Objective Assessment of Technical Skills in Otorhinolaryngology–Head and Neck Surgery Residents: A Systematic Review

Érika Mercier, MD1, Ségolène Chagnon-Monarque2, François Lavigne, MD, FRCSC1, and Tareck Ayad, MD, FRCSC1

No sponsorships or competing interests have been disclosed for this article.

Abstract

Objectives. The primary goal is the indexation of validated methods used to assess surgical competency in otorhinolaryngology–head and neck surgery (ORL-HNS) residents. Secondary goals include assessment of the reliability and validity of these tools, as well as the documentation of specific procedures in ORL-HNS involved.

Data Sources. MEDBASE, OVID, Medline, CINAHL, and EBM, as well as the printed references, available through the Université de Montréal library.

Review Methods. The PRISMA method was used to review digital and printed databases. Publications were reviewed by 2 independent reviewers, and selected articles were fully analyzed to classify evaluation methods and categorize them by procedure and subspecialty of ORL-HNS involved. Reliability and validity were assessed and scored for each assessment tool.

Results. Through the review of 30 studies, 5 evaluation methods were described and validated to assess the surgical competency of ORL-HNS residents. The evaluation method most often described was the combined Global Rating Scale and Task-Specific Checklist tool. Reliability and validity for this tool were overall high; however, considerable data were unavailable. Eleven distinctive surgical procedures were studied, encompassing many subspecialties of ORL-HNS: facial plastics, general ear-nose-throat, laryngology, otology, pediatrics, and rhinology.

Conclusions. Although assessment tools have been developed for an array of surgical procedures, involving most ORL-HNS subspecialties, the use of combined checklists has been repeatedly validated in the literature and shown to be easily applicable in practice. It has been applied to many ORL-HNS procedures but not in oncologic surgery to date.

Keywords

resident, residency, assessment tool, assessment method, evaluation, evaluation tool, evaluation method, reliability, validity, surgical competency, competency

Received May 14, 2017; revised August 2, 2017; accepted August 7, 2017.

Over the past few years, the conventional framework of residency, where competency is established through time spent rotating on different surgical services, has been gradually replaced with a curriculum based on the objective acquisition of program-specific competencies.1,2 This change is happening in various countries across the world, including the United States, where the Milestones program was implemented by the Accreditation Council for Graduate Medical Education. In Canada, the Royal College of Physicians and Surgeons is gradually implementing the Competence by Design program.4 The Intercollegiate Surgical Curriculum Program in the United Kingdom has already integrated the Procedure Based Assessment as part of its surgical programs.5 These programs have in common the role of providing assessment of trainees’ operative technical skills for routine surgical procedures until certification. For surgical residency programs such as otolaryngology, these new curriculums imply the acquisition of clinical and technical skills. In the current system, it is customary for the assessment of a resident’s technical skills to be evaluated informally by attending surgeons during operating time or retrospectively at the end of a resident’s rotation. However, this method does not ensure that residents receive reproducible, consistent, and standardized feedback to help them identify their strengths and weaknesses.6 For program directors, this method also does not provide a means of objectively and accurately ensuring...
the acquisition of varied surgical skills.7 To palliate this problem, various assessment tools have been progressively developed and validated to provide a traceable and dependable way to evaluate surgical skills in otolaryngology residents.8 Despite this progress, many standard otolaryngology procedures are devoid of specific evaluation methods.9 The main objective is to review the available literature to identify the methods currently available and validated to assess surgical competency in otolaryngology residents. We also aim at reporting the reliability and validity of these assessment tools, as well as identifying specific procedures or areas of subspecialties in otolaryngology for which the development of surgical assessment tools might still be needed.

