



AI-BASED COBB ANGLE MEASUREMENT MATCHES CLINICIANS IN SUSPECTED ADOLESCENT SCOLIOSIS

Şafak Sayar¹, Onur Aksoy²

¹Biruni University Faculty of Medicine, Department of Orthopedics and Traumatology, İstanbul, Türkiye

²Biruni University Faculty of Medicine, Department of Radiology, İstanbul, Türkiye

ABSTRACT

Objective: The Cobb angle is the reference standard for quantifying curve magnitude in adolescent idiopathic scoliosis, but manual measurements are subject to interobserver variability. Artificial intelligence (AI)-based systems are entering radiological practice, yet independent validation in adolescents remains limited. To compare Cobb angle measurements of an AI-based system, developed on adult radiographs, with those of an orthopedic surgeon and a radiologist.

Materials and Methods: Posteroanterior whole-spine radiographs of 130 consecutive adolescents (63 females and 67 males; mean age, 13.8 years) referred for scoliosis assessment were analyzed retrospectively. The Cobb angle of the primary curve was measured independently and in a blinded manner by an AI system (hScoliosis v1.0, Hevi AI, İstanbul, Türkiye), an orthopedic surgeon, and a radiologist. Agreement was assessed using the intraclass correlation coefficient (ICC), Bland-Altman analysis, Cohen's kappa (κ).

Results: Mean Cobb angles were $14.98^\circ \pm 9.18^\circ$ (orthopedic surgeon), $11.91^\circ \pm 9.29^\circ$ (radiologist) and $14.64^\circ \pm 9.75^\circ$ (AI). Agreement between the AI system and the orthopedic surgeon was excellent (ICC 0.954) and similar in magnitude to the agreement between the two clinicians (ICC 0.877). The mean difference was -0.33° (limits of agreement -5.93° to $+5.26^\circ$), and the absolute difference was 5° or less in 93.8% of patients, never exceeding 10° . Agreement at the 25° brace threshold was almost perfect ($\kappa=0.849$). The radiologist's measurements were systematically lower than those of the other observers.

Conclusion: AI-based Cobb angle measurement agreed with an experienced orthopedic surgeon at least as closely as two clinicians agreed with each other, despite the algorithm having been developed on adult radiographs, and it may serve as a reliable adjunct in adolescents.

Keywords: Scoliosis, artificial intelligence, deep learning, observer variation, reproducibility of results

INTRODUCTION

Adolescent idiopathic scoliosis (AIS) is the most common form of spinal deformity in the pediatric population, affecting approximately 2-3% of adolescents^(1,2). The Cobb angle, measured on standing posteroanterior spinal radiographs, remains the reference standard for quantifying curve magnitude and constitutes the principal parameter guiding clinical decision-making in AIS^(3,4). Established management thresholds are defined by this measurement: a Cobb angle of 10° is required for the diagnosis of scoliosis, curves of 25° or greater are generally considered for brace treatment, and curves approaching $45-50^\circ$ are evaluated for surgical intervention⁽⁵⁾. Given these thresholds, the accuracy and reproducibility of Cobb angle measurement directly influence the clinical trajectory of the individual patient.

Despite its central role, manual Cobb angle measurement is subject to well-documented intra- and inter-observer variability^(6,7). The measurement requires subjective identification of the end vertebrae-the vertebrae most tilted from the horizontal at the superior and inferior limits of the curve-as well as manual construction of angular reference lines. Reported inter-observer variability for manual measurement ranges from 3° to 10° ⁽⁷⁻⁹⁾, and a difference of at least 5° is conventionally required before a genuine change in curve magnitude can be inferred^(6,9). This inherent variability, together with the time required for manual measurement in high-volume clinical settings, has motivated interest in automated approaches.

Recent advances in artificial intelligence (AI) and deep learning have enabled the development of software capable of automatically detecting vertebral landmarks and computing

Address for Correspondence: Şafak Sayar, Biruni University Faculty of Medicine, Department of Orthopedics and Traumatology, İstanbul, Türkiye

E-mail: safaksayar@gmail.com

ORCID ID: orcid.org/0000-0003-1293-4436

Received: 21.07.2026 **Accepted:** 25.08.2026 **Epub:** 23.10.2026

Cite this article as: Sayar Ş, Aksoy O. AI-based Cobb angle measurement matches clinicians in suspected adolescent scoliosis. J Turk Spinal Surg. [Epub Ahead of Print]



spinal parameters from radiographs⁽¹⁰⁻¹²⁾. Such systems offer the potential for rapid, consistent, and observer-independent measurement, which may reduce workload and standardize reporting. However, before automated measurements can be incorporated into routine clinical practice, their agreement with measurements obtained by experienced clinicians must be rigorously validated. Prior validation studies have frequently been limited by restriction to radiographically confirmed scoliosis cohorts, comparison against a single human observer, or evaluation of consistency rather than absolute agreement⁽¹³⁻¹⁵⁾. The present study was designed to evaluate the agreement and reliability of an AI-based spinal measurement system for Cobb angle measurement in a consecutive cohort of adolescents referred for scoliosis assessment. The system evaluated had been developed on an exclusively adult dataset and had not previously been assessed in a skeletally immature population. AI-derived measurements were compared against those obtained independently by an orthopedic surgeon and a radiologist, thereby permitting assessment of both AI-human agreement and the underlying inter-observer variability between the two human disciplines. It was hypothesized that AI-based measurement would demonstrate agreement with the orthopedic surgeon comparable to, or exceeding, the agreement observed between the two human observers.

