The Reliability of Clinical Tonsil Size Grading in Children

Divjot S. Kumar, BSc; Dianne Valenzuela, BSc, BA; Frederick K. Kozak, MD, FRCS; Jeffrey P. Ludemann, MD, FRCS; J. Paul Moxham, MD, FRCS; Jane Lea, MD, FRCS; Neil K. Chadha, MBChB (Hons), MPHe, FRCS

Enlargement of the palatine tonsils is associated with substantial ill health consequences in children, reliable monitoring and documentation of tonsil size is necessary in clinical settings. Tonsil grading scales potentially allow clinicians to precisely record and communicate changes in tonsil size, but their reliability in a clinical setting has not been studied.

OBJECTIVE To assess the interobserver and intraobserver reliability of the Brodsky and Friedman tonsil size grading scales and a novel 3-grade scale.

DESIGN, SETTING, AND PARTICIPANTS Cross-sectional study between June 2012 and August 2013 at a tertiary pediatric otolaryngology outpatient clinic at British Columbia Children’s Hospital. We recruited 116 children, aged 3 to 14 years, with no major craniofacial abnormalities. For each child, 2 separate tonsil assessments (with at least a 5-minute interval in between) were conducted by 4 independent observers: 2 staff pediatric otolaryngologists, 1 otolaryngology trainee (fellow or resident), and 1 medical student. Each observer assessed and graded tonsil sizes using 3 different scales.

MAIN OUTCOMES AND MEASURES Interobserver and intraobserver reliabilities were assessed by deriving the intraclass correlation coefficients (ICCs) and Pearson correlation coefficients, respectively. To discount for any asymmetric scores, all data analysis was conducted on the left tonsil measurement only.

RESULTS Mean interobserver reliability was highest for the Brodsky grading scale (ICC, 0.721; Cronbach α, 0.911), followed by the Friedman grading scale (ICC, 0.647; Cronbach α, 0.879) and the 3-grade scale (ICC, 0.599; Cronbach α, 0.857). The mean intraobserver reliabilities for the Brodsky, Friedman, and modified 3-grade scales were 0.954, 0.932, and 0.927, respectively.

CONCLUSIONS AND RELEVANCE The Brodsky grading scale offered the highest interobserver and intraobserver reliability when compared with the Friedman and novel 3-grade scales. The results of this study would support the uniform use of the Brodsky scale for future clinical and research work.
knowledge, the reliability of existing grading systems in a clinical setting has not previously been formally studied. Understandably, doubts may arise pertaining to the relationship between tonsil size and health outcomes when there is an initial degree of uncertainty in measuring tonsil size. There is, therefore, a need to compare existing tonsillar grading scales and assess their reliability and reproducibility in a clinical setting.

Among the currently used grading scales, the 2 most widely adopted are (1) the Brodsky grading scale, in which the tonsils are assigned a grade from 1 to 4, depending on the percentage of oropharyngeal airway occupied by the tonsils; and (2) the Friedman grading scale, which classifies tonsil size using the location of the tonsils relative to surrounding structures in the oral cavity such as the anterior tonsillar pillars. Each of these scales can be used to assess individual tonsil sizes separately.

A recent study by Ng et al measured the reliability of scoring methods, including the Brodsky grading scale and 2 modified versions of this scale using 3 and 5 grades. It was found that the Brodsky grading scale and their 5-grade scale had higher interobserver reproducibility than their 3-grade scale, despite the latter being intuitively simpler to use because it had a wider grade interval. Their study, however, relied on observer measurements from video recordings of tonsillar examination in children. The video recordings were made with a fiberoptic endoscope in cooperative children and therefore are not necessarily reflective of the ability to assess tonsil size in a real-life clinical setting with direct oral inspection. Thus, to obtain a clinically sound measurement of the reliability of various tonsil grading scales, a study comparing them must be performed in an actual clinical setting.

In this study, we aimed to investigate the intraobserver reliability and interobserver reliability of 3 different tonsil grading scales: the Brodsky grading scale, the Friedman grading scale, and a novel modified 3-grade scale that was designed in our institutions. We hypothesized that a 3-grade scale may provide greater intraobserver and interobserver reliability when compared with the other tonsil grading scales as a result of its increased simplicity.

