Global Surgical Education - Journal of the Association for Surgical Education (2024) 3:83
https://doi.org/10.1007/s44186-024-00284-z

REVIEW

The role of artificial intelligence, performance metrics, and virtual reality in neurosurgical education: an umbrella review

Jason M. Harley[1,2,3,4] · Tiah Tawakol[1] · Sayed Azher[1] · Andrea Quaiattini[3,5] · Rolando Del Maestro[1,6]

Received: 14 December 2023 / Revised: 12 July 2024 / Accepted: 31 July 2024 / Published online: 10 August 2024

Abstract

Purpose Virtual reality (VR) and artificial intelligence (AI) play a pivotal role in surgical education. VR is a transformative tool, providing a low-risk environment for trainees to hone their skills. AI, in conjunction with VR, holds promise in addressing contemporary challenges in surgical education by mitigating human bias in evaluations, identifying at-risk residents, and enhancing surgical training through AI-produced performance metrics. Due to the rapid growth of literatures in VR and AI, it is essential to systematically evaluate and reflect on the current evidence regarding AI's role in assessing trainees' performance.
Method As such, an umbrella review was conducted focused on neurosurgery due to its high-stakes nature and active literature. We synthesize evidence from four systematic reviews and meta-analyses.
Results Key findings reveal that AI metrics effectively measure neurosurgical performance in VR environments, with a focus on time, kinematics, and force as dominant metrics. Moreover, the integration of AI-enhanced VR demonstrates potential in addressing critical challenges faced by surgical educators, including a shortage of surgeons and financial constraints. The review underscores the need for a more unified understanding of metrics and recommends further research to explore non-technical skills and the delivery of personalized feedback in AI-enhanced VR settings. As the field matures, the exploration of virtual assets, such as digital twins or other atypical patient case scenarios, presents a promising avenue for a more comprehensive and diverse range of training experiences in neurosurgery.
Conclusions Ultimately, this review outlines a bright future for the synergistic application of VR and AI metrics in neurosurgical education, with untapped potential in underexplored areas and a trajectory towards real-world implementation through both VR and Augmented Reality.

Keywords Neurosurgery · Surgical education · Simulation · Artificial intelligence · Machine learning · Virtual reality

Introduction and rationale

Medical and surgical education face a myriad of contemporary challenges that artificial intelligence (AI) stands to help address by reducing or removing human bias from instructor evaluations, identifying at-risk residents, improving written documents, supporting clinical decision-making, and helping trainees prepare for exams [1–3]. We define AI in surgical education as an intelligent system/program that acts to fulfill or support the fulfillment of educational tasks traditionally performed exclusively by surgical educators by making decisions in a manner similar to educators and

✉ Jason M. Harley
jason.harley@mcgill.ca

[1] Department of Surgery, Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada

[2] Research Institute of the McGill University Health Centre, Montreal, QC, Canada

[3] Institute of Health Sciences Education, Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada

[4] Steinberg Centre for Simulation and Interactive Learning, Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada

[5] Schulich Library of Physical Sciences, Life Sciences, and Engineering, McGill University, Montreal, QC, Canada

[6] Neurosurgical Simulation and Artificial Intelligence Learning Centre, Department of Neurology and Neurosurgery, Montreal Neurological Institute and Hospital, McGill University, Montreal, QC, Canada

providing customized adaptation, including performance assessment, personalized feedback, and error mitigation to surgical learners [4]. This definition differentiates between rule-based and non-rule based AI, where AI systems are given explicit rules programmed by humans in the former (e.g., using decision trees), but not in the latter (e.g., unsupervised machine learning) [4]. In other words, an AI is explicitly told when and how to respond to a medical resident's attempts to resect a brain tumor based on predetermined scoring for incoming instrument handling data in rule-based AI but must determine how to respond in non-rule-based AI. This means that when we speak about AI in surgical education, we can actually be referring to an AI that is the product of both rule and non-rule-based AI. For example, the Virtual Operative Assistant estimates a competence percentage score and binary expertise classification based on four metrics evaluated in two steps (step 1, safety: mean bleeding rate and maximum bipolar; step 2, instrument movement: mean instrument tip separation distance and mean bipolar) [5]. These four metrics were selected through both non-rule-based AI (machine learning: statistical, forward, and backward support vector machine feature selection) and human consultation [6]. Rule-based AI is sufficient, however, to examine the identified metrics and their thresholds and provide learners with automated audiovisual feedback and the right corrective instruction.

AI is a rapidly advancing technology and a topic of increasing focus in surgical education research as educators, researchers, and trainees seek to harness its potential to improve training and patient care. In the last five years, a growing body of literatures has supported the use of AI in surgical training programs [7]. AI in surgical education has shown to be highly accurate in characterizing surgical skill sets [1] and differentiating trainee expertise using various metrics. Performance metrics in surgical education are a set of measurements by which a plan or process can be assessed and that quantifies these elements of performance [8]. A recent review [7] found that the majority of surgical education studies that used AI took place in simulations: a focus of the current article.

Simulations have been widely used in various fields, including aviation and the military, to train and certify individuals in their respective professions [10]. In-person simulations are frequently used to train and evaluate healthcare learners and often involve actors role-playing to replicate real-life medical scenarios [11]. The goal of simulation is to engage learners and increase their preparedness for treating patients in a low-risk environment. Learners can practice tasks repeatedly and receive feedback on their performance in simulations, allowing them to make mistakes and learn from them [12]. Simulation-based education is known to be expensive, with personnel and equipment resources being the primary costs [13]. As such, this type of simulation may be a less viable option in inadequately funded healthcare systems and pose difficulty in its scalability. Simulations have been widely used in various fields, including aviation and the military, to train and certify individuals in their respective professions [9]. In-person simulations are frequently used to train and evaluate healthcare learners and often involve actors role-playing to replicate real-life medical scenarios [10]. The goal of simulation is to engage learners and increase their preparedness for treating patients in a low-risk environment. Learners can practice tasks repeatedly and receive feedback on their performance in simulations, allowing them to make mistakes and learn from them [11]. Simulation-based education is known to be expensive, with personnel and equipment resources being the primary costs [12]. As such, this type of simulation may be a less viable option in inadequately funded healthcare systems and pose difficulty in its scalability.

