Performance metrics in mastoidectomy: A systematic review

Fahd Al-Shahrestani, MD (1); Mads Sølvsten Sørensen, MD, DMSc (1); Steven Arild Wuyts Andersen, MD, PhD (1,2)

1. Department of Otorhinolaryngology—Head & Neck Surgery, Rigshospitalet, Copenhagen, Denmark.
2. The Simulation Centre at Rigshospitalet, Copenhagen Academy for Medical Education and Simulation (CAMES), Centre for HR, the Capital Region of Denmark.

Correspondence: Fahd Al-Shahrestani, MD, Department of Otorhinolaryngology—Head & Neck Surgery, Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen, Denmark. E-mail: f.l.shahrestani@gmail.com.

Conflicts of interests: None.
Financial disclosures: None.

Full citation: Al-Shahrestani F, Sørensen MS, Andersen SA. Performance metrics in mastoidectomy training: A systematic review. Eur Arch Otorhinolaryngol. 2019 Mar;276(3):657-664.
DOI: 10.1007/s00405-018-05265-9

Objective: To investigate validity evidence, and strengths and limitations of performance metrics in mastoidectomy training. Methods A systematic review following the PRISMA guidelines. Studies reporting performance metrics in mastoidectomy/ temporal bone surgery were included. Data on design, outcomes, and results were extracted by two reviewers. Validity evidence according to Messick's framework and level of evidence were assessed.

Results: The search yielded a total of 1085 studies from the years 1947–2018 and 35 studies were included for full data extraction after abstract and full-text screening. 33 different metrics on mastoidectomy performance were identified and ranked according to the number of reports. Most of the 33 metrics identified had some amount of validity evidence. The metrics with most validity evidence were related to drilling time, volume drilled per time, force applied near vital structures, and volume removed.

Conclusions: This review provides an overview of current metrics of mastoidectomy performance, their validity, strengths and limitations, and identifies the gap in validity evidence of some metrics. Evidence-based metrics can be used for performance assessment in temporal bone surgery and for providing integrated and automated feedback in virtual reality simulation training. The use of such metrics in simulation-based mastoidectomy training can potentially address some of the limitations in current temporal bone skill assessment and ease assessment in repeated practice. However, at present, an automated feedback based on metrics in VR simulation does not have sufficient empirical basis and has not been generally accepted for use in training and certification.

Level of evidence: 2A.

Keywords: Simulation-based training, temporal bone surgery, metrics, objective assessment, automatic evaluation

INTRODUCTION

High quality, evidence-based surgical skills and procedural training is essential to achieve the highest level of safe surgery and patient care.

Surgical skill training has traditionally been taught by the surgical trainer, who assumes the role of a coach and guides the surgical trainee towards competency by gradually increasing the surgical challenges [1]. Development of surgical skills requires surgical exposure, but this can be limited by ethical and legal concerns over patient safety, and financial costs related to training. As a result, the way we teach surgeons needs to adapt. Simulation-based training can be a part of the solution as it has proven a valuable and effective instrument for training and assessment in surgical education [2–5], including in otorhinolaryngology (ORL) [6, 7]. In general, simulation-based training has demonstrated large and positive effects on knowledge, skills, and behaviors, in addition to some effects on patient-related outcomes [8]. Furthermore, simulation-based training can provide a safe learning environment with a range of difficulty levels, clinical variations, and the opportunity for individualized learning [9, 10].

At many ORL training institutions, trainees practice temporal bone surgery in dissection labs, on surgical boot camps, or traditional temporal bone courses, using human cadaveric temporal bones (CTB), or plastic/plaster temporal bones (PTB) [11]. Such physical models for temporal bone training are expensive and especially human CTB are becoming scarce due to extensive regulation and fewer donations. In addition, the most common organization of temporal bone training such as boot camps and intensive training courses are isolated, single-instance practice opportunities that do not allow for distributed and repeated practice. From an educational point of view, this could be problematic as massed practice leads to poorer learning outcomes [12].

