F(4, 92) = 1.81$, $p = 0.134$, partial $\eta 2 = 0.07$. AI2EXPERT improved significantly on the ICEMS score at the fifth repetition of the task compared to the baseline, 0.39, 95%CI [0.09–0.69], $p = 0.008$ (mean difference, 95%CI [lower bound upper bound], p-value). AI-assisted learning demonstrated a significantly greater improvement compared to human-mediated learning in the first training session. Specifically, AI2EXPERT achieved a significantly higher score at the end of this session (0.23, 95%CI [0.002–0.45], $p = 0.049$) when compared to EXPERT2AI. Unexpectedly, students' performance declined in the second session when they received human expert-mediated training after the AI session. In particular, the ICEMS score for AI2EXPERT significantly declined at the fifth repetition of the task compared to the baseline (−0.42, 95%CI [−0.67 – (−0.18)], $p < 0.001$). Real-time AI-mediated training significantly improved students' performance in the second session, after their exposure to human-mediated training in the first session, which yielded no significant improvement. EXPERT2AI showed a significant increase in their score at the end of the second session (0.51, 95% CI [0.15–0.86], $p = 0.003$). During the second training session, EXPERT2AI had significantly higher scores compared to AI2EXPERT in both the fourth and fifth repetitions of the task with a mean difference of 0.44 (95% CI [0.11, 0.76], $p = 0.011$), and 0.67 (95% CI [0.43, 0.91], $p < 0.001$), respectively. These results demonstrate superior learning outcomes with AI-mediated training compared to human-mediated training.

Learning transfer to the test task (Fig. 1b)

EXPERT2AI achieved a higher ICEMS score in the test subpial resection task compared to AI2EXPERT at the end of the second training session, 0.06 95%CI [0.01–0.11]. $p = 0.013$. Both groups showed significant improvement in their scores in the second training session compared to the first, regardless of the intervention order. AI2EXPERT and EXPERT2AIs had significantly higher scores at the end of the second session when compared to their performance at the end of the first session in the test task 0.31 95%CI [0.14–0.48], $p = 0.002$, 0.41 95%CI [0.21–0.61], $p < 0.001$, respectively.

Skill retention

No statistically significant changes were observed in the ICEMS score from the end of the first session to the beginning of the second session for AI2EXPERT (0.01, 95%CI [−0.26–0.28], $p = 0.96$) and EXPERT2AI (0.04, 95%CI [−0.11–0.18], $p = 0.60$). These results may indicate that students were able to retain the information they have acquired when moving on to the second session regardless of the instruction they received.

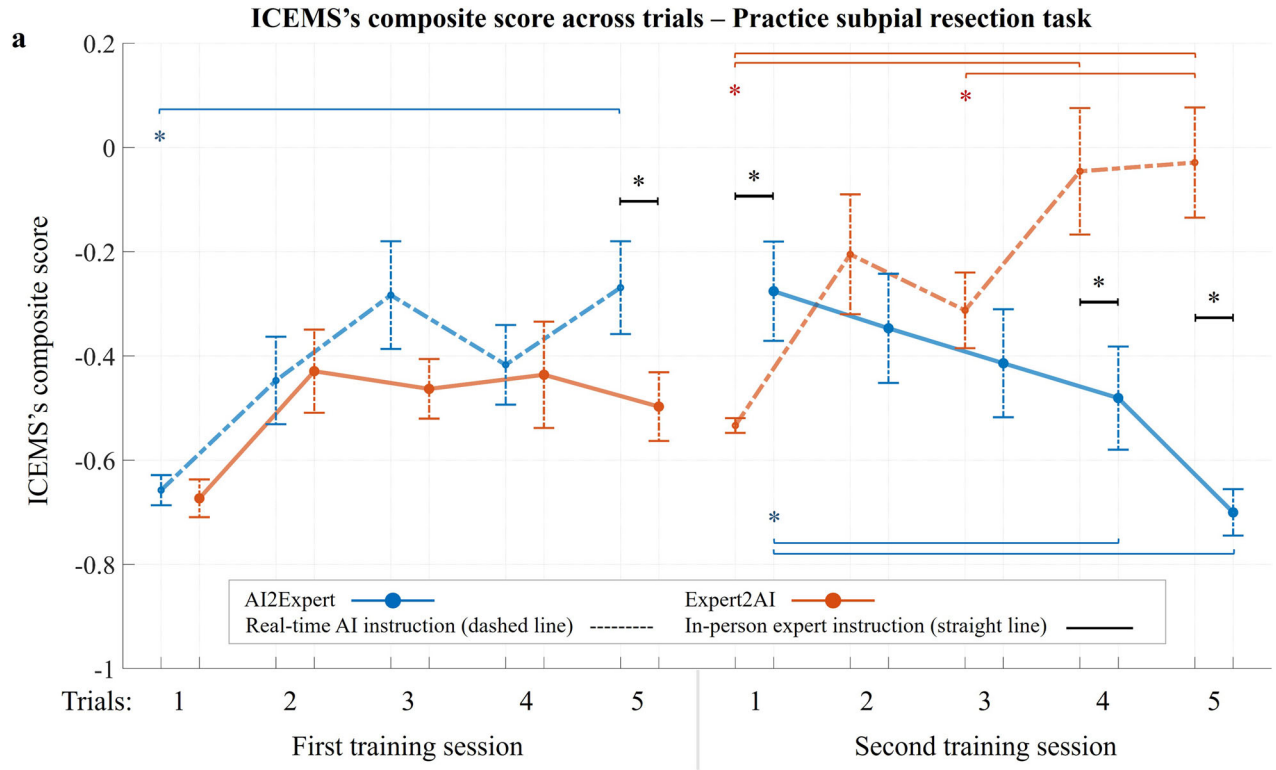
OSATS outcomes (Fig. 2a)

The OSATS ratings on the test task at the end of each training session were analyzed and compared between groups and within groups between sessions. An average score of the two raters was calculated for each task. Superior learning outcomes were observed with the AI feedback in terms of Respect for Tissue, Hemostasis, Economy of Movement, Flow, and the Overall Score in the first session compared to the human-mediated feedback. These significant differences disappeared in the second session, when students who initially received human-mediated feedback also began receiving AI feedback. In the Overall score, students who received real-time AI feedback in the first session (median, 5.25) achieved significantly higher OSATS ratings compared to those who received in-person expert instruction (4.5), with a median difference of 0.75, $z = -2.56$, $p = 0.01$. This

Table 1 | Demographics

	AI2Expert ($n = 12$)	Expert2AI ($n = 13$)	All Participants ($n = 25$)
Number of weeks between training sessions Median (range)	6.5 (6–14)	7 (4–9)	7 (4–14)
Mean age ± SD (range)	21.3 ± 3.7 (19–27)	21.8 ± 2.8 (18–31)	21.6 ± 3.2 (18–31)
Male/Female	5/7	8/5	13/12
Handedness (Right/Left/Ambidextrous)	11/1/0	12/0/1	23/1/1
Year in medical school:			
Preparatory year/1st/2nd/3rd/4th	2/10/0/0/0	3/6/1/3/0	5/16/1/3/0
Level of interest in surgery, median (range)	5 (1–5)	5 (3–5)	4 (1–5)
Completed surgical rotation (Y/N)	0/11	2/11	2/23
Medical School:			
McGill University	5	5	10
University of Montreal	1	7	8
University of Sherbrooke	3	1	4
University of Laval	3	0	3
Playing video games (Y/N)	5/7	9/4	14/11
Playing musical instruments (Y/N)	5/7	6/7	11/14
Previous activities that require hand dexterity	8/4	6/7	14/11
Previously used virtual reality simulation (Y/N)	0/12	0/13	0/25

Study participant demographics.



