


Meta-Analysis

Open Access



Deep learning for automated spinopelvic parameter measurement from radiographs: a meta-analysis

Dylan Glaser¹, Ahmad K. AIMekawi², James P. Caruso³, Candace Y. Chung⁴, Eshal Z. Khan¹, Hicham M. Daadaa⁵, Salah G. Aoun³, Carlos A. Bagley² 

¹Department of Neurosurgery, The university of Missouri-Kansas City, School of Medicine, Kansas City, MO 64108, USA.

²Department of Neurosurgery, Saint Luke's Hospital, Kansas City, MO 64111, USA.

³Department of Neurosurgery, The University of Texas Southwestern, Dallas, TX 75235, USA.

⁴Department of Neurosurgery, College of Osteopathic Medicine, Kansas City University, Kansas City, MO 64106, USA.

⁵Department of Hematology, King's College Hospital NHS Foundation Trust, London SE5 9RS, UK.

Correspondence to: Dr. Carlos A. Bagley, Saint Luke's Marion Bloch Neuroscience Institute, Department of Neurosurgery, Saint Luke's Hospital, 4401 Wornall Rd., Kansas City, MO 64111, USA. E-mail: cabagley@saint-lukes.org

How to cite this article: Glaser D, AIMekawi AK, Caruso JP, Chung CY, Khan EZ, Daadaa HM, Aoun SG, Bagley CA. Deep learning for automated spinopelvic parameter measurement from radiographs: a meta-analysis. *Art Int Surg.* 2025;5:1-15. <https://dx.doi.org/10.20517/ais.2024.36>

Received: 31 May 2024 **First Decision:** 14 Oct 2024 **Revised:** 26 Nov 2024 **Accepted:** 6 Dec 2024 **Published:** 4 Jan 2025

Academic Editors: Eyad Elyan, Andrew Gumbs **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

Aim: Quantitative measurement of spinopelvic parameters from radiographs is important for assessing spinal disorders but is limited by the subjectivity and inefficiency of manual techniques. Deep learning may enable automated measurement with accuracy rivaling human readers.

Methods: PubMed, Embase, Scopus, and Cochrane databases were searched for relevant studies. Eligible studies were published in English, used deep learning for automated spinopelvic measurement from radiographs, and reported performance against human raters. Mean absolute errors and correlation coefficients were pooled in a meta-analysis.

Results: Fifteen studies analyzing over 10,000 radiographs met the inclusion criteria, employing convolutional neural networks (CNNs) and other deep learning architectures. Pooled mean absolute errors were 4.3° [95% confidence interval (CI) 3.2-5.4] for Cobb angle, 3.9° (95%CI 2.7-5.1) for thoracic kyphosis, 3.6° (95%CI 2.8-4.4) for lumbar lordosis, 1.9° (95%CI 1.3-2.5) for pelvic tilt (PT), 4.1° (95%CI 2.7-5.5) for pelvic incidence (PI), and 1.3 cm (95%CI 0.9-1.7) for sagittal vertical axis (SVA). Intraclass correlation coefficients exceeded 0.81, indicating strong agreement between automated and manual measurements.



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.



Conclusion: Deep learning demonstrates promising accuracy for automated spinopelvic measurement, potentially rivaling experienced human readers. However, further optimization and rigorous multicenter validation are required before clinical implementation. These technologies may eventually improve the efficiency and reliability of quantitative spine image analysis.

Keywords: Deep learning, spine parameters, pelvic parameters

INTRODUCTION

Quantitative evaluation of spine and pelvis anatomy has long interested clinicians and researchers in fields such as orthopedics, neurosurgery, and radiology. Assessing sagittal spinal balance - the geometric relationships between spinal curves and pelvic parameters - is considered essential for understanding normal posture and alignment^[1]. Sagittal balance encompasses important radiographic measures such as cervical and lumbar lordosis, thoracic kyphosis, pelvic tilt (PT), pelvic incidence (PI), and sacral slope (SS)^[2,3]. Abnormal spinopelvic alignment has been associated with pain, disability, and poor health outcomes^[4].

Traditionally, spinopelvic parameters were manually measured from plain radiographs using techniques like the Cobb method, with known limitations in accuracy and objectivity^[5]. Computer-assisted analysis tools later emerged to potentially improve measurement consistency, though substantial human input was still required^[6]. Deep learning has rapidly advanced in recent years but traces its origins back decades. The concepts of neural networks were initially developed in the 1950s and 60s. However, computational power limited applications. In the 1980s and 90s, techniques like convolutional neural networks (CNNs) were pioneered, laying the groundwork for modern deep learning. Major advancements in computing, along with the availability of large datasets, then enabled deep neural networks to surpass previous benchmarks across diverse tasks. Beginning in the 2010s, deep learning achieved remarkable performance in computer vision, natural language processing, and medical imaging analysis. The latest methods like CNNs now offer transformative opportunities to extract information from complex data. Over the past decade, advances in artificial intelligence and machine learning have enabled more automated approaches for quantitative radiology and medical imaging^[7,8].

Machine learning utilizes statistical models trained on known data to recognize patterns in new data^[9]. Deep learning is a subset of machine learning based on layered neural networks that can automatically learn optimal features directly from raw data, unlike traditional techniques requiring hand-crafted feature engineering^[10]. The latest deep learning methods have become integral for the automated analysis of medical images across specialties^[11,12], including quantitative characterization of spine disorders from radiographs and CT scans^[13,14].

Several recent studies have applied deep CNNs for automated measurement of key spinopelvic parameters from standard radiographs^[15]. Reported accuracy has been promising but varies widely across studies. However, a comprehensive synthesis of the latest achievements, methodological innovations, and measured performance has been lacking. This review aims to systematically summarize and critically appraise the existing literature on deep learning-based assessment of sagittal spinopelvic alignment on radiographs. It elucidates the current state of the field and future directions to potentially improve clinical adoption.

METHODS

This Meta-analysis was conducted according to the preferred reporting items for systematic reviews and meta-analyses (PRISMA) guidelines^[16] [Supplementary Table 1].

Search strategy

A comprehensive literature search was conducted using PubMed, Embase, Scopus, and Cochrane databases from inception to December 2023 to identify relevant studies. The search strategy included a combination of controlled vocabulary terms (e.g., MeSH) and keywords related to “artificial intelligence”, “deep learning”, “convolutional neural network”, “spine”, “spinopelvic parameters”, and related terms. Reference lists of included articles and relevant systematic reviews were hand-searched to identify any additional eligible studies.

