Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A novel terminal vertebral axis angle for assessing fusion after single-level anterior cervical corpectomy and fusion: a retrospective diagnostic accuracy study

Xin Xu
Affiliated 960th Hospital of PLA, Shandong Second Medical University
Linyang Song
Qilu Hospital of Shandong University
Junlin Han
Affiliated 960th Hospital of PLA, Shandong Second Medical University
Ang Li
Affiliated 960th Hospital of PLA, Shandong Second Medical University
Haozhe Zhang
Affiliated 960th Hospital of PLA, Shandong Second Medical University
Xuanlin Liu
Affiliated 960th Hospital of PLA, Shandong Second Medical University
Ruoxian Song
ruoxiansong@sina.com
Affiliated 960th Hospital of PLA, Shandong Second Medical University

Research Article

Keywords: Anterior cervical corpectomy and fusion, Terminal vertebral axis angle, Cervical fusion, Dynamic radiography, Diagnostic accuracy, Interspinous motion

Posted Date: November 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11779102/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Reliable assessment of fusion after anterior cervical corpectomy and fusion (ACCF) is challenging when anatomical structures are obscured by implants or overlapping bone. We evaluated the feasibility, image quality, observer reliability, and diagnostic accuracy of a terminal vertebral axis angle (TVAA) derived from terminal-vertebral osseous landmarks.

Methods

This retrospective diagnostic accuracy study included 132 patients after single-level ACCF. Dynamic TVAA change (ΔTVAA), distance-based and angular interspinous motion (ΔISM-dis and ΔISM-ang), and segmental Cobb motion (ΔCobb) were compared with CT-defined extragraft bridging bone (Grades 0–1, non-complete fusion; Grade 2, complete fusion). Receiver operating characteristic analyses, reliability assessment, internal bootstrap resampling, and exploratory landmark-visibility analyses were performed.

Results

Forty of 132 patients (30.3%) had CT-defined non-complete fusion. TVAA was measurable in 132 of 136 patients undergoing feasibility assessment (97.1%). Under a post hoc acceptable-quality criterion requiring a total visibility score ≥ 14/16 and at least one clearly visible endpoint on each reference axis, 125 of 132 studies (94.7%) qualified; 59 (44.7%) fulfilled the stricter eight-landmark high-confidence criterion. Interobserver and intraobserver intraclass correlation coefficients for ΔTVAA were 0.943 and 0.957. Its AUC was 0.910 (95% CI 0.859–0.962), with sensitivity 82.5% and specificity 85.9% at the exploratory cutoff > 1.38°. AUC was higher than for ΔCobb (0.798; P = 0.001) but not significantly different from ΔISM-dis (0.885; P = 0.287) or ΔISM-ang (0.878; P = 0.176). Exploratory ΔTVAA AUCs were 0.912 (0.859–0.965) in the acceptable-quality group (n = 125) and 0.913 (0.836–0.991) in the strict high-confidence group (n = 59).

Conclusions

TVAA showed good discrimination and high observer reliability for CT-defined fusion after single-level ACCF. Its diagnostic performance was not demonstrably superior to established ISM measurements. Image-quality stratification supports further evaluation of its technical applicability, but the proposed quality criterion and diagnostic cutoff require independent validation.

Clinical trial registration

Not applicable.

1. Background

Anterior cervical corpectomy and fusion (ACCF) provides direct decompression and anterior column reconstruction for selected patients with degenerative cervical myelopathy [1, 2]. Three-dimensional (3D)-printed vertebral bodies have expanded reconstructive options beyond conventional titanium mesh cages [3]. Regardless of implant type, reliable assessment of postoperative osseous fusion is important because radiographic nonunion may affect construct stability, although it does not always indicate symptomatic pseudarthrosis [4].

Fusion criteria vary across studies, including bridging bone, graft–vertebral radiolucency, and residual motion on dynamic radiographs [5]. Multiplanar computed tomography (CT) depicts osseous continuity and is commonly used when plain radiographs are inconclusive [6]. Extragraft bridging bone (ExGBB) has shown diagnostic value in studies with surgical confirmation [7, 8]. However, CT evaluates morphology rather than dynamic stability, making flexion–extension radiographs a useful complementary examination. Because these modalities describe different aspects of the reconstructed segment, a dynamic method should be judged against an explicit reference standard without assuming that radiographic motion and anatomical nonunion are interchangeable.

Dynamic assessment generally relies on interspinous motion (ISM) and segmental Cobb angle changes. Quantitative ISM has been investigated for detecting pseudarthrosis, with adequate adjacent-segment motion recognized as important [9]. Alternative angular methods have also been described after anterior cervical discectomy and fusion [10]. More recently, Wang et al. assessed distance-based and angular ISM specifically after ACCF [11]. Despite these developments, small residual movements can be difficult to measure reproducibly when osseous landmarks are poorly defined. Such uncertainty may be particularly relevant when proposed decision boundaries differ by only a few degrees or millimeters. A useful measurement should therefore demonstrate not only discrimination against a reference standard but also acceptable technical feasibility and agreement between readers.

After corpectomy, anterior implants and overlapping vertebral contours may obscure conventional reference structures. Different reconstruction devices may further alter postoperative radiographic appearance [12, 13]. An angle derived from landmarks on the cranial and caudal terminal vertebrae could provide an alternative means of measuring residual motion, but its feasibility and diagnostic performance require evaluation. Separating the ability to obtain a numerical measurement from the confidence with which its component landmarks are visualized is particularly important for transparent assessment of a new radiographic technique.

We therefore developed the terminal vertebral axis angle (TVAA) using anterior vertebral body corners and posterior spinolaminar cortical turning points. Its construction does not depend on the superior endplate of the caudal terminal vertebra, which may be partly obscured by anterior reconstruction. This study evaluated TVAA feasibility, observer reliability, and diagnostic accuracy for CT-defined fusion after

single-level ACCF, compared TVAA with established dynamic measures, and examined the stability of an exploratory diagnostic threshold.

2. Methods

2.1 Study design and patient selection

This single-center retrospective diagnostic accuracy study was conducted at the Affiliated 960th Hospital of PLA, Shandong Second Medical University, and was reported with reference to STARD guidance [14]. Electronic medical records and institutional PACS images of patients undergoing single-level ACCF between January 2020 and June 2024 were screened. The study population was determined by the available eligible cases during this interval; no formal a priori sample size calculation was undertaken. The precision of diagnostic estimates was described with 95% confidence intervals.

Eligible patients underwent C4, C5, or C6 corpectomy with C3–C5, C4–C6, or C5–C7 reconstruction, respectively, for cervical spondylotic myelopathy, ossification of the posterior longitudinal ligament, or mixed degenerative pathology. Patients were required to have at least 24 months of follow-up and postoperative reconstructed CT and flexion–extension lateral radiographs obtained within 90 days of each other. Both three-dimensional (3D)-printed artificial vertebral bodies and titanium mesh cages were eligible. Exclusion criteria were multilevel or other-level corpectomy, confounding additional or previous cervical surgery, nondegenerative indications, follow-up < 24 months, inadequate or unavailable imaging, or a CT–radiograph interval > 90 days.

