On Extended Reality Objective Performance Metrics for Neurosurgical Training

Alessandro Iop[1(✉)], Olga Viberg[1], Adrian Elmi-Terander[3,4], Erik Edström[2,4], and Mario Romero[1]

[1] EECS, KTH Royal Institute of Technology, Stockholm, Sweden
{aiop,oviberg,marior}@kth.se
[2] Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden
[3] Department of Surgical Sciences, Uppsala University, Uppsala, Sweden
[4] Capio Spine Center Stockholm, Löwenströmska Hospital, Stockholm, Sweden

Abstract. The adoption of Extended Reality (XR) technologies for supporting learning processes is an increasingly popular research topic for a wide variety of domains, including medical education. Currently, within this community, the metrics applied to quantify the potential impact these technologies have on procedural knowledge acquisition are inconsistent. This paper proposes a practical definition of standard metrics for the learning goals in the application of XR to surgical training. Their value in the context of previous research in neurosurgical training is also discussed. Objective metrics of performance include: spatial accuracy and precision, time-to-task completion, number of attempts. The objective definition of what the learner's aims are enables the creation of comparable XR systems that track progress during training. The first impact is to provide a community-wide metric of progress that allows for consistent measurements. Furthermore, a measurable target opens the possibility for automated performance assessments with constructive feedback.

Keywords: Extended Reality · Surgical Simulation · Neurosurgical Education · Procedural Knowledge · Performance Metrics

1 Introduction

Surgical education is lengthy. It takes medical students from textbooks in the classroom to performing surgeries on live patients. There are currently numerous opportunities for surgical *residents*, i.e. doctors in training, to acquire nominal knowledge on human anatomy and physiology by means of both analog and digital tools. Conversely, hands-on practice of operations tends to be limited prior to the learner stepping into a real operating room (OR). In the traditional educational path, gaining the know-how about a specific operation happens directly

Supported by InfraVis, Swedish Research Council grant 2021-00181.

O. Viberg et al. (Eds.): EC-TEL 2023, LNCS 14200, pp. 573–579, 2023.
https://doi.org/10.1007/978-3-031-42682-7_44

in front of a live patient: residents start by observing a senior surgeon performing the operation, then they operate on their own while being under supervision of an expert until they gain the independence to operate on their own. Before entering the OR, opportunities for honing manual dexterity and hand-eye coordination are rare and expensive. Residents might be able to gain access to animal and/or human cadavers in a handful of occasions and with limited possibilities for repeated training sessions. In this context, what is lacking is the possibility for residents to quickly and independently assess their own performance after a surgical *simulation*, as well as compare it with the performance of experts or their own past performance, with the aim of analyzing their own learning curve over time. In recent years, researchers in medicine and engineering have come together to address this lack of resources for surgical practice by proposing numerous novel tools and environments targeting residents as individual learners who wish to hone their skills in a risk-free, controlled and accessible environment.

Table 1. Systematic review of metrics used in XR cranial neurosurgical training classified by educational setting. Based on [11].

Metric type	*Learning*	*Practicing*	*Skill assessment*	Total (freq.)
Space-time	[10]	[2,6,12,22,27]	[1,3,8,16,18,19,24,25]	14 (54%)
Force	[10]	[6,22]	[1,3,5,8,18,19,24,26]	11 (42%)
Outcome	[4,10,14]	[6,7,22,27]	[1,3,5,8,16,19,21,24–26]	17 (65%)
Qualitative	[13,17,20]	[2,15,23]	[16,26]	8 (31%)