MATERIALS AND METHODS

Study Design and Ethical Approval

This single-center retrospective study was approved by the Biruni University Clinical Research Ethics Committee (approval no: 2024-BİAEK/19-48, date: 30.03.2026). All procedures were conducted in accordance with the principles of the Declaration of Helsinki. Because the study was based on the retrospective analysis of anonymized, previously acquired radiographs, the requirement for informed consent was waived. The study was reported in accordance with the strengthening the reporting of observational studies in epidemiology (STROBE) statement and, for the reliability-specific components, with the guidelines for reporting reliability and agreement studies (GRRAS)^(16,17).

Study Population

The radiographic archive of the department of orthopedics and traumatology was retrospectively reviewed for adolescents referred for scoliosis assessment. Consecutive patients who underwent standing posteroanterior whole-spine radiography were considered for inclusion. The inclusion criteria were an age between 10 and 18 years and the availability of a standing posteroanterior whole-spine radiograph of adequate diagnostic quality. The exclusion criteria were a history of previous spinal surgery, the presence of spinal instrumentation, congenital vertebral anomaly, neuromuscular scoliosis, and radiographs of insufficient quality for reliable measurement. A total of 130 patients met the eligibility criteria and were included in the final analysis.

It should be noted that the cohort was defined by referral for scoliosis assessment rather than by a confirmed diagnosis of scoliosis. Consequently, patients with a major curve below the 10° diagnostic threshold were retained in the cohort, reflecting the spectrum of patients encountered in routine screening practice and avoiding spectrum bias.

Radiographic Measurement

For each radiograph, the Cobb angle of the primary (major) curve was measured independently by three observers: an AI-based spinal measurement system, an orthopedic surgeon with 10 years of clinical experience, and a radiologist with 15 years of clinical experience.

The two human observers performed manual measurements on the Synapse picture archiving and communication system (PACS) workstation (Fujifilm, Tokyo, Japan) using the standard Cobb method. In this method, the end vertebrae were identified as the vertebrae most tilted from the horizontal at the superior and inferior limits of the curve, and the angle was constructed between lines drawn along the upper endplate of the superior end vertebra and the lower endplate of the inferior end vertebra. In patients with double curves, the primary curve was measured by all three observers. The AI system likewise designates a primary curve in its output, and this value was used for all analyses; secondary curves were not evaluated. Each observer was blinded to the measurements of the other observers and to the output of the AI system. To minimize systematic ordering effects, radiographs were presented in a randomized sequence for each observer. Each radiograph was measured once by each human observer; intra-observer reliability was not assessed.

AI-based measurements were obtained using hScoliosis version 1.0 (Hevi AI, İstanbul, Türkiye), a deep learning application for the automated measurement of spinal parameters on full-spine radiographs. At the time of the study the software was a research-mode prototype; Conformité Européenne marking under the European Union Medical Device Regulation was pending and the software was not commercially available for clinical use. According to the technical documentation provided by the manufacturer, the system performs automated anatomical landmark (keypoint) detection on full-spine radiographs and derives coronal and sagittal spinopelvic parameters from the detected landmarks. Measurement was fully automated; manual selection of the end vertebrae and adjustment of the measurement lines were neither possible nor performed at any stage. The same software version was used throughout the study period, and no updates were applied during data collection. For all analyses, the absolute value of the Cobb angle was used. A representative example of the automated output is shown in Figure 1.

Development and internal validation of the algorithm were performed by the manufacturer and are summarized here on the basis of the technical documentation provided. The model was developed using radiographs of adult patients acquired at more than 15 Turkish and international centers on EOS (EOS

