JAMA Surgery | Original Investigation | AI IN SURGERY

Artificial Intelligence–Augmented Human Instruction and Surgical Simulation Performance
A Randomized Clinical Trial

Bianca Giglio, MSc; Abdulmajeed Albeloushi, MD; Ahmad Kh. Alhaj, MD; Mohamed Alhantoobi, MD, MSc; Rothaina Saeedi, MD; Vanja Davidovic, BHSc; Abicumaran Uthamacumaran, BSc; Recai Yilmaz, MD, PhD; Jason Lapointe, DEC; Neevya Balasubramaniam, MD; Trisha Tee, MSc; Ali M. Fazlollahi, MD, MSc; José A. Correa, PhD; Rolando F. Del Maestro, MD, PhD

IMPORTANCE How the Intelligent Continuous Expertise Monitoring System, an artificial intelligence tutoring system, might be best optimized for surgical training is unknown.

OBJECTIVE To determine the effects of artificial intelligence–augmented personalized expert instruction vs intelligent tutoring alone on surgical performance, skill transfer, and affective-cognitive responses.

DESIGN, SETTING, AND PARTICIPANTS This single-blinded randomized clinical trial was conducted among a volunteer sample of medical students in preparatory, first, or second year without prior use of a virtual reality surgical simulator (NeuroVR) at the McGill Neurosurgical Simulation and Artificial Intelligence Learning Centre in Montreal, Quebec, Canada. Cross-sectional data were collected from March to September 2024, and per-protocol data analysis was conducted in March 2025.

INTERVENTION During simulated surgical procedures, trainees received 1 of 3 feedback methods. Group 1 received only intelligent tutor instruction (control). The 2 intervention arms included group 2, which received expert feedback in identical words to the intelligent tutor, and group 3, which received artificial intelligence data–informed personalized expert feedback.

MAIN OUTCOMES AND MEASURES The coprimary outcomes included change in overall surgical performance across practice resections and skill transfer to a complex realistic scenario, measured by artificial intelligence–calculated composite expertise score (range, −1.00 [novice] to 1.00 [expert]). Secondary outcomes included emotional and cognitive demands, measured via questionnaires.

RESULTS In this randomized clinical trial, the final analysis included 87 medical students (46 [53%] women; mean [SD] age, 22.7 [4.0] years), with 30, 29, and 28 participants in groups 1, 2, and 3, respectively. Group 3 achieved significantly higher scores than group 1 across several trials, including trial 5 (mean difference, 0.26; 95% CI, 0.09-0.43; $P = .01$) and the realistic task (mean difference, 0.20; 95% CI, 0.06-0.34; $P = .02$). Group 3 also achieved significantly better scores than the other 2 groups in certain metrics, such as bleeding and injury risk. Emotions and cognitive load demonstrated significant differences.

CONCLUSIONS AND RELEVANCE In this randomized clinical trial, personalized expert instruction resulted in enhanced surgical performance and skill transfer compared with intelligent tutor instruction, highlighting the importance of human input and participation in artificial intelligence–based surgical training.

TRIAL REGISTRATION ClinicalTrials.gov Identifier: NCT06273579

Visual Abstract
Invited Commentary
Supplemental content

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Bianca Giglio, MSc, Neurosurgical Simulation and Artificial Intelligence Learning Centre, Department of Neurology and Neurosurgery, Montreal Neurological Institute and Hospital, McGill University, 300 Rue Léo-Pariseau, Ste 2210, Montreal, QC H2X 4B3, Canada (bianca.giglio@mail.mcgill.ca).

JAMA Surg. doi:10.1001/jamasurg.2025.2564
Published online August 6, 2025.

Although expert surgical technical skill is linked with improved patient outcomes, training novices to master these skills remains challenging.[1-3] Current surgical teaching models lack standardization[4-7] and rely on qualitative performance assessments by human experts rather than quantitative performance data.[8] Artificial intelligence (AI) tutoring systems have the potential to address these shortcomings due to their ability to process and analyze large, complex datasets, exceeding human capacity for pattern recognition.[9-13] The goal of these technologies is to create standardized AI-enhanced surgical curricula to improve trainee bimanual skills, thereby achieving better patient outcomes.[14-19]

In a randomized clinical trial (RCT), the Virtual Operative Assistant (VOA) intelligent tutoring system effectively augmented surgical performance on a virtual reality (VR) simulator via post hoc AI-selected metric feedback.[9,10] The VOA lacks the ability to assess real-time surgical performance and deliver continuous intraoperative instruction, limiting its educational utility in the dynamic operating room environment. The Intelligent Continuous Expertise Monitoring System (ICEMS) addresses the necessity for real-time application by using a multi-algorithm approach to assess bimanual surgical skills at 0.2-second intervals and provide continuous, action-oriented verbal feedback.[11] Built based on quantifiable, AI-derived metrics that enable continuous performance scored from −1.00 (novice) to 1.00 (expert),[11] the ICEMS demonstrates explainability and transparency critical to educator and learner engagement.[20-22] The ICEMS can be integrated into any VR surgical simulator, including the NeuroVR (CAE Healthcare). This system has been validated for its ability to accurately differentiate surgical expertise levels, track skill acquisition throughout a neurosurgical training program,[11] and serve as a pedagogical tool for risk assessment, coaching, and error detection.[12]

Another RCT demonstrated that ICEMS feedback yielded enhanced learning outcomes compared with expert feedback during a simulated surgical task.[12] Instructors in this study were blinded to the ICEMS error data and depended on qualitative observation rather than the quantitative evaluations offered by the ICEMS. A cohort study investigating VR surgical skill acquisition found that an AI-enhanced curriculum resulted in unintended consequences that negatively impacted some efficiency metrics, indicating a potential necessity for human expert input.[13] A randomized crossover trial assessed the effect of using both ICEMS and expert instruction methodologies in succession and found that ICEMS feedback significantly improved surgical performance following expert instruction.[23] These results suggest that AI-enhanced curricula may benefit from collaboration between human educators and intelligent tutors.

This study aimed to investigate the effect of AI-augmented human instruction—where human surgical educators were provided with quantitative ICEMS performance data—on learners' technical skill acquisition during simulation training. We hypothesized that expert instructors supported by quantitative AI data to deliver continuous personalized instruction would be more effective at improving learning and transfer of surgical technical skills among trainees compared with AI instructors, while also resulting in lower negative emotions and cognitive load.[24,25]

Key Points

Question Does artificial intelligence–augmented personalized expert instruction improve surgical performance, skill transfer, and affective-cognitive responses compared to intelligent tutoring alone?

Findings In this randomized clinical trial of 88 medical students, trainees achieved significantly higher performance scores when tutored by a human educator providing personalized feedback based on artificial intelligence error data than by an intelligent tutor alone.

Meaning Providing human educators with artificial intelligence performance data to tailor feedback improves learning outcomes in surgical simulation training.