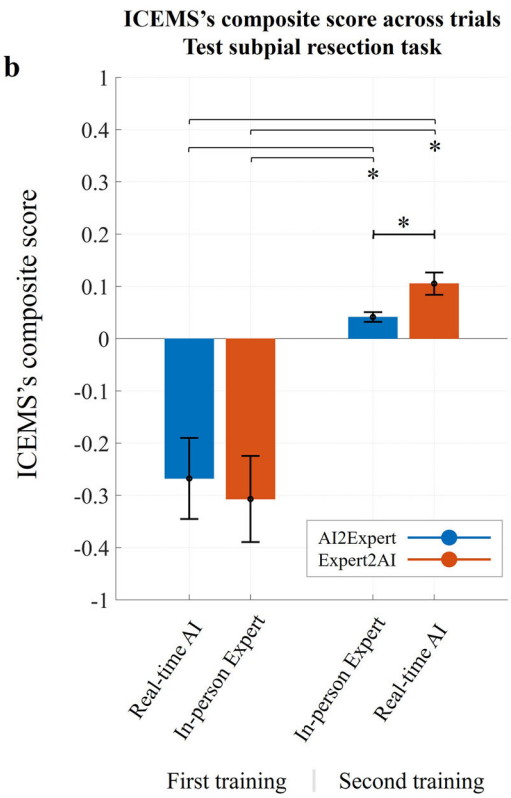


Fig. 1 | ICEMS composite score across tasks. Groups are color-coded. Horizontal lines represent statistically significant differences ($p < 0.05$). Vertical bars represent standard error. **a** ICEMS's composite score across practice subpial resection tasks. **b** ICEMS's composite score across test subpial resection tasks.

significant difference disappeared at the end of the second training session when the groups switched to the next feedback intervention (AI2EXPERT: 4.5 vs 5), 0.5, $z = 1.71$, $p = 0.09$. AI2EXPERT achieved significantly higher scores compared to EXPERT2AI at the end of the first training session in Respect for tissue (5.5 vs 4.5), 1, $z = -2.02$, $p = 0.04$; Hemostasis (6 vs 4), 2, $z = -2.45$, $p = 0.01$; Economy of movement (4.5 vs 3.63), 0.87, $z = -3.05$, $p = 0.002$; and Flow (5.25 vs 4.5), 0.75, $z = -2.69$, $p = 0.006$. Students in EXPERT2AI achieved significant improvement in Instrument Handling (first session: 4.5 vs 5), $z = 2.35$, $p = 0.019$, and Economy of Movement (3.63 vs 4.25), $z = 2.27$, $p = 0.023$ when they switched to real-time AI instruction.

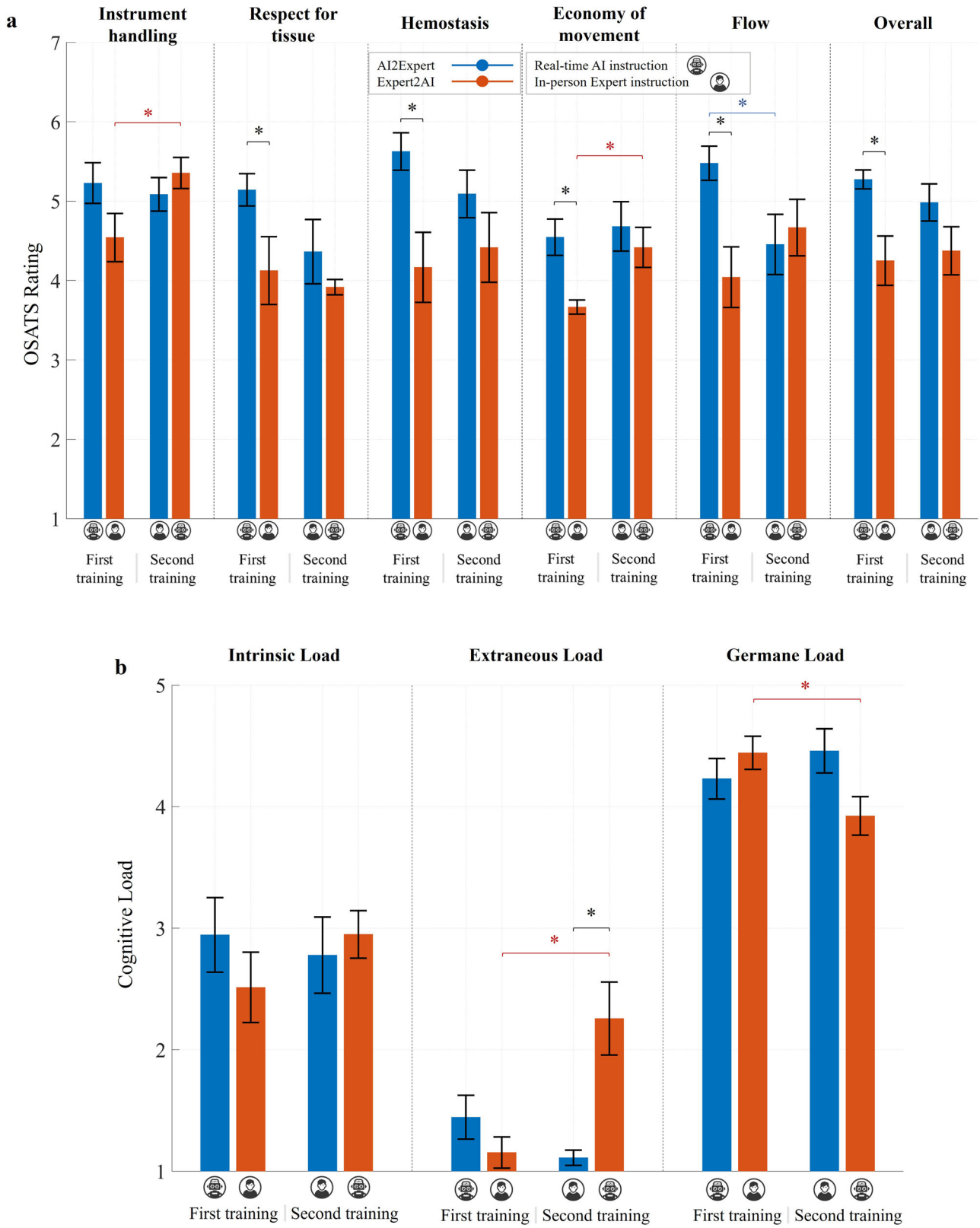


Fig. 2 | OSATS ratings and cognitive load. Groups are color-coded. Horizontal lines represent statistically significant differences ($p < 0.05$). Vertical bars represent standard error. **a** OSATS ratings by blinded experts on the test tasks. **b** Cognitive load during the first and second training sessions.

