Effect of Artificial Intelligence-Augmented Human Instruction on Feedback Frequency and Surgical Performance During Simulation Training

Vanja Davidovic, MSc,[†] Bianca Giglio, MSc,[†] Abdulmajeed Albeloushi, MD,[†,‡] Ahmad Kh. Alhaj, MD,[†,‡] Mohamed Alhantoobi, MD, MSc,[†,§] Rothaina Saeedi, MD,[†,‡] Sabrina Deraiche, DEC,[†,‖] Recai Yilmaz, MD, PhD,[†,¶] Trisha Tee, MSc,[†,††] Ali M. Fazlollahi, MD, MSc,[†,‡‡] Matthew Ha, MSc,[§§] Abicumaran Uthamacumaran, BSc,[†] Neevya Balasubramaniam, MD,[†,‡‡] José A. Correa, PhD,[¶¶] and Rolando F. Del Maestro, MD, PhD[†,‡]

[†]Neurosurgical Simulation and Artificial Intelligence Learning Centre, Department of Neurology and Neurosurgery, Montreal Neurological Institute and Hospital, McGill University, Montreal, QC, Canada; [‡]Department of Neurology and Neurosurgery, Montreal Neurological Institute and Hospital, McGill University, Montreal, QC, Canada; [§]Department of Neurosurgery, Hamilton General Hospital, McMaster University Medical Centre, Hamilton, ON, Canada; [‖]Faculté de médecine, Pavillon Roger-Gaudry, Université de Montréal, Montreal, QC, Canada; [¶]Division of Neurosurgery and Pediatrics, Children's National Medical Center, Washington, DC, United States of America; [††]Herbert Wertheim College of Medicine, Florida International University, Miami, FL, United States of America; [‡‡]Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada; [§§]Department of Surgery, McGill University Health Center, McGill University, Montreal, QC, Canada; and [¶¶]Department of Mathematics and Statistics, McGill University, Montreal, QC, Canada

OBJECTIVE: To determine whether personalized feedback from a human instructor receiving artificial intelligence (AI) error data will result in reduced feedback frequency and improvement of surgical skill compared to AI instruction. As feedback was only provided following AI error detection, a reduced feedback frequency is associated with fewer errors in performance. We hypothesized that AI-augmented personalized instruction would result in reduced feedback frequency and improvement in technical skill.

DESIGN: This cross-sectional cohort study was a follow-up of a randomized controlled trial, which utilized the NeuroVR, an immersive virtual reality neurosurgical simulator. Participants were stratified by year in medical school and block randomized to receive one of 3 educational interventions as they performed simulated procedures on the NeuroVR: AI tutor instruction, scripted human instruction, and AI-augmented personalized instruction. Performance was assessed by the feedback frequency and technical skill performance metrics. ClinicalTrials.gov ID: NCT06273579.

SETTING: Neurosurgical Simulation and Artificial Intelligence Learning Centre, McGill University, Montreal, Canada.

PARTICIPANTS: Volunteer sample of medical students from 4 Quebec universities in preparatory, first, or second year without prior use of the NeuroVR. Eighty-eight students participated in the study with 87 included in the final analysis; 1 was excluded due to technical issues.

Funding: This work was supported by a Brain Tumor Research Grant from the Brain Tumor Foundation of Canada, a Medical Education Research Grant from the Royal College of Physicians and Surgeons of Canada, the Franco Di Giovanni Foundation, and the Montreal Neurological Institute and Hospital. Vanja Davidovic, Trisha Tee, and Abicumaran Uthamacumaran were supported by a Canada Graduate Scholarships – Master's program. Trisha Tee was also supported by a Mitacs Accelerate Internship Grant. Recai 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 of Canada, Boucherville, Quebec, Canada. 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.

Correspondence: Inquiries to Vanja Davidovic, MSc, Neurosurgical Simulation and Artificial Intelligence Learning Centre, Department of Neurology and Neurosurgery, Montreal Neurological Institute and Hospital, McGill University, 300 Rue Léo-Pariseau, Suite 2210, Montreal, QC H2X 4B3, Canada; e-mail: vanja.davidovic@mail.mcgill.ca

  1931-7204/$30.00 https://doi.org/10.1016/j.jsurg.2025.103743

RESULTS: By the third repetition, the AI-augmented personalized instruction group received significantly fewer total instructions (incidence rate ratio [IRR], 1.50 [95% CI, 1.16-1.94] instructions; $p < 0.001$), and high aspirator force instructions (IRR, 1.71 [95% CI, 1.15-2.55] instructions; $p = 0.002$), compared to the second repetition. Compared to AI tutor instruction, AI-augmented personalized instruction resulted in a significantly lower rate of healthy tissue removal ($p = 0.01$), instrument tip separation distance (mean ratio, 1.25 [95% CI, 1.05-1.50] mm; $p = 0.008$), and aspirator force (mean ratio, 1.68 [95% CI, 1.23-2.31] N; $p < 0.001$). AI-augmented personalized instruction showed a significant improvement from baseline in all subsequent repetitions for all performance metrics.

CONCLUSIONS: This cohort study demonstrated that AI-augmented personalized instruction resulted in less frequent feedback, indicating fewer errors in trainee performance, and an improvement in simulated surgical skills. (J Surg Ed 82:103743. © 2025 The Authors. Published by Elsevier Inc. on behalf of Association of Program Directors in Surgery. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/))

KEY WORDS: artificial intelligence-augmented instruction, surgical simulation, surgical education, neurosurgical virtual reality, intelligent tutoring system, performance metrics

COMPETENCIES: Artificial Intelligence, Simulation, Neurosurgery, Learning Feedback, Technical Skills, Virtual Reality

ABBREVIATIONS: AI, artificial intelligence; ICEMS, Intelligent Continuous Expertise Monitoring System; VR, virtual reality; 3D, 3-dimensions; STROBE, Strengthening the Reporting of Observational Studies in Epidemiology; MLASE, Machine Learning to Assess Surgical Expertise; PGY, postgraduate year; ANOVA, analysis of variance; GLMM, generalized linear mixed model; IRR, incidence rate ratio; CI, confidence interval; LLM, large language model

INTRODUCTION

Mastery of surgical technical skill is essential to mitigate the risk of surgical errors.[1-5] The current pedagogical model for surgical training involves the constant interplay between the educator and the trainee in a dynamic operative environment.[6] These real-time communications rely on the subjective observations of human instructors for continuous assessment and immediate, personalized, actionable feedback to guide technical skill development and error mitigation.[7] However, there may be discrepancies between instructors' interpretations of proper surgical techniques, leading to difficulties comparing trainee skill level between multiple evaluators, let alone multiple institutions.[7] This reliance on subjective, qualitative performance data highlights a lack of objective, standardized instructional methodologies and assessments of surgical trainee performance.[7-10] Intelligent tutoring systems utilizing artificial intelligence (AI) to provide personalized and adaptive instructions to learners may help overcome these limitations due to their capacity to process and analyze large quantities of data to objectively assess performance.[11-16] These systems can quantify trainee performance and detect subtle errors that a human instructor may fail to notice and mitigate, such as a high force application that can lead to healthy tissue damage if left uncorrected.[7]

Intelligent tutoring systems have shown potential in teaching trainees surgical techniques and evaluating their competency using a data-driven approach in simulation environments.[11,12,17] A randomized controlled trial (RCT) utilizing the Virtual Operative Assistant intelligent tutoring system, employing only posthoc AI feedback, significantly improved simulated surgical performance.[12,15] This system lacks the capacity to continuously monitor intraoperative skills or provide real-time feedback, posing a disadvantage to its application in an operating room environment. The Intelligent Continuous Expertise Monitoring System (ICEMS) is a multi-algorithm AI system specifically designed to address these issues by employing quantitative data to continuously assess trainee performance and provide continuous, real-time instructions to mitigate and reduce trainee errors based on real-time risk detection.[16] Developed using a long short-term memory network and based on objective, AI-derived metrics, the ICEMS can be used to detect errors in surgical performance.[16] The ICEMS was trained on neurosurgeons' (experts) and medical students' (novices) operative data and demonstrated a granular differentiation across levels of expertise, and has shown face, content, construct, and predictive validity.[16,18] The NeuroVR, a high-fidelity virtual reality (VR) surgical simulator equipped with haptic feedback for brain tumor resection procedures, was used to develop the ICEMS.[19] The ICEMS can be applied to any simulation system.[16]

An RCT demonstrated that the ICEMS improved simulated surgical performance more than skilled instructors, indicating the pedagogical utility of the system.[11] Another crossover RCT found that trainee performance was significantly improved when instructed by a skilled educator first and then followed by ICEMS instruction.[17] Although this intelligent tutoring system can provide

objective feedback, it is limited to delivering specific verbal instructions, while human educators can provide context and personalize their feedback. In a previous cohort study, this limited variety of possible feedback instructions led to unintended outcomes in an AI-enhanced curriculum, which negatively impacted trainee performance efficiency.[20] The results of these studies suggest that combining a skilled instructor and an AI tutor would allow for the contextualization of AI error data and optimize trainee performance. A recent RCT found that AI-augmented personalized instruction resulted in enhanced ICEMS scores on a simulated subpial brain tumor practice resection scenario compared to AI tutor instruction and scripted human instruction, along with an improved transfer of surgical technical skills to a realistic simulated scenario.[21] These results highlight that personalized expert instruction results in enhanced surgical performance and skill transfer compared with intelligent tutor instruction, emphasizing the critical role of human engagement and contribution in artificial intelligence-based surgical training.

