Expert sampling of VR simulator metrics for automated assessment of mastoidectomy performance

Steven Arild Wuyts Andersen, MD, PhD (1,2); Peter Trier Mikkelsen, MSc (3), Mads Sølvsten Sørensen, MD, DMSc (1)

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. 3. The Alexandra Institute, Aarhus, Denmark.

Correspondence: Steven Andersen, MD, PhD, Department of Otorhinolaryngology—Head & Neck Surgery, Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen, Denmark. E-mail: stevenarild@gmail.com.

Financial disclosures: None.
Conflicts of interest: None.

Full citation: Andersen SA, Mikkelsen PT, Sørensen MS. Expert sampling of VR simulator metrics for automated assessment of mastoidectomy performance. Laryngoscope. 2019 Sep;129(9):2170–2177.
DOI: 10.1002/lary.27798

OBJECTIVE: Often assessment of mastoidectomy performance requires time-consuming manual rating. Virtual reality (VR) simulators offer potentially useful automated assessment and feedback but should be supported by validity evidence. We aimed to investigate simulator-metrics for automated assessment based on the expert performance approach, comparison with an established assessment tool, and the consequences of standard-setting.

METHODS: The performances of 11 experienced otosurgeons and 37 otorhinolaryngology residents. Participants performed three mastoidectomies in the Visible Ear Simulator. Nine residents contributed additional data on repeated practice in the simulator. 129 different performance metrics were collected by the simulator and final-product files were saved. These final-products were analyzed using a modified Welling Scale by two blinded raters.

RESULTS: 17 metrics could discriminate between resident and experienced surgeons' performances. These metrics mainly expressed various aspects of efficiency: Experts demonstrated more goal-directed behavior and less hesitancy; and they used less time, and selected large and sharp burrs more often. The combined metrics-based score (MBS) demonstrated significant discriminative ability between experienced surgeons and residents with a mean difference of 16.4 % (95 % CI [12.6–20.2], $p<<0.001$). A pass/fail score of 83.6 % was established. The MBS correlated poorly with the final-product score but excellently with the final-product score per time.

CONCLUSION: The metrics-based score mainly reflected efficiency components of the mastoidectomy procedure and even though it could have some uses in self-directed training, it fails to measure and encourage safe routines. Supplemental approaches and feedback are therefore required in VR simulation training of mastoidectomy.

Level of Evidence: 2b.

Key-words: temporal bone surgery, mastoidectomy, virtual reality surgical simulation, automated assessment, simulation-based training.

INTRODUCTION

Virtual reality (VR) simulation training has been established as a useful adjunct to traditional cadaveric dissection training for trainees in temporal bone surgery. Several VR temporal bone surgical simulators have been developed[1–5] and are found at many otorhinolaryngology training departments[6] where they allow trainees self-directed and repeated practice, serving the individual trainee's needs. It is well-known that deliberate and repeated practice is key in developing competency and expertise.[7] Nonetheless, ensuring sufficient skills and proficiency by self-directed VR simulation training is a challenge: learners cannot be trusted to have adequate insights into their own performance, resulting in a learning curve plateau at an unsatisfactory level.[8,9]

Automated assessment and feedback for self-directed training is found out of the box in many VR surgical simulators. However, most built-in metrics and scoring systems provided by the developers have little evidence, and often, subsequent scientific validation of metrics-based assessment and feedback is needed.[10–12] In temporal bone surgery, performance metrics in VR simulation of mastoidectomy have been part of the research agenda since early development of such systems.[13] Metrics and scoring systems have been further refined in several of the available VR temporal bone surgical simulators and studies have investigated various aspects of validity evidence of metrics for automated assessment and feedback[2,14–17]. Concurrently, a number of structured, objective assessment tools for mastoidectomy performance have been reported.[18] Common to these assessment tools is that they require multiple experts to perform either lengthy direct observation, or end-of-procedure final-product analysis. Consequently, such external assessment is inconvenient in the setting of self-directed practice whereas simulator-based automated assessment and feedback could provide a feasible alternative. Solid validity evidence for this automated simulation is needed before implementation into the training curriculum and should be assessed in relation to a contemporary validity framework such as Messick's.[19]

Most of the proposed mastoidectomy performance metrics do not have obvious standards or cut-off values for proficiency. To define these, different approaches can be considered. A consensus-based approach, which has worked fine in the development of objective, structured assessment tools,[20] can be difficult considering the complexity of simulator-gathered performance metrics such as volume removed per minute, path length,

and forces applied. Also, it can be argued that experts do not always practice what they teach and furthermore, it has been demonstrated that expert behavior changes in a VR simulation environment compared with human cadaveric temporal bone dissection.[21] In the expert performance framework[22], the superior performance of experts is used in the design of learning experiences. This expert performance model has previously been used for one approach to automatic scoring of mastoidectomy performance in VR temporal bone surgical simulation.[23] One goal of such automated assessment could be mastery learning[24] for self-directed training. However, the change in expert behavior in the simulation environment is a concern and any simulation-based automatic assessment should be validated against established assessment tools.

In this study, we used the expert performance approach to investigate automated scoring of mastoidectomy based on metrics from a VR temporal bone simulator compared with traditional final-product assessment. Further we aimed to explore the consequences of using such metrics-based assessment for standard-setting and in repeated practice.

MATERIAL AND METHODS

Setting and participants

The contributing 11 experienced otosurgeons were recruited by invitation of selected national and international colleagues, mainly from the Nordic countries. The first author visited each of the experts at their home institution in the period March 2015 to May 2016, bringing a fully working simulator running on a laptop for data collection. The experts were recruited to represent different schools of temporal bone surgery and all were sufficiently competent in mastoidectomy and temporal bone surgery to be considered well-experienced by their peers. Background data and characteristics are provided in Table I.

Data on resident performances were collected during the national Danish temporal bone courses in 2016 and 2017 in accordance with our study on the effect of structured simulation training during these courses.[25] 37 post-graduate years 2 to 5 residents contributed as reported previously; 9 further completed additional repeated and distributed practice which in the current study was used in analysis of effects of repeated practice. All residents had very limited experience with simulated and no experience with real-life temporal bone surgery as this requires subspecialized training in Denmark.

Study design

Experienced surgeons and residents were asked to perform 3 identical procedures consisting of a complete anatomical mastoidectomy and posterior tympanotomy in the VR simulator according to standard course instructions.[26] The first procedure was guided by the simulator-integrated tutor-function, which greenlights the volume to be drilled in correspondence with each step of the on-screen guide of the procedure. For the following two procedures, the experienced group was not guided except for a single static screenshot of the completed procedure that served as a reference, whereas residents continued to have access to the step-by-step guide as a learning resource but not the simulator-integrated tutor-function with greenlighting.