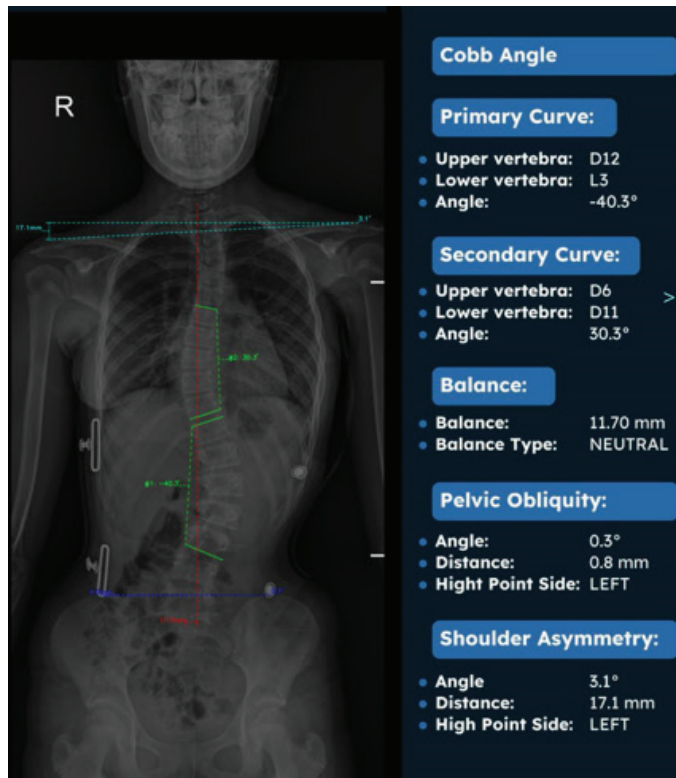


Figure 1. Representative output of the artificial intelligence (AI)-based measurement system (hScoliosis v1.0, Hevi AI, İstanbul, Türkiye) on a standing posteroanterior whole-spine radiograph of an adolescent female patient. The software automatically identifies vertebral landmarks and reports the major and secondary curve Cobb angles together with coronal balance, pelvic obliquity and shoulder asymmetry. In this patient the major curve Cobb angle was 40.3° by the AI system, 38.0° by the orthopedic surgeon and 39.0° by the radiologist. The difference of 2.3° between the AI system and the orthopedic surgeon corresponds closely to the mean absolute difference of 2.25° observed across the whole cohort; the case was therefore selected as representative of typical rather than optimal performance. Only the major curve Cobb angle was analyzed in the present study. All patient identifying information has been removed from the image

Imaging, Paris, France), stitched, and long-cassette radiographic systems, with a development dataset comprising 4,685 coronal and 4,204 sagittal examinations. Performance was assessed on held-out test sets of 200 coronal and 200 sagittal examinations excluded from model development. In this internal validation, keypoint detection on coronal radiographs achieved a mean F1 score of 0.94 (precision 0.94, recall 0.93) with a mean average precision of 0.95, and on sagittal radiographs a mean F1 score of 0.93 (precision 0.92, recall 0.94) with a mean average precision of 0.95. The mean absolute error of the derived coronal measurements was 2.8° for the Cobb angle, 1.8 mm for coronal balance, 0.2° for pelvic obliquity and 0.8° for shoulder asymmetry; for the sagittal measurements it was 1.8 mm for sagittal balance, 3.7° for thoracic kyphosis, 4.0° for lumbar lordosis, 4.1° for pelvic incidence, 1.5° for pelvic tilt and 3.1° for sacral slope. Because the development dataset consisted exclusively of adult examinations, the present study constitutes

an evaluation of the algorithm in a population outside its development domain.

Statistical Analysis

Statistical analysis was performed using IBM SPSS Statistics version 27 (IBM Corp., Armonk, NY, USA). Continuous variables were summarized as mean $\pm$ standard deviation, together with the median and range. Categorical variables were expressed as frequencies and percentages.

Agreement between measurement methods was assessed using the intraclass correlation coefficient (ICC), calculated using a two-way mixed-effects model with absolute agreement, and single-measures ICC values are reported. Pairwise agreement was assessed for the AI system versus the orthopedic surgeon, the AI system versus the radiologist, and the orthopedic surgeon versus the radiologist. ICC values were interpreted according to the criteria of Koo and Li⁽¹⁸⁾, in which values below 0.50 indicate poor reliability, values between 0.50 and 0.75 moderate reliability, values between 0.75 and 0.90 good reliability, and values above 0.90 excellent reliability.

Agreement was further evaluated using Bland-Altman analysis, with calculation of the mean difference (bias) and the 95% limits of agreement (defined as the mean difference $\pm$ 1.96 standard deviations)⁽¹⁹⁾. A one-sample t-test was used to assess whether the mean difference between each pair of methods differed significantly from zero. The presence of proportional bias was examined by regressing the difference between paired measurements on their mean. In addition, the absolute difference between paired measurements was categorized as 5° or less, greater than 5° up to 10°, and greater than 10°, in order to express agreement in clinically interpretable terms. The Pearson correlation coefficient was calculated to assess the linear association between measurement methods.

No a priori sample size calculation was performed; all consecutive patients meeting the eligibility criteria within the study period were included. A post-hoc calculation indicated that, for two raters at a two-sided alpha of 0.05, 30 subjects would be required to reject a null hypothesis ICC of 0.70 in favor of an alternative value of 0.90 with 90% power; the achieved power with 130 subjects therefore exceeded 99%.