Among 188 screened patients, 52 failed eligibility criteria. Landmark feasibility was evaluated in 136 patients, of whom four lacked identifiable TVAA landmarks. The remaining 132 formed the diagnostic accuracy cohort. CT Grades 0, 1, and 2 comprised 18, 22, and 92 patients, respectively (Fig. 1).

2.2 Dynamic radiographic measurements

Postoperative lateral cervical radiographs were obtained in maximal voluntary flexion and extension, and measurements were made using institutional picture archiving and communication system tools. Cranial and caudal terminal vertebrae were C3/C5, C4/C6, and C5/C7 for C4, C5, and C6 corpectomy, respectively.

2.2.1 Terminal vertebral axis angle

TVAA was defined by four osseous landmarks on each lateral radiograph (Fig. 2): a, the anterosuperior corner of the cranial terminal vertebral body; b, the superior cortical turning point where the cranial lamina meets the base of the spinous process; c, the anteroinferior corner of the caudal terminal vertebral body; and d, the corresponding inferior lamina–spinous process base cortical turning point of the caudal terminal vertebra. Posterior points b and d were defined at the spinous process base rather than its tip. Lines ab and cd formed the cranial and caudal reference axes, respectively. TVAA was the

relative angle between these axes and was measured separately during flexion and extension. Dynamic motion was calculated as $\Delta TVAA = |TVAA_{extension} - TVAA_{flexion}|$, with a consistent angular convention in both positions. This construction does not require delineation of the superior endplate of the caudal terminal vertebra.

2.2.2 Conventional dynamic measurements

Three established parameters served as comparators: distance-based interspinous motion (ΔISM-dis), angular interspinous motion (ΔISM-ang), and segmental Cobb angular motion (ΔCobb). For ΔISM-dis, the distance between corresponding spinous process calibration points on the cranial and caudal terminal vertebrae was measured in flexion and extension; their absolute difference was recorded in millimeters. For ΔISM-ang, lines connected the midpoint of each anterior fixation screw head with the corresponding posterior spinous process calibration point, and the absolute change in the angle between the cranial and caudal lines was recorded. Segmental Cobb angles were formed by the superior endplate of the cranial terminal vertebra and inferior endplate of the caudal terminal vertebra; their absolute flexion–extension difference defined ΔCobb. These measurements followed published ACCF principles [11, 12]. Published ΔISM-dis > 1.50 mm and ΔISM-ang > 1.75° thresholds [11] and fixed ΔCobb thresholds > 4° and > 2° were assessed separately from exploratory within-cohort cutoffs.

2.3 CT reference standard

Sagittal and coronal reconstructed CT images were graded using a study-defined classification based on extragraft bridging bone (ExGBB), consistent with established cervical fusion assessment principles [6–8]. Grade 0 indicated no definite ExGBB; Grade 1 indicated partial or discontinuous bridging; and Grade 2 indicated continuous ExGBB across the reconstruction without definite interruption (Fig. 3). Two CT readers assessed images independently while blinded to dynamic radiographic measurements; 14 discordant assessments were adjudicated by a third experienced reader. Initial grades were retained for agreement analyses, and adjudicated grades were used as the reference classification. Grade 2 was defined as CT-complete fusion and Grades 0–1 as CT-non-complete fusion. This imaging definition was not interpreted as surgically verified pseudarthrosis.

2.4 Landmark visibility and measurement reliability

Each landmark a–d was independently assessed on flexion and extension images using a three-point ordinal scale: 0, not reliably identifiable; 1, identifiable with uncertainty; and 2, clearly identifiable. TVAA measurability required all eight scores to be ≥ 1. Strict high-confidence visualization required all eight scores to equal 2. To explore the effect of image quality on diagnostic performance, a post hoc acceptable TVAA measurement-quality criterion was defined as a total visibility score ≥ 14/16, with at least one clearly identifiable endpoint (score 2) on each of four axes: ab and cd in flexion, and ab and cd in extension. Additional exploratory score thresholds of ≥ 14/16 without the axis rule and ≥ 15/16, and the joint strict-quality criterion for all four methods, were also assessed. These nested definitions characterize different degrees of landmark visibility; they are not interchangeable.

Both readers classified images according to the acceptable TVAA criterion to assess interobserver agreement. The conventional comparators were assessed using their own method-specific strict-quality flags, which were not directly equivalent to the TVAA scoring scale. Numerical measurement availability was distinguished from image-quality confidence: all four dynamic parameters had recorded values in all 132 patients included in the primary paired ROC analysis. The 136-patient TVAA landmark-identification proportion was therefore not directly compared as a like-for-like feasibility measure with the conventional strict-quality proportions.

Continuous-measurement reliability was evaluated in 36 patients representing different corpectomy levels and CT fusion grades. Two readers, blinded to CT grading, independently measured ΔTVAA, ΔISM-dis, ΔISM-ang, and ΔCobb; Reader 1 repeated the measurements after two weeks. Intraclass correlation coefficients (ICCs) with 95% confidence intervals used a two-way random-effects, absolute-agreement, single-measurement model [15]. Bland–Altman analysis calculated the mean paired observer difference (Reader 2 minus Reader 1) and 95% limits of agreement (mean ± 1.96 SD of differences) [16]. Unweighted Cohen's κ was used to assess individual landmark visibility ratings and classification under the acceptable-quality rule; CT grades were assessed using unweighted and weighted κ.

2.5 Classification reproducibility

The exploratory cutoff of ΔTVAA > 1.38° for CT-non-complete fusion was applied unchanged to the 36-patient reliability subset. Interobserver comparisons used the initial measurements of the two readers, and intraobserver comparisons used Reader 1's initial and two-week repeat measurements. Percentage agreement and Cohen's κ were reported; percentile confidence intervals for κ were estimated using 5,000 paired patient-level bootstrap resamples (random seed 2026). This assessed repeatability of threshold classification, not independent validation of the cutoff.

2.6 Statistical analysis

All analyses were conducted in R version 4.5. Continuous variables were summarized as mean ± SD, or median and interquartile range when appropriate, and categorical variables as n (%). Between-group comparisons used Welch's unequal-variance t test with corresponding 95% confidence intervals for mean differences and Mann–Whitney U tests as complementary rank-based analyses. Categorical variables were compared using Pearson χ^2 or Fisher's exact tests as appropriate. Across the three CT grades, one-way ANOVA and Kruskal–Wallis tests assessed overall differences; global tests were not interpreted as evidence that every pairwise contrast was significant.

With CT Grades 0–1 designated as the positive reference outcome and Grade 2 as the negative outcome, receiver operating characteristic (ROC) curves and AUCs with 95% DeLong confidence intervals were calculated for each dynamic parameter in the 132-patient cohort. Correlated AUCs were compared using paired DeLong tests [17]. Exploratory cutoffs maximizing the Youden index (sensitivity + specificity − 1) were used to estimate in-sample sensitivity, specificity, predictive values, and accuracy. The primary ΔTVAA decision boundary (> 1.38°) was retained without subgroup-specific optimization for

sensitivity and exploratory subgroup analyses. No multiplicity correction was applied to the exploratory comparisons.

