Reproducibility of Clinical Grading of Tonsillar Size

Siu Kwan Ng, FRCSEd(ORL); Dennis Lip Yen Lee, FRCSEd(ORL); Albert Martin Li, MD; Yun Kwok Wing, FRCPsych; Michael Chi Fai Tong, MD

Objective: To determine the reproducibility of the Brodsky grading scale and the modified 3-grade and 5-grade scales in reporting the size of the tonsils.

Design: Retrospective review of 60 video recordings of tonsil examination by 12 independent observers with different clinical backgrounds and various levels of training. The sizes of the tonsils were graded using different grading scales.

Setting: Tertiary care university hospital.

Participants: The video recordings were chosen from an ongoing epidemiologic study of sleep-related breathing disorder in children in Hong Kong.

Main Outcomes Measures: The intraobserver and interobserver reproducibility of each grading scale was determined using intraclass correlation. An intraclass correlation coefficient (ICC) exceeding 0.75 was set a priori to indicate an acceptable level of reliability.

Results: The mean intraobserver ICCs for the Brodsky grading scale and the modified 3-grade and 5-grade scales were 0.858, 0.830, and 0.865, respectively. The mean interobserver ICCs for the Brodsky grading scale and the modified 3-grade and 5-grade scales were 0.763, 0.739, and 0.783, respectively.

Conclusion: The Brodsky grading scale and the modified 5-grade scale achieved acceptable intraobserver and interobserver reproducibility.

Arch Otolaryngol Head Neck Surg. 2010;136(2):159-162

Adenotonsillar enlargement is the major etiologic factor of obstructive sleep apnea in otherwise healthy children.1,2 The American Academy of Pediatrics recommends adenotonsillectomy as the first-line treatment for most pediatric obstructive sleep apnea.2 Although this is well accepted conceptually, there is no specific definition of what constitutes larger-than-normal tonsils or adenoids. Part of the difficulty is that there is a complex and dynamic interaction among other airway variables, and there is no absolute cutoff for larger than normal size. Another major problem is the lack of a generally accepted, reliable, and valid grading system. This is particularly relevant for purposes of research that necessitates reproducible assessment of the size of the tonsils and adenoids.

Recently, Parikh et al3 proposed an endoscopic grading system for adenoid size based on the anatomic structures (including torus tubaris, vomer, and soft palate) in contact with adenoid tissue. Although that grading system does not have equal grade intervals of adenoid size, it was shown to have high interobserver reproducibility.

For tonsillar size, several grading systems have been adopted in previous studies.4-10 Among them, the most well-known and accepted grading scale was proposed by Brodsky.9 On the Brodsky grading scale, the size of the tonsils is categorized as 1 of 5 grades based on the percentage of oropharyngeal airway occupied by the 2 tonsils. The oropharyngeal airway is denoted by the linear distance between the 2 anterior tonsillar pillars. Despite its popularity, to our knowledge, the reproducibility of such a grading scale has not been assessed.

Theoretically, the more meticulous is the grading, the more precise is the description. At the same time, grading is expected to be less reproducible given the intrinsic error of repeated measurements. Therefore, a good grading scale needs to strike the best balance between precision and reproducibility.

The objectives of this study were to assess the reproducibility of the Brodsky grading scale and of modifications to the Brodsky grading scale. The modifications involved conversion to 3-grade and 5-grade scales by varying the cutoff values.
STUDY DESIGN

This study is a part of an ongoing epidemiologic study of sleep-related breathing disorder in children in Hong Kong that was approved by the ethics committee of The Chinese University of Hong Kong.11,12 The recruited children underwent examination of the tonsils and adenoids using flexible endoscopes (Olympus P4; Olympus, Tokyo, Japan). On examination of the tonsils, the tongue was in the neutral natural position in the mouth and was gently pressed onto the floor of the mouth. The appearance of the tonsils was video recorded using the flexible endoscope, which was placed inside the oral cavity. A random sample of 60 such video recordings was obtained for this study.

This study recruited 12 independent observers, including 2 otolaryngology specialists, 2 otolaryngology residents, 2 pediatric specialists, 2 pediatric residents, 2 family physicians, and 2 interns. They reviewed the video recordings separately and visually assessed the size of the tonsils, as gauged by their medial extension compared with the width of the oropharyngeal airway (defined as the linear distance between the anterior tonsillar pillars at the midtonsillar level). To aid in visual judgment and to account for some cases of asymmetric tonsils, the independent observers were asked to assess the tonsillar size in the following manner. The medial extension of the tonsil on one side was estimated first. The percentage of horizontal distance of the ipsilateral hemiopharynx at the midtonsillar level (from the anterior tonsillar pillar to the midline of the uvula) occupied by that tonsil ranges from 0% (if it remains at and lateral to the anterior tonsillar pillar) to 100% (if it reaches the midline). The result was recorded. This was repeated for the contralateral tonsil. Then, the size of both tonsils (represented by the horizontal distance of the medial extension) as a percentage of the whole oropharynx (represented by the distance between the anterior tonsillar pillar at the midtonsillar pole) was derived by taking the mean of these 2 values. Having the estimated percentage in mind, the independent observers then decided on the tonsillar grade using the different grading scales, which were recorded on a data sheet.