**Methods**

**Database Review**

To answer the aforementioned questions, we conducted a systematic review following the guidelines of the PRISMA statement. An extensive research was conducted on January 18, 2016, with the help of a professional health librarian affiliated to our institution. All relevant published studies were extracted from 5 databases (MEDBASE, OVID, Medline, CINAHL, and EBM), as well as the printed literature available through the Université de Montréal library. The terminology used for the search was as follows: Aptitude Tests OR Operative Competence OR Clinical Competence OR Comprehensive Assessment OR Surgical Skill OR Surgical Aptitude OR Aptitude Test OR Dexterity Test OR General Competence OR Surgical Performance OR Measuring Competence OR Drilling Skill OR Clinical Competence OR Educational Measurement OR Competency-Based Education AND Otolaryngology OR Otolaryngology OR Sinusitis/Surgery OR Rhinitis/Surgery OR Otologic Surgical Procedures OR Bronchoscopy/Standards OR Laryngoscopy/Standards OR Otologic Surgical Procedures OR Mastoid OR Nasal Septum/Surgery OR Paranasal Sinuses/Surgery OR Rhinoplasty/Standards OR Thyroidectomy/Standards OR Tonsillectomy/Standards OR Ear/Surgery OR Nose/Surgery OR Larynx/Surgery AND Internship and Residency OR Residency OR Internship OR Resident OR Education, Medical, Graduate OR Trainee OR Training OR Graduate.

This search strategy resulted in the extraction of a total of 1453 studies from the available digital literature and 14 articles from printed literature. Two reviewers checked the references for duplicates, and a total of 1416 original articles were retained for analysis. All article titles and abstracts were reviewed independently by E´.M. and S.C.M. and were retrieved if both reviewers considered the study to be relevant. Any disagreement between the two reviewers was solved by discussion to reach consensus or by a third party who had final say (T.A.), if necessary. Criteria used for inclusion of studies were as follows: research written in English or French, published after 1990 and containing information on assessment tools for resident evaluation in otorhinolaryngology–head and neck surgery (ORL-HNS).

Assessment tools from other surgical specialties were not considered in this review. Exclusion criteria were as follows: not medical education, not ear-nose-throat specific, not specific to assessment of surgical skills, not a study/not the desired type of research paper (editorial, opinion essay, open letter), not related to or applied to residents, and not available in full text (article could not be retrieved in its entirety).

**Full Manuscript Analysis**

Application of these criteria resulted in 43 abstracts identified as possible relevant studies. These 43 full-text articles were then assessed for eligibility. To identify relevant publications, the articles were evaluated independently by the same 2 reviewers (E´.M., S.C.M.), and the following aspects were assessed and analyzed: the type of article, the nature of the tool described in the paper, the otolaryngology subspecialty involved, the surgical procedure for which the tool was developed, the overall risk of bias, and the main conclusions of the authors. In cases where only abstracts of the original articles were available through regular research in databases and search engines, an advanced search was conducted by the health studies librarian to gain access to the missing article. Through this process, 13 studies did not meet the inclusion criteria and were excluded for the following reasons: 1 study assessed surgeon’s accreditation; 1 determined the surgical aptitudes of otolaryngology applicants; 1 did not validate an assessment tool; 1 was not applied to residents; 1 was only available in Japanese, even though the abstract was in English; 4 were not specific to ORL-HNS procedures; 3 were only abstracts of oral presentations; and 1 focused on evaluating the components that contribute to measurement error but was not a true analysis of the scale or its usefulness as an evaluation tool. In the end, 30 publications met our inclusion criteria and were included in our final analysis. The flowchart covering the search and analysis of available literature is described in Figure 1.

**Data Extraction**

For each of the 30 articles retained, the following information was obtained: year and location of publication, type of article, evaluation method, number and type of participants, subspecialty of ORL-HNS, surgical procedure involved, as well as general conclusions. Subspecialties of ORL-HNS included facial plastics, general otolaryngology, laryngology, otology, pediatrics, rhinology, skull base/neurotology, and oncology. Studies were then organized by method of evaluation used and by type of procedure involved. Data were indexed and are organized in Appendix 1 (available in the online version of the article).