Internal sampling variability was evaluated with 2,000 non-stratified patient-level bootstrap resamples, each of size 132, using random seed 2026. AUC and Youden-optimal cutoff were re-estimated in each resample. When Youden indices tied, the threshold maximizing sensitivity and then specificity was selected; the decision boundary was defined by the midpoint between adjacent observations. The mean AUC, percentile intervals, and cutoff distribution were summarized. This resampling assessed sampling stability but did not provide an independent or optimism-corrected validation of cutoff-based performance.

2.7 Sensitivity and subgroup analyses

Sensitivity analyses assessed CT Grade 0 versus Grade 2 after excluding Grade 1, a CT–radiograph interval ≤ 30 days, adequate dynamic effort (C2–C7 range of motion ≥ 20° and upper adjacent interspinous motion ≥ 4 mm), and a subset meeting high-confidence criteria for all three conventional comparator methods (n = 118). The last subset did not require strictly clear TVAA landmarks. Exploratory visibility-quality analyses separately examined a score-only threshold ≥ 14/16 (n = 129), the acceptable-quality definition (n = 125), a score threshold ≥ 15/16 (n = 110), strict TVAA high confidence (n = 59), and strict quality across all four tests (n = 53). The AUCs of TVAA and comparator methods were additionally compared within selected patient-matched visibility subsets. These quality analyses were conducted post hoc.

Further exploratory subgroups were defined by corpectomy level (C4, C5, or C6), implant type (3D-printed artificial vertebral body or titanium mesh cage), and the absence of implant subsidence ≥ 3 mm. Subgroup AUCs, DeLong confidence intervals, and sensitivity and specificity at the unchanged ΔTVAA cutoff > 1.38° were reported descriptively. Tests were two-sided, with $P < 0.05$ used for statistical significance.

2.8 Ethics approval

The Ethics Committee of the Affiliated 960th Hospital of PLA, Shandong Second Medical University approved this retrospective study (approval No. 2024009) and waived written informed consent. The study was conducted in accordance with the Declaration of Helsinki and its later amendments.

3. Results

3.1 Patient characteristics

A total of 188 patients were screened, 52 were excluded before feasibility assessment, and four of the remaining 136 could not undergo TVAA measurement. The final diagnostic cohort comprised 132 patients (Fig. 1). Their mean age was 56.41 ± 8.20 years, 80 were men (60.6%), and mean follow-up was 36.54 ± 7.12 months. Corpectomy was performed at C4 (n = 32), C5 (n = 58), or C6 (n = 42); 65 patients

received a 3D-printed artificial vertebral body and 67 a titanium mesh cage. The mean CT–dynamic radiograph interval was 16.57 ± 18.48 days. CT grading identified 18 Grade 0, 22 Grade 1, and 92 Grade 2 cases, corresponding to 40 non-complete and 92 complete fusions. Current smoking (35.0% vs 14.1%; P = 0.006), osteoporosis/low bone mass (40.0% vs 4.3%; P < 0.001), and lower adjacent vertebral Hounsfield units (277.10 ± 51.67 vs 313.87 ± 40.85; P < 0.001) were more frequent or pronounced in non-complete fusion. Other characteristics are provided in Table 1.

3.2 Measurability and landmark-visibility quality

TVAA landmarks were identifiable in 132 of 136 feasibility assessments (97.1%), and all four numerical dynamic measurements were available in the final 132-patient cohort. Based on eight landmark ratings, 129 patients (97.7%) had a total score ≥ 14/16, and 125 (94.7%) fulfilled the composite acceptable-quality definition requiring both score ≥ 14 and the reference-axis endpoint rule. A score ≥ 15/16 was observed in 110 patients (83.3%). Strict high-confidence visualization in both positions was achieved in 59 patients (44.7%); 53 (40.2%) met the strict quality criteria for TVAA and all three comparators simultaneously (Table 2).

The acceptable-quality criterion was met in 125 of 132 cases (94.7%) according to Reader 1 and 124 (93.9%) according to Reader 2. The readers agreed on 129 cases (97.7%; Cohen's κ = 0.788). Among individual TVAA landmarks, both-position clear visibility was highest for landmark c (116/132, 87.9%) and lowest for landmark d (101/132, 76.5%). Strict-quality flags for ΔISM-dis, ΔISM-ang, and ΔCobb were present in 125/132 (94.7%), 124/132 (93.9%), and 126/132 (95.5%) patients, respectively; their quality definitions differed from those for TVAA and were not compared as equivalent rates. Per-landmark and per-view proportions and individual visibility κ estimates are shown in Table 2.

3.3 Measurement and CT reader agreement

In the 36-patient reliability subset, interobserver ICCs were 0.943 (95% CI 0.896–0.972) for ΔTVAA, 0.964 (0.934–0.982) for ΔISM-dis, 0.956 (0.914–0.977) for ΔISM-ang, and 0.890 (0.790–0.941) for ΔCobb. Corresponding intraobserver ICCs were 0.957 (0.918–0.978), 0.981 (0.962–0.990), 0.984 (0.969–0.992), and 0.925 (0.861–0.962), respectively (Table 2). For ΔTVAA, mean interobserver bias was 0.090°, with 95% Bland–Altman limits of agreement − 0.498° to 0.677° (Supplementary Fig. S1; Supplementary Table S2).

The initial CT readers assigned identical fusion grades in 118/132 cases (89.4%; unweighted κ = 0.791). Linear-weighted κ was 0.844 (95% CI 0.751–0.918), and quadratic-weighted κ was 0.897 (0.826–0.947). Fourteen discordant cases required third-reader adjudication (Supplementary Table S1).

3.4 Radiographic motion across CT fusion grades

All four dynamic radiographic parameters differed overall across CT Grades 0, 1, and 2 (Kruskal–Wallis P < 0.001 for each; Table 3; Fig. 4). Mean ΔTVAA decreased from 2.61 ± 0.74° in Grade 0 to 1.70 ± 0.51° in Grade 1 and 0.97 ± 0.46° in Grade 2. The corresponding ΔISM-dis values were 2.94 ± 0.81, 1.99 ± 0.77,

and 1.16 ± 0.52 mm; ΔISM-ang values were 3.67 ± 1.13°, 2.51 ± 0.92°, and 1.47 ± 0.66°; and ΔCobb values were 3.94 ± 1.27°, 2.67 ± 1.03°, and 1.88 ± 1.04°, respectively.

In the binary reference-standard comparison, mean ΔTVAA was 2.11 ± 0.77° for non-complete versus 0.97 ± 0.46° for complete fusion, with a mean difference of 1.14° (95% Welch CI 0.88–1.40; P < 0.001). Differences were also significant for ΔISM-dis (2.42 ± 0.91 vs 1.16 ± 0.52 mm), ΔISM-ang (3.03 ± 1.16° vs 1.47 ± 0.66°), and ΔCobb (3.24 ± 1.29° vs 1.88 ± 1.04°), each P < 0.001. Overall C2–C7 motion, upper adjacent-segment ISM, and fused-segment height loss were not statistically different between the binary groups (Table 3).

