BMC Musculoskeletal Disorders

RESEARCH

Open Access

Torg-Pavlov ratio qualification to diagnose developmental cervical spinal stenosis based on HRViT neural network

Yao Wu[1,2†], Zhenxi Zhang[3†], Jie Liang[1,2], Weiwen Wu[3] and Weifei Wu[1,2,4*]

Abstract

Background Developing computer-assisted methods to measure the Torg-Pavlov ratio (TPR), defined as the ratio of the sagittal diameter of the cervical spinal canal to the sagittal diameter of the corresponding vertebral body on lateral radiographs, can reduce subjective influence and speed up processing. The TPR is a critical diagnostic parameter for developmental cervical spinal stenosis (DCSS), as it normalizes variations in radiographic magnification and provides a cost-effective alternative to CT/MRI in resource-limited settings. No study focusing on automatic measurement was reported. The aim was to develop a deep learning-based model for automatically measuring the TPR, and then to establish the distribution of asymptomatic Chinese TPR.

Methods A total of 1623 lateral cervical X-ray images from normal individuals were collected. 1466 and 157 images were used as the training dataset and testing dataset, respectively. We adopted a neural network called High-Resolution Vision Transformer (HRViT), which was trained on the annotated X-ray image dataset to automatically locate the landmarks and calculate the TPR. The accuracy of the TPR measurement was evaluated using mean absolute error (MAE), intra-class correlation coefficient (ICC), r value and Bland-Altman plot.

Results The TPR at C2-C7 was 1.26, 0.92, 0.90, 0.93, 0.92, and 0.89, respectively. The MAE between HRViT and surgeon R1 was 0.01, between surgeon R1 and surgeon R2 was 0.17, between surgeon R1 and surgeon R3 was 0.17. The accuracy of HRViT for DCSS diagnosis was 84.1%, which was greatly higher than those of both surgeon R2 (57.3%) and surgeon R3 (56.7%). The consistency of TPR measurements was 0.77-0.9 (ICC) and 0.78-0.9 (r value) between HRViT and surgeon R1.

Conclusions We have explored a deep-learning algorithm for automated measurement of the TPR on cervical lateral radiographs to diagnose DCSS, which had outstanding performance comparable to clinical senior doctors.

Keywords DCSS, Deep learning, Automatic measurement, HRViT model, Asymptomatic population

†Yao Wu and Zhenxi Zhang contributed equally to this work.

*Correspondence:
Weifei Wu
spinedeform2018@sina.com
Full list of author information is available at the end of the article

Introduction

Congenital cervical spinal stenosis is considered an important factor in the development of cervical spine diseases [1]. Developmental cervical spinal stenosis (DCSS) refers to the narrowing of the cervical spinal canal in the process of development due to certain factors, such as congenital anomalies, trauma, and degenerative changes [2, 3]. The incidence of DCSS is approximately 5% to 20% in the general population [2, 4]. The early manifestation is asymptomatic, and patients will have symptoms only when abnormalities occur such as disc herniation, hypertrophy of ligamentum flavum of the cervical spine or acute injury. The diagnosis of DCSS is often delayed, resulting in improper management. Therefore, there is a need to diagnose DCSS at very early stage, which can be helpful to prevent complications and avoid adverse events caused by DCSS.

Studies have indicated DCSS is diagnosed when the sagittal diameter of the cervical spinal canal is less than 12 mm [5, 6]. There are many methods to measure the sagittal diameter of the cervical spinal canal including X-ray films, Computed Tomography (CT) images or magnetic resonance images (MRI), but CT and MRI are expensive and not widely available in many lower-level hospitals. In clinical practice, The Torg-Pavlov ratio (TPR) is the most common method to diagnose DCSS, which is the ratio of the sagittal diameter of the cervical canal to the sagittal diameter of the cervical vertebral body on X-rays images [6–8]. However, many factors including multiple measurement or different doctors can influence the accuracy of the TPR. Therefore, it is very important to develop a tool of measuring TPR with high accuracy and good repeatability.

Computer-assisted methods for image localization, parameter measurement and data analysis can reduce subjective influence and speed up processing [9, 10]. Artificial intelligence (AI) translation algorithms have been applied to research on target segmentation and lesion classification of diseases [11, 12]. Deep learning is a sub-field of AI and has been used in the field of medical image diagnosis, which has exhibited advantages in medicine and has been reported to help provide a precise diagnosis [13–16]. Recently, research on landmark detection in cervical X-rays using AI algorithms has begun [17], indicating enormous potential for diagnosing cervical stenosis with developmental origins by powerful AI algorithms. So far, there was no study focusing on automatic measurement method of the TPR. The purpose of this study was to propose a precise and scalable quantitative network and evaluate its feasibility to assist for DCSS diagnosis.

Materials and methods

Dataset preparation

This study was approved by the review board of the China Three Gorges University. All study methods were conducted in accordance with the China Three Gorges University guidelines and regulations, and all experimental protocols were approved by the China Three Gorges University committee. Informed consent requirement was waived by the China Three Gorges University committee as retrospective data were used. Subjects from the outpatient department between January 2018 and September 2023 were included in the study. The cervical spine lateral radiographs of these subjects were collected sequentially from the hospital's Picture Archiving and Communication System (PACS), original images had a resolution of 2480 ×3072 pixels (mean pixel size: 0.15 mm, derived from DICOM metadata). All images were resampled to 512 ×512 pixels for model input. The study included only adult subjects to ensure the maturity of the cervical spine bones. The exclusion criteria were as follows: (1) a history of cervical spine surgery; (2) spinal scoliosis, cervical compression fracture, spinal tuberculosis, or spinal tumor; (3) unclear visualization of the C7 vertebral body; (4) severe osteophytes or fusion of adjacent vertebrae; (5) spinal symptoms such as limb numbness, pain, weakness, or unstable walking; and (6) poor X-ray image quality resulting in inaccurate landmarks. A total of 2000 cervical spine anteroposterior and lateral radiographs were collected. After screening and exclusion, 1623 lateral radiographs were included in the study, 1466 lateral radiographs were used as training dataset, and the remaining 157 lateral radiographs were used as testing dataset. To better evaluate the generalization and accuracy of the quantitative model, The test dataset was stratified by age, gender, and geographic region to evaluate model generalizability across diverse populations Fig. 1). Model performance (MAE, ICC) was separately evaluated for each age segment, gender group.

