COMPUTER-ASSISTED ALGORITHMS IMPROVE RELIABILITY OF KING CLASSIFICATION AND COBB ANGLE MEASUREMENT OF SCOLIOSIS

Ian A.F. Stokes, PhD and David D. Aronsson, MD

Department of Orthopaedics and Rehabilitation,
University of Vermont,

Correspondence: Ian A. Stokes
Department of Orthopaedics and Rehabilitation,
University of Vermont, Stafford Hall 434
Burlington, VT 05405-0084, USA.
Tel: +1 802 656 4249; Fax: +1 802 656 4247;
E-mail: Ian.Stokes@uvm.edu

Submitted to Spine, September 2004

Acknowledgements: This work was supported in part by NIH R01 AR 56432. The radiographic images were kindly provided to us by Dr L. Lenke. We acknowledge Randall Risinger, MD, Ketan Davae, MD, and Katherine Clark, BS who marked the radiographs. Richard Single, PhD advised on the statistical analysis of the data.
ABSTRACT

Study Design:
Inter- and intra-observer reliability study of improved method to evaluate radiographs of patients with scoliosis.

Objectives:
To determine the reliability of a computer-assisted measurement protocol for evaluating Cobb angle and King et al. classification.

Summary of Background Data:
Evaluation of scoliosis radiographs is inherently unreliable because of technical and human judgmental errors. Objective, computer-assisted evaluation tools may improve reliability.

Methods:
Posteroanterior pre-operative radiographic images of 27 patients with adolescent idiopathic scoliosis were each displayed on a computer screen. They were marked three times in random sequence by each of five evaluators (observers) who marked seventy standardized points on the vertebrae and sacrum in each radiograph. Coordinates of these points were automatically analyzed by a computer program (Spine, 2002; 27(24): 2801-2805) that identified curves, calculated Cobb angles, and generated the King et al. classification. The inter- and intra-observer variability of the Cobb angle and King et al. classification evaluations were quantified and compared with values obtained by unaided observers.

Results:
Average Cobb angle intra-observer standard deviation was 2.0 degrees for both the thoracic and lumbar curves (range 0.1 to 8.3 degrees for different curves). Inter-observer reliability was 2.5 degrees for thoracic curves and 2.6 degrees for lumbar curves. Among the five
observers there was an inverse relationship between repeatability and time spent marking images, and no correlation with image quality or curve magnitude. Kappa values for the variability of the King et al. classification averaged 0.85 (intra-observer).

Conclusions:
Variability of Cobb measurements compares favorably with previously published series. The classification was more reliable than achieved by unaided observers evaluating the same radiographs. The same principles may be applicable to other radiographic measurement and evaluation procedures.

Keywords: Scoliosis; Radiographs; Computer-assisted; Classification; Cobb-angle

Key Points:
1. Computerized algorithms can assist in the complex task of identifying end vertebrae of curves in radiographs of patients with scoliosis, measuring curve magnitudes, and in applying complex classification rules, potentially overcoming human judgmental errors.
2. A tool employing a computerized algorithm to facilitate Cobb angle measurements and King-Classifications was found to have reliability superior to that achieved by unaided individuals.
3. Computer-assisted evaluation was performed equally well by lesser-trained individuals, and the time spent marking the films (not an observer’s experience in managing patients with scoliosis) was the major determinant of accuracy and reliability.
Mini-Abstract:
A computer-assisted method to evaluate PA radiographs of scoliosis deformity permitted more reliable measurements of Cobb angle than reported for unaided traditional measurements, and better reliability in King Classification than unaided observers classifying the same radiographs. Apparently the standardized measurement reduced the technical errors, and the formal algorithms reduced the judgmental errors.
INTRODUCTION

The management of patients with idiopathic scoliosis relies heavily on radiographic measurements to identify coronal and sagittal curves, to detect progression of deformity, and to assist in the planning of conservative and surgical treatment. The Cobb angle has become the basis for quantifying scoliosis curve magnitude. Studies of inter- and intra-observer variability in measurement of this angle [1-6] have revealed that the errors in radiographic measurements are typically +/- 5 degrees, and this is comparable with thresholds of change that can influence treatment decisions [3]. The sources of the errors may include incorrect selection of the most tilted endplates, random errors in drawing lines across the endplates, and systematic errors due to inaccurately manufactured protractors [3].