The adoption of anatomically realistic *phantom* models, as well as of XR technologies, such as augmented and virtual reality, are two of the most prolific avenues of research from this perspective [9]. By combining virtual with real imagery, and integrating it with accurate replicas of human anatomy, numerous possibilities for interactive, realistic and adaptable simulation scenarios are enabled. One of the greatest benefits of employing these accessible and easy-to-develop emerging technologies is arguably the versatility of the resulting learning environments to fit different types of surgical practices. Furthermore, through sensors and automatic data collection, assessing simulation performance improvements over time is greatly facilitated by modern applications. A consequence of this research windfall is the need for common learning-measuring criteria. To enable a full exploration of the versatility, scalability and accessibility of newly developed XR tools for surgical education, there needs to be a common set of metrics that quantify simulation performance as a measure of *procedural knowledge* acquisition and transfer. With widespread adoption of standard performance metrics by the research community, it may be possible to compare learning achievements both across different technologies and between different actors, potentially distantly located and with varying degrees of expertise. In the present paper, we propose a set of metrics for the assessment of outcomes

in XR simulations of surgical operations, specifically in the field of cranial neurosurgery. From this perspective, this narrower field of research comes with a partially different set of challenges when compared to, for instance, spinal neurosurgery, and, as we have previously shown, has so far been relatively uncharted territory for educational XR technologies [11]. Nevertheless, our proposed set of metrics can be applied to other types of surgery where procedural knowledge acquisition and transfer involves precise and efficient hand-eye coordination and manual dexterity.

2 Survey of the Domain of Practice

In a recent systematic review, we surveyed the adoption of XR technologies, with varying degrees of augmentation, in cranial neurosurgical education [11]. There, we defined education as a combination of learning, practicing and skill assessment, with the goal of acquiring the necessary knowledge to successfully perform a surgical procedure. Table 1 highlights the variability in the metrics considered among the 26 studies that measured user performance. Studies are grouped by education type according to the definition above, while performance metrics are arbitrarily categorized based on complexity, degree of aggregation and collection method (automatic vs. non-automatic) into the following:

- **Space-time** metrics include time, position and orientation measures in the surgical performance, i.e. kinematic measures that can be related to both the surgical simulation as a whole or only part of it. Examples are time to completion, instrument position, entry point location.
- **Force** metrics include measures of forces applied by test subjects onto the instrument or onto the apparatus being used as a proxy for the patient (e.g. a phantom). Examples are bandwidth, ratio and sum of the forces applied.
- **Outcome** metrics include frequencies and patterns of accuracy, errors, precision and consistency in the simulation. Metrics derived from comparisons with the intended outcome or an existing benchmark are also included. Examples are number of attempts, frequency of complications and success rate.
- **Qualitative** metrics include non-automatic evaluations made by experimenters or experts to evaluate surgical simulation performance, according to either standard or arbitrary criteria. Examples are grades based on scales, answers to questions and scores given to operative dictation.

As shown in Table 1, despite a broad arbitrary categorization of performance metrics considered in the studies, there is no clear consensus on a common type of metric. There appears to be notable variance in the frequency distribution of the single metric categories considered here. The outcome category, while being the most frequent (65%), still falls short of being labeled as "widely adopted". This is surprising, considering that simple measures of success, accuracy and precision fall into this category. Possible reasons may be the nature and scale of the research questions investigated, as well as the alternative adoption of more

qualitative methods for assessing performance, i.e. the fourth category of metrics. In other words, not all the studies present data related to, for instance, time to completion or distance to the target, because an approximate estimate of outcome was performed by experts. This qualitative category, on the other hand, is the least representative among the 26 studies (31%). It involves non-automatic assessment from senior surgeons grading through validated forms, ensuring systematic scoring. Finally, quantity and variability of metrics also noticeably vary between different types of education under scrutiny. In particular, while only a handful of distinct "learning" studies employ performance metrics at all (n=7), the "skill assessment" studies present considerably richer data (n=12). In the latter case, more than one category of metric is often considered simultaneously.

3 PARENT Metrics for Objective Assessment

As previously discussed, one of the many benefits of adopting XR technologies in surgical education is their versatility and scalability in different settings. By enabling asynchronous, distributed, and independent procedural knowledge acquisition and transfer, the learning opportunities for a resident surgeon increase compared to traditional education. Given the traditional co-located, synchronous learning in this field, portable and relatively inexpensive XR technologies can thus complement training through automatically assessed performance metrics. Such metrics need not be limited to low-dimension measures of e.g. kinematics and forces; "advanced" computed metrics can also be considered, such that multiple "simple" ones can be aggregated into meaningful indexes. Furthermore, the absence of a teacher during this mediated learning experience entails that subjective evaluations of surgical performance are not scalable in space and time. That is, the need for an expert surgeon assessing and grading simulation performance is unwarranted by the ability for residents to practice "anytime, anywhere".

