Methodological and Statistical Problems in Uvulopalatopharyngoplasty Research

A Follow-up Study

Uchechukwu C. Megwalu, MD; Jay F. Piccirillo, MD

Objectives: To review the published literature on uvulopalatopharyngoplasty (UPPP) and assess the methodological quality of the research and compare it with a similar article published in 1995; and to determine what, if any, improvement in the methodological quality of the research resulted during the ensuing 10 years.

Design: Methodological and statistical evaluation of the published literature on UPPP. Thirty articles representing the clinical studies on UPPP and related procedures written from January 1996 to August 2005 were reviewed. Only articles reporting polysomnography data were included.

Results: Overall, the articles demonstrated fair methodological and statistical quality. Compared with the previous review by Schechtman et al, there was a slight increase in the number of articles that discussed statistical power and reported confidence intervals. There were increases in the mean sample size, the percentage of randomized controlled studies, the number of end points, and the use of validated subjective outcome measures; longer mean follow-up time; and more complete reporting of age and sex information. There was no increase in the percentage of published studies that used a prospective study design. None of the studies that required minimum acceptable baseline values of objective sleep parameter measures for enrollment indicated the use of separate screening and baseline assessments. There were 7 different definitions of sleep apnea and 17 different definitions of success in treatment.

Conclusions: There has been an overall improvement in the quality of the articles published on UPPP since 1995. Several areas still need improvement: use of more prospective studies, decrease in number of end points, use of separate screening and baseline assessments, and consensus in the definitions of sleep apnea and success.

Editorials, and letters were excluded from the review. Articles were excluded if they contained information about snorers who may not have had sleep apnea, or if they lacked appropriate baseline data (ie, apnea index, apnea-hypopnea index, or respiratory disturbance index). Articles were also excluded if (1) the patient population numbered less than 10, (2) they did not report postoperative sleep study results, or (3) the patients had already been described in another study. Thirty articles representing the clinical studies on UPPP and related procedures remained: 12 on UPPP,14-165 on laser-assisted UPPP,17-203 on UPPP combined with tongue reduction,21-236 on UPPP combined with maxillofacial surgery,24-291 on UPPP without tonsillectomy,301 comparing UPPP tongue reduction,21-236 on UPPP combined with maxillofacial surgery,24-291 on UPPP without tonsillectomy,301 comparing UPPP with lateral pharyngoplasty,311 comparing UPPP with transpalatal advancement pharyngoplasty,32 and 1 on modified UPPP.33

A standard data collection form was created to capture information on the key methodological problems identified by Schechtman et al.3 We reviewed the articles separately using the data collection form. After independent review, answers were evaluated, and where there was a discrepancy between us, consensus was reached through discussion.

Subjective outcome measures included measures of snoring, sleepiness, general well-being, and reports of patient symptoms. Validated subjective outcome measures were defined as measures that have demonstrated construct validity and internal consistency. We defined an end point as any measure of outcome that was included in the statistical analysis of outcome. We did not include variables that were evaluated but not reported.

### RESULTS

#### INADEQUATE SAMPLE SIZE AND LITTLE STATISTICAL POWER

Among the 30 articles, the mean sample size for each article was 55 (median size, 46; range of sample size, 13-277). Among the 30 articles, only 3 (10%) discussed statistical power. Although 26 of the 30 UPPP articles (87%) reported P values, only 4 (13%) presented confidence bounds.24 Of these 4, none incorporated confidence bounds in the interpretation of the results.

Eight studies (27%) used a control group to compare one treatment with another. Of the 8 prospective studies that used control groups, only 2 (25%) included randomization for treatment assignment.

#### INADEQUATE FOLLOW-UP

Among the 30 UPPP articles, 27 (90%) provided some information about the length of follow-up, 13 (43%) provided fixed follow-up times, 4 (13%) provided mean follow-up times, 3 (10%) provided minimum follow-up times, 4 (13%) provided a range of follow-up times, and 3 (10%) provided both mean follow-up times and a range of follow-up times. One article reported the follow-up time as “a couple of months,” which we assumed to be 2 months. Four articles provided both short- and long-term follow-up times. Using the shorter time in the articles with both short- and long-term follow-ups, the overall mean follow-up time was 5.1 months and the median was 4 months; the range was 1 to 18 months in the 27 articles that provided the necessary data.

#### RESULTS WITH UNCERTAIN GENERALIZABILITY

Among the 30 UPPP articles, 13 (43%) were definitely prospective, 11 (37%) were definitely retrospective, 2 (7%) were probably prospective, 1 (3%) was probably retrospective, and 3 (10%) were indeterminate. A mean of 26% of patients was lost to follow-up (range, 0%-76%). Seventeen of the 30 studies (57%) had more than 20% of patients lost to follow-up.

#### QOL MEASURES AND MULTIPLE END POINTS

Twenty of the articles (67%) used subjective outcome measures of improvement, such as measures of snoring and excessive daytime sleepiness. Of these 20 articles, 10 (50%) used validated subjective measures.

In the 30 UPPP articles, the mean number of end points was 7.3, and the range of end points was 1 to 33. Five articles (17%) reported 1 or 2 end points, 14 (47%) had...