VR simulation platform and data sampling

The freeware Visible Ear Simulator[27] which has been thoroughly described in previous publications[5,28] was used in this study. For the purpose of expert sampling, an experimental version of the Visible Ear Simulator (version 2.1) was developed. This version of the simulator was designed to record a range of metrics and derivatives during the procedure. These metrics were a priori identified by the research group as potentially relevant and technically possible at current (see Supplemental Digital Content, Table A4). Some metrics were compound data such as time drilled, others were more specific such as for example vector data for stroke path length, while some metrics were even recorded at the individual voxel level such as visibility, force, drill type, and drill size. Some metrics such as collision data were directly available for analysis. However, the complexity of the remaining data required two supporting analysis tools to be developed: the first tool could visualize the individual metric recordings such as drilling vectors and heat maps of for example drilling force. This tool was used to verify the accuracy and alignment of the recorded metrics. The second tool could for each recording summarize and export overall metrics and averages of the high-complexity (granular) data. In addition to simulator-gathered metrics, the final product of each procedure was saved and scored (final-product score) using a modified Welling Scale for final-product analysis[29] by two expert raters (SA and MS) blinded to participant, simulator tutoring, expertise, procedure number, time

used, and metrics score. The inter-rater reliability (Cohen's κ) was 0.67 similar to previous reports.[29,30]

Data analysis, outcomes and statistics

Data were analyzed in SPSS (SPSS Inc., IL, USA) version 23 for MacOS X. Linear mixed models were used to account for repeated measurements and hierarchical data. In the analysis of the final-product performance the model included level (experts and residents), tutored session, and rater, as fixed factors. In the analysis of which metrics to use in a combined score based on the expert performance model, these were identified as those that: A) could discriminate between residents and experienced surgeons ($p<0.10$) with B) experienced surgeons performing better than residents, and C) performance improving with repetition, and D) metrics that did not directly overlap. A cut-off score for each metric and the final-product performance was established as the upper 80 % CI bound of experienced surgeons in session 3 (if a higher value indicated a better performance) or the lower 80 % CI bound of experienced surgeons in session 3 (if a lower value indicated a better performance) meaning that the cut-off scores were set as the 10 % best of the experienced surgeons' performance. For each metric, the individual metric score was calculated as a percentage of the cut-off value; if performance exceeded the cut-off value the score was set at 100 %. Next, factor analysis (principal components analysis) was used to reduce the number of dimensions and classify score components with a coefficient >0.40 due to correlation between metrics. This was used to provide a weighted metrics-based score (MBS). A pass/fail score cut-off for this weighed score was established as the upper 80 % CI bound of the experienced group. For the investigation on the effects of repeated practice, performances were included only if a minimum volume was removed, corresponding to the number of voxels the experienced surgeons would minimally remove in the procedure. Validity evidence was included in relation to the five sources of validity evidence in the framework of Messick (Appendix, table A1).

Ethics

The regional ethics committee for the Capital Region of Denmark deemed this study to be exempt (H-15002506). All participants volunteered for the study and signed informed consent.

RESULTS

Final-product performance

The experienced surgeons performed significantly better than residents in both the final-product score (FPS) (Table 2A) and FP score per minute (Table 2B). For both groups, simulator-integrated tutoring had significant effects on performance: a positive effect on the FPS and a negative effect on the FPS per minute. Both groups demonstrated learning curves and increased performance with repeated practice (Supplemental digital content, Figure A2, means plot of FPA and FPA/min performance). For the FPS, a cut-off score (for non-tutored sessions) of 19.5 points was established. For the FPS per minute, a cut-off score of 0.82 was established. The consequence of this standard-setting (Supplemental digital content, Table A3) was that by the third procedure, 60 % of the experienced surgeons' performances passed the standard for FPS and 30 % the standard for FPS per minute. For the residents, the corresponding pass rates were 5.4 % and 2.7 %.

Simulator metrics performance

129 different metrics were investigated (Supplemental digital content, Table A4). 17 metrics fulfilled the inclusion criteria and were selected for the final metrics-based scoring model (Table 3).

Next, consequences of standard setting of the cut-off scores (pre-defined as the 10 % best of performances of the experienced surgeons) were explored to ensure validity. For three of the metrics, this cut-off resulted in very few experienced surgeons passing this standard in the third procedure: the average force (20 % passing), the number of drill jumps (0 % passing), and the average force when drilling with fine diamond burrs (10 %). Therefore, the cut-off for these metrics were instead set at the mean performance of the experienced group in the third procedure.

Finally, a components analysis classified the 17 metrics within 5 dimensions, which we have named "*Time and force efficiency*", "*Burr size efficiency*", "*Hesitancy*", "*Burr type efficiency*" and "*Goal-directed behavior*" based on their main metric components (Table 3). Each of the dimensions demonstrated statistically significant ability to discriminate between the experienced surgeons and residents, and consequently they were weighted equally in the calculation of the combined metrics-based score (MBS).

The MBS demonstrated significant discriminative validity between experienced surgeons and residents (mean difference 16.4 %, 95 % CI [12.6 %–20.2 %], $p<<0.001$, linear mixed models). Both experienced surgeons and residents demonstrated a learning curve for the MBS (Table 4). A cut-off score of 83.6 % (linear mixed models, cut-off mean performance of experienced surgeons in their 3rd procedure) was established. The consequence of this standard was that 5.4 % of residents and 60.0 % of experienced surgeons having a passing performance in the 3rd procedure (Table 4).

Effects of repeated practice

The minimum volume to be removed was based on the least amount of volume removed by any experienced surgeon 1,222,963 voxels. Consequently, 31/140 (22.1 %) of the resident performances in the repeated practice dataset were considered insufficient drillings and were excluded in the following analysis. Improved performance with repeated practice was found for most of the 17 individual metrics (Table A4). However, throughout all the procedures, residents used too much force on fine diamond burrs and small burrs compared with experienced surgeons and did not demonstrate much improvement. The MBS demonstrated a traditional negatively accelerated learning curve with repeated practice (Figure 1, top) and the number of total passing procedures increased with repeated practice as well (Figure 1, bottom).

The MBS demonstrated a poor correlation with the final-product score ($r^2=0.09$). Nonetheless, the MBS was found to be well-correlated with the final-product score per minute ($r^2=0.40$) (Figure 2), with intersection of the passing-standard for the MBS and FPS per minute closely to the linear fit.

DISCUSSION

In this study, we used the expert performance approach to investigate a large number of simulation-based metrics for potential automated assessment of mastoidectomy performance. We found that only 17 out of 129 metrics could discriminate between resident and experienced surgeons' performances with experienced surgeons having the better performance thereby providing validity evidence. The remaining metrics, did either not discriminate between resident and experienced surgeon performance or did not improve with repeated practice. For five metrics (metric #87, #94, #102, #117, #118,

Supplemental Table A4) residents were found to perform better than experienced surgeons, but these for these metrics, there is either no meaningful interpretation (for example the number of voxels removed with 5 mm burrs) or—for collisions with vital structures (chorda and digastric muscle)—related to either simulator fidelity or that residents did not reach these structures and therefore made no collisions.

The 17 included metrics mainly expressed various aspects of efficiency such as hesitancy, goal-directed behavior, time consumption, and more use of large and sharp burrs. This corroborates studies on other VR temporal bone simulators[14,15]—namely that experienced surgeons drill more efficiently than residents by using less time, applying more force, and using larger burrs. The large "efficiency" component of the metric-based score explains the poor correlation between the mainly "efficiency" metrics-based score and the final-product score, which includes several safety-related items that are not considered in the MBS. Moreover, this MBS efficacy bias can explain the excellent correlation with the final-product score *per time*, which fundamentally reflects drilling efficiency. Likely, inadequate simulator fidelity also contributes to the lack of discriminative validity for these safety-related metrics. Indeed, data indicate that novices and experts alike regularly fail instructions and inadvertently expose critical structures such as the facial nerve or the dura in the simulator. Visual cues are important for this and it seems expert performance in simulation requires higher graphic fidelity and haptic realism.[21]

