Cervical vertebrae maturation method morphologic criteria: Poor reproducibility

Trenton S. Nestman, Steven D. Marshall, Fang Qian, Nathan Holton, Robert G. Franciscus, and Thomas E. Southard
Brighton, Colo, and Iowa City, Iowa

Introduction: The cervical vertebrae maturation (CVM) method has been advocated as a predictor of peak mandibular growth. A careful review of the literature showed potential methodologic errors that might influence the high reported reproducibility of the CVM method, and we recently established that the reproducibility of the CVM method was poor when these potential errors were eliminated. The purpose of this study was to further investigate the reproducibility of the individual vertebral patterns. In other words, the purpose was to determine which of the individual CVM vertebral patterns could be classified reliably and which could not. Methods: Ten practicing orthodontists, trained in the CVM method, evaluated the morphology of cervical vertebrae C2 through C4 from 30 cephalometric radiographs using questions based on the CVM method. The Fleiss kappa statistic was used to assess interobserver agreement when evaluating each cervical vertebrae morphology question for each subject. The Kendall coefficient of concordance was used to assess the level of interobserver agreement when determining a "derived CVM stage" for each subject. Results: Interobserver agreement was high for assessment of the lower borders of C2, C3, and C4 that were either flat or curved in the CVM method, but interobserver agreement was low for assessment of the vertebral bodies of C3 and C4 when they were either trapezoidal, rectangular horizontal, square, or rectangular vertical; this led to the overall poor reproducibility of the CVM method. These findings were reflected in the Fleiss kappa statistic. Furthermore, nearly 30% of the time, individual morphologic criteria could not be combined to generate a final CVM stage because of incompatible responses to the 5 questions. Intraobserver agreement in this study was only 82%, on average, when the inconclusive stagings were excluded as disagreements. Intraobserver agreement was worse (44%) when the inconclusive stagings were included as disagreements. For the group of subjects that could be assigned a CVM stage, the level of interobserver agreement as measured by the Kendall coefficient of concordance was only 0.45, indicating moderate agreement. Conclusions: The weakness of the CVM method results, in part, from difficulty in classifying the vertebral bodies of C3 and C4 as trapezoidal, rectangular horizontal, square, or rectangular vertical. This led to the overall poor reproducibility of the CVM method and our inability to support its use as a strict clinical guideline for the timing of orthodontic treatment. (Am J Orthod Dentofacial Orthop 2011;140:182-8)

Accurate prediction of craniofacial growth has proven to be deceptively problematic. Whereas the study of craniofacial growth has shown growth patterns that might apply to the general population, these patterns are not as reliable when predicting growth at the individual level.\(^1\)\(^-\)\(^3\) Still, the pursuit of growth prediction methods continues, because of the potentially significant value of a simple and accurate prediction scheme during the diagnosis and treatment of patients with skeletal discrepancies.

One important aspect of craniofacial growth prediction involves the assessment of a patient’s skeletal age, which could aid in timing orthodontic treatment with the facial growth spurt, particularly in Angle Class II patients. Recently, a method to assess skeletal maturation as it relates to facial growth was devised by Lamparski,\(^4\) who created a set of standards for cervical vertebrae maturation (CVM) using lateral cephalograms and correlated this to hand-wrist radiographs. He reported that his set of standards was as accurate as the hand-wrist method, without additional radiation exposure to the patient. Subsequently, numerous authors have studied the relationship between CVM and skeletal maturity based on hand-wrist
radiographs,5-16 and others have examined the correlation between CVM and mandibular growth.7-23

Careful review of the literature on the CVM method showed potential methodologic errors that might influence the reported reproducibility of this method. Most of these studies reported interobserver and intraobserver reproducibility levels of greater than 90%,5,8-11,13 with the exception of Kucukkeles et al.7 who found that 2 of the 3 intraobserver tests of reproducibility were 45% and 65%. Most studies reporting high levels of reproducibility used tracings of lateral cephalograms, rather than actual radiographs, to determine CVM stages.4-9,16 Some latitude must be granted to a person in tracing cephalometric radiographs, since this is not an exact science, and cephalometric radiographs traced by another person might introduce a level of uniformity into the staging process that is not found if each observer performs an independent analysis. Therefore, the use of standardized tracings might inflate the levels of reproducibility reported in the literature. Furthermore, the observers performing the tests of reproducibility are often the authors themselves. It is possible that they have a “research-level” understanding of the CVM method that could overstate the reported levels of reproducibility.

