

Perspective

Open Access



Redefining precision: the current and future roles of artificial intelligence in spine surgery

Ryan W. Turlip, Harmon S. Khela, Mert Marcel Dagli, Daksh Chauhan, Yohannes Ghenbot, Hasan S. Ahmad, Jang W. Yoon

Department of Neurosurgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.

Correspondence to: Dr. Jang W. Yoon, Department of Neurosurgery, Perelman School of Medicine, University of Pennsylvania, 800 Spruce Street, Philadelphia, PA, 19107 USA. E-mail: jang.yoon@penmedicine.upenn.edu

How to cite this article: Turlip RW, Khela HS, Dagli MM, Chauhan D, Ghenbot Y, Ahmad HS, Yoon JW. Redefining precision: the current and future roles of artificial intelligence in spine surgery. *Art Int Surg* 2024;4:324-30. <https://dx.doi.org/10.20517/ais.2024.29>

Received: 17 May 2024 **First Decision:** 23 Sep 2024 **Revised:** 2 Oct 2024 **Accepted:** 15 Oct 2024 **Published:** 24 Oct 2024

Academic Editor: Rafael D. de la Garza-Ramos **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

The integration of artificial intelligence (AI) into spine surgery presents a transformative approach to preoperative and postoperative care paradigms. This paper explores the application of AI within spine surgery, focusing on diagnostic and predictive applications. AI-driven analysis of radiographic images, facilitated by machine learning (ML) algorithms such as convolutional neural networks (CNNs), potentially promises enhanced accuracy in identifying spinal pathologies and deformities; by combining these techniques with patient-specific data, predictive modeling can guide and inform diagnosis, prognosis, surgery selection, and treatment. Postoperatively, AI can leverage data from digital wearable technology to enhance the quantity and quality of outcome measures surgeons use to define and understand treatment success or failure. Still, challenges such as internal and external validation of AI models remain pertinent. Future directions include incorporating continuous variables from digital biomarkers and standardizing reporting metrics for AI studies in spine surgery. As AI continues to evolve, transparent validation frameworks and adherence to reporting guidelines will be crucial for its successful integration into clinical practice.

Keywords: Artificial intelligence, adult spinal deformity, radiographic imaging, machine learning, predictive modeling, objective outcomes



© The Author(s) 2024. **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.



INTRODUCTION

The integration of artificial intelligence (AI) into spine surgery has given rise to significant improvements in patient safety, peri-operative decision making, and clinical outcomes^[1]. As new technological innovations herald faster, more efficient, and more accurate AI models, it is imperative for surgeons to understand the impact of AI on current treatment paradigms and where spine surgeons' focus should lie as we assist in the development of AI-enabled personalized and precision medicine.

At the cornerstone of clinical advancement with AI are machine learning (ML) models, capable of identifying and extracting patterns from large datasets and making predictions based on learned trends. As the availability of data grows, ML model performance continues to improve; therefore, the advancement of AI in medicine is uniquely tied to our ability to provide these models with accurate and pertinent datapoints. In this perspective, we provide a brief historical outline of current ML and AI applications in spine surgery. We then offer our thoughts on where the future of AI and spine surgery lies, and how the unique relationship between model accuracy and data volume will shape the future of how AI is implemented in clinical contexts.

CURRENT AI APPLICATIONS IN SPINE SURGERY

One of the earliest and most compelling uses of ML in spine surgery has been the use of models to automatically decipher radiographic images. For example, the classification of lumbar disc degeneration from 2-dimensional magnetic resonance image (MRI) using ML has now reached levels comparable to expert radiologists^[1-3]. The morphology of the discs is first described according to their pathological features and classified according to the standardized grading system proposed by Pfirrmann *et al.*^[4]. A convolutional neural network (CNN) is then used to extract image features from the training data set to make predictions based on the radiologists' interpretations. CNNs, a specialized subtype of deep learning (DL) algorithms, parallel the architecture of human visual cortex processing and rely on unsupervised pattern recognition to classify images. CNN-based models for image classification are typically validated through a combination of k-fold cross-validation on training data and then tested on independent and external datasets to ensure generalizability. Other groups have also explored the use of generative models to create image-to-image translations of the musculoskeletal system^[5,6]. Clinically, this can provide a means to correct poor image resolution or blurriness due to patient motion during image acquisition.