Spinal curve pattern classifications that rely on radiographic measures are used in surgical planning for patients with adolescent idiopathic scoliosis, to select fusion levels [7]. The King et al. classification [8] is still the most widely used in surgical planning, although it was originally developed specifically for procedures employing Harrington instrumentation. It defines five thoracic scoliosis curve types, and an additional group called 'miscellaneous', based on measurements on standing radiographs, and can include measurements from lateral bending films. The King et al. classification relies on subjective identification and measurement of the radiographic features including the apical and end vertebrae of curves, vertebral endplate tilt angles, and the origin and alignment of the central sacral line. It also requires individual interpretation and memory of the classification criteria. Errors in identifying these radiographic landmarks and employing the resulting measurements in identifying the pattern of deformity provide numerous opportunities for both technical and
judgmental errors producing inter- and intra-observer variability [9]. Empirical studies [10-13] of repeat classification by the King et al. method have demonstrated problems with reliability.

A computer assisted algorithm [9] intended to minimize human involvement in Cobb angle measurement and in King et al. classification identified potential sources of classification errors. Specifically, it was reported that the classification was often unreliable when a radiographic measurement was close to a threshold value used to distinguish between two curve types. For example, when the apical vertebra of a lumbar curve was close to the central sacral line, repeated evaluations differed as to whether the curve crossed the midline, and consequently the classification was inconsistent. Digital radiography can facilitate software-assisted evaluation of radiographs to replace traditional 'pencil and ruler' measurement methods and evaluations that rely on subjective assessment and memory of classification criteria.

The purpose of this report was to determine the reliability of the previously published [9] computer-assisted protocol for evaluating Cobb angle and King et al. classification of radiographs of patients with idiopathic scoliosis who were candidates for surgery. Some possible influences on reliability, including the experience of individual using the computer-assisted tool, the magnitude of the scoliosis and the image quality were investigated.
METHODS

Posteroanterior radiographs taken pre-operatively of twenty-seven patients with adolescent idiopathic scoliosis that had been selected previously for studies of classification reliability [10,14] were used in the present study, thus permitting direct comparisons. The average Cobb angle in this series was reported as 64 degrees (range 45 to 105 degrees). For the present study, the radiographic images had been digitized and were displayed in a randomized sequence on a computer screen and marked three times by each of five observers. The radiographic images were provided as a 'PowerPoint' (Microsoft, Redwood WA, USA) file, from which grayscale image files were extracted. The image sizes were 925 pixels high by typically 475 pixels wide, i.e. each pixel was about 1 mm on the original spinal radiographs that were about 900 mm high. This format simulated digital radiography, although the pixel resolution was somewhat less. Seventy standardized radiographic landmarks were marked on each image using custom software and a computer 'mouse' to 'click' on selected points whose coordinates were then stored. The landmarks were the corners of the vertebral bodies (extremes of the endplate images) from T-1 to L-5 and two symmetrical landmarks on the sacrum (used to obtain the 'central sacral line') (Figure 1). The images were supplied with the anatomical level T-1 identified.

The five observers were a pediatric orthopaedic surgeon member of the Scoliosis Research Society, a non-clinician researcher having many years experience of marking spinal radiographs in a research context, an orthopaedic resident, a musculoskeletal radiologist, and a pre-medical student having no experience with spinal radiographs. The latter three evaluators had not used this computer-assisted tool prior to the present study. They learned how to use
it by having it demonstrated to them, emphasizing that they should identify and mark the corners of the vertebral body images and symmetrical points on the pelvis. This supervised instruction took about 30 minutes, including the time required to learn how to start the program and to identify locations of image files, etc. Subsequently the evaluators used the tool without supervision and without discussion of the findings. The time of completion of processing of each film was recorded by the computer, from which the time taken to process individual images was derived.