Methods

This parallel-design, single-blinded, 3-arm RCT, approved by the McGill University Health Centre Research Ethics Board, was registered at ClinicalTrials.gov on February 16, 2024 (NCT06273579), and follows the Consolidated Standards of Reporting Trials-Artificial Intelligence (CONSORT-AI)[26] guideline and the Machine Learning to Assess Surgical Expertise checklist.[27] Participants provided written informed consent. The trial protocol and statistical analysis plan are available in Supplement 1.

Participants

Participants were recruited between March and September 2024 for a single 90-minute surgical simulation session at the Neurosurgical Simulation and Artificial Intelligence Learning Centre in Montreal, Quebec, Canada (**Table**). A sample size calculation for a repeated measures analysis of variance (ANOVA) with a between-participants factor was conducted using G*Power version 3.1 (Heinrich-Heine-Universität Düsseldorf).[28] A power of 0.9, an effect size of 0.3, an α error probability of 0.05, and a correlation among repeated measures of 0.5[12] yielded a total of 87 participants, with 29 participants in each of 3 groups. Volunteer sampling was used to attain the desired sample size. Recruitment information was disseminated via student groups, social media, and word of mouth. Inclusion criteria consisted of enrollment in preparatory, first, or second year at 1 of 4 Quebec medical schools. The exclusion criterion was previous NeuroVR experience.

Randomization

Students were stratified based on their year in medical school and block randomized to 1 of 3 intervention arms with an allocation ratio of 1:1:1 using random number sequences generated by Random.org.[29] The participant recruitment flowchart is outlined in **Figure 1**.

Table. Demographic Characteristics of Included Study Participants

Characteristic	No. (%)			
	Group 1: AI tutor instruction (n = 30)	Group 2: expert instruction (n = 29)	Group 3: personalized expert instruction (n = 28)	All participants (N = 87)
Age, mean (SD), y	21.8 (2.4)	22.6 (4.4)	23.9 (4.8)	22.7 (4.0)
Sex				
Female	18 (60)	16 (55)	12 (43)	46 (53)
Male	12 (40)	13 (45)	15 (54)	40 (46)
Prefer not to say	0	0	1 (4)	1 (1)
Gender				
Woman	18 (60)	16 (55)	12 (43)	46 (53)
Man	12 (40)	13 (45)	15 (54)	40 (46)
Prefer not to say	0	0	1 (4)	1 (1)
Undergraduate medical training level				
Preparatory	9 (30)	8 (28)	8 (29)	25 (29)
First	15 (50)	14 (48)	13 (46)	42 (48)
Second	6 (20)	7 (24)	7 (25)	20 (23)
Institution				
McGill University	11 (37)	15 (52)	14 (50)	40 (46)
Université de Montréal	12 (40)	7 (24)	6 (21)	25 (29)
Université de Sherbrooke	4 (13)	6 (21)	7 (25)	17 (20)
Université Laval	3 (10)	1 (3)	1 (4)	5 (6)
Handedness				
Right	28 (93)	25 (86)	24 (86)	77 (89)
Left	2 (7)	3 (10)	4 (14)	9 (10)
Ambidextrous	0	1 (3)	0	1 (1)
Interest in pursuing surgery, mean (SD)[a]	4.0 (0.9)	4.1 (1.0)	3.9 (1.0)	4.0 (1.0)
Completed surgical rotation, clerkship, or shadowing				
Yes	12 (40)	10 (34)	11 (39)	33 (38)
No	18 (60)	19 (66)	17 (61)	54 (62)
Plays video games				
Yes	8 (27)	9 (31)	13 (46)	30 (34)
No	22 (73)	20 (69)	15 (54)	57 (66)
Played musical instruments in last 5 y				
Yes	9 (30)	9 (31)	13 (46)	30 (34)
No	21 (70)	20 (69)	15 (54)	56 (64)
Participated in activities that require hand dexterity				
Yes	8 (27)	12 (41)	11 (39)	31 (36)
No	22 (73)	17 (59)	17 (61)	56 (64)
Previously used VR surgical simulation				
Yes	1 (3)	2 (7)	5 (18)	8 (9)
No	29 (97)	27 (93)	23 (82)	79 (91)

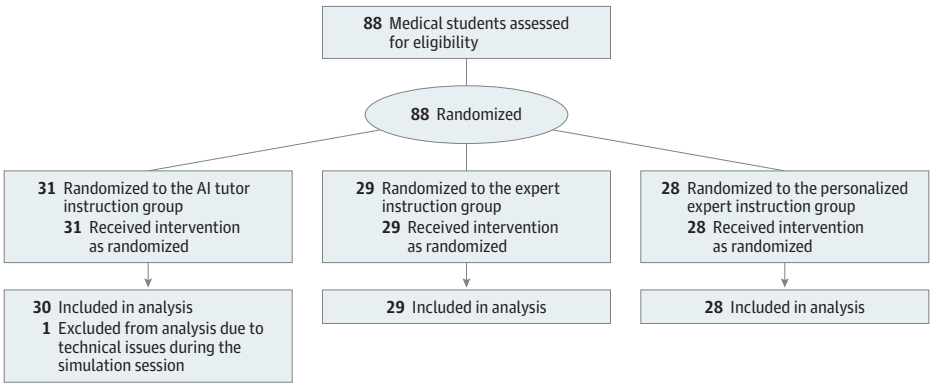
Abbreviations: AI, artificial intelligence; VR, virtual reality.

[a] Self-reported on a 5-point Likert scale, with 1 indicating less interest and 5 indicating more interest.

Simulation Session

All tasks were performed on the NeuroVR, a surgical simulator that simulates a subpial brain tumor resection procedure in a 3-dimensional VR environment.[30] The session consisted of 2 scenarios that have demonstrated face, content, and construct validity[11,16,31]: (1) a practice subpial resection task (eFigure 1 in Supplement 2) and (2) a realistic subpial brain tumor resection (eFigure 2 in Supplement 2).[32] These tasks involved the use of bipolar forceps and an ultrasonic aspirator, each attached to a haptic handle, to completely resect the abnormal tissue while minimizing bleeding and damage to the surrounding healthy tissue.[15,33] Participants performed six 5-minute practice tasks to assess their learning, followed by a 13-minute realistic task to assess skill transfer to a more complex procedure. A 5-minute rest period was afforded to participants between each task.

Figure 1. Participant Recruitment Flowchart



AI indicates artificial intelligence.

Study Procedure

Prior to the simulation session, participants read and signed an informed consent form. They then completed a pretrial questionnaire recording demographic information and self-reported baseline emotions using the Medical Emotions Scale (MES) (eFigure 3 in Supplement 2).[34] Following the performance of 6 practice tasks, participants completed a peritrial questionnaire to assess the strength of emotions elicited during training using the MES. After the realistic task, students filled out a posttrial questionnaire that recorded emotions after training using the MES and self-reported cognitive load using the Cognitive Load Index (CLI).[35] Participants and instructors were blinded to group assignments and study outcomes. The study procedure is outlined in eFigure 4 in Supplement 2.