As DL algorithms became more prevalent, they have gradually been implemented to automatically determine spinal landmarks to calculate deformity parameters. DL models are trained on large datasets to identify and classify complex phenomena through non-linear analysis in artificial neurons, similar in structure to the mammalian brain^[7]. The automated analysis of the Cobb angle to describe the severity of scoliotic curvature has been addressed through several DL techniques^[8-10]. Korez *et al.* also used DL to identify anatomical landmarks in X-ray images and measure spinopelvic parameters, finding no difference between DL and manual identification^[11].

The transformative capability of AI can expedite diagnosis and treatment planning, and has the potential to standardize surgical treatment strategies for various spinal pathologies after taking patient-specific factors into account. Widespread implementation, however, faces substantial ethical challenges as the prospect of removing human interpretation may lead to more patient distrust in conclusions. It is unlikely, then, that human radiologists will be replaced by AI technology; instead, their diagnostic accuracy will be improved as models continue to advance.

The advent of AI-powered predictive modeling also holds immense promise in the realm of personalized precision medicine. By assimilating vast repositories of patient data, including demographic information, comorbidities, and procedural specifics, AI algorithms can generate prognostic models tailored to individual patients, ushering in a new era where therapeutic decisions are guided by each patient's unique physiology. This is particularly important for patient risk stratification, where clinical variables can be used as inputs (predictors) for the potential of operative complications. Pellisé *et al.* trained a random forest algorithm with clinical variables from 1,612 patients with adult spinal deformity (ASD) and identified age, surgical invasiveness, and deformity magnitude as potential risk factors for major complications^[12]. Predictive models, such as random forest algorithms for complication risk stratification, undergo internal validation through cross-validation and are, at times, externally validated using datasets from different clinical settings to evaluate model transferability. In the study by Pellisé *et al.*, internal validation was performed with an 80%/20% split between training/testing groups, measuring model performance through the observed area under the receiver operating characteristic curve (AUC) and the Brier score^[12]. Ames *et al.* augmented this approach by applying unsupervised hierarchical clustering to classify ASD based on patient demographics and radiographic measurements with the goal of constructing a risk-benefit grid as a preoperative tool for decision making^[13].

Current work continues to build upon existing outcomes prediction and postoperative prognostication. ML has been implemented to assess the likelihood of surgical site infection, major intra-operative complications, hospital length of stay, or the necessity of blood transfusion after surgery^[14-17]. This has led to the development of universal prediction models trained retrospectively on large patient registries, such as the American College of Surgeons National Surgical Quality Improvement Project (ACS-NSQIP) database. The ACS-NSQIP developed an online calculator for morbidity and mortality risk, but reports demonstrated poor predictive performance^[18]. Other groups have used the available ACS-NSQIP patient data as a resource to train their own models, with early indications of clinical efficacy at predicting outcomes^[19,20]. Fully unsupervised models have extensive utility to revolutionize personalized care and stratify risk; however, deploying under-validated AI tools can lead to inaccurate diagnoses or inappropriate treatment recommendations, so caution is needed to ensure patient safety.

Lastly, an emerging implementation of ML and AI has been in the realm of outcomes assessment. Traditionally, evaluation of surgical outcomes relies on physician interpretation of radiographic imaging combined with patient questionnaires assessing changes in patient mobility, pain, and quality of life. These patient-reported outcome measures (PROMs) offer valuable insight into patients' own interpretation of their health status and physical function. However, these methods contain inherent subjectivity and often lack the precision and reliability needed for precise and actionable insights^[21,22]. More recently, there has also been a trend toward utilizing digital biomarkers and data-driven outcomes measurements in conjunction with traditional PROMs. Objective measurements of patient mobility obtained from patient smartphones, smartwatches, or other biometric wearables can add additional unbiased insight into patient function^[23-26]. The quantitative and continuous features of these data are well suited for integration with data-driven statistical and ML techniques, and they have enabled surgeons to better quantify changes in patient mobility after surgery and to predict which patients may be better suited to recover from a particular pathology^[24,25].

FUTURE DIRECTIONS

The use of accelerometer and GPS information is a relatively novel concept, and more complex ML predictive models have yet to be applied. The incorporation of such models could significantly improve the accuracy of patient assessments by providing real-time, continuous data that captures a patient's functional mobility in their everyday life. This can lead to a more detailed understanding of a patient's functional

baseline status and postoperative recovery, resulting in tailored personalized medicine. While many analyses of mobility data have been retrospective in nature, upon the growth of adequate datasets, predictive models may be able to accurately identify subtle changes in mobility-related complications or improvements earlier than would be possible with traditional assessments.