The stored coordinates were input to a published computer algorithm [9] that used derived vertebral positions and endplate tilt angles to identify scoliosis curves, their apices, end-vertebrae, and their Cobb angles, using a strict rule-based approach. The classification algorithm (Figure 2) implemented the King et al. classification using the published rules [8] incorporating the Cobb measurements and other criteria (positions of vertebrae, and endplate tilt angles). For Type 1 and Type 2 curves, the King et al. classification distinguishes between these classes based either on relative Cobb angle magnitudes or the ‘flexibility index’ obtained from lateral bending films. The criteria to distinguish these curve types as stated in King et al. [8] are ambiguous, requiring alternate algorithms in automated classification [9]. Here, the Cobb angle criterion was employed, not the flexibility index (Figure 2). The radiographs were assumed to be aligned with the vertical, so the central sacral line was considered to be parallel to the film edge, and passing through the midpoint of the two sacral landmarks.

The quality of each radiograph was evaluated subjectively by two of the observers who rated
each image by using a scale from 0 to 10, where a score of 10 would indicate that all
landmark points were easily identifiable, and 0 would indicate that none were visible.

Statistical Methods

The intra-observer repeatability of the Cobb angle measurements for each curve/observer
combination was calculated as the sample standard deviation of each observer’s three
measurements of each curve. Similarly, the inter-observer reliability for each curve/trial
combination was calculated as the sample standard deviation of the five observers’
measurements in each trial. These values were then averaged across the 27 patients. For the
purpose of statistical analyses of the Cobb angle data, each patient was assumed to have
provisionally two curves, that were analyzed separately. If two or more curves were present,
the main thoracic curve was designated as the upper curve, and the thoraco-lumbar or lumbar
curve was designated as the lower curve. If an upper thoracic curve was present it was
omitted from the Cobb angle reliability analyses.

The variability in the King et al. classification was assessed by the kappa statistic [16] that
measures the proportion of consistent classifications in two sets of observations, corrected for
the observed frequency of each class. This statistic was calculated for paired sets of
classifications by each observer (intra-observer repeatability) or between observers (inter-
observer reliability), using all combinations of paired observations. The resulting values were
averaged over combinations of pairs (between or within observers) to provide an overall
measure of inter- and intra-observer variability.
RESULTS

Intra-observer repeatability in Cobb angle evaluation.

Cobb angle variability between measurements of each individual patient was characterized by average sample standard deviations of 2.0 degrees for both the upper and lower curves (Table 1a). The greatest individual measurement error (difference from the overall mean for that curve) was 8.3 degrees. There was no significant correlation between Cobb angle repeatability for each patient and the image quality score ($R^2=0.17$ for upper curves; $R^2=0.02$ for lower curves). The image quality evaluation was judged to be reliable, based on a correlation with $R^2 = 0.62$ between the numerical scores assigned by the two independent observers.

Inter-observer reliability in Cobb angle evaluation.

The standard deviations of the samples of repeated observations averaged 2.5 degrees for upper curves, and 2.6 degrees for lower curves. There was no trend of the reliability increasing or decreasing over trials (Table 1b) and no evidence of systematic differences between observers.

Intra-observer repeatability in King et al. classifications:

Kappa values for the intra-observer repeatability of the King et al. classification averaged 0.85. The range was from 0.81 to 0.88 for the five observers (Table 2a). Among the five observers the Cobb angle and King classification repeatability did not correlate with experience in managing patients with scoliosis, but there was an inverse relationship between the rate of marking the radiographic images and the repeatability (Figure 3). The Spearman
rank correlation coefficient between the number of films marked per hour and the Cobb angle repeatability was 1.0 \( (p<0.001) \) and it was 0.82 \( (p<0.05) \) for the correlation between films marked per hour and the kappa statistic for classification repeatability.

Inter-observer reliability in King et al. classifications:

The overall inter-observer kappa values increased from 0.72 to 0.91 over the three series of measurement. The average inter-observer Kappa was 0.82. (Table 2b). Thirteen patients were consistently classified by all five observers in all three trials. Of the remaining 14 patients who were classified inconsistently, four different causes of the inconsistency were identified:

1. Inconsistent detection of an upper thoracic curve (King-Type assignment as variably 5 or 4; 3 or 4; or 3 or 5) - this occurred in a total of four patients in this study.