Among the studies included in the qualitative synthesis, an additional assessment of the reliability and validity of the evaluation tools was performed by the same 2 reviewers (É.M., S.C.M.), based on the evaluation elements highlighted by Gail M. Sullivan, in an editorial on the validity of assessment instruments.10 Evaluation of reliability was
based on internal consistency, interrater consistency, and intrarater consistency, as determined, respectively, by the Cronbach’s alpha value or Pearson’s correlation coefficient, the Kappa value, and the intraclass correlation coefficient. Analysis of validity consisted of appraisal of the instrument’s content, the response process, as defined by the relation between the subject’s thoughts and the test results; the relationship of the instrument to other variables (eg, a previous study or a predefined gold standard); and the consequences, defined as the pass-or-fail performance in the instance of a cutoff performance score.

For analysis of reliability, each article was scored according to each of these criteria as low, medium, or high by the 2 reviewers (E.M., S.C.M.) according to Cronbach’s alpha and kappa values. Values <0.49 were marked low, values between 0.50 and 0.75 were medium, and values ≥0.75 were marked high. When information about these criteria was unavailable, the item was marked N/A. The measure of the components of validity as described by Sullivan10 remain subjective. As such, general guidelines were used to classify the articles into high, medium, and low categories. High scores were attributed to articles with the presence of objective values confirming elements of high validity scores, the use of validated methods to develop tools (eg, Delphi method), extensive description of a quality methodology, or evidence in the manuscript that the component of validity described was fully in compliance with the description made by Sullivan. Low scores were attributed if no validated or objective method was used to develop the new tool, in the presence of objective evidence of low validity scores, or in the absence of elements in compliance with the descriptions by Sullivan for each component. Finally, medium scores were attributed if a mix of elements from high and low scores was present or if elements did not satisfactorily fit the criteria to qualify into one category or the other.

Results

Study of Surgical Competency Evaluation Methods

Five evaluation methods were described throughout the 30 articles reviewed. They are the combined Global Rating Scale (GRS) and Task-Specific Checklist (TSC), automated analysis of the surgeon’s motions, simulations on animal models, virtual reality technology, and written practical examinations. The combined checklist tool was by far the most used evaluation method in the literature, described in 25 of 30 articles reviewed. It is composed of 2 complementary checklists, the first being the GRS, originally based on the Objective Structured Assessment of Technical Skill, developed by Martin et al,11 which has been repeatedly used and validated in different specialties.12-15 It comprises a set of general skills applicable to any surgical procedure from whichever specialty, which must be mastered to become a proficient surgeon.16 The second scale is the TSC, which comprises a list of skills and steps specific to a surgical procedure that are necessary to the safe and autonomous completion of the procedure.17

These tools can be used either in live surgery cases or in a simulation setting. The checklists cited in this review were used to assess surgical competency in residents in both situations. The second-most prevalent evaluation method, the automated analysis of motion, comprises devices that record and analyze micromovements of the surgeon’s hand during a procedure and examine patterns of motion associated with skill and experience.18 Simulation on animal models includes completing procedures on cadavers and/or live anesthetized animals, to assess the surgical skills of the primary surgeon.19 Virtual reality has been used to simulate surgical situations and cases, enabling the resident to explore techniques safely and redo procedures as needed, while allowing automated or direct staff evaluation.20 Finally, written examinations imply having residents complete a standardized set of printed or computed-based questions, ranging from drawing a step of the procedure to providing short answers or descriptions.21 A graph analysis of these evaluation methods as well as their representation throughout available literature is presented in Figure 2.

Indexation of Surgical Procedures on Which Residents Are Evaluated

As mentioned previously, the type of surgical procedure on which residents were evaluated and the area of ORL-HNS involved for each assessment tool were indexed during the full-article analysis. These data were used to put forth procedures for which validated assessment tools have been developed and, in the same way, highlight any gaps in the
existing literature on the subject. The complete indexation of represented procedures and corresponding areas of ORL-HNS are shown in Figure 3. Eleven procedures were found to already have at least 1 specific evaluation tool, from 6 areas of ORL-HNS: facial plastics, general ear-nose-throat, laryngology, otology, pediatrics, and rhinology. The procedures most often used to assess residents’ levels of surgical skill were temporal bone dissection and endoscopic sinus surgery, with 8 and 7 studies focusing on these procedures, respectively.

