Development, Validation, and Testing of Artificial Intelligence Systems for Assessment and Training of Simulated Surgical Technical Skills

Muhammed Recai Yilmaz

Experimental Surgery Program, Department of Neurology and Neurosurgery,
Faculty of Medicine and Health Sciences, McGill University, Montreal, Quebec, Canada

January 2024

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Table of Contents

Abstract

In surgery, technical skills are of paramount importance. Since the subject is human life and well-being, poor surgical performance and techniques can result in high patient morbidity and mortality and costly clinical outcomes. Learning surgical skills is a lengthy and stressful endeavour where trainees need to engage in patient care as they are given increased responsibility under the supervision of expert surgical educators. This apprenticeship model faces challenges in outlining, assessing, and teaching the composites of surgical expertise in an objective and standardized way. To tackle this issue, new technologies and developments are being implemented in surgical education to establish a data-driven competency-based quantifiable framework.

Virtual reality surgical simulators are developed to realistically replicate a variety of surgical tasks, from simple to complex, while collecting vast amounts of data from the performance of the trainees and surgeons. This data utilized by artificial intelligence systems allows for accurate assessment of surgical performance, tailored feedback, and error mitigation assistance.

This thesis work encompasses the development, validation, and testing of a variety of feedback systems to demonstrate the utility of artificial intelligence-powered training systems for the assessment and teaching of bimanual surgical skills. Chapter 1 discusses the current needs in surgical education, existing teaching systems, and the increasing popularity of artificial intelligence applications in medical education. Chapter 2 outlines the development and predictive validation of the Intelligent Continuous Expertise Monitoring System (ICEMS), a multialgorithm artificial intelligence application with real-time surgical bimanual skill assessment, tailored

feedback, and risk detection ability. Chapter 3 provides a spatial analysis during a simulated tumor resection surgery to outline the importance of spatial awareness and feedback. Chapter 4 involves a randomized controlled trial comparing four feedback protocols, including no-feedback, to assess the efficacy of numeric, visual, and visuospatial feedback during simulation training. Finally, Chapter 5 outlines another randomized controlled trial to compare the efficacy of the real-time intelligence assistance provided by the ICEMS with in-person expert instruction in teaching simulated subpial tumor resection skills to demonstrate the future utility of artificial intelligence in real-time training.

Simulations equipped with artificial intelligence allow for a high-fidelity application in surgical education, providing accurate assessment and quantification of skills, tailored real-time feedback, and error mitigation. This development may shape the future of surgical training across all procedural medicine. I hope that the objective, standardized, and efficient training provided with these systems may be widely implemented in the future to help trainees master their skills and become more competent before performing real-life procedures, leading to improved patient outcomes.

Résumé

En chirurgie, les compétences techniques sont de la plus haute importance. Puisque le sujet est la vie et le bien-être humain, une mauvaise performance chirurgicale et des techniques inadéquates peuvent entraîner une morbidité et une mortalité élevées des patients ainsi que des résultats cliniques coûteux. L'apprentissage des compétences chirurgicales est un processus long et stressant où les stagiaires doivent s'engager dans les soins aux patients tout en se voyant confier des responsabilités accrues sous la supervision d'éducateurs chirurgicaux experts. Ce modèle d'apprentissage est confronté à des défis pour définir, évaluer et enseigner les composantes de l'expertise chirurgicale de manière objective et standardisée. Pour aborder cette question, de nouvelles technologies et développements sont mis en œuvre dans l'éducation chirurgicale pour établir un cadre quantifiable basé sur la compétence et piloté par les données.

Les simulateurs chirurgicaux en réalité virtuelle sont développés pour reproduire de manière réaliste une variété de tâches chirurgicales, des plus simples aux plus complexes, tout en collectant de vastes quantités de données sur la performance des stagiaires et des chirurgiens. Ces données utilisées par des systèmes d'intelligence artificielle permettent une évaluation précise de la performance chirurgicale, des retours personnalisés et une assistance à la mitigation des erreurs.

Ce travail de thèse englobe le développement, la validation et les tests d'une variété de systèmes de retour pour démontrer l'utilité des systèmes de formation alimentés par l'intelligence artificielle pour l'évaluation et l'enseignement des compétences chirurgicales bimanuelles. Le chapitre 1 examine les besoins actuels en éducation chirurgicale, des systèmes d'enseignement existants et de la popularité croissante des applications d'intelligence artificielle dans l'éducation

médicale. Le chapitre 2 décrit le développement et la validation prédictive du Système Intelligent de Surveillance Continue de l'Expertise (ICEMS), une application d'intelligence artificielle multi algorithme avec évaluation en temps réel des compétences chirurgicales bimanuelles, des retours personnalisés et une capacité de détection des risques. Le chapitre 3 fournit une analyse spatiale lors d'une chirurgie simulée de résection de tumeur pour souligner l'importance de la conscience spatiale et des retours. Le chapitre 4 implique un essai contrôlé randomisé comparant quatre protocoles de retour, y compris sans retour, pour évaluer l'efficacité des retours numériques, visuels et visuospatiaux lors de la formation en simulation. Enfin, le chapitre 5 décrit un autre essai contrôlé randomisé pour comparer l'efficacité de l'assistance en temps réel fournie par l'ICEMS avec l'instruction en personne par un expert dans l'enseignement des compétences de résection de tumeur sous-piale simulée pour démontrer l'utilité future de l'intelligence artificielle dans la formation en temps réel.

Les simulations équipées d'intelligence artificielle permettent une application d’haute-fidélité dans l'éducation chirurgicale, offrant une évaluation précise et une quantification des compétences, des retours personnalisés en temps réel et une mitigation des erreurs. Ce développement pourrait façonner l'avenir de la formation chirurgicale dans toute la médecine procédurale. J'espère que la formation objective, standardisée et efficace fournie par ces systèmes pourra être largement mise en œuvre à l'avenir pour aider les stagiaires à maîtriser leurs compétences et à devenir plus compétents avant de réaliser des procédures réelles, conduisant à de meilleurs résultats pour les patients.

Özet

Cerrahide teknik beceri büyük bir öneme sahiptir. Konu insan hayatı olduğundan, kötü cerrahi performans ve teknikleri yüksek hasta morbidite ve mortalitesi ile sonuçlanabilir ve maliyetli klinik sonuçlara neden olabilir. Cerrahi becerilerin öğrenilmesi uzun ve stresli bir çabadır. Cerrahi asistanlar, uzman cerrahi eğitmenlerinin gözetiminde artan sorumluluk altında cerrahi eğitimlerini alırlar. Bu usta-çırak modeli, cerrahide teknik yeterliliği objektif ve standart bir şekilde tanımlama, değerlendirme ve öğretme konularında yetersiz kalır. Bu sorunu ele almak için cerrahi teknik eğitiminde veri odaklı ve yetkinlik tabanlı ölçülebilir bir çerçeve oluşturmak için yeni teknolojiler ve gelişmeler uygulanmaktadır.

Sanal gerçeklik cerrahi simülatörleri çeşitli cerrahi operasyonları gerçekçi bir şekilde simüle ederken asistanların ve cerrahların performansından büyük miktarda veri toplamaktadır. Bu veriler yapay zeka sistemleri tarafından kullanıldığında cerrahi performansın değerlendirmesine, kişiye özel geri bildirimlere ve hata giderme yardımına olanak tanır.

Bu tez çalışması, bimanuel cerrahi becerilerin değerlendirilmesi ve öğretimi için yapay zeka destekli eğitim sistemlerinin faydalılığını göstermek için çeşitli geri bildirim sistemlerinin geliştirilmesi, onayı ve test edilmesini kapsamaktadır. Bölüm 1 cerrahi eğitimdeki mevcut ihtiyaçları, mevcut öğretim sistemlerini ve tıp eğitiminde yapay zeka uygulamalarının artan popülerliğini tartışmaktadır. Bölüm 2, Intelligent Continuous Expertise Monitoring System (ICEMS) adlı çoklu algoritma yapay zeka uygulamasının geliştirilmesi ve bu sistemin asistanların performansını tahmin etme yeteneğinin test edilmesini içerir. Bu sistem gerçek zamanlı cerrahi bimanuel beceri değerlendirmesi, kişiye özel geri bildirim ve risk tespit yeteneğine sahiptir. Bölüm 3, simülasyon ameliyatı sırasında uzaysal farkındalığın ve geri

bildirimin önemini açıklamak için üç boyutlu analiz sunar. Bölüm 4, simülasyon eğitimi sırasında sayısal, görsel ve görsel-uzaysal geri bildirimlerin verimliliğini değerlendirmek için geribildirim protokollerini içeren bir randomize kontrollü deney içerir. Son olarak, Bölüm 5, yapay zekanın gerçek zamanlı eğitimdeki gelecekteki kullanımını göstermek üzere başka bir randomize kontrollü deney içerir. Bu deney yüzeyel subpial tümör rezeksiyon becerilerini öğretmede insan ile yapay zekanın eğitmenler olarak karşılastılmasını içerdi. ICEMS tarafından sağlanan gerçek zamanlı yardım bir uzman kişi tarafından öğretilme ile karşılaştırıldı.

Yapay zeka ile donatılmış simülasyon sistemleri, cerrahi eğitimde kaliteli bir uygulama sunar ve becerilerin değerlendirmesi, kişiye özel geri bildirim ve hata engelleme gibi faydalar sağlayabilir. Bu gelişme, tıpta cerrahi eğitimin geleceğini şekillendirebilir ve asistanların gerçek hayatta cerrahide sorumluluk almadan çok daha önce becerilerini yeterli seviyeye ulaştırmalarına, ustalaşmalarına ve dolayısıyla ameliyat sırasında ve sonrasında daha iyi sonuçlara ulaşmalarına yardımcı olabilir.

Acknowledgements

This PhD work marks a time period of great changes in my life and is a collection of ‘what else can be done’. It is a journey of scientific discovery and personal growth from day one with the help of amazing individuals at the Neurosurgical Simulation and Artificial Intelligence Learning Centre. This journey was shaped around the trust and freedom I was provided by my supervisor, Dr. Rolando Del Maestro, and it has my answers to the question he asked from time to time ‘What else would you do if you had no limitations?’.

My journey started with a plan for ‘a short observership in a research lab’ which evolved into a master’s and, later, to this PhD project involving cutting-edge applications, patent filing, and publications in reputed journals. Dr. Del Maestro’s example, support, patience, and careful guidance was critical throughout. This path included my baby steps starting from scratch learning English while updating my supervisor as to what percentage I was able to understand what he and people in the lab say. I started with 20% and I vividly recall the joy that lit up his face when I said that I understood 99% of what was being said. The stories I had throughout this journey could fill a full book, but respecting the limits of this short section will only allow me to make small notes.

This work would not have been possible without the tremendous support provided by Dr. Del Maestro, my colleagues, and my supervisory committee at McGill University including my committee chair, Dr. Rahul Gawri and former chair, Dr. Hadil Al-Jallad, my co-supervisor, Dr. Carlo Santaguida, and members Dr. Gregory Berry, Dr. Jason Harley, Dr. Derek Rosenzweig, Dr. Mohammad Maleki, and Dr. Jeffery Hall. I was privileged to work closely with Dr. Alexander Winkler-Schwartz and have him both as a colleague and friend. Having his great example a few

steps ahead of me, and his insights and help were a relief at times when there was so much to learn.

Our lab had so many brilliant graduate students and researchers whose presence was important in creating an enriching learning environment, such as Nicole Ledwos, Dr. Nykan Mirchi, Dr. Vincent Bissonnette, Dan Huy Tran, Aiden Reich, Sommer Christie, Sharif Natheir, Ali Fazlollahi, Dr. Lucy Luo, Dr. Ahmad Alsayegh, Dr. Mohamad Bakhaidar, Nour Abou Hamdan, Dr. Abdulrahman Almansouri, Puja Pachchigar, Trisha Tee, Bianca Giglio and Vanja Davidovic. Working with them while witnessing their progress towards a bright future was a great pleasure. I would like to thank Dr. Jose Andres Correa, Sharon Turner, Angie Giannakopoulos, and many others at McGill University who provide critical assistance whenever the students need. I would also like to thank all the people at the Montreal Neurological Institute and Hospital for their service to humanity and for making this place a worldwide leader on many fronts.

I am grateful to my parents for their endeavors to raise me and my brothers Sezai and Zekai to fulfill our potential. Life took us on a path that we would never imagine. Some stories we had were so extraordinary that they could only be straight from books or movies. While I was starting a new life chapter in Canada, they did the same in Germany. My mother's life-long curiosity to learn and her and my father's efforts to quickly adapt to a new environment set the baseline of the family and the expectations from their children.

I thank Canada for opening its doors for me to meet the lovely people inside and for accepting me as a Canadian. I will carry the core memories I gathered in this special place wherever I go. I thank dear Pamela Miller, her son Minister of Crown, the Honorable Marc Miller, and the people in his office for making sure the progress of my path to become a

Canadian. I thank all countries that opened the doors for displaced people and welcomed them. I thank all displaced people who bring and adapt their enriching background to their new home and those who stand up for the truth no matter where they are.

I thank dear Pam Del Maestro for her care, proof-reading my writings from time to time, and sending her delicious muffins to the lab with Dr. Del Maestro. The life-long love they have for each other as a couple is an example of what younger generations can look for.

Finally, I appreciate spending a long time in the beautiful city of Montreal, with all its diversity and welcoming environment. I am greatly thankful to my friends and people in the Turkish community whose support and 'familyship' were very important to make Montreal feel like home while doing our best to thank and give back to the people of this city. While doing this PhD, I had the privilege of co-founding the Community Engagement Project (CEP), a non-profit organization that brings newcomers and students in Montreal together to help meet the needs of vulnerable populations, such as the homeless, in our city. I would like to extend my gratitude to all the cooks and volunteers of the CEP for reaching more than 1500 people over the last 2 years. Most importantly, I thank dear Sourour Harfouch, RD for her companionship and exceptional leadership to grow this meaningful initiative into what it is today.

I look forward to the opportunities to come in my lifetime to work with amazing people around me to dedicate our efforts to great causes both academic and humanitarian.

I dedicate this work to all displaced people who had to leave their homes and start a new life chapter in unfamiliar lands.

This work is dedicated to honoring the sacrifice from comfort of those who speak up for the truth no matter where they are, their resilience in the face of all challenges,
and their unwritten stories.

Contribution to Knowledge

In this work, Chapters 2, 3, and 4 outline original scholarships that have been published in peer-reviewed scientific journals. Chapter 5 outlines an original scholarship and is presently being peer-reviewed. These works have contributed to advancing methodologies and knowledge regarding surgical education as follows:

Chapter 2 constitutes the development and the predictive validation of the Intelligent Continuous Expertise Monitoring System (ICEMS), which is a multialgorithm artificial intelligence system. This system is a first-of-its-kind application with real-time assessment, tailored feedback, and risk detection abilities during simulated brain tumor resection tasks. It can be used to inform the surgeon about risks during the operation or teach surgery to a trainee while assessing their skill levels. The work in this chapter involved two main objectives, 1- the development of the ICEMS's three modules: performance assessment, feedback, and risk detection; and 2- the predictive validation of the first module on the performance of 26 neurosurgery residents. The ICEMS made a performance assessment in 0.2-second intervals during two simulated subpial tumor resection tasks and was able to differentiate between expertise levels and detect the trainee year in their neurosurgical training program based on their bimanual surgical skills. The efficacy of the second and third modules was tested in Chapter 6. This development shaped the research at our lab, the Neurosurgical Simulation and Artificial Intelligence Learning Centre in the coming years. I hope that it will also inspire applications in all procedural medicine where performance data is available.

Chapter 3 outlined a spatial analysis of non-dominant hand skills. Non-dominant hand skills are critical to increase efficiency in performance by assisting the dominant hand and

controlling bleeding. In a previous work published in 2019 in Jama Open Network by Winkler-Schwartz et. al., we assessed 6600 performance metrics during a simulated brain tumor resection task. The majority of significant differences were observed with metrics related to non-dominant hand between skilled and less-skilled participants, highlighting the importance of non-dominant hand skills in expertise. The work in this chapter demonstrated the differences between skilled and novice level performance in a three-dimensional space. Neurosurgeons used their bipolar more precisely, in contact with pia matter around the tumoral region. Skilled participants have a better spatial awareness with their instrument utilization, a skillset that trainees need to master.

Chapter 4 involved a four-parallel-arm randomized controlled trial to compare different post-hoc feedback methodologies: no-feedback, numerical, visual, and visuospatial feedback. Providing information with more engaging visual and visuospatial information resulted in higher performance scores while any type of feedback provided a better performance improvement in comparison to no feedback. This work demonstrated efficient methodologies in feedback delivery to maximize trainee skills acquisition in learning bimanual surgical skills during simulated tumor resections which will help shape the future paradigm in simulation training.

Chapter 5 outlined a three-parallel-arm randomized controlled trial to compare the efficiency of real-time and post-hoc intelligent feedback to in-person human expert-mediated instruction in teaching simulated surgical skills. Real-time feedback provided by the ICEMS achieved significantly better learning outcomes in performance scores, no-significantly different scores when rated by blinded experts using Objective Structures Assessment of Technical Skills (OSATS) rating, and significantly higher cognitive extraneous load. Real-time intelligent tailored feedback may provide more efficient training than learning in-person with expert instructors with

actionable and tailored instructions in a patient risk-free environment, saving experts' time and improving learning outcomes. This work demonstrated the testing of the remaining modules outlined in Chapter 2 related to ICEMS's feedback delivery and risk detection.

Contribution of Authors

Chapter 2: Continuous Monitoring of Surgical Bimanual Expertise Using Deep Neural Networks in Virtual Reality Simulation

R. Yilmaz: Conceptualization, Acquisition and interpretation of data, Methodology, Writing - Original Draft, Critical revision of the manuscript for important intellectual content, Applied machine learning, Statistical analysis, Visualization, **A. Winkler-Schwartz:** Conceptualization, Acquisition and interpretation of data, Critical revision of the manuscript, Statistical analysis, **N. Mirchi:** Conceptualization, Acquisition and interpretation of data, Critical revision of the manuscript, **A. Reich, S. Christie, N. Ledwos, and A. M. Fazlollahi:** Conceptualization, Critical revision of the manuscript . **D. Huy Tran:** Conceptualization, Critical revision of the manuscript, Visualization, **C. Santaguida**: Conceptualization, Critical revision of the manuscript, Funding, Supervision, **A. J. Sabbagh, K. Bajunaid:** Acquisition and interpretation of data, Critical revision of the manuscript, **R. Del Maestro:** Conceptualization, Acquisition and interpretation of data, Methodology, Writing - Original Draft, Critical revision of the manuscript, Funding, Supervision.

Chapter 3: Nondominant Hand Skills Spatial and Psychomotor Analysis During a Complex Virtual Reality Neurosurgical Task-A Case Series Study

R. Yilmaz: Conceptualization, Acquisition and interpretation of data, Methodology, Writing - Original Draft, Critical revision of the manuscript, Statistical analysis, Visualization, **N. Ledwos:** Conceptualization, Writing - Original Draft, Critical revision of the manuscript, **R. Sawaya:** Conceptualization, Methodology, Critical revision of the manuscript, Visualization, **A. Winkler-**

Schwartz: Conceptualization, Acquisition and interpretation of data, Methodology, Critical revision of the manuscript, **N. Mirchi, V. Bissonnette, A. M. Fazlollahi, M. Bakhaidar, A. Alsayegh, A. J. Sabbagh, K. Bajunaid:** Conceptualization, Critical revision of the manuscript, **R. Del Maestro:** Conceptualization, Methodology, Writing - Original Draft, Critical revision of the manuscript, Supervision, Administration, Funding.

Chapter 4: Effect of Feedback Modality on Simulated Surgical Skills Learning using Automated Educational Systems– A Four-Arm Randomized Control Trial

R. Yilmaz: Conceptualization, Data acquisition, Methodology, Data analysis, Writing the original draft, Critical revision of the manuscript, Statistical analysis, Visualization, Coding. **A. M. Fazlollahi:** Conceptualization, Data acquisition, Critical revision of the manuscript, **A. Winkler-Schwartz:** Conceptualization, Statistical analysis, Critical revision of the manuscript, **A. Wang:** Conceptualization, Pilot study, Critical revision of the manuscript, **H. H. Makhani:** Conceptualization, Pilot study, Data analysis, Critical revision of the manuscript, **M. Bakhaidar, A. Alsayegh, D. Huy Tran:** Conceptualization, Critical revision of the manuscript, **C. Santaguida:** Conceptualization of the study, Critical revision of the manuscript, Funding, and Supervision, **R. F. Del Maestro:** Conceptualization, Methodology, Writing the original draft, Critical revision of the manuscript, Funding, Administration, Supervision.

Chapter 5: Surgical Skills Training Using Real-Time Artificial Intelligence vs Human Instruction – A Randomized Controlled Trial

R. Yilmaz: Conceptualization, Participant recruitment, Methodology, Development and implementation of the intelligent system, Video feedback, Data analysis, Writing the original draft, Critical revision of the manuscript, Statistical analysis, All codes used in this study, **M.**

Bakhaidar, A. Alsayegh, Conducting in-person training, Critical revision of the manuscript, Blinded expert rating, **N. Abou Hamdan:** Conceptualization, Participant recruitment, Critical revision of the manuscript, **A. M. Fazlollahi:** Conceptualization, Methodology, Writing the original draft, Participant recruitment, Critical revision of the manuscript, **I. Langleben:** Conceptualization, Participant recruitment, Critical revision of the manuscript, **T. Tee:** Methodology, Statistical analysis, Writing the original draft, Critical revision of the manuscript, **A. Winkler-Schwartz:** Development of the feedback systems, Critical revision of the manuscript, **D. Laroche:** Real-time data transfer from the simulator, Technical assistance to ensure the proper work of the simulator, **C. Santaguida:** Conceptualization, Critical revision of the manuscript, Funding, Supervision, **R. Del Maestro:** Conceptualization, Methodology, Development, Writing the original draft, Critical revision of the manuscript, Funding, Administration, Supervision.

Funding

I received a grant for Doctoral studies from the Fonds de recherche du Quebec–Sante, a Medical Education Research Grant from the Royal College Physicians and Surgeons of Canada, a Max Binz Fellowship from McGill University Internal Studentships, and a Brain Tumour Research Grant from the Brain Tumour Foundation of Canada. The work at the Neurosurgical Simulation and Artificial Intelligence Learning Centre is supported by the Franco Di Giovanni Foundation, the Montreal English School Board, AO Foundation, and the Montreal Neurological Institute and Hospital. The National Research Council of Canada, Boucherville, Quebec, Canada provided a prototype of the NeuroVR and technical assistance which made this work possible.

List of Abbreviations

AI: Artificial intelligence

VR: Virtual reality

RCT: Randomized controlled trial

OSATS: Objective Structured Assessment of Technical Skills

ICEMS: Intelligent Continuous Expertise Monitoring System

LSTM: Long-short term memory network

RMSE: Root-mean squared error

ANOVA: Analysis of variance

AGI: Artificial general intelligence

GAI: Generative artificial intelligence

LLM: Large language models

Chapter 1 - Introduction

Surgical Practice

Surgery requires critical care of tissues by expert hands. In ancient times, surgical practice was rudimentary in the form of sorcery mixed with religious beliefs and traditions.[1] Hippocrates (460 - 375 BC) introduced the concept of detailed observation and documentation into practice and is traditionally called the 'Father of Medicine'.[2] Renaissance figures such as Andreas Vesalius (1524 - 1564) marked a shift towards a more scientific approach and a more detailed understanding of human anatomy.[3] Up until 1754, barbers were involved in surgical procedures due to their skillset with sharp instruments.[4] The evolution of surgical practice to today's advanced training, techniques, and technology-assisted applications came with great improvements in patient care.[5-7]

Despite ever-improving practice, avoidable surgical errors remain a significant factor contributing to patient morbidity and mortality as well as the burden of costs on healthcare systems.[8,9] Among these, technical errors are the most common class of errors.[8,10] Lack of technical competence in surgery correlates to poor patient outcomes with high rates of adverse events, reoperation, and readmission.[11-13] Technical skills are especially important in disciplines that involve complex procedures, such as neurosurgery,[10,14] where technical mastery needs to be obtained predominantly during neurosurgical training and maintained throughout practice to achieve optimal patient outcomes.

Surgical Education

Apprenticeship has always been the mainstream teaching and learning methodology in surgical practice. This methodology has evolved from a simple ‘see one, do one, teach one’ approach to more structured institutionalized programs with advanced training protocols.[7,15] Besides high medical knowledge requirements, surgery is an application of technical and practical skills. Therefore, unlike non-procedural medicine, obtaining technical competency is an integral part of surgical apprenticeship.[16]

The current surgical training paradigm is based on fixed-length training programs, referred to as residency. This paradigm involves the assumption that sufficient time spent in clinics and exposure to surgical procedures would ensure adequate technical capacity for trainees to perform patient cases on their own.[17] However, this approach falls short in assessing and teaching composites of surgical mastery in an objective and standardized way. These limitations relate to the challenges in quantifying performance and defining surgical technical expertise, including what would be considered ‘excellent’. As a result, the technical competency of the graduate surgeons varies greatly and this variation influences patient outcomes.[13] Additionally, learning in clinical contexts and the involvement of residents during surgery increase risks to patients.[18] This learning is heavily dependent on the presence of expert surgeon educators and their skills not only in surgery but also in teaching. As such, lack of adequate supervision can be a contributing factor to resident mistakes, undermining patient safety.[19]

To address the current challenges in surgical training and as a response to the growing public demand, a competency-based approach is being integrated into medical and surgical training.[20,21] Implementation of developing technologies in surgical training helps to address the concerns and enhance competency-based curricula.

Surgical Simulation

Surgical simulation has been provided as a means to replicate surgical tasks, allowing learners to practice and acquire technical skills in patient risk-free environments.[22-24] Historically, the concept of simulation in surgery was grounded in the use of physical models and anatomical cadaveric dissection.[25] Today's simulation platforms integrate a variety of techniques and tools and simulate diverse procedures ranging from simpler tasks, such as knot tying and suturing, to more complex ones.[26-28]

Some simulation models are based on physical models such as box trainers[29,30], 3D-printed tissues[31,32], or manikins.[33] Box trainers used for laparoscopic surgery allow the manipulation of real objects often using real operative instruments [34] although, some of the elements of the real tissues, such as bleeding, may be missing. Some evidence was shown that learning with these simulation systems may be transferable to improved operative performance during laparoscopic cholecystectomy.[35] Another commonly used simulation approach is by using biological tissues such as cadavers,[36,37] placenta,[38,39] or animals.[40,41] The use of fresh tissues provides realism however these settings require preparation before each practice and access to tissues when they are fresh can be challenging.[42,43]

A variety of simulation technologies exist. High-fidelity simulations separate themselves from lower-fidelity simulations by providing a more realistic, immersive environment [42] along with many advantages such as the capacity to collect data from users/trainees.[27] Although high-fidelity simulators may cost more[44] and may not provide significant benefits over lower-fidelity simulators in some cases,[45] the ability to collect data and quantify surgical performance may give higher-fidelity simulators unique advantages in the future.

Simulations based on digital settings, such as augmented reality,[46] and virtual reality,[47] are gaining popularity. With the growing interest, advancing computational resources, and the increasing popularity and promises of artificial intelligence, access to data is becoming more critical.[48,49] Virtual reality (VR) (Figure 1) simulation may differentiate itself from other simulation systems and play a crucial role in future surgical training, as they serve as high-fidelity platforms with a vast amount of data recorded during surgical performance.[50] The

Figure 1. The NeuroVR virtual reality neurosurgery simulator, the prototype (left) and the commercial version (right).

integration of artificial intelligence with VR simulation allows for quantifying skills, risk detection, and tailored feedback during realistically simulated complex surgical procedures.[51,52]

Simulation Realism & Validation

Many simulation models have been developed. One critical challenge for these systems is to meet the expectations of their end users: medical professionals and trainees. Hence validity assessments including face, content, and construct validity are common important considerations. Face validity refers to the ability of the simulation model to visually replicate the real task. Components of face validity may include the visual realism of the tissues simulated, tissue texture and color, the realism of the instruments, and the surrounding operative environment which all may affect user engagement and the perceived realism of the task.[42]

Content validity refers to the relevance of the simulation model for the intended use. If the purpose of the simulator is to teach a surgical procedure, or how to use certain instruments, the task being offered should sufficiently challenge the students and have them experience very closely what they would experience in a real-life setting. Haptic systems, as an example, allow the simulator to replicate the tactile feeling one would perceive while performing a real surgical procedure, while, in fact, they interact with simulated instruments and tissues. Although both face and content validities are essential, they alone do not necessarily indicate the effectiveness of the simulator.

Construct validity is used to assess the extent to which a simulator meaningfully and accurately measures what it is supposed to measure. One way to assess construct validity is to evaluate whether the simulation exhibits predictive validity, meaning that it can differentiate between those who are experienced and those who are inexperienced in the given task. If the

construct validity can be established, this indirectly suggests that the simulation system may benefit inexperienced users by allowing them to practice on the simulator and improve their skill levels to the experienced level. In this thesis work, Chapter 3 involved the face, content, and construct validity of a simulated neurosurgical procedure involving the subpial resection of a brain tumor.

Two additional terms may be important: predictive validity and concurrent validity. Predictive validity refers to the ability to ability to accurately assess and forecast future performance, therefore track progress. The work in Chapter 2, involved the demonstration of predictive validity of the Intelligent Continuous Expertise Monitoring System (ICEMS). This system was able to predict neurosurgical trainee skills acquisition throughout their residency.

Concurrent validity can be provided when the outcomes and measurements obtained during the simulation task can be compared to a gold-standard assessment conducted simultaneously. The significant linear relation between the OSATS scores and the ICEMS scores outlined in Chapter 4 can be an example of concurrent validity. In this example, concurrent validity means that achieving higher scores in ICEMS system's measurement reflects the improvement in the gold-standard OSATS rating.

Simulation and Neurosurgery

Neurosurgery is among the most critical, technically challenging surgical disciplines. As a leading center, the goal at the Neurosurgical Simulation and Artificial Intelligence Learning Centre is to utilize simulation and artificial intelligence technologies to enhance trainee neurosurgical skills acquisition. For this purpose, several virtual reality simulators were developed.[53,54] Most notably, the NeuroVR simulation (CAE Healthcare, Montreal, Canada) (Figure 1) was built in collaboration with a group of engineers and doctors to realistically simulate a variety of neurosurgical procedures, particularly cranial.[27,55] This platform provides 3D visualization of the simulated tissue via a binocular, resembling an intraoperative microscope commonly used during neurosurgical procedures. The two haptic handles provide force feedback, allowing for a realistic feeling of contact with the simulated environment while using simulated neurosurgical instruments.[56] Another simulation system, the Sim-Ortho virtual reality

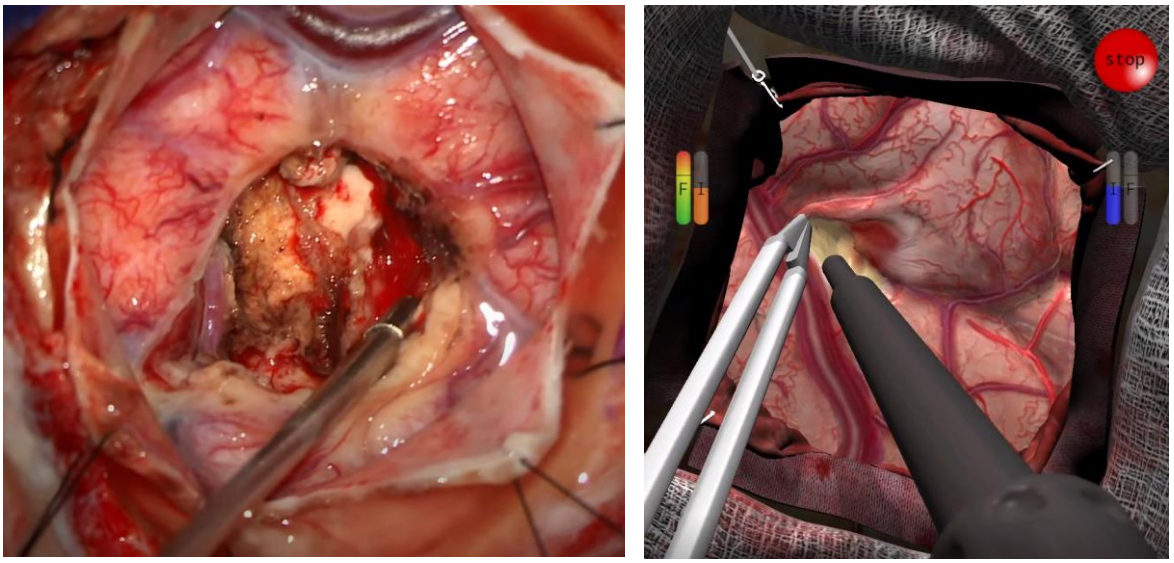


Figure 2. Subpial resection of brain tumors; real operation (left, microneurosurgery.org) and simulation (right, the NeuroVR simulator-CAE Healthcare).

platform simulates very common spinal procedures such as anterior cervical discectomy and fusion.[53] This simulation platform combines two robotic arms to create sufficient haptic force, simulating much stiffer structures like bone.

One of the essential skills to acquire in neurosurgery is the subpial resection of brain tumors and tissues (Figure 2). The subpial resection technique was initially developed for epilepsy surgery to remove epileptogenic brain regions.[57] It is also a common approach in removing brain tumors close to the pial surface.[47] This technique requires preserving the pia mater, the delicate protective layer of the brain, while removing the underlying cortical areas. Expert execution of this technique is important in a variety of neurosurgical procedures[58,59] to remove the abnormal areas completely while preserving the adjacent tissue critical for neurologic function. To enable practicing this skill in patient risk-free realistically simulated settings, the subpial tumor resection tasks were developed and integrated into the NeuroVR platform.[47] The goal of these tasks involved subpial removal of the tumors completely while minimizing bleeding and injury to the surrounding healthy tissue, using a simulated ultrasonic aspirator in the dominant hand for tumor removal and a bipolar forceps in the non-dominant hand for supportive movements and coagulating bleeding tissues.[60]

Objective Assessment of Technical Skills

A data-driven approach is crucial to achieve an objective assessment of technical skills. Virtual reality simulation platforms collect large data from various aspects of surgery, such as instrument utilization, force applied, instrument activation, tissue location, and amount of bleeding, tumor removed, and damage to healthy tissues, making a data-driven objective assessment of technical skills possible.

Neurosurgery involves interaction with delicate tissues. As such gentle utilization of instruments indicates the expertise level of the surgeon while high force applied correlates to poor operative outcomes.[61,62] Using instrument force data recorded during simulation performance, methodologies were developed to assess the expertise level of the trainee and provide feedback. Force pyramid and force heatmap models allowed 2D and 3D visualization of force applied by surgical instruments and the distribution on various critical tissues.[63,64] These models differentiated between expertise levels while also providing insights on hand postures suitable for optimal surgical performance.[65] The work in Chapter 3 involved a similar approach for spatial assessment of non-dominant hand skills and objective feedback.

Performance metrics were developed as the standard means of measurement to quantify safety, quality, efficiency, bimanual dexterity, and instrument movement during simulated operative procedures.[66-68] Using the raw data recorded by the NeuroVR platform, 50 recordings per second, 6600 performance metrics were assessed during a single tumor resection task.[50] However, traditional statistical methodologies fall short in making a comprehensive global assessment of expertise using such vast data. Therefore, more advanced approaches, such as AI, are required.

The integration of AI enables the concept of ‘intelligent systems’, where large data processing and providing high-fidelity assessment and feedback are possible.[69] Applications in surgical training introduce the ability of these systems to involve in teaching like a human instructor. They can engage in training like an independent intelligence and a decision maker aiming to enhance trainee learning outcomes. According to Merriam-Webster dictionary, the word ‘intelligence’ means ‘the ability to learn or understand or to deal with new or trying situations’. Although the use of the term ‘intelligence’ was subject to some premature claims

involving superficial or rule-based applications, the future of fully independent intelligence systems in surgical education and practice is undeniable.

Artificial Intelligence from a Technical Perspective

The area of artificial intelligence (AI) covers a variety of methodologies and techniques to make machines (or computers) capable of decision-making.[48] The detailed technical background of AI is out of the scope of this work; however, to understand how AI works, familiarity with several concepts is important. This section will outline a few concepts that may help the readers to better contextualize the AI methodology used in Chapter 2.

First, AI is computation, specifically through correlation. All AI systems work with algorithms that process input data to predict certain features (outputs). When an algorithm identifies correlations and patterns that sufficiently reveal the output, accurate decisions are made. The process of learning these correlations is a repetitive process referred to as 'algorithm training'. During the training process, algorithms adjust their internal parameters to minimize the difference between their predictions and the actual outputs. All algorithms undergo the process of training along with a validation process, where a smaller separate portion of the data is used to double-check whether during the training process, the algorithms' accuracy, indeed, is improving. Ideally, this training/validation step is followed by a testing step, evaluating the algorithms' performance on an independent dataset to estimate how well they generalize and perform in real-world situations.

AI can be applied in decision-making where a simple rule-based approach does not suffice for accurate results.[70] AI systems have background algorithms based on certain mathematical structures that allow them to learn and perform complex tasks accurately. These background structures vary from simpler statistical or distance analyses [50] to more complex multilayered neural networks (also known as deep neural networks).[51] As an example of simpler models, Naïve Bayes is a commonly used machine learning classifier where the decision is based on a statistical distribution, and the decision is the statistically most likely answer. Another example is K-nearest neighbors, which provides decisions based on the closest neighbors with the most resemblance, with K representing the number of neighbors to take into account. For example, if K=5 in a dichotomous decision, a new prediction will be made by examining the five most resembling occasions, and the decision will be based on the representation that is supported by at least three out of the five instances.