Further, advanced mobility metrics can add potential value for patient prognostication. As previously mentioned, groups are beginning to engineer universal prognostic models for outcome prediction trained on large data registries^[19,20]. Although still in their infancy, accurate prognostic models could transform patient management by offering more realistic recovery trajectories, customizing patient care, or identifying high risk for adverse outcomes. There are still challenges that limit the widespread implementation of such models, ranging from access to generalizable datasets, cost-effectiveness for stable implementation, or ethical concerns.

Mobility metrics are not the only AI application that is challenged with limited data availability. Access to high-quality, standardized data sets is one of the greatest challenges to overall AI and ML implementation, especially within spine surgery, given the varied and nuanced model inputs spanning complex patient presentations, operative courses, and radiographic imaging. To address this challenge, there is a growing movement toward the creation of standardized, multi-center datasets that include patients from several geographic areas and socioeconomic groups. Other groups such as the ACS are refining their existing patient registries to integrate additional data from the electronic health record. Together, these datasets and registries aim to provide a foundation for training more accurate and generalizable AI models that can be deployed across various clinical settings.

Patient selection is another area of current clinical practice that stands to benefit from future AI and ML integration. The art of understanding which patients will benefit from certain procedures is not easily replicated with frameworks and rules that can be directly input into computerized programs. However, as CNNs and ML algorithms continue to grow in computational ability, they can potentially identify relationships between datapoints that are otherwise unnoticeable to the un-aided human mind; in this way, future AI and ML models can augment surgeons' clinical practice and assist in identifying certain patient characteristics that are indicative of patients likely to benefit from specific surgical interventions.

While AI technologies like predictive modeling and image analysis hold promise in decision making, their potential intra-operative impact is already apparent^[1,7]. AI-assisted intra-operative tools, such as robotics, navigation systems, and mixed reality, have the potential to significantly enhance the surgeon's ability to execute procedures with high precision, particularly in minimally invasive and percutaneous surgeries. These technologies allow for real-time guidance and adjustment during complex procedures, reducing the margin of error. However, while AI can minimize the risk of intra-operative errors, it cannot fully replace the human element of adaptability and judgment. Surgeons must remain vigilant in managing unforeseen intra-operative variables and complications, as AI systems, though highly advanced, still require human oversight to ensure patient safety and the proper handling of unexpected challenges.

Although surgeon experience is regarded as a significant factor in decision making, there have been attempts to apply mathematical and data-driven approaches to surgical decision making^[27]. Lewandrowski *et al.* recently used the Rasch model to determine the choice of procedure for endoscopic lumbar decompression^[27]. The Rasch model is a logistic function analyzing categorical data, such as questionnaire responses, to find the relative difficulty of a task, and it has been widely established in education, marketing, and health economics^[28]. However, it was found that there was still disagreement among surgeons regarding

the ability to achieve adequate clinical outcomes, indicating that increased granularity through additional metrics is needed to overcome the disordered responses^[27].

Despite the promising advancements of AI in spine surgery, a significant limitation in the current literature is the lack of external validation of many studies. Most models are only internally validated on the same data from which they were derived, raising concerns about model generalizability to larger patient populations or different clinical settings. It was estimated that only 5% of published articles on prognostic models included an external validation framework^[29]. Without external validation, it is difficult to ensure that these AI models will perform reliably in diverse environments, further limiting their clinical application. This issue is compounded by the scarcity of randomized controlled trials (RCTs) investigating AI in spine surgery, which are essential for evaluating long-term effectiveness and accuracy.

Due to the lack of standardized reporting metrics for AI studies, it is imperative to create clear guidelines through which the risk of bias and the potential utility of these models can be evaluated. AI studies that focus primarily on diagnostic applications using medical imaging should adhere to the Checklist for Artificial Intelligence in Medical Imaging (CLAIM)^[30]. The forthcoming Standards for Reporting of Diagnostic Accuracy Studies for AI (STARD-AI), an AI-specific adaptation of the established STARD guidelines, is also under development. Upon its release, it is expected to be indexed on the Enhancing the QUALity and Transparency Of health Research (EQUATOR) Network, addressing similar methodological issues as those covered by CLAIM^[31].

