**Title:** Automated Assessment of Cortical Mastoidectomy Performance in Virtual Reality

**Running title:** Automated Mastoidectomy Assessment

**Authors**

Sudanthi Wijewickrema [a], Benjamin James Talks [a †] Jesslyn Lamtara [a], Jean-Marc Gerard [a], and Stephen O'Leary [a]

[a] Department of Surgery (Otolaryngology), University of Melbourne, Royal Victorian Eye and Ear Hospital, Level 5/ 32 Gisborne Street, East Melbourne, Victoria 3002, Australia.

† Population Health Sciences Institute, Newcastle University, Newcastle Upon Tyne, Tyne and Wear, NE1 7RU, United Kingdom

**Corresponding Author:** Dr Benjamin James Talks, Department of Surgery (Otolaryngology), University of Melbourne, Royal Victorian Eye and Ear Hospital, Level 5/ 32 Gisborne Street, East Melbourne, Victoria 3002, Australia.
Email: ben.talks@newcastle.ac.uk

**Acknowledgements:** Benjamin Talks would like to thank the University of Birmingham for his Student Development Scholarship, which supported his involvement in this project.

**Conflict of Interest Statement:** the authors have no potential conflicts of interest with respect to the research, authorship, and/ or publication of this article.

**Data Sharing:** the data that support the findings of this study are available from the corresponding author upon reasonable request.

**Funding Statement:** this research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Author contribution statement:**

Sudanthi Wijewickrema: conceptualization, methodology, formal analysis, writing – original draft, and writing – review and editing. Benjamin Talks: conceptualization, investigation, writing – original draft, and writing review and editing. Jesslyn Lamtara: project administration, data curation, and investigation. Jean-Marc Gerard: methodology, investigation, writing – review and editing. Stephen O'Leary: methodology, supervision, and writing – review and editing. We confirm that the manuscript has been read and approved by all named authors and that there are no persons who satisfied the criteria for authorship but are not listed.

DR BENJAMIN JAMES TALKS (Orcid ID : 0000-0003-4896-2324)

Article type : Original Article



**Abstract**

Introduction

Cortical mastoidectomy is a core skill that Otolaryngology trainees must gain competency in. Automated competency assessments have the potential to reduce assessment subjectivity and bias, as well as reducing the workload for surgical trainers.

Objectives

This study aimed to develop and validate an automated competency assessment system for cortical mastoidectomy.

Participants

Data from 60 participants (Group 1) was used to develop and validate an automated competency assessment system for cortical mastoidectomy. Data from 14 other participants (Group 2) was used to test the generalisability of the automated assessment.

Design

Participants drilled cortical mastoidectomies on a virtual reality temporal bone simulator. Procedures were graded by a blinded expert using the previously validated Melbourne Mastoidectomy Scale: a different expert assessed procedures by Groups 1 and 2. Using data from Group 1, simulator metrics were developed to map directly to the individual items of this scale. Metric value thresholds were calculated by comparing automated simulator metric values to expert scores. Binary scores per item were allocated using these thresholds. Validation was performed using random sub-sampling. The generalisability of the method was investigated by performing the automated assessment on mastoidectomies performed by Group 2, and correlating these with scores of a second blinded expert.

Results

The automated binary score compared to the expert score per item had an accuracy, sensitivity, and specificity of 0.9450, 0.9547, and 0.9343 respectively for Group 1; and 0.8614, 0.8579, and 0.8654 respectively for Group 2. There was a strong correlation between the total scores per participant assigned by the expert and calculated by the automatic assessment method for both Group 1 ($r = 0.9144$, $p < 0.0001$), and Group 2 ($r = 0.7224$, $p < 0.0001$).

Conclusion

This study outlines a virtual reality-based method of automated assessment of competency in cortical mastoidectomy, which proved comparable to the assessment provided by human experts.

**Key Words:** Virtual reality; surgical training; competency-based assessment; automated assessment; temporal bone surgery; surgical simulation

**Conflict of Interest Statement:** the authors have no potential conflicts of interest with respect to the research, authorship, and/ or publication of this article.

**Data Sharing:** the data that support the findings of this study are available from the corresponding author upon reasonable request.

**Key Points**

- Cortical mastoidectomy is a core-competency in Otolaryngology.
- Validated scoring systems have been created for assessment of cortical mastoidectomy.
- Virtual reality simulators present an ideal platform for competency assessments, able to both present a standardised task and record detailed performance metrics.
- We developed a voxel based anatomical method of virtual reality-based automatic assessment of cortical mastoidectomy.
- The automated assessment method was comparable with assessment by human experts.

## Introduction

Competency-based training has become standard across surgical training programmes, requiring trainees to achieve a certain skill level at a pre-defined set of tasks before progressing to the next stage of their training (1). Cortical mastoidectomy is one such task in Otolaryngology training, involving the removal of mastoid air cells as part of the management of chronic otitis media, or as the first step of cochlear implant surgery and various lateral skull base operations. To facilitate the assessment of technical skills in cortical mastoidectomy, validated scoring systems have been developed (2–7). However, these assessment scales are still limited by potential subjectivity, bias, and human errors by the assessor, not to mention the time and associated financial cost of employing an expert grader (8). Automated assessment of trainee performance offers an intuitive solution to these shortcomings.