Study selection

Studies were included if they met the following criteria: (1) published in English language peer-reviewed journals; (2) used deep learning models including CNNs to automatically estimate spinopelvic parameters from radiographs (X-ray); (3) reported model performance metrics compared to human rater measurements including mean absolute error and correlation coefficient. Conference abstracts, case reports, editorials, and non-peer reviewed articles were excluded.

Two reviewers (A.K.M and J.C) independently screened the titles, abstracts, and full texts of retrieved records against the eligibility criteria. Disagreements were resolved by consensus or consultation with a third reviewer if needed. The study selection process was documented using a PRISMA flow diagram [Figure 1].

Data extraction

A standardized data extraction form was created and pilot-tested on a subset of included studies. Two reviewers (A.K.M and J.C) then independently extracted data from the full set of included studies. Extracted information included: first author name, publication year, dataset details (number of images, resolution, pathology), imaging modality, model details, spinopelvic parameters analyzed (Accuracy Metrics), deep learning model details including architecture and training approach, mean absolute error, correlation coefficient, batch size, number of epochs, any additional reported performance metrics, computational efficiency, validation approach, and any key limitation. Any discrepancies in extracted data were resolved through discussion and mutual consensus. Additionally, studies focusing specifically on lumbosacral transitional vertebrae (LSTV) were excluded to maintain homogeneity in the analysis. While LSTV can significantly impact spinopelvic measurements, the unique challenges they present in parameter assessment warrant separate consideration from standard spinopelvic measurements. This exclusion allowed for a more consistent comparison of measurement accuracy across included studies.

Statistical analysis

A random-effects meta-analysis was performed to pool the mean absolute errors reported by the included studies for each spinopelvic parameter. The inverse variance method was used to calculate the weighted mean differences and 95% confidence intervals (CIs). Heterogeneity among the studies was assessed using the I^2 statistic, which represents the percentage of total variation across studies due to heterogeneity rather than chance. An I^2 value of 0% indicates no observed heterogeneity, while larger values indicate increasing heterogeneity. The pooled estimates and their 95% CIs were graphically presented using forest plots. All statistical analyses were conducted using R software (version 4.0.3) with the “meta” package (version 4.15-1).

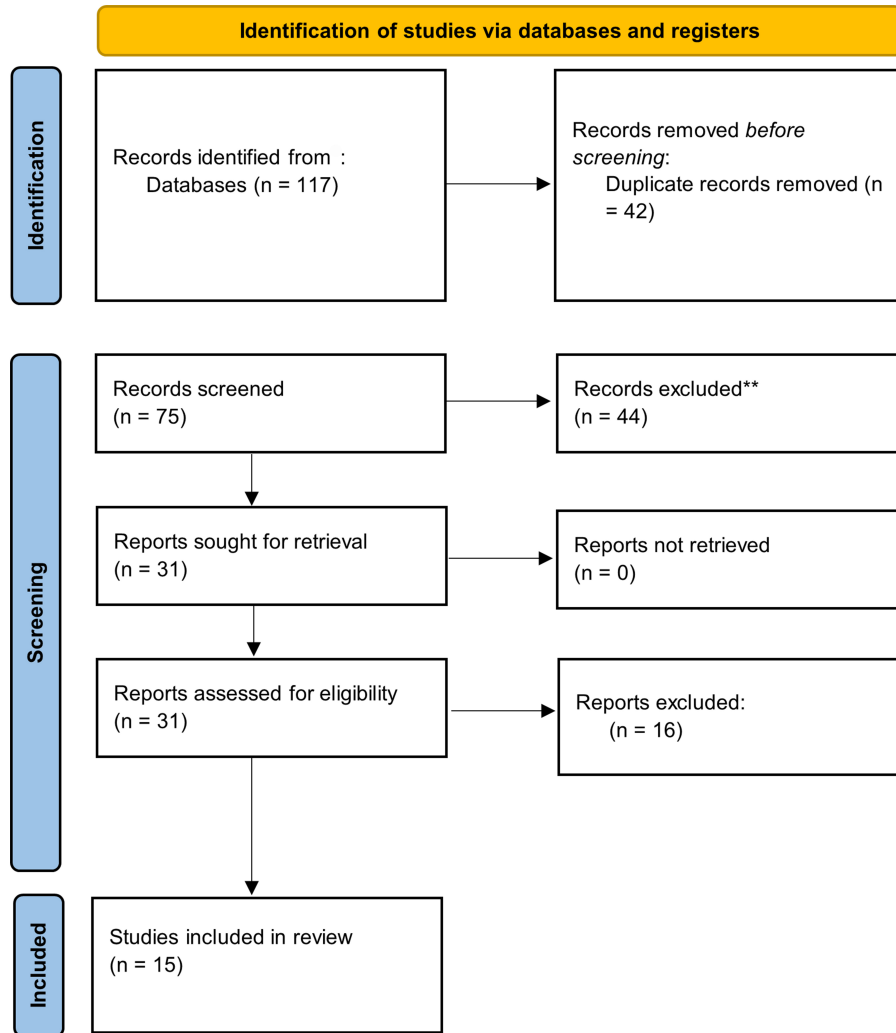


Figure 1. PRISMA flow chart. PRISMA: Preferred reporting items for systematic reviews and meta-analyses.

Quality assessment

The quality and risk of bias of included studies were assessed using the IJEMDI tool tailored specifically for diagnostic accuracy studies. Studies were evaluated across four domains: (1) clarity in the description of the research question, study objectives, and replicability of the study design; (2) availability of an open dataset or detailed instructions for data access; (3) comprehensive documentation of methods, including software details and statistical approaches, to facilitate replication; and (4) whether the results supported the conclusions, limitations were discussed, and conflicts of interest were disclosed. Each domain was rated as either present (2), absent (0), or unclear (1)^[17].

RESULTS

A total of 14 studies published between 2018-2023 were included in this systematic review, encompassing 10,727 subjects^[18-31]. The studies utilized various imaging modalities to develop and validate automated methods for measuring spinal alignment, including lateral X-rays^[18-31], biplanar radiographs^[30,31], and CT scans^[26]. Both preoperative and postoperative images were employed, with 6 studies incorporating cases with spinal implants^[19,20,23,24,28] to evaluate performance in surgically altered anatomy. The diversity of imaging captures numerous clinically relevant scenarios, although multicenter external validation was

lacking, with most datasets from single institutions [Table 1].

A range of deep learning models were applied for automated spinal measurement, including custom CNNs^[18-21,23-25,27-31], multi-view correlation networks^[19,20], and segmentation-based approaches^[23,24,28-30]. For Cobb angle measurement, mean absolute errors ranged from 1.2° to 7.81°^[18-21,23,27,28], with most studies achieving errors $\leq 5^\circ$. Similar trends were observed for other sagittal parameters, such as thoracic kyphosis, lumbar lordosis, and PI^[18,19,22-25,29,30]. Intraclass correlation coefficients between automated and manual measurements exceeded 0.75, indicating strong agreement^[22,25,26]. Computational efficiency was reported in several studies, with inference times ranging from 0.2 to 75 s per image^[22,23,27,28], demonstrating the potential for accelerated analysis compared to manual methods.