3.5 Primary diagnostic accuracy

ΔTVAA had an AUC of 0.910 (95% DeLong CI 0.859–0.962) in differentiating CT-defined non-complete from complete fusion. AUCs were 0.885 (0.819–0.950) for ΔISM-dis, 0.878 (0.813–0.943) for ΔISM-ang, and 0.798 (0.716–0.879) for ΔCobb (Table 4; Fig. 5). Paired DeLong comparisons indicated higher discrimination for ΔTVAA than ΔCobb (P = 0.001), but not versus ΔISM-dis (P = 0.287) or ΔISM-ang (P = 0.176).

The exploratory ΔTVAA cutoff > 1.38° yielded a sensitivity of 82.5% (33/40), specificity of 85.9% (79/92), positive predictive value of 71.7%, negative predictive value of 91.9%, and apparent accuracy of 84.8%. At their respective Youden-derived cutoffs, ΔISM-dis (> 1.64 mm), ΔISM-ang (> 1.92°), and ΔCobb (> 2.26°) had sensitivities of 80.0%, 80.0%, and 80.0% and specificities of 85.9%, 81.5%, and 66.3%, respectively. Diagnostic results for published ISM and fixed Cobb thresholds are shown in Table 4.

3.6 Bootstrap assessment and threshold reproducibility

Across 2,000 patient-level bootstrap resamples, the mean ΔTVAA AUC was 0.909 (95% percentile interval 0.850–0.955). Bootstrap mean AUCs were 0.885 (0.814–0.940) for ΔISM-dis, 0.879 (0.812–0.938) for ΔISM-ang, and 0.797 (0.707–0.874) for ΔCobb. The ΔTVAA cutoff distribution had a mean of 1.26°, median of 1.31°, and 95% percentile interval of 1.08°–1.62° (Table 5).

At ΔTVAA > 1.38° in the 36-patient repeatability sample, interobserver classification agreement was 32/36 (88.9%; κ = 0.780, 95% bootstrap CI 0.565–0.944), and intraobserver agreement after two weeks was 34/36 (94.4%; κ = 0.888, 95% bootstrap CI 0.717–1.000; Supplementary Table S4).

3.7 Sensitivity, quality-stratified and subgroup analyses

In sensitivity analyses, ΔTVAA AUCs were 0.964 (95% CI 0.923–1.000; n = 110) for CT Grade 0 versus Grade 2, 0.919 (0.866–0.971; n = 115) for a CT–radiograph interval ≤ 30 days, 0.907 (0.852–0.962; n = 118) with adequate dynamic effort, and 0.921 (0.875–0.966; n = 118) in the comparator-quality-restricted subset. The last subset required high-confidence measurements for the three conventional comparators, not strict high-confidence TVAA visibility (Table 5; Supplementary Fig. S2).

Exploratory visibility-based analyses yielded ΔTVAA AUCs of 0.908 (95% CI 0.855–0.960; n = 129) with a total score ≥ 14/16, 0.912 (0.859–0.965; n = 125) under the acceptable-quality composite criterion, 0.934 (0.891–0.977; n = 110) with a total score ≥ 15/16, 0.913 (0.836–0.991; n = 59) with strict high-confidence TVAA visualization, and 0.896 (0.806–0.986; n = 53) with strict quality across all four methods. At the unchanged > 1.38° cutoff, sensitivity and specificity were 84.2% (32/38) and 85.1% (74/87) in the acceptable-quality subset and 83.3% (10/12) and 80.9% (38/47) in the strict TVAA high-confidence subset. Paired ΔTVAA versus ΔISM-dis AUC comparisons within these same patients were not statistically significant (n = 125, P = 0.526; n = 59, P = 0.295); corresponding ΔISM-ang comparisons yielded P = 0.355 and P = 0.056. These subgroup analyses were exploratory (Table 5; Supplementary Table S3; Supplementary Fig. S2).

Across corpectomy levels C4, C5, and C6, ΔTVAA AUCs were 0.896, 0.894, and 0.939, respectively. By implant type, AUCs were 0.864 for 3D-printed artificial vertebral bodies and 0.947 for titanium mesh cages; among patients without subsidence ≥ 3 mm (n = 99), the AUC was 0.920. These subgroup results are descriptive and do not establish between-subgroup differences. Full estimates and 95% confidence intervals are given in Table 5 and Supplementary Table S3 and visualized in Supplementary Fig. S2.

4. Discussion

This study examined the diagnostic accuracy and measurement properties of a terminal-vertebral osseous landmark angle after single-level ACCF. ΔTVAA showed good discrimination for CT-defined fusion status (AUC 0.910) and high reader reliability. Its AUC was significantly higher than conventional segmental Cobb motion but did not differ significantly from established distance- or angle-based ISM. In exploratory visibility-based analyses, discrimination remained numerically similar across nested imaging-quality strata. These findings support further investigation of TVAA as a complementary radiographic measure while emphasizing the need to distinguish image availability, confidence in landmark identification, and precision of the resulting measurements.

4.1 Diagnostic performance and anatomical basis

Residual segmental motion is a practical marker in dynamic assessment of cervical fusion. Wang et al. demonstrated the value of distance- and angle-based ISM following ACCF [11]. Nevertheless, small measurements may be vulnerable to landmark and reader variation: Martin et al. noted variability in dynamic radiographic fusion assessments [18], while Wang et al. described a modified cervical motion measurement with improved reproducibility [19].

ΔTVAA decreased from CT Grade 0 to Grade 2, consistent with less residual motion in segments exhibiting more extensive bridging. However, its numerical AUC advantage over ISM did not reach statistical significance. TVAA's use of anterior cortical corners and posterior spinolaminar turning points may reduce dependence on obscured terminal endplates, but the current data cannot establish a mechanism of superiority. Its significant comparison with Cobb motion suggests potential utility when

conventional reference lines are difficult to construct. Unlike the more familiar endplate or spinous-tip references, TVAA uses cortical turning points near the spinous process bases, potentially altering the effects of overlap and projection. However, our analysis did not compare the physical visibility of these structures using directly equivalent criteria; the proposed anatomical advantage remains a design rationale rather than a proven mechanism.

The comparison of correlated ROC curves also has an important inferential limitation. A non-significant DeLong test against either ISM method does not demonstrate equivalence or non-inferiority; these would require explicit margins and adequately powered analyses. Conversely, the statistically significant difference from ΔCobb establishes a difference in discrimination within this dataset, not a general advantage in clinical decision-making. AUC summarizes ranking over all possible thresholds, whereas clinicians must act at a specific decision boundary. Therefore, discrimination, classification accuracy, and reproducibility of the chosen threshold are separate properties that should inform the interpretation of TVAA.