AI-enhanced virtual simulations are a promising application of AI in surgical education that leverages many of AI's strengths, such as its potential to provide immediate, personalized feedback, with the potential scalability and flexibility of virtual simulations (VS). VS are software applications that can run on computers or head-mounted displays [10] and provide surgical trainees the opportunity to practice skills in a virtual environment using interactive digital spaces, tools, and characters, including AI-driven patients and healthcare professionals, to mimic elements of the real world [13]. As with AI, there is growing interest to integrate VS into surgical training [14]. While many researchers distinguish VS as the most general category of such environments, virtual reality (VR) has also been broadly defined in the literature as encompassing simulators such as screen-based VR simulators (i.e., mixed or lightly augmented reality), screen-based virtual worlds (game-like environments), and immersive VR environments (often requiring a headset)—all of which intend to simulate natural settings [15, 16]. For the purposes of this article and conventions in the surgical education literature, we will adopt VR as our general term to refer to such educational technologies for surgical training. VR in surgical training has shown to be a promising approach to surgical training, supplementing traditional models and allowing trainees to acquire both technical and non-technical skills [17]. Studies in the field of orthopedic surgery show that VR training can be more effective than traditional training methods, suggesting it improves surgical skill acquisition and achieves shorter procedure times [18, 19]. In a randomized controlled trial conducted on obstetrics and gynecology (OB/GYN) trainees performing a laparoscopic procedure, the group that received 6 h of VR training prior to being in the operating room saw superior skill transfer and significantly shorter operating times compared to the control group that received no VR training [20]. This subsequently led to Denmark implementing VR laparoscopic training requirements for all OB/GYN residents [17].

AI-enhanced VR is not only a promising educational technology for surgical education, but a tool with potential to help address pressing challenges in surgical education, such as a dire shortage of surgeons [21, 22] which not only persists but is increasing in urgency [23, 24]. A reduction of residents' work hours, a lack of educators available to provide them with valuable operating room time, and financial restrictions on surgical training programs are additional challenges that call for innovation in surgical education [25]. AI-enhanced VR has the potential to help address these and other challenges with flexibility, scalability, and personalized performance feedback.

Objectives

While research on VR and AI rapidly proliferate, there is a crucial need to systematically assess and critically reflect on current evidence on the use of AI to provide metrics on trainees' performance. We elected to examine this topic in the context of neurosurgery because it is a discipline with an active literature on the topic and is an especially high-stakes field that can benefit from enhanced training programs with feedback and personalized evaluations [26, 27]. A preliminary search of the literature helped us identify more than one recent systematic review on our topic—indeed, one was published the very year we conducted our search [28]. Therefore, rather than duplicate efforts, we decided to conduct an umbrella review to identify potential additional prior meta-analyses, systematic reviews, and scoping reviews and synthesize findings to provide a comprehensive overview of ongoing efforts and available evidence. Umbrella reviews are overarching reviews that aggregate findings from several reviews that address specific questions and are therefore best suited for topics which are already addressed in systematic and/or meta-analyses [29, 30]. An umbrella review allowed us to explore the literature from a different lens than any of the previous individual systematic reviews and ask different research questions, including the following:

RQ1: How is neurosurgical performance being measured in VR environments?

RQ2: What does (a) each review conclude about the efficacy of performance metrics and (b) what can be concluded collectively?

RQ3: What is the role of AI in training and assessment?

Methods

Search strategy

A search was conducted in September 2023 to identify systematic reviews, scoping reviews and meta-analyses that reported on the use of VR environments and AI in neurosurgical education. The search was structured around the concepts of AI and neurosurgery, a full search strategy can be found in Online Resource materials. A combination of database-specific publication filters and keywords were applied to limit to the search to reviews. The databases searched were Ovid MEDLINE, Ovid Embase Classic + Embase and Scopus. The database time coverage was between 1946 (inception) to September 20, 2023. A manual search was also conducted by looking up keywords in PubMed.

Inclusion and exclusion criteria

A list of inclusion criteria was generated by researchers. To be included in the umbrella review, studies needed to meet all the below criteria or they were excluded from consideration (e.g., studies would be excluded if they examined augmented reality or orthopedic surgery).

1. Be in English
2. Examine virtual reality
3. Directly or indirectly include artificial intelligence
4. Mention neurosurgery and neurosurgical procedures
5. Provide an assessment of surgical skills
6. Be a systematic review, scoping review, or meta-analyses as defined by the authors

Systematic review selection

An initial subset of articles was screened independently by both reviewers and inter-rater agreement was calculated. Inter-rater agreement was calculated to be 93.88% and was deemed acceptable [32]. All identified reviews were uploaded into EndNote. Two reviewers conducted the screening process for all yielded studies. An initial subset of articles was screened independently by both reviewers and inter-rater agreement was calculated. Inter-rater agreement was calculated to be 93.88% and was deemed acceptable [31]. All disagreements were addressed and resolved through verbal deliberation.

Results

The search yielded a total of 195 articles with duplicates. A manual search was performed, and an additional four articles were identified. Once duplicates were removed, a total of 145 records remained. These records were screened using the abovementioned inclusion and exclusion criteria and 139 records were excluded. The remaining six records underwent full-text screening and were assessed for eligibility. After full-text screening, two records were excluded, leaving four systematic review articles in the final qualitative synthesis (see Online Resource Materials for study quality assessment). One systematic review article was excluded because it focused on augmented reality, not VR [32]. A second article was excluded because it focused on orthopedics, with limited information relating to neurosurgical use of VR [33].

