Measurement of Thoracolumbar Kyphosis After Burst Fracture: Evaluation of Intraobserver, Interobserver, and Variability of 4 Measurement Methods

Jerome G. Enad, MD, Joseph B. Slakey, MD, and Patrick S. McNulty, MD

Abstract
There are various methods for measuring kyphosis after thoracolumbar burst fracture. The reliability and reproducibility of these methods are not well defined. In the study reported here, we examined 4 commonly used measurement methods in order to determine intraobserver variability, interobserver variability, and variability between measurement methods. All 4 methods were found to be accurate and reproducible when used by 4 observers on 2 occasions. One method, in comparison with the others, tended to overestimate degree of kyphosis. Understanding the methods for measuring kyphotic deformity after thoracolumbar burst fracture is essential in making decisions about prognosis and treatment.

Management of thoracolumbar burst fractures is influenced by both clinical and radiographic findings. Surgical treatment consisting of open reduction, fixation, and arthrodesis confers the advantage of deformity correction, early mobilization, and reduced reliance on orthotic bracing and is considered in the event of neurologic deficit or when the initial sagittal deformity is large or progresses during the course of nonoperative treatment. However, several studies have also demonstrated that severity of initial or residual kyphotic deformity does not correlate with symptoms at follow-up, thereby bringing into question the efficacy of radiographic indications for surgery. Additionally, the numerous methods described for measuring thoracolumbar kyphosis after burst fracture suggest that no single method is widely accepted. Kuklo and colleagues examined 5 different methods for measuring thoracolumbar fracture kyphosis and found 1 that significantly exhibited higher interobserver and intraobserver reliability. This method measured from the superior endplate of the vertebral body 1 level above the injured vertebra to the inferior endplate of the vertebral body 1 level below. In that study, however, the absolute values obtained with each individual method were not compared with the values obtained with the other methods. Thus, though a single method may produce the most reliable and repeatable result, it should also be known whether that result may be significantly different from that obtained with other methods or whether each method that is proved to be accurate yields similar measurement values. If decision making depends in part on the absolute or relative value of the measured kyphosis, it is important to recognize whether a single method, in comparison with other methods, consistently results in a significantly different measured value.

In the study reported here, we evaluated 4 commonly used methods in order to determine intraobserver variability, interobserver variability, and variability between measurement methods.

Materials and Methods
Sixteen lateral spine standard cassette x-rays of single-level thoracolumbar burst fractures (range, T10–L2) treated at our institution were selected from the radiology department files. X-rays were selected for clarity, adequate rotation, and centering of the vertebral level on the film. Four high-quality hard copies of the x-rays were made and numbered by an independent assistant who did not perform the measurements. The films were renumbered and measurements retaken 3 weeks after the first iteration.

The observers were instructed on the 4 methods most commonly used at our institution: (1) measuring from...
the superior endplate of the vertebral body 1 level above the injured vertebral body to the inferior endplate of the vertebral body 1 level below, (2) measuring from the superior endplate of the vertebral body 1 level above the injured vertebral body to the inferior endplate of the vertebral body 1 level below, (3) measuring the angle between the posterior vertebral body 1 level above the injured vertebral body and 1 level below the injured vertebral body, (4) measuring from the superior endplate 1 level above the injured vertebral body and the superior endplate 1 level below the injured vertebral body. Illustration by Jerome G. Enad, MD.

Figure. Four methods used for fracture kyphosis measurement: (1) measuring from the superior endplate of the vertebral body 1 level above the injured vertebral body to the inferior endplate of the vertebral body 1 level below, (2) measuring from the superior endplate of the vertebral body 1 level above the injured vertebral body to the inferior endplate of the injured vertebral body, (3) measuring the angle between the posterior vertebral body 1 level above the injured vertebral body and 1 level below the injured vertebral body, (4) measuring from the superior endplate 1 level above the injured vertebral body and the superior endplate 1 level below the injured vertebral body. Illustration by Jerome G. Enad, MD.