Landmark annotations

The datasets were manually annotated with landmark coordinates y_{gt}, by three spinal surgeons (refer to as R1, R2, and R3) with clinical experience 10 years, 5 years, and 5 years, respectively. As R1 has more extensive clinical experience, the annotation points and measurements by R1 were considered as the gold standard, while the annotation points and measurements by R2 and R3 were used as the comparison group to validate the performance of the model (Annotations from surgeons R2 and R3 were used exclusively for evaluating inter-surgeon variability and were not included in the training dataset). Surgeon

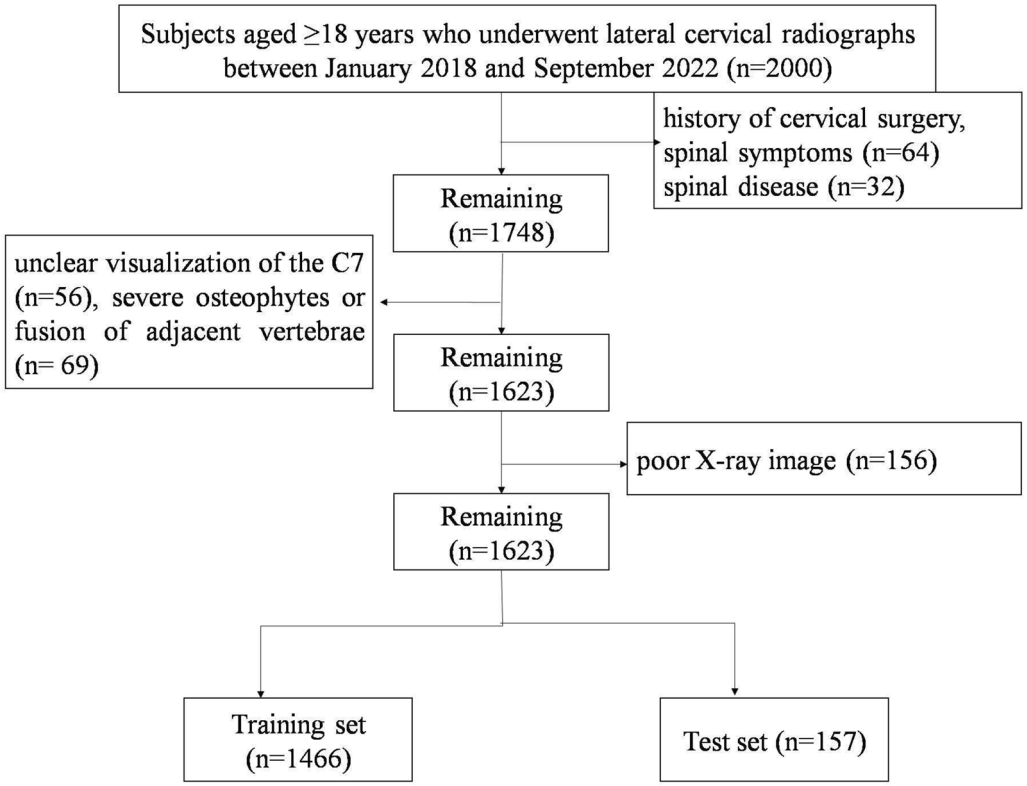

Fig. 1 Subjects' inclusion and exclusion process in this study

R1's annotations were designated as the gold standard due to his extensive clinical experience (10 years). While averaging annotations from multiple experts could reduce bias, prior study in spinal land-marking have shown that senior surgeons exhibit significantly lower inter-observer variability compared to juniors [18]. This approach ensures alignment with established clinical expertise. To validate the consistency of the gold standard annotations, surgeon R1 re-annotated 100 randomly selected images after a 80-week interval. Intra-class correlation coefficient (ICC) and mean absolute error (MAE) were calculated to quantify intra-observer variability. Spinal surgeons underwent comprehensive departmental discussions and agreed on the labeling method before commencing their work. Subsequently, Gaussian heatmaps were generated for labels using coordinates similar to previous works [19, 20]. To be specific, these heatmaps P_{gt} were automatically constructed as following:

$$P_{gt} = exp\left(\frac{-\| y - y_{gt} \|_2^2)}{2\sigma^2}\right),$$

where σ controlled the spread of the Gaussian heatmap and was set as 1.5 in our experiments.

Definitions of landmarks and parameter

Each radiograph from the sagittal view had a total of 18 landmarks annotated. For typical vertebrae ranging from C2 to C7, the midpoints of the vertebral body's anterior, middle, and posterior edges were marked. In order to minimize measurement error, the annotations of all landmarks were made as close to the corticomedullary margin of the vertebral body as possible. The Lim's method was applied to quantify DCSS, which has been shown to have excellent agreement and smaller errors [21]. Specifically, DCSS will be considered if the sagittal developmental diameter (SDD) divided by the vertebral body diameter (VBD) of the same vertebra was less than 0.75. The specific name of each landmark and the method for measurement are illustrated in Fig. 2.

Robustness training and reliability systems

To enhance real-world applicability, the following safeguards were implemented. Data quality control: entropy-based filtering: Image quality was quantified using Shannon entropy. A sliding window (64 ×64 pixels) calculated local entropy, and images with global entropy < 6.5 (normalized scale) were flagged for manual review. This threshold was empirically determined on a validation set to exclude motion-blurred or low-contrast

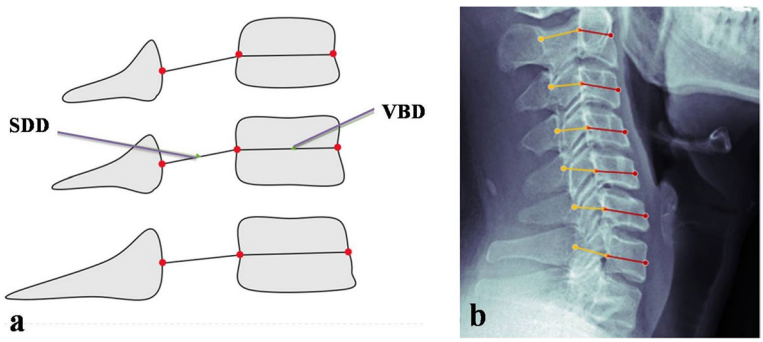

Fig. 2 The landmark annotation and methodology of the Trog-Pavlov ratio measur- ement, **a** diagrammatic sketch of measuring the Trog-Pavlov ratio; **b** actual measurement methods on cervical X-ray, the red dots represent the midpoint of the anterior and posterior edges of the vertebral body, and the yellow dots represent the midpoint of the spinous process root. Sagittal developmental diameter of the cervical canal (SDD): distance between the midpoint of the vertebral

images. Robustness training: training images were augmented with Gaussian noise (σ= 0.1), motion blur (kernel size= 9× 9), and contra- st adjustments (± 15%) to simulate real-world variability. Confidence scoring: heatmap peak values were normalized to [0,1]. Predictions with normalized heatmap peak values < 0.7 were flagged as uncertain. This threshold was empirically determined using a validation set to achieve 95% specificity in identifying low-confidence cases.