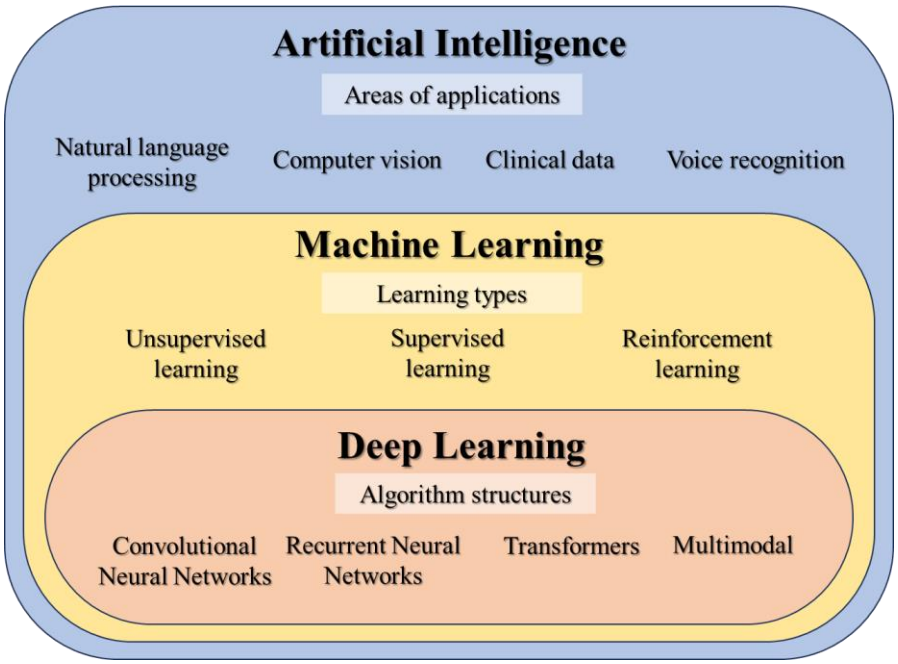


Figure 3. Some important concepts in artificial intelligence applications

On the contrary to simpler machine learning algorithms, complex multilayered AI structures necessitate high computational power. As computational resources develop and become widely available, the popularity of these designs increases, allowing for high-fidelity applications.[48] One important algorithmic structure is artificial neural networks, which employ computational units called nodes (neurons) similar to how the biological brain works. Deep neural networks are formed by using multilayer artificial neural networks, allowing a higher capacity to learn and store more complex concepts for decision-making. The majority of today's exciting AI applications are based on deep neural networks, whether for analyzing clinical data, image or video (computer vision), voice, or language (large language models, such as ChatGPT).[48]

Another concept to outline is the output of algorithms. Outputs can be designed to enable algorithms to predict/decide in various ways. The algorithms are referred to as 'classifiers' when their intended use is to predict specific classes. On the other hand, in cases where the intended output is not a class but rather a numerical value, these designs are known as 'regression models', allowing for more granular numerical decision-making.[51]

The training of algorithms can follow three main learning methodologies: supervised learning, unsupervised learning, and reinforcement learning. A supervised learning approach is used when the data is labeled, meaning that the desired outputs are defined. This methodology often leads to more interpretable predictions, as the predictions of the algorithm will align with the definitions. Unsupervised learning methodology, on the other hand, is used without labeling or supervision to reveal the inherent structures and patterns within the data. This approach yields clusters of data points with common patterns and identifies the differences between these

clusters. However, human interpretation may be required to make meaningful definitions for each of these data clusters.

A mixed approach called semi-supervised learning is used when only portions of the data is labeled. In this case, first, algorithms are pre-trained using the labeled portions of the data in a supervised fashion. The remaining unlabeled data is labeled by the pre-trained algorithm and then used for further training. Reinforcement learning is the training process achieved by introducing feedback in the form of either rewards or penalties. An algorithm makes decisions to maximize a cumulative reward. This approach is widely used in gaming, robotics, and autonomous systems.[71]

To put into context, the AI application introduced in Chapter 2, the Intelligent Continuous Expertise Monitoring System (ICEMS), involved supervised learning. For performance assessment, the input values of the algorithm were 16 performance metrics representing safety, quality, efficiency and bimanual dexterity during the simulated performance. The output values were expertise level of the surgeon/trainees measured between a score of 1 (expert) vs -1 (novice), representing a regression model. More detailed information can be found in the related sections.

Artificial Intelligence to Assess Surgical Performance

AI is reshaping medical applications. AI is widely implemented for assisting clinical diagnosis using imaging,[72] arrhythmia detection using ECG data,[73] colonic polyp and adenoma detection, [74] predicting patient outcomes in spine disorders [75] , and screening or early detection of diabetic retinopathy. [76] The common ground for all these applications is the availability of large datasets and patient parameters that have distinctive characteristics to be differentiated.

However, surgical expertise is a subjective and abstract notion. While an expert surgeon may perform at a less-skilled level at times, a less-skilled trainee may manage parts of the procedure expertly. Differently from other AI applications in medicine, the abstract notion of expertise and having no reliable example of 100% expert in technical skills makes quantifying competency in surgery challenging. This challenge led us to be creative to deal with this issue, as discussed below.

AI offers a variety of algorithm structures and parameters, and development methodologies as discussed above.[48] The choice with all available structures and functionalities determines the success of the algorithm in its desired usage. One may think that AI would never fail if it were not for the limited background AI engineering, misinterpretation, and the use of limited or biased data, all of which may involve human judgment, leading to deviations from the desired usage and causing unintended effects.[77,78]

As a part of this PhD work, we have defined an AI methodology to quantify surgical expertise and made a related patent application for 'Methods and systems for continuous monitoring of task performance' [52] in collaboration with McGill University. We proposed using recurrent neural networks, an AI structure that is suitable for assessing data with time dependencies[79], such as surgical performance[80]. Currently, measuring expertise is based on surgical experience: intraoperative exposure and years in practice. Our application is based on a safe assumption that a person who has long-time exposure to neurosurgical operations, such as a neurosurgeon, would have technical expertise greater than someone who has never been in a neurosurgical operating room, such as a medical student. Following this very logic, our system assigned a neurosurgeon's performance with a score of '1' and a medical student's performance with a score of '-1', indicating that all neurosurgeon performance data should be considered

better than medical students by our AI system. Using these two end-skill levels helped to minimize the aforementioned overlaps that exist between expert and less skilled performance. Additionally, similar to statistical sample size calculation, having the most distinctive groups possible granted success using a smaller sample size.[81] Making predictions on a continuous expertise scale between expert (1) and novice (-1) levels in real-time allowed the applications described in Chapter 2 and Chapter 5.

AI algorithms may have very complex structures, often referred to as the 'black box', that makes the interpretation difficult.[82] However, in high-stakes environments such as surgery, it is important to avoid black box problem and use interpretable approaches.[83,84] Therefore, our application involved making predictions on features that are relevant to surgical practice, and easy to understand and learn by trainees.

Learners' Cognitive Load

Cognitive load is the mental effort of a trainee to process and retain information.[85,86] It is an important consideration while designing a curriculum to maximize learning efficiency and ensure that the trainees remember the information long-term. Since its inception in the 1980s, cognitive load theory has become an instrumental framework that takes human cognitive architecture into account in educational psychology and instructional design.[87] Key components of cognitive load theory include 1- intrinsic cognitive load, 2- extraneous cognitive load, and 3- germane cognitive load.[88]

Intrinsic cognitive load refers to the inherent complexity of the task. For surgical trainees, depending on the surgical domain or the training stage, the task at hand may naturally be complex. Extraneous cognitive load refers to the unnecessary cognitive load caused by the way

the information is delivered to the trainees. More clear and concise manner with minimized redundancy and distractions may increase the efficiency in training.[89,90] Finally, germane cognitive load is the effort required for learners to integrate new information into their existing knowledge. Germane load may indicate a deeper understanding of the subject matter and cognitive absorption.[91] Cognitive overload occurs when these cognitive components combined exceed the learner's capacity.

Cognitive overload may limit the amount of information understood by the trainee, increase stress and anxiety, cause cognitive fatigue and a decline in motivation.[92] Novice trainees are especially at higher risk of cognitive load.[85] Hence, cognitive load is an essential element in designing guidelines in health professional education.[90]

Randomized Controlled Trials for Designing an Effective Curricula

Designing effective curricula involves careful planning and a thorough consideration of various components, as discussed above, along with the integration of technology. Randomized controlled trials (RCTs) are considered the gold standard methodology in advancing medical science to establish evidence-based advancements.[93] Several RCTs were conducted to assess the impact of surgical simulation training on global rating scores and operative time in a variety of procedures such as endoscopic, laparoscopic, endovascular, although a significant effect on patient outcome was limited.[22] Larsen et al. reported that a virtual reality surgical training resulted in significantly lower operating time compared to the control group with no VR training.[94] Zendejas at al. reported the only study with improved patient outcomes where training laparoscopic skills for total extraperitoneal inguinal hernia repair using a box trainer resulted in

fewer postoperative complications such as urinary retention, seroma, hematoma, or wound infection and it decreased hospital stay durations.[95]

An AI-integrated curriculum for surgical technical skills training may help to address the needs of the competency-based approach in surgical training and provide objective assessment and efficient learning. A series of important considerations can be evaluated through RCTs to provide a generalizable evidence-based causality and inference which may inform the future surgical technical skills training curricula. The RCTs in this PhD work assessed, first, whether feedback is essential to improve skills as opposed to practice alone with no feedback. Second, they explored the impact of the manner in which feedback is conveyed, investigating if more engaging feedback information results in better learning outcomes. Lastly, the research assessed the utility of the integration of real-time AI which may have benefits over traditional expert-mediated training. Another RCT was conducted concurrently with this PhD work, where the ICEMS was used to assess learning outcomes while the students received post-hoc AI-enhanced expert-benchmark feedback. This post-hoc feedback resulted in better learning outcomes when compared to remote-expert instruction.[96]

Thesis Goal and Objectives

The overall goal of this thesis work is to develop, validate, and test an objective and standardized assessment and teaching methodology for training surgical bimanual skills, using virtual reality simulation and AI. This work utilizes the resources available at the Neurosurgical Simulation and Artificial Intelligence Learning Centre, such as the NeuroVR platform and the subpial resection simulation tasks.

The overall research hypothesis is:

- Advanced goal-oriented actionable AI feedback would significantly improve learning outcomes in simulated bimanual tumor resection skills training.

The research objectives are:

1- Development and predictive validation of the Intelligent Continuous Expertise Monitoring System, a real-time AI system for assessment of surgical skills, tailored feedback, and risk detection.

2- Spatial analysis of expert-level non-dominant hand skills, and the development and construct validity of a spatial feedback methodology during a complex brain tumor resection task.

3- Comparing computer-assisted numeric, visual, and visuospatial feedback to no-feedback to explore optimal feedback methodologies in teaching technical skills in a randomized controlled trial.

4- Comparing the efficacy of real-time AI instruction to in-person expert instruction in teaching bimanual surgical skills in a randomized controlled trial.

References

1. Woods, M. & Woods, M.B. *Ancient medicine: from sorcery to surgery*, (Twenty-First Century Books, 2000).
2. Orfanos, C. From Hippocrates to modern medicine. *Journal of the European Academy of Dermatology and Venereology* **21**, 852-858 (2007).
3. O'Malley, C.D. *Andreas vesalius of Brussels, 1514-1564*, (Univ of California Press, 1964).
4. Robinson, J.O. The Barber-Surgeons of London. *Archives of Surgery* **119**, 1171-1175 (1984).
5. Stulberg, J.J.*, et al.* Adherence to Surgical Care Improvement Project Measures and the Association With Postoperative Infections. *JAMA* **303**, 2479-2485 (2010).
6. Ingraham, A.M., Richards, K.E., Hall, B.L. & Ko, C.Y. Quality Improvement in Surgery: the American College of Surgeons National Surgical Quality Improvement Program Approach. *Advances in Surgery* **44**, 251-267 (2010).
7. Cameron, J.L. William Stewart Halsted: Our Surgical Heritage. *Annals of Surgery* **225**(1997).
8. Healey, M.A., Shackford, S.R., Osler, T.M., Rogers, F.B. & Burns, E. Complications in Surgical Patients. *Archives of Surgery* **137**, 611-618 (2002).
9. Van Den Bos, J.*, et al.* The $17.1 Billion Problem: The Annual Cost Of Measurable Medical Errors. *Health Affairs* **30**, 596-603 (2011).
10. Leape, L.L.*, et al.* The Nature of Adverse Events in Hospitalized Patients. *New England Journal of Medicine* **324**, 377-384 (1991).
11. Birkmeyer, J.D.*, et al.* Surgical Skill and Complication Rates after Bariatric Surgery. *New England Journal of Medicine* **369**, 1434-1442 (2013).
12. Fecso, A.B., Szasz, P., Kerezov, G. & Grantcharov, T.P. The Effect of Technical Performance on Patient Outcomes in Surgery. *Annals of Surgery* **265**, 492-501 (2017).
13. Stulberg, J.J.*, et al.* Association Between Surgeon Technical Skills and Patient Outcomes. *JAMA Surgery* (2020).
14. Rolston, J.D.*, et al.* Medical errors in neurosurgery. *Surg Neurol Int* **5**, S435-S440 (2014).
15. Rodriguez-Paz, J.M.*, et al.* Beyond “see one, do one, teach one”: toward a different training paradigm. *Postgraduate Medical Journal* **85**, 244-249 (2009).
16. Sonnadara, R.R.*, et al.* Reflections on Competency-Based Education and Training for Surgical Residents. *Journal of Surgical Education* **71**, 151-158 (2014).
17. Long, D.M. Competency-based residency training: the next advance in graduate medical education. *Acad Med* **75**, 1178-1183 (2000).
18. Baisiwala, S., Shlobin, N.A., Cloney, M.B. & Dahdaleh, N.S. Impact of Resident Participation During Surgery on Neurosurgical Outcomes: A Meta-Analysis. *World Neurosurgery* **142**, 1-12 (2020).
19. Gupta, R.*, et al.* Neurosurgical Resident Error: A Survey of U.S. Neurosurgery Residency Training Program Directors' Perceptions. *World Neurosurgery* **109**, e563-e570 (2018).
20. Cadieux, M.*, et al.* Implementation of competence by design in Canadian neurosurgery residency programs*. *Medical Teacher* **44**, 380-387 (2022).
21. Rabski, J.E., Saha, A. & Cusimano, M.D. Setting standards of performance expected in neurosurgery residency: A study on entrustable professional activities in competency-based medical education. *The American Journal of Surgery* **221**, 388-393 (2021).
22. Meling, T.R. & Meling, T.R. The impact of surgical simulation on patient outcomes: a systematic review and meta-analysis. *Neurosurgical Review* **44**, 843-854 (2021).

23. Gélinas-Phaneuf, N. & Del Maestro, R.F. Surgical Expertise in Neurosurgery: Integrating Theory Into Practice. *Neurosurgery* **73**, S30-S38 (2013).
24. Seymour, N.E.*, et al.* Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Annals of surgery* **236**, 458-464 (2002).
25. Holomanova, A., Ivanova, A., Brucknerova, I. & Benuska, J. Andreas Vesalius-the reformer of anatomy. *Bratislavske lekarske listy* **102**, 48-54 (2001).
26. Bowyer, M.W. & Fransman, R.B. Simulation in General Surgery. in *Comprehensive Healthcare Simulation: Surgery and Surgical Subspecialties* (eds. Stefanidis, D., Korndorffer Jr, J.R. & Sweet, R.) 171-183 (Springer International Publishing, Cham, 2019).
27. Delorme, S., Laroche, D., DiRaddo, R. & Del Maestro, R.F. NeuroTouch: A Physics-Based Virtual Simulator for Cranial Microneurosurgery Training. *Operative Neurosurgery* **71**, ons32-ons42 (2012).
28. Dunkin, B., Adrales, G.L., Apelgren, K. & Mellinger, J.D. Surgical simulation: a current review. *Surgical Endoscopy* **21**, 357-366 (2007).
29. Guedes, H.G.*, et al.* Virtual reality simulator versus box-trainer to teach minimally invasive procedures: A meta-analysis. *International Journal of Surgery* **61**, 60-68 (2019).
30. van Empel, P.J.*, et al.* Validation of a new box trainer-related tracking device: the TrEndo. *Surgical Endoscopy* **26**, 2346-2352 (2012).
31. Kröger, E., Dekiff, M. & Dirksen, D. 3D printed simulation models based on real patient situations for hands-on practice. *European Journal of Dental Education* **21**, e119-e125 (2017).
32. Liu, Y.*, et al.* Fabrication of cerebral aneurysm simulator with a desktop 3D printer. *Scientific Reports* **7**, 44301 (2017).
33. Alsaad, A.A., Davuluri, S., Bhide, V.Y., Lannen, A.M. & Maniaci, M.J. Assessing the performance and satisfaction of medical residents utilizing standardized patient versus mannequin-simulated training. *Advances in Medical Education and Practice* **8**, 481-486 (2017).
34. Dehabadi, M., Fernando, B. & Berlingieri, P. The use of simulation in the acquisition of laparoscopic suturing skills. *International Journal of Surgery* **12**, 258-268 (2014).
35. Hamilton, E.C.*, et al.* Comparison of video trainer and virtual reality trainingsystems on acquisition of laparoscopic skills. *Surgical Endoscopy And Other Interventional Techniques* **16**, 406-411 (2002).
36. James, H.K., Chapman, A.W., Pattison, G.T.R., Griffin, D.R. & Fisher, J.D. Systematic review of the current status of cadaveric simulation for surgical training. *British Journal of Surgery* **106**, 1726-1734 (2019).
37. Gnanakumar, S.*, et al.* Effectiveness of Cadaveric Simulation in Neurosurgical Training: A Review of the Literature. *World Neurosurgery* **118**, 88-96 (2018).
38. de Oliveira, M.M.R.*, et al.* Learning brain aneurysm microsurgical skills in a human placenta model: predictive validity. *J Neurosurg* **128**, 846-852 (2018).
39. Ribeiro de Oliveira, M.M.*, et al.* Face, content, and construct validity of human placenta as a haptic training tool in neurointerventional surgery. *J Neurosurg* **124**, 1238-1244 (2016).
40. Alsayegh, A., Bakhaidar, M., Winkler-Schwartz, A., Yilmaz, R. & Del Maestro, R.F. Best Practices Using Ex Vivo Animal Brain Models in Neurosurgical Education to Assess Surgical Expertise. *World Neurosurgery* (2021).
41. Winkler-Schwartz, A.*, et al.* Creating a Comprehensive Research Platform for Surgical Technique and Operative Outcome in Primary Brain Tumor Neurosurgery. *World Neurosurg* **144**, e62-e71 (2020).
42. Oliveira, M.M.*, et al.* Face, Content, and Construct Validity of Brain Tumor Microsurgery Simulation Using a Human Placenta Model. *Operative Neurosurgery* **12**, 61-67 (2016).

43. Winkler-Schwartz, A.*, et al.* Creating a Comprehensive Research Platform for Surgical Technique and Operative Outcome in Primary Brain Tumor Neurosurgery. *World Neurosurgery* **144**, e62-e71 (2020).
44. Delisle, M. & Hannenberg, A.A. Alternatives to High-Fidelity Simulation. *Anesthesiology Clinics* **38**, 761-773 (2020).
45. Lefor, A.K., Harada, K., Kawahira, H. & Mitsuishi, M. The effect of simulator fidelity on procedure skill training: a literature review. *Int J Med Educ* **11**, 97-106 (2020).
46. Kim, Y., Kim, H. & Kim, Y.O. Virtual Reality and Augmented Reality in Plastic Surgery: A Review. *Arch Plast Surg* **44**, 179-187 (2017).
47. Sabbagh, A.J.*, et al.* Roadmap for Developing Complex Virtual Reality Simulation Scenarios: Subpial Neurosurgical Tumor Resection Model. *World Neurosurgery* **139**, e220-e229 (2020).
48. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444 (2015).
49. Leary, D.E.O. Artificial Intelligence and Big Data. *IEEE Intelligent Systems* **28**, 96-99 (2013).
50. Winkler-Schwartz, A.*, et al.* Machine Learning Identification of Surgical and Operative Factors Associated With Surgical Expertise in Virtual Reality Simulation. *JAMA Netw Open* **2**, e198363 (2019).
51. Yilmaz, R.*, et al.* Continuous monitoring of surgical bimanual expertise using deep neural networks in virtual reality simulation. *npj Digital Medicine* **5**, 54 (2022).
52. Yilmaz, R., Winkler-Schwartz, A., Mirchi, N., Del Maestro, R. Methods and systems for continuous monitoring of task performance. (2022).
53. Ledwos, N.*, et al.* Virtual Reality Anterior Cervical Discectomy and Fusion Simulation on the Novel Sim-Ortho Platform: Validation Studies. *Oper Neurosurg (Hagerstown)* **20**, 74-82 (2020).
54. Varshney, R.*, et al.* The McGill simulator for endoscopic sinus surgery (MSESS): a validation study. *Journal of Otolaryngology - Head & Neck Surgery* **43**, 40 (2014).
55. Brunozzi, D., McGuire, L.S. & Alaraj, A. NeuroVR™ Simulator in Neurosurgical Training. in *Comprehensive Healthcare Simulation: Neurosurgery* (ed. Alaraj, A.) 211-218 (Springer International Publishing, Cham, 2018).
56. Bugdadi, A.*, et al.* Is Virtual Reality Surgical Performance Influenced by Force Feedback Device Utilized? *J Surg Educ* **76**, 262-273 (2019).
57. Daniel, T.L., Michael, R.S., Jacqueline, A.F., Michael, J.O. & Connor. The syndrome of frontal lobe epilepsy. *Neurology* **45**, 780 (1995).
58. Esquenazi, Y.*, et al.* The Survival Advantage of "Supratotal" Resection of Glioblastoma Using Selective Cortical Mapping and the Subpial Technique. *Neurosurgery* **81**(2017).
59. Spencer, S.S.*, et al.* Multiple Subpial Transection for Intractable Partial Epilepsy: An International Meta-analysis. *Epilepsia* **43**, 141-145 (2002).
60. Yilmaz, R.*, et al.* Nondominant Hand Skills Spatial and Psychomotor Analysis During a Complex Virtual Reality Neurosurgical Task-A Case Series Study. *Oper Neurosurg (Hagerstown)* **23**, 22-30 (2022).
61. Sugiyama, T.*, et al.* Forces of Tool-Tissue Interaction to Assess Surgical Skill Level. *JAMA Surgery* **153**, 234-242 (2018).
62. Albakr, A., Baghdadi, A., Singh, R., Lama, S. & Sutherland, G.R. Tool-Tissue Forces in Hemangioblastoma Surgery. *World Neurosurgery* **160**, e242-e249 (2022).
63. Sawaya, R.*, et al.* Virtual Reality Tumor Resection: The Force Pyramid Approach. *Operative Neurosurgery* **14**, 686-696 (2017).
64. Azarnoush, H.*, et al.* The force pyramid: a spatial analysis of force application during virtual reality brain tumor resection. **127**, 171 (2016).
65. Sawaya, R.*, et al.* Development of a performance model for virtual reality tumor resections. *Journal of Neurosurgery JNS* **131**, 192-200 (2018).

66. Alotaibi, F.E.*, et al.* Neurosurgical Assessment of Metrics Including Judgment and Dexterity Using the Virtual Reality Simulator NeuroTouch (NAJD Metrics). *Surgical Innovation* **22**, 636-642 (2015).
67. Bissonnette, V.*, et al.* Artificial Intelligence Distinguishes Surgical Training Levels in a Virtual Reality Spinal Task. *J Bone Joint Surg Am* **101**, e127 (2019).
68. AlZhrani, G.*, et al.* Proficiency Performance Benchmarks for Removal of Simulated Brain Tumors Using a Virtual Reality Simulator NeuroTouch. *Journal of Surgical Education* **72**, 685-696 (2015).
69. Nagi, F.*, et al.* Applications of Artificial Intelligence (AI) in Medical Education: A Scoping Review. *Stud Health Technol Inform* **305**, 648-651 (2023).
70. Amisha, Malik, P., Pathania, M. & Rathaur, V.K. Overview of artificial intelligence in medicine. *Journal of Family Medicine and Primary Care* **8**(2019).
71. Li, Y. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274* (2017).
72. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H. & Aerts, H.J.W.L. Artificial intelligence in radiology. *Nature Reviews Cancer* **18**, 500-510 (2018).
73. Hannun, A.Y.*, et al.* Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine* **25**, 65-69 (2019).
74. Pu, W.*, et al.* Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* **68**, 1813 (2019).
75. Müller, D.*, et al.* Development of a machine-learning based model for predicting multidimensional outcome after surgery for degenerative disorders of the spine. *European Spine Journal* **31**, 2125-2136 (2022).
76. Gulshan, V.*, et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402-2410 (2016).
77. Nelson, G.S. Bias in artificial intelligence. *North Carolina medical journal* **80**, 220-222 (2019).
78. Kahneman, D., Slovic, P. & Tversky, A. *Judgment under uncertainty: Heuristics and biases*, (Cambridge university press, 1982).
79. Lipton, Z.C., Berkowitz, J. & Elkan, C. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019* (2015).
80. Yilmaz, R., Winkler-Schwartz, A., Mirchi, N., Reich, A. & Del Maestro, R. O51: ARTIFICIAL INTELLIGENCE UTILIZING RECURRENT NEURAL NETWORKS TO CONTINUOUSLY MONITOR COMPOSITES OF SURGICAL EXPERTISE. *British Journal of Surgery* **108**(2021).
81. Cohen, P.R. *Empirical methods for artificial intelligence*, (MIT press Cambridge, MA, 1995).
82. Castelvecchi, D. Can we open the black box of AI? *Nature News* **538**, 20 (2016).
83. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**, 206-215 (2019).
84. Rudin, C. Why black box machine learning should be avoided for high-stakes decisions, in brief. *Nature Reviews Methods Primers* **2**, 81 (2022).
85. Sweller, J. Cognitive load during problem solving: Effects on learning. *Cognitive Science* **12**, 257-285 (1988).
86. Young, J.Q., Van Merrienboer, J., Durning, S. & Ten Cate, O. Cognitive Load Theory: implications for medical education: AMEE Guide No. 86. *Med Teach* **36**, 371-384 (2014).
87. Sweller, J. CHAPTER TWO - Cognitive Load Theory. in *Psychology of Learning and Motivation*, Vol. 55 (eds. Mestre, J.P. & Ross, B.H.) 37-76 (Academic Press, 2011).
88. Sweller, J. Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load. *Educational Psychology Review* **22**, 123-138 (2010).
89. Mayer, R.E. Cognitive Theory of Multimedia Learning. in *The Cambridge Handbook of Multimedia Learning* (ed. Mayer, R.E.) 43-71 (Cambridge University Press, Cambridge, 2014).

90. Van Merriënboer, J.J.G. & Sweller, J. Cognitive load theory in health professional education: design principles and strategies. *Medical Education* **44**, 85-93 (2010).
91. Debue, N. & van de Leemput, C. What does germane load mean? An empirical contribution to the cognitive load theory. *Frontiers in Psychology* **5**(2014).
92. Andersen, S.A.W., Mikkelsen, P.T., Konge, L., Cayé-Thomasen, P. & Sørensen, M.S. The effect of implementing cognitive load theory-based design principles in virtual reality simulation training of surgical skills: a randomized controlled trial. *Advances in Simulation* **1**, 20 (2016).
93. Stanley, K. Design of Randomized Controlled Trials. *Circulation* **115**, 1164-1169 (2007).
94. Christian, R.L*., et al.* Effect of virtual reality training on laparoscopic surgery: randomised controlled trial. *BMJ* **338**, b1802 (2009).
95. Zendejas, B*., et al.* Simulation-based mastery learning improves patient outcomes in laparoscopic inguinal hernia repair: a randomized controlled trial. *Ann Surg* **254**, 502-509; discussion 509-511 (2011).
96. Fazlollahi, A.M*., et al.* Effect of Artificial Intelligence Tutoring vs Expert Instruction on Learning Simulated Surgical Skills Among Medical Students: A Randomized Clinical Trial. *JAMA Network Open* **5**, e2149008-e2149008 (2022).

Chapter 2 - Continuous Monitoring of Surgical Bimanual Expertise Using Deep Neural Networks in Virtual Reality Simulation

Preface

Chapter 1 discussed the need for AI in high-fidelity surgical performance assessment, feedback, and error mitigation. In this chapter, we outline the development of the Intelligent Continuous Expertise Monitoring System (ICEMS), a deep learning methodology using recurrent neural networks, for real-time assessment of surgical performance during two simulated subpial tumor resection tasks. This chapter involves two main goals, 1- the development of the ICEMS's three modules: performance assessment, feedback, and risk detection; and 2- the predictive validation of the first module throughout a neurosurgical training program, on performance data of 26 neurosurgery residents. The work in this chapter lays the foundation of future works by quantifying surgical performance, allowing for tailored feedback, facilitating objective comparison of skills, and tracking trainee learning. The manuscript was published as:

Yilmaz R, Winkler-Schwartz A, Mirchi N, Reich A, Christie S, Tran DH, Ledwos N, Fazlollahi AM, Santaguida C, Sabbagh AJ, Bajunaid K, Del Maestro R. Continuous monitoring of surgical bimanual expertise using deep neural networks in virtual reality simulation. NPJ Digital Medicine. 2022 Apr 26;5(1):54. https://doi.org/10.1038/s41746-022-00596-8

Abstract

In procedural-based medicine, technical ability can be a critical determinant of patient outcomes. Psychomotor performance occurs in real-time, hence a continuous assessment is necessary to provide action-oriented feedback and error avoidance guidance. We outline a deep learning application, the Intelligent Continuous Expertise Monitoring System (ICEMS), to assess surgical bimanual performance at 0.2-second intervals. A long-short term memory network was built using neurosurgeon and student performance in 156 virtually simulated tumor resection tasks. Algorithm predictive ability was tested separately on 144 procedures by scoring the performance of neurosurgical trainees who are at different training stages. The ICEMS successfully differentiated between neurosurgeons, senior trainees, junior trainees, and students. Trainee average performance score correlated with the year of training in neurosurgery. Furthermore, coaching and risk assessment for critical metrics were demonstrated. This work presents a comprehensive technical skill monitoring system with predictive validation throughout surgical residency training, with the ability to detect errors.

Introduction

The mastery of technical skills is of fundamental importance in medicine and surgery as technical errors can result in poor patient outcomes. [1-3] The learning of bimanual psychomotor skills still largely follows an apprenticeship model: one defined by a trainee completing a fixed-length residency working closely with instructors. Technical skills education is transitioning from this time-focused approach to competency-based quantifiable frameworks. [4,5]

Surgical trainees are considered competent when they can perform specific surgical procedures safely and efficiently, encompassing knowledge, judgement, technical and social skills to solve familiar and novel situations to provide adequate patient care. [6] The focus on "adequate" rather than "excellent" or "expert" patient care relates to challenges in outlining, assessing, quantifying, and teaching the composites of surgical expertise. To provide competency-based frameworks for complex psychomotor technical skills, advanced platforms need to be created which provide objective feedback during training along with error mitigation systems. [7] These frameworks need to be transparent and based on quantifiable objective metrics. [8,9]

A technically challenging operative procedure in surgery involves the subpial resection of brain tumors adjacent to critical cortical structures. [10] Neurosurgical graduates are expected to be proficient in this complex bimanual skill which includes minimizing injury to adjacent normal tissues and hemorrhage from subpial vessels. Technical errors in this procedure can result in significant patient morbidity. [10,11] Our group developed complex realistic virtual reality tumor resection tasks to aid learners in the mastery of this skill. [12,13] Exploiting these simulations on the NeuroVR platform with haptic feedback (CAE Healthcare, Montreal, Canada) we quantified

multiple components of the bimanual psychomotor skills used to expertly perform these tasks. Utilizing this data post-hoc, we developed expert performance benchmarks to which learner scores were compared and machine learning algorithms to classify participants into pre-defined expertise categories. [8,14,15] Limitations of these applications were the inability of ongoing assessment and error detection and improving performance during the task by providing continuous feedback.

Most surgical skills learning occurs in the operating room, with the surgeon instructor continuously evaluating trainee performance and providing coaching to improve performance with a particular focus on preventing surgical errors which may cause patient injury. This assessment occurs in real-time and is relevant to the precise action being performed by the trainee and the risks associated with that action. To mimic the role of expert operative instructors, we developed an artificial intelligence (AI) deep learning application, the Intelligent Continuous Expertise Monitoring System (ICEMS). The ICEMS was developed with two objectives: 1)- to make a continuous assessment of psychomotor skills to detect less-skilled performance during surgery, 2)- to provide ongoing action-oriented feedback and risk notifications.

This paper outlines the development of the ICEMS (Figure 1) and provides predictive validation evidence that enables future studies to explore its efficacy in simulation training. To our knowledge, this application is the first continuous bimanual technical skill assessment using deep learning with the predictive validation on surgical trainee performance throughout a residency program. [16]

Results

Participants and data

Neurosurgeons, neurosurgical fellows, neurosurgical residents, and medical students from McGill University were invited to participate. Neurosurgeons and medical students were categorized as experts (n = 14) and novices (n = 12), respectively. Neurosurgical fellows and residents were allocated a priori into two groups based on their previous operative exposure: seniors (4 neurosurgical fellows and 10 neurosurgical residents in years 4-6,) and juniors (10 neurosurgical residents in years 1-3) (Table). Each participant performed two different simulated subpial tumor resection tasks a total of six times, resulting the data from 300 attempts in total (Figure 2). The simulated scenarios and were described previously (Figure 3) [8,12]. Data was recorded in a single time point. No data-exclusion was applied. Mean age [SD] was, for experts: 45.9 [8], for seniors: 32.3 [2.1], for juniors: 29.8 [3.2] and for novices: 24 [1.3]. Trainee number of complete subpial tumor resections performed (mean [min-max]) was, seniors: 14.7 [0-45], juniors: 1 [0-7] (Supplementary Table 3).

AI design and development

The definition of expertise in surgical technical skills is challenging since surgical performance involves continuous interplay between multiple factors. [17] However, the composites of expertise are present in the performance of expert professionals. We developed the Intelligent Continuous Expertise Monitoring System in this context by training a Long-Short Term Memory (LSTM) network to learn operative surgical expertise from the difference between expert and novice surgical skills considering the continuous flow of the performance. The algorithm was trained with both end skill levels with more than 700 minutes of operative performance with a data entry at 0.2-second intervals (with over 200,000 data points of analysis).

A surgical performance is a combination of multiple intraoperative interactions. An appropriate assessment requires considering these tasks being carried out within the flow of the performance. LSTM networks, as a type of recurrent neural network, allowed for the evaluation of each time point in relation with the previous time points, giving the ability to consider sequences in movements. [18-20]

Sixteen performance metrics were extracted at 0.2-second increments from the simulation data (Figure 4). Metrics included features related to bimanual technical skills such as instruments tip separation distance, force applied by each instrument and velocity and acceleration of each instrument as well as operative factors such as tumor removed, control of bleeding and damage to healthy tissue. An LSTM algorithm was built by inputting these 16-performance metrics utilizing only expert/neurosurgeon (n=14) and novice/medical student (n=12) performance data on 84 and 72 tasks, respectively. The algorithm was structured as a regression model quantifying expertise level as a continuous variable from expert/skilled level (a score of 1.00) to novice/less-skilled level (a score of -1.00). To avoid overfitting, root-mean-squared-error (RMSE) values on the three separate datasets were monitored (Supplementary Table 1). Detailed information about algorithm structure and development can be found in Online Methods and supplementary data.

Quantifying skills

The performance of 24 trainees (on 144 tasks) in different years of neurosurgery training (Table) was used to assess the algorithm's predictive validation. All 300 participant trials were scored by the trained LSTM algorithm at 0.2 second intervals between '1.00'(skilled) and '-1.00'(less-skilled). An average performance score was calculated for each task (Supplementary Figure 5). Participants' mean scores were calculated across six trials for statistical comparisons.

Group average surgical performance scores were; experts, 0.509; 95% CI [0.424 0.593]; seniors, 0.258; 95% CI [0.114 0.402]; juniors, -0.11; 95% CI [-0.358 0.139]; and novices, -0.398; 95% CI [-0.545 -0.251]. No outliers were found, as assessed by boxplot. Only a trial data that belongs to a fifth attempt of a neurosurgeon was missing, no imputation was made. Average performance score was normally distributed for each expertise group as determined by Shapiro-Wilk test ($p > .05$). Levene's test showed equality of variances, based on median ($p = .083$).

The average performance score was significantly different between expertise groups, $F(3,46) = 33.927$, $p < .001$, as determined by a one-way ANOVA. Tukey-Kramer post-hoc test of between groups differences revealed that the expert group scored significantly higher than seniors (mean difference: .251 95%CI [.004-.497], $p = .045$) and juniors scored significantly higher than novices (mean difference: .289 95%CI [.009-.568], $p = .04$) in average performance score. The ICEMS also differentiated between surgical trainee groups with seniors scoring significantly higher than juniors (mean difference: .367 95%CI [.097-.638], $p = .004$) (Figure 5). In a linear regression analysis resident year of training in neurosurgery statistically predicted the average performance score, $F(1, 22) = 9.81$, $p = .005$ and accounted for 30.8% of the variation in the average score with adjusted $R2 = 27.7\%$, a large size effect according to Cohen (1988). [21] Average performance score increased by 0.092, 95% CI [.031-.153] per training-year (Figure 6). The ability of the ICEMS to continuously assess surgical performance during the surgical task is demonstrated in videos outlining a neurosurgeon [video-1] and a medical student performance [video-2] (video legend: Supplementary Figure 3).

Coaching and risk detection

A major application of the ICEMS is to provide continuous personalized action-oriented feedback helping trainees modify their bimanual psychomotor movements to expert level

performance and provide critical information to mitigate errors. Three algorithms provided continuous expert-level coaching for (1) aspirator utilization, (2) bipolar forceps utilization and (3) bimanual coordination. [8,15,22] These algorithms provided the ability to revise instrument utilization to expert level continuously. Two other algorithms demonstrated ongoing risk detection capacity for (4) bleeding and (5) healthy tissue injury. [8,23] RMSE values obtained for training, validation and testing of these algorithms are available in Supplementary Table 1.

Although, the validation of these modules in practice for coaching and risk detection will be the object of future studies, we outline the video performance of these algorithms on a senior [video-3] and a junior resident operation [video-4] (video legend: Supplementary Figure 4). Learning from the difference between expert and novice performance, the ICEMS reproduces some components of intelligent assessment and coaching similarly provided by expert surgical instructors in the operating room.

Discussion

The transition towards competency-based quantifiable frameworks for evaluation and teaching of surgical technical skills is resulting in the development of high-fidelity virtual reality simulators to aid this learning transformation. These systems provide trainees with repetitive opportunities for experiential learning in patient risk-free environments without limitations imposed by the availability of expert surgical instructors or patient cases. [24-26] We demonstrate an artificial intelligence application to enable these platforms to function as objective autonomous intelligent training platforms with the ability to continuously track psychomotor learning as surgical trainees transition along the spectrum from novice to expert performance.