Categorical agreement was evaluated at the clinically relevant thresholds of 10° (diagnosis of scoliosis) and 25° (consideration of brace treatment) using Cohen's kappa coefficient (κ). A p-value below 0.05 was considered statistically significant.

RESULTS

Patient Characteristics

A total of 130 adolescents were included in the analysis. Of these, 63 (48.5%) were female and 67 (51.5%) were male. The mean age was 13.83 $\pm$ 2.37 years (range, 10-18 years). According to the measurements of the orthopedic surgeon, the major curve was below 25° in 115 patients (88.5%), between 25° and 40° in 12 patients (9.2%), and above 40° in 3 patients (2.3%).

A major curve of 10° or greater, corresponding to the radiographic threshold for the diagnosis of scoliosis, was present in 90 patients (69.2%), whereas 40 patients (30.8%) had a curve below this threshold (Table 1).

Cobb Angle Measurements

The mean Cobb angle was 14.98±9.18° (range, 4.0-51.0°) for the orthopedic surgeon, 11.91±9.29° (range, 1.0-45.0°) for the radiologist, and 14.64±9.75° (range, 3.2-53.9°) for the AI system (Table 1). Pearson correlation coefficients were 0.956 between the AI system and the orthopedic surgeon, 0.948 between the AI system and the radiologist and 0.926 between the two human observers ($p<0.001$ for all).

Reliability

The AI system demonstrated excellent agreement with the orthopedic surgeon, with a single-measures ICC of 0.954 [95%

confidence interval (CI), 0.936-0.967]. Agreement between the AI system and the radiologist was also excellent (ICC=0.910; 95% CI, 0.615-0.964), whereas agreement between the two human observers was good (ICC=0.877; 95% CI, 0.542-0.949). For the three observers analyzed together the single-measures ICC was 0.914 (95% CI, 0.811-0.954) (Table 2).

Bland-Altman Analysis

The mean difference between the AI system and the orthopedic surgeon was -0.33±2.85°, with 95% limits of agreement ranging from -5.93° to +5.26°. This difference did not differ significantly from zero ($t=-1.331$, $p=0.186$). In contrast, the mean difference between the AI system and the radiologist was +2.74±3.10° (95% limits of agreement, -3.35° to +8.82°), and that between the orthopedic surgeon and the radiologist was +3.07±3.56° (95% limits of agreement, -3.92° to +10.05°); both differences were statistically significant ($p<0.001$ for both) (Table 2, Figure 2A-C).

Regression of the paired differences on their corresponding means revealed a small but statistically significant proportional bias between the AI system and the orthopedic surgeon ($B=0.062$; 95% CI, 0.010-0.115; $\beta=0.205$; $p=0.019$). No significant proportional bias was detected between the AI system and the radiologist ($B=0.050$; $\beta=0.153$; $p=0.082$) or between the two human observers ($B=-0.012$; $\beta=-0.031$; $p=0.723$).

Magnitude of Measurement Differences

The mean absolute difference was 2.25±1.77° between the AI system and the orthopedic surgeon, 3.20±2.62° between the AI system and the radiologist and 3.75±2.84° between the two human observers, with corresponding maxima of 8.50°, 10.60° and 13.00°.

The absolute difference between the AI system and the orthopedic surgeon was 5° or less in 122 patients (93.8%) and between 5° and 10° in the remaining 8 (6.2%); no patient exhibited a difference exceeding 10°. For the AI system versus the radiologist the corresponding figures were 104 (80.0%), 24 (18.5%) and 2 (1.5%) patients (Table 2).

Table 1. Patient characteristics and Cobb angle measurements (n=130)

Variable	Value
Age, years	13.83±2.37 (10-18)
Sex, female	63 (48.5)
Sex, male	67 (51.5)
Curve magnitude†	
<25°	115 (88.5)
25-40°	12 (9.2)
>40°	3 (2.3)
Cobb angle ≥10°†	90 (69.2)
Cobb angle <10°†	40 (30.8)
Cobb angle, degrees	
Orthopedic surgeon	14.98±9.18; 13.0 (4.0-51.0)
Radiologist	11.91±9.29; 9.0 (1.0-45.0)
AI system	14.64±9.75; 11.4 (3.2-53.9)

Values are mean ± standard deviation (range), n (%), or mean ± standard deviation; median (range). †: According to the measurements of the orthopedic surgeon, AI: Artificial intelligence

Table 2. Reliability and agreement between measurement methods (n=130)

Comparison	ICC (95% CI)	Mean difference ± SD, °	95% LOA, °	Proportional bias, B (p)	Absolute difference, mean ± SD (max), °	≤5°, n (%)
AI vs. orthopedic surgeon	0.954 (0.936-0.967)	-0.33±2.85	-5.93 to +5.26	0.062 (0.019)	2.25±1.77 (8.50)	122 (93.8)
AI vs. radiologist	0.910 (0.615-0.964)	+2.74±3.10	-3.35 to +8.82	0.050 (0.082)	3.20±2.62 (10.60)	104 (80.0)
Orthopedic surgeon vs. radiologist	0.877 (0.542-0.949)	+3.07±3.56	-3.92 to +10.05	-0.012 (0.723)	3.75±2.84 (13.00)	101 (77.7)
All three observers	0.914 (0.811-0.954)	-	-	-	-	-

