Improving neurosurgical operative performance through virtual reality simulation and artificial intelligence

Alexander Winkler-Schwartz

Integrated Program in Neuroscience Department of Neurology and Neurosurgery Faculty of Medicine, McGill University, Montreal, Quebec, Canada

November 2022

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

Table of Contents

Abstract

Surgical interventions carry substantial risks to patients and increased costs to health care systems. This risk is particularly true for brain tumor surgery, due to the fragile nature of the central nervous system. Given this issue, the demand for an objective demonstration of surgical competence from stakeholders ranging from patients' rights groups, governmental organizations, insurance agencies and hospital administrations is increasing. Similarly, as residency programs continue to move towards a competency-based curriculum, there is an increasing need for assessment of resident technical skills. As such, the use of virtual reality surgical simulators has been explored as a means of providing objective assessments in surgery. These simulators can track all movements and forces of simulated instruments, generating enormous datasets in the process. Artificial intelligence/machine learning systems lend themselves well to the analysis of such large datasets and their application to evaluate performance on virtual reality simulators has led to an increase in the volume and complexity of publications which bridge the fields of computer science, medicine and education. However, there remains a paucity of evidence that operative rehearsal enhances surgical performance in oncological neurosurgery.

The purpose of this dissertation is to address several unanswered questions about the role of simulation training in neurosurgery and to lay the groundwork for a future randomized controlled trial. As such, we outline the relevant background in the training of neurosurgeons, as well as the rise of simulation and artificial intelligence in medicine. We reviewed the literature on studies involving AI in surgical simulation and discovered that important communication gaps exist between medicine/education and computer-science/engineering. Consequently, we sought to bridge these gaps by introducing standardized reporting guidelines. These guidelines were applied as we sought to assess the feasibility of machine learning to assess surgical skills in

neurosurgical simulation and demonstrated that participants could be classified into 4 levels of expertise with 90% accuracy. Findings here may serve as a precursor to a machine learning powered experimental training arm in a future RCT. Furthermore, to understand the limitations surgical evaluation by supervision, we compared expert ratings with computer generated metrics on a simulated oncological neurosurgical procedure. Most notably, force exerted on tissues, an important aspect of neurosurgical technique, was less well captured by the expert evaluators. Finally, we outlined a framework whereby the primary outcome of the RCT could be evaluated, focusing on extent of resection of brain and tumor as assessed on MRI and visual inspection via tumor fluorescence, in addition to a further secondary outcome involving surgical movement capture.

Computerized simulation technology and artificial intelligence systems represent the latest iteration in man's quest to best understand and improve the world around him. Applied to neurosurgery, these innovations introduce significant technological and philosophical disruptions to an otherwise conservative and judicious field. Given its inherently high-stakes nature, neurosurgery represents a litmus test for medicine at large. Simply put, if these technologies can be successfully applied in the neurosurgical context, they can be applied to other domains in medicine as well.

It is my hope the innovations outlined in this thesis will have a positive impact on patients afflicted with neurosurgical conditions, as well as any patient undergoing an interventional procedure within the medical system.

Résumé

Les interventions chirurgicales exposent les patients à des risques importants et imposent des coûts élevés aux systèmes de santé. C’est particulièrement le cas pour celles visant le traitement des tumeurs cérébrales, en raison de la fragilité du système nerveux central. Par conséquent, les parties prenantes, comme les groupes de défense des droits des patients, les organismes gouvernementaux, les assureurs et les administrations hospitalières, exigent de plus en plus une démonstration objective des compétences chirurgicales. De même, puisque les programmes de résidence continuent de mettre l’accent sur les compétences, l’évaluation des aptitudes techniques des résidents revêt une importance accrue. C’est pourquoi l’utilisation de simulateurs chirurgicaux de réalité virtuelle a été explorée afin de fournir des évaluations objectives en chirurgie. Ces simulateurs peuvent reproduire tous les mouvements ainsi que la force des instruments simulés, générant au passage de vastes ensembles de données. Les systèmes d’apprentissage automatique ou par l’intelligence artificielle se prêtent bien à l’analyse de ces grands ensembles de données. Leur utilisation dans l’évaluation de la performance des simulateurs de réalité virtuelle a entraîné une augmentation du volume et de la complexité des publications, qui combinent dorénavant l’informatique, la médecine et l’éducation. Malgré cela, il existe encore peu de données probantes de grande qualité (niveau 1) montrant l’incidence de la répétition opératoire sur la performance chirurgicale dans le domaine de la neurochirurgie oncologique.

L’objectif général de ma thèse consiste à la fois à mieux comprendre le rôle de la simulation et de l’intelligence artificielle dans la formation des résidents en neurochirurgie et à préparer le terrain pour un futur essai contrôlé randomisé. Le chapitre 1 décrit les aspects pertinents de la formation des neurochirurgiens ainsi que l’essor de la simulation et de

l’intelligence artificielle en médecine. Le chapitre 2 présente une revue de littérature afin d’établir une méthodologie cohérente et des lignes directrices en matière de rapports dans les études traitant d’intelligence artificielle et de simulation chirurgicale. Le chapitre 3 met les éléments précédents en application. Nous y démontrons qu’il est possible d’utiliser l’apprentissage automatique pour évaluer les compétences chirurgicales dans le cadre d’une simulation neurochirurgicale. Les résultats obtenus pourraient préparer à l’utilisation future d’un group expérimental optimisé par l’apprentissage automatique et destiné aux études cliniques randomisées (ÉCR). Au chapitre 4, nous avons cherché à établir une échelle d’évaluation visuelle fiable pouvant servir de critère d’évaluation secondaire dans une ÉCR. Enfin, le chapitre 5 décrit un cadre permettant d’étoffer le critère d’évaluation principal de l’ÉCR, notamment l’étendue de la résection du cerveau et de la tumeur évaluée par IRM et par l’inspection visuelle en fonction de la fluorescence de la tumeur, et l’ajout d’un critère secondaire relatif à la capture du mouvement chirurgical.

La technologie de simulation informatisée et les systèmes d’intelligence artificielle constituent les plus récentes avancées dans la quête de l’être humain pour mieux comprendre et améliorer le monde qui l’entoure. Appliquées à la neurochirurgie, ces innovations entraînent des bouleversements technologiques et philosophiques importants dans un domaine par ailleurs conservateur et empreint de rationalisme. Compte tenu des risques élevés inhérents à la neurochirurgie, l’utilisation de ces technologies dans un tel contexte constitue un test décisif pour la médecine dans son ensemble. Autrement dit, si ces technologies peuvent être appliquées avec succès dans le contexte neurochirurgical, elles peuvent également l’être dans les autres disciplines médicales.

J’espère que les innovations décrites dans la présente thèse auront une incidence positive sur les patients atteints de troubles neurologiques ainsi que sur tout patient du système de santé devant subir une intervention chirurgicale.

Acknowledgements

With the completion of this PhD, I will successfully bring to a close over two and a half decades of formal education. This section must invariably be briefer than I wish it to be. To make it any longer would give the false impression that I have adequately thanked all those who have accompanied me in my journey from undifferentiated layperson to PhD-wielding neurosurgeon. These words will always fall short of capturing the depth of my gratitude.

My father, George, deserves the credit as the person whose unwavering support truly allowed me to reach my full potential. It was through him that my curiosity was always nourished and allowed to grow. It was through him that I discovered my love of the brain; a passion that has remained the central theme of my life. It has led me to medicine and a career in neurosurgery. It has been the lens through which I have best understood the scientific and spiritual dimensions of my existence. His love and attention are the best example of what it means to be a fantastic father, and how this can allow a life to flourish. I owe my father everything.

One of my father's final gifts to me was to bring me into contact with Dr. Del Maestro. I consider myself extremely lucky to have been at the receiving end of his intellect and deep kindness. This PhD would not be possible without his careful guidance. A true renaissance man, through him I have learnt what is required to be a whole person: one who nourishes esthetic as well as intellectual pursuits. I do not know of any neurosurgeon more dedicated to his patients, researcher more dedicated to his work, or mentor more dedicated to his students. Through our countless interactions both in and out of the lab, I have learnt not just how to conduct research and medicine, but how to carry myself in life.

I would also like to thank Dr. Lajoie, my co-supervisor, for her input throughout the PhD, as well as Hamed Azarnoush, Aiden Reich, Dan Tran, Vincent Bissonnette, Nicole Ledwos, Nykan Mirchi, Robin Sawaya, Fahad Alotaibi, Gmaan Al-Zhrani and other members of the Neurosurgical Simulation and Artificial Intelligence Learning Center. Recai Yilmaz, from the lab, has been an inspiration to work with and I have enjoyed watching him begin his foray into the much-deserved upper tiers of academic accomplishment.

I am extremely grateful to my friends and family, without whom life is not worth living. The Maniakas/Kehayia's have been particularly instrumental during the earlier years in medicine. My mother, Petra, has ensured that I not forget about the creative dimensions in my life. My long-time friends Evan, Wolf, Chris, Adrian and George have shared in all the highs and lows that life has to offer. The older I get the more I continue appreciate them. I thank the Franklins for their bottomless coffee supply and intellectually vigorous discussions at the dinner-table during my self-imposed impromptu writing retreats on Texada Island.

Finally, I cannot express enough my gratitude for Yasmine, my best friend, fellow surgeon and researcher, intellectual, wife and mother of Zayna. My earliest memories of conducting medical research were us, sitting on the floor, surrounded by scattered papers, teaching ourselves how to conduct a t-test. It has been quite a ride since then, and you have been my pillar of support throughout. My success is most definitely your success.

Contribution to Knowledge

Chapters 2, 3, 4 and 5 constitute original scholarship and have been published in peer-reviewed journals. This work has contributed to the advancement of surgical education within neurosurgery in the following ways:

Chapter 2 outlines the development of the Machine Learning to Assess Surgical Expertise (MLASE) checklist, which researchers can utilize when producing and reviewing virtual reality manuscripts involving machine learning to assess surgical expertise. A standardized approach in the reporting of these publications will allow researchers from these fields to form a better shared understanding of the burgeoning field of machine learning assisted surgical education. To further clarify the need for such a checklist, all published literature utilizing artificial intelligence methodologies in the context of virtual reality surgical education was reviewed and interesting differences in the quality of reporting between medical education and computer science journals were described. As expected, the medical education journals proved stronger in discussion quality and weaker in areas related to study design. The opposite trends were observed in computer science journals. At the time of publication, given the paucity on guidelines on the subject of artificial intelligence applications in virtual reality surgical education, this study represented the first attempt to create an organized reporting and evaluation structure for this type of research.

Chapter 3 is the first study to demonstrate the ability of machine learning algorithms to classify surgical expertise into four groups with high accuracy using fewer than ten performance measures. Of the 50 participants (comprising 250 simulated operations), only 5 were misclassified. Notwithstanding that the simulated task was a neurosurgical procedure performed on the NeuroVR, one of the most advanced surgical simulators available, the novel methodology outlined has broad applicability in any circumstance in which technical performance is measured.

Chapter 4 is the first study to concurrently compare participants' ratings obtained from observation alone with their computationally measured performance and operative complications. This provides interesting insights on the strengths and limitations of visual rating scales in neurosurgery. As residency programs continue to move towards a competency-based curriculum, there is an increasing need for assessment of resident technical skills. Visual rating scales remain convenient tools for generating organized formative assessments. A theoretical limitation of visual rating scales is the risk of rater subjectivity in skills assessment. Furthermore, little information exists on the ability of rating scales to capture subtler aspects of performance, including force applied by instruments during a procedure. Those metrics relating to instrument force and patient safety (brain volume removed, blood loss) were captured by the fewest number of visual rating scale components. The implications of these findings are that important aspects of technical performance, particularly those related to patient safety, are not well captured by simply observing the operation. These findings could only be possible by using the novel methodology described in the study.

Chapter 5 outlines a comprehensive research framework which allows for the study of technical performance and operative outcomes in oncological neurosurgery. This platform relies on a cost-effective alginate-based artificial brain tumor incorporated into an ex vivo calf brain within a controlled operative environment. This represents the first instance where an artificial tumor has been created based on biomechanical properties of human specimens obtained at time of resection. Operative performance can be assessed both via recordings from the surgical microscope and ceiling mounted camera, and by movements generated from instrument-mounted fiducials, while operative “success” can be assessed at time of surgery by presence of residual tumor, via ultraviolet fluorescence or ultrasound. Finally, MRI of residual tumor, as well as the location and volume of grey and white matter resected in 0.003 mm^3 increments present an opportunity for a precise quantification of operative outcome. This framework can offer the clinician, learner, or researcher the ability to carry out operative rehearsal, teaching, or studies involving intraparenchymal brain tumor surgery in a controlled laboratory environment and represents a crucial step in the understanding and training of expertise in neurosurgery.

Contribution of authors

Chapter 2: Artificial Intelligence in Medical Education: Best Practices Using Machine Learning to Assess Surgical Expertise in Virtual Reality Simulation

*__A. Winkler-Schwartz:__ Responsible for content as primary author, Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing, Visualization, Project administration, Funding acquisition. ***V. Bisonnette**: Methodology, Formal analysis, Investigation, Writing - review & editing, Visualization. **N. Mirchi**: Investigation, Writing - review & editing, Visualization. **N. Ponndurai**: Writing - review & editing. **R. Yilmaz**: Writing - review & editing. **N. Ledwos**: Writing - review & editing. **S. Siyar**: Writing - review & editing. **H. Azarnoush**: Writing - review & editing. **B. Karlik**: Writing - review & editing. **R. Del Maestro**: Conceptualization, Methodology, Writing - review & editing, Resources, Supervision, Funding acquisition.

*co-first authorship

Chapter 3: Machine Learning Identification of Surgical and Operative Factors Associated With Surgical Expertise in Virtual Reality Simulation

***A. Winkler-Schwartz:** Responsible for content as primary author, Conceptualization, Formal analysis, Writing - original draft, Writing - review & editing, Visualization, Funding acquisition. ***R. Yilmaz**: Conceptualization, Investigation, Formal Analysis, Writing – original draft, Writing - review & editing. **N. Mirchi**: Writing - review & editing. **V. Bisonnette**: Conceptualization, Formal analysis, Writing - review & editing. **N. Ledwos**: Conceptualization, Writing - review & editing. **S. Siyar**: Conceptualization, Writing - review & editing. **H. Azarnoush**: Conceptualization, Writing - review & editing, Supervision. **B. Karlik**: Formal analysis, Writing - review & editing. **R. Del Maestro**: Conceptualization, Writing - review & editing, Resources, Supervision, Funding acquisition.

*co-first authorship

Chapter 4: A Comparison of Visual Rating Scales and Simulated Virtual Reality Metrics in Neurosurgical Training: A Generalizability Theory Study

A. Winkler-Schwartz: Responsible for content as primary author, Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization, Project administration, Funding acquisition.

I. Marwa: Investigation, Writing - review & editing. **K. Bajunaid:** Writing - review & editing**.** **M. Mullah:** Writing - review & editing**. F. Alotaibi:** Writing - review & editing**. A. Bugdadi:** Writing - review & editing**. R. Sawaya:** Writing - review & editing**. A. Sabbagh:** Writing - review & editing**.**

R. Del Maestro: Conceptualization, Writing - review & editing, Resources, Supervision.

Chapter 5: Creating a Comprehensive Research Platform for Surgical Technique and Operative Outcome in Primary Brain Tumor Neurosurgery

A. Winkler-Schwartz: Responsible for content as primary author, Conceptualization, Methodology, Formal analysis, Investigation, Resources, Writing - original draft, Writing - review & editing, Visualization, Project administration, Funding acquisition. **R. Yilmaz:** Methodology, Validation, Formal analysis, Investigation, Writing - review & editing. **D. Tran:** Methodology, Investigation, Writing - review & editing. **H. Gueziri:** Software, Visualization. **B. Ying:** Methodology, Resources, Writing - review & editing. **M. Tuznik:** Investigation, Data curation, Writing - review & editing. **V. Fonov:** Software, Writing - review & editing. **L. Collins:** Conceptualization, Methodology, Software, Writing - review & editing, Funding acquisition. **D. Rudko:** Methodology, Investigation, Resources, Writing - review & editing, Supervision. **J. Li:** Conceptualization, Methodology, Writing - review & editing. **P. Debergue:** Methodology, Validation, Formal analysis, Investigation, Writing - review & editing. **V. Pazos:** Methodology, Validation, Formal analysis, Investigation, Writing - review & editing. **R. Del Maestro:** Conceptualization, Methodology, Resources, Supervision, Funding acquisition.

List of figures and tables

Figures

Supplementary figures

Tables

Chapter 1 - Introduction

Adverse Events in Surgery

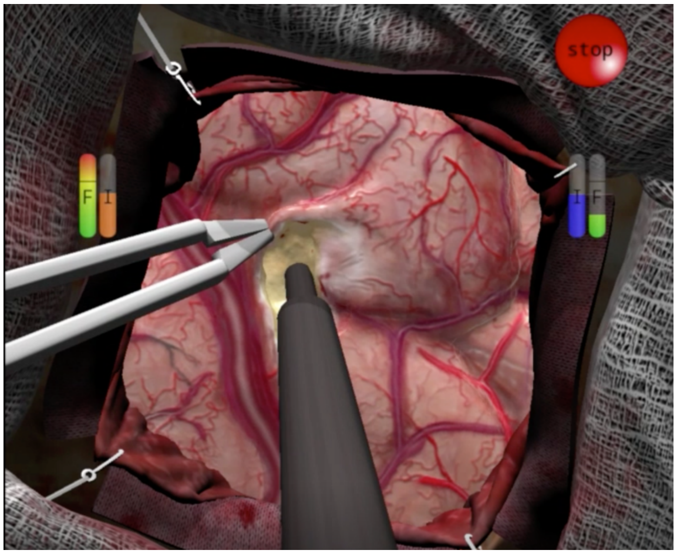
Despite improvements in healthcare technology and delivery systems, medical errors continue to be a cause of significant morbidity and mortality (1). Medical errors cost a staggering 17.1 billion dollars annually in the United States, with roughly one-third directly relating to surgery (2). Complications were once considered an unavoidable consequence of surgery, however increasingly the influence of surgeon technical skill on patient outcomes has come under scrutiny. Lack of technical competence or knowledge was identified as a contributing factor in 41% of 444 surgical malpractice claims from four insurers covering approximately 21,000 physicians in the United States (3). In another prospectively conducted study involving 9,830 patients undergoing surgical procedures over a 12-month period, 63.5% of the errors recorded were due to an "error of technique", classified as a difficulty with the execution of a given procedure, rather than an inappropriate procedure itself (4). Finally, a systematic review on the subject concluded that superior technical skills in surgery had a positive impact on patient outcomes (5).

Operating on the Nervous System

Harvey Cushing outlined the adverse consequences of errors in the neurosurgical operating room at the turn of the previous century (6), and although technology has improved, neurosurgery remains one of the most technically challenging surgical specialties, with high potential for morbidity and mortality. Indeed, there are few surgeries in medicine which capture the attention of the medical community and lay public quite as effectively as operating on the organic seat of consciousness itself. Neurosurgeons are called to operate on the brain due to various pathologies, including brain tumors, and are considered a core competency of any neurosurgical graduate (7, 8). Despite the relatively low incidence of 22 cases per 100,000 people (with approximately half representing malignant primary tumors), the health care costs associated with patients burdened by brain tumors are high (9). Both extent of resection (10) and post-operative functional performance status (11) are strong predictors of overall survival in the setting of malignant primary brain tumors (glioblastoma) and are heavily influenced by intra-operative factors.

The subpial resection technique, while initially pioneered for epilepsy surgery, is recognized as an essential skill for the removal of primary brain tumors (12). To achieve this, the pia mater, an extremely delicate and thin layer of connective tissue adherent to the surface of the cortex, is preserved while the underlying cortex is removed with a suction device. Figure 1 demonstrates this technique as seen in the NeuroVR neurosurgical simulation platform. This accomplishes two goals: (1) the surgeon can remain immediately adjacent to tumoral margins, staying within the relatively avascular surrounding tissue, maximizing brain-tumor interface identification, and (2) minimize damage to surrounding healthy anatomy, such as adjacent eloquent cortex and vascular structures. Given that technical errors constitute roughly one third of mistakes in neurosurgical practice (13-15), and may be an under-reported phenomenon (16), it is crucial that this technique is learnt and applied appropriately.

Figure 1.1 The subpial resection technique as seen in the NeuroVR platform



The Training of Neurosurgical Residents

Currently, learning the subpial resection technique occurs solely in the clinical context, exposing patients to increased risk of morbidity. This apprenticeship based model has been present for over a century, when Dr. Hallstead first introduced the concept of a surgical residency which included progressive graded clinical responsibility (17). Unfortunately, as residents learn a new skill (novice and advanced-beginner Dreyfus stages) they are liable to cause the most patient injury (18). A meta-analysis on the impact of resident involvement on neurosurgical patient outcomes demonstrated a slight increase in complications secondary to resident involvement (OR 1.14; CI 1.03-1.25, p = 0.02) (19). The propensity for errors during training should be theoretically offset by interventions on the part of the intraoperative surgical instructor, though this is not always the case. A lack of supervision was felt to be a contributing factor in 9.5% of neurosurgical resident errors according to surveyed program directors (20). In cases of adequate supervision, a prospectively gathered patient registry demonstrated no significant differences in mortality and complications in intracranial tumor surgery when the surgical operator was a supervised trainee or a staff neurosurgeon (21). However, it should be noted that there were no trainees below their 4th year of training, and over 85% were in their final two years of training. Technical errors were reported as most common among neurosurgical trainees (20), and the most junior neurosurgical residents (years 1-3 in training) demonstrated higher force variability in their surgical instruments during live surgeries (seen as a marker of an inability to regulate force exerted on tissue), compared with staff neurosurgeons (22). Given the delicate nature and functional importance of the nervous tissue, excess force applied to neural structures causes damage, as has been demonstrated in animal models of central nervous system injury (23, 24). In the insurance malpractice study quoted in the first introductory paragraph, trainees, which included surgical interns, residents or fellows, had a contributing factor in 46% of the claims (3).

Partially as a response to growing demand for public accountability in the training of physicians, the training of surgical residents is undergoing another revolution (17). The new framework, termed Competency Based Education in the United States (25) and Competency by Design in Canada (26) emphasizes the demonstration of competence in a given professional activity rather than simply a time-based advancement through residency training (where time-spent was considered as a surrogate for competence). In this mastery-based model, a trainee's progress is to some extent dependent on achieving a given standard or "cut-off". For assessment, both systems take a criteria-referenced approach, whereby an individual is compared to a defined standard as opposed to peer performance (27). However, robust psychometric surgical performance and patient outcome data is a prerequisite to a sanguine discussion about what constitutes 'adequate' surgical performance in a competency-based training system.

Quantification of Technical Skills

As it currently stands, neurosurgical resident's technical abilities are evaluated by their surgical instructor within a clinical context. Because of their convenience, visual rating scales may be utilized to generate organized formative (i.e. feedback) assessments. Numerous rating scales have been studied, however among the most popular has been Objective Structured Assessment of Technical Skills (OSATS), originally developed in 1997 (28). Follow-up studies across numerous surgical specialties have demonstrated the OSATS to be reliable and valid across repeated settings (29).

In the case of the Royal College of Physicians and Surgeons of Canada, the accepted visual rating scale is the O-SCORE (30). However, although the O-SCORE is considered psychometrically sound, in the sense that the rating scale was generalizable across the two surgical specialties involved in its conception (general and orthopedic surgery), this does not necessarily mean that it captures elements important for patient safety and operative success in a neurosurgical operation. Technical skills, while complementary between specialties, may be discrete in some cases, such as the bimanual microsurgical technique with operative microscope which is commonly used in neurosurgery. While the OSATS has been utilized within a neurosurgical clinical context (31), as we will see in chapter 4, visual rating scales may have important limitations in capturing elements of performance crucial to neurosurgery, notably force (32).

Tissue-Based Models in Neurosurgery

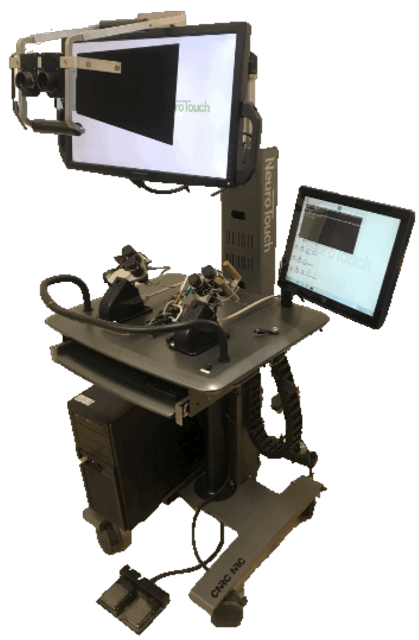
Unlike patient-based surgery, animal or human cadavers permit the evaluation of trainee psychomotor performance in an environment that poses no risk to patients. In a US based survey involving 65% of neurosurgical training programs, over 90% used animal or cadaveric laboratory dissections in their resident curricula (33). Human cadavers have been a cornerstone of medicine for thousands of years, and these permit the study of human anatomy in its most authentic form. In neurosurgery, cadavers have routinely been used for spinal and cranial surgery (34). Through the innovative use of pump systems, traditional limitations such as lack of active bleeding states and cerebrospinal fluid flow can be overcome (35, 36). Furthermore, the unrealistically high stiffness of nervous tissues in formaldehyde treated cadavers can be circumvented with novel embalming methods (37). Despite these benefits, human cadaver procurement is often costly and highly regulated. As such, given their low cost and relative ease of access, ovine, porcine, and bovine models have been proposed as useful alternatives in the evaluation and training of neurosurgical residents (38, 39). Similarities in anatomy between humans, ovine (40, 41), porcine (42-45) , and bovine (46, 47) brain make these animals ideal candidates for cranial neurosurgical operations. Moreover, the speed of harvest of fresh animal cadavers may obviate the need for the use of any preservation technique, preventing biomechanical changes to nervous tissue. Ex vivo models may be used as such, without modification, to allow the trainee to develop a familiarity with tissue handling or to demonstrate a particular surgical approach, or may be modified through the addition of synthetic or organic compounds to re-create a pathological state, such as a ruptured aneurysm or tumor (48). While human cadavers and animal models remain useful adjuncts in neurosurgical resident education, these remain single use and may suffer from the same pitfalls in inferring psychomotor skills through observation as real surgery (49, 50).

Virtual Reality Simulation

Computer-based simulation technology remains an interesting alternative as it offers the promise of integrating realism, pathology and active bleeding states while offering the possibility of the quantitation of psychomotor skills in an environment void of distractions with no risk to patients. The SARS-CoV-2 global pandemic with resulting reduction in the number of elective surgical cases has further increased the necessity for simulation systems to supplement training (51, 52). Broadly speaking, simulation systems can be thought to lie on a continuum between low and high-fidelity based on the degree to which the simulated task or object is captured in a realistic fashion (engineering fidelity) and whether the task captures the specific behaviours required in "real life" (psychological fidelity) (53) (though it should be noted that much controversy on the definition of these terms still exists) (54).

A recent systematic review has identified ten different virtual reality simulation systems available for neurosurgery (55). Among these, the most popular, by volume of publications, is the NeuroTouch. This system was developed in 2009 by a team of 50 experts from the National Research Council of Canada in collaboration with an advisory network of neurosurgeons from 23 Canadian teaching hospitals. In 2016 the distribution rights have been acquired by CAE Healthcare (Montreal, Quebec, Canada), and the system is now sold as the NeuroVR (Figure 1.2).

Figure 1.2 The NeuroVR Neurosurgical Simulator



The NeuroVR consists of a stereoscope through which a three-dimensional image is projected to the user. The user interacts with the simulation object through a bimanual haptic rendering system. Both hands are used together, mimicking the bimanual tasks faced in a live operating environment. A finite element rendering method generates the force felt by the user. The finite element method numerically models the structures in the scenario and applies a physics-based approach permitting mechanical characterization of the simulated structures. Although more computationally demanding (56), the finite element method allows for a more veridical representation of tissue deformation compared to voxel-based systems (57), although the latter has been incorporated successfully into spine-surgery simulation systems (58). As such, the attention to detail and resources utilized in the creation of the NeuroVR make it the most advanced high-fidelity simulator available for neurosurgery (59).

In essence, the system is designed to replicate an open neurosurgical procedure as viewed through an operative microscope. In the case of a brain tumor resection task, the individual holds a simulated ultrasonic aspirator and bipolar electrocautery in the dominant and non-dominant hand, respectively, and are activated by a foot pedal. Although endoscopic scenarios have been developed (endoscopic sinus surgery, and endoscopic third ventriculostomy), their applications are currently limited (60-62). Further limitations of the system are an inability to drastically change the operators' perspective vis-à-vis the patient, though small angle changes are permitted. Additionally, simulating non-microscope based open surgery remains limited as three-dimensional depth perception is lost when not looking through the stereoscope.

Although two recent literature reviews have failed to conclude that high-fidelity simulators confer significant educational benefit over their low fidelity counterparts, it is often the case that high-fidelity systems are computerized, and thus by their nature capture a greater number of performance data points (63, 64). This means that these systems, and the potential educational benefit conferred, will be preferentially favoured by advances in data-sciences.

Quantifying Performance using Metrics

Metrics are best considered as standards of measurement by which the safety, quality, efficiency, and progress of the procedure can be quantitatively assessed. These metrics should also encompass the essential psychomotor and cognitive aspects underlying performance. Raw data from the NeuroVR involving instrument movement, force applied to tissues, change in tissue volume and blood loss can be transformed into performance metrics.

The first major foray into metric development for the NeuroVR system was by Alotaibi et al, enabling the evaluator to quantify an individual's performance on the simulator (65). For a brain tumor resection task, these included: blood loss, brain and tumor removed, distance travelled by the dominant and non-dominant instrument tips, dominant and non-dominant instrument overall and maximum force applied, efficiency index (time spent interacting with tumor over total operative time), path length index (ultrasonic aspirator tip path length in tumor over total ultrasonic aspirator tip path length), coordination index (time both instruments are in use, over time suction instrument in use alone), and bimanual forces ratio (average force of ultrasonic aspirator when used with suction over average force of ultrasonic aspirator used without suction). These metrics permitted the comparison of performance of individuals along the continuum of surgical practice between medical-student, resident, fellow and consultant physician. Although grouping individuals by year of training remains a somewhat poor surrogate for the underlying construct of surgical ability, significant differences in metrics were found between consultant neurosurgeons, senior residents (years 4-6 after medical school graduation) and junior residents (years 1-3 after medical school graduation) on a spherical brain tumor resection task (66). Interestingly, even though considerable variability in performance was observed between individuals, within-person performance variability in the consultant group was very low for identical tumors, and may indicate attainment of the autonomous stage of learning (described as the final stage of motor learning) as proposed by Fitts and Posner (67). Finally, in a brain tumor resection scenario, medical-student applicants to a neurosurgical residency program could be segregated into three performance groupings (68).

A further application of metrics is in technical skills training. The aforementioned metrics can be organized into Tiers, structured as simple outcome metrics (Tier 1: blood loss, quantity of brain and tumor removed), information related to each independent hand (Tier 2: distance travelled by either instrument, forces applied, activation of instruments), or high order computations relating to the use of both hands in concert and overall operative efficiency (Advanced Tier 2: instrument tip separation, bimanual forces ratio, efficiency, coordination and path index). Alternatively, they can be structured into clinically relevant categories such as Safety, Quality, Efficiency, Bimanual and Cognitive Interactive Skills (65). Insights gained from metric performance were incorporated into schema to best understand psychomotor performance in virtual reality (VR) brain tumor resections (69) and early applications involved the aggregation of staff performance metrics to create benchmarks for learners (70). Most importantly, metrics can be used to identify and address specific, individual weaknesses in performance, and underlies the concept of "Technical Abilities Customized Training" (TACT) (68) which may prove particularly useful in cases of struggling trainees. However, a significant limitation of this method was the inability to account for the inter-related nature of the metrics while giving feedback tailored to an individual's performance. It is in such instances where more advanced data analytic techniques, such as artificial intelligence, have the potential to be useful.

Artificial Intelligence

Artificial intelligence (AI) refers to computer-based systems exhibiting human-like intelligent behavior and was first coined in 1956 during a summer research seminar at Dartmouth University. The field emerged following the growth of electronic computerized systems following World War 2 and represented an eclectic mix of ideas and theories from mathematics, biology, psychology, linguistics and philosophy (71). Early applications included work on simple board-games, such as Checkers, as a means for understanding human decision making and intelligence. However, gradually these advances made their way into the medical field. Initial efforts were aimed at supporting clinical decision making, as typified by the MYCIN system.

MYCIN was one of the first clinical decision computer programs developed in the 1970's to guide antibiotic therapy for hospitalized patients (72). Such knowledge-driven AI systems relied on expert opinions and accepted medical literature. However, the increasing processing power coupled with ease of data storage at the end of the 20th century ushered a shift towards data-driven analytic techniques, including machine learning (73). This "bottom-up" approach relies instead on the data itself to generate novel hypotheses and address clinical problems.