Over the past 20 years, several virtual reality (VR) simulators have been developed for high-fidelity representation of various procedures in ORL [12] including mastoidectomy. These enable self-directed learning, reducing the need of faculty for direct instruction [13] and making simulator practice more convenient and accessible for the individual trainee. Nonetheless, timely feedback is important for the development of competency regardless of whether the training modality is VR simulation, cadaveric dissection, or supervised surgery. Feedback can be formative with the purpose of adjusting performance during the procedure, or summative to provide a score typically at the end of the procedure (assessment). Summative feedback can be used to longitudinally monitor progression and to set standards for competency.

Both formative and summative feedback require data on the performance, and the simulated environment provides an optimal setting for gathering such performance data—simulator metrics—that can be used for providing the trainee with objective feedback and assessment. Some metrics in temporal bone surgery such as measurement of drilling time or violation of structures can be directly observed by a human assessor, but many metrics are difficult to quantify outside a computer-based VR simulation environment. These metrics could include force applied on the drill, the amount of bone removed per second, etc. Such computer-recorded metrics and derived scores for performance assessment and feedback must, however, be valid and reliable before implementation for routine use and high-stake assessment. A recent systematic review on the assessment of performance for mastoidectomy identified current assessment tools and their evidence [14]. However, simulation-based metrics for the assessment of mastoidectomy performance has not been reviewed and with many proposed metrics in the literature, there is a need to investigate the current validity evidence for their use in feedback and assessment.

In this systematic review, we aim to investigate simulation-based metrics in temporal bone surgery and the current validity evidence supporting their use for feedback and assessment of mastoidectomy performance.

METHODS

Eligibility criteria

We included all studies on performance metrics of mastoidectomy/temporal bone surgery. Any training modality was accepted including cadaveric temporal bones (CTB), artificial temporal bones, VR simulation, or supervised surgery. Types of studies eligible for inclusion were randomized and non-randomized trials, observational studies, feasibility studies, and technical descriptions. We excluded commentaries, letters to the editor, conference abstracts and reviews.

All types of participants were considered, including all types of trainees such as medical students, interns, residents, and more experienced surgeons such as fellows and consultants. If the study was interventional, all types of interventions were accepted.

Search methods

We followed the PRISMA guidelines for systematic reviews [15]. The search strategy was designed to access both published and unpublished English literature by searching the

following databases: Medline (Pubmed), Embase/Ovid, Cochrane Library, PLoS, BMC, OpenGrey, Google Scholar, DOAJ. See Online Appendix I for a description of the databases. For all databases, we used the combination of search terms provided in Fig. 1. Furthermore, the following MeSH terms were used: "otology", "temporal bone", "assessment, education" and reference lists of the articles gathered from the extensive search were also examined. The search was last updated on 18 March 2018.

Study selection

Results were imported into reference management software, EndNote (Thomsen Reuters, USA). Two reviewers (FA and SA) independently searched the databases and included articles that met the criteria by initially screening titles and abstracts, and finally screened the full text articles as indicated in Fig. 2. Duplicates were identified by software, confirmed by a reviewer, and excluded. Articles identified through reference lists were also considered for data collection based on their title. All discrepancies were resolved by a discussion between the reviewers, and studies were excluded only in case of consensus.

Data extraction

The two reviewers extracted the following data using a modified data extraction form from the Cochrane Review Group (Online Appendix II): basic data related to the study; eligibility; methods; risk of bias assessment; participant characteristics; intervention type; information on training, assessment or simulator design; simulator metrics and outcomes; results; supporting evidence for metrics or assessment according to Messick's framework for validity evidence; and level of evidence. Forms were piloted with five randomly chosen full-text articles to ensure consist- ency of the interpretation of data fields between reviewers. Study authors were contacted in cases of missing information.