4.2 CT fusion versus functional stability

CT and flexion–extension radiographs assess complementary attributes: osseous continuity and movement under voluntary loading. Studies comparing CT, plain radiography, and surgical exploration have documented disagreement between modalities [20, 21], and recent evidence continues to emphasize heterogeneity in fusion criteria [22]. Our reference classification used ExGBB continuity; it should not be equated with surgically confirmed pseudarthrosis. The intermediate motion observed for Grade 1 is compatible with incomplete bridging but cannot establish longitudinal progression. Excluding Grade 1 increased AUC, partly because distinguishing the extreme grades is a less challenging task. The apparent separation should therefore not be translated into claims about detecting clinically important pseudarthrosis in routine practice. Moreover, morphology–motion discordance can arise from incomplete maturation, instrumentation-related constraints, or limits of voluntary radiographic motion; our cross-sectional data cannot determine their relative contributions.

A further source of uncertainty is the interval between the index test and reference-standard examination. Although CT and dynamic radiography were obtained within 90 days, osseous bridging may evolve during postoperative follow-up; consequently, measurements performed on different dates may not describe precisely the same biological state. The shorter-interval sensitivity analysis provides a useful check, but cannot fully remove this concern. Moreover, defining Grades 0 and 1 together creates a clinically heterogeneous positive category: a segment with no bridging bone may differ from one showing partial continuity in both motion and future fusion potential. Reporting the individual CT grades and the Grade 0-versus-Grade 2 analysis therefore complements, rather than replaces, the primary binary analysis.

4.3 Landmark visibility, quality strata, and observer agreement

TVAA measurements could be obtained in 132 of the 136 feasibility cases (97.1%), but obtaining a numerical value did not guarantee unequivocally clear landmarks. Only 59 of 132 patients (44.7%) fulfilled the strict criterion that all four landmarks (a–d) be clearly identifiable in both flexion and extension. The post hoc acceptable-quality criterion was intentionally less restrictive: it allowed a limited number of uncertain landmarks while requiring at least one clearly visible endpoint for each reference axis. Under this definition, 125 patients (94.7%) qualified, and the two observers agreed on quality classification in 129 of 132 cases ($\kappa = 0.788$). This establishes repeatability of a dichotomized image-quality judgment, not independent proof that angular measurements are equally precise under both visibility standards. The conventional comparators used method-specific quality flags, so numerical proximity between their quality proportions and the acceptable TVAA proportion cannot be interpreted as equivalence of landmark visibility.

The AUC for ΔTVAA was 0.912 in the acceptable-quality subset and 0.913 in the 59-patient strict high-confidence subset, compared with 0.910 in the full cohort. These similar point estimates suggest that detectable discrimination was not confined to a single visibility stratum in this sample, but do not establish non-inferiority or equivalence between strata. The groups are nested, the quality definitions were evaluated post hoc, and the strict high-confidence subset contained only 12 CT-non-complete cases. In same-patient paired comparisons, neither the 125-patient nor the 59-patient analysis showed statistically significant superiority of TVAA over ISM-dis or ISM-ang. Moreover, the 118-patient subset defined by high-quality conventional comparator measurements should not be construed as a TVAA high-visibility group; a stricter all-four-method subset comprised 53 patients. These distinctions avoid conflating numerical availability with diagnostic validity or image clarity.

Measurement reliability needs interpretation beyond high ICCs. TVAA interobserver and intraobserver ICCs were 0.943 and 0.957, respectively, yet its Bland–Altman 95% limits of agreement were − 0.498° to 0.677°. At an exploratory decision boundary of 1.38°, observer differences of this magnitude may change classification for patients whose values lie near the threshold. Four of 36 patients were classified differently between readers, and two differed between repeated measurements by the same reader. Standardized landmark selection, repeat readings in uncertain images, and an independently derived indeterminate range may improve practical use, but none of these procedures was validated here. Prior work on automated cervical landmark measurement underscores the potential value of reproducible annotation rules [23]; any sagittal radiograph also remains limited in representing three-dimensional intervertebral motion [24].

4.4 Interpretation of the exploratory threshold

The > 1.38° cutoff was selected and evaluated in the same cohort, potentially inflating apparent performance. Data-driven selection of optimal thresholds can introduce bias, especially with relatively few positive cases [25, 26], and the selected cutoff depends on the optimization approach [27]. Bootstrap resampling yielded a cutoff median of 1.31° and 95% interval of 1.08°–1.62°, indicating appreciable uncertainty. These distributions are neither external validation nor optimism-corrected validation of the full selection process. Accordingly, the proposed threshold remains exploratory and

should be interpreted alongside image quality, other radiographs, CT when appropriate, and symptoms. A clinically useful decision rule would require external assessment of classification accuracy, explicit handling of values near the boundary, and evaluation of potential false-positive consequences before recommending routine adoption.

Selecting a cutoff solely by the maximum Youden index also implicitly assigns equal importance to sensitivity and specificity, although the consequences of false-positive and false-negative classifications may differ in postoperative care. An incorrectly elevated TVAA might prompt unnecessary additional imaging or concern, whereas an incorrectly low value could offer false reassurance when further evaluation is warranted. Neither the relative clinical costs of these errors nor a threshold-specific management pathway was evaluated here. Furthermore, positive and negative predictive values depend on the prevalence of CT-defined non-complete fusion, which was 30.3% in this selected cohort. Evaluation in a broader follow-up population should therefore examine calibration of classification risks and the practical consequences of proposed decision rules, not only their AUC.

4.5 Clinical context, limitations, and future research

AUCs varied numerically across implant types and levels, but subgroup samples were small and no formal effect modification was established. Smoking, osteoporosis/low bone mass, and lower vertebral CT attenuation were more frequent or pronounced in non-complete fusion. Previous ACCF research has linked such factors and implant geometry to subsidence after 3D-printed reconstruction [28, 29] or titanium mesh cage implantation [30]. Subsidence and fusion, however, are distinct outcomes, and these observational comparisons do not establish causality. The numerical difference in AUC between implant groups might reflect chance, case mix, or imaging geometry; with no formal interaction test, it should not be used to claim implant-specific superiority of the measurement.

TVAA could provide an additional quantitative measure from routine dynamic radiographs, but predictive values depend on fusion prevalence and should not be extrapolated uncritically. Imaging nonunion alone does not mandate reoperation, as demonstrated by work relating dynamic findings to revision after anterior cervical fusion [31]. The intended role of TVAA is therefore complementary classification of radiographic motion, not a standalone clinical decision on reoperation or CT replacement.

Several limitations warrant consideration. This retrospective single-center study relied on imaging-based ExGBB rather than surgical verification, combined absent and partial bridging in the primary reference-positive category, and excluded four patients whose TVAA landmarks could not be identified. Although numerical measurements were present for all four dynamic tests in the 132-patient cohort, some underlying landmarks were uncertain; image-quality scores for the comparators were not harmonized with TVAA ratings. The acceptable-quality rule and visibility-stratified comparisons were developed after inspection of the dataset, and the overlapping strata and multiple exploratory analyses leave room for selection and multiplicity effects. The strict TVAA high-confidence subset included only 12 reference-positive cases, limiting precision of its AUC and sensitivity estimates. Independent measurement repeatability was assessed in only 36 patients, and this sample contained too few images failing the

acceptable-quality criterion to reliably compare precision between high- and lower-quality examinations. No formal a priori sample size calculation was performed; the sample was defined by available eligible records, and 95% confidence intervals quantify precision rather than establish adequate power for non-inferiority.