Machine learning (ML), a subset of AI, includes supervised and unsupervised algorithms aimed at self-improving data processing. Supervised algorithms utilize labeled data with known outcomes, while unsupervised algorithms operate on unsorted, unlabeled data. Supervised algorithms can be sub-divided into algorithms who sort data using continuous (regression subtype) data or categorical (classification subtype). Examples of Regression algorithms include Linear, Multiple and Polynomial Regression which aim to find a line, plane or curve, respectively, to best fit data points. Examples of classification algorithms include: Support Vector Machines, whereby data is parsed by hyperplanes and Discriminant Analysis works by projecting data points into a new plane where data categorization is maximised. Naïve Bayes assigns data using a probabilistic approach. Nearest Neighbor algorithms group data by assuming that points that cluster close together are related. Decision Tree algorithms are made of decision and leaf nodes, which serve to recursively split a data set until all datapoints are accounted for. Because K-Nearest Neighbor and Decision Tree algorithms can utilize categorical and continuous data, they can be classified as either Regression or Classification algorithms. Importantly, for the supervised algorithm to effectively "learn", it must be "trained" on a dataset where the outcome of interest is known. The algorithm will thus go through numerous iterations of training until the data is parsed in a (programmer-defined) satisfactory manner. One advantage of this method is the trained algorithm integrates datapoints and their relationships in ways which may not be obvious to researchers and clinicians. For example, a trained supervised algorithm may utilize clinical information to categorize patients into pre-defined diagnostic categories, as in the case where researchers utilized over 6,000 genes to categorize 77 patients into cured versus fatal/refractory lymphoma (74).

By contrast, unsupervised algorithms cluster data with no reliance on a priori definitions, and in doing so may group the data in a manner which highlights the existence of previously unknown clinical entities or patterns. Unsupervised clustering algorithms group unlabelled data by relying on a specified number of pre-specified groups, such as K-Means Clustering, or by probabilistic means, as in Hidden Markov Models, to name a few. A recent example of this was the stratification of type 2 diabetes into five discrete clusters in terms of disease progression and risk of complications with K-means clustering using six clinical variables on close to 9,000 patients with a new diagnosis of adult-onset diabetes (75).

Artificial Neural Networks (ANN) represent another subset of Machine Learning and can be supervised or unsupervised (76). ANNs take their inspiration from how information is propagated by neurons in the brain. Fundamentally, ANNs are made up of a minimum three layers: an input, a hidden and an output layer. Each layer contains nodes with their own linear regression model. This node receives a weighted input from the previous layer and when a threshold is reached, the node “turns on” and serves as the input for nodes of a subsequent layer. The algorithm optimizes itself by adjusting the relative weights between the nodes in the system. Although definitions vary, by convention the term Deep Learning is used when there are more than three hidden layers.

The choice of which algorithm to use ultimately depends on its desired usage. The assumption among traditional medical education researchers that these technologies focus too narrowly on solving practical problems as opposed to generating new educational theory itself is unfortunately misguided (77). The use of Supervised Machine Learning algorithms, including ANNs with a single intervening hidden layer, allow for an understanding of the relationship between input and output variables, and may generate novel insights into surgical expertise itself; advancing conceptual frameworks of psychomotor skills acquisition (69). Such systems have seen applications in a VR spine simulation involving an anterior cervical discectomy and fusion (ACDF). This common neurosurgical procedure involves removing an intervertebral disc through an anterior approach and has been adequately simulated within the Sim-Ortho Platform (78). ANNs were able to classify surgeons participating on a VR ACDF scenario into 3 groups with a testing accuracy of at least 80% using between 9 and 16 metrics across domains of safety, motion and efficiency (79, 80). While ANNs with numerous hidden layers present a challenge in interpretation and represents the "black box" problem of Machine Learning (81), they can be useful tools for monitoring performance in real time. Such systems may provide "on demand" live feedback and may even serve in surgical error detection and prevention (82).

The SARS-CoV-2 Pandemic and the need for Intelligent Tutoring Systems

The large datasets generated by computer-based simulation performance has proved a boon for ML applications in surgical education applications and research. For example, performance metric data can be used to train algorithms to sort individuals according to pre-defined categories of expertise. These can be used in formative (i.e. feedback-generating) settings or summative (i.e. high-stakes exam) contexts. The information as to how a particular individual was sorted into a group remains unique for that given performance, and can thus provide a pre-defined, yet individually tailored instructional design for self-improvement on a given surgical task (83).

The SARS-CoV-2 pandemic has significantly impacted resident training (84) and has sparked a renewed interest in the role of simulation technology to address the training gap in neurosurgery (85-87). Hospital response to the crisis has caused a decline by as much as 70% in neurosurgery resident reported operative volume (88, 89). Surgical exposure has been limited by the cancellation of cases outright, by staff surgeons performing the cases themselves to maximize operative time, or by redeployment of residents to COVID-units (90, 91). Although knowledge-based aspects of neurosurgical training, such as didactic teaching and conferences, can be well adapted to online learning platforms (92), surgical skills, which rely on active participation in the operative theatre, cannot.

Importantly, intelligent tutoring systems for Virtual-Reality Simulation systems in neurosurgery have already been developed (83). Such systems serve as a closed loop, feeding performance information back to a trainee with the goal of increasing the chances that they will be categorized in a more skilled group. Feedback can be given concurrently or post-hoc; which may subserve different educational goals. When feedback is given at the conclusion of a task, the participant develops a wholistic understanding of the consequences of their operative decisions, as has been demonstrated in a virtual ACDF scenario (83). This post-hoc feedback may facilitate residents progression towards the autonomous stage of motor learning (93), an emblematic psychomotor feature of staff performance (67). By contrast, continuous systems track and provide real-time feedback as a task is being completed, and have been shown to categorize participant neurosurgical VR performance with a high degree of accuracy (94). Continuous feedback may be best suited for error prevention (93), and may have interesting future application as real-time error prediction tools as well.

A randomized controlled trial has demonstrated that virtual feedback use can be superior to instructor based feedback (95). Reproducing a trained algorithm to tutor a trainee on another machine can be done quickly and at relatively lower cost, in contrast to the years it may take for a single skilled instructor to develop the competencies at a given task. The use of such a system may serve to mitigate the added time constraint faced by staff surgeons during the pandemic, and the time constraints of a busy academic practice (96).

Surgical Skill Transfer in Neurosurgery

Despite the aforementioned advances in data-sciences and the growing number of simulation systems available for neurosurgery (97, 98), the evidence that such systems confer any advantage in the neurosurgical operating theatre can only be inferred by findings in other surgical specialties (99, 100). Though there is a willingness to incorporate simulation technology in the training of neurosurgical trainees (101), few studies relate practice on a simulator directly to operative outcomes in neurosurgery (102).

One study examined the effects simulation rehearsal in cannulation of the ventricle by use of an external ventricular drain (EVD). Though this does not constitute an “open” neurosurgical procedure, such procedures are often conducted by residents at the patient’s bed side, and thus provide a means of associating outcome (catheter placement) with a single user’s intervention. Sixteen neurosurgical residents participated in an interventional, cohort trial to examine the effect of simulated EVD practice on both simulated and real insertion attempts. Twelve individuals self-reported (though cross-validated with chart and imaging review) EVD attempts before and after simulated intervention were recorded. Interestingly, practice was associated with increased likelihood of successful ventricle cannulation on the first attempt (82% vs 94% OR, 4.74;95% CI, 1.10-20.4;P= 0.04), though curiously cannulation of the ipsilateral lateral ventricle (the ideal target for such a procedure) decreased post intervention (103). Another study randomized 25 patients undergoing neurosurgical aneurysm clipping to pre-operative clip placement rehearsal on a simulated system, or standard of care. All surgeries were performed by the same two staff surgeons. Though the authors state limitations in sample size, there was a significant decrease in time per clip used in patients in the intervention group (104). Furthermore, Oliveira et al. conducted a study examining whether practice on a placental aneurysm simulation improved intra-operative errors in neurosurgical trainees. Eight trainees with at least 6 months of microsurgical (but no neuro-vascular) experience were assigned to either undergo 20 hours of practice on placental brain aneurysm model or standard residency training. Video-recordings of their performance on two live aneurysm clippings were evaluated by two blinded raters. Through the sample size was small, the placental practice group demonstrated a significant decrease in intra-operative errors compared to the group undergoing standard training (105).

A high-quality randomized controlled trial (RCT) on whether practice on a simulator improves performance in open oncological neurosurgery would serve to address the lack of level 1 evidence of surgical skill transfer in neurosurgery, while also investigating for the first time the benefit of simulation in “real-life” oncological neurosurgery. In such a trial, trainees would be randomized to either receive standard residency training versus standard training in addition to simulated operative rehearsal on the NeuroVR platform. Ideally, rates of operative complications and extent of tumor resection could then be compared between the groups, before and after randomization.

Solutions and Limitations of a Randomized Controlled Trial

An RCT involving surgical skill transfer from the simulator to open intracranial oncological neurosurgery would suffer from practical constraints related to surgeon, patient, and environmental factors. In addition to posing limitations in the implementation of an RCT, the aforementioned factors have the potential to increase variability in the trial data.

Canadian neurosurgical residents undergo training in one of twelve academic centers, spanning over 5,000 kilometers of territory, with each accepting between one and five residents per year. This poses a problem for any RCT design as sufficient sample size can only be achieved by recruiting from multiple centers. Not all centers have access to a NeuroVR, limiting "on-site" simulated operative rehearsal.

Within the province of Quebec there are four neurosurgical training programs, with one further program in neighboring Ottawa. McGill University, home to the development of the NeuroVR and site of the Neurosurgical Simulation and Artificial Intelligence Learning Center, is also the largest program in the region with approximately 12-15 residents in total, while the others may have between 2 and 5 residents. All are within a three-hour driving commute. Thus, a realistic option is to train and test at McGill University. The type, frequency, timing, and even motivational properties of human-based feedback may influence participant learning and introduce training variability within the VR training intervention group (106), and would be expected to occur even if based on highly accurate metrics. Furthermore, it may be impossible to design a priori coaching protocols which provide a standardized feedback script given the sheer number of possible combinations of performance metrics. For these reasons, it is advantageous to utilize a computer-based coaching system, which can account for metric complexity while standardizing coaching. Artificial intelligence-based feedback systems may be helpful in this regard and are discussed in further detail in Chapter 3.

To test whether the simulation intervention was effective in the clinical environment, the trainees must perform the operations themselves. As it stands, trainees in the first few years of residency benefit the most from simulation training, while being those who would be the most likely to cause patient harm if allowed to operate independently (20). Furthermore, in a teaching hospital environment, a surgical instructor may rightly take over portions of the operation in cases where there is concern for patient harm, thus obfuscating the relative contributions of the trainee in the outcome of the case. Both of these realities mean that junior trainee's contributions may not be adequately captured in an RCT based in a traditional OR environment. Furthermore, patients suffering from intracranial brain tumors are an extremely heterogenous group, both in terms of tumor pathology, location, as well as underlying functional status and comorbidities (21). As such, surgical outcomes may be heavily influenced by patient, rather than operator factors. Finally, variability in a live OR environment may degrade trainee performance via distraction (107). Conducting the surgeries in a standardized environment with a brain tumor model keeping tumoral factors constant would address these limitations. All participants would conduct the surgical procedures in the same operative environment, void of distractions. The use of an animal model obviates patient safety concerns due to inexperienced operators, allows for a standardized surgical procedure between participants, and creates a 1-to-1 association of surgeon to outcome. Outcome measurements can be obtained by observing process in which the surgery is conducted, as well as the surgical end-product. Process can be evaluated by various means, including standardized visual feedback ratings and movement-tracking of instruments. Chapter 4 outlines the development of a visual rating system, while Chapter 5 outlines in greater detail the rationale for, and development of, the animal brain tumor model, including operative outcome measurements.

Overall Goal and Thesis Research Objectives

The overall goals of this Thesis are to lay the groundwork for an RCT for surgical skill transfer in open oncological neurosurgery, specifically:

1. Review the literature for virtual reality and machine learning and develop a framework to guide future research in the field (Chapter 2)
2. Use machine learning to identify factors to accurately classify participants by level of expertise in virtual reality neurosurgery (Chapter 3)
3. Assess the utility of a visual rating scale in capturing surgical performance as compared to simulated metrics (Chapter 4)
4. To develop a "sandbox" operative environment whereby surgical movements and operative outcome can be captured (Chapter 5)

Chapter 2 – Artificial Intelligence in Medical Education: Best Practices Using Machine Learning to Assess Surgical Expertise in Virtual Reality Simulation

Preface

In Chapter 1 we outlined the potential utility in training and assessment of AI technology in high-fidelity simulation. The use of AI in this context bridges the fields of medicine, education, computer science and data-sciences. Important epistemological beliefs underlie each discipline, and while there is overlap when it comes to the acceptance of the scientific method, important knowledge gaps must be crossed to create a shared understanding and language for robust critique and fruitful collaboration. The manuscript presented in this chapter both outlines the existing knowledge gap between computer-science and health education and proposes an evidence-based framework for best practices in the conduct and reporting of AI in VR simulation across these disciplines. This is an important precursor to a study examining the utility of AI in the evaluation and training of neurosurgical residents. The manuscript was published as: **Winkler-Schwartz A**, Bissonnette V, Mirchi N, Ponnudurai N, Yilmaz R, Ledwos N, Siyar S, Azarnoush H, Karlik B, Del Maestro RF. Artificial Intelligence in Medical Education: Best Practices Using Machine Learning to Assess Surgical Expertise in Virtual Reality Simulation. J Surg Educ. 2019 Nov-Dec;76(6):1681-1690. doi: 10.1016/j.jsurg.2019.05.015. Epub 2019 Jun 13. PMID: 31202633.

ABSTRACT

Objective: Virtual reality simulators track all movements and forces of simulated instruments, generating enormous datasets which can be further analyzed with machine learning algorithms. These advancements may increase the understanding, assessment and training of psychomotor performance. Consequently, the application of machine learning techniques to evaluate performance on virtual reality simulators has led to an increase in the volume and complexity of publications which bridge the fields of computer science, medicine, and education. Although all disciplines stand to gain from research in this field, important differences in reporting exist, limiting interdisciplinary communication and knowledge transfer. Thus, our objective was to develop a checklist to provide a general framework when reporting or analyzing studies involving virtual reality surgical simulation and machine learning algorithms. By including a total score as well as clear subsections of the checklist, authors and reviewers can both easily assess the overall quality and specific deficiencies of a manuscript.

Design: The Machine Learning to Assess Surgical Expertise (MLASE) checklist was developed to help computer science, medicine, and education researchers ensure quality when producing and reviewing virtual reality manuscripts involving machine learning to assess surgical expertise.

Setting: This study was carried out at the McGill Neurosurgical Simulation and Artificial Intelligence Learning Centre.

Participants: The authors applied the checklist to 12 articles using machine learning to assess surgical expertise in virtual reality simulation, obtained through a systematic literature review.

Results: Important differences in reporting were found between medical and computer science journals. The medical journals proved stronger in discussion quality and weaker in areas related to study design. The opposite trends were observed in computer science journals.

Conclusions: This checklist will aid in narrowing the knowledge divide between computer science, medicine, and education: helping facilitate the burgeoning field of machine learning assisted surgical education.

INTRODUCTION

The assessment and training of the complex psychomotor skills necessary to perform surgical procedures is critical to safe patient outcomes. As such, virtual reality simulators are being utilized to understand, evaluate, and train these skills.(1) Simulation platforms allow for the quantification of multiple aspects of surgical performance in safe environments. The combination of virtual reality simulators and machine learning has the potential to significantly augment current methods of surgical training.

In computer science, machine learning is a subset of artificial intelligence utilizing algorithms (such as classifiers) which give computers the capacity to "learn" patterns when provided with data. Broadly speaking, classifiers can be either supervised or unsupervised. Supervised classifiers use data which has been identified by the researchers' a priori to generate predictive models to identify novel unlabeled data. In its simplest application in an educational context this implies identifying "expert" and "nonexpert" participant data, thus generating models capable of categorizing individuals into these groups and, ostensibly, assessing expertise. Supervised classifiers lend themselves well to circumstances where groups can be clearly defined. Unsupervised algorithms require no a priori data labeling. Please refer to Table 2.1 for the definitions of relevant terms.

Increasingly, the application of artificial intelligence techniques to evaluate performance on virtual reality simulators has led to an increase in the volume and complexity of publications which bridge the fields of computer science, medicine, and education. Although all disciplines stand to gain from research in this field, important differences in reporting exist, limiting interdisciplinary communication and knowledge transfer. A standardized approach in the reporting of these publications will allow researchers from these fields to form a better shared understanding of the burgeoning field of machine learning assisted surgical education. As such, our goal is to diminish this gap by producing a framework known as the Machine Learning to Assess Surgical Expertise (MLASE) checklist which researchers can utilize when producing and reviewing virtual reality manuscripts involving machine learning to assess surgical expertise. By including a total score as well as clear subsections of the checklist, authors and reviewers can both easily assess the overall quality and specific deficiencies of a manuscript. The framework complements existing guidelines for best practices in reporting experimental design in medical education.(2) To our knowledge, this is the first attempt to create a conceptual structure to ensure quality of virtual reality studies utilizing machine learning to assess surgical skills.

In the manuscript we outline the MLASE checklist and apply it to publications obtained through a systematic literature review on the use of machine learning to assess surgical expertise in virtual reality simulation.

TABLE 2.1. Definitions in the Context of Artificial Intelligence and Machine Learning

Keyword	Definition
Artificial intelligence	Intelligence demonstrated by a machine able to make decisions in a manner similar to human intelligence.
Machine learning	A sub-branch of artificial intelligence where machines process data and learn on their own, without constant human supervision.
Metric	A measurement to quantitate performance.
Feature	Input data that is fed to the artificial intelligence algorithm.
Label	A determinant of the class to which a data point belongs to in the classification process. Usually applied to a dataset in the context of supervised learning. In the context of surgical simulation, an individual's data could be labelled as "expert" or "novice".
Classifier	A machine learning algorithm which sorts data into predefined categories.
Accuracy	
$\frac{\sum STrue\ Positive\ \ \sum True\ Nega}{\sum Total\ population}$	A measure of ability of machine learning to correctly classify new data.
Sensitivity	
$(\sum True\ Positive\ Positive\ \ )$ $/(\sum True\ Positive$ $+ (\sum False\ Negative)$	A measure of how many positive condition predictions are actually true positives.
Specificity	
$(\sum STrue\ Positive\ \ \ )$ $/(\sum True\ Negative$ $+ \sum Total\ population)$	A measure of how many negative condition predictions are true negatives.

METHODS

MLASE Checklist

Upon consultation with interdisciplinary groups of physicians, computer scientists, engineers, and specialists in artificial intelligence, we developed the "Machine Learning to Assess Surgical Expertise" (MLASE) checklist comprised of 20 essential key elements when reporting studies using machine learning algorithms to assess technical skills in virtual reality surgical simulators. The key elements were divided into 4 sections: Study Design, Data Structure, Supervised Machine Learning and Discussion Quality (Table 2.2).

Study Design

This section contains 5 elements: Literature Review, Sample Size, Expertise Definition, Simulator Description and Simulated Tasks Description.

TABLE 2.2. Machine Learning to Assess Surgical Expertise (MLASE) Checklist

Section	**Element**	**Yes**
Study design (5 points)	1. Is relevant literature on the use of artificial intelligence in simulation provided? 2. Is the sample size clearly stated (including number of groups and number of participants in each group)? 3. Is a definition of each group of expertise provided? 4. Is the simulator described? 5. Are the surgical tasks to be performed outlined?	
Data structure (6 points)	6. Is raw data acquisition described? 7. Is feature extraction mentioned? 8. Is an effort made to normalize the data? 9. Is feature selection mentioned? 10. Is the count of features used by the algorithm clearly stated? 11. Are the final selected features clearly described?	
Supervised machine learning (5 points)	12. Is the type of the classifier used mentioned and justified (either by comparing multiple classifiers or citing relevant literature)? 13. Is the mechanism of the classifier explained or is a relevant source provided? 14. Is an effort made to clearly describe the methods used to train and test the algorithm? 15. Is the accuracy of the classifier mentioned? 16. Is the sensitivity and specificity mentioned?	
Discussion quality (4 points)	17. Are efforts made to explain the educational rationale of the features used by the algorithm? 18. Is the educational application of classifiers in the context of surgical simulation stated, specifically its use as a summative or formative assessment tool? 19. Are methodological limitations discussed, including those pertaining to any above-points? 20. Are the future directions discussed?	
Total Score = /20		

The checklist contains 20 elements, separated into 4 sections. A point is awarded for every element completed in the article. The total score is calculated by adding the total number of elements checked.

Literature Review. A relevant literature review on the previous use of similar machine learning algorithms to evaluate skill level should be presented. An effort should be made to situate the current manuscript in the context of previous publications.

Sample Size. The number of groups and participant numbers per group should be clearly stated. In virtual reality surgical education trials, it is often easier to recruit nonexpert (medical student and junior resident) rather than expert (physician consultant) members. As such, algorithms using a dataset obtained from such groups may be biased to incorrectly categorize a new expert participant. Furthermore, as with statistical tests, certain algorithms function poorly with little input data. Thus, the sample size must be appropriate for the algorithm used. Potential pitfall: Having unbalanced groups will skew the algorithms' predictive ability towards the largest group, limiting its future predictive ability. Having a small sample size may be inappropriate for the algorithm used.

Expertise Definition. When utilizing supervised algorithms, judgments concerning what constitutes "expert" performance create algorithms which recapitulate the human assumptions that underlie them. A clear definition of each group is critically important, specifically what constitutes an "expert" vs a "nonexpert". For example, the algorithm accuracy may differ substantially if first year medical students are considered novices, compared to third year residents. Potential pitfall: The outcome of a supervised algorithm classifying process will vary according to the researchers' definition of expertise.

Simulator Description. A description of the simulator hardware and software tools used, type of data recorded, and the experimental environment setting should be elaborated. If available, previous publications outlining the aforementioned items can be cited instead.

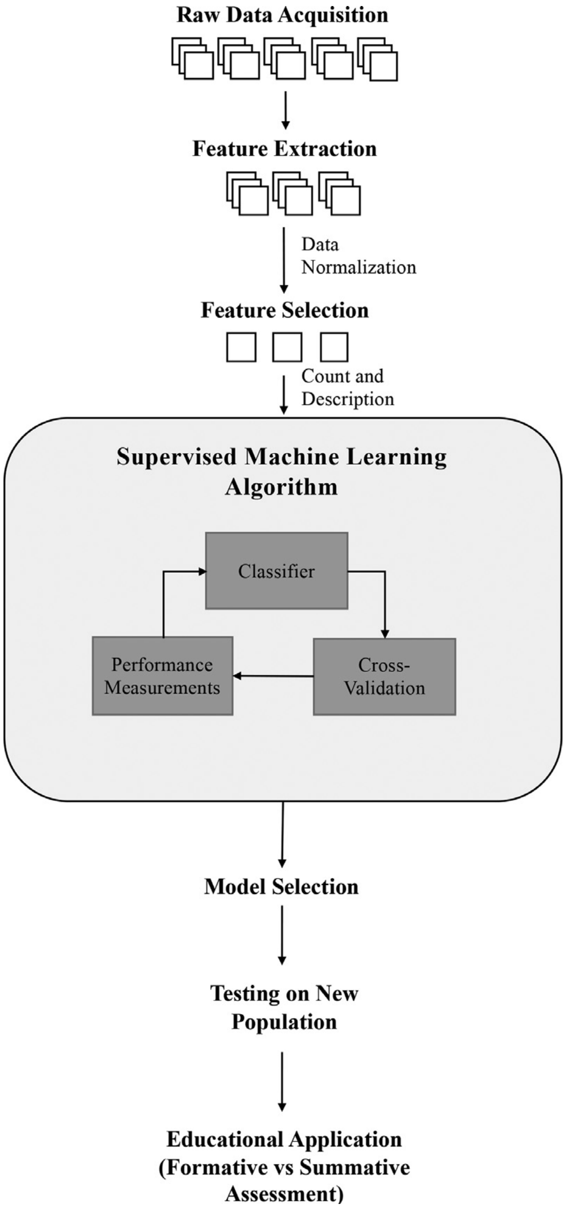
Potential pitfall: Study reproducibility can only be achieved with a clear description of the simulator platform utilized.

Simulated Tasks Description. Due to the variety of simulated scenarios on a given virtual reality system, an adequate description of surgical task should be provided. Potential pitfall: A lack of clear description of the simulated task may impact study reproducibility, applicability, and pedagogical insights. A broad overview of the following 3 sections can be found in Figure 2.1.

Data Structure

The Data Structure section contains 6 elements: Raw Data Acquisition, Feature Extraction, Data Normalization, Feature Selection, Count, and Description of Final Features selected.

FIGURE 2.1. A broad overview of the application of machine learning technology in virtual reality surgical simulation according to the Machine Learning to Assess Surgical Expertise (MLASE).



Raw data acquired from the simulator is transformed into a format which can be inputted into the machine learning algorithm via feature extraction and selection. Following this, an iterative process involving cross-validation is utilized in which the machine learning classifier is optimized. Once a final model is selected it is retrained on the entire study dataset. After this, educational applications of the model can be tested in novel populations.

Raw Data Acquisition. The process of raw data acquisition should be briefly outlined. The most important information to provide is the fundamental structure of the data yielded by the simulator during a simulated task. Notable examples include positional data every second, and applied force vectors in 3 dimensions. Potential pitfall: As in any statistical test, general description of the nature of the data acquired is essential to best understand the functioning of the algorithm and the potential educational benefits

Feature Extraction. Raw data from virtual reality simulators is often very complex, repetitive, and with varying degrees of ‘signal to noise’ ratio. Feature extraction is a method that reduces the dimensionality of a dataset by manipulating raw data, however this can be accomplished by many ways.(3) One can automatically reduce the dimensionality of data by using statistical procedures such as principal component analysis. Alternatively, data can be combined by experienced individuals to generate features in which there may be an a priori hypothesis in distinguishing between experts and novices, such as force applied close to a structure felt to be critical in an operation.(4) Potential pitfall: Failure to provide relevant input will force the algorithm to find patterns in features which may be irrelevant to surgical competency. This may, in addition to limiting the educational use of the model, negatively impact the accuracy and efficiency of the machine learning algorithm.

Data Normalization. Various features generated from feature extraction may be scaled differently, as such, feature normalization should be carried out before providing them as inputs into the algorithm. Potential pitfall: Failure to normalize data will result in diminished accuracy of the classification process.

Feature Selection. Feature selection is a method that highlights the most relevant features and eliminates those that are causing noise. Statistical methods can select only those features showing significant differences between groups (2 sample t test, for example), and are thus most likely to aid the algorithm classifying process. Numerous other feature selection techniques exist, however these are beyond the scope of this article. (5) Potential pitfall: Improper feature selection will negatively impact an algorithm's classification ability.

Count of Final Features Selected. It is of critical importance to include the final count of features used by the algorithm. Including an abundance of features may reduce the algorithms' predictive accuracy by adding noise (i.e., irrelevant information not helpful in the classification process) or by overfitting (a process in which an algorithm is able to detect small differences between groups on a study dataset at the expense of not capturing larger trends which are useful in classifying a novel dataset). Potential pitfall: If the number of final features is too large for a given sample size, the algorithm may appear to be extremely accurate using the study dataset, however its ability to make accurate predictions in a novel dataset may be compromised.

Description of Final Features Selected. We recognize that it may be impractical for authors to describe every final feature in detail if many were included in the final algorithm. However, efforts should be made to apply broad categories, such as features relating to force, movement, tissue removed, to name a few. Potential pitfall: Not including an adequate description of final features may miss interesting insights concerning surgical performance which may serve as the basis for trainee feedback.

Supervised Machine Learning

The Supervised Machine Learning section comprises 5 elements: Type of Classifier and Justification, Mechanism of the Classifier, Training and Testing Set, Accuracy, and Sensitivity/Specificity.

Type of Classifier and Justification. Various supervised machine learning classifiers, such as hidden Markov models, support vector machines, and artificial neural networks have been used to assess surgical expertise level in simulation studies.(4) Authors should not only state the type of classifier employed, but should also provide the rationale for their choice. Such justification can be provided by citing a relevant study using a classifier in a similar context. Potential pitfall: It is important to consider the variability of classifier performance depending on the surgical task. For instance, a classifier can accurately predict the expertise level in a laparoscopic surgery task but perform poorly in a brain tumor resection task. Therefore, an alternative would be to compare the performance of multiple classifiers and select the most accurate for a given task.

Mechanism of the Classifier. The manuscript should include a simple explanation with regards to how the machine learning algorithm works or refer the reader to a source that does so. Potential pitfall: Since artificial intelligence is a novel field in medicine, additional clarification may be necessary, thereby allowing the medical community to gain knowledge on this highly technical topic.

Training and Testing Set. In cases of supervised machine learning, training datasets consist of participant data where groups of expertise have been defined by the researchers. The algorithms' performance in a testing dataset will determine its ability to judge whether novel data will be classified as expert or nonexpert (or various gradations in between). Since this represents a crucial aspect of algorithm development, efforts should be made to clearly describe the process of training and testing. Two common methods are described.(6) If the sample size from each group of expertise is large enough, the sample can be divided in 2 subsamples where one is used for training and the other for testing. However, when the sample size is smaller, many different subsamples of training and testing sets can be used and aver aged to obtain the accuracy. This process is known as cross-validation. Many cross-validation methods exist and are beyond the scope of this publication.(6) Potential pitfall: Failure to provide a clear explanation of the training and testing sets does not allow the reader to understand and evaluate the methodology of the study. Ultimately, cross-validation is a technique used to estimate the accuracy of many models and select the one that is most likely to perform well on a new dataset. However, cross-validation is not an exact measurement of a model's accuracy in real-life application. Therefore, assumptions should not be made about the generalizability of a model that performs well in cross-validation.

Accuracy. Accuracy can be defined as the number of correct predictions made by the machine learning algorithm on all the predictions made (see Table 2.1). Accuracy is a key element because it evaluates the overall ability of the classifier to predict expertise level with a given set of features.

Sensitivity and Specificity. The engineering and medical literature differs based on their reporting of test success. Whereas the engineering community reports in terms of accuracy and equal error rates, these may be less intuitive to medical readers themselves familiar with sensitivity and specificity. For this reason, it is important to discuss sensitivity and specificity when reporting studies in medical journals. Potential pitfall: Authors should mention the percentage of experts and novices misclassified as it may assist readers in understanding whether the authors' conclusions for the use of the algorithm are justified. For example, a highly sensitive but poorly specific algorithm, namely one which misclassifies many nonexperts as expert, would be incompatible with its application as a summative assessment tool. If study design allows, another option is to present a full confusion matrix, which is similar in structure to a 2-by-2 table commonly used in medicine.(7)

Discussion Quality

The Discussion Quality section contains 4 elements: Educational Application of Machine Learning, Educational Rationale of the Selected Features, Methodological Limitations and Future Directions.

Educational Application of Machine Learning. Authors should clearly state the educational aim of their use of machine learning. Classifiers are designed to categorize data, thus lending themselves well as a summative assessment tool. As such, machine learning can be used as a summative assessment tool to evaluate a surgeon's performance. Although more challenging to execute in practice, machine learning can also be involved in formative assessment by facilitating feedback to help surgeons improve their skills. Both types of assessment have different requirements.(8) Potential pitfall: Summative assessment can have a drastic impact on surgeons' success, hence they require extremely high accuracy and reproducibility. On the other hand, formative assessment requires an understanding of the specific features used by the algorithm to help surgeons improve their technical skills.

Educational Rationale of the Selected Features. Authors should clearly describe why the chosen features are important in an educational context. Potential pitfall: Overly abstract features (such as eye movement) may serve well as summative assessments, however if the intended use is for a formative assessment then the chosen features must be easily teachable.

Methodological Limitations. Authors should always address the limitations of their study. Specifically, the shortcomings of the use of machine learning in surgical skill assessment should be outlined.

Future Directions. Future directions should be mentioned. This benefits the medical education community as it provides the reader with a clear understanding of how the field may continue to evolve.

Literature Review

In order to evaluate the current status of articles on the subject using our checklist, we performed a systematic review involving artificial intelligence or machine learning to distinguish experts and novices using virtual surgical simulators in the Medline, Embase, and Web of Science databases. Investigations were included if: (1) its purpose was to assess surgical skill, (2) employing a supervised machine learning algorithm, and (3) on tasks performed on a virtual reality simulator.

Two authors (V.B., N.M.) individually reviewed and scored each article using the MLASE checklist. The article was awarded 1 point for each element of the checklist. If differing article scores were present, an attempt was made by the 2 reviewers to come to a consensus. If none was obtained, then consensus was achieved with the remaining authors. Scores were compiled in a table and analyzed using descriptive statistics. Inter-rater reliability between the reviewers was calculated with Cohen's Kappa.

