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

Vanja Davidovic, BHSc

Department of Surgical and Interventional Sciences

Faculty of Medicine and Health Sciences

McGill University

Montreal, Canada

August 2025

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

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 ..... **xiv**
Abbreviations ..... **xvi**
Thesis Introduction **17**
Background **18**
Instructional Methods in Surgical Education 18
Simulation in Surgical Education 22
Simulation in Neurosurgical Education 23
Performance Assessment in Surgery 25
Performance Assessment Algorithms 26
Intelligent Tutoring Systems 28
Rationale **30**
The Study Hypothesis **31**
The Study Objectives **31**
The Study **33**
Highlights 35
Abstract 35

Keywords 36
Introduction 37
Methods 39
Participants 39
Study Procedure and Simulation Session 40
Interventions 41
Group 1: AI Tutor Instruction 41
Group 2: Scripted Human Instruction 41
Group 3: AI-Augmented Personalized Instruction 42
Performance Metric Extraction 42
Outcome Measures 43
Statistical Analysis 44
Results 45
Feedback Frequency Across Simulated Practice Subpial Resections 46
Technical Skill Performance Across Simulated Practice Subpial Resections 47
Technical Skill Transfer to the Simulated Realistic Subpial Resection 49
Discussion 49
Limitations 52
Conclusion 52
Thesis Summary 54
Contributions to Original Knowledge 54
Discussion 54
Limitations 57
Future Directions 58
Conclusion 58
References 60
Figures 70
Figure 1 70
Figure 2 71
Figure 3 72
Figure 4 73

Figure 5 74
Figure 6 75
Figure 7 76
Tables..... 77
Table 1. Metrics assessed by the ICEMS in hierarchical order, and their corresponding instructions. 77
Table 2. Demographic characteristics of included study participants. 78
Table 3. Incidence rates and standard errors of instructions for six ICEMS metrics received during the second to fifth repetitions of the practice resection scenario. 79
Table 4. Estimated geometric means and standard errors of technical skill performance metrics over six repetitions of the practice resection scenario. 80
Table 5. Estimated marginal means and standard errors of technical skill performance metrics during the realistic resection scenario. 81

ABSTRACT

Background

Current practices for teaching surgical technical skills rely on the subjective observations of human instructors, underscoring a need for objective instructional methodologies and performance assessments that are standardized across teaching institutions. Thus, our team developed the Intelligent Continuous Expertise Monitoring System (ICEMS), an artificial intelligence (AI) system that uses quantitative data to continuously assesses trainee performance and provide real-time verbal feedback. Using the NeuroVR simulation platform, a randomized controlled trial from our center found that AI-augmented personalized instruction resulted in enhanced ICEMS scores on a simulated subpial resection scenario compared to AI tutor instruction and scripted human instruction.

Objectives

The objective of this study is to determine whether AI-augmented personalized instruction will result in a reduced feedback frequency and be more effective in improving surgical technical skill acquisition in a simulated surgical scenario compared to intelligent tutor instruction alone.

Methods

The number of feedback instructions that resulted from each instructional method was extracted from the ICEMS and analyzed. Feedback focused on 6 predetermined, AI-derived metrics: healthy tissue injury risk, bleeding risk, high instrument tip separation distance, high force applied with the bipolar forceps, low force applied with the bipolar forceps, and high force applied with the ultrasonic aspirator. In addition, participant performance was assessed through technical skill performance metrics, recorded by the NeuroVR simulation platform, including the rate of healthy tissue removal (mm^3/t), total volume of blood lost (mm^3), instrument tip

separation distance (mm), force applied with bipolar forceps (N), and force applied with ultrasonic aspirator (N). The mean of each performance metric was calculated for each repetition of the simulated scenario.

Results

The analysis included 522 practice scenarios and 87 realistic scenarios. By the third repetition of the practice scenario, the AI-augmented personalized instruction group received significantly fewer total instructions (incidence rate ratio (IRR), 1.50 [95% CI, 1.16 to 1.94] instructions; $P <$ .001), and instructions relating to high aspirator force application (IRR, 1.71 [95% CI, 1.15 to 2.55] instructions; $P = .002$) compared to the second repetition. Compared to AI tutor instruction, AI-augmented personalized instruction resulted in improved technical skill performance, including a significantly lower rate of healthy tissue removal ($P = .01$), instrument tip separation distance (mean ratio, 1.25 [95% CI, 1.05 to 1.50] mm; $P = .008$), and aspirator force (mean ratio, 1.68 [95% CI, 1.23 to 2.31] N; $P < .001$) by the third repetition of the practice scenario. The AI-augmented personalized instruction group showed a significant improvement from baseline in all subsequent repetitions for all five performance metrics.

Conclusion

Artificial intelligence-augmented personalized instruction resulted in less frequent feedback and an improvement in simulated surgical skills, providing further evidence for the critical role that human educators play in an intelligent operating room environment.

RÉSUMÉ

Contexte

Les méthodes courantes d'enseignement des compétences techniques chirurgicales sont fondées sur les observations subjectives des instructeurs humains. Cela souligne l'importance des méthodologies pédagogiques objectives et des évaluations de la performance standardisées dans l'ensemble des établissements d'enseignement. Ainsi, notre équipe a développé *l'Intelligent Continuous Expertise Monitoring System* (ICEMS), un système d'intelligence artificielle (IA) qui utilise des données quantitatives pour évaluer les performances des stagiaires en continu et pour fournir des instructions verbales en temps réel. Avec l'aide de la plateforme de simulation *NeuroVR*, un essai contrôlé randomisé mené par notre centre a démontré que des instructions personnalisées augmentées par l'IA entraînaient de meilleurs résultats d'ICEMS dans un scénario simulé de résection sous-piale, comparativement aux instructions fournies par un tuteur IA et aux instructions humaines scriptées.

Objectifs

L'objectif de cette étude est de déterminer si des instructions personnalisées augmentées par l'IA entraîneront une réduction dans la fréquence de rétroactions et amélioreront, avec plus d'efficacité, l'acquisition des compétences techniques chirurgicales dans un scénario chirurgical simulé, comparativement aux instructions fournies uniquement par un tuteur IA.

Méthodes

Le nombre de rétroactions générées par chaque méthode d'enseignement a été extrait de l'ICEMS et analysé. Les instructions étaient concentrées sur six indicateurs prédéterminés, dérivés de l'IA : le risque de lésion des tissus sains, le risque de saignement, la distance de séparation élevée des pointes des instruments, une force trop élevée appliquée avec la pince

bipolaire, une faible trop force appliquée avec la pince bipolaire et une force trop élevée appliquée avec l'aspirateur à ultrasons. De plus, les performances des participants ont été évaluées à l'aide d'indicateurs de compétences techniques, enregistrées par la plateforme de simulation *NeuroVR*, notamment : le taux d'élimination des tissus sains (mm^3/t), le volume total de sang perdu (mm^3), la distance de séparation des pointes des instruments (mm), la force appliquée avec une pince bipolaire (N) et la force appliquée avec un aspirateur à ultrasons (N). La moyenne de chaque mesure de performance a été calculée pour chaque répétition du scénario simulé.

Résultats

L'analyse comprenait 522 scénarios d'entraînement et 87 scénarios réalistes. Dès la troisième répétition du scénario d'entraînement, le groupe recevant des instructions personnalisées augmentées par l'IA avait reçu une diminution significative du nombre de rétroactions (rapport de taux d'incidence (TRI) 1,50 [IC de 95 % 1,16 à 1,94] instructions ; $p < 0,001$), ainsi que moins de rétroactions concernant l'application d'une force d'aspiration excessive (TRI 1,71 [IC de 95 % 1,15 à 2,55] instructions ; $p = 0,002$) comparativement à la deuxième répétition. Par rapport aux instructions fournies par le tuteur IA, nous avons observé une amélioration des performances techniques, notamment : une réduction significative du taux d'élimination des tissus sains ($p = 0,01$), de la distance de séparation des pointes des instruments (rapport moyen 1,25 [IC de 95 % 1,05 à 1,50] mm ; $p = 0,008$) et de la force d'aspiration (rapport moyen 1,68 [IC de 95 % 1,23 à 2,31] N ; $p < 0,001$) lors de la troisième répétition du scénario d'entraînement. Le groupe ayant reçu des instructions personnalisées augmentées par l'IA a montré une amélioration significative par rapport à la ligne de base pour les cinq indicateurs de performances, au cours des répétitions suivantes.

Conclusion

L’instruction personnalisée augmentée par l'intelligence artificielle a entraîné une diminution significative du nombre de rétroactions fournies aux stagiaires ainsi qu’une amélioration des compétences chirurgicales en simulations, apportant des preuves supplémentaires qui illustrent le rôle essentiel des éducateurs humains dans un environnement opératoire intelligente.

DEDICATION AND PREFACE

To women and girls with ADHD: Your mind may work differently, but it will always be your greatest asset. Embrace your difference, as it's a source of your brilliance.

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

This work was presented at McGill University's Neurosurgical Research Day in Montreal, QC, Canada on June 20, 2025.

ACKNOWLEDGEMENTS

The successful completion of this thesis would not have been possible without the invaluable contributions of several individuals. First and foremost, thank you to my supervisor and mentor, Dr. Rolando F. Del Maestro – your unwavering support has meant more than I can say. Thank you for seeing something in me and pushing me to see it too, for putting up with me as I suggest projects that are way beyond the scope of a master's degree, for your advice about research and about life, and for never giving up on me or my project when I was ready to give up. I cherish everything you have taught me over these past two years and I'm grateful for the opportunity to have learned from your expertise.

Thank you to my research advisory committee – Dr. Julio Flavio Fiore Jr., Dr. Carlo Santaguida, and Dr. Gregory Berry – for taking time out of their busy schedules to provide guidance and encouragement, and for challenging me throughout my degree. I'd also like to thank Dr. Jeffrey Atkinson for reviewing my thesis and providing stimulating comments that improved the quality of my final thesis.

To my colleagues at the Neurosurgical Simulation and Artificial Intelligence Learning Centre, I feel honoured to have worked with such a talented group of individuals. Through countless insightful discussions, I was lucky enough to learn from the vast breadth of knowledge that exists among all of you. Thank you to Trisha Tee, for your mentorship, your company in and out of the lab, and for being my friend. I feel very lucky to have met you and to have learned from you. To Bianca Giglio, who was always willing to lend a helping hand and listen to my brainstorms, and to Dr. Recai Yilmaz, who's guidance I would have been lost without: thank you for the enormous part you played in my success. Thank you to Matthew Ha and Sabrina Deraiche, who were key contributors to my thesis and who I'm now lucky enough to call my friends. To my colleagues,

Dr. Ali Fazlollahi, Dr. Neevya Balasubramaniam, Abicumaran Uthamacumaran, Dr. Abdulmajeed Albeloushi, Dr. Ahmad Alhaj, Dr. Mohamed Alhantoobi, and Dr. Rothaina Saeedi, your dedication, drive, and inquisitive minds inspire me.

I would not be where I am without my family. To my dad, Miloš Davidović, who has helped me realize my full potential and who never made me doubt that he would drop everything if I needed him. I'm grateful that I inherited your fix-it spirit. To my mom, Dragana Davidović, thank you for your gentle and unconditional love. Growing up, watching your success as you shattered the glass ceiling has truly been the greatest inspiration. Both of you sacrificed so much to come to Canada, and for you I am eternally grateful. Thank you for always showing me how proud you are of me; I'm proud of you, too. Mnogo vas volim. To my siblings, Marko and Danka Davidović, thank you for helping me throughout my academic career, even when you probably had no idea what I was talking about. Marko, your perseverance and your dedication to staying true to yourself leaves me in awe. Danka, I wanted to be just like you growing up, and I still do. I couldn't have asked for better role models in my life. I don't say it enough, but I love you both a lot. And to Oli, my sweet, fluffy boy who sometimes drives me crazy, but who is also my pride and joy, thank you for your comfort and love.

I'm grateful for the unwavering support of my friends over these past two years. To Olivia Ardilliez, thank you for being my biggest cheerleader, no matter what I'm doing or how good I am at it. You've been by my side for over a decade, and I can't imagine life without you. To Olivia Thomsen and Grace Bennett, thank you for always believing in me, even when I didn't believe in myself. Your unconditional love and support mean more than I can say. To Jordyn Gattie, thank you for always being honest with me, even when you know it's not what I want to

hear. I'm grateful for your patience and uplifting energy, and I'm so very proud of you. And to Abigail Francis, thank you for proofreading my work and keeping me sane when I needed it. Finally, I can't write acknowledgements without mentioning one of my biggest drivers. At graduation, the science teachers from my high school chose one student to present with the NHS Science Award: "presented to the student who through aptitude, curiosity, and tenacity in the senior scientists has shown the most potential to make a positive impact on the world of scientific inquiry after graduation." To these teachers, who sparked my love of STEM and who probably don't realize the impact they made on me, as corny as it may sound, your belief in me as a scientist made me believe in myself. I'm forever grateful for your mentorship. You are amazing at your jobs and your students are incredibly lucky to have you in their corner. You may not remember or realize the impact you made on my life, but I certainly do, so thank you.

AUTHOR CONTRIBUTIONS

The structure of this thesis follows a manuscript-based format, and the authors of the manuscript have made substantial contributions to finalizing this work. The author's contributions are detailed using the Contributor Roles Taxonomy (CRediT) format.[1,2] The following statements outline the specific contributions to this research project made by each individual.

Vanja Davidovic: Contributed to conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, resources, visualization, writing – original draft, and writing – review & editing.

Bianca Giglio: Contributed to conceptualization, investigation, methodology, project administration, resources, and writing – review & editing.

Dr. Abdulmajeed Albeloushi: Contributed to investigation.

Dr. Ahmad Kh. Alhaj: Contributed to investigation.

Dr. Mohamed Alhantoobi: Contributed to investigation.

Dr. Rothaina Saeedi: Contributed to investigation.

Sabrina Deraiche: Contributed to formal analysis and resources.

Dr. Recai Yilmaz: Contributed to conceptualization, data curation, methodology, resources, and software.

Trisha Tee: Contributed to conceptualization and methodology.

Dr. Ali M. Fazlollahi: Contributed to conceptualization and methodology.

Matthew Ha: Contributed to conceptualization and methodology.

Abicumaran Uthamacumaran: Contributed to conceptualization.

Dr. Neevya Balasubramaniam: Contributed to conceptualization, methodology, and resources.

Widad Safih: Contributed to resources.

Dr. José A. Correa: Contributed to formal analysis.

Dr. Rolando F. Del Maestro: Contributed to conceptualization, data curation, funding acquisition, methodology, project administration, resources, supervision, writing – original draft, and writing – review & editing.