See Fig. 1 for the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) flow diagram [34] (Fig. 1).

Data were extracted into an Excel table. See Online Resource Material for the data extraction sheet headings. The quality of each review was rated independently by two researchers using the Critical Appraisal Skills Program (CASP) checklist for systematic reviews [35]. This checklist evaluates the validity of the results of each review and its quality. It looks at research questions, papers included in the systematic review, authors' quality assessment of studies, the precision of results, and how results can be applied to help the local population. The CASP checklist was completed for each study. Regardless of their quality assessment, all studies were included in the final qualitative synthesis. See Online Resource Materials for Critical Appraisal Skills Program (CASP) checklist results for each included systematic review. Furthermore, an evaluation of included reviews' data collection methodologies and risk of bias assessment use was conducted according to the PRIOR guidelines [36]. Results of these are shown below in Table 1. Responses to the full PRIOR guidelines can be found in the Online Resource Material.

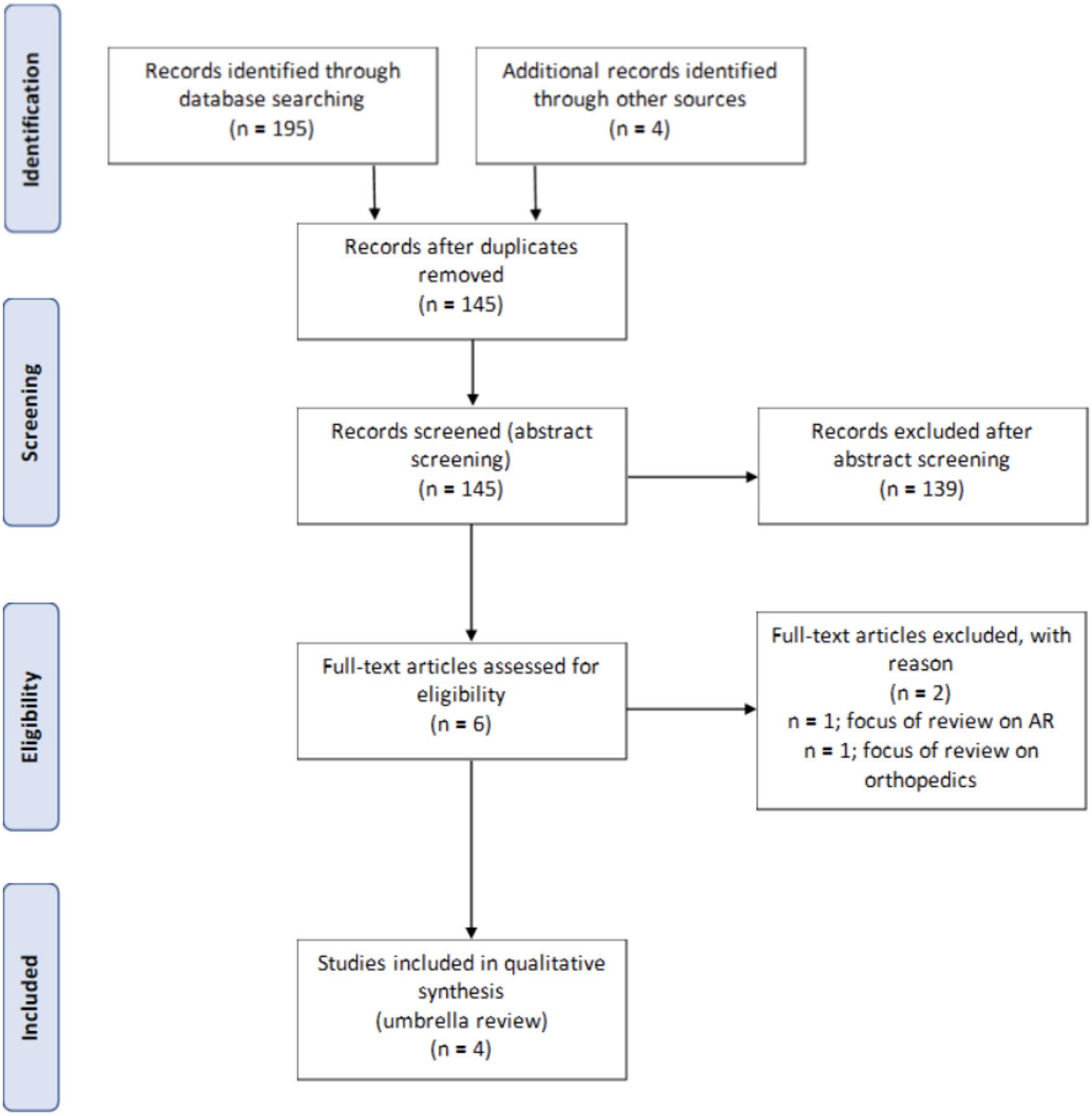
Identification
Records identified through database searching (n = 195)
Additional records identified through other sources (n = 4)
Records after duplicates removed (n = 145)
Screening
Records screened (abstract screening) (n = 145)
Records excluded after abstract screening (n = 139)
Eligibility
Full-text articles assessed for eligibility (n = 6)
Full-text articles excluded, with reason (n = 2)
n = 1; focus of review on AR
n = 1; focus of review on orthopedics
Included
Studies included in qualitative synthesis (umbrella review) (n = 4)

Fig. 1 PRISMA flowchart for the paper selection process. [33, 34]. Excluded articles at the full text screening stage included work by Ruikar and colleagues (2018) and Barsom and colleagues [32, 33]. See Results for textual description