In the repeated practice data, an insufficient volume of bone was drilled in many residents' performances and the average FPS of these performances were significantly lower than that of adequately drilled performances (mean score of 14.0 vs. 17.5, $p<<0.001$). This substantiates that novices demonstrate poor self-assessment of their own performance in self-directed mastoidectomy training.[9] More concerning was that the insufficiently drilled performances had a significantly higher mean FPS per minute (0.80 vs. 0.69, $p=0.03$) and mean MBS (83.5 % vs. 77.2 %, $p=0.01$) than the sufficiently drilled performances. This stresses the need for defining a safe volume for mastoidectomy in the simulator and also to institute other mechanisms to prevent that automated assessment rewards efficiency to such an extent that it potentially can cause overhasty and dangerous behavior in self-directed practice. This type of risky behavior and poor self-assessment of

novices has also been demonstrated in other simulation-based technical skills training such as endobronchial ultrasound with needle aspiration.[10]

In a large systematic review and meta-analysis, feedback was found to be an essential for effective simulation-based training.[31] Consistent with this, we also found that feedback by the simulator-integrated tutor-function increased the FPS and the number of passing procedures. However, the simulator-integrated tutoring also reduced efficiency and resulted in lower mean FPS per minute as well as MBS most likely because the tutoring caused trainees to be more meticulous and careful. Such cognitive engagement is key for developing true expertise[7] and should be encouraged, supported, and ultimately rewarded during assessment.

Simulator metrics can be important for different reasons: currently only 17 metrics demonstrated discriminative validity and should be used for automated scoring that supports aspects of learning and performance. The remaining metrics cannot discriminate residents from experts but still may have value for providing real-time formative feedback to support a safe performance. With improved simulator fidelity, these metrics may at some point also get discriminative properties and add safety-related items to automated assessment of mastoidectomy performance.

Some of the strengths of this study are the large number of performances by experienced surgeons and residents and that these performances were also assessed using an established assessment tool. Further, this study also investigated consequences of standard setting of the metrics-based assessment and other sources of validity evidence according to the contemporary validity framework of Messick. The major weaknesses of the study relate to the performance of the experienced surgeons: first of all, the experienced surgeons were chosen to represent a spectrum of surgical traditions. Next, each of the experienced surgeons also demonstrated a learning curve in simulation and optimally standard setting should have been based on data from the individual plateau phase of this learning curve. This also resulted in 40 % of experts not having a passing performance based on the MBS in their third performance. Having experts contributing with more performances was not feasible considering the number of experts and the substantial amount of time they committed for simulation drillings for this study.

Consequently, we chose to set standards using the level of the 10 % best performance of the experienced surgeons for most of the metrics, and adjusted this level for three of the

metrics, as a balance between the challenging yet attainable and investigated the consequences of this both in terms of the individual metrics and effects of repeated practice by residents to validate this standard. Finally, the minimum volume criteria did not consider which area of the temporal bone the voxels where drilled as this was not possible.

In the lens of the expert performance framework[22], our study has by different metrics captured some aspects of expert performance in mastoidectomy, which also provide insights into the mechanisms underlying expertise. This data on expert behavior can be used to inform future development of simulator fidelity, instructions, feedback, and training goals in simulation-based training of temporal bone surgery. In current VR temporal bone simulators, automated assessment by a general metrics-based score (MBS) for summative feedback can only serve as a minor learning support in self-directed training because important aspects of a safe mastoidectomy performance such as avoiding injuries to vital structures is not sufficiently ensured by the MBS alone. Real time formative feedback, simulator-integrated tutoring, structured guiding, and other learning supports need to be considered in temporal bone surgical training. Going forward, research on simulator metrics in mastoidectomy should focus on refining automated assessment in relation to different volumes/key areas of the temporal bone[16,23] including the safe volume for drilling and providing formative feedback.[17] Also, there is limited data on the number of VR simulation procedures needed to prove consistency of mastoidectomy performance. Establishing this would be a key step for future proficiency-based simulation training and mastery learning in temporal bone surgery.

CONCLUSION

A metrics-based score (MBS) for automated assessment of mastoidectomy performance in VR temporal bone surgical simulation was investigated based on the expert performance approach including the consequences of establishing a credible pass/fail standard, and validity evidence was collected. Even though the MBS demonstrated a traditional learning curve with repeated practice, the MBS mainly evaluates efficiency-related components of performance and fails to reflect other key elements of a safe mastoidectomy performance. Other learning supports such as formative feedback and

simulator-integrated tutoring need to be considered in the implementation of self-directed VR simulation training of temporal bone surgery.

Acknowledgements:

The authors would like to thank Drs. Michael McKenna (US), Juha Silvola (NO), Karin Strömbäck (SE), Michael Gaihede (DK), Kjell Tvetterås (DK), Lars Vendelbo (DK), Christian Faber (DK), Jens Wanscher (DK), Sven-Eric Stangerup (DK) and Per Cayé-Thomasen (DK) for contributing to the expert sampling.

REFERENCES

1. Morris D, Sewell C, Barbagli F, Salisbury K, Blevins NH, Girod S. Visuo-haptic simulation of bone surgery for training and evaluation', IEEE Comput Graph Appl 2006;26:48–57.

2. Zirkle M, Roberson DW, Leuwer R, Dubrowski, A. Using a virtual reality temporal bone simulator to assess otolaryngology trainees. Laryngoscope 2007;117:258–63.

3. O'Leary, S. J.; Hutchins, M. A.; Stevenson, D. R et al. Validation of a networked virtual reality simulation of temporal bone surgery. Laryngoscope 2008;118:1040–6.

4. Wiet GJ, Rastatter JC, Bapna S, Packer M, Stredney D, Welling DB. Training otologic surgical skills through simulation-moving toward validation: a pilot study and lessons learned. J Grad Med Educ 2009;1:61–6.

5. Sorensen MS, Mosegaard J, Trier P. The visible ear simulator: a public PC application for GPU-accelerated haptic 3D simulation of ear surgery based on the visible ear data. Otol Neurotol 2009;30:484–487.

6. Frithioff A, Sørensen MS, Andersen SAW. European status on temporal bone training: a questionnaire study. Eur Arch Otorhinolaryngol. 2018 Feb;275(2):357-363.

7. Ericsson KA. Deliberate Practice and the Acquisition and Maintenance of Expert Performance in Medicine and Related Domains. Academic Medicine. 2004;79(Supplement):S70-S81.

8. Andersen SA, Konge L, Cayé-Thomasen P, Sørensen MS. Learning Curves of Virtual Mastoidectomy in Distributed and Massed Practice. JAMA Otolaryngol Head Neck Surg. 2015 Oct;141(10):913-8.