Many studies reporting CVM reproducibility levels had small sample sizes, some of which appeared to be significantly reduced from larger samples, so that the overall randomness of the sample was in question.17,18,22 In many instances, the same sample was used repeatedly in subsequent studies, and the authors failed to test their results on separate, larger, and more random samples. In addition, some studies reported values of reproducibility with the Pearson correlation coefficient, which is a measure of association between 2 normally distributed interval types of variables. However, a more stringent measure of association for use with categorical data is recommended for measuring agreement between judges.24-27

The reproducibility of the CVM method was recently assessed by Gabriel et al.,1 who attempted to use methods to “eliminate the methodologic shortcomings of previous studies.” In that study, 30 subjects (15 boys, 15 girls) were randomly selected from the University of Iowa Facial Growth Study. Their age range was 10 to 16 years. Thirty lateral cephalograms (1 for each subject) were placed in random order and presented to 10 private-practice orthodontists who served as judges and were asked to assign a CVM stage (1-6) to each subject. The judges were then asked to stage the same subjects 3 weeks later. From these results, intraobserver and interobserver levels of agreement were evaluated. The authors found that interobserver agreement for CVM staging among practicing orthodontists was below 50%; on average, the clinicians agreed with their own staging only 62% of the time (ranging from a high of 80% for 1 clinician to a low of 43% for 2 clinicians); and the reproducibility of trained clinicians was significantly below the level purported in the literature. Based on these results, the authors could not support the use of the CVM method as a strict clinical guideline for timing orthodontic treatments.

For any measurement to be of value, it must be reproducible. But why does the CVM method demonstrate poor reproducibility? The purpose of this study was to expand on the study by Gabriel et al. to determine why the reproducibility of CVM staging was poor. Specifically, our objective was to answer the following questions. What was the level of agreement between the observers regarding specific cervical vertebral morphology? Were certain morphologic features of the cervical vertebrae more consistently classified the same by different observers? Were certain morphologic features of the cervical vertebrae less consistently classified the same by different observers? Was it possible to assign a CVM stage to each subject based on the observers’ responses, or were there combinations of morphologic features that would make it impossible to stage the subject? In all cases when a CVM stage could be assigned, what was the level of agreement between the observers regarding the derived CVM stage for each subject (interobserver agreement)? Finally, how well did the orthodontists agree with their own previous stagings of the same subjects?

MATERIAL AND METHODS

The sample used in this study was the same sample used by Gabriel et al.,1 who randomly selected subjects with untreated longitudinal growth records from the Iowa Facial Growth Study. Thirty lateral cephalograms of good quality with complete visualization of cervical vertebrae C1 through C4 were selected for 15 white boys and 15 white girls. The lateral cephalograms were scanned at 600 dpi for placement into a presentation as high-resolution images in TIF format to maintain the original radiographic quality (Fig 1). The lateral cephalograms were cropped to include cervical vertebrae C1 to C4 and to eliminate any additional information such as stage of dentition that might bias the observer.

The observers in this study included the same 10 private-practice orthodontists who participated in the study by Gabriel et al. The orthodontists had between 7 and 40 years of clinical orthodontic experience at the time of our study (mean, 21.2 years of clinical experience). The observers did not participate in the design or construction of the research project. Each observer was trained in the CVM method and given definitions of CVM morphology to be used at any time during the...
study (Tables I and II, Fig 2). The observers were then shown a PowerPoint (Microsoft, Redmond, Wash) presentation containing lateral cephalograms of the 30 subjects previously described.

For each image, the following 5 questions regarding cervical vertebrae morphology (used to stage the cephalograms according to the CVM method) were asked:

1. Is the lower border of C2 best described as flat or curved?
2. Is the lower border of C3 best described as flat or curved?
3. Is the lower border of C4 best described as flat or curved?
4. Is the vertebral body of C3 best described as trapezoidal, rectangular horizontal, square, or rectangular vertical?
5. Is the vertebral body of C4 best described as trapezoidal, rectangular horizontal, square, or rectangular vertical?