Concerning article quality from our CASP assessment: Three of four articles addressed a clearly focused research question [28, 37, 38]. All articles looked for the right type of papers (e.g., designs) based on their research questions. We considered that only half of the articles did enough to assess quality of the included studies [28, 38], and only one of them reported results with precision [38]. Overlap between systematic reviews concerning their primary studies was lacking in places, and potentially lower than might be expected for systematic reviews conducted between 2019 and 2023. The extent of overlap between primary studies can be found in a table in the Online Resources Materials. While differences in years accounts for some of this, we believe that variance in search strategy, including some reviews only using a single database also contributed. Searching more than one database is the gold standard for systematic reviews [39–42]. We did not consider Google Scholar a database because results are impacted by Google's search algorithm and are therefore not necessarily reproducible. Further, other databases listed in Table 1, such as Medline and Pubmed, have so much overlap they don't meet the criteria to be considered as separate [42]. Our use of the PRIOR guidelines further highlighted that only half of the systematic reviews used a bias assessment tool [28, 38]. All reviews used PRISMA and/or PICOS guidelines. Accordingly, we assessed the quality of systematic reviews included in our umbrella review to be mixed.

RQ 1: How is neurosurgical performance being measured in VR environments?

The four systematic reviews described and discussed different metrics (see Table 2), but we found that the majority reported on performance metrics that could be reliably classified as being related to *time* (time to task completion or procedure-specific time to task completion) *kinematics* (surgical tool movement; velocity, jerk acceleration), and *force* (maximum force, sum of forces, and duration of excessive forces applied) [28, 37, 38, 43]. See Table 3 for all metric classification groupings.

RQ 2: What does (a) each review conclude about the efficacy of performance metrics and (b) what can be concluded collectively?

The review conducted by Chan et al. [37] concluded that the metrics they identified such as time to completion of surgical procedure, distance to target, total distance travelled, and surgical tool movement were able to distinguish expert neurosurgeons from junior residents, as well as predict participants' training levels. More specifically, force metrics and looking at the maximum force applied during a particular procedure allowed researchers to accurately discern between expertise levels [37]. The review conducted by Davids et al. [38] used metrics to assess improvements in performance following simulation practice. Their review highlighted significant improvements in procedural-specific skill acquisition, as well as procedural-specific speed and time improvements when using various forms of simulation [38]. The review by Titov et al. [28] concluded that metrics can adequately distinguish experts from novices; specifically, force applied, use of both instruments, more touches to the adjacent structures, tool acceleration and dissector jerks. Lee and Wong's [43] review included studies that assessed the ability of different metrics to distinguish novice from expert surgeons; however, these were not discussed in the review. Collectively, all three of the systematic reviews that discussed the efficacy of performance metrics concluded that metrics can be used to distinguish experts from novices (two) or assess improvements in performance following simulation practice (one).

RQ 3: What is the role of AI in training and assessment?

Currently, the primary use of AI in neurosurgical VR is to assess skills, provide information about the learner, and predict and classify training levels amongst participants. When looking into feedback being provided to learners based on their assessment, only one study [5] included in the review by Titov et al. [28], described the use of a virtual operative assistant (VOA) to facilitate feedback being given to learners. In this study, the VOA was shown to be a powerful AI tool, capable of generating similar cognitive and emotional responses as an expert tutor within learners [5]. Students who were trained by the VOA also showed better outcomes, such as higher expertise and objective structured assessment of technical skills (OSATS) scores when compared to expert-tutored and control groups [5].

Discussion: recommendations and limitations

We used an umbrella review to analyze current systematic reviews and provide a comprehensive understanding of the currently available use of AI metrics in neurosurgical VR and evidence of their effectiveness. A current challenge to synthesizing this literature is the use and inconsistent labelling of a variety of metrics. For example, Chan et al. [37] defined the time to complete brain resections as "Time", while Davids et al. [38] defined this same metric as "Speed" and Lee and Wong [43] described this as "Duration of operation". While the descriptions of these metrics matched, they were labelled differently. As a result, we classified them under the same domain of "Time". Such labelling variations

Table 1 Assessment of included reviews' data collection methodology and risk of bias assessment tool utilization

Studies	Methods utilized to collect data	Risk of bias assessment method used
Chan et al. [37]	This study utilized PRISMA guidelines. MEDLINE and PubMed databases were searched using MeSH terms. Two independent reviewers conducted title, abstract, and full text screening while a third reviewer acted as the arbitrator. They conducted thematic analysis to classify automated performance metrics and then used this to guide data extraction from included studies	No risk of bias tool was utilized by the authors
Davids et al. [38]	A multiplatform database search was conducted on the OVID platform with a variety of included databases. An additional PubMed search was also conducted. Furthermore, the authors elected to do another extended search in PubMed, Ovid-Medline, HDAS, and SCOPUS. Post-search, Rayyan web platform was utilized to conduct screening by two blinded researchers with adequate experience in the topic, following PRISMA and PICOS-guided methods. Following screening, articles were imported into Endnote and an arbitrator (senior author) was utilized to resolve ties and conflicts during article selection post-screening	The authors utilized two tools for their risk of bias assessment: The Medical Education Research Study Quality Instrument Checklist and The Cochrane risk of bias tool. In using both tools, the same arbitrator was utilized to resolve disagreements through discussion
Titov et al. [28]	This study used PRIMSA guidelines to conduct their review. Two authors performed a literature search and imported results into Mendeley. The search spanned Medline, PubMed, and Google Scholar. For Google Scholar, the authors took the first 200 relevant results. After removal of duplicates, and screening title/abstracts, full text was analyzed by the two authors. This was guided by PICOS guidelines. Disagreements were resolved through consensus. The same two authors conducted data extraction while a third author acted as the arbitrator for disagreements	The authors utilized two tools for this risk of bias assessment: The Medical Education Research Study Quality Instrument and the American Society of Clinical Oncology guidelines for the levels of evidence
Lee and Wong [43]	The authors searched PubMed/Medline, Google Scholar, and Cochrane. Two reviewers conducted the literature search as well as the title, abstract, and full-text screening. This review followed the PRISMA guidelines	No risk of bias tool was utilized by the authors

