MULTI-INSTITUTIONAL DEVELOPMENT OF A MASTOIDECTOMY PERFORMANCE EVALUATION INSTRUMENT

Thomas Kerwin[1], Brad Hittle[1], Don Stredney[1], Paul De Boeck[2], Gregory Wiet[3]

[1] Interface Lab, Ohio Supercomputer Center, Columbus, Ohio, United States

[2] Department of Psychology, Ohio State University, Columbus, Ohio, United States

[3] Department of Otolaryngology, Ohio State University, Columbus, Ohio, United States

CORRESPONDING AUTHOR:
Thomas Kerwin[1]

1224 Kinnear Rd., Columbus, Ohio, 43212, United States

Email address: kerwin@osc.edu

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2931v1 | CC BY 4.0 Open Access | rec: 18 Apr 2017, publ: 18 Apr 2017

Multi-Institutional Development of a Mastoidectomy Performance Evaluation Instrument

Abstract

Objective

A method for rating surgical performance of a mastoidectomy procedure that is shown to apply universally across teaching institutions has not yet been devised. This work describes the development of a rating instrument created from a multi-institutional consortium.

Design

Using a participatory design and a modified Delphi approach, a multi-institutional group of expert otologists constructed a 15 element task-based checklist for evaluating mastoidectomy performance. This instrument was further refined into a 14 element checklist focusing on the concept of safety after using it to rate a large and varied population of performances.

Setting

Twelve Otolaryngological surgical training programs in the United States.

Participants

14 surgeons from 12 different institutions took part in the construction of the instrument.

Results

By using 14 experts from 12 different institutions and a literature review, individual metrics were identified, rated as to the level of importance and operationally defined to create a rating scale for mastoidectomy performance. Initial use of the rating scale showed modest rater agreement. The operational definitions of individual metrics were modified to emphasize “safe” as opposed to “proper” technique. A second rating instrument was developed based on this feedback.

Conclusions

Using a consensus building approach with multiple rounds of communication between experts is a feasible way to construct a rating instrument for mastoidectomy. Expert opinion alone using a Delphi method provides face and content validity evidence, however, this is not sufficient to develop a universally acceptable rating instrument. A continued process of development and

experimentation to demonstrate evidence for reliability and validity making use of a large population of raters and performances is necessary to achieve universal acceptance.

KEY WORDS

mastoidectomy, assessment

COMPETENCIES

Medical Knowledge, Practice-Based Learning and Improvement

INTRODUCTION

Skill assessment is essential to all types of training, and otologic surgery is no exception. In addition to providing evidence that a basic level of skill proficiency has been achieved, accurate feedback can accelerate learning[1]. Surgical residency programs currently use a variety of tools to assess trainees, and no single tool has emerged as the "gold standard". At a minimum, a good assessment tool must be reliable, feasible, fair, objective, and valid[2]. The time-honored assessment currently used by the American Board of Otolaryngology (ABOto) and the Accreditation Council for Graduate Medical Education (ACGME) is based upon both the accumulation of “adequate” case numbers during training and also the attestation of the specific residency program director where the resident trained. Notwithstanding, there is little evidence of the reliability or validity of the current assessment regimen.

A universally applicable set of metrics that can be agreed upon and used for assessment of technical skill in performing a mastoidectomy has not been developed or adopted. In order to develop such an assessment tool, care must be given to formulate and validate that tool taking into account differences between training programs. Assessment tools must be designed based on what the measurement instrument will be used for and what specific inferences will be made based on the results[3]. There is need for an instrument for both user feedback in training and for determining the level of an individual’s performance (novice, intermediate or expert) in terms of technical performance of a mastoidectomy with facial recess approach.

In this work, we describe the creation and evolution of a set of metrics specifically for determining the level of an individual's performance in mastoidectomy. We used a broad-based consortium of surgeons at different institutions in consecutive feedback steps so that the instrument can be universally applied to all temporal bone dissection performances regardless of institution or background.

PREVIOUS WORK

Rating instruments for mastoidectomy have been developed by other groups, but they do not include such a broad base of expert input. A recent review of the current instruments for measuring mastoidectomy performance by Sethia et al. discusses each of the instruments in greater detail4.

A group at Johns Hopkins developed an instrument based on the work of Martin et al.[5] for mastoidectomy performance containing both a Task-Based Checklist (TBC) and Global Rating Scale (GRS)[6]. Both of the scales use a list of evaluation items with ratings of one to five. Work by Laeeq et al.[7] and Awad et al.[8] show some validity evidence for that instrument but in only a small number of institutions.