Interventions

The ICEMS continuously assessed each participant's performance at 0.2-second intervals and calculated expertise scores based on the following performance metrics: healthy tissue injury risk, bleeding risk, instrument tip separation distance, bipolar forceps force, and ultrasonic aspirator force.[11,12] An error was defined as a difference of more than 1 standard deviation from the expert benchmark for more than 1 second.[11,12]

All participants completed a practice task during which they did not receive feedback to establish a baseline. They proceeded to perform their second through fifth repetitions of the practice task while receiving intraoperative instruction, with the feedback delivery method varying between groups. No post hoc feedback was provided. All groups completed a sixth practice task without feedback, serving as a summative assessment. Finally, they completed the realistic brain tumor resection task without feedback to assess skill transfer to a more complex scenario.

All instructors were neurosurgical residents who underwent evaluation by a senior consultant and were identified as competent for their ability to train novices during simulated subpial resection procedures.

Group 1 (Control): AI Tutor Instruction

The control group received real-time verbal feedback delivered by the ICEMS when a metric error was detected.

Group 2: Expert Instruction

One experimental group received in-person, real-time verbal feedback from 1 of 2 neurosurgical residents (A.K.A., postgraduate year [PGY] 4; or M.A., PGY 5) based on ICEMS error detection. The ICEMS alerted the instructor via colored indicators when a metric error was detected, and the instructor delivered feedback to the trainee using the exact wording provided by the ICEMS (eTable in Supplement 2).

Group 3: Personalized Expert Instruction

One experimental group received AI-augmented, in-person, real-time verbal feedback from a neurosurgical resident (A.A., PGY 4) based on ICEMS error detection. The ICEMS alerted the instructor via colored indicators when a metric error was detected, and the instructor delivered tailored, personalized feedback to the trainee without restriction to ICEMS wording.

Outcome Measures

The first coprimary outcome was trainee learning and overall surgical performance on NeuroVR practice tasks, as scored by the ICEMS, which assessed each participant's performance in 0.2-second intervals. The second coprimary outcome was trainee technical skill transfer to a realistic resection task on the NeuroVR, as scored by the ICEMS.

The secondary outcome was trainees' self-reported affective-cognitive responses.[36] These included the strength of emotions elicited before, during, and after training and cognitive load after training. Emotions and cognitive load were measured via questionnaires using the MES[34] on a 7-point Likert scale and the CLI[35] on a 5-point Likert scale.

Statistical Analysis

Within-group differences from baseline in practice task scores and MES scores were compared using a mixed-model 1-way ANOVA. Between-group comparisons at each time point for

practice task scores and MES scores were conducted using a mixed-model 2-way analysis of covariance (ANCOVA), with baseline performance as a covariate. Realistic task scores and CLI scores were compared using a 1-way ANOVA. Post hoc pairwise comparisons were adjusted for multiple testing using the Tukey method for between-group differences and the Šidák method for within-group differences. Assumptions of normality and homogeneity of errors, as well as the presence of outliers, were investigated with graphical analyses of model residuals. Outlier observations were removed. Means from Likert items were computed prior to analysis of emotions and cognitive load. All statistical hypothesis tests were 2-sided and performed at a significance level of 0.05. Statistical analyses and score predictions were performed using R version 4.4.3 (R Foundation).[37] Data analysis was conducted in March 2025.

Results

A total of 88 medical students enrolled in Quebec medical schools were block randomized according to their year of study, with 31 in the AI tutor instruction group (group 1), 29 in the expert instruction group (group 2), and 28 in the personalized expert instruction group (group 3). Data from 1 participant in group 1 were excluded from analysis due to technical issues that occurred during the simulation session. The ICEMS assessed data from 87 participants (46 female students [53%]; mean [SD] age, 22.7 [4.0] years), including 522 practice resections and 87 realistic resections (Table). No statistically significant differences between groups were found in baseline demographic information.

Performance Across Practice Subpial Resection Trials

Following the baseline assessment (trial 1), the mean composite expertise scores were −0.58 (95% CI, −0.68 to −0.49) for group 1, −0.60 (95% CI, −0.70 to −0.50) for group 2, and −0.55 (95% CI, −0.65 to −0.44) for group 3. Group 3 significantly outperformed group 1 during trial 4 (mean difference, 0.26; 95% CI, 0.09-0.43; $P = .01$) and trial 5 (mean difference, 0.26; 95% CI, 0.09-0.43; $P = .01$). Although group 3 generally achieved higher mean scores than group 2 across practice trials, these differences were not statistically significant. During trial 5, group 2 significantly outperformed group 1 (mean difference, 0.23; 95% CI, 0.04-0.41; $P = .02$), indicating that the presence of a human instructor may play a role in improving trainee surgical performance. No statistically significant differences between groups were observed during trial 6. The only group whose mean expertise score surpassed the novice threshold of 0 was group 3 in trial 4. For within-group differences, all groups demonstrated statistically significant improvements in their scores from baseline across practice trials (**Figure 2**A).

Scores for individual ICEMS metrics used for competency training were also assessed. Personalized expert instruction, the intervention delivered to group 3 participants, largely resulted in metric scores closer to expert benchmarks than the other 2 interventions. From trials 3 to 6, group 3 achieved significantly lower bleeding risk than both groups 1 and 2 and lower injury risk than group 1 (Figure 2B and C). For aspirator force, group 3 significantly outperformed group 2 during trials 4 and 6 and group 1 during trials 3 and 4 (Figure 2D). Within-group comparisons revealed that all groups improved significantly from their baseline performance on several metrics, with group 3 showing the most consistent improvement.

Performance During Realistic Subpial Resection Task

After completion of the realistic task, the mean composite expertise scores were −0.35 (95% CI, −0.45 to −0.24) for group 1, −0.32 (95% CI, −0.43 to −0.21) for group 2, and −0.14 (95% CI, −0.25 to −0.04) for group 3. Group 3 significantly outperformed both group 1 (mean difference, 0.20; 95% CI, 0.06-0.34; $P = .02$) and group 2 (mean difference, 0.18; 95% CI, 0.03-0.32; $P = .049$), underscoring better skill transfer (**Figure 3**A). Group 3 also outperformed group 1 on both risk assessment metrics, achieving significantly lower bleeding risk (mean difference, 0.11; 95% CI, 0.05-0.16; $P < .001$) and injury risk (mean difference, 0.03; 95% CI, 0.01-0.04; $P = .009$) (Figure 3B and C). No statistically significant differences between groups were found for aspirator force, bipolar force, and tip separation distance (Figure 3D-F).