Automated assessment of surgical performance is typically implemented through methods such as tool, hand, and/or eye motion tracking and muscle contraction analysis (8). Data is usually extracted from sensors and/or video using computer vision (9) and machine learning is used to analyse the data and provide a performance valuation (8). Performance assessment in virtual reality (VR) is an alternative that has grown in popularity in recent years (10–12). VR simulators hold promise for integration into competency assessments as they can present trainees with a standardised surgical task and collect detailed metrics on performance. VR-based automated assessment could provide trainees with valuable feedback on their current skill level and support further self-directed "deliberate practice" (13).

Many simulator metrics have been previously proposed for assessing mastoidectomy performance (14). These can be grouped into measures of surgical technique (drill force, velocity, and burr size) (12,15), comparison of voxel removal with an expert data set (11), and voxel removal in anatomically defined regions (16). Sewell et al. proposed 20 simulator metrics for assessing mastoidectomy performance: 15 were measures of drilling and suctioning technique, whilst 5 were voxel-based measures of appropriate bone removal or facial nerve damage (15). However, all metrics were compared to a study-specific global score rather than a validated mastoidectomy assessment scale. Andersen et al. investigated a

further 129 metrics collected by the Visible Ear Simulator, 17 of which (including time, force, burr size, hesitancy, and burr type) were able to distinguish between novices and experts in temporal bone surgery and subsequently contribute to automated assessment (12). Kerwin et al. described two voxel-based algorithms to compare the volume of bone removed by trainees to an expert data set for anatomically segmented regions of a temporal bone (11). However, the effectiveness of these methods was only tested against 5 expert-rated criteria. Finally, Andersen et al. looked at whether the volume of bone removed within segmented sections representing steps of the mastoidectomy operation correlated with an expert-graded modified Welling Scale (16). They did not find a correlation between volumes of bone removed inside and outside of the operative steps and operative performance.

To our knowledge, none of the existing VR-based automated assessments align fully with validated surgical assessment scales. Therefore, although they provide important information on performance, these methods are not able to provide detailed feedback on different aspects of performance that typically define surgical competence.

## Objectives

The main objective of this study was to use a voxel-based anatomical approach to develop a VR-based method of automatic assessment for cortical mastoidectomy, able to emulate the assessment of human experts. We aimed to provide detailed feedback on each stage of the procedure by mapping a specific metric, in terms of voxel removal in an anatomical region, to each item of a validated assessment scale (2).

## Materials and Methods

### Ethical Considerations

Ethics approval was obtained by the Human Ethics Committee of the Royal Victorian Eye and Ear Hospital (Group 1: #19/1419HL; Group 2: #19/1441HL and #16/1300H). All participants provided signed consent.

### Setting

We use the University of Melbourne VR temporal bone surgery simulator (Figure 1) as our platform. The simulator comprises virtual models of human temporal bones, a haptic device (SensAble PHANTOM Omni) that provides the user with a virtual surgical drill and delivers tactile feedback, and a MIDI controller that is used as an input device to change environment variables such as magnification level and burr size. Depth perception is achieved through NVIDIA 3D vision technology.

**Participants**

Data from Group 1 was used to develop and validate an automated assessment method for cortical mastoidectomy. This comprised 60 surgeries from 10 Otolaryngology consultants (experts), 10 Otolaryngology registrars (intermediates), and 40 University students with an interest in surgery (novices).

Data from Group 2 was used to test the generalisability of the automated assessment. This was comprised of 35 surgeries from 4 experts (12 surgeries), 4 intermediates (11 surgeries), and 6 novices (12 surgeries).

**Study Procedure**

After a 5-minute familiarisation period, participants performed a cortical mastoidectomy on the simulator. As the students had no prior surgical experience, they were shown a 15-minute video tutorial on how to perform a mastoidectomy first. No form of guidance or feedback was provided during the procedure. All procedures were recorded by the simulator and using screen-capture software for later grading.

**Performance Assessment**

Video recordings of all procedures were evaluated by a blinded expert, using the Melbourne Mastoidectomy Scale (MMS), an end-product dissection scale designed for cortical mastoidectomy (2). This scale was validated on a VR simulator: its inter-rater reliability (between 3 expert graders) was shown to be high ($r = 0.921$, $p < 0.0001$) and its ability to differentiate between skill levels (novice, intermediate and expert) was also high ($p < 0.0001$) (2). The MMS comprises 20 items, outlined in Table 1. To ensure that points are not awarded

for incomplete dissection, it has dependencies between items (e.g., if a structure has not been identified, points are not awarded for avoiding damage of that structure).