The Welling Scale (WS1) uses final product analysis (FPA) for evaluating a complete mastoidectomy with facial recess performed in the temporal bone lab9,10. It defines binary items that are summed to provide an overall score.

As seen in the survey results from Butler et al.[10], even though a set of common evaluation items for mastoidectomy can be created, there exist many differences between the importance given to those items by experts from different institutions. Additional care must be given to develop and evaluate instruments that can be used broadly at all institutions. In order to create such an instrument, an attempt was made by Wan et al.[11] to use a modified Delphi method to find consensus on which items should be incorporated into a TBC. The Hopkins scale was also developed using a Delphi method, but included only Johns Hopkins faculty members in the process.

The Wan et al. study received responses from 88 members of the American Neurotology Society or American Otological Society on criteria important to a successful temporal bone dissection. Based on those responses, a list of criteria ordered by importance was created and used in this study.

Materials and Methods

In order to create a consensus-based, cross-institutional rating instrument to measure surgical performance we started with the list of assessment items from Wan et al.[11] These items were then further refined using a Delphi method described in detail below with an expert group consisting of 14 fellowship-trained otologists from 12 different institutions (Table 1). In this refinement, the individual items from the Wan study were more explicitly defined to encourage a uniform interpretation for determining success or failure for each item. This list was then reviewed by all individuals in the same group of experts by means of an online survey.

In a first round, members of the consortium were asked to rank the 5 most important and 6 least important metrics on the list. Results of the survey showed 24 metrics with additional suggestions (Table 2).

A face-to-face meeting for active discussion regarding each metric, its overall importance and an agreed upon operational definition was convened with the members of the expert group. In this meeting, each metric was presented separately along with any comments that were made within the survey context. An example of a metric result and discussion is presented in Figure 1.

Next, the experts were asked to assign an importance measure to each metric, as follows:

- **Pass/Fail (P/F)**: Critical metrics that, if any one is violated, there is an automatic failure. Violations of these metrics will result in serious morbidity to the patient.

- **High**: Dangerous, if violated could potentially result in morbidity to the patient.
- **Medium**: Potential complication that requires intervention and could be rectified or managed without significant morbidity to the patient.
- **Low**: Potential complication which does not require intervention – poor technique.

Then, in a second round, experts were asked to identify which items were needed to be competent in order to be considered novice level (ready to operate on patient under supervision), intermediate level (ready for minimal supervision – PGY 4/5 level), advanced level (practice independently at fellowship trained level). Using the following criteria:

- **Novice level**. (competency on each of the high importance areas and no Fs). (ready for cadaveric lab)
- **Intermediate level**. (competency on all of high and medium items and no Fs). (ready for Supervised OR experience).
- **Advanced level**. (expert on all metrics and no Fs) (ready for independent surgery, does not need supervision).

The results of the above two rounds are listed in Table 3 as original and final relative importance. The items listed as P/F (Pass/Fail) include those items for which if they were not achieved, the global performance automatically resulted in a failing score regardless of performance on any other metric. The items listed as High priority were those items with conditions to be fulfilled to be considered as a novice operator, the items listed as Medium are items with conditions to be fulfilled to be considered as an intermediate and the items listed as low are necessary conditions for an advanced level operator. By implication the absence of more important violations is necessary as well for each of the three levels.

At this point, under IRB approval from The Ohio State University Office of Responsible Research, we performed a study using our previously developed temporal bone dissection simulator[12,13] across the 12 institutions. This resulted in sixty-six mastoidectomy performances for review. They covered a wide distribution of skill levels: medical students, PGY (Post-Graduate Year) 2-5, fellows and attending physicians. This set comprised 36 sessions collected from faculty and 30 collected from residents and students. Each of twelve expert reviewers, all considered experts in otologic surgery, was assigned eleven grading tasks (individual mastoidectomy performances). They were blinded to the identity of the subject performing the dissection and did not review their own performances. This resulted in two sets of ratings using the instrument for each virtual mastoidectomy in the testing set. After examining the statistical measures from this trial, a moderately low level of agreement among raters was seen (over half the interclass correlation[12] (ICC) values were below 0.4, which is considered poor agreement).

As a result of the relatively weak inter-rater agreement, we concluded that perhaps this may be due to poor agreement on the operational definition of each metric and how it should be scored. As a result, an additional face-to-face Delphi process was undertaken to discuss the poor agreement scores. It was the consensus of the group that the operational definitions of each item were a source of continued variability in how they should be interpreted. The group recommended further refinement based on the premise that they would be used to identify "safe" as opposed to "proper"

surgical technique. It was recognized that there are various opinions as to what constitutes "proper" technique. The consensus was that there would be greater agreement if the operational definition of individual metrics could be judged on the basis of its "safety". Specifically, if a particular style of technique was not one that a particular rater recognized as "proper", it could still be judged on whether or not it was considered high risk i.e., not safe. Based on this discussion, a second set of assessment items was developed. Additionally, at the suggestion of the expert group, the two items in the original list that concerned the external auditory canal were combined into one. The result of this second discussion group was the development of a second set of metrics encompassing a list of 15 items. The individual items for both metric sets can be seen and compared in Table 4.