However, this study did not investigate how AI-augmented personalized expert instruction influenced the frequency of feedback instructions, nor the differences in trainee technical performance between groups. Our study builds on this previous investigation,[21] using a cohort study design to evaluate these 2 components. We hypothesized that participants receiving AI-augmented personalized instruction would (1) receive a significantly lower number of feedback instructions compared to those receiving AI tutor instruction, and (2) show a significantly better response to these instructions through improvement in technical skill performance compared to those receiving AI tutor instruction.

METHODS

Participants

We conducted a planned secondary analysis using retrospective data from a previous RCT involving 87 medical students at the Neurosurgical Simulation and Artificial Intelligence Learning Centre, McGill University, Montreal, Canada from March to September 2024.[21] Students were recruited for a single 90-minute surgical simulation session with no follow-up. Medical students enrolled in their preparatory, first, or second year at one of 4 Quebec institutions were considered eligible for the study. The exclusion criterion was previous experience with the NeuroVR, the VR simulator used in this study. A sample size calculation with 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 resulted in a total of 87 participants, with 29 participants in each of 3 groups. Recruitment materials were distributed through student groups, social media, and word of mouth. Each participant performed the same simulated procedure with a different instructional method. This study was approved by the McGill University Health Centre Research Ethics Board, Neurosciences-Psychiatry and was registered on ClinicalTrials.gov on February 16, 2024 (NCT06273579). All participants signed an approved informed consent form prior to commencing the study. Participants did not receive any benefits or compensation for their participation. This report follows the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)[22] guidelines for cohort studies and the Machine Learning to Assess Surgical Expertise (MLASE) checklist.[23]

Study Procedure and Simulation Session

Following voluntary enrollment, students were stratified according to year in medical school and block randomized to one of 3 intervention arms with a 1:1:1 allocation ratio. All participants received standardized written and verbal instructions outlining the use of the instruments, the goal of the task, and how the session would proceed. Students were not made aware of the trial's purpose and assessment metrics at any point. The study utilized the NeuroVR (CAE Healthcare, Montreal, Canada), a high-fidelity VR neurosurgical simulator, on which participants performed simulated subpial brain tumor resection procedures.[19,24] The platform has previously demonstrated face, content, and construct validity.[24-26] The simulation tasks involved the use of an ultrasonic aspirator and bipolar forceps, each equipped with haptic feedback, to completely resect a simulated tumor while minimizing bleeding and damage to nonpathological tissue.[25,27] All participants completed six 5-minute practice subpial resection scenarios to assess their learning (Figure 1), followed by a 13-minute realistic scenario to assess skill transfer to a more complex procedure (Figure 2). Between each repetition, a rest period of 5 minutes was afforded to participants.[28,29]

Interventions

Participants performed their first practice resection scenario without feedback to establish their baseline performance level. The second to fifth repetitions of the practice scenario served as a formative assessment, during which participants received feedback only when an error was identified by the ICEMS. Feedback methods differed between the 3 groups. Participants then proceeded to perform a sixth repetition of the practice scenario without feedback as a summative assessment of their performance. Trainees then completed 1 repetition of the realistic scenario to assess skill transfer to a more complex scenario. The study procedure is outlined in

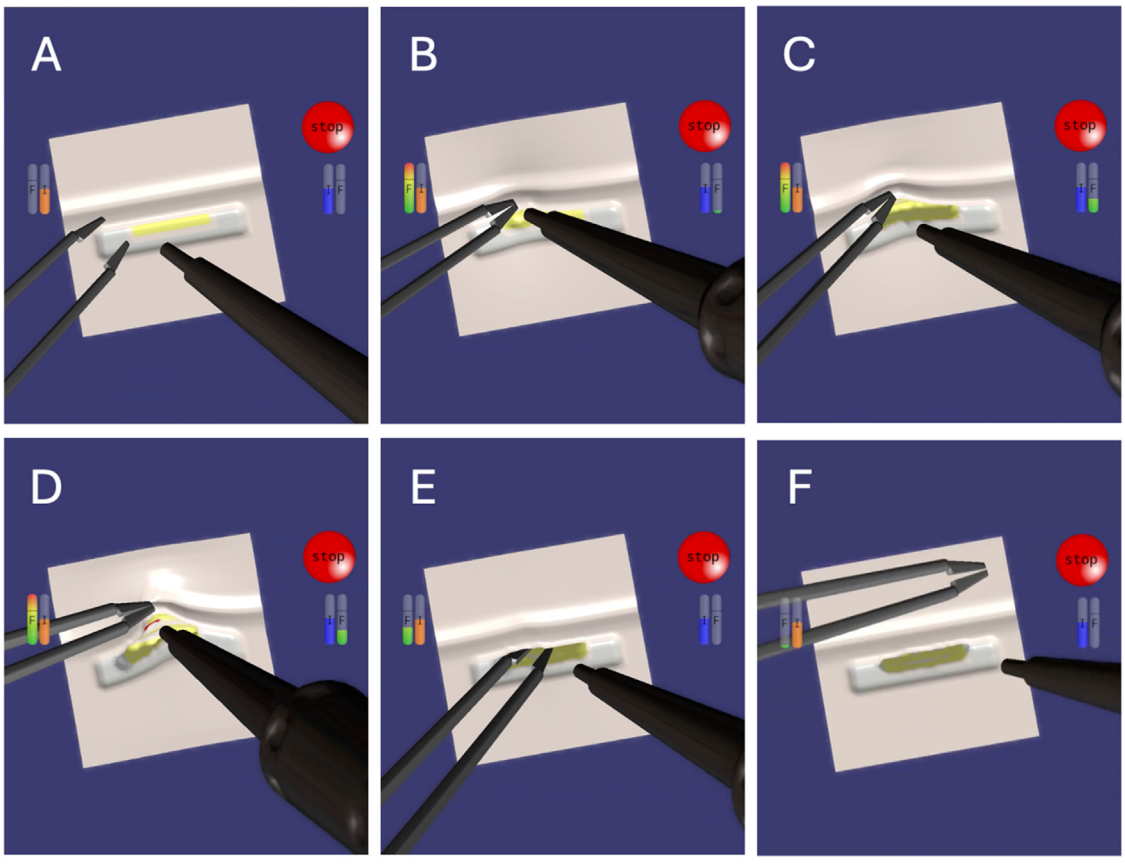

FIGURE 1. Practice subpial tumor resection scenario. (A) Start of the practice subpial tumor resection scenario. Yellow area represents the tumor and white area represents healthy brain tissue. Instrument on the left is the bipolar forceps, and instrument on the right is the ultrasonic aspirator. (B) Participant lifts the pia using the bipolar forceps to expose the underlying tumor, and the ultrasonic aspirator to resect the tumor. (C) Appearance following resection of the superficial tumor. Deeper tumor areas shown by remaining yellow tissue. (D) Participant exposes deep cerebral vessel (red). (E) Participant uses the bipolar forceps to cauterize a bleeding point. (F) Complete resection of the tumor.

Figure 3. Participants and instructors were blinded to group assignments and study outcomes.

The instructors were senior neurosurgical residents with experience in clinical and simulated subpial resection procedures. A senior neurosurgical consultant with extensive involvement in VR neurosurgical simulation and clinical subpial operations identified these instructors as competent to train novices during these simulated procedures.

Group 1: AI Tutor Instruction

Group 1 received real-time verbal feedback from the ICEMS upon metric error detection.

Group 2: Scripted Human Instruction

Group 2 received instructions from one of 2 neurosurgical residents (M.A., postgraduate year [PGY] 5; A.K.A., PGY 4) upon metric error detection by the ICEMS. Prompted by the ICEMS using colored indicators, instructors provided real-time verbal feedback using the same wording as the ICEMS (Table 1).

Group 3:AI-Augmented Personalized Instruction

Group 3 received instructions from a neurosurgical resident (A.A., PGY 4) upon metric error detection by the ICEMS. Prompted by the ICEMS using colored indicators, the instructor provided real-time personalized verbal feedback in their own words based on the trainee's manipulations.

Instructions were provided based on preselected metrics: healthy tissue injury risk, bleeding risk, instrument tip separation distance, bipolar forceps force, and ultrasonic aspirator force. The metrics followed a hierarchy, as employed in previous studies; if more than 1 error occurred simultaneously, instructions for the metric higher in the hierarchy would be prioritized.[11,16] The feedback instructions provided in groups 1 and 2 and the hierarchical order of these metrics are outlined in Table 1.