2. Inconsistent detection of the lumbar curve crossing the midline (King-Type either 2 or 3) - this occurred in two patients in this study.

3. Inconsistent identification of the thoraco-lumbar curve apex level (King-Type assigned variably as Miscellaneous or Type 1, or as Type 5) - this occurred in four patients in this study.

4. Inconsistent calculation of the relative Cobb angle magnitudes of the upper and lower curve (assigned either Type 1 or Type 2) - this occurred in four patients in this study.
DISCUSSION

Human observers can make technical and judgmental errors in evaluating spinal radiographs, thereby reducing accuracy and reliability. The findings of this study indicate that the task of identifying scoliosis curves in radiographs and subsequently classifying the curve type is more reliable with the assistance of a computerized tool. With use of this tool, standardized measurement procedures can reduce the technical errors, and formal objective algorithms can reduce the judgmental errors. Clinical experience was not found to be a factor in determining an individual observer’s reliability, instead the time spent in selecting the radiographic landmarks was found to be a significant factor. Relative to manual marking and classification, any additional time spent (potentially by a lesser trained person) could have worthwhile benefits in more accurate monitoring of curve progression and in treatment planning.

The reliability of the King-classification obtained in this study was superior to previously published series (Table 3). This was despite the fact that the classification was performed without pre-marking the radiographs, as in some previous studies [10,12]. Pre-marking has been identified as a significant factor in facilitating classification [11]. In the present study, the observers did not need to be coached or trained in the classification groupings, nor to memorize them, as they only had to identify and mark the vertebral body and sacral landmarks. These comparisons between series, including the direct comparison with the study of Lenke et al. [10] using the same radiographs used in this study but pre-marked with Cobb angles and the central sacral line, suggests that judgmental (rather than technical measurement errors) predominate in the King et al. classification. Further evidence for this comes from the
finding that the Cobb angle variability inherent in the computer-assisted method is comparable
with that reported previously.

It has been noted [9] that there is ambiguity as to the use of lateral bending measurements in
distinguishing between Type 1 and Type 2 curves in the King-Classification, and the
classification of these curve types is frequently performed based on the relative curve
magnitudes present in a standing radiograph, as in this study. The validity of the computer-
assisted algorithm was reported [9] by testing it against the examples given in King et al. [8].

The reproducibility of the Cobb angle measures obtained here appears equal to or better than
previously reported [1-6,17]. However, direct comparisons cannot be made with the previous
studies since different radiographs were evaluated, and differing statistical methods have been
used in those studies to evaluate Cobb angle reproducibility. Some published reports pre-
marked the end vertebrae, and some preselected good quality films, or those having smaller
curve magnitudes than in the pre-surgical group studied here. Oda et al. [1] reported that five
surgeons, measuring fifty radiographs had an average error of 9 degrees (calculated as twice
the standard deviation) and that the main error source was in identifying end vertebrae.
Morrissy et al. [3] reported repeated measurements by four surgeons of 48 ‘good quality’
 radiographs of patients having Cobb angle in the range 20 to 40 degrees. When the end
 vertebrae were not pre-selected, the standard deviation of paired differences was 2.4 degrees.
 Carman et al [4] reported an average difference of 3.8 degrees (95% of differences less than
 8.0 degrees) in repeated measurements by five readers of eight radiographs. They inferred
 from analysis of variance components that the overall standard deviation was 2.97 degrees.
These findings indicated that a change in a Cobb angle measurement of less than 10 degrees can not be interpreted with confidence as a real change. Goldberg et al. [5] reported interobserver variability of 2.5 degrees and intraobserver reliability of 1.9 degrees a study by four evaluators of the primary curve identified in thirty radiographs. They also reported that the interclass correlation coefficient for the Cobb angle was 0.98. Ylikoski et al. [6] studied 30 consecutive untreated patients having mean Cobb 24.4 degrees. Two readers used a specially designed angle measuring instrument (‘Plurimeter’) and found that the inter-observer standard deviation was 2.8 degrees and the intra-observer standard deviation was 1.8 degrees. In the present study of patients with larger (pre-operative) scoliosis the average sample standard deviations of the Cobb angle were (intra-observer) 2.0 degrees for upper and lower curves, and (inter-observer) 2.5 and 2.6 degrees for upper and lower curves respectively.