Quality assessment

The quality of the papers was evaluated based on integrity and validity evidence. Level of evidence was evaluated according to the Oxford Centre of Evidence-based Medicine (CEBM 'Levels of Evidence 1'—Online Appendix III). Finally, the validity of the assessment tools according to Messick's framework was evaluated based on a scoring system [16] (Online Appendix IV). The framework consists of five aspects: content, response process, internal structure, relation to other variables and consequences. The two main validity measures considered were "reliability" (interrater and/or internal consistency) and "discriminative

evidence" (whether these tools generated scores that could differentiate trainees with different levels of skills). Reliability and discriminative evidence can be categorized under "internal structure" and "relations to other variables", respectively [16]. Although we assessed all included studies based on the five categories, the main focus was on metrics with high validity evidence in relation to "internal structure" and "relation to other variables". Any disagreement was resolved by discussion.

RESULTS

From a total of 1085 studies identified, 35 were included for analysis, with publication dates ranging from 1947 to 2018. Table 1 provides an overview of study characteristics. The two most frequent study designs were observational and interventional trials. 94% of the studies considered VR TB simulation, whereas CTB was the second-most reported modality. 63% of the included studies had residents as study subjects, 49% experts, and 43% medical students (some studies included several groups). See Online Appendix VI for an annotated list of all included studies.

A total of eight controlled trials (seven randomized and one non-randomized) were identified and assessed for risk of bias (Online Appendix VII). Blinding of participants is often not possible when assessment consists of direct observation during an ongoing procedure, and we have, therefore, not considered the trials for risk of bias due to blinding of participants. Only two trials were deemed of low risk of bias. The rest of the trials had unclear or high risk of bias primarily regarding random sequence generation and allocation concealment as this most often was not detailed.

Table 2 provides an overview of the metrics described and used in the included studies. A total number of 33 metrics were identified and ranked after validity evidence. Validity evidence was based on study conclusions, weighting of number of studies, number of participants, and level of evidence. As an example, metrics considered with substantial validity evidence had several studies of significant conclusion, a high number of participants, high level of evidence, and few studies with contradictory validity evidence.

The most well-established metrics, based on these criteria were: (1) time, (2) volume removed per second/efficiency, (3) force applied near vital structures, and (4) volume removed compared to reference/Euclidian Algorithm. These four metrics have clear evidence supporting a correlation between performance and expertise by comparison of two or more levels of experience, for example, novices, intermediates and experts. The metrics with least validity were: volume removed, distance between each 1000 voxels removed, pct. circular

strokes and identification of structures in correct order. Poor validity evidence was a result of either contradiction between strong studies showing various positive results, or results supported by too few studies with a small number of participants. Online Appendix V provides a full list of identified metrics and the associated results, validity and level of evidence.

DISCUSSION

This systematic review is the first on performance metrics in mastoidectomy and the identified metrics were classified according to evidence-based frameworks and cur- rent validity evidence. Several metrics were supported by a substantial amount of validity evidence: time, volume removed per second/efficiency, force applied near vital structures, and volume removed compared to reference. Most evidence was related to reliability and the ability to discriminate novice from expert performance. Overall, there was a predominance of observational studies and trials.

The metric supported by most validity evidence reflects different aspect of time—for example, drilling time, non- drilling time, and total time—with a total of 11 studies and 263 subjects (Online Appendix V). Time is a metric of fine granularity, which enables good discriminative precision. Moreover, time is easily measured either automatically by the simulator or manually by an observer. Time to completion was found to have substantial discriminative validity for different levels of experience. However, most of the included studies compared only performance between novices and experts, and only two studies compared three levels of expertise (novices, trainees and experts) [17, 18]. Validity on internal structure has been investigated by comparing VR TB with CTB performance, thereby establishing reliability. For validity evidence on process response, variation in time to completion on different modalities was found: one study demonstrated that experts used less time in VR mastoidectomy than in CTB mastoidectomy [19]. The explanation for this was identified by investigating another metric—the "number of burr changes"—because it was found that changing burr was easier and faster in the VR environment compared to cadaveric dissection. Metrics can, therefore, be correlated, which should be considered. Although supported by some validity studies, time and efficiency such as "volume removed per time unit" are essentially non-specific metrics, which do not address any unequivocal component of competency on their own. Compared to procedure-specific metrics such as proximity/collisions with limiting structures and "volume removed compared to reference"

that has the purpose of adjusting performance during the procedure, time and efficiency are not useful for formative feedback.