The NeuroVR platform (previously NeuroTouch, CAE Healthcare, Montreal, Canada) used in this study is a high-fidelity virtual reality neurosurgical simulator that allows 3D visual and haptic interaction in a hyper-realistic simulated surgical environment. [13] This platform was developed by a team of engineers from the National Research Council of Canada with expert inputs from 23 international training hospitals. Extended realism was provided by the 3D microscopic visualization through a binocular, and two haptic handles to allow bimanual simultaneous movement. Tumor physical properties were adjusted using data from multiple primary human brain tumor specimens. [27] Haptic tuning was applied based on the feedback from neurosurgeons. [12] Human brain tissue and bleeding mechanics were implemented including pulsation of blood vessels. A brain tumor surgery intraoperative audio recording was added to increase background auditory realism. The vast dataset generated by this platform allowed for the development of comprehensive intelligent systems. [8,9]

Studies involving real-time surgical technical skills assessment demonstrated supportive results; however, these studies were restricted to one-handed virtual reality systems during a steerable needle task, epidural needle insertion or drilling a simulated femur. [28-30] Most operative procedures involve the coordinated interactions of both hands, each employing a different instrument to accomplish an operative goal. The major roles of expert operative room surgical instructors are to assess trainees' bimanual skills and help them improve their skills to safely carry out procedures to decrease patient morbidity and mortality. [31] This is crucial especially for high-risk medical procedures. Our group has focused on developing an LSTM network to mirror the role of surgical instructors in assessing bimanual performance involving high-risk complex neurosurgical procedures like the subpial resection. Previous real-time assessment applications utilized small datasets, included engineering students or non-identified participants and have not

validated or tested their algorithms on appropriate learner performance. [16,28-30,32] In contrast, the ICEMS was developed utilizing neurosurgeon/expert and medical student/novice performance, and its was tested using the data from neurosurgical trainees who are at different stages of training.

Our framework offers several advantages. First, the ICEMS was trained as a regression model with the two-end skill level performance, providing a continuous expertise scale from novice to expert level. This allowed a more granular performance assessment from the previous applications[8] and tracking of learning throughout the years of residency training from medical school training to years of practice. Second, we developed our system utilizing two simulated tasks that require the same bimanual surgical technique. This approach offers a more generalizable assessment of this surgical technique across different tasks.

One of the drawbacks of deep learning applications is the 'black box' problem where complexity of the analysis (1) limits the interpretability of the assessment and (2) makes providing relevant information for feedback difficult. To overcome these issues; (1) our assessment system was built on relevant features that are easy to understand and learn. Based on our previous studies, we implemented features representing dominant and non-dominant hand movement and force applied, bimanual cognitive, tissue and bleeding information, and safety metrics. (2) Separate algorithms were trained to work in reverse and provide ongoing feedback for the very features that the assessment was made on. We demonstrate a methodology to generate feedback for any essential performance metric and provide five example features for coaching and risk detection (Supplementary Figure 2).

In previous self-tutoring frameworks, the proposed coaching was based on expert level classification or pre-recorded expert parameters such as videos, benchmarks, or milestones. [9,33-35]

In contrast to determining feedback based on expertise group classification or static parameters, the ICEMS produces dynamic feedback for each performance metric by separate algorithms. This involves revising predictions to the highest expert performance level for specific metrics continuously throughout the task, and this revised information can be used as feedback for trainees or any level of performance including expert groups. An action-oriented personalized coaching is provided for specific metrics.

The continuous evaluation done by the ICEMS can be utilized either in real-time to produce visual, auditory, and haptic clues to enhance performance during the task, or to make a summative assessment and provide feedback after task completion. Both learners and instructors can be provided with post-hoc performance videos flagged with the exact time frame(s) of less-skilled performance (see the videos provided in Results). This AI-generated information outlining the reasons for less-skilled assessment may improve trainee self-directed performance and help educators improve learner skills.

Experts may demonstrate performance features that are similar to that of less-skilled level performance. These common features may be due to the intrinsic characteristics of human bimanual performance, the simulated task, or the limits in recording data. For this reason, the ICEMS was built using expert level performance in comparison to novice performance to differentiate expert specific features. Our results have shown that these expert specific patterns were increasing throughout trainee-years in training.

Expert surgeons develop and implement autonomous motor activity defined as 'psycho-motor skills script' with increasing surgical knowledge. [15] Our system allows trainees to have constant awareness of their level of performance as visualized on a less-skilled to expert scale. By self- modifying their bimanual psychomotor movements with the capacity for unlimited

repetitions to achieve expert performance trainees may more quickly develop a "psychomotor skills script" associated with muscle memory that expert surgeons develop and maintain. This may allow trainees to be more prepared when faced with similar procedures in the operating room. [15,36,37]

Our system is developed in the context of surgical simulation using the extensive information recorded by a specific virtual reality simulator. However, this methodology can be useful beyond the scope of surgical simulation and applicable to any technical performance where the necessary data is available. Intraoperative surgical instrument tracking systems are being developed. [27] Future surgical operative rooms may benefit from this application by the integration of AI and intraoperative data recording systems/instruments. [38,39] Surgical operative rooms may evolve into intelligent operating rooms outfitted with a series of evaluating and intelligent tutoring platforms focused on enhancing safe operative performance and thus improving patient outcomes. [40]

Studies have demonstrated that technical skills may correlate with surgical outcomes. [2,41] Improvement in technical skills may improve the outcome, hence, current attempts in simulation training are focused on enhancing trainee technical skills acquisition. However, it remains to be explored if training with intelligent simulation systems can improve patient outcomes.

Deep learning applications require larger datasets. [19] Complex patient cases often require surgeons who have specific expertise in these operative procedures. Surgical trainees acquire these skills operating with limited number of experts, but in multiple repetition of patient cases. Intelligent systems can be developed in a similar way that the trainees learn, using information from limited number of experts but involving multiple occasions of a surgical procedure. This study involved data from 14 neurosurgeons (experts) each repeating the simulation tasks a total

of six times, allowing an assessment of 83 expert trial data. If the number of experts is limited, the number of task repetitions performed by each surgeon can be increased to develop accurate and generalizable intelligent systems. This approach may provide a feasible and reproducible method in the intelligent assessment of different surgical skills. Should the data size be limited, data augmentation methodologies can help to increase data size and achieve reliable predictions. [42]Intelligent systems can be continuously improved with more data available. Applications with real-time assessment, coaching and risk detection ability may promote the use of these systems, provide access to new data, and allow further improvement of these systems.

This study has several limitations. Our simulation does not reproduce many of the complex and dynamic learning interactions occurring in modern operating rooms and variables such as the view angle, surgeon instrument choice and instrument intensities were controlled. As simulation platforms advance and incorporate more detailed real-life interactions, more comprehensive assessments can be generated by the ICEMS. For training this supervised deep learning application, each data point of the performance of expert and novices was given the same score (expert: 1.00, novice: -1.00) throughout the task, allowing the algorithms to learn both extremes of the skill spectrum. However, individuals may not always perform in line with their expertise levels. In other words, skilled individuals may perform closer to less-skilled level in certain parts of the procedure and vice versa. Nevertheless, the magnitude of the data allowed algorithms to learn from the two end-skill levels and our system provided a granular differentiation across expertise levels as well as between trainee levels. We defined trainee expertise level based on operative exposure or year in training. However, trainee skill levels may not be completely consistent with these parameters and many other factors may also affect trainee technical skill, including trainee inherent ability or the type of exposure to operative

skills[23] (Supplementary Table 3). By quantifying skills, our application addresses an important limitation for future studies to track trainee learning and explore trainee learning patterns. [43] Our study involved a small number of participants from a single institution. With a broader cohort, the generalizability of our model can be increased.

This work, being limited to a previously collected data, provided a validation for the assessment module. An ongoing randomized control trial (ClinicalTrials.gov Identifier: NCT05168150) is addressing the efficiency and validation of coaching and risk detection modules by providing feedback to trainees while tracking their improvement by the assessment module.

As newer technologies[44] and techniques such as reservoir computing[45,46] become available, further progress can be made in the applications of continuous technical skill assessment, feedback and operative risk detection using newer and existing datasets.

With the ongoing pandemic, limiting human contact became an essential practice and the present educational paradigms are being re-evaluated. [47] Virtual reality simulators provided with assessment and coaching modules are self-practicing intelligent tools which may aid trainees and educators navigate the ever-evolving landscape that learners will face.

This work presents a technical skills continuous assessment application built using expert surgeon data, with predictive validity across a training program on surgical trainee performance. [16,35] This deep learning application demonstrated a granular differentiation across expertise and between resident levels. The ICEMS offers a generalizable and objective continuous assessment of surgical bimanual skills which may have implications in the assessment and training of procedural interventions.

Methods

Setting

Data of this consecutive retrospective case series study was collected at a single time point between March 2015 to May 2016, with no follow-up. Neurosurgeons, neurosurgical fellows, and residents from one Canadian university were invited to participate in this study at the Neurosurgical Simulation and Artificial Intelligence (AI) Learning Centre, McGill University. Medical students who expressed interest in neurosurgery or were rotating on the neurosurgical service were also invited to take part. Participant data was anonymized. All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Declaration of Helsinki. [48] This study was approved by the McGill University Health Centre Research Ethics Board, Neurosciences-Psychiatry and all participants signed an approved consent form before trial participation. This report adheres to guidelines for best practices in reporting studies on machine learning to assess surgical expertise in virtual reality simulation, reporting observational studies and the reporting of studies developing and validating a prediction model, as applicable. [49-52]

Simulation

Participants carried out a simulated subpial tumor resection 5 times followed by a simulated complex brain tumor resection (Figure 3), employing a simulated ultrasonic aspirator in the dominant hand and a simulated bipolar forceps in the non-dominant hand, using the NeuroVR high-fidelity simulation platform (CAE Healthcare, Montreal, Canada). These tasks were designed to replicate the high-risk complex subpial brain tumor resection task. [12] Participants were given verbal and written instructions to remove the tumor completely while

minimizing bleeding and injury to surrounding tissue. Simulation data was recorded by the NeuroVR platform in 0.02-second increments (50-recording per second).

Performance metrics

Before any processing, the raw data underwent interpolation to regularize the timing of data points. Sixteen performance metrics were extracted from raw simulation data, at 0.2 second intervals, based on our previous studies, representing five essential aspects of the operative performance: safety, quality, efficiency, bimanual cognitive and movement. [14,23,31,53-59] Although deep learning does not require metric extraction, The ICEMS is developed as a training and feedback tool, therefore particular attention is given to develop the system on features which a trainee can understand and learn. The performance metrics are listed in Figure 4.

Data preparation before AI application

The data comprised a total of 156 tasks (neurosurgeons: 84 tasks, medical students: 72 tasks) was randomly divided into three different subsets as training (70%, a total of 107 tasks), validation (15%, a total of 24 tasks) and testing (15%, a total of 24 tasks) dataset, to provide independent verification and validation (Figure 2). [60] Each individual's performance data was always kept in the same subset. The performance metrics were normalized by z-score normalization, using the mean and standard deviation values based on the training set. Since the algorithm was designed as a 'regression' model where the output feature is predicted as a continuous variable, the categories of expertise levels were transformed into numbers where neurosurgeons (experts) and medical students (novices) were represented as '1' and '-1' respectively, at 0.2-second intervals. Assessment could be as frequent as 0.02 seconds (50 decisions a second) however we limited the decisions to 0.2 seconds (5 decisions per second) as more frequent decisions may overwhelm human perception. Considering the z-score

normalization, ‘1’ and ‘-1’ represented one standard deviation above and below the mean performance, these values determined the two ends of performance (expert versus novice) of neurosurgical skill. This arrangement allowed not only detecting the two end levels of surgical performance but also the assessment of performance spectrum in between.

Algorithm design and AI training

Long-short term memory (LSTM) network is favorable for time-series performance analysis where long-term relations are important. [18-20] We utilized a supervised learning technique and designed our algorithm as a regression model. Our LSTM network was designed to minimize the computational burden (Supplementary Figure 1). The algorithm is composed of the first input sequence layer, two unidirectional LSTM layers, a fully connected layer, and a regression layer. Two dropout layers were used, after each LSTM layer, to help avoid overfitting. The number of nodes used for LSTM layers was calculated by adding one (1) to the number of input metrics (performance metrics). Sequence-to-sequence supervised learning was used. More complex designs can be developed, and the performance can be compared to our design. During the training, Adam (adaptive moment estimation) optimizer was utilized with a starting learning rate of 1e-3, decreased by x0.1 every 25 epochs. Minibatch size was 18, determined as the number of trials in the training set (108) divided by the number of repeats per person (6). Shuffling was used at every epoch. Training was performed with 1000 epochs monitoring root-mean-squared-error values visually (Supplementary Table 1), using NVIDIA GeForce GTX 660 (6.0Gbps).

Assessing trainee performance

The trained algorithm was used to make an assessment at 0.2-second intervals considering 16 performance metrics. Assessment was made as a continuous variable from ‘1’

expert level to '-1' novice level while any score above '1' or below '-1' was also allowed. The data from 24 neurosurgical trainee participants (six trials per participant) on 144 tasks was used to test the algorithm performance. An average score was calculated for each task and task scores were averaged across six trials for each participant.

Statistics

A one-way ANOVA and the post-hoc analysis were conducted to compare the average performance score of experts, senior trainees, junior trainees, and novices. A linear regression analysis was conducted to compare trainee average score to that trainee year of training. All data analysis, algorithm training and statistics were carried out using MATLAB (The MathWorks Inc.) release 2020a and IBM SPSS Statistics, Version 27 by codes written by the authors.

Providing coaching and risk assessment

Three algorithms were developed to provide expert level coaching related to (1) aspirator force utilization, (2) bipolar forceps force utilization, and (3) instrument tip separation distance, outputting these features. While making the predictions for expert level coaching, the expertise level was inputted as an expert '1' throughout the task. Two other algorithms had output predictions for bleeding and non-tumor tissue injury risks. While making the predictions for risk assessment, the expertise level was inputted aligned with the expertise level of the user (expert: '1', seniors: '0.33', juniors: '-0.33', medical student: '-1'). More detailed information about input and output features can be found at the Supplementary Table 2. A future study may address the testing and validation of coaching and risk detection modules of the ICEMS.

Data availability

A sample raw simulation data file is available online. [61]

Code availability

The codes written by the authors can be found online.

References

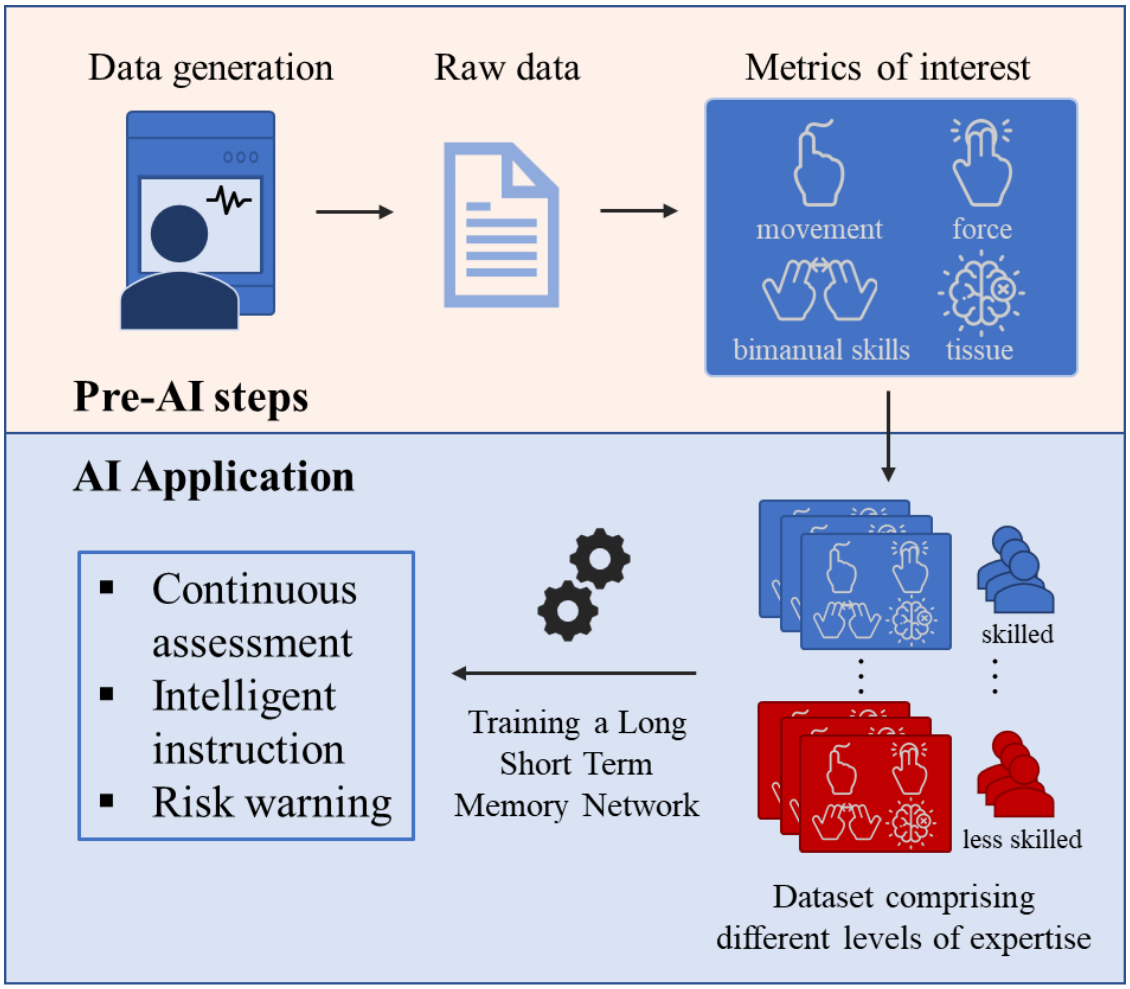
1. Anderson, O., Davis, R., Hanna, G.B. & Vincent, C.A. Surgical adverse events: a systematic review. *The American Journal of Surgery* **206**, 253-262 (2013).
2. Stulberg, J.J.*, et al.* Association Between Surgeon Technical Skills and Patient Outcomes. *JAMA Surgery* (2020).
3. Regenbogen, S.E.*, et al.* Patterns of Technical Error Among Surgical Malpractice Claims: An Analysis of Strategies to Prevent Injury to Surgical Patients. *Annals of Surgery* **246**, 705-711 (2007).
4. Gélinas-Phaneuf, N. & Del Maestro, R.F. Surgical Expertise in Neurosurgery: Integrating Theory Into Practice. *Neurosurgery* **73**, S30-S38 (2013).
5. Brightwell, A. & Grant, J. Competency-based training: who benefits? *Postgraduate Medical Journal* **89**, 107 (2013).
6. Ericsson, K.A. & Charness, N. Expert performance: Its structure and acquisition. *American Psychologist* **49**, 725-747 (1994).
7. Samuel, B.T., Benjamin, K.H. & Aaron, A.C.-G. Editorial. Innovations in neurosurgical education during the COVID-19 pandemic: is it time to reexamine our neurosurgical training models? *Journal of Neurosurgery JNS* **133**, 14-15 (2020).
8. Winkler-Schwartz, A.*, et al.* Machine Learning Identification of Surgical and Operative Factors Associated With Surgical Expertise in Virtual Reality Simulation. *JAMA Netw Open* **2**, e198363 (2019).
9. Mirchi, N.*, et al.* The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLOS ONE* **15**, e0229596 (2020).
10. Hebb, A.O., Yang, T. & Silbergeld, D.L. The sub-pial resection technique for intrinsic tumor surgery. *Surgical neurology international* **2**, 180-180 (2011).
11. Santiago, G.-R. & Hugues, D. Surgical management of World Health Organization Grade II gliomas in eloquent areas: the necessity of preserving a margin around functional structures. *Neurosurgical Focus FOC* **28**, E8 (2010).
12. Sabbagh, A.J.*, et al.* Roadmap for Developing Complex Virtual Reality Simulation Scenarios: Subpial Neurosurgical Tumor Resection Model. *World Neurosurgery* **139**, e220-e229 (2020).
13. Delorme, S., Laroche, D., DiRaddo, R. & Del Maestro, R.F. NeuroTouch: A Physics-Based Virtual Simulator for Cranial Microneurosurgery Training. *Operative Neurosurgery* **71**, ons32-ons42 (2012).
14. AlZhrani, G.*, et al.* Proficiency Performance Benchmarks for Removal of Simulated Brain Tumors Using a Virtual Reality Simulator NeuroTouch. *Journal of Surgical Education* **72**, 685-696 (2015).
15. Bugdadi, A.*, et al.* Automaticity of force application during simulated brain tumor resection: testing the Fitts and Posner model. *Journal of surgical education* **75**, 104-115 (2018).
16. Chan, J.*, et al.* A systematic review of virtual reality for the assessment of technical skills in neurosurgery. *Neurosurgical Focus* **51**, E15 (2021).
17. Norman, G.R.*, et al.* Expertise in Medicine and Surgery. in *The Cambridge Handbook of Expertise and Expert Performance* (eds. Williams, A.M., Kozbelt, A., Ericsson, K.A. & Hoffman, R.R.) 331-355 (Cambridge University Press, Cambridge, 2018).
18. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **9**, 1735-1780 (1997).
19. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444 (2015).

20. Lipton, Z.C., Berkowitz, J. & Elkan, C. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019* (2015).
21. Cohen, J. *Statistical power analysis for the behavioral sciences*, (Academic press, 2013).
22. Sawaya, R.*, et al.* Virtual Reality Tumor Resection: The Force Pyramid Approach. *Operative Neurosurgery* **14**, 686-696 (2017).
23. Winkler-Schwartz, A.*, et al.* Bimanual psychomotor performance in neurosurgical resident applicants assessed using NeuroTouch, a virtual reality simulator. *Journal of surgical education* **73**, 942-953 (2016).
24. Lohre, R.*, et al.* Effectiveness of Immersive Virtual Reality on Orthopedic Surgical Skills and Knowledge Acquisition Among Senior Surgical Residents: A Randomized Clinical Trial. *JAMA Network Open* **3**, e2031217-e2031217 (2020).
25. Seymour, N.E.*, et al.* Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Annals of surgery* **236**, 458-464 (2002).
26. Grantcharov, T.P.*, et al.* Randomized clinical trial of virtual reality simulation for laparoscopic skills training. *British journal of surgery* **91**, 146-150 (2004).
27. Winkler-Schwartz, A.*, et al.* Creating a Comprehensive Research Platform for Surgical Technique and Operative Outcome in Primary Brain Tumor Neurosurgery. *World Neurosurgery* (2020).
28. Ershad, M., Rege, R. & Fey, A.M. Adaptive Surgical Robotic Training Using Real-Time Stylistic Behavior Feedback Through Haptic Cues. *arXiv preprint arXiv:2101.00097* (2020).
29. Fekri, P., Dargahi, J. & Zadeh, M. Deep Learning-Based Haptic Guidance for Surgical Skills Transfer. *Frontiers in Robotics and AI* **7**(2021).
30. Vaughan, N. & Gabrys, B. Scoring and assessment in medical VR training simulators with dynamic time series classification. *Engineering Applications of Artificial Intelligence* **94**, 103760 (2020).
31. Sawaya, R.*, et al.* Development of a performance model for virtual reality tumor resections. *Journal of Neurosurgery JNS* **131**, 192-200 (2018).
32. Forestier, G.*, et al.* Surgical motion analysis using discriminative interpretable patterns. *Artificial Intelligence in Medicine* **91**, 3-11 (2018).
33. Chartrand, G.*, et al.* Self-directed learning by video as a means to improve technical skills in surgery residents: a randomized controlled trial. *BMC Medical Education* **21**, 91 (2021).
34. Sadeghi Esfahlani, S.*, et al.* Development of an Interactive Virtual Reality for Medical Skills Training Supervised by Artificial Neural Network. in *Intelligent Systems and Applications* (eds. Bi, Y., Bhatia, R. & Kapoor, S.) 473-482 (Springer International Publishing, Cham, 2020).
35. Castillo-Segura, P., Fernández-Panadero, C., Alario-Hoyos, C., Muñoz-Merino, P.J. & Delgado Kloos, C. Objective and automated assessment of surgical technical skills with IoT systems: A systematic literature review. *Artificial Intelligence in Medicine* **112**, 102007 (2021).
36. Charlin, B., Boshuizen, H.P.A., Custers, E.J. & Feltovich, P.J. Scripts and clinical reasoning. *Medical Education* **41**, 1178-1184 (2007).
37. Gioia, D.A. & Poole, P.P. Scripts in Organizational Behavior. *Academy of Management Review* **9**, 449-459 (1984).
38. Zareinia, K.*, et al.* A Force-Sensing Bipolar Forceps to Quantify Tool–Tissue Interaction Forces in Microsurgery. *IEEE/ASME Transactions on Mechatronics* **21**, 2365-2377 (2016).
39. Davids, J.*, et al.* Automated Vision-Based Microsurgical Skill Analysis in Neurosurgery Using Deep Learning: Development and Preclinical Validation. *World Neurosurgery* **149**, e669-e686 (2021).
40. Levin, M.*, et al.* Surgical data recording in the operating room: a systematic review of modalities and metrics. *British Journal of Surgery* (2021).
41. Birkmeyer, J.D.*, et al.* Surgical Skill and Complication Rates after Bariatric Surgery. *New England Journal of Medicine* **369**, 1434-1442 (2013).

42. Wen, Q.*, et al.* Time series data augmentation for deep learning: A survey. *arXiv preprint arXiv:2002.12478* (2020).
43. Fazlollahi, A.M.*, et al.* Effect of Artificial Intelligence Tutoring vs Expert Instruction on Learning Simulated Surgical Skills Among Medical Students: A Randomized Clinical Trial. *JAMA Network Open* **5**, e2149008-e2149008 (2022).
44. Biamonte, J.*, et al.* Quantum machine learning. *Nature* **549**, 195-202 (2017).
45. Fan, H., Jiang, J., Zhang, C., Wang, X. & Lai, Y.-C. Long-term prediction of chaotic systems with machine learning. *Physical Review Research* **2**, 012080 (2020).
46. Seoane, L.F. Evolutionary aspects of reservoir computing. *Philosophical Transactions of the Royal Society B: Biological Sciences* **374**, 20180377 (2019).
47. Mirchi, N., Ledwos, N. & Del Maestro, R.F. Intelligent Tutoring Systems: Re-Envisioning Surgical Education in Response to COVID-19. *Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques*, 1-3 (2020).
48. World Medical, A. World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. *JAMA* **310**, 2191-2194 (2013).
49. Winkler-Schwartz, A.*, et al.* Artificial Intelligence in Medical Education: Best Practices Using Machine Learning to Assess Surgical Expertise in Virtual Reality Simulation. *J Surg Educ* **76**, 1681-1690 (2019).
50. Cheng, A.*, et al.* Reporting guidelines for health care simulation research: extensions to the CONSORT and STROBE statements. *Advances in Simulation* **1**, 25 (2016).
51. Moons, K.G.*, et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* **162**, W1-73 (2015).
52. Vandenbroucke, J.P.*, et al.* Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *PLoS Med* **4**, e297 (2007).
53. Alotaibi, F.E.*, et al.* Assessing Bimanual Performance in Brain Tumor Resection With NeuroTouch, a Virtual Reality Simulator. *Operative Neurosurgery* **11**, 89-98 (2015).
54. Alotaibi, F.E.*, et al.* Neurosurgical Assessment of Metrics Including Judgment and Dexterity Using the Virtual Reality Simulator NeuroTouch (NAJD Metrics). *Surgical Innovation* **22**, 636-642 (2015).
55. Azarnoush, H.*, et al.* Neurosurgical virtual reality simulation metrics to assess psychomotor skills during brain tumor resection. *International Journal of Computer Assisted Radiology and Surgery* **10**, 603-618 (2015).
56. Azarnoush, H.*, et al.* The force pyramid: a spatial analysis of force application during virtual reality brain tumor resection. **127**, 171 (2016).
57. Khalid, B.*, et al.* Impact of acute stress on psychomotor bimanual performance during a simulated tumor resection task. *Journal of Neurosurgery JNS* **126**, 71-80 (2017).
58. Bissonnette, V.*, et al.* Artificial Intelligence Distinguishes Surgical Training Levels in a Virtual Reality Spinal Task. *J Bone Joint Surg Am* **101**, e127 (2019).
59. Mirchi, N.*, et al.* Artificial Neural Networks to Assess Virtual Reality Anterior Cervical Discectomy Performance. *Oper Neurosurg (Hagerstown)* (2019).
60. Brian, J.T., Marjorie, A.D. & Christina, D.M. Verification and validation of neural networks: a sampling of research in progress. in *Proc.SPIE*, Vol. 5103 (2003).
61. Yilmaz, R. SubPialResection101-KFMC_scenario.xml_2015-Oct-22_14h06m26s_log.csv. https://doi.org/10.6084/m9.figshare.15132507.v1. (2021).

Figures

Figure 1: Outline of the application. Raw data acquired from the simulator is used to calculate relevant features, metrics of interest. Data obtained from participants who are at different stages of expertise is used to train a LSTM network. The trained algorithm provided continuous assessment, intelligent instructions, or risk warnings, depending on the output feature selected. Multiple algorithms are trained to demonstrate potential applications of the ICEMS.



AI: artificial intelligence

Figure 2: Flow diagram. AI: artificial intelligence. One trial data belonging to a neurosurgeon was not available.

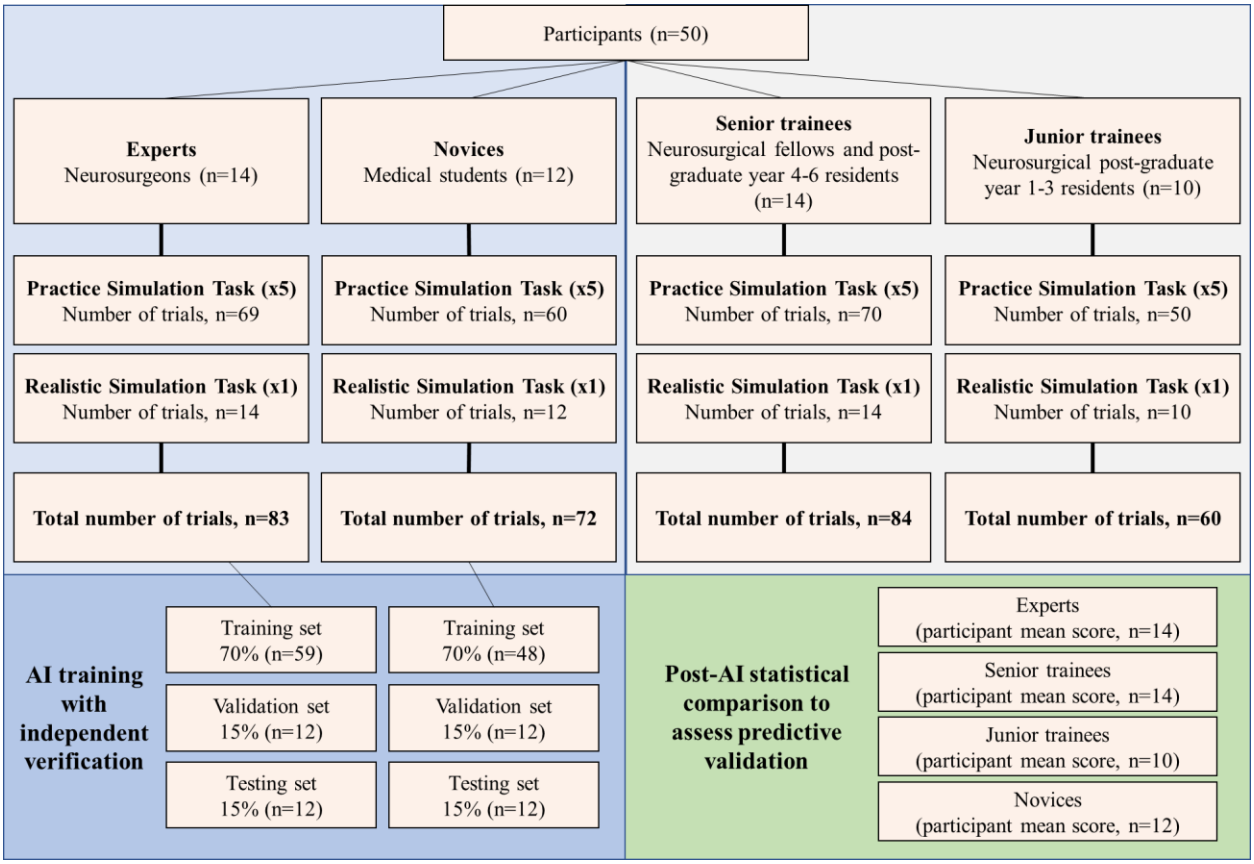


Figure 3: Simulated tumor resection tasks. Participants carried out two simulated tumor resection tasks, the simulated subpial tumor resection (**a**, **b** and **c**) 5 times and the simulated complex brain tumor operation (**d**, **e** and **f**) once, employing a simulated ultrasonic aspirator in the dominant hand and a simulated bipolar forceps in the non-dominant hand. Both instruments were activated by separate pedals. These tasks were designed with bleeding capacity to replicate the high-risk complex subpial brain tumor resection. (**f**) demonstrates cauterization using the bipolar forceps.

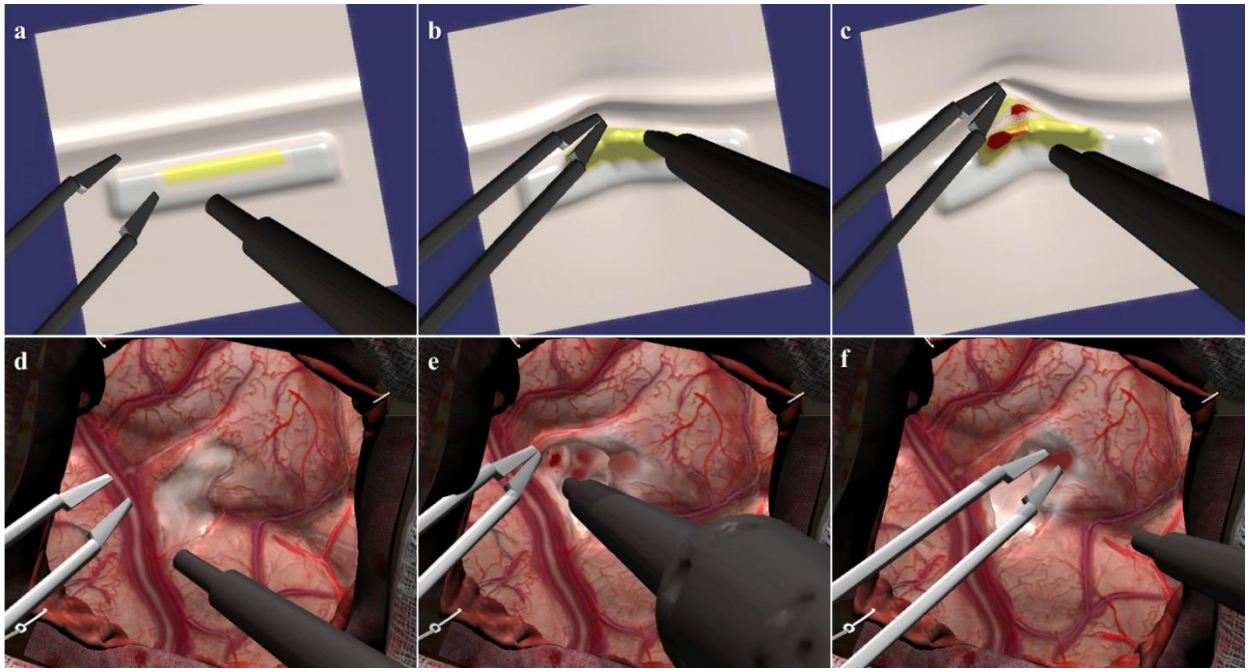


Figure 4: Performance metrics. Sixteen performance metrics from five categories: safety, quality, efficiency, bimanual cognitive and movement, were extracted from the raw data. An LSTM network was trained inputting the 16-performance metrics, predicting expertise. The LSTM network was structured as regression model to predict expertise as a continuous variable from 1 (expert) to -1 (novice). **Unit abbreviations:** N: Newton, mm: millimeter, t: time (0.02 seconds).

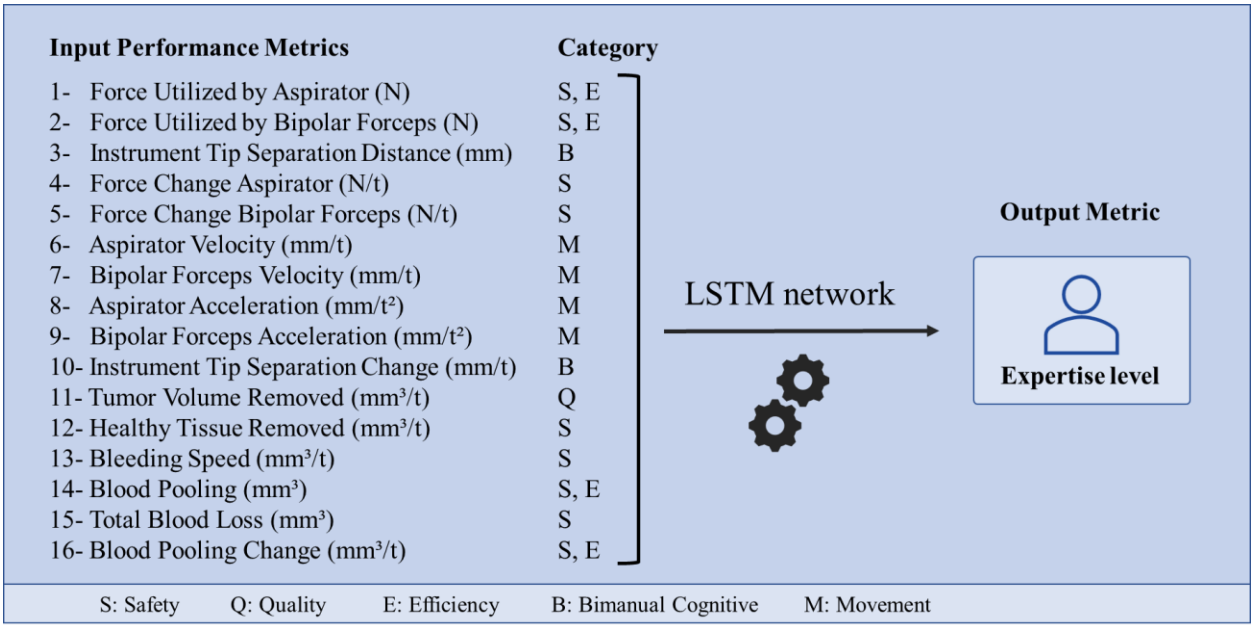


Figure 5: Average score of groups. When the performance of the participants was scored by the ICEMS, the average scores were: for experts (neurosurgeons, n=14) 0.509; 95% CI [0.424 0.593], for seniors (n=14) 0.258; 95% CI [0.114 0.402], for juniors (n=10) -0.11; 95% CI [-0.358 0.139], and for novices (medical students, n=12) -0.398; 95% CI [-0.545 -0.251]. Skilled and less skilled performance are represented in the y-axis by scores closer to '1' and '-1', respectively. Bars represent standard errors.