RESULTS

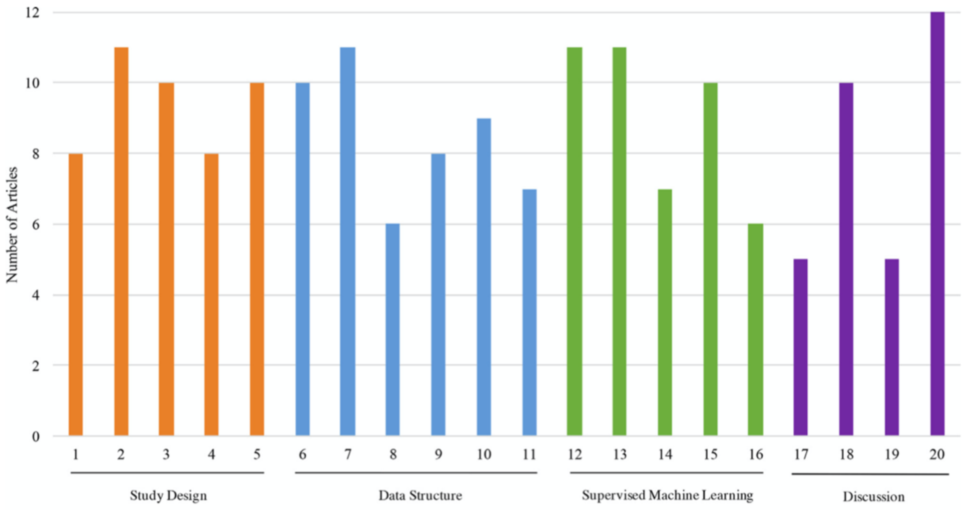
A total of 2642 articles were identified. Following review of abstracts and titles, 84 articles involving simulation and artificial intelligence or machine learning were assessed. A total of 54 articles were excluded as they did not involve virtual reality surgical simulation. Of the remaining 30 articles, 21 were removed as they did not meet all the elements of the inclusion criteria. Three further articles were identified through manual searches of Google Scholar and Cochrane databases for a total of 12 articles.

TABLE 2.3. Articles Assessed on the Use of Artificial Intelligence to Classify Expertise in Virtual Reality Surgical Simulation

Journal Category	Year Published	Classifier	Authors
Medical	2018	Naïve Bayes and support vector machines	Ershad et al.(9)
	2012	Decision tree	Kerwin et al.(10)
	2011	Hidden Markov models	Rhienmora et al.(11)
	2010	Linear discriminant analysis and artificial neural network	Richstone et al.(12)
	2010	Naïve Bayes, hidden Markov models and logistic regression	Sewell et al.(13)
	2005	Fuzzy	Huang et al.(14)
Engineering	2011	Support vector machines and hidden Markov models	Loukas et al.(15)
	2011	Hidden Markov models	Liang et al.(16)
	2011	Support vector machines and decision tree	Jog et al.(17)
	2007	Fuzzy	Hajshirmohammadi et al.(18)
	2006	Hidden Markov models	Megali et al.(19)
	2003	Hidden Markov models	Murphy et al.(20)

These 12 articles (9-20) utilizing machine learning to assess surgical expertise in simulation were reviewed using the MLASE checklist (Table 2.3). Inter-rater reliability between the 2 reviewers was calculated with an observed agreement of 80% (Cohen's Kappa = 0.56). Six of the articles were published in medical and 6 in computer science or engineering journals. The results are summarized in Table 2.4 and Figure 2.2. The global average score for all articles was 73%. This can be further divided into sections where Study Design, Data Structure, Supervised Machine Learning, and Discussion Quality scored 78, 71, 75, and 67%, respectively.

Figure 2.2 Machine Learning to Assess Surgical Expertise Element Number (1 to 20)



The authors applied the Machine Learning to Assess Surgical Expertise (MLASE) checklist to 12 articles on obtained from a systematic review involving artificial intelligence or machine learning to distinguish experts and novices using virtual reality surgical simulators

TABLE 2.4. Results of Assessment of Articles Using the Machine Learning to Assess Surgical Expertise Checklist

	All	**Journal Type** **All Medical**	**Engineering**
MLASE section score	**Score/percentage mean (max — min)**	**MLASE section score (max — min)**	**Score/percentage mean max — min)**
Study design	3.92(5 — 2)/78(100 — 40)	3.83(5 — 2)/77(100 — 40)	4.00(5 — 2)/80(100 — 40)
Data structure	4.25(6 — 2)/71(100 — 33)	4.17(6 — 2)/69(100 — 33)	4.33(6 — 2)/72(100 — 33)
Supervised machine learning	3.75(5 — 2)/75(100 — 40)	3.5(5 — 2)/70(100 40)	4.00(5 — 3)/80(100 — 60)
Discussion	2.67(4 — 1)/67(100 — 25)	3.17(4 — 2)/79(100 — 50)	2.17(4 — 1)/54(100 — 25)
Overall	14.58(18 — 11)/73(90 — 55)	14.67(18 — 11)/73(90 — 55)	14.50(17 — 12)/73(85 — 60)

The 3 lowest scoring elements were: explaining the educational rationale of the selected features (element 17, 5/12, 42% articles) explaining the methodological limitations (element 19, 5/12, 42% articles), normalization of the data (element 9, 6/12, 50% articles), and mentioning the specificity and sensitivity of the algorithm (element 16, 6/12, 50% articles).

We also analyzed articles based on their journal category. Articles from medical and engineering journals both scored 73% overall. Medical articles scored lowest in Data Structure (69%) and Supervised Machine Learning (70%) and highest in Discussion Quality (79%) whereas engineering articles scored lowest in Discussion Quality (54%) and highest in Supervised Machine Learning (80%) and Study Design (80%).

DISCUSSION

Our results indicate that a conceptual framework has the potential of improving the quality of future manuscripts. Though our checklist was tested on articles using machine learning assessing surgical expertise employing virtual reality simulation, we believe the MLASE checklist is also applicable to benchtop simulators, augmented reality, or any other studies which digitize physical surgical performance and use machine learning methods to assess surgical expertise.

Although manuscripts published in medical and computer science journals received, on average, the same overall MLASE total score, important differences in the subsections were noted. We identified the Discussion Quality section of the MLASE checklist as one which will require the most attention from computer scientists wishing to publish in the field of medicine. In medical journals, more detail is required in the Data Structure and Supervised Machine Learning section. The MLASE checklist makes it possible for researchers from these differing communities to ensure their publications reach the widest possible audience. Furthermore, this manuscript may serve as a guide for journal editors and reviewers to ensure that best practices in applying machine learning methodologies in a surgical-simulation context are adhered to. As such, improvements in reporting practices will ultimately facilitate interdisciplinary communication and knowledge transfer, helping to advance this field of research.

Further Suggestions for Future Authors and Reviewers to Enhance the Quality of Manuscripts

Following our article review, we identified new elements which may further enhance the quality of future manuscripts. Firstly, some studies18,19 attempt to increase their sample size by allowing the same surgeon to perform a procedure several times. When such methods are used, it is crucial to explain how each trial is used in the analysis. Often, explanations are vague and it is unclear if different trials from the same surgeon were part of both, the training and testing sample. This would lead to overfitting of the algorithm as performance measures extracted from different trials of the same surgeon are highly correlated. This may hinder an algorithm's ability to accurately classify a new participant. Secondly, if sample size permits, having a third dataset excluded from the initial testing and training to run the chosen model may yield information regarding its generalizability. Thirdly, as an increasingly holistic understanding of expertise continues to be developed (i.e., one which is not based solely on the number of years of practices or on the number of procedures completed), supervised algorithms' predictive abilities will continue to improve. Finally, there are potential educational benefits in describing the individuals that were misclassified by the algorithm, particularly if the same participant is misclassified by different algorithms.

Limitations

The objective of the MLASE checklist is to provide a general framework when reporting or analyzing these studies in the future. However, we acknowledge that the checklist is not extensive and further elements can be added to enhance the quality of a study. The checklist only presents the 20 elements deemed essential to bridge the knowledge gaps in different communities studying the use of artificial intelligence in surgical education. The MLASE checklist was designed and evaluated using only supervised machine learning articles. The MLASE checklist can be applied to studies utilizing unsupervised learning algorithms, however these algorithms do not necessarily always require feature extraction and feature selection.

Future Directions

Artificial intelligence systems may be developed to not only classify participants according to surgical expertise but to coach trainees to a defined surgical standard. These systems will allow for the conduct of studies to further elaborate the proper approach in using this technology in the teaching of psychomotor skills. Regardless of what the future holds, a clear understanding of surgery, artificial intelligence methodologies, and educational best practices will be crucial to the ultimate success of these systems.

CONCLUSIONS

The MLASE checklist was developed to help computer science, medical, and education researchers ensure quality when producing and reviewing virtual reality manuscripts involving the use of machine learning to assess surgical expertise in virtual reality simulation. We believe our checklist will narrow the knowledge divide between computer science, medicine, and education helping facilitate the burgeoning field of machine learning assisted surgical education.

REFERENCES

1. Gallagher AG, Ritter EM, Champion H, Higgins G, Fried MP, Moses G, et al. Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. Annals of Surgery. 2005;241(2):364.
2. Cook DA, Beckman TJ, Bordage G. Quality of reporting of experimental studies in medical education: a systematic review. Medical Education. 2007;41(8):737-45.
3. Ding S, Zhu H, Jia W, Su C. A survey on feature extraction for pattern recognition. Artificial Intelligence Review. 2012;37(3):169-80.
4. Vedula SS, Ishii M, Hager GD. Objective assessment of surgical technical skill and competency in the operating room. Annual Review of Biomedical Engineering. 2017;19:301-25.
5. Chandrashekar G, Sahin F. A survey on feature selection methods. Computers & Electrical Engineering. 2014;40(1):16-28.
6. Kubat M. An introduction to machine learning: Springer; 2017.
7. Watson RA. Use of a machine learning algorithm to classify expertise: Analysis of hand motion patterns during a simulated surgical task. Academic Medicine. 2014;89(8):1163-7.
8. Chauvin SW. Applying educational theory to simulation-based training and assessment in surgery. Surgical Clinics. 2015;95(4):695-715.
9. Ershad M, Rege R, Fey AM. Meaningful assessment of robotic surgical style using the wisdom of crowds. International Journal of Computer Assisted Radiology and Surgery. 2018;13(7):1037-48.
10. Kerwin T, Wiet G, Stredney D, Shen H-W. Automatic scoring of virtual mastoidectomies using expert examples. International Journal of Computer Assisted Radiology and Surgery. 2012;7(1):1-11.
11. Rhienmora P, Haddawy P, Suebnukarn S, Dailey MN. Intelligent dental training simulator with objective skill assessment and feedback. Artificial Intelligence in Medicine. 2011;52(2):115-21.
12. Richstone L, Schwartz MJ, Seideman C, Cadeddu J, Marshall S, Kavoussi LR. Eye metrics as an objective assessment of surgical skill. Annals of Surgery. 2010;252(1):177-82.
13. Sewell C, Morris D, Blevins NH, Dutta S, Agrawal S, Barbagli F, et al. Providing metrics and performance feedback in a surgical simulator. Computer Aided Surgery. 2008;13(2):63-81.
14. Huang J, Payandeh S, Doris P, Hajshirmohammadi I. Fuzzy classification: towards evaluating performance on a surgical simulator. Studies in Health Technology and Informatics. 2005;111:194-200.
15. Loukas C, Georgiou E. Multivariate autoregressive modeling of hand kinematics for laparoscopic skills assessment of surgical trainees. IEEE Transactions on Biomedical Engineering. 2011;58(11):3289-97.
16. Liang H, Shi MY, editors. Surgical skill evaluation model for virtual surgical training. Applied Mechanics and Materials; 2011: Trans Tech Publ.
17. Jog A, Itkowitz B, Liu M, DiMaio S, Hager G, Curet M, et al., editors. Towards integrating task information in skills assessment for dexterous tasks in surgery and simulation. 2011 IEEE International Conference on Robotics and Automation; 2011: IEEE.
18. Hajshirmohammadi I, Payandeh S. Fuzzy set theory for performance evaluation in a surgical simulator. Presence: Teleoperators and Virtual Environments. 2007;16(6):603-22.
19. Megali G, Sinigaglia S, Tonet O, Dario P. Modelling and evaluation of surgical performance using hidden Markov models. IEEE Transactions on Biomedical Engineering. 2006;53(10):1911-9.

20. Murphy TE, Vignes CM, Yuh DD, Okamura AM, editors. Automatic motion recognition and skill evaluation for dynamic tasks. Proc Eurohaptics; 2003: Citeseer.

Chapter 3 - Machine Learning Identification of Surgical and Operative Factors Associated With Surgical Expertise in Virtual Reality Simulation

Preface

Chapter 3 outlines a study evaluating the role of AI algorithms in evaluating neurosurgical performance on the NeuroVR simulation platform, and represents the precursor of an AI powered feedback system which may be used in the VR training experimental arm of the RCT. The study was designed as a proof-of-concept to test whether AI algorithms can distinguish participant level of training by simulated performance alone. At publication, this manuscript was the first to develop a highly accurate AI algorithm which could divide performance into more than three groupings in simulation. In addition, and in keeping with the spirit of the framework in Chapter 2, it utilizes the concept of ‘Explainable AI’ as a means of circumventing the “Black Box” AI problem. By utilizing methodologies to understand how participants were grouped into their respective categories, we retain the ability to understand what constitutes domain-specific expertise itself. Put another way, one can understand the specific psychomotor metrics which underlie staff versus resident neurosurgical operative performance. These insights may serve to further the development of conceptual frameworks of motor learning and mastery. This manuscript was published as:

Winkler-Schwartz A*, Yilmaz R*, Mirchi N, Bissonnette V, Ledwos N, Siyar S, Azarnoush H, Karlik B, Del Maestro R. Machine Learning Identification of Surgical and Operative Factors Associated With Surgical Expertise in Virtual Reality Simulation. JAMA Netw Open. 2019 Aug 2;2(8):e198363. doi: 10.1001/jamanetworkopen.2019.8363. PMID: 31373651. *co-first authors

ABSTRACT

Importance: Despite advances in the assessment of technical skills in surgery, a clear understanding of the composites of technical expertise is lacking. Surgical simulation allows for the quantitation of psychomotor skills, generating data sets that can be analyzed using machine learning algorithms.

Objective: To identify surgical and operative factors selected by a machine learning algorithm to accurately classify participants by level of expertise in a virtual reality surgical procedure.

Design, setting and participants: Fifty participants from a single university were recruited between March 1, 2015, and May 31, 2016, to participate in a case series study at McGill University Neurosurgical Simulation and Artificial Intelligence Learning Centre. Data were collected at a single time point and no follow-up data were collected. Individuals were classified a priori as expert (neurosurgery staff), seniors (neurosurgical fellows and senior residents), juniors (neurosurgical junior residents), and medical students, all of whom participated in 250 simulated tumor resections.

Exposures: All individuals participated in a virtual reality neurosurgical tumor resection scenario. Each scenario was repeated 5 times

Main outcomes and measures: Through an iterative process, performance metrics associated with instrument movement and force, resection of tissues, and bleeding generated from the raw simulator data output were selected by K-nearest neighbor, naive Bayes, discriminant analysis, and support vector machine algorithms to most accurately determine group membership.

Results: A total of 50 individuals (9 women and 41 men; mean [SD] age, 33.6 [9.5] years; 14 neurosurgeons, 4 fellows, 10 senior residents, 10 junior residents, and 12 medical students) participated. Neurosurgeons were in practice between 1 and 25 years, with 9 (64%) involving a predominantly cranial practice. The K-nearest neighbor algorithm had an accuracy of 90% (45 of 50), the naive Bayes algorithm had an accuracy of 84% (42 of 50), the discriminant analysis algorithm had an accuracy of 78% (39 of 50), and the support vector machine algorithm had an accuracy of 76% (38 of 50). The K-nearest neighbor algorithm used 6 performance metrics to classify participants, the naive Bayes algorithm used 9 performance metrics, the discriminant analysis algorithm used 8 performance metrics, and the support vector machine algorithm used 8 performance metrics. Two neurosurgeons, 1 fellow or senior resident, 1 junior resident, and 1 medical student were misclassified.

Conclusions and Relevance: In a virtual reality neurosurgical tumor resection study, a machine learning algorithm successfully classified participants into 4 levels of expertise with 90% accuracy. These findings suggest that algorithms may be capable of classifying surgical expertise with greater granularity and precision than has been previously demonstrated in surgery.

Key Points

Question: Can a machine learning algorithm differentiate participants according to their stage of practice in a complex simulated neurosurgical task?

Findings: In this case series study, 50 individuals (14 neurosurgeons, 4 fellows, 10 senior residents, 10 junior residents, and 12 medical students) participated in 250 simulated tumor resections. An accuracy of 90% was achieved using 6 performance features by a K-nearest

neighbor algorithm and 2 neurosurgeons, 1 fellow or senior resident, 1 junior resident, and 1 medical student were misclassified.

Meaning: The findings suggest that machine learning algorithms may be capable of classifying surgical expertise with greater granularity and precision than has been previously demonstrated in surgery.

INTRODUCTION

Despite technological advances in artificial intelligence and machine learning, delivery of health care is mediated largely by direct interaction between physician and patient. This scenario is particularly true for surgical interventions, which carry substantive patient risks and increased costs to health care systems.(1) As a consequence, the burgeoning field of surgical data science represents efforts to improve interventional health care through increased data collection, quantification, and analysis. (2) Similarly, the use of virtual reality simulators has been explored as a means of providing objective assessment of technical ability in medicine, with the added benefit of retaining realism, pathology, and active bleeding states in a controlled laboratory setting. These systems generate vast data sets that quickly challenge traditional statistical methods. Artificial intelligence and machine learning systems lend themselves well to the analysis of large data sets generated in surgical procedures in 2 important ways: first, by uncovering previously unrecognized patterns, they can expand the understanding of the composites of technical expertise and surgical error, and second, by grouping participants according to technical ability, they offer novel avenues for training and feedback in health care.

We sought to study the operative factors selected by a series of machine learning algorithms to

most accurately classify participants by level of expertise in a virtual reality surgery. Using an advanced high-fidelity neurosurgical simulator allows participants to conduct a complex open neurosurgical brain tumor resection task in a risk-free environment.(3, 4) Our group has extensive experience in virtual reality surgical simulation; several studies have demonstrated that performance measures obtained from simulation can differentiate technical skills both between and within groups of expertise.(5-9) Given the task complexity and the abundance of data generated during the simulated operation, we hypothesized that machine learning algorithms could identify previously unrecognized performance measures, as well as differentiate participants according to their stage of medical practice.

METHODS

Study Participants

All neurosurgeons, neurosurgical fellows, and neurosurgical residents from a single Canadian university were invited between March 1, 2015, and May 31, 2016, to participate in the trial. Medical students rotating on a neurosurgical service or having expressed interest in being contacted for trials were invited. Data were collected at a single time point and no follow-up data were collected.

Participants were classified a priori as expert (neurosurgery staff), seniors (neurosurgery fellows and residents in years 4-6), juniors (neurosurgery residents in years 1-3), and medical students. All participants signed an approved Montreal Neurological Institute and Hospital Research Ethics Board consent form before trial participation. 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.(10) The study received local ethics board approval at the Montreal Neurological Institute and Hospital. This report is structured according to guidelines for best practices in reporting studies on machine learning to assess surgical expertise in virtual reality simulation.(11, 12)

Study Design

The Simulator

The NeuroVR (CAE Healthcare) is a high-fidelity neurosurgical simulator designed to recreate the visual and haptic experience of resecting a human brain tumor through an operative microscope. The platform was developed in 2012 by a team from the National Research Council of Canada in collaboration with an advisory network of surgeons from 23 Canadian and international teaching hospitals. (4) Care was taken to provide the most realistic sensory feedback for the user by incorporating physical properties of human primary brain tumors.(4) As such, the attention to detail and resources used in the creation of the NeuroVR make it one of the most advanced high-fidelity simulators available for neurosurgery.(3)

The Virtual Reality Tumor Resection Task

The trial was carried out at the McGill Neurosurgical Simulation and Artificial Intelligence Learning Centre in a controlled laboratory environment void of distractions. A human intrinsic subpial brain tumor resection task was designed by neurosurgeons with extensive experience in neuro-oncologic and epilepsy neurosurgery. The subpial technique is a challenging bimanual psychomotor skill acquired in neurosurgery and is primarily used in epilepsy and oncologic surgery, where preservation of adjacent eloquent structures is of paramount importance.(13) Participants were given written and verbal instructions that the goal of the scenario was removal of the cortical tumor using the ultrasonic aspirator without damaging adjacent normal brain tissue and vessels. A bipolar instrument could be used to lift and retract the pial membrane to gain access to the tumor and cauterize possible bleeding points. Participants performed the scenario 5 times; however, for the analysis these tasks were averaged and not treated separately. The duration of the resection procedure was limited to 3 minutes.(14) Video 1 is a sample video of the task and Video 2 is a 3-dimensional tumoral reconstruction.

Statistical Analysis

Raw Data Obtained From the Simulator

After each trial, the NeuroVR provides a comma separated value (CSV) file containing, in 20-millisecond increments, the activation, force applied, tip position, and angle of each instrument; the volume of tumor and surrounding healthy tissues removed; blood loss; and whether a given instrument was in contact with the tumor, a blood vessel, or healthy tissue. MATLAB, release 2018a (The MathWorks Inc) was used to process the data into operative performance metrics that can be used by a machine learning algorithm. Interpolation was used to render the data regular and fill occasional missing data points (due to slight fluctuations in computer processing). Figure 3.1 in the Supplement has further examples.

Performance Metric Extraction

To begin, raw data were transformed into performance metrics to be used by the algorithm, with the intention of generating operative measurements that would be easily interpretable by teachers and students of surgery. This process includes transforming instrument movement from the original x, y, and z coordinates into 3-dimensional representations of velocity (first derivative of position), acceleration (first derivative of velocity), and jerk (first derivative of acceleration), as well as the separation between instrument tips. The acceleration and tip distance variables were further refined to reflect the rate of change while the instruments were speeding up and slowing down as well as converging and diverging. The rate of change in volume of tumor and healthy tissue, as well as the rate of change of bleeding, and the number of attempts to stop bleeding were generated. Next, the aforementioned variables were extracted during 3 operative conditions: during the course of the whole scenario, during the tumor resection (ie, only when the ultrasonic aspirator was activated with decreasing tumor volume), and during blood suctioning (ie, when the ultrasonic aspirator was not active and while blood in the operative view was decreasing). Finally, the mean, median, and maximum values of all metrics in all conditions were obtained. Table 3.1 lists all 270 metrics generated. Examples among the total 270 possible metrics generated include mean aspirator force while resecting the tumor, maximum rate of bleeding during the course of the whole scenario, and median tip distance while suctioning blood. Performance measures of the 5 scenarios were averaged together for each participant.

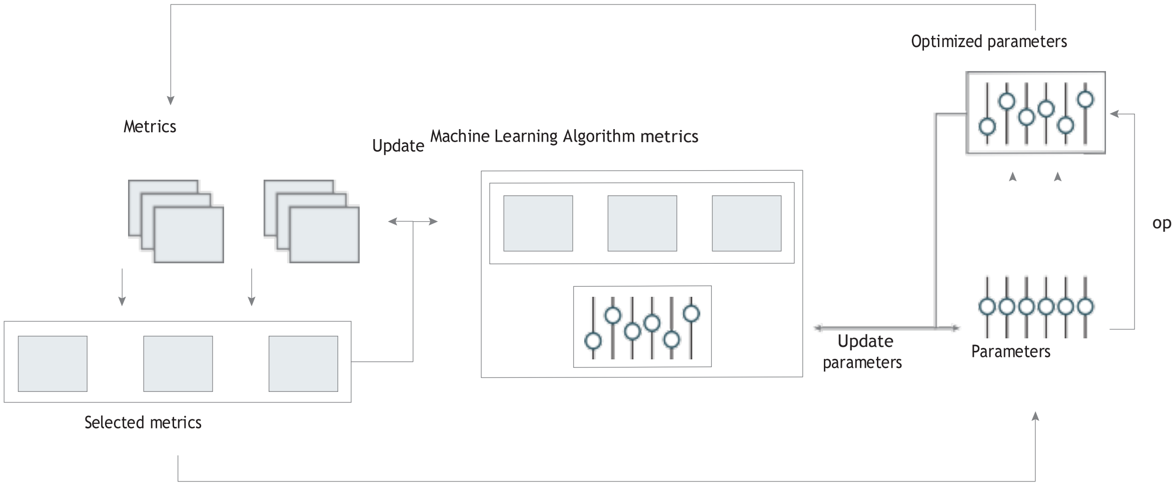
Metric Reduction and Normalization

Metrics failing to demonstrate a significant (P < .05) difference on a 2-sided t test between any 2 groups were excluded. No corrections for multiple tests were done, as the t tests were performed for data-reductive purposes. Subsequent inclusion of the metrics in the algorithm corrects for the probability of type I error at this stage. Metrics were normalized via z score transformation to ensure optimal algorithm functioning.

Iterative Loop

The following steps involve a repetitive process whereby algorithm optimization and final performance metric selection occur. The process is outlined in Figure 3.1. Forward (starting with 1 and increasing in number) metric selection was performed by randomly adding metrics and backward (starting with the maximum and decreasing in number) metric selection was performed by randomly removing metrics. Calculation of accuracy was accomplished by leave-1-out cross-validation. Leave- 1-out validation involves training the machine learning algorithm on the entire participant data set except for 1 individual, whose group membership is then estimated. The process is repeated with different individuals excluded until all participants have been classified. The total number of correctly classified individuals represents the overall accuracy of a given algorithm. No external data set was used to obtain the algorithm accuracy.

Figure 3.1. The Process of Generating a Final Optimized Machine Learning Algorithm With a Set of Selected Metrics



For algorithm optimization, each machine learning algorithm has a defined set of parameters by which it functions, the adjustment of which will modify its overall performance. An analogy for these parameters is the statistical methods that underlie P value adjustments (eg, Bonferroni and Benjamini-Hochberg). MATLAB, release 2018a (MathWorks Inc) was used to modify the intrinsic properties of 4 machine learning algorithms (K-nearest neighbor, naive Bayes, discriminant analysis, and support vector machine).

Algorithms Used

Four classifier algorithms were used: K-nearest neighbor, naive Bayes, discriminant analysis, and support vector machine. Parameter optimizations were carried out using functions included in MATLAB, release 2018a, as well as code written by us.(15-19)

RESULTS

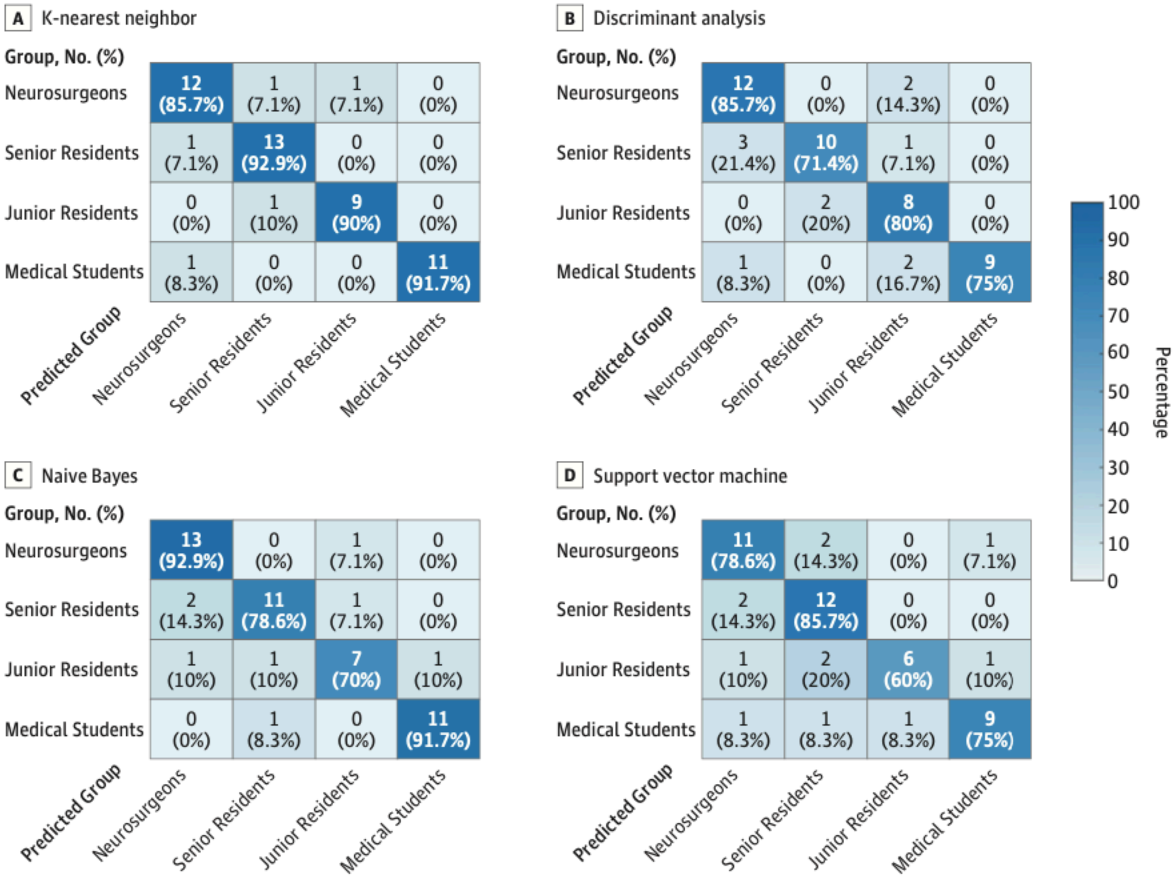
Participant Characteristics

A total of 50 individuals (14 neurosurgeons, 4 fellows, 10 senior residents, 10 junior residents, and 12 medical students) participated in 250 simulated tumor resections. Demographic information is presented in Table 3.2. Consultant neurosurgeon subspecialization covered a wide breadth of practice, with most (9 [64%]) primarily involved in cranial surgery. A total of 7 neurosurgeons (50%), 10 senior residents (69%), 6 junior residents (60%), and 3 medical students (25%) indicated that they had used a surgical simulator previously.

Machine Learning Ability to Classify Participants

The K-nearest neighbor algorithm had an accuracy of 90% (45 of 50), the naive Bayes algorithm had an accuracy of 84% (42 of 50), the discriminant analysis algorithm had an accuracy of 78% (39 of 50), and the support vector machine algorithm had an accuracy of 76% (38 of 50). Figure 3.2 presents details on individual misclassification. Although beyond the scope of the initial hypothesis, in response to misclassifications between medical students and neurosurgeons, the algorithm optimization process was repeated with an emphasis on preventing misclassification between neurosurgeons and medical students, with resulting accuracies ranging between 88% (44 of 50) and 72% (36 of 50). This was accomplished by allowing the algorithm optimization process to stop if no misclassifications between neurosurgeons and medical students occurred, in addition to attaining a desired accuracy. Figure 3.2 in the Supplement has further information regarding the individual misclassifications of these algorithms.