MISSING DATA AND MISSING OR INCONSISTENT DEFINITIONS

As mentioned in the “Inadequate Follow-up” subsection, 3 articles (10%) did not provide mean follow-up data, and 6 articles (20%) did not specify whether the study was prospective or retrospective. Three articles (10%) did not provide sufficient information to determine conclusively whether the study was prospective or retrospective; however, there were 3 other articles in which the status could be determined from the context, although there was no clear statement as to whether the research was prospective or not. Two articles (7%) did not provide information on the age of the population evaluated, and 1 article (3%) did not provide information on the sex distribution of the sample.

The definition of obstructive sleep apnea (OSA) was not specified in 16 articles (53%). Among the articles that defined OSA, there were 7 different definitions (Table 1). Of the 30 articles, 26 articles (87%) defined criteria for success in treatment. Among these articles, there were 17 different definitions (Table 2). Three articles had separate criteria for “success” and “cure.”

Required minimum acceptable baseline values of outcome measures for enrollment were defined in 13 of the 30 studies (43%). Of these 13, none indicated the use of separate screening and baseline assessments.


In Table 3, we compare our findings in 2005 with those of Schechtman et al.3 Compared with their study, the percentage of articles that discussed statistical power increased, although the 95% confidence interval (CI) around this increase suggests that this difference is not statistically significant (95% CI, −3.4 to 35.2). The mean sample size increased, and thus the power of the studies increased. The percentage of articles reporting CIs increased, although the 95% CI around this increase suggests that this difference also is not statistically significant (95% CI, −2.6 to 23.8). Confidence intervals are important to report because they provide information about the precision of the results; CIs are also helpful in the interpretation of clinical research results.34 Both the percentage of controlled studies and that of randomized controlled studies increased, although the 95% CIs demonstrate that this difference is not statistically significant (95% CIs, −3.4 to 35.2 and −5.0 to 55.0, respectively). The mean follow-up time, use of prospective studies, and the mean number of end points increased, although none of these increases were statistically significant (95% CIs, −7.7 to 31.0 and −4.1 to 3.1, respectively). There was a statistically significant increase in the use of subjective outcome measures (95% CI, 16.4-50.2), especially in the use of validated subjective outcome measures. There was also a statistically significant improvement in the reporting of age and sex information (95% CIs, 7.8-42.8 and 20.1-54.3, respectively). Slightly more articles cited in the current study defined OSA than in the earlier study,3 although this difference was not significant (95% CI, −18.2 to 29.6). It should be noted that there is still no consensus on the definition of sleep apnea. In both reviews, none of the studies that required minimum acceptable baseline values of outcome measures for enrollment indicated the use of separate screening and baseline assessments.

COMMENT

The purpose of this study was to assess the quality of articles that have been published on UPPP in the decade since the 1995 study by Schechtman et al3 and to deter-

Table 3. Comparisons Between the 1995 Study* and the Current Study

<table>
<thead>
<tr>
<th>Key Methodological Problems</th>
<th>1995 Study (n=37 Articles)</th>
<th>Current Study (n=30 Articles)</th>
<th>Difference (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size and statistical power: discussed statistical power</td>
<td>0</td>
<td>10.0</td>
<td>10.0 (−0.7 to 20.7)</td>
</tr>
<tr>
<td>Confidence bounds: reported CIs</td>
<td>2.7</td>
<td>13.3</td>
<td>10.6 (−2.6 to 23.8)</td>
</tr>
<tr>
<td>Uncontrolled studies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controlled studies</td>
<td>10.8</td>
<td>26.7</td>
<td>15.9 (−3.4 to 35.2)</td>
</tr>
<tr>
<td>Randomized controlled studies</td>
<td>0.0</td>
<td>25.0</td>
<td>25.0 (−5.0 to 55.0)</td>
</tr>
<tr>
<td>Adequate follow-up: follow-up information</td>
<td>81.1</td>
<td>90.0</td>
<td>8.9 (−7.7 to 25.5)</td>
</tr>
<tr>
<td>Results with uncertain generalizability: prospective studies</td>
<td>43.0</td>
<td>50.0</td>
<td>7.0 (−17.0 to 31.0)</td>
</tr>
<tr>
<td>Multiple end points: mean number of end points</td>
<td>6.8</td>
<td>7.3</td>
<td>0.5 (−4.1 to 3.1)</td>
</tr>
<tr>
<td>QOL: used validated subjective outcome measures</td>
<td>0.0</td>
<td>33.3</td>
<td>33.3 (16.4 to 50.2)</td>
</tr>
<tr>
<td>Missing data and missing definitions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information on age</td>
<td>68.0</td>
<td>93.3</td>
<td>25.3 (7.8 to 42.8)</td>
</tr>
<tr>
<td>Information on sex</td>
<td>59.5</td>
<td>96.7</td>
<td>37.2 (20.1 to 54.3)</td>
</tr>
<tr>
<td>Defined sleep apnea</td>
<td>41.0</td>
<td>46.7</td>
<td>5.7 (−18.2 to 29.6)</td>
</tr>
<tr>
<td>Biased baseline data: separate screening and baseline assessments</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Abbreviations: CI, confidence interval; QOL, quality of life.