Proportional bias is the unstandardized regression coefficient of the paired difference on the paired mean. The mean difference did not differ from zero for AI versus the orthopedic surgeon ($p=0.186$) and differed from zero for the remaining two comparisons ($p<0.001$ for both). Pearson correlation coefficients were 0.956, 0.948 and 0.926, respectively ($p<0.001$ for all). ICC: Intraclass correlation coefficient (single measures, two-way mixed-effects model, absolute agreement), LOA: Limits of agreement, SD: Standard deviation, AI: Artificial intelligence, CI: Confidence interval

Categorical Agreement at Clinical Thresholds

At the 25° brace-treatment threshold the orthopedic surgeon and the AI system agreed in 126 of 130 cases ($\kappa=0.849$). Taking the orthopedic surgeon as the reference, the AI system identified 13 of 15 patients with a curve of 25° or greater (sensitivity, 86.7%) and 113 of 115 below this threshold (specificity, 98.3%). Agreement at the same threshold was $\kappa=0.884$ both between the AI system and the radiologist and between the two human observers (Table 3).

At the 10° diagnostic threshold, agreement between the AI system and the orthopedic surgeon was moderate ($\kappa=0.609$), with 74 of 90 patients classified as having scoliosis by both methods (sensitivity, 82.2%) and 33 of 40 patients classified as not having scoliosis by both (specificity, 82.5%). Agreement between the AI system and the radiologist at this threshold was also moderate ($\kappa=0.609$), whereas agreement between the two human observers was the lowest of the three comparisons ($\kappa=0.480$), with discordant classification in 35 of 130 patients (26.9%) (Table 3).

Subgroup Analysis: Radiographically Confirmed Scoliosis

In the subgroup of 90 patients with a major curve of 10° or greater according to the orthopedic surgeon, the mean Cobb angle was $18.48 \pm 8.98^\circ$ for the orthopedic surgeon, $14.53 \pm 9.98^\circ$ for the radiologist, and $17.76 \pm 10.18^\circ$ for the AI system. Agreement between the AI system and the orthopedic surgeon remained excellent (ICC=0.949; 95% CI, 0.922-0.966), as did agreement between the AI system and the radiologist (ICC=0.901; 95% CI, 0.498-0.964). The mean difference between the AI system and the orthopedic surgeon was $-0.71 \pm 3.01^\circ$ ($p=0.027$), and the mean absolute difference was $2.48 \pm 1.84^\circ$, with a maximum of 8.50° . The mean difference between the AI system and the radiologist was $+3.23 \pm 3.29^\circ$ ($p<0.001$).

DISCUSSION

In the present study, AI-based Cobb angle measurement demonstrated excellent agreement with an experienced orthopedic surgeon, with a single-measures ICC of 0.954 and a mean absolute difference of 2.25° . The mean difference did