After the first round of observations, the video recordings were shuffled, and the independent observers reviewed the video recordings the next day. The same estimation procedures were repeated. Therefore, each observer had 2 sets of observations for the same collection of video recordings. The intraobserver and interobserver reproducibility of the different grading scales for tonsillar size was determined.

TONSILLAR SIZE GRADING SCALES

In this study, 3 grading scales were used. These include the Brodsky grading scale and the modified 3-grade and 5-grade scales.

Brodsky Grading Scale

The Brodsky grading scale comprised the following 5 grades: grade 0 (tonsils within the tonsillar fossa), grade 1 (tonsils just outside of the tonsillar fossa and occupy 25% of the oropharyngeal width), grade 2 (tonsils occupy 26%-50% of the oropharyngeal width), grade 3 (tonsils occupy 51%-75% of the oropharyngeal width), and grade 4 (tonsils occupy >75% of the oropharyngeal width).

Modified 3-Grade Scale

The modified 3-grade scale comprised the following 3 grades: grade 1 (tonsils occupy ≤33% of the oropharyngeal width), grade 2 (tonsils occupy 34%-66% of the oropharyngeal width), and grade 3 (tonsils occupy >66% of the oropharyngeal width).

Modified 5-Grade Scale

The modified 5-grade scale comprised the following 5 grades: grade 1 (tonsils occupy ≤20% of the oropharyngeal width), grade 2 (tonsils occupy 21%-40% of the oropharyngeal width), grade 3 (tonsils occupy 41%-60% of the oropharyngeal width), grade 4 (tonsils occupy 61%-80% of the oropharyngeal width), and grade 5 (tonsils occupy >80% of the oropharyngeal width).

STATISTICAL ANALYSIS

The intraobserver and interobserver reproducibility for the different grading scales was determined using the statistical method of intraclass correlation.13 Cronbach α and the intraclass correlation coefficient (ICC) were calculated. An ICC of 0 indicates complete lack of agreement beyond chance, and an ICC of 1 indicates perfect agreement between the sets of observations. An ICC exceeding 0.75 was set a priori to indicate an acceptable level of reliability.13,14

RESULTS

The intraobserver reproducibility of the tonsillar size estimation for the 12 independent observers is given in Table 1. The intraobserver ICC exceeded 0.75 for all independent observers using all grading scales except for 1 observer. This observer’s ICC was below 0.75 using all grading scales. The mean intraobserver ICCs for the Brodsky grading scale, modified 3-grade scale, and modified 5-grade scale were 0.858, 0.830, and 0.865, respectively. The interobserver reproducibility of the tonsillar size estimation for the 12 independent observers is given in Table 2. Because each observer had 2 sets of observations, the interobserver ICC was calculated for both sets of data. The mean interobserver ICCs for the 2 sets of observations using the Brodsky grading scale, modified 3-grade scale, and modified 5-grade scale were 0.763, 0.739, and 0.783, respectively.

COMMENT

A prerequisite for a useful clinical grading system is good reproducibility of the results, including observations made at different times by the same rater or observations made by different observers. In addition, physicians with various backgrounds should be able to successfully use it. Therefore, 12 independent observers with different clinical backgrounds and various levels of training were invited to join our study. The specialties chosen were those most frequently involved in managing diseases of the tonsillar area.

The independent observers were asked to evaluate the tonsil on each side separately and then to arrive at a composite grade. We believe that this method of evaluation...
is essential to maximize reproducibility. To explain this point, imagine a hypothetical patient in whom the 2 tonsils were situated on the same side adjoining each other, with the oropharyngeal airway free on the other side. The percentage of oropharyngeal airway occupied by the tonsils would be more easily and accurately estimated than in a patient with normal anatomy in which the oropharyngeal space is situated between the 2 tonsils. This is particularly relevant for patients with asymmetric tonsils. By evaluating the tonsil on each side separately, the tonsils-free space comparison is facilitated.

Our results indicated that the intraobserver (except for 1 observer) reproducibility was acceptable for all 3 scales. The interobserver reproducibility for the Brodsky grading scale and for the modified 5-grade scale was also acceptable, but the interobserver reproducibility for the modified 3-grade scale was just below the level of acceptable reliability. Therefore, while the intraobserver ICCs were similar among the 3 scales, the Brodsky grading scale and the modified 5-grade scale yielded better interobserver ICCs than the modified 3-grade scale. This seemed counterintuitive because a modified 3-grade scale has a wider grade interval and supposedly can accommodate greater error of repeated measurements. It is difficult to give a concrete explanation. We speculate that our eyes may be more sensitive to halves and quarters than to thirds. Still, it is elusive to find that the modified 5-grade scale has an advantage of allowing an error range at the important quarters. To illustrate this point, consider a patient in whom the tonsils appear to occupy about 50% of the oropharyngeal airway. If the observer’s impression at the first view is slightly greater than 50% and at the second view is slightly less than 50%, that would translate to grade 3 for the first view and grade 2 for the second view using the Brodsky grading scale. In contrast, it would translate to grade 3 (ie, tonsils occupy 41%-60% of the oropharyngeal width) for both views using the modified 5-grade scale. While these speculations might not sound convincing to all readers, our study findings indicated that the Brodsky grading scale and the modified 5-grade scale were comparable and seemed to provide more reproducible results than the modified 3-grade scale.