**Assessment of the Reliability and Validity of the Evaluation Methods**

As stated by Gail M. Sullivan’s criteria, the assessment of reliability was based on internal consistency and inter- and intrarater consistency for each evaluation tool. Internal consistency serves to confirm that the tool tests for the concept of interest, while intrarater and interrater consistency test respectively the outcomes of having only 1 versus many raters use the tool, to determine the reproducibility of the results. Each of these criteria was ranked as high, moderate, or low. Internal consistency was overall high, with 15 tools ranking high, 3 ranking moderate, and none scoring low. However, 12 papers did not present sufficient data to draw useful conclusions in the internal consistency criteria. Interrater consistency was ranked high in 7 articles, moderate in 6, and low in 3. Fourteen articles have insufficient information to permit useful analysis. More important, very little information was available to assess intrarater consistency, as 26 articles of 30 had insufficient data. Two articles were rated high for this item, 2 were rated moderate, while none were rated low. Aforementioned data are depicted in Figure 4.

Assessment of validity was done in the same manner, with the following criteria: content, response process, relationship with other variables, and consequences. “Content” encompasses the quality of the process and the steps leading to the creation of the tool, including the people and concepts involved. The “response process” evaluates if the subject’s test results reflect his or her actions. It also encompasses the quality of the rater’s training as well as the instructions given to the subjects prior to evaluation. The “relationship with other variables” criterion relates to the way that the evaluation tool compares with similar tools or a gold standard. Last, “consequences” refers to the capacity of the evaluation tool to reflect better scores after additional
training. The majority of articles scored high on the content criterion, with 19 ranking high, 8 moderate, and 1 low. Data were unavailable for 2 studies. On the response process criterion, 10 evaluations tools were rated high, 4 moderate, and 2 low. However, 14 studies did not have sufficient data for useful analysis. In regard to the relationship with other variables, as many as 17 studies had insufficient data for analysis. Within the ones assessed, 9 tools scored high, 3 moderate, and 1 low. Finally, the consequences criterion was rated high on 13 studies; no studies were rated moderate or low. Last, data were insufficient for 17 studies. This information is summarized in **Figure 5**.

Similar analyses were conducted but this time only with the studies in which the primary assessment method was the combined GRS and TSC tool, to gain perspective on the inherent reliability and validity of this widely used assessment tool. The GRS and TSC checklists are the most reviewed and, possibly, the most promising prospective resident evaluation methods. They are straightforward, inexpensive, readily available, and easily applicable to an academic practice, where time is often a limiting factor.\textsuperscript{23-25}

These analyses put forth the results of the validity and reliability extracted from their corresponding studies. They show favorable outcomes overall, with 67.7% of studies displaying high reliability and 74.5% exhibiting high validity, within the studies displaying sufficient data for analysis. Specifically, the reliability scores were predominantly high on all 3 predefined criteria—internal consistency, interrater consistency, and intrarater consistency, with 14, 7, and 1 studies rating high, respectively. Comparatively, only 3, 5, and 1 studies were rated moderate, while a mere 2 studies were ranked low in the interrater consistency measure. However, 7, 10, and 22 studies had insufficient data in each of the 3 categories. These results are summarized in graph form in **Figure 6**. In the same manner, validity scores were predominantly high, with content, response process, relationship to other variables, and consequences criteria scoring high in 16, 8, 6, and 11 studies, respectively. In the same order, 6, 8, 6, and 0 scored moderate, while only 1 scored low in the content and response process criteria. Finally, 1, 12, 15, and 13 studies had insufficient data for analyses. **Figure 7** summarizes these data.
Discussion