Figure 3.2. Individual Misclassifications by Machine Learning Algorithms



A K-nearest neighbor

Group, No. (%)	Neurosurgeons	Senior Residents	Junior Residents	Medical Students
Neurosurgeons	12 (85.7%)	1 (7.1%)	1 (7.1%)	0 (0%)
Senior Residents	1 (7.1%)	13 (92.9%)	0 (0%)	0 (0%)
Junior Residents	0 (0%)	1 (10%)	9 (90%)	0 (0%)
Medical Students	1 (8.3%)	0 (0%)	0 (0%)	11 (91.7%)

B Discriminant analysis

Group, No. (%)	Neurosurgeons	Senior Residents	Junior Residents	Medical Students
Neurosurgeons	12 (85.7%)	0 (0%)	2 (14.3%)	0 (0%)
Senior Residents	3 (21.4%)	10 (71.4%)	1 (7.1%)	0 (0%)
Junior Residents	0 (0%)	2 (20%)	8 (80%)	0 (0%)
Medical Students	1 (8.3%)	0 (0%)	2 (16.7%)	9 (75%)

C Naive Bayes

Group, No. (%)	Neurosurgeons	Senior Residents	Junior Residents	Medical Students
Neurosurgeons	13 (92.9%)	0 (0%)	1 (7.1%)	0 (0%)
Senior Residents	2 (14.3%)	11 (78.6%)	1 (7.1%)	0 (0%)
Junior Residents	1 (10%)	1 (10%)	7 (70%)	1 (10%)
Medical Students	0 (0%)	1 (8.3%)	0 (0%)	11 (91.7%)

D Support vector machine

Group, No. (%)	Neurosurgeons	Senior Residents	Junior Residents	Medical Students
Neurosurgeons	11 (78.6%)	2 (14.3%)	0 (0%)	1 (7.1%)
Senior Residents	2 (14.3%)	12 (85.7%)	0 (0%)	0 (0%)
Junior Residents	1 (10%)	2 (20%)	6 (60%)	1 (10%)
Medical Students	1 (8.3%)	1 (8.3%)	1 (8.3%)	9 (75%)

Table 3.1. Performance Metrics Generated From Raw Simulator Data

Metric No.	Measurement	Instrument	Performance Measure Associated With Movement, Force, Bleeding, or Tissue	Operative Condition
1	Maximum	Aspirator	Acceleration of instrument	Over whole scenario
2	Maximum	Aspirator	Force of instrument	Over whole scenario
3	Maximum	Aspirator	Change in force of instrument	Over whole scenario
4	Maximum	Aspirator	Jerk of instrument	Over whole scenario
5	Maximum	Aspirator	Rate of slowing down of instrument	Over whole scenario
6	Maximum	Aspirator	Rate of speeding up of instrument	Over whole scenario
7	Maximum	Aspirator	Velocity of instrument	Over whole scenario
8	Maximum	Bipolar	Acceleration of instrument	Over whole scenario
9[a,b]	Maximum	Bipolar	Force of instrument	Over whole scenario
10	Maximum	Bipolar	Change in force of instrument	Over whole scenario
11	Maximum	Bipolar	Jerk of instrument	Over whole scenario
12	Maximum	Bipolar	Rate of slowing down of instrument	Over whole scenario
13	Maximum	Bipolar	Rate of speeding up of instrument	Over whole scenario
14	Maximum	Bipolar	Velocity of instrument	Over whole scenario
15	Maximum	NA	Bleeding speed	Over whole scenario
16	Maximum	NA	Change in bleeding speed	Over whole scenario
17	Maximum	NA	Blood in view	Over whole scenario
18	Maximum	NA	Change in blood in view	Over whole scenario
19	Maximum	NA	Increase in bleeding speed	Over whole scenario
20	Maximum	NA	Converging of instrument tips	Over whole scenario
21	Maximum	NA	Diverging of instrument tips	Over whole scenario
22	Maximum	NA	Increased blood in view	Over whole scenario
23	Maximum	NA	Decreasing bleeding rate	Over whole scenario
24	Maximum	NA	Decrease in blood in view	Over whole scenario
25	Maximum	NA	Tip distance of instruments	Over whole scenario
26	Maximum	NA	Change in tip distance of instruments	Over whole scenario
27	Maximum	NA	Change in volume of brain tissue	Over whole scenario
28	Maximum	NA	Total blood emitted	Over whole scenario
29	Maximum	NA	Change in total blood emitted	Over whole scenario
30	Maximum	NA	Change in volume of tumor	Over whole scenario
31	Mean	Aspirator	Acceleration of instrument	Over whole scenario
32	Mean	Aspirator	Force of instrument	Over whole scenario
33	Mean	Aspirator	Change in force of instrument	Over whole scenario
34[a]	Mean	Aspirator	Jerk of instrument	Over whole scenario
35	Mean	Aspirator	Rate of slowing down of instrument	Over whole scenario
36	Mean	Aspirator	Rate of speeding up of instrument	Over whole scenario
37	Mean	Aspirator	Velocity of instrument	Over whole scenario
38	Mean	Bipolar	Acceleration of instrument	Over whole scenario
39[a]	Mean	Bipolar	Force of instrument	Over whole scenario
40	Mean	Bipolar	Change in force of instrument	Over whole scenario
41	Mean	Bipolar	Jerk of instrument	Over whole scenario
42	Mean	Bipolar	Rate of slowing down of instrument	Over whole scenario
43	Mean	Bipolar	Rate of speeding up of instrument	Over whole scenario
44[a]	Mean	Bipolar	Velocity of instrument	Over whole scenario
45[b]	Mean	NA	Bleeding speed	Over whole scenario
46	Mean	NA	Change in bleeding speed	Over whole scenario
47	Mean	NA	Blood in view	Over whole scenario
48[b]	Mean	NA	Change in blood in view	Over whole scenario
49	Mean	NA	Increase in bleeding speed	Over whole scenario

Table 3.1. Performance Metrics Generated From Raw Simulator Data (continued)

Metric No.	Measurement	Instrument	Performance Measure Associated With Movement, Force, Bleeding, or Tissue	Operative Condition
50	Mean	NA	Converging of instrument tips	Over whole scenario
51[a]	Mean	NA	Diverging of instrument tips	Over whole scenario
52	Mean	NA	Increased blood in view	Over whole scenario
53	Mean	NA	Decreasing bleeding rate	Over whole scenario
54	Mean	NA	Decrease in blood in view	Over whole scenario
55[a,c]	Mean	NA	Tip distance of instruments	Over whole scenario
56	Mean	NA	Change in tip distance of instruments	Over whole scenario
57	Mean	NA	Change in volume of brain tissue	Over whole scenario
58	Mean	NA	Total blood emitted	Over whole scenario
59	Mean	NA	Change in total blood emitted	Over whole scenario
60[a,b,d]	Mean	NA	Change in volume of tumor	Over whole scenario
61	Median	Aspirator	Acceleration of instrument	Over whole scenario
62	Median	Aspirator	Force of instrument	Over whole scenario
63[c,d]	Median	Aspirator	Change in force of instrument	Over whole scenario
64	Median	Aspirator	Jerk of instrument	Over whole scenario
65	Median	Aspirator	Rate of slowing down of instrument	Over whole scenario
66	Median	Aspirator	Rate of speeding up of instrument	Over whole scenario
67[c]	Median	Aspirator	Velocity of instrument	Over whole scenario
68	Median	Bipolar	Acceleration of instrument	Over whole scenario
69[a]	Median	Bipolar	Force of instrument	Over whole scenario
70	Median	Bipolar	Change in force of instrument	Over whole scenario
71	Median	Bipolar	Jerk of instrument	Over whole scenario
72	Median	Bipolar	Rate of slowing down of instrument	Over whole scenario
73	Median	Bipolar	Rate of speeding up of instrument	Over whole scenario
74	Median	Bipolar	Velocity of instrument	Over whole scenario
75	Median	NA	Bleeding speed	Over whole scenario
76	Median	NA	Change in bleeding speed	Over whole scenario
77	Median	NA	Blood in view	Over whole scenario
78	Median	NA	Change in blood in view	Over whole scenario
79	Median	NA	Increase in bleeding speed	Over whole scenario
80	Median	NA	Converging of instrument tips	Over whole scenario
81	Median	NA	Diverging of instrument tips	Over whole scenario
82	Median	NA	Increased blood in view	Over whole scenario
83	Median	NA	Decreasing bleeding rate	Over whole scenario
84	Median	NA	Decrease in blood in view	Over whole scenario
85	Median	NA	Tip distance of instruments	Over whole scenario
86	Median	NA	Change in tip distance of instruments	Over whole scenario
87	Median	NA	Change in volume of brain tissue	Over whole scenario
88	Median	NA	Total blood emitted	Over whole scenario
89	Median	NA	Change in total blood emitted	Over whole scenario
90	Median	NA	Change in volume of tumor	Over whole scenario
91	Maximum	Aspirator	Acceleration of instrument	While removing tumor
92	Maximum	Aspirator	Force of instrument	While removing tumor
93	Maximum	Aspirator	Change in force of instrument	While removing tumor
94	Maximum	Aspirator	Jerk of instrument	While removing tumor
95	Maximum	Aspirator	Rate of slowing down of instrument	While removing tumor
96	Maximum	Aspirator	Rate of speeding up of instrument	While removing tumor
97	Maximum	Aspirator	Velocity of instrument	While removing tumor
98	Maximum	Bipolar	Acceleration of instrument	While removing tumor
99	Maximum	Bipolar	Force of instrument	While removing tumor

Table 3.1. Performance Metrics Generated From Raw Simulator Data (continued)

Metric No.	Measurement	Instrument	Performance Measure Associated With Movement, Force, Bleeding, or Tissue	Operative Condition
100	Maximum	Bipolar	Change in force of instrument	While removing tumor
101	Maximum	Bipolar	Jerk of instrument	While removing tumor
102	Maximum	Bipolar	Rate of slowing down of instrument	While removing tumor
103	Maximum	Bipolar	Rate of speeding up of instrument	While removing tumor
104	Maximum	Bipolar	Velocity of instrument	While removing tumor
105	Maximum	NA	Bleeding speed	While removing tumor
106[b]	Maximum	NA	Change in bleeding speed	While removing tumor
107	Maximum	NA	Blood in view	While removing tumor
108	Maximum	NA	Change in blood in view	While removing tumor
109	Maximum	NA	Increase in bleeding speed	While removing tumor
110	Maximum	NA	Converging of instrument tips	While removing tumor
111	Maximum	NA	Diverging of instrument tips	While removing tumor
112	Maximum	NA	Increased blood in view	While removing tumor
113	Maximum	NA	Decreasing bleeding rate	While removing tumor
114	Maximum	NA	Decrease in blood in view	While removing tumor
115	Maximum	NA	Tip distance of instruments	While removing tumor
116	Maximum	NA	Change in tip distance of instruments	While removing tumor
117	Maximum	NA	Change in volume of brain tissue	While removing tumor
118	Maximum	NA	Total blood emitted	While removing tumor
119	Maximum	NA	Change in total blood emitted	While removing tumor
120	Maximum	NA	Change in volume of tumor	While removing tumor
121	Mean	Aspirator	Acceleration of instrument	While removing tumor
122	Mean	Aspirator	Force of instrument	While removing tumor
123[c]	Mean	Aspirator	Change in force of instrument	While removing tumor
124	Mean	Aspirator	Jerk of instrument	While removing tumor
125[c]	Mean	Aspirator	Rate of slowing down of instrument	While removing tumor
126	Mean	Aspirator	Rate of speeding up of instrument	While removing tumor
127	Mean	Aspirator	Velocity of instrument	While removing tumor
128	Mean	Bipolar	Acceleration of instrument	While removing tumor
129[a]	Mean	Bipolar	Force of instrument	While removing tumor
130	Mean	Bipolar	Change in force of instrument	While removing tumor
131	Mean	Bipolar	Jerk of instrument	While removing tumor
132	Mean	Bipolar	Rate of slowing down of instrument	While removing tumor
133	Mean	Bipolar	Rate of speeding up of instrument	While removing tumor
134	Mean	Bipolar	Velocity of instrument	While removing tumor
135	Mean	NA	Bleeding speed	While removing tumor
136	Mean	NA	Change in bleeding speed	While removing tumor
137	Mean	NA	Blood in view	While removing tumor
138	Mean	NA	Change in blood in view	While removing tumor
139	Mean	NA	Increase in bleeding speed	While removing tumor
140	Mean	NA	Converging of instrument tips	While removing tumor
141	Mean	NA	Diverging of instrument tips	While removing tumor
142	Mean	NA	Increased blood in view	While removing tumor
143	Mean	NA	Decreasing bleeding rate	While removing tumor
144	Mean	NA	Decrease in blood in view	While removing tumor
145	Mean	NA	Tip distance of instruments	While removing tumor
146	Mean	NA	Change in tip distance of instruments	While removing tumor
147	Mean	NA	Change in volume of brain tissue	While removing tumor
148	Mean	NA	Total blood emitted	While removing tumor
149	Mean	NA	Change in total blood emitted	While removing tumor

Table 3.1. Performance Metrics Generated From Raw Simulator Data (continued)

Metric No.	Measurement	Instrument	Performance Measure Associated With Movement, Force, Bleeding, or Tissue	Operative Condition
150	Mean	NA	Change in volume of tumor	While removing tumor
151	Median	Aspirator	Acceleration of instrument	While removing tumor
152	Median	Aspirator	Force of instrument	While removing tumor
153	Median	Aspirator	Change in force of instrument	While removing tumor
154	Median	Aspirator	Jerk of instrument	While removing tumor
155	Median	Aspirator	Rate of slowing down of instrument	While removing tumor
156	Median	Aspirator	Rate of speeding up of instrument	While removing tumor
157[c]	Median	Aspirator	Velocity of instrument	While removing tumor
158	Median	Bipolar	Acceleration of instrument	While removing tumor
159[d]	Median	Bipolar	Force of instrument	While removing tumor
160	Median	Bipolar	Change in force of instrument	While removing tumor
161	Median	Bipolar	Jerk of instrument	While removing tumor
162	Median	Bipolar	Rate of slowing down of instrument	While removing tumor
163	Median	Bipolar	Rate of speeding up of instrument	While removing tumor
164	Median	Bipolar	Velocity of instrument	While removing tumor
165	Median	NA	Bleeding speed	While removing tumor
166	Median	NA	Change in bleeding speed	While removing tumor
167	Median	NA	Blood in view	While removing tumor
168[b,d]	Median	NA	Change in blood in view	While removing tumor
169	Median	NA	Increase in bleeding speed	While removing tumor
170	Median	NA	Converging of instrument tips	While removing tumor
171	Median	NA	Diverging of instrument tips	While removing tumor
172	Median	NA	Increased blood in view	While removing tumor
173	Median	NA	Decreasing bleeding rate	While removing tumor
174	Median	NA	Decrease in blood in view	While removing tumor
175	Median	NA	Tip distance of instruments	While removing tumor
176[b]	Median	NA	Changeintipdistanceofinstruments	While removing tumor
177	Median	NA	Change in volume of brain tissue	While removing tumor
178	Median	NA	Total blood emitted	While removing tumor
179	Median	NA	Change in total blood emitted	While removing tumor
180	Median	NA	Change in volume of tumor	While removing tumor
181	Maximum	Aspirator	Acceleration of instrument	While suctioning blood
182	Maximum	Aspirator	Force of instrument	While suctioning blood
183	Maximum	Aspirator	Change in force of instrument	While suctioning blood
184	Maximum	Aspirator	Jerk of instrument	While suctioning blood
185	Maximum	Aspirator	Rate of slowing down of instrument	While suctioning blood
186	Maximum	Aspirator	Rate of speeding up of instrument	While suctioning blood
187	Maximum	Aspirator	Velocity of instrument	While suctioning blood
188	Maximum	Bipolar	Acceleration of instrument	While suctioning blood
189[d]	Maximum	Bipolar	Force of instrument	While suctioning blood
190	Maximum	Bipolar	Change in force of instrument	While suctioning blood
191	Maximum	Bipolar	Jerk of instrument	While suctioning blood
192	Maximum	Bipolar	Rate of slowing down of instrument	While suctioning blood
193	Maximum	Bipolar	Rate of speeding up of instrument	While suctioning blood
194[d]	Maximum	Bipolar	Velocity of instrument	While suctioning blood
195	Maximum	NA	Bleeding speed	While suctioning blood
196	Maximum	NA	Change in bleeding speed	While suctioning blood
197	Maximum	NA	Blood in view	While suctioning blood
198	Maximum	NA	Change in blood in view	While suctioning blood
199	Maximum	NA	Increase in bleeding speed	While suctioning blood

Table 3.1. Performance Metrics Generated From Raw Simulator Data (continued)

Metric No.	Measurement	Instrument	Performance Measure Associated With Movement, Force, Bleeding, or Tissue	Operative Condition
200	Maximum	NA	Converging of instrument tips	While suctioning blood
201	Maximum	NA	Diverging of instrument tips	While suctioning blood
202	Maximum	NA	Increased blood in view	While suctioning blood
203	Maximum	NA	Decreasing bleeding rate	While suctioning blood
204	Maximum	NA	Decrease in blood in view	While suctioning blood
205	Maximum	NA	Tip distance of instruments	While suctioning blood
206	Maximum	NA	Change in tip distance of instruments	While suctioning blood
207	Maximum	NA	Change in volume of brain tissue	While suctioning blood
208	Maximum	NA	Total blood emitted	While suctioning blood
209	Maximum	NA	Change in total blood emitted	While suctioning blood
210	Maximum	NA	Change in volume of tumor	While suctioning blood
211	Mean	Aspirator	Acceleration of instrument	While suctioning blood
212	Mean	Aspirator	Force of instrument	While suctioning blood
213	Mean	Aspirator	Change in force of instrument	While suctioning blood
214	Mean	Aspirator	Jerk of instrument	While suctioning blood
215	Mean	Aspirator	Rate of slowing down of instrument	While suctioning blood
216	Mean	Aspirator	Rate of speeding up of instrument	While suctioning blood
217	Mean	Aspirator	Velocity of instrument	While suctioning blood
218	Mean	Bipolar	Acceleration of instrument	While suctioning blood
219	Mean	Bipolar	Force of instrument	While suctioning blood
220	Mean	Bipolar	Change in force of instrument	While suctioning blood
221	Mean	Bipolar	Jerk of instrument	While suctioning blood
222	Mean	Bipolar	Rate of slowing down of instrument	While suctioning blood
223	Mean	Bipolar	Rate of speeding up of instrument	While suctioning blood
224	Mean	Bipolar	Velocity of instrument	While suctioning blood
225	Mean	NA	Bleeding speed	While suctioning blood
226	Mean	NA	Change in bleeding speed	While suctioning blood
227	Mean	NA	Blood in view	While suctioning blood
228	Mean	NA	Change in blood in view	While suctioning blood
229	Mean	NA	Increase in bleeding speed	While suctioning blood
230	Mean	NA	Converging of instrument tips	While suctioning blood
231	Mean	NA	Diverging of instrument tips	While suctioning blood
232	Mean	NA	Increased blood in view	While suctioning blood
233	Mean	NA	Decreasing bleeding rate	While suctioning blood
234	Mean	NA	Decrease in blood in view	While suctioning blood
235[d]	Mean	NA	Tip distance of instruments	While suctioning blood
236	Mean	NA	Change in tip distance of instruments	While suctioning blood
237	Mean	NA	Change in volume of brain tissue	While suctioning blood
238	Mean	NA	Total blood emitted	While suctioning blood
239	Mean	NA	Change in total blood emitted	While suctioning blood
240	Mean	NA	Change in volume of tumor	While suctioning blood
241	Median	Aspirator	Acceleration of instrument	While suctioning blood
242	Median	Aspirator	Force of instrument	While suctioning blood
243	Median	Aspirator	Change in force of instrument	While suctioning blood
244	Median	Aspirator	Jerk of instrument	While suctioning blood
245	Median	Aspirator	Rate of slowing down of instrument	While suctioning blood
246	Median	Aspirator	Rate of speeding up of instrument	While suctioning blood
247	Median	Aspirator	Velocity of instrument	While suctioning blood
248	Median	Bipolar	Acceleration of instrument	While suctioning blood

Table 3.1. Performance Metrics Generated From Raw Simulator Data (continued)

Metric No.	Measurement	Instrument	Performance Measure Associated With Movement, Force, Bleeding, or Tissue	Operative Condition
249	Median	Bipolar	Force of instrument	While suctioning blood
250[d]	Median	Bipolar	Change in force of instrument	While suctioning blood
251	Median	Bipolar	Jerk of instrument	While suctioning blood
252	Median	Bipolar	Rate of slowing down of instrument	While suctioning blood
253	Median	Bipolar	Rate of speeding up of instrument	While suctioning blood
254	Median	Bipolar	Velocity of instrument	While suctioning blood
255	Median	NA	Bleeding speed	While suctioning blood
256	Median	NA	Change in bleeding speed	While suctioning blood
257	Median	NA	Blood in view	While suctioning blood
258	Median	NA	Change in blood in view	While suctioning blood
259	Median	NA	Increase in bleeding speed	While suctioning blood
260	Median	NA	Converging of instrument tips	While suctioning blood
261	Median	NA	Diverging of instrument tips	While suctioning blood
262	Median	NA	Increased blood in view	While suctioning blood
263	Median	NA	Decreasing bleeding rate	While suctioning blood
264	Median	NA	Decrease in blood in view	While suctioning blood
265[b]	Median	NA	Tip distance of instruments	While suctioning blood
266	Median	NA	Change in tip distance of instruments	While suctioning blood
267	Median	NA	Change in volume of brain tissue	While suctioning blood
268	Median	NA	Total blood emitted	While suctioning blood
269	Median	NA	Change in total blood emitted	While suctioning blood
270	Median	NA	Change in volume of tumor	While suctioning blood

Abbreviation: NA, not applicable.

a Performance metric selected by naive Bayes algorithm.

b Performance metric selected by support vector machine algorithm.

c Performance metric selected by K-nearest neighbor algorithm.

d Performance metric selected by discriminant analysis algorithm.

Machine Learning Optimized Parameters

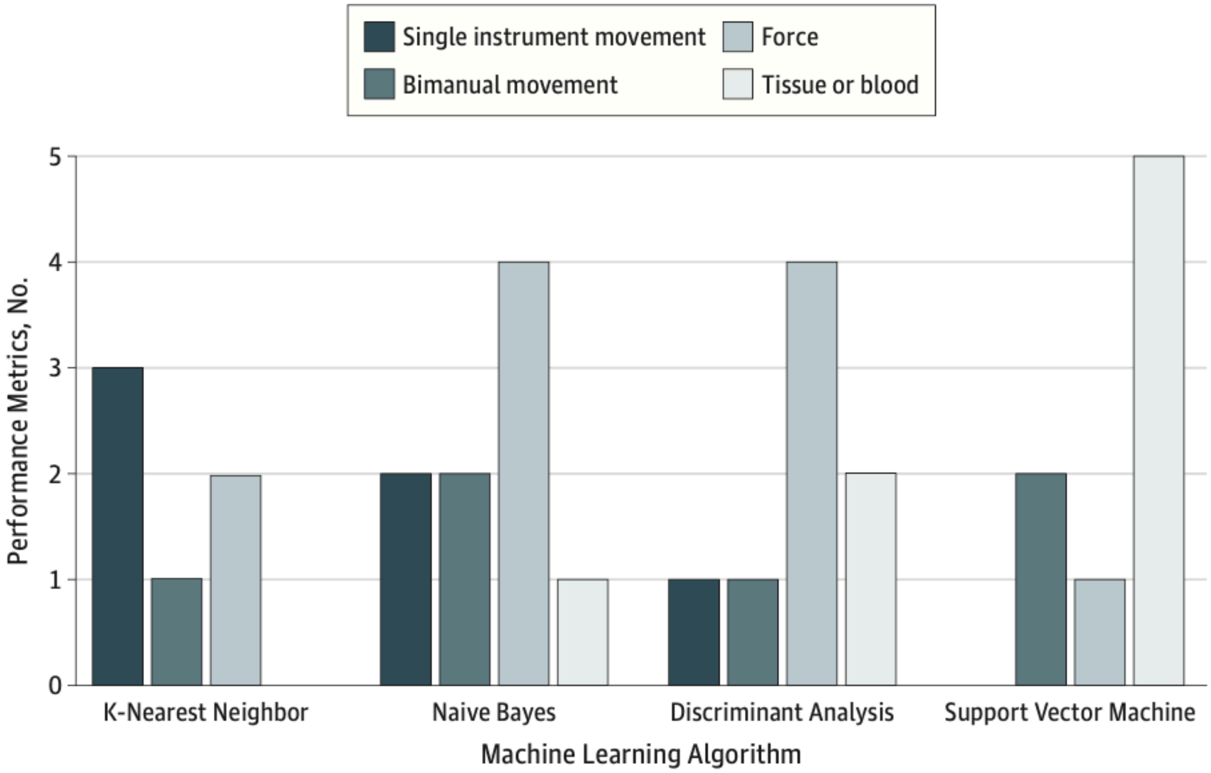
The final K-nearest neighbor algorithm used included 2 neighbors with a cosine distance calculation. Novel data points were classified into the more skilled group in cases when 2 neighbors were from differing groups.

The best-performing naive Bayes algorithm used gaussian (normal) kernel smoothing with a width of 0.31408. The final discriminant analysis algorithm used a δ value of 0.00068926 and a γ value of 0.99808 with a pseudo-linear discriminant type. The final support vector machine algorithm used a gaussian kernel function with the formula $G(xj, xk)= \exp (-||xj - xk||\chi2)$. Box constraint was 0.12958 and kernel scale was 3.1667 using the 1-vs-all coding method (in which 1 group is compared with all others).

Performance Metrics Selected by Machine Learning Algorithm

Of the 270 performance metrics generated from raw data, 122 were selected after reduction and normalization. The K-nearest neighbor algorithm used 6 performance metrics to classify participants (55, 63, 67, 123, 125, and 157), the naive Bayes algorithm used 9 performance metrics (9, 34, 39, 44, 51, 55, 60, 69, and 129), the discriminant analysis algorithm used 8 performance metrics (60, 63, 159, 168, 189, 194, 235, and 250), and the support vector machine algorithm used 8 performance metrics (9, 45, 48, 60, 106, 168, 176, and 265) (Table 3.1). Performance metrics selected by the algorithms spanned the following 4 principal domains: movement associated with a single instrument, both instruments used in concert, force applied by the instruments, and tissue removed or bleeding caused (Figure 3.3).

Figure 3.3. Number of Performance Metrics Selected for By 4 Different Machine Learning Algorithms



Performance metrics are categorized as those involving movements of 1 or both instruments, force applied to the underlying structures and damage to underlying brain, blood loss, and quantity of tumor removed.

Table 3.2. Demographic Information of Participants

Characteristic	Staff Neurosurgeons (n = 14)	Fellows or Senior Residents (n = 14)	Junior Residents (n = 10)	Medical Students (n = 12)
Age, median (range), y	45 (33-59)	33 (29-35)	30 (27-38)	23 (23-26)
Sex, No. (%)				
Male	14 (100)	13 (93)	8 (80)	6 (50)
Female	0	1 (7)	2 (20)	6 (50)
Total No. of years of practice, median (range)	12.5 (1-25)	NA	NA	NA
Neurosurgical subspecialty, No. (%)				
Spine	5 (36)	NA	NA	NA
Oncology and epilepsy	4 (29)	NA	NA	NA
Skull base	2 (14)	NA	NA	NA
Pediatrics	2 (14)	NA	NA	NA
Cerebrovascular	1 (7)	NA	NA	NA

Abbreviation: NA, not applicable.

DISCUSSION

In this prospective study using a high-fidelity virtual reality simulated neurosurgical brain tumor resection procedure, we sought to assess whether machine learning algorithms could select performance measures to classify participants according to their level of neurosurgical training. This study comes at a time of ever-increasing time pressure facing physician-educators to balance their commitment to patients and learners.(20) In parallel, in the United States the search continues for a reliable means of examining Part III of the Maintenance of Certification, namely, the assessment of knowledge, judgment, and skills unique to surgical and procedurally oriented medical specialties.(21, 22) Both require an objective, consistent, transparent, and defendable means of summative and formative assessments of psychomotor ability.

Simulators, while affording learners the opportunity to safely develop technical skills during the particularly dangerous and error-prone early phases of skill acquisition, do not obviate the need for learner feedback, which is often given by skilled instructors.(23) Furthermore, although simulation has been incorporated into the certification process of the American Board of Surgery and the American Board of Anesthesiology, the former relies on human evaluators while the latter is meant only to stimulate self-reflection.(22, 24) Simulation-based technical skills training informed by artificial intelligence feedback systems may offer a solution.

As innovations in artificial intelligence continue, so do the efforts to maintain human understanding of the algorithm classification process. This field has been termed transparent or explainable artificial intelligence.(25) By understanding the performance data used by the algorithm to render its decision, it is possible to design systems to deliver on-demand assessments at the convenience of the examinee and with minimal input from skilled instructors. Such systems may be subject to continuous improvement as increasing participant data are collected and integrated into the algorithm.

We found that the best-performing machine learning algorithm used as few as 6 performance metrics to successfully classify 45 of 50 participants into 1 of 4 groups of expertise. Although we chose to limit the performance measures to those that could be easily interpreted by a user, theoretically higher accuracies may be attained by including more abstruse metrics. Nevertheless, to our knowledge, no previous study using artificial intelligence to evaluate performance has demonstrated the ability to identify 4 groups in open surgery.(26-37)

Limitations

Insofar as technical skills measured on a simulator are reflective of operating skill in the real world, our findings outline a novel approach to understanding technical expertise in surgery. Although 4 different machine learning algorithms were used, there still exists the possibility that all algorithms are overfitted to our data set, limiting their performance when faced with novel data.(38) As such, these algorithms must be tested on an independent data set before making final conclusions about their accuracy. Furthermore, in 3 of 4 algorithms a single medical student was categorized as a neurosurgeon. In response to this misclassification, we sought to limit misclassifications between these 2 groups in the algorithm optimization process as a proof of concept. Although this modification came at a cost of reduced overall accuracy, explicitly preventing misclassifications between certain groups may be desirable in high-stakes certification examinations.

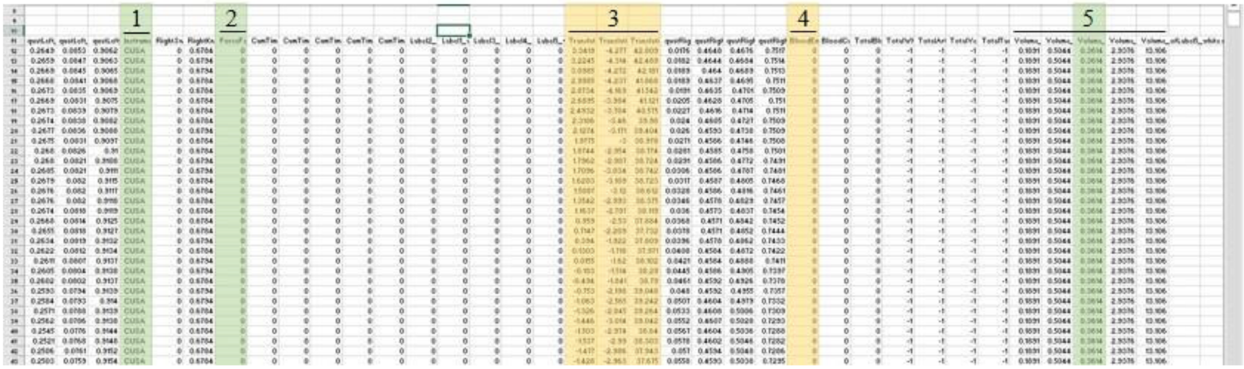
In addition, it is challenging to define populations of surgeons, fellows, and residents with equivalent skill to allow accurate classification. Neurosurgeon skill level was based on being a certified surgeon and resident skill level was based on their educational year, which does not adequately take into account subspecialization or other construct-validated objective assessments of skill sets. A more comprehensive evaluation of participants with an emphasis on demonstrated skills across assessment domains (eg, visual rating scales and training evaluations or assessment of visuospatial abilities) may result in improved algorithm performance.

CONCLUSIONS

Our study demonstrates the ability of machine learning algorithms to classify surgical expertise with greater granularity and precision than has been previously demonstrated. Although the task involved a complex neurosurgical tumor resection task, the protocol outlined can be applied to any digitized platform to assess performance in a setting in which technical skill is paramount.