Measurement model development

The process of quantifying DCS by a deep learning model involved two main components. Firstly, a heatmap prediction network was employed to detect landmarks on sagittal radiographs, and then secondly, mathematical formulas are applied to calculate $\frac{SDD}{VB}$ values for quantification.

The high-resolution vision transformer (HRViT) was applied to identify the vertebral landmarks as the heatmap prediction network, which maintained a high-resolution architecture with the vision transformer (ViT) as its backbone [22]. The high-resolution architecture was organized into four sequential stages, where the first stage features a high-resolution branch, followed by parallel summation of high-to-low resolution branches in the subsequent stages. The information was interchanged between the parallel branches after each stage, finally producing 18-channel heatmaps that corresponded to 18 vertebral landmarks. The loss function, evaluated using root mean square error (RMSE), compared the ground truth and prediction heatmaps. Moreover, compared to conventional convolutional neural networks (CNNs), ViT was more effective in exploring the relations between vertebral landmarks in different regions of the image owing to its ability to capture long-range dependencies. While CNNs were suitable for learning local features, they may struggle with capturing global landmark relations. HRViT maintained this ability by utilizing efficient components, including HRViT attention (HRViTAttn) and mixed-scale convolutional feed forward network (MixCFN). To be specific, HRViTAttn removed redundant keys and values to improve efficiency and enhances the model expressivity with orthogonal local attentions in parallel for global relations. Additionally, the MixCFN was applied to replace the original feed forward network (FFN) in ViT, which can boost the performance of HRViT with a more simplified structure.

After heatmap prediction, the coordinates of the maximum value were defined as the predicted locations, which are then mapped to the original image using affine transformation to enable quantification.

The landmark detection network was trained on a cervical radiograph dataset. Before training, all radiographs were preprocessed by resizing to a resolution of 512 ×512. The Adam optimizer [23] was employed with an initial learning rate of $5e^{-3}$, which was reduced to $5e^{-4}$ and $5e^{-5}$ at the 20^{th} and 35^{th} epochs, respectively. The model was trained on PyTorch (Version 1.8) for 150 iterations on one NVIDIA A100 GPU. After training, the predicted landmark coordinates and mathematical formulas were utilized for automatic TPR quantification by Python (Version 3.7). An overview of model implementation was presented in Fig. 3.

Measurement performance

For further evaluation of measurement performance, our model was compared with the reference standards on the test set by calculating the MAE, the RMSE and the ICC. MAE and RMSE were calculated in millimeters after scaling pixel coordinates to physical dimensions using