In order to propose a set of metrics that can reach consensus across multiple domains of expertise, their usefulness should be balanced with their scalability. While metrics that suit a specific surgical operation are effective in providing the necessary data to inform reliable evaluations, this approach is very sensitive to small variations in the learning scenario. This means that, for instance, the total volume of tumor removed may be relevant in tumor resection tasks, but not applicable at all in ventriculostomy tasks. On the other hand, metrics that are too abstract for the scenario may fall short of being informative enough for a learner aiming at assessing their own performance by comparing it to the intended one (e.g. as performed by experts) or to their past performances. A simple grade on an arbitrary and opaque A–F scale is an example: the grade does not tell why the performance was graded as such or what the student may have done correctly or not. The following proposed metrics should therefore act as a concrete origin for the automatic collection and aggregation of relevant performance indicators. If kept agnostic to the specific surgical procedure, they can be robust enough to enable cross-domain comparisons, and sufficient for a preliminary real-time assessment.

- **P***recision of distances and angles*: how close the measured values are to each other, i.e. their variability across multiple simulation trials. This can be inferred by calculating Euclidean and angular distances between surgical instruments at equivalent time frames in two or more trials.
- **A***ccuracy of distances and angles*: how close the measured values are to the intended (target) value, i.e. their correctness for each simulation trial, inferred by comparing against benchmark baselines.
- **R***ate of success*: ratio between the number of successful simulation trials and total number of trials. It is complementary to the rate of error, the ratio between the number of unsuccessful trials and total number of trials. A clear definition of a threshold between success and error is warranted here.
- **E***rrors of measurement*: robustness of the hardware in measuring performance indicators, expressed as the minimal detectable difference between two distinct observations over the range of values across all observations.
- **N***umber of attempts*: count of simulation trials. This metric needs a clear definition distinguishing a re-start from a continuation of a previous attempt.
- **T***ime to completion*: total time elapsed between the start and the end of a single simulation trial. A clear definition for procedure start and end, either as a location in space and/or a moment in time, is warranted here.

4 Conclusions and Future Work

Growing research on XR technologies in neurosurgical training calls for consensus one learning metrics. By observing current trends in the field and carefully balancing scalability and efficacy, we have proposed six concrete metrics for objective quantification of procedural knowledge acquisition and transfer. Consensus over them and, ultimately, their adoption throughout the broader field of surgical education will potentially enable more impactful research results that are comparable across different application domains. In future research, these metrics may afford rigorous and quantitative comparison between participant populations, simulated procedures, and XR tools. For validation, we plan to disseminate them in future workshops and surveying the research community.

References

1. Alotaibi, F.E., et al.: Assessing bimanual performance in brain tumor resection with NeuroTouch, a virtual reality simulator. Neurosurgery **11**, 89–98 (2015)
2. Ansaripour, A., et al.: P56 virtual reality simulation in neurosurgical training: a single blinded randomised controlled trial & review of all available training models. J. Neurol. Neurosurg. Psychiatry **90**(3), e38 (2019)
3. Azarnoush, H., et al.: Neurosurgical virtual reality simulation metrics to assess psychomotor skills during brain tumor resection. Int. J. Comput. Assist. Radiol. Surg. 603–618 (2015)
4. Azimi, E., Molina, C., Chang, A., Huang, J., Huang, C.-M., Kazanzides, P.: Interactive training and operation ecosystem for surgical tasks in mixed reality. In: Stoyanov, D., et al. (eds.) CARE/CLIP/OR 2.0/ISIC -2018. LNCS, vol. 11041, pp. 20–29. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01201-4_3