For ML multivariable prediction models, whether diagnostic or prognostic, the recently published Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis + Artificial Intelligence (TRIPOD + AI) provides a structured protocol for reporting predictive algorithms^[32]. Despite the advancements since the initial 2015 TRIPOD statement, which has shown promise in improving methodological transparency^[32,33], substantial gaps persist that hinder the broader integration of AI in clinical practice^[34]. As AI prediction algorithms become more pervasive in spine surgery, internal and external validation frameworks are necessary to appraise model performance, ensuring the variability in different patient populations is reflected to enhance surgical precision.

CONCLUSION

The integration of AI and ML into spine surgery represents a transformative shift toward precision medicine, offering enhanced diagnostic and prognostic capabilities. With the advances in automated radiographic imaging, patient risk stratification, outcomes prediction, and personalized medicine, future work promises to tailor treatment to individual patients more accurately. Despite the promising achievements so far, the field must address challenges in data accuracy by expanding training datasets and implementing robust validation frameworks. As AI becomes more prevalent in spine surgery, successful integration has the power to refine surgical decision making and improve patient outcomes.

DECLARATIONS

Authors' contributions

Original draft preparation, methodology, conceptualization: Turlip RW

Original draft presentation, conceptualization: Khela HS

Review and editing, supervision: Dagli MM, Ghenbot Y, Ahmad HS

Review and editing, validation: Chauhan D

Review and editing, supervision, conceptualization: Yoon JW

Critical writing: Turlip RW, Khela HS, Dagli MM, Chauhan D, Ghenbot Y, Ahmad HS, Yoon JW

Availability of data and materials

Not applicable.

Financial support and sponsorship

None.

Conflicts of interest

All authors declared that there are no conflicts of interest. The Academic Editor Rafael D. de la Garza-Ramos will no longer serve as the journal's academic editor upon the expiration of his term in December 2024.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright REFERENCES

© The Author(s) 2024.

- Galbusera F, Casaroli G, Bassani T. Artificial intelligence and machine learning in spine research. *JOR Spine* 2019;2:e1044. DOI PubMed PMC
- Ruiz-España S, Arana E, Moratal D. Semiautomatic computer-aided classification of degenerative lumbar spine disease in magnetic resonance imaging. *Comput Biol Med* 2015;62:196-205. DOI PubMed
- 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 PubMed
- Pfirrmann CW, Metzdorf A, Zanetti M, Hodler J, Boos N. Magnetic resonance classification of lumbar intervertebral disc degeneration. *Spine* 2001;26:1873-8. DOI PubMed
- Johnson PM, Drangova M. Conditional generative adversarial network for 3D rigid-body motion correction in MRI. *Magn Reson Med* 2019;82:901-10. DOI PubMed
- Galbusera F, Bassani T, Casaroli G, et al. Generative models: an upcoming innovation in musculoskeletal radiology? A preliminary test in spine imaging. *Eur Radiol Exp* 2018;2:29. DOI PubMed PMC
- Charles YP, Lamas V, Ntilikina Y. Artificial intelligence and treatment algorithms in spine surgery. *Orthop Traumatol Surg Res* 2023;109:103456. DOI PubMed
- 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
- Zhang J, Li H, Lv L, Zhang Y. Computer-aided cobb measurement based on automatic detection of vertebral slopes using deep neural network. *Int J Biomed Imaging* 2017;2017:9083916. DOI PubMed PMC
- 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
- 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
- Pellisé F, Serra-Burriel M, Smith JS, et al; International Spine Study Group, European Spine Study Group. Development and validation of risk stratification models for adult spinal deformity surgery. *J Neurosurg Spine* 2019;31:587-99. DOI PubMed
- Ames CP, Smith JS, Pellisé F, et al; European Spine Study Group, International Spine Study Group. Artificial intelligence based hierarchical clustering of patient types and intervention categories in adult spinal deformity surgery: towards a new classification scheme that predicts quality and value. *Spine* 2019;44:915-26. DOI
- Lee MJ, Cizik AM, Hamilton D, Chapman JR. Predicting surgical site infection after spine surgery: a validated model using a prospective surgical registry. *Spine J* 2014;14:2112-7. DOI PubMed
- Scheer JK, Smith JS, Schwab F, et al; International Spine Study Group. Development of a preoperative predictive model for major complications following adult spinal deformity surgery. *J Neurosurg Spine* 2017;26:736-43. DOI PubMed

