

**Effect of Artificial Intelligence-Augmented Human Instruction on Surgical Simulation
Performance: A Randomized Controlled Trial**

Bianca Giglio, BSc



Department of Surgical and Interventional Sciences

Faculty of Medicine and Health Sciences

McGill University

Montreal, Canada

May 2025

A thesis submitted to McGill University in partial fulfillment of the requirement of the degree of
Master of Science.

© Bianca Giglio 2025

TABLE OF CONTENTS

Abstract	v
Background	v
Objectives	v
Methods	v
Results	vi
Conclusion	vi
Résumé	vii
Contexte	vii
Objectifs	vii
Méthodes	vii
Résultats	viii
Conclusion	ix
Dedication and Preface	x
Acknowledgements	xi
Author Contributions	xiii
Abbreviations	xv
Thesis Introduction	1
Background	3
History of Neurosurgical Education	3
Simulation in Surgical Education	6
Simulation in Neurosurgical Education	10
Performance Assessment in Surgery: Foundations	12
Performance Assessment in Surgery: Innovations	14
Intelligent Tutoring Systems	16
The Study Objectives	20
The Study Hypothesis	20
The Study	21
Key Points	23
Abstract	24
Introduction	26

Methods	27
Participants	28
Randomization	28
Simulation Session	28
Study Procedure	29
Interventions	29
Group 1 (Control): AI Tutor Instruction	30
Group 2 (Experimental): Expert Instruction	30
Group 3 (Experimental): Personalized Expert Instruction	31
Outcome Measures	31
Statistical Analysis	31
Results	32
Performance Across Practice Subpial Resection Trials	32
Performance During Realistic Subpial Resection Task	33
Emotions and Cognitive Load	34
Discussion	34
Limitations	37
Conclusion	37
Thesis Summary	39
Discussion	39
Limitations and Future Directions	43
Conclusion	45
References	46
Tables	61
Table 1. ICEMS Metrics and Commands	61
Table 2. Demographic Characteristics of Included Study Participants	62
Figures	63
Figure 1. The Medical Emotions Scale	63
Figure 2. NeuroVR Practice Subpial Tumour Resection Scenario	64
Figure 3. NeuroVR Realistic Subpial Tumour Resection Scenario	65
Figure 4. Participant Recruitment Flowchart	66

Figure 5. Flow Chart of Events in Randomized Controlled Trial	67
Figure 6. Performance Assessment Across Practice Trials	68
Figure 7. Performance Assessment During Realistic Task	69
Figure 8. Emotions and Cognitive Load Throughout Simulation Training	70

ABSTRACT

Background

With current surgical teaching models' lack of standardization and reliance on subjective assessments by human experts rather than quantitative performance data, training novices to master surgical technical skills remains challenging. To mitigate this issue, we developed an artificial intelligence (AI) application known as the Intelligent Continuous Expertise Monitoring System (ICEMS) capable of assessing bimanual surgical skills at 0.2-second intervals and providing continuous, action-oriented verbal feedback.

Objectives

The objective of this study is to determine the effect of AI-augmented personalized expert instruction versus AI tutor instruction alone on surgical performance, skill transfer, and affective-cognitive responses.

Methods

A multi-institutional randomized controlled trial was conducted wherein medical students performed subpial brain tumour resection tasks on the NeuroVR and received real-time feedback on their performance. Students were stratified based on their year in medical school and block randomized to one of three groups. Group 1 received AI tutor instruction delivered by the ICEMS, group 2 received expert feedback in identical words to the ICEMS, and group 3 received AI data-informed personalized expert feedback. Trainees performed six practice subpial resection tasks to assess learning followed by a complex realistic brain tumour resection scenario to assess skill transfer. The ICEMS quantitatively evaluated trainee performance. Participants

self-reported emotions before, during, and after training and cognitive load after training via questionnaires.

Results

Eighty-seven medical students from four Quebec institutions were randomly assigned to the AI instruction (n = 30), expert instruction (n = 29), and personalized expert instruction (n = 28) groups. The ICEMS assessed and scored 522 practice resections and 87 realistic resections. During the practice tasks, personalized expert instruction resulted in significantly greater expertise scores than AI tutor instruction across several trials, including trial 5 (mean difference, 0.26 [95% CI, 0.09 to 0.43]; $P = 0.01$). During the realistic task, the personalized instruction group had significantly higher expertise scores than both the AI tutor instruction (mean difference, 0.20 [95% CI, 0.06 to 0.34]; $P = 0.02$) and expert instruction (mean difference, 0.18 [95% CI, 0.03 to 0.32]; $P = 0.049$) groups. The personalized expert instruction group also achieved significantly higher scores than the other two groups in certain metrics, such as bleeding and injury risk. Emotions and cognitive load demonstrated significant differences.

Conclusion

Personalized expert instruction resulted in enhanced surgical performance and skill transfer compared with intelligent tutor instruction, highlighting the importance of human input and active participation in AI-based surgical training.

RÉSUMÉ

Contexte

Le manque de standardisation des modèles actuels d'enseignement chirurgical et leur dépendance aux évaluations subjectives d'experts humains plutôt qu'à des données quantitatives sur la performance rendent l'apprentissage des compétences techniques chirurgicales difficile pour les novices. Pour remédier à ce problème, nous avons développé une application d'intelligence artificielle (IA) appelée *Intelligent Continuous Expertise Monitoring System* (ICEMS), capable d'évaluer les compétences chirurgicales bimanuelles à des intervalles de 0,2 seconde et de fournir un retour verbal continu orienté vers l'action.

Objectifs

L'objectif de cette étude est de déterminer l'effet d'une instruction personnalisée par un expert, augmentée par l'IA, comparé à une instruction par le tuteur IA seul, sur la performance chirurgicale, le transfert des compétences et les réponses affectivo-cognitives.

Méthodes

Un essai contrôlé aléatoire multi-institutionnel a été mené, dans lequel des étudiants en médecine ont réalisé des tâches de résection tumorale sous-piale sur le simulateur NeuroVR tout en recevant un retour en temps réel sur leur performance. Les étudiants ont été classifiés en fonction de leur année d'études en médecine et répartis aléatoirement par blocs dans l'un des trois groupes. Le premier groupe a reçu une instruction par le tuteur IA via l'ICEMS. Le deuxième groupe a bénéficié d'un retour d'expert utilisant les mêmes termes que l'ICEMS. Le

troisième groupe a reçu un retour personnalisé d'un expert informé par les données de l'IA. Les participants ont réalisé six tâches de résection sous-piale pour évaluer leur apprentissage, suivies d'un scénario réaliste de résection tumorale complexe pour évaluer le transfert de compétences. L'ICEMS a quantitativement évalué la performance des participants. Ces derniers ont également auto-évalué leurs émotions avant, pendant et après la formation, ainsi que leur charge cognitive après l'entraînement via des questionnaires.

Résultats

Quatre-vingt-sept étudiants en médecine provenant de quatre institutions québécoises ont été répartis aléatoirement dans les groupes instruction par IA ($n = 30$), instruction par expert ($n = 29$) et instruction personnalisée par expert ($n = 28$). L'ICEMS a évalué et noté 522 résections d'entraînement et 87 résections réalistes. Pendant les tâches d'entraînement, l'instruction personnalisée par expert a permis d'obtenir des notes d'expertise significativement plus élevées que l'instruction par tuteur IA sur plusieurs essais, notamment l'essai 5 (différence moyenne : 0,26 [IC 95 %, 0,09 à 0,43] ; $P = 0,01$). Lors de la tâche réaliste, le groupe ayant reçu une instruction personnalisée a obtenu des notes d'expertise significativement plus élevées que le groupe tuteur IA (différence moyenne : 0,20 [IC 95 %, 0,06 à 0,34] ; $P = 0,02$) et le groupe instruction par expert (différence moyenne : 0,18 [IC 95 %, 0,03 à 0,32] ; $P = 0,049$). Le groupe instruction personnalisée par expert a également obtenu des notes significativement supérieures aux deux autres groupes pour certains critères, notamment le risque de saignement et le risque de blessure. Des différences significatives ont également été observées concernant les émotions et la charge cognitive des participants.

Conclusion

L'instruction personnalisée par expert a conduit à une amélioration des performances chirurgicales et du transfert de compétences, par rapport à l'instruction par tuteur IA, soulignant l'importance de l'apport humain et de la participation active dans la formation chirurgicale basée sur l'IA.

DEDICATION AND PREFACE

To women in male-dominated fields: your voices matter and your contributions are indispensable. Keep turning the tides.

This thesis is original work by the candidate and is structured in a manuscript-based format.

An abstract of this study, entitled “Efficiency of Verbal Intelligent Tutor Instruction in Neurosurgical Simulation: A Randomized Controlled Trial,” was presented at the Canadian Conference for the Advancement of Surgical Education (C-CASE) in Toronto, ON, Canada on October 17, 2024 and published in the *Canadian Journal of Surgery* on February 6, 2025.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my heartfelt appreciation to my supervisor, Dr. Rolando Del Maestro. Thank you for giving me the opportunity to work in your lab and trusting me to carry out this project. I'm endlessly inspired by your dedication to neurosurgical education, to the history of medicine, to art, to love. You taught me to strive for excellence and nothing less in all that I do. As you say, "People always make room for excellence"—I will carry that with me always. It has been the honour of my life to learn from someone as accomplished as you and I look forward to continuing to work together in the future.

I would also like to acknowledge the members of my research advisory committee—Dr. Jacques Lapointe, Dr. Amir Hooshir, and Dr. Housseem-Eddine Gueziri—for their support and guidance throughout my graduate journey.

I am deeply grateful to the members of the Neurosurgical Simulation and Artificial Intelligence Learning Centre with whom I've had the opportunity to work. To Nour Abou Hamdan, who took me under her wing three years ago: working on your thesis project with you was what first sparked my passion for research. To Neevya Balasubramaniam, my collaborator on countless projects from the very beginning: you are one of the kindest and most hardworking people I know, and you will make a fantastic neurosurgeon. To my colleagues throughout the years, Vanja Davidovic, Abicumaran Uthamacumaran, Dr. Recai Yilmaz, Trisha Tee, Ali Fazlollahi, Dr. Abdulmajeed Albeloushi, Dr. Ahmad Alhaj, Dr. Mohamed Alhantoobi, Dr. Rothaina Saeedi, Puja Pachchigar, and Chinyelum Agu: your intelligence, dedication, passion, and hard work inspire me every day.

I would not be where I am without my family's unending love and support. To my dad, Marco Giglio, who, all my life, made sure that I had everything I could ever need. I never doubt

that you're proud of me, and I'm proud of you too. To my mom, Dimitra Demerson, the first person to ever believe in me. Even back when it was just elementary school math competitions and high school Brain Bees, you always made me feel like I could be someone. Everything I am is because of you. To my sister and best friend, Chiara Giglio, who brings colour and light to my life. To my brother, Franco Giglio, who lets me hug him sometimes. To Hershey, Kisses, Loki, Fuji, and Albi, for being the cutest and most perfect critters in the whole world.

And to my love, Jason Lapointe, who spent weeks working on this project with me—not because he stands to gain anything, but because I am lucky enough to be loved by him. I was in a very different place when I started this degree—before I met you. Now, tenderness and serenity and joy are everywhere, and I give in to them.

AUTHOR CONTRIBUTIONS

The candidate led this trial and contributed to all aspects of the study, including design of the study protocol, instructor training, data acquisition, statistical analysis, results interpretation, and manuscript writing. The candidate had full access to all the data and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Dr. Abdulmajeed Albeloushi, Dr. Ahmad Kh. Alhaj, and Dr. Mohamed Alhantoobi contributed to the instruction of participants.

Dr. Rothaina Saeedi contributed to the acquisition of data.

Vanja Davidovic contributed to the study conceptualization and design, assisted in running trials, and provided statistical input.

Abicumaran Uthamacumaran and Jason Lapointe assisted with data interpretation and statistical analysis.

Dr. Recai Yilmaz contributed to the study conceptualization and design and assisted with data interpretation.

Neevya Balasubramaniam, Widad Safih, and Sabrina Deraiche assisted with participant recruitment.

Trisha Tee contributed to the study conceptualization and design and provided insight on the acquisition, analysis, and interpretation of affective and cognitive data.

Ali M. Fazlollahi provided insight on the trial design, methods, and data analysis.

Dr. José A. Correa contributed to statistical analysis.

Dr. Rolando F. Del Maestro supervised the project plan and was primarily in charge of the study's overall direction. He acquired the ethics approval, provided the neurosurgical expertise and resources necessary to design, conduct, and interpret the results of the study,

assisted in recruitment, contributed to the preparation of the manuscript, and offered ongoing support at every step of the trial.