Performance Metric Extraction

During the second to fifth repetitions, the ICEMS recorded the number of instructions given for each

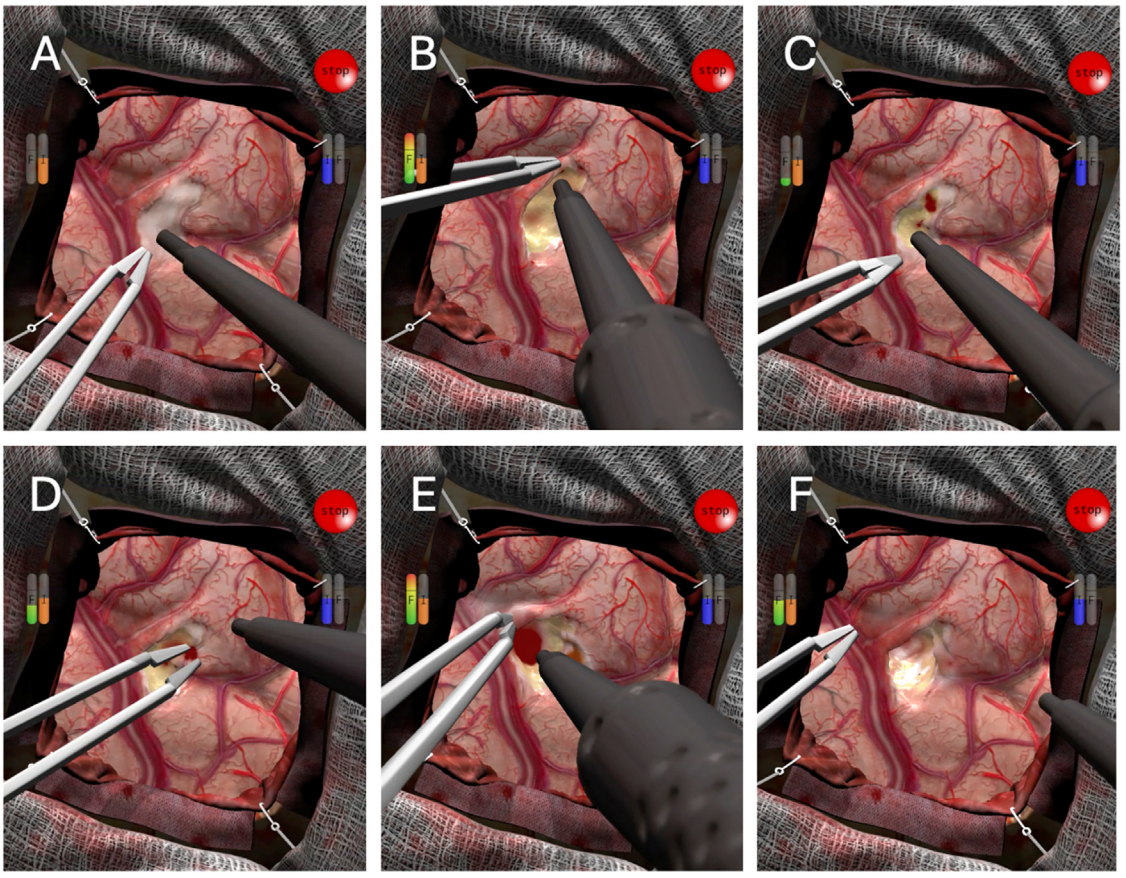

FIGURE 2. Realistic subpial tumor resection scenario. (A) Start of the realistic subpial tumor resection scenario. Off-white area represents the tumor and pink area represents healthy brain tissue. Instrument on the left is the bipolar forceps, and instrument on the right is the ultrasonic aspirator. (B) Participant lifts the pia using the bipolar forceps to expose the underlying tumor, and the ultrasonic aspirator to resect the tumor. (C) Participant causes minor bleeding from the tumor while using the ultrasonic aspirator. (D) Participant uses the bipolar forceps to cauterize a bleeding point. (E) Participant causes major bleeding from the healthy tissue while using the ultrasonic aspirator. (F) Complete resection of the tumor.

ICEMS metric: healthy tissue injury risk, bleeding risk, high instrument tip separation distance, high bipolar force, low bipolar force, and high aspirator force.[11,16] Following the completion of a repetition, the number of instructions given to a participant in total and for each metric was summed. The average number of instructions was calculated for each group for each formative repetition of the practice scenario. During every repetition,

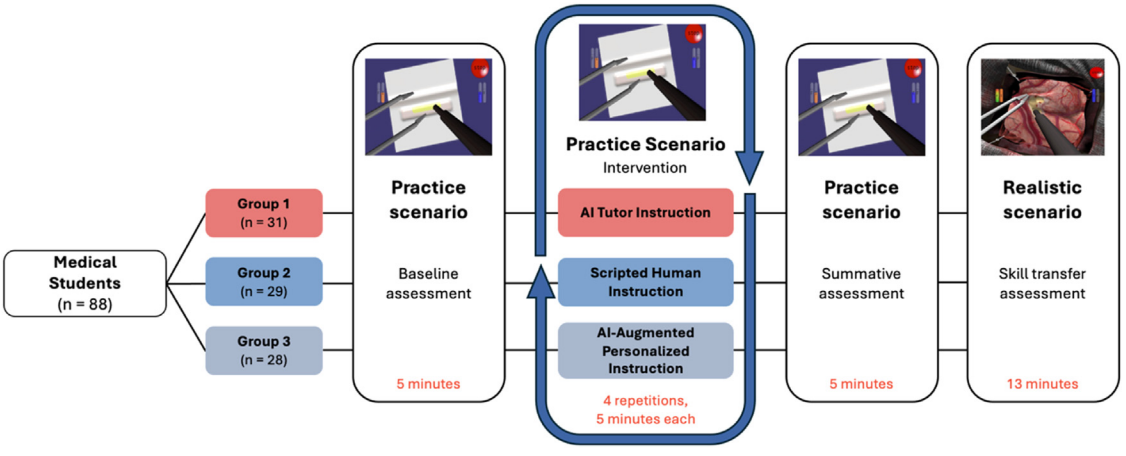

FIGURE 3. Flow diagram. Eighty-eight students were randomly allocated into 3 intervention groups. Abbreviation: AI, artificial intelligence.

TABLE 1. Metrics Assessed by the ICEMS in Hierarchical Order, and Their Corresponding Instructions[16,‡]

Metric	ICEMS Instruction
1. Healthy tissue injury risk	"Try to avoid damaging the healthy brain surrounding the tumor."
2. Bleeding risk	"Careful control of bleeding will improve your performance."
3. Instrument tip separation distance	"Keeping your instruments closer together will improve your performance."
4. High bipolar force application	"Try to decrease the amount of force you are applying with your bipolar."
5. Low bipolar force application	"You can improve your performance by applying more force with your bipolar."
6. High aspirator force application	"Try to decrease the amount of force you are applying with your aspirator."

Abbreviation: ICEMS, Intelligent Continuous Expertise Monitoring System.
‡The right column shows the instructions given to group 1 and 2 upon metric error detection by the ICEMS. If the ICEMS identifies more than one error simultaneously, it is programmed to provide instruction on the metric higher in the hierarchy.

the NeuroVR recorded participant technical skill performance data in 20-millisecond increments (50 recordings per second; $t = 20$ ms), including rate of healthy tissue removal (mm^3/t), total volume of blood lost (mm^3), instrument tip separation distance in the 3D space (mm), force applied with bipolar forceps (N), and force applied with ultrasonic aspirator (N).[30] These performance metrics were selected based on their relation to the feedback instructions given during the session to assess their effectiveness. The raw data were collected, and an average was calculated for 4 of the 5 technical skill performance metrics for each participant in each repetition. Only the final value of total blood volume lost was utilized for each repetition, rather than taking an average. AI tutor-automated feedback provision and data visualization were performed using MATLAB (The MathWorks Inc., Natick, Massachusetts, USA) release 2024b.

Outcome Measures

The first coprimary outcome of this study was the number of instructions that trainees received in total and for each ICEMS metric during each of the formative practice subpial resection scenarios. The second coprimary outcome was trainee technical skill performance during the practice scenarios and realistic scenario, measured using the 5 performance metrics recorded by the NeuroVR.

Statistical Analysis

Between- and within-group comparisons of the mean number of feedback instructions received over the second to fifth repetitions of the practice resection scenario were conducted using generalized linear mixed (GLMM) Poisson regression models for count data. Model assumptions and the presence of possible outliers or influential observations were investigated using graphical analyses of simulated residuals. Posthoc pairwise comparisons were adjusted using the Šidák method for between-group differences and the Bonferroni correction for within-group differences. Results are reported as incidence rate ratios (IRR) and 95% confidence intervals (CI).