16. Durand WM, DePasse JM, Daniels AH. Predictive modeling for blood transfusion after adult spinal deformity surgery: a tree-based machine learning approach. *Spine* 2018;43:1058-66. [DOI](#) [PubMed](#)
17. Wang SK, Wang P, Li ZE, et al. Development and external validation of a predictive model for prolonged length of hospital stay in elderly patients undergoing lumbar fusion surgery: comparison of three predictive models. *Eur Spine J* 2024;33:1044-54. [DOI](#) [PubMed](#)
18. Sebastian A, Goyal A, Alvi MA, et al. Assessing the performance of national surgical quality improvement program surgical risk calculator in elective spine surgery: insights from patients undergoing single-level posterior lumbar fusion. *World Neurosurg* 2019;126:e323-9. [DOI](#) [PubMed](#)
19. Goyal A, Ngufor C, Kerezoudis P, McCutcheon B, Storlie C, Bydon M. Can machine learning algorithms accurately predict discharge to nonhome facility and early unplanned readmissions following spinal fusion? Analysis of a national surgical registry. *J Neurosurg Spine* 2019;31:568-78. [DOI](#) [PubMed](#)
20. Broda A, Sanford Z, Turcotte J, Patton C. Development of a risk prediction model with improved clinical utility in elective cervical and lumbar spine surgery. *Spine* 2020;45:E542-51. [DOI](#) [PubMed](#)
21. Churrua K, Pomare C, Ellis LA, et al. Patient-reported outcome measures (PROMs): a review of generic and condition-specific measures and a discussion of trends and issues. *Health Expect* 2021;24:1015-24. [DOI](#) [PubMed](#) [PMC](#)
22. Dansie EJ, Turk DC. Assessment of patients with chronic pain. *Br J Anaesth* 2013;111:19-25. [DOI](#) [PubMed](#) [PMC](#)
23. Mobbs RJ. From the subjective to the objective era of outcomes analysis: how the tools we use to measure outcomes must change to be reflective of the pathologies we treat in spinal surgery. *J Spine Surg* 2021;7:456-7. [DOI](#) [PubMed](#) [PMC](#)
24. Ahmad HS, Yang AI, Basil GW, et al. Developing a prediction model for identification of distinct perioperative clinical stages in spine surgery with smartphone-based mobility data. *Neurosurgery* 2022;90:588-96. [DOI](#) [PubMed](#)
25. Chauhan D, Ahmad HS, Subtirelu R, et al. Defining the minimal clinically important difference in smartphone-based mobility after spine surgery: correlation of survey questionnaire to mobility data. *J Neurosurg Spine* 2023;39:427-37. [DOI](#) [PubMed](#)
26. Boaro A, Leung J, Reeder HT, et al. Smartphone GPS signatures of patients undergoing spine surgery correlate with mobility and current gold standard outcome measures. *J Neurosurg Spine* 2021;35:796-806. [DOI](#) [PubMed](#) [PMC](#)
27. Lewandrowski KU, Alvim Fiorelli RK, Pereira MG, et al. Polytomous rasch analyses of surgeons' decision-making on choice of procedure in endoscopic lumbar spinal stenosis decompression surgeries. *Int J Spine Surg* 2024;18:164-77. [DOI](#) [PubMed](#) [PMC](#)
28. Lorio M, Martinson M, Ferrara L. Paired comparison survey analyses utilizing rasch methodology of the relative difficulty and estimated work relative value units of CPT® code 72729. *Int J Spine Surg* 2016;10:40. [DOI](#) [PubMed](#) [PMC](#)
29. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J* 2021;14:49-58. [DOI](#) [PubMed](#) [PMC](#)
30. 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](#)
31. 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](#)
32. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024;385:e078378. [DOI](#) [PubMed](#) [PMC](#)
33. Collins GS, Reitsma JB, Altman DG, Moons KG; TRIPOD Group. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. The TRIPOD Group. *Circulation* 2015;131:211-9. [DOI](#) [PubMed](#) [PMC](#)
34. Zamanipoor Najafabadi AH, Ramspek CL, Dekker FW, et al. TRIPOD statement: a preliminary pre-post analysis of reporting and methods of prediction models. *BMJ Open* 2020;10:e041537. [DOI](#) [PubMed](#) [PMC](#)