An overview of the steps we took to construct the metrics and the reasons behind them can be seen in Table 5.

DISCUSSION

As with clinical care, it is important that clear and rigorous evidence exists to objectively appraise the efficacy of our educational programs.[15] Subjective determinations by program directors or trainee self-reporting of number of procedures must evolve into more evidence based assessments. This requires a concerted effort to develop outcome measures that are agreed upon and universally translatable. For assessments to be valid, they must accumulate validity evidence in a number of areas including content evidence, response process, internal structure, relations with other variables and consequences.[16] Our metrics demonstrate "content evidence" based on the nature of the development process noted above. The next validation steps include the demonstration of a sufficiently high intra-rater agreement and the relationship with an external criterion for the quality of a performance.

We have followed the process outlined by Dauphinee and Wood-Dauphinee for developing evidenced-based medical education.[15] This involves defining the parameters to be measured, measuring those parameters, and benchmarking those parameters to assess educational outcomes. As noted by our work, the effort to define outcome measures with an acceptable level of content validity is in itself often painstaking, especially if the goal includes universal acceptance. Studies conducted at one institution often are fraught with subjective bias and low sample sizes.[15] This makes dissemination of recommendations and guidelines for assessment problematic.

Our attempt at developing a specific set of metrics for a procedure as specialized as mastoidectomy has proven extremely challenging. In mastoid surgery, there are a number of assessment tools in existence today, none of which provide broad enough acceptance and universality.[4] It is the goal of this research to continue the process of painstakingly refining the metrics established and the rating process so that they can show the validity evidence necessary to make assessments that correlate with clinical performance.

Identifying, defining and applying metrics so that they can be universally useful and still provide sufficient information to make valid decisions based on their use is difficult even at the early stages.

For the next steps we are necessarily subject to many sources of possible assessor error[17]. These include possible drift in assessor interpretation of individual metrics, individual performance expectations, and lack of familiarity of being an assessor as opposed to a trainer. These sources of rating error are multiplied with the expanded number of assessors. These sources however, can be mitigated in the future by making a concerted effort to provide good operational definitions of each metric , careful training of assessors (perhaps a group session where a standardized performance is rated and discussed within the context of the group), and monitoring of the assessor's performance as suggested by Gallagher et al.[17]

In the future, we will use our new set of metrics to accumulate additional validity evidence. Emphasizing safety as the global concept in defining and administering the items is one way we can make our operational definitions more widely applicable. We are currently investigating defining our measurement scales in terms of three separate axes: bone removal, tool control and violations of structures (Table 6). These can function as distinct subscales. Measurement scales for skill mastery will be built for each axis such that the performances of trainees can be evaluated in terms of descriptive and normative mastery levels. The descriptive levels are specific positions on the measurement scales while the normative levels are levels that must be reached to be considered an independent expert (expert level) or an intermediate level trainee (intermediate level). The approach to be used is a two-fold extension of item response theory (IRT).[18-20] IRT is a family of statistical measurement models that has become the standard for the measurement of skills in an educational and training context. IRT scores are model-based descriptive mastery levels. Additionally, we are designing a methodology to easily "train the raters" so that consistency in interpretation and application of the metrics is plausible.

Conclusion

Our work moves closer to the goal of developing a universally acceptable and applicable set of performance metrics for mastoid surgery. We have used an extensive participatory process to formulate a list of metrics based on literature review, multiple rounds of expert feedback, and continued refinement. Based on our methodology, we feel that our results demonstrate significant content validity. Our results demonstrate considerable input of diverse expert opinion but still need to be supplemented with other types of validity in a multi-institute context.

Acknowledgements

This work was supported by The National Institute for Deafness and other Communication Disorders, National Institutes of Health, USA, R01DC011321.