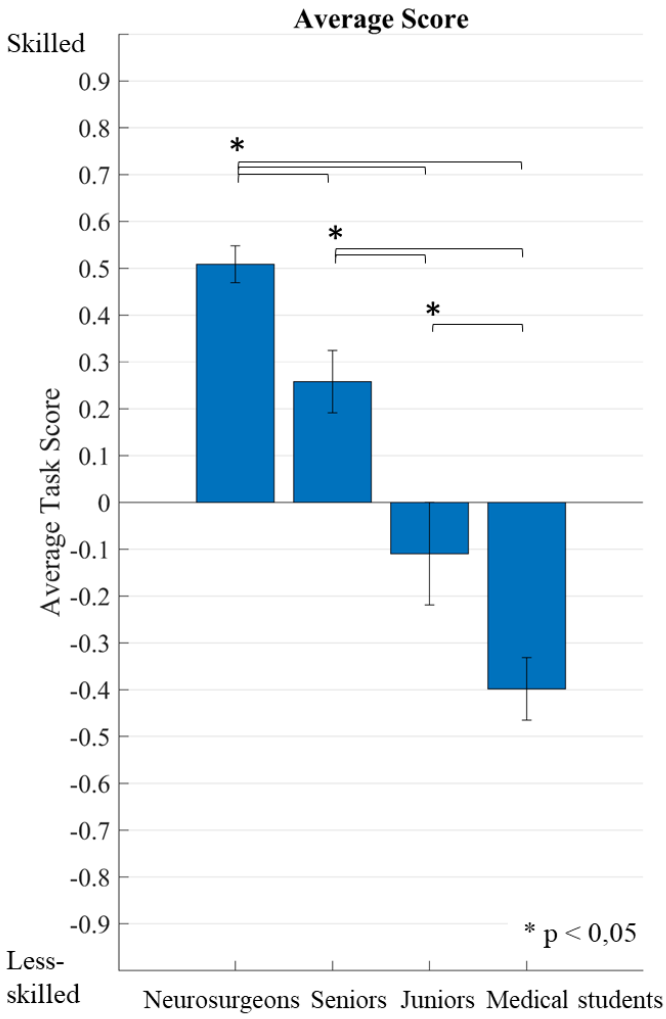
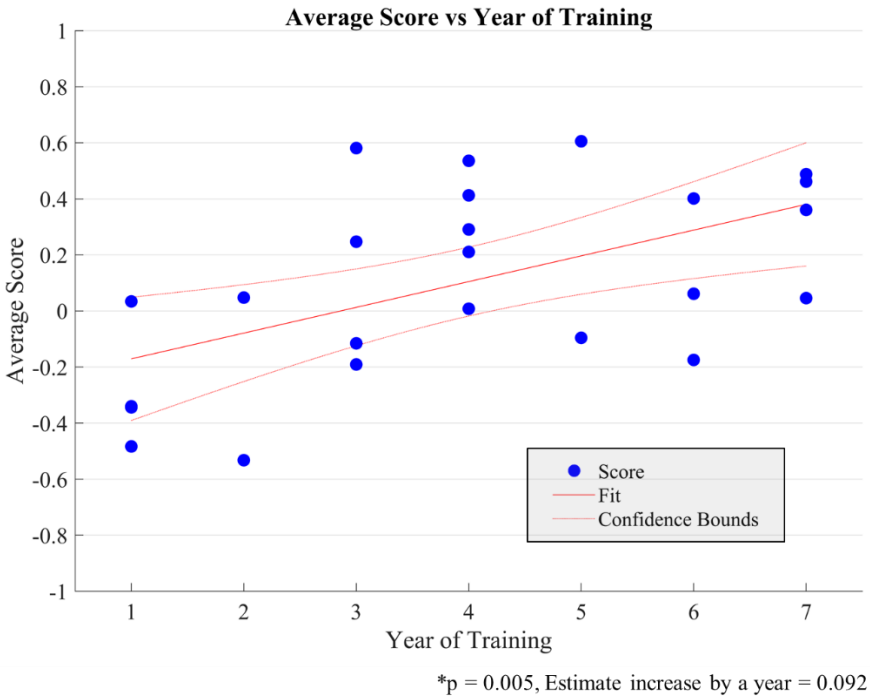


Figure 6: Average score versus year of training in neurosurgery. The average score yielded a significant correlation with the trainees' year of training (p = 0.005), increased by 0.092 per training-year, with a linear regression analysis. Blue dots represent the average score of each trainee, x axis represents year of training in neurosurgery. Resident participants' neurosurgery training program was six years. Neurosurgical fellows were considered in 7th year in training.



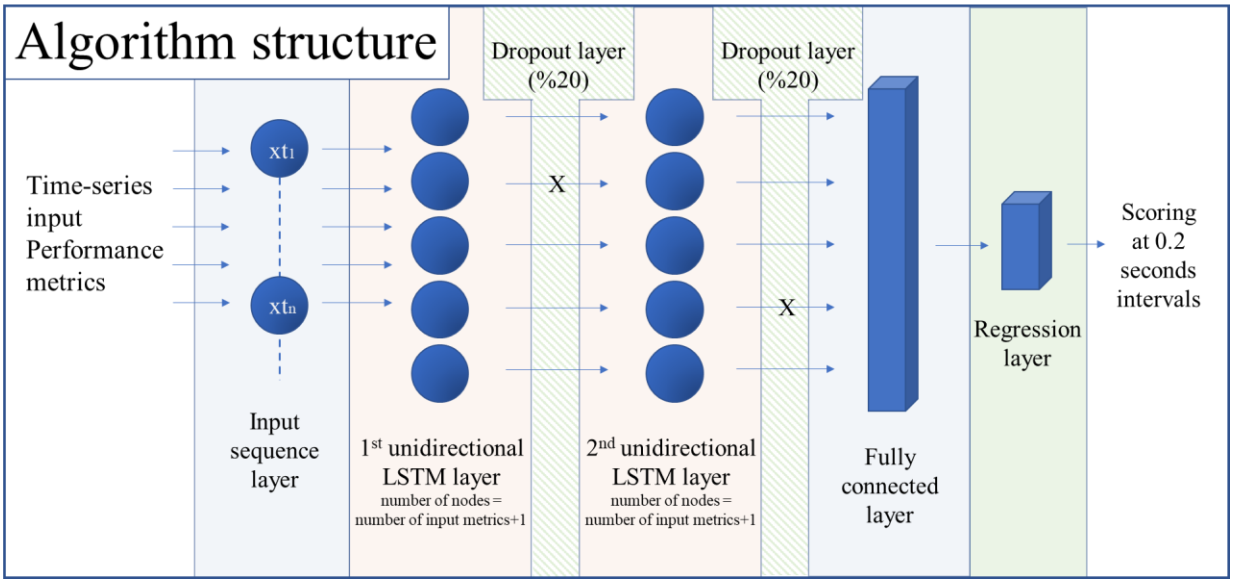
*p = 0.005, Estimate increase by a year = 0.092

Table: Residents' demographics. Twenty-four neurosurgical trainees participated in the study: 4 neurosurgical fellows, 10 senior residents (post-graduate year 4-6), 10 junior residents (post graduate year 1-3).

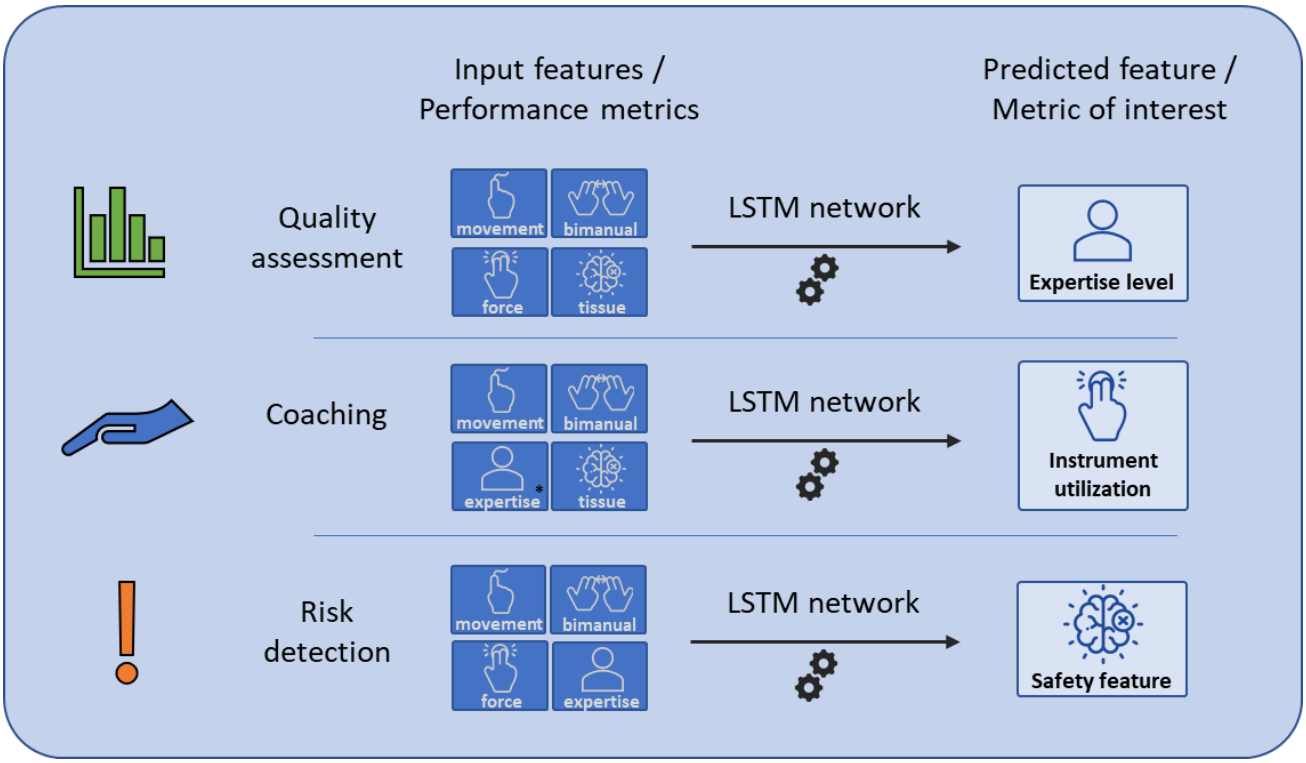
	Post Graduate Year of Training	Number of trainees
Neurosurgical Fellows	7	4
Neurosurgical Senior Residents	6	3
	5	2
	4	5
Neurosurgical Junior Residents	3	4
	2	2
	1	4
Total		24

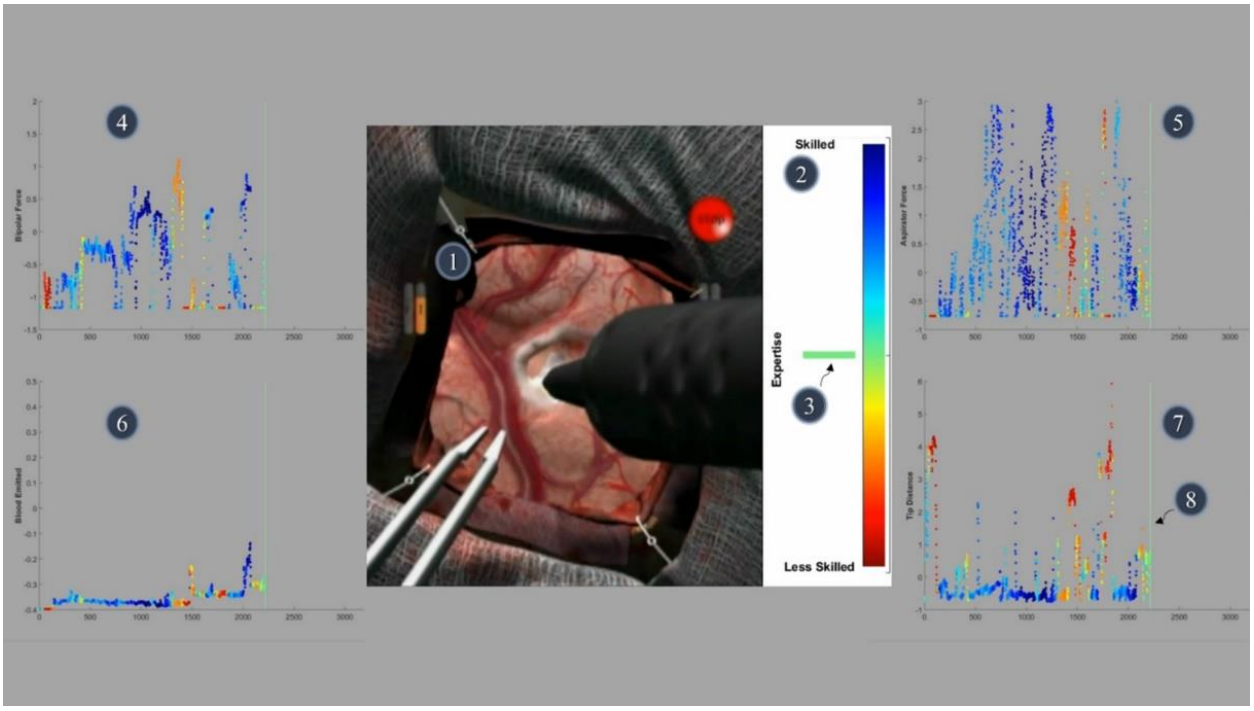
Supplementary information

Supplementary Figure 1: Algorithm structure.



Supplementary Figure 2: Applications of the ICEMS. Our system can be used for three applications. When the expertise level is defined as the output feature, a quality assessment of the performance can be made. When a feature relating instrument utilization or operative factor is output, coaching can be provided (*expertise is inputted as the expert level). When a safety metric is defined as the output, a risk detection algorithm can be developed.

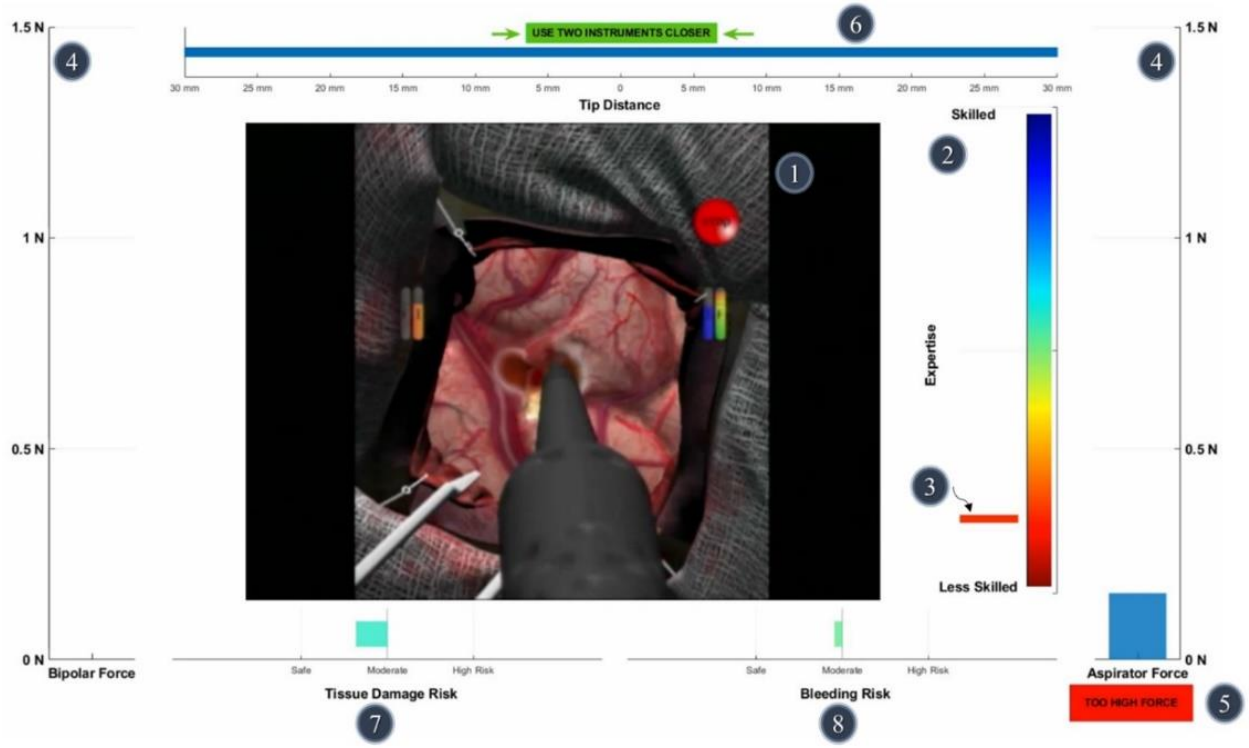




Supplementary Figure 3: Legend of Supplementary Video-1 and Supplementary Video-2.

This video represents the expertise assessment made by the ICEMS in relation to 4 of the 16 critical performance metrics inputted to the algorithm. Middle screen **(1)** shows the user view during the virtual reality surgical task. The color bar **(2)** represents the assessment made from skilled -blue- to less skilled -red- levels of expertise, shown by the colored indicator **(3)** at 0.2-second intervals. Four scatter plots, for each critical features including aspirator force **(5)**, bipolar force **(4)**, tip distance **(7)** and blood emitted **(6)**, represent how the expertise assessment relates to these metrics. In these graphs, each dot represents an expertise assessment made by the ICEMS by its color (according to the color bar **(2)**), at each 0.2-second intervals. Colored dots are drawn according to the expertise level determined by the algorithm as the time progress, same color as **(3)** and the colored time indicator **(8)**. x-axis show the number of decisions made. During this >10min task more 3000 assessments were made. y-axis show the z-score values for each performance metric. Higher values indicate higher force applied at **(4)** and **(5)** with bipolar and

aspirator, respectively. High values indicate high bleeding rate at **(6)** and instrument tip separation at **(7)**. The colored time indicator **(8)** proceeds on the y-axis.

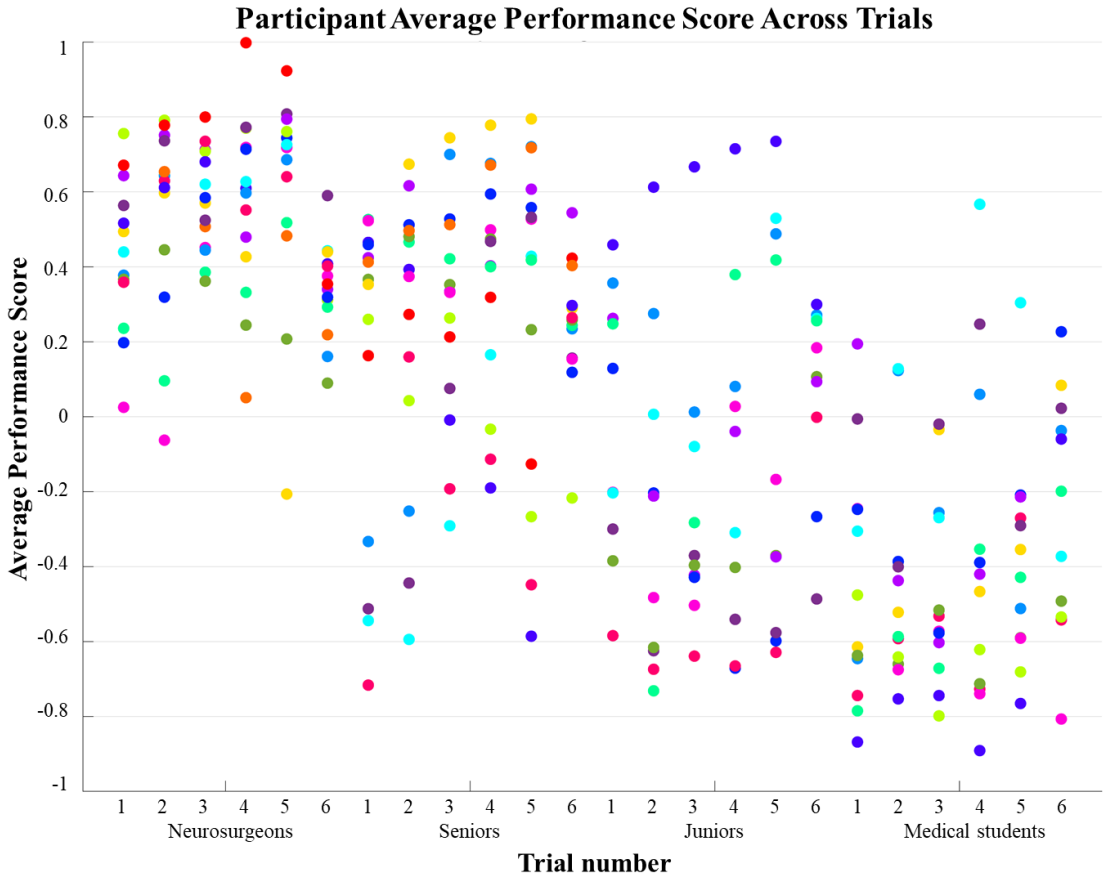


Supplementary Figure 4: Legend of Supplementary Video-3 and Supplementary Video-4.

The ICEMS composed of three modules: assessment, coaching and risk detection. Middle screen **(1)** shows the user view during surgical performance. The color bar **(2)** represents the assessment module where the assessment is made at 0.2-second intervals between skilled -blue- and less skilled -red- levels and shown by the colored indicator **(3)**. In this example coaching is provided for three critical metrics: aspirator force utilized, bipolar forceps force utilized and instrument tip separation distance. The bars **(4)** show the amount of force applied by bipolar **(4-left)** and aspirator **(4-right)**. Two background algorithms calculate the expected force applied for expert level instrument utilization. If the expected value is one standard deviation below the actual value a warning **(5)** 'too high force' is given. If the expected value is one standard deviation above the actual value a warning 'use bipolar/aspirator more efficiently' is given. The top bar **(6)** shows the distance between the tip of the two instruments. A background algorithm calculates the

expected tip separation distance for expert level instrument utilization. If the expected level is one standard deviation below the actual value, a warning ‘use two instruments together’ is shown. The risks related to two critical features were detected: tissue (healthy brain) damage risk **(7)**, and bleeding risk **(8)**. The moderate risk level equals the average risk achieved by all individuals within our dataset, where z-score=0. Higher values indicated behaviour with high risk and lower values indicated safe behaviour.

Supplementary Figure 5: Participant Average Performance Score Across Trials. X-axis represents the trial numbers from first to sixth repeat for each expertise group. Trial number 1 to 5 belongs to the practice trial while trial number 6 indicates the realistic scenario. Y-axis represents the average performance score. Participant scores at each task are indicated with a colored dot. The same color was utilized within the same expertise group for each participant. Data that belongs to a neurosurgeon for the fifth repeat was not available.



Supplementary Table 1: Root-mean-squared-error (RMSE) values obtained. A total of six algorithms were trained for assessment, coaching and risk detection. The training accuracy was monitored by root-mean-squared error (RMSE) values. During algorithm training, overfitting happens when the model fits a dataset too closely preventing accurate prediction on a new dataset (low generalizability). To avoid this problem the separate validation dataset was used to monitor the training progress. Training was acceptable when the RMSEs for training and validation datasets decreased in tandem and stay aligned by the end of the training. After the training was complete, the separate testing dataset was used to check the final state of the training. Having no gold standard, close values for training, validation and testing were targeted to help reject overfitting.

	TRAINING RMSE	VALIDATION RMSE	TESTING RMSE
Expertise	0.7065	0.7097	0.7525
Force Aspirator	0.8169	1.0648	0.8994
Force Bipolar	0.6859	0.7939	0.6647
Tip Distance	0.6611	0.7166	0.6786
Tissue Damage	0.8791	0.8377	0.7998
Bleeding	0.6910	0.6200	0.6099

RMSE: root-mean-squared-error

Supplementary Table 2: Input and output features (metric of interest) for each trained algorithm. Colors indicate the three categories of application: (1- green) expertise assessment, (2- blue) coaching, and (3- red) risk assessment.

Assessment made	Input features	Output feature	Explanation
Performance assessment	**1**, **2**, **3**, **4**, **5**, **6**, **7**, **8**, **9**, **10**, **11**, **12**, **13**, **14**, **15**, **16**	**0**- Expertise	This algorithm makes a global expertise assessment given the 16 input features.
Aspirator utilization	**0***, **2**, **3**, **5**, **6**, **7**, **8**, **9**, **10**, **11**, **12**, **13**, **14**, **15**, **16** *inputted as always '1'. Excluded **4** because the output feature is a parent feature.	**1**- Force Utilized by Aspirator	This algorithm predicts the amount of force utilization expected for expert level ('1'), for specific action being performed. Predicted values can be used to coach a trainee. The output metric represents a safety and efficiency feature. A trainee is expected use aspirator efficiently while avoiding high forces.
Bipolar forceps utilization	**0***, **1**, **3**, **4**, **6**, **7**, **8**, **9**, **10**, **11**, **12**, **13**, **14**, **15**, **16** *inputted as always '1'. Excluded **5** because the output feature is a parent feature.	**2**- Force Utilized by Bipolar Forceps	This algorithm predicts the amount of force utilization expected for expert level ('1'), for specific action being performed. Predicted values can be used to coach a trainee. The output metric represents a safety and efficiency feature. A trainee is expected use bipolar forceps efficiently while avoiding high forces.
Bimanual instrument utilization	**0***, **1**, **2**, **4**, **5**, **6**, **7**, **8**, **9**, **11**, **12**, **13**, **14**, **15**, **16** *inputted as always '1'. Excluded **10** because the output feature is a parent feature.	**3**- Instrument Tip Separation Distance	This algorithm predicts how close the two instruments should be during the action being performed for expert level ('1'). Predicted values can be used to coach a trainee. The output metric represents the bimanual cognitive. A trainee is expected use the tips of the two instruments closely together.
Bleeding risk detection	**0***, **1**, **2**, **3**, **4**, **5**, **6**, **7**, **8**, **9**, **10**, **11**, **12** *inputted aligned with the expertise level of the user (expert: '1', seniors: '0.33', juniors: '-0.33', medical student: '-1'). Excluded **14**, **15**, **16** because the output feature is a closely related feature.	**13**- Bleeding Speed	This algorithm predicts bleeding rate during the action being performed. The output metric represents a safety feature.
Tissue damage risk detection	**0***, **1**, **2**, **3**, **4**, **5**, **6**, **7**, **8**, **9**, **10**, **11**, **13**, **14**, **15**, **16** *inputted aligned with the expertise level of the user (expert: '1', seniors: '0.33', juniors: '-0.33', medical student: '-1'). No feature exclusion was made.	**12**- Healthy Tissue Removed	This algorithm predicts damage to the healthy surrounding tissue during the action being performed. The output metric represents a safety feature.
Number coded features: 0- Expertise, **1**- Force Utilized by Aspirator, **2**- Force Utilized by Bipolar Forceps, **3**- Instrument Tip Separation Distance, **4**- Force Change Aspirator, **5**- Force Change Bipolar Forceps, **6**- Aspirator Velocity, **7**- Bipolar Forceps Velocity, **8**- Aspirator Acceleration, **9**- Bipolar Forceps Acceleration, **10**- Instrument Tip Separation Distance Change, **11**- Tumor Volume Removed, **12**- Healthy Tissue Removed, **13**- Bleeding Speed, **14**- Blood Pooling, **15**- Total Blood Loss, **16**- Blood Pooling Change			

Supplementary Table 3: Trainee self-reported subpial resection operative experience and trainee average ICEMS scores. Trainees reported the number of subpial procedures they had been involved in, including epilepsy cases, and frontal, temporal, and occipital brain tumor surgical procedures. Right-most column shows the trainee average expertise score rated across six simulation trials by the ICEMS. Neurosurgical fellows were considered in 7th year in training.

	Trainee ID	Post Graduate Year of Training	Number of times assisted in subpial resection	Number of times carried out partial subpial resection	Number of times carried out complete subpial resection	Average Score
Neurosurgical Fellows	1	7	110	85	32	0.488
	2	7	4	4	4	0.461
	3	7	31	22	34	0.361
	4	7	36	4	6	0.046
Neurosurgical Senior Residents	5	6	0	0	2	-0.175
	6	6	24	16	12	0.401
	7	6	0	20	45	0.062
	8	5	24	20	7	-0.096
	9	5	133	130	5	0.605
	10	4	18	3	0	0.291
	11	4	2	1	0	0.413
	12	4	3	1	0	0.008
	13	4	57	30	10	0.536
	14	4	2	2	0	0.211
Neurosurgical Junior Residents	15	3	3	1	0	-0.191
	16	3	8	35	7	-0.115
	17	3	15	12	3	0.581
	18	3	0	0	0	0.247
	19	2	3	0	0	-0.532
	20	2	1	0	0	0.048
	21	1	0	0	0	-0.340
	22	1	0	0	0	0.034
	23	1	1	2	0	-0.344
	24	1	24	0	0	-0.483

Chapter 3 - Nondominant Hand Skills Spatial and Psychomotor Analysis During a Complex Virtual Reality Neurosurgical Task-A Case Series Study

Preface

Bimanual dexterity is an important skillset for surgeons to complete operative procedures efficiently and achieve desired outcomes. Non-dominant hand plays a role in assisting the dominant hand to optimize task execution and achieving hemostasis. This work outlined the development and construct validity of a spatial assessment system concerning non-dominant hand skills. The goals were 1- to demonstrate the spatial awareness of experts in their control of bipolar forceps instrument during a complex subpial brain tumor resection task, 2- to provide learners an objective demonstration of this skill to guide them to achieve expert-level non-dominant hand instrument utilization. The findings of this study helped the development of feedback systems in the randomized controlled trial in Chapter 4. The manuscript was published as:

Yilmaz R, Ledwos N, Sawaya R, Winkler-Schwartz A, Mirchi N, Bissonnette V, Fazlollahi AM, Bakhaidar M, Alsayegh A, Sabbagh AJ, Bajunaid K, Del Maestro R. Nondominant Hand Skills Spatial and Psychomotor Analysis During a Complex Virtual Reality Neurosurgical Task—A Case Series Study. Operative Neurosurgery. 2022 Jul 1;23(1):22-30. DOI: 10.1227/ons.0000000000000232

Abstract

BACKGROUND: Virtual reality surgical simulators provide detailed psychomotor performance data, allowing qualitative and quantitative assessment of hand function. The non-dominant hand plays an essential role in neurosurgery in exposing the operative area, assisting the dominant hand to optimize task execution, and hemostasis. Outlining expert level non-dominant hand skills may be critical to understand surgical expertise and aid learner training.

OBJECTIVES: (1) To provide validity for the simulated bimanual subpial tumor resection task, and (2) To utilize this simulation in qualitative and quantitative evaluation of non-dominant hand skills for bipolar forceps utilization.

METHODS: In this case-series study, 45 right-handed participants performed a simulated subpial tumor resection utilizing simulated bipolar forceps in the non-dominant hand for assisting the surgery and hemostasis. A 10-item questionnaire was used to assess task validity. The non-dominant hand skills across four expertise levels (neurosurgeons, senior trainees, junior trainees, and medical students) were analyzed by two visual models and performance metrics.

RESULTS: Neurosurgeon median (range) overall satisfaction with the simulated scenario was 4.0/5.0 (2.0-5.0). The visual models demonstrated a decrease in high force application areas on pial surface with increased expertise level. Bipolar-pia mater interactions were more focused around the tumoral region for neurosurgeons and senior trainees. These groups spent more time using the bipolar while interacting with pia. All groups spent significantly higher time in the left-upper pial quadrant than other quadrants.

CONCLUSIONS: This work introduces new approaches for the evaluation of non-dominant hand skills which may help surgical trainees by providing both qualitative and quantitative feedback.

Introduction

In surgery, the interaction of dominant and non-dominant hands is essential to accomplish operative goals. [1-3] This bimanual psychomotor performance can be constrained by the skill level of the non-dominant hand. [4-7] The mastery of non-dominant hand skills in assisting the dominant hand to optimize task execution, exposing operative regions and hemostasis is critical for learners to perform surgical tasks safely and efficiently. [8] In brain tumor surgery, understanding expert level non-dominant hand skills necessary to perform complex procedures such as the subpial resection is lacking. Outlining these skills is crucial in revealing the composites of surgical expertise to provide trainees with personalized feedback to help improve non-dominant hand skills.

This study first focused on assessing face and content validity of a simulated complex virtual reality subpial tumor resection scenario. Then, utilizing data from this simulation platform, we investigated the non-dominant hand skills essential for successful completion of the task. Visual and quantitative models developed in this work were used to explore differences in non-dominant hand skills between skilled and less-skilled groups. Our research questions were: (1) Do visual models regarding force and time utilization indicate differences in non-dominant hand skills between expertise groups? (2) How efficiently and precisely do skilled groups use their non-dominant hand in comparison to less-skilled groups during tumor resection? (3) What are some common non-dominant hand skill features that exist across skilled and less-skilled groups, and those acquired with increasing training and expertise?

Methods

Subjects

A consecutive case series of 50 participants from a single Canadian university enrolled in this retrospective study between March 2015-May 2016 at a single time point, with no follow-up. Data was anonymized. Five left-handed were removed from the analysis due to differing instrument utilization between right- and left-handed participants[3,9] (Figure-1). The remaining 45 right-handed participants were classified a-priori as neurosurgeons (13), seniors (3 neurosurgical fellows and 9 senior residents (post graduate year 4-6)), juniors (9 junior residents (post graduate year 1-3)) and medical students (11) (Table-1). All participants signed a consent form approved by the university ethics board.

Simulation Scenario

The NeuroVR platform (CAE Healthcare, Canada) allowed users to interact with a 3D-operative environment through a microscope while providing haptic feedback on contact with the simulated tissues (Figure-2, 2A). [10-13] The simulated scenario involved a previously described complex brain tumor subpial resection procedure (Figure-2, 2B). [14,15] The tumor was placed under pia mater, adjacent to critical brain areas such as a main blood vessel, and motor and sensory strips (Figure-2, 2F). The task was performed utilizing a simulated ultrasonic aspirator in the dominant hand to resect tumor, and a bipolar forceps in the non-dominant hand for supportive movements such as lifting or moving pia mater to assist the dominant hand, and cauterizing bleeding tissues (Figure-2, 2D and 2E, see video). Both instruments were activated by foot pedals. Background sounds of mechanical ventilation and heartbeats were included. Participants were instructed to remove the tumor completely within thirteen minutes[15] while

minimizing bleeding and damage to adjacent structures. No feedback was provided during or after each task.

Participant Rating of the Task

After task completion, participants completed a 10-item questionnaire to assess the face and content validity of the subpial resection (Table-2). Participants were asked to rate their neurosurgical simulation experience and satisfaction on a 5-point Likert-scale. A median score of ≥3.0 was deemed sufficient for the face and content validity and overall satisfaction. [16]

Performance Data

The NeuroVR platform provides a csv (comma-separated-value) file containing the coordinates of instrument tip location, instrument activation, force application, tumor and tissue volumes, bleeding, and instrument-tissue contacts. Force application and distance were measured in Newtons (N) and millimeters (mm), respectively. Data was recorded at 20-millisecond increments (50 recordings/second).

Force Heatmap and Time Scatter Models

3D tumor and brain mesh models were extracted from the simulator software. Brain pia mater surface was divided into four quadrants, numbered counter-clockwise starting from the right upper quadrant (Q1, Q2, Q3, Q4) [17] with the center of the tumor, represented on the pial surface as the reference point (Figure-2, 2C). Visualization of bipolar force application on pia mater was provided by two models: (1) 3D Force heatmap was created to represent bipolar force application on pia mater, (2) Time scatter was generated to demonstrate spatial distribution of bipolar-pia mater interactions. For both models, mean values (force or time) were recorded at each grid point (pixel) on pia mater for each task. These values were averaged for each group to

generate heatmap and scatter models for four groups. Force and time quantities were represented with color scales. For the time scatter model, only grid points with bipolar contact were shown.

Performance Metrics

Bipolar non-dominant hand skills were assessed by two groups of metrics. The first group focused on performance while using the bipolar to assist the dominant hand while resecting tumor. Metrics in this category included total time spent interacting with pia mater, average force application on pial surface, total force application on pial surface, bipolar average tip distance from the center reference, and bipolar precision. The bipolar precision metric was based on the standard error values of bipolar tip distance from the center reference, assessing the distance variation of the bipolar-pia mater interactions.

The second group metrics explored non-dominant hand utilization in the four pial quadrants during the entire task. At each quadrant, three performance metrics were calculated; percentage time spent while the bipolar was in contact with pial surface, average force application and total force application.

Statistical Analyses

The first group of metrics were compared between four expertise groups. The second group of metrics were compared between four quadrants within each expertise groups. Outliers were observed by boxplot. No data exclusion was made. Normality of data distribution was determined by Shapiro-Wilk test for each metric ($p>.05$). Statistical analysis was performed by one-way ANOVA for normally distributed metrics. Levene's test was used to check equality of variances, based on median ($p>.05$). One-way ANOVA was followed by Tukey-Kramer or Games-Howell post-hoc tests in case of equal and unequal variances, respectively. For non-

normally distributed metrics, Kruskal-Wallis test was followed by Dunn's (1964) post-hoc procedure with Bonferroni correction for multiple tests. $P<0.05$ was considered statistically significant. MATLAB (MathWorks) r.2021a and IBM SPSS Statistics v.27 were used for the analyses. This study is reported in line with the PROCESS Guideline. [18]

Results

The participant demographics and previous simulation experience are outlined in Table-1.

Rating of the Task

The simulated subpial tumor resection median scores and ranges for face and content validity on a 5-point Likert-scale as well as participants' satisfaction with the simulated task are listed in Table-2. Neurosurgeons rated the overall visual realism of the simulated task a median [range] of 4.0 [1.0-5.0]. The sensory realism was rated 3.0 [2.0-5.0] with the feel of the simulated pia rated (4.0 [2.0-4.0]) higher than the simulated tumor (3.0 [2.0-4.0]). Neurosurgeons also agreed with using the simulator for technical skills training (4.0 [2.0-5.0]) with 85% recommending integrating simulation into training. These results were consistent with face and content validity. [16]

Force Heatmap and Time Scatter Models

In the 3D force heatmap model, all groups had high bipolar pia force application in Q2, the left upper pial region (Figure-3). Higher force applied areas are shown in red. Neurosurgeons and seniors demonstrated smaller red areas when compared to junior and medical student groups. In the bipolar time scatter model, red areas represented grid points with average time spent greater than one-second. All groups had red areas mainly at the left upper quadrant.