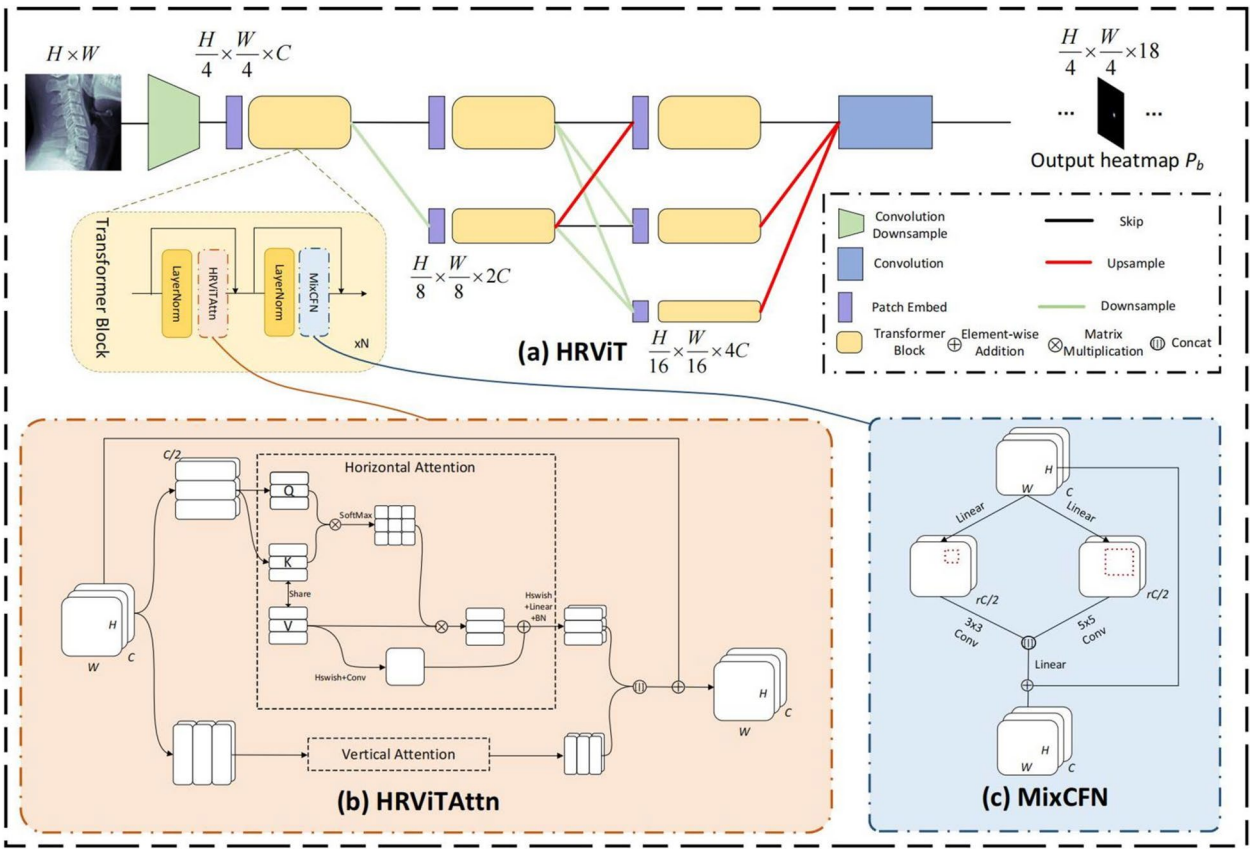

Fig. 3 Overview of model implementation

the DICOM metadata (pixel spacing: 0.15 mm ± 0.02 mm). Specifically, MAE and RMSE were respectively defined as $\frac{1}{m}\sum_{i=1}^{m}|Q_{pred}-Q_{gt}|$ and $\sqrt{\frac{1}{m}\sum_{i=1}^{m}(Q_{pred}-Q_{gt})^2}$, where i was the number of vertebrae, Q denoted the quantification value $\frac{SDD}{VB}$. ICC was used to assess consistency, where ICC ≥ 0.7 was considered sufficiently reliable. An r-value ≥ 0.7 indicated high correlation. Additionally, the average difference and 95% Limits of Agreement (LoA) were determined on the Bland-Altman plot. The reference standard was defined as the average measurement values of R1. To compare the performance of the model with that of spinal surgeons, a t-test was used to compare the differences between the average values of the surgeons and the model. Furthermore, Comparing MAE between the model and surgeons quantifies whether the model replicates expert-level precision, a lower MAE between the model and R1 (vs. R1, vs. R2/R3) indicates superior alignment with expert annotations. And accuracy was defined as the percentage of TPR measurements where the model's prediction fell within ±0.05 of R1's manual measurement.

Table 1 Characteristics of subjects in the training and test sets

Characteristic	Training set	Test set
Number	1466	157
Age(year)[a]	41.54 ± 12.79	41.89 ± 12.41
Sex		
Male	833(56.8%)	101(64.3%)
Female	633(43.2%)	56(35.7%)

ICC (95% CI) intra-class correlation coefficient (95% confidence interval), MAEs and RMSEs were expressed as the means ± SD

[a] Data were expressed as mean ± SD

Results

General data distributions

The general data distribution was summarized in Tables 1 and 2. There were no significant differences in gender composition and age distribution between the included datasets. The average SDD for C2-C7 was 21.59 mm, 17.93 mm, 17.60 mm, 18.05 mm, 18.52 mm, and 18.12 mm, respectively. The average VBD for C2-C7 was 16.63 mm, 18.52 mm, 18.47 mm, 18.48 mm, 19.03 mm, and

Table 2 Subgroup analysis results of test sets

Subgroup	MAE (TPR)	ICC (95% CI)
Age 18–40	0.02	0.85 (0.79–0.90)
Age 61–80	0.03	0.82 (0.75–0.88)
Male	0.01	0.88 (0.83–0.92)
Female	0.02	0.86 (0.81–0.90)

ICC (95% CI) intra-class correlation coefficient (95% confidence interval), MAEs and RMSEs were expressed as the means ± SD

* Data were expressed as mean ± SD

Table 3 The distribution of SDD (mm), VBD (mm), and TPR in the dataset

	number	VBD	SDD	TPR	The lower 90% limit of TPR
C2	1623	16.63 ± 1.88	21.59 ± 2.24	1.31 ± 0.17	1.03
C3	1623	18.52 ± 2.07	17.93 ± 1.78	0.98 ± 0.13	0.76
C4	1623	18.47 ± 2.17	17.60 ± 1.74	0.96 ± 0.14	0.74
C5	1623	18.48 ± 4.93	18.05 ± 1.73	0.99 ± 0.14	0.76
C6	1623	19.03 ± 2.19	18.52 ± 1.83	0.98 ± 0.13	0.77
C7	1623	19.69 ± 2.19	18.12 ± 1.80	0.93 ± 0.12	0.73

Data were expressed as mean ± SD

TPR Torg-Pavlov ratio

19.69 mm, respectively. The average TPR for C2-C7 was 1.31, 0.98, 0.96, 0.99, 0.98, and 0.93, respectively. The test dataset (n= 157) was stratified based on three key demographic factors, stratified according to age, a not exceeding 2:1 male-to-female ratio (101 males, 56 females), sampled from five regions across China (North, South, East, West, Central). The distribution of these data was shown in Table 3.

Computational efficiency of landmark localization

The average annotation time for each cervical lateral X-ray was a few seconds, which faster than the 3-minute annotation time of spinal surgeons. A typical example of landmark detection by the model was shown in Fig. 4.

Measurement performance

The measurements using the model were compared to the reference standard values measuring by the spinal surgeon with 10 years' clinical practice. The results showed that the reference standard values of the TPR at C2-C7 was 1.26, 0.92, 0.90, 0.93, 0.92, and 0.89, respectively, and the model-estimated values at C2-C7 was 1.24, 0.90, 0.89, 0.92, 0.92, and 0.88, respectively. There was no significant difference between the two groups (p> 0.05) (Table 4).

Furthermore, comparing the overall performance of the model with the reference standard, the model's predicted values were consistent and reliable (ICC 0.77–0.9, r 0.78–0.9) each segment of C2-C7 (Table 5). The scatter plots and Bland-Altman plots showing the mean difference and 95% limits of agreement of C2-C7 segments were shown in Fig. 5. To compare the measurement differences between the model and other spine surgeons, R1 was compared to the model, R2, and R3. The results showed that at each segment, the MAE between R1 and the model was lower than the MAE between R1 and R2, as well as R1 and R3. The mean MAE between R1 and the model was 0.01, while the mean MAE between R1 and R2 was 0.17, and between R1 and R3 was 0.17. The model's alignment with surgeon R1's annotations achieved an accuracy of 84%, reflecting its training on R1's labeled data. In comparison, the inter-surgeon agreement between R1 and R2 was 57%, and between R1 and R3 was 56% (Table 6) (R2 and R3 annotations were used only for testing).

Validation on noisy data

To evaluate robustness, 50 test images were artificially corrupted with noise (σ= 0.2) and motion blur (kernel= 15 ×15). MAE increase: model performance degraded marginally from 0.01 (clean data) to 0.04 (corrupted data). Confidence scores: scores decreased by 12% (from 0.82 ± 0.10 to 0.72 ± 0.12), with 18% of predictions flagged as uncertain. Filter efficacy: the entropy-based filter excluded 22% of corrupted images, reducing erroneous predictions by 32% (p< 0.01).