**Automation of the MMS**

As the first step in the automation process, we classified the different types of items in the MMS. Consultation with expert surgeons resulted in the identification of 3 types of items, based on what they assessed. The first of these, is when a landmark/anatomical structure is broadly exposed (e.g., temporal line). The second item type is when an anatomical structure is skeletonised (e.g., adequate exposure of the sigmoid sinus) and the third is based on the damage caused to an anatomical structure (e.g., middle fossa plate identification without damage). We call these 3 types 'exposure-based', 'skeletonisation-based', and 'damage-based' respectively.

Then, we determined how an expert surgeon marked each item type. For exposure-based items, a minimum amount of bone had to be removed from a given region of the temporal bone. Additionally, greater importance was allocated to particular parts of that region. For skeletonisation-based items, the decision was based on whether a minimum amount of the anatomical structure was skeletonised anywhere along it. Damage-based items were scored on the amount of damage caused to an anatomical structure.

To map how an expert scored items to our automatic assessment scale, we identified simulator metrics that defined these items.

For exposure-based items, we asked an expert surgeon to drill the region of the temporal bone required to identify the landmark associated with that MMS item. Examples are shown in Figure 2, where dark green, light green, and blue green denote the regions that need to be removed to expose the external ear canal, dura, and sigmoid sinus respectively. Next, we determined the relative importance of drilling each voxel by analysing the mastoidectomies drilled by the expert participants in Group 1 of our study and calculating how many experts drilled each voxel of the temporal bone. Figure 2b shows the heatmap that denotes this: the colour ranges from dark blue to red showing voxels that were drilled by an increasing number of surgeons from one to all. From this data we assigned weights to each voxel; the higher the number of experts that drilled a given voxel, the higher its weight. Finally, to determine the

simulator metric for an exposure-based MMS item, we defined a corresponding temporal bone region for each item and then calculated the weighted average of the voxels drilled by expert participants in Group 1 for each region.

The expert assessor awarded marks for skeletonisation of an anatomical structure if they could see any part of that structure through a thin layer of bone. To generate a corresponding skeletonisation-based simulator metric, we first obtained a thin layer (of bone) around the structure (using dilation, the thickness of which was determined by this expert surgeon; Figure 3a). The adequacy of exposure was then quantified as the percentage of voxels drilled in the bone layer thus obtained.

Good temporal bone surgical technique identifies and exposes an anatomical structure, but does not damage it. To calculate the simulator-based metric for damage-based items, we calculated the number of voxels belonging to an anatomical structure that were drilled (Figure 3b).

The simulator metrics defined above provided values over a range, while the MMS items were scored dichotomously, as either 0 or 1. Therefore, we determined the threshold values of the metrics where an expert's score transitioned from 0 to 1 (or vice versa). To this end, we fitted a sigmoid function to the data for each item. This function is defined as $y = 1/(1 + e\hat{}b(-x + a))$, where $y$ is the expert assessment (0 or 1), $x$ is the simulator metric value and $a$ and $b$ are the coefficients of the sigmoid function denoting the shift and slope respectively. Non-linear least squares fitting with trust region optimization was used for fitting this function (17). The slope was constrained to be in the range of [0, 1] and the shift was constrained to be positive. The threshold value was then considered to be the shift $a$ of the function (value of the simulator metric $(x)$ at $y = 0.5$). This method ensures that the outliers in the data are ignored and the middle of the overlapping region (where there are both 0 and 1 assessments due to human subjectivity) are considered as the threshold (Figure 4a).

For damage-based items, the metric value and expert assessment are inversely related (the less damage the better), so we constrained the slope to be in the range [-1, 0] (Figure 4b).

As mentioned above, the scoring of some items depends on whether another item has been scored as 1 or not (e.g., if the incus has not been identified, the item that checks damage to the incus is scored as a 0 regardless of whether there is any damage or not). We removed such dissection data from the dependent items to avoid errors in the subsequent steps.

**Data Usage (Group 1)**

We observed that even the novice group completed some parts of the procedure (e.g., drilling of MacEwan's triangle) consistently well. As such, the range of data available for the development of the automated assessment method was unbalanced. To ensure that unsatisfactory assessments (binary ratings of '0') were available for all items in the assessment scale, we generated some synthetic data, as is commonly done as a method of data augmentation in similar situations (18,19). To this end, one of the authors performed 31 additional procedures with varying degrees of completeness of the different steps of the procedure, to supplement the original dataset of 60 surgeries by Group 1. The blinded expert who assessed them was not made aware of the synthetic nature of these procedures.

We split the data from Group 1 randomly into sets of 80% and 20% for training and testing respectively. As we used the expert procedures to calculate weights for exposure-based items, to avoid bias, we included all expert dissections in the training set. We constrained the splitting of the dataset to ensure that at least one of each class (0 and 1) was available for each MMS item in the training set. We generated 20 such random splits using repeated random sub-sampling (20) to ensure robustness of the developed method.

We scored all items as 0 or 1 for the data in each test set based on the thresholds calculated using the corresponding training set. To account for dependencies, if an independent item was scored as 0, we scored its corresponding dependent items as 0, regardless of their actual score.