Reproducibility of grading systems is essential for effective communication. For practical clinical usefulness, a good scale should also show significant association with polysomnographic findings if other relevant variables (eg, nasal obstruction and facial skeletal problems) are controlled for. Further studies are required to evaluate this aspect of reproducibility of clinical grading.

In conclusion, the Brodsky grading scale and the modified 5-grade scale demonstrated acceptable intraobserver and interobserver reproducibility. They are preferred over the modified 3-grade scale because of their

---

### Table 1. Intraobserver Reproducibility of Tonsillar Size Estimation

<table>
<thead>
<tr>
<th>Independent Observer</th>
<th>Brodsky Grading Scale</th>
<th>Modified 3-Grade Scale</th>
<th>Modified 5-Grade Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cronbach α</td>
<td>ICC</td>
<td>Cronbach α</td>
</tr>
<tr>
<td>Otorhinolaryngology specialist 1</td>
<td>0.941</td>
<td>0.889</td>
<td>0.927</td>
</tr>
<tr>
<td>Otorhinolaryngology specialist 2</td>
<td>0.929</td>
<td>0.866</td>
<td>0.947</td>
</tr>
<tr>
<td>Pediatric specialist 1</td>
<td>0.954</td>
<td>0.913</td>
<td>0.957</td>
</tr>
<tr>
<td>Pediatric specialist 2</td>
<td>0.949</td>
<td>0.904</td>
<td>0.968</td>
</tr>
<tr>
<td>Family physician 1</td>
<td>0.917</td>
<td>0.833</td>
<td>0.872</td>
</tr>
<tr>
<td>Family physician 2</td>
<td>0.731</td>
<td>0.581a</td>
<td>0.616</td>
</tr>
<tr>
<td>Otorhinolaryngology resident 1</td>
<td>0.991</td>
<td>0.982</td>
<td>0.970</td>
</tr>
<tr>
<td>Otorhinolaryngology resident 2</td>
<td>0.936</td>
<td>0.865</td>
<td>0.932</td>
</tr>
<tr>
<td>Pediatric resident 1</td>
<td>0.876</td>
<td>0.777</td>
<td>0.880</td>
</tr>
<tr>
<td>Pediatric resident 2</td>
<td>0.908</td>
<td>0.827</td>
<td>0.909</td>
</tr>
<tr>
<td>Intern 1</td>
<td>0.942</td>
<td>0.892</td>
<td>0.985</td>
</tr>
<tr>
<td>Intern 2</td>
<td>0.984</td>
<td>0.967</td>
<td>0.953</td>
</tr>
<tr>
<td>Mean</td>
<td>...</td>
<td>0.858</td>
<td>...</td>
</tr>
</tbody>
</table>

Abbreviation: ICC, intraclass correlation coefficient.

---

### Table 2. Interobserver Reproducibility of Tonsillar Size Estimation

<table>
<thead>
<tr>
<th>Variable</th>
<th>Brodsky Grading Scale</th>
<th>Modified 3-Grade Scale</th>
<th>Modified 5-Grade Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cronbach α</td>
<td>ICC</td>
<td>Cronbach α</td>
</tr>
<tr>
<td>First set of observations</td>
<td>0.980</td>
<td>0.761</td>
<td>0.977</td>
</tr>
<tr>
<td>Second set of observations</td>
<td>0.981</td>
<td>0.764</td>
<td>0.977</td>
</tr>
<tr>
<td>Mean</td>
<td>...</td>
<td>0.763</td>
<td>...</td>
</tr>
</tbody>
</table>

Abbreviation: ICC, intraclass correlation coefficient.

**a** Intraclass correlation coefficient (ICC) less than 0.75. All other ICCs exceed 0.75, indicating acceptable reproducibility.

**b** Because each observer had 2 sets of observations, the ICC was calculated for both sets of data.

**c** All ICCs exceed 0.75, indicating acceptable reproducibility.
better reproducibility and more precise description of tonsillar size. Their use in research of sleep-related breathing disorder is feasible.

Submitted for Publication: February 22, 2009; final revision received May 22, 2009; accepted June 16, 2009.

Correspondence: Michael Chi Fai Tong, MD, Department of Otorhinolaryngology–Head and Neck Surgery, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin NT, Hong Kong SAR (mtong@ent.cuhk.edu.hk).

Author Contributions: All authors had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Study concept and design: Ng and Li. Acquisition of data: Ng, Lee, and Wing. Analysis and interpretation of data: Wing and Tong. Drafting of the manuscript: Ng and Li. Critical revision of the manuscript for important intellectual content: Ng, Lee, Wing, and Tong. Obtained funding: Li and Wing. Administrative, technical, and material support: Ng, Wing, and Tong. Study supervision: Wing and Tong.

Financial Disclosure: None reported.

REFERENCES