Neurosurgeons and seniors demonstrated more focused bipolar-pia-mater-interactions around the tumoral region. Results from both models indicate that higher skilled groups were using their non-dominant hand gentler (lower force) and more centralized around the tumor. These parameters were statistically investigated with the performance metrics in the next sections.

Psychomotor Analysis

Quantitative analysis demonstrated that the neurosurgeons ($p=.048$) and seniors ($p=.047$) spent significantly more bipolar time interacting with pia than medical students, and these groups applied less average bipolar force to pial surface than the medical student group ($p=.039$ and $p<.001$, respectively) (Figure-4). Seniors' average force application to pial surface was significantly lower than juniors' ($p=.015$). Total bipolar force application on pial surface was not significantly different between groups (Figure-4).

Subpial tumor resection necessitates using the bipolar in contact with the pia mater around the tumor operational field. Bipolar contacts on the pia were more centralized for neurosurgeons within 12 millimeters radius from the tumor center while in other groups, many bipolar contacts outside of this radius were identified (Figure-5, 5A). The average bipolar distance from the center reference increased from neurosurgeons to less-skilled groups and was significantly smaller in neurosurgeons than medical students ($p=.020$, Figure-5, 5B). In addition, neurosurgeons were more precise with their bipolar contacts on pia (lower distance variation from the center reference) during tumor resection compared to medical students ($p=.017$, Figure-5, 5C).

Quadrant Metrics

Non-dominant hand instrument utilization was evaluated for each quadrant by three different metrics. In terms of time spent, the most favorable quadrant was Q2 for all groups having a statistically higher time spent than other quadrants. Neurosurgeons were the only group who spent time in Q3 as much as Q2, with no statistical significance between these two quadrants (p=.107). Neurosurgeons spent significantly more time in Q3 than Q4 (p=.023) (Figure-6).

All groups had the highest average force application in Q2 however, for juniors and medical students no significant differences between quadrants were found (Figure-6). Average force application was significantly higher in Q2 than Q4 for neurosurgeons, and it was significantly higher in Q2 than Q1 for seniors. A significantly higher total force application was observed in Q2 than in any other quadrant for all groups, except that for neurosurgeons there was no statistical difference between Q2 and Q3. Neurosurgeons applied significantly more total force at Q3 than Q4 (p=.027).

Discussion

Virtual reality neurosurgical simulation, with haptic feedback, provides large datasets allowing a quantitative assessment of simulated surgical performance. [12,14] These datasets have been utilized to evaluate surgical bimanual psychomotor performance metrics and differentiate skilled and less-skilled groups. [1,3,8,9,19,20] The goal of these applications was to determine the critical bimanual technical skills that underlie expert performance and to create personalized feedback systems that allow surgical instructors develop formative and summative educational platforms.

This study focused on non-dominant hand skills involved in bipolar forceps utilization during a complex simulated subpial tumor resection task and the validation of this virtual reality surgical procedure. Neurosurgeons' rating of overall visual, sensory and color realism with the scenario was consistent with face and content validity and this simulation's construct validity has previously been demonstrated. [21]

Our visual models and quantitative analysis have shown that skilled groups use the bipolar more focally around the tumoral region with lower force application. Average force application decreased from medical students to seniors where seniors had significantly less average bipolar force application than both juniors and medical students. Neurosurgeons applied slightly higher forces than seniors and significantly lower forces than medical students. This discontinuous technical skill pattern (increase in force after a decreasing trend) may be caused by overly cautious phases involve during training. Neurosurgeons may continue modulating their instrument utilization after training and slightly increase their forces with increasing competence. Total bipolar pial force application was not statistically different between groups. Less-skilled groups had higher average force application over less time resulting in similar total force application to skilled groups who used lower average forces over longer times. Similarly in previous studies, lower instrument force utilization reflected higher expertise level. [9,22,23] All groups contacted the bipolar at the left upper pial region the most, indicating that this quadrant may offer the most ergonomic hand position.

Intraoperative surgery involves many components encompassing knowledge, judgement, and technical skills. Surgeons develop and maintain a muscle memory, defined as 'psychomotor skills script', which allows for being more prepared and efficient in the operating room. [24] Our results demonstrated this script exists with the non-dominant hand in the neurosurgeon group,

where they utilize bipolar more precisely and focused around the tumoral region. A previous study in laparoscopic radical prostatectomy outlined three deficiencies noted between trainees and experienced surgeons, one of which is the lack of synchronized non-dominant hand movements. [25] Virtual reality simulators allow for outlining less-skilled technical skills and provide trainees feedback to improve. The cognitive load during training can be decreased by mastering non-dominant hand skills separately. Trainees may more quickly obtain expert level psychomotor skills by spatial awareness and self-modifying their movements with unlimited repetitions in risk-free simulated environments. [26,27] By revealing expert level skills, this study enables providing trainees with both visual and quantitative feedback to master their non-dominant hand skills. The efficacy of such automated feedback methods can be compared to traditional expert mediated instruction.

Focused non-dominant hand skill training may provide significant advantages. A laparoscopic randomized controlled trial demonstrated that non-dominant hand training improved dominant hand function skills, a phenomenon known as intermanual transfer of skill learning. [7] Functional near-red spectroscopy investigations of non-dominant hand use resulted in bilateral sensory-motor cortex activation while dominant hand use only localized to the contralateral hemisphere, suggesting non-dominant hand training activates critical bilateral brain regions involved in motor control which may improve dominant hand motor system function. [28] Other investigations demonstrate a greater competence of the non-dominant limb/hemisphere to rely on sensory input, thus improving motor function. [29] Further studies outlining the interdependence of dominant and non-dominant hand function in surgery seem warranted. Previously, a support vector machine algorithm was used to differentiate expertise in subpial resection procedures in which 3 of the 4 performance metrics selected by the algorithm were

related to bipolar utilization (mean acceleration, maximum force and instrument tip separation). [2] In another study, 16 of the total of 31 performance metrics chosen by four machine learning algorithms to differentiate expertise into four groups were related to bipolar use, [1] consistent with the important role of the non-dominant hand in neurosurgical procedures.

Virtual reality simulation is a cutting-edge technology; however, these systems fail to represent many elements of the dynamic operating room environment. In this study users were not able to change the view angle, or the instruments. These limitations may affect participants' surgical performance and their conception of realism. A critical bipolar skill to be mastered is cauterization of bleeding vessels. Due to limits in data acquisition skills such as cauterization, and tissue related analyses such as pial retraction and deformation were not studied in this study. As simulation platforms advance and provide more detailed real-life interactions and data, more comprehensive assessments can be done. Our study included a post-hoc analysis with metrics representing an overall assessment of the performance. Ongoing works focus more action-oriented assessments of non-dominant hand skills using advance methodologies, such as deep learning. [30] Instrument utilization differ between right- and left-handed individuals carrying out virtual reality procedures, where instrument utilization of left-handed participants is usually a mirrored version of those of right-handers. [3,9] Having small number left-handed participants excluded from the study prevented exploring their bipolar-pia interactions. The small cohort involved from one institution may have limited the detection of differences between groups involving some metrics. With a broader cohort the generalization of results can be increased.

Conclusion

This work introduces novel qualitative and quantitative approaches for outlining the non-dominant hand skills involved during a virtual reality tumor resection. These visual models and performance metrics provide objective assessment of technical skills. Such systems may aid in the future development of competency-based training curricula.

References

1. 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.
2. 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.
3. Sawaya R, Bugdadi A, Azarnoush H, et al. Virtual Reality Tumor Resection: The Force Pyramid Approach. *Operative Neurosurgery.* 2017;14(6):686-696.
4. Peters M. Prolonged practice of a simple motor task by preferred and nonpreferred hands. *Perceptual and Motor Skills.* 1976;43(2):447-450.
5. Peters M. Handedness: effect of prolonged practice on between hand performance differences. *Neuropsychologia.* 1981;19(4):587-590.
6. Haaland E, Hoff J. Non-dominant leg training improves the bilateral motor performance of soccer players. *Scandinavian journal of medicine & science in sports.* 2003;13(3):179-184.
7. Nieboer TE, Sari V, Kluivers KB, Weinans MJN, Vierhout ME, Stegeman DF. A randomized trial of training the non-dominant upper extremity to enhance laparoscopic performance. *Minimally Invasive Therapy & Allied Technologies.* 2012;21(4):259-264.
8. Alotaibi FE, AlZhrani GA, Sabbagh AJ, Azarnoush H, Winkler-Schwartz A, Del Maestro RF. Neurosurgical Assessment of Metrics Including Judgment and Dexterity Using the Virtual Reality Simulator NeuroTouch (NAJD Metrics). *Surgical Innovation.* 2015;22(6):636-642.
9. Azarnoush H, Siar S, Sawaya R, et al. The force pyramid: a spatial analysis of force application during virtual reality brain tumor resection. 2016;127(1):171.
10. AlOtaibi F, Al Zhrani G, Bajunaid K, Winkler-Schwartz A, Azarnoush H. Assessing Neurosurgical Psychomotor Performance: Role of Virtual Reality Simulators, Current and Future Potential. SOJ Neurol 2 (1), 1-7. *Assessing Neurosurgical Psychomotor Performance: Role of Virtual Reality Simulators, Current and Future Potential.* 2015.
11. Azarnoush H, Alzhrani G, Winkler-Schwartz A, et al. Neurosurgical virtual reality simulation metrics to assess psychomotor skills during brain tumor resection. *International Journal of Computer Assisted Radiology and Surgery.* 2015;10(5):603-618.
12. Delorme S, Laroche D, DiRaddo R, Del Maestro RF. NeuroTouch: A Physics-Based Virtual Simulator for Cranial Microneurosurgery Training. *Operative Neurosurgery.* 2012;71(suppl_1):ons32-ons42.
13. Gélinas-Phaneuf N, Choudhury N, Al-Habib AR, et al. Assessing performance in brain tumor resection using a novel virtual reality simulator. *International journal of computer assisted radiology and surgery.* 2014;9(1):1-9.
14. Sabbagh AJ, Bajunaid KM, Alarifi N, et al. Roadmap for Developing Complex Virtual Reality Simulation Scenarios: Subpial Neurosurgical Tumor Resection Model. *World Neurosurgery.* 2020;139:e220-e229.
15. 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.
16. Schout BMA, Hendrikx AJM, Scheele F, Bemelmans BLH, Scherpbier AJJA. Validation and implementation of surgical simulators: a critical review of present, past, and future. *Surgical Endoscopy.* 2010;24(3):536-546.

17. Sawaya R, Alsideiri G, Bugdadi A, et al. Development of a performance model for virtual reality tumor resections. *Journal of Neurosurgery JNS.* 2018;131(1):192-200.
18. Agha RA, Sohrabi C, Mathew G, et al. The PROCESS 2020 Guideline: Updating Consensus Preferred Reporting Of CasE Series in Surgery (PROCESS) Guidelines. *International Journal of Surgery.* 2020;84:231-235.
19. Bissonnette V, Mirchi N, Ledwos N, Alsidieri G, Winkler-Schwartz A, Del Maestro RF. Artificial Intelligence Distinguishes Surgical Training Levels in a Virtual Reality Spinal Task. *J Bone Joint Surg Am.* 2019;101(23):e127.
20. Mirchi N, Bissonnette V, Ledwos N, et al. Artificial Neural Networks to Assess Virtual Reality Anterior Cervical Discectomy Performance. *Oper Neurosurg (Hagerstown).* 2019.
21. Yilmaz R, Winkler-Schwartz A, Mirchi N, Reich A, Del Maestro R. O51: ARTIFICIAL INTELLIGENCE UTILIZING RECURRENT NEURAL NETWORKS TO CONTINUOUSLY MONITOR COMPOSITES OF SURGICAL EXPERTISE. *British Journal of Surgery.* 2021;108(Supplement_1).
22. Sugiyama T, Lama S, Gan LS, Maddahi Y, Zareinia K, Sutherland GR. Forces of tool-tissue interaction to assess surgical skill level. *JAMA surgery.* 2018;153(3):234-242.
23. Sugiyama T, Gan LS, Zareinia K, Lama S, Sutherland GR. Tool-Tissue Interaction Forces in Brain Arteriovenous Malformation Surgery. *World Neurosurgery.* 2017;102:221-228.
24. Bugdadi A, Sawaya R, Olwi D, et al. Automaticity of force application during simulated brain tumor resection: testing the Fitts and Posner model. *Journal of surgical education.* 2018;75(1):104-115.
25. Gupta R, Cathelineau X, Rozet F, Vallancien G. Feedback from operative performance to improve training program of laparoscopic radical prostatectomy. *J Endourol.* 2004;18(9):836-839.
26. Kaufman DM, Mann KV. Teaching and learning in medical education: how theory can inform practice. *Understanding medical education: Evidence, theory and practice.* 2010;5:57-60.
27. McGaghie WC. Mastery Learning: It Is Time for Medical Education to Join the 21st Century. *Academic Medicine.* 2015;90(11):1438-1441.
28. Lee SH, Jin SH, An J. The difference in cortical activation pattern for complex motor skills: A functional near- infrared spectroscopy study. *Scientific Reports.* 2019;9(1):14066.
29. Bravi R, Cohen EJ, Martinelli A, Gottard A, Minciacchi D. When Non-Dominant Is Better than Dominant: Kinesiotape Modulates Asymmetries in Timed Performance during a Synchronization-Continuation Task. *Frontiers in Integrative Neuroscience.* 2017;11(21).
30. Yilmaz R, Winkler-Schwartz A, Reich A, Del Maestro R. SP2.1.1Continuous Monitoring and Assessment of Surgical Technical Skills Using Deep Learning. *British Journal of Surgery.* 2021;108(Supplement_7).

Tables and Figures

Table-1. Demographics of Participants. Represented formula: Mean +/- Standard Deviation (Range), *PGY: Post graduate year.

	Neurosurgeons (n=13)	Senior Trainees (Neurosurgical fellows and residents *PGY 4-6, n=12)	Junior Trainees (Neurosurgical residents *PGY 1-3, n=9)	Medical Students (n=11)
Mean age	45 +/- 7.5 (33-59)	32 +/- 2.3 (29-35)	30 +/- 3.3 (37-38)	24 +/- 1.3 (23-26)
Men/women	13/0	12/0	7/2	5/6
Number of complete subpial resections performed	210 +/- 286 (0-800)	10 +/- 14 (0-45)	0.7 +/- 2 (0-7)	0 +/- 0
Number of partial subpial resections performed	27 +/- 82 (0-300)	20 +/- 35 (0-130)	4 +/- 11 (0-35)	0 +/- 0
Used simulator previously	6 (46%)	9 (75%)	5 (56%)	3 (27%)
Used Neuro-VR previously	5 (38%)	8 (67%)	5 (56%)	1 (9%)

Table-2. Participant Rating of the Simulated Subpial Resection Task (median (range)).

	Neurosurgeons	Senior Trainees	Junior Trainees
Sensory realism of the 'feel' of the simulated **pia** (1-completely unrealistic, 5-completely realistic)	4.0 (2.0 – 4.0)	3.0 (1.0 – 4.0)	4.0 (2.0 – 4.0)
Sensory realism of the 'feel' of the simulated **tumor** (1-completely unrealistic, 5-completely realistic)	3.0 (2.0 – 4.0)	3.0 (1.0 – 4.0)	4.0 (3.0 – 4.0)
Color of the simulated tumor (1-completely unrealistic, 5-completely realistic)	4.0 (2.0 – 5.0)	4.0 (3.0 – 5.0)	4.0 (3.0 – 5.0)
Overall visual realism of the simulation task (1-completely unrealistic, 5-completely realistic)	4.0 (1.0 – 5.0)	4.0 (1.0 – 5.0)	4.0 (2.0 – 5.0)
Overall sensory realism (the feel of the different tissues) of this simulation task (1-completely unrealistic, 5-completely realistic)	3.0 (2.0 – 5.0)	3.0 (1.0 – 4.0)	3.0 (2.0 – 4.0)
If this simulator was available in my program, I would use this simulation scenario for training of the technical skills simulated. (1-completely disagree, 5-completely agree)	4.0 (2.0 – 5.0)	4.0 (2.0 – 5.0)	4.0 (1.0 – 5.0)
Would you recommend integrating simulation training (using virtual reality operative simulation) into a curriculum during neurosurgery training program as a mandatory block? (YES/NO)	85% YES	67% YES	78% YES
Difficulty of the simulated tumor resection scenario (1- very easy, 5-very hard)	4.0 (3.0 – 5.0)	3.0 (2.0 – 4.0)	4.0 (2.0 – 5.0)
Self-rating of performance on the simulated scenario on a scale of 5 (1-very poor, 5-excellent)	2.5 (1.0 – 4.0)	3.0 (2.0 – 4.0)	3.0 (1.0 – 3.0)
Overall satisfaction with the simulated task (1-completely unsatisfied, 5-completely satisfied)	4.0 (2.0 – 5.0)	4.0 (2.0 – 4.0)	4.0 (1.0 – 5.0)

Figure-1. Study participants.

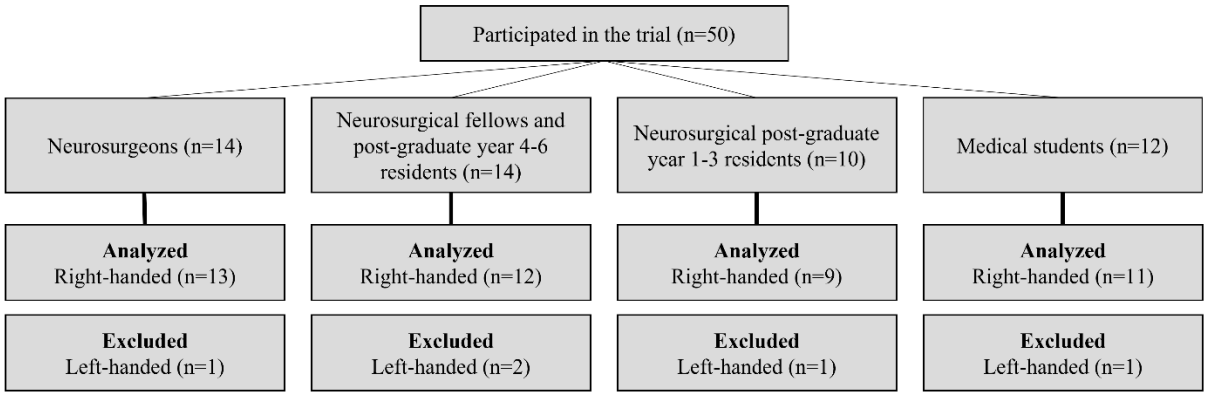


Figure-2. Scenario setting. **A-** The NeuroVR platform. **B-** User view of the scenario. Whitish color tumor is present in the simulated motor cortex with the simulated Rolandic Vein separating the motor and sensory cortex, major blood vessel lies very close to the tumor at the left side. **C-** 3D model of the tumor, from the top view, with four quadrants in counter-clockwise order. The tumor (middle) and blood vessel (red) lie under pia mater. **D-** The instruments. Users interact with the virtual environment with the simulated ultrasonic aspirator held in the dominant hand to resect tumor, and the bipolar held in the non-dominant hand to assist in exposing the tumor and to cauterize possible bleeding points. Both instruments were activated by foot pedals. **E-** A demonstration of bipolar use. Participants are required the pull the pia mater off the tumor from the brain-tumor edge to allow access the tumor regions beneath. **F-** An illustration of use of a real ultrasonic aspirator and bipolar on the simulated tumor resection scenario- 3D side view. The tumor (grey) and blood vessel (red) are seen under the pia mater (blue).

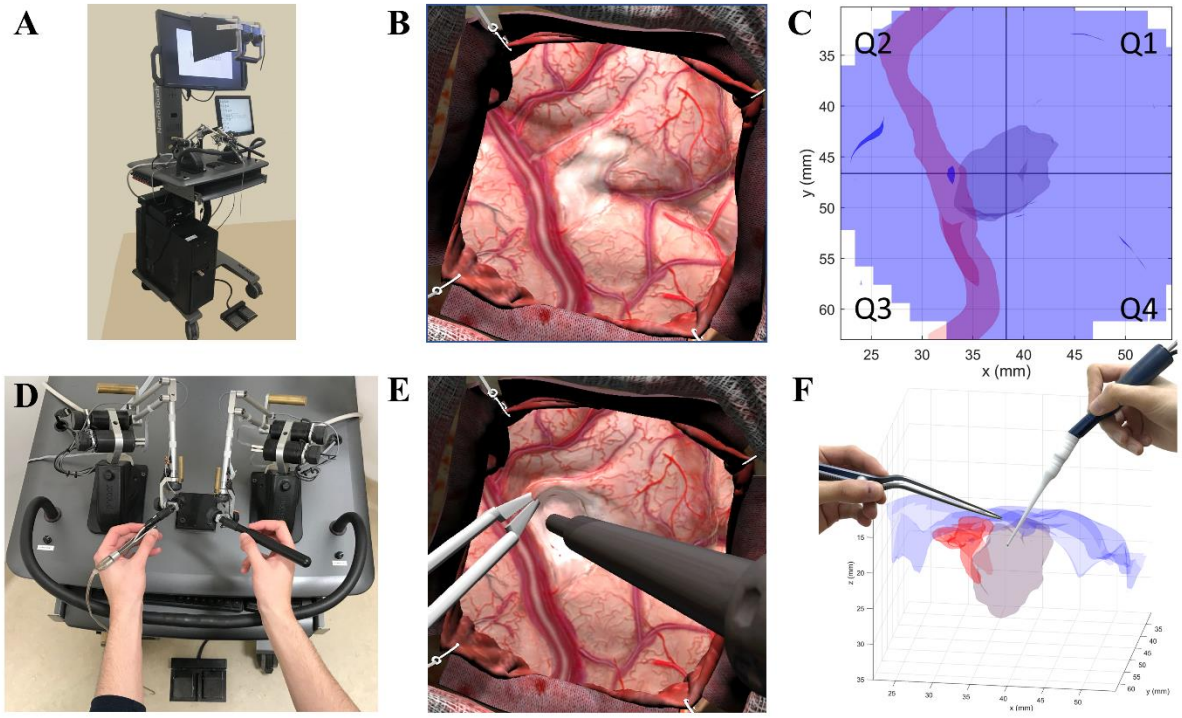


Figure-3. Spatial distribution of bipolar-pia mater interactions. **A-** Force heatmap: blue surface represents the pia mater. Force application (Newtons) is averaged across participants within each group and shown in four quadrants according to the color scale. **B-** Time scatter: represents two-dimensional (x-y) bipolar-pia matter interactions. Each grid point is colored according to the average time spent (second) when the instrument is in contact with the pia mater at that location. Only points in which time spent is greater than zero are shown.

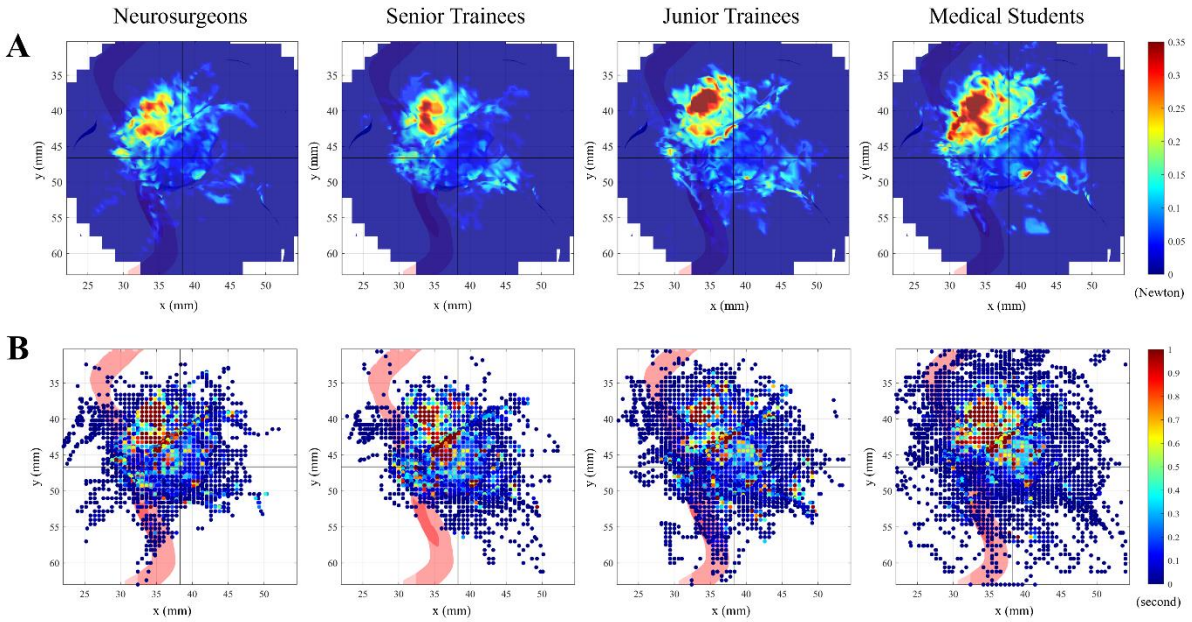


Figure-4. Psychomotor evaluation. The performance metrics included: time spent on pia (seconds), average force application on pia (Newtons), and total force application on pia (Newtons) while resecting tumor. Groups were neurosurgeons (NS, n=13), senior trainees (ST, n=12), junior residents (JT, n=10), medical students (MS, n=11). Values represent mean. Bars represent standard errors. Horizontal lines represent statistically significant differences ($p<0.05$).

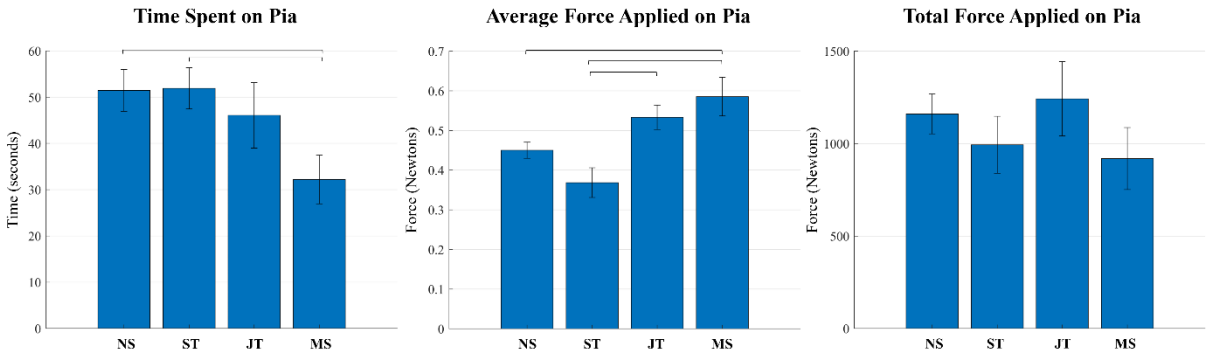


Figure-5. Bipolar precision analysis. **A-** Bipolar tip distance from the center for each group. X dimension represents mm distance from the center reference, Y dimension represents the percentage of being on a particular distance range (0.1mm). **B-** Average bipolar distance (mm) from the center of the tumor for each group. **C-** Precision of bipolar use: Standard error of bipolar distance from the center during the time while the tumor is being resected was calculated to find precision for each individuals' bipolar function. * $p<0.05$. NS: Neurosurgeons, ST: Senior trainees, JT: Junior trainees, MS: Medical students.

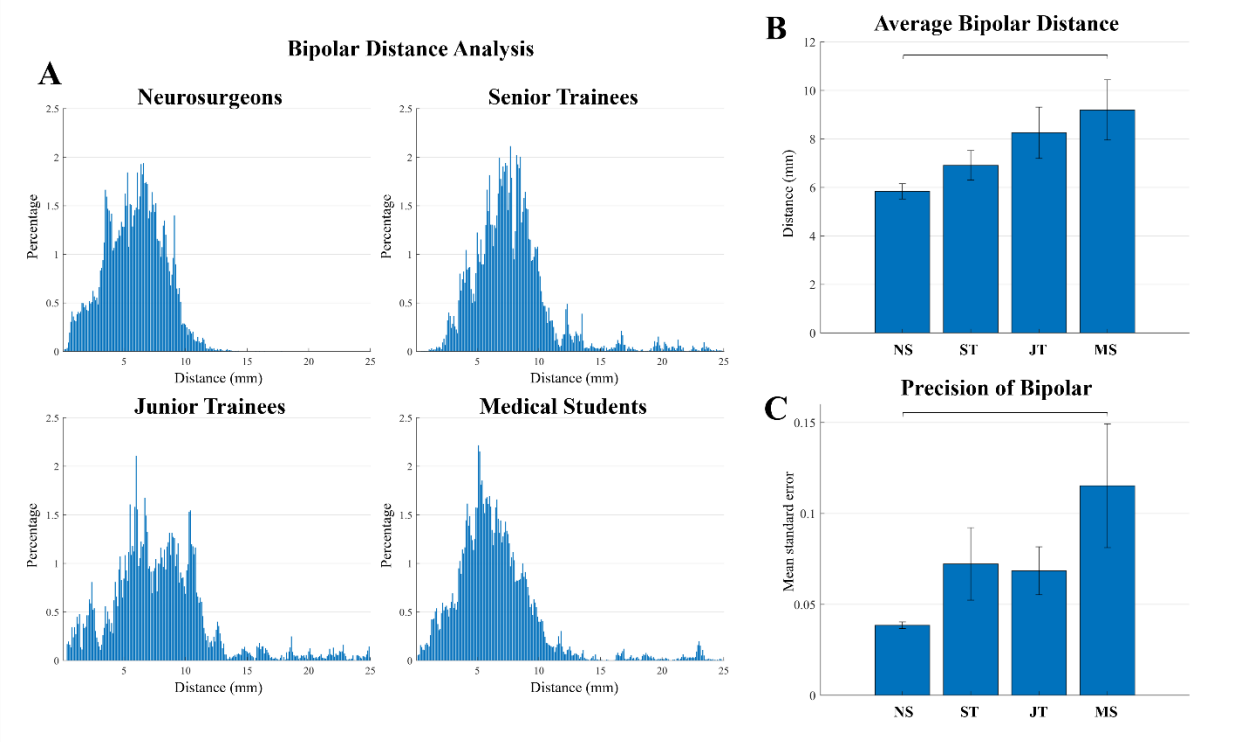
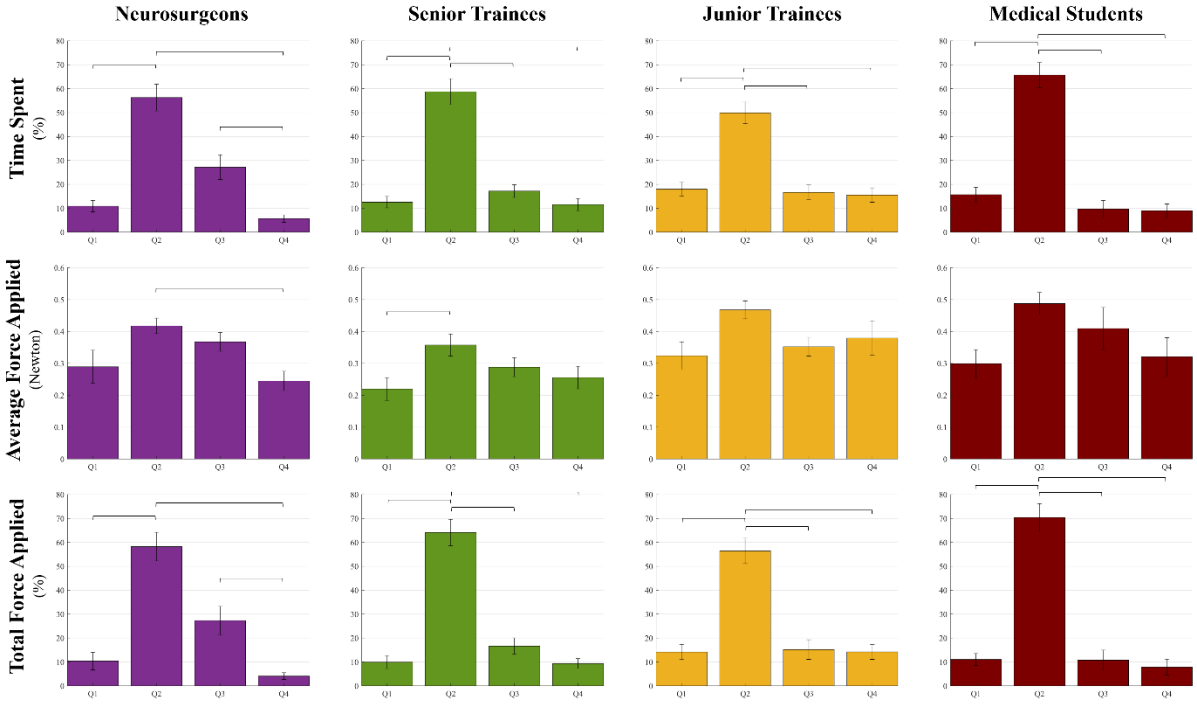


Figure-6. Quadrant metrics. Percentage time spent (%), average force application (Newtons), and total force application (percentagewise) were calculated per quadrant. Values represent mean. Bars represent standard errors. Horizontal lines indicate statistically significant differences ($p<0.05$).



Video legend. The simulated subpial tumor resection task. Users interact with the 3D environment. The whitish area at the center represents the tumor, which is adjacent to critical brain areas such as the main blood vessel at left hand side, and motor and sensory strips around. Tissues had bleeding capacity. Ultrasonic aspirator, at the dominant hand, was used to remove the tumor while bipolar, at the non-dominant hand, was used to assist the dominant hand and cauterize bleeding tissues. Sounds of mechanical ventilation and heart monitor were included.

Chapter 4 – Effect of Feedback Modality on Simulated Surgical Skills Learning using Automated Educational Systems– A Four-Arm Randomized Control Trial

Preface

Optimizing feedback information is essential to maximize efficacy in teaching surgical technical skills. This randomized controlled trial was the first feedback trial we have conducted at the Neurosurgical Simulation and Artificial Intelligence Learning Centre, integrated with the NeuroVR platform, to understand the benefits of providing feedback in comparison to no feedback and compare between different feedback modalities. This study integrated the spatial feedback information outlined in Chapter 3 in addition to the colored feedback. The results of this study would inform the future feedback applications in surgical simulation training that would come with the integration of artificial intelligence to increase training engagement of the trainees and improve their learning outcomes. The manuscript was published as:

Yilmaz, R., Fazlollahi, A. M., Winkler-Schwartz, A., Wang, A., Hassan Makhani, H., Alsayegh, A., Bakhaidar, M., Huy Tran, D., Santaguida, C., Del Maestro, R. F. Effect of Feedback Modality on Simulated Surgical Skills Learning Using Automated Educational Systems– A Four-Arm Randomized Control Trial. Journal of Surgical Education (2023).

Abstract

Objective: To explore optimal feedback methodologies to enhance trainee skill acquisition in simulated surgical bimanual skills learning during brain tumor resections.

Hypotheses: 1- Providing feedback results in better learning outcomes in teaching surgical technical skill when compared to practice alone with no tailored performance feedback. 2- Providing more visual and visuospatial feedback results in better learning outcomes when compared to providing numerical feedback.

Design: A prospective four-parallel-arm randomized controlled trial.

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

Participants: Medical students (n=120) from four Quebec medical schools.

Results: Participants completed a virtually simulated tumor resection task five times while receiving one of four feedback based on their group allocation: (1) practice-alone without feedback, (2) numerical feedback, (3) visual feedback, and (4) visuospatial feedback. Outcome measures were participants' scores on 14-performance metrics and the number of expert benchmarks achieved during each task. There were no significant differences in the first task which determined baseline performance. A statistically significant interaction between feedback allocation and task repetition was found on the number of benchmarks achieved, $F(10.558, 408.257) = 3.220$, $p<.001$. Participants in all feedback groups significantly improved their performance compared to baseline. The visual feedback group achieved significantly higher number of benchmarks than the practice-alone group by the third repetition of the task, $p=0.005$,

95%CI [0.42 3.25]. Visual feedback and visuospatial feedback improved performance significantly by the second repetition of the task, p=0.016, 95%CI [0.19 2.71] and p=0.003, 95%CI [0.4 2.57], respectively.

Conclusion: Simulations with autonomous visual computer assistance may be effective pedagogical tools in teaching bimanual operative skills via visual and visuospatial feedback information delivery.

Introduction

In medical education, advancing educational technologies promise to support trainee learning. [1] Among these, computer-assisted tools, such as artificial intelligent tutors, emerged as appropriate candidates to guide independent learning, and some offered advantages over traditional learning. [2] In surgical education, simulation platforms equipped with automated feedback systems allow learners to practice their bimanual surgical skills in a risk-free environment without the need for supervision. [3, 4] This liberates instructors' time to be invested in other aspects of patient care or surgical education such as mentorship. A key technical advantage of these computer-assisted systems is their ability to differentiate the expertise level of surgeons with granularity and precision. [3, 5] This not only presents new perspectives to understand the composites of expertise, but increases efficiency in trainee learning by providing quantifiable learning objectives, for which specific feedback and actionable goals can be directed to improve performance. [2] In addition, these systems can provide trainees with detailed visuospatial information about their bimanual performance which may increase their three-dimensional appreciation of surgical performance on anatomical structures. [6]

In medical education, extensive research is conducted to design effective curricula. [7-9] Teaching methodologies focus on increasing trainee engagement in learning while the students efficiently master their skills. Although quantifying surgical bimanual skills serves the purpose of providing objective feedback, this data can be presented to learners in a variety of formats such as numerical, visual, spatial, video, haptic and auditory. [3, 10, 11] However, because of the relative recency of these educational tools in surgical simulation training, more research is needed to evaluate the effectiveness of various feedback modalities to maximize efficiency in teaching technical skills. This randomized control trial investigated the effect of four feedback protocols

including numerical, visual, and visuospatial feedback along with practice alone with no tailored performance feedback, as a control, to evaluate the rate of acquisition of technical skills of medical students. The objectives were: 1- To explore the effect of feedback to the learning rate in surgical simulation training in comparison with practice without feedback. 2- To determine how more visual and spatial feedback modalities compare with numerical feedback.