Students in AI2EXPERT experienced a significant drop in Flow when they switched to in-person human instruction (5.5 vs 5), $z = -2.2$, $p = 0.028$. There was poor agreement between the two raters, with an intraclass correlation coefficient (ICC) value of 0.07.

Specific learning outcomes (Fig. 3)

Specific learning outcomes were assessed across five performance metrics: tissue injury risk, bleeding risk, aspirator force applied, bipolar force applied, and instrument tip separation distance, to outline potential reasons behind

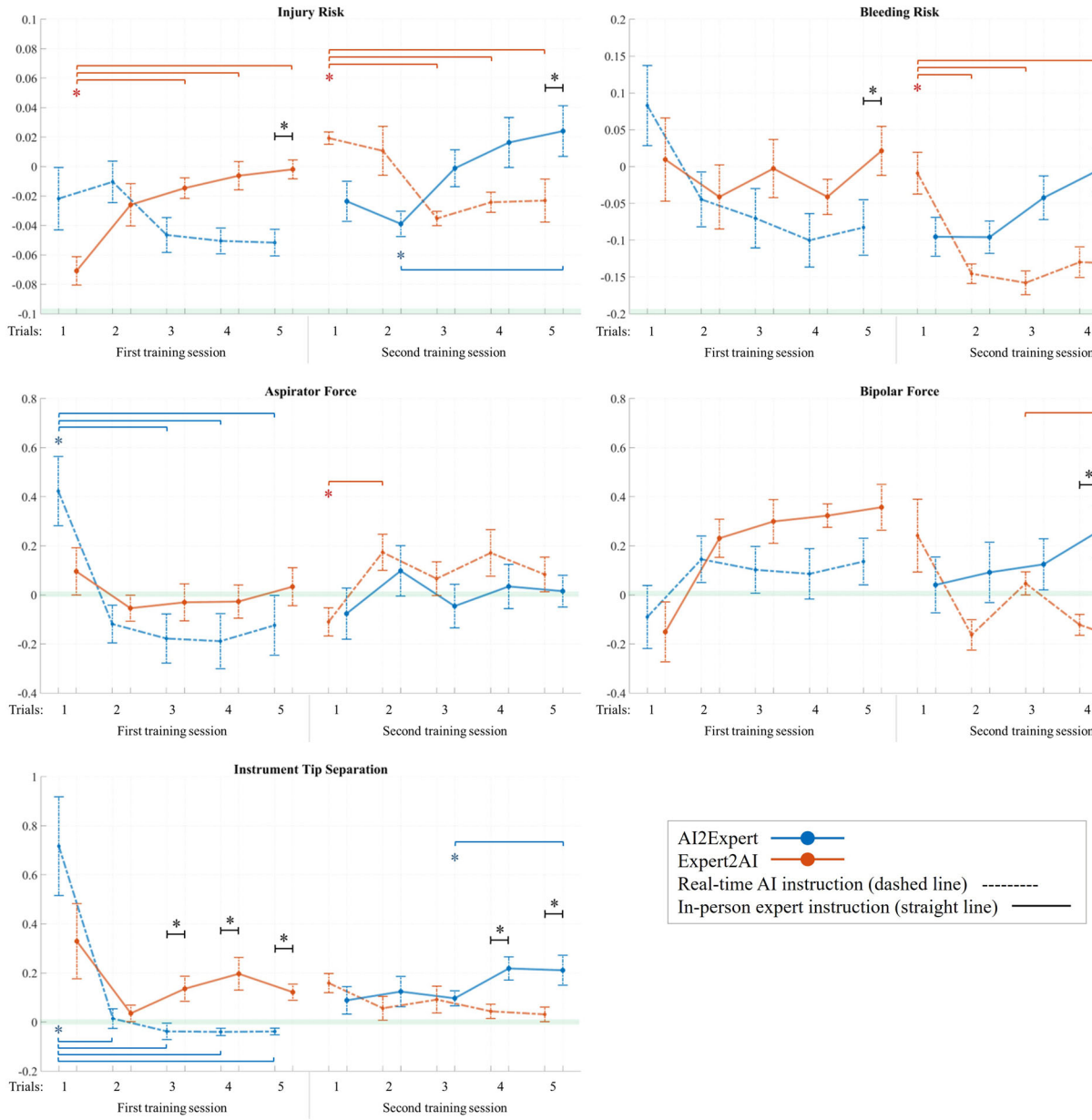


Fig. 3 | Five learning outcomes. Groups are color-coded. Horizontal lines represent statistically significant differences ($p < 0.05$). Black lines represent significant differences between groups. Vertical bars represent standard error. The green area represents the ideal. For risk assessment, this means lower risk. For instrument utilization, this means no (zero) difference from the expert level. Y-axis represents the standard deviation from the mean.

the differences between groups in the overall score and OSATS ratings. These performance metrics were assessed by the ICEMS continuously and an average metric score was calculated for each metric for statistical comparison.

Injury risk

In all tumor resection surgeries, avoidance of injury to adjacent structures is of paramount importance. In brain tumor resection, reducing injury risk is a critical determinant of patient neurologic outcome. Overall, AI feedback resulted in a decreased injury risk score, whereas human-mediated feedback led to an increase in this score. In the first training session, AI2EXPERT with real-time AI intervention achieved significantly lower injury risk scores in the fourth and fifth repetition of the task compared to EXPERT2AI, −0.04, 95%CI [−0.07 −0.02], $p = 0.003$ and −0.05, 95%CI [−0.07 to −0.03], $p < 0.001$, respectively. Without real-time AI intervention, there was a significant increase in the injury risk score for EXPERT2AI by the third repetition of the task which reached a mean difference of 0.07, 95%CI [0.03–0.11], $p = 0.001$ in the fifth repetition of the task when compared to the baseline performance. In the second training session, there was a significant decline in the injury risk score for EXPERT2AI which reached a mean difference of −0.042, 95%CI [−0.09–0.00], $p = 0.043$, at the fifth repetition of the task. Without AI assistance, AI2EXPERT experienced a significant increase in injury risk score from the second to the fifth repetition of the task in the second training session, 0.06, 95%CI [0.00–0.12], $p = 0.032$.

Bleeding risk

Bleeding avoidance is a critical skill for improving patient outcomes. There were no significant changes observed within groups in the first training session. EXPERT2AI had significantly higher bleeding risk in the fifth repetition of the task compared to AI2EXPERT group, 0.10, 95%CI [0.0–0.20], $p = 0.049$. In the second training session, a significant decline was seen in the bleeding risk score for EXPERT2AI which reached a mean difference of the fifth repetition of the task from baseline 0.12, 95%CI [0.02–0.23], $p = 0.03$. There was a significant difference in the bleeding risk

score between the two groups in the fourth repetition of the task in the second training session 0.12, 95%CI [0.05–0.20], $p = 0.002$.