Emotions and Cognitive Load

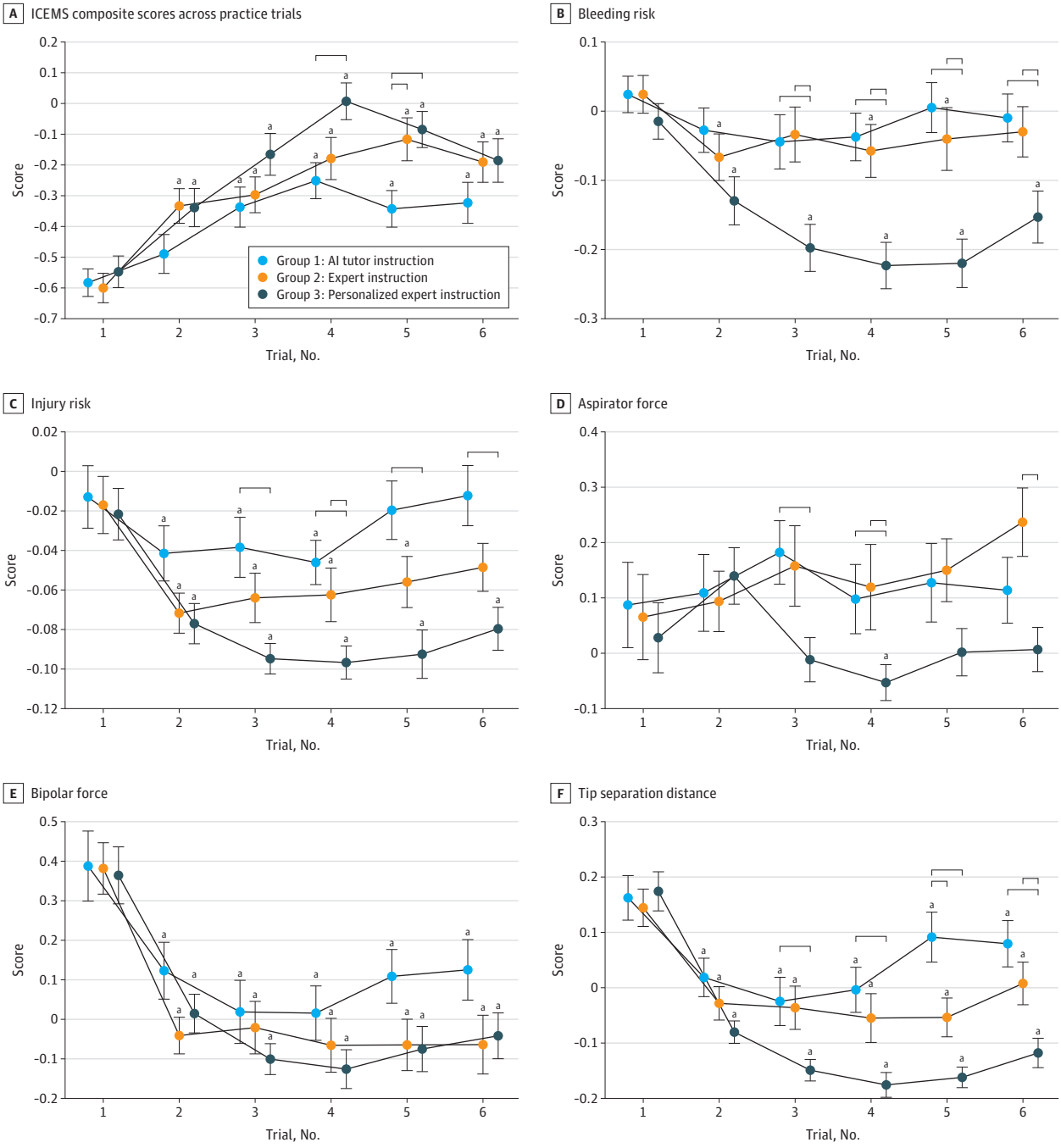
Group 3 reported significantly greater levels of negative activating emotions (eg, frustration) than group 1 after the trial (mean difference, 0.42; 95% CI, 0.01-0.82; $P = .04$). No between-group differences were observed for positive deactivating and negative deactivating emotional categories. The only group that experienced a statistically significant increase in positive deactivating emotions (eg, relief) was group 1 following the practice trials (**Figure 4**A). All 3 groups experienced statistically significant increases in both negative activating and negative deactivating emotions (eg, disappointment) after the practice trials, although these differences only persisted for group 3 after the realistic task (Figure 4B and C). Pairwise comparisons for cognitive load indicated that group 3 had a significantly higher intrinsic cognitive load compared to group 1 (mean difference, 0.56; 95% CI, 0.18-0.94; $P = .02$). No differences in extraneous or germane cognitive load were found (Figure 4D).

Discussion

To the authors' knowledge, this RCT is the first study that assesses the pedagogical utility of augmenting personalized expert instruction with AI error data to improve surgical training. Intelligent tutors that provide action-oriented feedback for assessment, coaching, and risk mitigation are adaptable to any surgical or technical specialty dependent on bimanual psychomotor expertise.[9,11] The main challenge in incorporating these technologies in surgical education paradigms is harnessing both the human instructor's expertise and the AI platform's real-time data processing to maximize student engagement and learning.[9,11]

Consistent with our hypothesis, the findings of this RCT demonstrate that AI-augmented personalized expert instruction yields improved surgical performance and skill transfer