ABBREVIATIONS

CBME	Competency-Based Medical Education
OSATS	Objective Structured Assessment of Surgical Technical Skill
VR	Virtual Reality
AI	Artificial Intelligence
ICEMS	Intelligent Continuous Expertise Monitoring System
LSTM	Long Short-Term Memory
RCT	Randomized Controlled Trial
3D	Three-Dimensions
EPA	Entrustable Professional Activities
PGY	Post-Graduate Year
VOA	Virtual Operative Assistant
STROBE	Strengthening The Reporting of Observational Studies in Epidemiology
MLASE	Machine Learning to Assess Surgical Expertise
GLMM	Generalized Linear Mixed Model
IRR	Incidence Rate Ratio
CI	Confidence Interval
ANOVA	Analysis Of Variance
LLM	Large Language Model
EGM	Estimated Geometric Mean
EMM	Estimated Marginal Mean

INTRODUCTION

Being one of the most perilous, dynamic medical fields, the surgical domain is associated with medical errors.[3–10] Patient outcomes depend on a surgeon's technical and non-technical skills,[8–21] particularly in the field of neurosurgery, where errors can result in significant patient morbidity and mortality.[9,10,13,22–26] Due to the correlation between surgical errors and complication rates, a greater emphasis has been placed on assessing performance throughout surgical residency programs by defining measurable competencies.[27,28] These program curricula are shifting from the traditional, Halstedian, time-based approach towards the adoption of a competency-based medical education (CBME) curriculum.[29,30] However, assessing these competencies has proven difficult due to a lack of standardized, structured assessments.[11,27,31] Global rating scales, such as the Objective Structured Assessment of Surgical Technical Skill (OSATS),[32] have been adopted; however, they have been criticized for their reliance on qualitative data, introducing subjectivity and bias due to the variability between individual educators.[33]

Innovations such as virtual reality (VR) simulators have shown great potential in addressing these limitations.[34–40] By replicating the visual, auditory, and haptic feedback of particularly challenging procedures, VR provides an immersive, controlled environment that can effectively prepare trainees before they enter the operating room.[13,28,36–38,41] Virtual reality simulators capture large amounts of quantitative performance data in real-time, much of which cannot be assessed by human instructors, such as the economy of movement or volume of blood lost.[34] And yet, these systems still tend to rely on human instructors to assess trainee performance and provide feedback, maintaining the aforementioned issue of subjectivity.[34,41] Feeding this data into artificial intelligence (AI) technologies opens the door to the objective, structured assessment of skills and logging of performance overtime.[18,42–46] AI systems can identify deficits in trainee performance and provide actionable feedback accordingly, thereby supplementing

current surgical teaching practices.[41–43,47,48] Our team has designed one such intelligent tutoring application, which can be integrated into a VR surgical platform to train bimanual psychomotor skills.[47]

The Intelligent Continuous Expertise Monitoring System (ICEMS) uses a long short-term memory (LSTM) network to continuously assess trainee performance in 0.2-second intervals and provide continuous, real-time verbal feedback to improve trainee performance and mitigate errors.[47] It has been validated for its ability to assess performance,[47] and outperformed human expert instructors in teaching technical skills on a neurosurgical simulation platform, validating it's use as an intelligent tutoring system.[42] However, studies have suggested that combining intelligent tutors with human instructors may be beneficial, as human instructors can contextualize the errors identified by the algorithm.[41,49] A recent randomized controlled trial (RCT) sought to investigate this combination; however, the study focused on the ICEMS's performance assessment, rather than other aspects of trainee performance.[50] This thesis aims to investigate the impact of AI-augmented personalized instruction on the frequency of feedback instructions provided and on trainee technical skill performance. These findings can be used to inform the adoption of AI performance assessment and intelligent tutoring functionalities into existing surgical residency program curricula.

BACKGROUND

Instructional Methods in Surgical Education

Situational awareness, technical skills, and interpersonal skills are only some of the factors that make up a well-rounded, expert surgeon.[8–21] The ability to draw on large bodies of knowledge, confront uncertainty, and problem solve are key to positive patient outcomes and delivering safe, high-quality care.[8,11,12,17–21] There is a strong association between the application of these skills

and surgical outcomes[8–10,12,13,20,21]; surgery sees a high rate of preventable complications, and many of these are due to surgical errors.[3–10] A lack of technical skills has been linked with poor surgical outcomes,[8–10,12,13] accounting for approximately 25% of surgical complication rates.[10] Furthermore, the United States sees up to 98 000 annual deaths due to preventable medical errors.[51] In addition to these findings, up to 42% of residents don't feel adequately prepared to perform procedures on their own[27,52,53]; the 80-hour work week restriction limiting the diversity of cases they're exposed to.[6,27] Therefore, it is imperative that residents demonstrate competence in order to be prepared for independent clinical practice.[29] This is particularly important in neurosurgery, as these highly invasive procedures are especially vulnerable to medical errors.[26] For instance, a prospective study investigating errors in neurosurgical procedures found that 75% were preventable.[22] The surgical field has seen some significant innovations, ranging from the discovery of anesthesia to surgical robotics. However, the methods for teaching surgical residents have remained mostly unchanged.[16,27]

Surgical residency was founded in 1890 at Johns Hopkins Hospital by Dr. William Halsted based on the "see one, do one, teach one" approach.[54] In this model, trainees observe an attending surgeon's performance, then are expected to perform the operation based on what they learned, and afterwards teach their peers these skills.[27,29] The approach is based on apprenticeship, where trainees work closely with an expert surgeon gaining incremental responsibility, and it is assumed that an accumulation of knowledge (ie, a longer time spent practicing) inherently means better surgical skills, though this is not always necessarily the case.[34,37,55] This method of training surgical residents has been criticized, as it involves on-the-job training of residents while in the operating room, posing a potential risk to patient safety and ethical concerns.[6,30,31,34,40,56] This is

especially so given the abundance of alternative methods for skill development that are currently available.[30]

The modern paradigm follows a CBME model, which necessitates that residents reach a certain level of competency before progressing in their surgical residency program.[29,30] Defined learning objectives guide residents in their skill development.[29,30] In Canada, CBME is implemented through the Competence by Design (CBD) model.[57] Effective, personalized coaching, specifically in-the-moment coaching. has been identified as a critical component to the success of CBD.[57–59] It is supported by a detailed framework that facilitates conversations between learners and instructors, known as the R2C2 model.[59] Despite this push for improved feedback and guidance, surgical teaching methodologies remain very similar to the apprenticeship model, wherein residents continue to receive the majority of their training in the operating room, with the added aspects of in-the-moment coaching and formally assessing trainee performance before entrusting them with further responsibilities.[27,60] In this approach, residents may be limited in the diversity of cases they see,[27,61] and training relies on the presence of an instructor, limiting how often trainees can practice and demanding a lot of time from attending surgeons whose secondary focus is the resident's learning – the primary being providing quality care to their patients.[27,60] The risk posed to patient safety also remains.[30,31,34] Therefore, there is a clear need to expand surgical teaching methods beyond the confines of the operating room.

Studies have shown that a variety of instructional methods should be used when teaching surgical trainees, such as learning by performing tasks, self-reflection, modelling behaviours of experienced surgeons, self-directed study, and more.[62] However, in practice, the limitations imposed by learning in the operating room make these approaches incompatible to the current

teaching paradigm, as implementing these methods introduces concerns of time constraints and patient safety.[62–64]

Previous research has also supported the concept of deliberate practice,[13,29,35,55,65,66] wherein trainees practice technical skills through tasks with well-defined goals while receiving real-time feedback to improve their performance.[55] For the practice to be deliberate, there must be the opportunity to continually train, repeating the task many times to refine performance, and overtime be faced with new challenges to overcome.[55] This method of skills acquisition involves complex cognitive processes that help trainees avoid becoming automated in how they perform a task, and they instead veer towards mastery.[55] The use of these principles is not evident in teaching practices in the human operating room due to difficulties in applying these systems in these complex surgical environments. Repeating steps is difficult since prolonging a procedure for teaching purposes creates unnecessary risks to patient safety.[63,64] Furthermore, deliberate practice requires trainees to challenge themselves,[55] but applying this facet in the human operating room would involve a resident attempting to deal with an operative issue that may be above their current ability, introducing another potential for patient harm.[13] To bypass these limitations and allow surgical residents to benefit from the application of deliberate practice, surgical simulators can be employed. Simulators provide a risk-free environment where trainees can be assessed on specific criteria, receive real-time feedback, refine challenging techniques, and repeat a task indefinitely.[13,28,36,55,56,61] In addition, simulators lend themselves well to the current surgical education curriculum, as they provide a platform for CBD in-the-moment coaching without necessitating the presence of a human instructor.

Simulation in Surgical Education

Simulation involves the reproduction of real-life experiences, immersing the user in the simulation environment.[13,36,67] It has proven to be useful for training in many domains, such as aviation training,[68] though its applications in the surgical field are only beginning to be developed.[69] Simulation can allow trainees to acquire skills in a risk-free, controlled environment, rather than in the operating room.[13,34,36,41,55] A large benefit of simulators is their ability to be assessed for face and content validity (ie, realism of the simulation setting and how applicable the system is as a teaching tool, respectively).[34] Validated simulators can be used for training essential surgical skills, decreasing the emphasis placed on training done in the operating room if adequate alternative models are available.

Simulation includes live animal models, human cadaver models, synthetic models, and VR systems.[34,70] Applications of all these models have been created for teaching surgical residents, but most of their training is still done on patients in the operating room.[60] Live animal models include *in vivo* and *ex vivo* models.[70] These are considered high-fidelity simulations, as the biological tissue is similar to that of a human, so trainees can practice all aspects of an operation.[70] However, they come with disadvantages, most notably their high costs, limitations in repetition, and ethical concerns.[70] Human cadavers are considered the gold standard for surgical simulation due to their high-fidelity.[70] However, these models may not always be conducive for practicing certain procedures (eg, decreased tissue quality in embalmed cadavers), are expensive to obtain and maintain, have limited availability, and are not re-usable.[70] Furthermore, synthetic simulators, such as benchtop and laparoscopic box simulators and manikins are typically low-fidelity, with new developments creating more high-fidelity options.[70] The utility of these models in developing surgical skills has been proven numerous times, and they are currently used in

many surgical training curricula, though these, too, have their limitations.[70] High-fidelity synthetic models are expensive and not readily available, while low-fidelity options predominately focus on individual techniques, rather than the interaction of many skills.[34,70] Surgical competence requires that residents are able to apply multiple skills simultaneously; an expert does not usually focus on a solitary skill while operating.[8,15,17–19,30] Additionally, both high- and low-fidelity synthetic models are just that – synthetic – meaning their realism is inherently limited.[70] Finally, VR simulators create realistic, immersive environments in which trainees can practice a variety of procedures on a single system.[70] They allow for the repetition of a procedure and are often considered high-fidelity, as the trainee manipulates realistic, computer-generated images while receiving haptic feedback.[67,70] These systems entirely remove the risk to patients,[13,14,29,34,36,41] allowing trainees to focus on their technical skill development, address their weaknesses, and challenge themselves by attempting more complex procedures.[13,36,39,41,55] They have proven to be useful in improving surgical skills in a variety of surgical fields.[34,42,43] Virtual reality simulators also collect performance data, allowing for objective and quantitative assessment and eliminating the need for supervision; trainees can receive feedback from the simulator itself.[34,40,42–46,71] The main disadvantage of VR simulation is its high cost; however, these systems are becoming more and more cost-effective as new research developments are made.[40,70]

Simulation in Neurosurgical Education

Of all the surgical fields, neurosurgery is known for being particularly high stakes, as even small errors can result in significant patient morbidity and mortality.[9,10,13,22–25] This characteristic makes this specialty a good candidate for the development of simulation-based training of technical skills. Trainees can practice outside of the high-stakes operating room environment,

creating an optimal setting for deliberate practice while mitigating patient harm. As such, our team has developed and validated VR and *ex vivo* neurosurgical simulators.[72–75]

The NeuroVR (CAE Healthcare, Montreal, Canada), developed by a team of researchers at the National Research Council Canada, is a high-fidelity, VR simulator that recreates the audiovisual and haptic experience of neurosurgical procedures.[72,76] Offering interactive neurosurgical and spinal procedures, the NeuroVR includes realistic anatomical structures, haptics, and physical and physiological responses that allow for a more immersive training experience.[72] The system consists of a microscope, which allows for 3D visualization, as well as two instruments – bipolar forceps and an ultrasonic aspirator – attached to haptic handles and activated by foot pedals.[72] The ultrasonic aspirator is used by a trainee to suction blood and resect the abnormal tissue, while the bipolar forceps allow for better visualization of the surgical field and can be used to cauterize bleeding points.[72] By consulting with expert neurosurgeons, the NeuroVR's face and content validity has been established. In addition, it's construct validity has been determined through the use of machine learning to evaluate simulator data.[48,73,76]

A subpial resection procedure involves the use of bipolar forceps to retract the pia mater in order to then resect abnormal tissue using an ultrasonic aspirator.[77] The NeuroVR houses two scenarios for this procedure: a simpler practice scenario used to acquire the necessary technical skills for this technique (Figure 1), and a realistic scenario to assess the transfer of these skills to a more complex procedure (Figure 2). In these scenarios, trainees must remove the glioma-like abnormal tissue while minimizing bleeding and damage to the surrounding healthy tissue.

Virtual reality simulators such as the NeuroVR are limited by the fact that they cannot fully replicate the sensation of handling biological tissue.[71] As such, an *ex vivo* model was developed by our group to provide a more realistic setting for trainee learning, while still maintaining a

risk-free environment.[75] This model utilizes a calf brain – an affordable and available model that is anatomically similar to a human pediatric brain.[75] This model has previously demonstrated face and content validity.[75]

Performance Assessment in Surgery

The modern paradigm for surgical residency training follows the CBME model, wherein residents are required to meet defined learning objectives to progress in their training and prepare for independent clinical practice.[29,30] In Canada, CBME is implemented through the CBD model, and these competencies are assessed using entrustable professional activities (EPAs).[57,78] EPAs are key tasks that a resident can be trusted to perform independently once competence has been demonstrated.[78] A surgical resident's progress in the residency program is measured by the successful completion of these EPAs. EPAs help standardize the assessment of core competencies in the CBME model; however, they lack objective, structured, and specific criteria for evaluating these domains, making them prone to subjectivity and bias. In addition, EPAs generally focus on the successful completion of steps in a procedure, rather than assessing a trainee's grasp of specific, essential technical skills.[79,80]

The Objective Structured Assessment of Technical Skills (OSATS) is just one of many global rating scales developed to evaluate technical skill performance, often regarded as the gold standard.[32] Using this scale, evaluators rate seven domains of performance using a 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.[32] While the OSATS solves the issue of specificity seen with EPAs when evaluating technical skill performance,[80] indicating set qualities to observe in a trainee's performance, this reliance on qualitative data introduces subjectivity and bias.[27,34,62] Although employing multiple evaluation

methods from multiple evaluators can be beneficial to a trainee's development,[81] assessments relying on qualitative data pose difficulties in generalizing grading across multiple instructors, let alone multiple institutions, and may make it difficult for trainees to determine whether their performance is improving overtime. Thus, a need for an objective, structured, and specific assessment of trainee technical skills is identified.