Table 2 Metrics extracted from each study that were used to classify surgeon skill level

Metrics used to measure neurosurgical performance	Review article			
	Chan et al. [37]	Davids et al. [38]	Titov et al. [28]	Lee and Wong [43]
Distance	Distance to target—distance between an implanted device's final placement and surgeon's intended target, used to represent accuracy Total distance travelled—used to represent efficiency			
Time	Task completion—time to complete brain resections, ventriculostomies Time of contact Time under fluoroscopy (shorter times represent efficiency, longer times could represent safety prioritization)	Speed—defined as procedural-specific time to task completion (outcomes observed: time per clip used (used to close off aneurysms in brain), improvement in time saved/efficiency, total time of pedicle screw placement)		Duration of operation
Kinematics	Surgical tool movement data: Velocity Acceleration Jerk (first derivative of acceleration)		Instrument motion: observed use of both instruments—bimanual work Velocity: observed lower velocity in the Z-direction during all operations and in the Y-direction, observed lower dissector velocity Jerks: observed dissector jerks Acceleration: slower tool acceleration, higher delays between two consecutive burr accelerations	Tool path length, total tip path length Efficient use of aspirator Ultrasonic aspirator path length Ultrasonic aspirator bimanual forces ratio
Force	Maximum force applied Cumulative sum of forces (use of force histograms, force pyramids—those with higher skill and improved task performance applied less force)		Force applied: observed force during surgery like on spinal dura, left posterior longitudinal ligament, C5 vertebra, etc	Duration of excessive forces applied Maximum and sum of forces used by instruments
Blood loss	Blood loss—volume of blood loss during virtual procedure (top surgeons experienced less blood loss		Blood loss	Total blood loss
Volume of resection	Volume and extent of resection – % of brain tumor and normal tissue removed, vol over time quantifies efficiency; experts resect less tumor but had least amount of normal tissue removed			Volume of tissues removed Tumor percentage resected Total simulated normal brain volume removed Extent of resection
Accuracy		Accuracy – defined as learners' acquisition of procedural skill using simulation		

Table 2 (continued)

Metrics used to measure neurosurgical performance	Review article			
	Chan et al. [37]	Davids et al. [38]	Titov et al. [28]	Lee and Wong [43]
Safety		Safety: outcomes for safety were based on aneurysm clipping time, ratio of clip attempt to clip usage, reduction in surgical error, comfort levels in patients, accuracy of pedicle screw placement	Safety: Observed in touches to adjacent structures by suction	
Skill		Knowledge and procedural skill of students: OSAT, PPDIS and ODS scores used to assess this domain. Other outcomes were measures of understanding 3D images, successful lumbar punctures after 1 year of training, etc		

Table 3 Combined list of all metrics extracted from the 4 studies, classified by domain

All metrics	Metrics defined by reviews
Distance	Distance to target
	Total distance travelled
Time	Task completion time
	Time of contact
	Time under fluoroscopy
	Speed
	Duration of operation
Kinematics	Velocity
	Acceleration
	Jerk
	Instrument motion
	Tool path length
	Total tip path length
	Efficient use of aspirator
	Ultrasonic aspirator path length
	Ultrasonic aspirator bimanual forces ratio
Force	Maximum force applied
	Cumulative sum of forces
	Duration of excessive forces applied
	Maximum and sum of forces used by instruments
Blood loss	Volume of total blood loss
Volume of resection	Volume of tissues removed
	Tumor percentage resected
	Normal tissue percentage resected
	Extent of resection
Accuracy	Accuracy
Safety	Safety
Skill	Knowledge

can create ambiguity within the scientific community when it comes to understanding which metrics should be used to assess surgeon performance. We contribute to supporting a more united understanding of metrics by classifying them into nine domains, three of which were especially dominant metric domains for assessing surgical performance: kinematics, force, and time. While only highlighted in two of the four analyzed reviews, we propose that "safety" is an equally important if not the most important domain to consider when assessing surgeon performance with metrics. Indeed, recent literature has indicated that surgical safety emerged as being the most related to trainee expertise in a surgical environment when VR surgical simulations are assessed with artificial neural networks [44–46]. Another critical observation we made from our umbrella review was that while most reviews identified a time metric, "efficiency" is an important time-related metric we felt was not given full consideration. Efficiency not only includes task completion time but should also consider the quality of the operative performance. In other words, efficiency is important because speed without consideration of surgical outcome holds limited value.

Concerning efficacy, all three of the systematic reviews that discussed the efficacy of performance metrics concluded that metrics can be effectively used to distinguish experts from novices or assess improvements in performance following simulation practice. We recommend that researchers draw on our metric classification, as the literature continues to develop, to conduct a meta-analysis to examine the relative efficacy of each of these metrics to weigh the relative evidence of each and to do so for different procedures and specialties. In the meantime, our review suggests that time, kinematics, and force are especially widely used and generally effective metrics to assess surgical performance.

We noted several important gaps in the literature while conducting this review. Notably, that non-technical skills receive very little attention in the development and testing of surgical performance metrics. Given that non-technical skills, such as communication and collaboration, have been linked to both neurosurgical errors [47] and patient compliance and outcomes [48] and that these are also increasingly identified as core competencies for surgeons [49], AI-enhanced VR should also be used to scaffold such skills. A recent scoping review of AI and VR in doctor-patient risk communication [16] identified three neurosurgery studies [50–52], none of which formally assessed trainees' communication skills (with or without AI metrics). Surgical education should leverage the affordances of VR and AI to not only perform technical skills but communicate with members of the inter-professional OR team and patients. Surgical trainees can practice such skills by having conversations with virtual agents, referred to as non-playable characters in videos games, using either pre-selected text options or, where natural language processing (NLP) allows, speaking or typing their response. Research in the broader artificial intelligence in education community has demonstrated that such interactions can be effective for learning and collaborating, such as establishing goals using NLP [53, 54].