ABBREVIATIONS

3D	Three-Dimensional
AI	Artificial Intelligence
ANCOVA	Analysis of Covariance
ANOVA	Analysis of Variance
AR	Augmented Reality
CBME	Competency-Based Medical Education
CI	Confidence Interval
CLI	Cognitive Load Index
EPA	Entrustable Professional Activity
ICEMS	Intelligent Continuous Expertise Monitoring System
LSTM	Long Short-Term Memory Network
MES	Medical Emotions Scale
ML	Machine Learning
OR	Operating Room
OSATS	Objective Structured Assessment of Technical Skills
PGY	Post-Graduate Year
RCT	Randomized Controlled Trial
SD	Standard Deviation
VOA	Virtual Operative Assistant
VR	Virtual Reality

INTRODUCTION

Surgery is one of the most complex and high-stakes fields in medicine, necessitating expertise in a range of competencies including technical skills, anatomical knowledge, clinical reasoning, leadership, and communication.^{1,2} A significant share of patient harm incidents is attributable to technical errors in surgical procedures.³ This is especially true in neurosurgery, where even minor errors can result in patient mortality, morbidity, decreased quality of life, and impaired functioning.⁴ Consequently, quality surgical education is a key determinant of positive patient outcomes.⁵⁻⁷

At present, surgical residency programs follow the apprenticeship-based training model first established by Dr. William Halsted.^{8,9} However, as the field of surgical education transitions towards a competency-based framework, the limitations associated with the Halstedian model are becoming increasingly evident. The lack of structured, standardized surgical curricula¹⁰⁻¹³ coupled with the absence of hands-on skill development opportunities outside the operating room (OR)⁸ raise concerns about patient safety. Although the adoption of competency-based medical education (CBME) has led to the implementation of surgical performance assessment methods such as the Objective Structured Assessment of Technical Skills (OSATS), this framework is based on qualitative visual ratings, making it prone to evaluator bias and subjectivity.¹⁴ The demand for structured, standardized surgical curricula, opportunities for risk-free technical skill development, and objective, quantitative performance assessment is more critical than ever.

Innovative technologies like virtual reality (VR) and artificial intelligence (AI) are having an increasingly significant impact across many disciplines, with surgical education being no exception. These advancements are expanding the scope of training to support learners in attaining mastery. VR surgical simulators replicate the tactile, visual, and auditory conditions of

real-world operative procedures, providing learners with controlled, risk-free environments for technical proficiency training. These simulators capture extensive amounts of data through user interaction, measuring various parameters such as instrument position, healthy tissue injury risk, and volume of blood loss.¹⁵ AI algorithms can be employed to analyze this data and assess metric-specific performance based on predetermined expert benchmarks.¹⁶ Based on this AI-based performance assessment, intelligent tutoring systems can provide targeted metric feedback to improve trainee performance.^{16,17} When combined, VR surgical simulators and intelligent tutors can offer an optimal setting for deliberate practice. To adapt to the growing demand for competency-based surgical curricula, our team developed an intelligent tutoring system that can be integrated into VR surgical simulators to train bimanual psychomotor skills.¹⁶

The Intelligent Continuous Expertise Monitoring System (ICEMS) is a deep learning application that continuously assesses surgical performance in 0.2-second intervals, calculates an expertise score on a scale of -1.00 (novice) to 1.00 (expert), and provides real-time action-oriented verbal feedback to improve trainee performance and mitigate errors.¹⁶ The ICEMS has been validated as a tool for performance assessment¹⁶ and as an intelligent tutoring system for technical skill acquisition during a simulated neurosurgical task, with AI-tutored students significantly outperforming those tutored by an expert instructor.¹⁸ However, the effect of integrating of human expert instruction and intelligent tutoring on surgical technical skill acquisition has not previously been studied. This thesis aims to assess the utility of enhancing human expert instruction with AI performance data for the teaching of technical skills in surgical simulation. The findings reported here can inform surgical residency programs on curriculum design and the most effective strategies for incorporating performance assessment algorithms and intelligent tutoring systems into the training paradigm.

BACKGROUND

History of Surgical Education

In 1890, after being made the first chief of surgery at Johns Hopkins Hospital, Dr. William Stewart Halsted established the first surgical residency program in the United States.⁸ Halsted's training paradigm is known as the "see one, do one, teach one" model and involves trainees undergoing rapid evolutions from observers to participants to teachers.⁹ After observing an operation, residents are expected to be capable of performing and then teaching said procedure.⁸ As a result of its success, this framework was soon adopted throughout the United States and other countries, including Canada. Inspired by Halsted's model, Dr. William Edward Gallie implemented the first surgical training program in Canada at the Toronto General Hospital in 1931.¹⁹ Residency programs transformed surgical education globally, providing trainees with opportunities for apprenticeship and experiential learning.⁸

Nevertheless, modern advancements in surgical education research are bringing to light the limitations of this traditional approach to resident training. In the absence of standardized, structured surgical curricula,¹⁰⁻¹³ residents are forced to rely on opportunistic learning, wherein their exposure to surgical procedures and techniques is dictated by case availability.²⁰ As such, they may not be able to observe and later perform certain less common, higher-risk procedures if the opportunities do not arise.²¹ Considering that the attending surgeon's primary focus in the operating room (OR) is the care and safety of the patient, meeting the trainee's learning objectives may be relegated to a secondary role.¹¹ During a real operation, temporal and logistical constraints also play a role in hindering the OR's viability as an educational setting.¹¹ The operation cannot be unnecessarily prolonged and steps cannot be repeated for demonstration purposes.¹¹ Furthermore, resident working hour restrictions and an increased focus on non-

operative duties like research limit the amount of time dedicated to gaining hands-on experience.^{20,21} Finally, with minimal opportunities to hone their technical skills outside the OR, residents are constrained to practicing on patients, which presents hazards to patient safety.⁸ This raises important ethical concerns regarding OR-based teaching.

According to a 2017 report by RiskAnalytica commissioned by the Canadian Patient Safety Institute, a total of 5.7 million medical incidents will occur in Canadian hospitals over the following 30 years, resulting in the deaths of 713000 patients and costing the healthcare system 39 billion dollars.²² Another report from the Canadian Institute for Health Information states that approximately 1 in 17 hospital stays in Canada from 2023 to 2024 resulted in patient harm, with procedural errors accounting for 18% of these cases.³ Surgery is a high-stakes, high-risk branch of medicine with errors having the potential to result in patient mortality and morbidity, reduced quality of life, and increased economic burdens on the healthcare system. There are numerous factors underpinning the success of an operative procedure, including but not limited to bimanual psychomotor skills, anatomical knowledge, situational awareness, communication, leadership, and teamwork.^{1,2} Studies have shown that technical performance is a critical determinant of patient outcomes in surgery.⁵⁻⁷ Neurosurgery, in particular, is a highly specialized surgical discipline that necessitates years of knowledge and skill acquisition to attain competence. Even minor errors during neurosurgical procedures can severely impair functioning. In fact, one study reported that neurosurgeons face the highest proportion of malpractice claims per year of any medical specialty.²³ Moreover, the most prevalent errors in neurosurgery are technical in nature.⁴

The high incidence of surgical errors—especially surgical technical errors—and the resulting patient harm suggest that the traditional Halstedian framework for surgical education may require significant reform. Particularly in the case of more technically complex surgical

specialties like neurosurgery, a shift from apprenticeship-based to competency-based training is warranted. For competency-based training to be effective, surgical curricula must be structured and standardized to ensure a consistent and comprehensive educational experience. Surgical simulation is a feasible solution, providing trainees with opportunities to develop their technical skills outside the OR at no cost to patient safety.^{9,21}

Simulation in Surgical Education

Simulation is the replication of real-life scenarios to produce authentic and immersive settings.²⁴ In the context of surgical training, simulation provides learners with controlled, risk-free environments for technical skill development, allowing them to refine and strengthen their expertise before operating on patients.^{24,25} Surgical simulation can be broadly categorized into two types: organic and inorganic.²⁴ Inorganic simulators are further divided into synthetic and virtual models.²⁶

Organic simulators encompass any simulator that utilizes biological tissues such as cadaveric models, live animals, and *ex vivo* animal tissues.²⁶ While these models replicate anatomical structures and tactile sensations with high fidelity to human operations, they are costly, non-reusable, and pose ethical challenges.²⁶ Synthetic models are typically lower in fidelity and constructed from materials such as plastic, rubber, and latex, making them a cheaper alternative.²⁴ Examples include three-dimensional (3D)-printed models, benchtop simulators, and manikins.^{24,26} These models often prioritize cost-effectiveness and portability at the expense of realism, positioning them as useful tools for training specific skills like hand-eye coordination and dexterity, yet insufficient for capturing the full spectrum of competencies required for operative procedures.²⁶ With the rapid evolution of innovative technologies driving progress in medicine, virtual simulators are continuously being developed and optimized for simulation-based surgical education. Virtual simulators consist mainly of virtual reality (VR) and augmented reality (AR). VR combines visual, auditory, and haptic stimuli to immerse users in a fully computer-generated environment that mimics real-world conditions.²⁷ AR, on the other hand, bridges the virtual and physical worlds by overlaying digital components onto real images.²⁷ AR-based surgical simulators can be used as navigational tools during live surgery²⁸ or for skills

training in conjunction with physical simulators.^{27,29} By contrast, due to their confinement to virtual environments, VR simulators are employed for surgical training. Although generally costly, virtual simulators provide high-fidelity environments and allow for repetitive practice. Moreover, virtual simulators operate on computer systems capable of collecting large amounts of performance data, facilitating the data analysis process.³⁰

Surgical simulators are commonly assessed for their viability as educational tools through validation studies, particularly those evaluating face, content, and construct validity. Face and content validity are subjective evaluation methods involving completion of questionnaires by participants.³¹ Face validity describes a model's realism and how accurately it resembles the task it aims to simulate, whereas content validity relates to its capacity to improve users' skills during training.³¹ Construct validity is an objective measure of a simulator's ability to reliably differentiate between individuals with varying skill levels.³¹ Studies measuring learning and skill transfer from simulated environments to real-world surgical settings can also be used to determine a simulator's utility for surgical training. For example, one study investigated the transferability of simulated laparoscopic surgical skills to the OR through a prospective, blinded controlled trial.³² One group of residents underwent laparoscopic skills training on a VR simulator, while the other group received no training.³² All participants proceeded to perform a real laparoscopic cholecystectomy and were evaluated by an experienced surgeon.³² The group that received training performed significantly better than the group that received no training, indicating that laparoscopic skills acquired during VR simulation transferred to the OR.³²

Simulation-based training is grounded in several learning theories. Deliberate practice refers to a systematic approach to skill acquisition involving repetitive performance, targeted feedback by an expert supervisor, and well-defined learning objectives.³³ This approach aligns

well with simulation-based surgical training, where trainees can repeatedly practice procedures under expert supervision until competence is achieved. Furthermore, practicing on simulators enables trainees to develop their skills through experiential learning without exposing patients to the safety risks involved in OR-based training.³⁴

When designing simulation-based surgical training curricula, it is important to consider affective and cognitive demands to foster an environment conducive to learning while minimizing external distractions. Although human thinking and learning were traditionally considered purely cognitive activities, recent research has highlighted the previously-overlooked relationship between emotions and learning-related outcomes.³⁵ According to the control-value theory of achievement emotions, learners' emotions can be influenced by their perceived control over a given activity and the value they ascribe to it.³⁶ In turn, these emotions may impact learner interest, engagement, motivation, and overall performance.³⁶ Emotions are classified by valence (pleasantness) and activation (physiological arousal), yielding four distinct emotional categories: positive activating, positive deactivating, negative activating, and negative deactivating (Figure 1).³⁷ As core dimensions of emotion, valence and activation function by either promoting or hindering motivation and action, respectively.³⁸ Thus, positive activating emotions (e.g., pride, curiosity) are positively correlated with achievement, while negative deactivating emotions (e.g., disappointment, boredom) are generally considered detrimental to learning outcomes.³⁸ Positive deactivating emotions (e.g., relaxation, relief) and negative activating emotions (e.g., frustration, fear) often result in more variable behavioural responses that may support or impede learning, depending on the intensity of these emotions and the surrounding learning context.³⁹

Beyond emotional factors, the cognitive effort required to process new information also plays an important role in determining learning outcomes. Cognitive load theory postulates that

learners have limited cognitive resources, which must therefore be directed efficiently to support learning.⁴⁰ Instruction influences how information is processed by generating three forms of cognitive load within learners: intrinsic, extraneous, and germane.⁴¹ Intrinsic load refers to the inherent difficulty of the learning material and is influenced by the learner's prior knowledge, the number of instructional elements, and the complexity of their interactions.⁴¹ In contrast, germane load involves the mental processes required to organize new information into schemas.⁴¹ Put simply, intrinsic load pertains to the demands of performing a task, while germane load reflects the mental effort invested in learning from a task.⁴¹ Finally, extraneous load arises from factors external to the learning material, such as inadequate guidance and poorly-designed educational settings.⁴¹

To better understand the affective and cognitive responses elicited in learners during training, validated instruments have been developed to measure both dimensions. The Medical Emotions Scale (MES) is a tool designed to measure the intensity of emotions experienced by learners in medical education contexts.³⁷ The MES assesses seven positive activating, two positive deactivating, seven negative activating, four negative deactivating, and two non-valence emotions.³⁷ The Cognitive Load Index (CLI) is a 10-item instrument developed to quantitatively measure intrinsic, extraneous, and germane cognitive load.⁴² In this study, all participants filled out MES and CLI questionnaires to assess the affective and cognitive demands associated with each educational intervention.