**Generalisability of the Automated Assessment (Group 2)**

To test how our method of automated assessment, which was based on the assessments of one human expert compared to that of a different expert, new data was analysed (Group 2). The automated assessment method was trained using all the data from Group 1 (as opposed to

the previous analysis where data was separated into different sets for training and testing). Then, we calculated the automatic scores for the 35 additional surgeries performed by the 14 participants in Group 2 using this model. We then asked a second independent expert, who was not involved in the assessments used in the development stage, to assess these surgeries and compared the expert's assessments with the corresponding automatic scores.

**Main Outcome Measures**

We compared the automatic scores for each surgery with the corresponding expert assigned scores using Pearson's correlation coefficient. We used the root-mean-squared (RMS) difference to compare the total scores per participant assigned by the expert and calculated by the automatic assessment. We used a significance level of 0.05 and used MATLAB R2020a (Mathworks, Natick, USA) for all implementations.

## Results

The automated binary score when compared to the expert assigned score per item for the 20 repetitions of random sub-sampling of Group 1 are shown in a confusion matrix in Figure 5a. The accuracy, sensitivity, and specificity of the method was 0.9450, 0.9547, and 0.9343 respectively. There was a strong and significant correlation between the total scores per participant assigned by the expert and calculated by the automatic assessment for Group 1, $r = 0.9144$, $p < 0.0001$, 95% confidence interval = [0.8957, 0.9299]; Figure 5b shows the correlation results; the RMS (root-mean-squared) difference between the scores was 1.5581 points (out of 20).

A strong significant correlation between the independent expert and automatic scores was also observed from the analysis of mastoidectomies performed by Group 2, $r = 0.7224$, $p < 0.0001$, 95% confidence interval = [0.6849, 0.7560]. Accuracy, sensitivity, and specificity were 0.8614, 0.8579, and 0.8654 respectively; the RMS difference was 2.7098.

## Discussion

We outlined a VR-based method of automated assessment of competency in cortical mastoidectomy, which proved comparable to the assessment provided by two human experts.

In contrast to previous VR-based automated surgical assessment methods (11,12,21), this method is the first to fully align its scoring system with a validated surgical assessment scale. The one-to-one mapping of the automatic assessment to the scale items makes it possible to provide detailed feedback to the trainees on their performance. The total score, in conjunction with a suitable cut-off value, can be used in competency-based training, where the next level of training is introduced only after a certain level of competence is reached. This enables this method to be easily integrated into existing surgical curricula, which will reduce the workload of human experts in surgical training and support self-directed surgical training.

A limitation of this method is that the model developed here is only valid for the specimen it was developed on. To extend it to other specimens, the relevant regions on those specimens will have to be identified first. We will explore how our previous work on anatomical registration of temporal bone regions (22) can be extended for this purpose. Second, the model developed here will have to be adapted to suit other specimens (e.g., the threshold values determined for the original specimen may not be valid for a new specimen). To this end, we will investigate the use of transfer learning techniques (23), to avoid the need to collect data on each new specimen. Additionally, the number of participants in this study was relatively small.

Furthermore, as this method was specifically designed for cortical mastoidectomy, it cannot be used in other surgeries. However, the concepts developed here are easily transferable to other surgical procedures and domains. It should also be tested what the benefits of this form of assessment are to the learning process in practice.

## Conclusion

This study outlines a VR-based method of automated assessment of cortical mastoidectomy, which proved comparable to assessment by human experts. As this method maps simulator metrics directly to the items of a validated assessment scale, it can provide detailed performance feedback to trainees. Automated assessment will reduce the workload of experts in surgical training and support self-directed practice.

## References

1. Bhatti N, Cummings C. Viewpoint: Competence in surgical residency training:

Defining and raising the bar. Acad Med. 2007;82:569–73.

2. Talks B, Lamtara J, Wijewickrema S, Gerard J-M, Mitchell-Innes A, O'Leary S. The Melbourne Mastoidectomy Scale: validation of an end-product dissection scale for cortical mastoidectomy. Clin Otolaryngol. 2020;45(5):746–53.
3. Butler N, Wiet G. Reliability of the Welling Scale (WS1) for rating temporal bone dissection performance. Laryngoscope. 2007;117:1803–8.
4. Zirkle M, Taplin M, Anthony R, Dubrowski A. Objective assessment of temporal bone drilling skills. Ann Otol Rhinol Laryngol. 2007;116(11):793–8.
5. Mowry S, Woodson E, Gubbels S, Cargrae M, Hansen M. A simple assessment tool for evaluation for cadaveric temporal bone dissection. Laryngoscope. 2018;128:451–5.
6. Pisa J, Gousseau M, Mowat S, Westerberg B, Unger B, Hochman J. Simplified summative temporal bone dissection scale demonstrates equivalence to existing measures. Ann Otol Rhinol Laryngol. 2018;127(1):51–8.
7. Laeeq K, Bhatti N, Carey J, Della Santina C, Limb C, Niparko J, et al. Pilot testing of an assessment tool for competency in mastoidectomy. Laryngoscope. 2009;119:2402–10.
8. Levin M, McKechnie T, Khalid S, Grantcharov T, Goldenberg M. Automated methods of technical skill assessment in surgery: A systematic review. J Surg Educ. 2019;76(6):1629–39.
9. Law H, Ghani K, Deng J. Surgeon technical skill assessment using computer vision based analysis. In: MLHC. 2017. p. 88–99.
10. Pan J, Chang J, Yang X, Liang H, Zhang J, Qureshi T, et al. Virtual reality training and assessment in laparoscopic rectum surgery. Int J Med Robot. 2015;11(2):194–209.
11. Kerwin T, Wiet G, Stredney D, Shen H. Automatic scoring of virtual mastoidectomies using expert examples. Int J Comput Assist Radiol Surg. 2012;7(1):1–11.
12. Andersen S, Mikkelsen P, Sørensen M. Expert sampling of VR simulator metrics for automated assessment of mastoidectomy performance. Laryngoscope. 2019;129(9):2170–7.
13. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med. 2004;79(10):S70-81.
14. Al-Shahrestani F, Sørensen M, Andersen S. Performance metrics in mastoidectomy training: a systematic review. Eur Arch Otorhinolaryngol. 2019;276(3):657–64.
15. Sewell C, Morris D, Blevins N, Agrawal S, Dutta S, Barbagli F, et al. Validating metrics for a mastoidectomy simulator. Stud Heal Technol Inf. 2007;125:421–6.