The metric “force applied near vital structures” (such as the sigmoid sinus, dura and facial nerve) was also sup- ported by a substantial amount of validity evidence. For this metric, three or more different levels of expertise can be significantly differentiated [17, 18, 20–22]. Consistency in scoring the same participant by different raters demonstrated high inter-rater reliability [22]. Additionally, the metric is also supported by content validity evidence. Studies have, however, approached this metric differently: some have investigated the force applied directly on the vital structures, whereas others have studied the force applied on bone voxels close to the structure.

As examples of metrics that have a limited validity evidence, percentage of circular strokes and number of burr changes are currently only supported by expert opinion. Further trials are needed to gather better validity evidence data for these metrics. Finally, several of the proposed metrics need clearer definitions as well as a hypothesis for the expected outcome in relation to, for example, novice/expert performance. A minor limitation to this study was classifying the different metrics, due to different definitions or names of the same metrics.

It is also important to consider that even though metrics are supported by validity evidence, they are often specific for the simulator or context: if, for example, the transparency function of a VR simulator is very poor technically, neither novices nor experts will be able to thin the bone over the dura without unintended soft tissue exposure. This can result in many injuries to the dura in both groups, causing a metrics such as identify dura without exposing it difficult to achieve in first place and second result in poor discriminative power. Graphic detail and haptic realism are, therefore, essential for realistic performance in any simulation-based training modality.

This should be considered when results from different simulators are pooled and analysed. Moreover, it adds another modification to the interpretation of a given metric: the validity of a metric in test may be low, not because it is not surgically important or clinically irrelevant, but simply because current simulator fidelity is too low.

Most of the described metrics we have found in the literature have not been exposed to structured and evidence- based assessment, currently limiting their use in training and certification. Structured assessment requires metrics assessing different categories—ability, skill, task and procedure. Satava et al. [23] have proposed a way to categorize performance metrics according to (1) ability—the natural state or condition of being capable, aptitude; (2)

skill—a developed proficiency or dexterity in some art, craft, or the like; (3) task—a piece of work to be done, a difficult or tedious undertaking; (4) procedure—a series of steps taken to accomplish an end [23]. The identified metrics in this systematic review were mainly distributed in the categories of ability, skill, and task, because the procedure was predefined by inclusion criterion to be mastoidectomy.

Another requisite for structured assessment would be establishing evidence-based cut-off scores to determine a pass-fail standard or multiple levels of competency [24]. Trainees can be divided in the following five levels of competency: novice, competent, proficient, expert and master [23]. These levels of overall competency suggest a framework that allows for the measuring of progress in performance, when advancing from one level to another. Further work is needed to validate the five different levels and the performance, in terms of a combined metrics-based score, required before advancing between the levels.

In VR simulation, metrics-based scoring may provide a standardized and objective performance assessment and summative feedback. Moreover, metrics can support automatic formative feedback throughout the training process, forming the basis for a self-directed mastoidectomy training [10]. Integrating metrics for self-directed training could provide a cost-effective setup for learning basic temporal bone skills, most likely optimizing the time needed to reach the necessary competency for progression to supervised surgery, and reducing dependency on expensive human mentor-based evaluation [10] in the initial stages of training. This would also align with increased focus on patient safety because trainees can effectively practice in a safe and controlled simulation-based environment before commencing supervised surgery. Altogether, evidence-based performance metrics could promote self-directed, repeated practice of temporal bone surgery if used for validated feedback and assessment.

CONCLUSION

A large number of performance metrics in mastoidectomy was identified and the current validity evidence supporting these was investigated. Several metrics had substantial or moderate validity evidence and these metrics could potentially be used for objective and automated feedback and assessment in self-directed simulation-based temporal bone training. Currently, automated feedback based on performance metrics have insufficient empirical basis and has not been generally accepted for use in training and certification, and further research is needed to translate simulation-based performance metrics into valid and reliable assessment.