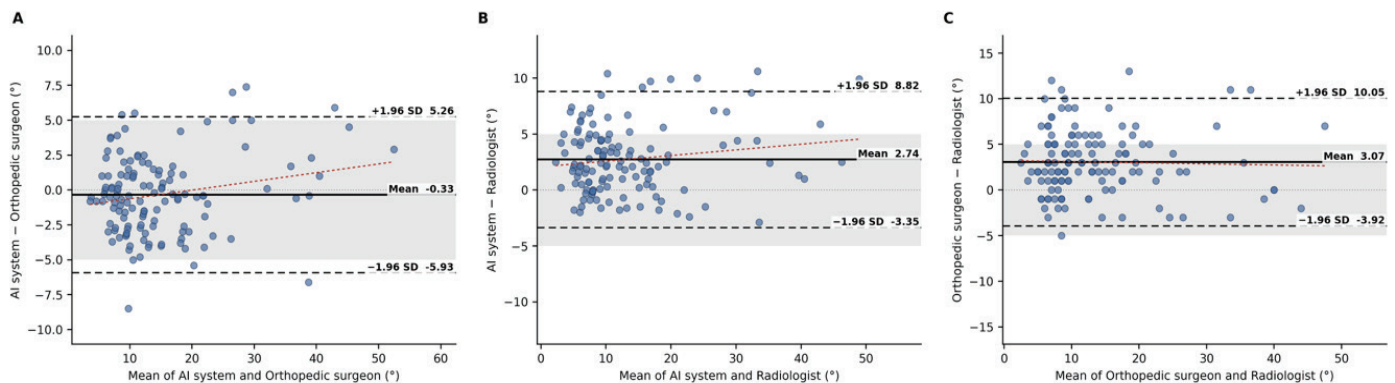


Figure 2. Bland-Altman plots of agreement between measurement methods. (A) AI system versus orthopedic surgeon; (B) AI system versus radiologist; (C) orthopedic surgeon versus radiologist. In each panel the solid line indicates the mean difference and the dashed lines the 95% limits of agreement; the shaded band denotes the $\pm 5^\circ$ range conventionally regarded as clinically acceptable, and the dotted line represents the regression of the difference on the mean. Mean differences and limits of agreement were -0.33° (-5.93° to $+5.26^\circ$) in panel A, $+2.74^\circ$ (-3.35° to $+8.82^\circ$) in panel B and $+3.07^\circ$ (-3.92° to $+10.05^\circ$) in panel C. Proportional bias was statistically significant only in panel A ($B=0.062$; $p=0.019$). AI: Artificial intelligence

Table 3. Categorical agreement at clinical decision thresholds (n=130)

Threshold/comparison	Concordant cases, n (%)	κ	Sensitivity, %	Specificity, %
25° (brace treatment)				
AI vs. orthopedic surgeon	126 (96.9)	0.849	86.7	98.3
AI vs. radiologist	127 (97.7)	0.884	-	-
Orthopedic surgeon vs. radiologist	127 (97.7)	0.884	-	-
10° (diagnosis of scoliosis)				
AI vs. orthopedic surgeon	107 (82.3)	0.609	82.2	82.5
AI vs. radiologist	104 (80.0)	0.609	-	-
Orthopedic surgeon vs. radiologist	95 (73.1)	0.480	-	-

Sensitivity and specificity are reported with the measurement of the orthopedic surgeon taken as the clinical reference; they are not reported for the remaining comparisons, in which no observer was designated as a reference standard. AI: Artificial intelligence, κ : Cohen's kappa coefficient ($p<0.001$ for all)

not differ significantly from zero, and the absolute difference was 5° or less in 93.8% of patients, with no case exceeding 10° . The point estimate for AI-surgeon agreement exceeded that between the two human observers ($ICC=0.877$), but the CIs overlapped and no formal comparison of ICC values was undertaken; the two should therefore be regarded as being of a similar order rather than as demonstrating superiority of automated measurement.

The accuracy observed here is consistent with that reported for other deep learning-based Cobb angle algorithms. One study reported a mean absolute error of $1.97 \pm 1.57^\circ$ and an ICC of 0.981, with 95.9% of angles within 5° of the reference⁽¹²⁾; in a multi-expert reader study the mean deviation of an automated algorithm was 4.17° , smaller than the mean intra-reader discrepancy of 5.16° ⁽²⁰⁾. The values obtained here fall within this range. An important distinction is that many published algorithms have been evaluated on internal validation subsets drawn from the development dataset, whereas the present evaluation was conducted independently, at a center that did not contribute to that dataset. The manufacturer's internal validation, on a held-out set of 200 adult coronal examinations, reported a mean absolute error of 2.8° ; the 2.25° observed here against the orthopedic surgeon is of comparable magnitude, indicating that development performance was reproduced in an independent adolescent cohort.

A further consideration specific to the system evaluated here is that its development dataset consisted exclusively of adult examinations, whereas the present cohort comprised adolescents between 10 and 18 years of age. Vertebral morphology, endplate ossification and overall spinal proportions differ between skeletally immature and mature patients, and an algorithm trained on adult radiographs cannot be assumed a priori to transfer to a pediatric population. The excellent agreement observed in the present study therefore provides evidence that the landmark detection approach generalizes across this age boundary, at least for the coronal major-curve Cobb angle. This observation should not, however, be extrapolated to the sagittal and pelvic parameters produced by the same system, which were not evaluated here and for which the manufacturer's internal error estimates were larger.

The disagreement observed between the two human observers is also consistent with the historical literature. In an early analysis of fifty radiographs measured on six occasions by four orthopedic surgeons, the 95% confidence limit for interobserver variability was 7.2° when observers selected the end vertebrae themselves and 6.3° when these were pre-specified, while intraobserver variability fell from 4.9° to 2.8° under increasingly standardized conditions⁽⁶⁾. Subjective selection of the end vertebrae is therefore a dominant source of measurement error, which plausibly explains the present findings: an automated system applies a fixed, deterministic rule for landmark identification and eliminates one of the principal sources of manual variability. The interobserver difference reported

here nonetheless derives from a single pair of clinicians and represents one realization of clinician variability rather than an estimate of its range; multi-rater studies have reported variability of comparable or greater magnitude^(6,7).

A systematic difference was observed between the radiologist and both other observers, the radiologist recording consistently lower values (mean difference 3.07° from the orthopedic surgeon and 2.74° from the AI system). Regression of the paired differences on their means showed no proportional bias in either comparison, indicating a constant offset across the angular range rather than an error increasing with curve magnitude. Combined with the high linear correlation between the two clinicians ($r=0.926$), this pattern is more consistent with a systematic difference in the convention applied to end vertebra selection than with random imprecision, and suggests that the AI system and the orthopedic surgeon applied comparable conventions.