9. Andersen SA, Konge L, Mikkelsen PT, Cayé-Thomasen P, Sørensen MS. Mapping the plateau of novices in virtual reality simulation training of mastoidectomy. Laryngoscope. 2017 Apr;127(4):907-914.
10. Konge L, Annema J, Clementsen P, Minddal V, Vilmann P, Ringsted C. Using virtual-reality simulation to assess performance in endobronchial ultrasound. Respiration. 2013;86(1):59-65.
11. Thomsen AS, Kiilgaard JF, Kjaerbo H, la Cour M, Konge L. Simulation-based certification for cataract surgery. Acta Ophthalmol. 2015 Aug;93(5):416-21.
12. Hovgaard LH, Andersen SAW, Konge L, Dalsgaard T, Larsen CR. Validity evidence for procedural competency in virtual reality robotic simulation, establishing a credible pass/fail standard for the vaginal cuff closure procedure. Surg Endosc. 2018;32(10):4200–4208.
13. Sewell C, Morris D, Blevins N, Barbagli F, Salisbury K. Quantifying risky behavior in surgical simulation. Stud Health Technol Inform. 2005; 111:451-457.
14. Khemani S, Arora A, Singh A, Tolley N, Darzi A. Objective Skills Assessment and Construct Validation of a Virtual Reality Temporal Bone Simulator. Otology & Neurotology. 2012;33(7):1225-1231.
15. Ioannou I, Zhou Y, Wijewickrema S, et al. Comparison of Experts and Residents Performing a Complex Procedure in a Temporal Bone Surgery Simulator. Otol Neurotol. 2017;38(6):e85-e91.
16. Kerwin T, Stredney D, Wiet G, Shen H-W. Virtual mastoidectomy performance evaluation through multi-volume analysis. Int J CARS. 2012;8(1):51-61.
17. Wijewickrema S, Piromchai P, Zhou Y, et al. Developing effective automated feedback in temporal bone surgery simulation. Otolaryngol Head Neck Surg. 2015;152(6):1082-1088.
18. Sethia R, Kerwin TF, Wiet GJ. Performance Assessment for Mastoidectomy. Otolaryngol Head Neck Surg. 2017 Jan;156(1):61-69.
19. American Educational Research Association APA, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing. Standards for educational and psychological testing. 2014. American Educational Research Association, Washington, DC.

20. Wan D, Wiet GJ, Welling DB, Kerwin T, Stredney D. Creating a cross-institutional grading scale for temporal bone dissection. Laryngoscope. 2010;120(7):1422-1427.
21. Ioannou I, Avery A, Zhou Y, Szudek J, Kennedy G, O'Leary S. The effect of fidelity: How expert behavior changes in a virtual reality environment. Laryngoscope. 2014;124(9):2144-2150.
22. Causer J, Barach P, Williams AM. Expertise in medicine: using the expert performance approach to improve simulation training. Med Educ. 2014;48(2):115-123.
23. Kerwin T, Wiet G, Stredney D, Shen H-W. Automatic scoring of virtual mastoidectomies using expert examples. Int J CARS. 2012;7(1):1-11.
24. Yudkowsky R, Park YS, Lineberry M, Knox A, Ritter EM. Setting Mastery Learning Standards. Academic Medicine. 2015;90(11):1495-1500.
25. Andersen SA, Foghsgaard S, Cayé-Thomasen P, Sørensen MS. The Effect of a Distributed Virtual Reality Simulation Training Program on Dissection Mastoidectomy Performance. Otol Neurotol. 2018; October 9 [Epub ahead of print].
26. The Visible Ear Simulator Dissection Manual, p. 35–47. Available from https://ves.alexandra.dk/system/files/download/VES_version3.2.zip. [Accessed 3 November 2018].
27. The Visible Ear Simulator, software. Available from https://ves.alexandra.dk/forums/ves3-ready. [Accessed 3 November 2018].
28. Trier P, Noe KØ, Sørensen MS, Mosegaard J. The visible ear surgery simulator. Stud Health Technol Inform. 2008; 132:523-5.
29. Andersen SA, Cayé-Thomasen P, Sørensen MS. Mastoidectomy performance assessment of virtual simulation training using final-product analysis. Laryngoscope. 2015 Feb;125(2):431-5.
30. Butler NN, Wiet GJ. Reliability of the Welling scale (WS1) for rating tem- poral bone dissection performance. Laryngoscope 2007;117:1803–1808.
31. Cook DA, Hamstra SJ, Brydges R, et al. Comparative effectiveness of instructional design features in simulation-based education: Systematic review and meta-analysis. Med Teach. 2013;35(1):e867-e898.

TABLES

Table 1. Participant background (experienced surgeons)

	Mean	Median	Range
Age, years	56	59	38–65
Gender	10 males / 1 female		
Handedness	100 % right handed		
Years as a specialist	20	20	8–30
Number of procedures			
Tympanoplasty	932	500	100–2000
Cholesteatoma surgery	600	500	50–2000
Stapes surgery	575	400	0–3000
Other middle ear surgery	57	50	0–300
Cochlear implantation	232	200	0–700
Other implantable hearing devices	50	0	0–200
Schwannoma surgery	191	0	0–1500
Other surgical neurotology	8	0	0–50
Experience with temporal bone surgical simulation (5-point Likert Scale)	1.2	1	0–4

Table 2A. Final-product score. Estimated marginal means.

	Tutored	Non-tutored	*Mean*	
Residents	16.0 (15.2–16.8)	13.3 (12.7–13.9)	14.7 (14.1–15.2)	p<<0.001
Experienced surgeons	20.9 (19.8–22.1)	18.2 (17.2–19.2)	19.6 (18.6–20.5)	
Mean	18.5 (17.6–19.3)	15.8 (15.1–16.4)		
	p<<0.001			

Table 2B. Final-product score per minute. Estimated marginal means.

	Tutored	Non-tutored	*Mean*	
Residents	0.34 (0.30–0.37)	0.48 (0.43–0.54)	0.41 (0.38–0.44)	p<<0.001
Experienced surgeons	0.55 (0.50–0.61)	0.70 (0.64–0.77)	0.63 (0.57–0.69)	
Mean	0.45 (0.41–0.48)	0.59 (0.54–0.65)		
	p<<0.001			

Table 3. Included metrics.

		Results of factor analysis (correlation coefficient)				
Metric	**Cut-off value**	Component 1 "Time and force efficiency"	Component 2 "Burr size efficiency"	Component 3 "Hesitancy"	Component 4 "Burr type efficiency"	Component 5 "Goal directed behaviour"
Drilling time (minutes)	<21.0 min	-0.630				
Voxels removed per minute	>65592 voxels/min	0.727				
Average force (N)	>0.72 N	0.913				
Percentage of voxels drilled while obscured	<0.29 %				0.482	
Percentage of voxels removed using sharp burrs	>91.6 %		-0.419		-0.654	
Percentage of voxels removed using fine diamond burrs	<4.2 %				0.830	
Number of jumps >5 mm	<31				0.422	0.498
Average force on sharp burrs	>0.79 N	0.894				
Average force on fine diamond burrs	<0.28 N				0.732	
Time not drilling and not in contact with bone (s)	<219 s			0.870		
Time not drilling but in contact with bone (s)	<70 s			0.794		
Percentage of time drilling and in contact with bone	>61 %			-0.897		
Percentage of voxels removed with small size burrs (0.5–2 mm)	<0.64 %		0.648			0.464
Percentage of voxels removed with medium size burrs (3–4 mm)	<12.7 %		0.912			
Percentage of voxels removed with large size burrs (5–7 mm)	>86.4 %		-0.941			
Average force on small size burrs (0.5–2 mm)	<0.23 N				0.496	0.476
Number of collisions with incus	<1.17					0.568
Estimated marginal mean score novices (95 % CI)		70.2 % (68.7 % –71.8 %)	63.1 % (60.9 %–65.3 %)	59.1 % (56.7 %–61.4 %)	58.7 % (56.9 %–60.5 %)	57.5 % (55.7–59.3 %)
Estimated marginal mean score experienced surgeons (95 % CI)		79.1 % (76.2 %–82.0 %)	80.7 % (76.6 %–84.9 %)	70.5 % (66.1 %–74.9 %)	79.7 % (76.3 %–83.0 %)	80.5 % (77.1 %–83.9 %)
Significance of discriminative ability		p<0.001	p<0.001	p<0.005	p<0.001	p<0.001