Each observer was given a recording sheet and asked to answer the same 5 questions for each of the 30 subjects (a total of 150 responses for each of the 10 observers). The observers had unlimited time to make their evaluations.

Statistical analysis

The Fleiss kappa statistic was used to assess interobserver agreement when evaluating each of the 5 vertebral morphology questions for each subject. This measure calculates the degree of agreement over that which would be expected by chance. The Fleiss kappa varies between 0 (no agreement) and 1 (perfect agreement). Coefficients of 0.4 to 0.6 are generally considered to indicate moderate agreement. The Kendall coefficient of concordance (Kendall’s W) was used to assess interobserver agreement when determining a “derived CVM stage” for each subject. Kendall’s W varies between 0.0 (no agreement) and 1.0 (maximum agreement). Kendall’s W values of 0.4 to 0.6 are generally considered to indicate moderate agreement. The alpha level of each test was set at 0.05. Intraobserver agreement was evaluated by comparing the staging results made by Gabriel et al1 with those of the same authors on the same subjects in this study.

RESULTS

There were a total of 1350 interobserver evaluations for each of the 5 questions regarding cervical vertebrae morphology, and the number of disagreements between evaluators steadily increased from questions 1 through 5.
Table III. Fleiss kappa values for questions 1 through 5

<table>
<thead>
<tr>
<th>Question</th>
<th>Fleiss kappa value</th>
<th>Statistical level of agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.47</td>
<td>Moderate</td>
</tr>
<tr>
<td>2</td>
<td>0.65</td>
<td>Substantial</td>
</tr>
<tr>
<td>3</td>
<td>0.63</td>
<td>Substantial</td>
</tr>
<tr>
<td>4</td>
<td>0.39</td>
<td>Fair</td>
</tr>
<tr>
<td>5</td>
<td>0.34</td>
<td>Fair</td>
</tr>
</tbody>
</table>

The Fleiss kappa statistic of the reliability of interobserver agreement showed "moderate" to "substantial" agreement for questions 1 through 3, but questions 4 and 5 were rated as "fair" (Table III). In other words, questions dealing with the lower borders of C2 through C4 as flat or curved resulted in far better interobserver agreement than questions dealing with the vertebral bodies of C3 and C4 described as trapezoidal, rectangular horizontal, square, or rectangular vertical. This finding was further illustrated by the number of subjects judged unanimously by the orthodontists (Fig 3). For example, for question 1, there was unanimous agreement among the observers for 20 of the 30 subjects. For question 2, there was unanimous agreement for 17 of the 30 subjects. For question 3, there was unanimous agreement for 14 of the 30 subjects. However, only 7 of the 30 subjects were judged unanimously for question 4, and only 4 of the 30 subjects were judged unanimously for question 5.

In most cases, a final CVM stage could be derived from the 5 questions for a certain subject. However, for 88 of the 300 subject evaluations, a final CVM stage could not be determined because of incompatible responses to the 5 questions. For example, assume in a subject that the orthodontist defines the lower border of C2 as flat and at the same time describes the lower border of C3 as curved. This scenario is impossible under CVM guidelines, which require the curvature of C2 to develop before any curvature is evident in C3; this would result in an inconclusive CVM stage. On average, the orthodontists agreed in this study with their previous stagings only 62% of the time, when the inconclusive stagings were excluded as disagreements. When the inconclusive stagings were included, on average, the orthodontists agreed with their previous stagings only 44% of the time.

For the group of subjects who could be assigned a CVM stage, the level of interobserver agreement measured by the Kendall's W was only 0.45, indicating moderate agreement. When observers disagreed regarding a subject's CVM stage, they did so by a difference of as many as 3 stages, as illustrated in Figure 4.

DISCUSSION

The principal finding of this CVM method study was that interobserver agreement is high for assessing the lower borders of C2, C3, and C4 as either flat or curved, but interobserver agreement is low for assessing the vertebral bodies of C3 and C4 as either trapezoidal, rectangular horizontal, square, or rectangular vertical; this leads to the overall poor reproducibility of the CVM method. Reproducibility of the CVM method is poor, whether CVM staging is performed by using grouped criteria or is derived from individual morphologic criteria, as was done in this study.