Between-group comparisons of the mean values of the technical skill performance metrics at each repetition of the practice resection scenario were conducted using a two-way mixed model analysis of variance (ANOVA). Repeated measures ANOVA was used to investigate within-group differences of the mean values of the technical skill performance metrics at each repetition of the practice resection scenario. One-way ANOVA was used to compare the mean values of each technical skill performance metric in the realistic resection scenario. Assumption of errors of ANOVA models, including normality, homogeneity of variance, and the presence of possible outliers or influential observations were assessed by graphical examination of model residuals. Posthoc pairwise comparisons of mean differences were adjusted using the Šidák method for between-group differences and the Bonferroni correction for within-group differences. When model residuals did not show evidence of having a Normal (Gaussian) distribution, a natural logarithmic transformation of the values was used as the model outcome to stabilize the variability. A robust linear mixed model approach to the ANOVA was used when the assumption of homogeneity of variance or residuals was violated. In cases where the normality or homogeneity of variance assumptions were drastically violated, we used the Kruskal-Wallis test at each repetition for between-group analysis, followed by Dunn's test with Bonferroni correction for multiple comparisons, and the Friedman test was used for within-group analysis, followed by the Nemenyi test. Results are reported as estimated mean differences and 95% CI and, in cases where a log transformation was used, as estimated ratios of geometric means and 95% CI.

Data analysis was performed using R Statistical Software (v4.3.3; R Core Team 2024)[31] from February to May 2025. All codes were written by the authors. ANOVAs and Poisson GLMM were implemented using

TABLE 2. Demographic Characteristics of Included Study Participants

	Group 1 AI tutor instruction (*n* = 30)	Group 2 Scripted human instruction (*n* = 29)	Group 3 AI-augmented personalized instruction (*n* = 28)	All Participants (*n* = 87)
Age, mean ± SD (range)	21.8 ± 2.4 (18-27)	22.6 ± 4.4 (18-38)	23.9 ± 4.8 (19-37)	22.7 ± 2.4 (18-38)
Sex				
Female	18	16	12	46
Male	12	13	15	40
Prefer not to say	0	0	1	1
Gender				
Woman	18	16	12	46
Man	12	13	15	40
Prefer not to say	0	0	1	1
Undergraduate medical training level				
Preparatory	9	8	8	25
First	15	14	13	42
Second	6	7	7	20
Institution				
McGill University	11	15	14	40
Université de Montréal	12	7	6	25
Université de Sherbrooke	4	6	7	17
Université Laval	3	1	1	5
Handedness				
Right	28	25	24	77
Left	2	3	4	9
Ambidextrous	0	1	0	1
Interest in pursuing surgery, mean (range)[‡]	4 (2-5)	4.1 (2-5)	3.9 (2-5)	4 (2-5)
Completed surgical rotation/ clerkship/shadowing				
Yes	12	10	11	33
No	18	19	17	54
Plays video games				
Yes	8	9	13	30
No	22	20	15	57
Played musical instruments in last 5 years				
Yes	9	9	13	31
No	21	20	15	56
Participated in activities that require hand dexterity				
Yes	8	12	11	31
No	22	17	17	56
Previously used VR surgical simulation				
Yes	1	2	5	8
No	29	27	23	79

Abbreviations: AI, artificial intelligence; VR, virtual reality.
[‡]Rated on a 5-point Likert scale, with 1 indicating less interest and 5 indicating more interest.

the *lme4* [32] and *glmmTMB* [33] R packages, respectively. The GLMM analysis of simulated residuals was implemented using the *DHARMa* R package.[34] The robust linear mixed model approach to ANOVA was done using the *robustlmm* R package.[35]

RESULTS

Eighty-eight medical students from 4 Quebec universities participated in the study. Participants were stratified according to year in medical school and block

randomized to one of 3 groups. There were 31 students in the AI tutor instruction group (group 1), 29 in the scripted human instruction group (group 2), and 28 in the AI-augmented personalized instruction group (group 3). Due to technical issues that arose during the simulation session, data from 1 participant in group 1 were excluded from the analysis. Data from 87 participants (46 [53%] women, 40 [46%] men, 1 [1%] unspecified; mean [SD] age, 22.7 [4.0] years) were available for analysis, including 522 practice scenarios and 87 realistic scenarios (Table 2).

Feedback Frequency Across Simulated Practice Subpial Resections

All groups began receiving instructions in the second repetition of the practice scenario. In total, over the second to fifth repetitions, group 1 received 1464 instructions, group 2 received 1183 instructions, and group 3 received 728 instructions. Figure 4 outlines the number of instructions received by the groups for each metric in each repetition of the practice scenario where feedback was given (repetitions 2 to 5). The incidence rates of instructions for each metric can be found in Table 3. Group 1 received significantly more instructions in total compared to group 3 by the second repetition (IRR, 1.52 [95% CI, 1.03-2.23] instructions; p = 0.03). Group 2 received significantly more feedback instructions in total compared to group 3 by the third repetition of the practice scenario (IRR, 1.74 [95% CI, 1.15-2.63] instructions; p = 0.001) (Fig. 5A). Pertaining to bleeding risk instructions, group 1 received significantly more feedback compared to group 3 in the second (IRR, 5.55 [95% CI, 1.22-25.15] instructions; p = 0.01), fourth (IRR, 5.02 [95% CI, 1.09-23.19] instructions; p = 0.03), and fifth (IRR, 8.20 [95% CI, 1.73-38.77] instructions; p = 0.001) repetitions, and group 2 received significantly more feedback than group 3 in the second (IRR, 4.89 [95% CI, 1.06-22.53] instructions; p = 0.04) and fifth (IRR, 6.81 [95% CI, 1.41-32.80] instructions; p = 0.006) repetitions (Fig. 5C). Group 1 received significantly more instructions relating to high bipolar force than group 3 in the fourth repetition of the practice scenario (IRR, 2.13 [95% CI,

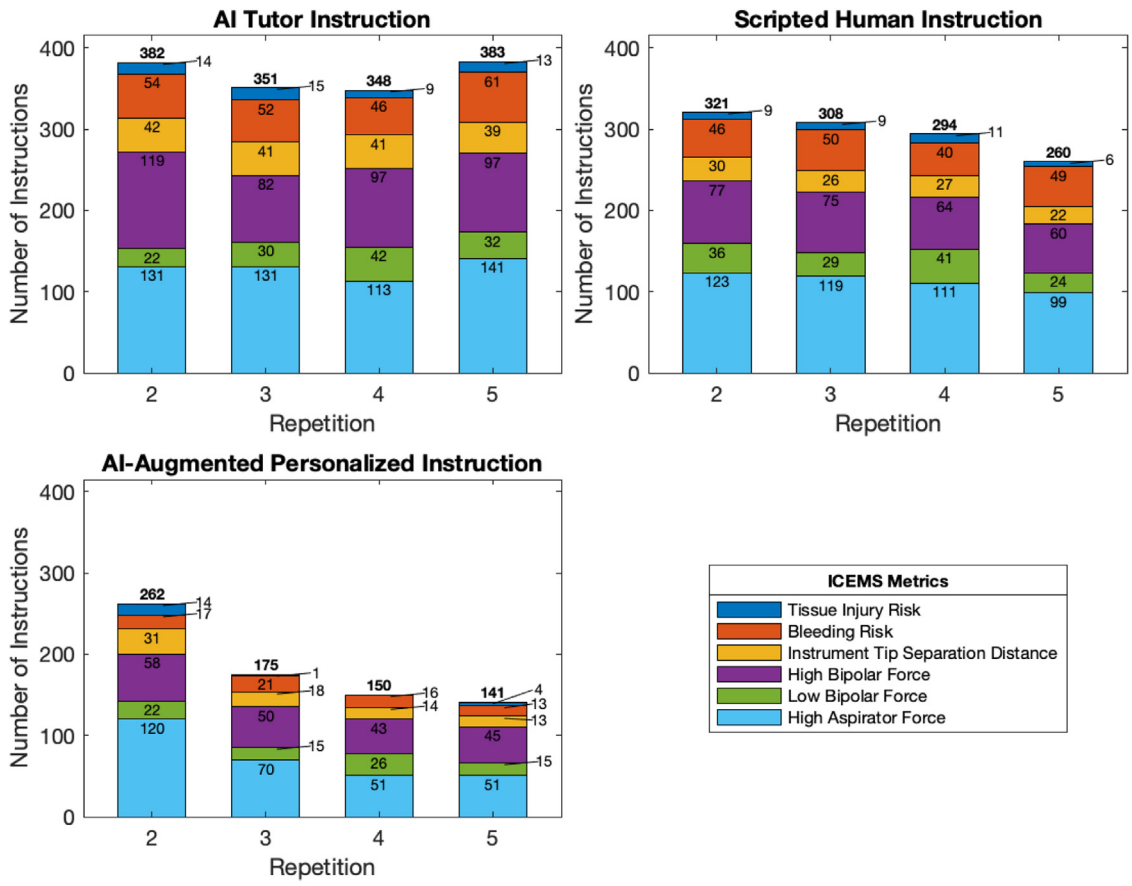

FIGURE 4. Number of instructions that each group received for each ICEMS metric across the second to fifth repetitions of the practice scenario. Metrics are color coded (see legend). X-axis represents the repetition number. Colored stacked bars represent the number of instructions for each ICEMS metric. Instructions were given upon metric error detection by the ICEMS. Total number of instructions are indicated in bold above each bar. AI, artificial intelligence; ICEMS, Intelligent Continuous Expertise Monitoring System.