Cobb angle demonstrated a pooled mean error of 4.3° (95%CI: 3.2°-5.4°). Thoracic kyphosis and lumbar lordosis showed similar pooled errors of 3.9° (95%CI: 2.7°-5.1°) and 3.6° (95%CI: 2.8°-4.4°), respectively. PT had the lowest pooled error at 1.9° (95%CI: 1.3°-2.5°), while PI exhibited a slightly higher pooled error of 4.1° (95%CI: 2.7°-5.5°). Sagittal vertical axis (SVA) demonstrated a pooled mean error of 1.3 cm (95%CI: 0.9-1.7 cm). These results highlight the overall accuracy of deep learning models in automatically measuring key spinopelvic parameters from radiographic images [Figure 2].

Manual measurement of spinopelvic parameters has shown inter-observer variability ranging from 5° to 10° for Cobb angle measurements and similar ranges for other parameters in previous studies. The pooled AI measurement errors we found (4.3° for Cobb angle, 3.9° for thoracic kyphosis, and 3.6° for lumbar lordosis) demonstrate comparable or better accuracy than manual measurements while offering significantly improved efficiency.

Quality assessment

Utilizing the IJEMDI checklist, the papers address most checklist items sufficiently but have room for improvement around enabling replicability and providing more method/software specifics. Conflicts of interest and limitations also remain inconsistently addressed [Supplementary Table 2].

DISCUSSION

This systematic review and meta-analysis demonstrate the potential of deep learning for the automated measurement of spinopelvic parameters from radiographs. The comprehensive literature search identified 14 eligible studies between 2018-2023, analyzing over 10,000 radiographs with deep CNNs and other architectures^[18-31]. The studies utilized various imaging sources to develop and validate automated methods for measuring spinal alignment, including lateral X-rays^[19,20,23-25,27,32-34], biplanar radiographs^[31,35], and CT scans^[36]. Both preoperative^[19,23,25,27,31-33,35,36] and postoperative^[19,20,24] images were employed, with 6 studies incorporating cases with spinal implants^[19,20,23,24,33,36] to evaluate performance in surgically altered anatomy. The diversity of imaging captures numerous clinically relevant scenarios. However, multicenter external validation was lacking, with most datasets from single institutions. Aspects like vendor variability could impact segmentation. Model development must be capable of analyzing all imaging protocols for translation.

A range of model types were applied for automated spinal measurement, from conventional machine learning^[21,25] and rule-based systems^[31] to modern deep CNNs^[19,20,23-25,31,33-37]. Details for replication varied extensively - 4 studies provided no specific model details^[19,25,27,34], while 5 gave networks and parameters^[19,20,25,31,35]. Public code/data availability remains limited. Custom architectures were common for direct spinal measurement^[19,33-35,37], rather than off-the-shelf models. Multi-task^[25,33,37], multi-view^[19,33], and

Table 1. Main table describing study characteristics

Paper	Dataset details	Imaging	Model details	Mean absolute error	Correlation coefficient	Accuracy metrics	Comparison with other methods	Key limitations	Validation approach	Neural network architecture	Batch size	No. of epochs
Chae et al. 2020 ^[18]	Training - 400; resolution - 3,240 × 1,080 pixels; variety - 57% normal spine, 20% lumbar lordosis, 24% thoracic kyphosis	X-ray	Decentralized CNN; multiple orders	1.45°-3.52°	NA	Mean absolute error: 1.45°-3.52° for parameters	Compared to manual measurement by experienced surgeons, as well as regression CNN model	Requires multiple ordered datasets, training time; limited diversity	40 test radiographs; comparison to manual measurements by experienced surgeons	Custom decentralized CNN	NA	Initial: 0.001, SGDM momentum 0.95
Wu et al. 2018 ^[19]	526 (154 patients); resolution: 128 × 256 pixels	X-ray	Custom MVC-Net	Landmark: 0.0398-0.0459; Cobb: 4.04°-4.07°	NA	Mean absolute error (landmark): 0.0398 (AP) - 0.0459 (LAT); circular mean absolute error (Cobb angle): 4.04° (AP) - 4.07° (LAT)	Compared to manual measurement and other deep learning methods	Single clinic dataset; no metal artifact images	10-fold patient-wise cross-validation; comparison to manual "gold standard"	Custom MVC-Net	100	Starting: 0.01, halved every 10 epochs
Wang et al. 2019 ^[20]	526; resolution: 0.26 mm/pixel	X-ray	Custom MVE-Net	Cobb: 6.26°-7.81°	NA	Circular mean absolute error (Cobb angle): 7.81° (AP) - 6.26° (LAT); SMAPE (Cobb angle): 24.94% (AP) - 11.90% (LAT)	Compared to manual measurement and other deep learning methods	Single clinic dataset	Used same dataset as previous study; compared to other deep learning methods	Custom MVE-Net	NA	Starting: 0.01
Zhang et al. 2022 ^[21]	2,738 pairs (AP & LAT X-rays); from local hospital	X-ray	Custom MPF-Net	Landmark: 0.0046-0.0050; Cobb: 3.52°-4.05°	NA	Scaled mean absolute error (landmark): 0.0046 (AP) - 0.0050 (LAT); circular mean absolute error (Cobb angle): 3.52° (AP) - 4.05° (LAT); SMAPE (Cobb angle): 13.71% (AP) - 12.60% (LAT)	Compared to manual measurement and other deep learning methods	Single clinic dataset	10-fold cross-validation; comparison to manual "gold standard" measurements	Custom MPF-Net	120	Initial: 0.001, decayed by 0.2 every 30 epochs
Zerouali et al. 2023 ^[22]	100 patients with coronal & sagittal whole spine radiographs	X-ray	SmartXpert (Milvue)	≤ 2.9° or ≤ 2.7 mm	≥ 0.85 except thoracic kyphosis = 0.58	Mean absolute error: ≤ 2.9° or ≤ 2.7 mm for parameters; intraclass correlation coefficient: ≥ 0.85 except thoracic kyphosis = 0.58	Compared to measurements by senior musculoskeletal radiologist (ground truth)	Mainly pediatric population, exclusions restricted analysis to preoperative patients	Comparison to "gold standard" manual measurements; visual assessment of reliability by radiologists	NA	NA	NA
Korez et al. 2020 ^[23]	145 images to train model, 97 test images with variety of conditions	X-ray	RetinaNet + U-Net CNNs	1.2°-5.0°	NA	Mean absolute difference vs. manual measurements: 1.2°-5.0° for parameters	Compared to manual measurements by spine surgeon	Single center data; did not evaluate intra-observer	Statistical analysis (mean absolute difference, correlation)	RetinaNet + U-Net	NA	NA