Simulation in Neurosurgical Education

Given the nervous system's fragility and central role in bodily functioning, neurosurgery is among the most complex, high-stakes surgical disciplines, where errors have the potential to result in significant patient morbidity and mortality.⁴³ Thus, neurosurgery represents an excellent surgical subspecialty for the assessment and implementation of simulation-based surgical training. To support the deliberate practice of neurosurgical technical skills in controlled, risk-free environments, our team has contributed to the development and validation of both VR and *ex vivo* neurosurgical simulators.^{15,25,44,45}

Developed by a team of experts from the National Research Council Canada, the NeuroVR (CAE Healthcare, Montreal, Canada) is an immersive, high-fidelity VR surgical simulation platform that offers a range of interactive neurosurgical and spinal procedure scenarios.¹⁵ Realistic anatomical structures, haptic sensations (e.g., tissue resistance), and physical and physiological responses (e.g., heartbeat, bleeding) are all represented with accuracy to real human surgical conditions.¹⁵ The simulation is viewed through a neurosurgical microscope, which enables 3D visualization of the simulated tissues.¹⁵ The simulator is equipped with bipolar forceps and an ultrasonic aspirator, each attached to a haptic handle.¹⁵ These instruments are also represented virtually within the simulation and reflect the user's physical manipulations on screen in real time.¹⁵ The ultrasonic aspirator removes pathological tissue and suctions blood, while the bipolar forceps are used to retract tissue for improved visualization and to cauterize bleeding points.¹⁵ The NeuroVR has demonstrated face, content, and construct validity, establishing its viability as an educational tool.²⁵

Two of the scenarios offered by the NeuroVR allow users to practice subpial resection, a critical neurosurgical technique for the management of brain tumours and epilepsy. During this

procedure, the bipolar forceps are used to lift the pia mater while the ultrasonic aspirator resects the underlying pathological tissue.⁴⁶ The NeuroVR features a simpler “practice” subpial resection scenario for developing the core competencies involved in performing this technique (Figure 2) and a more complex “realistic” scenario used to evaluate how well trainees apply, retain, and transfer acquired skills (Figure 3).⁴⁴ In both scenarios, users are tasked with removing a glioma-like human primary brain tumour while minimizing bleeding and damage to healthy tissue.

Although VR simulators like the NeuroVR can achieve high degrees of realism, they are unable to fully replicate the experience of handling biological tissue, which remains essential for developing surgical competence. As such, an *ex vivo* calf brain model for subpial resection—intended to function as an intermediary between the virtual operating environment of the NeuroVR and the real-world OR—was developed.⁴⁵ The calf brain was chosen due to its affordability, availability, and anatomical similarity to the human pediatric brain.⁴⁵ This model exhibited face and content validity according to evaluations by neurosurgical residents, fellows, and expert consultants.⁴⁵

Performance Assessment in Surgery: Foundations

Growing critiques of traditional apprenticeship-based surgical training frameworks have catalyzed the adoption of competency-based medical education (CBME) principles in residency programs worldwide. Performance assessment methodologies are key components of CBME, facilitating the process of evaluating trainee competence and ensuring readiness for independent clinical practice. In an effort to bridge the gap between CBME and clinical practice, the Royal College of Physicians and Surgeons of Canada integrated entrustable professional activities (EPAs) into residency programs.^{47,48} EPAs are defined as tasks or responsibilities residents are permitted to perform autonomously only once they have demonstrated sufficient competence, as assessed by attending surgeons.⁴⁷⁻⁵⁰ While EPAs establish benchmarks for a resident's abilities at each stage of training,⁴⁷ they lack the structure offered by criteria-based performance assessment rating scales.

Multiple rating scales for evaluating surgical technical skills exist, but the Objective Structured Assessment of Technical Skills (OSATS) is often recognized as the gold standard due to its widespread use in surgical training.⁵¹ The OSATS comprises seven categories, each rated by the evaluator on a 5-point Likert scale: respect for tissue, time and motion, instrument handling, knowledge of instruments, use of assistants, flow of operation and forward planning, and knowledge of specific procedure.⁵² This scale is relevant and adaptable to any surgical specialty.

Unfortunately, the use of the OSATS and similar rating scales involves limitations. Without quantitative measures for metrics such as amount of force applied, volume of blood loss, and width of resection margins, the OSATS is based entirely on qualitative visual ratings. As with any assessment method reliant on human observation, this renders it vulnerable to evaluator

bias and subjectivity.¹⁴ Moreover, studies have reported significant institutional and regional variations in surgical practice, citing differences in surgical protocols (e.g., perioperative blood transfusion use)⁵³ and operative factors (e.g., negative margin rates and lymph node yields).⁵⁴ Such variations may further bias the scoring process and contribute to inconsistencies in OSATS ratings, diminishing the generalizability of results across institutions and regions. These limitations emphasize the need for more objective, data-driven systems to evaluate surgical performance by breaking it down into multiple quantifiable metrics.

Performance Assessment in Surgery: Innovations

Artificial intelligence (AI) is a term that encompasses all forms of computer simulation of human intelligence through learning, reasoning, problem-solving, predicting, decision-making, and more.⁵⁵ Machine learning (ML) is a branch of AI that refers to the process of computers learning from and recognizing patterns within the input data.⁵⁶ Three major types of ML include supervised, unsupervised, and reinforcement learning.⁵⁷ Supervised ML algorithms use labelled data to map the input to the output, allowing them to make predictions about unseen data.⁵⁷ Supervised ML has established its utility in various facets of surgery including preoperative planning and decision-making,^{58,59} instrument detection and tracking,^{60,61} postoperative outcome prediction,⁶²⁻⁶⁴ technical skill assessment,^{16,65-67} and training.^{16-18,68}

With the recent shift of surgical training paradigms from apprenticeship-based to competency-based, innovations in surgical performance assessment are emerging. Current assessment methods rely on qualitative observation by human experts rather than more quantitative measures. AI's capacity to process and analyze large, complex datasets render it suitable for quantifying surgical technical skills. Studies have employed various supervised machine learning algorithms for performance assessment including artificial neural network, naïve Bayes, linear discriminant analysis, logistic regression, support vector machine, k-nearest neighbour, and random forest models.^{16,65-67} A key challenge in applying these AI technologies lies in determining how performance can be quantified. The primary approach used to achieve this is through the extraction of specific metrics.⁶⁹ In surgical training, metrics function as tools for measurement, providing standards of reference by which performance, efficiency, and progress can be assessed. Many operative procedures necessitate multiple bimanual psychomotor skills. As such, metrics provide expert benchmarks to measure individual skills and subsequently

allow for surgical performance to be inferred and extrapolated.²⁵ As such, our team developed a surgical skill assessment system known as the Intelligent Continuous Expertise Monitoring System (ICEMS).¹⁶

The ICEMS is a deep learning application that quantitatively assesses surgical performance at 0.2-second intervals and calculates an expertise score on a scale of -1.00 (novice) to 1.00 (expert).¹⁶ To develop this system, performance data on a NeuroVR subpial brain tumour resection task was collected from 12 medical students and 14 neurosurgeons classified as novices and experts, respectively.¹⁶ Sixteen performance metrics comprising instrument handling (e.g., acceleration, velocity, and force application for each instrument) and risk assessment (e.g., healthy tissue injury risk) were extracted from the raw data.¹⁶ This metric data was then inputted into a long short-term memory (LSTM) network to train it to learn surgical expertise from the difference between novice and expert performance.¹⁶ The system has demonstrated predictive validity for its ability to accurately differentiate between neurosurgeons, senior neurosurgery residents (post-graduate year [PGY] 4 to 6), junior neurosurgery residents (PGY 1 to 3), and medical students based on their surgical performance.¹⁶ The ICEMS is granular, objective, and has demonstrated aptitude for scoring hundreds of virtual reality surgical simulation tasks in randomized controlled trials (RCTs).^{18,68}

Intelligent Tutoring Systems

Intelligent tutoring systems are autonomous pedagogical agents that employ AI algorithms to provide guidance and instruction to learners.⁷⁰⁻⁷² Designed to simulate the role of a human educator, intelligent tutoring systems evaluate learner progress and adapt their content to individual learners' knowledge and skillsets.⁷² Numerous intelligent tutoring systems have been developed for improving clinical and surgical competencies including things like diagnostic reasoning, emergency management, and procedural skills.⁷³⁻⁷⁶ In the context of surgical technical skills training, intelligent tutoring systems can quantitatively assess trainee performance, identify errors according to AI-derived metrics, and provide action-oriented feedback accordingly.^{16,17}

In 2020, the Virtual Operative Assistant (VOA), an intelligent tutoring system for teaching bimanual psychomotor surgical skills in surgical simulation, was developed by our group.¹⁷ The VOA uses NeuroVR performance data to calculate learner scores on AI-derived metrics and leverages a linear support vector machine to classify their performance as either novice or expert.¹⁷ Following the completion of a simulation task, text-based, auditory, and video-based instructions are provided to the trainee based on their metric scores.¹⁷ The VOA assesses and trains four AI-selected metrics: bipolar forceps force application, rate of bleeding, instrument tip separation distance, and bipolar forceps acceleration.¹⁷ The system adheres to the mastery learning model,⁷⁷ ensuring learners reach competency in the safety metrics before moving onto instrument movement metrics.¹⁷ In an RCT, the VOA yielded superior learning outcomes for the post-hoc teaching of AI-selected metrics compared with human expert instruction.⁶⁸ However, the VOA lacks the capacity to continuously evaluate operative performance and provide real-time instruction. As such, this intelligent tutor fails to replicate the

learner-educator dynamics in the human operating room, wherein the attending surgeon monitors the resident's performance in real time and intervenes to correct errors as they arise.¹⁶

To address these shortcomings, a new intelligent tutoring system known as the Intelligent Continuous Expertise Monitoring System (ICEMS) was developed for training surgical technical skills.¹⁶ Compatible with any VR surgical simulator, the ICEMS prioritizes deliberate practice in a controlled simulation environment. The ICEMS uses a deep learning algorithm known as an LSTM network to continuously assess intraoperative performance at 0.2-second intervals.¹⁶ Based on this performance evaluation, the system proceeds to provide real-time verbal instruction on five AI-selected metrics.¹⁶ These metrics are divided into two categories: risk assessment—comprising healthy tissue injury risk and bleeding risk—and coaching—consisting of instrument tip separation distance, bipolar forceps force application, and ultrasonic aspirator force application.¹⁶ Expert benchmarks were established for each of the five metrics during the ICEMS's development, when NeuroVR performance data from 14 experts was inputted into an LSTM network.¹⁶ When the system detects an error in one of the metrics—defined as the participant's score deviating by more than standard deviation from the expert benchmark for over 1 second—it delivers real-time verbal feedback to prompt error correction.¹⁶

Explainability and transparency are important features of AI that allow stakeholders to trust the learning potential of these technologies.⁷⁸⁻⁸⁰ However, the processes underlying the decision-making of AI systems are often obscure to humans, leading to what is known as the black box problem.⁸¹ During the ICEMS's development, specific metrics which distinguish expert from novice performance were identified and an LSTM network was trained using labelled metric data. The system's algorithm structure uses metric scores to calculate a composite expertise score. These attributes of the ICEMS underscore the system's explainability and

transparency, allowing us to understand both how it was built and the logic behind its decision-making.

A previous RCT demonstrated that trainees tutored by the ICEMS demonstrated superior learning outcomes compared with those tutored by a human instructor.¹⁸ However, in this study, human educators were not provided with quantitative error data, making their instructions susceptible to subjectivity. This limitation has highlighted the need for further investigation into the integration of quantitative AI error data and personalized expert instruction during surgical simulation training. A subsequent randomized cross-over trial sought to further explore both instructional modalities in separate training sessions.⁸² Two groups received feedback from either the ICEMS tutor or a human expert instructor during the first training session before crossing over to the alternate intervention in a follow-up session.⁸² The surgical performance of trainees who received expert instruction followed by AI instruction improved significantly across both sessions, while trainees who received AI instruction first exhibited significant skill decay during the second session.⁸² These findings suggest that human expert instruction may provide trainees with foundational knowledge and skills, while intelligent tutors serve to refine these skills. However, the possible unintended consequences of AI tutoring during training⁸³ make evident that careful incorporation of intelligent tutoring systems into surgical curricula based on empirical evidence from well-designed studies is essential. A cohort study investigating the effects of intelligent tutoring on learning outcomes in surgical simulation training reported that AI tutoring unintentionally hindered trainee performance on several efficiency metrics.⁸³ The study concluded that human expert input may be imperative when designing AI-enhanced surgical curricula.⁸³

To date, no investigations have been conducted to evaluate the effects of harnessing and combining the strengths of both human and AI tutoring methodologies within a single surgical simulation training session. We sought to assess the pedagogical utility of augmenting human expert instruction with intelligent tutoring systems' quantitative performance assessment capabilities.

THE STUDY OBJECTIVES

To the best of our knowledge, the effect of augmenting human expert instruction with quantitative AI performance data on surgical skill acquisition during simulation training remains unexplored. Therefore, the primary objective of this thesis is to evaluate the effect of AI-augmented personalized expert instruction—where human surgical educators are provided with quantitative ICEMS performance data—compared with AI tutor instruction on trainees’ surgical performance, improvement, and skill transfer during simulation training. The secondary objective is to investigate how these instructional modalities influence trainees’ affective and cognitive responses.