TABLE 3. Incidence Rates and Standard Errors of Instructions for 6 ICEMS Metrics Received During the Second to Fifth Repetitions of the Practice Resection Scenario[‡]

Group	Incidence Rate (IR) [SE]						
	Healthy Tissue Injury Risk	Bleeding Risk	Instrument Tip Separation Distance	High Bipolar Force	Low Bipolar Force	High Aspirator Force	All Metrics Combined
Repetition 2							
1	0.15 [0.07]	0.88 [0.28]	1.03 [0.22]	3.24 [0.50]	0.59 [0.15]	3.84 [0.47]	12.81 [1.18]
2	0.07 [0.04]	0.78 [0.26]	0.79 [0.19]	2.26 [0.38]	1.00 [0.22]	3.94 [0.49]	10.24 [0.96]
3	0.11 [0.07]	0.16 [0.07]	0.83 [0.20]	1.68 [0.31]	0.65 [0.16]	3.92 [0.50]	8.45 [0.84]
Repetition 3							
1	0.16 [0.08]	0.85 [0.27]	1.00 [0.22]	2.24 [0.37]	0.80 [0.18]	3.84 [0.47]	11.63 [1.09]
2	0.07 [0.04]	0.84 [0.28]	0.68 [0.17]	2.20 [0.37]	0.81 [0.19]	3.81 [0.48]	9.83 [0.93]
3	0.008 [0.009]	0.20 [0.08]	0.48 [0.14]	1.45 [0.28]	0.44 [0.13]	2.29 [0.34]	5.65 [0.61]
Repetition 4							
1	0.093 [0.05]	0.75 [0.24]	1.00 [0.22]	2.64 [0.42]	1.12 [0.23]	3.32 [0.42]	11.56 [1.08]
2	0.091 [0.05]	0.67 [0.23]	0.71 [0.18]	1.87 [0.33]	1.14 [0.24]	3.55 [0.45]	9.38 [0.90]
3	$1.24E^{-11}$ [$3.17E^{-7}$]	0.15 [0.07]	0.38 [0.12]	1.24 [0.25]	0.77 [0.18]	1.67 [0.28]	4.84 [0.55]
Repetition 5							
1	0.14 [0.07]	0.99 [0.32]	0.95 [0.21]	2.64 [0.42]	0.85 [0.19]	4.14 [0.50]	12.58 [1.16]
2	0.050 [0.03]	0.83 [0.27]	0.58 [0.15]	1.76 [0.31]	0.67 [0.16]	3.17 [0.42]	8.30 [0.81]
3	0.033 [0.02]	0.12 [0.06]	0.35 [0.11]	1.30 [0.26]	0.44 [0.13]	1.67 [0.28]	4.55 [0.52]

Abbreviation: ICEMS, Intelligent Continuous Expertise Monitoring System.
[‡]Estimates are from the between-group statistical analyses of these ICEMS metrics.

1.03-4.39] instructions; p = 0.04) (Fig. 5E). By the fourth repetition of the practice scenario, groups 1 and 2 both received significantly more high aspirator force feedback instructions than group 3 (IRR, 1.99 [95% CI, 1.10-3.61] instructions; p = 0.01; and IRR, 2.13 [95% CI, 1.17-3.87] instructions; p = 0.004, respectively) (Fig. 5G). Only group 3 received significantly fewer instructions across repetitions of the practice scenario. By the third repetition, this group received significantly fewer total instructions (IRR, 1.50 [95% CI, 1.16-1.94] instructions; p < 0.001) and instructions relating to aspirator force (IRR, 1.71 [95% CI, 1.15-2.55] instructions; p = 0.002) compared with the second repetition (Figs. 5A and G).

Technical Skill Performance Across Simulated Practice Subpial Resections

Learning curves were assessed for the 5 technical skill performance metrics. The estimates for each metric analyzed using parametric statistical methods can be found in Table 4. No statistically significant differences were observed between the groups at baseline performance (first repetition) in all 5 performance metrics. Dunn's tests with Bonferroni correction indicated that group 3 demonstrated a significantly lower rate of healthy tissue removal compared to group 1 by the third (p = 0.01) repetition of the practice scenario, and that group 3 had significantly less bleeding than group 1 in the second (p = 0.02) and fourth (p = 0.02) repetitions of the practice scenario (Figure 6A and B). In addition, a pairwise test with Šidák adjustment indicated that group 3 demonstrated a significantly lower instrument tip separation distance by the second repetition of the practice scenario compared with group 1 (mean ratio, 1.25 [95% CI, 1.05-1.50] mm; p = 0.008) and compared with group 2 in the third (mean ratio, 1.20 [95% CI, 1.00-1.44] mm; p = 0.049) and fourth (mean ratio, 1.22 [95% CI, 1.02-1.46] mm; p = 0.03) repetitions of the practice scenario. This same statistical test found that group 2 had a significantly lower instrument tip separation distance than group 1 during the fifth repetition (mean ratio, 1.33 [95% CI, 1.11-1.59] mm; p < 0.001; Fig. 6C). No other statistically significant differences were observed between group 1 and 2 in the other performance metrics. A pairwise test with Šidák adjustment also indicated that, by the third repetition, group 3 used significantly less force with the ultrasonic aspirator than group 1 (mean ratio, 1.68 [95% CI, 1.23-2.31] N; p < 0.001) and group 2 (mean ratio, 1.50 [95% CI, 1.09-2.06] N; p = 0.007; Fig. 6E). No statistically significant differences were found between the groups in the force applied using the bipolar forceps (p > 0.05), as indicated by a