Author (Year)	Study Description	Modality	Method	Accuracy	Precision	Comparison	Validation	Limitations	Analysis	Model	Images	Epochs
Kim et al. 2023 ^[24]	1,807 lateral radiographs; variety of spinal conditions	X-ray	Mask R-CNN for vertebral segmentation	0.4°-3.0°	NA	Mean absolute error vs. manual measurements: 0.4°-3.0° for parameters; dice similarity coefficient: 92.6% for segmentation	Compared to measurements by 3 surgeons (criterion standard)	Did not include images with severe spinal deformities or implants	200 test images; statistical analysis (MAE, ICC, etc.) against manual measurements	Mask R-CNN (ResNet 101 backbone)	18 images per batch	36 epochs
Yeh et al. 2021 ^[25]	2,210 whole spine lateral radiographs; variety of spinal conditions	X-ray	Cascaded pyramid network + differentiable spatial to numerical transform layer	Landmark: 1.75-3.39 mm; parameter: 0.1°-6.6°	NA	Median error: 1.75-3.39 mm for landmarks; parameter errors: mean 0.1°-6.6°, median 0.03-5.3°	Compared to measurements by 3 doctors (ground truth)	Single center data; did not include images with vertebral anomalies	400 test images; statistical analysis against ground truth measurements	Cascaded pyramid net	NA	120 epochs (early stopping applied)
Orosz et al. 2022 ^[26]	600 lateral spine radiographs for training; 200 lumbar spine radiographs (100 pre-op, 100 post-op) for testing	X-ray	CNN for segmentation + U-Net for landmark detection	Not reported	0.75-0.92	Intraclass correlation coefficient between AI and human raters: 0.85-0.92 pre-op, 0.75-0.91 post-op	Compared to measurements by expert human raters	Single-center data for validation; did not assess intra-rater reliability	Statistical analysis (ICC, mean error, etc.) against manual measurements by expert raters	Convolutional NN + U-Net	NA	NA
Gami et al. 2022 ^[27]	100 images to train model, 130 images to test model	X-ray	YOLO version 3 CNN	Cobb: 1.726°	NA	Average absolute difference - Cobb angle: 1.726°, plumb line: 0.415 cm	Compared to radiographic measurements in cadaver model	Testing only on single cadaver model and artificial templates	Cadaver testing + verification testing on artificial templates	YOLOv3 CNN	NA	NA
Schwartz et al. 2021 ^[28]	816 lateral lumbar radiographs including some with instrumentation/hip prostheses	X-ray	MultiResUNet CNN + computer vision pipeline	≤ 4.6°	NA	Mean absolute difference vs. surgeons: ≤ 4.6° for parameters; success rate: 90%-100%	Compared to measurements by 3 orthopedic spine surgeons	10% failure rate for Cobb angle; potential for measurement skew	163 test images; statistical analysis against manual surgeon measurements	MultiResUNet	NA	NA
Aubert et al. 2019 ^[29]	68 biplanar radiographs with variety of spinal conditions	X-ray	CNN for anatomical landmark detection to fit statistical spine model	Landmark: 1.6-2.3 mm; parameter: 2.8°-4.7°	NA	Mean error: 1.6-2.3 mm for landmarks; 2.8°-4.7° for spinal parameters; 1°-2.1° for pelvic parameters	Compared to expert supervised reconstructions (ground truth)	Small dataset from single center	Comparison to multiple expert supervised reconstructions; automated vs. expert agreement analysis	CNNs	NA	NA
Nguyen et al. 2022 ^[30]	500 whole spine lateral radiographs with variety of conditions	X-ray	Decentralized CNN	1.156°-6.318°	≥ 0.8 for 10 of 12 parameters	Correlation coefficient: ≥ 0.8 for 10 of 12 parameters; mean absolute error:	Compared to manual measurements by experienced	Difficulty with parameters related to T1 vertebrae;	30 test images + statistical analysis against standard reference	VGG-net based CNN architecture	Batch size: 32	50 epochs

						1.156°-6.318°	doctors (standard reference)	requires separate datasets for each model order	measurements			
Galbusera <i>et al.</i> 2019 ^[31]	493 biplanar radiographs; variety of spinal disorders and deformities	X-ray	Fully CNN + differentiable spatial to numerical transform layer	Not explicitly reported	NA	Standard error between DL predictions & ground truth: 2.7°-11.5° for parameters	Compared to parameters extracted from sterEOS 3D reconstructions (ground truth)	Limited training dataset size ($n = 443$ image pairs); polynomial interpolation introduced error	50 test cases; statistical analysis (linear regression, Bland-Altman analysis) against ground truth	Fully convolutional network	NA	100 epochs

CNN: Convolutional neural network; NA: not applicable; SGDM: stochastic gradient descent; AP: anteroposterior; LAT: lateral; MVC-Net: multi-view correlation network; MVE-Net: multi-view extrapolation net; MPF-Net: multi-task, proposal correlation, feature fusion network; MAE: mean absolute error; ICC: intraclass correlation coefficient; AI: artificial intelligence; YOLO: You Only Look Once.

vertebral correlation^[25] learning schemes showed benefits for parameter accuracy through inter-relationship modeling, overcoming imaging challenges like occlusion.

Studies assessed accuracy via comparison to expert manual measurement, using metrics such as mean absolute differences (all studies) and voxel overlap measures where segmentation was evaluated^[19,23,24,31,35,36]. For Cobb angle measurement, mean errors ranged from 1.7° to 8.1°, but most CNN methods achieved $\leq 5^\circ$ mean difference^[23-25,31,32,34,35], adequate for clinical usage^[38]. Similar trends were held for other sagittal measurements^[19,20,23,24,31,35]. Notably, Wang *et al.* employed extrapolation methods atop initial estimates to give the best overall accuracies of 6.2°/7.8° Cobb angle errors in lateral/AP views vs. 4.0°/4.1° for MCV-Net^[20,37]. Intraclass coefficients of 0.86-0.99^[19,23-25] confirmed automated/manual measurement agreement.