Aspirator force

The use of the surgical aspirator to remove brain tissue requires a balance where insufficient force prevents effective removal of tissue, while excessive force can lead to removal of tissue beyond the target depth. Both feedback interventions resulted in close values to expert level, which was defined as a score of zero (zero difference from the expected expert value in the ICEMS output). There was a significant decline in the aspirator force for AI2EXPERT which reached a mean difference of −0.55 95%CI [−1.08 to −0.02], $p = 0.04$ in the fifth repetition of the task from baseline in the first training session. There was a significant increase in the aspirator force for EXPERT2AI in the second repetition from baseline in the second training session 0.28, 95%CI [0.01 to −0.56], $p = 0.04$. Neither group had a significant change in aspirator force with expert feedback.

Bipolar force

In this paradigm, the bipolar forceps is used to provide visualization by retracting tissue to be aspirated or cauterization. Insufficient force application prevents adequate tissue retraction while excessive force can injure the adjacent brain. There were no significant differences within and between groups in the first training session. Expert feedback resulted in increased bipolar force utilization in the second training session, while AI feedback had the opposite effect. There was a significant decline in bipolar force score in EXPERT2AI from the third repetition of the task to the fifth repetition −0.25, 95%CI [−0.44 to −0.05], $p = 0.009$. There was a significant difference in the bipolar force score between the two groups at the fourth and fifth repetitions of the task in the second training session, 0.38, 95%CI [0.14–0.62], $p = 0.005$ and 0.53, 95%CI [0.27–0.80], $p < 0.001$, respectively.

Instrument coordination

The separation between instruments held in each hand is a key metric of surgical skill[7]: expert surgeons typically work with both instruments close together in a highly coordinated fashion. AI feedback resulted in instrument coordination significantly closer to the expert level when compared to human-mediated feedback. There was a significant decline in instrument tip separation in AI2EXPERT from baseline in the first training session that reached a mean difference of −0.76, 95%CI [−1.4–(−0.06)], $p = 0.03$, in the fifth repetition of the task. EXPERT2AI had no significant changes in the first training session across the five repetitions of the task. EXPERT2AI had a significantly higher instrument tip separation score in the fifth repetition of the task at the end of the first training session, 0.16, 95%CI [0.08–0.24], $p = <0.001$. EXPERT2AI had a significantly lower instrument tip separation distance in the fifth repetition of the task at the end of the second training session.

Cognitive load (Fig. 2b)

It is important to optimize cognitive load to maximize learning without overloading the trainees with redundancy and distractions[13]. Trainees' cognitive load was measured through self-reporting questionnaires (Supplementary Data). Intrinsic load refers to the natural complexity of the task while extraneous load is linked to the unnecessary difficulty in the way the feedback information is delivered. Germane load measures the effort used to integrate new information into knowledge[14]. There were no significant cognitive load differences between groups in the first session. Within-group analysis during the second session revealed that students in EXPERT2AI perceived significantly a higher extraneous load (1 vs 1.67), $z = 3.01$, $p = 0.003$ and a significantly lower germane load (4.25 vs 4), $z = -1.97$, $p = 0.048$, after switching from in-person expert instruction to real-time AI feedback. In the between-group comparison, students in EXPERT2AI reported significantly higher extraneous load compared to those in AI2EXPERT in the second training session (1.67 vs 1), $z = -3.74$, $p < 0.001$, indicating that the feedback provided by the AI system caused significantly higher mental stress.

Discussion

To the best of our knowledge, this study is the first to investigate the surgical skills taught by a real-time AI tutor combined with in-person human instruction over two training sessions in a simulation setting, outlining a comprehensive assessment of learning outcomes assessed by both an AI system and blinded experts' OSATS rating.

In the washout period of 6–8 weeks, our results demonstrated that neither the expert nor the ICEMS group had a significant decrease in skills as compared to their last performance in the first training session, suggesting that skills obtained from both the expert and the ICEMS training are maintained over the two months between trials.

The ICEMS feedback system, including its video feedback component, was designed for students with minimal to no prior knowledge of tumor resection. Given this focus and the depth of instruction an expert can provide, we initially hypothesized that AI feedback would be most beneficial in the first session. However, by the second session, students would likely gain more skills from a human instructor, as human instruction is not constrained in scope. Contrary to our hypothesis, we found AI training to be unexpectedly superior to expert training in both the first and second sessions. Introducing expert instruction early in the training phase was associated with a higher performance score at the end of the second training session. Students exhibited significant decay in skills after switching to expert training in the second session. Although both groups improved on the test task from the first to the second session, EXPERT2AI had significantly greater improvement compared to AI2EXPERT. When evaluated by human raters using the OSATS scale, the OSATS scores were greater in AI2EXPERT compared to EXPERT2AI in the first session and then remained stable in both groups.

Experts are not provided with explicit pedagogical instructions or training and are free to provide focused personalized feedback. AI feedback, however, has been designed and validated to provide consistent, continuous monitoring and feedback that results in both overall and subdomain-specific performance improvement. Unlike AI, experts can use their experience to deconstruct a complex surgical procedure into multiple steps and teach effective strategies and techniques for executing each of these procedural steps. By chunking information, experts provide a general understanding of the subpial tumor resection task, and trainees gain insight into the surgeon's approach to this procedure. The greatest gains from AI feedback, and the most diminished returns from expert feedback, occur in the second session. The AI system is capable of detecting quantitative metrics, that may be imperceptible to humans, thus reducing the benefit of human feedback in more highly trained learners[15]. Taken together, these results suggest that continuous AI monitoring may be highly important for the more experienced learners, and warrant prospective validation in the real operative setting.

Nonetheless, both groups achieved significantly higher performance scores on the test task at the end of their second session compared to those at the end of their first session. Although this work did not involve longitudinal simulation training, the findings suggest that combining real-time AI-instructed and in-person expert-instructed sessions in longitudinal simulation training settings can be beneficial in improving surgical skills in a high-fidelity simulated task. Students may benefit from increased exposure to training, regardless of the type or order of intervention. Further, despite AI2EXPERT's significant decrease in performance scores by the end of the second training session, they still significantly improved their skills on the test task over the two training sessions. These results suggest that AI2EXPERT's expert-mediated learning in the second training session was not reflected in the ICEMS score during the practice sessions but was more reflected during the test task. EXPERT2AI achieved significantly higher performance scores than AI2EXPERT in the test task at the end of the second session. This finding suggests that training with a real-time AI-monitored feedback system following expert-mediated training enhances trainees' transfer of learning.

One argument to the differences between learning outcomes can be the reliance that the students may have developed on either expert or AI

feedback. Therefore, learning is affected by both the presence of one method of instruction and the absence of the other. For unbiased assessment, our study implemented the test task with neither expert nor AI feedback to measure the skills acquired by the students when no feedback is available.