Figure 2. Performance Assessment Across Practice Trials



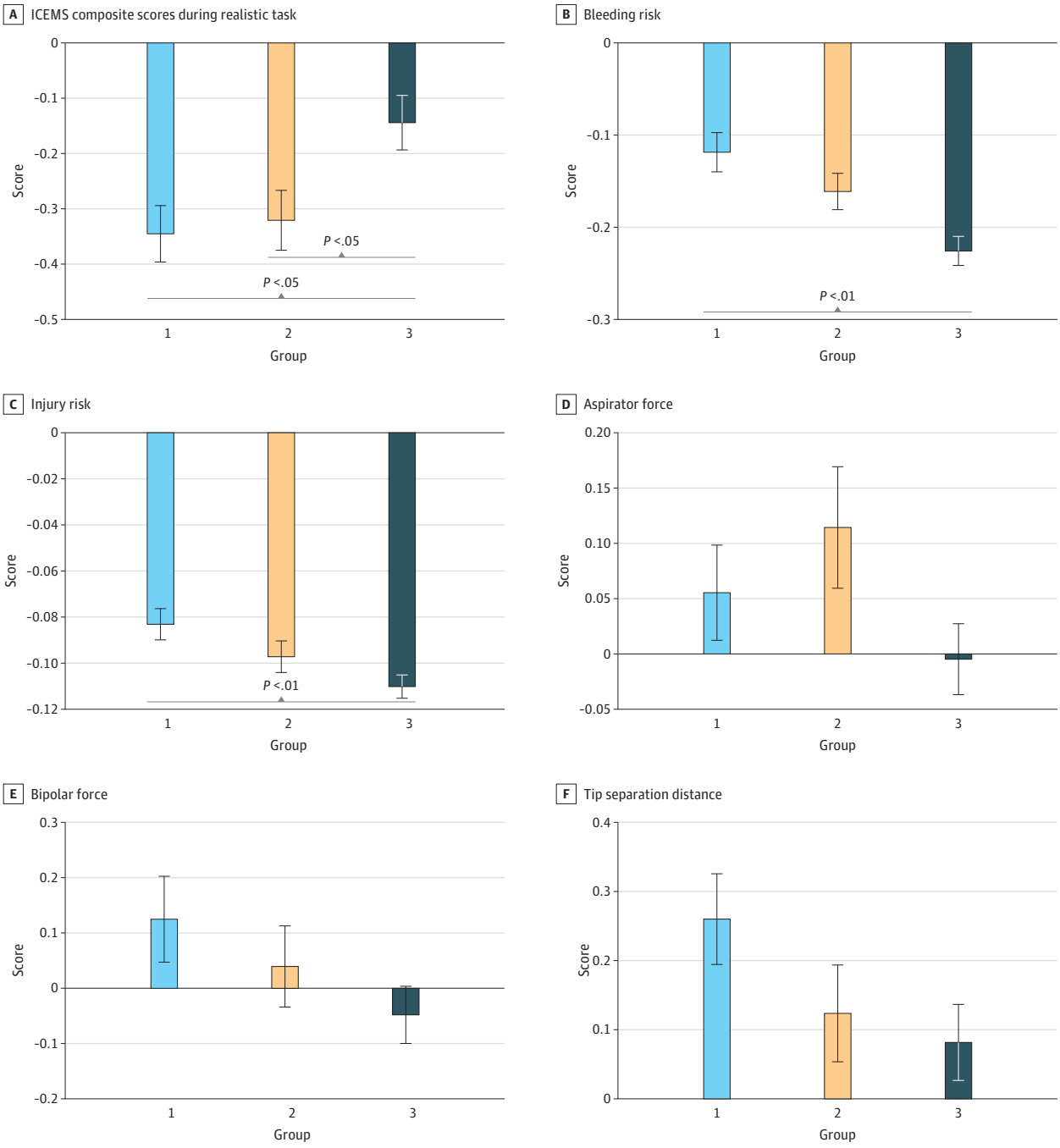
Intelligent Continuous Expertise Monitoring System (ICEMS) composite expertise scores (A) and 5 performance metrics across all practice trials (B-F). Groups are color coded (see key). The x-axis represents the trial number and the y-axis represents the composite or metric score calculated by the ICEMS. Points represent group means and error bars represent standard errors.

Horizontal brackets indicate statistically significant differences between groups ($P < .05$) during a given trial. AI indicates artificial intelligence.

[a]Indicates statistically significant differences from the baseline ($P < .05$) for that group.

compared with AI tutor instruction. The expert instruction group exhibited results superior to AI tutor instruction but inferior to AI-augmented personalized expert instruction and failed to significantly improve skill transfer. The ICEMS's capacity to supply quantitative data on individual risk assessment and coaching metrics facilitates the understanding of these results by providing explainability and transparency.[20-22] The ICEMS was developed using a long short-term memory network trained on performance data from experts (neurosurgeons) and novices (medical students).[11] Its algorithm

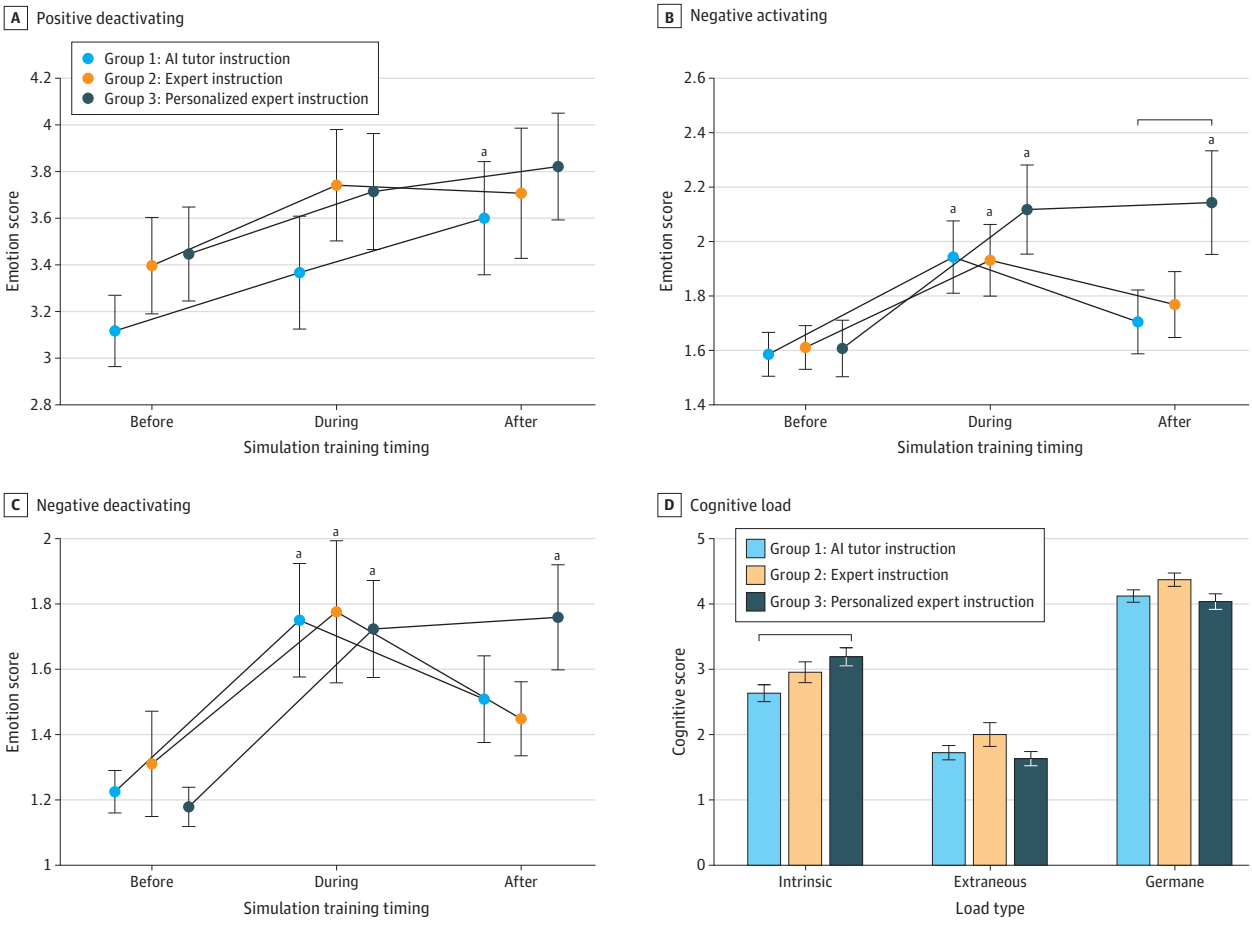
Figure 3. Performance Assessment During Realistic Task



Intelligent Continuous Expertise Monitoring System (ICEMS) composite expertise scores (A) and 5 performance metrics during the realistic task (B-F). The x-axis represents the group and the y-axis represents the composite or metric score calculated by the ICEMS. Colored bars represent group means and error bars represent standard errors. Horizontal lines with *P* values indicate statistically significant differences between groups ($P < .05$) during the realistic task.