Methods

Setting

This four-parallel arm randomized controlled trial (trial registration: ISRCTN17590019) was conducted at the Neurosurgical Simulation and Artificial Intelligence Learning Centre, McGill University, Montreal, Canada. Medical students in first to fourth year from four universities in the Province of Quebec were invited to participate in the trial. Data was collected between July 2019 – October 2020, in 60-minute simulation sessions with no follow up. One hundred and twenty medical students participated in the trial, and no exclusion criteria were applied. No changes were made to the methods after trial commencement. An online random number generator was used to determine participant group allocation. Study procedures were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Declaration of Helsinki. [12] COVID-19 public health measurements and the Montreal Neurological Institute and Hospital's protocols were followed to ensure participants' and researchers' safety during the conduct of the study. The time frame of the trial was predetermined with no restrictions on the number of simulation sessions that could take place. The trial participation was terminated with the restrictions imposed by changes to public health protocols due to COVID-19 pandemic in October 2020 while the number of participants

sufficed a statistical power of .99 for between- within-group interaction. This study was approved by the McGill University Health Centre Research Ethics Board, Neurosciences-Psychiatry. An approved consent form was signed by all study participants before trial participation. All participants filled a pre-questionnaire related to demographics and previous simulation experience and surgical exposure (Table-1). A post-questionnaire was completed after the trial for the rating of the simulation learning (Supplementary Table-2). This report adheres to guidelines for the reporting of multi-arm parallel group randomized trials, extension of the CONSORT 2010 Statement. [13] Study interventions involved no harm to participants. Participants were informed that their information will be anonymized, and despite the careful measures taken to avoid the chance that they may be identified, their trial performance would have no influence on their academic evaluation.

Simulation setting

The NeuroVR (previously NeuroTouch) neurosurgical simulation platform (CAE Healthcare, Montreal, Canada) with haptic feedback was utilized. [14] The haptic feedback integrated in the instrument handles was to provide a more realistic experience for all participants of the study regardless of the feedback interventions they receive for learning. This haptic technology allowed the integration of learning feedback on instrument force utilization for the study groups as trainees interact with delicate tissues during the simulated performance. The simulated task was previously developed to replicate the subpial resection of brain tumors. [15] Participants performed this simulated subpial tumor resection task five times with five minutes given to complete each task. The simulated scenario included the subpial resection of a yellow rectangular tumor (Figure-2) using a simulated ultrasonic aspirator and bipolar forceps to completely remove the tumor within the time limit while minimizing damage to the surrounding

tissue which mimics the adjacent normal gyrus. [5, 16] Both instruments were activated using pedals. Part of the tumor was placed under healthy brain tissue where lifting this simulated pial layer using the bipolar was necessary to gain access and remove the remaining underlying tumor. A blood vessel was incorporated into the simulation adjacent to the distal tumor wall and bleeding resulted from injury to this vessel. Bleeding was controlled utilizing the cauterizing function of the bipolar forceps (Figure-2e). The NeuroVR platform recorded performance data in 20-millisecond increments (50 recording per second) involving time, the information of force applied by the two instruments, instrument tip location, amount of tissue and tumor removed, amount of bleeding, and pedal activation.

Expert level benchmarks

Expert level benchmarks were developed using previously validated 14-performance metrics [3, 5, 17], described in the Results section. The data used to develop these benchmarks was previously available in our center and was recorded during 14 neurosurgeons' performance on the same simulated tumor resection task. Using this dataset, expert mean, and standard deviation values were calculated for each performance metrics to define the limits of the expert level benchmark. A metric score between one standard deviation above and below the mean was considered within the benchmark for that task.

Feedback setting

Four feedback protocols included (1) practice alone with no tailored performance feedback, (2) numerical feedback, (3) visual feedback, and (4) visuospatial feedback. All participants received standard verbal and written instructions before the start of the trial including how to use the simulator handles to carry out the simulated procedure and the feedback

information they would be provided with. All participants were also informed concerning the 14-performance metrics that would be used to assess their performance. The data recorded by the simulator was used to calculate participants' metric scores and determine whether they are within the benchmarks. Participants were given five-minutes between the tasks either to rest or receive the feedback information corresponding their group allocation. After each task, participants in Group-1 (n=30) received no tailored performance feedback. In Group-2 (n=30), participants received a printed copy of their performance scores on the 14 metrics that was compared with expert level benchmarks (Supplementary Figure-1). Any performance score falling above or below the expert benchmark was indicated with a letter 'H' (higher) or 'L' (lower), respectively. In Group-3 (n=29), participants received a screen-based graphical representation of their performance scores on the 14 metrics. The graphics were green colored for each performance metrics if participant's score was within the benchmark, yellow if their score was between one and two standard deviations of the benchmark, or red if their score was outside two standard deviations of the benchmark (Supplementary Figure-2). The graphics were also represented in purple for any performance score that was better than the benchmark. Participants in Group-4 (n=31) received the same colored-graphical demonstration but additionally, they were shown two 3D spatial models that showed the anatomical structures of the tumor and pial surface. The amount of force applied on these tissues by the ultrasonic aspirator and the bipolar were shown according to the color scale ranging from red to blue, where red indicated a higher force applied (Supplementary Figure-3). For all groups, the number of benchmarks achieved was calculated across five repetitions of the task. Automated feedback during the trial, data analysis and visualization were performed using MATLAB (The MathWorks Inc.) release 2021a. All codes were written by the authors.

Hypotheses

(1) Participants in feedback groups will achieve significantly higher number of benchmarks than those who practice without feedback. (2) Participants who receive visual and visuospatial feedback will achieve significant improvement earlier across the five repetitions of the task than those who receive only numerical information.

Statistical analysis

A priori sample size calculation, with a statistical power of 0.9, an effect size of 0.3, a correlation of 0.5 among repeated measures, and an alpha error probability of 0.05 for between groups comparison yielded a requirement of 25 participants in each group, and 100 participants in total. The participation of 120 students provided an achieved statistical power of .95. Two-way mixed ANOVA explored the interaction of feedback group assignment (between-groups) and task repetition (within-groups) on participants number of benchmarks achieved. There were no outliers, as assessed by visual examination of studentized residuals for values greater than ±3. Data was normally distributed, as visually assessed by Normal Q-Q Plot. Levene's test showed homogeneity of variances, based on median ($p>.05$), and Box's test demonstrated homogeneity of covariances, $p=.948$. Mauchly's test of sphericity indicated that the assumption of sphericity was violated for the two-way interaction, $\chi2(9) = 34.92$, $p < .001$. The results with Greenhouse-Geisser correction are reported. Differences between-feedback groups were investigated using one-way ANOVA. Within-feedback group differences were analyzed using one-way repeated measures ANOVA. Between-feedback group posthoc analyses were done using Tukey HSD or Games Howell tests depending on the homogeneity or heterogeneity of variances, respectively. Within-group posthoc analyses were done using Bonferroni posthoc tests. Cohen's d effect sizes

were reported for post hoc comparisons. [18] The variable 'number of benchmarks achieved' was assumed as a ratio variable, having the meaningful zero point (no success). As such, our analyses were done using parametric statistical tests described above. Non-parametric equivalent statistical analysis was also reported in the supplementary data (Supplementary Figure-4). Statistical analysis was done using IBM SPSS Statistics, Version 27.

Results

Participants

Participants' average age (mean [SD, min-max]) was 23.1 [3.6, 18-44] years and participant handedness was 108/10/2 (right-handed/left-handed/ambidextrous) (Table-1). Five participants previously used virtual reality simulation.

Data and performance metrics

Data from 120 participants, from a total of 600 trials, was available for analysis (Figure-1, Flow diagram) and there was no missing data. Participants' performance progress was tracked across five repetitions of the task on 14-performance metrics from four categories (1) safety, (2) quality, (3) efficiency, and (4) bimanual cognitive. Safety category included six metrics: (1) brain volume removed (cc), (2) amount of blood loss (cc), (3) maximum force applied with dominant hand (N), (4) maximum force applied with non-dominant hand (N), (5) sum of forces applied with dominant hand (N), (6) sum of forces applied with non-dominant hand (N). Quality category included only (7) tumor percentage removed. Efficiency category included 4 metrics: (8) total tip path length dominant hand (mm), (9) total tip path length non-dominant hand (mm), (10) path length index, and (11) efficiency index. Bimanual cognitive category included (12)

average instrument tips separation distance (mm), (13) coordination index, and (14) bimanual forces ratio. Descriptions of the performance metrics can be found on Supplementary Table-1.

Learning curves

No statistical difference was found between groups at baseline performance (p=0.121). There was a statistically significant interaction between the feedback group allocation and the number of repetitions of the task on the number of benchmarks achieved, F (10.558, 408.257) = 3.220, $p < .001$, effect size (partial $\eta 2$) = .077, $\varepsilon = .88$ (Figure-3). Group-3 made the quickest improvement where the number of benchmarks achieved was significantly higher than Group-1 by the third repetition of the task (p=0.005, 95%CI [0.42 3.25], effect size (Cohen's d)=0.878). Group-4 outperformed Group-1 by the fourth repetition of the task (p=0.002, 95%CI [0.54 3.00], effect size=1.035) while Group-2 did not outperform Group-1 within the five repetitions. In the final repetition of the task, Group-4 achieved 9.19 ± 1.66 (mean ± standard deviation) of the 14 benchmarks, Group-3 achieved 9.10 ± 1.82, Group-2 achieved 8.40 ± 2.06 while Group-1 achieved 7.30 ± 1.69 of the 14 benchmarks. Group-3 and Group-4 improved significantly from their baseline performance by the second repetition of the task (p=0.016, 95%CI [0.19 2.71], effect size=0.746; and p=0.003, 95%CI [0.4 2.57], effect size=0.885, respectively). Group-2 improved significantly from their baseline performance by the third repetition of the task (p=0.004, 95%CI [0.42 3.04], effect size=0.886) while Group-1 had no statistically significant improvement during the five repetitions.

Learning curves were also assessed for the 14-performance metrics. In the fifth repetition of the task, around 90% of participants in all groups, including no-tailored-feedback group, were within the tumor percentage removed benchmark (Figure-4). All groups removed significantly

more tumor in the fifth repetition of the task compared to baseline performance ($p<0.05$) (Figure-5a). With only feedback groups, participants achieved the benchmarks >50% of the time with the metrics healthy tissue removed and instrument tip separation distance. Group-1 caused significantly more healthy tissue damage than Group-3 in the third to fifth repetitions of the task ($p=0.002$ 95%CI [0.03 0.16], effect size=0.998) (Figure-5b). Participants in Group-4 had a statistically significant lower instrument tip separation distance (using the two instruments together) than Group-1 at the fourth and fifth repetitions of the task ($p<0.001$ 95%CI [-4.97 -1.21], effect size=1.133), and this was also observed in participants in Group-3 from the second to fifth repetitions of the task ($p=0.029$ 95%CI [-5.81 -0.23], effect size=0.862) (Figure-5c). Group-3 and Group-4 improved significantly in efficiency index by the second repetition of the task (($p<0.001$ 95%CI [0.12 0.25], effect size=1.780) and ($p<0.001$ 95%CI [0.08 0.21], effect size=1.432), respectively) while the remaining groups improved significantly by the third repetition (Figure-5d). The learning curves and statistical comparison of the metric scores of the remaining 10 performance metrics can be found in Supplementary Figure-5.

In the post-questionnaire 5-point Likert scale, participants rated their simulation learning experience (Supplementary Table-2). Students' rating in Group-3 and Group-4 for the question 'How beneficial do you think the simulator and training system is for learning about surgery?' was 5.0 [3-5] (median [range]) while in Group-2 and Group-1, it was 4.0 [3-5]. Participants in feedback groups rated 'How beneficial was it to your performance to know which metrics you were being assessed on?' 5.0 [3-5] while no-tailored-feedback group rated 4.0 [2-5].

Discussion

In surgery, advanced computer technologies allow for the collection of vast amounts of data concerning technical skill, accurate skill assessment, and provide error detection and tailored feedback. [3-5, 19, 20] These systems used in virtual reality simulation training have been shown to enhance learner skills, and provide more efficient training than remote post-hoc human instruction. [2]

To put this work in context, providing trainees with efficient training feedback while challenging them in realistically replicated operative tasks required a series of components. First, virtual reality platforms with realistic surgical procedures and extensive data recording capacity were developed. [14, 21-24] Second, performance metrics encompassing critical features concerning the surgical procedure such as safety, efficiency, and performance quality along with bimanual dexterity and movement were developed to differentiate expertise groups and outline expert level performance benchmarks. [17, 25, 26] Spatial analysis of surgical performance using 3D tumor and tissue models has demonstrated differences between expert and novice level performances. [6, 27] Third, artificial intelligence methodologies were employed to provide a comprehensive performance assessment and outline performance metrics critical to achieve expert level performance. [5, 28, 29] Fourth, feedback systems provided trainees with expert level performance benchmarks to improve bimanual skills, based on virtual reality artificial intelligence platforms. [3, 4] After completing these steps, the current work explored the educational utility of these systems in improving trainee skills. We explored the efficacy of various instruction modalities by comparing numerical, visual, and visuospatial feedback.

In this study, the training sessions were organized based on time (number of repetitions) rather than defining a specific target proficiency level that trainees to achieve. This decision was influenced by the diverse training outcomes assessed and the time required for trainees to achieve proficiency in all 14-expert level benchmarks was unknown. Based on the results seen in Figure-4, achieving all 14 benchmarks would have been very challenging in a single training session, even for groups who received more efficient learning feedback.

Although Group-3 and Group-2 received the same metric information except for the application of color, Group-3 performed significantly better than Group-2 during the third repetition of the task. Additionally, Group-3 outperformed baseline performance in the second repetition of the task while Group-2 did not achieve the same success. The link between human color perception and psychological functioning is well studied. [30] In achievement contexts, such as education or athletic contests, psychologists have suggested that different colors cue learners' emotions and cognition which yields behavioral changes that can either optimize or impair performance. [31, 32] Our results indicated that the colored visualization of the feedback information is critical in achieving more efficient training. In the future, computer assisted teaching systems including artificial intelligence applications may benefit from incorporating visually enriched feedback methodologies, which provides a more engaging learning feedback to maximize trainee surgical skill acquisition. [2, 4, 33] Similar training applications can provide benefits across different procedural medical disciplines.

In this study, 14-performance metric benchmarks were utilized to assess the simulated surgical performance and track improvement across the five repetitions of the tumor resection task. Some of the 14-performance metrics showed improvement for all groups regardless of feedback (Figure-5 and Supplementary Figure-5) because they may have epitomized some of the

obvious goals of this surgical task. As such, all participants removed significantly more tumor (tumor percentage removed), achieved greater efficiency (efficiency index) and used their non-dominant hand more efficiently (coordination index, instrument tip separation distance) in the fifth repetition of the task (Figure-5). However, feedback provided faster learning for the intervention groups and better performance improvement.

Although some of the performance metrics were expected to improve, the goal with some of the other metrics such as brain volume removed, was to stay within the benchmark (Figure-5b) and to remove more tumor while not damaging the healthy tissue. Both Group-1 and Group-3 removed the same amount of tumor, around 80%, while Group-3 harmed significantly less healthy tissue, used their dominant hand more precisely (lower total tip path length), and had significantly lower scores in instrument tip separation during the fifth repetition of the task. These results may indicate that feedback is necessary to achieve an appreciation of the complex interplay between multiple factors during tumor surgery to meet the goals of the task more safely and efficiently.

Real-time intelligent systems are being developed and tested in surgical bimanual skills training using virtual reality simulation. [3, 34] Although this study has shown visual systems to be efficient for post-hoc feedback, in future directions of this work, auditory instructions may be an alternative for real-time feedback applications to prevent visual distractions. Systems with audio, visual, and video feedback are combined in our current trials (ClinicalTrials.gov, NCT05168150) to provide engaging feedback information to trainees which may improve the amount of information received by trainees and their skill acquisition. [35]

The tailored information provided by the intelligent systems is important; however, the major advantage of computer systems in skills acquisition may be achieved by optimal combinations of visual and auditory feedback components (e.g., video). In a randomized controlled trial involving the resection of a simulated brain tumor resection task, participants were instructed by the Virtual Operative Assistance on four performance metrics selected by a support vector machine algorithm along with feedback demonstration videos. [2] Participants improved on a composite score based on 16 performance metrics, and on eight of these 16 metrics changed significantly without receiving specific metric-based instructions. [36] Although the mechanism behind this extended effect is currently under investigation, a possible explanation is the breadth of extrinsic information contained in the feedback video demonstrations. [10] The ability to use both visual and auditory information may be the main advantage of these feedback systems in skills acquisition. To optimize the effectiveness in new feedback applications, it may be imperative to prioritize the pedagogical aspect of technical skill training and integrate informative, engaging, and easy to understand feedback information with the intelligent training systems.

This study has several limitations. (1) The training outcome in our simulation setting was limited to bimanual skills improvement. However, surgical operative room involves many other factors which can affect surgeon's performance and patient outcomes. Developing surgical simulation systems may provide a more immersive surgical training experience in the future. (2) Surgical trainees may be the most relevant trainee cohort for the testing of surgical training simulators. However, this study recruited medical students, a study cohort that may provide some advantages while also imposing limitations. Learning experience may differ as expertise develops. [37] Medical students' different interest level and procedural knowledge compared to

surgical trainees may affect their surgical training interaction and skill acquisition, however, their limited experience provides a greater room for improvement in skill acquisition, a scenario closer to that of a fresh surgical trainee who has just started training. Additionally, a medical student cohort provides a large number of participants to obtain statistical power, which is difficult to obtain with the limited number of surgical trainees available. For these reasons, medical students may be a better cohort than surgical trainees especially for the development and testing phases of simulation and training systems. Once these systems are well established, their efficacy in teaching and assisting surgical trainee cohorts should be confirmed in multi-institution trials. (3) Cognitive overload may limit the amount of information understood by the trainee. Cognitive load theory in education suggests that an optimal learning environment finds a balance between learners' intrinsic cognitive capacity, their motivation, and the extrinsic load of the instructional milieu. [38] Novice medical learners are also demonstrated to be at greater risk of overload in surgical simulation training. [39] In this application, training involved one session, in which learners sequentially removed five tumors, and were expected to improve on fourteen performance metrics. The amount of information needed to master these 14 performance features in one session may overwhelm trainee cognitive capacity and limit skill acquisition. Cognitive overload may have limited the amount of improvement especially with the participants in Group-4 since providing extra visuospatial information to this group did not achieve better results. One can speculate that the ability of trainees in Group-4 to adequately review the complex additional visual and spatial information available to them in only the limited five-minute feedback session may have been difficult. This could have resulted in increased trainee stress, leaving less time for critical learning methods such as self-reflection and improvement planning. [40] Results of Group-3 may support this conclusion as this group made a faster improvement without the 3D spatial

information, having a significantly greater number of benchmarks achieved than the baseline by the third repetition of the task. To prevent cognitive overload, longitudinal training settings with structured training goals in multiple sessions and/or different instruction methodologies may provide a better performance improvement. [41, 42] These longitudinal settings may integrate visual and visuospatial feedback to achieve efficient learning settings as outlined in this study and help to assess and compare retention of skills. (4) Our focus in this study was to maximize efficiency in learning with visual assistance. This study did not incorporate tailored auditory, video, tactile feedback, or other possible feedback modalities. Computer systems may incorporate different feedback mechanisms, not being limited to visual feedback, while the feedback can be adjusted to user preference. Future studies may compare different feedback modalities and explore multimodal learning. [43] Using the haptic technology of the simulator, tailored tactile feedback, such as vibration, can be implemented to inform the trainee when they apply too much force on delicate tissues.

In conclusion, this randomized controlled trial allowed the comparison of different post-hoc feedback modalities in surgical technical skills learning in the simulated environment. Simulations with autonomous visual and visuospatial feedback assistance provided trainees with a more effective way to master their bimanual operative skills.

References

1. Mirchi N, Ledwos N, Del Maestro RF. Intelligent Tutoring Systems: Re-Envisioning Surgical Education in Response to COVID-19. Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques. 2020:1-3.DOI: 10.1017/cjn.2020.202.
2. Fazlollahi AM, Bakhaidar M, Alsayegh A, Yilmaz R, Winkler-Schwartz A, Mirchi N, et al. Effect of Artificial Intelligence Tutoring vs Expert Instruction on Learning Simulated Surgical Skills Among Medical Students: A Randomized Clinical Trial. JAMA Network Open. 2022;5(2):e2149008-e.DOI: 10.1001/jamanetworkopen.2021.49008.
3. Yilmaz R, Winkler-Schwartz A, Mirchi N, Reich A, Christie S, Tran DH, et al. Continuous monitoring of surgical bimanual expertise using deep neural networks in virtual reality simulation. npj Digital Medicine. 2022;5(1):54.DOI: 10.1038/s41746-022-00596-8.
4. 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.
5. Winkler-Schwartz A, Yilmaz R, Mirchi N, Bissonnette V, Ledwos N, Siyar S, 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.
6. Yilmaz R, Ledwos N, Sawaya R, Winkler-Schwartz A, Mirchi N, Bissonnette V, et al. Nondominant Hand Skills Spatial and Psychomotor Analysis During a Complex Virtual Reality Neurosurgical Task-A Case Series Study. Oper Neurosurg (Hagerstown). 2022;23(1):22-30.DOI: 10.1227/ons.0000000000000232.
7. Stojan J, Haas M, Thammasitboon S, Lander L, Evans S, Pawlik C, et al. Online learning developments in undergraduate medical education in response to the COVID-19 pandemic: A BEME systematic review: BEME Guide No. 69. Med Teach. 2022;44(2):109-29.DOI: 10.1080/0142159x.2021.1992373.
8. Schwab B, Hungness E, Barsness KA, McGaghie WC. The Role of Simulation in Surgical Education. Journal of Laparoendoscopic & Advanced Surgical Techniques. 2017;27(5):450-4.DOI: 10.1089/lap.2016.0644.
9. Stefanidis D, Sevdalis N, Paige J, Zevin B, Aggarwal R, Grantcharov T, et al. Simulation in surgery: what's needed next? Ann Surg. 2015;261(5):846-53.DOI: 10.1097/sla.0000000000000826.
10. Fazlollahi AM. Artificial intelligence tutoring compared to expert instruction in surgical simulation training: A randomized controlled trial: McGill University; 2021.
11. Rangarajan K, Davis H, Pucher PH. Systematic Review of Virtual Haptics in Surgical Simulation: A Valid Educational Tool? J Surg Educ. 2020;77(2):337-47.DOI: 10.1016/j.jsurg.2019.09.006.
12. Association WM. World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. JAMA. 2013;310(20):2191-4.DOI: 10.1001/jama.2013.281053.
13. Juszczak E, Altman DG, Hopewell S, Schulz K. Reporting of Multi-Arm Parallel-Group Randomized Trials: Extension of the CONSORT 2010 Statement. JAMA. 2019;321(16):1610-20.DOI: 10.1001/jama.2019.3087.
14. Delorme S, Laroche D, DiRaddo R, Del Maestro RF. NeuroTouch: A Physics-Based Virtual Simulator for Cranial Microneurosurgery Training. Operative Neurosurgery. 2012;71(suppl_1):ons32-ons42.DOI: 10.1227/NEU.0b013e318249c744.
15. Sabbagh AJ, Bajunaid KM, Alarifi N, Winkler-Schwartz A, Alsideiri G, Al-Zhrani G, et al. Roadmap for Developing Complex Virtual Reality Simulation Scenarios: Subpial Neurosurgical Tumor Resection Model. World Neurosurgery. 2020;139:e220-e9.DOI: https://doi.org/10.1016/j.wneu.2020.03.187.

16. Ledwos N, Mirchi N, Yilmaz R, Winkler-Schwartz A, Sawni A, Fazlollahi AM, et al. Assessment of learning curves on a simulated neurosurgical task using metrics selected by artificial intelligence. Journal of Neurosurgery. 2022:1-12.DOI: 10.3171/2021.12.JNS211563.
17. Alotaibi FE, AlZhrani GA, Sabbagh AJ, Azarnoush H, Winkler-Schwartz A, Del Maestro RF. Neurosurgical Assessment of Metrics Including Judgment and Dexterity Using the Virtual Reality Simulator NeuroTouch (NAJD Metrics). Surgical Innovation. 2015;22(6):636-42.DOI: 10.1177/1553350615579729.
18. Lee DK. Alternatives to P value: confidence interval and effect size. Korean J Anesthesiol. 2016;69(6):555-62.DOI: 10.4097/kjae.2016.69.6.555.
19. Alkadri S, Ledwos N, Mirchi N, Reich A, Yilmaz R, Driscoll M, et al. Utilizing a multilayer perceptron artificial neural network to assess a virtual reality surgical procedure. Computers in Biology and Medicine. 2021;136:104770.DOI: https://doi.org/10.1016/j.compbiomed.2021.104770.
20. Reich A, Mirchi N, Yilmaz R, Ledwos N, Bissonnette V, Tran DH, et al. Artificial Neural Network Approach to Competency-Based Training Using a Virtual Reality Neurosurgical Simulation. Oper Neurosurg (Hagerstown). 2022;23(1):31-9.DOI: 10.1227/ons.0000000000000173.
21. Ribeiro de Oliveira MM, Nicolato A, Santos M, Godinho JV, Brito R, Alvarenga A, et al. Face, content, and construct validity of human placenta as a haptic training tool in neurointerventional surgery. J Neurosurg. 2016;124(5):1238-44.DOI: 10.3171/2015.1.JNS141583.
22. Winkler-Schwartz A, Yilmaz R, Tran DH, Gueziri HE, Ying B, Tuznik M, 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.
23. Bowyer MW, Fransman RB. Simulation in General Surgery. In: Stefanidis D, Korndorffer Jr JR, Sweet R, editors. Comprehensive Healthcare Simulation: Surgery and Surgical Subspecialties. Cham: Springer International Publishing; 2019. p. 171-83.
24. Ledwos N, Mirchi N, Bissonnette V, Winkler-Schwartz A, Yilmaz R, Del Maestro RF. Virtual Reality Anterior Cervical Discectomy and Fusion Simulation on the Novel Sim-Ortho Platform: Validation Studies. Oper Neurosurg (Hagerstown). 2020;20(1):74-82.DOI: 10.1093/ons/opaa269.
25. AlZhrani G, Alotaibi F, Azarnoush H, Winkler-Schwartz A, Sabbagh A, Bajunaid K, et al. Proficiency Performance Benchmarks for Removal of Simulated Brain Tumors Using a Virtual Reality Simulator NeuroTouch. Journal of Surgical Education. 2015;72(4):685-96.DOI: https://doi.org/10.1016/j.jsurg.2014.12.014.
26. Alotaibi FE, AlZhrani GA, Mullah MAS, Sabbagh AJ, Azarnoush H, Winkler-Schwartz A, et al. Assessing Bimanual Performance in Brain Tumor Resection With NeuroTouch, a Virtual Reality Simulator. Operative Neurosurgery. 2015;11(1):89-98.DOI: 10.1227/NEU.0000000000000631.
27. Sawaya R, Bugdadi A, Azarnoush H, Winkler-Schwartz A, Alotaibi FE, Bajunaid K, et al. Virtual Reality Tumor Resection: The Force Pyramid Approach. Operative Neurosurgery. 2017;14(6):686-96.DOI: 10.1093/ons/opx189.
28. Bissonnette V, Mirchi N, Ledwos N, Alsidieri G, Winkler-Schwartz A, Del Maestro RF. Artificial Intelligence Distinguishes Surgical Training Levels in a Virtual Reality Spinal Task. J Bone Joint Surg Am. 2019;101(23):e127.DOI: 10.2106/jbjs.18.01197.
29. Mirchi N, Bissonnette V, Ledwos N, Winkler-Schwartz A, Yilmaz R, Karlik B, et al. Artificial Neural Networks to Assess Virtual Reality Anterior Cervical Discectomy Performance. Oper Neurosurg (Hagerstown). 2020;19(1):65-75.DOI: 10.1093/ons/opz359.
30. Elliot AJ. Color and psychological functioning: a review of theoretical and empirical work. Frontiers in Psychology. 2015;6.DOI: 10.3389/fpsyg.2015.00368.
31. Elliot AJ, Maier MA. Color Psychology: Effects of Perceiving Color on Psychological Functioning in Humans. Annual Review of Psychology. 2014;65(1):95-120.DOI: 10.1146/annurev-psych-010213-115035.

32. Hill RA, Barton RA. Red enhances human performance in contests. Nature. 2005;435(7040):293-.DOI: 10.1038/435293a.
33. Roh TH, Oh JW, Jang CK, Choi S, Kim EH, Hong C-K, et al. Virtual dissection of the real brain: integration of photographic 3D models into virtual reality and its effect on neurosurgical resident education. Neurosurgical Focus. 2021;51(2):E16.DOI: 10.3171/2021.5.FOCUS21193.
34. Yilmaz R, Winkler-Schwartz, A., Mirchi, N., Del Maestro, R. Continuously Monitoring Tools to Assess and Enhance Human Performance While Providing Error Avoidance Systems, patent No. 05001770-883USPR. . 2020
35. Howie EE, Dharanikota H, Gunn E, Ambler O, Dias R, Wigmore SJ, et al. Cognitive Load Management: An Invaluable Tool for Safe and Effective Surgical Training. Journal of Surgical Education. 2023;80(3):311-22.DOI: https://doi.org/10.1016/j.jsurg.2022.12.010.
36. Ahmed Z, Lau CHH, Poole M, Arshinoff D, El-Andari R, White A, et al. Canadian Conference for the Advancement of Surgical Education (C-CASE) 2021: Post-Pandemic and Beyond Virtual Conference Abstracts. Canadian Journal of Surgery. 2021;64(6 Suppl 1):S65-S79.DOI: 10.1503/cjs.018821.
37. Daley BJ. Novice to Expert: An Exploration of How Professionals Learn. Adult Education Quarterly. 1999;49(4):133-47.DOI: 10.1177/074171369904900401.
38. Sweller J. Cognitive Load During Problem Solving: Effects on Learning. Cognitive Science. 1988;12(2):257-85.DOI: https://doi.org/10.1207/s15516709cog1202_4.
39. Andersen SA, Mikkelsen PT, Konge L, Cayé-Thomasen P, Sørensen MS. Cognitive Load in Mastoidectomy Skills Training: Virtual Reality Simulation and Traditional Dissection Compared. J Surg Educ. 2016;73(1):45-50.DOI: 10.1016/j.jsurg.2015.09.010.
40. Marcovitch S, Jacques S, Boseovski JJ, Zelazo PD. Self-Reflection and the Cognitive Control of Behavior: Implications for Learning. Mind, Brain, and Education. 2008;2(3):136-41.DOI: https://doi.org/10.1111/j.1751-228X.2008.00044.x.
41. Andersen SAW, Mikkelsen PT, Konge L, Cayé-Thomasen P, Sørensen MS. The effect of implementing cognitive load theory-based design principles in virtual reality simulation training of surgical skills: a randomized controlled trial. Advances in Simulation. 2016;1(1):20.DOI: 10.1186/s41077-016-0022-1.
42. Chan A, Singh S, Dubrowski A, Pratt DD, Zalunardo N, Nair P, et al. Part versus whole: a randomized trial of central venous catheterization education. Advances in Health Sciences Education. 2015;20(4):1061-71.DOI: 10.1007/s10459-015-9586-0.
43. Baykan Z, Naçar M. Learning styles of first-year medical students attending Erciyes University in Kayseri, Turkey. Advances in Physiology Education. 2007;31(2):158-60.DOI: 10.1152/advan.00043.2006.

Tables and Figures

Table-1: Participant Characteristics.

	Group 1 Practice alone with no feedback (n=30)	**Group 2** No visual feedback (n=30)	**Group 3** Visual feedback (n=29)	**Group 4** Visuospatial feedback (n=31)	**All Participants** (n=120)
Mean age +/- SD (range)	23.6 +/- 4.8 (19-44)	22.8 +/- 3.3 (19-31)	22.4 +/- 2.6 (19-28)	23.6 +/- 3.5 (18-33)	23.1 +/- 3.6 (18-44)
Male/Female	18/12	18/12	18/11	17/14	71/49
Handedness (Right/Left/Ambidextrous)	27/3/0	28/2/0	24/4/1	29/1/1	108/10/2
Medical School:					
McGill University	24	22	21	25	92
University of Montreal	5	6	4	5	20
University of Sherbrooke	1	2	3	1	7
University of Laval	0	0	1	0	1
Year in medical school:					
1st	16	21	18	20	75
2nd	10	6	7	8	31
3rd	3	2	2	1	8
4th	1	1	2	2	6
Level of interest in surgery, median (range)	4 (2-5)	4 (1-5)	4 (1-5)	4 (1-5)	4 (1-5)
Completed surgical rotation (Y/N)	2/28	1/29	2/27	2/29	7/113
Playing video games:					
Not at all	12	13	13	13	51
Occasionally (< 2 hours per week)	9	9	7	9	34
Often (2- 10 hours per week)	6	8	6	6	26
Very often (> 10 hours per week)	3	0	3	3	9
Playing musical instruments:					
I don't play any musical instrument	11	14	9	17	51
Yes, I am at beginner level	6	4	6	3	19
Yes, I am at intermediate level	6	7	8	6	27
Yes, I am at advanced level	6	4	4	5	19
Yes, I am at master level	1	0	2	0	3
Previously used virtual reality simulation (Y/N)	1/29	2/28	0/29	2/29	5/115

Table-2: Performance Metrics.

Category	Performance Metric (unit)	Description
Safety	Brain volume removed (cc)	Total amount of healthy tissue (white surrounding tissue) removed
	Amount of blood loss (cc)	Total amount of bleeding
	Maximum force applied with dominant hand (N)	Maximum force amount utilized by ultrasonic aspirator
	Maximum force applied with non-dominant hand (N)	Maximum force amount utilized by bipolar
	Sum of forces applied with dominant hand (N)	Total force utilized by ultrasonic aspirator during the whole procedure
	Sum of forces applied with non-dominant hand (N)	Total force utilized by bipolar during the whole procedure
Quality	Tumor percentage removed (%)	Total volume of tumor removed
Efficiency	Total tip path length dominant hand (mm)	Total trace length of the tip of ultrasonic aspirator
	Total tip path length non-dominant hand (mm)	Total trace length of the tip of bipolar
	Path length index (ratio)	Tip trace rate in which ultrasonic aspirator was active (in contact with tissues)
	Efficiency index (ratio)	Time rate in which ultrasonic aspirator was active
Bimanual cognitive	Average instrument tips separation distance (mm)	Average distance between tips of the instruments during the whole procedure
	Coordination index (ratio)	Time rate while both instruments were used together vs bipolar used alone
	Bimanual forces ratio (ratio)	Ultrasonic aspirator force ratio while both instruments were used together compared to ultrasonic aspirator used alone

Figure-1: Flow diagram. One hundred and twenty students were randomly allocated into four different feedback groups including practice-alone with no-feedback group. No participant/data was excluded from the analysis.

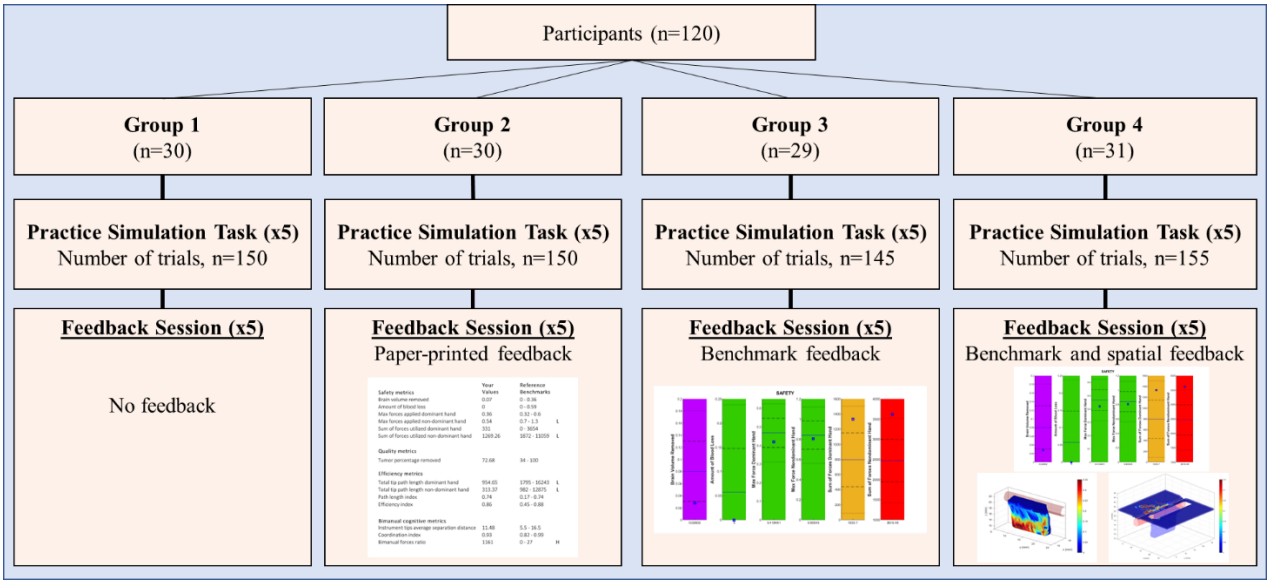


Figure-2: Simulated Scenario. The virtually simulated task involved the subpial resection of a rectangular yellow tumor using an ultrasonic aspirator in the dominant hand and a bipolar forceps in the non-dominant hand (a). The goal of the task was to remove the tumor completely while minimizing injury to surrounding tissues (b). There was a blood vessel with ability to bleed, located posterior to the tumor (c). Any damage to this blood vessel resulted in bleeding (d). Ultrasonic aspirator was used to aspirate the blood (d) and bipolar was used to cauterize the bleeding vessel (e). The appearance of the tissue after successful cauterization (f).