AI tools can monitor surgical performance, which can then inform the surgical educator about the trainee's performance. A hybrid training model would combine the advantages of AI and in-person instruction simultaneously, capturing the instances that may not be obvious to the naked eye and providing them to the expert instructor to guide their instructions. As such, future studies may integrate the ICEMS in assisting surgical educators in delivering optimal feedback to students and compare this hybrid model to either system alone.

Human judgment may not align perfectly with surgical performance assessments made by AI systems. AI versus human-mediated assessments may each have their unique advantages, and different key aspects of surgical performance may have been prioritized in their evaluations. Despite these differences, our results have demonstrated that both scores increase simultaneously as students learn and acquire skills, and that an increase in the ICEMS score does not come at the expense of the skills evaluated by OSATS. Our work provided a detailed report on surgical skills assessment as outlined by the ICEMS on six outcomes. Through comprehensive studies using the data obtained during this RCT, human decisions using OSATS can be compared to the decisions made by the ICEMS. Such investigations may highlight areas unique to human perception, which can inform future AI systems. These systems can be trained based on expert insights and function under their supervision, a process referred to as human-in-the-loop[16]. Human-in-the-loop applications may provide increased interpretability in high-stakes surgical decision-making, continuous improvement, and oversight[17,18].

Although more evidence is needed to establish AI tools as an important component of modern surgical training[19], our findings suggest that integrating intelligent systems with traditional expert mediated apprenticeship would further augment the present competency-based model. Virtual reality simulation provides an immersive medium for trainees to become familiar with the steps, actions, tools, and anatomy relating to a procedure before entering the operating room[20]. Learning these requires understanding optimal techniques and approaches to perform each step successfully. Traditionally, instructional modules in virtual reality simulators have accompanied narrative descriptions or expert demonstrations to achieve this goal but they lacked the capacity for ongoing personalized feedback or performance risk prediction. However, as these systems can become productized in a portable and distributed fashion, the primary rate-limiting step to VR simulation has been the requirement for expert coaching —limiting the global adoption of standalone VR trainers. Our work demonstrates a significant unmet need to use VR systems as an integrated data capture and feedback system to substantially improve surgical performance without human expert intervention.

AI and expert-guided learning may occur concurrently in the future. This study examined how both AI and expert instruction methodologies complement each other, as well as the changes in students' performance in the presence versus absence of each teaching methodology. Our work does not propose a standardized curriculum. Students may always alternate between training interventions based on their individual needs or preferences, as well as the availability of experts. Integrating real-time performance monitoring in residency training requires support from institutions, faculties, and residents. This commitment would involve fostering a culture of continuous improvement and openness to innovative educational methods. The institutional benefits of investing in technological infrastructure must be justified by their added value to reduce operational costs while enhancing education, faculty engagement, and patient outcomes.

Surgical educators play a key role throughout the integration process. Their understanding of program needs, knowledge of training gaps, and critical appraisal of existing evidence enable them to determine the role of real-time AI tutoring in their program's curriculum. They can be the champions that advocate for the integration of intelligent systems in simulation training and investigate their outcomes. In return, residents need to have protected time in their already busy schedules to actively engage with these systems, take part in quality control studies, and complete qualitative surveys.

AI-guided technical training is not limited to the simulation environment. With advances in intraoperative data acquisition, continuous performance monitoring may soon become the norm. It is therefore important to consider the regulatory process required for their implementation in the OR. AI/ML-based software and medical devices pertaining to diagnostic specialties have experienced a leap in regulatory approval by the FDA compared to surgical specialties which require more complex and dynamic AI/ML algorithms due to the nature of real-time intraoperative decision-making[21]. In general, the FDA classifies medical devices based on their potential risks[22]. Class I devices are categorized as low risk and require notification only in terms of their approval pathway, while Class II and Class III are categorized as moderate and high risk and require the 510(k) and premarket approval (PMA) pathways, respectively[22]. The International Medical Device Regulators Forum (IMDRF), guiding the FDA, established that the level of regulatory examination of software as a medical device (SaMD) application and their algorithms should be based on the risk of harm and their clinical evaluation accomplished via a valid clinical association between SaMD output and targeted clinical conditions, analytical validation and clinical validation[23,24]. A recent review of the FDA/CE list failed to identify any approved AI-powered surgical simulation software for appraisal and feedback, however, it is postulated that such devices could potentially be commercially available without any regulatory approval process due to lack of direct device-patient interaction[25]. AI-powered surgical simulators should undergo a regulatory process, potentially designed by the IMDRF and the FDA, to ensure patient safety, transparency, cyber security, and improved quality of training which indirectly impacts patients[25].

In simulation training, the ICEMS system can be further optimized to deliver efficient feedback in a variety of simulations to provide trainees with comprehensive training of the necessary skills. This system can be implemented in intermediary more realistic settings such as animal and placenta models where surgical instruments data are recorded, and relevant real-time feedback can be provided[26–28]. Further investigation is needed to outline whether training using these platforms improve real-life intraoperative skills.

The integration of the ICEMS into the surgical operating room would be possible with access to real-time surgical performance data, which is currently limited. Computer vision could enable intraoperative applications such as surgical video assessment where AI is employed to monitor surgical performance to detect and track surgical instruments[29], recognize surgical actions, phases, and gestures[30], and predict the amount of hemorrhage[31]. The convenient nature of intraoperative cameras may allow the development of smart intraoperative cameras equipped with a system similar to the ICEMS' feedback to monitor surgical performance and provide assistance during surgery to enhance efficiency and patient safety[32,33].

This work involved limitations worth noting. The simulation learning environment may not replicate the stressful nature of real patient cases. Therefore, trainee attention and response to the instructions, as well as learning engagement, may be limited. Both the ICEMS and OSATS ratings evaluate the skillset of the students more than the procedural outcome. Trainees may achieve high ratings on both scoring systems without completely resecting the tumor. In future studies, the ratings of the technical skills students received can be compared to the procedural outcomes such as the amount of tumor removed and spatial information[34,35], to outline whether students utilize correct instrument techniques while achieving desired outcomes.

The participants in this study included 25 medical students. This cohort represents trainees with little to no experience in surgery, similar to surgical trainees at the start of their training. Their lack of knowledge and experience minimized variation in baseline skill levels and allowed for greater room for improvement, resulting in significant differences between

groups even with a smaller sample size. However, more nuanced operative skills may require resident participation to demonstrate the utility of AI platforms in teaching these skills during residency training. Although the current version of the ICEMS system was designed specifically for trainees with no knowledge, such systems also need to be developed and tested for trainees who have some level of experience, helping them to further advance technical skillset mastery.

Technical skills learning may require carefully designed, structured, longitudinal curricula. Integration of AI into surgical training may yield significant benefits when combined with in-person expert-guided instructions. The future of surgical training may need optimal methodologies for AI integration, achieving the best learning outcomes with long-term retention and successful skill transfer.