Another notable gap in research and practice is the relatively narrow use of AI. While our umbrella review highlights the promise that AI metrics hold for assessing performance, very few studies took advantage of the potential for AI to provide feedback to trainees. This is particularly unfortunate as it fails to capitalize on an important component of simulation that can be easily delivered and at scale. Further, failing to provide feedback to support learning and deliberate practice positions such technologies as exclusively research rather than educational tools one conducts research on to improve education [55]. Given the use of simulations as typically formative rather than summative (high-stakes) educational activities and skill assessments, it is reasonable to assume that AI metrics in VR environments will serve a similar role. If that is the case, surgical educators will find that the accuracy of metrics is only one core aspect. *How* the feedback is delivered to learners, simulation and surgical educators, and program directors will matter. Future programs of research should consider not only comparing AI to human feedback, but also conducting research to better understand what the right type of feedback looks like—something that is expected to vary not just by whether the person receiving feedback is a trainee or educator. Dimensions to consider include the timing, frequency, emotional and motivational messaging, personalized recommendations for further practice based on observed weaknesses and strengths, as well as user experience considerations, such as graphical summaries and learner dashboards [56]. In addition to considering how feedback is delivered, surgical educators should closely examine *how* it is received. Recent research has shown, for example, that an AI-enhanced curriculum for bimanual surgical skills resulted in unintended changes that improved performance in safety but negatively affected some efficiency metrics [57].

Related to the above limitation of AI metrics being dominantly shared with researchers rather than trainees is the potential for such metrics to eventually be deployed in real-time during non-simulated training, such as interactions with patients or parts of surgical procedures. A recent review of current applications of AI in the operating room only identified nine studies, but a wide range of potential applications including procedure duration prediction, gesture recognition, intraoperative cancer detection, intraoperative video analysis, workflow recognition, an endoscopic guidance system, knot-tying, and automatic registration and tracking of the bone in orthopedic surgery [58]. While such research is in the very early phases and more appropriate for augmented than virtual reality, it reinforces the need for feedback to be received by and tested by stakeholders other than researchers.

Finally, one of the advantages of VR simulations is the use of virtual assets and associated opportunities to evaluate trainees on not only a range of procedures, but a range of patients [59]. When it comes to neurosurgery, the use of digital twins [60, 61] of typical or atypical patient cases stands to provide trainees with a greater range of experiences and associated difficulty levels. Such training would likely require a more longitudinal research design than most studies to date have explored (for good reason), but such a direction holds educational promise as the field matures.

This umbrella review was limited by our ability to only draw upon four systematic review studies to address our questions. Further, the quality of these systematic reviews was mixed. While this was sufficient to conduct an umbrella review, a greater number of studies would both reflect a more mature field and potentially yield additional insights.

Conclusion

This umbrella review provides an overview of the current literature supporting the use of VR in neurosurgical training. We shared and applied a unifying classification of AI metrics for evaluating neurosurgery and identified time, kinematics, and force as dominant metrics. Concerning efficacy, all three of the systematic reviews that discussed the efficacy of performance metrics concluded that they can be effectively used to distinguish experts from novices or assess improvements in performance following simulation practice. This article also highlighted some key gaps and future directions in surgical education research and practice. Proximal future directions included developing AI metrics to assess non-technical skills, sharing personalized feedback with trainees on their performance and assessing how feedback can be best delivered with attention to how it is received by different learners. We also encouraged surgical educators to take advantage of the flexibility of virtual assets to present varied and diverse cases to trainees and explore their learning curves not only on typical procedures, but also on increasingly challenging or atypical patient cases. Digital twin technology could be used to support this direction. More distal future directions include implementing AI metrics in practice, but through the use of augmented reality rather than VR. Overall, our review suggests the future is bright when it comes to VR and AI metrics for neurosurgical education and there are many under-explored and promising places to take rapidly evolving technologies.

Summary box

- Virtual reality and artificial intelligence in surgical education have the potential to address contemporary challenges in surgical education.
- Due to the rapid growth of literature in VR and AI, it is essential to systematically evaluate and reflect on the current evidence regarding AI's role in assessing trainees' performance.
- The use of AI in distinguishing expertise levels and assessing performance improvements in neurosurgical VR is evident, with time, kinematics, and force identified as dominant metric domains.
- Current literature lacks focus on non-technical skills, the delivery and reception of AI feedback, and exploring how to leverage virtual assets to practice on a diverse range of patient cases.
- The future of VR and AI metrics in neurosurgical education is promising, with potential advancements in assessing non-technical skills, delivering personalized feedback, and incorporating augmented reality for real-time applications.

Where to find more information

- A systematic review of virtual reality for the assessment of technical skills in neurosurgery—https://pubmed.ncbi.nlm.nih.gov/34333472/
- Simulation for skills training in neurosurgery: a systematic review, meta-analysis, and analysis of progressive scholarly acceptance—https://pubmed.ncbi.nlm.nih.gov/32944808/
- Neurosurgical skills analysis by machine learning models: systematic review—https://pubmed.ncbi.nlm.nih.gov/37191734/
- Virtual reality and augmented reality in the management of intracranial tumors: A review—https://pubmed.ncbi.nlm.nih.gov/30642663/

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s44186-024-00284-z.

Acknowledgements Not applicable for this review article.