*a Schechtman et al.3 1995.

b The percentage of the controlled studies that was randomized.
mine if there has been an improvement in the methodological quality of the research. Overall, it seems that there has been modest improvement in the quality of the published literature. Notable areas include more randomized studies and the use of subjective outcome measures in assessing the efficacy of treatment modalities. It is well known that the results of objective tests often correlate poorly with symptoms of sleep apnea and the functional impairments associated with it. Although objective measures of sleep apnea severity are important, disease-specific health status and QOL measures need to be evaluated to give clinical relevance to results. Although there has been an increase in the use of validated subjective outcome measures (eg, the Epworth Sleepiness Scale), the relative infrequent use of validated QOL measures remains a weakness in the UPPP literature. Several disease-specific QOL measures are now available. The OSA Patient-Oriented Severity Index,36 the Calgary Sleep Apnea QOL index,37 and the Functional Outcomes of Sleep Questionnaire38 have been shown to be valid disease-specific, health-related QOL measures for OSA.

Several methodological areas still need improvement. There was no appreciable increase in the proportion of the studies that were prospective. Well-performed retrospective cohort studies are very valuable; however, they are more prone to bias than prospective studies. Also, the mean percentage of patients lost to follow-up was 26%, with 57% of the studies having more than 20% of patients lost to follow-up. This high number of patients lost to follow-up highlights the inherent challenge in longitudinal studies. Complete follow-up is difficult to achieve and requires a lot of effort, but it is necessary to minimize bias and achieve a high level of confidence in the results. Patients lost to follow-up introduce bias into the study because the reason for failure to follow up may be linked to the outcome, resulting in a study population that may not be typical of the larger population. Consequently, it may be difficult to assess the generalizability of results. This problem is more prevalent in retrospective studies because they require follow-up as an entry criterion since patients missing are excluded.

There has been an increase in the mean number of end points. Large numbers of end points increase the incidence of type I or “false-positive” errors (the probability of wrongly claiming significance). At the $P = .05$ level of significance, a statistical test has a 5% probability of claiming significance when none exists. When more than 1 statistical test is performed, the probability of wrongly claiming significance in at least 1 of the tests exceeds 5%. In fact, when 10 statistically independent tests are performed, the chance of at least 1 test being significant when in fact there is no statistically significant difference is 40%. Given this, significant results in some of the studies with large numbers of end points may have occurred by chance alone. Multiple end points are not necessarily bad, but they have to be managed appropriately, for example by (1) defining the primary end point, (2) statistical correction, and/or (3) demonstrating consistency across all end points.

None of the studies that required minimum baseline values of sleep parameter measures for enrollment indicated the use of separate screening and baseline assessments. This problem is almost universal in the sleep apnea literature. When a minimum baseline value of an objective measure is required as a criterion for enrollment in a study, it is important to perform a baseline assessment separate from the screening assessment because results on screening assessments may reflect day-to-day variability. Setting minimum baseline values for enrollment biases the baseline data toward higher values because patients with lower values are not included in the study. If the screening values are also used as baseline values, baseline values will be biased estimates of the true values. Thus, posttreatment values will tend to be lower than pretreatment values even when there is no therapeutic effect. This effect is due to regression toward the mean.

Other areas that need improvement include the definition of OSA and the definition of criteria for treatment success. Most of the articles we reviewed did not define OSA. In addition, there were 7 different definitions for OSA. This suggests that there has been no improvement in this area over the past 10 years. Similarly, there were 17 different definitions of treatment success. These problems are prevalent in the sleep apnea literature and are not limited to the UPPP literature. This highlights the need for consensus on the definition of OSA and the definition of success when evaluating the efficacy of therapeutic modalities for OSA.

This study has its limitations. There were occasional disagreements between the numbers of end points initially reported by each of us on independent review. However, consensus was reached on further review by both of us. To match the report by Schechtman et al,3 we reviewed only UPPP articles that included polysomnography outcomes. There may be high-quality studies focusing on subjective outcomes, which we did not review. Therefore, this review may underestimate the use of patient-based subjective measures in sleep apnea research.

In summary, there has been an improvement in the overall quality of the articles published on UPPP since 1995. However, certain areas still need improvement. Clinical researchers, journal reviewers, and editors should insist on higher methodological and statistical standards in an effort to improve the care of patients with OSA. In addition, expert opinion and thought leaders should exclude from their reviews articles and suggestions that do not demonstrate high standards.

Submitted for Publication: March 27, 2007; final revision received November 27, 2007; accepted December 3, 2007.

Correspondence: Jay F. Piccirillo, MD, Clinical Outcomes Research Office, Department of Otolaryngology—Head and Neck Surgery, Washington University School of Medicine, Campus Box 8115, 660 S Euclid Ave, St Louis, MO 63110 (piccirilloj@ent.wustl.edu).

Author Contributions: Both 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: Megwalu and Piccirillo. Acquisition of data: Megwalu and Piccirillo. Analysis and interpretation of data: Megwalu. Drafting of the manuscript: Megwalu. Critical revision of the manuscript for impor-