Methods

This multi-institutional instructor-blinded crossover randomized trial was approved by McGill University Health Centre Research Ethics Board, Neurosciences–Psychiatry. The inclusion criterion for participants was enrollment in a medical program in Canada. The exclusion criterion was prior participation in the virtual reality simulator trial, the NeuroVR, used in this study. All participants provided informed consent prior to trial participation. This report follows the Consolidated Standards of Reporting Trials involving AI (CONSORT-AI)[36,37]. This trial was conducted as an extension of trial registration: NCT05168150, clinicaltrials.gov. The trial started in February 2022 and ended in April 2022 after recruiting all students who registered to participate. The study involved no harm to participants.

Simulation training

Participants received training across two sessions in a controlled laboratory environment at the Neurosurgical Simulation and Artificial Intelligence Learning Centre, McGill University. After providing written consent, participants reviewed standardized instructions on how to use the virtual reality simulator and completed a background information questionnaire. They performed simulated brain tumor resection tasks using the NeuroVR virtual reality simulator (CAE Healthcare, Montreal, Canada). The NeuroVR is a high-fidelity simulator that recreates the auditory, visual, and haptic experience of interacting with brain tissues realistically and enables repetitive practice of selected neurosurgical tasks[20,38]. At each training session, participants first completed a practice brain tumor resection task five times. After the practice sessions, they performed a test task, a brain tumor resection that included lifelike bleeding and tissue modeling[38], for summative assessment. They had 5 min to complete each practice resection and 13 min for the test task. The time limit was defined based on the previous feedback from neurosurgeons and trainees[9,34]. Participants were randomly allocated into two cross-over groups to receive a study intervention (described below) during practice tasks. Upon completion of the first training session, participants were asked to return within 6–8 weeks to undergo the same training with the alternate intervention. The initial practice task served as the baseline.

Study interventions

Participants in both interventions received real-time instructions during the 5 min they performed the resection and, in addition, had an extra 5 min to receive post-hoc debriefing and feedback between each attempt. The first group, AI2EXPERT ($n = 12$), received AI instruction during their first training session followed by expert feedback in the second training session. Conversely, the second group, EXPERT2AI ($n = 13$), started with expert feedback and received AI feedback in the second training session. The specific educational content of each intervention is described below.

AI instruction

In this phase, participants received real-time auditory instructions given by the ICEMS – a multialgorithm intelligent tutoring system trained on time-series data using a deep learning model known as a long short-term memory network[5]. The ICEMS's ability of performance assessment, risk detection, and coaching along with its predictive validity on trainee performance were previously demonstrated[5]. This model evaluates surgical performance, predicts risk five times per second, and can provide auditory instructions to change participants' behavior.

In this study, the ICEMS system had five teaching objectives: minimizing the risk of (1) bleeding and (2) healthy tissue damage, while optimizing (3) dominant hand force, (4) non-dominant hand force, and (5) bimanual coordination. If the difference between the participant's score and the predicted expert-level score (by the ICEMS) exceeded a predefined threshold (0.5 for risk metrics and 1 for coaching metrics), an error was identified, and the system automatically delivered real-time verbal warnings or instructions. Participants received ongoing tutoring from the ICEMS in all practice tasks except for the baseline task.

ICEMS training also included personalized automated post-hoc feedback. Following each practice task, including the baseline, the algorithm selected an error footage from the video recording of the participant's performance, categorized it based on the learning objective, and presented it as a 10-s clip followed by a metric-specific instructional video including an expert demonstration of that objective.

Expert instruction

A panel of two expert instructors, highly experienced in simulation training and brain tumor resection surgery, was created. One instructor was present in person during the practice tasks, except the baseline, to instruct each participant in the expert instruction group. To ensure standardized instruction, the instructors completed a workshop and utilized a modified Promoting Excellence and Reflective Learning in Simulation (PEARLS) debriefing script[39]. Throughout the practice tasks, the expert instructor provided verbal feedback to the students, using their judgment and expertise to make recommendations. Following each practice task, the expert demonstrated relevant strategies and techniques on the simulation as they considered appropriate.

Outcome measures

The primary outcome measure was the composite performance score assessed by the ICEMS for practice and test tasks, averaged across each task. Our secondary outcome measures were the ratings on six OSATS items on the test task, including 5-items and an overall score. The OSATS ratings were provided by experts who were blinded to the participant group and ICEMS assessment. Additionally, scores on 5 ICEMS metrics were tracked and analyzed. Finally, cognitive load, including intrinsic, extraneous, and germane load, was measured through questionnaires.

Randomization and sample size

Students were randomized into 2 intervention sequences with an allocation ratio of 1:1, using an internet-based random number generator without stratification[40]. Group allocation was concealed by the study coordinator, and instructors were notified of appointment times in advance for scheduling purposes. The participant recruitment flowchart is outlined in Fig. 4. Using the study primary outcome, the ICEMS score during the practice tasks across two training sessions, we conducted a power analysis. With a power of 0.95, an effect size of 0.3, and a correlation of 0.5 among repeated measures, the analysis yielded a total sample size of 22 participants (11 participants per group) for the between-within-group interaction in a two-group randomized crossover trial[41,42]. Analysis was conducted based on intention-to-treat.

Statistical analysis

The interaction between the ICEMS score and the trials at each training session was assessed by a two-way mixed model ANOVA. Shapiro Wilk test was used to observe normality ($p > 0.05$). Outliers were imputed using the nearest non-outlier value[43]. The composite score across five repetitions of the practice task was compared using one-way repeated measures ANOVA or Friedman's test depending on normality. Mauchly's test indicated that the assumption of sphericity was met for two-way

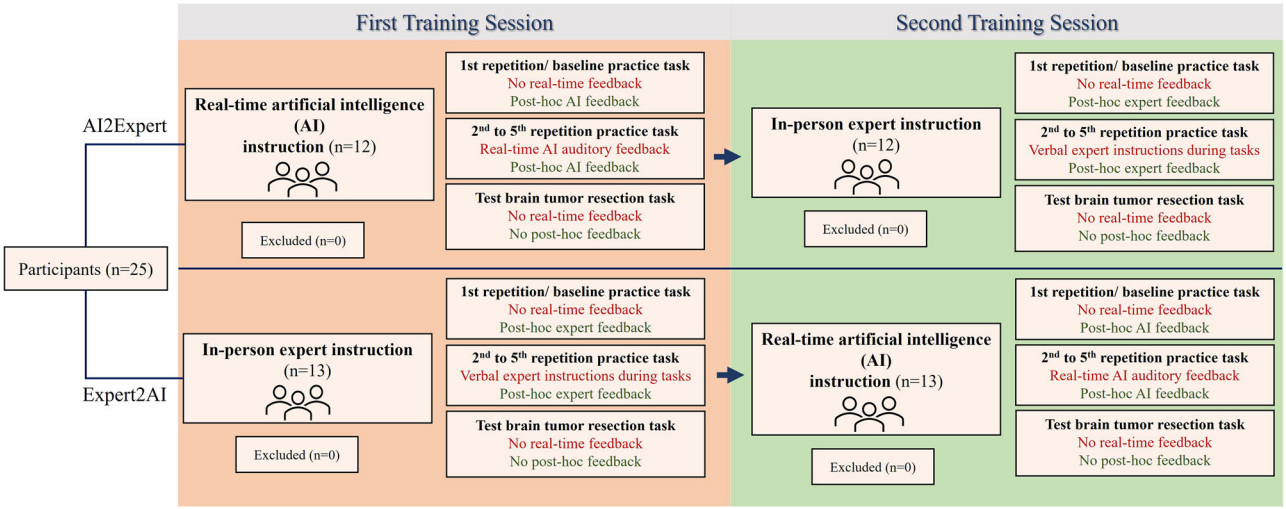


Fig. 4 | Flowchart. Study interventions and flowchart.