Further constraints include variation in patient effort, positioning, and radiographic projection, the interval between CT and dynamic radiography, possible changes in fusion status over time, and lack of direct correlation with symptoms, functional recovery, revision surgery, or longitudinal fusion progression. The 2,000-resample bootstrap evaluated sampling variability but did not provide optimism-corrected validation of the 1.38° cutoff. Prospective multicenter studies should predefine landmark-quality thresholds and diagnostic objectives, use comparable quality criteria across methods, measure repeatability in deliberately sampled low-confidence images, and assess whether TVAA offers meaningful incremental information when combined with established ISM measures. The diagnostic threshold and image-quality rule should be independently validated before routine clinical adoption.

5. Conclusions

A terminal vertebral axis angle based on cranial and caudal osseous landmarks demonstrated good diagnostic discrimination and high observer reliability for CT-defined fusion after single-level ACCF. ΔTVAA outperformed segmental Cobb motion in this cohort, whereas differences relative to established interspinous motion measurements were not statistically significant. Exploratory quality-stratified analyses yielded similar AUC estimates under the acceptable measurement-quality definition and the stricter eight-landmark visibility criterion, although these results do not establish equivalent measurement precision or external validity. TVAA may be considered a complementary dynamic radiographic index; prospective, independently validated assessment of landmark quality, diagnostic thresholds, and clinical utility is required.

Abbreviations

3D
Three-dimensional
ACCF
Anterior cervical corpectomy and fusion
AUC
Area under the receiver operating characteristic curve
AVB
Artificial vertebral body
BMI
Body mass index
CI
Confidence interval

CSM
Cervical spondylotic myelopathy
CT
Computed tomography
DR
Digital radiography
ExGBB
Extragraft bridging bone
HU
Hounsfield unit
ICC
Intraclass correlation coefficient
IQR
Interquartile range
ISM
Interspinous motion
ISM-ang
Angular interspinous motion
ISM-dis
Distance-based interspinous motion
LoA
Limits of agreement
NPV
Negative predictive value
OPLL
Ossification of the posterior longitudinal ligament
PACS
Picture archiving and communication system
PPV
Positive predictive value
ROC
Receiver operating characteristic
ROM
Range of motion
SD
Standard deviation
STARD
Standards for Reporting Diagnostic Accuracy Studies
TMC
Titanium mesh cage

TVAA
Terminal vertebral axis angle
ΔCobb
Dynamic change in segmental Cobb angle
ΔISM-ang
Dynamic change in angular interspinous motion
ΔISM-dis
Dynamic change in distance-based interspinous motion
ΔTVAA
Dynamic change in terminal vertebral axis angle

Declarations

Ethics approval and consent to participate

This retrospective study was approved by the Ethics Committee of the Affiliated 960th Hospital of PLA, Shandong Second Medical University (Approval No. 2024009). The requirement to obtain written informed consent was waived by the ethics committee due to the retrospective nature of the study. All procedures were performed in accordance with the ethical standards of our institution and with the 1964 Declaration of Helsinki and its later amendments.

Consent for publication

Not applicable. The radiographic and CT examples are de-identified and contain no direct personal identifiers.

Competing interests

The authors declare no competing interests.

Funding

This study received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

X.X. and R.S. conceived and designed the study. X.X. and L.S. contributed equally to this work. X.X. performed the statistical analyses and was the primary contributor to drafting the manuscript. L.S. contributed to clinical data collection and management, figure preparation, and manuscript organization.

J.H. and A.L. assisted with clinical data acquisition and radiographic measurements. H.Z. and X.L. contributed to imaging assessment, data verification, and preparation of figures and tables. R.S. supervised the study and critically revised the manuscript. All authors contributed to the interpretation of the results, reviewed and approved the final manuscript, and agreed to be accountable for all aspects of the work.

Acknowledgement

We acknowledge the valuable support of Lili Yu in statistical consultation.

Data Availability

The aggregate findings and analyses are provided in the manuscript and supplementary tables. Access to patient-level imaging and clinical records is governed by institutional and ethical data-protection requirements.

References

1. Fehlings MG, Evaniew N, Ter Wengel PV, et al. AO Spine Clinical Practice Recommendations for Diagnosis and Management of Degenerative Cervical Myelopathy: Evidence Based Decision Making - A Review of Cutting Edge Recent Literature Related to Degenerative Cervical Myelopathy. Global Spine J. 2025;15(5):2585–2593. doi:10.1177/21925682251331050.
2. Verma A, Kumar A, Anand U, et al. Anterior Cervical Discectomy and Fusion vs. Anterior Cervical Corpectomy and Fusion for 2-Level Degenerative Cervical Myelopathy: A Systematic Review and Meta-Analysis. JBJS Rev. 2025;13(3). doi:10.2106/JBJS.RVW.24.00204.
3. Li J, Zhang J, Wang B, et al. Comparison of Titanium Mesh Cage, Nano-Hydroxyapatite/Polyamide Cage, and Three-Dimensional-Printed Vertebral Body for Anterior Cervical Corpectomy and Fusion. Spine (Phila Pa 1976). 2025;50(2):88–95. doi:10.1097/BRS.0000000000005126.
4. Leven D, Cho SK. Pseudarthrosis of the Cervical Spine: Risk Factors, Diagnosis and Management. Asian Spine J. 2016;10(4):776–786. doi:10.4184/asj.2016.10.4.776.
5. Oshina M, Oshima Y, Tanaka S, Riew KD. Radiological Fusion Criteria of Postoperative Anterior Cervical Discectomy and Fusion: A Systematic Review. Global Spine J. 2018;8(7):739–750. doi:10.1177/2192568218755141.
6. Lin W, Ha A, Boddapati V, Yuan W, Riew KD. Diagnosing Pseudoarthrosis After Anterior Cervical Discectomy and Fusion. Neurospine. 2018;15(3):194–205. doi:10.14245/ns.1836192.096.
7. Song KS, Piyaskulkaew C, Kim HJ, Mesfin A, Park SM, Riew KD. Anterior cervical fusion assessment using reconstructed computed tomographic scans: surgical confirmation of 254 segments. Spine (Phila Pa 1976). 2013;38(25):2171–2177. doi:10.1097/BRS.0000000000000017.