REFERENCES

1. Reznick RK (1993) Teaching and testing technical skills. Am J Surg 165 (3):358-36
2. Cook DA (2014) How much evidence does it take? A cumulative meta-analysis of outcomes of simulation-based education. Medical education 48 (8):750-760. doi:10.1111/medu.12473
3. Draycott TJ, Crofts JF, Ash JP, Wilson LV, Yard E, Sibanda T, Whitelaw A (2008) Improving neonatal outcome through practical shoulder dystocia training. Obstet Gynecol 112 (1):14-20. doi:10.1097/AOG.0b013e31817bbc61
4. McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB (2011) Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. Acad Med 86 (6):706-711. doi:10.1097/ACM.0b013e318217e119
5. Vaughan N, Dubey VN, Wainwright TW, Middleton RG (2016) A review of virtual reality based training simulators for orthopaedic surgery. Medical engineering & physics 38 (2):59-71. doi:10.1016/j.medengphy.2015.11.021
6. Andersen SA, Foghsgaard S, Konge L, Caye-Thomasen P, Sorensen MS (2016) The effect of self-directed virtual reality simulation on dissection training performance in mastoidectomy. The Laryngoscope 126 (8):1883-1888. doi:10.1002/lary.25710
7. Fried MP, Sadoughi B, Gibber MJ, Jacobs JB, Lebowitz RA, Ross DA, Bent JP, 3rd, Parikh SR, Sasaki CT, Schaefer SD (2010) From virtual reality to the operating room: the endoscopic sinus surgery simulator experiment. Otolaryngology--head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery 142 (2):202-207. doi:10.1016/j.otohns.2009.11.023
8. Cook DA, Hatala R, Brydges R, Zendejas B, Szostek JH, Wang AT, Erwin PJ, Hamstra SJ (2011) Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA 306 (9):978-988. doi:10.1001/jama.2011.1234
9. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ (2005) Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Medical teacher 27 (1):10-28. doi:10.1080/01421590500046924
10. Sachdeva AK, Buyske J, Dunnington GL, Sanfey HA, Mellinger JD, Scott DJ, Satava R, Fried GM, Jacobs LM, Burns KJ (2011) A new paradigm for surgical procedural training. Curr Probl Surg 48 (12):854-968. doi:10.1067/j.cpsurg.2011.08.003
11. Frithioff A, Sorensen MS, Andersen SAW (2018) European status on temporal bone training: a questionnaire study. European archives of oto-rhino-laryngology : official journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS) : affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery 275 (2):357-363. doi:10.1007/s00405-017-4824-0
12. Bhutta MF (2016) A review of simulation platforms in surgery of the temporal bone. Clinical otolaryngology : official journal of ENT-UK ; official journal of Netherlands Society for Oto-Rhino-Laryngology & Cervico-Facial Surgery 41 (5):539-545. doi:10.1111/coa.12560
13. Arora A, Hall A, Kotecha J, Burgess C, Khemani S, Darzi A, Singh A, Tolley N (2015) Virtual reality simulation training in temporal bone surgery. Clinical otolaryngology : official journal of ENT-UK ; official journal of Netherlands Society for Oto-Rhino-Laryngology & Cervico-Facial Surgery 40 (2):153-159. doi:10.1111/coa.12352
14. Sethia R, Kerwin TF, Wiet GJ (2016) Performance Assessment for Mastoidectomy: State of the Art Review. Otolaryngology--head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery. doi:10.1177/0194599816670886
15. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 151 (4):264-269, W264