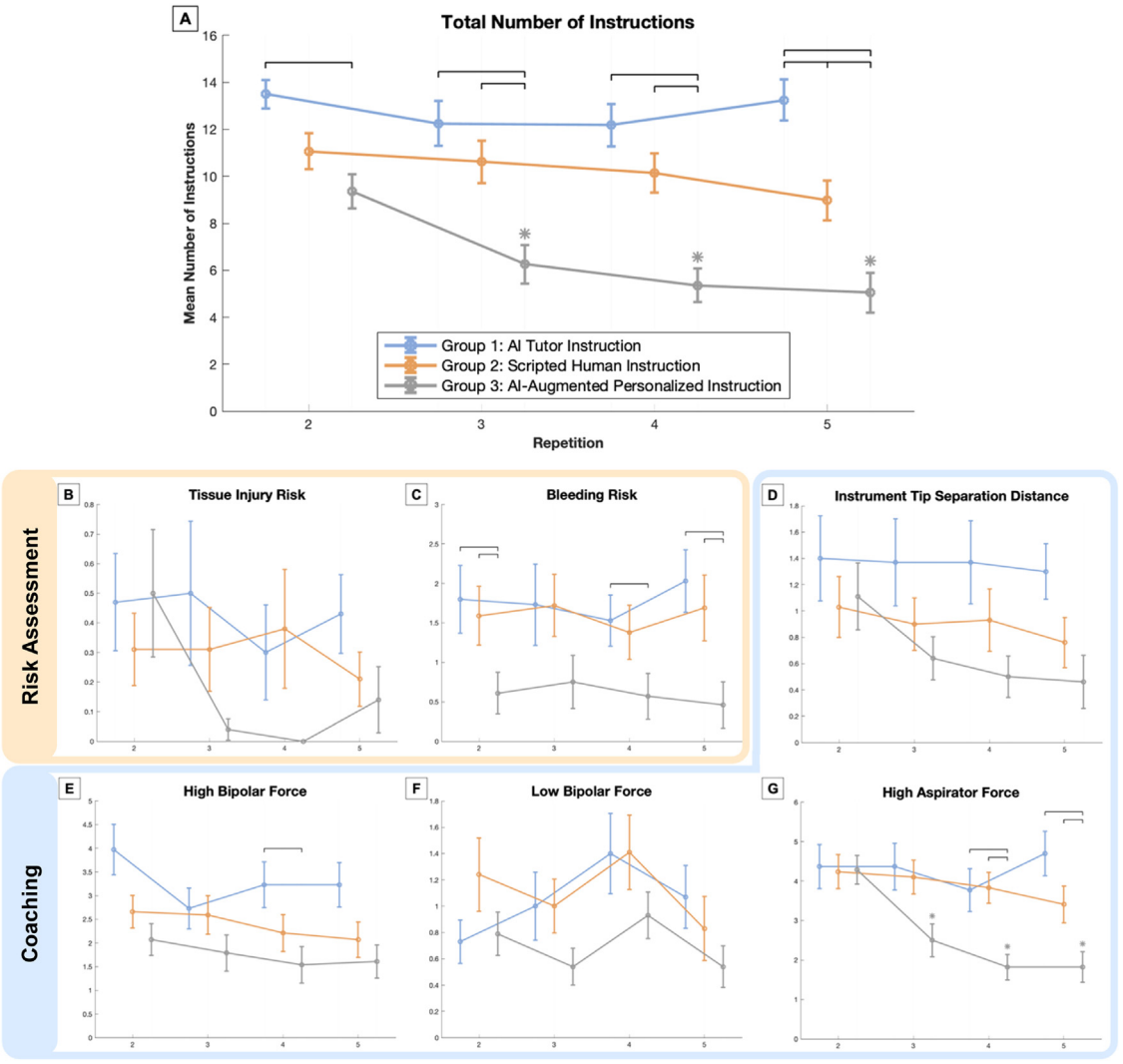

FIGURE 5. The mean total number of instructions and mean number of instructions in each ICEMS performance metric across the second to fifth repetitions of the practice scenario. Groups are color coded (see legend). The X-axis represents the repetition number. Points represent group means and error bars represent standard errors. Black horizontal brackets indicate statistically significant differences between groups ($p < 0.05$) during a given repetition. Asterisks indicate statistically significant differences from the baseline ($p < 0.05$) for that group. AI, artificial intelligence; ICEMS, Intelligent Continuous Expertise Monitoring System.

robust linear mixed model regression (Fig. 6D). Compared to baseline performance, by the second repetition of the practice scenario, group 1 significantly decreased the distance between their instruments (mean difference, 2.28 [95% CI, 0.57-3.99] mm; $p = 0.001$). This finding was also observed for group 2 (mean difference, 3.47 [95% CI, 2.30-4.64] mm; $p < 0.001$) and group 3 (mean ratio, 1.55 [95% CI, 1.36-1.77] mm; $p < 0.001$; Fig. 6C), as indicated by a pairwise test with Bonferroni adjustment. This same test found that, compared to baseline performance, groups 1, 2, and 3 significantly lowered the force applied with the bipolar forceps (mean difference, 0.08 [95% CI, 0.02-0.13] N; $p < 0.001$; mean difference, 0.13 [95% CI, 0.07-0.18] N; $p < 0.001$; and mean ratio, 1.38 [95% CI, 1.13-1.67] N; $p < 0.001$, respectively) by the second repetition (Fig. 6D). Nemenyi test showed that group 3 also achieved a significantly lower rate of healthy tissue removal ($p < 0.001$) and volume of blood lost ($p < 0.001$) by the second repetition compared to baseline, and a pairwise test with Bonferroni correction found that this group demonstrated a lower force applied with the ultrasonic aspirator (mean difference, 0.05 [95% CI, 0.02-0.07] N; $p < 0.001$) by the second repetition compared to baseline (Figs. 6A, B and E). Other improvements from baseline performance and between specific trials are shown in Figure 6.

TABLE 4. Estimated Geometric Means and Standard Errors of Technical Skill Performance Metrics Over 6 Repetitions of the Practice Resection Scenario[‡]

Group	Estimated Geometric Mean (EGM) [SE]		
	Instrument Tip Separation Distance (mm)	Force Applied With Bipolar Forceps (N)	Force Applied With Ultrasonic Aspirator (N)
Repetition 1			
1	11.79 [0.61]	0.43 [0.02][§]	0.16 [0.01]
2	11.84 [0.63]	0.43 [0.02][§]	0.15 [0.01]
3	11.76 [0.63]	0.44 [0.02][§]	0.16 [0.01]
Repetition 2			
1	9.52 [0.50]	0.36 [0.02][§]	0.13 [0.01]
2	8.33 [0.44]	0.30 [0.02][§]	0.11 [0.01]
3	7.59 [0.41]	0.32 [0.02][§]	0.11 [0.01]
Repetition 3			
1	8.84 [0.46]	0.33 [0.02][§]	0.13 [0.01]
2	8.20 [0.43]	0.31 [0.02][§]	0.12 [0.01]
3	6.84 [0.37]	0.27 [0.02][§]	0.08 [0.01]
Repetition 4			
1	8.78 [0.46]	0.33 [0.02][§]	0.13 [0.01]
2	7.69 [0.41]	0.29 [0.02][§]	0.11 [0.01]
3	6.32 [0.34]	0.26 [0.02][§]	0.07 [0.01]
Repetition 5			
1	9.87 [0.51]	0.36 [0.02][§]	0.14 [0.01]
2	7.42 [0.39]	0.28 [0.02][§]	0.11[0.01]
3	6.62 [0.36]	0.27 [0.02][§]	0.07 [0.01]
Repetition 6			
1	9.82 [0.51]	0.37 [0.02][§]	0.14 [0.01]
2	8.32 [0.44]	0.30 [0.02][§]	0.12 [0.01]
3	7.10 [0.38]	0.28 [0.02][§]	0.08 [0.01]

[‡]Only metrics analyzed using parametric statistical methods are included. Estimates are from the between-group statistical analyses of these technical skill performance metrics.
[§]Estimated marginal mean (EMM) [SE].

Technical Skill Transfer to the Simulated Realistic Subpial Resection

The estimates for each metric analyzed using parametric statistical methods can be found in Table 5. Following the completion of the realistic scenario, a pairwise test with Šidák adjustment indicated that group 3 applied significantly less force with the ultrasonic aspirator than group 1 (mean difference, 0.04 [95% CI, 0.01-0.07] N; p = 0.01) (Figure 7E). No other statistically significant differences were found between the groups.

DISCUSSION

To the authors' knowledge, this cohort study is the first investigation to demonstrate the pedagogical impact of AI-augmented personalized instruction on the frequency of feedback instructions and on the results of these specific surgical instructions on trainee surgical performance. A previous RCT used ICEMS scores to explore the effect of the 3 different instructional methods utilized in this investigation on trainee skill acquisition and skill transfer.[21] This study builds on this investigation, focusing on the frequency of instructions provided and their impact on changes in technical skill.

Consistent with our first hypothesis, participants receiving AI-augmented personalized instruction received fewer total instructions compared with AI tutor and scripted human instruction. Since feedback was only provided when the ICEMS detected an error, a reduced feedback frequency can be associated with trainees making fewer performance errors. Thus, fewer feedback instructions suggests that AI-augmented instructions may be more comprehensible and provide more clarity to trainees to understand how to correct errors in their performance. In the second repetition, when trainees first began receiving feedback, the AI-augmented personalized instruction group received significantly fewer instructions compared to those trained by the AI tutor, providing evidence for this methodology's immediate efficacy for teaching trainees how to correct errors. This trainee group was the only 1 to receive significantly fewer