Statistical Analysis
Data were analyzed with a statistical software program, SPSS 7.5 for Windows (SPSS Inc, Chicago, Ill). Analyses of variance (ANOVAs) for repeated measures were performed to calculate the intraclass correlation coefficient, $\rho$ (rho), for estimating the intraobserver and interobserver reliability as described by Winer. The intraobserver reliability assessed the reproducibility of each observer for each measurement method. In this study, each observer measured the same x-ray twice for each method. The interobserver reliabilities were obtained to assess overall agreement among the 4 observers for all methods and for each method. For analyzing interobserver reliability, the first measurement of each observer was entered into the ANOVA. The ANOVA generalized linear model for repeated measures was performed to determine the effect of the different methods on the intraobserver and interobserver reliability. The Duncan multiple range test was used to compare the mean values obtained from each method to determine variability between methods. (The test con-

trols only for the type 1 comparison-wise error rate, not the experiment-wise error rate). Further, a post hoc power analysis for a sample size of 16 subjects (ie, x-rays) was performed to estimate the power of the difference in mean values obtained between methods.

RESULTS

Intraobserver Reliability
Reproducibility for each observer was high in comparison of each of the 4 methods (Table I). The intraclass coefficient ($\rho$) varied from .979 to .996 for observer 1, from .874 to .939 for observer 2, from .978 to .992 for observer 3, and from .987 to .991 for observer 4. The intraclass coefficients were most consistent for method 1 (mean $\rho$, .978), measuring from the superior endplate of the vertebral body 1 level above the injured vertebral body to the inferior endplate of the vertebral body 1 level below. This was followed by method 4 (mean $\rho$, .978), measuring from the superior endplate of the vertebral body 1 level above the injured vertebral body to the superior endplate of the vertebral body 1 level below. Next was method 3 (mean $\rho$, .959), measuring the angle between the posterior vertebral body 1 level above and below the injured vertebra. Method 2, measuring from the superior endplate of the vertebral body 1 level above the injured vertebral body to the inferior endplate of the injured vertebral body, produced the lowest intraclass coefficients overall (mean $\rho$, .954).

Interobserver Reliability
Paired comparisons between observers had low variability, according to $\rho$ intraclass correlation coefficients for each measurement method (Table II). Method 1 (mean $\rho$, .962) had the best interobserver reliability. Next were method 3 (mean $\rho$, .948) and method 4 (mean $\rho$, .945), followed by method 2 (mean $\rho$, .898). All methods resulted in excellent reliability ($\rho$>.80).

Measurement Differences Between Methods
According to the Duncan multiple range test for variability ($\alpha = .05$), mean values for each x-ray using methods 1, 3, and 4 were not significantly different from each other. The mean value for each x-ray using method 2 was significantly different from ($P<.001$) and higher in absolute value than that using method 1, 3, or 4 (mean difference, 8.0°-9.8°). A post hoc power analysis, performed to detect a large effect difference between values from method 2 and values from methods 1, 3, and 4, yielded 91% power at a .05 significance level.

DISCUSSION
The first goal of this study was to determine the best measurement method in terms of intraobserver reproducibility and interobserver reliability for thoracolumbar burst fracture. The second goal was to detect any discernible difference in values obtained from using each of the methods. Our data suggest that each of the 4 methods used to measure posttraumatic kyphosis in thoracolumbar burst fracture is reproduc-
Measure of Thoracolumbar Kyphosis After Burst Fracture

In a similar study, examining 5 methods for measuring fracture kyphosis, Kuklo and colleagues\textsuperscript{16} noted that 1 method had superior intraobserver and interobserver reliability. This method—measuring from the superior endplate of the vertebral body 1 level above the injured vertebral body to the inferior endplate of the vertebral body 1 level below—was identical to method 1 described in our study. In our study, we also found that this method had the most consistent intraobserver correlation coefficient and the best paired-comparison interobserver reliability of all methods examined.