When staging is derived from individual morphologic criteria, a corresponding CVM stage could not be derived for nearly 30% of the total sample; this suggests that CVM staging based on a visual analysis of the cervical
vertebrae is clinically unreliable. For subjects who could be staged, the Kendall's W for interobserver staging agreement was 0.45, indicating only moderate agreement among the judges. Several studies reported higher levels of interobserver agreement than those in our study. Franchi et al. found 96.6% agreement between observers. Similarly, Ozer et al. reported interobserver agreement levels of 98%. Uysal et al. showed interobserver reliability coefficients between 0.955 and 0.987, whereas Chang et al. found reliability coefficients of 0.85 and 0.90. The high levels of interobserver agreement reported in these studies might be overstated because of the use of third-person tracings to stage the lateral cephalograms. In addition, the authors of these studies also were the observers in the tests of interobserver agreement.

The decreased level of interobserver agreement for questions 4 and 5 might be because, for questions 1 through 3, there were only 2 answers from which to choose (flat or curved), whereas, for questions 4 and 5, there were 4 (trapezoid, rectangular horizontal, square, and rectangular vertical). Nevertheless, the significant difference between the levels of agreement for questions 4 and 5 compared with questions 1 through 3 suggests that not all variables assessed in the CVM method are equally reliable. In addition, Gabriel et al. found that the vertebrae of certain subjects appeared to be easier to stage than others. In this study, we concurred with this finding (Fig 4). For some subjects, the range of staging differed by as many as 3 cervical stages. Other subjects were staged identically by all 10 orthodontists. Hassel and Farman reported that skeletal maturation is a continuous process and that each maturation stage blends into the next. Therefore, it might be difficult to differentiate between borderline subjects, whereas other subjects can be readily staged. In addition, certain lateral cephalograms could have a greater level of radiographic “noise,” making staging difficult because of decreased image clarity.

Intraobserver agreement was made by comparing the CVM staging results from Gabriel et al. by using grouped morphologic criteria with the same orthodontists and

![Image: Fig 3. Interobserver agreement: A, question 1: is the lower border of C2 best described as flat or curved? B, question 2: is the lower border of C3 best described as flat or curved? C, question 3: is the lower border of C4 best described as flat or curved? D, question 4: is the vertebral body of C3 best described as trapezoidal, rectangular horizontal, square, or rectangular vertical? and E, question 5: is the vertebral body of C4 best described as trapezoidal, rectangular horizontal, square, or rectangular vertical? ]
images from our study by using individual morphologic criteria (based on responses to questions 1–5). Intraobserver agreement levels between these 2 times ranged from 50% to 88.2%. As was true for interobserver agreement, the levels of intraobserver agreement in this study were low compared with the levels reported in previous studies. Franchi et al reported 100% intraobserver agreement, Chang et al reported agreement of 92%, and Hassel and Farman reported intraobserver agreement of 95%. Kucukkeles et al were among the few who reported intraobserver agreement levels similar to those that we found in this study. Their agreement levels for 3 observers were 90%, 65%, and 45%, but no explanation was given for the poor levels of intraobserver agreement. Most of these studies involved the use of tracings for CVM staging. In addition, the authors of these studies also served as the observers in the tests of agreement, introducing a potential bias into their results.

However, even if a patient's skeletal age can be determined precisely, the value of this information could be limited by 2 factors: the facial growth spurt, if it occurs at all, might be only slightly greater than the patient's normal facial growth rate throughout adolescence, and peak mandibular growth might not directly relate to enhanced anterior positioning of the mandible in every patient. What do these findings mean to clinical orthodontic practice? The CVM method is currently being promoted and marketed as a measurement tool for timing orthopedic treatment. But, to be useful, a measurement must be reproducible: it must give you the same reading (or close to the same reading) every time you make the measurement. If it does not—and the CVM method does not—then your decision to begin treatment based on this measurement will depend on which time you make the measurement.

CONCLUSIONS

The weakness of the CVM method results, in part, from difficulty in classifying the vertebral bodies of C3 and C4 as trapezoidal, rectangular horizontal, square, or rectangular vertical. This leads to its overall poor reproducibility and our inability to support its use as a strict clinical guideline for the timing of orthodontic treatment.

REFERENCES