primarily uses risk assessment metrics to calculate composite scores, with a secondary focus on coaching metrics.[11] The instructor's capacity to continuously modify individual feedback based on AI data in the AI-augmented personalized expert instruction group was shown to be particularly beneficial for risk mitigation. All trials were recorded, and studies are being carried out to determine which commands elicited the best responses among participants. The expert instruction group's outperformance of the AI instruction group in some trials suggests that the mere presence of a human instructor using identical words to the ICEMS may play a role in improved student engagement.[38] Other human factors, such as nonverbal cues and adaptive communication, may also influence student learning outcomes, but this requires further

Figure 4. Emotions and Cognitive Load Throughout Simulation Training



Groups are color coded (see key). The y-axis represents the Medical Emotions Scale or Cognitive Load Index score. Horizontal brackets indicate statistically significant differences between groups ($P < .05$). Error bars represent standard errors. Positive deactivating emotions (A) include relaxation and relief, negative activating emotions (B) include frustration and fear, and negative deactivating emotions (C) include disappointment and boredom. Cognitive load (D) consists of intrinsic, extraneous, and germane load. AI indicates artificial intelligence.

[a]Indicates statistically significant differences from the baseline ($P < .05$) for that group.

investigation.[39,40] During summative assessment trial 6, the personalized expert instruction group significantly outperformed the AI instruction group in bleeding and injury risk. However, no statistically significant between-group differences were found in the ICEMS composite scores in this trial. This may be attributed to learner fatigue.[41] Future studies should evaluate the mental fatigue of participants via self-report questionnaires.

A previous RCT conducted at our laboratory involving ICEMS tutoring was unable to demonstrate statistically significant between-group differences during the realistic task.[12] In this study, we provide evidence that AI-augmented personalized expert instruction more effectively improves skill transfer to a realistic scenario than AI tutor instruction and expert instruction. This realistic task more closely approximates a real brain tumor resection and involves similar competencies. Studies assessing whether this finding holds true for skill transfer from VR simulation to more realistic operating room environments using ex vivo animal models are in development.[42-45]

Unlike other RCTs conducted at our center,[10,12] this investigation did not include post hoc instruction, but resulted in equivalent increases in ICEMS expertise scores. In this RCT, continuous intraoperative action-directed feedback based on quantitative AI data was a critical determinant of learning. The impact of intraoperative combined with post hoc instruction in simulation curriculum design needs further assessment.

Our secondary hypothesis is not supported by the results. All 3 groups experienced significant increases in negative deactivating emotions, such as disappointment, during the trial. Posttrial levels of negative activating emotions, such as fear, were significantly greater in the personalized expert instruction group compared with the AI instruction group. Negative activating emotions often result in variable behavioral responses that may support or impede learning. The personalized expert instruction group's superior performance during the realistic task highlights the potential role of these emotions in supporting learning in this context.[46,47] The personalized expert instruction group reported significantly higher intrinsic

cognitive load than the AI instruction group, indicating increased mental effort required to understand the complexity of the variable instructions.[48,49] Research focused on negative activating emotions and cognitive load may help optimize learning with intelligent tutors.

Consistent with tenets of learning theory,[50-52] providing human instructors with quantitative AI performance data and allowing them to use their expertise to tailor and contextualize feedback leads to improved learning. Increased intraoperative educator-student engagement in this learning paradigm based on quantitative learner performance data may be the critical element explaining this study's findings. This RCT and our crossover study[23] results suggest that the optimization of surgical curricula designed to improve technical skill acquisition would involve experts initially providing critical context to operative procedure goals. In subsequent training sessions, educators would then leverage quantitative AI error data to deliver action-oriented feedback. This study helps provide pathways toward the overarching goal of creating an intelligent operating room using intraoperative intelligent tutoring systems capable of assessing and training learners while minimizing errors during human surgical procedures.

Limitations

Intelligent tutoring systems on VR simulation platforms do not encompass the full range of competencies involved in the dynamic interplay between trainee and educator in the operating room.[36] This study involved small cohorts of junior medical students with minimal surgical experience from only 4 institutions, limiting the generalizability of the results to other groups of learners. However, participants' inexperience resulted in a steeper learning trajectory, making it easier to detect trends in their learning curves. Understanding how medical students can attain AI-derived benchmarks of more advanced learners has offered insights into the optimization of surgical intelligent tutoring systems.[9-13] Although studies involving neurosurgical residents are in preparation, the limited number of available residents may result in an inability to achieve sufficient power to detect statistically significant differences. Finally, the applicability of these results to human surgical environments was beyond the scope of this research project but requires further investigation.

Conclusions

In this RCT, AI-augmented personalized expert instruction resulted in superior surgical performance and skill transfer compared with AI tutor instruction. These findings highlight the importance of human input and active participation in AI-based surgical training and provide an investigative platform for the further integration of intelligent tutoring systems in novel student-centered surgical curricula.

ARTICLE INFORMATION

Accepted for Publication: June 4, 2025.

Published Online: August 6, 2025. doi:10.1001/jamasurg.2025.2564

Author Affiliations: Neurosurgical Simulation and Artificial Intelligence Learning Centre, Department of Neurology and Neurosurgery, Montreal Neurological Institute and Hospital, McGill University, Montreal, Quebec, Canada (Giglio, Albeloushi, Alhaj, Alhantoobi, Saeedi, Davidovic, Uthamacumaran, Yilmaz, Lapointe, Balasubramaniam, Tee, Fazlollahi, Del Maestro); Department of Neurology and Neurosurgery, Montreal Neurological Institute and Hospital, McGill University, Montreal, Quebec, Canada (Albeloushi, Alhaj, Saeedi); Department of Neurosurgery, Hamilton General Hospital, McMaster University Medical Centre, Hamilton, Ontario, Canada (Alhantoobi); Division of Neurosurgery and Pediatrics, Children's National Medical Center, Washington, DC (Yilmaz); Faculty of Science and Engineering, Université Laval, Quebec City, Quebec, Canada (Lapointe); Faculty of Medicine and Health Sciences, McGill University, Montreal, Quebec, Canada (Balasubramaniam, Fazlollahi); Florida International University Herbert Wertheim College of Medicine, Miami (Tee); Department of Mathematics and Statistics, McGill University, Montreal, Quebec, Canada (Correa).

Author Contributions: Ms Giglio had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: Giglio, Albeloushi, Davidovic, Yilmaz, Balasubramaniam, Tee, Fazlollahi, Del Maestro.
Acquisition, analysis, or interpretation of data: Giglio, Albeloushi, Alhaj, Alhantoobi, Saeedi, Davidovic, Uthamacumaran, Yilmaz, Lapointe, Correa, Del Maestro.
Drafting of the manuscript: Giglio, Correa, Del Maestro.
Critical review of the manuscript for important intellectual content: All authors.
Statistical analysis: Giglio, Davidovic, Uthamacumaran, Lapointe, Correa, Del Maestro.
Obtained funding: Del Maestro.
Administrative, technical, or material support: Yilmaz, Balasubramaniam, Fazlollahi, Del Maestro.
Supervision: Del Maestro.