Our study has some potential sources of variability. A standard radiographic method was not used to obtain images of the vertebral burst fracture. Of note, these injury films were each taken in the supine position and do not account for the load-bearing effects on the spine during upright posture. We did not examine fracture type based on the Denis classification of burst fractures.\textsuperscript{22} Quality of duplicated x-rays may have distorted interpretation of endplate and posterior body landmarks. Our selection criteria allowed us to minimize these technical concerns but left us with only 16 eligible x-rays. This sample size was calculated by our statistician to be sufficient to yield a power of more than 90% in determining reliability and reproducibility of each measurement method and the differences between methods.

**CONCLUSIONS**

Invariably, the amount of posttraumatic kyphotic deformity is only one factor in determining overall treatment of thoracolumbar burst fracture. If this criterion is to be considered, however, the method for measuring the angle of kyphosis must be reproducible, accurate, and well defined. Our study demonstrated that each of the 4 methods described is reliable and reproducible. Method 1, measuring from the superior endplate of the vertebral body 1 level above the injured vertebral body to the inferior endplate of the vertebral body 1 level below, is the most consistent in terms of intraobserver and interobserver reliability, and its continued use is recommended. Method 2, measuring from the superior endplate of the vertebral body 1 level above the injured vertebral body to the inferior endplate of the injured vertebral body, has the lowest intraobserver and interobserver reliability. Method 2 also produces consistently higher values in comparison with the other methods, and its continued use is discouraged.

Some authors have suggested that the initial amount of fracture kyphosis is related to prognosis,\textsuperscript{1-4,23} but there is no clear association between kyphotic deformity and reported pain at follow-up.\textsuperscript{5-9} However, accurately measuring the angle of kyphosis remains important both for documentation of clinical assessment and for examination of parameters related to treatment outcomes. Therefore, one should strive to use a measurement method that is reliable and reproducible.

**AUTHORS’ DISCLOSURE STATEMENT AND ACKNOWLEDGMENTS**

The authors report no actual or potential conflict of interest in relation to this article.

The views expressed in this article are those of the authors and do not reflect the official policy or position of the US Department of the Navy, the US Department of Defense, or the US government.

The authors acknowledge the contributions of Dr. Jeffrey Knapp and Dr. Anthony Lapinsky as participating investigators in data collection.

---

**Table I. Intraclass Correlation Coefficients (\(r\)) for Intraobserver Reliability for Each Observer Using Each of the 4 Measurement Methods**

<table>
<thead>
<tr>
<th>Observer</th>
<th>Method 1</th>
<th>Method 2</th>
<th>Method 3</th>
<th>Method 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.996</td>
<td>.989</td>
<td>.979</td>
<td>.993</td>
</tr>
<tr>
<td>2</td>
<td>.931</td>
<td>.874</td>
<td>.938</td>
<td>.939</td>
</tr>
<tr>
<td>3</td>
<td>.992</td>
<td>.978</td>
<td>.931</td>
<td>.985</td>
</tr>
<tr>
<td>4</td>
<td>.991</td>
<td>.976</td>
<td>.987</td>
<td>.979</td>
</tr>
</tbody>
</table>

**Table II. Paired Interobserver Reliability Calculated Using Intraclass Correlation Coefficients (\(r\)) Given a Specified Measurement Method**

<table>
<thead>
<tr>
<th>Observers</th>
<th>Method 1</th>
<th>Method 2</th>
<th>Method 3</th>
<th>Method 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 &amp; 2</td>
<td>.935</td>
<td>.815</td>
<td>.932</td>
<td>.929</td>
</tr>
<tr>
<td>1 &amp; 3</td>
<td>.995</td>
<td>.956</td>
<td>.973</td>
<td>.977</td>
</tr>
<tr>
<td>1 &amp; 4</td>
<td>.989</td>
<td>.979</td>
<td>.974</td>
<td>.962</td>
</tr>
<tr>
<td>2 &amp; 3</td>
<td>.939</td>
<td>.857</td>
<td>.948</td>
<td>.938</td>
</tr>
<tr>
<td>2 &amp; 4</td>
<td>.926</td>
<td>.817</td>
<td>.903</td>
<td>.901</td>
</tr>
<tr>
<td>3 &amp; 4</td>
<td>.991</td>
<td>.962</td>
<td>.957</td>
<td>.962</td>
</tr>
</tbody>
</table>
REFERENCES