Table 4. Metrics-based score (MBS) performance

		Procedure 1	Procedure 2	Procedure 3
Residents	MBS, estimated marginal means (95 % CI)	56.5 % (53.7 %–59.4 %)	61.4 (58.6 %–64.3 %)	67.2 % (64.4–70.1 %)
	Number of passing performances (%)	0 (0.0 %)	2 (5.3 %)	2 (5.4 %)
Experienced surgeons	MBS, estimated marginal means (95 % CI)	72.9 % (68.9 %–76.9 %)	77.8 % (73.8 %–81.8 %)	83.6 % (79.6 %–87.7 %)
	Number of passing performances (%)	2 (18.2 %)	4 (36.4 %)	6 (60.0 %)

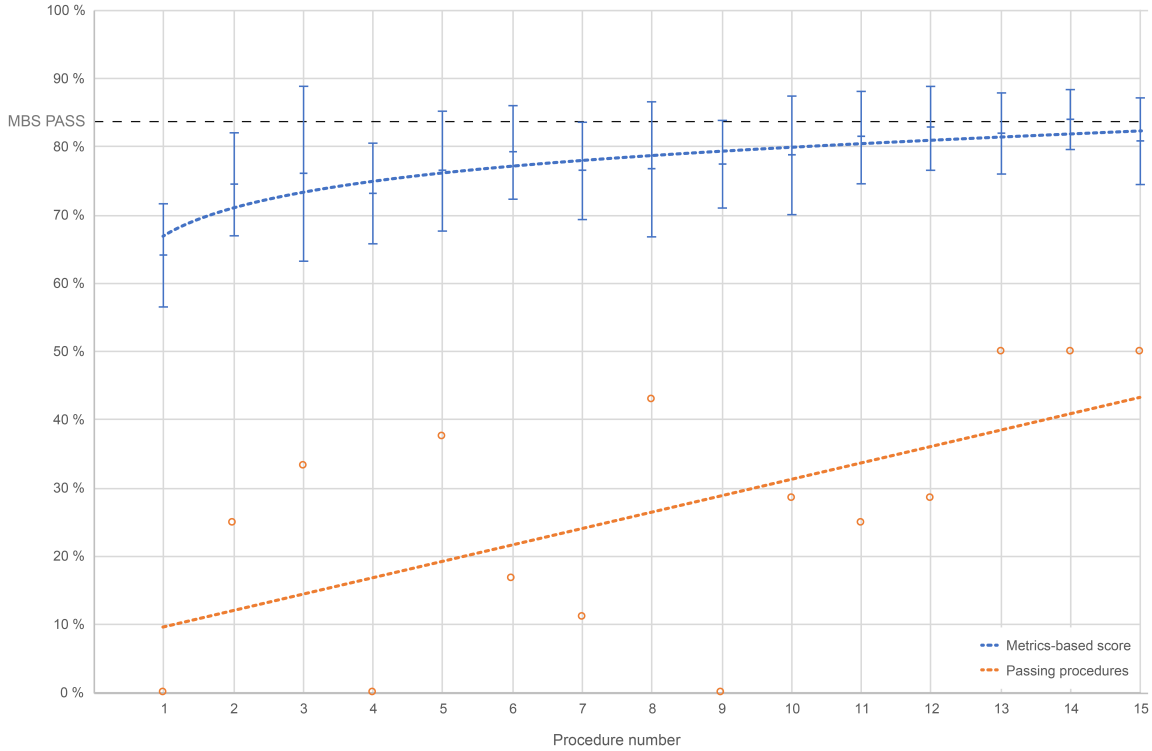

Figure 1. Means plot and fits of the metrics-based score (top) and the total number of passing performances (bottom) for resident performances. Bars indicate 95 % confidence intervals. The metrics-based score (MBS) pass/fail standard is set at 83.6 %.

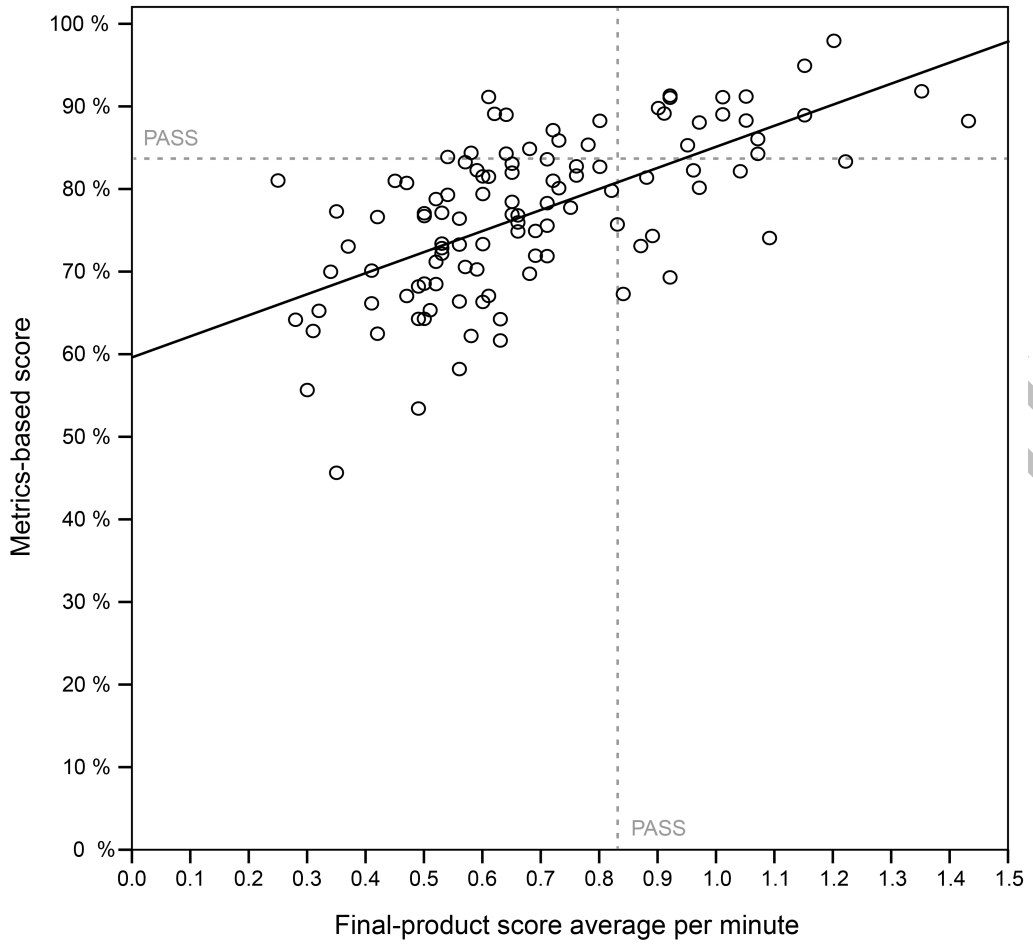

Figure 2. Correlation between the metrics-based score and the final-product score average per minute.

Supplementary figures/tables

Table A1. Messick's framework of validity applied to the metrics-based score.

Source of evidence	Method
Content	Inclusion of metrics described in the relevant literature and reports.
Response process	Objective metrics recorded by the simulator thereby avoiding rating bias. Verification of metrics using a visualization tool.
Internal structure	Components analysis for weighting of items.
Relationship to other variables	Comparison of scores achieved by experts and residents. Comparison with final-product analysis score and final-product score per min.
Consequences	Consequences of pass/fail standards. Consequences of repeated practice.