Methods

This nonexperimental, cross-sectional study was conducted in the outpatient clinic in British Columbia Children's Hospital Division of Pediatric Otolaryngology. Ethics approval for this study was obtained from the University of British Columbia Children's and Women's Research Ethics Board. The study was discussed with each parent and child, and written informed consent and assent were obtained prior to data collection. Over the span of June through August 2012 and June through August 2013, we recruited 116 children, aged 3 to 14 years, attending the pediatric otolaryngology clinic. This age range was chosen because it includes the age at which tonsil size has been found to peak in children. Exclusion criteria included craniofacial abnormalities, history of tonsillectomy, and congenital disorders.

For each child, 4 independent observers, at various levels of seniority, visually assessed and measured tonsil size using the Brodsky scale, Friedman scale, and the modified 3-grade scale. All authors of this study took part as observers, but at any one time, there were only 4 observers chosen. This included 1 medical student (either D.S.K. or D.V.), 2 staff otolaryngologists (from among F.K.K., J.P.L., J.P.M., J.L., N.K.C.), and 1 fellow or resident. We included observers with different clinical backgrounds to validate the reliability of the tonsil grading scales irrespective of the level of clinical experience of the user.

Tonsil assessments were conducted by visually inspecting the tonsils and recording a corresponding score using the 3 grading scales. To be as precise as possible, we asked the observers to distinguish between and specify the right and left tonsil grades when the tonsils were considered asymmetrical. The observer then waited 5 to 10 minutes and repeated the assessment. Thus, there were 4 sets of paired observations (4 observers assessing twice) for each child seen. Furthermore, each observer was blinded to their own scores for repeated measurements and to the scores of other observers.

Tonsil size data were entered into a database, and the intraclass reliability was assessed by deriving the intraclass correlation coefficients (ICCs) and Cronbach α, which are statistical measures of interrater reliability. Interrater reliability was derived using the Pearson correlation coefficient. It was predetermined that an ICC of greater than 0.75 would indicate an acceptable reliability level.

In this study, 3 grading scales were applied: the Brodsky grading scale (Table 1), the Friedman grading scale (Table 2), and the novel 3-grade scale (Table 3). The 3-grade scale was created using existing classifications that were divided into 3 grades. Because this scale uses structural references to classify tonsil size, it can be easily applied in clinical settings in which the tonsils are transiently seen in an oral examination. Furthermore, we hypothesized that reducing the number of grades, the simplicity of the scale, and the higher interobserver reproducibility would provide greater intraobserver and interobserver reliability when compared with the other tonsil grading scales as a result of its increased simplicity.

Table 1. Brodsky Grading Scale

<table>
<thead>
<tr>
<th>Grade</th>
<th>Airway, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>≤25</td>
</tr>
<tr>
<td>2</td>
<td>26-50</td>
</tr>
<tr>
<td>3</td>
<td>51-75</td>
</tr>
<tr>
<td>4</td>
<td>&gt;75</td>
</tr>
</tbody>
</table>

Table 2. Friedman Grading Scale

<table>
<thead>
<tr>
<th>Grade</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No tonsils seen</td>
</tr>
<tr>
<td>I</td>
<td>In tonsillar fossa</td>
</tr>
<tr>
<td>II</td>
<td>Visible beyond anterior pillars</td>
</tr>
<tr>
<td>III</td>
<td>Extended 3/4 of way to midline</td>
</tr>
<tr>
<td>IV</td>
<td>Completely obstructing airway; “kissing tonsils”</td>
</tr>
</tbody>
</table>

Table 3. Three-Grade Scale

<table>
<thead>
<tr>
<th>Grade</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Not visible or within tonsillar fossa—“small tonsils”</td>
</tr>
<tr>
<td>II</td>
<td>Enlarged beyond tonsillar pillars but not close to uvula or midline—“medium tonsils”</td>
</tr>
<tr>
<td>III</td>
<td>Extending close to uvula or beyond—“large tonsils”</td>
</tr>
</tbody>
</table>

This study was discussed with each parent and child, and written informed consent and assent were obtained prior to data collection. The study was discussed with each parent and child, and written informed consent and assent were obtained prior to data collection.
grade classifications would allow for reduced variability and increased consistency and reliability.