Data availability This is a review article rather than an original research article. Accordingly, there is no original research data to share.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Kirubarajan A, Young D, Khan S, Crasto N, Sobel M, Sussman D. Artificial intelligence and surgical education: a systematic scoping review of interventions. J Surg Educ. 2022;79(2):500–15.
2. Patel VL, Shortliffe EH, Stefanelli M, Szolovits P, Berthold MR, Bellazzi R, et al. The coming of age of artificial intelligence in medicine. Artif Intell Med. 2009;46(1):5–17.
3. Kirubarajan A, Taher A, Khan S, Masood S. Artificial intelligence in emergency medicine: a scoping review. J Am Coll Emerg Physicians Open. 2020;1(6):1691–702.
4. Bilgic E, Gorgy A, Young M, Abbasgholizadeh-Rahimi S, Harley JM. Artificial intelligence in surgical education: considerations for interdisciplinary collaborations. Surg Innov. 2022;29(2):137–8.
5. 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 Netw Open. 2022;5(2): e2149008.
6. Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Maestro RFD. The virtual operative assistant: an explainable artificial intelligence tool for simulation-based training in surgery and medicine. PLoS ONE. 2020;15(2): e0229596.

7. Bilgic E, Gorgy A, Yang A, Cwintal M, Ranjbar H, Kahla K, et al. Exploring the roles of artificial intelligence in surgical education: A scoping review. Am J Surg. 2022;224(1 pt A):205–16.
8. Alsayegh A, Bakhaidar M, Winkler-Schwartz A, Yilmaz R, Del Maestro RF. Best practices using ex vivo animal brain models in neurosurgical education to assess surgical expertise. World Neurosurg. 2021;155:e369–81.
9. Ayaz O, Ismail FW. Healthcare simulation: a key to the future of medical education - a review. Adv Med Educ Pract. 2022;13:301–8.
10. Azher S, Cervantes A, Marchionni C, Grewal K, Marchand H, Harley JM. Virtual simulation in nursing education: headset virtual reality and screen-based virtual simulation offer a comparable experience. Clin Simul Nurs. 2023;1(79):61–74.
11. Zhang J, Lu V, Khanduja V. The impact of extended reality on surgery: a scoping review. Int Orthop. 2023;47(3):611–21.
12. Al-Elq AH. Simulation-based medical teaching and learning. J Family Community Med. 2010;17(1):35–40.
13. Harley JM, Bilgic E, Lau CHH, Gorgy A, Marchand H, Lajoie SP, et al. Nursing students reported more positive emotions about Training during COVID-19 after using a virtual simulation paired with an in-person simulation. Clin Simul Nurs. 2023. https://doi.org/10.1016/j.ecns.2023.04.006.
14. Wu Q, Wang Y, Lu L, Chen Y, Long H, Wang J. Virtual simulation in undergraduate medical education: a scoping review of recent practice. Front Med (Lausanne). 2022;9: 855403.
15. Bracq MS, Michinov E, Jannin P. Virtual reality simulation in nontechnical skills training for healthcare professionals: a systematic review. Simul Healthc. 2019;14(3):188–94.
16. Antel R, Abbasgholizadeh-Rahimi S, Guadagno E, Harley JM, Poenaru D. The use of artificial intelligence and virtual reality in doctor-patient risk communication: a scoping review. Patient Educ Couns. 2022;105(10):3038–50.
17. Nassar AK, Al-Manaseer F, Knowlton LM, Tuma F. Virtual reality (VR) as a simulation modality for technical skills acquisition. Annals of Medicine and Surgery. 2021;1(71): 102945.
18. Cevallos N, Zukotynski B, Greig D, Silva M, Thompson RM. The utility of virtual reality in orthopedic surgical training. J Surg Educ. 2022;79(6):1516–25.
19. Hasan LK, Haratian A, Kim M, Bolia IK, Weber AE, Petrigliano FA. Virtual reality in orthopedic surgery training. Adv Med Educ Pract. 2021;12:1295–301.
20. Larsen CR, Soerensen JL, Grantcharov TP, Dalsgaard T, Schouenborg L, Ottosen C, et al. Effect of virtual reality training on laparoscopic surgery: randomised controlled trial. BMJ. 2009;338: b1802.
21. Hoyler M, Finlayson SRG, McClain CD, Meara JG, Hagander L. Shortage of doctors, shortage of data: a review of the global surgery, obstetrics, and anesthesia workforce literature. World J Surg. 2014;38(2):269–80.
22. Williams TE, Satiani B, Thomas A, Ellison EC. The impending shortage and the estimated cost of training the future surgical workforce. Ann Surg. 2009;250(4):590–7.
23. Elkbuli A, Sutherland M, Sanchez C, Liu H, Ang D, McKenney M. The shortage of trauma surgeons in the US. Am Surg. 2022;88(2):280–8.
24. Ellison EC, Pawlik TM, Way DP, Satiani B, Williams TE. Ten-year reassessment of the shortage of general surgeons: Increases in graduation numbers of general surgery residents are insufficient to meet the future demand for general surgeons. Surgery. 2018;164(4):726–32.
25. Mao RQ, Lan L, Kay J, Lohre R, Ayeni OR, Goel DP, et al. Immersive virtual reality for surgical training: a systematic review. J Surg Res. 2021;268:40–58.
26. Mofatteh M. Neurosurgery and artificial intelligence. AIMS Neurosci. 2021;8(4):477–95.
27. Satapathy P, Hermis AH, Rustagi S, Pradhan KB, Padhi BK, Sah R. Artificial intelligence in surgical education and training: opportunities, challenges, and ethical considerations - correspondence. Int J Surg. 2023;109(5):1543–4.
28. Titov O, Bykanov A, Pitskhelauri D. Neurosurgical skills analysis by machine learning models: systematic review. Neurosurg Rev. 2023;46(1):121.
29. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Info Libr J. 2009;26(2):91–108.
30. Schaepkens SPC, Veen M, de la Croix A. Is reflection like soap? a critical narrative umbrella review of approaches to reflection in medical education research. Adv Health Sci Educ Theory Pract. 2022;27(2):537–51.
31. Belur J, Tompson L, Thornton A, Simon M. Interrater reliability in systematic review methodology: exploring variation in coder decision-making. Sociological Methods & Research. 2021;50(2):837–65.
32. Barsom EZ, Graafland M, Schijven MP. Systematic review on the effectiveness of augmented reality applications in medical training. Surg Endosc. 2016;30(10):4174–83.
33. Ruikar DD, Hegadi RS, Santosh KC. A systematic review on orthopedic simulators for psycho-motor skill and surgical procedure training. J Med Syst. 2018;42(9):168.
34. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372: n71.
35. Critical Appraisal Skills Programme (2022) CASP Systematic Review Checklist [Internet]. Critical Appraisal Checklists (cited 2023 Oct 6). Available from https://casp-uk.net/casp-tools-checklists/
36. Gates M, Gates A, Pieper D, Fernandes RM, Tricco AC, Moher D, et al. Reporting guideline for overviews of reviews of healthcare interventions: development of the PRIOR statement. BMJ. 2022;378: e070849.
37. 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.
38. 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. Neurosurg Rev. 2021;44(4):1853–67.
39. Lefebvre C, Glanville J, Briscoe S, Littlewood A, Marshall C, Metzendorf MI, et al. Searching for and selecting studies. In: Cochrane Handbook for Systematic Reviews of Interventions [Internet]. John Wiley & Sons, Ltd; 2019 (cited 2024 Jul 3). pp. 67–107. Available from: https://onlinelibrary.wiley.com/doi/abs/https://doi.org/10.1002/9781119536604.ch4
40. Ewald H, Klerings I, Wagner G, Heise TL, Stratil JM, Lhachimi SK, et al. Searching two or more databases decreased the risk of missing relevant studies: a metaresearch study. J Clin Epidemiol. 2022;149:154–64.
41. Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. Syst Rev. 2017;6(1):245.
42. Cowan S. Search Engines vs. Google Scholar vs. Library Databases | Leddy Library [Internet]. Search Engines vs. Google Scholar vs. Library Databases (cited 2024 Jul 3). Available from https://leddy.uwindsor.ca/get-help/guides/search-engines-vs-google-scholar-vs-library-databases
43. Lee C, Wong GKC. Virtual reality and augmented reality in the management of intracranial tumors: a review. J Clin Neurosci. 2019;62:14–20.
44. Reich A, Mirchi N, Yilmaz R, Ledwos N, Bissonnette V, Tran DH, et al. Artificial neural network approach to competency-based