interaction for the first training session, $\chi2(9) = 12.58$, $p = 0.184$, and it was violated for the second training session $\chi2(9) = 20.86$, $p = 0.014$. Values with Greenhouse-Geisser correction were reported for violation of the assumption of sphericity for two-way interactions and repeated measure analyses. Between-group comparisons at each repetition of the task and the composite score on the test task were done using independent samples t-test. Within-group differences between 1st and 2nd training sessions for the performance on the test task were analyzed using paired samples t-tests. Levene's test showed heterogeneity of variances, based on median ($p < 0.05$), and Box's test demonstrated violation of homogeneity of covariances, $p < 0.001$. Pairwise comparisons were adjusted by Bonferroni correction for multiple tests. Assessment of OSATS scores was made using non-parametric tests: Mann–Whitney U test for between-group comparison, and Wilcoxon test for within-group comparisons between sessions. IBM SPSS Statistics, Version-27 was used to conduct statistical analyses. Figures were created using MATLAB, version R2023b (MathWorks, Natick, MA).

Data availability

The dataset is available from the corresponding author on a reasonable request. A sample raw simulation data file can be accessed online.

Code availability

The codes used in this study are available from the corresponding author on a reasonable request.

Received: 29 July 2024; Accepted: 7 August 2025;
Published online: 03 November 2025

References

1. Stephens, E. H. & Dearani, J. A. On becoming a master surgeon: role models, mentorship, coaching, and apprenticeship. *Ann. Thorac. Surg.* **111**, 1746–1753, https://doi.org/10.1016/j.athoracsur.2020.06.061 (2021).
2. Stulberg, J. J. et al. Association between surgeon technical skills and patient outcomes. *JAMA Surg*. https://doi.org/10.1001/jamasurg.2020.3007 (2020).
3. Fecso, A. B., Szasz, P., Kerezov, G. & Grantcharov, T. P. The effect of technical performance on patient outcomes in surgery. *Ann. Surg.* **265**, 492–501, https://doi.org/10.1097/SLA.0000000000001959 (2017).
4. Leape, L. L. et al. The nature of adverse events in hospitalized patients. *N. Engl. J. Med.* **324**, 377–384, https://doi.org/10.1056/nejm199102073240605 (1991).
5. Yilmaz, R. et al. Continuous monitoring of surgical bimanual expertise using deep neural networks in virtual reality simulation. *npj Digital Med.* **5**, 54. https://doi.org/10.1038/s41746-022-00596-8 (2022).
6. Winkler-Schwartz, A. et al. Machine learning identification of surgical and operative factors associated with surgical expertise in virtual reality simulation. *JAMA Netw. Open* **2**, e198363, https://doi.org/10.1001/jamanetworkopen.2019.8363 (2019).
7. Mirchi, N. et al. The virtual operative assistant: an explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLOS ONE* **15**, e0229596, https://doi.org/10.1371/journal.pone.0229596 (2020).
8. Fazlollahi, A. M., et al. Effect of artificial intelligence tutoring vs expert instruction on learning simulated surgical skills among medical students: a randomized clinical trial. *JAMA Netw. Open* **5**, e2149008, https://doi.org/10.1001/jamanetworkopen.2021.49008 (2022).
9. Yilmaz, R. et al. Effect of feedback modality on simulated surgical skills learning using automated educational systems– a four-arm randomized control trial. *J. Surg. Educ.* **81**, 275–287, https://doi.org/10.1016/j.jsurg.2023.11.001 (2024).
10. Fazlollahi, A. M. et al. AI in surgical curriculum design and unintended outcomes for technical competencies in simulation training. *JAMA Netw. Open* **6**, e2334658–e2334658, https://doi.org/10.1001/jamanetworkopen.2023.34658 (2023).
11. Chang, A., Moreno, T., Feaster, W. & Ehwerhemuepha, L. Towards artificial and human intelligence in hybrid healthcare. In *Hybrid Healthcare* (eds. Al-Razouki, M. & Smith, S.) 7-16 (Springer International Publishing, 2022).
12. Jarrahi, M. H., Lutz, C. & Newlands, G. Artificial intelligence, human intelligence and hybrid intelligence based on mutual augmentation. *Big Data Soc.* **9**, 20539517221142824, https://doi.org/10.1177/20539517221142824 (2022)
13. Sweller, J. Cognitive load during problem solving: Effects on learning. *Cognit. Sci.* **12**, 257–285, https://doi.org/10.1016/0364-0213(88)90023-7 (1988).
14. Sweller, J. Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educ. Psychol. Rev.* **22**, 123–138, https://doi.org/10.1007/s10648-010-9128-5 (2010).
15. Rothermel, L. D. & Lipman, J. M. Estimation of blood loss is inaccurate and unreliable. *Surgery* **160**, 946–953, https://doi.org/10.1016/j.surg.2016.06.006 (2016).
16. Wu, X. et al. A survey of human-in-the-loop for machine learning. *Future Gener. Comput. Syst.* **135**, 364–381, https://doi.org/10.1016/j.future.2022.05.014 (2022).