Performance Assessment Algorithms

AI systems are made to simulate human intelligence and reasoning.[82] In the case of machine learning, a branch of AI, these applications have progressed as far as computers learning and acquiring human intelligence by identifying patterns in datasets and predicting outcomes based on input data.[82] Deep learning is a subset of machine learning, in which multiple neural networks simulate human decision making.[82] There are three main methods by which machine learning algorithms can learn: unsupervised learning, supervised learning, and reinforcement learning.[83] Supervised learning involves feeding labelled data into a machine learning algorithm, allowing it to recognize patterns within the dataset to make accurate, informed predictions when provided with new data.[83] AI has shown great promise in many aspects of the surgical field, including but not limited to, predicting patient outcomes,[84,85] detecting pathologies,[86,87] and surgical skill assessment[44,45,47] and training.[42,43,46,47]

AI has proven to be a valuable tool for finding hidden patterns within datasets, helping researchers understand ambiguous findings. This feature is particularly intriguing for surgical training, where there is a dearth of objective, quantitative assessments of surgical technical skill.[11,27,31] Measuring an expert surgeon's psychomotor skills and understanding the components that make up expert-level performance has proven difficult, making assessing these skills and teaching them to a novice difficult as well.[88] Given that AI can process and analyze large,

complex datasets, these algorithms may be helpful in quantifying skills, thereby paving the way to the development of objective assessments.

Difficulties with understanding the decision-making processes of unsupervised machine learning algorithms are known as the “Black Box” problem, and they lead to hesitations in integrating AI into the surgical field.[89] Surgery is a very high stakes setting, thus, not understanding the reasoning used by innovative technologies could have detrimental effects to outcomes. This is unhelpful when trying to provide feedback to improve trainee performance, as well. However, there are methods to circumvent this. Extracting key features of expert performance and feeding them into a supervised learning algorithm can allow developers some control over the model’s decision making.[48,90] These features can be used as benchmarks of expert performance, providing specific metrics as a standard of reference against which trainees can be assessed. Using this method, our group developed the ICEMS, a deep learning application for assessing surgical performance.

The ICEMS was developed by collecting data from 12 medical students (novices) and 14 neurosurgeons (experts) performing a subpial brain tumor resection procedure on the NeuroVR.[47] Subsequently, sixteen performance metrics associated with instrument handling (eg, velocity, acceleration) and risk assessment (eg, healthy tissue injury risk) that distinguished expert performance from novices were extracted.[47] This metric data was then labelled and used to train an LSTM model to differentiate expertise level among participants.[47] As such, the ICEMS quantitatively assesses performance at 0.2-second intervals, providing an expertise score ranging from -1.00 (novice) to 1.00 (expert) based on these sixteen metrics.[47] The ICEMS has demonstrated predictive validity by accurately distinguishing between the surgical performance of medical students, junior residents (post-graduate year [PGY] 1 to 3), senior residents (PGY 4

to 6), and neurosurgeons.[47] The method by which this system was developed allows us to understand the logic behind its decision making, making it a practical tool for scoring trainee surgical simulation performance. In fact, it has been used to score participant performance in a variety of RCTs.[42,43]

Intelligent Tutoring Systems

Intelligent tutoring systems are computer-based systems that leverage AI techniques to provide feedback to learners.[91,92] These systems can simulate one-one-one learning, helping trainees acquire skills and knowledge relevant to their field.[91,92] In the medical field, intelligent tutoring systems have proven useful in a variety of areas, including clinical reasoning, diagnoses, treatment planning, and skills training.[91] In surgery, these systems can be used to quantitatively assess performance, continuously identify errors, and provide real-time feedback to mitigate these errors.[46,47]

The Virtual Operative Assistant (VOA) was created by our group in 2020 to teach bimanual psychomotor skills.[46] By applying a linear support vector machine and AI-derived metrics to process NeuroVR performance data, the VOA calculates learner scores and classifies trainee performance as expert or novice.[46] With this scoring, it provides trainees with instructions to improve their performance following their completion of a simulation task.[46] The VOA provides instructions on four AI-selected metrics: two safety metrics – bipolar forceps force application and rate of bleeding, and two instrument movement metrics – instrument tip separation distance and bipolar forceps acceleration.[46] Trainees are required to master the metrics pertaining to safety before moving on to the instrument movement metrics.[46] In an RCT, medical students trained using the VOA outperformed those taught by a remote human expert instructor, indicating the VOA's utility as a surgical teaching tool.[43] However, as previously discussed, principles of

deliberate practice suggest that surgical training applications that provide real-time assessment and feedback are preferable for learner acquisition of knowledge and skills,[55] as they better mimic the dynamic operating room environment and trainee-instructor relationship. The VOA lacks the ability to continuously assess performance and provide real-time feedback to trainees, limiting its pedagogical utility.

Due to these shortcomings, the ICEMS was developed by our group.[47] This application uses an LSTM to intraoperatively assess and score trainee NeuroVR performance in five AI-selected metrics every 0.2 seconds.[47] These metrics consist of two safety metrics – tissue injury risk and bleeding risk – and three coaching metrics – instrument tip separation distance, bipolar forceps force application, and ultrasonic aspirator force application.[47] Using the NeuroVR performance data of 14 neurosurgeons, an LSTM established expert benchmarks that could then be used to detect errors in trainee performance.[47] The system provides real-time, auditory feedback to trainees following metric error detection; an error is defined as a trainee's score differing from the expert benchmark by one standard deviation for more than one second.[47]

A previous RCT ran by our group demonstrated the educational utility of the ICEMS, as medical students taught by the ICEMS outperformed those taught by an in-person expert human instructor.[42] However, the expert instructors were not provided with AI-derived performance data during this trial, relying solely on their observations, making their instructions vulnerable to subjectivity. Additionally, a randomized cross-over trial investigated these two teaching methods in separate training sessions. Students first learned from the ICEMS or an in-person expert instructor before crossing over to receive the other instructional method.[93] Trainees who first received AI instruction followed by expert instruction showed a decline in their surgical performance after the two sessions, while the surgical performance of those taught by an expert

instructor before receiving AI instruction significantly improved.[93] This study explored the effect of the subsequent application of these instructional methods, but not their combination. The findings suggest that human expert instruction and AI instruction may each provide trainees with knowledge pertaining to different aspects of surgical expertise, pointing to the potential utility in combining these two teaching methods. In addition, a cohort study investigating the unintended effects of AI instruction on surgical performance showed that intelligent tutoring may lead to suboptimal outcomes in several efficiency-related metrics,[49] indicating that human experts may be necessary to contextualize the feedback provided by intelligent tutoring systems to create an optimal learning environment. A recent RCT from our center sought to combine the strengths of AI instruction with human expert instruction in a simulation environment. This investigation found that AI-augmented personalized instruction enhanced ICEMS scores and resulted in improved skill transfer to a more realistic simulated scenario compared to AI tutor instruction alone. These results emphasize the critical role that human educators play in AI-based surgical teaching.[50] However, this study did not investigate the effect of augmenting human instruction with AI-derived error data on the frequency of instructions provided and technical skill improvement. As such, we aim to assess the pedagogical impact of AI-augmented personalized instruction on the frequency of feedback instructions and on the results of these specific surgical instructions on trainee surgical performance using the NeuroVR simulation platform.

RATIONALE

Proper surgical technical skills are essential for safe operative procedures,[8–10,12,13] but they are often acquired in high-stakes, stressful environments that are not conducive to methods of deliberate practice.[6,13,30,31,34,40,55,56] This challenge in surgical resident training is exacerbated by a lack of objective, standardized assessments of performance within surgical residency program

curricula.[11,27,31] Virtual reality simulators combined with intelligent tutoring systems offer a promising alternative method for acquiring and assessing surgical skills, and are slowly becoming more common for teaching surgical trainees.[34–40] A previous study found that AI-augmented personalized instruction during simulated VR brain tumor resection procedures improved trainee surgical performance and skill transfer compared to AI tutor instruction alone.[50] However, this study focused on the ICEMS-derived performance scores and did not consider other aspects of performance. Therefore, the effect of this instructional methodology on the frequency of feedback instructions and on technical skill performance remains unknown.

THE STUDY HYPOTHESIS

Our primary hypothesis is that AI-augmented personalized instruction will lead to a significantly lower number of feedback instructions compared with AI tutor instruction. Our secondary hypothesis is that AI-augmented personalized instruction will result in superior technical skills compared with AI tutor instruction. These hypotheses are based on adult learning theories that highlight the importance of personalized learning and contextualization to optimize learning outcomes.[94–96]

THE STUDY OBJECTIVES

To the best of our knowledge, there are no previous investigations into the prospect of combining AI-derived quantitative data with human expert instruction on the frequency of feedback provided to a trainee, nor on the improvement in a trainee's technical skill performance, in a simulation environment.

Therefore, the first coprimary objective of this thesis is to determine the effect of providing human expert instructors with ICEMS quantitative performance data on the number of instructions received by each trainee during simulation training. The second coprimary objective

is to determine the effect of this same instructional methodology on trainee technical skill level, measured by their performance in various metrics recorded by the NeuroVR system.

THE STUDY

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

Authors: Vanja Davidovic, BHSc[1], Bianca Giglio, MSc[1], Abdulmajeed Albeloushi, MD[1,2], Ahmad Kh. Alhaj, MD[1,2], Mohamed Alhantoobi, MD, MSc[1,3], Rothaina Saeedi, MD[1,2], Sabrina Deraiche, DEC[1,4], Recai Yilmaz, MD, PhD[1,5], Trisha Tee, MSc[1,6], Ali M. Fazlollahi, MD, MSc[1,7], Matthew Ha, MSc[8], Abicumaran Uthamacumaran, BSc[1], Neevya Balasubramaniam, MD[1,7], José A. Correa, PhD[9], Rolando F. Del Maestro, MD, PhD[1]

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. Faculté de médecine, Université de Montréal, Pavillon Roger-Gaudry, 2900 Edouard Montpetit Blvd, Montreal, QC H3T 1J4
5. Children's National Medical Center, Division of Neurosurgery and Pediatrics, 111 Michigan Ave NW, Washington, D.C. 20010, United States of America

6. Florida International University Herbert Wertheim College of Medicine, 11200 SW 8th St AHC2, Miami, FL 33199, United States of America
7. Faculty of Medicine and Health Sciences, McGill University, 3605 Rue de la Montagne, Montreal, QC, Canada H3G 2M1
8. Department of Surgical and Interventional Sciences, McGill University, Montreal General Hospital, 1650 Cedar Avenue, T5-110, Montreal, QC H3G 1A4
9. Department of Mathematics and Statistics, McGill University, 805 Sherbrooke St W, Montreal, QC, Canada H3A 1Y2

Corresponding author: Vanja Davidovic, 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; e-mail: vanja.davidovic@mail.mcgill.ca.

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 *Journal of Surgical Education* on July 14, 2025.

HIGHLIGHTS

- Intelligent tutoring systems result in limited surgical skill improvement
- Learners acquire simulated surgical skills when taught by human instructors
- Feedback frequency decreases when personalized instruction is provided
- Instructions informed by artificial intelligence data improve surgical performance

ABSTRACT

Objective: To determine whether personalized feedback from a human instructor receiving artificial intelligence (AI) error data will result in reduced feedback frequency and improvement of surgical skill compared to AI instruction. We hypothesized that AI-augmented personalized instruction would result in reduced feedback frequency and improvement in technical skill.

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

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

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

Results: By the third repetition, the AI-augmented personalized instruction group received significantly fewer total instructions (incidence rate ratio [IRR], 1.50 [95% CI, 1.16 to 1.94] instructions; $P < .001$), and high aspirator force instructions (IRR, 1.71 [95% CI, 1.15 to 2.55] instructions; $P = .002$), compared to the second repetition. Compared to AI tutor instruction, AI-augmented personalized instruction resulted in a significantly lower rate of healthy tissue removal ($P = .01$), instrument tip separation distance (mean ratio, 1.25 [95% CI, 1.05 to 1.50] mm; $P = .008$), and aspirator force (mean ratio, 1.68 [95% CI, 1.23 to 2.31] N; $P < .001$). AI-augmented personalized instruction showed a significant improvement from baseline in all subsequent repetitions for all performance metrics.

Conclusions: This cohort study demonstrated that AI-augmented personalized instruction resulted in less frequent feedback and an improvement in simulated surgical skills.

KEYWORDS: artificial intelligence-augmented instruction; surgical simulation; surgical education; technical skill; neurosurgical virtual reality, performance metrics

INTRODUCTION

Mastery of surgical technical skill is essential to mitigate the risk of surgical errors.[8–10,12,13] The current pedagogical model for surgical training involves the constant interplay between the educator and the trainee in a dynamic operative environment.[60] These real-time communications rely on the subjective observations of human instructors for continuous assessment and immediate, personalized, actionable feedback to guide technical skill development and error mitigation.[34] This reliance on subjective, qualitative performance data highlights a lack of objective, standardized instructional methodologies and assessments of surgical trainee performance.[11,27,31,34] Intelligent tutoring systems utilizing artificial intelligence (AI) to provide personalized and adaptive instructions to learners may help overcome these limitations due to their capacity to process and analyze large quantities of data to objectively assess performance.[42–47]

Intelligent tutoring systems have shown potential in teaching trainees surgical techniques and evaluating their competency using a data-driven approach in simulation environments.[42,43,97] A randomized controlled trial (RCT) utilizing the Virtual Operative Assistant intelligent tutoring system, employing only post-hoc AI feedback, significantly improved simulated surgical performance.[43,46] This system lacks the capacity to continuously monitor intraoperative skills or provide real-time feedback. The Intelligent Continuous Expertise Monitoring System (ICEMS) is a multi-algorithm AI system specifically designed to address these issues by employing quantitative data to continuously assess trainee performance and provide instructions to mitigate and reduce trainee errors based on real-time risk detection.[47] Developed using a long short-term memory network and based on objective, AI-derived metrics, the ICEMS can be used to detect

errors in surgical performance.[47] The ICEMS was trained on neurosurgeons' (experts) and medical students' (novices) operative data and demonstrated a granular differentiation across levels of expertise, and has shown face, content, construct, and predictive validity.[47,98] The NeuroVR, a high-fidelity virtual reality (VR) surgical simulator equipped with haptic feedback for brain tumor resection procedures, was used to develop the ICEMS.[72] The ICEMS can be applied to any simulation system.[47]

An RCT demonstrated that the ICEMS improved simulated surgical performance more than skilled instructors, indicating the pedagogical utility of the system.[42] Another crossover RCT found that trainee performance was significantly improved when instructed by a skilled educator first and then followed by ICEMS instruction.[97] Although this intelligent tutoring system can provide objective feedback, it is limited to delivering specific verbal instructions, while human educators can provide context and personalize their feedback. In a previous cohort study, this limited variety of possible feedback instructions led to unintended outcomes in an AI-enhanced curriculum, which negatively impacted trainee performance efficiency.[49] The results of these studies suggest that combining a skilled instructor and an AI tutor would allow for the contextualization of AI error data and optimize trainee performance. A recent RCT from our center found that AI-augmented personalized instruction resulted in enhanced ICEMS scores on a simulated subpial brain tumor practice resection scenario compared to AI tutor instruction and scripted human instruction, along with an improved transfer of surgical technical skills to a realistic simulated scenario.[50] These results highlight that personalized expert instruction results in enhanced surgical performance and skill transfer compared with intelligent tutor instruction,

emphasizing the critical role of human engagement and contribution in artificial intelligence-based surgical training.