16. Andersen S, Cayé-Thomasen P, Sørensen M. Mastoidectomy performance assessment of virtual simulation training using final-product analysis. Laryngoscope. 2015;125:431–5.
17. Byrd R, Schnabel R, Shultz G. Approximate solution of the trust region problem by minimization over two-dimensional subspaces. Math Program. 1988;40:247–63.
18. Fawaz H, Forestier G, Weber J, Idoumghar L, Muller P. Data augmentation using synthetic data for time series classification with deep residual networks. arXiv Prepr arXiv 180802455. 2018;
19. Frid-Adar M, Amitai J, Goldberger J, Greenspan H. Synthetic data augmentation using GAN for improved liver lesion classification. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). Washingtom DC; 2018. p. 289–93.
20. Kuhn M, Johnson K. Applied predictive modeling. Springer, New York; 2013.
21. Sewell C, Morris D, Blevins N, Dutta S, Agrawal S, Barbagli F, et al. Providing metrics and performance feedback in a surgical simulator. Comput Aided Surg. 2008;13(2):63–81.
22. Zhou Y, Ioannou I, Wijewickrema S, Bailey J, Piromchai P, Gregor K, et al. Transfer learning of a temporal bone performance model via anatomical feature registration. In: 2014 22nd International Conference on Pattern Recognition. 2014. p. 1916–21.
23. Margolis A. A literature review of domain adaptation with unlabeled data. Tec Rep. 2011;1–42.



**Tables**

Table 1. The Melbourne Mastoidectomy Scale. Region-based items are denoted by * (2).

| | Definition | Disagree | Agree |
|---|---|---|---|
| **MacEwans Triangle defined as** | | | |
| 1. Temporal line * | Cortex removed along the temporal line, delineating the superior limit of dissection. | 0 | 1 |
| 2. Posterior external auditory canal wall * | Cortex removed behind the posterior wall of the external auditory canal, defining the anterior limit of dissection. | 0 | 1 |

| | | | |
|---|---|---|---|
| 3. Sigmoid sinus * | Cortex removed over the suspected course of the sigmoid sinus, from the temporal line towards the mastoid tip, defining the posterior limit of dissection. | 0 | 1 |
| **Middle fossa plate** | | | |
| 4. Identified * | Partial exposure/clear identification of the middle fossa plate. | 0 | 1 |
| 5. Adequately exposed *[4] | Skeletonised middle fossa plate from sinodural angle to tegmen tympani without overhanging cortex. | 0 | 1 |
| 6. Identified without minor damage [4] | No small holes in the middle fossa plate. | 0 | 1 |
| 7. Identified without major damage [4] † | No large holes in the middle fossa plate or drilling of the underlying dura. | 0 | 1 |
| **Sigmoid sinus** | | | |
| 8. Identified * | Partial exposure/ clear identification of the sigmoid sinus. | 0 | 1 |
| 9. Adequately exposed *[8] | Skeletonised sigmoid sinus from sinodural angle towards mastoid tip, without overhanging cortex. | 0 | 1 |
| 10. Identified without damage [8] † | No holes in the overlying bone or direct drilling of the sigmoid sinus. | 0 | 1 |
| 11. Sinodural angle defined *[8] | Sharp angle between the exposed sigmoid sinus and middle fossa plate. | 0 | 1 |
| **External auditory canal** | | | |
| 12. Canal wall preserved | Grossly skeletonised external canal wall. | 0 | 1 |
| 13. Posterior canal wall adequately thinned *[12] | Precisely skeletonised external canal wall on at least 130 degrees. | 0 | 1 |
| 14. Canal wall thinned with no holes [13] | No holes in the external canal wall. | 0 | 1 |
| **Mastoid antrum** | | | |
| 15. Antrum opened * | Drilling to open the mastoid antrum with exposure of lateral semi-circular canal. | 0 | 1 |

| | | | |
|---|---|---|---|
| 16. Antrum opened with no damage of the semicircular canals [15] † | All the semicircular canals remain intact, with no holes. | 0 | 1 |
| 17. Incus identified * | The entire superior edge of short process of the incus is visible. | 0 | 1 |
| 18. Incus identified without damage [17] | No drilling or disruption of the ossicular chain. | 0 | 1 |
| **Facial nerve** | | | |
| 19. Vertical section identified * | The vertical section of the facial nerve is visible. | 0 | 1 |
| 20. Identified with no damage [19] † | No exposure of facial nerve sheath. | 0 | 1 |
| **TOTAL SCORE** | | **/20** | |