training using a virtual reality neurosurgical simulation. Operative Neurosurgery. 2022;23(1):31.

45. 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.
46. 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. Comput Biol Med. 2021;136: 104770.
47. Hartley BR, Hong C, Elowitz E. Communication in neurosurgery-the tower of babel. World Neurosurg. 2020;133:457–65.
48. Hartley BR, Elowitz E. Barriers to the enhancement of effective communication in neurosurgery. World Neurosurg. 2020;133:466–73.
49. Frank JR, Snell L, Sherbino J. CanMEDS 2015 Physician Competency Framework; 2015.
50. Eisenmenger LB, Wiggins RH, Fults DW, Huo EJ. Application of 3-dimensional printing in a case of osteogenesis imperfecta for patient education, anatomic understanding, preoperative planning, and intraoperative evaluation. World Neurosurg. 2017;107:1049.e1-1049.e7.
51. van de Belt TH, Nijmeijer H, Grim D, Engelen LJLPG, Vreeken R, van Gelder MMHJ, et al. Patient-specific actual-size three-dimensional printed models for patient education in glioma treatment: first experiences. World Neurosurg. 2018;117:e99–105.
52. Perin A, Galbiati TF, Ayadi R, Gambatesa E, Orena EF, Riker NI, et al. Informed consent through 3D virtual reality: a randomized clinical trial. Acta Neurochir (Wien). 2021;163(2):301–8.
53. Azevedo R, Bouchet F, Duffy M, Harley J, Taub M, Trevors G, et al. Lessons learned and future directions of MetaTutor: leveraging multichannel data to scaffold self-regulated learning with an intelligent tutoring system. Front Psychol. 2023. https://doi.org/10.3389/fpsyg.2022.813632.
54. Harley JM, Taub M, Azevedo R, Bouchet F. Let's set up some subgoals: understanding human-pedagogical agent collaborations and their implications for learning and prompt and feedback compliance. IEEE Trans Learn Technol. 2018;11(1):54–66.
55. Ericsson KA, Nandagopal K, Roring RW. Toward a science of exceptional achievement: attaining superior performance through deliberate practice. Ann N Y Acad Sci. 2009;1172:199–217.
56. Harley JM, Lajoie SP, Frasson C, Hall NC. Developing emotion-aware, advanced learning technologies: a taxonomy of approaches and features. Int J Artif Intell Educ. 2017;27(2):268–97.
57. Fazlollahi AM, Yilmaz R, Winkler-Schwartz A, Mirchi N, Ledwos N, Bakhaidar M, et al. AI in surgical curriculum design and unintended outcomes for technical competencies in simulation training. JAMA Netw Open. 2023;6(9): e2334658.
58. Birkhoff DC, van Dalen ASHM, Schijven MP. A review on the current applications of artificial intelligence in the operating room. Surg Innov. 2021;28(5):611–9.
59. Conigliaro RL, Peterson KD, Stratton TD. Lack of diversity in simulation technology: an educational limitation? Simul Healthc. 2020;15(2):112–4.
60. Paul G, Hamdy Doweidar M (Eds.) Digital Human Modeling and Medicine: The Digital Twin. In: Digital Human Modeling and Medicine (Internet). Academic Press; 2023 (cited 2023 Dec 8). p. iii. Available from https://www.sciencedirect.com/science/article/pii/B9780128239131010010
61. Jones D, Snider C, Nassehi A, Yon J, Hicks B. Characterising the digital twin: a systematic literature review. CIRP J Manuf Sci Technol. 2020;29:36–52.