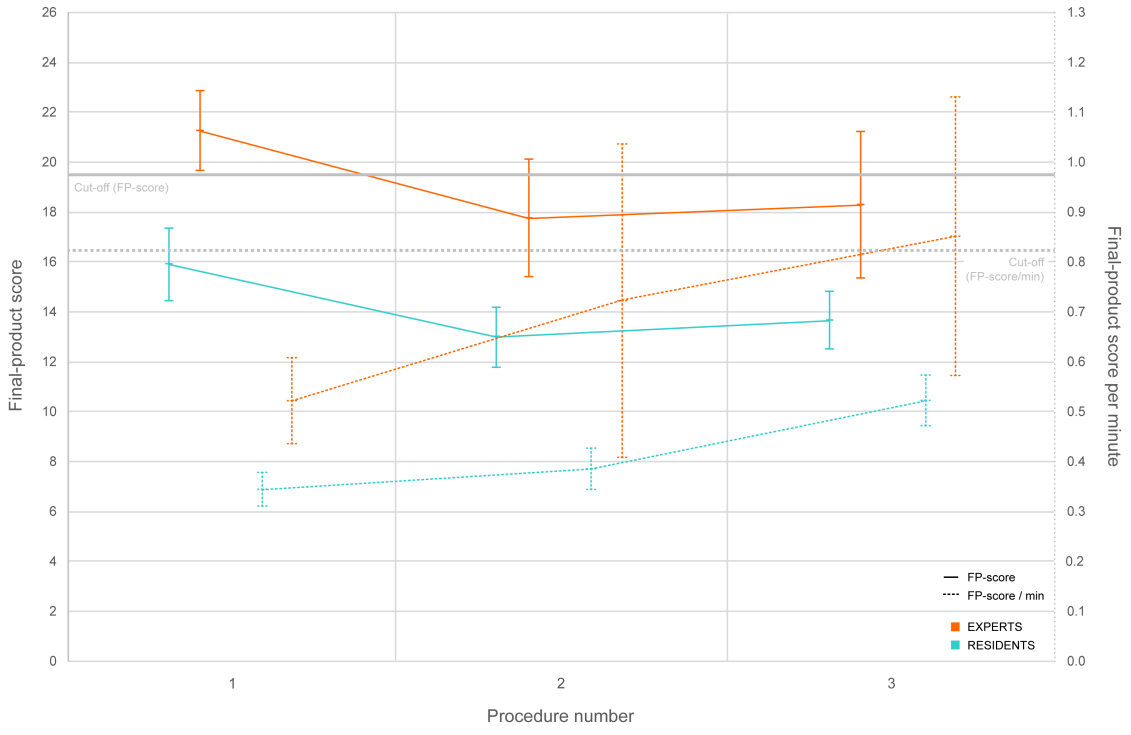

Figure A2. Means plot of final-product score and final-product score per minute for experts and novices by procedure number.

Table A3. Consequences of standard-setting (final-product assessment)

		Final-product score		Final-product score per minute	
		Fail	Pass	Fail	Pass
		Count (%)	*Count (%)*	*Count (%)*	*Count (%)*
Residents	Procedure 1	26 (70.3 %)	11 (29.7 %)	37 (100.0 %)	0 (0.0 %)
	Procedure 2	36 (94.7 %)	2 (5.3 %)	38 (100.0 %)	0 (0.0 %)
	Procedure 3	35 (94.6 %)	2 (5.4 %)	36 (97.3 %)	1 (2.7 %)
Experienced surgeons	Procedure 1	2 (18.2 %)	9 (81.8 %)	11 (100.0 %)	0 (0.0 %)
	Procedure 2	8 (72.7 %)	3 (27.3 %)	9 (81.8 %)	2 (18.2 %)
	Procedure 3	4 (40.0 %)	6 (60.0 %)	7 (70.0 %)	3 (30.0 %)