A statistically significant proportional bias was detected between the AI system and the orthopedic surgeon ($B=0.062$; $p=0.019$). The clinical magnitude of this effect is, however, limited: across the range of curves observed in this cohort, the predicted difference attributable to proportional bias varied by less than 3° , remaining below the 5° threshold conventionally required to infer a genuine change in curve magnitude. Nevertheless, this finding merits attention in cohorts containing a greater proportion of severe deformities, particularly since automated measurement error has been reported to increase substantially in surgical AIS populations with extreme curves⁽²¹⁾. Categorical agreement differed substantially between the two thresholds examined. At the 25° brace-treatment threshold agreement between the AI system and the orthopedic surgeon was almost perfect ($\kappa=0.849$), whereas at the 10° diagnostic threshold it was only moderate ($\kappa=0.609$). Agreement between the two clinicians at this threshold was the lowest of all comparisons ($\kappa=0.480$), with discordant classifications in more than a quarter of patients. The reduction is therefore not specific to the automated system; it reflects the instability of dichotomizing a continuous measurement at a threshold around which many cases cluster. Categorization near the diagnostic threshold should accordingly not rest upon a single measurement, whether obtained manually or automatically. The sensitivity and specificity values reported here were derived using the orthopedic surgeon's measurement as the clinical reference and therefore quantify concordance with that observer; because the Cobb angle has no universally accepted gold standard, they are not estimates of true diagnostic accuracy. When the analysis was restricted to the 90 patients with radiographically confirmed scoliosis, agreement between the AI system and the orthopedic surgeon remained excellent ($ICC=0.949$), with a mean absolute difference of 2.48° . This indicates that the overall result was not an artifact of including patients with curves below the diagnostic threshold.

Study Limitations

Several limitations should be acknowledged. This was a single-center study of a single AI system, and the findings cannot be generalized to other platforms. Each radiograph was measured only once by each observer, so intra-observer reliability could not be assessed and the contribution of each clinician's internal variability to the observed interobserver difference remains unknown. Only one orthopedic surgeon and one radiologist participated; the interobserver variability reported here therefore reflects a single pair of raters rather than the range existing among clinicians, and a design with several raters per discipline would have permitted a more robust characterization. The distribution of curve magnitudes was skewed toward mild deformities, with 88.5% of curves below 25° and only three patients exceeding 40°; precision for large curves is correspondingly low and the findings should not be extrapolated to severe deformity, in which algorithm performance remains to be established. Only the coronal primary-curve Cobb angle was evaluated. Radiographs of inadequate diagnostic quality were excluded, which may have overestimated performance relative to unselected practice. Skeletal maturity was not recorded, so it could not be determined whether performance varies across stages of skeletal development. Finally, the Cobb angle has no absolute reference standard independent of observer judgment; all agreement, sensitivity and specificity values reported here describe concordance between observers rather than diagnostic accuracy against an external truth.

Within these constraints, AI-based measurement may be employed as a reliable adjunct for Cobb angle assessment in adolescents referred for scoliosis evaluation, with the potential to reduce measurement workload and improve consistency of reporting between observers and across serial examinations. Automated output should nevertheless remain subject to clinician review, particularly near the diagnostic threshold and for deformities of substantial magnitude. The software evaluated is moreover a research-mode prototype without regulatory clearance; these results characterize its measurement performance and do not constitute an endorsement of clinical deployment in its current form.

CONCLUSION

In adolescents referred for scoliosis assessment, AI-based Cobb angle measurement showed excellent agreement with an experienced orthopedic surgeon, with a mean absolute difference of 2.25° and no case differing by more than 10°. Agreement was of a similar order to that between the two clinicians and was almost perfect at the 25° brace-treatment threshold, but only moderate at the 10° diagnostic threshold, where the two clinicians also agreed poorly. AI-based measurement may therefore serve as a reliable adjunct in this setting, provided that automated output remains subject to clinician review near the diagnostic threshold.

Ethics

Ethics Committee Approval: This single-center retrospective study was approved by the Biruni University Clinical Research Ethics Committee (approval no: 2024-BİAEK/19-48, date: 30.03.2026).

Informed Consent: Because the study was based on the retrospective analysis of anonymized, previously acquired radiographs, the requirement for informed consent was waived.

Footnotes

During the preparation of this work the authors used Claude (Anthropic, San Francisco, CA, USA) to assist with drafting and language editing of the manuscript, with the preparation of the Bland-Altman figures, and with completion of the STROBE and GRRAS reporting checklists. All measurements were performed by the authors, and all statistical analyses were carried out by the authors using IBM SPSS Statistics version 27; every value reported in the manuscript was verified by the authors against the original statistical output. The study design, data collection and interpretation of the findings were undertaken solely by the authors. After carefully reviewing and editing the content as necessary, full responsibility for the publication's content is taken by the authors. This incorporation of AI tool usage primarily affected manuscript language, figure preparation and checklist completion; it did not affect the study data, the statistical results or their interpretation.