CHAPTER 3 – SUPPLEMENTARY FIGURES

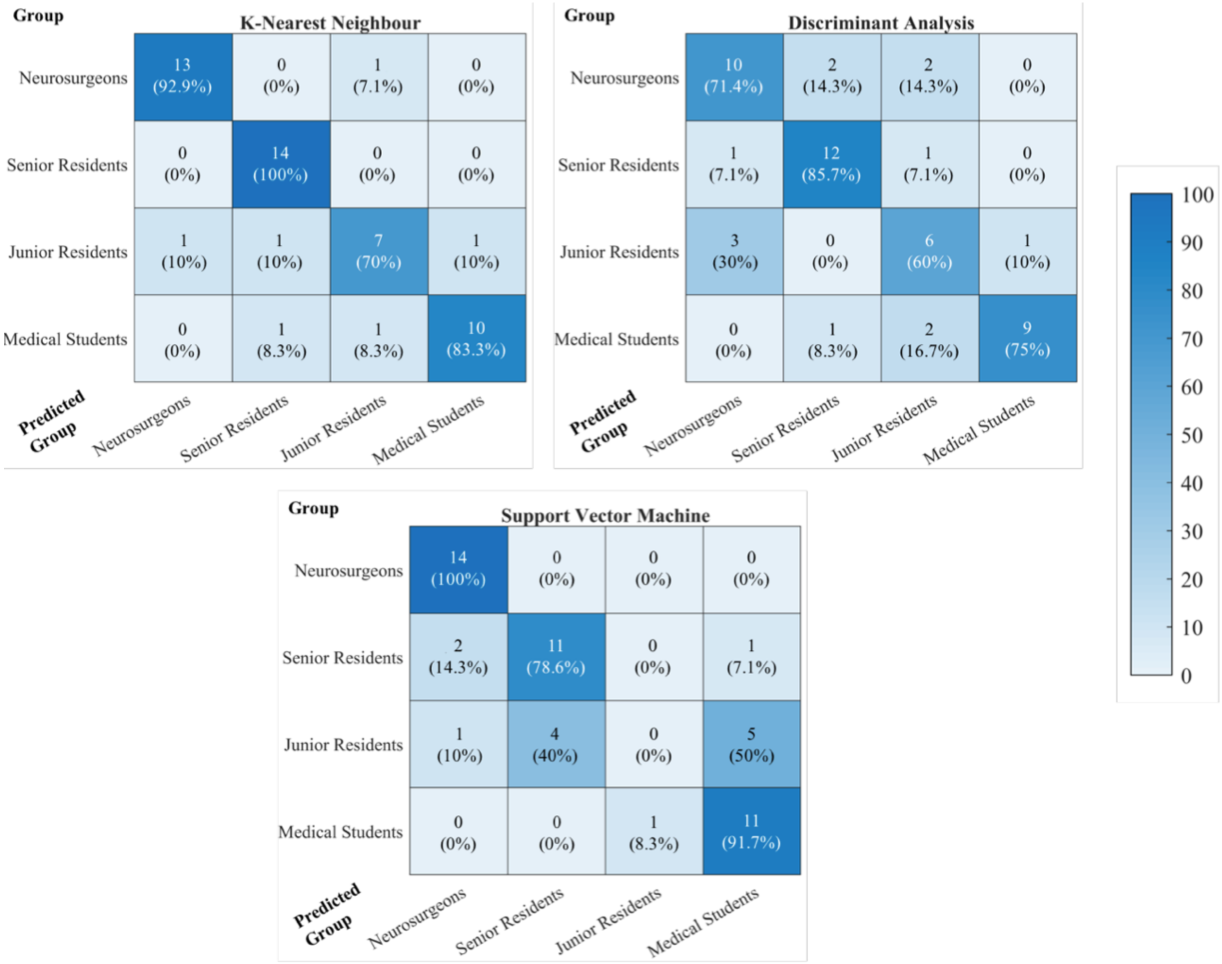
Supplementary Figure 3.1. Screen Capture of the Comma Separated Value File Representing the Output of the Simulator



11	qsvtLeft	qsvtLeft	qsvtLeft	Instrum	RightIn	RightKn	ForceF	CumTim	CumTim	CumTim	CumTim	CumTim	Lobe2_	Lobe1_	Lobe3_	Lobe4_	Lobe5_	Transl	Transl	Transl	qsvRig	qsvRig	qsvRig	qsvRig	BloodE	BloodC	TotalBl	TotalV	TotalAr	TotalVo	TotalTu	Volume_	Volume_	Volume_	Volume_	Volume_ofLobe5_white
12	0.2643	0.0853	0.9062	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	3.3419	-4.277	42.003	0.0176	0.4640	0.4676	0.7517	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
13	0.2659	0.0847	0.9063	CUSA	0	0.6794	0	0	0	0	0	0	0	0	0	0	0	3.2245	-4.314	42.469	0.0182	0.4644	0.4684	0.7514	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
14	0.2669	0.0845	0.9065	CUSA	0	0.6794	0	0	0	0	0	0	0	0	0	0	0	3.0985	-4.272	42.101	0.0183	0.464	0.4683	0.7513	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
15	0.2668	0.0841	0.9068	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	2.9981	-4.237	41.868	0.0183	0.4637	0.4695	0.7511	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
16	0.2673	0.0835	0.9069	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	2.8734	-4.183	41.342	0.0191	0.4635	0.4701	0.7509	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
17	0.2669	0.0831	0.9075	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	2.6835	-3.964	41.121	0.0205	0.4628	0.4705	0.751	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
18	0.2673	0.0833	0.9079	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	2.4932	-3.784	40.515	0.0227	0.4616	0.4714	0.7511	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
19	0.2674	0.0838	0.9082	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	2.3106	-3.45	39.56	0.024	0.4605	0.4721	0.7509	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
20	0.2677	0.0836	0.9088	CUSA	0	0.6794	0	0	0	0	0	0	0	0	0	0	0	2.1274	-3.171	39.404	0.026	0.4593	0.4738	0.7505	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
21	0.2675	0.0831	0.9097	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	1.9775	-3	38.978	0.0271	0.4586	0.4746	0.7506	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
22	0.268	0.0826	0.91	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	1.8744	-2.954	38.774	0.0281	0.4585	0.4758	0.7501	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
23	0.268	0.0821	0.9106	CUSA	0	0.6794	0	0	0	0	0	0	0	0	0	0	0	1.7962	-2.987	38.724	0.0291	0.4586	0.4772	0.7491	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
24	0.2685	0.0821	0.911	CUSA	0	0.6794	0	0	0	0	0	0	0	0	0	0	0	1.7096	-3.034	38.742	0.0306	0.4586	0.4787	0.7481	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
25	0.2679	0.082	0.9115	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	1.6203	-3.169	38.723	0.0317	0.4587	0.4805	0.7468	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
26	0.2676	0.082	0.9117	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	1.5087	-3.12	38.612	0.0328	0.4586	0.4816	0.7461	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
27	0.2676	0.082	0.9118	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	1.3542	-2.933	38.375	0.0346	0.4576	0.4829	0.7457	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
28	0.2674	0.0818	0.9119	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	1.1637	-2.701	38.119	0.036	0.4573	0.4837	0.7454	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
29	0.2668	0.0814	0.9125	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	0.959	-2.53	37.884	0.0368	0.4571	0.4842	0.7452	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
30	0.2651	0.0818	0.9127	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	0.7147	-2.209	37.732	0.0378	0.4571	0.4852	0.7444	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
31	0.2634	0.0813	0.9132	CUSA	0	0.6794	0	0	0	0	0	0	0	0	0	0	0	0.394	-1.922	37.609	0.0396	0.4578	0.4862	0.7433	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
32	0.2622	0.0812	0.9134	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	0.1003	-1.718	37.371	0.0408	0.4584	0.4872	0.7422	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
33	0.2611	0.0807	0.9137	CUSA	0	0.6794	0	0	0	0	0	0	0	0	0	0	0	0.0155	-1.62	38.102	0.0421	0.4584	0.4888	0.7411	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
34	0.2605	0.0804	0.9138	CUSA	0	0.6794	0	0	0	0	0	0	0	0	0	0	0	-0.153	-1.518	38.23	0.0445	0.4586	0.4905	0.7397	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
35	0.2602	0.0802	0.9137	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	-0.434	-1.841	38.79	0.0461	0.4592	0.4926	0.7376	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
36	0.2593	0.0794	0.9139	CUSA	0	0.6794	0	0	0	0	0	0	0	0	0	0	0	-0.753	-2.198	39.048	0.048	0.4592	0.4955	0.7357	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
37	0.2584	0.0793	0.914	CUSA	0	0.6794	0	0	0	0	0	0	0	0	0	0	0	-1.063	-2.565	39.242	0.0507	0.4604	0.4979	0.7332	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
38	0.2571	0.0788	0.9139	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	-1.326	-2.941	39.264	0.0533	0.4606	0.5006	0.7309	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
39	0.2562	0.0786	0.9138	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	-1.446	-3.014	39.042	0.0552	0.4607	0.5028	0.7293	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
40	0.2545	0.0776	0.9144	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	-1.303	-2.974	38.84	0.0561	0.4604	0.5036	0.7288	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
41	0.2521	0.0768	0.9148	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	-1.537	-2.99	38.503	0.0578	0.4602	0.5046	0.7282	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
42	0.2506	0.0761	0.9152	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	-1.417	-2.986	37.943	0.057	0.4594	0.5048	0.7286	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106
43	0.2503	0.0759	0.9154	CUSA	0	0.6784	0	0	0	0	0	0	0	0	0	0	0	-1.428	-2.963	37.675	0.0558	0.4590	0.5038	0.7295	0	0	0	-1	-1	-1	-1	0.1891	0.5044	0.3614	2.9376	13.106

Column 1, 2, 3, 4 and 5 represent instrument type, force applied by instrument (in newton), instrument coordinates in the X, Y and Z planes, total blood emitted and blood in current frame and total tumor volume remaining, respectively. Examples of performance metrics include: ultrasonic aspirator force while resecting tumor (calculated by combining columns 1, 2 and 5) and tip distance while suctioning blood (calculated by combining columns 3 and 4).

Supplementary Figure 3.2. Individual Misclassifications of Machine Learning Algorithms Emphasizing no Misclassifications Between Neurosurgeons and Medical Students



K-Nearest Neighbour

Group / Predicted Group	Neurosurgeons	Senior Residents	Junior Residents	Medical Students
Neurosurgeons	13 (92.9%)	0 (0%)	1 (7.1%)	0 (0%)
Senior Residents	0 (0%)	14 (100%)	0 (0%)	0 (0%)
Junior Residents	1 (10%)	1 (10%)	7 (70%)	1 (10%)
Medical Students	0 (0%)	1 (8.3%)	1 (8.3%)	10 (83.3%)

Discriminant Analysis

Group / Predicted Group	Neurosurgeons	Senior Residents	Junior Residents	Medical Students
Neurosurgeons	10 (71.4%)	2 (14.3%)	2 (14.3%)	0 (0%)
Senior Residents	1 (7.1%)	12 (85.7%)	1 (7.1%)	0 (0%)
Junior Residents	3 (30%)	0 (0%)	6 (60%)	1 (10%)
Medical Students	0 (0%)	1 (8.3%)	2 (16.7%)	9 (75%)

Support Vector Machine

Group / Predicted Group	Neurosurgeons	Senior Residents	Junior Residents	Medical Students
Neurosurgeons	14 (100%)	0 (0%)	0 (0%)	0 (0%)
Senior Residents	2 (14.3%)	11 (78.6%)	0 (0%)	1 (7.1%)
Junior Residents	1 (10%)	4 (40%)	0 (0%)	5 (50%)
Medical Students	0 (0%)	0 (0%)	1 (8.3%)	11 (91.7%)

Overall accuracy of k-nearest neighbor, discriminant analysis and support vector machine algorithms are 88%, 74% and 72%, respectively. Naïve Bayes algorithm is omitted as there were no misclassifications between medical students and neurosurgeons noted initially.

REFERENCES

1. Anderson O, Davis R, Hanna GB, Vincent CA. Surgical adverse events: a systematic review. The American Journal of Surgery. 2013;206(2):253-62.
2. Maier-Hein L, Vedula SS, Speidel S, Navab N, Kikinis R, Park A, et al. Surgical data science for next-generation interventions. Nature Biomedical Engineering. 2017;1(9):691-6.
3. Alaraj A, Tobin MK, Birk DM, Charbel FT. Simulation in neurosurgery and neurosurgical procedures. The Comprehensive Textbook of Healthcare Simulation: Springer; 2013. p. 415-23.
4. 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.
5. Alotaibi FE, AlZhrani GA, Mullah MA, 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.
6. 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.
7. Bajunaid K, Mullah MAS, Winkler-Schwartz A, Alotaibi FE, Fares J, Baggiani M, et al. Impact of acute stress on psychomotor bimanual performance during a simulated tumor resection task. Journal of Neurosurgery. 2017;126(1):71-80.
8. Winkler-Schwartz A, Bajunaid K, Mullah MA, Marwa I, Alotaibi FE, Fares J, et al. Bimanual psychomotor performance in neurosurgical resident applicants assessed using NeuroTouch, a virtual reality simulator. Journal of Surgical Education. 2016;73(6):942-53.
9. Bugdadi A, Sawaya R, Olwi D, Al-Zhrani G, Azarnoush H, Sabbagh AJ, et al. Automaticity of force application during simulated brain tumor resection: testing the Fitts and Posner model. Journal of Surgical Education. 2018;75(1):104-15.
10. Organization WH. World Medical Association Declaration of Helsinki-Ethical principles for medical research involving human subjects. Bulletin of the World Health Organization. 2001;79(4):373-4.
11. Winkler-Schwartz A, Bissonnette V, Mirchi N, Ponnudurai N, Yilmaz R, Ledwos N, et al. Artificial intelligence in medical education: best practices using machine learning to assess surgical expertise in virtual reality simulation. Journal of Surgical Education. 2019;76(6):1681-90.
12. Cheng A, Kessler D, Mackinnon R, Chang T, Nadkarni V, Hunt E. International Network for Simulation-based Pediatric Innovation Research and Education (INSPIRE). Reporting Guidelines Investigators Reporting guidelines for health care simulation research: extensions to the CONSORT and STROBE statements. 2016;11(4):238-48.
13. Hebb AO, Yang T, Silbergeld DL. The sub-pial resection technique for intrinsic tumor surgery. Surgical Neurology International. 2011;2.
14. Bugdadi A, Sawaya R, Bajunaid K, Olwi D, Winkler-Schwartz A, Ledwos N, et al. Is virtual reality surgical performance influenced by force feedback device utilized? Journal of surgical education. 2019;76(1):262-73.
15. MathWorks. fitcsvm: Train support vector machine (SVM) classifier for one-class and binary classification [Available from: https://www.mathworks.com/help/stats/fitcsvm.html.

16. MathWorks. fitcdiscr: Fit discriminant analysis classifier [Available from: https://www.mathworks.com/help/stats/fitcdiscr.html.

17. MathWorks. fitcnb: Train multiclass naive Bayes model [Available from: https://www.mathworks.com/help/stats/fitcnb.html.

18. MathWorks. fitcknn: Fit k-nearest neighbor classifier [Available from: https://www.mathworks.com/help/stats/fitcknn.html.

19. MathWorks. fitcecoc: fit multiclass models for support vector machines or other classifiers [Available from: https://www.mathworks.com/help/stats/fitcecoc.html.

20. Spencer J. Learning and teaching in the clinical environment: ABC of learning and teaching in medicine. BMJ. 2003;326(7389):591-4.

21. Specialties ABoM. Steps toward initial certification and MOC [Available from: https://www.abms.org/boardcertification/.

22. Ross BK, Metzner J. Simulation for maintenance of certification. Surgical Clinics. 2015;95(4):893-905.

23. Barry Issenberg S, Mcgaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Medical Teacher. 2005;27(1):10-28.

24. Epstein RM. Assessment in medical education. New England journal of medicine. 2007;356(4):387-96.

25. Abdul A, Vermeulen J, Wang D, Lim BY, Kankanhalli M, editors. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. Presented at: 2018 CHI Conference on Human Factors in Computing Systems;; 2018; Montreal, QC.

26. Ershad M, Rege R, Fey AM. Meaningful assessment of robotic surgical style using the wisdom of crowds. International journal of computer assisted radiology and surgery. 2018;13(7):1037-48.

27. Kerwin T, Wiet G, Stredney D, Shen H-W. Automatic scoring of virtual mastoidectomies using expert examples. International journal of computer assisted radiology and surgery. 2012;7(1):1-11.

28. Rhienmora P, Haddawy P, Suebnukarn S, Dailey MN. Intelligent dental training simulator with objective skill assessment and feedback. Artificial intelligence in medicine. 2011;52(2):115-21.

29. Richstone L, Schwartz MJ, Seideman C, Cadeddu J, Marshall S, Kavoussi LR. Eye metrics as an objective assessment of surgical skill. Annals of surgery. 2010;252(1):177-82.

30. Loukas C, Georgiou E. Multivariate autoregressive modeling of hand kinematics for laparoscopic skills assessment of surgical trainees. IEEE transactions on biomedical engineering. 2011;58(11):3289-97.

31. Jog A, Itkowitz B, Liu M, DiMaio S, Hager G, Curet M, et al., editors. Towards integrating task information in skills assessment for dexterous tasks in surgery and simulation. 2011 IEEE International Conference on Robotics and Automation; 2011: IEEE.

32. Hajshirmohammadi I, Payandeh S. Fuzzy set theory for performance evaluation in a surgical simulator. Presence: Teleoperators and Virtual Environments. 2007;16(6):603-22.

33. Megali G, Sinigaglia S, Tonet O, Dario P. Modelling and evaluation of surgical performance using hidden Markov models. IEEE Transactions on Biomedical Engineering. 2006;53(10):1911-9.

34. Murphy TE, Vignes CM, Yuh DD, Okamura AM, editors. Automatic motion recognition and skill evaluation for dynamic tasks. Proc Eurohaptics; 2003: Citeseer.

35. Sewell C, Morris D, Blevins NH, Dutta S, Agrawal S, Barbagli F, et al. Providing metrics and performance feedback in a surgical simulator. Computer Aided Surgery. 2008;13(2):63-81.

36. Huang J, Payandeh S, Doris P, Hajshirmohammadi I. Fuzzy classification: towards evaluating performance on a surgical simulator. Studies in health technology and informatics. 2005;111:194-200.

37. Liang H, Shi MY, editors. Surgical skill evaluation model for virtual surgical training. Applied Mechanics and Materials; 2011: Trans Tech Publ.

38. Deo RC. Machine learning in medicine. Circulation. 2015;132(20):1920-30.

Chapter 4 - A Comparison of Visual Rating Scales and Simulated Virtual Reality Metrics in Neurosurgical Training: A Generalizability Theory Study

Preface

Chapter 3 outlined the utility of AI in determining technical expertise in neurosurgery. In this chapter, the development of a visual rating scale for neurosurgical performance was outlined. Such a scale can be used to evaluate surgical performance on the animal model, and serve as an outcome measure for the RCT. Furthermore, because claims of veridical measurement are often made by proponents of either computerized simulation systems or expert raters, it was considered useful to compare these measurements directly to one another using Generalizability Theory, a robust psychometric technique, to best understand their respective strengths and weaknesses in the measurement of surgical performance in neurosurgery. As such, this was the first study at the time of publication to compare simulation data to expert visual ratings of performance. The manuscript was published as:

Winkler-Schwartz A, Marwa I, Bajunaid K, Mullah M, Alotaibi FE, Bugdadi A, Sawaya R, Sabbagh AJ, Del Maestro R. A Comparison of Visual Rating Scales and Simulated Virtual Reality Metrics in Neurosurgical Training: A Generalizability Theory Study. World Neurosurg. 2019 Jul;127:e230-e235. doi: 10.1016/j.wneu.2019.03.059. Epub 2019 Mar 15. PMID: 30880209.

ABSTRACT

Background: Adequate assessment and feedback remains a cornerstone of psychomotor skills acquisition, particularly within neurosurgery where the consequence of adverse operative events is significant. However, a critical appraisal of the reliability of visual rating scales in neurosurgery is lacking. Therefore, we sought to design a study to compare visual rating scales with simulated metrics in a neurosurgical virtual reality task.

Methods: Neurosurgical faculty rated anonymized participant video recordings of the removal of simulated brain tumors using a visual rating scale made up of seven composite elements. Scale reliability was evaluated using generalizability theory, and scale subcomponents were compared with simulated metrics using Pearson correlation analysis.

Results: Four staff neurosurgeons evaluated 16 medical student neurosurgery applicants. Overall scale reliability and internal consistency were 0.73 and 0.90, respectively. Reliability of 0.71 was achieved with two raters. Individual participants, raters, and scale items accounted for 27%, 11%, and 0.6% of the data variability. The hemostasis scale component related to the greatest number of simulated metrics, whereas respect for no-go zones and tissue was correlated with none. Metrics relating to instrument force and patient safety (brain volume removed and blood loss) were captured by the fewest number of rating scale components.

Conclusions: To our knowledge, this is the first study comparing participant's ratings with simulated performance. Given rating scales capture less well instrument force, quantity of brain volume removed, and blood loss, we suggest adopting a hybrid educational approach using visual rating scales in an operative environment, supplemented by simulated sessions to uncover potentially problematic surgical technique.

INTRODUCTION

As residency programs continue to evolve toward a competency-based curriculum, there is an increasing need for assessment of resident technical skills. Adequate assessment and feedback remain a cornerstone of psychomotor skills acquisition, particularly within neurosurgery where the consequence of adverse operative events is great.(1) Visual rating scales remain convenient tools for generating organized formative assessments. Different rating scales for surgery have been developed, including the Objective Structured Assessment of Technical Skills (OSATS), which has been used previously in a neurosurgical context.(2-4) A theoretical limitation of visual rating scales is the risk of rater subjectivity in skills assessment. Furthermore, little information exists on the ability of rating scales to capture subtler aspects of performance, including instrument force applied during a procedure. This last point is particularly important because consistent evidence from the neurosurgical simulation literature suggests that applied force differentiates levels of expertise.(5-10) In addition, a recent study found that excess force applied during live neurosurgical operations is associated with increased intraoperative bleeding.(11)

The objective of the project was to conduct a generalizability study to better understand the use of a visual rating scale of operative performance in neurosurgery and to compare it with computerized metrics generated during a virtual reality neurosurgical operative procedure. We hypothesize that both methods will measure the same underlying construct, namely, surgical performance.

MATERIALS AND METHODS

Subjects

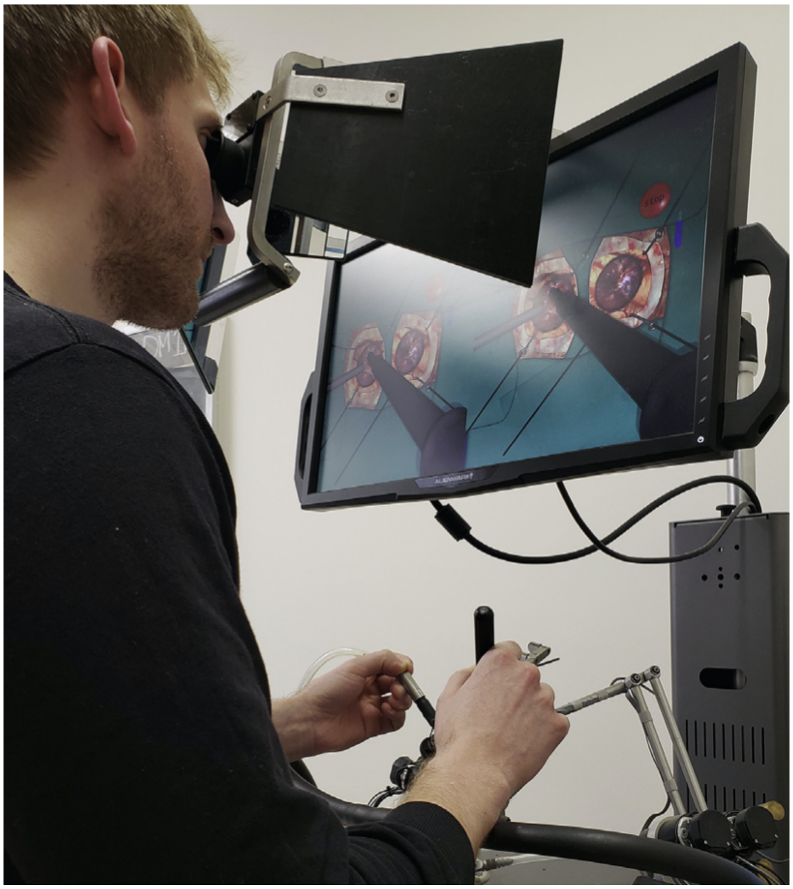
Medical student applicants to a single Canadian neurosurgery program in 2015 were recruited to participate in a trial involving a simulated brain tumor resection task.(8) Sixteen of the 17 applicants participated, comprising over 70% of the national neurosurgical applicant pool for that study year.(11) Data were collected at a single time point within the Neurosurgical Simulation and Artificial Intelligence Learning Centre in a controlled laboratory environment void of distracting noise. No follow-up data were collected. All students signed an approved university ethics board

consent form before trial participation. All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2008.

High Fidelity Simulator and Brain Tumor Resection Task

Participant performance during an established virtual reality brain tumor resection task(5) was assessed using construct-validated metrics(10, 12) for the NeuroVR (CAE Healthcare, Montreal, Quebec, Canada) simulation platform providing real-time visual and haptic feedback. The results of this analysis are available in a previous publication.(8) Participants were instructed to remove sequentially 6 spherical tumors of identical stiffness and glioma-like color while minimizing damage to simulated normal tissue. Tumor stiffness (Young modulus = 9 kPa) was higher than that of the surrounding normal tissue (Young modulus = 3 kPa) to facilitate the ability of participants to differentiate the tumor-normal tissue interface. The task was completed with an ultrasonic aspirator and suction device held in the dominant and nondominant hand, respectively. See Figure 4.1 for example.

Figure 4.1. One of the authors performs a simulated brain tumor resection task on the NeuroVR neurosurgical simulation platform



Performance Video Recording and Rating Scale

Graphical representation of the virtual surgical environment is delivered via computer monitor via the NeuroVR graphic card port. Each eye is presented with an offset view of the operative field, therefore recreating the stereoscopy of a neurosurgical microscope. A high-resolution recording of the virtual reality operation from the perspective of the user was obtained by directing the graphical output in parallel to a DVD recording device. To reduce potential bias, anonymized participant video recordings were shared with four neurosurgical faculty from two institutions and rated using a modified OSATS Global Rating Scale.(13) The scale is made up of seven composite elements (respect for tissue, economy of movement, instrument handling, overall flow, hemostasis, respect for normal brain, and overall score) measured on a 10-point

Likert scale. The scale was produced by the authors after collection of the simulated performance data. Neurosurgical faculty serving as evaluators were not privy to the scale components prior to its use as an evaluation tool in the study.

Statistical Analysis

We report descriptive statistics as counts and percentages for categorical variables. For continuous variables, means and standard deviations are used. Continuous variables include visual rating scale items (respect for tissues, economy of movement, instrument handling, flow, hemostasis, no-go zones, and overall score) and demographic information (number of neurosurgery elective weeks undertaken and number of surgical skin closures performed). Categorical variables include demographic information (previous exposure to simulators). Generalizability theory was used to evaluate scale reliability. G_String with urGENOVA (McMaster Education Research, Innovation & Theory Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada) was used to generate variance components of participants, raters, and scale items, and their interactions. Because all raters evaluated every participant, the study design was considered fully crossed. Simulated metrics from a previous publication(8) were compared with the rating scale subcomponents using Pearson correlation coefficient analysis. For ease of analysis and to further compare the rating scale with simulated metrics, single composite scores for both were created. A rating scale total score was generated by adding individual subcomponents together (range, 7-70). The performance metrics of efficiency index, bimanual forces ratio, suction coordination index, and path length index were combined to create a total metric score (range 0-8). Scores of 0, 1 and 2 were assigned to a given performance metric if an individual achieved below the 25th percentile, between the 25th and 50th percentile, and above the 50th percentile, respectively, compared to their peers. These four metrics were selected because these have shown to best differentiate performance between groups (6) and within groups.(8) Missing data, if present, were replaced with means. All statistical analyses were completed using STATA version 13.0 (StataCorp., LLC, College Station, Texas, USA).

Table 4.1. Descriptive Statistics of Rating Scale Subcomponents

	Rating	Range	
Scale Item	**Mean ± Standard Deviation**	**Minimum**	**Maximum**
Respect for tissues	4.28 ± 1.80	1	8
Economy of movement	4.12 ± 1.82	1	8
Instrument handling	3.90 ± 1.71	1	8
Flow	4.10 ± 1.78	2	9
Hemostasis	4.30 ± 2.22	1	9
No-go zones	4.75 ± 2.10	1	9
Overall	3.81 ± 1.59	1	9

Four staff neurosurgeons used the rating scale in 64 observations in 16 medical student applicants to neurosurgery residency at McGill University.

RESULTS

Four staff neurosurgeons evaluated 16 medical students for a total of 64 observations. Table 4.1 includes a descriptive analysis, demonstrating use of the full range of the Likert scale. Demographic information is available in a previous publication(8) and can be summarized as follows: 7 out of 16 participants (43%) previously used a simulator, the mean number of neurosurgery elective weeks was 11.2 +/- 4.6 (range, 4-22), and the mean number of surgical skin closures was 10.9 +/- 6.3 (range, 1-25). Five observations across 3 participants were missing and were replaced with means. Additionally, one reviewer failed to complete the overall scale subcomponent for all participants, representing 16 missing observations. As a result, the overall scale subcomponent was excluded from the generation of the composite rating scale score.

Generalizability theory analysis demonstrated relative g coefficients corresponding to an overall reliability of 0.73 and internal consistency of 0.90. A decision study was conducted, demonstrating that scale reliability of 0.71 can be achieved with only 2 raters (relative error coefficient, and keeping item facet fixed). A single rater failed to complete the overall scale subcomponent of the visual rating scale; therefore, their scores were replaced with those for the group mean.

Table 4.2. Sources of Variance in Scores

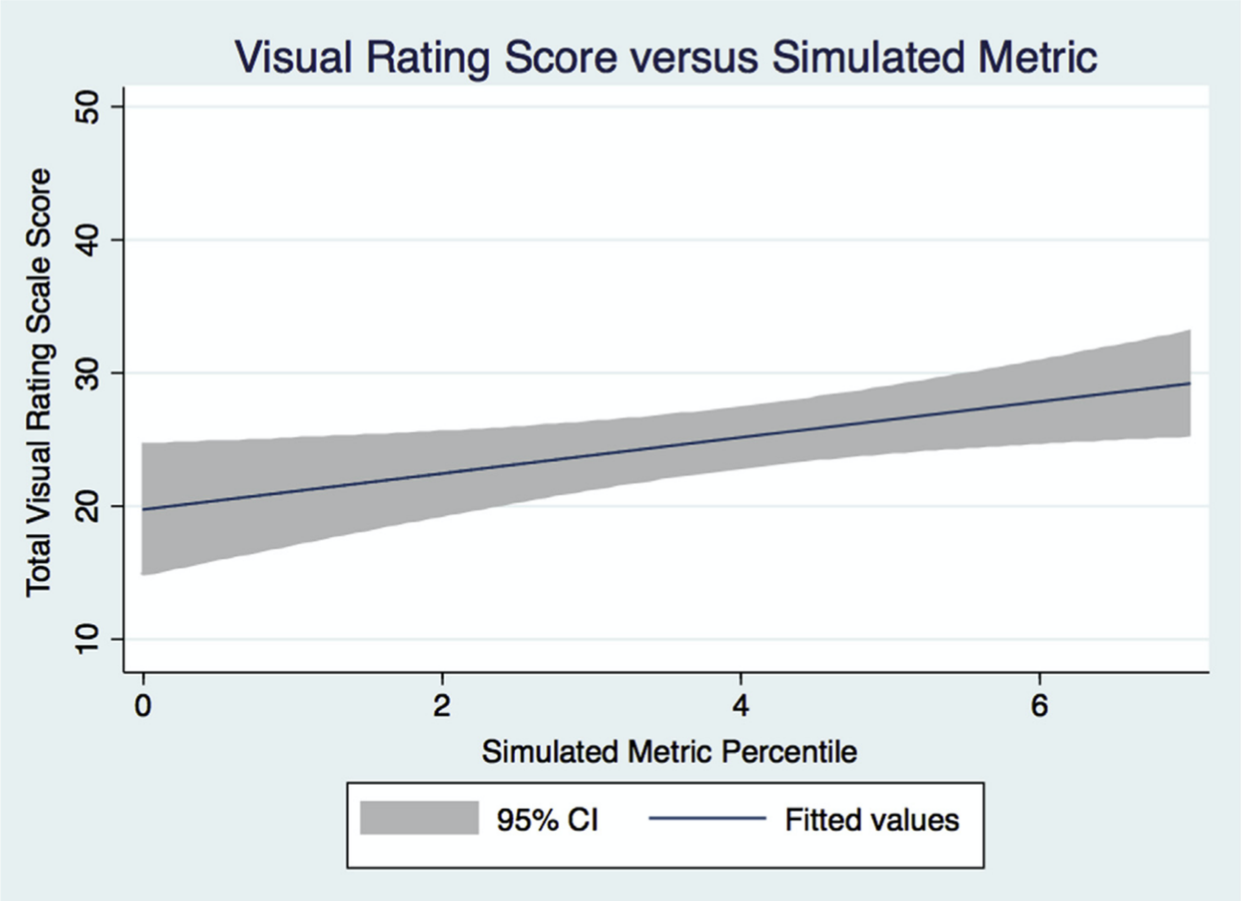
Effect	***df***	**T Score**	**Sum of Squares**	**Mean Squares**	**Variance Component**	**Variance (%)**
Participants	15	530.80	530.81	35.39	0.98	27.0
Raters	3	156.21	156.21	52.07	0.39	10.9
Items	6	36.453	36.45	6.08	0.02	0.6
Interactions						
Participants × raters	45	962.96	275.95	6.13	0.75	20.6
Participants × items	90	817.47	250.21	2.78	0.47	13.0
Raters × items	18	241.94	49.280	2.74	0.11	3.2
Participants × raters × items	270	1541.06	242.15	0.90	0.90	24.7

Table 4.2 displays the variance components associated with the score. Greatest and least sources of data variance were explained by individual participants and individual rating scale items, respectively.

Table 4.3 represents comparison of the visual rating scale subcomponents with known individual simulated metrics using Pearson correlation analysis. The low variance in the scale items (0.6%) and significant interitem correlation among the scale subcomponents justified the creation of a summative total score for the scale. The scale components with no significant relation to any metrics were respect for no-go zones and respect for tissue (however it should be noted that, even though not significant, they are both negatively correlated with brain volume removed).

The scale component which is significantly correlated with the greatest number of simulated metrics is hemostasis, in which positive correlation is seen for efficiency index, suction coordination index, path length index, tumor percentage removed, brain volume removed, sum of forces in the dominant hand, and maximum force dominant hand, and a significant negative correlation with blood loss. The other scale subcomponents have statistically significant correlation with efficiency index, coordination index, path length index, and sum of forces in dominant hand. The only two visual rating scale components that have a significant negative correlation with the bimanual force ratio are instrument handling and overall score. Those metrics relating to instrument force (sum of forces in nondominant hand, maximum force in dominant hand, and maximum force in nondominant hand) and patient safety (brain volume removed and blood loss) were captured by the fewest number of scale subcomponents. Finally, composite total of visual rating scale score and composite total simulated metric score demonstrated a significant positive correlation (Pearson correlation, 0.31; P = 0.01). (Figure 4.2). The mean total simulated metric score was 4 ± 2.1.

Figure 4.2. Comparison of composite total visual rating scale score versus composite simulated metric score



Pearson correlation = 0.31, P = 0.01. Composite total of scale obtained by summing scale subcomponents (range, 6-60). Composite total of simulated metrics corresponded to below 25th percentile, between 25th and 50th percentile, and above 50th percentile performance on efficiency index, bimanual forces ratio, suction coordination index, and path length index (range, 0-8). Note, overall score subcomponent not included in composite score. CI, confidence interval.