THE STUDY HYPOTHESIS

Our primary hypothesis is that AI-augmented personalized expert instruction will result in superior surgical performance, technical skill acquisition, and skill transfer among trainees compared with AI tutor instruction. This hypothesis is grounded in adult learning theories that emphasize the benefits of personalized learning and contextualization on learning outcomes.⁸⁴⁻⁸⁶ Our secondary hypothesis is that the AI-tutored group will experience stronger negative emotions and increased cognitive demands compared with both instructor-led groups. While research in this area remains limited, one study reported that learning with non-human tutors imposes greater cognitive demands than with human ones.⁸⁷ As suggested in the literature, this may be due to human tutors’ capacity to interpret emotional signals and respond strategically to mediate learner emotions and cognitive load.⁸⁸

THE STUDY

Effect of Artificial Intelligence-Augmented Human Instruction on Surgical Simulation

Performance: A Randomized Controlled Trial

Authors: Bianca Giglio, BSc¹; Abdulmajeed Albeloushi, MD^{1,2}; Ahmad Kh. Alhaj, MD^{1,2}; Mohamed Alhantoobi, MD, MSc^{1,3}; Rothaina Saeedi, MD^{1,2}; Vanja Davidovic, BHSc¹; Abicumaran Uthamacumaran, BSc¹; Recai Yilmaz, MD, PhD^{1,4}; Jason Lapointe, DEC^{1,5}; Neevya Balasubramaniam, DEC^{1,6}; Trisha Tee, MSc^{1,7}; Ali M. Fazlollahi, MSc^{1,6}; José A. Correa, PhD⁸; Rolando F. Del Maestro, MD, PhD¹

Affiliations:

1. 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, Canada H2X 4B3
2. Department of Neurology and Neurosurgery, Montreal Neurological Institute and Hospital, McGill University, 3801 Rue University, Montreal, QC, Canada H3A 2B4
3. Department of Neurosurgery, Hamilton General Hospital, McMaster University Medical Centre, 237 Barton St E., Hamilton, ON, Canada L8L 2X2
4. Children's National Medical Center, Division of Neurosurgery and Pediatrics, 111 Michigan Ave NW, Washington, D.C. 20010, United States of America
5. Faculty of Science and Engineering, Université Laval, Pavillon Alexandre-Vachon, 1045 Av. de la Médecine, Quebec City, QC, Canada G1V 0A6

6. Faculty of Medicine and Health Sciences, McGill University, 3605 Rue de la Montagne, Montreal, QC, Canada H3G 2M1
7. Florida International University Herbert Wertheim College of Medicine, 11200 SW 8th St AHC2, Miami, FL 33199, United States of America
8. Department of Mathematics and Statistics, McGill University, 805 Sherbrooke St W, Montreal, QC, Canada H3A 1Y2

The preceding work has been augmented with additional information and materials to reflect the requirements for thesis submission for a Master of Science.

This manuscript was submitted for review to *JAMA Surgery* on April 7, 2025.

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.

ABSTRACT

Importance: To understand how to optimize the Intelligent Continuous Expertise Monitoring System, an artificial intelligence tutoring system, for surgical training.

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

Design: Single-blinded randomized controlled trial. Cross-sectional data collected from March to September 2024 and per-protocol analysis conducted in March 2025.

Setting: McGill University's Neurosurgical Simulation and Artificial Intelligence Learning Centre.

Participants: Volunteer sample of Quebec medical students in preparatory, first, or second year without prior use of NeuroVR.

Intervention: During simulated surgical procedures, trainees received one of three feedback methods. Group 1 received only intelligent tutor instruction (control). Two intervention arms included group 2, receiving expert feedback in identical words to the intelligent tutor, and group 3, receiving artificial intelligence data-informed personalized expert feedback.

Main Outcomes and Measures: 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: Final analysis included 87 medical students (46 [53%] women, 40 [46%] men, 1 [1%] unspecified; age mean [SD], 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 to 0.43]; $P = 0.01$) and the realistic task (mean difference, 0.20 [95% CI, 0.06 to 0.34]; $P = 0.02$). Group 3 also achieved significantly better scores than the other two groups in certain metrics, such as bleeding and injury risk. Emotions and cognitive load demonstrated significant differences.

Conclusions and Relevance: In this randomized controlled trial, artificial intelligence-augmented 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: Registered on ClinicalTrials.gov on February 16, 2024 (name: Efficiency of Verbal Intelligent Tutor Instruction in Surgical Simulation; number: NCT06273579; URL: clinicaltrials.gov/study/NCT06273579).

INTRODUCTION

Though expert surgical technical skill is linked with improved patient outcomes, training novices to master these skills remains challenging.⁵⁻⁷ Current surgical teaching models lack standardization¹⁰⁻¹³ and rely on qualitative performance assessments by human experts rather than quantitative performance data.¹⁴ 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.^{16-18,68,83} The goal of these technologies is to create standardized AI-enhanced surgical curricula to improve trainee bimanual skills, thereby achieving better patient outcomes.^{70-72,89-91}

In a randomized controlled 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.^{17,68} 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 employing a multi-algorithm approach to assess bimanual surgical skills at 0.2-second intervals and provide continuous, action-oriented verbal feedback.¹⁶ Built based on quantifiable, AI-derived metrics that enable continuous performance scoring from -1.00 (novice) to 1.00 (expert),¹⁶ the ICEMS demonstrates explainability and transparency critical to educator and learner engagement.⁷⁸⁻⁸⁰ The ICEMS can be integrated into any VR surgical simulator including the NeuroVR. This system has been validated for its ability to accurately differentiate surgical expertise levels, track skill acquisition throughout a neurosurgical training program,¹⁶ and serve as a pedagogical tool for risk assessment, coaching, and error detection.¹⁸

Another RCT demonstrated that ICEMS feedback yielded enhanced learning outcomes compared with expert feedback during a simulated surgical task.¹⁸ Instructors in this study were blinded to the ICEMS error data and depended on qualitative observation rather than on 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.⁸³ A randomized cross-over 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.⁸² 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 resulting in lower negative emotions and cognitive load.^{87,88}

METHODS

This parallel-design single-blinded three-arm RCT was approved by the McGill University Health Centre Research Ethics Board, Neurosciences-Psychiatry. The study was registered at ClinicalTrials.gov on February 16, 2024 (NCT06273579). This report follows the

Consolidated Standards of Reporting Trials involving Artificial Intelligence (CONSORT-AI)⁹² and the Machine Learning to Assess Surgical Expertise (MLASE) checklist.⁹³

Participants

Participants were recruited between March and September 2024 for a single 90-minute surgical simulation session without follow-up at McGill University's Neurosurgical Simulation and Artificial Intelligence Learning Centre (Table 2). A sample size calculation for a repeated measures analysis of variance (ANOVA) with a between-subjects factor was conducted using G*Power version 3.1.⁹⁴ 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 0.5¹⁸ yielded a total of 87 participants, with 29 in each of three groups. Volunteer sampling was used to attain the desired sample size. Recruitment information was disseminated via student groups, social media, and word of mouth. The inclusion criteria consisted of enrollment in preparatory, first, or second year at one of four Quebec medical schools. The exclusion criterion consisted of previous use of the NeuroVR.

Randomization

Students were stratified based on their year in medical school and block randomized to one of three intervention arms with an allocation ratio of 1:1:1 using random number sequences generated by Random.org.⁹⁵ Participant recruitment flowchart is outlined in Figure 4.

Simulation Session

All tasks were performed on the NeuroVR (CAE Healthcare), a validated surgical simulator that simulates a subpial brain tumour resection in a 3D VR environment.^{15,25} The

session consisted of two scenarios: a practice subpial resection task (Figure 2) and a realistic subpial brain tumour resection (Figure 3).⁴⁴ 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.^{90,96} Participants performed six 5-minute practice tasks to assess their learning and technical skill acquisition, 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.

Study Procedure

Prior to the simulation session, participants read and signed an informed consent form. They then completed a pre-trial questionnaire recording demographic information and self-reported baseline emotions using the Medical Emotions Scale (MES) (Figure 1).³⁷ Following the performance of six practice tasks, participants completed a peri-trial questionnaire to assess the strength of emotions elicited during training using the MES. After the realistic task, students filled out a post-trial questionnaire that recorded emotions after training using the MES and self-reported cognitive load using the Cognitive Load Index (CLI).⁴² Random allocation sequence generation, participant enrollment, and assignment to interventions were conducted by the study coordinator. Participants and instructors were blinded to group assignments and study outcomes. Study procedure is outlined in Figure 5.

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 utilization, and ultrasonic aspirator force utilization.^{16,18} An error is defined as a difference of more than one standard deviation from the expert benchmark for over 1 second.^{16,18}

All participants began by completing a practice resection 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 resection without feedback, serving as a summative assessment. Finally, they completed the realistic brain tumour resection task without feedback to assess skill transfer to a more complex scenario.

Group 1 (Control): AI Tutor Instruction

The control group received real-time verbal feedback delivered by the ICEMS when a metric error was detected. Table 1 contains the verbatim instructions delivered by the AI tutor for each metric.

Group 2 (Experimental): Expert Instruction

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

Group 3 (Experimental): 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 coloured 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, 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, scored by the ICEMS.

The secondary outcome was trainees' self-reported affective-cognitive responses.⁹⁷ These include 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 MES³⁷ on 7-point Likert scales and the CLI⁴² on 5-point Likert scales.

Statistical Analysis

Within-group differences from baseline in practice task scores and MES scores were compared using a mixed-model one-way ANOVA. Between-group comparisons at each timepoint for practice task scores and MES scores were conducted using a mixed-model two-way analysis of covariance (ANCOVA), with baseline performance as a covariate. Realistic task scores and CLI scores were compared using a one-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. Observations with studentized residuals exceeding an absolute value of 3 were deemed outliers and removed. Means from Likert items were computed prior to analysis of emotions and cognitive load. All statistical hypothesis tests were two-sided and performed at a significance level of 0.05. Statistical analyses and score predictions were performed using R version 4.4.3.⁹⁸ Data analysis was conducted in March 2025.

RESULTS

Eighty-eight medical students enrolled in one of four medical schools across the province of Quebec 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 one participant in group 1 was excluded from analysis due to technical issues that occurred during the simulation session. The ICEMS assessed data from 87 participants (46 [53%] women, 40 [46%] men, 1 [1%] unspecified; age mean [SD], 22.7 [4.0] years; 25 [29%] preparatory year, 42 [48%] first year, 20 [23%] second year), including 522 practice resections and 87 realistic resections (Table 2).

Performance Across Practice Subpial Resection Trials

The mean composite expertise scores from the baseline assessment (trial 1) 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 to 0.43]; $P = 0.01$) and trial 5 (mean difference, 0.26 [95% CI,

0.09 to 0.43]; $P = 0.01$). 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 to 0.41]; $P = 0.02$), 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 6A).

Scores for individual ICEMS metrics used for competency training were also assessed. AI-augmented personalized expert instruction, the intervention delivered to group 3 participants, largely resulted in metric scores closer to expert benchmarks than the other two 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 6B-C). For aspirator force, group 3 significantly outperformed group 2 during trials 4 and 6 and group 1 during trials 3 and 4 (Figure 6D). 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

The mean composite expertise scores from the realistic task 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 to 0.34]; $P = 0.02$) and group 2 (mean difference, 0.18 [95% CI, 0.03 to 0.32]; $P = 0.049$), underscoring better skill transfer (Figure 7A). Group 3 also outperformed group 1 on both risk

assessment metrics, achieving significantly lower bleeding risk (mean difference, 0.11 [95% CI, 0.05 to 0.16]; $P < 0.001$) and injury risk (mean difference, 0.03 [95% CI, 0.01 to 0.04]; $P = 0.009$) (Figure 7B-C). No statistically significant differences between groups were found for aspirator force, bipolar force, and tip separation distance (Figure 7D-F).

Emotions and Cognitive Load

Group 3 reported significantly greater levels of negative activating emotions (e.g., frustration) than group 1 after the trial (mean difference, 0.42 [95% CI, 0.01 to 0.82]; $P = 0.04$). 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 (e.g., relief) was group 1 following the practice trials (Figure 8A). All three groups experienced statistically significant increases in both negative activating and negative deactivating emotions (e.g., disappointment) after the practice trials, although these differences only persisted for group 3 after the realistic task (Figure 8B-C). Post-hoc 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 to 0.94]; $P = 0.02$). No differences in extraneous or germane cognitive load were found (Figure 8D).

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.^{16,17} 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.^{16,17}

Consistent with our hypothesis, the findings of this RCT demonstrate that AI-augmented personalized expert instruction yields improved surgical performance and skill transfer 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.⁷⁸⁻⁸⁰ The ICEMS was developed using a long short-term memory network trained on performance data from experts (neurosurgeons) and novices (medical students).¹⁶ Its algorithm primarily uses risk assessment metrics to calculate composite scores with a secondary focus on coaching metrics.¹⁶ 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. 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.⁹⁹ Other human factors, such as non-verbal cues and adaptive communication, may also influence student learning outcomes, but this requires further investigation.^{100,101} 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. Future studies should evaluate the mental fatigue of participants via self-report questionnaires.

A previous RCT conducted at our lab involving ICEMS tutoring was unable to demonstrate statistically significant between-group differences during the realistic task.¹⁸ 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. Studies assessing whether this finding holds true for skill transfer from VR simulation to more realistic OR environments using *ex vivo* animal models are in development.^{45,102-104}

Unlike other RCTs conducted at our centre,^{18,68} 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 three groups experienced significant increases in negative deactivating emotions, such as disappointment, during the trial. Post-trial levels of negative activating emotions such as fear were significantly greater in the personalized expert instruction group compared with the AI instruction group, highlighting their potential role in supporting learning in this context.^{36,39} The personalized expert instruction group had significantly higher intrinsic cognitive load than the AI instruction group, indicating increased mental effort required to understand the complexity of the variable instructions.^{41,105} Research focused on negative activating emotions and cognitive load may help optimize learning with intelligent tutors.

Consistent with tenets of learning theory,⁸⁴⁻⁸⁶ 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 our findings. This RCT and our cross-over study⁸² 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.⁹⁷ Our study involved small cohorts of medical students with minimal surgical experience from only four institutions, limiting the generalizability of the results to other groups of learners. However, understanding how medical students can achieve AI-derived benchmarks of more advanced learners has offered insights into the optimization of surgical intelligent tutoring systems.^{16-18,68,83} Finally, the applicability of these results to human surgical environments was beyond the scope of this research project but requires further investigation.

CONCLUSION

In this randomized controlled trial, 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-centred surgical curricula.

THESIS SUMMARY

Discussion

To determine the effect of AI-augmented personalized expert instruction versus intelligent tutoring on surgical skill acquisition during simulation training, a parallel-design single-blinded randomized controlled trial was conducted with 88 participants from four Quebec medical schools. Participants performed subpial resection tasks on the NeuroVR simulator and received one of three educational interventions based on their group assignment. Participants' simulated surgical performance, technical skill acquisition, and skill transfer, as well as their emotions and cognitive load, were measured.