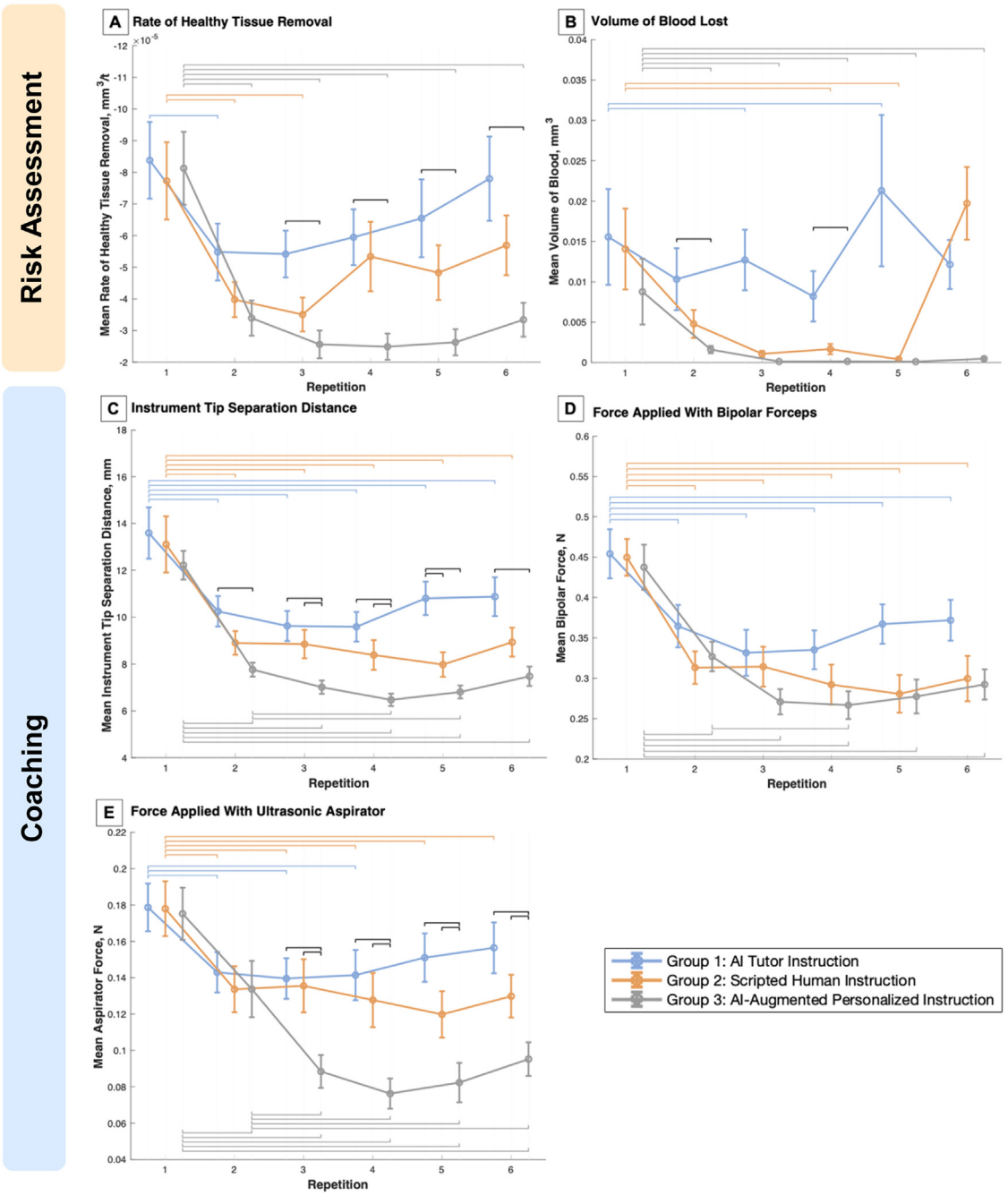

FIGURE 6. The learning curves of 5 technical skill performance metrics across 6 repetitions of the practice scenario. Groups are color coded (see legend). The X-axis represents the repetition number. Points represent group means and error bars represent standard errors. Since the simulator records the metrics at a frequency of 50 Hz, the unit of time (t) is equal to 20 ms. Black horizontal brackets indicate statistically significant differences between groups ($p < 0.05$) during a given repetition. Within-group differences are represented by horizontal brackets in the respective color for that group ($p < 0.05$). AI, artificial intelligence.

total instructions throughout the session. This finding is consistent with the concept that personalized instructions provided sufficient context to be actionable in real time, as this group consistently had fewer errors in performance.

The AI-augmented personalized instruction group had significantly lower values compared to the AI tutor instruction group for both risk metrics assessed and two of the 3 coaching metrics, consistent with our second hypothesis. Except for instructions pertaining to bipolar

TABLE 5. Estimated Marginal Means and Standard Errors of Technical Skill Performance Metrics During the Realistic Resection Scenario[‡]

Group	Estimated Marginal Mean (EMM) [SE]		
	Instrument Tip Separation Distance (mm)	**Force Applied With Bipolar Forceps (N)**	**Force Applied With Ultrasonic Aspirator (N)**
1	15.51 [1.28]	0.40 [0.02]	0.12 [0.01]
2	13.92 [1.30]	0.34 [0.02]	0.11 [0.01]
3	12.16 [1.32]	0.34 [0.02]	0.08 [0.01]

[‡]Only metrics analyzed using parametric statistical methods are included. Estimates are from the between-group statistical analyses of these technical skill performance metrics.

force, all the instructions given directed participants to decrease the technical skill performance metric values (Table 1). Thus, achieving lower values in these metrics suggests a better understanding of the feedback instructions. The absence of significant differences in the bipolar force applied may be attributable to participants receiving 2 instructions for this metric − 1 to increase and another to decrease the force applied with the bipolar forceps throughout the session. This may have allowed participants in all groups to learn the ideal amount of force application with this instrument. Employing a similar methodology for teaching the other metrics, where trainees are made aware of appropriate changes to performance as much as they are made aware of errors, has proven beneficial for surgical skill acquisition and may be an avenue to explore in future studies involving the ICEMS.[36]

The scripted human instruction group did not consistently receive fewer instructions or exhibit significantly better technical skill performance compared with the AI tutor instruction group. These groups received instructions using identical wording (Table 1), suggesting that the instructions programmed into the ICEMS may not provide sufficient information to allow trainees to learn to consistently correct errors. This is further supported by the AI-augmented personalized instruction group's fewer instructions received and outperformance of both other groups in several metrics. Investigations into the instructions that elicited the most appropriate changes in performance are presently being conducted using a series of Large Language Models (LLMs) to further optimize both the ICEMS and human expert instructions to enhance performance outcomes.[37]

The results indicate that correcting performance relating to a high amount of force applied with the ultrasonic aspirator may be most effectively accomplished with AI-augmented instructions. This group received fewer instructions for this metric and outperformed both other groups during the summative assessment in repetition 6, and group 1 during the realistic scenario by using significantly less force. Studies focused on exploring the utility of LLMs to understand the reasons for the success of AI-augmented personalized instructions for this particular metric may further enhance the actionable vocabulary of the ICEMS.

These findings have supported the hypothesis that providing skilled instructors with AI-generated error data to facilitate the provision of personalized, continuous, contextualized feedback improves learning in a simulated surgical environment. Further research is required to determine whether these findings can be generalized to more realistic surgical settings, and such studies using an *ex vivo* animal model are currently underway.[38-41] This cohort study demonstrates the potential for AI-augmented personalized instruction to optimize trainee assessment, teaching, and error mitigation in the operating room environment, helping to lay the foundations for the development of future intelligent human operating rooms powered by AI technology.

LIMITATIONS

Intelligent tutoring systems cannot completely replicate the communication interchange between a surgical educator and a learner in complex human operating settings.[42] This study was carried out using a small sample of medical students in their preparatory, first, or second year, and findings cannot be generalized to senior medical students or surgical residents. However, the results of a series of simulation studies has demonstrated that using medical students with minimal surgical experience has provided valuable insights.[11,12,17,20] Investigations using neurosurgical residents, fellows, and neurosurgeons are in preparation involving *ex vivo* models, but the limited number of participants available may limit the ability of these studies to achieve sufficient power to detect statistically significant differences unless multiple teaching centers are involved.

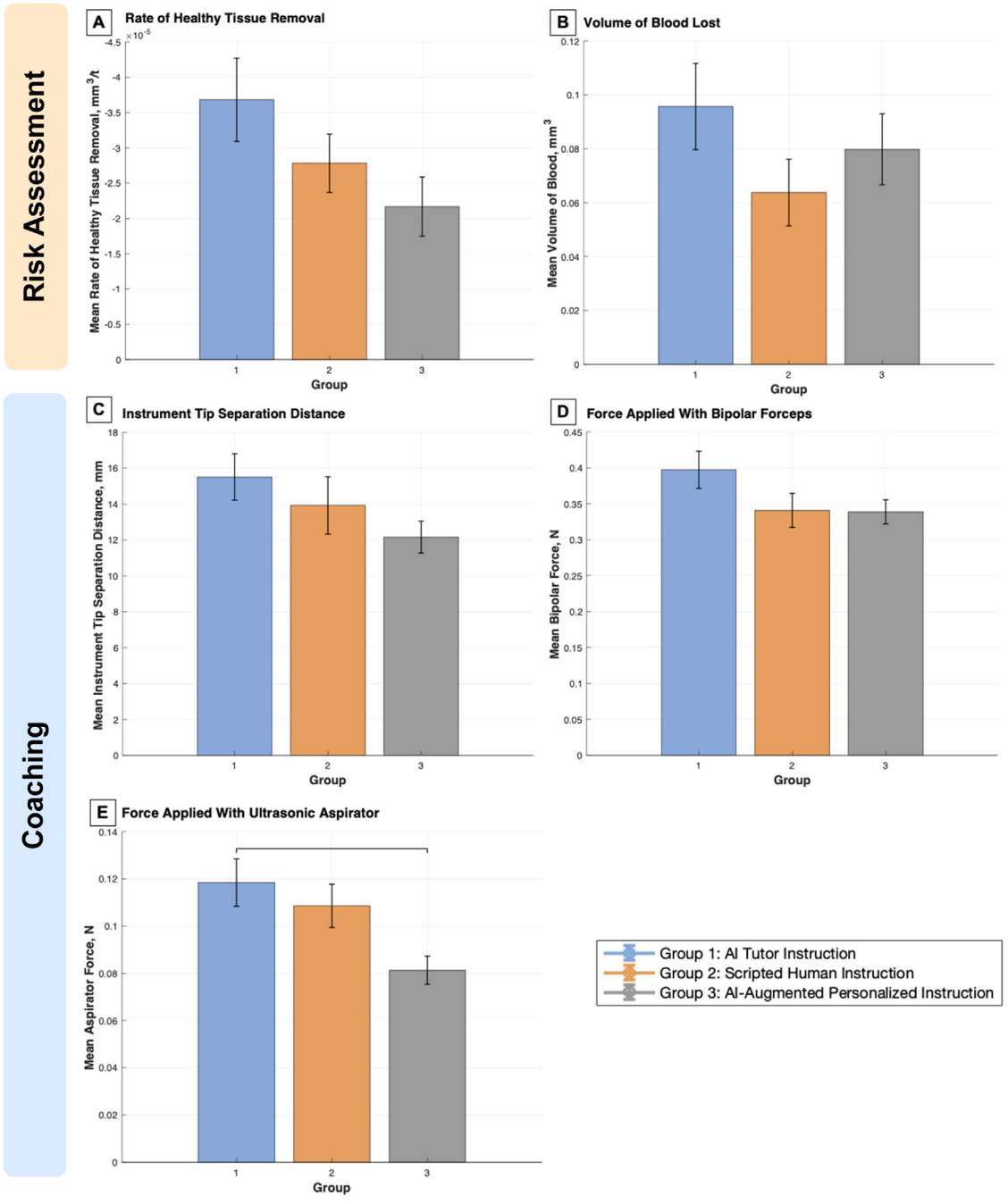

FIGURE 7. Five technical skill performance metrics during the realistic scenario. Groups are color coded (see legend). The X-axis represents the group. Colored bars represent group means and error bars represent standard errors. Since the simulator records the metrics at a frequency of 50 Hz, the unit of time (t) is equal to 20 ms. Black horizontal brackets indicate statistically significant differences between groups ($p < 0.05$). AI, artificial intelligence.