DISCUSSION

Based on studies of technical performance in neurosurgery, we have recently introduced a conceptual framework to understand surgical expertise in neurosurgery.(14) Although it is clear that many non-technical factors, such as clinical decision-making, contribute to expertise, having a framework allows one to better structure research questions relating to the interaction of cognitive and motor domains and how these contribute to operative outcomes, particularly at a challenging juncture in the surgery. In keeping with this, this study aims to further clarify how one may adequately assess technical skills in neurosurgery and to better establish the role for, and limitations of, visual rating scales. There are a number of strengths related to the visual rating scale. The scale demonstrated overall reliability for as few as two raters.

Table 4.3. Comparison of Scale Subcomponents with Known Simulated Metrics

	Bimanual Cognitive				Quality	Safety					
								Instrument Force			
								Dominant		Non-Dominant	
	Efficiency Index	Path Length Index	Suction Coordination Index	Bimanual Forces Ratio	Tumor Percentage Removed	Brain Volume Removed	Blood Loss	Sum of Forces	Max Force	Sum of Forces	Max Force
Hemostasis		0.41	0.47			0.39	-0.51	0.45	0.35		
Overall		0.43	0.46	-0.29	0.44			0.52			
Instrument	0.41	0.33	0.33	-0.37				0.31			-0.29
Economy	0.44	0.36	0.39		0.29			0.39			
Flow					0.47			0.52			
Respect											
Nogo											

The main effect for participant's variance component accounts for the largest percentage (27%) of total variability, allowing for generalization of the findings to future potential participants. In assessment, it is desirable for a given scale to capture a large component of variability from participants.(15) Interestingly, these findings recapitulate variability observed in a previous study in the same population, whereby participant performance segregated into three discrete groups: high, middle, and low performers.(8)

The percentage of total subsection variability for participants by items interaction effect (13%) shows that the ratings of participants differ somewhat across scale items, averaged over raters, suggesting perhaps that each item of the scale measures a different aspect of performance. Another interpretation of these findings is that scale performance within an individual is not uniform (i.e., candidates may differ in their relative strengths and weaknesses). In our previous publication, we introduced the concept of Technical Abilities Customized Training in neurosurgery, whereby a custom psychomotor intervention tailored to the individual needs of a particular learner is carried out.(8) Given this, it may be interesting to repeat the current study with neurosurgical residents and faculty to evaluate whether performance, as judged by the visual rating scale, becomes more uniform with increasing experience.

The scale item's main effect variance component has a rate of 0.6% in total variance, indicating that participant's ratings on subcomponents of the scale were similar. A low (3.2%) variance component for rater by item interaction effect shows that individual item scales were scored similarly by a given rater.

Weaknesses associated with the scale include the high variance (20.6%) component for participant by rater interaction, suggesting that a given rater scores a given candidate more leniently or severely than other raters. This difference, however, may be accounted for by the fact that the four raters came from four different neurosurgical subspecialties (spine, oncology, epilepsy, and trauma), in addition to perhaps differing comfort with rating participants through video recordings of a simulated performance. Further rater training and calibration may also reduce variability. (16) Additionally, the participant, item, and rater interaction plus further unmeasured sources of variation were high, indicating that up to roughly one quarter of the variability is not explained by the factors measured in the study.

Other observations include a moderate variance component for the rater main effect (10.9%), suggesting that some raters are more lenient than others in their scoring across all candidates (i.e., hawks vs. doves). Although not the first case of an OSATs inspired checklist's use in neurosurgery, (2-4) this study provides interesting insights on the strengths and limitations of visual rating scales. Contrary to our initial hypothesis, aspects of the visual rating scale specifically included to capture adverse events (avoidance of no-go zone and respect for tissue) were not associated with damage to healthy simulated brain. Furthermore, the visual rating scale was not able to properly determine force characteristics exerted by the participants. This may be because of the 2-dimensional nature of the video recordings. Although arduous, this limitation could be overcome by providing the evaluators a means of viewing the surgical video in stereoscopy. Similar to the participants using the NeuroVR, evaluators of live surgery obtain a stereoscopic image of the surgical field through the operative microscope, and as such may be better suited to judge deformations in tissue caused by force exerted by a trainee. Interestingly, the positive correlation in the composite scores between the visual rating scale and simulated metrics suggests that both methods may broadly be measuring the same underlying construct, namely technical surgical performance.

The NeuroVR platform allows measurement of force application by individual instruments in all simulated tumor areas, therefore providing a comprehensive 3-dimensional representation of force

application during simulated tumor operations. Our group has exploited this information to develop the pyramid and surgical fingerprint concepts, which have contributed to our understanding of the detrimental influence of force application in specific tumor regions.(17, 18) These results would suggest that force may be a crucial element to closely monitor during neurosurgical operations. In a recent study by Sugiyama et al.,(11) using force profiles measured by specialized bipolar instruments during neurosurgical operations was associated with increased odds of intraoperative bleeding. As such, this rating scale may be used to evaluate performance in an operative setting. However, as previously mentioned, if instructors and trainees would like to better understand force applied during a surgical procedure, simulation technology should be used as an adjunct. These findings come at an important time as resident training is not only being seen as a responsibility of accreditation bodies throughout the world but is increasingly coming under the guise of quality improvement.(19) Simply put, better methods of assessment and training can help reduce patient harm.

Limitations

There are several limitations to this study. First, having raters from various neurosurgical subspecialties may have contributed to differing ratings of individuals. It may not always be feasible to have raters from the same subspecialty available to rate participants. Therefore, this represents a real-world application of this rating scale. Future improvements may lie in selecting a homogeneous rater population more familiar with the evaluated procedure.

Second, this study only includes medical students; however, our previous work with simulation suggests that medical students and junior residents share many similar psychomotor characteristics.(5)

Third, by virtue of the study design, a performance during a real operation was not rated; however, this has been previously demonstrated by others to be feasible. (3) Finally, by design, no causal relationship can be inferred between the rating scale and simulated metrics; however, the appropriate correlation, as observed for example between the hemostasis subcomponent and blood loss on the simulator, suggests that a similar underlying construct may be evaluated by both systems.

CONCLUSIONS

The visual rating scale can reliably be administered by as few as two raters and seems to reflect operative performance as measured on the simulator. However, force exerted during the neurosurgical operation and the quantity of brain volume removed and blood loss were less well captured by the visual rating scale. To our knowledge, this is the first study to be able to concurrently compare participant's ratings with their computationally measured performance and operative complications. We suggest adopting a hybrid educational approach using visual rating scales in an operative environment, supplemented by simulated training sessions to uncover potentially problematic surgical technique.

REFERENCES

1. Jensen RL, Alzhrani G, Kestle JR, Brockmeyer DL, Lamb SM, Couldwell WT. Neurosurgeon as educator: a review of principles of adult education and assessment applied to neurosurgery. Journal of Neurosurgery. 2017;127(4):949-57.
2. Aldave G, Hansen D, Briceño V, Luerssen TG, Jea A. Assessing residents' operative skills for external ventricular drain placement and shunt surgery in pediatric neurosurgery. Journal of Neurosurgery: Pediatrics. 2017;19(4):377-83.
3. Hadley C, Lam SK, Briceño V, Luerssen TG, Jea A. Use of a formal assessment instrument for evaluation of resident operative skills in pediatric neurosurgery. Journal of Neurosurgery: Pediatrics. 2015;16(5):497-504.
4. Sarkiss CA, Philemond S, Lee J, Sobotka S, Holloway TD, Moore MM, et al. Neurosurgical skills assessment: measuring technical proficiency in neurosurgery residents through intraoperative video evaluations. World Neurosurgery. 2016;89:1-8.
5. Bajunaid K, Mullah MAS, Winkler-Schwartz A, Alotaibi FE, Fares J, Baggiani M, et al. Impact of acute stress on psychomotor bimanual performance during a simulated tumor resection task. Journal of Neurosurgery. 2017;126(1):71-80.
6. 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.
7. 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.
8. Winkler-Schwartz A, Bajunaid K, Mullah MA, Marwa I, Alotaibi FE, Fares J, et al. Bimanual psychomotor performance in neurosurgical resident applicants assessed using NeuroTouch, a virtual reality simulator. Journal of Surgical Education. 2016;73(6):942-53.
9. Gélinas-Phaneuf N, Choudhury N, Al-Habib AR, Cabral A, Nadeau E, Mora V, 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.
10. Alotaibi FE, AlZhrani GA, Mullah MA, 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.
11. 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):234-42.
12. Azarnoush H, Alzhrani G, Winkler-Schwartz A, Alotaibi F, Gelinas-Phaneuf N, Pazos V, 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):603-18.
13. Aggarwal R, Grantcharov T, Moorthy K, Milland T, Darzi A. Toward feasible, valid, and reliable video-based assessments of technical surgical skills in the operating room. Annals of Surgery. 2008;247(2):372-9.
14. Sawaya R, Alsideiri G, Bugdadi A, Winkler-Schwartz A, Azarnoush H, Bajunaid K, et al. Development of a performance model for virtual reality tumor resections. Journal of Neurosurgery. 2018;131(1):192-200.
15. Streiner DL, Norman GR, Cairney J. Health Measurement Scales: A Practical Guide to Their Development and Use. Oxford, England, UK: Oxford University Press; 2015.

16. Feldman M, Lazzara EH, Vanderbilt AA, DiazGranados D. Rater training to support high-stakes simulation-based assessments. Journal of Continuing Education in the Health Professions. 2012;32(4):279-86.

17. 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. 2018;14(6):686-96.

18. Azarnoush H, Siar S, Sawaya R, Al Zhrani G, Winkler-Schwartz A, Alotaibi FE, et al. The force pyramid: a spatial analysis of force application during virtual reality brain tumor resection. Journal of Neurosurgery. 2016;127(1):171-81.

19. Pang P, Raslan AM, Selden NR. Improving performance by improving education. Quality and Safety in Neurosurgery. Cambridge, Massachusetts: Elsevier, Academic Press; 2018. p. 213-24.

Chapter 5 - Creating a Comprehensive Research Platform for Surgical Technique and Operative Outcome in Primary Brain Tumor Neurosurgery

Preface

Chapter 4 elucidated the role of a visual rating scale in neurosurgery. However, applying such a scale, as well as directed educational interventions directly in a live-operative context may introduce unknown and unwanted risks to patients. As such, the current chapter outlines the development of a brain tumor model involving an ex vivo animal brain with incorporated tumour which may serve as a "middle-way" model; capturing the essence of live oncological neurosurgery, while still retaining the lab-based environmental control inherent to the simulated tasks outlined in previous chapters. At publication, it was the first model created using biomechanical reference data obtained from real patient brain tumors. The animal model would serve as the condition by which one could evaluate the value of VR simulation training in the context of the RCT. Furthermore, real time movements of surgical instruments captured via instrument mounted fiducials, as well as extent of resection measured "bedside" or by magnetic resonance image can serve as quantifiable outcomes measures in the RCT. The manuscript was published as:

Winkler-Schwartz A, Yilmaz R, Tran DH, Gueziri HE, Ying B, Tuznik M, Fonov V, Collins L, Rudko DA, Li J, Debergue P, Pazos V, Del Maestro R. Creating a Comprehensive Research Platform for Surgical Technique and Operative Outcome in Primary Brain Tumor Neurosurgery. World Neurosurg. 2020 Dec;144:e62-e71. doi: 10.1016/j.wneu.2020.07.209. Epub 2020 Aug 3. PMID: 32758649.

ABSTRACT

Background: The operative environment poses many challenges to studying the relationship between a surgical act and patient outcomes in intracranial oncological neurosurgery. The authors sought to develop a framework whereby neurosurgical performance and operative outcomes could be precisely quantified in a controlled setting.

Methods: Stiffness of an alginate hydrogel-based tumor was modified with differing concentration of cross-linking agent calcium sulfate until similar biomechanical properties to human primary brain tumors measured at resection were achieved. Artificial tumor was subsequently incorporated into an ex vivo animal brain as a final model. MRI enhancement and ultraviolet (UV) fluorescence was achieved by incorporating gadolinium and fluorescein solution, respectively. Video from operative microscope, ceiling cameras, and recordings of instrument-mounted fiducials within a surgical suite environment captured operative performance.

Results: Twenty-four rheometer measurements were conducted on alginate hydrogels containing 10, 11 and 12 millimolar concentrations of calcium sulfate. Sixty-eight stiffness measurements were conducted on 8 patient tumor samples. No differences were found between alginate and brain tumor stiffness values (Kruskal-Wallis $\chi^2(4) = 9.187$, $p = 0.057$). Tumor was identified on UV fluorescence and ultrasound. Volume and location of resected white and grey matter and residual tumor could be quantified in 0.003 mm^3 increments using 7-Tesla MRI. Ultrasonic aspirator and bipolar movement data could be successfully transformed into performance metrics.

Conclusion: This framework can offer clinicians, learners, and researchers the ability to carry out operative rehearsal, teaching, or studies involving brain tumor surgery in a controlled

laboratory environment and represents a crucial step in the understanding and training of expertise in neurosurgery.

INTRODUCTION

The removal of brain tumors, a skill expected of neurosurgical graduates, confers significant risk to patients and remains among the most technically challenging procedures within medicine. There exist little more than anecdotal accounts of the most effective ways of teaching oncological neurosurgery.

A research framework relating technical performance to operative outcomes in oncological neurosurgery is key to answer questions relating to the training of neurosurgeons and the examination of operative techniques and technologies. It must account for tumor variability, ensure an adequate means of capturing operative performance in real time, and have an accurate means of evaluating operative success. Unfortunately, such a platform does not yet exist.

The influence of a single surgeon's technical performance on operative outcomes can be best understood by creating a convincing operative mimic of a brain tumor surgery including controlling for tumor size, location, stiffness, bleeding, and environmental factors. Such a mimic in a standardized operating room represents an ideal setting to conduct interventional experiments while obviating patient safety concerns.

Little effort has been made to recreate the known tactile and imaging properties of human brain tumors when creating artificial mimics. Various substances have served as artificial brain tumors, from fibrin glue(1), silicone(2, 3), polymer resins(4-6), food-grade gels(7-9), autologous animal organs(10) and polyvinyl alcohol(11-13). While authors have described imaging(11, 14, 15) and biomechanical properties(16) of these artificial tumors, none have explicitly developed tumors using real human brain tumors as a reference standard. None offer a platform designed to recreate a human brain tumor operative experience while including a means of capturing and quantifying operative performance relative to post-operative outcome.

Alginate hydrogels are polymers derived from algae whose biomedical applications are expanding(17), including drug delivery(18), wound dressing and tissue engineering(19). These polymers, while initially liquid-like, can form solid hydrated materials which can be adjusted to required stiffness at room temperature by varying the concentration of ionic cross-linking agents. These polymers are biocompatible and ideal substances to inject into ex or in-vivo tissues, including brain. Compared to the fibrin glue and agarose gels, alginate hydrogels feature enhanced capacity to develop appropriate stiffness and mechanical toughness(17). Moreover, alginate hydrogels exhibit viscoelastic behavior as brain tissues(20). Furthermore, unlike the polyvinyl alcohol, desired stiffness can be accomplished in a mild condition without the need for freeze-thaw cycles.

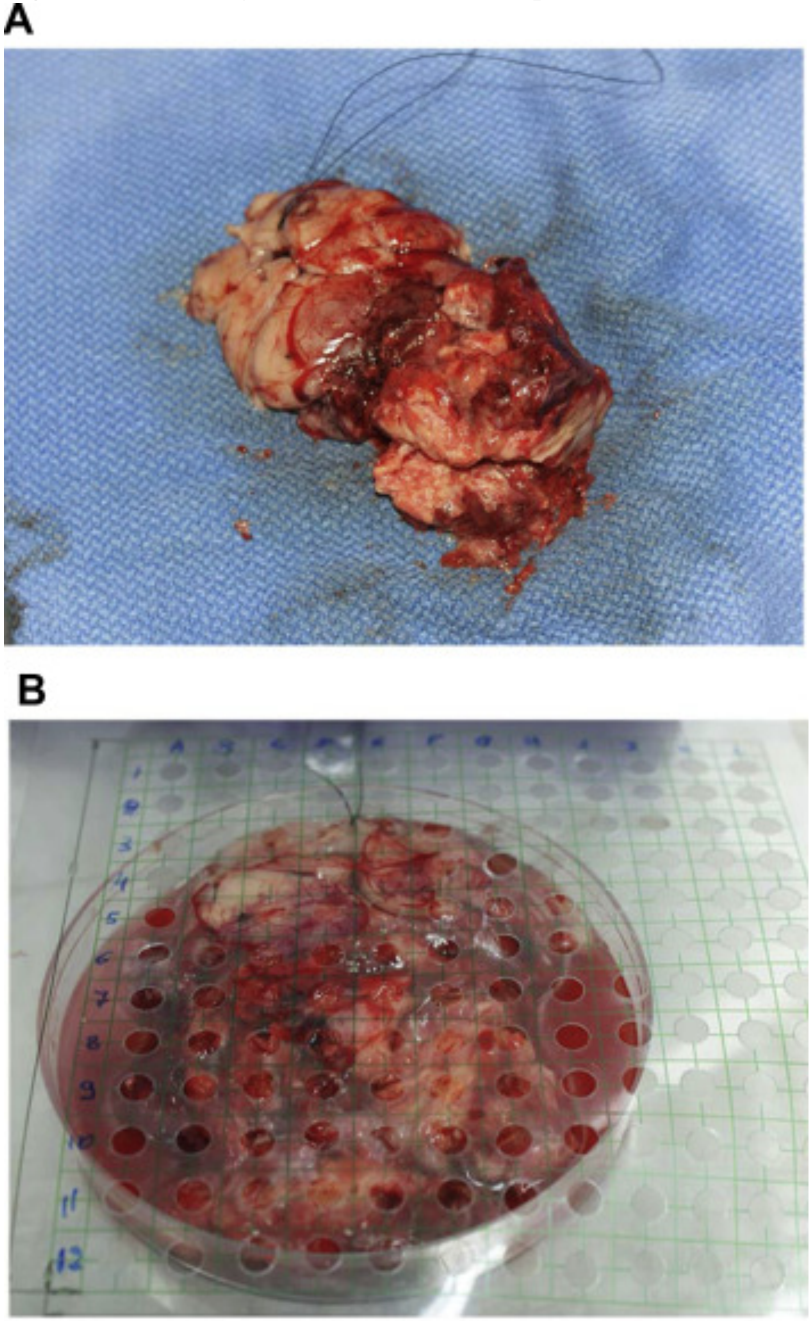
The authors sought to develop a framework whereby surgical performance and operative outcome could be precisely quantified. This was accomplished by creating an artificial brain tumor using an alginate hydrogel incorporated into an animal brain with similar biomechanical and imaging characteristics to human brain tumors. This allows for accurate pre- and post-operative magnetic resonance and ultrasound-based imaging. Operative performance was captured via video-recordings from operative microscope, ceiling mounted cameras, and instrument-mounted fiducials within a surgical environment.

METHODS

Establishing Biomechanical Properties of Human Primary Brain Tumors

Human tumor biomechanical characterization was carried out using a 2-mm diameter flat-punch portable indenter, designed and assembled at the National Research Council of Canada (Boucherville, QC, Canada). The device was attached to a fixed arm above an elevated plate onto which the surgical sample was placed. Following excision of brain tumor tissue, the specimens were placed in a saline solution just above 0^{o} C to limit structural changes to the tissue. Although biomechanical properties of human brain tissue do not change substantially within the first 6 hours post resection(21), we transported specimens to the pathology lab within 5 minutes, with biomechanical measurements begun within 15 minutes from resection (Figure 5.1, upper panel). Specimens were divided into quadrants using pre-cut guides, and measurements were repeated twice (Figure 5.1, lower panel). Following the biomechanical characterization, samples were directed to the pathology department for clinical assessment. Histopathological diagnosis of each specimen was obtained via the clinical pathological report. Any specimen results that were not confirmed to contain tumor were removed from the analysis.

Figure 5.1. Primary human brain tumor specimens.



***A**) Primary human brain tumor with reference suture immediately after resection. (**B**) Primary human brain tumor with superimposed grid. Stiffness was measured by guiding the flat punch through the precut circular hole in the grid.*

Informed consent was obtained from each patient whose tumor was used. The local University Health Centre Research Ethics Board, Neurosciences-Psychiatry approved the studies to carry out the human tumor stiffness characterizations.

Development of Artificial Tumor

Biomechanical properties of the hydrogel were characterized using the Model HR-2 Rheometer (TA Instruments, Delaware, United States). Each testing session represented a new tumor. No tumors were tested more than a single time.

A 2% weight by volume (W/V) Algin I-1G Alginate (KIMICA Corporation, Tokyo, Japan) served as the basis for the artificial tumor. Deionized water at room temperature was mixed with the requisite powder of alginate (kept in standard refrigeration for preservation purposes) and was allowed to passively mix over a period of three days until full dissolution in standard residential-grade refrigeration at 4 degrees Celsius. We chose to carry out all biomechanical experiments within 1 week of alginate gel creation to avoid any potential degradation. Desired tumor stiffness was obtained by varying the input of calcium sulfate solution in the alginate-calcium sulfate mixture. To imitate the post-gadolinium hyper-intensity seen in many high-grade primary brain tumors, 10 times dilution of gadolinium solution, Gadobutrol (Bayer AG, Leverkusen, Germany) was added to the deionized water to reach a final concentration of 1 millimolar (mM) per litre in the hydrogel. Tumor fluorescence under ultraviolet light was achieved by adding to the deionized water a fluorescein solution extracted from an “invisible ink” marker (iPang UV light Pen, iPang Co.-Ltd, South Korea).

The final composition of 1,100 microliters (µL) tumor consisted of 1,000 µL of 2% W/V alginate gel, 71 µL of deionized H_2O with yellow and red coloring (Club House, McCormick & Company, Inc, MD, United States), 13 µL of 1 M concentration calcium sulfate, 11 µL of 100 μM gadolinium solution and 5 μL of fluorescein solution. This corresponded to a final calcium sulfate concentration of 12 mM within the hydrogel.

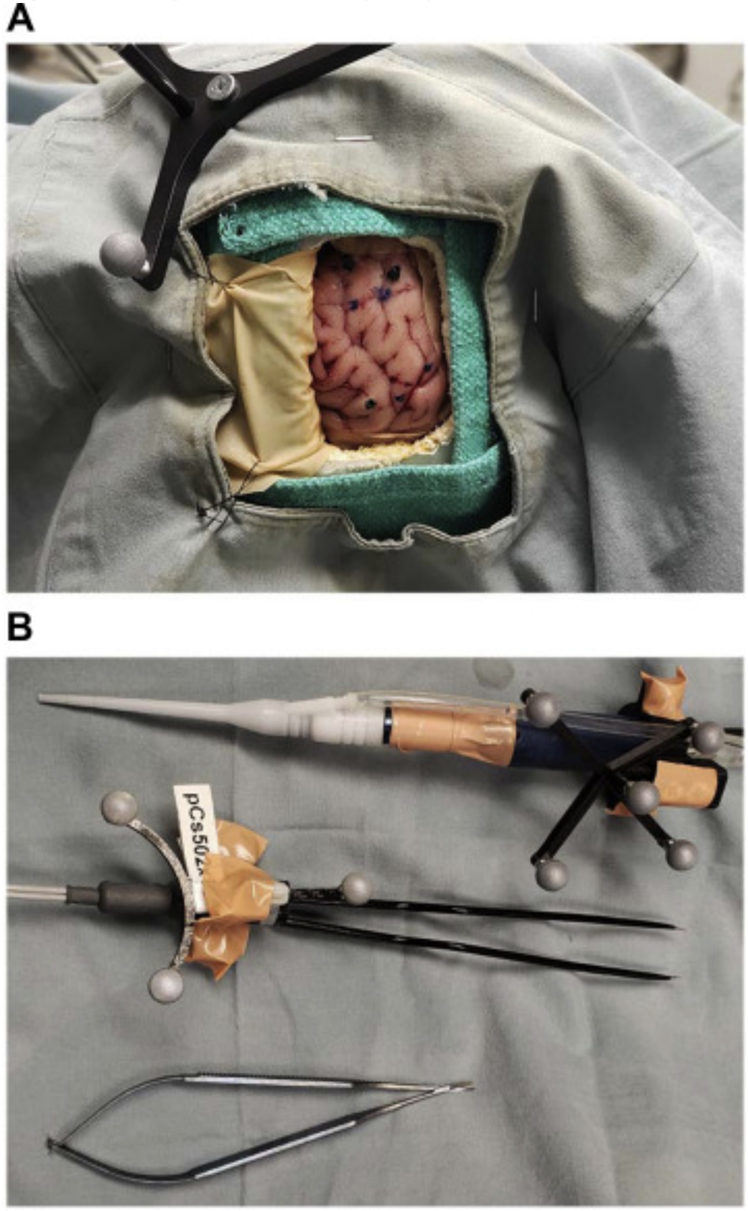
Ex vivo Animal Brain

Cranially extricated, fresh calf brains were obtained. These were chosen due to their abundant availability, cost, size (approximately 300 grams, small enough to fit in an animal 7 Tesla MRI coil) and morphological similarity to human brain(22). Bovine simulation platforms have also been described for microsurgical training in neurosurgery(2, 23-25).

Alginate hydrogels were injected at a 30-degree angle at a subcortical depth of 5-7 mm in the longest continuous frontal gyrus, typically the second frontal gyrus. To give a healthy margin for tumor solidification, 30 minutes was allowed to elapse prior to operation. The tumor was shaped by hydro-dissection at the time of injection.

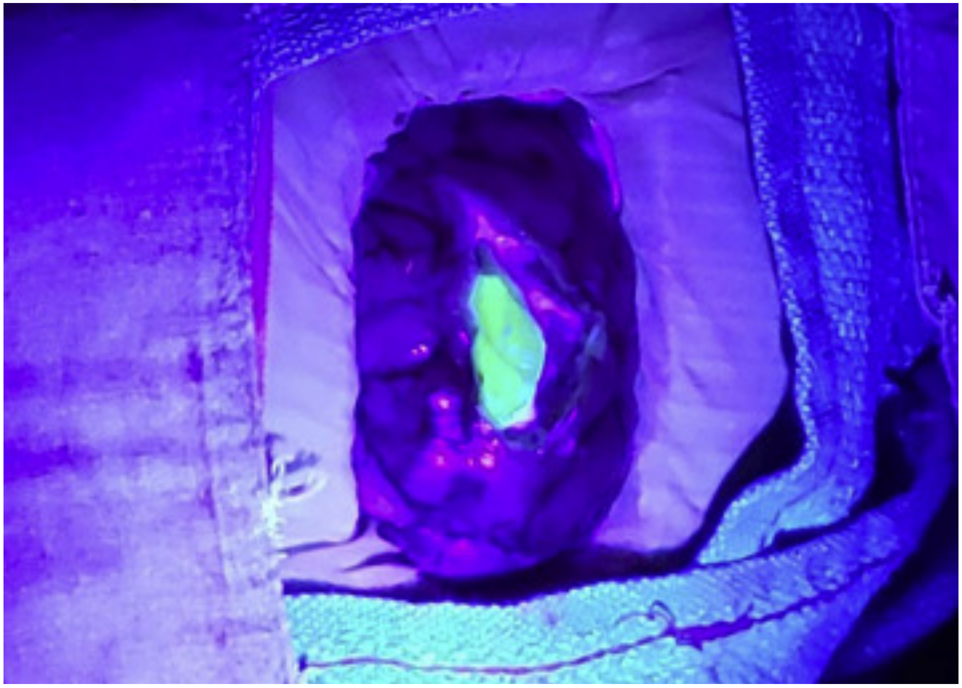
The *ex vivo* brain with injected tumor was placed into a 3D printed (Ultimaker S5, Ultimaker, Utrecht, Netherlands) form fitted holder underneath a plastic cranium with pre-cut window mimicking an off-midline craniotomy view squared off with towels and drapes (Figure 5.2, upper panel). Micro scissors, bipolar electrocautery and an ultrasonic aspirator (Stryker, Kalamazoo, MI, USA) were available (Figure 5.2, lower panel). Incorporated fluorescein solution made for avid tumor fluorescence under ultraviolet black light (Figure 5.3). The operation was conducted in an animal operative suite equipped with an OPMI Pico surgical microscope (Carl Zeiss Co., Oberkochen, Germany). To record the operation, HDMI outputs from the microscope and a ceiling mounted camera were routed through an HDMI recording device (HDML-Cloner box turbo HCB-988BT, Cloner Alliance Inc.). A sample operation of a subpial resection technique can be viewed in Video 1 and key images can be seen in Figure 5.4.

Figure 5.2. Operative view of artificial brain tumor and ex vivo calf brain.



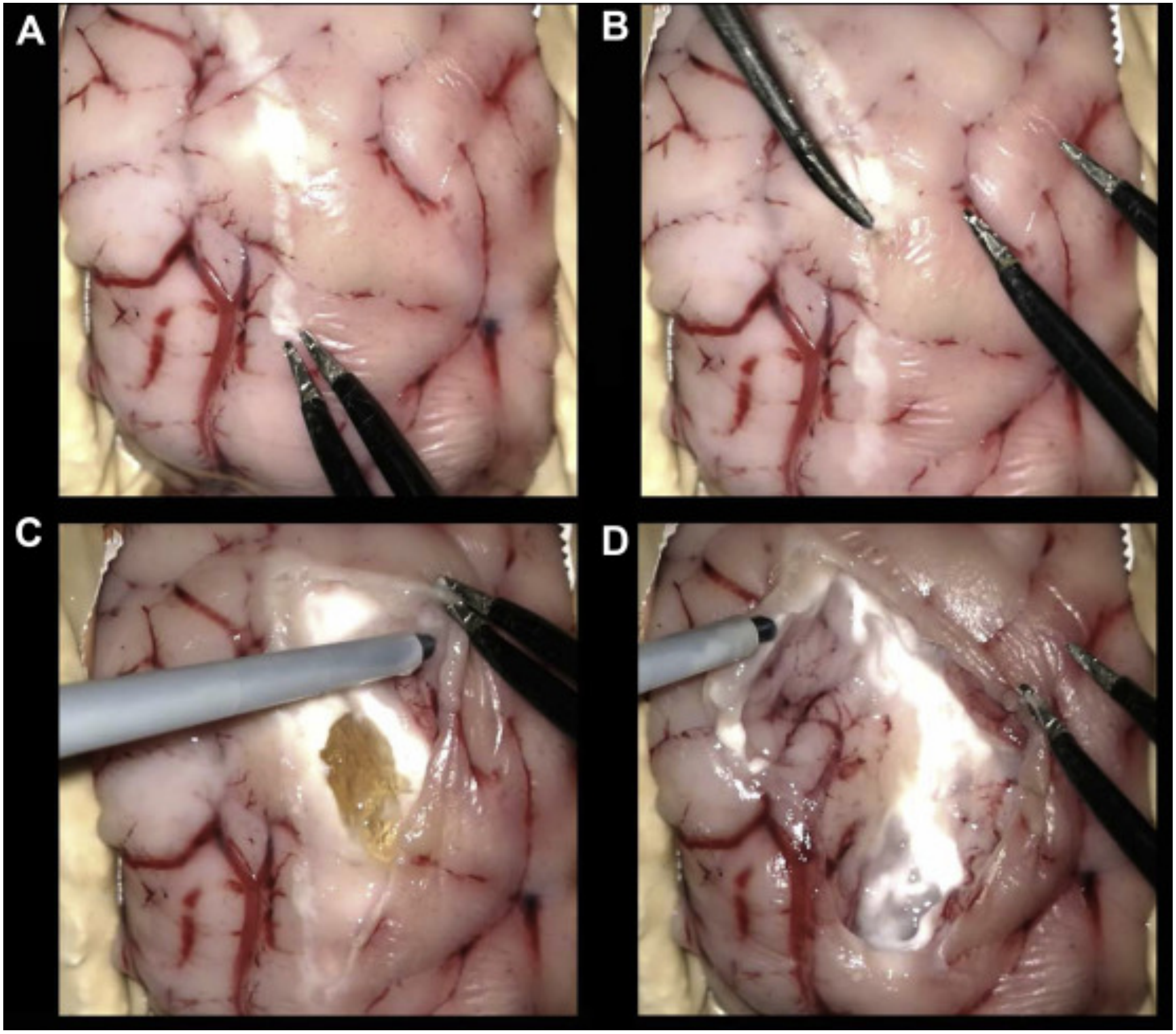
*(**A**) Operative view of draped surgical field for ex vivo calf brain containing an alginate hydrogel artificial tumor. Ink on the brain surface outlines the limits of surgical resection.* Black spheres *represent tips of fiducial markers in the brain used to merge the pre- and postoperative images. (*Upper Left*) A fixed reference arm with fiducial markers can be seen. (**B**) Microscissors, ultrasonic aspirator, and bayonetted bipolar electrocautery device with custom 3-dimensional printed mounted fiducial markers to all for movement capture. Surgical tape was used to cover any reflective metal surfaces.*

Figure 5.3. Photograph of an alginate tumor with incorporated florescent solution shown fluorescing under ultraviolet light.