In this study, consistent with our hypothesis, AI-augmented personalized expert instruction resulted in superior surgical performance, technical skill acquisition, skill transfer compared with AI tutor instruction and expert instruction. The personalized expert instruction group achieved significantly higher composite expertise scores than the AI tutor instruction group in trials 4 and 5 and the expert instruction group in trial 5 (Figure 6A). These differences did not remain significant in the summative assessment (trial 6), which may reflect fatigue experienced by learners, although this requires further investigation. AI-augmented personalized expert instruction was particularly effective at improving safety outcomes, yielding lower bleeding risk and healthy tissue injury risk relative to the other two groups (Figure 6B-C). The only metric that did not demonstrate any statistically significant between-group differences is bipolar force application (Figure 6E). This may be due to the calculation method for this metric, which aggregates both high and low force values, potentially obscuring differences at either extreme. Moreover, while the scores of all three groups significantly improved from baseline across multiple trials and metrics, the personalized expert instruction group exhibited the most

consistent improvement (Figure 6A-F). The personalized expert instruction group's significant outperformance of both other groups during the realistic task highlights that this instructional methodology resulted in not only effective learning but also superior skill transfer to a more realistic scenario (Figure 7A). These results are in accordance with adult learning theories, which postulate that personalized and contextualized learning are associated with improved learning outcomes.⁸⁴⁻⁸⁶ Whereas the AI tutor instruction and expert instruction groups received general metric-specific commands in response to errors (Table 1), the personalized expert instruction group was provided with more detailed explanations and tailored feedback aligned with each learner's manipulations.

In a post-trial questionnaire, participants were asked to indicate their preferred instructional modality for learning surgical technical skills: independently without feedback, with feedback from a human instructor, with feedback from an intelligent tutoring system, or with feedback from both a human instructor and an intelligent tutoring system. Among 87 respondents, 70 (80.5%) expressed a preference for receiving feedback from both sources. These results reinforce the notion that combining intelligent tutoring and human expert instruction may result in optimal learning outcomes.

Both the AI tutor instruction and expert instruction groups received the same instructional commands, with the sole difference being the presence of a human instructor in the latter. Nonetheless, the expert instruction group significantly outperformed the AI tutor instruction group during practice trial 5 (Figure 6A). Potential explanations for this outcome include increased engagement associated with human-led instruction,⁹⁹ lack of trust in AI systems,¹⁰⁶ and the influence of human factors such as adaptive communication and non-verbal cues.^{100,101} This should be investigated further in future studies.

Contrary to our hypothesis, intelligent tutoring did not result in higher levels of negative emotions and cognitive load among participants compared with human expert tutoring. In fact, the AI-augmented personalized expert instruction group reported significantly stronger negative activating emotions after the trial than the AI tutor instruction group—a difference largely driven by elevated levels of fear (Figure 8B). The personalized expert instruction group also experienced significantly higher levels of intrinsic cognitive load than the AI tutor instruction group (Figure 8D). This suggests that the former perceived the learning material as more complex,^{41,105} potentially attributable to the human educator’s ability to dynamically modify instructional commands and provide contextual explanations of the operative procedure. Students in their preparatory, first, or second year of medical school who participated in this study were relatively inexperienced and had minimal surgical experience. According to cognitive load theory, experienced learners expend fewer cognitive resources than novices because their stronger knowledge base reduces the amount of new information they need to process.¹⁰⁷ Consequently, the significant differences in intrinsic load observed in this study may not have emerged with a group of more experienced participants.

In a previous RCT conducted at our centre, expert instructors were not provided with AI performance data and relied on qualitative observation when teaching surgical technical skills during simulation training.¹⁸ As a result, the AI-tutored group significantly outperformed the expert-tutored group.¹⁸ In contrast, findings from the present study indicate that the personalized expert instruction group—where educators tailored their feedback according to real-time AI error data—significantly outperformed the AI-tutored group. This reversal underscores the value of quantitative performance assessments in surgical training. Residency programs stand to benefit from implementing AI-driven quantitative performance assessment methods, which may

streamline the development of surgical technical skills and support trainees on their path to mastery.

Limitations and Future Directions

It is important to acknowledge that this study focuses on only one aspect of surgical competence: technical expertise. In practice, surgical competence spans multiple other competencies including leadership, communication, and clinical decision-making. Future studies should incorporate non-technical skills training to better approximate the many dimensions involved in real human surgical procedures. In addition, although the NeuroVR is a high-fidelity simulator that has shown face, content, and construct validity,²⁵ virtual reality surgical simulators cannot fully emulate the experience of manipulating biological tissues.⁹⁷ Other types of simulators like *ex vivo* animal models⁴⁵ can act as intermediaries to mediate skill transfer from VR to real-world clinical environments.

The use of voluntary sampling introduces potential bias: given that participants were self-selected, highly motivated, and expressed an interest in surgical disciplines, the sample obtained may not be representative of the target population. Furthermore, participants were students in their preparatory, first, or second year of medical school from institutions in Quebec, with minimal surgical experience. This limits the generalizability of these results to more advanced learners (e.g., residents) or to students from institutions in other provinces or countries. Despite these limitations, the rationale behind the inclusion of junior medical students is that they exhibit a steeper learning trajectory than intermediate or expert performers, making it easier to detect trends in their learning curves. Nevertheless, the overarching aim of this study and similar investigations is to inform the implementation of structured surgical simulation curricula within residency programs. Studies involving neurosurgical residents should be conducted to determine whether the educational methodologies employed in the present study effectively support learning at the residency level. However, with the limited numbers of available residents, it may

not be possible to achieve sufficient power to detect statistically significant differences between groups unless large numbers of training centres have access to both the simulator and the intelligent tutoring system used in this study. Many study participants also reported that English was not their first language. With 54% of participants enrolled in a French-language institution, one way to address this limitation would be to offer the option of completing the trial in French.

Future research should focus on optimizing the integration of quantitative AI error data and personalized expert instruction in the training of surgical procedures. All instructions provided by experts during the present study were recorded. This will enable the incorporation of a library of personalized instructions within the ICEMS to more closely replicate the adaptive, individualized nature of one-on-one human expert instruction. These personalized instructions can be assessed in randomized controlled trials to determine their effectiveness in enhancing current training methodologies. To advance toward the wider objective of improving patient safety, incorporating intelligent tutoring systems into *ex vivo* simulation models would facilitate the development of an ICEMS-equipped operating room. Such a system would allow educators to deliver AI-augmented instruction in more realistic operative scenarios and evaluate the applicability of findings from prior simulation-based investigations.

Conclusion

In this randomized controlled trial, AI-augmented personalized expert instruction resulted in improved surgical performance, technical skill acquisition, 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 integration of intelligent tutoring systems in novel student-centred surgical curricula. This study also contributes to the broader goal of creating an intelligent operating room. This optimized educational environment—combining the knowledge of expert surgical educators and the quantitative error detection capabilities of intelligent tutoring systems—will enable intraoperative learner assessment and training, ultimately minimizing errors during human surgical procedures and improving patient outcomes.

References

1. Agha RA, Fowler AJ, Sevdalis N. The role of non-technical skills in surgery. *Ann Med Surg.* 2015;4(4):422-427. doi:10.1016/j.amsu.2015.10.006
2. Collins JW, Dell'Oglio P, Hung AJ, Brook NR. The Importance of Technical and Non-technical Skills in Robotic Surgery Training. *Eur Urol Focus.* 2018;4(5):674-676. doi:10.1016/j.euf.2018.08.018
3. Canadian Institute for Health Information. *Patient harm in Canadian hospitals? It does happen.* <https://www.cihi.ca/en/patient-harm-in-canadian-hospitals-it-does-happen>
4. Stone S, Bernstein M. Prospective Error Recording in Surgery: An Analysis of 1108 Elective Neurosurgical Cases. *Neurosurgery.* 2007;60(6):1075-1082. doi:10.1227/01.NEU.0000255466.22387.15
5. Rogers SO, 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
6. Fabri PJ, Zayas-Castro JL. Human error, not communication and systems, underlies surgical complications. *Surgery.* 2008;144(4):557-565. doi:10.1016/j.surg.2008.06.011
7. 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
8. Kotsis S, Chung K. Application of the “See One, Do One, Teach One” Concept in Surgical Training. *Plast Reconstr Surg.* 2013;131(5):1194-1201. doi:10.1097/PRS.0b013e318287a0b3

9. Han JJ, Patrick WL. See one—practice—do one—practice—teach one—practice: The importance of practicing outside of the operating room in surgical training. *J Thorac Cardiovasc Surg.* 2019;157(2):671-677. doi:10.1016/j.jtcvs.2018.07.108
10. Madani A, Vassiliou M, 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
11. 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
12. 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.
13. Mirza M, Koenig JF. Teaching in the Operating Room. In: Köhler TS, Schwartz B, eds. *Surgeons as Educators.* Springer; 2017:3-13.
14. 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
15. Delorme S, Laroche D, DiRaddo R, Del Maestro RF. NeuroTouch: A Physics-Based Virtual Simulator for Cranial Microneurosurgery Training. *Oper Neurosurg (Hagerstown).* 2012;71:ons32-ons42. doi:10.1227/NEU.0b013e318249c744
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. doi:10.1038/s41746-022-00596-8

17. Mirchi N, Bissonnette V, Yilmaz R, et al. 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
18. 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. doi:10.1038/s41598-024-65716-8
19. Dossani RH, Shaughnessy J, Kalakoti P, Nanda A. William Edward Gallie (1882–1959): father of the Gallie wiring technique for atlantoaxial arthrodesis. *J Neurosurg*. 2018;128(3):938-941. doi:10.3171/2016.12.JNS161224
20. Roberts NK, Williams RG, Kim MJ, Dunnington GL. The Briefing, Intraoperative Teaching, Debriefing Model for Teaching in the Operating Room. *J Am Coll Surg*. 2009;208(2):299-303. doi:10.1016/j.jamcollsurg.2008.10.024
21. Yanagawa B, Ribeiro R, Naqib F, Fann J, Verma S, Puskas J. See one, simulate many, do one, teach one: cardiac surgical simulation. *Curr Opin Cardiol*. 2019;34(5):571-577. doi:10.1097/HCO.0000000000000659
22. RiskAnalytica. *The Case for Investing in Patient Safety in Canada*. August 2017. <https://www.bcit.ca/files/health/pdf/risk-analytica-2017-investing-in-patient-safety-in-canada.pdf>
23. Jena AB, Seabury S, Lakdawalla D, Chandra A. Malpractice Risk According to Physician Specialty. *N Engl J Med*. 2011;365(7):629-636. doi:10.1056/NEJMsa1012370
24. Tan SSY, Sarker SK. Simulation in surgery: a review. *Scott Med J*. 2011;56(2):104-109. doi:10.1258/smj.2011.011098

25. Alotaibi FE, AlZhrani GA, Mullah MAS, et al. Assessing Bimanual Performance in Brain Tumor Resection With NeuroTouch, a Virtual Reality Simulator. *Oper Neurosurg (Hagerstown)*. 2015;11(1):89-98. doi:10.1227/NEU.0000000000000631
26. Badash I, Burt K, Solorzano CA, Carey JN. Innovations in surgery simulation: a review of past, current and future techniques. *Ann Transl Med*. 2016;4(23):453. doi:10.21037/atm.2016.12.24
27. Cao C, Cerfolio RJ. Virtual or Augmented Reality to Enhance Surgical Education and Surgical Planning. *Thorac Surg Clin*. 2019;29(3):329-337. doi:10.1016/j.thorsurg.2019.03.010
28. Rynio P, Witowski J, Kamiński J, Serafin J, Kazimierczak A, Gutowski P. Holographically-Guided Endovascular Aneurysm Repair. *J Endovasc Ther*. 2019;26(4):544-547. doi:10.1177/1526602819854468
29. Prasad K, Miller A, Sharif K, et al. Augmented-Reality Surgery to Guide Head and Neck Cancer Re-resection: A Feasibility and Accuracy Study. *Ann Surg Oncol*. 2023;30:4994-5000. doi:10.1245/s10434-023-13532-1
30. Bakhaidar M, Alsayegh A, Yilmaz R, et al. Performance in a Simulated Virtual Reality Anterior Cervical Discectomy and Fusion Task: Disc Residual, Rate of Removal, and Efficiency Analyses. *Oper Neurosurg (Hagerstown)*. 2023;25(4):e196-e205. doi:10.1227/ons.0000000000000813
31. Chawla S, Devi S, Calvachi P, Gormley WB, Rueda-Esteban R. Evaluation of simulation models in neurosurgical training according to face, content, and construct validity: a systematic review. *Acta Neurochir (Wien)*. 2022;164:947-966. doi:10.1007/s00701-021-05003-x

32. Prashanth AT, Lakshmikantha N, Lakshman K. The impact of virtual reality training on laparoscopic surgical skills; A prospective blinded controlled trial. *J Clin Invest Surg*. 2021;6(1):37-42. doi:10.25083/2559-5555
33. Rowse PG, Dearani JA. Deliberate Practice and the Emerging Roles of Simulation in Thoracic Surgery. *Thorac Surg Clin*. 2019;29(3):303-309. doi:10.1016/j.thorsurg.2019.03.007
34. Zigmont JJ, Kappus LJ, Sudikoff SN. Theoretical Foundations of Learning Through Simulation. *Semin Perinatol*. 2011;35(2):47-51. doi:10.1053/j.semperi.2011.01.002
35. Artino AR, Holmboe ES, Durning SJ. Can achievement emotions be used to better understand motivation, learning, and performance in medical education? *Med Teach*. 2012;34(3):240-244. doi:10.3109/0142159X.2012.643265
36. Pekrun R. The Control-Value Theory of Achievement Emotions: Assumptions, Corollaries, and Implications for Educational Research and Practice. *Educ Psychol Rev*. 2006;18:315-341. doi:10.1007/s10648-006-9029-9
37. 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
38. 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
39. 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