16. Ghaderi I, Manji F, Park YS, Juul D, Ott M, Harris I, Farrell TM (2015) Technical skills assessment toolbox: a review using the unitary framework of validity. Ann Surg 261 (2):251-262. doi:10.1097/SLA.0000000000000520
17. Khemani S, Arora A, Singh A, Tolley N, Darzi A (2012) Objective skills assessment and construct validation of a virtual reality temporal bone simulator. Otology & neurotology : official publication of the American Otological Society, American Neurotology Society [and] European Academy of Otology and Neurotology 33 (7):1225-1231. doi:10.1097/MAO.0b013e31825e7977
18. Zhao YC, Kennedy G, Hall R, O'Leary S (2010) Differentiating levels of surgical experience on a virtual reality temporal bone simulator. Otolaryngology--head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery 143 (5 Suppl 3):S30-35. doi:10.1016/j.otohns.2010.03.008
19. Ioannou I, Avery A, Zhou Y, Szudek J, Kennedy G, O'Leary S (2014) The effect of fidelity: how expert behavior changes in a virtual reality environment. The Laryngoscope 124 (9):2144-2150. doi:10.1002/lary.24708
20. Linke R, Leichtle A, Sheikh F, Schmidt C, Frenzel H, Graefe H, Wollenberg B, Meyer JE (2013) Assessment of skills using a virtual reality temporal bone surgery simulator. Acta otorhinolaryngologica Italica : organo ufficiale della Societa italiana di otorinolaringologia e chirurgia cervico-facciale 33 (4):273-281
21. Morris D, Sewell C, Barbagli F, Salisbury K, Blevins NH, Girod S (2006) Visuohaptic simulation of bone surgery for training and evaluation. IEEE computer graphics and applications 26 (6):48-57
22. Sewell C, Morris D, Blevins NH, Dutta S, Agrawal S, Barbagli F, Salisbury K (2008) Providing metrics and performance feedback in a surgical simulator. Computer aided surgery : official journal of the International Society for Computer Aided Surgery 13 (2):63-81. doi:10.3109/10929080801957712
23. Satava RM, Cuschieri A, Hamdorf J, Metrics for Objective Assessment of Surgical Skills W (2003) Metrics for objective Assessment. Surg Endosc 17 (2):220-226. doi:10.1007/s00464-002-8869-8
24. Konge L, Annema J, Clementsen P, Minddal V, Vilmann P, Ring- sted C (2013) Using virtual-reality simulation to assess perfor- mance in endobronchial ultrasound. Respir Int Rev Thorac Dis 86(1):59–65. https://doi.org/10.1159/000350428
25. Kerwin T, Wiet G, Stredney D, Shen H-W (2012) Automatic scoring of virtual mastoidectomies using expert examples. Int J Comput Assist Radiol Surg 7(1):1–11. https://doi.org/10.1007/ s11548-011-0566-4

Table 1. Characteristics of the included studies.

Study characteristics	Studies, n (%)	Participants, n
All studies	35 (100)	719
Study design		
Technical description	7 (20)	188
Observational	19 (54)	346
Expert opinion	1 (3)	0
Trials		
Randomized Controlled	7 (20)	175
Non-randomized Controlled	1 (3)	34
Training model		
Virtual Reality Temporal Bone	33 (94)	675
Cadaveric Temporal Bone	6 (17)	111
Plastic Temporal Bone	1 (3)	44
Patients	1 (3)	4
Participants		
Novices/Medical students	15 (43)	227
Residents	22 (63)	364
Experts	17 (49)	128

Table 1. An overview of all metrics in the included studies ranked by validity evidence.