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

METHODS

Participants

We conducted a planned secondary analysis using retrospective data from a previous RCT involving 87 medical students at the Neurosurgical Simulation and Artificial Intelligence Learning Centre, McGill University, Montreal, Canada from March to September 2024.[50] Students were recruited for a single 90-minute surgical simulation session with no follow-up. Medical students enrolled in their preparatory, first, or second year at one of four Quebec institutions were considered eligible for the study. The exclusion criterion was previous experience with the NeuroVR, the VR simulator used in this study. A sample size calculation with a power of 0.9, an effect size of 0.3, an α error probability of 0.05, and a correlation among repeated measures of 0.5 resulted in a total of 87 participants, with 29 participants in each of

three groups. Each participant performed the same simulated procedure with a different instructional method. This study was approved by the McGill University Health Centre Research Ethics Board, Neurosciences-Psychiatry and was registered on ClinicalTrials.gov on February 16, 2024 (NCT06273579). All participants signed an approved informed consent form prior to commencing the study. This report follows the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)[99] guidelines for cohort studies and the Machine Learning to Assess Surgical Expertise (MLASE) checklist.[100]

Study Procedure and Simulation Session

Following voluntary enrollment, students were stratified according to year in medical school and block randomized to one of three intervention arms with a 1:1:1 allocation ratio. All participants received standardized written and verbal instructions outlining the use of the instruments, the goal of the task, and how the session would proceed. Students were blinded to the trial's purpose and assessment metrics. The study utilized the NeuroVR (CAE Healthcare, Montreal, Canada), a validated, high-fidelity VR neurosurgical simulator, on which participants performed simulated subpial brain tumor resection procedures.[72,73] The simulation tasks involved the use of an ultrasonic aspirator and bipolar forceps, each equipped with haptic feedback, to completely resect a simulated tumor while minimizing bleeding and damage to non-pathological tissue.[48,101] All participants completed six 5-minute practice subpial resection scenarios to assess their learning (Figure 1), followed by a 13-minute realistic scenario to assess skill transfer to a more complex procedure (Figure 2). Between each repetition, a rest period of five minutes was afforded to participants.[74,102]

Interventions

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

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

Group 1: AI Tutor Instruction

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

Group 2: Scripted Human Instruction

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

Group 3: AI-Augmented Personalized Instruction

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

Instructions were provided based on pre-selected metrics: healthy tissue injury risk, bleeding risk, instrument tip separation distance, bipolar forceps force, and ultrasonic aspirator force. The metrics followed a hierarchy, as employed in previous studies; if more than one error occurred simultaneously, instructions for the metric higher in the hierarchy would be prioritized.[42,47] The feedback instructions provided in groups 1 and 2 and the hierarchical order of these metrics are outlined in Table 1.

Performance Metric Extraction

During the second to fifth repetitions, the ICEMS recorded the number of instructions given for each ICEMS metric: healthy tissue injury risk, bleeding risk, high instrument tip separation

distance, high bipolar force, low bipolar force, and high aspirator force.[42,47] Following the completion of a repetition, the number of instructions given to a participant in total and for each metric was summed. The average number of instructions was calculated for each group for each formative repetition of the practice scenario. During every repetition, the NeuroVR recorded participant technical skill performance data in 20-millisecond increments (50 recordings per second; t = 20 ms), including rate of healthy tissue removal (mm^3/t), total volume of blood lost (mm^3), instrument tip separation distance in the 3D space (mm), force applied with bipolar forceps (N), and force applied with ultrasonic aspirator (N).[103] These performance metrics were selected based on their relation to the feedback instructions given during the session to assess their effectiveness. The raw data were collected, and an average of 4 of the technical skill performance metrics was calculated for each participant in each repetition. Only the final value of total blood volume lost was utilized for each repetition, rather than taking an average. AI tutor-automated feedback provision and data visualization were performed using MATLAB (The MathWorks Inc., Natick, Massachusetts, USA) release 2024b.

Outcome Measures

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

Statistical Analysis

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

Between-group comparisons of the mean values of the technical skill performance metrics at each repetition of the practice resection scenario were conducted using a two-way mixed model analysis of variance (ANOVA). Repeated measures ANOVA was used to investigate within-group differences of the mean values of the technical skill performance metrics at each repetition of the practice resection scenario. One-way ANOVA was used to compare the mean values of each technical skill performance metric in the realistic resection scenario. Assumption of errors of ANOVA models, including normality, homogeneity of variance, and the presence of possible outliers or influential observations were assessed by graphical examination of model residuals. Post-hoc pairwise comparisons of mean differences were adjusted using the Šidák method for between-group differences and the Bonferroni correction for within-group differences. When model residuals did not show evidence of having a Normal (Gaussian) distribution, a natural logarithmic transformation of the values was used as the model outcome to stabilize the

variability. A robust linear mixed model approach to the ANOVA was used when the assumption of homogeneity of variance or residuals was violated. In cases where the normality or homogeneity of variance assumptions were drastically violated, we used the Kruskal-Wallis test at each repetition for between-group analysis, followed by Dunn's test with Bonferroni correction for multiple comparisons, and the Friedman test was used for within-group analysis, followed by the Nemenyi test. Results are reported as estimated mean differences and 95% CI and, in cases where a log transformation was used, as estimated ratios of geometric means and 95% CI.

Data analysis was performed using R Statistical Software (v4.3.3; R Core Team 2024)[104] from February to May 2025. All codes were written by the authors. ANOVAs and Poisson GLMM were implemented using the *lme4* [105] and *glmmTMB* [106] R packages, respectively. The GLMM analysis of simulated residuals was implemented using the *DHARMa* R package.[107] The robust linear mixed model approach to ANOVA was done using the *robustlmm* R package.[108]

RESULTS

Eighty-eight medical students from four Quebec universities participated in the study. Participants were stratified according to year in medical school and block randomized to one of three groups. There were 31 students in the AI tutor instruction group (group 1), 29 in the scripted human instruction group (group 2), and 28 in the AI-augmented personalized instruction group (group 3). Due to technical issues that arose during the simulation session, data from one participant in group 1 were excluded from the analysis. Data from 87 participants (46 [53%]

women, 40 [46%] men, 1 [1%] unspecified; mean [SD] age, 22.7 [4.0] years) were available for analysis, including 522 practice scenarios and 87 realistic scenarios (Table 2).

Feedback Frequency Across Simulated Practice Subpial Resections

All groups began receiving instructions in the second repetition of the practice scenario. In total, over the second to fifth repetitions, group 1 received 1464 instructions, group 2 received 1183 instructions, and group 3 received 728 instructions. Figure 4 outlines the number of instructions received by the groups for each metric in each repetition of the practice scenario where feedback was given (repetitions 2 to 5). The incidence rates of instructions for each metric can be found in Table 3. Group 1 (378 instructions) received significantly more instructions in total compared to group 3 (262 instructions) by the second repetition (IRR, 1.52 [95% CI, 1.03 to 2.23] instructions; $P = .03$). Group 2 (308 instructions) received significantly more feedback instructions in total compared to group 3 (175 instructions) by the third repetition of the practice scenario (IRR, 1.74 [95% CI, 1.15 to 2.63] instructions; $P = .001$) (Figure 5A). Pertaining to bleeding risk instructions, group 1 (54, 46, and 61 instructions, respectively) received significantly more feedback compared to group 3 (17, 16, and 13 instructions, respectively) in the second (IRR, 5.55 [95% CI, 1.22 to 25.15] instructions; $P = .01$), fourth (IRR, 5.02 [95% CI, 1.09 to 23.19] instructions; $P = .03$), and fifth (IRR, 8.20 [95% CI, 1.73 to 38.77] instructions; $P = .001$) repetitions, and group 2 (46 and 49 instructions, respectively) received significantly more feedback than group 3 (17 and 13 instructions, respectively) in the second (IRR, 4.89 [95% CI, 1.06 to 22.53] instructions; $P = .04$) and fifth (IRR, 6.81 [95% CI, 1.41 to 32.80] instructions; $P = .006$) repetitions (Figure 5C). Group 1 (97 instructions) received significantly more

instructions relating to high bipolar force than group 3 (43 instructions) in the fourth repetition of the practice scenario (IRR, 2.13 [95% CI, 1.03 to 4.39] instructions; $P = .04$) (Figure 5E). By the fourth repetition of the practice scenario, groups 1 (113 instructions) and 2 (111 instructions) both received significantly more high aspirator force feedback instructions than group 3 (51 instructions; IRR, 1.99 [95% CI, 1.10 to 3.61] instructions; $P = .01$; and IRR, 2.13 [95% CI, 1.17 to 3.87] instructions; $P = .004$, respectively) (Figure 5G). Only group 3 received significantly fewer instructions across repetitions of the practice scenario. By the third repetition, this group received significantly fewer total instructions (175 instructions; IRR, 1.50 [95% CI, 1.16 to 1.94] instructions; $P < .001$) and instructions relating to aspirator force (70 instructions; IRR, 1.71 [95% CI, 1.15 to 2.55] instructions; $P = .002$) compared with the second repetition (262 and 120 instructions, respectively; Figure 5A and 5G).

Technical Skill Performance Across Simulated Practice Subpial Resections

Learning curves were assessed for the five technical skill performance metrics. The estimates for each metric analyzed using parametric statistical methods can be found in Table 4. No statistically significant differences were observed between the groups at baseline performance (first repetition) in all five performance metrics. Dunn's tests with Bonferroni correction indicated that group 3 demonstrated a significantly lower rate of healthy tissue removal compared to group 1 by the third ($P = .01$) repetition of the practice scenario, and that group 3 had significantly less bleeding than group 1 in the second ($P = .02$) and fourth ($P = .02$) repetitions of the practice scenario (Figure 6A-B). In addition, a pairwise test with Šidák adjustment indicated that group 3 demonstrated a significantly lower instrument tip separation

distance by the second repetition of the practice scenario compared with group 1 (mean ratio, 1.25 [95% CI, 1.05 to 1.50] mm; $P = .008$) and compared with group 2 in the third (mean ratio, 1.20 [95% CI, 1.00 to 1.44] mm; $P = .049$) and fourth (mean ratio, 1.22 [95% CI, 1.02 to 1.46] mm; $P = .03$) repetitions of the practice scenario. This same statistical test found that group 2 had a significantly lower instrument tip separation distance than group 1 during the fifth repetition (mean ratio, 1.33 [95% CI, 1.11 to 1.59] mm; $P < .001$; Figure 6C). No other statistically significant differences were observed between group 1 and 2 in the other performance metrics. A pairwise test with Šidák adjustment also indicated that, by the third repetition, group 3 used significantly less force with the ultrasonic aspirator than group 1 (mean ratio, 1.68 [95% CI, 1.23 to 2.31] N; $P < .001$) and group 2 (mean ratio, 1.50 [95% CI, 1.09 to 2.06] N; $P = .007$; Figure 6E). No statistically significant differences were found between the groups in the force applied using the bipolar forceps ($P > .05$), as indicated by a robust linear mixed model regression (Figure 6D). Compared to baseline performance, by the second repetition of the practice scenario, group 1 significantly decreased the distance between their instruments (mean difference, 2.28 [95% CI, 0.57 to 3.99] mm; $P = .001$). This finding was also observed for group 2 (mean difference, 3.47 [95% CI, 2.30 to 4.64] mm; $P < .001$) and group 3 (mean ratio, 1.55 [95% CI, 1.36 to 1.77] mm; $P < .001$; Figure 6C), as indicated by a pairwise test with Bonferroni adjustment. This same test found that, compared to baseline performance, groups 1, 2, and 3 significantly lowered the force applied with the bipolar forceps (mean difference, 0.08 [95% CI, 0.02 to 0.13] N; $P < .001$; mean difference, 0.13 [95% CI, 0.07 to 0.18] N; $P < .001$; and mean ratio, 1.38 [95% CI, 1.13 to 1.67] N; $P < .001$, respectively) by the second repetition (Figure 6D). Nemenyi test showed that group 3 also achieved a significantly lower rate of healthy tissue removal ($P < .001$) and volume of blood lost ($P < .001$) by the second repetition compared to

baseline, and a pairwise test with Bonferroni correction found that this group demonstrated a lower force applied with the ultrasonic aspirator (mean difference, 0.05 [95% CI, 0.02 to 0.07] N; $P < .001$) by the second repetition compared to baseline (Figure 6A-B, 6E). Other improvements from baseline performance and between specific trials are shown in Figure 6.

Technical Skill Transfer to the Simulated Realistic Subpial Resection

The estimates for each metric analyzed using parametric statistical methods can be found in Table 5. Following the completion of the realistic scenario, a pairwise test with Šidák adjustment indicated that group 3 applied significantly less force with the ultrasonic aspirator than group 1 (mean difference, 0.04 [95% CI, 0.01 to 0.07] N; $P = .01$) (Figure 7E). No other statistically significant differences were found between the groups.

DISCUSSION

To the authors' knowledge, this cohort study is the first investigation to demonstrate the pedagogical impact of AI-augmented personalized instruction on the frequency of feedback instructions and on the results of these specific surgical instructions on trainee surgical performance. A previous RCT used ICEMS scores to explore the effect of the three different instructional methods utilized in this investigation on trainee skill acquisition and skill transfer.[50] This study builds on this investigation, focusing on the frequency of instructions provided and their impact on changes in technical skill.

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

The AI-augmented personalized instruction group had significantly lower values compared to the AI tutor instruction group for both risk metrics assessed and two of the three coaching metrics, consistent with our second hypothesis. Except for instructions pertaining to bipolar force, all the instructions given aim to decrease values in the technical skill performance metrics (Table 1). The absence of significant differences in the bipolar force applied may be attributable to participants receiving instructions to increase or decrease the force applied with the bipolar forceps throughout the session. This may have allowed participants in all groups to learn the ideal amount of force application with this instrument. Employing a similar methodology for teaching the other metrics, where trainees are made aware of appropriate changes to performance as much as they are made aware of errors, has proven beneficial for surgical skill acquisition and may be an avenue to explore in future studies involving the ICEMS.[109]

The scripted human instruction group did not consistently receive fewer instructions or exhibit significantly better technical skill performance compared with the AI tutor instruction group. These groups received instructions using identical wording, suggesting that the instructions programmed into the ICEMS may not provide sufficient information to allow trainees to learn to consistently correct errors. This is further supported by the AI-augmented personalized instruction group's fewer instructions received and outperformance of both other groups in several metrics. Investigations into the instructions that elicited the most appropriate changes in performance are presently being conducted using a series of Large Language Models (LLMs) to further optimize both the ICEMS and human expert instructions to enhance performance outcomes.[110]

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

These findings have supported the hypothesis that providing skilled instructors with AI-generated error data to facilitate the provision of personalized, continuous, contextualized feedback improves learning in a simulated surgical environment. Further research is required to

determine whether these findings can be generalized to more realistic surgical settings, and such studies using an *ex vivo* animal model are currently underway.[75,111–113] This cohort study demonstrates the potential for AI-augmented personalized instruction to optimize trainee assessment, teaching, and error mitigation in the operating room environment, helping to lay the foundations for the development of future intelligent human operating rooms powered by AI technology.