† These items represent major complications of the procedure and damage of the marked structures can class the dissection as unacceptable regardless of overall score.

‡ *Superscripted numbers (*[1-20]) represent the dependency of that item on a previous item on the scale denoted by the number.

**Figure Legends**

Figure 1. A surgeon performing an operation on the virtual reality temporal bone surgery simulator. A virtual temporal bone is displayed on the computer screen, which is viewed in 3D using NVIDIA 3D vision technology. A haptic device (shown as a drill on the screen) enables drilling and provides tactile feedback. A MIDI controller provides a convenient interface for changing settings such as the burr size and magnification level.

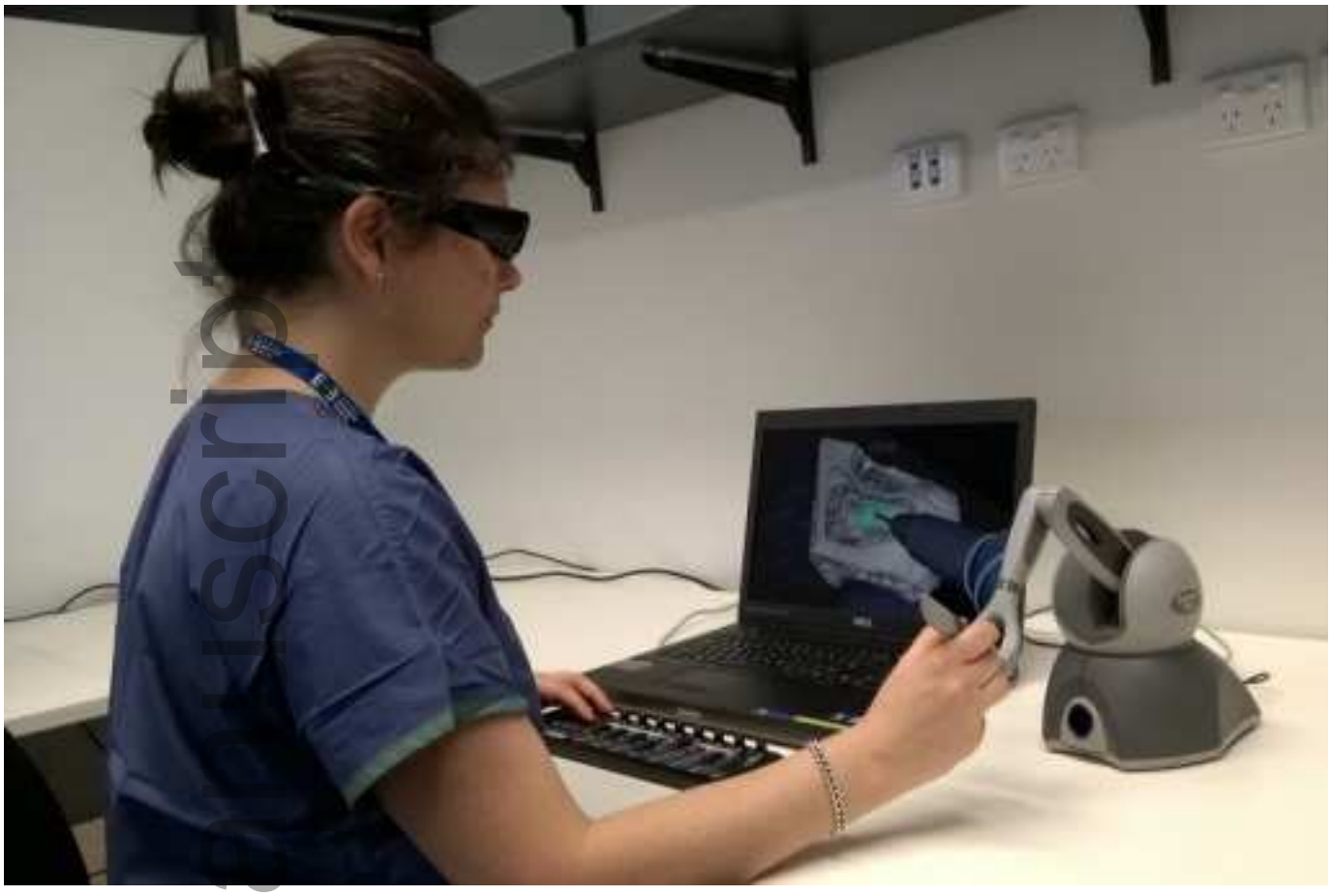
Figure 2. a) Regions defined by an expert surgeon for some exposure-based items. Different shades of green denote different temporal bone regions/MMS items (dark green was the temporal bone region drilled for exposure of the external ear canal; light-green for exposure of the dura; and blue-green for the sigmoid sinus). b) Heatmap illustrating the number of

expert surgeons that drilled a given voxel of the temporal bone. The colours vary from dark red, voxels drilled by all surgeons; to dark blue, those drilled by one surgeon.