Authorship Contributions

Concept: Ş.S., Design: Ş.S., O.A., Data Collection or Processing: Ş.S., O.A., Analysis or Interpretation: Ş.S., O.A., Literature Search: Ş.S., O.A., Writing: Ş.S.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

REFERENCES

1. Weinstein SL, Dolan LA, Cheng JC, Danielsson A, Morcuende JA. Adolescent idiopathic scoliosis. *Lancet*. 2008;371:1527-37.
2. Konieczny MR, Senyurt H, Krauspe R. Epidemiology of adolescent idiopathic scoliosis. *J Child Orthop*. 2013;7:3-9.
3. Cobb JR. Outline for the study of scoliosis. *Instr Course Lect*. 1948;5:261-75.
4. Herzenberg JE, Waanders NA, Closkey RF, Schultz AB, Hensinger RN. Cobb angle versus spinous process angle in adolescent idiopathic scoliosis. The relationship of the anterior and posterior deformities. *Spine (Phila Pa 1976)*. 1990;15:874-9.
5. Negrini S, Donzelli S, Aulisa AG, Czaprowski D, Schreiber S, de Mauroy JC, et al. 2016 SOSORT guidelines: orthopaedic and rehabilitation treatment of idiopathic scoliosis during growth. *Scoliosis Spinal Disord*. 2018;13:3.
6. Morrissy RT, Goldsmith GS, Hall EC, Kehl D, Cowie GH. Measurement of the Cobb angle on radiographs of patients who have scoliosis. Evaluation of intrinsic error. *J Bone Joint Surg Am*. 1990;72:320-7.
7. Carman DL, Browne RH, Birch JG. Measurement of scoliosis and kyphosis radiographs. Intraobserver and interobserver variation. *J Bone Joint Surg Am*. 1990;72:328-33.

8. Gstoettner M, Sekyra K, Walochnik N, Winter P, Wachter R, Bach CM. Inter- and intraobserver reliability assessment of the Cobb angle: manual versus digital measurement tools. *Eur Spine J.* 2007;16:1587-92.
9. Langensiepen S, Semler O, Sobottke R, Fricke O, Franklin J, Schönau E, et al. Measuring procedures to determine the Cobb angle in idiopathic scoliosis: a systematic review. *Eur Spine J.* 2013;22:2360-71.
10. Sun Y, Xing Y, Zhao Z, Meng X, Xu G, Hai Y. Comparison of manual versus automated measurement of Cobb angle in idiopathic scoliosis based on a deep learning keypoint detection technology. *Eur Spine J.* 2022;31:1969-78.
11. Maeda Y, Nagura T, Nakamura M, Watanabe K. Automatic measurement of the Cobb angle for adolescent idiopathic scoliosis using convolutional neural network. *Sci Rep.* 2023;13:14576.
12. Wang MX, Kim JK, Choi JW, Park D, Chang MC. Deep learning algorithm for automatically measuring Cobb angle in patients with idiopathic scoliosis. *Eur Spine J.* 2024;33:4155-63.
13. Berlin C, Adomeit S, Grover P, Dreischarf M, Halm H, Dürr O, et al. Novel AI-based algorithm for the automated computation of coronal parameters in adolescent idiopathic scoliosis patients: a validation study on 100 preoperative full spine X-rays. *Global Spine J.* 2024;14:1728-37.
14. Chen K, Stotter C, Klestil T, Mitterer JA, Lepenik C, Nehrer S. Fully automated measurement of Cobb angles in coronal plane spine radiographs. *J Clin Med.* 2024;13:4122.
15. Wong JC, Reformat MZ, Parent EC, Stampe KP, Southon Hryniuk SC, Lou EH. Validation of an artificial intelligence-based method to automate Cobb angle measurement on spinal radiographs of children with adolescent idiopathic scoliosis. *Eur J Phys Rehabil Med.* 2023;59:535-42.
16. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet.* 2007;370:1453-7.
17. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J Clin Epidemiol.* 2011;64:96-106.
18. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* 2016;15:155-63.
19. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;1:307-10.
20. Li K, Gu H, Colglazier R, Lark R, Hubbard E, French R, et al. Deep learning automates Cobb angle measurement compared with multi-expert observers. *BJR|Artificial Intelligence.* 2025;2:ubaf009.
21. Bonanno MC, Ducou le Pointe H, Gaume M, Durteste M, Benhabiles M, Cohen A, et al. Performance of a deep learning-based algorithm for automated measurements of Cobb angles on preoperative spine radiographs in adolescent idiopathic scoliosis. *Pediatr Radiol.* 2026;56:618-28.