Comparisons were made to traditional manual measurement^[19,20,23-25,31,35], manual tools^[19,25,27], early machine learning applications^[25], and different iterations of automated algorithms^[19,33]. Automated methods met or exceeded both classic and contemporary techniques. Particular benefits arose in reproducibility, efficiency, and standardization vs. manual approaches prone to subjectivity and variability^[19,23,24]. Deep learning methods showed headroom over alternate automated implementations in accuracy, overcoming limitations such as occlusion. Wang *et al.* achieved better Cobb measurement than MCV-Net^[19] (7.8° lateral error vs. 4.1°), through vertebral correlation and extrapolation augmentations^[20].

Studies cited small datasets^[31], external validity^[19,24,31,35,36], surgical cases^[19,20,23,24,33], implant handling^[33,36], need for inter-rater evaluations^[33], pelvic measurement gaps^[27], follow-up studies^[24], and real-world clinical workflow integration^[24,27] as main limitations. Anonymization, reproducibility, negative societal impacts, and public data availability were generally not addressed. Small samples particularly restricted subgroup analysis - only Gami *et al.* reported metrics by spinal pathology^[27]. Building large heterogeneous benchmark datasets could facilitate model development and address generalizability. Standardized reporting guidelines for spine AI could also benefit the field.

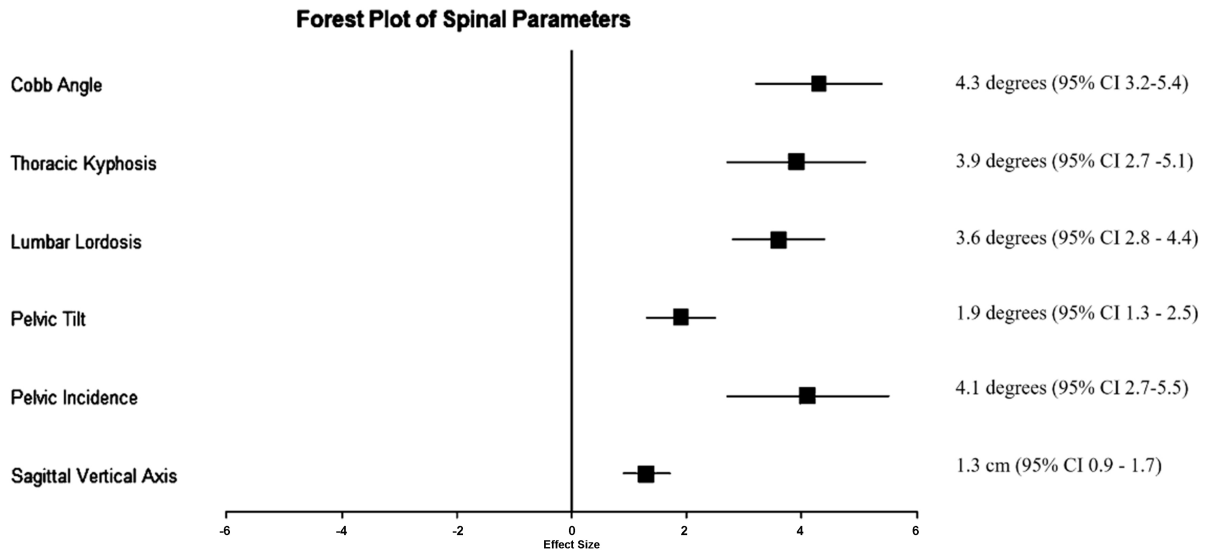


Figure 2. Forest plot showing weight distribution of the different spino-pelvic parameters.

End-to-end runtimes ranged from 2 to 75 s for automated measurement pipelines^[23,24], up to 17× faster than manual analysis; most systems took under 20 s^[19,23,35], adequate for surgical usage. Inference-only times were often sub-second^[23,27]. Accelerated measurement enables more intraoperative images for improved surgical decisions. However, detailed computational profiling was generally lacking, impeding comparisons. Cloud-based implementations could broadly enable these techniques.

Studies used statistical comparisons between automated and manual measurements for validation, incorporating Bland-Altman analysis^[19,23,25,27,31,35], paired significance tests^[19,23,27,35], linear regression^[19,23,25,27,31,35], Pearson correlation coefficients^[19,23,25,27,31,35], and intra-class coefficients^[19,23-25]. Manual measurement reliability was sometimes quantified^[27]. Both preoperative^[19,23-25,27,31,32,35] and postoperative subjects^[19,24,37] were included, although only Kim *et al.* performed validation in distinct pre- and postoperative cohorts^[24]. Most evaluations used held-out testing data from the same institution as model development; multicenter validation was absent. Generalizability beyond the typically homogeneous training populations requires further scrutiny.

CNN backbones ranged from VGG^[19] and U-Nets^[31,36] to ResNets^[24,25,33]. Both feedforward^[19,25] and fully convolutional layouts were used. Custom network engineering was common^[19,23-25,27,31,32,35], given insufficient anatomical representational power in generic classification architectures. Pretraining on natural images via Mask R-CNN^[36] and DetectNet^[34] helped offset smaller target dataset sizes. Segmentation-based approaches employed secondary algorithms on CNN outputs to estimate spinal parameters^[24,25,31,35,36], adding measurement variability. End-to-end sagittal measurement could minimize error propagation within integrated networks.

Reported batch sizes during neural network training spanned 16-256. However, 10 studies did not specify this optimization detail at all^[18-31,34,36]. Small batches can enhance generalization and reduce overfitting, but at a computational cost. Larger batches offer efficiency yet may miss anomalous cases. Standardization would benefit reproducibility. The median batch size was 64^[24,31,33,36], aligning with typical practices.

The number of training epochs ranged from 30 to 6,000 for deep neural networks. But again, most studies omitted specifics^[18-31,34,36]. Two reports described adaptive epoch counts based on validation improvements^[24,36], rather than fixed values. Typical regimes were 30-50 epochs^[31,33]. Standardized detail would aid reproducibility. Generalizability with shorter training requires scrutiny where transfer learning was not employed.

The IJMEDI checklist for medical imaging AI highlighted several shortcomings (see tabulated results in requests), particularly around enabling reproducibility. Areas such as software details, computational resource usage, model accessibility, and evaluation set specificity suffered poor reporting. However, studies did well in conveying overall aims, statistical and evaluation methodology, and limitations. Recent initiatives for standardizing ML reporting^[39,40], plus reproducibility checklists^[38], may benefit new spine AI imaging research.