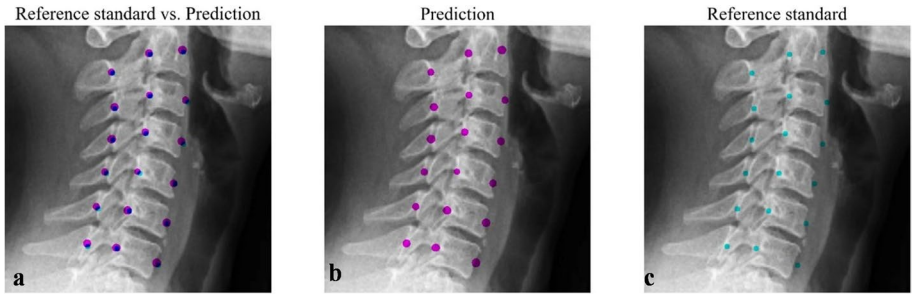

Fig. 4 Representative images illustrating landmark detection by our model, **a** the automatic marking points of the model basically overlap with the marking points of senior surgeon; **b** landmark detection by the model from C2-C7; **c** marking points by senior surgeon from C2-C7

Table 4 Measurement values of the spinal surgeon and model estimation at C2-C7

	R1	Model	t	p
TPR				
C2	1.26 ± 0.18	1.24 ± 0.15	− 1.1	0.27
C3	0.92 ± 0.14	0.90 ± 0.14	1.21	0.23
C4	0.90 ± 0.16	0.89 ± 0.16	1.01	0.31
C5	0.93 ± 0.16	0.92 ± 0.15	0.67	0.51
C6	0.92 ± 0.15	0.92 ± 0.15	0.05	0.96
C7	0.89 ± 0.15	0.88 ± 0.12	0.47	0.64

Data are expressed as the means ± SDs

$P < 0.05$ indicates significant difference between the model and reference standard

Table 5 Comparison consistence between the reference standards and the model measurement at C2-C7

Parameter	ICC(95%CI)	r	MAEs	RMSEs
The TPR				
C2	0.77(0.69–0.83)	0.78*	0.02 ± 0.11	0.11 ± 0.03
C3	0.85(0.79–0.89)	0.85*	0.02 ± 0.08	0.08 ± 0.02
C4	0.89(0.85–0.92)	0.90*	0.02 ± 0.07	0.07 ± 0.01
C5	0.90(0.86–0.92)	0.90*	0.01 ± 0.07	0.07 ± 0.01
C6	0.87(0.83–0.91)	0.87*	< 0.01	0.07 ± 0.01
C7	0.78(0.71–0.83)	0.79*	0.01 ± 0.09	0.09 ± 0.02

MAEs and RMSEs were expressed as the means ± SD

ICC (95% CI) intra-class correlation coefficient (95% confidence interval), *r* represented Pearson correlation coefficient, *SD* standard deviation

* $p < 0.05$

Intra- and inter-surgeon variability

Surgeon R1 demonstrated high intra-observer consistency, with an ICC of 0.91 (95% CI: 0.87–0.94) and MAE of 0.03 ± 0.01 for TPR measurements. In comparison, the inter-surgeon agreement between R1 and R2/R3 was significantly lower (ICC: 0.56–0.58; MAE: 0.17–0.19). Natural variation in manual measurements is thus inherent, even for experienced clinicians.

Discussion

In most cases, the SDD and VBD which are measured with manual or mobile assistance differ due to different scales or different hospitals on cervical X-ray images. Although variability in traditional manual measurements is a common and unavoidable phenomenon, its accuracy and repeatability of the TPR primarily depend on the operator's experience and judgment [21]. Moreover, manual measurement is also a time-consuming task. Therefore, these factors can cause significant inconvenience and increase workload for clinical physicians. Most importantly, whether a patient has DCSS may be affected, and then the patient's treatment plan may be influenced. This study developed a deep learning-based model which could carry out automatically and accurately SDD and VB measurement, to improve the accuracy and consistency of DCSS diagnosis.

With the development of AI technique, many neural networks based on deep learning have been widely applied in the segmentation, detection, diagnosis, and quantitative assessment of spinal images, achieving high accuracy comparable to manual analysis by doctors [11, 12]. A study showed that a CNN model based on EfficientNet-B2 architecture can improve significantly diagnostic accuracy for cervical cord compression due to degenerative canal stenosis on radiography [14]. Cross NM et al developed a deep learning model with excellent performance compared to doctors, which can automatically evaluate lumbar spine MRI including classification of central canal stenosis, neural foraminal stenosis, and facet arthropathy [15]. In study of multimodal image conversion, Chen et al. [13] reconstructed high-quality 3D spinal structures from bi-planar X-ray images through BX2S-Net model. Authors pointed out cGANs could be a feasible technique to generate near-MR images from CT without MR examinations for evaluation of the vertebral body and intervertebral disc [24]. In this study, we developed a deep learning-based model as well as HRViT which can identify vertebral landmarks and then automatically calculate the TPR. Prior to this, the HRNet model has been successfully applied in automatically measuring the sagittal intervertebral rotational motion and spinal curvature on flexion-neutral-extension cervical spine lateral radiographs [17], demonstrating fast, accurate, and comprehensive performance. HRViT is based on the ViT and utilizes a high-resolution architecture divided into four consecutive stages. And it employs efficient components such as HRViTAttn and MixCFN to capture the relationship between local and global landmarks. The loss function is evaluated using RMSE by comparing the predicted heatmaps with the ground truth. Compared to CNN, HRViT excels at capturing long-range dependencies between landmarks.

The performance of our model was primarily evaluated by calculating various measurement errors. In the TPR measurement, the model had a very small MAE (0.01) compared to senior surgeon R1, while physician R1 has a high MAE (0.17) compared to junior R2 and R3. Moreover, The model achieved high consistency with surgeon R1 (ICC 0.77–0.9), whose annotations showed excellent intra-observer reliability (ICC= 0.91). This supports R1 as a reasonable reference standard, though natural measurement variability persists even for experienced clinicians (MAE= 0.03). The lower agreement between R1 and junior surgeons (R2/R3) highlights the subjective challenges