Table A4 - Metrics selection

		Linear mixed model									*Selection*				
		Parameter estimates for procedure number				Parameter estimates for level			80 % CI of experts						
#	Description of metric	Model intercept	Procedure 1	Procedure 2	Procedure 3	Residents	Experts	p	Lower bound	Upper bound	Criterium A	Criterium B	Criterium C	Criterium D	Notes
1	Duration of simulation (seconds)	1608	1222	463	0	337	0	0.011	1443	1774	True	True	True	4	
2	Duration of simulation (minutes)	26.8	20.4	7.7	0	5.6	0	0.011	24.1	30.0	True	True	True	4	
3	Drilling time (seconds)	1388	1229	435	0	257	0	0.012	1259	1516	True	True	True	4	
4	Drilling time (minutes)	23.1	20.5	7.3	0	4.3	0	0.012	**21.0**	25.3	True	True	True	Included	
5	Number of voxels removed	1620643	395715	12708	0	-108479	0	0.079	1535684	1705603	True	False	False		
6	Voxels removed per minute (simulation time)	61249	-13794	-8813	0	-10568	0	<0.001	56906	**65592**	True	True	True	Included	
7	Voxels removed per minute (drilling time)	72892	-18753	-11335	0	-13383	0	<0.001	67315	78468	True	True	True	6	
8	Time not drilling (seconds)	211	73	94	0	92	0	0.125	124	297	False	True	False		
9	Time not drilling (minutes)	3.5	1.22	1.57	0	1.55	0	0.124	2.06	4.96	False	True	False		
10	Percentage of time not drilling	12.4	-2	0.32	0	1.6	0	0.403	9.4	15.4	False	False	False		
11	Average force (N)	0.72	-0.047	-0.039	0	-0.077	0	0.017	0.68	**0.76***	True	True	True	Included	* *Cut-off (mean) 0.72*
12	Total path length (mm)	36712	24763	2392	0	4272	0	0.337	30567	42857	False	True	True		
13	Path length per minute (simulation time)	1365	-33	-126	0	-70	0	0.599	1185	1545	False	True	False		
14	Path length per minute (drilling time)	1618	-356	-270	0	23	0	0.884	1318	1918	False	False	True		
15	Path length per second (simulation time)	22.7	-0.6	-2.1	0	-1.2	0	0.601	19.7	25.7	False	True	False		
16	Path length per second (drilling time)	27	-4.2	-4.4	0	0.3	0	0.923	22	32	False	False	False		
17	Number of voxels drilled while obscured	12329	4814	2880	0	1031	0	0.600	9863	14795	False	True	True		
18	Percentage of voxels drilled while obscured	0.43	0.12	0.17	0	0.24	0	0.039	**0.29**	0.58	True	True	True	Included	
19	Number of voxels removed using a 0.5 mm sharp burr							-			n/a	n/a	n/a		
20	Number of voxels removed using a 1 mm sharp burr	152	-82	-97	0	-53	0	0.057	7	297	True	True	False		
21	Number of voxels removed using a 2 mm sharp burr	603	1327	2263	0	1250	0	0.504	-1816	3022	False	True	False		
22	Number of voxels removed using a 3 mm sharp burr	16219	8612	6320	0	17301	0	0.212	-1990	34428	False	True	True		
23	Number of voxels removed using a 4 mm sharp burr	79607	34851	19852	0	48405	0	0.136	37071	122142	False	True	True		
24	Number of voxels removed using a 5 mm sharp burr	340062	58883	-6289	0	-165241	0	0.003	262051	418072	True	False	False		
25	Number of voxels removed using a 6 mm sharp burr	381034	-21038	-28024	0	-132038	0	0.059	279845	482222	True	True	False		
26	Number of voxels removed using a 7 mm sharp burr	599937	-4357	-28445	0	-117898	0	0.191	477039	722834	False	True	False		
27	Number of voxels removed using a 0.5 mm coarse diamond burr							-			n/a	n/a	n/a		
28	Number of voxels removed using a 1 mm coarse diamond burr	1709	-160	-447	0	-843	0	0.078	1017	2403	True	True	False		
29	Number of voxels removed using a 2 mm coarse diamond burr	8017	4698	-2833	0	1861	0	0.610	2838	13196	False	True	False		
30	Number of voxels removed using a 3 mm coarse diamond burr	52014	20497	-9987	0	22979	0	0.360	18537	85491	False	True	False		
31	Number of voxels removed using a 4 mm coarse diamond burr	72649	54355	-6768	0	48412	0	0.159	28357	116940	False	True	False		
32	Number of voxels removed using a 5 mm coarse diamond burr	64717	71340	3312	0	-20635	0	0.426	31064	98371	False	False	True		
33	Number of voxels removed using a 6 mm coarse diamond burr	23554	14454	11798	0	-15620	0	0.178	9070	38038	False	False	True		
34	Number of voxels removed using a 7 mm coarse diamond burr	830	15304	-971	0	1500	0	0.360	-2250	3910	False	True	False		
35	Number of voxels removed using a 0.5 mm fine diamond burr	-0.72	-1.98	40.6	0	82	0	0.147	-71	70	False	False	False		
36	Number of voxels removed using a 1 mm fine diamond burr	3074	-52	737	0	1790	0	0.053	1867	4282	True	False	False		
37	Number of voxels removed using a 2 mm fine diamond burr	7363	7139	-4722	0	8155	0	0.023	2225	12502	True	True	False		
38	Number of voxels removed using a 3 mm fine diamond burr	-533	34478	8688	0	30646	0	0.004	-13935	12870	True	True	True	48	
39	Number of voxels removed using a 4 mm fine diamond burr	2486	25987	43184	0	59108	0	0.015	-28123	33095	True	True	False		
40	Number of voxels removed using a 5 mm fine diamond burr	-7708	66770	11399	0	34169	0	0.024	-27541	12126	True	True	True	48	
41	Number of voxels removed using a 6 mm fine diamond burr	8052	-3160	-2075	0	28878	0	0.293	-17764	33867	False	False	True		
42	Number of voxels removed using a 7 mm fine diamond burr	4651	5107	-4765	0	1217	0	0.392	-2769	12071	False	True	False		
43	Total number of voxels removed using sharp burrs	1397684	78604	-34131	0	-322961	0	<0.001	1297720	1497647	True	False	False		
44	Total number of voxels removed using coarse diamond burrs	227760	180383	-5977	0	32256	0	0.572	152778	302741	False	True	False		
45	Total number of voxels removed using fine diamond burrs	474	136619	52738	0	175526	0	0.001	-64419	65367	True	True	True		
46	Percentage of voxels removed using sharp burrs	86.4	-11.5	-3.4	0	-15	0	<0.001	81.1	**91.6**	True	True	True	Included	
47	Percentage of voxels removed using coarse diamond burrs	13.6	5.7	-0.01	0	3.7	0	0.308	8.8	18.4	False	True	True		
48	Percentage of voxels removed using fine diamond burrs	0.2	5.8	3.4	0	11.1	0	<0.001	-3.8	**4.2**	True	True	True	Included	
49	Number of jumps >5 mm	31	42	7	0	45	0	<0.001	**20***	43	True	True	True	Included	**Cut-off (mean) 31*
50	Number of jumps >10 mm	12	15	3	0	12	0	<0.001	9	16	True	True	True	49	
51	Number of jumps >15 mm	5	5	1	0	3	0	0.005	4	6	True	True	True	49	
52	Number of jumps >20 mm	2.7	1.4	0.1	0	1	0	0.061	2.0	3.5	True	True	True	49	
53	Number of jumps >25 mm	1.8	-0.2	-0.1	0	0.3	0	0.275	1.5	2.2	False	False	True		
54	Average force on 0.5 mm sharp burr	0.005	-0.004	-0.003	0	0	0	>0.999	-0.001	0.01	False	False	True		
55	Average force on 1 mm sharp burr	0.15	0.003	-0.003	0	-0.012	0	0.017	0.007	0.02	True	False	False		
56	Average force on 2 mm sharp burr	0.035	0.073	0.01	0	-0.003	0	0.887	0.004	0.07	False	False	True		
57	Average force on 3 mm sharp burr	0.11	0.03	-0.03	0	0.08	0	0.087	0.05	0.18	True	True	False		
58	Average force on 4 mm sharp burr	0.23	0.1	0	0	0.063	0	0.203	0.16	0.29	False	True	True		
59	Average force on 5 mm sharp burr	0.38	0.029	-0.009	0	0.011	0	0.851	0.30	0.46	False	True	False		
60	Average force on 6 mm sharp burr	0.38	0.016	0.011	0	0.023	0	0.751	0.28	0.48	False	True	True		
61	Average force on 7 mm sharp burr	0.58	-0.04	0.016	0	-0.009	0	0.891	0.48	0.67	False	True	False		
62	Average force on 0.5 mm coarse diamond burr	0.007	0.005	-0.007	0	0	0	>0.999	-0.001	0.016	False	False	False		
63	Average force on 1 mm coarse diamond burr	0.074	0.016	-0.023	0	-0.039	0	0.029	0.049	0.10	True	False	False		
64	Average force on 2 mm coarse diamond burr	0.124	-0.004	0.01	0	-0.007	0	0.854	0.07	0.17	False	True	False		
65	Average force on 3 mm coarse diamond burr	0.26	0.017	0.006	0	-0.02	0	0.637	0.19	0.32	False	False	True		