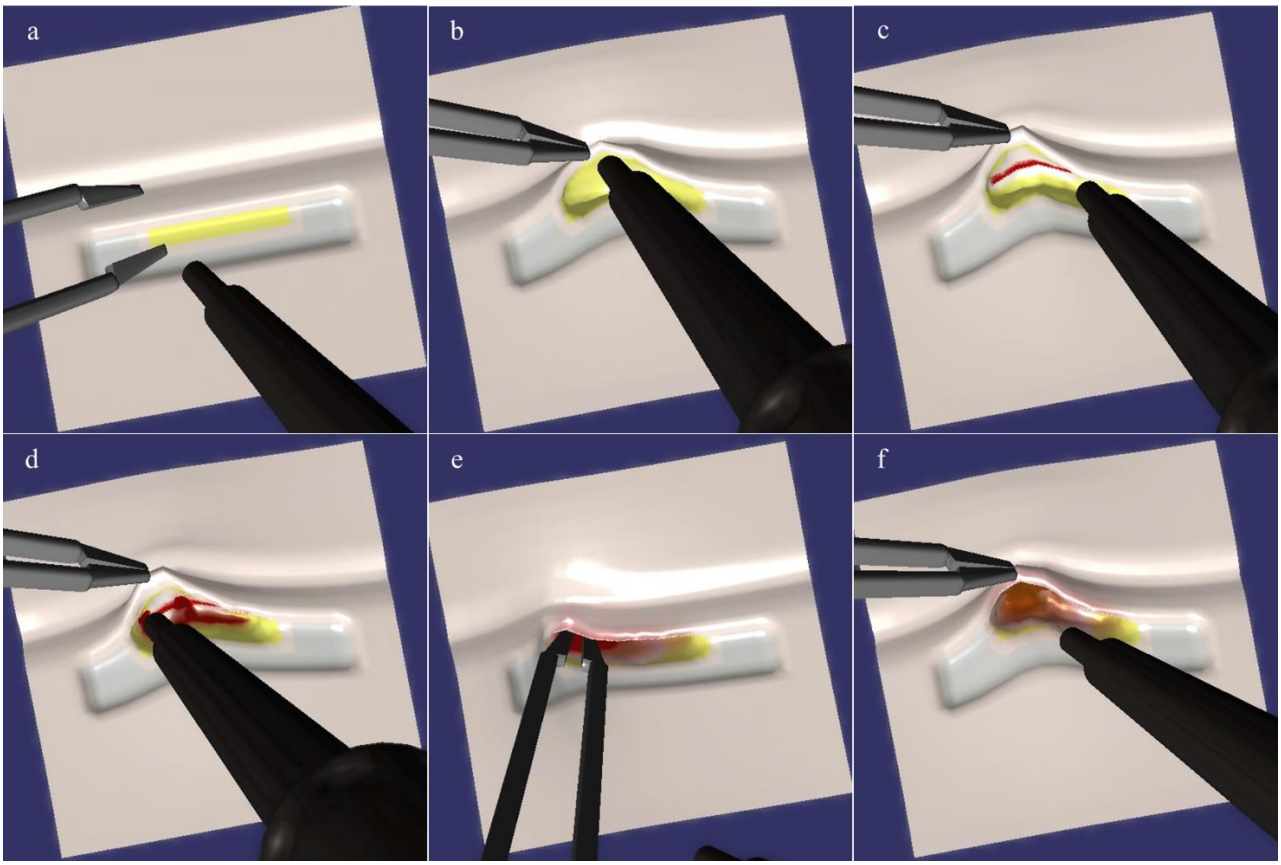


Figure-3: Number of Benchmarks Achieved. X-axis represents the four feedback groups. Each feedback group is color-coded (see the legend). Y-axis represents the average number of benchmarks achieved by each feedback group. *Horizontal lines represent statistically significant difference ($p<.05$). For within group differences, horizontal lines are represented with the respected color of the group. Vertical lines represent standard error bars. Group 3 and Group 4 improved significantly compared to the baseline performance by the second repetition. Group 2 improved significantly compared to baseline performance by the third repetition. Group 3 outperformed practice-alone Group 1 by the third repetition. Group 4 outperformed practice-alone Group 1 by the fourth repetition.

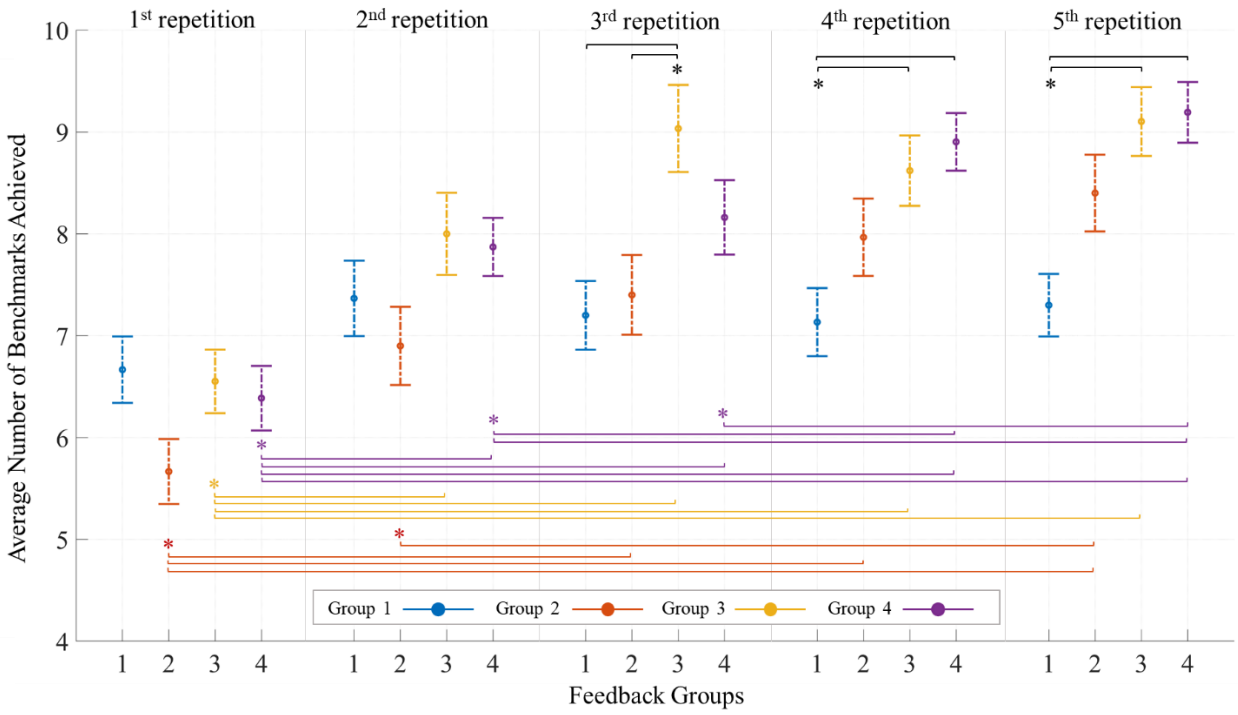


Figure-4: Percentage of Trainees who Achieved Benchmarks. X-axis shows each of the 14-performance metrics on which the trainees were assessed. Each feedback group is color-coded (see the legend). Y-axis represents the percentage of trainees who achieved the benchmarks. There are five percentages shown for each performance metric across five trials, from the first repetition of the task/baseline performance to the fifth repetition.

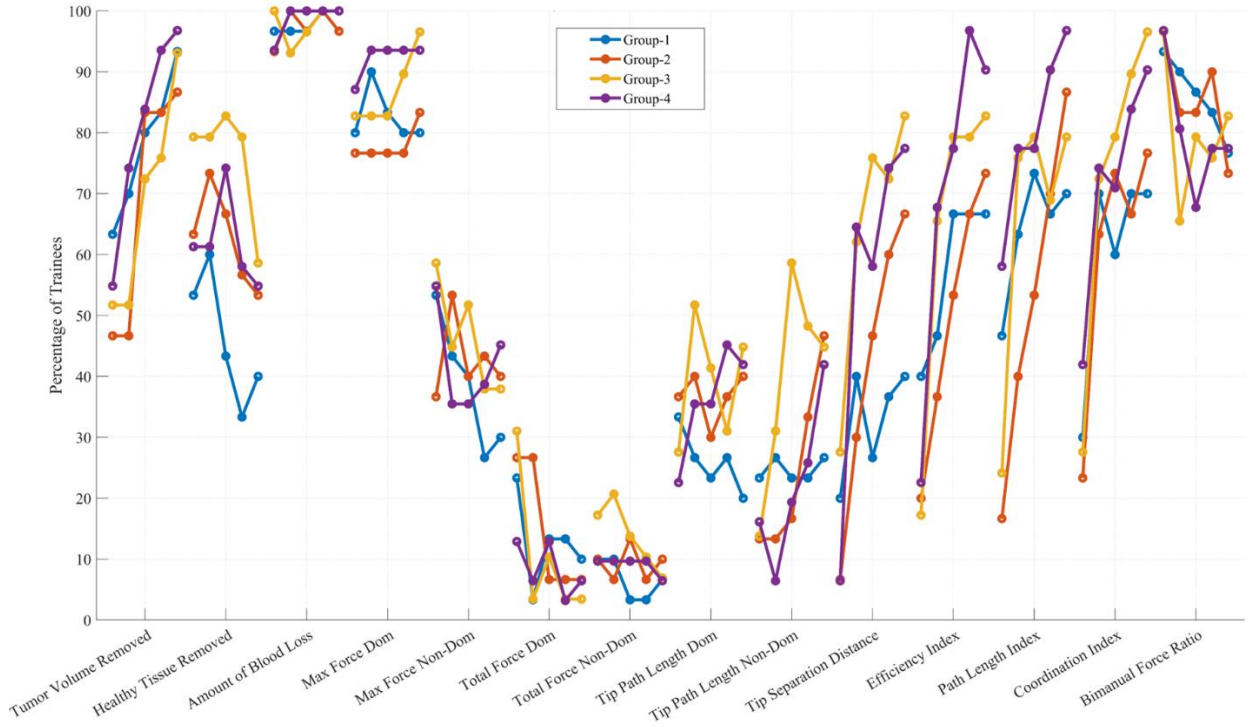
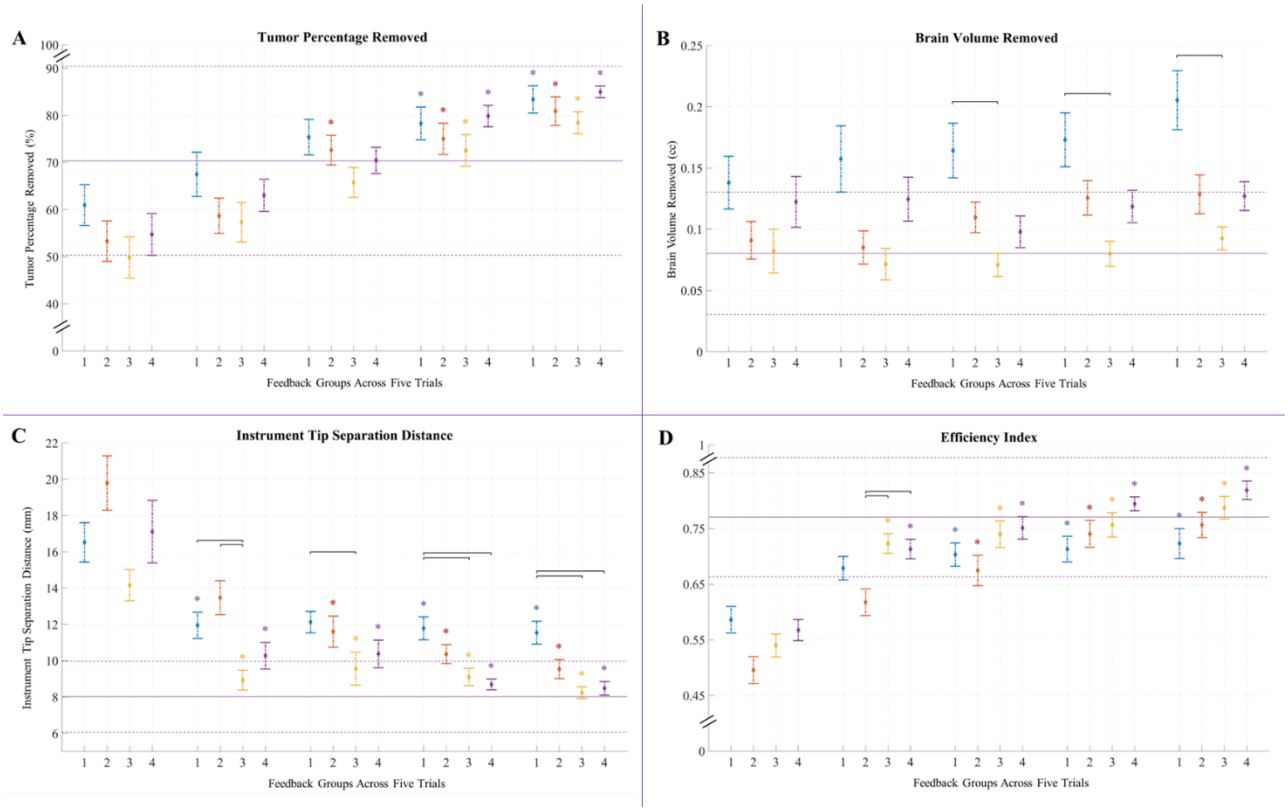


Figure-5: Performance Metrics Learning Curves. The learning curves of four performance metrics. X-axes represent the task repetition from the 1st repetition/baseline performance to the 5th repetition for the 4-feedback groups. The purple straight horizontal line indicates the mean expert value for each performance metric while the two dotted purple lines one standard deviation above and below the mean indicate the boundaries of the expert benchmark. *Asterisks indicate significantly different values from the 1st repetition/baseline performance of that group. Horizontal square brackets show significant differences between feedback groups at the same repetition of the task. Axis brakes were indicated along y-axis. The learning curves of the remaining 10 of the 14-performance metrics are shown in Supplementary Data.



Chapter 5 - Surgical Skills Training Using Real-Time Artificial Intelligence vs Human Instruction – A Randomized Controlled Trial

Preface

The development and promising predictive performance of the Intelligent Continuous Expertise Monitoring System across a neurosurgical residency training program inspired this randomized controlled trial. The ICEMS was able to assess performance and predict risks as outlined in Chapter 2. The very next question was 'Can the ICEMS teach?'. This study outlined the first time intelligent real-time feedback application in comparison to in-person human expert instruction in teaching surgical bimanual skills. Learning from the work in Chapter 4, feedback was crucial to achieve better learning outcomes. Hence this randomized controlled trial integrated an active controlled group where learning outcomes from the real-time intelligent system were compared to this active control group instead of a control group with no-feedback. The results of the study were to demonstrate the promising future of artificial intelligence in augmenting learning via real-time feedback and its comparable performance to traditional learning via in-person expert instruction.

The manuscript is in under peer-review:

Recai Yilmaz, Mohamad Bakhaidar, Ahmad Alsayegh, Nour Abou Hamdan, Ali M. Fazlollahi, Trisha Tee, Ian Langleben, Alexander Winkler-Schwartz, Denis Laroche, Carlo Santaguida, Rolando Del Maestro. Surgical Skills Training Using Real-Time AI vs Human Instruction – A Randomized Clinical Trial

Abstract

Question: How does real-time artificial intelligence feedback compare to in-person human instruction in teaching surgical bimanual skills?

Findings: In this randomized clinical trial involving 97 participants, students who are taught by a real-time artificial intelligence system achieved significantly better learning outcomes than those taught in-person by expert instructors. Learning from an artificial intelligence system caused a significantly higher cognitive load and it resulted in a similar transfer rate of skills to a more complex realistic surgical procedure.

Meaning: Real-time artificial intelligence feedback may provide efficient simulated surgical bimanual skills training, comparable to person-to-person instruction.

Importance: Teaching operative bimanual skills via the apprenticeship model faces challenges in outlining, assessing, and teaching the composites of surgical expertise. Artificial intelligence (AI) provides a precise real-time assessment of the action being performed and tailored feedback which can transform the way operating skills are taught.

Objective: To compare real-time AI feedback with in-person expert instruction in teaching simulated tumor resection skills.

Design: A double-blinded randomized clinical trial conducted between January and May 2022.

Setting: A multicenter study involving a single simulation training session.

Participants: Ninety-nine students who were enrolled in four Canadian medical schools participated. Two participants were excluded from the analysis due to technical problems faced during the simulation sessions.

Intervention: A 90-minute simulation training involving six tumor resections, a practice tumor resection five times followed by a realistic brain tumor resection with three feedback interventions: 1- AI auditory and audiovisual feedback, 2- in-person expert instruction, and 3- control group with no real-time feedback.

Main Outcome(s) and Measure(s): Improvement in the composite performance score (range, −1.00 to 1.00) in practice sessions and learning transfer to a more realistic task were quantified by a validated AI system, the Intelligent Continuous Expertise Monitoring System (ICEMS). Secondary outcomes were Objective Structured Assessment of Technical Skills (OSATS; range 1-7) rating on realistic tumor resection, rated by blinded experts, and self-reported cognitive load.

Results: Ninety-seven participants (mean [range] age: 21.3 [17-31], 60% women) completed the simulation training and were included in the analysis. Training with real-time AI feedback resulted in significantly better performance outcomes compared to both no real-time feedback and in-person instruction, .266, [95%CI .107 to .425], p<.001 and .332, [95%CI .173 to .491], p=.005, respectively. Learning from real-time AI caused a significantly higher cognitive-load $\chi 2(2)=3.173$, p=.005, and a similar transfer rate of skills $F(2, 94)=1.241$, p=.294 and OSATS ratings (4.30 vs 4.11) when compared to in-person training with expert instruction.

Conclusions and Relevance: Real-time intelligent feedback provided superior learning outcomes, with similar learning transfer and OSATS ratings compared to person-to-person instruction. Intelligent systems provide critical tailored, quantifiable feedback and actionable instructions for the mastery of bimanual operative skills.

Trial Registration: NCT05168150

Introduction

Surgery is a high-stakes intervention on delicate tissues that requires cautious care by expert hands. Intraoperative errors lead to high patient morbidity and mortality, which increase economic costs to society and contribute to physician burnout.[1-4] Learning such skills is a difficult and stressful endeavor, as surgeons and trainees must balance teaching/learning and maintaining patient safety in a dynamic operating room environment.[5,6]

Mastering surgical skills occurs during a comprehensive but often lengthy apprenticeship, known as residency. This apprenticeship involves intraoperative teachings, encompassing continuous interaction between the surgical educator and the learner, along with ongoing intraoperative assessment and feedback. However, the feedback provided is largely limited to instructional communication. This surgical teaching model lacks objectivity and standardization, has challenges in defining, evaluating, quantifying, and teaching the composites of surgical expertise, and may depend on the availability of patient cases.[7-9] As a result, surgical education is implementing a competency-based quantifiable framework.[10-12]

The Intelligent Continuous Expertise Monitoring System (ICEMS) augments instructional teachings during surgical technical skills training using artificial intelligence.[13,14] This intelligent system mimics the role of expert surgical instructors in the context of surgical simulation training, integrated into the NeuroVR (CAE Healthcare) simulator, an immersive virtual reality platform for performing brain tumor resections.[15,16] The ICEMS continuously assesses surgical performance in 0.2-second intervals and provides real-time instruction and risk detection. This system demonstrated a granular differentiation of skill levels between experts and residents, and between residents at different stages in their neurosurgery training program.[13] Although the

predictive ability of this system's continuous performance assessment is validated, its pedagogical utility and efficiency in teaching surgical bimanual skills via real-time instruction and risk detection remain unexplored.

This double-blinded prospective randomized controlled trial compared the efficacy of tailored intelligent feedback provided by ICEMS to that of in-person expert instruction in simulated surgical skills training. We hypothesized that learners provided with ICEMS real-time feedback will (1) achieve a similar improvement compared to those learning in-person with expert instructors, (2) achieve a similar improvement in the Objective Structure Assessment of Technical Skills (OSATS)[17] rating compared to those learning in-person with expert instructors, and (3) have a similar cognitive load compared to those learning in-person with expert instructors.

Methods

This study was approved by the McGill University Health Centre Research Ethics Board, Neurosciences-Psychiatry. This report followed the extensions of the CONSORT 2010 Statement, guidelines for the reporting of multi-arm parallel group randomized trials and interventions involving artificial intelligence.[18-20]

Participants

(Figure-1) Participants were recruited between January 2022 – March 2022, for a single 90-minute simulation session with no follow up. Inclusion criterion was enrollment in year one to four of a medical school program in Canada. Our exclusion criterion was previous experience in using our simulation platform, the NeuroVR (CAE Healthcare). Participants signed an approved consent form before the start of the trial. Public health measurements and the Montreal

Neurological Institute and Hospital's regulations related to COVID-19 pandemic were followed to ensure health safety. Methods remained unchanged after trial commencement. The study protocol was in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Declaration of Helsinki.[21] All participants completed two questionnaires; a pre-questionnaire related to demographics, previous simulation experience and surgical exposure and, a post-questionnaire to rate their cognitive load and simulation learning experience. Participants were informed that the study involved no harm to participants, that their information is anonymized. Participants were blinded to the study outcomes.

Randomization

Randomization was applied without stratification using an online random number generator.[22] Participants were allocated into three groups based on a random number generation between number one and three.

Simulation

All participants were given a standardized instruction sheet before the simulation session. The sessions were carried out in a controlled distraction-free environment. Two tumor resection tasks were performed; a practice subpial tumor resection task and a realistic brain tumor resection (Video).[16] Expert execution of subpial technique is important in a variety of neurosurgical procedures to remove culprit tissues while preserving the neurologic function.[23,24] The NeuroVR (CAE Healthcare, Canada) 3D neurosurgical simulation platform with two haptic handles was utilized to simulate the tasks.[15] Both tasks required using two instruments, an ultrasonic aspirator and a bipolar forceps, to completely remove the simulated tumor while minimizing bleeding and damage to surrounding healthy tissue.[25,26] Face and content validity of the simulation tasks were

previously demonstrated.[16,27] The time limit was five minutes for the practice task, and 13 minutes for the realistic tumor resection task.

Feedback was incorporated in two stages: during the task (real-time), and after the task (post hoc). Participants were randomly allocated into three groups, (1) post hoc-only feedback (active control), (2) real-time and post hoc intelligent instruction (ICEMS group), and (3) real-time and post hoc expert instruction (expert instructor group). Participants completed the practice task five times. The first repetition was completed without feedback during the performance to determine baseline. After completion of the baseline performance, participants received post hoc feedback based on their group allocation, as described in detail below. Five minutes was given for post hoc feedback for all groups. Finally, all participants performed a realistic brain tumor resection task once without feedback to assess skill transfer to this more complex simulated procedure.

Post hoc feedback group

Participants in this group received no real-time feedback during the tasks. After the baseline and after each task, participants were provided with post hoc feedback on their performance scores in comparison to expert benchmarks on five performance metrics, which included the same metrics listed in the next section. The goal was to meet all five benchmarks by the last repetition of the task.

Real-time artificial intelligence instruction

(Figure-2) Participants in this group received real-time auditory instructions given by the ICEMS.[13] The ICEMS assessed surgical performance at 0.2-second intervals on five performance metrics: (1) bleeding risk, (2) healthy tissue damage risk, (3) ultrasonic aspirator force utilization, (4) bipolar instrument force utilization and (5) using the two instruments together. Six auditory instructions (one instruction per performance metrics and two instructions

for bipolar high and low force utilization) were incorporated. ICEMS predicted expert level performance metrics in real-time based on the actions being performed by the learner. An error was identified when participant performance score differed more than one standard deviation from the expert level assessment of the ICEMS, for at least one second. Real-time auditory instructions were automatically delivered upon error identification during all practice tasks except the baseline performance.

Post hoc artificial intelligence instruction

(Figure-2) The participants' performance was video recorded. After the completion of each practice task, including the baseline performance, the ICEMS located the timing of specific errors using the performance data. The ICEMS cut these error footages from the entire performance video clip and demonstrated them to the participants. An error video-clip relating to each performance metrics, to a maximum total of six error video-clips were shown to the participant in the form of 10-second video-clips (see Supplementary information).

Real-time expert instruction

Two neurosurgery residents (M.B. and A.A., post-graduate year six) had standardized teaching experience in a recent simulation trial[28] and completed a training to achieve consultants' benchmarks during the simulated tasks. They provided in-person real-time instructions using a modified OSATS rating scale (see Supplementary Information) and a modified PEARLS debriefing script.[29] Instructors were blinded to the ICEMS assessment metrics. From the second repetition of the practice task to the fifth repetition, an expert instructor provided verbal instructions to the participant during the simulated tasks.

Post hoc expert instruction

After the completion of each practice task, including the baseline performance, the expert instructor had five minutes with the participant to outline any pertinent information to enhance performance. The expert instructors also had the option to personally demonstrate strategies and surgical techniques on the NeuroVR simulation on how to expertly perform the simulated subpial resection. To facilitate standardization, instructors followed learning objectives based on the OSATS rubric (Supplementary Information).

Outcome measures

All performance data was recorded along with the video recordings of each task. The primary outcome measure was the composite performance score quantified by the ICEMS during practice and realistic tumor resections. The ICEMS scored participants' performance between a score of -1 (novice) and 1 (expert) at 0.2-second intervals. An average composite-score was calculated for each repetition of the task for statistical comparisons. The video recordings of the realistic brain tumor resection task were rated by two blinded expert raters using the OSATS scale as previously described.[17,28] Cognitive load was assessed through a questionnaire before, during, and after the simulation exercises.[28]

Statistical analysis

Data was not normally distributed as assessed by Shapiro-Wilk's test ($p<.05$). Non-parametric statistical tests: Friedman's test and Kruskal-Wallis H test, were utilized. See Supplementary data for more detailed information.

Results

Participants and sample size

Ninety-nine medical students who were presently enrolled in four medical schools across the province of Quebec participated in this three-parallel-arm randomized controlled trial (Figure-1). Participant simulation performance data was recorded in one session without a follow-up. Data from two participants was excluded from the analysis due to technical issues faced during the simulated tasks. Mean participant age +/- SD (Range) was 21.3 +/-2.7 (17-31) years, and participant handedness was 89/7/1 (right-handed/left-handed/ambidextrous). Participants' level of interest in surgery was a median (range) of 4 (1-5) (Table). A sample size calculation for a power of .99 with an effect size of 0.25, 0.5 correlation among repeated measures, and with .85 non-sphericity correction epsilon, yielded 32 participants in each group, and 96 participants in total, for assessment of within- and between-group interaction. Data analysis was conducted based on intention-to-treat.

Between-feedback comparison

(Figure-3) There were no significant differences in the composite-score in the baseline performance, p=.421 among the three groups. There was a statistically significant interaction between feedback allocation and task repetition in a two-way mixed model ANOVA on the ICEMS composite score, $F(6.8, 319.5)=5.06$, p<.001, partial $\eta 2=.097$. In the third task, both the ICEMS and expert instruction groups outperformed post hoc feedback group, (.343, 95%CI [.182 .504], p<.001), and (.190, 95%CI [.052 .330], p=.049), respectively. In the fourth task, the ICEMS group outperformed post hoc feedback group, (.265, 95%CI [.061 .468], p=.019), while expert instruction group was not significantly different than post hoc feedback group, (.079, 95%CI [-.125 .284], p=.069). In the fifth task, the ICEMS group outperformed both post hoc and

expert instruction groups, .266, 95%CI [.107 .425], p<.001 and .332, 95%CI [.173 .491], p=.005, respectively.

Within-group learning curves

(Figure-3) The post hoc-only feedback group improved their performance in the fifth task compared to the baseline (.185, 95%CI [.039 .332], p=.009). The ICEMS group outperformed their baseline in the third, fourth, and fifth tasks; .295, 95%CI [.073 .516], p=.031, .350, 95%CI [.107 .593], p=.001, and .400, 95%CI [.180 .620], p<.001, respectively. The expert instruction group achieved a steep performance improvement in the composite-score where they outperformed their baseline performance in the second, third, and fourth tasks; .252, 95%CI [.070 .434], p=.001, .213, 95%CI [.054 .372], p=.027, .235, 95%CI [.051 .418], p=.016, after which they reached a plateau. There was a decrease in the composite-score and no significant difference was found between the fifth task and the baseline for this group, .138, 95%CI [.023 .253], p=.269.

Performance on the realistic task

(Figure-4a) The composite score on the realistic task was compared by a one-way ANOVA between feedback groups. Mean [95%CI] scores were -0.343 [-0.450 -0.236] for post hoc feedback group, -0.233 [-0.330 -0.136] for real-time AI group, and -0.263 [-0.371 -0.156] for expert instruction group. No statistically significant between groups differences were observed, $F(2, 94)=1.241$, p=.294.

Blinded expert OSATS rating

(Figure-4c) The OSATS rating (median score on a 7-point scale) of the realistic task involved five items and an overall score given by two blinded experts. An average of the ratings by two experts were calculated for each item. Participants in the ICEMS group (4.30) achieved a

significantly higher overall score than those in post hoc feedback group (3.47), p=.017. The overall score achieved by the participants in the expert instruction group (4.11) was not significantly different than both post hoc and the ICEMS groups, p=.137, and p=1, respectively. The ICEMS group (4.9) outperformed both post hoc (4.15) and expert instruction groups (3.69) in hemostasis, p=.017, and p<.001, respectively. The ICEMS group outperformed the post hoc feedback group in instrument handling (4.49 vs 3.57, p=.006), respect for tissue (4.26 vs 3.73, p=.015), and flow (4.26 vs 3.18, p=.002) while the expert instruction group outperformed the post hoc feedback group only in instrument handling (4.45 vs 3.57, p=.014). Hence, the ICEMS group achieved the best learning outcomes concerning hemostasis, respect for tissue, flow, and overall OSATS score. There was a significant correlation between the ICEMS's composite score and the average OSATS score given by two expert raters, Spearman's correlation coefficient =.224, p=.028. The correlation coefficient between the two expert raters was also significant, Spearman's correlation coefficient =.258, p=.011.

Cognitive load assessment

(Figure-4b) Intrinsic, extraneous, and germane load (median score o on a 5-point scale) were assessed through the Cognitive Load Index for cognitive demands on a 5-point Likert scale.[30] No significant differences were observed between groups in intrinsic and germane load; $\chi 2(2)=1.983$, p=.371, and $\chi 2(2)=3.732$, p=.155, respectively. Participants in ICEMS group (1.19) reported significantly higher extraneous load than those in expert-instruction group (1.13), p=.005, indicating increased cognitive difficulty experienced by the trainees in understanding ICEMS's instructions.

Discussion

To the best of our knowledge, this is the first randomized controlled trial that compares real-time intelligent instruction with in-person human expert instruction in teaching bimanual surgical skills in simulation training.[31,32] Our findings demonstrate superior learning outcomes using a real-time intelligent system compared to in-person expert instruction. These results are confirmed both when measured quantitively by the ICEMS and when assessed by blinded experts.

Previous simulation training methodologies typically involve repetitive practice of basic to complex tasks, often without feedback or with post hoc performance feedback.[28,33-37] In both intervention arms of this study, we aimed to replicate the real-time training engagement happening in the operating room where trainees receive ongoing assessment and instructions from expert surgeons. For the first time, an artificial intelligence-powered tutor provided trainees with real-time feedback and action-oriented instructions as they performed a simulated neurosurgical task.

Feedback is critical for skill acquisition, and the most effective modalities may depend on the surgical procedure being taught.[38-40] In training for complex procedures such as the subpial resection of brain tumors, practice without feedback has resulted in little to no improvement while post hoc feedback based on performance metrics benchmarks has resulted in significant improvement in learning.[28] Hence, our study utilized an active control group that received post hoc feedback.

Cognitive load is the mental exertion of a trainee to process and retain information.[41,42] In this trial, learning from the real-time intelligent instructions resulted in significantly higher

extraneous load, suggesting increased cognitive demand experienced by the trainees to understand the real-time auditory instructions and the post hoc video demonstrations. In future applications, it is important to minimize extraneous cognitive load to maximize learning.[43,44]

In this study, expert instructors had greater flexibility in their teaching engagement with students. Experts could provide learners with more surgical context concerning the procedure, share relevant strategies, and help students develop a plan to use the instruments and remove the tumor efficiently. The ICEMS provided direct instructions on five predetermined performance metrics with limited context about the surgery or the action being performed by the student. Despite the limitations of the intelligent system, the data-driven tailored approach provided more or similarly efficient training. With the advancing techniques in artificial intelligence and integration of large language models,[45] user engagement of intelligence systems may improve substantially.

The training in this study involved one session with no follow-up. Trainees instructed by the ICEMS system achieved a mean composite score of -0.2 in the fifth repetition of the task, indicating that there is still a big room for improvement. Perhaps, longitudinal training with multiple training sessions is needed to improve performance further.

Although this study was conducted in a simulation training setting, the applications of intelligent instruction and assistance may not be limited to simulation settings. Methodologies are being developed to accurately identify surgical steps, potentially assess intraoperative performance during surgery, and provide feedback using artificial intelligence.[46,47] Obtaining performance data during surgery in realistic operating settings using real surgical instruments may enable transitioning intelligent feedback systems to the real operating room to mitigate errors during surgery.[48,49]

A limitation of the ICEMS system is that continuous task assessment may not accurately reflect the procedural outcome.[50] In some cases, trainees may demonstrate correct instrument utilization techniques without removing sufficient tumor. Both ICEMS and OSATS assessments are more focused on instrument technique than the operative outcomes. Future intelligent systems may need to determine the quality of the operative goals achieved to help trainees reach expert-level procedural outcomes while using correct instrument techniques. The ICEMS currently uses six algorithms to evaluate surgical performance and provide feedback in real-time[13,14]. Future versions of this system may incorporate additional modules to evaluate the procedural progress, outcome, and spatial information.[27,51]

The trainees' skillset may affect learning and capacity for performance improvement. Our study involved medical students with little to no surgical exposure. Their limited procedural knowledge may have provided a greater room for improvement in tumor resection skill acquisition, and this may resemble the scenario of a surgical trainee who just started their surgical training. Further studies may test intelligent systems in helping trainees with more procedural knowledge and a better skillset. Human expert instructors may adapt their instructions based on the trainees' needs and skill level. To maximize efficiency in training, intelligent systems may assess the trainees' skill level and adapt more advanced instructions as their skillset progresses from novice to expert level.[42,52]

In summary, this randomized controlled trial demonstrated an effective use of a real-time intelligent system in teaching bimanual surgical tumor resection skills that is more efficient when compared to in-person instruction from human experts. Using data-driven performance monitoring and intelligent feedback may not only help to meet the needs of competency-based surgical training but also provide an effective tool to sustain technical mastery.

References

1. Stulberg, J.J*., et al.* Association Between Surgeon Technical Skills and Patient Outcomes. *JAMA Surgery* (2020).
2. Ng, R., Chahine, S., Lanting, B. & Howard, J. Unpacking the Literature on Stress and Resiliency: A Narrative Review Focused on Learners in the Operating Room. *J Surg Educ* **76**, 343-353 (2019).
3. Ahsani-Estahbanati, E., Doshmangir, L., Najafi, B., Akbari Sari, A. & Sergeevich Gordeev, V. Incidence rate and financial burden of medical errors and policy interventions to address them: a multi-method study protocol. *Health Services and Outcomes Research Methodology* **22**, 244-252 (2022).
4. Leung, A*., et al.* "First, do no harm": balancing competing priorities in surgical practice. *Acad Med* **87**, 1368-1374 (2012).
5. Gabrysz-Forget, F., Zahabi, S., Young, M., Nepomnayshy, D. & Nguyen, L.H.P. "It's a Big Part of Being Good Surgeons": Surgical Trainees' Perceptions of Error Recovery in the Operating Room. *Journal of Surgical Education* **78**, 2020-2029 (2021).
6. de Montbrun, S.L. & Macrae, H. Simulation in surgical education. *Clin Colon Rectal Surg* **25**, 156-165 (2012).
7. Haluck, R.S. & Krummel, T.M. Computers and Virtual Reality for Surgical Education in the 21st Century. *Archives of Surgery* **135**, 786-792 (2000).
8. Gélinas-Phaneuf, N. & Del Maestro, R.F. Surgical Expertise in Neurosurgery: Integrating Theory Into Practice. *Neurosurgery* **73**, S30-S38 (2013).
9. Brightwell, A. & Grant, J. Competency-based training: who benefits? *Postgraduate Medical Journal* **89**, 107 (2013).
10. Yilmaz, R*., et al.* Continuous monitoring of surgical bimanual expertise using deep neural networks in virtual reality simulation. *npj Digital Medicine* **5**, 54 (2022).
11. Yilmaz, R., Winkler-Schwartz, A., Mirchi, N., Del Maestro, R. Continuously Monitoring Tools to Assess and Enhance Human Performance While Providing Error Avoidance Systems, patent No. 05001770-883USPR. . (2020).
12. Delorme, S., Laroche, D., DiRaddo, R. & Del Maestro, R.F. NeuroTouch: A Physics-Based Virtual Simulator for Cranial Microneurosurgery Training. *Operative Neurosurgery* **71**, ons32-ons42 (2012).
13. Sabbagh, A.J*., et al.* Roadmap for Developing Complex Virtual Reality Simulation Scenarios: Subpial Neurosurgical Tumor Resection Model. *World Neurosurgery* **139**, e220-e229 (2020).
14. Mirchi, N*., et al.* The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLOS ONE* **15**, e0229596 (2020).
15. Fazlollahi, A.M*., et al.* Effect of Artificial Intelligence Tutoring vs Expert Instruction on Learning Simulated Surgical Skills Among Medical Students: A Randomized Clinical Trial. *JAMA Network Open* **5**, e2149008-e2149008 (2022).
16. Martin, J.A*., et al.* Objective structured assessment of technical skill (OSATS) for surgical residents. *British Journal of Surgery* **84**, 273-278 (1997).
17. Ledwos, N*., et al.* Assessment of learning curves on a simulated neurosurgical task using metrics selected by artificial intelligence. *Journal of Neurosurgery*, 1-12 (2022).
18. Winkler-Schwartz, A*., et al.* Machine Learning Identification of Surgical and Operative Factors Associated With Surgical Expertise in Virtual Reality Simulation. *JAMA Netw Open* **2**, e198363 (2019).