There is some disagreement as to whether the precision of Cobb angle measurements is substantially improved when the end vertebrae are preselected [1,3] or not [4,5]. In the present study the end vertebrae were selected automatically based on the values of endplate inclination calculated from the vertebral body landmarks. The accuracy of marking points on endplates was studied by Cheung et al. [18] who reported a coefficient of repeatability 0.8 mm and 1.3 mm in horizontal and vertical directions respectively, suggesting an angular error of about 2 degrees for the determination of each endplate inclination.

Here, the variability of the Cobb angle determination was not found to vary significantly with the radiographic quality. However, the image resolution (pixel size approximately one millimeter) was rather low, relative to original full-size films, and relative to that available in
digital radiographs.

Since a radiograph only records a patient’s spinal shape at an instant of time, repeated radiographs would introduce additional variability because of differing radiographic technique, postural sway, etc. [17]. For instance, Beauchamp et al. [19] reported diurnal variation of Cobb angle measurement. In the present, and most previous studies, the additional radiographic dose has precluded use of repeat radiographs, and this additional source of variability is ignored.

In the assignment of the King et al. classification, several factors have been noted previously as contributing to variability when a specific patient has a scoliosis deformity with features close to classification criteria [9]. These factors influenced the findings in the present study. They include the observed presence or absence of third (upper thoracic) curve, uncertainty as to whether a curve ‘crosses the midline’, and the relative magnitude of thoracic and lumbar curves. Problems arise when a patient has a spinal shape very close to any of these criteria. These kinds of problems could occur in alternate classifications. For example, the newer classification scheme developed by Lenke et al. has reliability characterized by kappa values in the range 0.64 to 0.89 [11,14,15], with lower values if the radiographs were not pre-marked (i.e. pre-measured). The ‘lumbar modifier’ used in this classification recognizes this possibility of a pedicle lying very close to the ‘cut-off’ point, and reliability in identifying this feature is relatively high (kappa statistic equal to 0.89 [15]). Nevertheless, factors such as variability in marking the central sacral line could still affect this judgement. The same kind of computer-assisted algorithmic approach as used in the present study could be applicable to
alternate classification systems with precisely-defined classification criteria, and might improve their reliability.

If the classification were taken as the sole factor in deciding the extent of a spinal arthrodesis for each patient, then the variation between observers and observations would alter the surgical plan. For instance, the difference between a Type 5 and either Type 2 or 3 classification (if the detection of an upper thoracic curve was inconsistent) would influence whether the upper curve were fused.

As digital imaging and computer-assisted medical decision-making become increasingly available, clinicians can increasingly turn to computerized tools to assist in analyzing, classifying and managing patients with adolescent idiopathic scoliosis. Computerized tools can be helpful in the automated interpretation of data, as well as its storage and display. For evaluation of radiographs of patients with scoliosis, it would be beneficial to replace the traditional pencil, ruler and protractor methods with interactive marking of landmark points and having the display software also include formal algorithms for measurement and classification. This can reduce technical errors, as well as the need for memorization of measurement and classification procedures.
Figure Captions:

Figure 1: Radiographic image showing marked points (four on each vertebral body, two on the sacrum) and the curve end-vertebrae and the Cobb angle values, as determined from the landmarks by the computerized algorithm. The curve pattern was automatically classified as Type 3.

Figure 2: Flowchart of the algorithm used in the computer-assisted classification. Reproduced from Stokes and Aronsson [9].