40. Chandler P, Sweller J. Cognitive Load Theory and the Format of Instruction. *Cogn Instr.* 1991;8(4):293-332. doi:10.1207/s1532690xci0804_2
41. 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
42. 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
43. Suri A, Patra DP, Meena RK. Simulation in neurosurgery: Past, present, and future. *Neurol India.* 2016;64(3):387-395. doi:10.4103/0028-3886.181556
44. Sabbagh AJ, Bajunaid K, 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
45. 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
46. Hebb A, Yang T, Silbergeld DL. The sub-pial resection technique for intrinsic tumor surgery. *Surg Neurol Int.* 2011;2:180. doi:10.4103/2152-7806.90714
47. Cadieux M, Healy M, Petrusa E, et al. Implementation of competence by design in Canadian neurosurgery residency programs. *Med Teach.* 2022;44(4):380-387. doi:10.1080/0142159X.2021.1994937

48. Rabski JE, Saha A, Cusimano MD. Setting standards of performance expected in neurosurgery residency: A study on entrustable professional activities in competency-based medical education. *Am J Surg*. 2021;221(2):388-393.
doi:10.1016/j.amjsurg.2020.12.014
49. Hackney L, O'Neill S, O'Donnell M, Spence R. A scoping review of assessment methods of competence of general surgical trainees. *Surgeon*. 2023;21(1):60-69.
doi:10.1016/j.surge.2022.01.009
50. Wagner JP, Lewis CE, Tillou A, et al. Use of Entrustable Professional Activities in the Assessment of Surgical Resident Competency. *JAMA Surg*. 2018;153(4):335-343.
doi:10.1001/jamasurg.2017.4547
51. Nazari T, Bogomolova K, Ridderbos M, et al. Global versus task-specific postoperative feedback in surgical procedure learning. *Surgery*. 2021;170(1):81-87.
doi:10.1016/j.surg.2020.12.038
52. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skills (OSATS) for surgical residents. *Br J Surg*. 1997;84:273-278. doi:10.1046/j.1365-2168.1997.02502.x
53. Bennett-Guerrero E, Zhao Y, O'Brien SM, et al. Variation in Use of Blood Transfusion in Coronary Artery Bypass Graft Surgery. *JAMA*. 2010;304(14):1568-1575.
doi:10.1001/jama.2010.1406
54. Schoppa DW, Rhoads KF, Ma Y, et al. Measuring Institutional Quality in Head and Neck Surgery Using Hospital-Level Data: Negative Margin Rates and Neck Dissection Yield. *JAMA Otolaryngol Head Neck Surg*. 2017;143(11):1111-1116.
doi:10.1001/jamaoto.2017.1694

55. Xu Y, Liu X, Cao X, et al. Artificial intelligence: A powerful paradigm for scientific research. *Innovation*. 2021;2(4):100179. doi:10.1016/j.xinn.2021.100179
56. Amin A, Cardoso SA, Suyambu J, et al. Future of Artificial Intelligence in Surgery: A Narrative Review. *Cureus*. 2024;16(1):e51631. doi:10.7759/cureus.51631
57. Pruneski JA, Pareek A, Kunze KN, et al. Supervised machine learning and associated algorithms: applications in orthopedic surgery. *Knee Surg Sports Traumatol Arthrosc*. 2023;31:1196-1202. doi:10.1007/s00167-022-07181-2
58. Dundar TT, Yurtsever I, Pehlivanoglu MK, et al. Machine Learning-Based Surgical Planning for Neurosurgery: Artificial Intelligent Approaches to the Cranium. *Front Surg*. 2022;9:863633. doi:10.3389/fsurg.2022.863633
59. Loftus TJ, Tighe PJ, Filiberto AC, et al. Artificial Intelligence and Surgical Decision-Making. *JAMA Surg*. 2020;155(2):148-158. doi:10.1001/jamasurg.2019.4917
60. Kitaguchi D, Lee Y, Hayashi K, et al. Development and Validation of a Model for Laparoscopic Colorectal Surgical Instrument Recognition Using Convolutional Neural Network-Based Instance Segmentation and Videos of Laparoscopic Procedures. *JAMA Netw Open*. 2022;5(8):e2226265. doi:10.1001/jamanetworkopen.2022.26265
61. Pan X, Bi M, Wang H, Ma C, He X. DBH-YOLO: a surgical instrument detection method based on feature separation in laparoscopic surgery. *Int J Comput Assist Radiol Surg*. 2024;19:2215-2225. doi:10.1007/s11548-024-03115-0
62. Wise ES, Hocking KM, Brophy CM. Prediction of in-hospital mortality after ruptured abdominal aortic aneurysm repair using an artificial neural network. *J Vasc Surg*. 2015;62(1):8-15. doi:10.1016/j.jvs.2015.02.038

63. Salati M, Migliorelli L, Moccia S, et al. A Machine Learning Approach for Postoperative Outcome Prediction: Surgical Data Science Application in a Thoracic Surgery Setting. *World J Surg.* 2021;45:1585-1594. doi:10.1007/s00268-020-05948-7
64. Shi W, Giuste FO, Zhu Y, et al. Predicting pediatric patient rehabilitation outcomes after spinal deformity surgery with artificial intelligence. *Commun Med.* 2025;5:1. doi:10.1038/s43856-024-00726-1
65. 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. doi:10.1016/j.combiomed.2024.108809
66. 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. doi:10.1016/j.combiomed.2022.106286
67. Reich A, Mirchi N, Yilmaz R, et al. Artificial Neural Network Approach to Competency-Based Training Using a Virtual Reality Neurosurgical Simulation. *Oper Neurosurg (Hagerstown).* 2022;23(1):31-39. doi:10.1227/ons.0000000000000173
68. Fazlollahi A, 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
69. Bissonnette V, Mirchi N, Ledwos N, Alsidieri G, Winkler-Schwartz A, Del Maestro RF. Artificial Intelligence Distinguishes Surgical Training Levels in a Virtual Reality Spinal Task. *J Bone Joint Surg Am.* 2019;101(23):e127(1-8). doi:10.2106/JBJS.18.01197

70. Balakrishnan S, Dakua SP, El Ansari W, Aboumarzouk O, Al Ansari A. Chapter 14: 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.
71. 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
72. 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
73. Treceño-Fernández D, Calabia-del-Campo J, Bote-Lorenzo ML, Gómez-Sánchez E, de Luis-García R, Alberola-López C. Integration of an intelligent tutoring system in a magnetic resonance simulator for education: Technical feasibility and user experience. *Comput Methods Programs Biomed*. 2020;195:105634. doi:10.1016/j.cmpb.2020.105634
74. Wang M, Sun Z, Jia M, et al. Intelligent virtual case learning system based on real medical records and natural language processing. *BMC Med Inform Decis Mak*. 2022;22:60. doi:10.1186/s12911-022-01797-7
75. Julian D, Smith R. Developing an intelligent tutoring system for robotic-assisted surgery instruction. *Int J Med Robot*. 2019;15(6):e2037. doi:10.1002/rcs.2037

76. Furlan R, Gatti M, Menè R, et al. A Natural Language Processing–Based Virtual Patient Simulator and Intelligent Tutoring System for the Clinical Diagnostic Process: Simulator Development and Case Study. *JMIR Med Inform.* 2021;9(4):e24073. doi:10.2196/24073
77. Winget M, Persky AM. A Practical Review of Mastery Learning. *Am J Pharm Educ.* 2022;86(10):8906. doi:10.5688/ajpe8906
78. 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. doi:10.1007/s00521-024-10868-x
79. 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
80. 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
81. Lang BH, Nyholm S, Blumenthal-Barby J. Responsibility Gaps and Black Box Healthcare AI: Shared Responsibilization as a Solution. *Digit Soc.* 2023;2:52. doi:10.1007/s44206-023-00073-z
82. 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

83. 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
84. Kaufman DM, Mann KV. Chapter 2: 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.
85. 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
86. 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
87. 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, USA. doi:10.1109/RO-MAN60168.2024.10731328
88. 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, QC, Canada. doi:10.1007/978-3-540-69132-7_10

89. 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
90. 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
91. 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
92. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med.* 2020;26:1364-1367. doi:10.1038/s41591-020-1034-x
93. 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
94. Faul F, Erdfelder E, Buchner A, Lang A-G. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behav Res Methods.* 2009;41:1149-1160. doi:10.3758/BRM.41.4.1149
95. Random.org. Accessed November 15, 2023. <https://www.random.org/>

96. 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;137:1160-1171. doi:10.3171/2021.12.JNS211563
97. 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
98. R Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing; 2025. <https://www.R-project.org/>
99. 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
100. 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
101. 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
102. 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
103. 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

104. 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
105. 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
106. Nazaretsky T, Mejia-Domenzain P, Swamy V, Frej J, Käser T. AI or Human? Evaluating Student Feedback Perceptions in Higher Education. Paper presented at: 19th European Conference on Technology Enhanced Learning; September 16-20, 2024; Krems, Austria. doi:10.1007/978-3-031-72315-5_20
107. Scheiter K, Gerjets P, Vollman B, Catrambone R. The impact of learner characteristics on information utilization strategies, cognitive load experienced, and performance in hypermedia learning. *Learn Instr.* 2009;19(5):387-401. doi:10.1016/j.learninstruc.2009.02.004

TABLES

Table 1. ICEMS Metrics and Commands

Metric	ICEMS Command
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.”

Metrics and their corresponding commands are in hierarchical order. The commands in the right column are the instructions that groups 1 and 2 received when an error was detected in their performance. Abbreviations: ICEMS, Intelligent Continuous Expertise Monitoring System.

Table 2. Demographic Characteristics of Included Study Participants

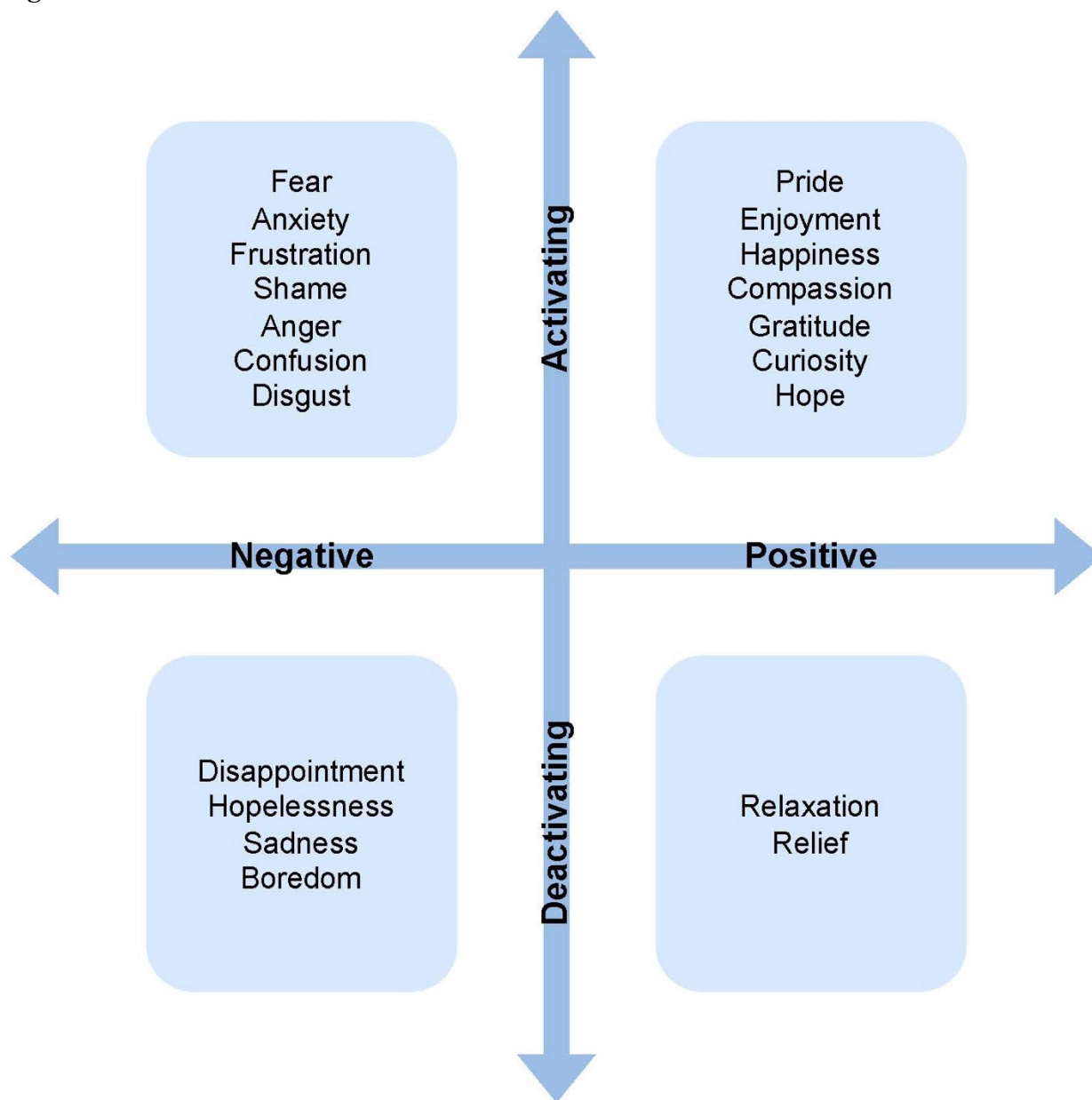
Characteristic	Group 1 AI tutor instruction (n = 30)	Group 2 Expert instruction (n = 29)	Group 3 Personalized expert instruction (n = 28)	All participants (n = 87)
Age, mean (SD)	21.8 (2.4)	22.6 (4.4)	23.9 (4.8)	22.7 (4.0)
Sex				
Female	18	16	12	46
Male	12	13	15	40
Prefer not to say	0	0	1	1
Gender				
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 (SD)^a	4.0 (0.9)	4.1 (1.0)	3.9 (1.0)	4.0 (1.0)
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 yrs				
Yes	9	9	13	30
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; SD, standard deviation; VR, virtual reality.