19. Leppink, J., Paas, F., Van der Vleuten, C.P.M., Van Gog, T. & Van Merriënboer, J.J.G. Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods* **45**, 1058-1072 (2013).
20. Association, W.M. World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. *JAMA* **310**, 2191-2194 (2013).
21. Juszczak, E., Altman, D.G., Hopewell, S. & Schulz, K. Reporting of Multi-Arm Parallel-Group Randomized Trials: Extension of the CONSORT 2010 Statement. *JAMA* **321**, 1610-1620 (2019).
22. Cheng, A.*, et al.* Reporting guidelines for health care simulation research: extensions to the CONSORT and STROBE statements. *Advances in Simulation* **1**, 25 (2016).
23. Dean, W.H.*, et al.* Intense Simulation-Based Surgical Education for Manual Small-Incision Cataract Surgery: The Ophthalmic Learning and Improvement Initiative in Cataract Surgery Randomized Clinical Trial in Kenya, Tanzania, Uganda, and Zimbabwe. *JAMA Ophthalmology* **139**, 9-15 (2021).
24. Ledwos, N.*, et al.* Assessment of learning curves on a simulated neurosurgical task using metrics selected by artificial intelligence. *Journal of Neurosurgery* **137**, 1160-1171 (2022).
25. Wang, Z. & Shen, J. Simulation training in spine surgery. *Journal of the American Academy of Orthopaedic Surgeons* **30**, 400-408 (2022).
26. Yari, S.S., Jandhyala, C.K., Sharareh, B., Athiviraham, A. & Shybut, T.B. Efficacy of a virtual arthroscopic simulator for orthopaedic surgery residents by year in training. *Orthopaedic journal of sports medicine* **6**, 2325967118810176 (2018).
27. Logishetty, K., Rudran, B. & Cobb, J.P. Virtual reality training improves trainee performance in total hip arthroplasty: a randomized controlled trial. *The Bone & Joint Journal* **101-B**, 1585-1592 (2019).
28. Sweller, J. Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science* **12**, 257-285 (1988).
29. Young, J.Q., Van Merrienboer, J., Durning, S. & Ten Cate, O. Cognitive Load Theory: implications for medical education: AMEE Guide No. 86. *Med Teach* **36**, 371-384 (2014).
30. Sweller, J. Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load. *Educational Psychology Review* **22**, 123-138 (2010).
31. Kiyasseh, D.*, et al.* A vision transformer for decoding surgeon activity from surgical videos. *Nature Biomedical Engineering*, 1-17 (2023).
32. Kiyasseh, D.*, et al.* Human visual explanations mitigate bias in AI-based assessment of surgeon skills. *npj Digital Medicine* **6**, 54 (2023).
33. Tran, D.H.*, et al.* Quantitation of Tissue Resection Using a Brain Tumor Model and 7-T Magnetic Resonance Imaging Technology. *World Neurosurgery* **148**, e326-e339 (2021).
34. Winkler-Schwartz, A.*, et al.* Creating a Comprehensive Research Platform for Surgical Technique and Operative Outcome in Primary Brain Tumor Neurosurgery. *World Neurosurgery* **144**, e62-e71 (2020).
35. Yilmaz, R.*, et al.* Nondominant Hand Skills Spatial and Psychomotor Analysis During a Complex Virtual Reality Neurosurgical Task-A Case Series Study. *Oper Neurosurg (Hagerstown)* **23**, 22-30 (2022).
36. Sawaya, R.*, et al.* Virtual Reality Tumor Resection: The Force Pyramid Approach. *Operative Neurosurgery* **14**, 686-696 (2017).
37. Yilmaz, R. SubPialResection101-KFMC_scenario.xml_2015-Oct-22_14h06m26s_log.csv. https://doi.org/10.6084/m9.figshare.15132507.v1. (2021).

Figures

Figure-1: Flow diagram.

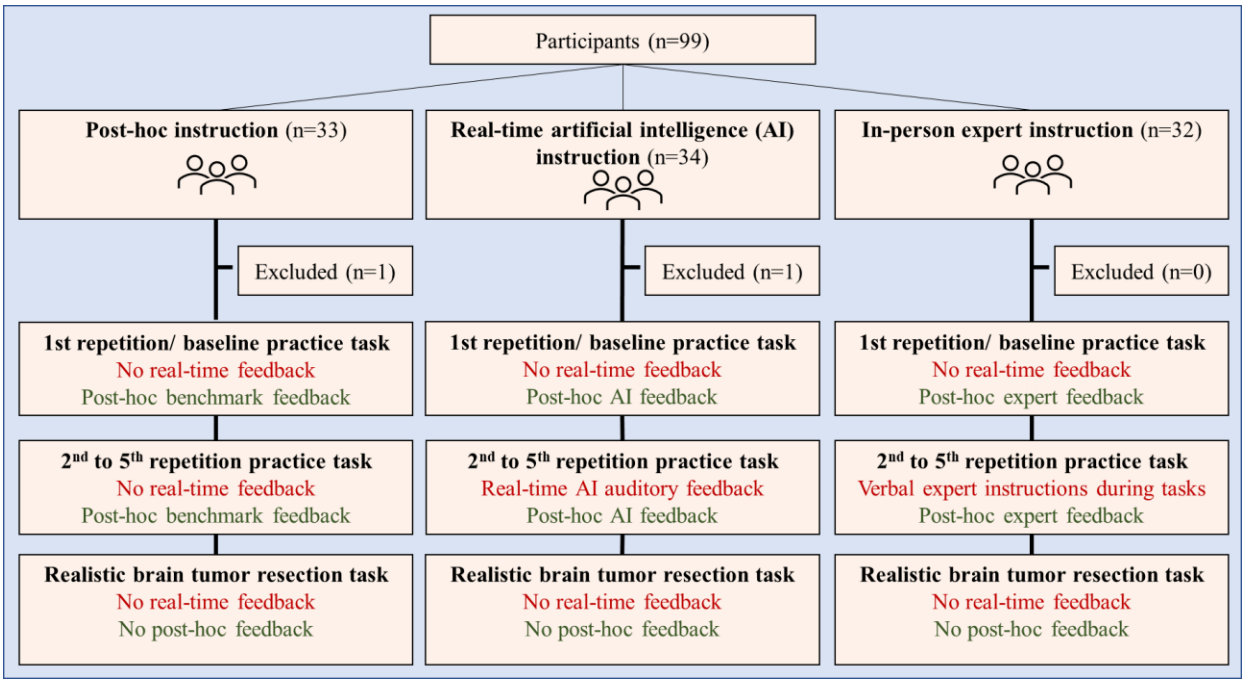
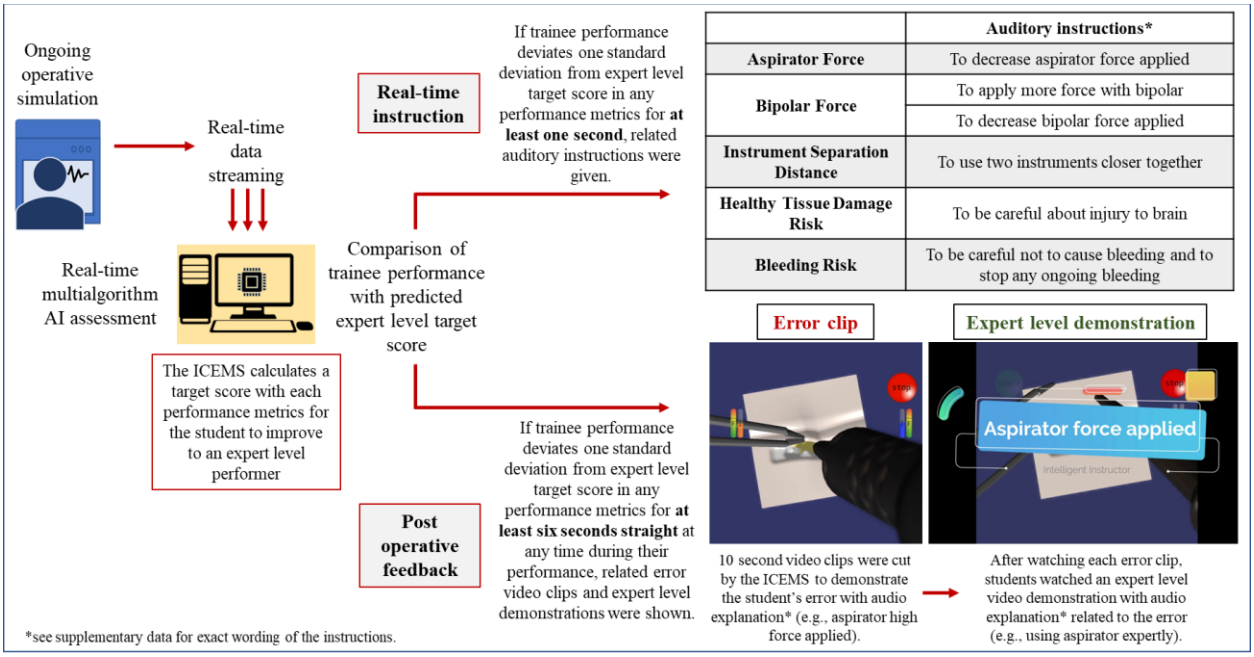


Figure-2: Real-time and post hoc ICEMS Feedback



	Auditory instructions*
Aspirator Force	To decrease aspirator force applied
Bipolar Force	To apply more force with bipolar
	To decrease bipolar force applied
Instrument Separation Distance	To use two instruments closer together
Healthy Tissue Damage Risk	To be careful about injury to brain
Bleeding Risk	To be careful not to cause bleeding and to stop any ongoing bleeding

Figure-3: ICEMS's composite-score across trials. Groups are color-coded (see the legend). X-axis represents the task repetition while Y-axis represents the ICEMS's composite score. The composite score is shown between -1 to 0; however, the maximum achievable score was +1. *Horizontal lines represent statistically significant differences ($p<.05$). For within-group differences, horizontal lines are represented with the respective color of the group. Vertical bars represent standard error.

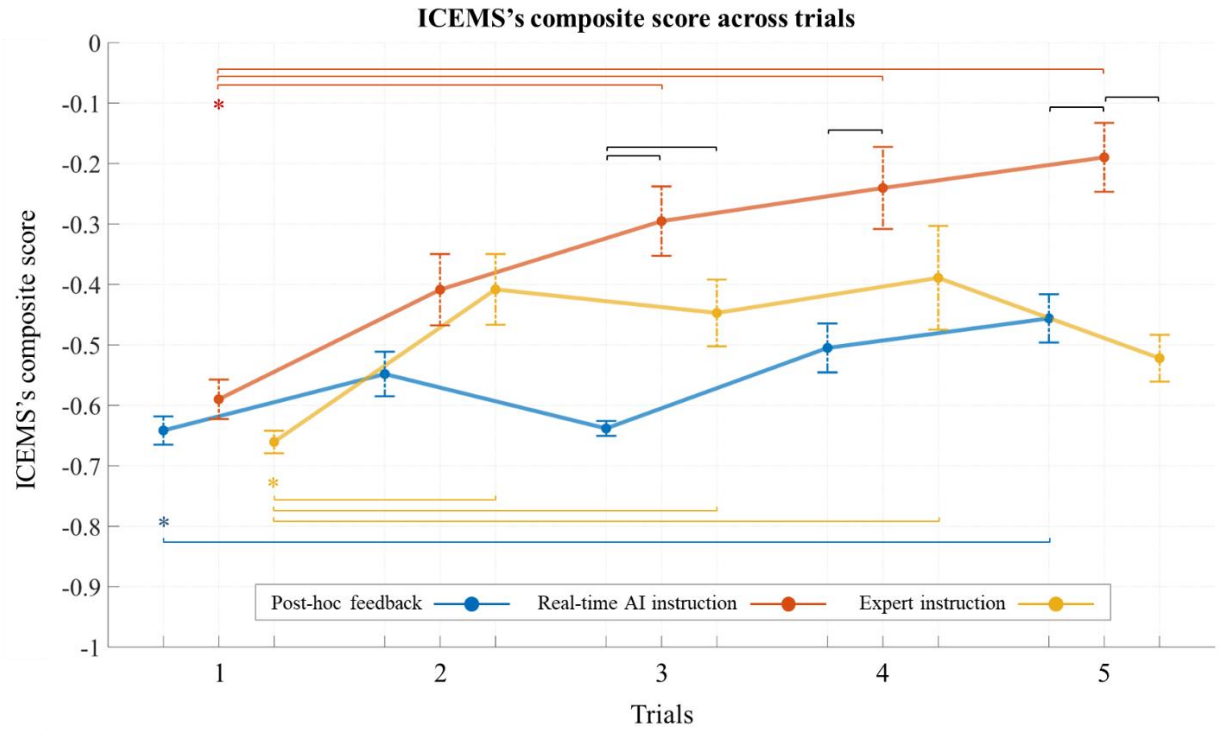
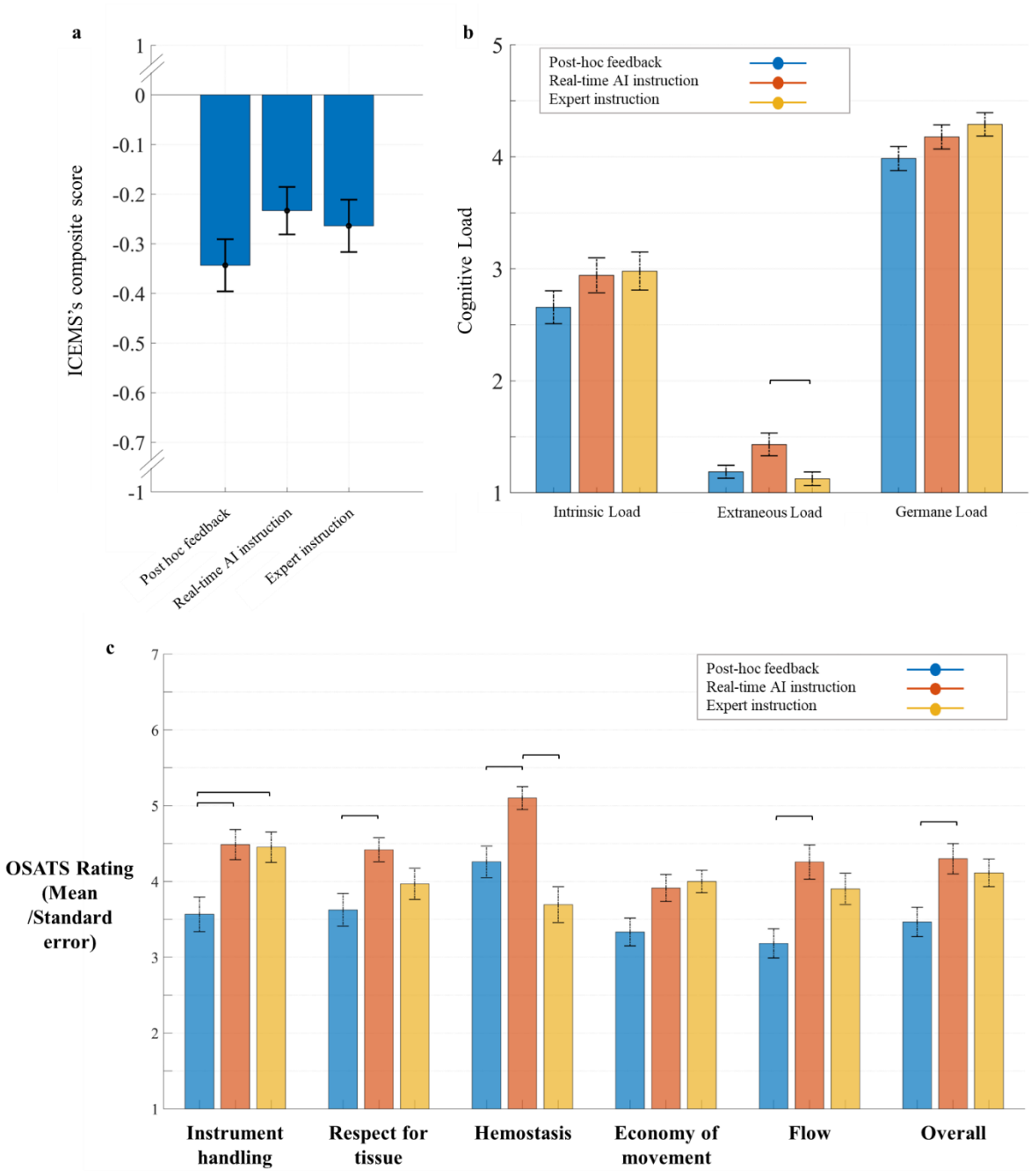


Figure-4: (a) ICEMS's composite-score in realistic task. The vertical bars represent standard errors. There was no significant difference between three feedback groups. **(b) Cognitive load.** Groups are color-coded (see the legend). The vertical bars represent standard errors. Participants who received real-time AI instruction reported significantly higher extraneous load than those received in-person expert instruction. There were no significant differences between groups concerning intrinsic load and germane load. **(c) Blinded expert OSATS rating.** Horizontal lines represent statistically significant differences ($p<.05$). Vertical bars represent standard error.


a
ICEMS's composite score
1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-1
Post hoc feedback
Real-time AI instruction
Expert instruction
b
Cognitive Load
5
4
3
2
1
Post-hoc feedback
Real-time AI instruction
Expert instruction
Intrinsic Load
Extraneous Load
Germane Load
c
OSATS Rating (Mean /Standard error)
7
6
5
4
3
2
1
Post-hoc feedback
Real-time AI instruction
Expert instruction
Instrument handling
Respect for tissue
Hemostasis
Economy of movement
Flow
Overall

Table

Table: Participant Characteristics.

	Group 1 Post-hoc feedback (n=32)	**Group 2** Real-time AI feedback (n=33)	**Group 3** Expert instruction (n=32)	**All Participants** (n=97)
Mean age +/- SD (range)	21.1 +/- 2.4 (19-26)	21.4 +/- 3.0 (17-31)	21.3 +/- 2.8 (17-31)	21.3 +/- 2.7 (17-31)
Male/Female	10/22	14/19	15/17	39/58
Handedness (Right/Left/Ambidextrous)	28/4/0	30/3/0	31/0/1	89/7/1
Year in medical school:				
Preparatory year	9	8	9	26
1st	20	23	13	56
2nd	3	1	6	10
3rd	0	0	4	4
4th	0	1	0	1
Level of interest in surgery, median (range)	4 (2-5)	4 (2-5)	4 (1-5)	4 (1-5)
Completed surgical rotation (Y/N)	0/31	1/33	3/29	4/93
Medical School:				
McGill University	15	16	12	43
University of Montreal	10	6	8	24
University of Sherbrooke	0	4	6	10
University of Laval	7	7	6	20
Playing video games:				
Not at all	18	24	17	59
1-5 hours per week	11	6	11	28
6-10 hours per week	2	1	2	5
>10 hours per week	1	2	2	5
Playing musical instruments (Y/N)	15/17	15/18	16/16	46/51
Previous activities that require hand dexterity	13/19	17/16	13/19	43/54
Previously used virtual reality simulation (Y/N)	1/31	0/33	0/32	1/96

Chapter 6 – Summary and Conclusions

General Findings

The overarching goal of this thesis work is to develop, validate, and test objective and standardized assessment and teaching methodologies for virtual reality surgical skills training, integrating AI.

Chapter 2 demonstrated the development and predictive validation of a real-time intelligent system, the Intelligent Continuous Expertise Monitoring System (ICEMS). This system allowed for conducting several randomized controlled trials at the Neurosurgical Simulation and Artificial Intelligence Learning Centre to investigate the integration of real-time AI instructions in training. The most efficient training is to be explored by modifying the feedback delivery of this system while using the same background ICEMS algorithms. The ICEMS's ability to differentiate between expertise levels and between trainees demonstrated the promising future of AI in designing surgical curricula while accurately tracking trainee progress as they master their technical skills.

The work in Chapter 3 provided spatial performance analysis during a simulated brain tumor resection. This work informed the randomized controlled trial (RCT) in Chapter 4 that the 3D feedback may be necessary for trainees to appreciate the spatial surgical environment they are interacting with. The randomized controlled trial in Chapter 4 involved the testing of feedback delivery to maximize the efficiency in learning. In accordance with the previous literature, learners who receive performance feedback improved their performance quicker than those who practice without feedback. Additionally, more engaging colored and visuospatial feedback helped trainees learn even more efficiently. These findings expanded our knowledge on

methodologies to design effective training, limiting cognitive overload while keeping the trainees engaged and providing action-oriented quantitative information.

Finally, Chapter 5 involved the testing of the ICEMS, our first-ever real-time intelligent assessment, instruction, and risk detection system. The results of this RCT have demonstrated that, in fact, real-time AI assistance may be more efficient in teaching surgical technical skills than traditional learning via human expert instruction. Although this initial application is way away from being perfect and may be missing many important considerations in surgical education, the promising results will make a series of RCTs possible to improve the ICEMS's ability to deliver feedback in the most informative, concise, clear, and engaging way.

Important Considerations

Virtual reality simulations offer many advantages, which made the applications and the AI integration in this PhD work possible. However, there are limitations worth noting. Virtual reality simulators may be missing many important factors present in a real operating room. This brings the question of whether the promising findings in our randomized controlled trials would be transferrable to the skillset in real operating rooms.

Further studies are needed to assess whether learning in a simulated environment can improve intraoperative performance. Considering the current stage of our research, applications in real operating theatre might be too challenging due to costs, safety and privacy concerns, and regulations. However, at the Neurosurgical Simulation and Artificial Intelligence Learning Centre, intermediary platforms between virtual reality simulation and the real operating room are being developed.[1] These platforms provide performing tasks on ex vivo animal models involving biologic tissues, with scenarios resembling real-life cases such as tumor resection or epilepsy

surgery. Optical cameras were used to monitor and collect data from instrument movements.[2] Additionally, the integration of pre-op and post-op MRI scans allowed for the measurement of the amount of tissue resected and remaining tumor.[3] Performing in such a realistic simulation environment may be more reflective of performance in real operating theatre, hence the transfer of skills to this more realistic setting can be measured. A future study may explore if performance improvement in virtual reality simulation using the systems outlined in this PhD work may result in performance improvement on an ex vivo simulation task. However, before this step, the ICEMS needs to be fine-tuned, using transfer learning, to accurately assess the performance during ex vivo simulations to allow for objective comparisons.

Surgery involves many steps from planning, execution, and follow-up. Intraoperative process is also a combination of many steps from the preparation of the patient, opening skin and skull, access to the area of interest, tumor removal, and closure of structures such as dura mater, and skin. Additionally, neurosurgery involves many technically challenging procedures close to neural structures such as the cerebellum, midbrain, pons, and medulla, nearby cranial nerves, and vasculatures. The cerebral cortex, the largest area of our brain, controls voluntary muscle control, sensation, memory, emotions, and executive functions. Neurosurgical interventions on or close to any of these important structures may result in permanent loss of patient functionality. Therefore, skillset mastery in procedures such as neurovascular, brainstem, posterior fossa, and transsphenoidal surgeries is of utmost importance. It is important to note the limitation that our work involved only one skill which was the subpial removal of the simulated tumors. This might be an important skill and trainees may not have other opportunities to practice tumor resection skills except in virtual reality simulation. Ideally, all skills that are part of neurosurgery should be

integrated into virtual reality simulation training in the future where a more comprehensive assessment and feedback can be provided.

AI is used to understand patterns in the data that can not be outlined using simple equations. This provides great advantages; however, AI may lack transparency and interpretability due to its complex structure, which relates to the 'black box' problem that is discussed in Chapter 1. To overcome this issue, human input may be necessary, which is often referred to as the 'human-in-the-loop' approach.[4] This approach is especially important to handle complex and uncertain situations in which AI may not produce meaningful and reliable information. In teaching surgical mastery, the inclusion of surgeon experts in the loop may benefit the systems to correctly interpret AI decisions and align the information for the students to understandable and relevant forms. Finally, AI algorithms may get over-tuned to biases the datasets contain. Larger multi-institutional datasets help to address this problem, allowing for more generalizable applications. Similar to other healthcare areas, surgery is a high-stakes domain, therefore careful considerations in AI applications and expert revision of these systems are of utmost importance.

Our feedback and AI models were based on expert data from a single institution which may include biases towards how the procedures are done in this specific location. Using multi-institutional datasets would help to overcome this issue. The expert dataset included only one female expert. This may cause a gender bias towards men in case there are inherent differences between men and women in the way the instruments are used. More balanced datasets help to develop more reliable models and would also inform about the effect of gender on surgical technical skills.

In a learning environment, it is important to provide students with information in a language that they are comfortable with. The randomized controlled trials (RCT) in this thesis work were conducted at McGill University, located in Montreal, that has both English and French speaking student body. The participating students were competent in English. The informed consent form was provided in both languages. However, the feedback provided in the RCTs were optimized for English speaking students and the recruiting researchers as well as expert instructors were English speakers. This feedback setting may have caused extra stress for those who are not as competent in English, however due to randomization, we have not expected this issue to influence our results. In the future, feedback systems can be implemented with different languages where trainees can engage in learning in a language of their choice.

The Promise of Artificial Intelligence in Surgery

'Artificial intelligence is the new electricity' says Andrew Angy. Without so much hesitation, it is fair to say that developing AI systems will continue to amaze us and shape the future of our lives and medicine. Following the idea in the quote, all electronic devices we use may become integrated into the intelligent systems that provide assistance and make our lives easier. Considering a surgical theatre, having the operating environment optimized by intelligent systems using the data from surrounding devices could maximize safety and efficiency. Such a futuristic setting would be beneficial for care providers, hospitals, insurance companies, and most importantly, patients.

Applications of AI in surgery may change the practice from preoperative planning to intraoperative guidance.[5,6] The field of robotic surgery has been considered a good candidate for exciting AI applications with access to multichannel surgical data from robotic arms. However,

despite the significant potential, limited information is available on AI's efficacy in improving patient safety in robot-assisted surgery.[7] Computational resources are improving and becoming more widely available, allowing for the analysis of extensive healthcare data. This facilitates the discovery of hidden knowledge, risk identification, and enhanced communication within the clinical field.[8] The integration of technology and AI into surgery may increase the autonomy of the systems to function with less input from the surgeon, towards fully automated diagnosis and care, similar to the development of self-driving cars, as outlined by Kitaguchi et al.[9] Ethical and legal considerations may become increasingly a hot topic in the implementation of AI into surgery.[10] The four medical ethics principles: autonomy, beneficence, nonmaleficence, and justice should be strongly considered before AI integration into the healthcare system.[11]

Integration of AI tools into surgical education requires careful planning. One of the immediate goals for these tools is to demonstrate their effectiveness in comparison to the traditional educational models with improved outcomes. Their integration may yield reduced operational costs and increased safety and efficiency. In this process, feedback from faculties and trainees may be of utmost importance to tune these systems to meet the needs of educators and trainees most efficiently. As these developments encounter ethical and regulatory considerations, topics such as the ownership of the data, the extent of commercial use of patient data, transparency, explainability, and potential biases these systems may have will become some key discussion points. Adaptation of these systems may benefit from building the technical infrastructure through pilot projects and testing.

High-Fidelity Simulation Data

One of the big limits of today's AI application is the limits with the data available. Simulation systems allow for the creation of realistic settings with minimal resources and costs with unlimited task repetitions. As such, simulation data can be used to train algorithms which may only need to be fine-tuned using small real-life samples before they are applied in real life. This may allow to overcome the limitations with access to data.

Especially in high-risk situations simulation data can be important. Fortunately, real-life intraoperative cases don't involve many accidents and errors. However, to train a system for an accident, related data needs to be shown to the algorithms to demonstrate 'what is an accident/error'. In different industries such as the aircraft industry, simulation systems are used to create accidental scenarios, so the nature and the consequences of the accident can be examined. Simulations help decrease the costs by digitizing the real-world and they also enable obtaining a vast data, which may not be possible to record from the real-world. For example, crashing a test airplane in real-life to study potential accidents is very costly, and the data recorded during the crash can be very limited. However, majority of the time, these accidents can be virtually simulated with low costs. As a more medical example, in surgery, accidents increase patient morbidity and mortality, therefore simulation systems may offer 'the only solution' that would produce large datasets related to accidents/errors. One advantage in the development of the ICEMS system was that we were able to have medical students with little to no experience in surgery to operate brain tumor resections, which cannot be done in real life. Such a setting allowed us to produce a dataset with a variety of mistakes that a very novice person would make. This enabled our AI system to examine the patterns during the poor handling of the instruments

and the execution of brain tumor surgery. Learning from this and comparing it to an expert-level performance the ICEMS enabled continuous tracking of trainee progress as they learn.

AI models can be used to develop simulations and vice versa, where simulation data can be used to train AI models. As an interesting futuristic thought, imagine a setting where AI produces the data it needs to learn from, similar to a person who is preparing him or herself for a task, imagining many possible scenarios that may happen, so that he/she is prepared. Being prepared would allow that person to handle these situations better and make meaningful interpretations, just like AI may become more and more accurate by simulating scenarios and subsequently learning from the simulated consequences.

Popular AI Domains

Some AI applications have increasing popularity. One noteworthy application is computer vision, which is a subtype of AI that mainly involves image and video analysis. Computer vision applications in medicine include the analysis of visual information such as X-rays, MRI data, pathology and histology slides, and pictures of various skin conditions. Surgical video assessment is one of the use cases of computer vision where surgical performance is monitored by algorithms for purposes such as instrument detection,[12] surgical procedure recognition, predicting hemorrhage,[13] and phase detection.[14,15] This application promises a future as intraoperative videos are rather easily collected through video cameras without interfering with the flow of surgery, and without compromising the sterile environment. Furthermore, cameras are already integrated into devices such as surgical microscopes. This convenient nature may allow the development of smart cameras in the future, integrated into real-surgical theatre, to monitor performance and provide safety assistance during surgery.[16]

Artificial general intelligence (AGI) refers to AI systems possessing broad intelligence across a variety of domains, similar to human cognitive skills. The majority of the time, these systems are trained using immense data available online. Specially trained AI models have specific capabilities in specific tasks; however, AGI models, such as large language models, provide flexible intelligence and diverse intellectual capabilities. Generative artificial intelligence (GAI) is AI that is capable of producing original work, including images, text, audio, or other media. These systems learn contexts and patterns from the existing content and generate resembling but authentic new instances. Generative AI is used commonly with large language models to create human-like interaction, as well as in generating art, images, and music.

Large language models (LLMs) such as Chat GPT (Chat Generative Pre-Trained Transformer) are groundbreaking developments, widely appreciated and utilized. These systems are AGI and GAI, and they learn how to think like humans from all the resources available online, interact with a person, and provide accurate information, majority of the time, in different domains. For example, ChatGPT passed the United States Medical Licensing Examination (USMLE),[17,18] and the bar exam,[19] and it would rank in the top 1% of microeconomics and macroeconomic exam takers,[20] although it has not been specifically trained in any of these areas. LLMs are being implemented in medicine in education, research and practice.[21] The applications of ChatGPT in medicine include drug development, medical literature reviewing, improving data analysis and personalized medicine.[22] Checco et al. demonstrated that AI was able to predict the outcome of peer-review and provide comments resembling the reviewer comments.[23] GPT-4 achieved a passing score in neurosurgical written board examination.[24]

One limitation in our simulation training application was the high cognitive load due to difficulties in understanding the explanations and instructions provided by the real-time system.

This may have happened due to missing context in AI's auditory instructions. Implementation of large language models may mitigate this problem by enabling human-like interaction. These systems may interact with the student verbally, communicate with them back and forth, like a surgeon supervisor talking with a resident when they operate, explaining how to skillfully perform the procedure and allowing the resident to ask questions when they need clarifications or have questions. Such interaction may reduce the cognitive load and increase trainee engagement.

Augmenting Access to Data

To enable a wider implementation of AI in surgery and medicine, one important consideration is to augment data collection. The transition of applications from analog to digital may allow for data collection. The majority of the time digital patient information is used momentarily to watch/monitor the situation (such as an intraoperative microscopic camera) and is lost afterward since hospitals do not have systems to store the data. A systematic approach to data collection and storage would increase the chance of possible AI implementations which would ultimately increase patient and clinicians' safety and improve outcomes. This can be achieved by increasing the collaborative work between clinicians, health governors, AI engineers, and industry.

As an example, intraoperative microscopes are routinely used in surgery, particularly in neurosurgery. Currently, these systems are used to allow neurosurgeons to visualize the operative area and enable bystanders to follow the procedure. The footage is rarely recorded, most of the time for training purposes. This footage may be used by an expert to rate the performance of the operator using a rating scale such as the OSATS to provide feedback. A more systematic approach with the recording of a high volume of cases would facilitate the training of AI models

to assess surgical performance and provide feedback. Further implementations, such as the prediction of patient outcomes based on surgical video recording would contribute to the clinical practice, providing deeper insights into what constitutes surgical technical expertise.

Potential Pitfalls

A multidisciplinary, both medical and engineering, expertise is necessary to utilize AI tools correctly to sufficiently meet the needs. Medical expertise is important to outline the needs correctly and guide engineers to develop and integrate tools in the most efficient ways for clinicians. AI applications involve tiny decisions that may influence the overall performance of the tools. It is possible that AI systems may be biased. These biases occur due to the biases in the data they are trained on as well as wrong interpretations or missing considerations. In an extended analysis of acquired skills in a simulation training, it was shown that trainees may not always progress towards optimal learning outcomes, or they may overrespond to some instructions.[28] The gap in medical knowledge among engineers and the gap in engineering knowledge among clinicians may cause to less reliable systems when they work without input from one another. Not following a systematic approach in AI model training may result in false results. Reporting of AI applications without separate testing may hide important background problems.

Ideally, all AI applications should follow a step where the models are separately tested on independent datasets to measure the generalizability of the system. Some applications miss this important step. Although these applications demonstrate high training and validation accuracies, these results can simply, and likely, be overfitting. Hence, their real-life applications might be limited, or contain biases. In certain disciplines, for example, due to limited data, proper

applications may not be possible. Nevertheless, being open about the limitations and potential biases of these systems may help the development of more reliable systems.

High computational resources are increasingly available to a wider population. There is a growing population with AI expertise. This increases the rate of developments in AI as more people are involved in this area. In the scientific area, the peer review process has limitations as more complex AI applications are introduced, and they may appear as a 'black box'. The review process may not provide the necessary expertise to outline shortcomings and biases that the AI application may involve. Quality check guidelines are being implemented for reporting AI applications.[25] These checklists help to ensure the quality of the work such as reporting of the clinical trials that involve AI and outline AI best practices.[26,27]

Conclusion

An all-in-one sophisticated AI integration was made into surgical simulation for technical skills training, with the ability to assess performance, provide feedback, and risk mitigation. This PhD work outlined important methodologies to improve learning by providing visuospatial information and increasing student engagement. This work may have paved the way for further research to explore the optimal real-time feedback assessment and feedback for surgical bimanual skills training.

References

1. Alsayegh, A., Bakhaidar, M., Winkler-Schwartz, A., Yilmaz, R. & Del Maestro, R.F. Best Practices Using Ex Vivo Animal Brain Models in Neurosurgical Education to Assess Surgical Expertise. *World Neurosurgery* (2021).
2. Winkler-Schwartz, A*., et al.* Creating a Comprehensive Research Platform for Surgical Technique and Operative Outcome in Primary Brain Tumor Neurosurgery. *World Neurosurgery* **144**, e62-e71 (2020).
3. Tran, D.H*., et al.* Quantitation of Tissue Resection Using a Brain Tumor Model and 7-T Magnetic Resonance Imaging Technology. *World Neurosurgery* **148**, e326-e339 (2021).
4. Wu, X*., et al.* A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* **135**, 364-381 (2022).
5. Zhou, X.Y., Guo, Y., Shen, M. & Yang, G.Z. Application of artificial intelligence in surgery. *Front Med* **14**, 417-430 (2020).
6. Jia, S*., et al.* Performance evaluation of an AI-based preoperative planning software application for automatic selection of pedicle screws based on computed tomography images. *Frontiers in Surgery* **10**(2023).
7. Moglia, A., Georgiou, K., Georgiou, E., Satava, R.M. & Cuschieri, A. A systematic review on artificial intelligence in robot-assisted surgery. *International Journal of Surgery* **95**, 106151 (2021).
8. Choudhury, A. & Asan, O. Role of Artificial Intelligence in Patient Safety Outcomes: Systematic Literature Review. *JMIR Med Inform* **8**, e18599 (2020).
9. Kitaguchi, D., Takeshita, N., Hasegawa, H. & Ito, M. Artificial intelligence-based computer vision in surgery: Recent advances and future perspectives. *Ann Gastroenterol Surg* **6**, 29-36 (2022).
10. Morris, M.X., Song, E.Y., Rajesh, A., Asaad, M. & Phillips, B.T. Ethical, Legal, and Financial Considerations of Artificial Intelligence in Surgery. *Am Surg* **89**, 55-60 (2023).
11. Farhud, D.D. & Zokaei, S. Ethical Issues of Artificial Intelligence in Medicine and Healthcare. *Iran J Public Health* **50**, i-v (2021).
12. Wang, Y., Sun, Q., Liu, Z. & Gu, L. Visual detection and tracking algorithms for minimally invasive surgical instruments: A comprehensive review of the state-of-the-art. *Robotics and Autonomous Systems* **149**, 103945 (2022).
13. Pangal, D.J*., et al.* Expert surgeons and deep learning models can predict the outcome of surgical hemorrhage from 1 min of video. *Scientific Reports* **12**, 8137 (2022).
14. Garrow, C.R*., et al.* Machine Learning for Surgical Phase Recognition: A Systematic Review. *Annals of Surgery* **273**(2021).
15. Birkhoff, D.C., van Dalen, A.S.H.M. & Schijven, M.P. A Review on the Current Applications of Artificial Intelligence in the Operating Room. *Surgical Innovation* **28**, 611-619 (2021).
16. Kiyasseh, D*., et al.* A multi-institutional study using artificial intelligence to provide reliable and fair feedback to surgeons. *Communications Medicine* **3**, 42 (2023).
17. Mbakwe, A.B., Lourentzou, I., Celi, L.A., Mechanic, O.J. & Dagan, A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. *PLOS Digit Health* **2**, e0000205 (2023).
18. Gilson, A*., et al.* How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med Educ* **9**, e45312 (2023).
19. Katz, D.M., Bommarito, M.J., Gao, S. & Arredondo, P. Gpt-4 passes the bar exam. *Available at SSRN 4389233* (2023).

20. Geerling, W., Mateer, G.D., Wooten, J. & Damodaran, N. ChatGPT has Aced the Test of Understanding in College Economics: Now What? *The American Economist* **68**, 233-245 (2023).
21. Shoja, M.M., Van de Ridder, J.M. & Rajput, V. The emerging role of generative artificial intelligence in medical education, research, and practice. *Cureus* **15**(2023).
22. Ruksakulpiwat, S., Kumar, A. & Ajibade, A. Using ChatGPT in Medical Research: Current Status and Future Directions. *J Multidiscip Healthc* **16**, 1513-1520 (2023).
23. Checco, A., Bracciale, L., Loreti, P., Pinfield, S. & Bianchi, G. AI-assisted peer review. *Humanities and Social Sciences Communications* **8**, 25 (2021).
24. Ali, R.*, et al.* Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations. *Neurosurgery* **93**, 1353-1365 (2023).
25. Liu, X.*, et al.* Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nature Medicine* **25**, 1467-1468 (2019).
26. Liu, X.*, et al.* Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine* **26**, 1364-1374 (2020).
27. Winkler-Schwartz, A.*, et al.* Artificial Intelligence in Medical Education: Best Practices Using Machine Learning to Assess Surgical Expertise in Virtual Reality Simulation. *J Surg Educ* **76**, 1681-1690 (2019).
28. Fazlollahi, A.M.*, et al.* AI in Surgical Curriculum Design and Unintended Outcomes for Technical Competencies in Simulation Training. *JAMA Network Open* **6**, e2334658-e2334658 (2023).