LIMITATIONS

Intelligent tutoring systems cannot completely replicate the communication interchange between a surgical educator and a learner in complex human operating settings.[114] This study was carried out using a small sample of medical students in their preparatory, first, or second year, and findings cannot be generalized to senior medical students or surgical residents. However, the results of a series of simulation studies has demonstrated that using medical students with minimal surgical experience has provided valuable insights.[42,43,49,97] Investigations using neurosurgical residents, fellows, and neurosurgeons are in preparation involving *ex vivo* models, but the limited number of participants available may limit the ability of these studies to achieve sufficient power to detect statistically significant differences unless multiple teaching centers are involved.

CONCLUSION

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

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

THESIS SUMMARY

Contributions to Original Knowledge

This study contributes to the understanding of the best practices for surgical education, specifically for providing intraoperative instructions to surgical trainees, in the following way:

1. To our knowledge, this study is the first to investigate the effect of augmenting human instruction using AI-derived error data on the frequency of feedback instructions and on trainee technical skill development in a VR simulation environment.

Discussion

This thesis aimed to investigate the effect of AI-augmented personalized instructions on feedback frequency and trainee technical skill performance. We conducted a secondary analysis of a single-blinded parallel-design RCT involving 88 medical students from four Quebec universities. These students performed subpial brain tumor resection procedures on the NeuroVR while receiving instructions, and the instructional methodologies differed according to their group allocation. The number of feedback instructions received by each trainee, as well as their technical skill performance metrics during the operation, were measured for analysis.

The first coprimary objective of this thesis was to determine the effect of providing human expert instructors with ICEMS quantitative performance data on the number of instructions received by each trainee during simulation training. Of the three intervention groups, only the AI-augmented personalized instruction group received significantly fewer total instructions and instructions relating to high aspirator force between the second repetition and the three subsequent formative repetitions of the practice scenario. This suggests that the instructions were clear and more actionable throughout the session than the other two instructional methodologies (Figure 5A, 5G). Additionally, consistent with our first hypothesis, the AI-augmented personalized

instruction group received fewer total instructions compared to the AI tutor instruction and scripted human instruction groups (Figure 5A). Instructions were only provided following metric error detection by the ICEMS; thus, receiving less instructions meant that less errors were made, suggesting a better understanding of the feedback provided. The AI-augmented personalized instruction group also received significantly fewer total instructions in the second repetition, when participants first began receiving feedback, compared to the AI tutor instruction group. This suggests that the instructions were immediately more comprehensible and more effective in reducing errors (Figure 5A).

The second coprimary objective of this thesis was to determine the effect of AI-augmented personalized instruction on trainee technical skill level, using metrics recorded by the NeuroVR system. AI-augmented personalized instruction was the only group to exhibit consistent technical skill improvement from baseline. This group demonstrated improved performance in all five technical skill metrics in all subsequent repetitions of the practice scenario, further implying that AI-augmented personalized instructions led trainees to better understand how to fix their errors (Figure 6). Furthermore, AI-augmented personalized instruction resulted in lower technical skill performance metric values compared to AI tutor instruction for both risk assessment metrics (Figure 6A-B) and two of the three coaching metrics (Figure 6C, 6E), consistent with our second hypothesis. Besides instructions relating to bipolar force application, the instructions directed participants to achieve lower values in the technical skill performance metrics. Therefore, achieving lower values in these metrics suggests a better understanding of the feedback provided. There are two instructions associated with the bipolar force application metric – one notifying trainees when they apply too much force, and one when they apply too little – as deviances in either direction pose a significant risk to patient safety (Table 1).[47,48] In comparison to the other

metrics, where only high values trigger the provision of an instruction, the dual instructions relating to bipolar force may have resulted in trainees learning the ideal amount of force to apply, providing a possible explanation for the lack of significant differences between the three intervention groups for this metric. This finding points to the potential benefit of applying operant conditioning theories to the ICEMS teaching methods to improve technical skill development, which have already proven useful for surgical skill acquisition.[109]

Scripted human instruction did not consistently result in a reduced feedback frequency, nor in an improvement in technical skill performance compared to AI tutor instruction (Figure 5, 6). These groups received instructions in the same wording (Table 1), implying that the feedback provided by the ICEMS may need to be modified in order for trainees to exhibit the intended changes to their technical skill performance. This is further supported by the AI-augmented personalized instruction group's outperformance of both groups in the frequency of feedback instructions and in technical skill performance. A study using Large Language Models (LLMs) is currently underway to understand the specific instructions that elicited the most appropriate responses by trainees.[110]

A previous RCT demonstrated that feedback provided by an expert instructor based solely on the instructor's observations was inferior to AI tutor instruction in improving surgical performance.[42] The results of a cross-over RCT showed that performance significantly improved when trainees first received expert instruction followed by AI tutor instruction.[97] Additionally, the primary investigation of the RCT outlined in this thesis determined that AI-augmented personalized instruction resulted in a significantly enhanced surgical performance compared to both other groups.[50] The finding from these three studies, as well as the results in this thesis, support the notion that resident technical skill development in the operating room could benefit from

combining human instructors with real-time quantitative data from an intelligent system. Implementing these systems could improve technical skill acquisition and assessment, posing a great advantage to surgical residency programs.

Limitations

A sample of medical students in their preparatory, first, or second year from four Québec universities was used for this study. These findings therefore cannot be generalized to the target population (ie, surgical residents), nor can they be applied to trainees in other institutions. However, the low availability of surgical residents poses difficulties in achieving statistical power in RCTs. Additionally, the junior medical students in this sample are still early in their medical careers, meaning they have less clinical experience than their senior counterparts. While this means that the findings cannot be applied to trainees at the residency level, this is a favourable characteristic for an RCT as it reduces the chance of introducing confounding factors that could make participants less comparable to one another (eg, prior experience in a surgical rotation). Finally, several previous simulation studies have employed a sample of junior medical students and shown meaningful results.[42,43,49,97] Nevertheless, future studies should be conducted with trainees at the residency level to determine whether similar conclusions are made, as the aim of this study and similar investigations is to inform the use of simulation for technical skill development in surgical residency programs.

While the NeuroVR platform has been previously validated,[47,98] the replication of neurosurgical procedures and the handling of biological tissues can only be simulated by these VR systems; it is never exactly the same as real-life procedures.[71] As such, further investigations are required to determine whether these findings can be applied to more realistic operating room environments. One such study is currently underway, utilizing a more realistic *ex vivo* animal model.

Future Directions

As previously discussed, this investigation demonstrated that AI-augmented personalized instruction resulted in lower technical skill performance metric values, suggesting a better understanding of the instructions provided. However, it has not been determined whether these technical skills are consistent with an expert-level performance. Future studies may wish to compare the technical abilities of trainees taught by AI-augmented personalized instruction to those of expert surgeons using the NeuroVR platform.

The AI-augmented personalized instruction group achieved a lower feedback frequency and lower performance metric values for ultrasonic aspirator force application compared to both other groups in summative assessment repetition six and compared to the AI tutor instruction group in the realistic scenario (Figure 5G, 6E, 7E). These findings suggest that AI-augmented personalized instructions provided details that were essential to correcting errors in this metric but were missing from the other groups' instructions. As a potential starting point for optimizing the teaching capabilities of the ICEMS, future studies can employ LLMs to identify the most effective instructions relating to this metric and later expand this investigation to include the other metrics. These findings could be used to inform the teaching of surgical skills in a simulated environment using the ICEMS, in the human operating room with an expert instructor, or for the development of future intelligent human operating rooms.

Conclusion

In this cross-sectional cohort study, AI-augmented personalized instruction resulted in a reduced feedback frequency and an improvement in technical skill performance compared to AI tutor instruction. These results show the value of employing quantitative data for surgical training and assessment in surgical residency programs. The potential benefits of augmenting human

instruction with AI-derived error data to rectify deficits in performance are outlined, and further investigations are required to determine their transferability to the human operating room environment.

REFERENCES

1. Brand A, Allen L, Altman M, Hlava M, Scott J. Beyond authorship: attribution, contribution, collaboration, and credit. *Learn Publ*. 2015;28(2):151-155. doi:10.1087/20150211

2. Allen L, O'Connell A, Kiermer V. How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship. *Learn Publ*. 2019;32:71-74. doi:10.1002/leap.1210

3. Gawande AA, Thomas EJ, Zinner MJ, Brennan TA. The incidence and nature of surgical adverse events in Colorado and Utah in 1992. *Surgery*. 1999;126(1):66-75. doi:10.1067/msy.1999.98664

4. Nuland SB. Mistakes in the Operating Room - Error and Responsibility. *N Engl J Med*. 2004;351(13):1281-1283.

5. Institute of Medicine (US) Committee on Quality of Health Care in America. *To Err Is Human: Building a Safer Health System*. (Kohn LT, Corrigan JM, Donaldson MS, eds.). National Academies Press (US); 2000. Accessed May 8, 2025. http://www.ncbi.nlm.nih.gov/books/NBK225182/

6. Rohrich RJ. "See One, Do One, Teach One": An Old Adage with a New Twist: *Plast Reconstr Surg*. 2006;118(1):257-258. doi:10.1097/01.prs.0000233177.97881.85

7. Brennan TA, Leape LL, Laird NM, et al. Incidence of Adverse Events and Negligence in Hospitalized Patients: Results of the Harvard Medical Practice Study I. *N Engl J Med*. 1991;324(6):370-376. doi:10.1056/NEJM199102073240604

8. Rogers Jr 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

9. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical Skill and Complication Rates after Bariatric Surgery. *N Engl J Med*. 2013;369(15):1434-1442. doi:10.1056/NEJMsa1300625

10. Stulberg JJ, Huang R, Kreutzer L, et al. Association Between Surgeon Technical Skills and Patient Outcomes. *JAMA Surg*. 2020;155(10):960-968. doi:10.1001/jamasurg.2020.3007

11. Hamdorf JM, Hall JC. Acquiring surgical skills. *Br J Surg*. 2000;87(1):28-37. doi:10.1046/j.1365-2168.2000.01327.x

12. Gawande AA, Zinner MJ, Studdert DM, Brennan TA. Analysis of errors reported by surgeons at three teaching hospitals. *Surgery*. 2003;133(6):614-621. doi:10.1067/msy.2003.169

13. Marcus H, Vakharia V, Kirkman MA, Murphy M, Nandi D. Practice Makes Perfect? The Role of Simulation-Based Deliberate Practice and Script-Based Mental Rehearsal in the

Acquisition and Maintenance of Operative Neurosurgical Skills. *Neurosurgery*. 2013;72:A124-A130. doi:10.1227/NEU.0b013e318270d010

14. Agha RA, Fowler AJ, Sevdalis N. The role of non-technical skills in surgery. *Ann Med Surg*. 2015;4(4):422. doi:10.1016/j.amsu.2015.10.006

15. 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

16. 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: A Guide for Academic Development and Teaching Excellence*. Springer, Cham; 2018:3-13. doi:10.1007/978-3-319-64728-9_1

17. Moulton C anne E, Regehr G, Mylopoulos M, MacRae H. Slowing Down When You Should: A New Model of Expert Judgment. *Acad Med*. 2007;82(10):S109-S116.

18. Moulton C anne, Regehr G, Lingard L, Merritt C, MacRae H. 'Slowing Down When You Should': Initiators and Influences of the Transition from the Routine to the Effortful. *J Gastrointest Surg*. 2010;14(6):1019-1026. doi:10.1007/s11605-010-1178-y

19. Ross KG, Lussier JW, Klein G. From the Recognition Primed Decision Model to Training. In: Betsch T, Haberstroh S, eds. *The Routines of Decision Making*. 1st ed. Psychology Press; 2004:424.

20. McQuillan P, Pilkington S, Allan A, et al. Confidential inquiry into quality of care before admission to intensive care. *Br Med J*. 1998;316(7148):1853-1858. doi:10.1136/bmj.316.7148.1853

21. Lighthall GK, Barr J, Howard SK, et al. Use of a fully simulated intensive care unit environment for critical event management training for internal medicine residents. *Crit Care Med*. 2003;31(10):2437-2443. doi:10.1097/01.CCM.0000089645.94121.42

22. 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

23. Dewan MC, Rattani A, Fieggen G, et al. Global neurosurgery: the current capacity and deficit in the provision of essential neurosurgical care. Executive Summary of the Global Neurosurgery Initiative at the Program in Global Surgery and Social Change. *J Neurosurg*. 2018;130(4):1055-1064. doi:10.3171/2017.11.JNS171500

24. The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada - ProQuest. Accessed May 7, 2025. https://www.proquest.com/docview/205022166?OpenUrlRefId=info:xri/sid:wcdiscovery&accountid=12339&sourcetype=Scholarly%20Journals

25. Bracco D, Favre JB, Bissonnette B, et al. Human errors in a multidisciplinary intensive care unit: a 1-year prospective study. *Intensive Care Med*. 2001;27:137-145. doi:https://doi-org.proxy3.library.mcgill.ca/10.1007/s001340000751

26. Bernstein M, Parrent AG. Complications of CT-guided stereotactic biopsy of intra-axial brain lesions. *J Neurosurg*. 1994;81(2):165-168.

27. Mirza M, Koenig JF. Teaching in the Operating Room. In: Köhler TS, Schwartz B, eds. *Surgeons as Educators: A Guide for Academic Development and Teaching Excellence*. Springer, Cham; 2018:137-160. doi:10.1007/978-3-319-64728-9_8

28. Madani A, Vassiliou MC, Watanabe Y, et al. What Are the Principles That Guide Behaviors in the Operating Room?: Creating a Framework to Define and Measure Performance. *Ann Surg*. 2017;265(2):255-267. doi:10.1097/SLA.0000000000001962

29. Kotsis SV, Chung KC. 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

30. Rajaratnam V, Rahman N, Dong C. Integrating instructional design principles into surgical skills training models: an innovative approach. *Ann R Coll Surg Engl*. 2021;103(10):718-724. doi:10.1308/rcsann.2020.7132

31. Curry JI. 'See one, practise on a simulator, do one' – the mantra of the modern surgeon. *S Afr J Surg*. 2011;49(1):4-6.

32. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997;84(2):273-278. doi:10.1046/j.1365-2168.1997.02502.x

33. Anderson DD, Long S, Thomas GW, Putnam MD, Bechtold JE, Karam MD. Objective Structured Assessments of Technical Skills (OSATS) Does Not Assess the Quality of the Surgical Result Effectively. *Clin Orthop*. 2016;474(4):874-881. doi:10.1007/s11999-015-4603-4

34. George EI, Skinner A, Pugh CM, Brand TC. Performance Assessment in Minimally Invasive Surgery. In: Köhler TS, Schwartz B, eds. *Surgeons as Educators: A Guide for Academic Development and Teaching Excellence*. Springer, Cham; 2018:53-91. doi:10.1007/978-3-319-64728-9_5

35. McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Does Simulation-Based Medical Education With Deliberate Practice Yield Better Results Than Traditional Clinical Education? A Meta-Analytic Comparative Review of the Evidence: *Acad Med*. 2011;86(6):706-711. doi:10.1097/ACM.0b013e318217e119

36. Issenberg SB, McGaghie WC, Hart IR, et al. Simulation Technology for Health Care Professional Skills Training and Assessment. *JAMA*. 1999;282(9):861-866. doi:10.1001/jama.282.9.861

37. Holmboe ES, Batalden P. Achieving the Desired Transformation: Thoughts on Next Steps for Outcomes-Based Medical Education. *Acad Med*. 2015;90(9):1215-1223. doi:10.1097/ACM.0000000000000779

38. Ahlberg G, Enochsson L, Gallagher AG, et al. Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. *Am J Surg*. 2007;193(6):797-804. doi:10.1016/j.amjsurg.2006.06.050

39. Barry Issenberg S, Mcgaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teach*. 2005;27(1):10-28. doi:10.1080/01421590500046924

40. Reznick RK, MacRae H. Teaching Surgical Skills — Changes in the Wind. *N Engl J Med*. 2006;355(25):2664-2669. doi:10.1056/NEJMra054785

41. Rogers MP, DeSantis AJ, Janjua H, Barry TM, Kuo PC. The future surgical training paradigm: Virtual reality and machine learning in surgical education. *Surgery*. 2021;169(5):1250-1252. doi:10.1016/j.surg.2020.09.040

42. 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

43. Fazlollahi AM, Bakhaidar M, Alsayegh A, et al. Effect of Artificial Intelligence Tutoring vs Expert Instruction on Learning Simulated Surgical Skills Among Medical Students: A Randomized Clinical Trial. *JAMA Netw Open*. 2022;5(2):e2149008. doi:10.1001/jamanetworkopen.2021.49008

44. 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.compbiomed.2024.108809

45. 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.compbiomed.2022.106286

46. Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Del Maestro RF. The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLoS ONE*. 2020;15(2):e0229596. doi:10.1371/journal.pone.0229596

47. 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

48. 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

49. 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

50. Giglio B, Albeloushi A, Alhaj AKh, et al. Artificial Intelligence–Augmented Human Instruction and Surgical Simulation Performance: A Randomized Clinical Trial. *JAMA Surg*. Published online August 6, 2025. doi:10.1001/jamasurg.2025.2564

51. America I of M (US) C on Q of HC in. Errors in Health Care: A Leading Cause of Death and Injury. In: Kohn LT, Corrigan JM, Donaldson MS, eds. *To Err Is Human: Building a Safer Health System*. National Academies Press (US); 2000. Accessed May 9, 2025. https://www.ncbi.nlm.nih.gov/books/NBK225187/

52. Smith CC, Gordon CE, Feller-Kopman D, et al. Creation of an Innovative Inpatient Medical Procedure Service and a Method to Evaluate House Staff Competency. *J Gen Intern Med*. 2004;19(5 Pt 2):510-513. doi:10.1111/j.1525-1497.2004.30161.x

53. Mason WTM, Strike PW. See one, do one, teach one—is this still how it works? A comparison of the medical and nursing professions in the teaching of practical procedures. *Med Teach*. 2003;25(6):664-666. doi:10.1080/01421590310001605705

54. Cameron JL. William Stewart Halsted. Our surgical heritage. *Ann Surg*. 1997;225(5):445-458. doi:10.1097/00000658-199705000-00002

55. Anders Ericsson K. Deliberate Practice and Acquisition of Expert Performance: A General Overview. *Acad Emerg Med*. 2008;15(11):988-994. doi:10.1111/j.1553-2712.2008.00227.x

56. Sroka G, Feldman LS, Vassiliou MC, Kaneva PA, Fayez R, Fried GM. Fundamentals of Laparoscopic Surgery simulator training to proficiency improves laparoscopic performance in the operating room—a randomized controlled trial. *Am J Surg*. 2010;199(1):115-120. doi:10.1016/j.amjsurg.2009.07.035

57. Frank JR, Karpinski J, Sherbino J, et al. Competence By Design: a transformational national model of time-variable competency-based postgraduate medical education. *Perspect Med Educ*. 2024;13(1):201-223. doi:10.5334/pme.1096

58. Richardson D, Landreville JM, Trier J, et al. Coaching in Competence by Design: A New Model of Coaching in the Moment and Coaching Over Time to Support Large Scale Implementation. *Perspect Med Educ*. 2024;13(1):33-43. doi:10.5334/pme.959

59. Lockyer J, Armson H, Könings KD, et al. In-the-Moment Feedback and Coaching: Improving R2C2 for a New Context. *J Grad Med Educ*. 2020;12(1):27-35. doi:10.4300/JGME-D-19-00508.1

60. Bezemer J, Cope A, Kress G, Kneebone R. Holding the Scalpel: Achieving Surgical Care in a Learning Environment. *J Contemp Ethnogr*. 2013;43(1):38-63. doi:10.1177/0891241613485905

61. Vozenilek J, Huff JS, Reznek M, Gordon JA. See One, Do One, Teach One: Advanced Technology in Medical Education. *Acad Emerg Med*. 2004;11(11):1149-1154. doi:10.1197/j.aem.2004.08.003

62. Pakkasjärvi N, Anttila H, Pyhältö K. What are the learning objectives in surgical training – a systematic literature review of the surgical competence framework. *BMC Med Educ*. 2024;24:1-22. doi:10.1186/s12909-024-05068-z

63. Campbell DA, Henderson WG, Englesbe MJ, et al. Surgical Site Infection Prevention: The Importance of Operative Duration and Blood Transfusion—Results of the First American College of Surgeons–National Surgical Quality Improvement Program Best Practices Initiative. *J Am Coll Surg*. 2008;207(6):810-820. doi:10.1016/j.jamcollsurg.2008.08.018

64. Procter LD, Davenport DL, Bernard AC, Zwischenberger JB. General Surgical Operative Duration Is Associated with Increased Risk-Adjusted Infectious Complication Rates and Length of Hospital Stay. *J Am Coll Surg*. 2010;210(1):60-65.e2. doi:10.1016/j.jamcollsurg.2009.09.034

65. 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

66. Kirkman MA. Deliberate Practice, Domain-Specific Expertise, and Implications for Surgical Education in Current Climes. *J Surg Educ*. 2013;70(3):309-317. doi:10.1016/j.jsurg.2012.11.011

67. Tan SSY, Sarker SK. Simulation in surgery: a review. *Scott Med J*. 2011;56(2):104-109. doi:10.1258/smj.2011.011098

68. Lee AT. *Flight Simulation: Virtual Environments in Aviation*. Routledge; 2017. doi:10.4324/9781315255217

69. Pedowitz RA, Marsh JL. Motor skills training in orthopaedic surgery: a paradigm shift toward a simulation-based educational curriculum. *J Am Acad Orthop Surg*. 2012;20(7):407-409. doi:10.5435/JAAOS-20-07-407

70. Badash I, Burtt 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

71. Palter VN, Grantcharov TP. Simulation in surgical education. *Can Med Assoc J*. 2010;182(11):1191-1196. doi:10.1503/cmaj.091743

72. Delorme S, Laroche D, DiRaddo R, F. Del Maestro R. NeuroTouch: A Physics-Based Virtual Simulator for Cranial Microneurosurgery Training. *Oper Neurosurg*. 2012;71(suppl_1):32-42. doi:10.1227/NEU.0b013e318249c744

73. Alotaibi FE, AlZhrani GA, Mullah MAS, et al. Assessing Bimanual Performance in Brain Tumor Resection With NeuroTouch, a Virtual Reality Simulator. *Oper Neurosurg*. 2015;11(1):89-98. doi:10.1227/NEU.0000000000000631

74. Sabbagh AJ, Bajunaid KM, Alarifi N, et al. Roadmap for Developing Complex Virtual Reality Simulation Scenarios: Subpial Neurosurgical Tumor Resection Model. *World Neurosurg*. 2020;139:e220-e229. doi:10.1016/j.wneu.2020.03.187

75. Almansouri A, Abou Hamdan N, Yilmaz R, et al. Continuous Instrument Tracking in a Cerebral Corticectomy Ex Vivo Calf Brain Simulation Model: Face and Content Validation. *Oper Neurosurg*. 2024;27(1):106-113. doi:10.1227/ons.0000000000001044

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

77. Hebb AO, 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

78. Tobin S. Entrustable Professional Activities in Surgical Education. In: *Advancing Surgical Education*. Springer, Singapore; 2019:229-238. doi:10.1007/978-981-13-3128-2_21

79. ten Cate O. Nuts and Bolts of Entrustable Professional Activities. *J Grad Med Educ*. 2013;5(1):157-158. doi:10.4300/JGME-D-12-00380.1

80. Brian R, Rodriguez N, Zhou CJ, et al. "Doing well": Intraoperative entrustable professional activity assessments provided limited technical feedback. *Surg Open Sci*. 2024;18:93-97. doi:10.1016/j.sopen.2024.02.008

81. Tabish SA. Assessment Methods in Medical Education. *Int J Health Sci*. 2008;2(2):3-7.

82. Xu Y, Liu X, Cao X, et al. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*. 2021;2(4):100179. doi:10.1016/j.xinn.2021.100179

83. Pruneski JA, Pareek A, Kunze KN, et al. Supervised machine learning and associated algorithms: applications in orthopedic surgery. *Knee Surg Sports Traumatol Arthrosc Off J ESSKA*. 2023;31(4):1196-1202. doi:10.1007/s00167-022-07181-2

84. Mofidi R, Duff MD, Madhavan KK, Garden OJ, Parks RW. Identification of severe acute pancreatitis using an artificial neural network. *Surgery*. 2007;141(1):59-66. doi:10.1016/j.surg.2006.07.022

85. Monsalve-Torra A, Ruiz-Fernandez D, Marin-Alonso O, Soriano-Payá A, Camacho-Mackenzie J, Carreño-Jaimes M. Using machine learning methods for predicting inhospital

mortality in patients undergoing open repair of abdominal aortic aneurysm. *J Biomed Inform*. 2016;62:195-201. doi:10.1016/j.jbi.2016.07.007

86. De Perrot T, Hofmeister J, Burgermeister S, et al. Differentiating kidney stones from phleboliths in unenhanced low-dose computed tomography using radiomics and machine learning. *Eur Radiol*. 2019;29:4776-4782. doi:10.1007/s00330-019-6004-7

87. Graffy PM, Liu J, O'Connor S, Summers RM, Pickhardt PJ. Automated segmentation and quantification of aortic calcification at abdominal CT: application of a deep learning-based algorithm to a longitudinal screening cohort. *Abdom Radiol*. 2019;44:2921-2928. doi:10.1007/s00261-019-02014-2

88. Azarnoush H, Alzhrani G, Winkler-Schwartz A, et al. Neurosurgical virtual reality simulation metrics to assess psychomotor skills during brain tumor resection. *Int J Comput Assist Radiol Surg*. 2015;10(5):603-618. doi:10.1007/s11548-014-1091-z

89. Binkley CE, Green BP. Does Intraoperative Artificial Intelligence Decision Support Pose Ethical Issues? *JAMA Surg*. 2021;156(9):809-810. doi:10.1001/jamasurg.2021.2055

90. Bissonnette V, Mirchi N, Ledwos N, Alsideiri G, Winkler-Schwartz A, Del Maestro RF. Artificial Intelligence Distinguishes Surgical Training Levels in a Virtual Reality Spinal Task. *J Bone Jt Surg*. 2019;101(23):e127. doi:10.2106/JBJS.18.01197

91. Balakrishnan S, Dakua SP, El Ansari W, Aboumarzouk O, Al Ansari A. Novel applications of deep learning in surgical training. In: De Pablos PO, Zhang X, eds. *Artificial Intelligence, Big Data, Blockchain and 5G for the Digital Transformation of the Healthcare Industry*. Vol 3. Information Technologies in Healthcare Industry. Academic Press; 2024:301-320. doi:10.1016/B978-0-443-21598-8.00021-X

92. Ma W, Adesope OO, Nesbit JC, Liu Q. Intelligent tutoring systems and learning outcomes: A meta-analysis. *J Educ Psychol*. 2014;106(4):901-918. doi:10.1037/a0037123

93. Yilmaz R, Fazlollahi A, Alsayegh A, Bakhaidar M, Del Maestro R. 428 Artificial Intelligence Training Versus In-person Expert Training in Teaching Simulated Tumor Resection Skills - A Cross-Over Randomized Controlled Trial. *Neurosurgery*. 2024;70(Supplement_1):129. doi:10.1227/neu.0000000000002809_428

94. 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

95. Shemshack A, Spector JM. A systematic literature review of personalized learning terms. *Smart Learn Environ*. 2020;7(1):33. doi:10.1186/s40561-020-00140-9

96. Wozniak K. Personalized Learning for Adults: An Emerging Andragogy. In: Yu S, Ally M, Tsinakos A, eds. *Emerging Technologies and Pedagogies in the Curriculum*. Bridging Human and Machine: Future Education with Intelligence. Springer Singapore; 2020:185-198. doi:10.1007/978-981-15-0618-5_11

97. Yilmaz R, Fazlollahi A, Alsayegh A, Bakhaidar M, Del Maestro R. 428 Artificial Intelligence Training Versus In-person Expert Training in Teaching Simulated Tumor Resection Skills - A Cross-Over Randomized Controlled Trial. *Neurosurgery*. 2024;70(Supplement_1):129-130. doi:10.1227/neu.0000000000002809_428

98. Yilmaz R, Ledwos N, Sawaya R, et al. Nondominant Hand Skills Spatial and Psychomotor Analysis During a Complex Virtual Reality Neurosurgical Task-A Case Series Study. *Oper Neurosurg Hagerstown Md*. 2022;23(1):22-30. doi:10.1227/ons.0000000000000232

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

100. 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

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

102. Bugdadi A, Sawaya R, Bajunaid K, et al. Is Virtual Reality Surgical Performance Influenced by Force Feedback Device Utilized? *J Surg Educ*. 2019;76(1):262-273. doi:10.1016/j.jsurg.2018.06.012

103. Yilmaz R, Fazlollahi AM, Winkler-Schwartz A, et al. Effect of Feedback Modality on Simulated Surgical Skills Learning Using Automated Educational Systems– A Four-Arm Randomized Control Trial. *J Surg Educ*. 2024;81(2):275-287. doi:10.1016/j.jsurg.2023.11.001