66	Average force on 4 mm coarse diamond burr	0.26	0.01	0.006	0	0.05	0	0.329	0.18	0.33	False	True	True		
67	Average force on 5 mm coarse diamond burr	0.17	0.11	0.02	0	-0.04	0	0.448	0.94	0.23	False	False	True		
68	Average force on 6 mm coarse diamond burr	0.1	0.05	0.03	0	-0.06	0	0.185	0.05	0.16	False	False	True		
69	Average force on 7 mm coarse diamond burr	0.006	-0.003	0.05	0	0.04	0	0.240	-0.05	0.07	False	False	False		
70	Average force on 0.5 mm fine diamond burr	-0.002	0.3	0.01	0	0.03	0	0.096	-0.026	0.023	True	True	True	80	
71	Average force on 1 mm fine diamond burr	0.14	-0.04	0.009	0	0.09	0	0.004	0.10	0.18	True	False	False		
72	Average force on 2 mm fine diamond burr	0.15	-0.002	-0.02	0	0.11	0	0.005	0.10	0.20	True	False	False		
73	Average force on 3 mm fine diamond burr	0.024	0.06	0.05	0	0.2	0	<0.001	-0.03	0.08	True	True	True	80	
74	Average force on 4 mm fine diamond burr	0.06	0.07	-0.005	0	0.17	0	0.001	-0.006	0.13	True	True	False		
75	Average force on 5 mm fine diamond burr	-0.012	0.14	0.04	0	0.114	0	0.011	-0.07	0.046	True	True	True	80	
76	Average force on 6 mm fine diamond burr	0.03	-0.0002	-0.02	0	0.03	0	0.362	-0.02	0.08	False	False	False		
77	Average force on 7 mm fine diamond burr	-0.009	0.023	0.008	0	0.03	0	0.191	-0.04	0.022	False	True	True		
78	Average force on sharp burrs	0.074	-0.047	-0.032	0	-0.075	0	0.032	0.69	**0.79**	True	True	True	Included	
79	Average force on coarse diamond burrs	0.43	0.005	0.035	0	0.037	0	0.221	0.39	0.48	False	True	False		
80	Average force on fine diamond burrs	0.28	0.023	0.02	0	0.15	0	<0.001	**0.24**	0.32	True	True	True	Included	**Cut-off (mean) 0.28*
81	Time not drilling and not in contact with bone (s)	272	235	154	0	98	0	0.020	**219**	326	True	True	True	Included	
82	Time not drilling but in contact with bone (s)	98	75	51	0	84	0	<0.001	**70**	126	True	True	True	Included	
83	Time drilling but not in contact with bone (s)	257	191	49	0	-31	0	0.294	219	294	False	False	True		
84	Time drilling and in contact with bone (s)	874	358	70	0	-9	0	0.892	786	963	False	False	True		
85	Percentage of time not drilling and not in contact with bone	0.19	0.02	0.04	0	0.03	0	0.045	0.16	0.21	True	True	False		
86	Percentage of time not drilling but in contact with bone	0.071	-0.077	0.006	0	0.04	0	<0.001	0.057	0.085	True	False	False		
87	Percentage of time drilling but not in contact with bone	0.17	0.03	0.001	0	-0.03	0	0.002	0.15	0.18	True	False	True		
88	Percentage of time drilling and in contact with bone	0.577	-0.042	-0.05	0	-0.044	0	0.046	0.55	**0.61**	True	True	True	Included	
89	Number of voxels removed with 0.5 mm burrs of any kind	14	-26	15	0	97	0	0.105	-63	90	False	False	False		
90	Number of voxels removed with 1 mm burrs of any kind	5261	-301	188	0	483	0	0.604	4009	6513	False	False	False		
91	Number of voxels removed with 2 mm burrs of any kind	13492	13216	-5255	0	14420	0	0.014	5834	21150	True	True	False		
92	Number of voxels removed with 3 mm burrs of any kind	65188	63639	5058	0	74118	0	0.017	25400	104975	True	True	True	108	
93	Number of voxels removed with 4 mm burrs of any kind	145474	115386	56406	0	167697	0	0.004	72311	218638	True	True	True	108	
94	Number of voxels removed with 5 mm burrs of any kind	400590	106920	8370	0	-156177	0	0.011	317046	484134	True	False	True		
95	Number of voxels removed with 6 mm burrs of any kind	407848	-9645	-18229	0	-122694	0	0.093	303011	512685	True	True	False		
96	Number of voxels removed with 7 mm burrs of any kind	596108	16248	-34043	0	-103356	0	0.251	473534	718683	False	False	False		
97	Percentage of voxels removed with 0.5 mm burrs of any kind	0.0008	-0.002	0.002	0	0.004	0	0.231	-0.005	0.006	False	False	False		
98	Percentage of voxels removed with 1 mm burrs of any kind	0.35	-0.1	-0.003	0	0.039	0	0.524	0.26	0.44	False	False	True		
99	Percentage of voxels removed with 2 mm burrs of any kind	0.82	0.031	-0.3	0	1.1	0	0.007	0.3	1.3	True	True	False		
100	Percentage of voxels removed with 3 mm burrs of any kind	4.3	1.3	0.2	0	5.5	0	0.008	1.4	7.2	True	True	True	108	
101	Percentage of voxels removed with 4 mm burrs of any kind	10.3	2	4.2	0	9.6	0	0.003	6.2	14.5	True	True	False		
102	Percentage of voxels removed with 5 mm burrs of any kind	24.3	6.6	0	0	-8.4	0	0.019	19.4	29.3	True	False	True		
103	Percentage of voxels removed with 6 mm burrs of any kind	25.4	-4.8	-2.1	0	-6.1	0	0.154	18.8	32.0	False	True	True		
104	Percentage of voxels removed with 7 mm burrs of any kind	34.8	-5.4	-2.7	0	-2.2	0	0.654	27.8	41.9	False	True	True		
105	Number of voxels removed with small size burrs (0.5–2 mm)	16275	12940	-5015	0	18165	0	0.011	10888	26948	True	True	False		
106	Number of voxels removed with medium size burrs (3–4 mm)	219564	178840	61332	0	230507	0	0.001	224031	375211	True	True	True		
107	Number of voxels removed with large size burrs (5–7 mm)	1388329	203862	-43661	0	-361628	0	<0.001	1350234	1533225	True	True	False		
108	Percentage of voxels removed with small size burrs (0.5–2 mm)	1.16	0.21	-0.31	0	1.15	0	0.008	**0.64**	1.62	True	True	True	Included	
109	Percentage of voxels removed with medium size burrs (3–7 mm)	14.35	3.36	4.45	0	15.5	0	<0.001	**12.65**	21.26	True	True	True	Included	
110	Percentage of voxels removed with large size burrs (5–7 mm)	84.48	-3.56	-4.13	0	-16.63	0	<0.001	77.47	**86.37**	True	True	True	Included	
111	Average force on small size burrs (0.5–2 mm)	0.23	0.029	0.003	0	0.109	0	<0.001	**0.226**	0.266	True	True	True	Included	
112	Average force on medium size burrs (3–4 mm)	0.44	0.020	-0.009	0	0.031	0	0.271	0.415	0.478	False	True	True		
113	Average force on large size burrs (5–7 mm)	0.75	-0.057	-0.041	0	-0.055	0	0.112	0.677	0.755	False	True	True		
114	Number of collisions with cerebellum	0.086	0.25	0.42	0	0.13	0	0.416	-0.11	0.28	False	True	False		
115	Number of collisions with cerebrum	0.38	0.01	0.15	0	-0.13	0	0.435	0.17	0.58	False	False	False		
116	Number of collisions with chorda	8.5	1.7	1.2	0	-1.79	0	0.220	6.5	10.5	False	False	True		
117	Number of collisions with digastric muscle	7.46	5.18	0.38	0	-3.3	0	0.006	6.0	8.9	True	False	True		
118	Number of collisions with dura	27.2	18	3.1	0	-7.6	0	0.032	22.8	31.6	True	False	True		
119	Number of collisions with facial nerve	5.1	1.6	2.5	0	1.6	0	0.309	3.0	7.3	False	True	False		
120	Number of collisions with incus	0.55	0.57	0.74	0	1.46	0	0.003	-0.06	**1.17**	True	True	True	Included	
121	Number of collisions with the inner ear	0.62	0.31	2.4	0	0.49	0	0.386	0.07	1.37	False	True	False		
122	Number of collisions with malleus	0.02	0.6	0.06	0	0.03	0	0.671	-0.056	0.10	False	True	True		
123	Number of collisions with stapes	0.21	-0.13	0.05	0	0.027	0	0.774	0.024	0.40	False	False	False		
124	Number of collisions with tympanic membrane	0.85	-0.018	0.094	0	0.47	0	0.319	0.19	1.50	False	False	False		
125	Number of collisions with vertebral artery	0	0	0.02	0	0	0	>0.999	-0.00003	0.00003	False	False	False		
126	Number of collisions with ear canal skin	8.9	1.9	2.7	0	2.2	0	0.287	6.2	11.7	False	True	False		
127	Total number of collisions	59.9	30	14.8	0	-6.2	0	0.445	50	70	False	False	True		
128	Total number of critical collisions	1.2	0.85	0.96	0	-46	0	0.224	0.75	1.63	False	False	False		
129	Number of collisions per minute	2.85	-0.36	0.09	0	-0.75	0	0.023	2.43	3.28	True	True	False		

Highlighted: Metric included in the model
Bold: Cut-off score

Table A4. Metrics selection.