8. Riew KD, Yang JJ, Chang DG, et al. What is the most accurate radiographic criterion to determine anterior cervical fusion? Spine J. 2019;19(3):469–475. doi:10.1016/j.spinee.2018.07.003.
9. Song KS, Piyaskulkaew C, Chuntarapas T, et al. Dynamic radiographic criteria for detecting pseudarthrosis following anterior cervical arthrodesis. J Bone Joint Surg Am. 2014;96(7):557–563. doi:10.2106/JBJS.M.00167.
10. Balouch E, Burapachaisri A, Woo D, et al. Assessing Postoperative Pseudarthrosis in Anterior Cervical Discectomy and Fusion (ACDF) on Dynamic Radiographs Using Novel Angular Measurements. Spine (Phila Pa 1976). 2022;47(16):1151–1156. doi:10.1097/BRS.0000000000004375.
11. Wang H, Xia T, Qu R, et al. Interspinous Motion Measurement Could Serve as a Quantitative Method for Assessing Bony Fusion After Anterior Cervical Corpectomy and Fusion. Spine (Phila Pa 1976). 2025;50(3):E39–E45. doi:10.1097/BRS.0000000000005101.
12. Wang H, Qu R, Xia T, et al. Artificial vertebral body versus titanium mesh cage for single-level anterior cervical corpectomy and fusion (ACCF): a propensity score matching analysis of fusion, implant subsidence and clinical outcomes. BMC Musculoskelet Disord. 2025;26:1060. doi:10.1186/s12891-025-09308-1.
13. Dong C, Zhao Z, Zhou J, et al. Three-dimensional printed artificial vertebral body versus titanium mesh cage and nanohydroxyapatite/polyamide-66 cage in anterior cervical corpectomy and fusion: a retrospective cohort study with a minimum 5-year follow-up. J Orthop Surg Res. 2026;21:573. doi:10.1186/s13018-026-07076-9.
14. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351:h5527. doi:10.1136/bmj.h5527.
15. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016;15(2):155–163. doi:10.1016/j.jcm.2016.02.012.
16. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307–310. doi:10.1016/S0140-6736(86)90837-8.
17. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–845. doi:10.2307/2531595.
18. Martin CT, Yoon ST, Alluri RK, et al. How Reliable is the Assessment of Fusion Status Following ACDF Using Dynamic Flexion-Extension Radiographs? Global Spine J. 2025;15(4):2450–2457. doi:10.1177/21925682241303107.
19. Wang J, Zhou H, Hu B, et al. A modified measurement method for functional spinal unit ROM of cervical spine. Sci Rep. 2025;15:7877. doi:10.1038/s41598-025-92658-6.
20. Ploumis A, Mehbod A, Garvey T, Gilbert T, Transfeldt E, Wood K. Prospective assessment of cervical fusion status: plain radiographs versus CT-scan. Acta Orthop Belg. 2006;72(3):342–346.
21. Buchowski JM, Liu G, Bunmaprasert T, Rose PS, Riew KD. Anterior cervical fusion assessment: surgical exploration versus radiographic evaluation. Spine (Phila Pa 1976). 2008;33(11):1185–1191.

doi:10.1097/BRS.0b013e318171927c.

22. Bilancia G, Contartese D, Delbello F, et al. Accuracy and reliability of radiological methods for assessing fusion rates in patients undergoing spinal arthrodesis and stabilization: a systematic review of the past 10 years. Front Surg. 2025;12:1692887. doi:10.3389/fsurg.2025.1692887.
23. Ham DW, Choi YS, Yoo Y, Park SM, Song KS. Measurement of interspinous motion in dynamic cervical radiographs using a deep learning-based segmentation model. J Neurosurg Spine. 2023;39(3):329–334. doi:10.3171/2023.5.SPINE23293.
24. Como CJ, LeVasseur CM, Oyekan AA, et al. Standard of Care Cervical Spine Flexion/Extension Radiograph Measurements Do Not Predict Multiplanar Intervertebral Motion. J Biomech Eng. 2025;147(5):054502. doi:10.1115/1.4068076.
25. Leeflang MMG, Moons KGM, Reitsma JB, Zwinderman AH. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. Clin Chem. 2008;54(4):729–737. doi:10.1373/clinchem.2007.096032.
26. Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. J Clin Epidemiol. 2006;59(8):798–801. doi:10.1016/j.jclinepi.2005.11.025.
27. Hassanzad M, Hajian-Tilaki K. Methods of determining optimal cut-point of diagnostic biomarkers with application of clinical data in ROC analysis: an update review. BMC Med Res Methodol. 2024;24:84. doi:10.1186/s12874-024-02198-2.
28. Mei J, Wang Z, Tian X, Liu Q, Sun L, Liu Q. Risk Factors for Early Subsidence of 3D-Printed Artificial Vertebral After Anterior Cervical Corpectomy and Fusion. World Neurosurg. 2025;193:770–780. doi:10.1016/j.wneu.2024.10.067.
29. Qu R, Wang H, Xia T, et al. Risk factors of early subsidence of 3D-printed artificial vertebral body following single-level anterior cervical corpectomy and fusion (ACCF): retrospective study of 98 patients. J Orthop Surg Res. 2025;20:633. doi:10.1186/s13018-025-06056-9.
30. Abudouaini H, Wu T, Liu H, Wang B, Chen H. The predictive value of Hounsfield units for titanium mesh cage subsidence after anterior cervical corpectomy and fusion. Front Surg. 2022;9:1012364. doi:10.3389/fsurg.2022.1012364.
31. Lambrechts MJ, D'Antonio ND, Karamian BA, et al. What is the role of dynamic cervical spine radiographs in predicting pseudarthrosis revision following anterior cervical discectomy and fusion? Spine J. 2022;22(10):1610–1621. doi:10.1016/j.spinee.2022.04.020.

Tables

Table 1 to 5 are available in the Supplementary Files section.