^a Measured using a 5-point Likert scale with 1 indicating less interest and 5 indicating more interest.

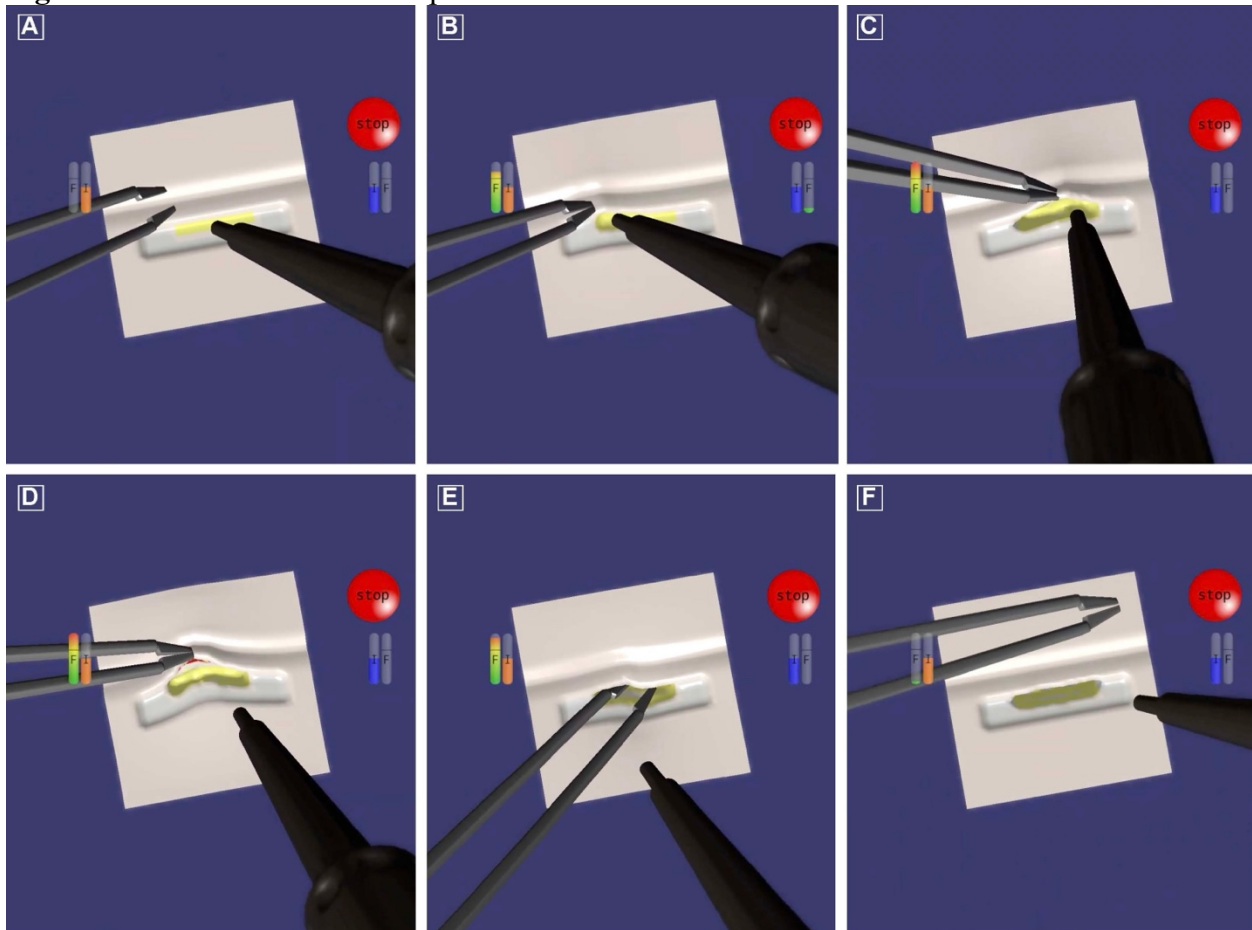
FIGURES

Figure 1. The Medical Emotions Scale



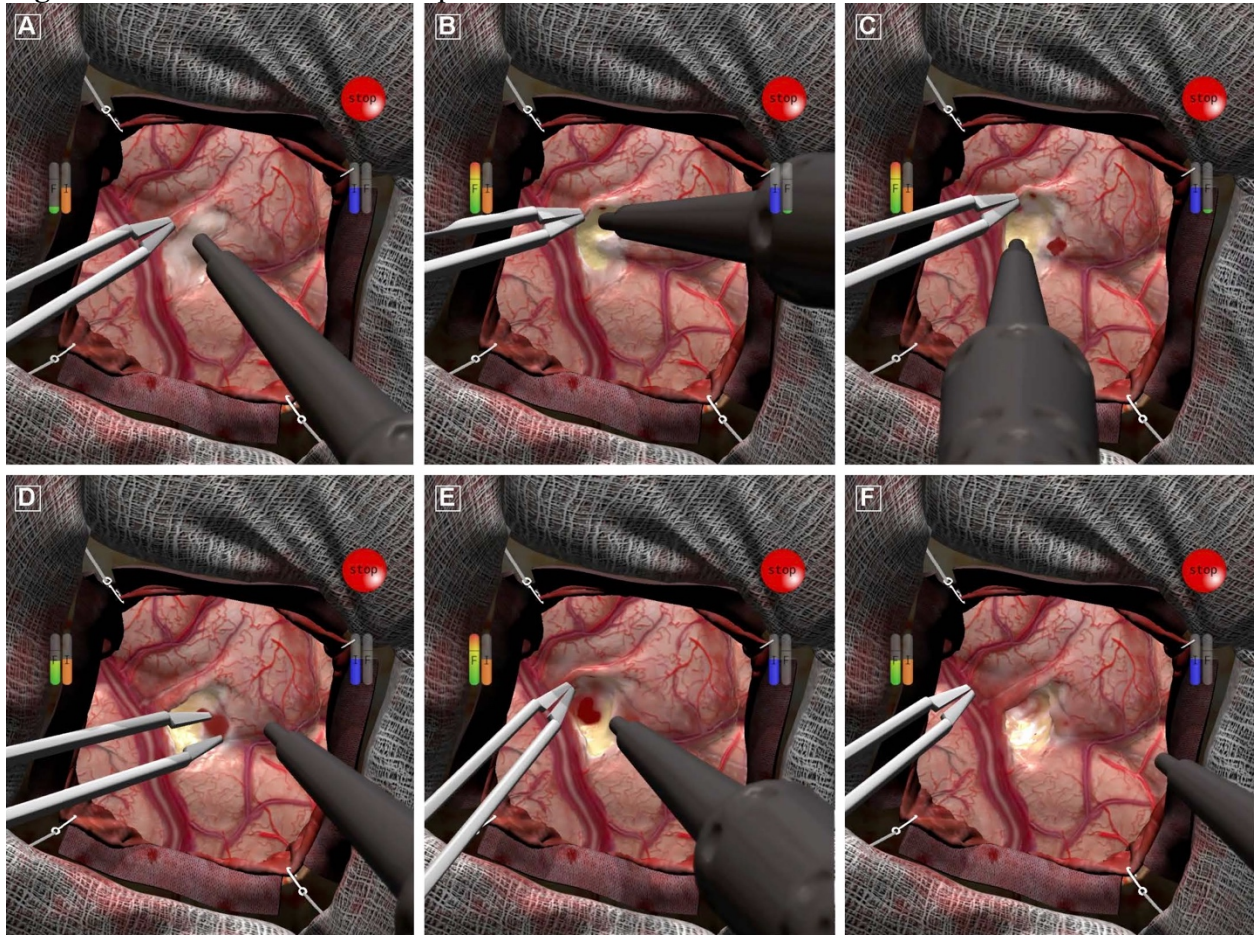
Adapted from Duffy et al.³⁷ X-axis represents valence and Y-axis represents activation. Two emotions (neutrality and surprise) were excluded from analysis as they do not have valence.

Figure 2. NeuroVR Practice Subpial Tumour Resection Scenario

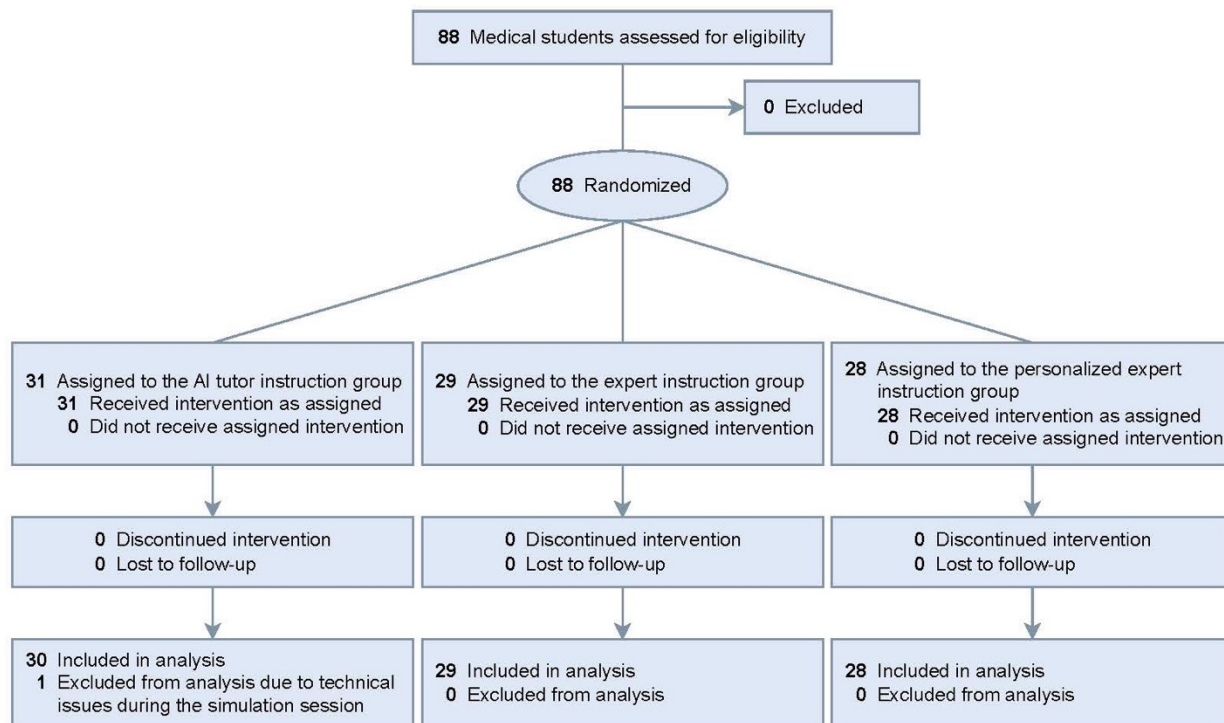


(A) Start of the practice subpial resection simulation scenario. Yellow section represents the tumour and white represents healthy tissue. Instrument on the left is the bipolar forceps and instrument on the right is the ultrasonic aspirator. (B) Participant uses the bipolar to lift the pia and the aspirator to resect the underlying tumour. (C) Appearance following resection of the superficial tumour. Yellow tissue remaining depicts deeper tumour areas. (D) Participant exposes the simulated deep cerebral vessel (red). (E) Participant cauterizes a bleeding point using the bipolar. (F) Complete resection of the tumour.

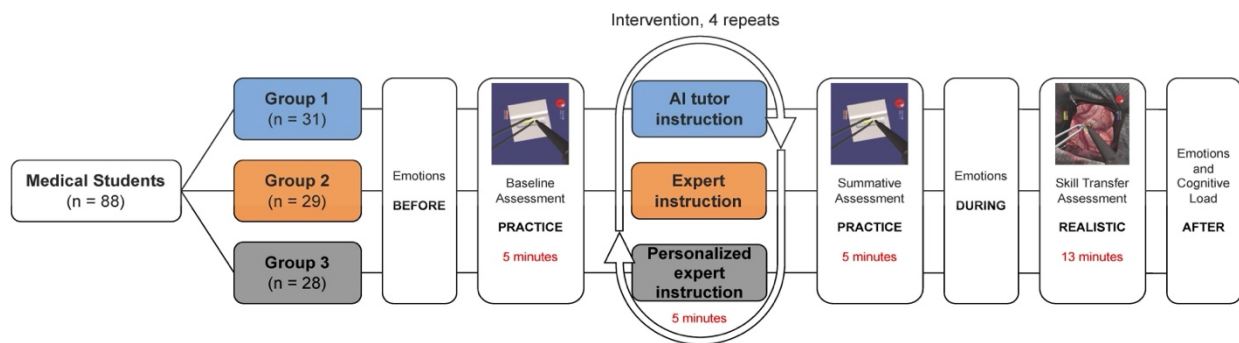
Figure 3. NeuroVR Realistic Subpial Tumour Resection Scenario



(A) Start of the realistic subpial resection simulation scenario. Off-white tissue represents the tumour. Instrument on the left is the bipolar forceps and instrument on the right is the ultrasonic aspirator. (B) Participant is using the bipolar to lift the pia and the aspirator to resect the underlying tumour. (C) Participant causes minor bleeding from the tumour while using the aspirator. (D) Participant cauterizes a bleeding point using the bipolar. (E) Participant causes significant bleeding by damaging the healthy tissue. (F) Complete resection of the tumour.