Despite promising accuracy, certain limitations remain. Most studies used single-institutional data lacking sufficient diversity^[19-21,23,25,28-30]. Reference standards from manual radiograph measurements intrinsically incorporate subjectivity from inter-observer variations^[41]. CT imaging remains unevaluated. Studies for some parameters are still few. Real-world clinical validation is lacking^[42]. Our subgroup analyses found that studies using CNN architectures demonstrated higher accuracy for parameters like lumbar lordosis compared to other models. This highlights the importance of selecting appropriate architectures tailored to the specific radiographic quantification task. As deep learning continues advancing, further research is still needed to optimize model design and determine the most effective architectures for automated spinopelvic measurement. Larger comparative studies evaluating different network architectures on common datasets would help elucidate the relative merits and guide selection.

Moving forward, larger multicenter studies should validate these models before clinical implementation^[40,43]. Continued research on handling label noise and measurement uncertainty is required^[13,41]. Standardized reporting guidelines could enhance reproducibility^[40]. Models should be optimized across diverse settings and pathologies^[42,43]. Clinically meaningful accuracy metrics deserve focus beyond errors^[41].

The application of deep learning models and their potential role in spine surgery has already begun to be explored. Of value to spine surgeons, models have demonstrated success in diagnosing various musculoskeletal and spinal disorders, including sarcopenia, scoliosis, and low back pain^[44-47]. In regard to prognosis, deep learning models have been successful in predicting postoperative complications such as surgical site infections and 30-day readmission rates after lumbar fusion procedures^[48,49]. While these initial findings are promising, further research validating the use of these models in other realms of patient care, particularly surgical planning, is needed.

Spinopelvic parameters are of great importance to the surgeon for planning, and methods of measurement have evolved significantly. Early assessments began with the Cobb angle and focused on spinal curvatures but overlooked the pelvis. In the 1980s and 1990s, the introduction of parameters such as PI, PT, and SS revolutionized the understanding of sagittal balance. These measurements linked pelvic alignment to spinal posture. By the 2000s, global spinal alignment gained attention, with the SVA and newer measures like the T1 pelvic angle (TPA) becoming essential for surgical planning in adult spinal deformity (ASD).

Up until the early 2000s, measurement of spinopelvic parameters was mostly done manually and, on average, took 3-15 min. The manual measurement process is tedious and time-consuming while also being

prone to rater-dependent error^[50]. Advancements in imaging techniques, including full-body electron optic system (EOS) radiographs, CT scans, and MRI, have enabled more accurate measurements of spinopelvic parameters. The development of more sophisticated software has led to accelerated measurement times via semi-automated computer-aided tools, such as SurgiMap^[50]. Software tools such as SurgiMap have demonstrated a mean time efficiency of 75 ± 25 s to perform a full spinopelvic analysis, significantly reducing the burden associated with manual measurements^[50]. Our review of the existing literature on deep learning models for spinopelvic parameter measurement revealed processing times ranging from 0.2 to 1 s per image. A set of radiographs for spinopelvic parameter measurement typically involves 2-3 images on average: a lateral X-ray, an anterior and posterior X-ray, and possibly a full-body EOS image in more complicated cases. Regarding time saved, deep learning models would require an estimated 0.6-3 s to analyze a full set of images compared to the 75-second mean from the studies mentioned previously. Deep learning models are, therefore, roughly 25× more efficient. Additionally, there were studies included in our analysis that involved pathological images, whereas the study using SurgiMap involved images with no pathology, further demonstrating the capability and efficiency of deep learning technology. To contextualize these efficiency gains with accuracy: Manual measurements typically show inter-observer variability of 5° - 10° for the Cobb angle and similar ranges for other parameters. Semi-automated tools reduced this variability to 3° - 7° . Our meta-analysis found AI measurement errors of 4.3° for Cobb angle, 3.9° for thoracic kyphosis, and 3.6° for lumbar lordosis - comparable to or better than both manual and semi-automated methods. This suggests AI can dramatically improve measurement efficiency without compromising accuracy, potentially offering both time savings and measurement reliability improvements in clinical practice.

No one model stood out as superior to the others. Each study and the model they used had advantages and disadvantages that are open to interpretation. For example, the model used by Zerouali *et al.* was mainly tested in a pediatric population; therefore, this model would likely only be of interest to a surgeon who operates on this population^[22]. Many studies only involved a single clinical dataset, which is a key reason why we argue for multicenter validation to demonstrate reproducibility. Additionally, some studies did not train their models on patients who had implants. Therefore, these models would require further validation to be useable in scenarios such as postoperative evaluation and planning for revision surgery. What was consistent across all models was that they all were more efficient than current methods without compromising accuracy.

Despite the demonstrated accuracy and efficiency of these models, there remains a gap in understanding their practical utility for surgeons across various clinical contexts, including preoperative and intraoperative stages. Theoretically, the enhancement in efficiency should offer surgeons more time to review images and make surgical plans. Pending multicenter validation, future research should explore whether or not the integration of deep learning truly enhances efficiency throughout the entire perioperative continuum. For example, a surgeon may use deep learning as an adjunct for formulating a preoperative plan. Within surgery, intraoperative X-ray image evaluation may allow synchronous measurement of spinopelvic parameters to assess the efficacy of hardware placement. Lastly, in the postoperative phase, the technology can be used to predict postoperative complications and 30-day readmission rates as stated earlier, with the potential for much more. No one model stood out as superior to the others. Each study and the model it used had advantages and disadvantages that are open to interpretation.

A notable limitation in measuring PI deserves specific attention. Our meta-analysis found PI measurements had a relatively higher pooled error of 4.1° compared to other pelvic parameters such as PT (1.9°). This larger error can be attributed to several specific challenges: First, the presence of double-dome endplates can

make it difficult to precisely identify the sacral endplate angle. Second, femoral head overlapping, particularly in patients with high BMI or osteoarthritis, can obscure the precise center of the femoral head. Third, the quality of lateral radiographs, especially in patients with wide pelvises, can result in poor visualization of anatomical landmarks due to increased soft tissue density. Fourth, metallic implants such as total hip replacements can create artifacts that interfere with landmark identification. These factors compound measurement uncertainty and likely contribute to the higher error rates observed for PI across studies. Future deep learning models should specifically address these challenges, perhaps through specialized preprocessing steps or architectural modifications designed to better handle landmark obscurity and anatomical variations.

As this technology continues to evolve, it is highly unlikely that it will not play a role in patient healthcare. It is of great importance for future research to ensure adequate ethical standards, as new concepts and technologies are often met with some resistance. Issues with accountability, transparency, and permissions could come into question by involving deep learning in the decision-making process. Therefore, the integration of deep learning technology should come as a complementary tool in the surgical decision-making processes, where surgeons can potentially optimize patient care pathways and improve overall clinical outcomes.