Conflict of Interest Disclosures: Dr Yilmaz reported a pending patent for methods and systems for continuous monitoring of task performance under the international patent application number WO2022077109A1. Ms Davidovic, Mr Uthamacumaran, and Ms Tee reported receiving Canada Graduate Scholarships—Master's program during the conduct of the study. Dr Del Maestro reported that the laboratory received the Brain Tumour Research Grant from the Brain Tumour Foundation of Canada, the Medical Education Research Grant from the Royal College of Physicians and Surgeons of Canada, the Mitacs Accelerate Grant, and funding from the Montreal Neurological Institute and Hospital and the Franco Di Giovanni Foundation; and provision of a prototype of the NeuroVR simulator used in this study to the laboratory from the National Research Council of Canada. No other disclosures were reported.

Funding/Support: This work was supported by a Brain Tumour Research Grant from the Brain Tumour Foundation of Canada, a Medical Education Research Grant from the Royal College of Physicians and Surgeons of Canada, a Mitacs Accelerate Grant, the Franco Di Giovanni Foundation, and the Montreal Neurological Institute and Hospital. Ms Davidovic, Mr Uthamacumaran, and Ms Tee were supported by Canada Graduate Scholarships—Master's program. Dr Yilmaz was supported by a grant from the Fonds de recherche du Québec—Santé for doctoral training and a Max Binz Fellowship from McGill University Internal Studentship. A prototype of the NeuroVR used in this study was provided by the National Research Council Canada.

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Meeting Presentation: An abstract of this paper was presented at the Canadian Conference for the Advancement of Surgical Education (C-CASE); October 17, 2024; Toronto, Ontario, Canada.

Data Sharing Statement: See Supplement 3.

Additional Contributions: The authors thank all medical students who participated in this study. The authors also thank Widad Safih, DEC (Faculty of Medicine, Université de Sherbrooke), and Sabrina Deraiche, DEC (Faculty of Medicine, Université de Montréal), for their help with trial participant recruitment.

Additional Information: The codes used in this study are available from the authors on request.

REFERENCES

1. Rogers SO Jr, Gawande AA, Kwaan M, et al. Analysis of surgical errors in closed malpractice claims at 4 liability insurers. *Surgery*. 2006;140(1):25-33. doi:10.1016/j.surg.2006.01.008

2. Fabri PJ, Zayas-Castro JL. Human error, not communication and systems, underlies surgical complications. *Surgery*. 2008;144(4):557-563. doi:10.1016/j.surg.2008.06.011

3. Fecso AB, Szasz P, Kerezov G, Grantcharov TP. The effect of technical performance on patient outcomes in surgery: a systematic review. *Ann Surg*. 2017;265(3):492-501. doi:10.1097/SLA.0000000000001959

4. Madani A, Vassiliou MC, Watanabe Y, et al. What are the principles that guide behaviors in the operating room? creating a framework to define and measure performance. *Ann Surg*. 2017;265(2):255-267. doi:10.1097/SLA.0000000000001962

5. Haluck RS, Krummel TM. Computers and virtual reality for surgical education in the 21st century. *Arch Surg*. 2000;135(7):786-792. doi:10.1001/archsurg.135.7.786

6. Cianciolo AT, Blessman J. "See one, do one, teach one?" a story of how surgeons learn. In: Köhler TS, Schwartz B, eds. *Surgeons as Educators*. Springer; 2017:3-13.

7. Mirza M, Koenig JF. Teaching in the operating room. In: Köhler TS, Schwartz B, eds. *Surgeons as Educators*. Springer; 2017:3-13.

8. Hiemstra E, Kolkman W, Wolterbeek R, Trimbos B, Jansen FW. Value of an objective assessment tool in the operating room. *Can J Surg*. 2011;54(2):116-122. doi:10.1503/cjs.032909

9. Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Del Maestro RF. The virtual operative assistant: an explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLoS One*. 2020;15(2):e0229596. doi:10.1371/journal.pone.0229596

10. 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. *JAMA Netw Open*. 2022;5(2):e2149008. doi:10.1001/jamanetworkopen.2021.49008

11. Yilmaz R, Winkler-Schwartz A, Mirchi N, et al. Continuous monitoring of surgical bimanual expertise using deep neural networks in virtual reality simulation. *NPJ Digit Med*. 2022;5(1):54. doi:10.1038/s41746-022-00596-8

12. 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. doi:10.1038/s41598-024-65716-8

13. 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. doi:10.1001/jamanetworkopen.2023.34658

14. Mirchi N, Ledwos N, Del Maestro RF. Intelligent tutoring systems: re-envisioning surgical education in response to COVID-19. *Can J Neurol Sci*. 2021;48(2):198-200. doi:10.1017/cjn.2020.202

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

16. Yilmaz R, Ledwos N, Sawaya R, et al. Nondominant hand skills spatial and psychomotor analysis during a complex virtual reality neurosurgical task-a case series study. *Oper Neurosurg (Hagerstown)*. 2022;23(1):22-30. doi:10.1227/ons.0000000000000232

17. Balakrishnan S, Dakua SP, El Ansari W, Aboumarzouk O, Al Ansari A. Novel applications of deep learning in surgical training. In: De Pablos PO, Zhang X, eds. *Artificial Intelligence, Big Data, Blockchain and 5G for the Digital Transformation of the Healthcare Industry: A Movement Toward More Resilient and Inclusive Societies*. Academic Press; 2023:301-320.

18. Vannaprathip N, Haddaway P, Schultheis H, Suebnukarn S. Intelligent tutoring for surgical decision making: a planning-based approach. *Int J Artif Intell Educ*. 2022;32:350-381. doi:10.1007/s40593-021-00261-3

19. Mousavinasab E, Zarifsanaiey N, Niakan Kalhori SR, Rakhshan M, Keikha L, Ghazi Saeedi M. Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interact Learn Environ*. 2021;29(1):142-163. doi:10.1080/10494820.2018.1558257

20. Eke CI, Shuib L. The role of explainability and transparency in fostering trust in AI healthcare systems: a systematic literature review, open issues and potential solutions. *Neural Comput Appl*. 2024;37:1999-2034. doi:10.1007/s00521-024-10868-x

21. Jung J, Lee H, Jung H, Kim H. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: a systematic review. *Heliyon*. 2023;9(5):e16110. doi:10.1016/j.heliyon.2023.e16110

22. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform*. 2021;113:103655. doi:10.1016/j.jbi.2020.103655

23. Yilmaz R, Fazlollahi A, Alsayegh A, Bakhaidar M, Del Maestro R. 428 Artificial intelligence training versus in-person expert training in teaching simulated tumor resection skills - a cross-over randomized controlled trial. *Neurosurgery*. 2024;70(suppl 1):129-130. doi:10.1227/neu.0000000000002809_428