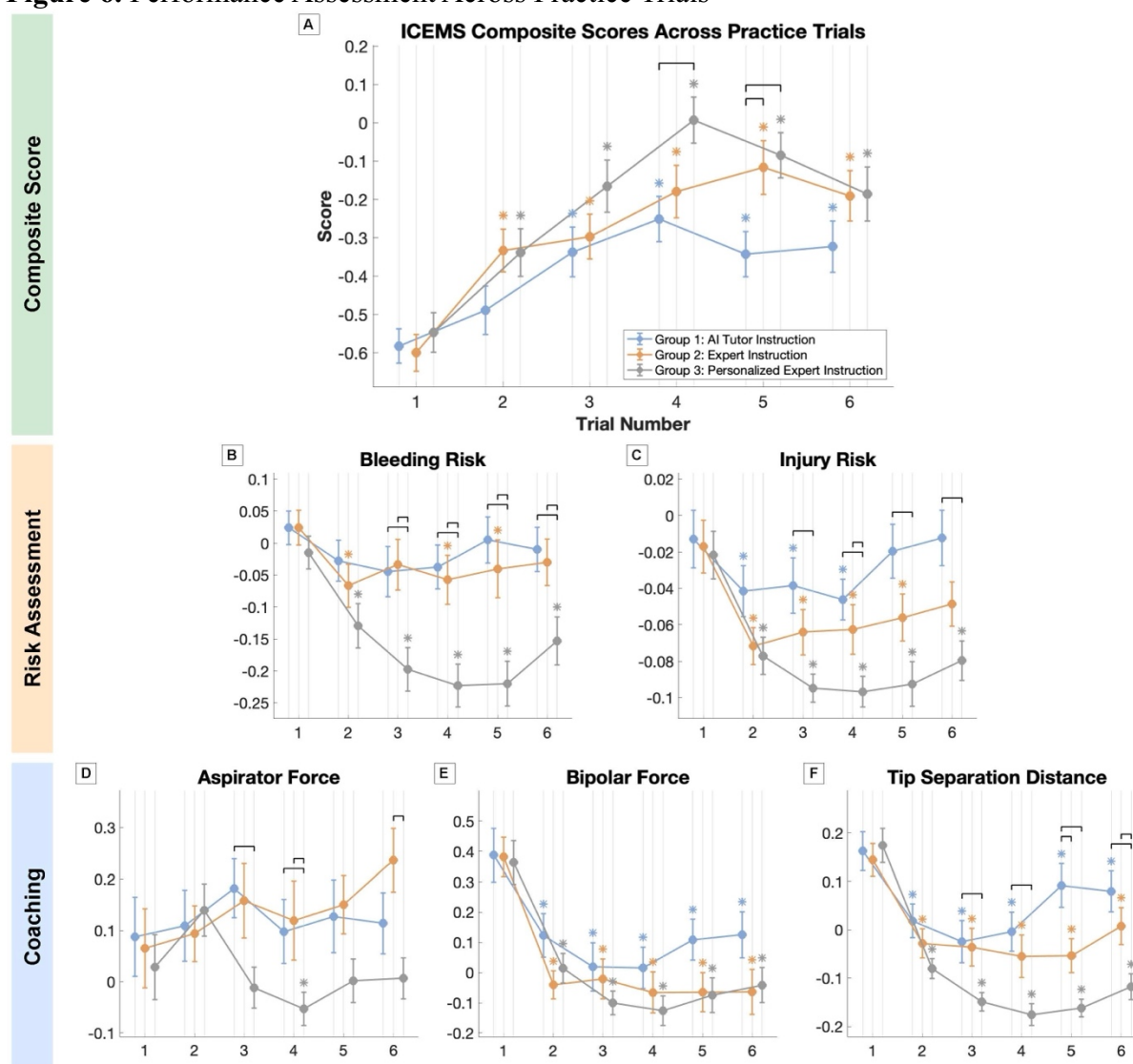
Figure 4. Participant Recruitment Flowchart

Abbreviations: AI, artificial intelligence.

Figure 5. Flow Chart of Events in Randomized Controlled Trial

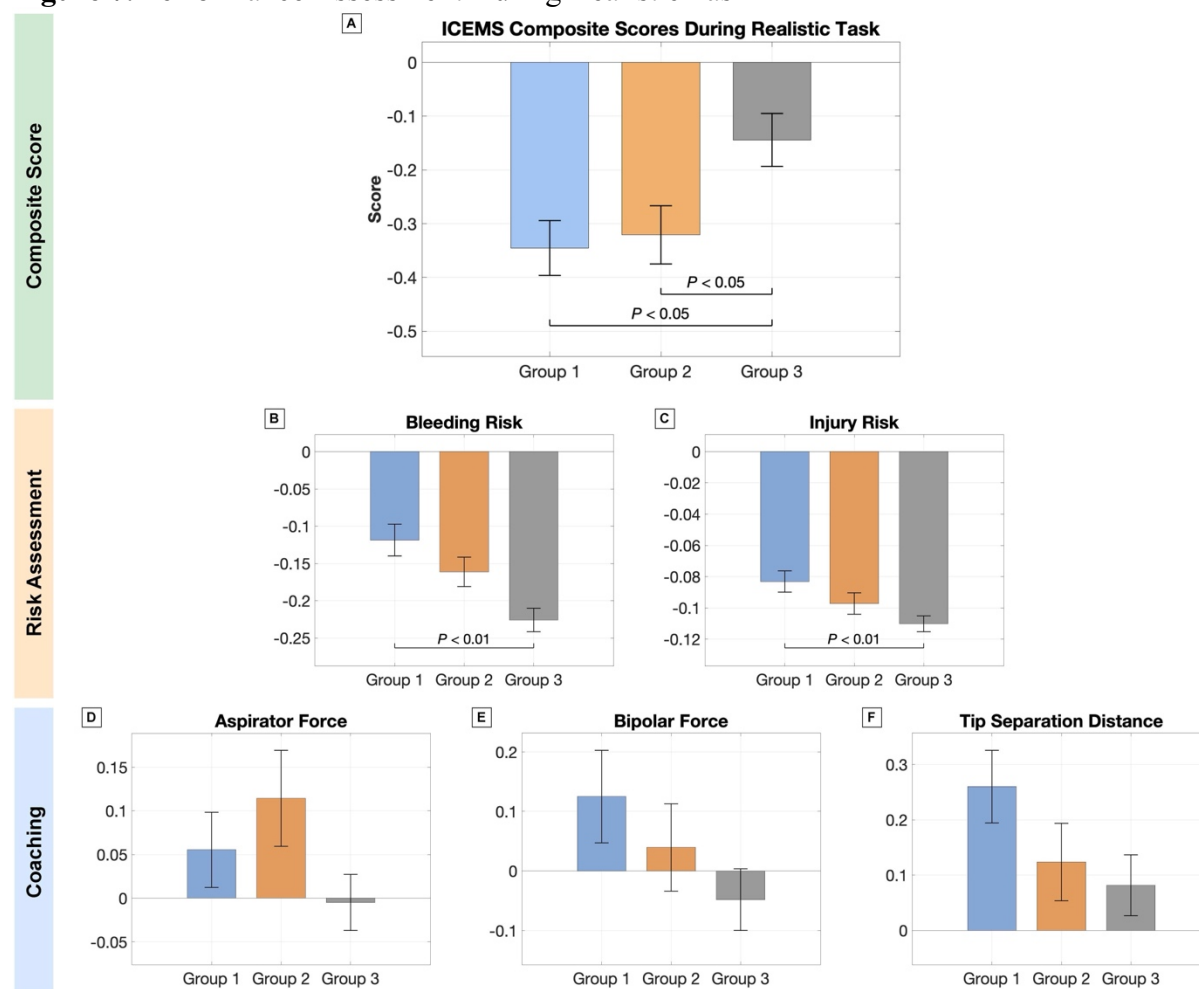
Abbreviations: AI, artificial intelligence.

Figure 6. Performance Assessment Across Practice Trials

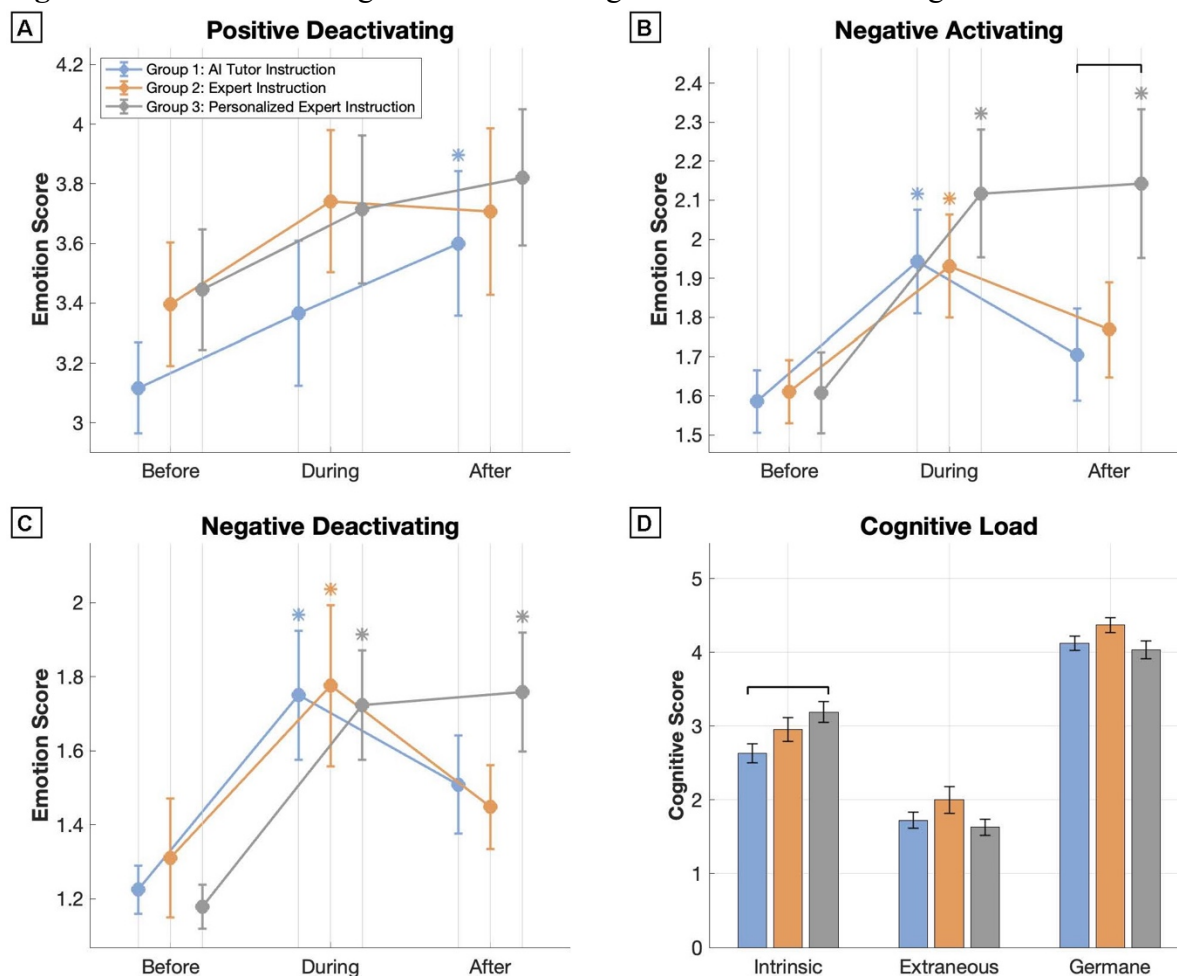


ICEMS composite expertise scores and five performance metrics across all practice trials. Groups are colour coded (see legend). X-axis represents the trial number and Y-axis represents the composite or metric score calculated by the ICEMS. Points represent group means and vertical bars represent standard errors. Black horizontal brackets indicate statistically significant differences between groups ($P < 0.05$) during a given trial. Asterisks indicate statistically significant differences from the baseline ($P < 0.05$) for that group. Error bars represent standard errors. Abbreviations: ICEMS, Intelligent Continuous Expertise Monitoring System; AI, artificial intelligence.

Figure 7. Performance Assessment During Realistic Task



ICEMS composite expertise scores and five performance metrics during realistic task. X-axis represents the group and Y-axis represents the composite or metric score calculated by the ICEMS. Colored bars represent group means and vertical bars represent standard errors. Black horizontal brackets indicate statistically significant differences between groups ($P < 0.05$) during the realistic task. Error bars represent standard errors. Abbreviations: ICEMS, Intelligent Continuous Expertise Monitoring System.

Figure 8. Emotions and Cognitive Load Throughout Simulation Training

Groups are colour coded (see legend). Y-axis represents the Medical Emotions Scale or Cognitive Load Index score. Black horizontal brackets indicate statistically significant differences between groups ($P < 0.05$). Asterisks indicate statistically significant differences from the baseline ($P < 0.05$) for that group. Error bars represent standard errors. (A) Positive deactivating emotions include relaxation and relief, (B) negative activating emotions include frustration and fear, (C) negative deactivating emotions include disappointment and boredom. (D) Cognitive load consists of intrinsic, extraneous, and germane load. Abbreviations: AI, artificial intelligence.