104. R Core Team. R: A Language and Environment for Statistical Computing. Published online 2024. https://www.R-project.org/

105. Bates D, Maechler M, Bolker B, et al. lme4: Linear Mixed-Effects Models using “Eigen” and S4. Published online March 26, 2025. Accessed June 19, 2025. https://github.com/lme4/lme4/

106. Brooks M, Bolker B, Kristensen K, et al. glmmTMB: Generalized Linear Mixed Models using Template Model Builder. Published online April 2, 2025. Accessed June 19, 2025. https://github.com/glmmTMB/glmmTMB

107. Hartig F, Lohse L, leite M de S. DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models. Published online October 18, 2024. Accessed June 19, 2025. http://florianhartig.github.io/DHARMa/

108. Koller M. robustlmm: Robust Linear Mixed Effects Models. Published online May 26, 2025. Accessed June 19, 2025. https://github.com/kollerma/robustlmm

109. Levy IM, Pryor KW, McKeon TR. Is Teaching Simple Surgical Skills Using an Operant Learning Program More Effective Than Teaching by Demonstration? *Clin Orthop*. 2016;474:945-955. doi:10.1007/s11999-015-4555-8

110. Du Y, Chen K, Zhan Y, et al. LMT++: Adaptively Collaborating LLMs with Multi-specialized Teachers for Continual VQA in Robotic Surgical Videos. *IEEE Trans Med Imaging*. Published online June 20, 2025. doi:10.1109/TMI.2025.3581108

111. 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

112. 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

113. 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

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

FIGURES

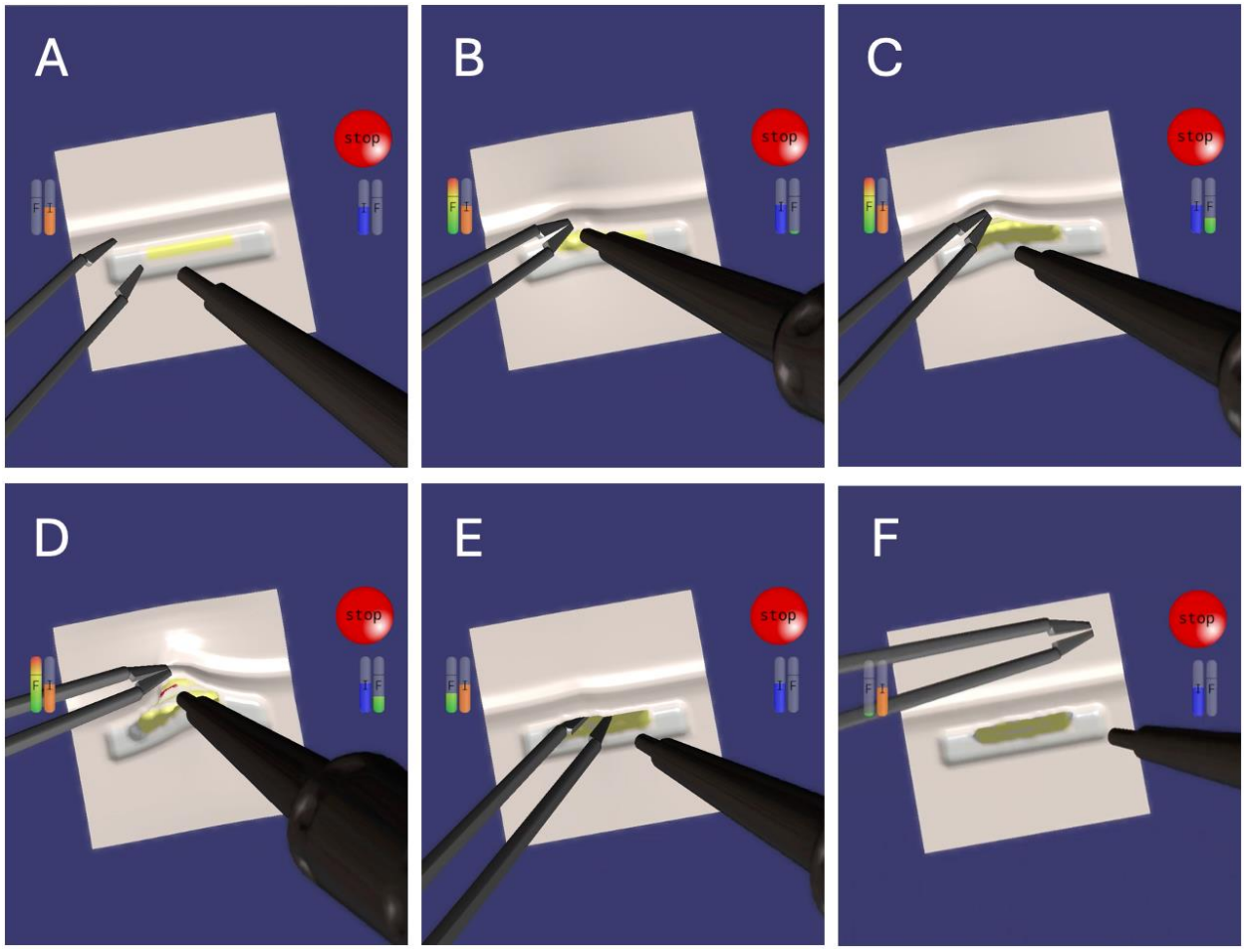

Figure 1. Practice subpial tumor resection scenario. (A) Start of the practice subpial tumor resection scenario. Yellow area represents the tumor and white area represents healthy brain tissue. Instrument on the left is the bipolar forceps, and instrument on the right is the ultrasonic aspirator. (B)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

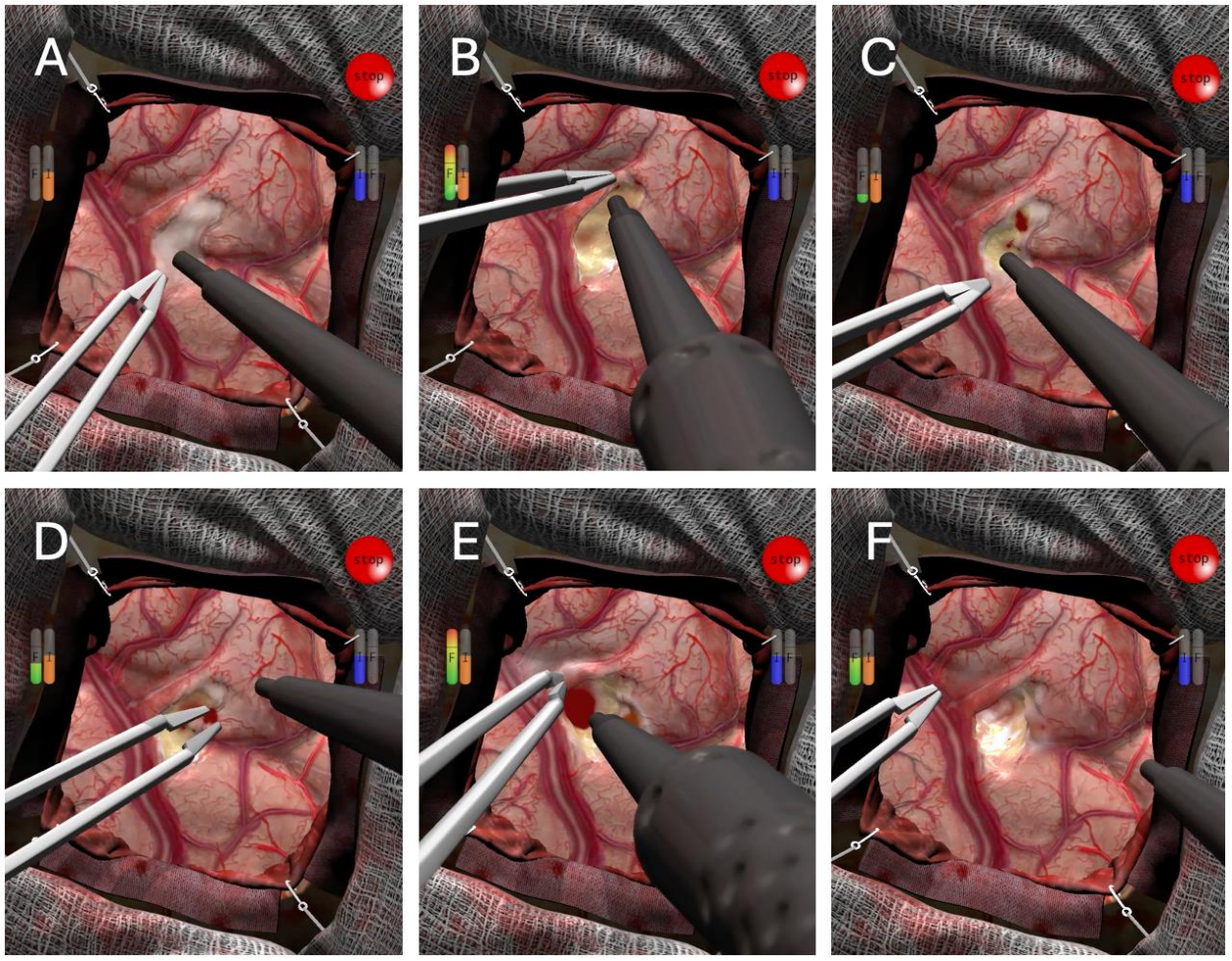

Figure 2. Realistic subpial tumor resection scenario. (A) Start of the realistic subpial tumor resection scenario. Off-white area represents the tumor and pink area represents healthy brain tissue. Instrument on the left is the bipolar forceps, and instrument on the right is the ultrasonic aspirator.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

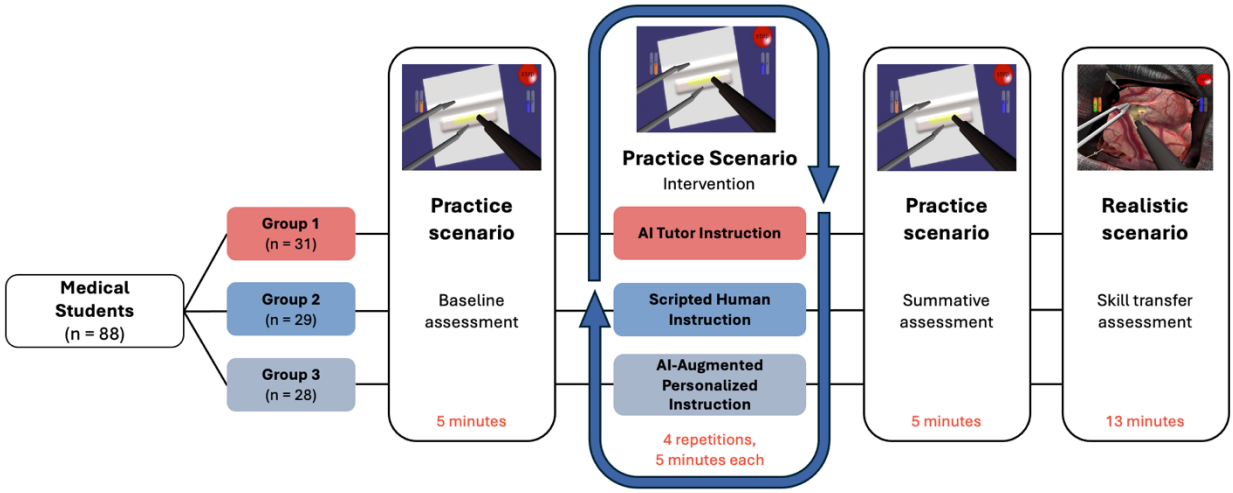

Figure 3. Flow diagram. Eighty-eight students were randomly allocated into 3 intervention groups. Abbreviation: AI, artificial intelligence.

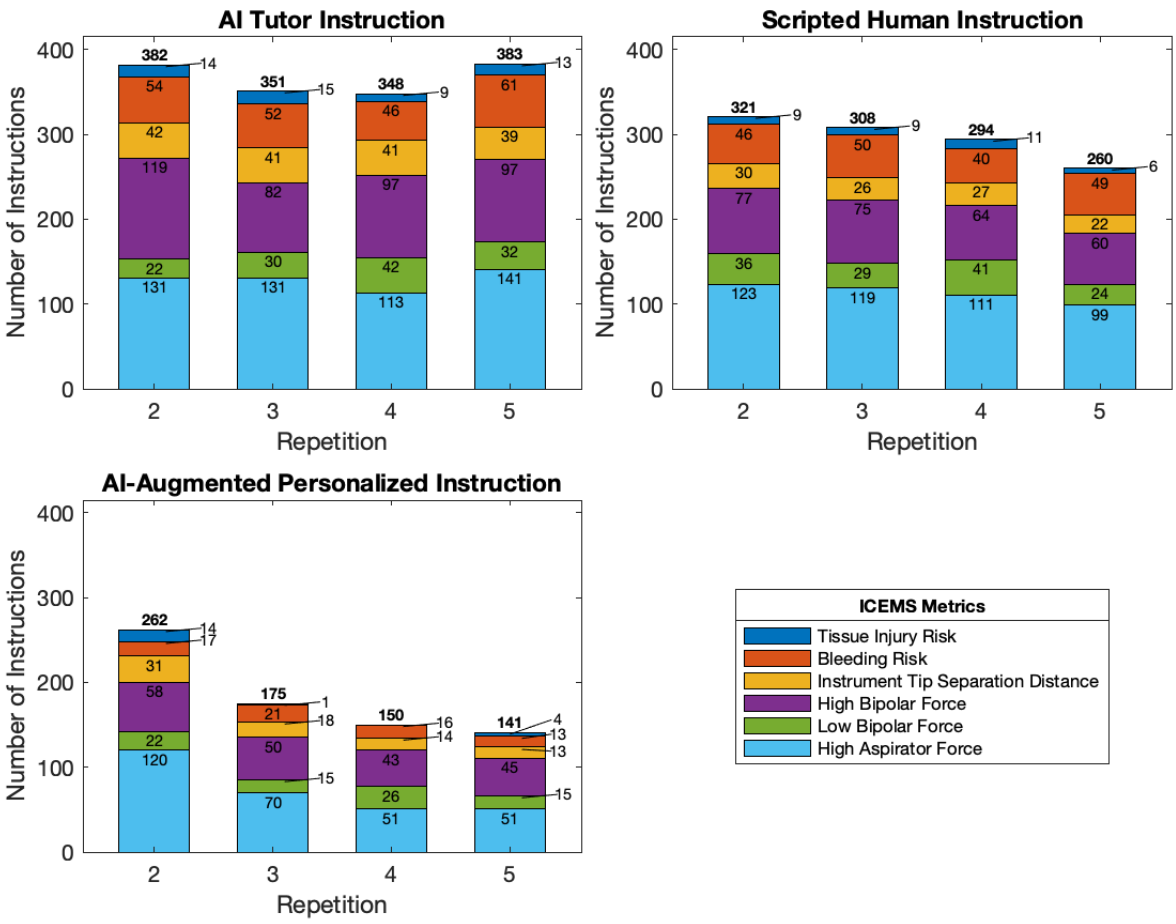

Figure 4. Number of instructions that each group received for each ICEMS metric across the second to fifth repetitions of the practice scenario. Metrics are color coded (see legend). X-axis represents the repetition number. Colored stacked bars represent the number of instructions for each ICEMS metric.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