	Content (0–3) Max (mean)	Process response (0–3) Max (mean)	Internal structure (0–3) Max (mean)	Relation to other variables (0–3) Max (mean)	Consequences (0–3) Max (mean)
Substantial validity evidence					
Time (total/drilling/non-drilling)	3 (2.1)	3 (2.1)	2 (1.1)	3 (1.9)	2 (1.6)
Volume removed per sec/efficiency	2 (1.9)	3 (2.1)	2 (1.2)	3 (1.8)	3 (1.2)
Force applied near vital structures[a]	3 (1.9)	3 (2.1)	2 (1.2)	3 (2.1)	3 (1.8)
Volume removed compared to reference/Euclidian Algorithm[b]	2 (2)	3 (2.4)	2 (1.4)	3 (2.2)	3 (1.5)
Moderate validity evidence					
Number of injuries on vital structures[a]	3 (2)	3 (2)	3 (1.3)	3 (2.1)	1 (1)
EMD Algorithm[c]	2 (2)	2 (2)	1 (1)	2 (2)	2 (2)
Time burr not visible	2 (2)	2 (2)	2 (1.3)	3 (1.8)	2 (2)
Angle of drilling	3 (2)	3 (2)	3 (1.9)	3 (2.5)	3 (1.9)
Burr diameter (mean/median)	3 (2.1)	3 (2)	3 (1.6)	3 (2.3)	2 (2)
Stroke distance	3 (2)	3 (2.1)	3 (1.5)	3 (2.2)	3 (2.3)
Identification of structures in correct time	2 (1.8)	2 (1.6)	3 (1.8)	3 (2.4)	3 (1.5)
Identification of vital structures[a]	2 (1.8)	2 (1.5)	3 (2)	3 (2.3)	2 (2)
Direct exposure[d]	3 (2.2)	3 (2.2)	3 (2)	3 (2.5)	2 (1.5)
Some validity evidence					
Stroke velocity	3 (1.9)	3 (2.2)	2 (1.2)	3 (2.1)	3 (1.8)
Inferred exposure[e]	3 (2)	3 (2)	2 (1.3)	3 (2)	2 (1.7)
Number of procedure steps completed (predefined in simulator)	2 (1.5)	2 (1.5)	2 (1.5)	3 (2.5)	1 (1)
Stroke duration	3 (2)	3 (2.2)	2 (1.2)	3 (2)	2 (2)
Strokes per second	3 (2.3)	3 (2.3)	1 (1)	3 (2)	1 (1)
Volume removed with drill obscuring view	3 (2.3)	3 (2)	3 (2)	3 (2.8)	2 (1.3)
Volume removed with correct drill	2 (2)	3 (3)	2 (2)	3 (3)	1 (1)
Volume removed with bone dust reducing visibility	2 (2)	3 (3)	2 (2)	3 (3)	3 (1.5)
Inter-tool distance	2 (2)	3 (3)	2 (1.7)	3 (2.7)	3 (1.6)
Little or no validity evidence					
Pct. straight strokes	3 (2)	2 (2)	2 (1.3)	3 (2.3)	2 (2)
Number of burr changes	3 (2.5)	3 (2.5)	1 (1)	2 (1.5)	3 (1.5)
Time spent with drill obscuring view	2 (2)	3 (2.5)	2 (1.5)	3 (2)	1 (1)
Stroke force	2 (2)	3 (3)	1 (1)	1 (1)	2 (1.5)
Number of bone movements	2 (2)	3 (3)	1 (1)	1 (1)	3 (1.6)
Number of magnification changes	2 (2)	3 (3)	1 (1)	1 (1)	3 (1.9)
Stroke path in relation to structures[a]	3 (2)	1 (1)	1 (1)	2 (2)	3 (2.2)
Identification of structures in correct order	2 (1.5)	1 (1)	1 (1)	2 (1.5)	1 (1)
Pct. circular strokes	3 (3)	2 (2)	1 (1)	2 (2)	1 (1)
Distance between each 1000 voxels removed ("drill jumping")	2 (2)	3 (2.8)	2 (1.5)	3 (2.8)	3 (2)
Volume removed	3 (1.9)	3 (1.9)	3 (1.3)	3 (2)	2 (1.3)

Validity evidence is based on study conclusions, weighting number of studies, number of participants and level of evidence (**Appendix V**).

a. Vital structures: The sigmoid sinus, brain, dura, ossicles, facial nerve, inner ear structures, external auditory canal

b. Euclidean algorithm: A calculation of whether "should-be removed" voxels are removed and "should-not be removed" voxels are not using expert performance as a reference.[26]

c. EMD: Uses the same principle as the Euclidian algorithm, and additionally takes into consideration if wrongly removed voxels are near an area where the expert removed voxels, and vice versa for wrongly kept voxels.[26]

d. Direct exposure: Removing voxels that exposes a vital structure.

e. Inferred exposure: Some voxels along a structure are removed and thereby the surgeon may infer the location of the other nearby points of the structure.