Limitations

This review has certain limitations. The literature search was restricted to studies published in English, potentially excluding some relevant non-English studies. Searches were limited to four databases, although additional sources were hand-searched. Study screening and data extraction were performed by only two reviewers. The meta-analysis combined studies using different deep learning architectures and imaging modalities, which may have introduced heterogeneity. Only mean absolute errors and correlation coefficients were synthesized, although various other accuracy metrics were reported in the studies.

An additional limitation that should be taken into consideration is that the included studies did not account for anatomic variations such as LSTV. The prevalence of LSTV varies widely within the literature, ranging anywhere from 3.3% to 35.6%. A recent study by Khalifé *et al.* demonstrated that patients with low-grade LSTV, defined as Castelv I and II, have similar alignments as PI-matched no-LSTV and, therefore, should have their measurements taken from S1. Patients with high-grade LSTV, defined as Castelv III and IV, have more kyphotic L5-S1 segments with more cranial lumbar apex and thoracolumbar inflection point and, therefore, should have their measurements taken from L5. Future studies involving machine learning models for measuring spinopelvic parameters may have to pre-identify patients with LSTV and manually input the starting point to account for these anatomic variations^[51].

Conclusion

In conclusion, the breadth of imaging, network architecture details, spine pathologies, and statistical validation encompassed within these studies support automated measurement of spinal curvature as viable for clinical integration pending minor reporting enhancements. Multicenter datasets and model access could additionally reinforce external validity and enable incremental developments in this space.

Overall, this review supports deep learning as a potentially transformative technique for automated spinopelvic measurement from radiographs pending rigorous multicenter validation. These AI technologies may eventually improve efficiency, accuracy, and reliability for quantitative spine image analysis.

DECLARATIONS

Authors' contributions

Manuscript writing and revision: Glaser D

Data collection, analysis, and manuscript revision: AlMekkawi AK, Caruso JP

Data contribution and manuscript revision: Chung CY, Khan EZ, Daadaa HM

Conceptualization, progress monitoring, and final manuscript revision: Aoun SG, Bagley CA

Availability of data and materials

The data are available from the corresponding author upon reasonable request.

Financial support and sponsorship

None.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2025.

REFERENCES

1. Le Huec JC, Thompson W, Mohsinaly Y, Barrey C, Faundez A. Sagittal balance of the spine. *Eur Spine J.* 2019;28:1889-905. DOI PubMed
2. Vrtovec T, Janssen MM, Likar B, Castelein RM, Viergever MA, Pernuš F. A review of methods for evaluating the quantitative parameters of sagittal pelvic alignment. *Spine J.* 2012;12:433-46. DOI PubMed
3. Legaye J, Duval-Beaupère G, Hecquet J, Marty C. Pelvic incidence: a fundamental pelvic parameter for three-dimensional regulation of spinal sagittal curves. *Eur Spine J.* 1998;7:99-103. DOI PubMed PMC
4. Glassman SD, Bridwell K, Dimar JR, Horton W, Berven S, Schwab F. The impact of positive sagittal balance in adult spinal deformity. *Spine.* 2005;30:2024-9. DOI PubMed
5. Maillot C, Ferrero E, Fort D, Heyberger C, Le Huec JC. Reproducibility and repeatability of a new computerized software for sagittal spinopelvic and scoliosis curvature radiologic measurements: Keops®. *Eur Spine J.* 2015;24:1574-81. DOI PubMed
6. Lafage R, Ferrero E, Henry JK, et al. Validation of a new computer-assisted tool to measure spino-pelvic parameters. *Spine J.* 2015;15:2493-502. DOI PubMed
7. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal.* 2017;42:60-88. DOI
8. Shen D, Wu G, Suk HI. Deep learning in medical image analysis. *Annu Rev Biomed Eng.* 2017;19:221-48. DOI PubMed PMC
9. Galbusera F, Casaroli G, Bassani T. Artificial intelligence and machine learning in spine research. *JOR Spine.* 2019;2:e1044. DOI PubMed PMC
10. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521:436-44. DOI PubMed
11. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. *Z Med Phys.* 2019;29:102-27. DOI PubMed
12. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health.* 2019;1:e271-97. DOI
13. Chea P, Mandell JC. Current applications and future directions of deep learning in musculoskeletal radiology. *Skeletal Radiol.* 2020;49:183-97. DOI PubMed
14. Jamaludin A, Lootus M, Kadir T, et al; Genodisc Consortium. ISSLS PRIZE IN BIOENGINEERING SCIENCE 2017: automation of reading of radiological features from magnetic resonance images (MRIs) of the lumbar spine without human intervention is comparable with an expert radiologist. *Eur Spine J.* 2017;26:1374-83. DOI
15. Lopez CD, Boddapati V, Lombardi JM, et al. Artificial learning and machine learning applications in spine surgery: a systematic