Figures

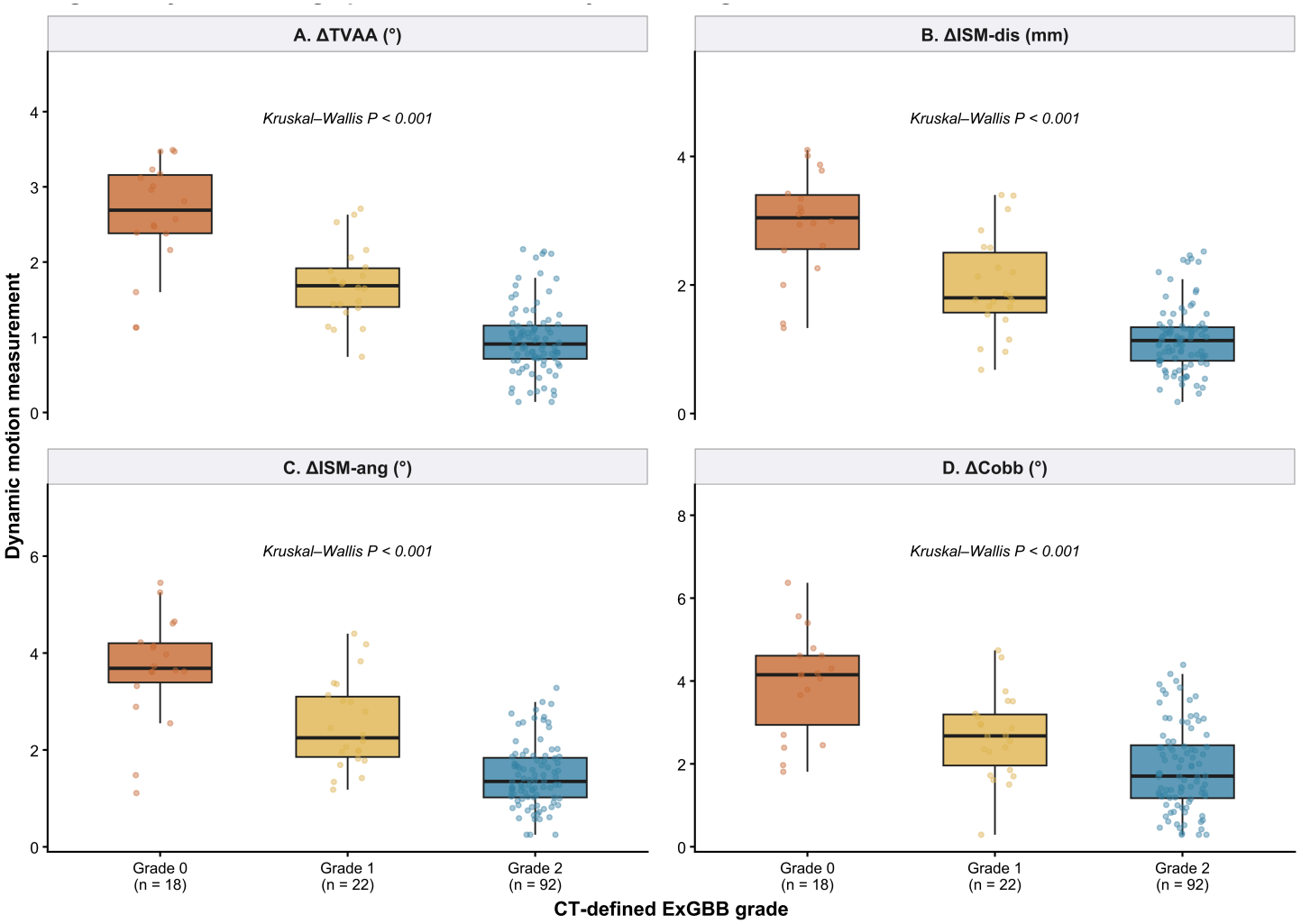

Figure 4

Dynamic radiographic measurements by CT fusion grade

Boxplots show (A) ΔTVAA, (B) ΔISM-dis, (C) ΔISM-ang, and (D) ΔCobb according to CT Grades 0 (n = 18), 1 (n = 22), and 2 (n = 92). Boxes indicate interquartile ranges and central lines medians; whiskers extend to the most extreme observations within 1.5 times the interquartile range, and points represent patient-level measurements. Overall group differences were assessed with Kruskal–Wallis tests (all P <0.001). Each panel uses the unit shown in its title. CT, computed tomography; TVAA, terminal vertebral axis angle; ISM, interspinous motion.

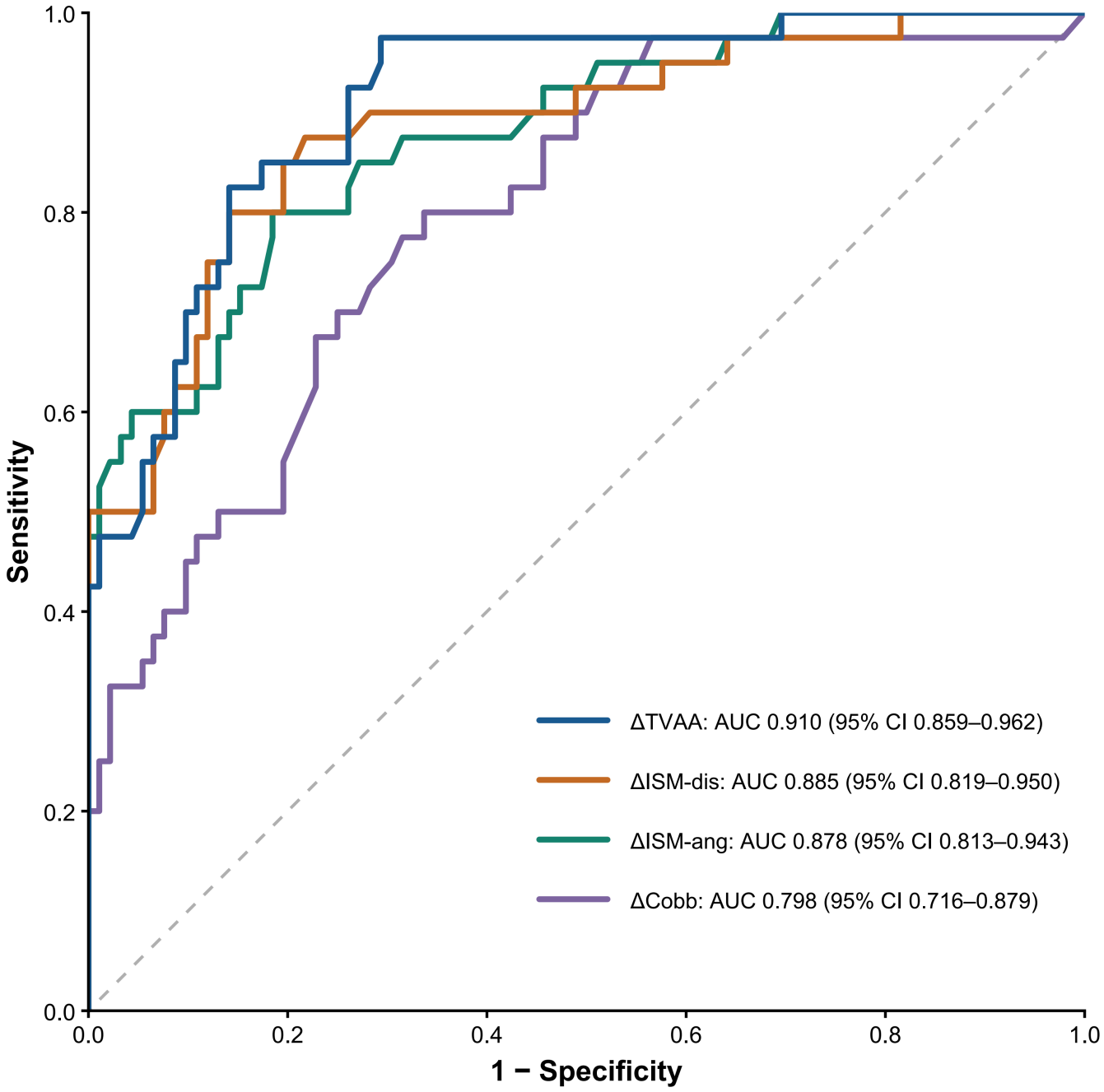

Figure 5

Diagnostic performance of dynamic radiographic measurements

ROC curves compare ΔTVAA (AUC 0.910; 95% CI 0.859–0.962), ΔISM-dis (0.885; 0.819–0.950), ΔISM-ang (0.878; 0.813–0.943), and ΔCobb (0.798; 0.716–0.879) for CT-defined non-complete fusion (Grades 0–1) versus complete fusion (Grade 2) in 132 paired patient observations. The gray diagonal indicates chance discrimination. AUC, area under the receiver operating characteristic curve; CT, computed tomography; TVAA, terminal vertebral axis angle; ISM, interspinous motion.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigureS1.png
- SupplementaryFigureS2.png
- SupplementaryTable14.docx
- Table15.docx