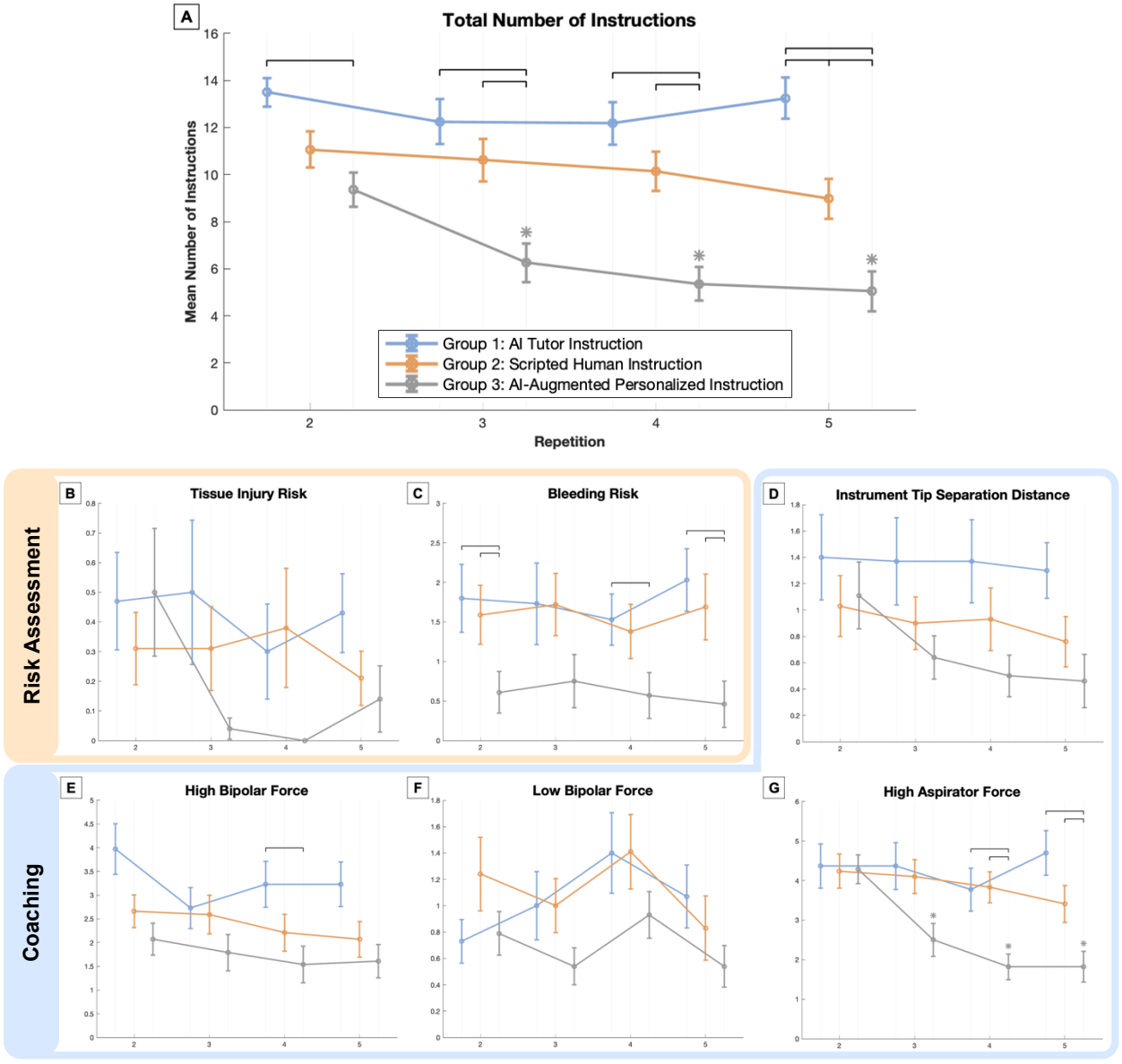

Figure 5. The mean total number of instructions and mean number of instructions in each ICEMS performance metric across the second to fifth repetitions of the practice scenario. Groups are color coded (see legend). The X-axis represents the repetition number. Points represent group means and error bars represent standard errors. Black horizontal brackets indicate statistically significant differences between groups ($P < .05$) during a given repetition. Asterisks indicate statistically significant differences from the baseline ($P < .05$) for that group. Abbreviations: AI, artificial intelligence; ICEMS, Intelligent Continuous Expertise Monitoring System.

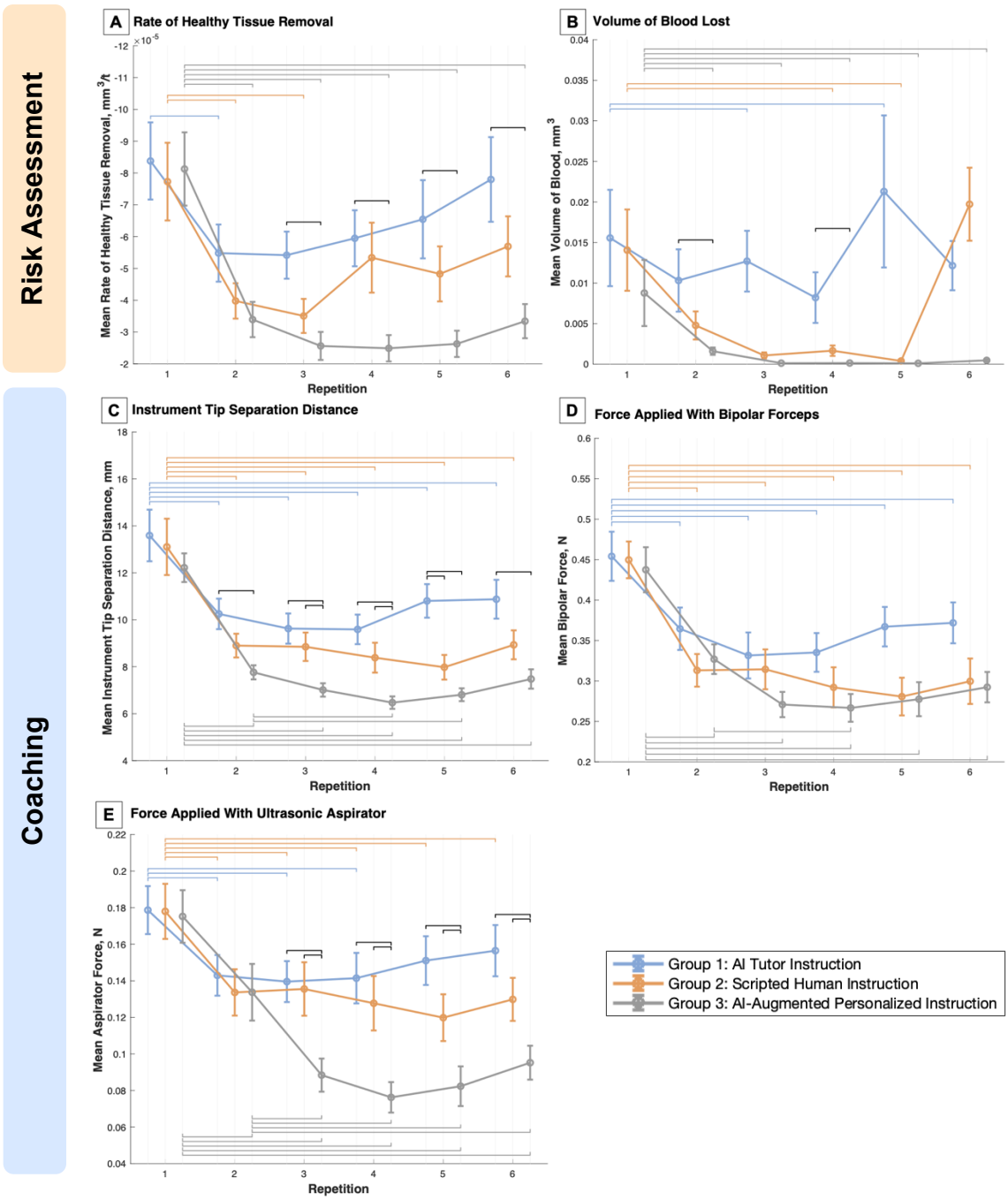

Figure 6. The learning curves of five technical skill performance metrics across six repetitions of the practice scenario. Groups are color coded (see legend). The X-axis represents the repetition number. Points represent group means and error bars represent standard errors. Since the simulator records the metrics at a frequency of 50 Hz, the unit of time (t) is equal to 20 ms. Black horizontal brackets indicate statistically significant differences between groups ($P < .05$) during a given repetition. Within-group differences are represented by horizontal brackets in the respective color for that group ($P < .05$). Abbreviations: AI, artificial intelligence.

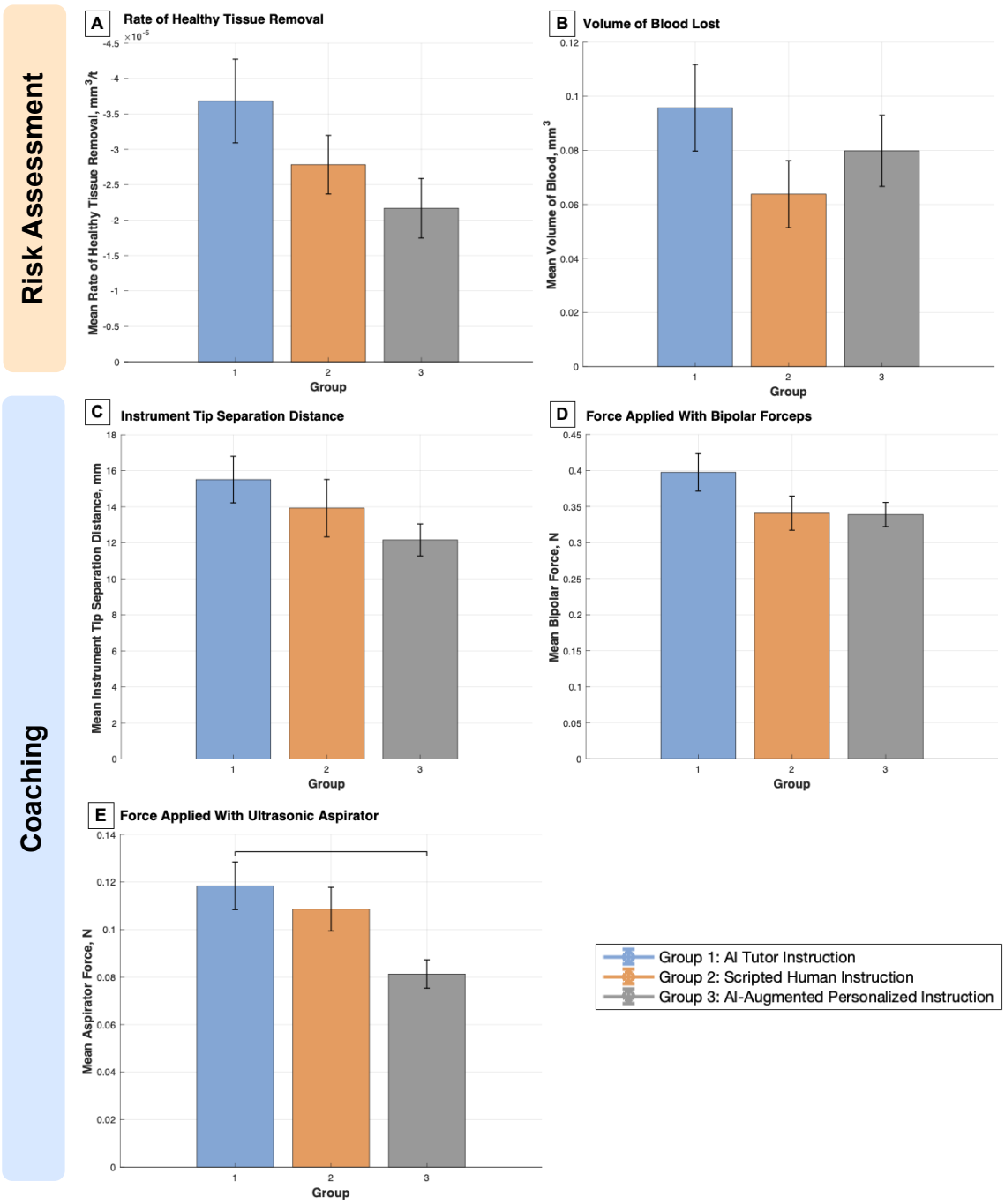

Figure 7. Five technical skill performance metrics during the realistic scenario. Groups are color coded (see legend). The X-axis represents the group. Colored bars represent group means and error bars represent standard errors. Since the simulator records the metrics at a frequency of 50 Hz, the unit of time (t) is equal to 20 ms. Black horizontal brackets indicate statistically significant differences between groups ($P < .05$). Abbreviations: AI, artificial intelligence.

TABLES

Metric	ICEMS Instruction
1. Healthy Tissue Injury Risk	"Try to avoid damaging the healthy brain surrounding the tumor."
2. Bleeding Risk	"Careful control of bleeding will improve your performance."
3. Instrument Tip Separation Distance	"Keeping your instruments closer together will improve your performance."
4. High Bipolar Force Application	"Try to decrease the amount of force you are applying with your bipolar."
5. Low Bipolar Force Application	"You can improve your performance by applying more force with your bipolar."
6. High Aspirator Force Application	"Try to decrease the amount of force you are applying with your aspirator."

Table 1. Metrics assessed by the ICEMS in hierarchical order, and their corresponding instructions.[47,a]

Abbreviation: ICEMS, Intelligent Continuous Expertise Monitoring System.

[a] The right column shows the instructions given to group 1 and 2 upon metric error detection by the ICEMS. If the ICEMS identifies more than one error simultaneously, it is programmed to provide instruction on the metric higher in the hierarchy.

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

Table 2. Demographic characteristics of included study participants.

Abbreviations: AI, artificial intelligence; VR, virtual reality

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

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

Table 3. Incidence rates and standard errors of instructions for six ICEMS metrics received during the second to fifth repetitions of the practice resection scenario.[a]

Abbreviation: ICEMS, Intelligent Continuous Expertise Monitoring System.

[a] Estimates are from the between-group statistical analyses of these ICEMS metrics.

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

Table 4. Estimated geometric means and standard errors of technical skill performance metrics over six repetitions of the practice resection scenario.[a]

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

[b] Estimated marginal mean (EMM) [SE].

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

Table 5. Estimated marginal means and standard errors of technical skill performance metrics during the realistic resection scenario.[a]

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