- review. *Global Spine J*. 2022;12:1561-72. [DOI](#) [PubMed](#) [PMC](#)
16. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Int J Surg*. 2021;88:105906. [DOI](#) [PubMed](#)
 17. Cabitza F, Campagner A. The IJMEDI checklist for assessment of medical AI. *Int J Med Inform*. 2021;153. [DOI](#)
 18. Chae DS, Nguyen TP, Park SJ, Kang KY, Won C, Yoon J. Decentralized convolutional neural network for evaluating spinal deformity with spinopelvic parameters. *Comput Methods Programs Biomed*. 2020;197:105699. [DOI](#) [PubMed](#)
 19. Wu H, Bailey C, Rasoulinejad P, Li S. Automated comprehensive adolescent idiopathic scoliosis assessment using MVC-Net. *Med Image Anal*. 2018;48:1-11. [DOI](#) [PubMed](#)
 20. Wang L, Xu Q, Leung S, Chung J, Chen B, Li S. Accurate automated Cobb angles estimation using multi-view extrapolation net. *Med Image Anal*. 2019;58:101542. [DOI](#)
 21. Zhang K, Xu N, Guo C, Wu J. MPF-net: an effective framework for automated cobb angle estimation. *Med Image Anal*. 2022;75:102277. [DOI](#)
 22. Zerouali M, Parpaleix A, Benbakoura M, Rigault C, Champsaur P, Guenoun D. Automatic deep learning-based assessment of spinopelvic coronal and sagittal alignment. *Diagn Interv Imaging*. 2023;104:343-50. [DOI](#) [PubMed](#)
 23. Korez R, Putzier M, Vrtovec T. A deep learning tool for fully automated measurements of sagittal spinopelvic balance from X-ray images: performance evaluation. *Eur Spine J*. 2020;29:2295-305. [DOI](#) [PubMed](#)
 24. Kim YT, Jeong TS, Kim YJ, Kim WS, Kim KG, Yee GT. Automatic spine segmentation and parameter measurement for radiological analysis of whole-spine lateral radiographs using deep learning and computer vision. *J Digit Imaging*. 2023;36:1447-59. [DOI](#) [PubMed](#) [PMC](#)
 25. Yeh YC, Weng CH, Huang YJ, Fu CJ, Tsai TT, Yeh CY. Deep learning approach for automatic landmark detection and alignment analysis in whole-spine lateral radiographs. *Sci Rep*. 2021;11:7618. [DOI](#) [PubMed](#) [PMC](#)
 26. Orosz LD, Bhatt FR, Jazini E, et al. Novel artificial intelligence algorithm: an accurate and independent measure of spinopelvic parameters. *J Neurosurg Spine*. 2022;37:893-901. [DOI](#)
 27. Gami P, Qiu K, Kannappan S, et al. Semiautomated intraoperative measurement of Cobb angle and coronal C7 plumb line using deep learning and computer vision for scoliosis correction: a feasibility study. *J Neurosurg Spine*. 2022;37:713-21. [DOI](#)
 28. Schwartz JT, Cho BH, Tang P, et al. Deep learning automates measurement of spinopelvic parameters on lateral lumbar radiographs. *Spine*. 2021;46:E671-8. [DOI](#)
 29. Aubert B, Vazquez C, Cresson T, Parent S, de Guise JA. Toward automated 3D spine reconstruction from biplanar radiographs using CNN for statistical spine model fitting. *IEEE Trans Med Imaging*. 2019;38:2796-806. [DOI](#) [PubMed](#)
 30. Nguyen TP, Jung JW, Yoo YJ, Choi SH, Yoon J. Intelligent evaluation of global spinal alignment by a decentralized convolutional neural network. *J Digit Imaging*. 2022;35:213-25. [DOI](#) [PubMed](#) [PMC](#)
 31. Galbusera F, Niemeier F, Wilke HJ, et al. Fully automated radiological analysis of spinal disorders and deformities: a deep learning approach. *Eur Spine J*. 2019;28:951-60. [DOI](#)
 32. Zhang T, Zhu C, Lu Q, Liu J, Diwan A, Cheung JPY. A novel tool to provide predictable alignment data irrespective of source and image quality acquired on mobile phones: what engineers can offer clinicians. *Eur Spine J*. 2020;29:387-95. [DOI](#) [PubMed](#)
 33. Horng MH, Kuok CP, Fu MJ, Lin CJ, Sun YN. Cobb angle measurement of spine from X-ray images using convolutional neural network. *Comput Math Methods Med*. 2019;2019:6357171. [DOI](#) [PubMed](#) [PMC](#)
 34. Sun H, Zhen X, Bailey C, Rasoulinejad P, Yin Y, Li S. Direct estimation of spinal Cobb angles by structured multi-output regression. In: Niethammer M, Styner M, Aylward S, Zhu H, Oguz I, Yap P, Shen D, editors. Information processing in medical imaging. Cham: Springer International Publishing; 2017. pp. 529-40. [DOI](#)
 35. Weng CH, Wang CL, Huang YJ, et al. Artificial intelligence for automatic measurement of sagittal vertical axis using ResUNet framework. *J Clin Med*. 2019;8:1826. [DOI](#) [PubMed](#) [PMC](#)
 36. H A, Prabhu GK. Automatic quantification of spinal curvature in scoliotic radiograph using image processing. *J Med Syst*. 2012;36:1943-51. [DOI](#) [PubMed](#)
 37. Zhang J, Lou E, Le LH, Hill DL, Raso JV, Wang Y. Automatic Cobb measurement of scoliosis based on fuzzy Hough Transform with vertebral shape prior. *J Digit Imaging*. 2009;22:463-72. [DOI](#) [PubMed](#) [PMC](#)
 38. Sounderajah V, Ashrafian H, Golub RM, et al; STARD-AI Steering Committee. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open* 2021;11:e047709. [DOI](#) [PubMed](#) [PMC](#)
 39. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Lancet Digit Health*. 2020;2:e537-48. [DOI](#) [PubMed](#) [PMC](#)
 40. Mongan J, Moy L, Kahn CE Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell*. 2020;2:e200029. [DOI](#) [PubMed](#) [PMC](#)
 41. Willemink MJ, Koszek WA, Hardell C, et al. Preparing medical imaging data for machine learning. *Radiology*. 2020;295:4-15. [DOI](#) [PubMed](#) [PMC](#)
 42. Ghaednia H, Lans A, Sauder N, et al. Deep learning in spine surgery. *Semin Spine Surg*. 2021;33:100876. [DOI](#)
 43. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17:195. [DOI](#) [PubMed](#) [PMC](#)
 44. Cho BH, Kaji D, Cheung ZB, et al. Automated measurement of lumbar lordosis on radiographs using machine learning and computer

- vision. *Global Spine J.* 2020;10:611-8. [DOI](#) [PubMed](#) [PMC](#)
45. Burns JE, Yao J, Chalhoub D, Chen JJ, Summers RM. A machine learning algorithm to estimate sarcopenia on abdominal CT. *Acad Radiol.* 2020;27:311-20. [DOI](#) [PubMed](#)
 46. Yang J, Zhang K, Fan H, et al. Development and validation of deep learning algorithms for scoliosis screening using back images. *Commun Biol.* 2019;2:390. [DOI](#) [PubMed](#) [PMC](#)
 47. Hu B, Kim C, Ning X, Xu X. Using a deep learning network to recognise low back pain in static standing. *Ergonomics.* 2018;61:1374-81. [DOI](#) [PubMed](#)
 48. Staartjes VE, de Wispelaere MP, Vandertop WP, Schröder ML. Deep learning-based preoperative predictive analytics for patient-reported outcomes following lumbar discectomy: feasibility of center-specific modeling. *Spine J.* 2019;19:853-61. [DOI](#) [PubMed](#)
 49. Hines AL, Barrett ML, Jiang HJ, Steiner CA. Conditions with the largest number of adult hospital readmissions by payer, 2011. In: Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. [PubMed](#)
 50. Valenzuela JG, Cirillo Toteria JI, Turkieltaub DH, Echaurren CV, Álvarez Lemos FL, Arriagada Ramos FI. Spino-pelvic radiological parameters: comparison of measurements obtained by radiologists using the traditional method versus spine surgeons using a semi-automated software (Surgimap). *Acta Radiol Open.* 2023;12:20584601231177404. [DOI](#) [PubMed](#) [PMC](#)
 51. Khalifé M, Lafage R, Daniels AH, et al; International Spine Study Group. Assessing abnormal proximal junctional angles in adult spinal deformity: a normative data approach to define proximal junctional kyphosis. *Spine.* 2025;50:103-9. [DOI](#) [PubMed](#)