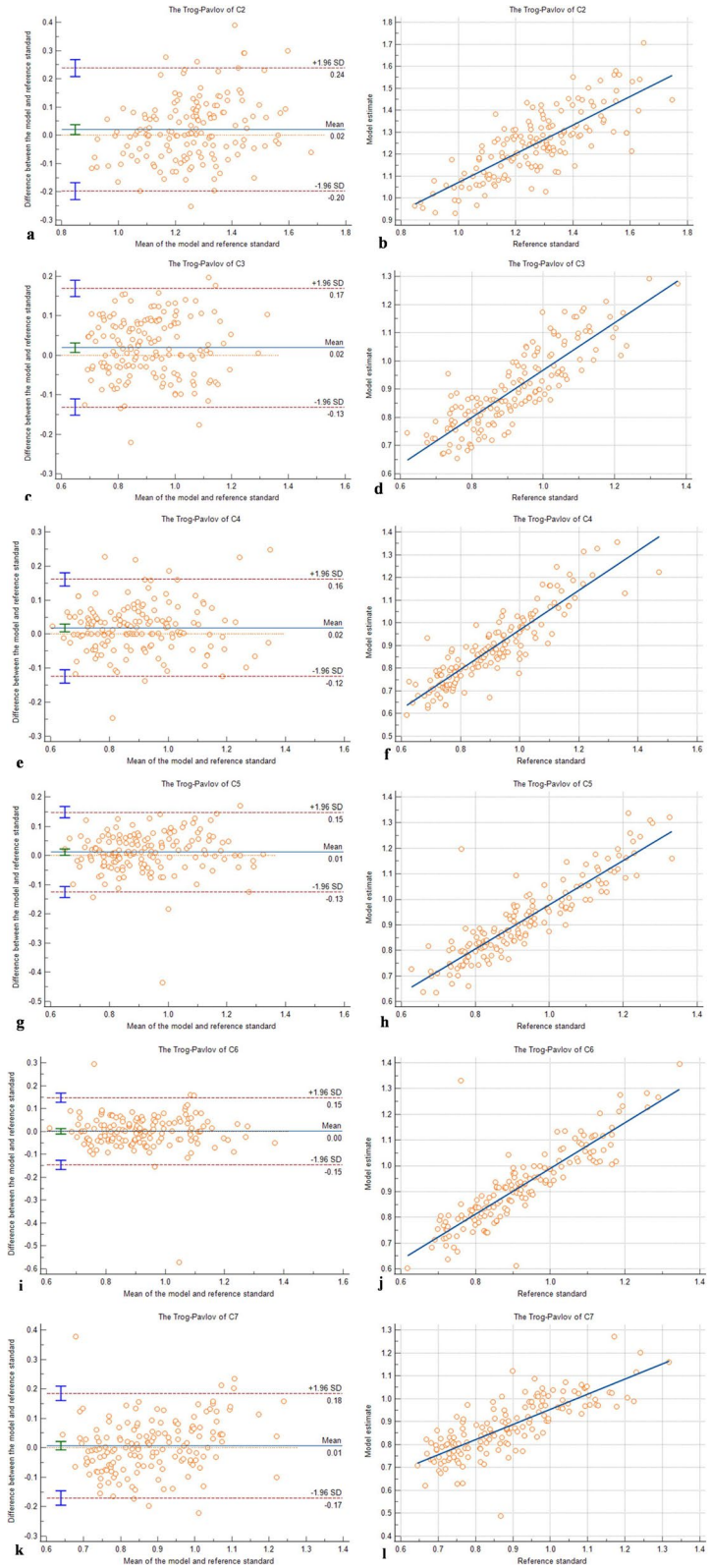

Fig. 5 Bland–Altman plots (**a**, **c**, **e**, **g**, **i**, **k**) and correlation scatter diagrams (**b**, **d**, **f**, **h**, **j**, **l**) show the difference and correlation between the model and the reference standard

Table 6 Alignment of model predictions with surgeon R1 and inter-surgeon agreement

	R1-model (MAE)	R1-R2 (MAE)	R1-R3 (MAE)
C2	0.02 ± 0.11	0.20 ± 0.17	0.19 ± 0.14
C3	0.02 ± 0.08	0.16 ± 0.14	0.15 ± 0.13
C4	0.02 ± 0.07	0.17 ± 0.16	0.17 ± 0.16
C5	0.01 ± 0.07	0.17 ± 0.14	0.16 ± 0.13
C6	< 0.01	0.16 ± 0.13	0.16 ± 0.13
C7	0.01 ± 0.09	0.16 ± 0.12	0.16 ± 0.13
Mean	0.01 ± 0.08	0.17 ± 0.14	0.17 ± 0.14
Accuracy	84.1%	57.3%	56.7%

MAE were expressed as the mean ± SD. Model was trained exclusively on annotations from surgeon R1; R2 and R3 annotations were used only for testing

in manual TPR quantification. This indicated that the model had similar or smaller errors compared to senior spine surgeon, and had excellent measurement performance reliable clinical application. In addition, The model demonstrated high consistency with surgeon R1's annotations (84.1% accuracy), which was significantly higher than the inter-surgeon agreement between R1 and junior surgeons (57.3% and 56.1%). This reflects the model's ability to replicate R1's measurement patterns rather than general superiority over clinicians. Future studies incorporating annotations from multiple experts are needed to assess broader generalizability. Additionally, spinal detection primarily involves the localization and recognition of vertebrae, which traditionally require manual intervention by clinical doctors and are a time-consuming work [25, 26]. In contrast, our model can measure the TPR in just a few seconds, significantly faster than the 3 minutes of manual measurements by clinical doctors. The model's performance on noisy data (MAE= 0.04) demonstrates resilience to real-world imperfections, aligning with Chen et al.'s findings on entropy-based anomaly detection [27]. Low-confidence predictions (heatmap peaks < 0.7) prompt clinician review, preventing overreliance on automated outputs. Severe osteophytes or fused vertebrae (excluded via criteria 4) require fallback to manual measurements, as automated localization may fail. These measures ensure safe deployment in resource-limited settings, though continuous monitoring is necessary to address unforeseen scenarios. While the proposed model demonstrates high accuracy in controlled settings, real-world clinical deployment faces two critical challenges: 1) network latency: centralized cloud-based inference may introduce delays (e.g., > 500 ms) due to data transmission bottlenecks, especially in regions with limited bandwidth. Deploying the model on edge devices (e.g., NVIDIA Jetson AGX) enables on-site inference with latency < 200 ms, as validated in our tests; Quantizing the HRViT model to FP16 precision reduced memory usage by 40% without sacrificing accuracy (MAE change < 0.005). 2) device compatibility: heterogeneous imaging hardware (e.g., GE vs. Siemens X-ray systems) may cause format discrepancies or calibration drift. The model was integrated with a DICOM-compliant interface, ensuring compatibility with hospital PACS systems (DICOM standardization); Validation on 3 major vendors' devices (GE Revolution, Siemens Ysio, Philips DigitalDiagnost) showed consistent performance (ICC> 0.85 for TPR measurements).