Bibliography

1. Catania AC. *Learning.* 2nd ed. Englewood Cliffs, N.J.: Prentice-Hall; 1984.

2. Shah J, Darzi A. Surgical skills assessment: an ongoing debate. *BJU Int.* 2001;88(7):655-660.
3. Michelson JD, Manning L. Competency assessment in simulation-based procedural education. *Am J Surg.* 2008;196(4):609-615.
4. Sethia R, Kerwin TF, Wiet GJ. Performance Assessment for Mastoidectomy: State of the Art Review. *Otolaryngol Head Neck Surg.* 2016.
5. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg.* 1997;84(2):273-278.
6. Francis HW, Masood H, Chaudhry KN, et al. Objective assessment of mastoidectomy skills in the operating room. *Otol Neurotol.* 2010;31(5):759-765.
7. Laeeq K, Bhatti NI, Carey JP, et al. Pilot testing of an assessment tool for competency in mastoidectomy. *Laryngoscope.* 2009;119(12):2402-2410.
8. Awad Z, Tornari C, Ahmed S, Tolley NS. Construct validity of cadaveric temporal bones for training and assessment in mastoidectomy. *Laryngoscope.* 2015;125(10):2376-2381.
9. Fernandez SA, Wiet GJ, Butler NN, Welling B, Jarjoura D. Reliability of surgical skills scores in otolaryngology residents: analysis using generalizability theory. *Eval Health Prof.* 2008;31(4):419-436.
10. Butler NN, Wiet GJ. Reliability of the Welling scale (WS1) for rating temporal bone dissection performance. *Laryngoscope.* 2007;117(10):1803-1808.
11. 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.
12. Wiet GJ, Bryan J, Dodson E, et al. Virtual temporal bone dissection simulation. *Studies in health technology and informatics.* 2000;70:378-384.
13. Wiet GJ, Stredney D, Kerwin T, et al. Virtual temporal bone dissection system: OSU virtual temporal bone system: development and testing. *Laryngoscope.* 2012;122 Suppl 1:S1-12.
14. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.* 1979;86(2):420-428.
15. Dauphinee WD, Wood-Dauphinee S. The need for evidence in medical education: the development of best evidence medical education as an opportunity to inform, guide, and sustain medical education research. *Acad Med.* 2004;79(10):925-930.
16. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Advances in health sciences education : theory and practice.* 2014;19(2):233-250.
17. Gallagher AG, O'Sullivan GC. *Fundamentals of surgical simulation : principles and practices.* London: Springer; 2012.
18. Linden WJvd. *Handbook of item response theory: Models, statistical tools, and applications* Boca Raton: CRC Press; 2016.
19. De Boek P, Wilson M. *Explanatory item response models. A generalized linear and nonlinear approach.* New York: Springer; 2004.
20. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. *Medical education.* 2010;44(1):109-117.

254 *Table 1: Participating Training Institutions*

Baylor University
Duke University
Henry Ford Hospital System
University of Iowa
University of Mississippi
Montefiore/Albert Einstein College of Medicine
Stanford University
University of California, Irvine
University of Cincinnati
University of Texas, Southwestern
The Ohio State University
Medical University of South Carolina

255

256

257

258 *Table 2: Results from survey on importance of individual metrics. 13 experts listed 5 items as high*
259 *importance and 6 items as low importance.*

Metric	Experts selecting as high importance	Experts selecting as low importance
Maintains visibility of tool while removing bone	6	1
Select appropriate burr type and size	4	2
Antrum entered	4	1
No violation of facial nerve canal	11	0
No violation of sigmoid sinus	3	1
Identifies tympanic segment of facial nerve	0	2
Does not drill on ossicle	5	1
Does not use excessive drill force near critical structures	6	0
Identifies chorda tympani	0	3
Drills in best direction (understanding of cutting edge)	3	3
Canal wall up	1	3
Identifies facial nerve at cochlearform process	0	4
Appropriate depth of cavity	0	3
Drills with broad strokes	1	3
No holes in EAC	2	2
Complete saucerization	2	4
Posterior external auditory canal wall thinned	2	2
Facial recess completely exposed	2	1
Identifies facial nerve at external genu	1	2
Low frequency of drill “jumps”	2	6
No holes in the tegmen	3	2
Use of diamond burr within 2mm of facial nerve	1	2
No cells remain on sinodural angle	0	10
Sinodural angle sharply defined	0	7
Other additional metric	1	0

260

261

262 *Table 3: Original and final distributions of metrics based on level of importance and which metric*
263 *expected to be achieved at each performance level.*

Metric	Importance proposed to experts.	*Final*
Maintains visibility of burr while removing bone	High	High
Excessive force will not be used near critical structures	High	High
Appropriate depth of cavity	Low	Low
No holes in tegmen	Low	Low
Select appropriate burr	Medium	Medium
Violation of the sigmoid sinus	Medium	Medium
Identification of chorda tympani nerve	High	Medium
Drill in best direction	Medium	Medium
External auditory canal wall will remain up	Medium	Medium
No holes in external auditory canal wall	Low	Medium
Complete saucerization	Medium	Medium
Posterior external auditory canal wall thinned appropriately	Medium	Medium
Violation of the facial nerve	P/F	P/F
Violation of the horizontal (lateral) semi-circular canal	P/F	P/F
Drill contact with ossicles	P/F	P/F
Violation of dura		P/F