The cortical surface had been unroofed to allow for direct tumor visualization.

Figure 5.4. Resection of an alginate tumor implanted in an ex vivo calf brain as viewed through an operative microscope.



*(**A**) Pial coagulation with bipolar device. (**B**) Pial cutting using bayonetted microscissors. (**C**) Subpial technique performed with ultrasonic aspirator and bipolar device. Tumor shown in* yellow. *(**D**) Postoperative view demonstrating complete resection of tumor and gyrus with adjacent sulcal banks seen on either side. White matter can be seen deep in the cavity.*

Surgical Movement Capture

Fiducials (Northern Digital Inc., ON, Canada) were attached to the bipolar and ultrasonic aspirator via custom 3D printed polylactic acid mounts. An optical tracking camera (FusionTrack 500, AtracsysLLC, Puidoux, Switzerland) captured movement of the bipolar and ultrasonic aspirator with reference to fixed fiducials mounted adjacent to the craniotomy window.

Imaging Characteristics of Artificial Tumor on MRI and Ultrasound

MRI was performed using the 7 T Bruker Pharmascan (Bruker Biosciences, Billerica, MA) ultra-high field system. Brains were housed in a cylindrical container and immersed in an MR-invisible fluorinated solution, FC-40 (Sigma Aldrich, St. Louis, Missouri), to remove background MRI signal. For radiofrequency excitation and reception, a 6 cm inner diameter volume resonator was used. The imaging protocol included a 3D steady-state free precession MRI sequence with an echo time (TE) of 5 milliseconds, repetition time (TR) of 10 milliseconds, a receiver bandwidth of 50 kHz and an excitation pulse flip angle of 30 degrees. The image acquisition matrix was selected to achieve an isotropic voxel resolution of 150 μm^3. The Field of View size was adjusted according to the dimensions of the calf brain. However, the voxel resolution remained constant for all samples. Anti-aliasing was activated along the 2nd phase-encode axis to prevent image aliasing. Twenty-four signal averages were collected leading to a total scan time of approximately 12 hours for the overnight histology-grade scan. To assist in alignment of pre- and post-operative images, plastic reference arrays were inserted into the brain at the margins of the craniotomy. Tumor enhancement was achieved by incorporating Gadobutrol (Bayer AG, Leverkusen, Germany) in the tumor.

Ultrasonic images of the tumor were also acquired using an HDI 5000 ultrasound with a phased array probe P4-7 (Philips, Amsterdam, Netherlands). The probe spatial position in relation to a fixed reference tool was obtained using the IBIS neuro-navigation system(26) with an optical tracking camera (Polaris, Northern Digital Inc., ON, Canada).

Statistical and Imaging Analysis

We report descriptive statistics as counts and percentages for categorical variables. For normally distributed continuous variables, means and standard deviations are utilized. Data analysis and visualization were conducted using Stata (StataCorp. 2013. Stata Statistical

Software: Release 13. College Station, TX: StataCorp LP) and MATLAB (Statistics Toolbox Release 2017b, The MathWorks, Inc., Natick, Massachusetts, United States), respectively. Ultrasound and MRI image post processing were completed using an open-source software application framework, 3D Slicer(27) (Version 4.10.01).

RESULTS

Mechanical Properties of Human Primary Brain Tumors

A total of 68 tumor stiffness measurements were conducted on eight patient samples. Biomechanical properties of these groups are outlined in Table 5.1.

Table 5.1. Biomechanical Properties of Human Primary Brain Tumors

Object	Force in kilopascal	
	Mean (SD; min-max)	Median (inter-quartile range)
Human Brain Tumors		
DNET (n=1)	3.31 (1.35; 1.81-6.61)	2.86 (2.80-3.27)
Oligodendroglioma (n=2)	6.74 (6.82; 1.14-23.07)	1.95 (1.52-12.51)
Anaplastic Astrocytoma (n=1)	1.37 (0.67; 0.50-2.38)	1.34 (0.72-1.88)
Glioblastoma (n=3)	4.14 (2.68; 0.44-9.85)	2.93 (2.36-5.98)

n = number. Min = minimum. Max = maximum. SD= standard deviation.
DNET = Dysembryoplastic Neuroepithelial Tumor

Mechanical Properties of Artificial Tumor as Compared to Primary Human Brain Tumors

Rheology was performed on pure alginate and calcium sulfate mixtures. Twenty-four rheometer testing sessions were conducted, each lasting one hour. Four sessions were excluded due to technical problems. Hydrogels reached a maximum stiffness in a mean of 81 seconds (range 36 – 130 seconds). For ease of analysis when comparing with hydrogel stiffness, brain tumors were grouped into two: glioblastoma and all other tumors. Table 5.2 outlines the biomechanical properties of the various hydrogel concentrations in relation to the primary brain tumor groups outlined. Due to the skewed nature of the data, Kruskal-Wallis was performed to compare the stiffness of the tumor groups with the 10, 11- and 12-mM concentrations of alginate hydrogel. No significant differences were found between any groups ($\chi^2(4) = 9.187$, $p = 0.057$).

Table 5.2. Biomechanical Properties of Alginate Hydrogels Compared to Primary Human Brain Tumors

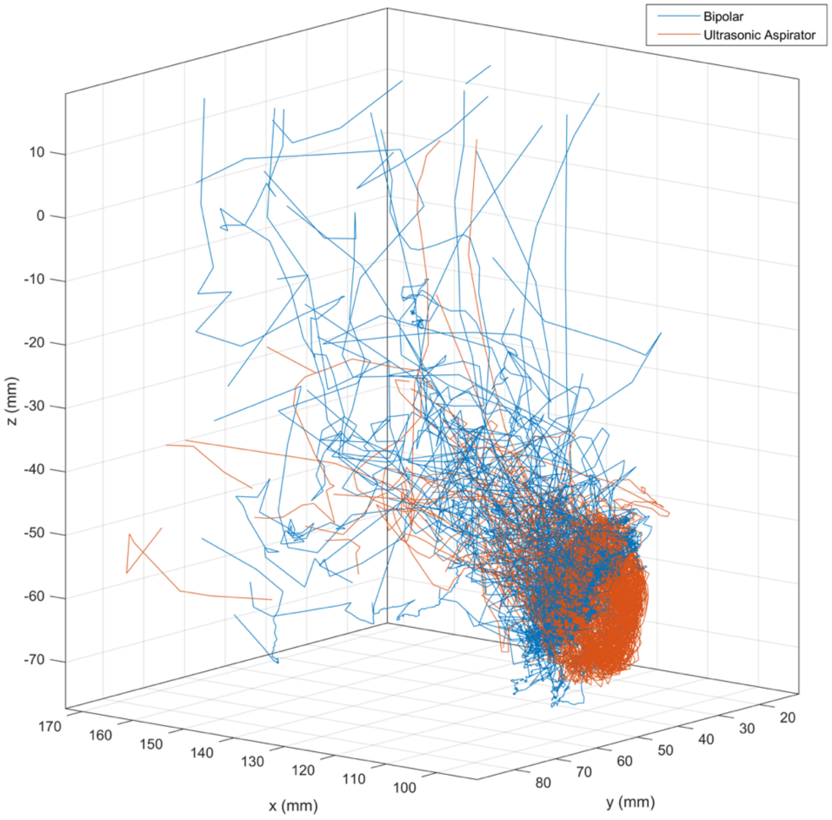
Object	Force, kilopascal	
	Mean (SD; min-max)	Median (inter-quartile range)
Alginate Hydrogels		
Calcium 10 µL (n=7)	2.01 (0.75; 1.08-3.18)	2.22 (1.14-2.46)
Calcium 11 µL (n=9)	2.85 (0.72; 1.80-3.90)	2.88 (2.43-3.27)
Calcium 12 µL (n=4)	2.24 (0.64; 1.80-3.18)	1.98 (1.83-2.64)
Human Brain Tumors		
Glioblastoma (n=3)	4.14 (2.68; 0.44-9.85)	2.93 (2.36-5.98)
Other (n=4)	3.82 (4.80; 0.50-23.07)	1.89 (1.26-3.26)

n = number. Min = minimum. Max = maximum. SD= standard deviation.

Surgical Movement Capture

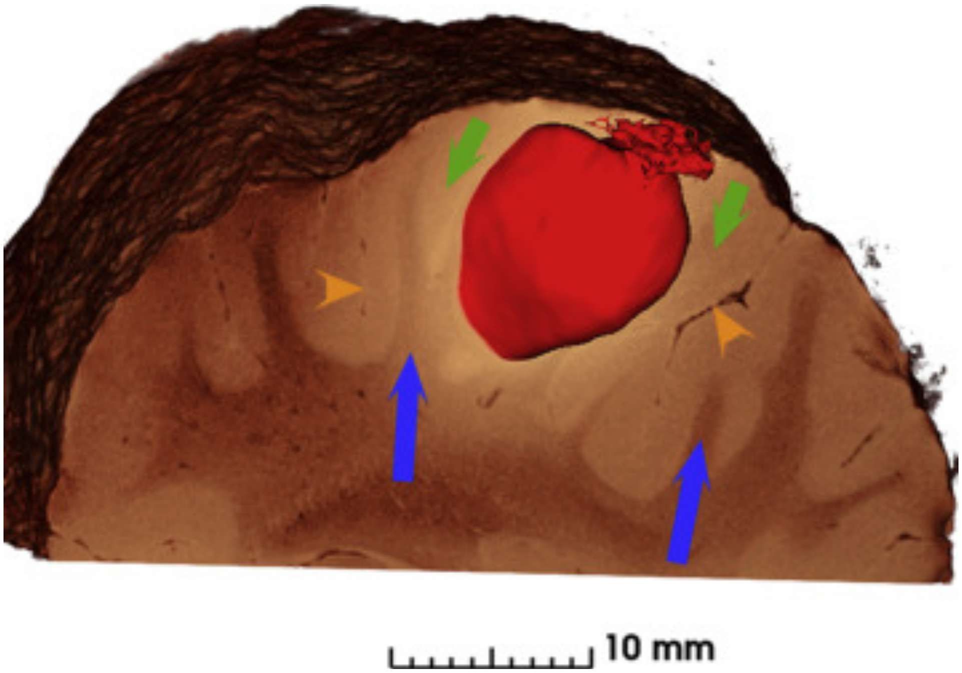
Movement data of the instruments could be successfully captured (Figure 5.5) and transformed into performance metrics using techniques previously described in neurosurgical virtual reality (VR) surgery(28-31). Mean aspirator and bipolar velocity were 16.8 millimeter/second (mm/s) and 11.35 mm/s, respectively. Mean aspirator and bipolar acceleration were 7.81 mm/s ^{2} and 5.21 mm/s ^{2}, respectively. Mean instrument tip distance was 9.95 millimeters.

Figure 5.5. Surgical motion capture.



*Graph showing 3-dimensional reconstruction of motion of the ultrasonic aspirator (*red*) and bipolar device (*blue*) during artificial brain tumor resection.*

Figure 5.6. Magnetic resonance image (7T) of alginate hydrogel tumor (red) in an ex vivo calf brain.

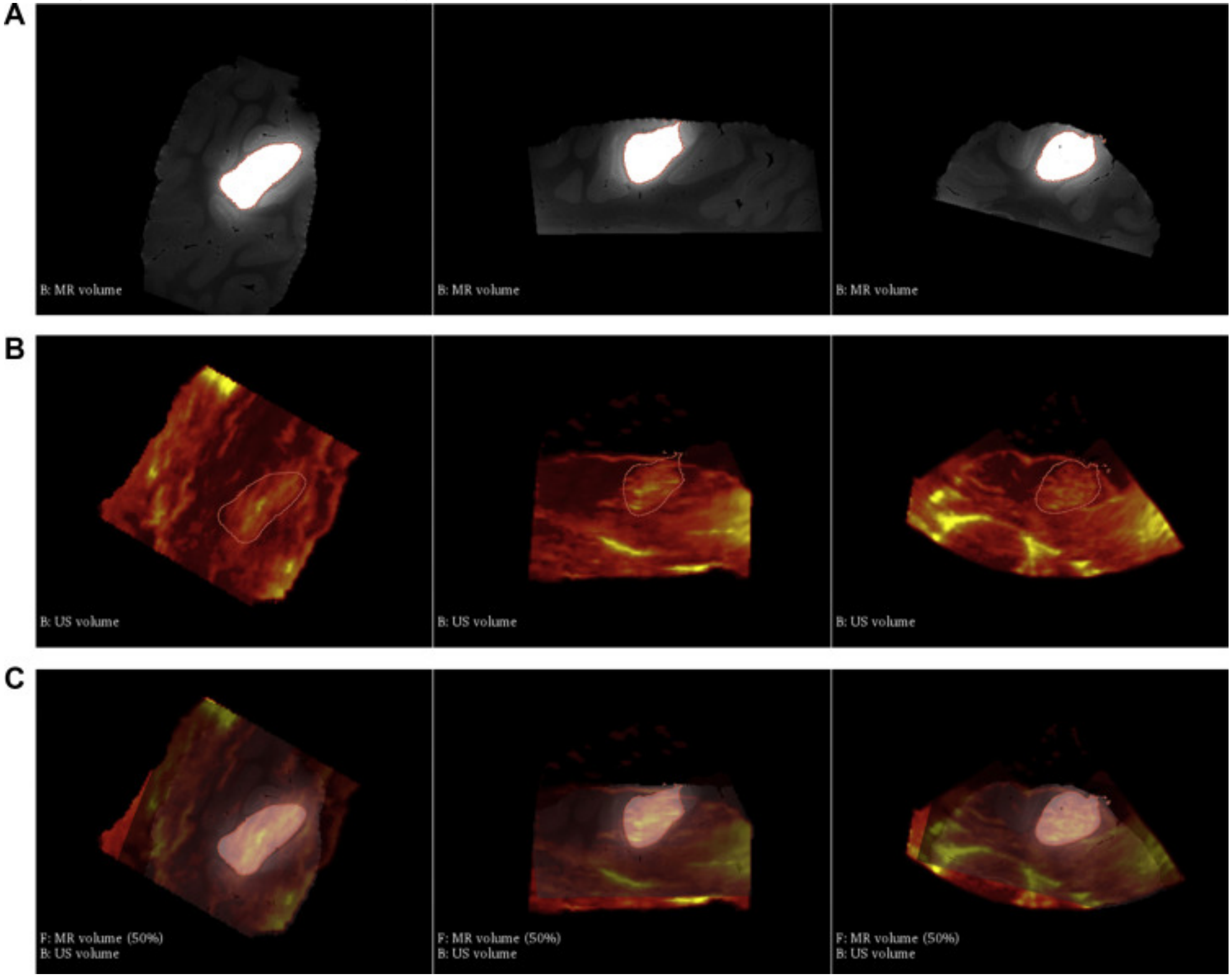


Long blue arrows indicate white matter; medium green arrows indicate gray matter; and short orange arrows indicate sulci.

Imaging Characteristics of Tumor

The tumor and surrounding brain architecture were well identified on MRI (Figure 5.6). MRI and ultrasound image registration allowed for ease of identification of the tumor on MRI and ultrasound (Figure 5.7). Using 3D slicer with Otsu thresholding and segmentation, the hyper-intense tumor, as well as grey and white matter could be quantified in 0.003 mm^3 increments.

Figure 5.7. A comparison of magnetic resonance images and ultrasound images of the alginate hydrogel tumor in an ex vivo calf brain.



*Sagittal, axial, and coronal views of the hyperintense alginate hydrogel tumor in ex vivo calf brain on (**A**) 7T magnetic resonance images, (**B**) ultrasound images, and (**C**) both overlaid. Tumor hyperintensity was achieved by incorporating gadolinium solution into the hydrogel tumor. The* red outline *of the tumor was added via postprocessing in 3D Slicer and visualized using IBIS (intraoperative brain imaging system).*

DISCUSSION

We have developed a comprehensive research framework which allows for the study of technical performance and operative outcomes in oncological neurosurgery. This platform relies on a cost-effective alginate-based artificial brain tumor incorporated into an ex vivo calf brain within a controlled operative environment. This represents the first instance where an artificial tumor has been created based on biomechanical properties of human specimens obtained at time of resection. Operative "performance" can be assessed both via recordings from the surgical microscope and ceiling mounted camera, and by movements generated from instrument-mounted fiducials, while operative "success" can be assessed at time of surgery by presence of residual tumor, via ultraviolet fluorescence or ultrasound. Finally, MRI of residual tumor, as well as the location and volume of grey and white matter resected in 0.003 mm^3 increments present an opportunity for a precise quantification of operative outcome.

Extensive work has been done to create meaningful surgical performance measures from raw movement data VR neurosurgery ranging from simple, user generated metrics(28-31), three dimensional representations of movement and force(32, 33), to artificial-intelligence assisted methodologies(34-36). The promise of performing similar analyses on live surgical movement in an experimental or even clinical environment may allow for a better understanding of surgical nuance, technical expertise and complication avoidance.

The development of simulation platforms such as the NeuroVR(37) (formerly NeuroTouch, CAE Healthcare, Montreal, Quebec, Canada) have made it possible to better understand the technical composites required to carry out intra-axial tumor resection. By allowing for the creation of complex tumor resection scenarios(38) and by integrating the physical properties of human brain tumors obtained at time of resection, the NeuroVR provides the most realistic computer-based recapitulation of an oncological neurosurgery to date.

Face, content and construct validity (28-31) for the NeuroVR have been demonstrated, however, the question of concurrent validity, that is, whether practice on the simulator improves performance in the "real world" is best addressed by conducting a randomized controlled trial.

Unfortunately, significant variability due to pathology, patient factors, the unpredictability of the operating room environment, and multi-surgeon input confounds the relationship between a surgical act and patient outcome and serves to decrease statistical power. Although a reduction in power can be traditionally addressed by increasing the number of recruits, notwithstanding the extra resources this may require, the operative setting presents ethical concerns for patient safety when experimental interventions are involved. A research ethics board may question the existence of equipoise in a study comparing traditional residency training (control group) with traditional residency training in addition to virtual reality operative rehearsal (experimental group). Additionally, the residents which may most benefit from operative rehearsal may be at a stage in their training where they may not yet safely conduct oncological neurosurgery with minimal supervisor assistance.

If successful, the proposed randomized controlled trial will be the first time simulated operative rehearsal has been demonstrated to influence surgical performance in an open neurosurgical procedure. Surgical “boot camps” involving simulators have already been developed for trainees early in a neurosurgical residency(39, 40). These residents, early in their learning curve, stand the most to benefit from the experiential training afforded by simulators(41) and one could envision that residency programs, hospital administrators and patient advocacy groups may make simulation training a mandatory precursor to participation in high risk operative cases early in training. Ultimately standardizing surgical education and training to expert, rather than competence level has the potential to reduce operative complications, leading to decreased patient morbidity and mortality and reduced medical care costs.

Limitations

Unless the preparations are performed in sterile environments, bacterial contamination can occur which can lead to biomechanical inconsistency of the gel beyond 2 weeks. Approaches to sterilize the alginate such as sterile filtration have been well established in literature, but were not pursued in this work. Three days between alginate gel creation and use is required therefore some foresight is necessary to integrate its use in a clinical/educational context. No active bleeding state exists in the current model. Although we achieved intracranial blood flow by the use a porcine brain(42, 43) with cannulated carotid arteries as described in human cadaveric studies(44), this animal model had significant limitations. The thick skull and relatively small brain volume made surgical access time-consuming and inefficient(45). Furthermore, the large porcine head was not able to fit inside the 7T MRI and removing the brain to allow for accurate scanning risked damaging the tissue.

The hydro-dissection caused by the tumor injection into brain prevents the creation of an indistinct brain-tumor border characteristic of primary brain tumors. In addition, the *ex vivo* nature of the model does not allow for the development of reactive gliosis (and the resulting subtle biomechanical changes) surrounding the tumor familiar to neurosurgeons. Ideally, tumors injected into live animals should be given time to develop a reactive gliosis, however the challenges and ethical concerns in maintaining animals during this period are substantial.

Operative outcomes have been defined purely in imaging terms. While the ex vivo nature of the brain samples precludes a functional assessment, extent of subcortical white matter injury may provide a useful imaging correlate, as some have suggested (46).

CONCLUSIONS

A comprehensive research framework to study operative expertise in oncological intracranial neurosurgery has been developed. This framework can offer the clinician, learner, or researcher the ability to carry out operative rehearsal, teaching, or studies involving intraparenchymal brain tumor surgery in a controlled laboratory environment and represents a crucial step in the understanding and training of expertise in neurosurgery.

REFERENCES

1. Aurich LA, Silva Junior LF, Monteiro FM, Ottoni AN, Jung GS, Ramina R. Microsurgical training model with nonliving swine head. Alternative for neurosurgical education. Acta cir. 2014;29(6):405-9.
2. Turan Suslu H, Ceylan D, Tatarli N, Hicdonmez T, Seker A, Bayri Y, et al. Laboratory training in the retrosigmoid approach using cadaveric silicone injected cow brain.[Erratum appears in Br J Neurosurg. 2014 Dec;28(6):823 Note: Bahri, Yasar [corrected to Bayri, Yasar]]. Br J Neurosurg. 2013;27(6):812-4.
3. Oliveira MM, Araujo AB, Nicolato A, Prosdocimi A, Godinho JV, Valle ALM, et al. Face, content, and construct validity of brain tumor microsurgery simulation using a human placenta model. Oper Neurosurg. 2015;12(1):61-7.
4. Ashour AM, Elbabaa SK, Caputy AJ, Gragnaniello C. Navigation-Guided Endoscopic Intraventricular Injectable Tumor Model: Cadaveric Tumor Resection Model for Neurosurgical Training. World Neurosurg. 2016;96:261-6.
5. Berhouma M, Baidya NB, Ismail AA, Zhang J, Ammirati M. Shortening the learning curve in endoscopic endonasal skull base surgery: a reproducible polymer tumor model for the trans-sphenoidal trans-tubercular approach to retro-infundibular tumors. Clin Neurol Neurosurg. 2013;115(9):1635-41.
6. Gragnaniello C, Nader R, van Doormaal T, Kamel M, Voormolen EH, Lasio G, et al. Skull base tumor model. Journal of neurosurgery. 2010;113(5):1106-11.
7. Kamp MA, Knipps J, Steiger H-J, Rapp M, Cornelius JF, Folke-Sabel S, et al. Training for brain tumour resection: a realistic model with easy accessibility. Acta Neurochir (Wien). 2015;157(11):1975-81.
8. Mashiko T, Oguma H, Konno T, Gomi A, Yamaguchi T, Nagayama R, et al. Training of Intra-Axial Brain Tumor Resection Using a Self-Made Simple Device with Agar and Gelatin. World Neurosurg. 2018;109:e298-e304.
9. Valli D, Belykh E, Zhao X, Gandhi S, Cavallo C, Martirosyan NL, et al. Development of a Simulation Model for Fluorescence-Guided Brain Tumor Surgery. Front. 2019;9:748-.
10. Csokay A, Papp A, Imreh D, Czabajszky M, Valalik I, Antalfi B. Modelling pathology from autolog fresh cadaver organs as a novel concept in neurosurgical training. Acta Neurochir (Wien). 2013;155(10):1993-5.
11. Chen SJ, Hellier P, Marchal M, Gauvrit JY, Carpentier R, Morandi X, et al. An anthropomorphic polyvinyl alcohol brain phantom based on Colin27 for use in multimodal imaging. Med Phys. 2012;39(1):554-61.
12. Chen SJ, Hellier P, Gauvrit JY, Marchal M, Morandi X, Collins DL. An anthropomorphic polyvinyl alcohol triple-modality brain phantom based on Colin27. Med Image Comput Comput Assist Interv. 2010;13(Pt 2):92-100.
13. Reinertsen I, Collins DL. A realistic phantom for brain-shift simulations. Med Phys. 2006;33(9):3234-40.
14. Lee J-S, Tailor A-RA, Lamki T, Zhang J, Ammirati M. Properties and Storage Methods of the Stratathane ST-504–Based Neurosurgical Tumor Model: Comprehensive Analysis. World Neurosurg. 2016;96:350-4.
15. Gragnaniello C, Gagliardi F, Chau AMT, Nader R, Siu A, Litvack Z, et al. Intracranial Injectable Tumor Model: Technical Advancements. Journal of Neurol Surg Part B. 2014;75(5):301-8.

16. Stewart DC, Rubiano A, Dyson K, Simmons CS. Mechanical characterization of human brain tumors from patients and comparison to potential surgical phantoms. PloS one. 2017;12(6):e0177561.

17. Lee KY, Mooney DJ. Alginate: Properties and biomedical applications. Progress in Polymer Science. 2012;37(1):106-26.

18. Li J, Weber E, Guth-Gundel S, Schuleit M, Kuttler A, Halleux C, et al. Tough Composite Hydrogels with High Loading and Local Release of Biological Drugs. Advanced Healthcare Materials. 2018;7(9):1701393.

19. Rowley JA, Madlambayan G, Mooney DJ. Alginate hydrogels as synthetic extracellular matrix materials. Biomaterials. 1999;20(1):45-53.

20. Zhao X, Huebsch N, Mooney DJ, Suo Z. Stress-relaxation behavior in gels with ionic and covalent crosslinks. J Appl Phys. 2010;107(6):63509-.

21. Garo A, Hrapko M, Van Dommelen J, Peters G. Towards a reliable characterisation of the mechanical behaviour of brain tissue: the effects of post-mortem time and sample preparation. Biorheology. 2007;44(1):51-8.

22. Schmidt MJ, Pilatus U, Wigger A, Kramer M, Oelschläger HA. Neuroanatomy of the calf brain as revealed by high-resolution magnetic resonance imaging. Journal of Morphology. 2009;270(6):745-58.

23. Gökyar A, Cokluk C. Using of Fresh Cadaveric Cow Brain in the Microsurgical Training Model for Sulcal-Cisternal and Fissural Dissection. Journal of neurosciences in rural practice. 2018;9(1):26-9.

24. Hicdonmez T, Hamamcioglu MK, Tiryaki M, Cukur Z, Cobanoglu S. Microneurosurgical training model in fresh cadaveric cow brain: a laboratory study simulating the approach to the circle of Willis. Surgical neurology. 2006;66(1):100-4; discussion 4.

25. Hicdonmez T, Hamamcioglu MK, Parsak T, Cukur Z, Cobanoglu S. A laboratory training model for interhemispheric-transcallosal approach to the lateral ventricle. Neurosurgical review. 2006;29(2):159-62.

26. Drouin S, Kochanowska A, Kersten-Oertel M, Gerard IJ, Zelmann R, De Nigris D, et al. IBIS: an OR ready open-source platform for image-guided neurosurgery. International journal of computer assisted radiology and surgery. 2017;12(3):363-78.

27. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J-C, Pujol S, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging. 2012;30(9):1323-41.

28. 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.

29. Alotaibi FE, AlZhrani GA, Mullah MA, Sabbagh AJ, Azarnoush H, Winkler-Schwartz A, et al. Assessing bimanual performance in brain tumor resection with NeuroTouch, a virtual reality simulator. Neurosurgery. 2015;11 Suppl 2:89-98; discussion

30. Winkler-Schwartz A, Bajunaid K, Mullah MA, Marwa I, Alotaibi FE, Fares J, et al. Bimanual Psychomotor Performance in Neurosurgical Resident Applicants Assessed Using NeuroTouch, a Virtual Reality Simulator. J Surg Educ. 2016;73(6):942-53.

31. 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. J Surg Educ. 2015;72(4):685-96.

32. Sawaya R, Bugdadi A, Azarnoush H, Winkler-Schwartz A, Alotaibi FE, Bajunaid K, et al. Virtual Reality Tumor Resection: The Force Pyramid Approach. Oper Neurosurg (Hagerstown). 2018;14(6):686-96.
33. Azarnoush H, Siar S, Sawaya R, Zhrani GA, Winkler-Schwartz A, Alotaibi FE, et al. The force pyramid: a spatial analysis of force application during virtual reality brain tumor resection. Journal of neurosurgery. 2017;127(1):171-81.
34. 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 Network Open. 2019;2(8):e198363-e.
35. Bissonnette V, Mirchi N, Ledwos N, Alsidieri G, Winkler-Schwartz A, Del Maestro RF, et al. Artificial Intelligence Distinguishes Surgical Training Levels in a Virtual Reality Spinal Task. JBJS. 2019;Latest Articles.
36. Siyar S, Azarnoush H, Rashidi S, Winkler-Schwartz A, Bissonnette V, Ponnudurai N, et al. Machine learning distinguishes neurosurgical skill levels in a virtual reality tumor resection task. Med Biol Eng Comput. 2020.
37. Delorme S, Laroche D, DiRaddo R, Del Maestro RF. NeuroTouch: a physics-based virtual simulator for cranial microneurosurgery training. Neurosurgery. 2012;71:ons32-ons42.
38. Sabbagh AJ, Bajunaid KM, Alarifi N, Winkler-Schwartz A, Alsideiri G, Al-Zhrani G, et al. Roadmap for Developing Complex Virtual Reality Simulation Scenarios: The Subpial Neurosurgical Tumor Resection Model. World Neurosurg. 2020.
39. Haji FA, Clarke DB, Matte MC, Brandman DM, Brien S, de Ribaupierre S, et al. Teaching for the Transition: the Canadian PGY-1 Neurosurgery 'Rookie Camp'. Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques. 2015;42(1):25-33.
40. Harrop J, Lobel DA, Bendok B, Sharan A, Rezai AR. Developing a neurosurgical simulation-based educational curriculum: an overview. Neurosurgery. 2013;73 Suppl 1:25-9.
41. Sadideen H, Kneebone R. Practical skills teaching in contemporary surgical education: how can educational theory be applied to promote effective learning? The American Journal of Surgery. 2012;204(3):396-401.
42. Regelsberger J. Surgery of the Brain and Spinal Cord in a Porcine Model2016. 165-73 p.
43. Regelsberger J, Eicker S, Siasios I, Hanggi D, Kirsch M, Horn P, et al. In vivo porcine training model for cranial neurosurgery. Neurosurgical review. 2015;38(1):157-63; discussion 63.
44. Aboud E, Al-Mefty O, Yaşargil MG. New laboratory model for neurosurgical training that simulates live surgery. Journal of neurosurgery. 2002;97(6):1367-72.
45. Sauleau P, Lapouble E, Val-Laillet D, Malbert CH. The pig model in brain imaging and neurosurgery. Animal. 2009;3(8):1138-51.
46. Duffau H, Peggy Gatignol ST, Mandonnet E, Capelle L, Taillandier L. Intraoperative subcortical stimulation mapping of language pathways in a consecutive series of 115 patients with Grade II glioma in the left dominant hemisphere. Journal of neurosurgery. 2008;109(3):461-71.

Chapter 6: Summary and conclusions

General findings

The overarching goal of my thesis is to better understand the role of simulation and artificial intelligence technology in the training of neurosurgical residents and to lay the groundwork for a future randomized controlled trial.

In Chapter 2 we reviewed the literature to develop consistent methodology and reporting guidelines in studies involving AI in surgical simulation. Firstly, a common language reduces the silo effect and improves the flow of information between research in medicine, computer science, and education. Secondly, guidelines highlight for reviewers from both medicine and computer science the pearls and pitfalls in each respective domain, and therefore whether conclusions drawn by authors are founded in proper methodology.

These guidelines were then applied in Chapter 3 where we demonstrated the feasibility of machine learning to assess surgical skills in neurosurgical simulation. At the time of publication, the methodology applied in this paper resulted in a greater degree of granularity in participant classification than had been reported in the literature until that point. In addition, improved categorization can be accomplished with only a handful of performance metrics, providing additional insight into expertise itself. Such technology may have direct implications for surgical education. Firstly, one can provide a participant with holistic feedback on a given performance, i.e. a senior year resident was categorized as a junior. Secondly, given such categorizations rely on a relatively small number of performance metrics, one can demonstrate to the trainee areas for targeted improvement, i.e. a participant was categorized as a junior because they exerted a high maximum force with their bipolar over the entire scenario. This concept underlies AI-powered feedback systems in VR neurosurgical simulation, such as the Virtual Operative Assistant and

Intelligent Continuous Expertise Monitoring System (83, 94). Such systems represent the experimental training arm in the RCT.

In Chapter 4 we sought to establish a reliable rating scale which can be used by evaluators to assess performance by observation alone. In the RCT, this provides a useful alternative measurement of expertise other than static outcomes such as brain tumor or brain volume removed during a surgery. While this raises an interesting question of whether a 'perfect outcome' performed 'imperfectly' is considered acceptable or not, in cases where the research environment is not within a live operating room, one cannot foresee which 'imperfect' techniques may translate to adverse consequences to patients and thus it is best to cast the net as wide as possible.

Finally, in chapter 5 we outlined a framework whereby the primary outcome of the RCT could be completed, notably the extent of resection of brain and tumor as assessed on 7-Tesla MRI. Further study has validated the 7-Tesla MRI as an accurate measurement tool in this context, with a near perfect correlation between predicted and actual weight of tissue resected (108). Furthermore, this study demonstrated the ability to capture movement during an artificial tumor resection task of real surgical instruments, and may allow for another means of evaluating participant improvement in the RCT.