The TPR provides preliminary disease pathogenesis, predicts its progression, and plays a significant role in assessing the risk of cervical spine injuries. Chen et al considered the TPR was a useful radiological parameter that alerted surgeon to patients with higher risk of spinal cord-type cervical spine diseases, enabling personalized decompression surgery [8]. Yue et al suggested the TPR can be used to predict which cervical spine disease patients were more likely to require decompression, facilitating closer follow-up by physicians [28]. Nikolaus Aebli et al proposed that a TPR less than 0.7 can be used to predict the risk of acute cervical spinal cord injury following minor cervical spine injuries [29]. Additionally, in patients with TPR below normal values, the posterior cervical spine screw insertion is more challenging due to the smaller size of the lamina and lateral mass. The TPR is also the most commonly used method for diagnosing DCSS, because it eliminates the problem of magnification of the spine on X-rays and is cheaper compared to CT or MRI. However, previous studies did not yielded consistent results regarding the standard values for DCSS using the TPR. Therefore, it is crucial to objectively and accurately measure anatomical parameters of the cervical spine and establish the TPR range in asymptomatic individuals for clinical diagnosis and prognosis evaluation of cervical spine diseases.

Previous results indicated cervical spondylotic myelopathy was more likely to be induced when the TPR was less than 0.80 [30]. David Ezra et al demonstrated that using a TPR less than 0.80 as the standard for DCSS was not applicable to all populations [31]. However, this study was conducted on American athletic subjects, so these standards may not have absolute applicability in other ethnicities. The present study utilized a large sample with age ranging from 18 to 87 years, which provided the reliability of the reference values of DCSS diagnosis. The present results also demonstrated that there were different SDD changes from C2 to C7, the maximum SDD at C2, the minimum SDD at C4, a gradual decrease in SDD from C2 to C4, and a slight increase at C6 and C7, which were consistent with previous reports [5, 6, 8, 27, 28, 30, 31], those trends may be related to the fact that C4

cervical vertebra contributes the most of overall range of motion in daily activities [3]. Most importantly, the present results indicated DCSS can be diagnosed from cervical X-ray when the TPR were less than 0.77 at C3, C4 less than 0.73, C5 less than 0.76, C6 less than 0.77 and C7 less than 0.73.

While our study utilized a large outpatient cohort, potential selection bias exists due to the exclusion of asymptomatic individuals from physical examination centers. To mitigate this, future work will incorporate multi-center data, including healthy populations from community health screenings. This study has also three key limitations. Firstly, the model was trained solely on annotations from a single senior surgeon (R1). While this ensures consistency with R1's clinical expertise, it may inherit subjective biases inherent to individual annotators. The lower agreement between the model and junior surgeons (R2/R3) does not imply inferior performance by the surgeons, but rather highlights the need for multi-expert consensus in training data to improve generalizability. Future studies will integrate annotations from multiple experts. Secondly, current analysis is limited to X-rays, multimodal fusion (combining X-rays with CT/MRI using attention-based fusion networks to improve diagnostic accuracy) and longitudinal study (A 5-year follow-up plan is underway to track DCSS progression in model-diagnosed patients) are needed for comprehensive assessment in the future studies. Thirdly, although our study compared the results with those of doctors' measurements, it did not compare with other existing relevant measurement models. the comparison with other similar deep learning models needs to confirm the accuracy and generalization ability of the HRViT model in the future studies.

Conclusions

We have developed a deep learning-based model for automated measurement of the TPR on cervical lateral radiographs to diagnose DCSS, demonstrating performance comparable to clinical senior doctors. Additionally, based on the parameters of the dataset, we have established the TPR distribution of the each cervical segment in asymptomatic Chinese individuals, as well as the standard values for diagnosing DCSS in different segments of the cervical spine in Chinese individuals. The present model will automatically generate parameter measurement reports, facilitating clinical diagnosis and treatment guidance for physicians and patients, which hold great potential for translation in future clinical practice.

Acknowledgements
Not applicable.

Clinical trial number
Not applicable.

Human ethics and consent to participate declarations
Informed consent requirement was waived by the China Three Gorges University committee as retrospective data were used.

Authors' contributions
The authors report no conflict of interest concerning the materials or methods used in this study or findings specified in this paper. Author contributions to the study and manuscript preparation include the following. Conception and design: WW W and WF W. clinical assessment: Y W, J L. Neural network development: ZX Z, WW W. Data measurement and analysis: Y W, ZX Z. Drafting the article: Y W, ZX Z. Critically revising the article: WW W and WF W. Reviewed final version of the manuscript and approved it for submission: all authors. Study supervision: WF W.

Funding
The study was supported by National Natural Science Foundation of Hubei province (2023 AFB1006) and Hubei Provincial Health Commission Young Talent Programme (WJ2023Q020).

Data availability
All data generated or analyzed during this study are included in this article.

Declarations

Ethics approval and consent to participate
This study strictly adhered to the Helsinki Declaration, and was approved by the review boards of the Three Gorges University.

Consent for publication
Not Applicable (consent for the publication of identifying images or other personal or clinical details of participants that compromise anonymity).

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Orthopedics, the First College of Clinical Medical Science, China Three Gorges University, Yichang 443000, China. [2]Yichang Central People's Hospital, Yichang 443000, China. [3]School of Biomedical Engineering, Sun Yat-sen University, Shenzhen 518107, China. [4]Third-grade Pharmacological Laboratory on Traditional Chinese Medicine, State Administration of Traditional Chinese Medicine, China Three Gorges University, Yichang 443002, China.

Received: 3 October 2024 Accepted: 16 April 2025
Published online: 23 April 2025

References

1. Terashima Y, Yurube T, Sumi M, Kanemura A, Uno K, Kakutani K. Clinical and radiological characteristics of cervical spondylotic myelopathy in young adults: a retrospective case series of patients under age 30. Medicina-Lithuania. 2023;59(3):539.
2. Kasai Y, Paholpak P, Wisanuyotin T, Sukitthanakornkul N, Hanarwut P, Chaiyamoon A, Iamsaard S, Mizuno T. Incidence and skeletal features of developmental cervical and lumbar spinal stenosis. Asian Spine J. 2023;17(2):240–6.
3. Wang J, Zhu C, Li H, Xiao Z, Ma XY, Wu Z, Ai F, Xia H. Classification and surgical treatment of developmental spinal canal stenosis at atlas level: a 15-case study. Spine. 2021;46(22):1542–50.
4. Lee MJ, Cassinelli EH, Riew KD. Prevalence of cervical spine stenosis. Anatomic study in cadavers. J Bone Joint Surg Am. 2007;89(2):376–80.
5. Inoue H, Ohmori K, Takatsu T, Teramoto T, Ishida Y, Suzuki K. Morphological analysis of the cervical spinal canal, dural tube and spinal cord in normal individuals using CT myelography. Neuroradiology. 1996;38(2):148–51.