Results

A total of 116 children were assessed by 4 independent raters with different training levels (2 staff pediatric otolaryngologists, 1 otolaryngology trainee [fellow or resident], 1 medical student). All data analyses were conducted on the left tonsil measurement only, to discount for asymmetric scores. The mean interobserver reliability of the 3 tonsil grading scales is provided in Table 4. Because there were 2 sets of assessments conducted by each observer, we have provided the interobserver reliability for each assessment, as well as the mean values. Mean interobserver reliabilities for the Brodsky, Friedman, and 3-grade scales were 0.721, 0.647, and 0.599, respectively.

Intraobserver reliability is given in Table 5 and was derived using the Pearson correlation coefficient. The mean intraobserver reliabilities for the Brodsky, Friedman, and 3-grade scale were 0.954, 0.932, and 0.927, respectively.

Discussion

Enlarged tonsils in children may be a contributing factor to airflow limitation and various clinical conditions, including OSA. Obstructive sleep apnea is characterized by intermittent pauses in breathing or episodes of complete airflow obstruction during sleep. A patient's suitability for adenotonsillectomy as a potential treatment for OSA may be influenced by objective physical examination findings such as tonsil size. Thus, precise monitoring and careful communication of tonsil size may be a critical component of care of these children. Many research studies on the interventions for OSA include measurement of tonsil size. Therefore, it is imperative to identify a reliable tonsil grading scale for valid conclusions to be drawn about the role of tonsil size in outcome.

To identify a tonsil scale that offers the highest interrater and intrarater reliability, we asked 4 tonsil raters with differing clinical backgrounds to visually grade tonsil size using 3 different tonsil scales: Brodsky, Friedman, and a novel 3-grade scale (designed by N.K.C.). Our study found interobserver reliability to be highest for the Brodsky grading scale, whereas the 3-grade scale had the lowest interobserver reliability. It is notable that all 3 scales exhibited less than the typical threshold for an acceptable reliability level based on an ICC level of 0.75 or greater, although the Brodsky scale was close (ICC = 0.721). Intraobserver reliability was also highest for the Brodsky scale and lowest for the modified 3-grade scale (except for the fellow or resident). It was somewhat unexpected that the modified 3-grade scale had the poorest interobserver reproducibility given its reduced number of levels, which we hypothesized might intrinsically reduce variability and offer higher reliability. This finding is supported by the study of Ng et al., who also demonstrated that the Brodsky grading scale had higher interobserver reliability than a 3-grade scale when using endoscopic tonsil videos. Ng et al. provide a possible explanation for these counterintuitive results by stating that through other life experience, we may be more proficient in distinguishing between halves and quarters than between thirds. Hence, a 3-grade scale would exhibit low interobserver reliability. Another possible explanation is the greater familiarity of physicians with the Brodsky and Friedman scales than the new 3-grade scale. We believe that this prior familiarity may have helped reduce opportunities for confusion when the Brodsky and Friedman grading scales were used, whereas the new 3-grade scale would be susceptible to such errors.

It is important to recognize that although tonsil grading systems may be reliable, the correlation of these scales with the actual severity of obstruction is not clear. Hypertrophied tonsils may represent one of the several possible etiolo-
gies behind airflow limitation and sleep apnea in children. Furthermore, flow limitation may not necessarily result in subjective obstructive symptoms in children. Clinicians should be aware of the limitations of using these grading scales and must consider parental reports of nighttime snoring and presence of other associated obstructive symptoms in order to diagnose OSA in children.

A potential limitation is our intraobserver reliability methodology. It is possible that the second rating by an observer could be influenced by the first assessment. To minimize the influence of observer memory, we asked the observers to tend to other patients in the outpatient clinic during the minimum 5-minute interval between the 2 assessments. This aimed to mimic a clinical setting while minimizing the influence of recall.

Other studies have recognized the importance of accurate tonsil grading and have explored alternate means of assessing tonsil size rather than standard visual grading. Studies have considered radiographic and photographic assessment of tonsil size make them impractical means of tonsil grading at present.

Conclusions

Visual objective examination of tonsil size, in which the clinician is free to expose the tonsils from different angles, is the pragmatic standard method used to assess eligibility for adenotonsillectomy. The results of the present study support the uniform use of the Brodsky scale to grade tonsil size in future clinical and research work.

REFERENCES