Figure 3: Intra-observer variability as a function of time spent by each observer marking the radiographs. Unfilled squares: the mean standard deviation for Cobb angle measurements. Filled triangles: the mean kappa statistic for King et al. classifications.
REFERENCES


[8] King HA, Moe JH, Bradford DS, Winter RB. The selection of fusion levels in


[16] Cohen J. A coefficient of agreement for nominal scales. Educational and


Table 1: Standard deviations of repeated measures of Cobb angles (degrees)

(a) Intra-observer repeatability

<table>
<thead>
<tr>
<th>Observer</th>
<th>Upper Curve</th>
<th>Lower Curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.1</td>
<td>2.1</td>
</tr>
<tr>
<td>2</td>
<td>1.8</td>
<td>1.8</td>
</tr>
<tr>
<td>3</td>
<td>1.9</td>
<td>1.7</td>
</tr>
<tr>
<td>4</td>
<td>2.3</td>
<td>2.4</td>
</tr>
<tr>
<td>5</td>
<td>2.0</td>
<td>2.2</td>
</tr>
<tr>
<td>Average</td>
<td>2.0</td>
<td>2.0</td>
</tr>
</tbody>
</table>

(b) Inter-observer reliability

<table>
<thead>
<tr>
<th>Trial</th>
<th>Upper Curve</th>
<th>Lower Curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.6</td>
<td>2.6</td>
</tr>
<tr>
<td>2</td>
<td>2.3</td>
<td>2.4</td>
</tr>
<tr>
<td>3</td>
<td>2.7</td>
<td>2.6</td>
</tr>
<tr>
<td>Average</td>
<td>2.5</td>
<td>2.6</td>
</tr>
</tbody>
</table>
Table 2. Kappa values for the King et al. Classifications

(a) Intra-observer repeatability:

<table>
<thead>
<tr>
<th>Observer</th>
<th>Classification Consistency (percent)</th>
<th>Kappa value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>88</td>
<td>0.84</td>
</tr>
<tr>
<td>2</td>
<td>90</td>
<td>0.87</td>
</tr>
<tr>
<td>3</td>
<td>90</td>
<td>0.88</td>
</tr>
<tr>
<td>4</td>
<td>85</td>
<td>0.81</td>
</tr>
<tr>
<td>5</td>
<td>90</td>
<td>0.87</td>
</tr>
<tr>
<td>Average</td>
<td>89</td>
<td>0.85</td>
</tr>
</tbody>
</table>

(b) Inter-observer reliability:

<table>
<thead>
<tr>
<th>Trial</th>
<th>Classification Consistency (percent)</th>
<th>Kappa value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>79</td>
<td>0.72</td>
</tr>
<tr>
<td>2</td>
<td>87</td>
<td>0.84</td>
</tr>
<tr>
<td>3</td>
<td>93</td>
<td>0.91</td>
</tr>
<tr>
<td>Average</td>
<td>86</td>
<td>0.82</td>
</tr>
</tbody>
</table>
Table 3. Mean Kappa Values obtained for King et al. Classification reliability in published studies, and in the present study.

<table>
<thead>
<tr>
<th></th>
<th>Inter-Observer</th>
<th>Intra-Observer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lenke et al. [10]</td>
<td>0.49</td>
<td>0.62</td>
</tr>
<tr>
<td>Cummings et al. [13]</td>
<td>0.44</td>
<td>0.64</td>
</tr>
<tr>
<td>Richards et al. [15]</td>
<td>0.61</td>
<td>0.81</td>
</tr>
<tr>
<td>Behensky et al. [12]</td>
<td>0.46</td>
<td>0.79</td>
</tr>
<tr>
<td>Present Study</td>
<td>0.82</td>
<td>0.85</td>
</tr>
</tbody>
</table>
Figure 1: Radiographic image showing marked points (four on each vertebral body, two on the sacrum) and the curve end-vertebrae and the Cobb angle values, as determined from the landmarks by the computerized algorithm. The curve pattern was automatically classified as Type 3.
Figure 2: Flowchart of the algorithm used in the computer-assisted classification. Reproduced from Stokes and Aronsson [9].
Figure 3: Intra-observer variability as a function of time spent by each observer marking the radiographs. Unfilled squares: the mean standard deviation for Cobb angle measurements. Filled triangles: the mean kappa statistic for King et al. classifications.