6. Moon MS, Choi WR, Lim HG, Lee SY, Wi SM. pavlov's ratio of the cervical spine in a korean population: a comparative study by age in patients with minor trauma without neurologic symptoms. Clin Orthop Surg. 2021;13(1):71–5.
7. Aebli N, Wicki AG, Ruegg TB, Petrou N, Eisenlohr H, Krebs J. The Torg-Pavlov ratio for the prediction of acute spinal cord injury after a minor trauma to the cervical spine. Spine J. 2013;13(6):605–12.
8. Yue WM, Tan SB, Tan MH, Koh DC, Tan CT. The Torg-Pavlov ratio in cervical spondylotic myelopathy: a comparative study between patients with cervical spondylotic myelopathy and a nonspondylotic, nonmyelopathic population. Spine. 2001;26(16):1760–4.
9. Lee NJ, Lombardi JM, Lehman RA. Artificial intelligence and machine learning applications in spine surgery. Int J Spine Surg. 2023;7(S1):S18–25.
10. Browd SR, Park C, Donoho DA. Potential applications of artificial intelligence and machine learning in spine surgery across the continuum of care. Int J Spine Surg. 2023;17(S1):S26–33.
11. Atasever S, Azginoglu N, Terzi DS, Terzi R. A comprehensive survey of deep learning research on medical image analysis with focus on transfer learning. Clin Imag. 2023;94:18–41.
12. Sistaninejhad B, Rasi H, Nayeri P. A review paper about deep learning for medical image analysis. Comput Math Method M. 2023;2023:7091301.
13. Chen Z, Guo L, Zhang R, Fang Z, He X, Wang J. BX2S-Net: learning to reconstruct 3D spinal structures from bi-planar X-ray images. Comput Biol Med. 2023;154:106615.
14. Tamai K, Terai H, Hoshino M, Tabuchi H, Kato M, Toyoda H, Suzuki A, Takahashi S, Yabu A, Sawada Y, et al. Deep learning algorithm for identifying cervical cord compression due to degenerative canal stenosis on radiography. Spine. 2023;48(8):519–25.
15. Bharadwaj UU, Christine M, Li S, Chou D, Pedoia V, Link TM, Chin CT, Majumdar S. Deep learning for automated, interpretable classification of lumbar spinal stenosis and facet arthropathy from axial MRI. Eur Radiol. 2023;33(5):3435–43.
16. Wang D, Sun Y, Tang X, Liu C, Liu R. Deep learning-based magnetic resonance imaging of the spine in the diagnosis and physiological evaluation of spinal metastases. J Bone Oncol. 2023;40:100483.
17. Yan Y, Zhang X, Meng Y, Shen Q, He L, Cheng G, Gong X. Sagittal intervertebral rotational motion: a deep learning-based measurement on flexion-neutral-extension cervical lateral radiographs. BMC Musculoskel Dis. 2022;23(1):967.
18. Gahlot N, Kunal K, Elhence A, Meena U, Gupta A, Netaji J, Swami D, Goyal M, Jamal A. Ten-segment classification has lowest inter/intra-observer reliability as compared to Schatzker, three-column and AO systems for Tibial Plateau Fractures: a comparison based on surgeons' experience. Arch Bone Jt Surg. 2023;11(4):256–61.
19. Pang S, Pang C, Su Z, Lin L, Zhao L, Chen Y, Zhou Y, Lu H, Feng Q. DGM-SNet: spine segmentation for MR image by a detection-guided mixed-supervised segmentation network. Med Image Anal. 2022;75:102261.
20. Payer C, Stern D, Bischof H, Urschler M. Integrating spatial configuration into heatmap regression based CNNs for landmark localization. Med Image Anal. 2019;54:207–19.
21. Lim JK, Wong HK. Variation of the cervical spinal Torg ratio with gender and ethnicity. Spine J. 2004;4(4):396–401.
22. Dosovitskiy A , Beyer L , Kolesnikov A ,et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[C]. International Conference on Learning Representations. 2021.
23. Kingma D , Ba J . Adam: A Method for Stochastic Optimization.Computer Science, 2014. https://doi.org/10.48550/arXiv.1412.6980.
24. Gotoh M, Nakaura T, Funama Y, Morita K, Sakabe D, Uetani H, Nagayama Y, Kidoh M, Hatemura M, Masuda T, et al. Virtual magnetic resonance lumbar spine images generated from computed tomography images using conditional generative adversarial networks. Radiography. 2022;28(2):447–53.
25. Chen Y, Gao Y, Li K, Zhao L, Zhao J. Vertebrae identification and localization utilizing fully convolutional networks and a hidden Markov model. IEEE Trans Med Imaging. 2020;39(2):387–99.
26. Liao H, Mesfin A, Luo J. Joint vertebrae identification and localization in spinal CT images by combining short- and long-range contextual information. IEEE Trans Med Imaging. 2018;37(5):1266–75.
27. Chen IH, Liao KK, Shen WY. Measurement of cervical canal sagittal diameter in Chinese males with cervical spondylotic myelopathy. Zhonghua Yi Xue Za Zhi (Taipei). 1994;54(2):105–10.
28. Summerfield SL. The relationship of developmental narrowing of the cervical spinal canal to reversible and irreversible injury of the cervical spinal cord in football players. An epidemiological study. J Bone Joint Surg Am. 1998;80(10):1554–5.
29. Ezra D, Slon V, Kedar E, Masharawi Y, Salame K, Alperovitch-Najenson D, Hershkovitz I. The torg ratio of C3–C7 in African Americans and European Americans: a skeletal study. Clin Anat. 2019;32(1):84–9.
30. Bajwa NS, Toy JO, Young EY, Ahn NU. Establishment of parameters for congenital stenosis of the cervical spine: an anatomic descriptive analysis of 1,066 cadaveric specimens. Eur Spine J. 2012;21(12):2467–74.
31. Ghogawala Z, Whitmore RG. Asymptomatic cervical canal stenosis: is there a risk of spinal cord injury? Spine J. 2013;13(6):613–4.

Publisher's Note