CONCLUSION

This cross-sectional cohort study demonstrated that artificial intelligence-augmented personalized instruction resulted in less frequent feedback and improved surgical technical skills.

These results continue to outline the importance of human educator engagement and the critical role they play in developing intelligent tutoring systems for surgical education applicable to the human operating room.

PREVIOUS PRESENTATIONS

Portions of this work were presented at the Canadian Conference for the Advancement of Surgical Education (C—CASE) in Toronto, ON, Canada, on October 17, 2024, and at McGill University's Neurosurgical Research Day in Montreal, QC, Canada on June 20, 2025.

DATA AVAILABILITY STATEMENT

The dataset is available from the authors on a reasonable request.

CODE AVAILABILITY STATEMENT

The codes used in this study are available from the authors on a reasonable 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. https://doi.org/10.1016/j.surg.2006.01.008.
2. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. *N Engl J Med*. 2013;369(15):1434–1442. https://doi.org/10.1056/NEJMsa1300625.
3. Stulberg JJ, Huang R, Kreutzer L, et al. Association between surgeon technical skills and patient outcomes. *JAMA Surg*. 2020;155(10):960–968. https://doi.org/10.1001/jamasurg.2020.3007.
4. Gawande AA, Zinner MJ, Studdert DM, Brennan TA. Analysis of errors reported by surgeons at three teaching hospitals. *Surgery*. 2003;133(6):614–621. https://doi.org/10.1067/msy.2003.169.
5. Marcus H, Vakharia V, Kirkman MA, Murphy M, Nandi D. Practice makes perfect? The role of simulation-based deliberate practice and script-based mental rehearsal in the acquisition and maintenance of operative neurosurgical skills. *Neurosurgery*. 2013;72:A124–A130. https://doi.org/10.1227/NEU.0b013e318270d010.
6. Bezemer J, Cope A, Kress G, Kneebone R. Holding the scalpel: achieving surgical care in a learning environment. *J Contemp Ethnogr*. 2013;43(1):38–63. https://doi.org/10.1177/0891241613485905.
7. George EI, Skinner A, Pugh CM, Brand TC. Performance assessment in minimally invasive surgery. Köhler TS, Schwartz B, editors. Surgeons as Educators: a Guide for Academic Development and Teaching Excellence, Cham: Springer; 2018:53–91. https://doi.org/10.1007/978-3-319-64728-9_5.
8. Hamdorf JM, Hall JC. Acquiring surgical skills. *Br J Surg*. 2000;87(1):28–37. https://doi.org/10.1046/j.1365-2168.2000.01327.x.
9. Mirza M, Koenig JF. Teaching in the operating room. Köhler TS, Schwartz B, editors. Surgeons as Educators: a Guide for Academic Development and Teaching Excellence, Cham: Springer; 2018:137–160. https://doi.org/10.1007/978-3-319-64728-9_8.
10. Curry JI. 'See one, practise on a simulator, do one' – the mantra of the modern surgeon. *S Afr J Surg*. 2011;49(1):4–6.
11. 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;14:15130. https://doi.org/10.1038/s41598-024-65716-8.
12. 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. https://doi.org/10.1001/jamanetworkopen.2021.49008.
13. Alkadri S, Del Maestro RF, Driscoll M. Unveiling surgical expertise through machine learning in a novel VR/AR spinal simulator: a multilayered approach using transfer learning and connection weights analysis. *Comput Biol Med*. 2024;179:108809. https://doi.org/10.1016/j.compbiomed.2024.108809.
14. Natheir S, Christie S, Yilmaz R, et al. Utilizing artificial intelligence and electroencephalography to assess expertise on a simulated neurosurgical task. *Comput Biol Med*. 2023;152:106286. https://doi.org/10.1016/j.compbiomed.2022.106286.
15. 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. https://doi.org/10.1371/journal.pone.0229596.
16. 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;5:54. https://doi.org/10.1038/s41746-022-00596-8.

17. 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(Supplement_1):129–130. https://doi.org/10.1227/neu.0000000000002809_428.

18. 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 Md*. 2022;23(1):22–30. https://doi.org/10.1227/ons.0000000000000232.

19. Delorme S, Laroche D, DiRaddo R, Del Maestro R. NeuroTouch: a physics-based virtual simulator for cranial microneurosurgery training. *Oper Neurosurg*. 2012;71(suppl_1):32–42. https://doi.org/10.1227/NEU.0b013e318249c744.

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

21. Giglio B, Albeloushi A, AKh Alhaj, et al. Artificial intelligence—augmented human instruction and surgical simulation performance: a randomized clinical trial. *JAMA Surg*. 2025;160(9):993–1003. https://doi.org/10.1001/jamasurg.2025.2564.

22. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med*. 2007;147(8):573–577. https://doi.org/10.7326/0003-4819-147-8-200710160-00010.

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

24. Alotaibi FE, AlZhrani GA, Mullah MAS, et al. Assessing bimanual performance in brain tumor resection with NeuroTouch, a virtual reality simulator. *Oper Neurosurg*. 2015;11(1):89–98. https://doi.org/10.1227/NEU.0000000000000631.

25. 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. https://doi.org/10.1001/jamanetworkopen.2019.8363.

26. Gélinas-Phaneuf N, Choudhury N, Al-Habib AR, et al. Assessing performance in brain tumor resection using a novel virtual reality simulator. *Int J Comput Assist Radiol Surg*. 2014;9(1):1–9. https://doi.org/10.1007/s11548-013-0905-8.

27. 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. https://doi.org/10.3171/2021.12.JNS211563.

28. 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. https://doi.org/10.1016/j.wneu.2020.03.187.

29. Bugdadi A, Sawaya R, Bajunaid K, et al. Is virtual reality surgical performance influenced by force feedback device utilized? *J Surg Educ*. 2019;76(1):262–273. https://doi.org/10.1016/j.jsurg.2018.06.012.

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

31. R Core Team. R: a language and environment for statistical computing. https://www.R-project.org/. Accessed June 19, 2025.

32. Bates D, Maechler M, Bolker B, et al. lme4: linear mixed-effects models using "Eigen" and S4. Accessed June 19, 2025. https://github.com/lme4/lme4/.

33. Brooks M, Bolker B, Kristensen K, et al. glmmTMB: generalized linear mixed models using template model builder. Accessed June 19, 2025. https://github.com/glmmTMB/glmmTMB.

34. Hartig F, Lohse L, leite M de S. DHARMa: residual diagnostics for hierarchical (multi-level /mixed) regression models. Accessed June 19, 2025. http://florianhartig.github.io/DHARMa/.

35. Koller M. robustlmm: robust linear mixed effects models. Accessed June 19, 2025. https://github.com/kollerma/robustlmm.

36. Levy IM, Pryor KW, McKeon TR. Is teaching simple surgical skills using an operant learning program more effective than teaching by demonstration? *Clin Orthop*. 2016;474:945–955. https://doi.org/10.1007/s11999-015-4555-8.

37. Du Y, Chen K, Zhan Y, et al. LMT++: adaptively collaborating LLMs with multi-specialized teachers for continual VQA in robotic surgical videos. *IEEE Trans Med Imaging*. 2025. https://doi.org/10.1109/TMI.2025.3581108.

38. 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*. 2024;27(1):106–113. https://doi.org/10.1227/ons.0000000000001044.

39. 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. https://doi.org/10.1016/j.wneu.2021.08.061.

40. 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. https://doi.org/10.1016/j.wneu.2020.12.141.

41. 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. https://doi.org/10.1016/j.wneu.2020.07.209.

42. 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. *Glob Surg Educ - J Assoc Surg Educ*. 2024;3:83. https://doi.org/10.1007/s44186-024-00284-z.