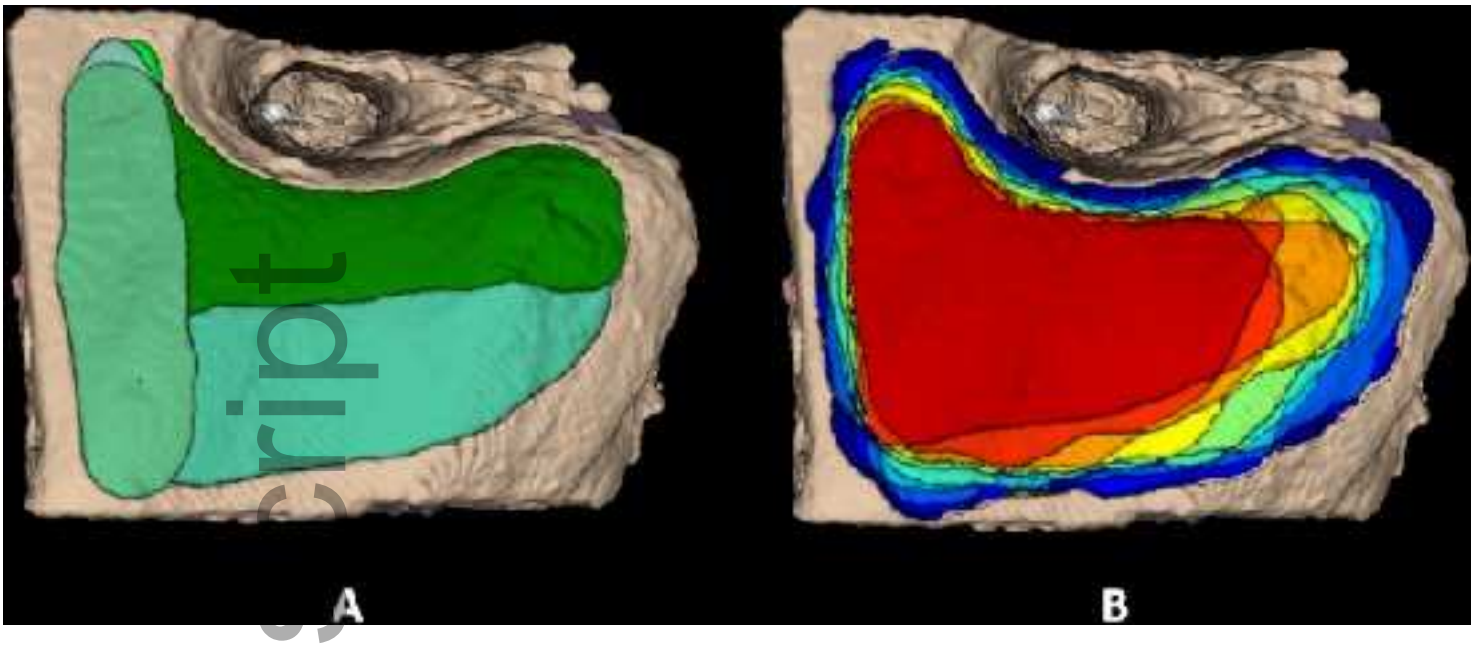
Figure 3. a) Regions around anatomical structures determined through dilation of the structure defined for some skeletonisation-based items. The different shades of blue denote the different regions. b) An example of damage caused to an anatomical structure (circled in red), used in the assessment of damage-based items.

Figure 4. Calculation of thresholds for a) an exposure-based item and b) a damage-based item.