264

265

266 *Table 4: Text of questions asked during mastoidectomy performance review. Question #10 was removed*
267 *for the second instrument, due to overlap with question #9.*

Number	Instrument 1	Instrument 2
1	Maintains visibility of burr while removing bone	Maintains safe view of the burr throughout the procedure
2	Excessive force will not be used near critical structures	Maintains safe force near critical structures throughout the procedure
3	Appropriate depth of cavity	Sufficient removal of mastoid air cells for proper visualization of deep structures
4	No holes in tegmen	Maintains integrity of tegmen
5	Select appropriate burr	Efficient and Safe burr selection
6	Violation of the sigmoid sinus	Maintains integrity of sigmoid sinus
7	Identification of chorda tympani nerve	Identifies chorda tympani nerve sufficiently to perform facial recess approach
8	Drill in best direction	Efficient and safe direction of drilling (parallel to critical structures)
9	External auditory canal wall will remain up	Sufficient thinning of posterior external auditory canal wall to visualize facial nerve
10	No holes in external auditory canal wall	
11	Complete saucerization	Sufficient saucerization for safe drilling
12	Posterior external auditory canal wall thinned appropriately	Avoids overthinning or holes in posterior auditory canal wall
13	Violation of the facial nerve	Maintains integrity of facial nerve
14	Violation of the horizontal (lateral) semi-circular canal	Maintains integrity of horizontal semi-circular canal
15	Drill contact with ossicles	Maintains integrity of ossicles
16	Violation of dura	Maintains integrity of dura

268

269

270 *Table 5: Steps taken to develop the instrument, in order, with a brief reason for each one.*

Step	**Reason**
Start with list of items from Wan et al.	Include a wide sample of surgical expertise
Survey to determine most and least important items	Remove very low priority items and establish broad levels of importance
Meeting to present survey results and define metrics	Revise item text based on consensus from experts
Classification of metrics for novice, intermediate, expert achievement level.	Reflect importance levels of items in the scoring of the instrument
Validation study using instrument	Test instrument
Revision of instrument focusing on safety	Attempt to increase interrater reliability

271

272 *Table 6: Metrics and Performance Axis for Assessment Strategy*

Metrics	Axis
Sufficient removal of mastoid air cells for proper visualization of deep structures	Bone Removal
Identifies chorda tympani nerve sufficiently to perform facial recess approach	Bone Removal
Sufficient thinning of posterior external auditory canal wall to visualize facial nerve	Bone Removal
Sufficient saucerization for safe drilling	Bone Removal
Avoids overthinning or holes in posterior auditory canal wall	Bone Removal
Maintains safe view of the burr throughout the procedure	Tool control
Maintains safe force near critical structures throughout the procedure	Tool control
Efficient and Safe burr selection	Tool control
Efficient and safe direction of drilling (parallel to critical structures)	Tool control
Maintains integrity of tegmen	Violation
Maintains integrity of sigmoid sinus	Violation
Maintains integrity of facial nerve	Violation
Maintains integrity of horizontal semi-circular canal	Violation
Maintains integrity of ossicles	Violation
Maintains integrity of dura	Violation

273

Figure 1: Example slide of an individual metric level of importance and operational definition discussion based on survey to group of experts.

1. Maintains visibility while removing bone.

Maintains visibility of tool while removing bone (Select if you agree)		
Answer Options	Response Percent	Response Count
The outside edge of the burr should be visible throughout the process (The undersurface is never visible)	66.7%	8
Visibility may be contextual. However, for some circumstances, it is acceptable to not see the burr.	25.0%	3
Visibility is essential for the novice. However, it is reasonable for a more experienced surgeon to disregard.	50.0%	6
Comment (please specify)		4
answered question		12
skipped question		1

These metrics differ depending on the skill of the surgeon in identifying where it is safe not to see the bur.
Novices need to see the bone-metal interface at all times, more experiences surgeons can be safe with more "invisible" bone removal.
The burr cuts on its side; the point should not be used; the last two boxes, I have problems with their implications.
Drilling without seeing the bur is dangerous even for experts. although visibility may be contextual in the OR on the simulator the skill should be demonstrated for both experts and novices in order to properly get a metric.
If the experts don't or can't do this on the simulator can you expect this from the novice?