The array of tools existing to assess surgical competency in residents is diverse and explores a variety of approaches. Newly developed tools, such as the automated analysis of motion tool and the virtual reality tool, integrate innovative technology to standardize, quantify, and objectify assessments. However, these methods are often costly, require specific equipment, and might get quickly outdated. They might not yet be suited for a widespread use in an everyday academic setting, because of lack of applicability. While a tool such as simulation on animal models can model reality very closely, it also has the limitations of high costs, the need for dedicated infrastructure, the inevitable anatomic differences with human specimens, and potential ethical issues. It might remain an interesting choice for teaching purposes but might not be appropriate for sequential monitoring of resident proficiency. The written practical examination has the advantage of being very applicable but tests better for knowledge than skill and thus might be a good complementary tool to another form of technical evaluation. These potential limitations might explain why few studies address these various methods. Finally, the GRS/TSC tool can allow for the evaluation of surgical technical competencies in an operating room setting; it is procedure specific and applicable in an academic setting; and it does not require any special equipment. These characteristics, combined with the relative simplicity of development, might explain why it is overrepresented in our review. It should be noted, however, that a limitation of this study is the potential presence of a publication bias, which would favor an overrepresentation of assessment tools showing favorable results in terms of acceptability, validity, and reliability. This could lead to overly optimistic conclusions about these assessment tools’ performances and so give a skewed appreciation of their performance. Also, many studies in our review used the checklists in simulation situations rather than actual operating room cases. This is another potential limitation of this review, as the results could introduce confounders. A simulation context does not include some of the specificities and challenges of live surgery.

As stated, the most prevalent techniques are temporal bone dissection and endoscopic sinus surgery. A possible explanation is that these procedures are relatively straightforward, with a commonly agreed-on set of steps to successfully complete the procedure, making them easier to
Within the scope of procedures that ORL-HNS residents can perform, oncologic surgical techniques, such as neck dissection, are myringotomies or tonsillectomies. However, some basic straightforward skill assessments for procedures such as myringotomies or tonsillectomies have been developed. Nevertheless, in light of this study, many oncologic surgical techniques, such as neck dissection, are within the scope of procedures that ORL-HNS residents should master during their training.

Conclusions

Medical education, like the practice of medicine itself, is a field of constant evolution and reinvention. The recent changes in the format of residency give residents the opportunity to tailor their residency to meet their individual learning needs and skills. An array of assessment tools has been developed over the last years to meet the new evaluation needs and to enable staff to monitor residents’ progress. Five major means of assessing surgical competency have emerged in this review: the combined GRS and TSC, the automated analysis of the surgeon’s hand motions, simulation on animal models, virtual reality, and written practical examination. Overall, these tests demonstrated high reliability and validity scores when analyzed according to Gail M. Sullivan’s criteria. On its own, the combined GRS and TSC tool, which was used in >83% of all assessment studies, presented validity and reliability results ranking predominantly in the high range, with publication biases being a possible limitation. Note that while this tool has been frequently adapted for many fundamental surgeries of ORL-HNS, it has not yet been used for training in oncologic surgeries. Nevertheless, in light of this study, many validated assessment methods are now available and may serve as templates on which to tailor new ways to assess resident progress and success in the future.

Acknowledgments

To Daniela Ziegler, librarian and health care reference and information specialist at Centre Hospitalier de l’Université de Montréal library and reference center, for her help in planning and executing a methodical and thorough review of paper and database references.

Author Contributions

Érika Mercier, substantial contributions to the conception and design of the work; the acquisition, analysis, and interpretation of data for the work, study conception and design, writing and technical editing of the manuscript, gave final approval of the version to be submitted and any revised version; Ségolène Chagnon-Monarque, substantial contributions to the conception and design of the work; the acquisition, analysis, and interpretation of data for the work, writing assistance, participated in writing and technical editing of the manuscript, collaboration in the general development of the project; François Lavigne, contributed to drafting the work and revising it critically for important intellectual content, writing assistance, agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved, general supervision of the research group, collaboration in the general development of the project, brainstorming and research orientation, proofreading; Tareck Ayad, contributed to drafting the work and revising it critically for important intellectual content, study conception and design, agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved, general supervision of the research group, brainstorming and research orientation, proofreading, gave final approval of the version to be submitted and any revised version.

Disclosures

Competing interests: None.
Sponsorships: None.
Funding source: None.

Supplemental Material

Additional supporting information is available in the online version of the article.

References