17. Rudin, C. Why black box machine learning should be avoided for high-stakes decisions, in brief. *Nat. Rev. Methods Prim.* **2**, 81, https://doi.org/10.1038/s43586-022-00172-0 (2022).
18. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215, https://doi.org/10.1038/s42256-019-0048-x (2019).
19. Guerrero, D. T., Asaad, M., Rajesh, A., Hassan, A. & Butler, C. E. Advancing surgical education: the use of artificial intelligence in surgical training. *Am. Surg.* **89**, 49–54, https://doi.org/10.1177/00031348221101503 (2022).
20. Delorme, S., Laroche, D., DiRaddo, R. & Del Maestro, R. F. NeuroTouch: a physics-based virtual simulator for cranial microneurosurgery training. *Operative Neurosurg.* **71**, ons32–ons42, https://doi.org/10.1227/NEU.0b013e318249c744 (2012).
21. Gupta, A. et al. Artificial intelligence: a new tool in surgeon's hand. *J. Educ. Health Promot* **11**, 93, https://doi.org/10.4103/jehp.jehp_625_21 (2022).
22. Sastry, A. Overview of the US FDA medical device approval process. *Curr. Cardiol. Rep.* **16**, 494. https://doi.org/10.1007/s11886-014-0494-3 (2014).
23. U.S. Food and Drug Administration. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). Department of Health and Human Services (United States). https://apo.org.au/node/228371 (2019).
24. Larson, D. B. et al. Regulatory frameworks for development and evaluation of artificial intelligence–based diagnostic imaging algorithms: summary and recommendations. *J. Am. Coll. Radiol.* **18**, 413–424, https://doi.org/10.1016/j.jacr.2020.09.060 (2021).
25. Park, J. J., Tiefenbach, J. & Demetriades, A. K. The role of artificial intelligence in surgical simulation. *Front. Med. Technol.* **4**, 1076755 (2022).
26. Almansouri, A. et al. Continuous instrument tracking in a cerebral corticectomy ex vivo calf brain simulation model: face and content validation. *Operat. Neurosurg.* https://doi.org/10.1227/ons.0000000000001044 (2024)
27. Winkler-Schwartz, A. et al. Creating a comprehensive research platform for surgical technique and operative outcome in primary brain tumor neurosurgery. *World Neurosurg.* **144**, e62–e71, https://doi.org/10.1016/j.wneu.2020.07.209 (2020).
28. Oliveira, M. M. et al. Face, content, and construct validity of brain tumor microsurgery simulation using a human placenta model. *Operat. Neurosurg.* **12**, 61–67, https://doi.org/10.1227/neu.0000000000001030 (2016).
29. Wang, Y., Sun, Q., Liu, Z. & Gu, L. Visual detection and tracking algorithms for minimally invasive surgical instruments: A comprehensive review of the state-of-the-art. *Robot. Autonomous Syst.* **149**, 103945, https://doi.org/10.1016/j.robot.2021.103945 (2022).
30. Garrow, C. R. et al. Machine learning for surgical phase recognition: a systematic review. *Ann. Surg.* **273**, 684–693 (2021)
31. Pangal, D. J. et al. Expert surgeons and deep learning models can predict the outcome of surgical hemorrhage from 1 min of video. *Sci. Rep.* **12**, 8137. https://doi.org/10.1038/s41598-022-11549-2 (2022).
32. Birkhoff, D. C., van Dalen, A. S. H. M. & Schijven, M. P. A review on the current applications of artificial intelligence in the operating room. *Surg. Innov.* **28**, 611–619, https://doi.org/10.1177/1553350621996961 (2021).
33. Kiyasseh, D. et al. A multi-institutional study using artificial intelligence to provide reliable and fair feedback to surgeons. *Commun. Med.* **3**, 42, https://doi.org/10.1038/s43856-023-00263-3 (2023).
34. Yilmaz, R. et al. Nondominant hand skills spatial and psychomotor analysis during a complex virtual reality neurosurgical task-a case series study. *Oper. Neurosurg.***23**, 22–30, https://doi.org/10.1227/ons.0000000000000232 (2022).
35. Alotaibi, F. E. et al. Neurosurgical assessment of metrics including judgment and dexterity using the virtual reality simulator NeuroTouch (NAJD Metrics). *Surg. Innov.* **22**, 636–642, https://doi.org/10.1177/1553350615579729 (2015).
36. Liu, X. et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.* **26**, 1364–1374, https://doi.org/10.1038/s41591-020-1034-x (2020).
37. Cheng, A. et al. Reporting guidelines for health care simulation research: extensions to the CONSORT and STROBE statements. *Adv. Simul.* **1**, 25. https://doi.org/10.1186/s41077-016-0025-y (2016).
38. Sabbagh, A. J. et al. Roadmap for developing complex virtual reality simulation scenarios: subpial neurosurgical tumor resection model. *World Neurosurg.* **139**, e220–e229, https://doi.org/10.1016/j.wneu.2020.03.187 (2020).
39. Eppich, W. & Cheng, A. Promoting Excellence and Reflective Learning in Simulation (PEARLS): development and rationale for a blended approach to health care simulation debriefing. *Simul. Healthcare* **10**, 106–115 (2015)
40. random.org. https://www.random.org/ Accessed January 1, 2024.
41. Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. G. Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**, 175–191, https://doi.org/10.3758/BF03193146 (2007).
42. Cohen, J. *Statistical power analysis for the behavioral sciences* (Routledge, 2013).
43. Tukey, J. W. *Exploratory data analysis* (Reading/Addison-Wesley, 1977)

Acknowledgements

The authors would like to thank the medical students who participated in this study. This work was supported by a Medical Education Research Grant from the Royal College of Physicians and Surgeons of Canada, a grant from the Fonds de recherche du Quebec–Sante for doctoral training, a Max Binz Fellowship from McGill University Internal Studentships, a Brain Tumor Foundation of Canada Brain Tumor Research Grant, along with the Franco Di Giovanni Foundation, and the Montreal Neurological Institute and Hospital. The National Research Council of Canada, Boucherville, Quebec, Canada provided a prototype of the NeuroVR used in this study. Portions of this work were presented at the 2023 Congress of Neurological Surgeons, Washington, D.C., USA, on September 10-13, 2023, and at the Annual Neurosurgery Research Day, at the Montreal Neurological Institute, Montreal, Canada on May 26, 2023.

Author contributions

R.Y. contributed to the design and conceptualization of the study, participant recruitment, methodology, development and implementation of the intelligent system, video feedback, data analysis, writing the original draft, critical revision of the manuscript for important intellectual content, statistical analysis, and visualization, and wrote all the codes used in this study. A.A. and M.B. contributed to conducting in-person expert-mediated training, critical revision of the manuscript for important intellectual content, and blinded expert OSATS rating of video performance. A.M.F. contributed to the conceptualization of the study, methodology, writing the original draft, participant recruitment, and critical revision of the manuscript for important intellectual content. N.A.H. contributed to the conceptualization of the study, development of the real-time feedback system, participant recruitment, and critical revision of the manuscript for important intellectual content. T.T. contributed to methodology, statistical analysis, writing the original draft, and critical revision of the manuscript for important intellectual content. A.S. contributed to writing the original draft, and critical revision of the manuscript for important intellectual content. D.L. contributed to the development of real-time data transfer from the simulator, technical assistance to ensure the proper work of the simulator, and critical revision of the manuscript for

important intellectual content. C.S. contributed to the conceptualization of the study, critical revision of the manuscript for important intellectual content, obtained funding, and supervision. D.A.D. contributed to the conceptualization of the study, writing the original draft, critical revision of the manuscript for important intellectual content. R.F.D.M. contributed to the conceptualization of the study, methodology, development, writing the original draft, critical revision of the manuscript for important intellectual content, obtained funding, administration, and supervision.

Competing interests

The authors declare no competing interests. The funding sources of this study were not involved in the design or conduct of the study; the collection, analysis, or interpretation of the data; the preparation or approval of the manuscript; or the decision to submit the manuscript for publication.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44387-025-00032-8.

Correspondence and requests for materials should be addressed to Recai Yilmaz.