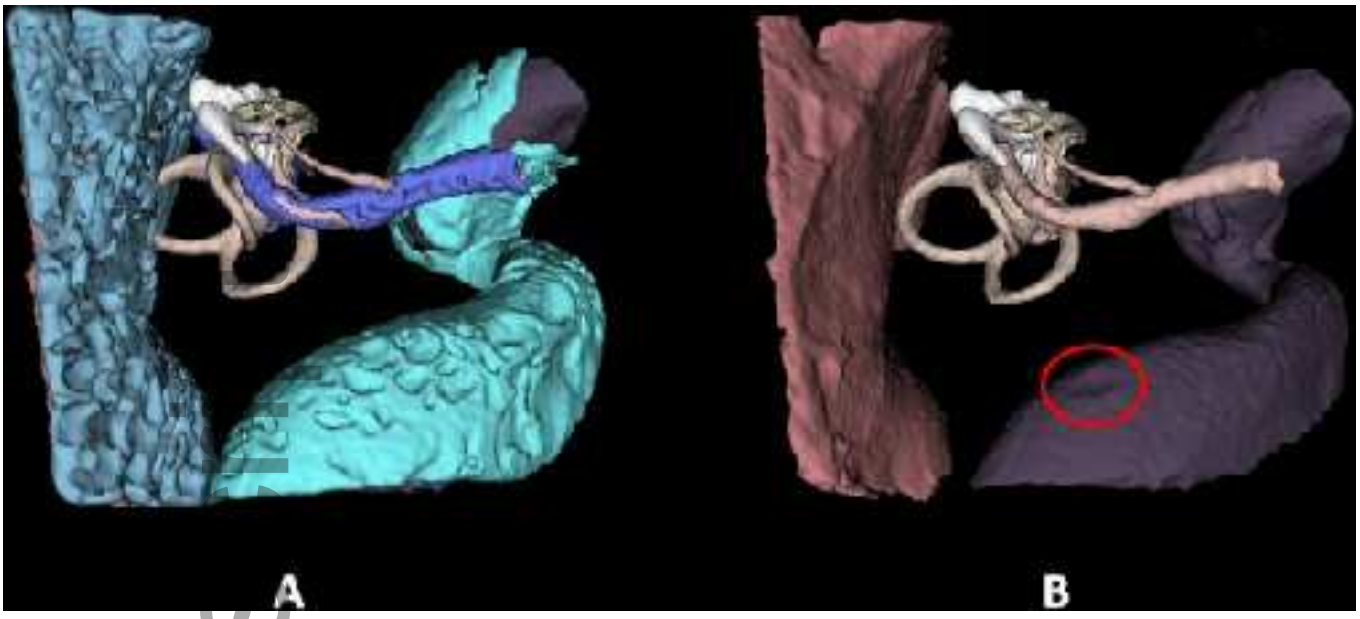
Figure 5. Validation results across the test sets of the 20 random sub-samplings: a) confusion matrix showing the per-item thresholding performance and b) comparison of the total expert and automatic scores. The red line denotes the ideal results, a 1:1 mapping between expert and automatic scores.



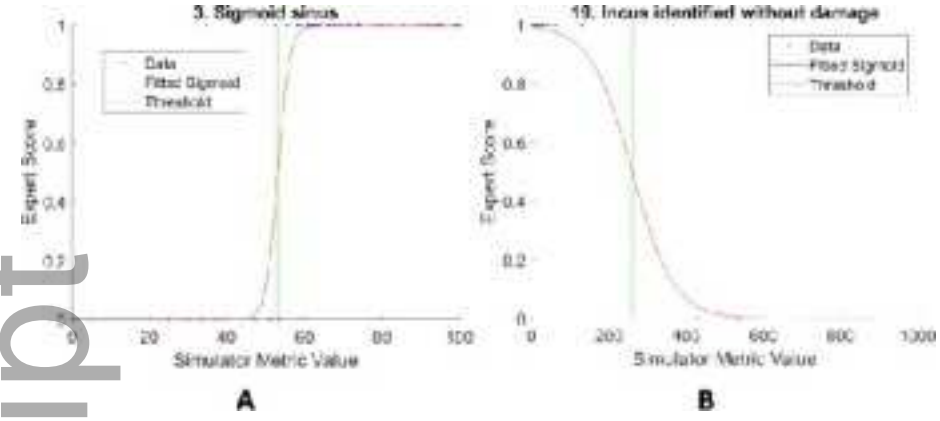
coa_13760_f1.jpg



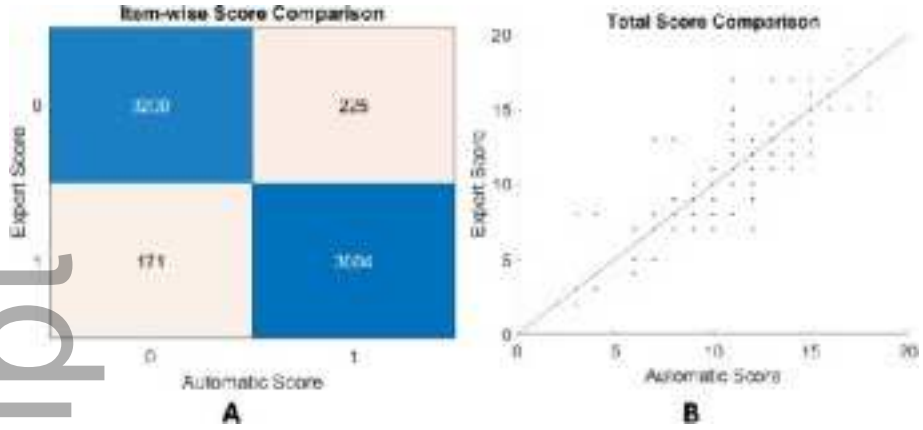

coa_13760_f2.tiff




coa_13760_f3.tiff




coa_13760_f4.tiff




coa_13760_f5.tiff