24. Lehman B, Matthews M, D'Mello S, Person N. What are you feeling? investigating student affective states during expert human tutoring sessions. Paper presented at: 9th International Conference on Intelligent Tutoring Systems; June 23-27, 2008; Montreal, Quebec, Canada. doi:10.1007/978-3-540-69132-7_10

25. Hei X, Zhang H, Tapus A. Exploring help-seeking behavior, performance, and cognitive load in individual tutoring: a comparative study between human tutors and social robots. Paper presented at: 2024 33rd IEEE International Conference on Robot and Human Interactive Communication; August 26-30, 2024; Pasadena, CA. doi:10.1109/RO-MAN60168.2024.10731328

26. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364-1374. doi:10.1038/s41591-020-1034-x

27. Winkler-Schwartz A, Bissonnette V, Mirchi N, et al. Artificial intelligence in medical education: best practices using machine learning to assess surgical expertise in virtual reality simulation. *J Surg Educ*. 2019;76(6):1681-1690. doi:10.1016/j.jsurg.2019.05.015

28. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav Res Methods*. 2009;41(4):1149-1160. doi:10.3758/BRM.41.4.1149

29. Random.org. Accessed November 15, 2023. https://www.random.org/

30. Delorme S, Laroche D, DiRaddo R, Del Maestro RF. NeuroTouch: a physics-based virtual simulator for cranial microneurosurgery training. *Neurosurgery*. 2012;71(1)(Suppl Operative):32-42. doi:10.1227/NEU.0b013e318249c744

31. Alotaibi FE, AlZhrani GA, Mullah MAS, et al. Assessing bimanual performance in brain tumor resection with NeuroTouch, a virtual reality simulator. *Neurosurgery*. 2015;11(1)(suppl 2):89-98. doi:10.1227/NEU.0000000000000631

32. Sabbagh AJ, Bajunaid KM, Alarifi N, et al. Roadmap for developing complex virtual reality simulation scenarios: subpial neurosurgical tumor resection model. *World Neurosurg*. 2020;139:e220-e229. doi:10.1016/j.wneu.2020.03.187

33. Ledwos N, Mirchi N, Yilmaz R, et al. Assessment of learning curves on a simulated neurosurgical task using metrics selected by artificial intelligence. *J Neurosurg*. 2022;137(4):1160-1171. doi:10.3171/2021.12.JNS211563

34. Duffy MC, Lajoie SP, Pekrun R, Lachapelle K. Emotions in medical education: examining the validity of the Medical Emotion Scale (MES) across authentic medical learning environments. *Learn Instr*. 2020;70:101150. doi:10.1016/j.learninstruc.2018.07.001

35. Leppink J, Paas F, Van der Vleuten CPM, Van Gog T, Van Merriënboer JJG. Development of an instrument for measuring different types of cognitive load. *Behav Res Methods*. 2013;45(4):1058-1072. doi:10.3758/s13428-013-0334-1

36. Harley JM, Tawakol T, Azher S, Quaiattini A, Del Maestro R. The role of artificial intelligence, performance metrics, and virtual reality in neurosurgical education: an umbrella review. *Global Surg Educ*. 2024;3:83. doi:10.1007/s44186-024-00284-z

37. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. Accessed March 25, 2025. https://www.R-project.org/

38. Netland T, von Dzengelevski O, Tesch K, Kwasnitschka D. Comparing human-made and AI-generated teaching videos: an experimental study on learning effects. *Comput Educ*. 2025;224:105164. doi:10.1016/j.compedu.2024.105164

39. Pogue LL, Ahyun K. The effect of teacher nonverbal immediacy and credibility on student motivation and affective learning. *Commun Educ*. 2006;55(3):331-344. doi:10.1080/03634520600748623

40. Siegle RF, Craig SD. The voice quality of pedagogical agents impacts learning and agent perceptions. *J Comput Assist Learn*. 2024;40:2278-2291. doi:10.1111/jcal.13027

41. Hopstaken JF, van der Linden D, Bakker AB, Kompier MAJ. A multifaceted investigation of the link between mental fatigue and task disengagement. *Psychophysiology*. 2015;52(3):305-315. doi:10.1111/psyp.12339

42. Almansouri A, Abou Hamdan N, Yilmaz R, et al. Continuous instrument tracking in a cerebral corticectomy ex vivo calf brain simulation model: face and content validation. *Oper Neurosurg (Hagerstown)*. 2024;27(1):106-113. doi:10.1227/ons.0000000000001044

43. Alsayegh A, Bakhaidar M, Winkler-Schwartz A, Yilmaz R, Del Maestro RF. Best practices using ex vivo animal brain models in neurosurgical education to assess surgical expertise. *World Neurosurg*. 2021;155:e369-e381. doi:10.1016/j.wneu.2021.08.061

44. Tran DH, Winkler-Schwartz A, Tuznik M, et al. Quantitation of tissue resection using a brain tumor model and 7-T magnetic resonance imaging technology. *World Neurosurg*. 2021;148:e326-e339. doi:10.1016/j.wneu.2020.12.141

45. Winkler-Schwartz A, Yilmaz R, Tran DH, et al. Creating a comprehensive research platform for surgical technique and operative outcome in primary brain tumor neurosurgery. *World Neurosurg*. 2020;144:e62-e71. doi:10.1016/j.wneu.2020.07.209

46. Pekrun R, Marsh HW, Elliot AJ, et al. A three-dimensional taxonomy of achievement emotions. *J Pers Soc Psychol*. 2023;124(1):145-178. doi:10.1037/pspp0000448

47. Tze V, Parker P, Sukovieff A. Control-value theory of achievement emotions and its relevance to school psychology. *Can J Sch Psychol*. 2022;37(1):23-39. doi:10.1177/08295735211053962

48. Howie EE, Dharanikota H, Gunn E, et al. Cognitive load management: an invaluable tool for safe and effective surgical training. *J Surg Educ*. 2023;80(3):311-322. doi:10.1016/j.jsurg.2022.12.010

49. Young JQ, Van Merriënboer J, Durning S, Ten Cate O. Cognitive load theory: implications for medical education: AMEE Guide No. 86. *Med Teach*. 2014;36(5):371-384. doi:10.3109/0142159X.2014.889290

50. Kaufman DM, Mann KV. Teaching and learning in medical education: how theory can inform practice. In: Swanwick T, ed. *Understanding Medical Education: Evidence, Theory and Practice*. 2nd ed. Wiley-Blackwell; 2013:7-29. doi:10.1002/9781118472361.ch2

51. Taylor DCM, Hamdy H. Adult learning theories: implications for learning and teaching in medical education: AMEE Guide No. 83. *Med Teach*. 2013;35(11):e1561-e1572. doi:10.3109/0142159X.2013.828153

52. Wozniak K. Personalized learning for adults: an emerging andragogy. In: Yu S, Ally M, Tsinakos A, eds. *Emerging Technologies and Pedagogies in the Curriculum*. Springer; 2020:185-198. doi:10.1007/978-981-15-0618-5_11