The highly publicized existence of "bad apple" surgeons, such as "Dr. Death", the American neurosurgical spine surgeon imprisoned for misconduct (109), or two British pediatric cardiac surgeons suspended for an unacceptably high rate of complications (110), has done much to erode public trust in the ability of the medical profession to self-regulate and train its own. Indeed, surgical operative volume during residency may influence patient outcomes after residency (111), and as such, society has a stake in the adequate training of surgeons. Unlike other psychomotor activities such as athletic or musical endeavours, surgery with its interplay between declarative knowledge, technical expertise, and highly variable patient factors, represents one of the greatest challenges for educational reformists. One may reasonably wonder as to whether technological advances in simulation and artificial intelligence can be leveraged to augment the apprenticeship-based model. However, extreme care should be taken in introducing reforms in a system which, by and large, has produced largely competent surgeons.

Groups across multiple domains are liable to misapply these technologies. Clinicians may falsely assume that an algorithm that offers to discern residents by level of training may be appropriate to use to select medical student applicants to a residency program, while computer-scientists may oversimplify the surgical task being simulated, and in doing so miss crucial details impacting patient care. Hospital administrators may demand that surgeons demonstrate proficiency in simulated performance with little bearing in real life. Given this, research is necessary to provide a clear understanding of the benefits and limitations of the application of these technologies within medical education.

Health professions education has substantial roots in the social sciences, and thus asserts that inquiry be firmly grounded in “theory”. A theory is a conceptual framework meant to guide “health professions researchers and educators as they navigate the practical implications for teaching, learning, and research” (112). However, one challenge is that there are numerous theoretical frameworks applicable to a given context. In some cases, the underlying principles of these theories may be speculative or, if "evidence-based," may not be directly relevant to the context at hand (113). With a myriad of terms in use, clarifications are often necessary (114). This situation can potentially cause a disconnect between a study's design and conclusions from the clinical reality.

The introduction of data-centered science into health education has directly challenged the theory-first approach in health-sciences research. In the case of simulation research, data-sets may be analyzed without a priori assumptions. For example, in Chapter 3, the metrics used by the AI algorithm to group participants could not have been known to researchers. However, the metrics selected may provide a window into surgical expertise itself and may consequently advance theoretical frameworks on mastery. Additionally, as emphasized in Chapter 2, significant differences exist in how computer scientists and clinicians conduct and report AI-related research. If this gap remains unaddressed, it could hinder progress in the field. Establishing transparent norms can foster high-quality, reproducible studies, furthering educational theory and encouraging interdisciplinary collaboration.

Limitations

While the initial goal of the thesis was to both lay the foundation for, as well as conduct a randomized controlled trial, the COVID-19 pandemic represented an unforeseen limitation. The pandemic has significantly impacted the conduct of clinical research. The added constraints on the healthcare systems have diverted resources to more urgent clinical resources (115). Furthermore, physical distancing measures enforced by governments and healthcare units make the conduct of studies which require close contact difficult. However, an advantage of VR based instruction in the context of medical education studies is that physical distancing measures can be respected.

Future directions, the Randomized Controlled Trial

If completed, the randomized controlled trial may help to answer the effect of VR rehearsal on neurosurgical performance among neurosurgical residents. I have included its outline below as a template for future research.

Specific Aims and Hypothesis

Specific Aim 1: To demonstrate improvement in volume of brain and tumor resected, disruption of brain/tumor boundary, blood loss, and time of surgery from the bovine task after VR intervention. Hypothesis: We hypothesize that VR intervention will lead to improvement in clinically relevant operative results on an in-vivo tumor resection task.

Specific Aim 2: To demonstrate improvement in score obtained during the bovine tumor resection on the Global Rating Scale (GRS) of Operative Performance after VR intervention, as outlined in Chapter 4. Hypothesis: We hypothesize that VR intervention will lead to improvement in the GRS during an in-vivo tumor resection task.

Methods

Study Design: A single-blinded 2-armed randomized control trial will be conducted to assess the effect of VR training on the primary outcomes (volume of brain and tumor resected, disruption of brain/tumor boundary, blood loss, time of operation), and secondary outcome (Global Rating Scale of Operative Performance) during a bovine brain-tumor resection task. Study population: For the trial we will recruit 40 neurosurgical residents across 5 training sites, with 18 already located in the testing city. The remaining participants are within 3 hours driving distance from the testing centre. Furthermore, participant involvement will not extend beyond a Saturday to Sunday period to further decrease inconvenience and improve enrollment.

Study Protocol: Randomization: Participants will undergo randomization within a rolling admission period to the experimental (VR training) or control (standard residency training) group.

Control Group: The control group will undergo standard residency training, which includes formal face-to-face feedback sessions every 2 weeks. Experimental Group: The experimental group will be trained through VR simulation rehearsal, the basis of which was outlined in Chapter 3.

Animal Model: Both groups will perform a previously established tumor resection task (116) on a bovine animal model, outlined in Chapter 5, before and after the experimental intervention. Preoperative and postoperative magnetic resonance imaging will be conducted on the animal brains to assess residual brain tumor and resected brain volume. Identical surgical instruments as utilized in the VR training will be provided, namely bipolar electrocautery, ultrasonic aspirator and a suction device. The task will be considered complete when the resident indicates that they have safely removed all the tumor.

Primary Outcome: Volume of brain and tumor resected will be determined by evaluation of the pre-and post-operative MRI, and aspiration content examination. Disruption of brain/tumor boundary will be completed by examining post-operative MRI. Secondary Outcome. Video-recordings obtained from the operative microscope of the tumor resection task will be rated by two blinded evaluators utilizing a modified Global Rating Scale (GRS) of Operative Performance.

Inclusion criteria: Neurosurgical residents at all years of training on full-time clinical service.

Exclusion criteria: Residents enrolled in a non-neurosurgery rotation or having taken vacation within the study time.

Endpoints and follow-up: Each resident will complete a total of 4 animal-tumor resection tasks (two pre- and two post-experimental condition).

Statistical approach and sample size calculations: Descriptive statistics will be provided in means and standard deviation or median and inter-quartile range. Between group differences in continuous dependent variables will be compared using two-sample t-test or Mann-Whitney-Wilcoxon rank-sum test depending on normality. Sample size: For the primary outcome, a delta of 0.9 can be detected between groups, assuming a total sample size of 40 (20 control, 20 experimental) with 80% power, a two-sided alpha of 0.05 and a standard deviation of 1.

Unresolved Considerations

An important consideration is whether to utilize concurrent or post-hoc feedback systems in the experimental arm. As mentioned in Chapter 1, these systems may serve different learning objectives and it is unclear which one would be superior in the context of the outlined RCT. Portions of the surgery where the consequences for errors are dramatic and immediately apparent, such as an aneurysm rupture, may benefit from a continuous monitoring system which may anticipate an error, as demonstrated by the Intelligent Continuous Expertise Monitoring System which utilizes a long-short term memory network to analyze data in 0.2 s intervals (94). This contrasts with aspects of the surgery where the consequences may be less dire, such as opening skin or dura, which may be amenable to post-hoc assessments such as the Virtual Operative Assistant (83).

Clearly, "real" feedback by an observant clinical instructor incorporates elements of both; while on the spot feedback is used to correct egregious errors or to ensure proper technique ("hold your needle driver like so"), there is also opportunity to allow the trainee to complete a task with minimal interference ("perform a craniotomy"). Interestingly, a recent RCT comparing instructor feedback to post-hoc AI-based feedback in a VR neurosurgery tumor resection task demonstrated better performance in the AI based feedback group (95).

Future directions, the Intelligent Operative Theatre

The insights gained on simulated operative performance may have applicability to the development of "smart" operative instruments. The metrics which can best differentiate groups of expertise may guide the development of instruments, with the explicit goal of measuring these factors in the operative environment. For example, in Chapter 3, performance metrics selected by the algorithms spanned the following 4 principal domains: movement associated with a single instrument, both instruments used in concert, force applied by the instruments, and tissue removed, or bleeding caused. Four of the total 270 metrics were selected by at least two algorithms. Two of those four were related to force, and, as we recall, force was an element which was poorly captured by visual rating scales. Force is a particularly important factor in neurosurgery, given the delicate tissues of the nervous system. Studies using a force sensing bipolar in neurosurgery demonstrated increased complications with higher force variability (117, 118). This implies that providing adequate feedback of force applied during live surgeries may benefit surgeons and especially trainees. AI-powered feedback systems could process force data in real time and alert surgeons of an imminent error, as demonstrated with the Intelligent Continuous Expertise Monitoring System.

Technological advances in operative instruments have already demonstrated favorable safety benefits. One example is the introduction of the “tissue select” mode in ultrasonic aspirator system in brain tumor surgery. Tissue select allows for the fragmentation of tumors while leaving surrounding structures intact (119). This safety feature has already translated to clinical benefits in brain tumor surgery (120). Similarly, the introduction of intelligent operative instruments would be a powerful tool in the neurosurgeon’s armamentarium to deliver safe care to patients.

Conclusions

Computerized simulation technology and artificial intelligence systems represent the latest iteration in man’s quest to best understand and improve the world around him. These innovations introduce significant technological and philosophical disruptions to the otherwise conservative and judicious field of medicine.

Ultimately, neurosurgery aims to treat individuals who unfortunately have encountered a life-changing diagnosis, and given its inherently high-stakes nature, represents a litmus test for medicine at large. By demonstrating the communication gap between engineering and medicine in AI simulation research, using an AI algorithm to distinguish groups of expertise in neurosurgical simulation, uncovering the limitations of a visual feedback system for neurosurgical performance, and developing an artificial brain tumor and instrument movement system, it is my hope the benefits and limitations of these technologies’ applications in the field of neurosurgical education were further elaborated.

References

1. Donaldson MS, Corrigan JM, Kohn LT. To err is human: building a safer health system: National Academies Press; 2000.
2. Van Den Bos J, Rustagi K, Gray T, Halford M, Ziemkiewicz E, Shreve J. The $17.1 billion problem: the annual cost of measurable medical errors. Health Aff (Millwood). 2011;30(4):596-603.
3. Rogers SO, Gawande AA, Kwaan M, Puopolo AL, Yoon C, Brennan TA, et al. Analysis of surgical errors in closed malpractice claims at 4 liability insurers. Surgery. 2006;140(1):25-33.
4. Fabri PJ, Zayas-Castro JL. Human error, not communication and systems, underlies surgical complications. Surgery. 2008;144(4):557-63; discussion 63-5.
5. Fecso AB, Szasz P, Kerezov G, Grantcharov TP. The Effect of Technical Performance on Patient Outcomes in Surgery: A Systematic Review. Annals of Surgery. 2017;265(3):492-501.
6. Latimer K, Pendleton C, Olivi A, Cohen-Gadol AA, Brem H, Quiñones-Hinojosa A. Harvey Cushing's open and thorough documentation of surgical mishaps at the dawn of neurologic surgery. Archives of surgery. 2011;146(2):226-32.
7. Rabski JE, Saha A, Cusimano MD. Setting standards of performance expected in neurosurgery residency: A study on entrustable professional activities in competency-based medical education. The American Journal of Surgery. 2021;221(2):388-93.
8. Haji FA, Steven DA. Readiness for Practice: A Survey of Neurosurgery Graduates and Program Directors. Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques. 2014;41(6):721-8.
9. Kutikova L, Bowman L, Chang S, Long SR, Thornton DE, Crown WH. Utilization and cost of health care services associated with primary malignant brain tumors in the United States. J Neurooncol. 2007;81(1):61-5.
10. Stummer W, Meinel T, Ewelt C, Martus P, Jakobs O, Felsberg J, et al. Prospective cohort study of radiotherapy with concomitant and adjuvant temozolomide chemotherapy for glioblastoma patients with no or minimal residual enhancing tumor load after surgery. J Neurooncol. 2012;108(1):89-97.
11. Chambless LB, Kistka HM, Parker SL, Hassam-Malani L, McGirt MJ, Thompson RC. The relative value of postoperative versus preoperative Karnofsky Performance Scale scores as a predictor of survival after surgical resection of glioblastoma multiforme. J Neurooncol. 2015;121(2):359-64.
12. Hebb AO, Yang T, Silbergeld DL. The sub-pial resection technique for intrinsic tumor surgery. Surg Neurol Int. 2011;2:180.
13. Stone S, Bernstein M. Prospective error recording in surgery: an analysis of 1108 elective neurosurgical cases. Neurosurgery. 2007;60(6):1075-80; discussion 80-2.
14. Rolston JD, Zygourakis CC, Han SJ, Lau CY, Berger MS, Parsa AT. Medical errors in neurosurgery. Surg Neurol Int. 2014;5(Suppl 10):S435-40.
15. Gozal YM, Aktüre E, Ravindra VM, Scoville JP, Jensen RL, Couldwell WT, et al. Defining a new neurosurgical complication classification: lessons learned from a monthly Morbidity and Mortality conference. Journal of Neurosurgery JNS. 2020;132(1):272.
16. Boström J, Yacoub A, Schramm J. Prospective collection and analysis of error data in a neurosurgical clinic. Clin Neurol Neurosurg. 2010;112(4):314-9.

17. Long DM. Competency based residency training: the next advance in graduate medical education. Acta Neurochir Suppl. 2001;78:153-8.
18. Dreyfus SE. The five-stage model of adult skill acquisition. Bulletin of science, technology & society. 2004;24(3):177-81.
19. Baisiwala S, Shlobin NA, Cloney MB, Dahdaleh NS. Impact of Resident Participation During Surgery on Neurosurgical Outcomes: A Meta-Analysis. World Neurosurg. 2020;142:1-12.
20. Gupta R, Moore JM, Adeeb N, Griessenauer CJ, Schneider AM, Gandhi CD, et al. Neurosurgical Resident Error: A Survey of U.S. Neurosurgery Residency Training Program Directors' Perceptions. World Neurosurg. 2018;109:e563-e70.
21. Vasella F, Velz J, Neidert MC, Henzi S, Sarnthein J, Krayenbühl N, et al. Safety of resident training in the microsurgical resection of intracranial tumors: Data from a prospective registry of complications and outcome. Sci Rep. 2019;9(1):954.
22. Sugiyama T, Lama S, Gan L, Maddahi Y, Zareinia K, Sutherland GR. Forces of tool-tissue interaction to assess surgical skill level. JAMA Surg. 2018;153(3):234-42.
23. Cernak I. Animal models of head trauma. NeuroRx : the journal of the American Society for Experimental NeuroTherapeutics. 2005;2(3):410-22.
24. Verma R, Virdi JK, Singh N, Jaggi AS. Animals models of spinal cord contusion injury. Korean J Pain. 2019;32(1):12-21.
25. Potts JR, 3rd. Assessment of Competence: The Accreditation Council for Graduate Medical Education/Residency Review Committee Perspective. Surg Clin North Am. 2016;96(1):15-24.
26. Van Melle E, Frank J, Brzeznia S, Gorman L. Competency by Design-Residency Education: A Framework for Program Evaluation. DRAFT Ottawa, Ontario, Canada: Royal College of Physicians and Surgeons of Canada. 2017.
27. Yudkowsky R, Park YS, Lineberry M, Knox A, Ritter EM. Setting mastery learning standards. Academic Medicine. 2015;90(11):1495-500.
28. Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. The British journal of surgery. 1997;84(2):273-8.
29. Ahmed K, Miskovic D, Darzi A, Athanasiou T, Hanna GB. Observational tools for assessment of procedural skills: a systematic review. American journal of surgery. 2011;202(4):469-80.e6.
30. Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): a tool to assess surgical competence. Academic Medicine. 2012;87(10):1401-7.
31. Niitsu H, Hirabayashi N, Yoshimitsu M, Mimura T, Taomoto J, Sugiyama Y, et al. Using the Objective Structured Assessment of Technical Skills (OSATS) global rating scale to evaluate the skills of surgical trainees in the operating room. Surgery Today. 2013;43(3):271-5.
32. Winkler-Schwartz A, Marwa I, Bajunaid K, Mullah M, Alotaibi FE, Bugdadi A, et al. A Comparison of Visual Rating Scales and Simulated Virtual Reality Metrics in Neurosurgical Training: A Generalizability Theory Study. World Neurosurg. 2019.
33. Kshettry VR, Mullin JP, Schlenk R, Recinos PF, Benzel EC. The role of laboratory dissection training in neurosurgical residency: results of a national survey. World Neurosurg. 2014;82(5):554-9.

34. Gnanakumar S, Kostusiak M, Budohoski KP, Barone D, Pizzuti V, Kirollos R, et al. Effectiveness of Cadaveric Simulation in Neurosurgical Training: A Review of the Literature. World Neurosurg. 2018;118:88-96.
35. Olabe J, Olabe J, Sancho V. Human cadaver brain infusion model for neurosurgical training. Surgical neurology. 2009;72(6):700-2.
36. Zada G, Bakhsheshian J, Pham M, Minneti M, Christian E, Winer J, et al. Development of a Perfusion-Based Cadaveric Simulation Model Integrated into Neurosurgical Training: Feasibility Based On Reconstitution of Vascular and Cerebrospinal Fluid Systems. Oper Neurosurg. 2018;14(1):72-80.
37. Benet A, Rincon-Torroella J, Lawton MT, Gonzalez Sanchez JJ. Novel embalming solution for neurosurgical simulation in cadavers. Journal of neurosurgery. 2014;120(5):1229-37.
38. Morosanu CO, Nicolae L, Moldovan R, Farcasanu AS, Filip GA, Florian IS. Neurosurgical cadaveric and in vivo large animal training models for cranial and spinal approaches and techniques - a systematic review of the current literature. Neurol Neurochir Pol. 2019;53(1):8-17.
39. Janowski M. Experimental Neurosurgery in Animal Models. In: Janowski M, editor. Neuromethods Humana Press; 2016. p. 69-89.
40. Grisham W. Resources for teaching Mammalian neuroanatomy using sheep brains: a review. Journal of undergraduate neuroscience education : JUNE : a publication of FUN, Faculty for Undergraduate Neuroscience. 2006;5(1):R1-R6.
41. Hoffmann A, Stoffel MH, Nitzsche B, Lobsien D, Seeger J, Schneider H, et al. The Ovine Cerebral Venous System: Comparative Anatomy, Visualization, and Implications for Translational Research. PloS one. 2014;9(4):e92990.
42. Habib CA, Utriainen D, Peduzzi-Nelson J, Dawe E, Mattei J, Latif Z, et al. MR imaging of the yucatan pig head and neck vasculature. J Magn Reson Imaging. 2013;38(3):641-9.
43. Sauleau P, Lapouble E, Val-Laillet D, Malbert CH. The pig model in brain imaging and neurosurgery. Animal. 2009;3(8):1138-51.
44. Schmidt V. Comparative anatomy of the pig brain: an integrative magnetic resonance imaging (MRI) study of the porcine brain with special emphasis on the external morphology of the cerebral cortex. 2015.
45. Weickenmeier J, Kurt M, Ozkaya E, Wintermark M, Pauly KB, Kuhl E. Magnetic resonance elastography of the brain: A comparison between pigs and humans. Journal of the Mechanical Behavior of Biomedical Materials. 2018;77:702-10.
46. Schmidt MJ, Pilatus U, Wigger A, Kramer M, Oelschläger HA. Neuroanatomy of the calf brain as revealed by high-resolution magnetic resonance imaging. Journal of Morphology. 2009;270(6):745-58.
47. Zdun M, Frąckowiak H, Kiełtyka-Kurc A, Kowalczyk K, Nabzdyk M, Timm A. The Arteries of Brain Base in Species of Bovini Tribe. The Anatomical Record. 2013;296(11):1677-82.
48. Alsayegh A, Bakhaidar M, Winkler-Schwartz A, Yilmaz R, Del Maestro RF. Best Practices Using Ex Vivo Animal Brain Models in Neurosurgical Education to Assess Surgical Expertise. World Neurosurg. 2021.
49. Gélinas-Phaneuf N, Del Maestro RF. Surgical Expertise in Neurosurgery: Integrating Theory Into Practice. Neurosurgery. 2013;73(supplement 1):S30-S8 10.1227/NEU.0000000000000115.

50. Marcus H, Vakharia V, Kirkman MA, Murphy M, Nandi D. Practice makes perfect? The role of simulation-based deliberate practice and script-based mental rehearsal in the acquisition and maintenance of operative neurosurgical skills. Neurosurgery. 2013;72 Suppl 1:124-30.
51. Zaed I, Tinterri B. Letter to the Editor: How is COVID-19 Going to Affect Education in Neurosurgery? A Step Toward a New Era of Educational Training. World Neurosurg. 2020;140:481-3.
52. Mirchi N, Ledwos N, Del Maestro RF. Intelligent Tutoring Systems: Re-Envisioning Surgical Education in Response to COVID-19. Can J Neurol Sci. 2020:1-3.
53. Maran NJ, Glavin RJ. Low- to high-fidelity simulation - a continuum of medical education? Medical education. 2003;37 Suppl 1:22-8.
54. van Merriënboer JJ, Sweller J. Cognitive load theory in health professional education: design principles and strategies. Medical education. 2010;44(1):85-93.
55. Chan J, Pangal DJ, Cardinal T, Kugener G, Zhu Y, Roshannai A, et al. A systematic review of virtual reality for the assessment of technical skills in neurosurgery. Neurosurg Focus. 2021;51(2):E15.
56. Marinkovic D, Zehn M. Survey of finite element method-based real-time simulations. Applied Sciences. 2019;9(14):2775.
57. Delorme S, Laroche D, DiRaddo R, Del Maestro RF. NeuroTouch: a physics-based virtual simulator for cranial microneurosurgery training. Neurosurgery. 2012;71(1 Suppl Operative):32-42.
58. 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. 2020.
59. Levine AI, DeMaria Jr S, Schwartz AD, Sim AJ. The comprehensive textbook of healthcare simulation: Springer Science & Business Media; 2013.
60. Rosseau G, Bailes J, del Maestro R, Cabral A, Choudhury N, Comas O, et al. The Development of a Virtual Simulator for Training Neurosurgeons to Perform and Perfect Endoscopic Endonasal Transsphenoidal Surgery. Neurosurgery. 2013;73(supplement 1):S85-S93 10.1227/NEU.0000000000000112.
61. Varshney R, Frenkiel S, Nguyen LH, Young M, Del Maestro R, Zeitouni A, et al. Development of the McGill simulator for endoscopic sinus surgery: a new high-fidelity virtual reality simulator for endoscopic sinus surgery. Am J Rhinol Allergy. 2014;28(4):330-4.
62. Breimer GE, Haji FA, Bodani V, Cunningham MS, Lopez-Rios A-L, Okrainec A, et al. Simulation-based Education for Endoscopic Third Ventriculostomy: A Comparison Between Virtual and Physical Training Models. Oper Neurosurg. 2016;13(1):89-95.
63. Lefor AK, Harada K, Kawahira H, Mitsuishi M. The effect of simulator fidelity on procedure skill training: a literature review. Int J Med Educ. 2020;11:97-106.
64. Norman G, Dore K, Grierson L. The minimal relationship between simulation fidelity and transfer of learning. Medical education. 2012;46(7):636-47.
65. 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). Surg Innov. 2015;22(6):636-42.
66. Alotaibi FE, AlZhrani GA, Mullah MA, Sabbagh AJ, Azarnoush H, Winkler-Schwartz A, et al. Assessing bimanual performance in brain tumor resection with NeuroTouch, a virtual reality simulator. Neurosurgery. 2015;11 Suppl 2:89-98; discussion

67. Bugdadi A, Sawaya R, Olwi D, Al-Zhrani G, Azarnoush H, Sabbagh AJ, et al. Automaticity of Force Application During Simulated Brain Tumor Resection: Testing the Fitts and Posner Model. J Surg Educ. 2017.

68. Winkler-Schwartz A, Bajunaid K, Mullah MA, Marwa I, Alotaibi FE, Fares J, et al. Bimanual Psychomotor Performance in Neurosurgical Resident Applicants Assessed Using NeuroTouch, a Virtual Reality Simulator. J Surg Educ. 2016.

69. Sawaya R, Alsideiri G, Bugdadi A, Winkler-Schwartz A, Azarnoush H, Bajunaid K, et al. Development of a performance model for virtual reality tumor resections. Journal of neurosurgery. 2018;1(aop):1-9.

70. 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. J Surg Educ. 2015;72(4):685-96.

71. Buchanan BG. A (very) brief history of artificial intelligence. Ai Magazine. 2005;26(4):53-.

72. Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Computers and biomedical research. 1975;8(4):303-20.

73. Peek N, Combi C, Marin R, Bellazzi R. Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes. Artif Intell Med. 2015;65(1):61-73.

74. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RCT, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Medicine. 2002;8(1):68-74.

75. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol. 2018;6(5):361-9.

76. Shahid N, Rappon T, Berta W. Applications of artificial neural networks in health care organizational decision-making: A scoping review. PloS one. 2019;14(2):e0212356-e.

77. Tolsgaard MG, Boscardin CK, Park YS, Cuddy MM, Sebok-Syer SS. The role of data science and machine learning in Health Professions Education: practical applications, theoretical contributions, and epistemic beliefs. Adv Health Sci Educ Theory Pract. 2020;25(5):1057-86.

78. 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.

79. 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.

80. 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.

81. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019;1(5):206-15.

82. 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).

83. 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.

84. Dedeilia A, Sotiropoulos MG, Hanrahan JG, Janga D, Dedeilias P, Sideris M. Medical and Surgical Education Challenges and Innovations in the COVID-19 Era: A Systematic Review. In Vivo. 2020;34(3 Suppl):1603-11.

85. Tomlinson SB, Hendricks BK, Cohen-Gadol AA. Editorial. Innovations in neurosurgical education during the COVID-19 pandemic: is it time to reexamine our neurosurgical training models? Journal of neurosurgery. 2020:1-2.

86. Haji FA. Simulation in Neurosurgical Education During the COVID-19 Pandemic and Beyond. Can J Neurol Sci. 2021;48(2):152-4.

87. Zaed I, Tinterri B. Letter to the Editor: How is COVID-19 Going to Affect Education in Neurosurgery? A Step Toward a New Era of Educational Training. World Neurosurg. 2020;140:481-3.

88. Kilgore MD, Scullen T, Mathkour M, Dindial R, Carr C, Zeoli T, et al. Effects of the COVID-19 Pandemic on Operative Volume and Residency Training at Two Academic Neurosurgery Centers in New Orleans. World Neurosurg. 2021;151:e68-e77.

89. Goyal N, Venkataram T, Singh V, Chaturvedi J. Collateral damage caused by COVID-19: Change in volume and spectrum of neurosurgery patients. J Clin Neurosci. 2020;80:156-61.

90. Pelargos PE, Chakraborty A, Zhao YD, Smith ZA, Dunn IF, Bauer AM. An Evaluation of Neurosurgical Resident Education and Sentiment During the Coronavirus Disease 2019 Pandemic: A North American Survey. World Neurosurg. 2020;140:e381-e6.

91. Zoia C, Raffa G, Somma T, Della Pepa GM, La Rocca G, Zoli M, et al. COVID-19 and neurosurgical training and education: an Italian perspective. Acta Neurochir (Wien). 2020;162(8):1789-94.

92. El-Ghandour NMF, Ezzat AAM, Zaazoue MA, Gonzalez-Lopez P, Jhawar BS, Soliman MAR. Virtual learning during the COVID-19 pandemic: a turning point in neurosurgical education. Neurosurgical Focus FOC. 2020;49(6):E18.

93. Lefroy J, Watling C, Teunissen PW, Brand P. Guidelines: the do's, don'ts and don't knows of feedback for clinical education. Perspect Med Educ. 2015;4(6):284-99.

94. 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.

95. 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.

96. Mirchi N, Ledwos N, Del Maestro RF. Intelligent Tutoring Systems: Re-Envisioning Surgical Education in Response to COVID-19. Can J Neurol Sci. 2021;48(2):198-200.

97. Patel EA, Aydin A, Cearns M, Dasgupta P, Ahmed K. A Systematic Review of Simulation-Based Training in Neurosurgery, Part 2: Spinal and Pediatric Surgery, Neurointerventional Radiology, and Nontechnical Skills. World Neurosurg. 2020;133:e874-e92.

98. Patel EA, Aydin A, Cearns M, Dasgupta P, Ahmed K. A Systematic Review of Simulation-Based Training in Neurosurgery, Part 1: Cranial Neurosurgery. World Neurosurg. 2020;133:e850-e73.

99. Cox T, Seymour N, Stefanidis D. Moving the needle: simulation's impact on patient outcomes. Surgical Clinics. 2015;95(4):827-38.
100. Brydges R, Hatala R, Zendejas B, Erwin PJ, Cook DA. Linking simulation-based educational assessments and patient-related outcomes: a systematic review and meta-analysis. Academic Medicine. 2015;90(2):246-56.
101. Ganju A, Aoun SG, Daou MR, El Ahmadieh TY, Chang A, Wang L, et al. The role of simulation in neurosurgical education: a survey of 99 United States neurosurgery program directors. World Neurosurg. 2013;80(5):e1-8.
102. Davids J, Manivannan S, Darzi A, Giannarou S, Ashrafian H, Marcus HJ. Simulation for skills training in neurosurgery: a systematic review, meta-analysis, and analysis of progressive scholarly acceptance. Neurosurgical review. 2020.
103. Yudkowsky R, Luciano C, Banerjee P, Schwartz A, Alaraj A, Lemole GMJ, et al. Practice on an Augmented Reality/Haptic Simulator and Library of Virtual Brains Improves Residents' Ability to Perform a Ventriculostomy. Simulation in Healthcare. 2013;8(1):25-31.
104. Chugh AJ, Pace JR, Singer J, Tatsuoka C, Hoffer A, Selman WR, et al. Use of a surgical rehearsal platform and improvement in aneurysm clipping measures: results of a prospective, randomized trial. 2017;126(3):838.
105. Oliveira MM, Ferrarez CE, Lovato R, Costa PV, Malheiros JA, Avellar L, et al. Quality assurance during brain aneurysm microsurgery—operative error teaching. World Neurosurg. 2019;130:e112-e6.
106. Wulf G, Shea C, Lewthwaite R. Motor skill learning and performance: a review of influential factors. Medical education. 2010;44(1):75-84.
107. Mentis HM, Chellali A, Manser K, Cao CGL, Schwaitzberg SD. A systematic review of the effect of distraction on surgeon performance: directions for operating room policy and surgical training. Surgical endoscopy. 2016;30(5):1713-24.
108. Tran DH, Winkler-Schwartz A, Tuznik M, Gueziri H-E, Rudko DA, Reich A, et al. Quantitation of Tissue Resection Using a Brain Tumor Model and 7-T Magnetic Resonance Imaging Technology. World Neurosurg. 2021;148:e326-e39.
109. Chapman JR, Wang JC, Wiechert K. How Should We Deal With "Black Swan" Surgeons in Spine Surgery? Global Spine Journal. 2019;9(4):365-7.
110. Wisheart JD, Dhasmana JP. The Bristol Affair: lessons to be learned. The Annals of thoracic surgery. 2001;71(4):1403-4.
111. Mowat A, Maher C, Ballard E. Surgical outcomes for low-volume vs high-volume surgeons in gynecology surgery: a systematic review and meta-analysis. American journal of obstetrics and gynecology. 2016;215(1):21-33.
112. Samuel A, Konopasky A, Schuwirth LWT, King SM, Durning SJ. Five Principles for Using Educational Theory: Strategies for Advancing Health Professions Education Research. Academic Medicine. 2020;95(4):518-22.
113. Colliver JA. Educational Theory and Medical Education Practice: A Cautionary Note for Medical School Faculty. Academic Medicine. 2002;77(12 Part 1):1217-20.
114. Varpio L, Paradis E, Uijtdehaage S, Young M. The Distinctions Between Theory, Theoretical Framework, and Conceptual Framework. Academic Medicine. 2020;95(7):989-94.
115. Sohrabi C, Mathew G, Franchi T, Kerwan A, Griffin M, Soleil C Del Mundo J, et al. Impact of the coronavirus (COVID-19) pandemic on scientific research and implications for clinical academic training – A review. Int J Surg. 2021;86:57-63.

116. Regelsberger J, Eicker S, Siasios I, Hanggi D, Kirsch M, Horn P, et al. In vivo porcine training model for cranial neurosurgery. Neurosurgical review. 2015;38(1):157-63; discussion 63.

117. Albakr A, Baghdadi A, Singh R, Lama S, Sutherland GR. Tool-Tissue Forces in Hemangioblastoma Surgery. World Neurosurg. 2022;160:e242-e9.

118. Baghdadi A, Lama S, Singh R, Hoshyarmanesh H, Razmi M, Sutherland GR. A data-driven performance dashboard for surgical dissection. Sci. 2021;11(1):15013.

119. Jallo GI. CUSA EXcel ultrasonic aspiration system. Neurosurgery. 2001;48(3):695-7.

120. Tang H, Zhang H, Xie Q, Gong Y, Zheng M, Wang D, et al. Application of CUSA Excel ultrasonic aspiration system in resection of skull base meningiomas. Chin J Cancer Res. 2014;26(6):653-7.