Mastoidectomy		Assessment		Training
Mastoid surgery		Metric		Virtual
Temporal bone surgery	AND	Evaluation	AND	Simulat*
		Education		Computer aided
		Scale		Dissection
		Scor*		Education
		Performance evaluation		
		Objective assessment		
		Automatic evaluation		

Figure 1. Search strategy.

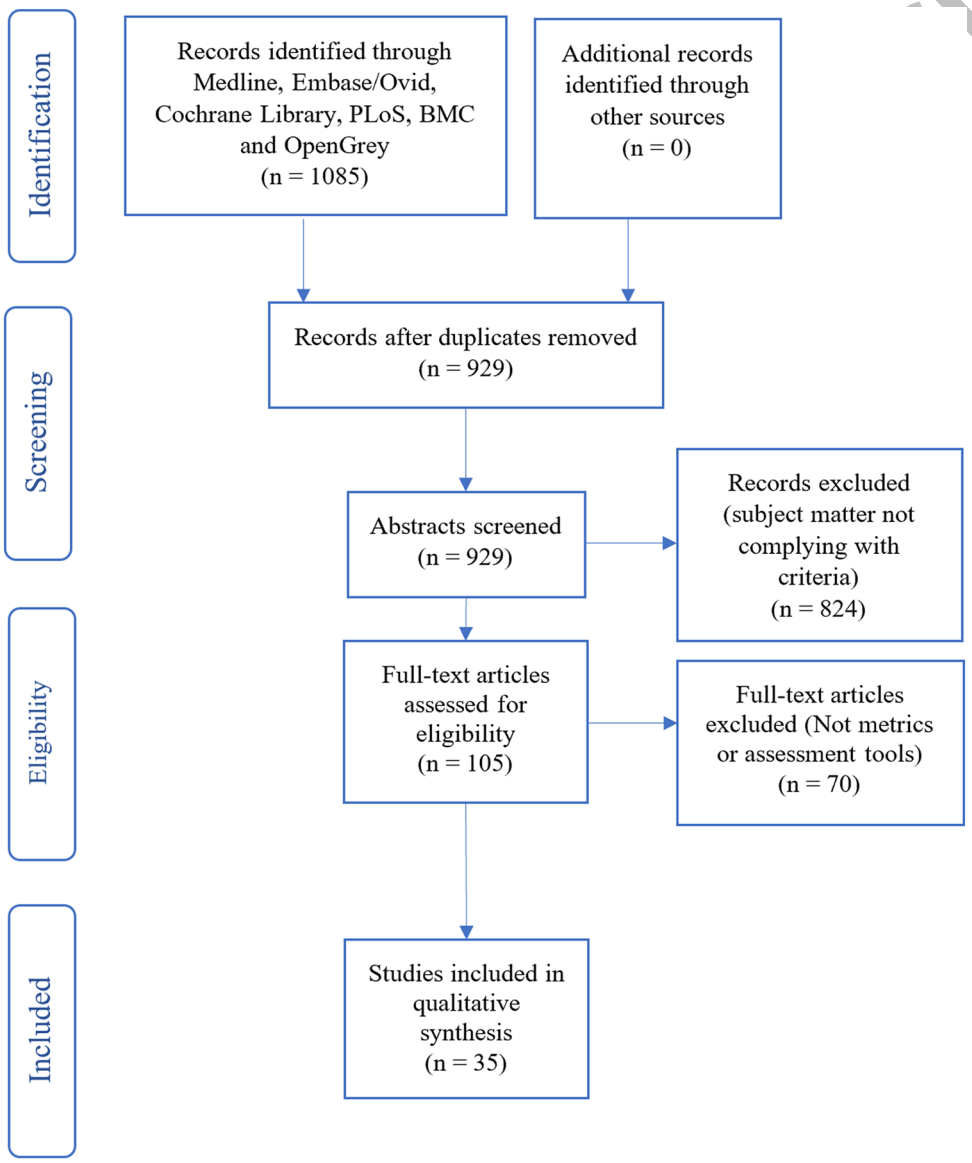

Figure 2. Flow diagram.