5. Bugdadi, A., et al.: Automaticity of force application during simulated brain tumor resection. J. Surg. Ed. 104–115 (2018)
6. Bugdadi, A., et al.: Is virtual reality surgical performance influenced by force feedback device utilized? J. Surg. Ed. 262–273 (2019)
7. Cutolo, F., et al.: A new head-mounted display-based AR system in neurosurg. ONC.: a study on phantom. Comput. Assist. Surg. 39–53 (2017)
8. Gelinas-Phaneuf, N., et al.: Assessing performance in brain tumor resection using a novel virtual reality simulator. Int. J. Comput. Assist. Radiol. Surg. 1–9 (2014)
9. Herur-Raman, A., et al.: Next-generation simulation-integrating extended reality technology into medical education. Front. Virtual Real. 115 (2021)
10. Holloway, T., et al.: Operator experience determines performance in a simulated computer-based brain tumor resection task. Int. J. Comput. Assist. Radiol. Surg. 1853–1862 (2015)
11. Iop, A., et al.: Extended reality in neurosurgical education: a systematic review. Sensors **22**(16), 6067 (2022)
12. Ledwos, N., et al.: Assessment of learning curves on a simulated neurosurgical task using metrics selected by AI. J. Neurosurg. 1–12 (2022)
13. Lin, W., Zhu, Z., He, B., Liu, Y., Hong, W., Liao, Z.: A novel virtual reality simulation training system with haptic feedback for improving lateral ventricle puncture skill. Virtual Real. 1–13 (2021)
14. Patel, A., et al.: Neurosurgical tactile discrimination training with haptic-based virtual reality simulation. Neurol. Res. **36**(12), 1035–1039 (2014)
15. Perin, A., et al.: The "stars-cascade" study: VR simulation as a new training approach in vascular neurosurgery. World Neurosurg. **154**, e130–e146 (2021)
16. Roitberg, B.Z., et al.: Evaluation of sensory and motor skills in neurosurgery applicants using a virtual reality neurosurgical simulator: the sensory-motor quotient. J. Surg. Educ. **72**(6), 1165–1171 (2015)
17. Ros, M., et al.: Applying an immersive tutorial in virtual reality to learning a new technique. Neuro-Chirurgie, 212–218 (2020)
18. Sawaya, R., et al.: Virtual reality tumor resection: the force pyramid approach. Oper. Neurosurg. **14**(6), 686–696 (2018)
19. Sawaya, R., et al.: Development of a performance model for virtual reality tumor resections. J. Neurosurg. 192–200 (2019)
20. Schirmer, C.M., et al.: Virtual reality-based simulation training for ventriculostomy: an evidence-based approach. Neurosurgery, 66–73 (2013)
21. Shakur, S.F., et al.: Usefulness of a VR percutaneous trigeminal rhizotomy simulator in neurosurgical training. Op. Neurosurg. 420–425 (2015)
22. Teodoro-Vite, S., et al.: A high-fidelity hybrid virtual reality simulator of aneurysm clipping repair with brain sylvian fissure exploration for vascular neurosurgery training. Simul. Healthc. 285–294 (2021)
23. Thawani, J., et al.: Resident simulation training in endoscopic endonasal surgery utilizing haptic feedback technology. J. Clin. Neurosci. 112–116 (2016)
24. Winkler-Schwartz, A., et al.: Bimanual psychomotor performance in neurosurgical resident applicants assessed using NeuroTouch, a virtual reality simulator. J. Surg. Educ. 942–953 (2016)
25. Winkler-Schwartz, A., et al.: Machine learning identification of surgical and operative factors associated with surgical expertise in virtual reality simulation. JAMA Netw. Open (2019)
26. Winkler-Schwartz, A., et al.: A comparison of visual rating scales and simulated virtual reality metrics in neurosurgical training: a generalizability theory study. World Neurosurg. **127**, e230–e235 (2019)

27. Yudkowsky, R., et al.: Practice on an augmented reality/haptic simulator and library of virtual brains improves residents' ability to perform a ventriculostomy. Simul. Healthc. **8**(1), 25–31 (2013)