ISSN: 2572-3243

Wenban AB, et al. J Musculoskelet Disord Treat 2026, 12:135
DOI: 10.23937/2572-3243.1510135
Volume 12 | Issue 1
Open Access

Journal of Musculoskeletal Disorders and Treatment

ORIGINAL RESEARCH ARTICLE

A Clinical Decision Framework for Evaluating Artificial Intelligence Tools in Spine Imaging: Ensuring Safe and Effective Conservative Care

Adrian B. Wenban[1]*, *Francine L. Denis*[2], *Sara M. Sandino*[3] *and Matisse N. Wenban*[4]

[1]*Cofounder, College of Advanced Healthcare Sciences, Spain*

[2]*Module Lead and Clinic Supervisor, Barcelona College of Chiropractic, Spain*

[3]*Private Practice, Rome Italy*

[4]*London, United Kingdom*

***Corresponding author:** *Adrian B. Wenban, Avenida Diagonal 355, 1, 2, 08037 Barcelona Spain, Tel: 00 34 637571210*

Abstract

Background: Artificial intelligence (AI) tools for spine imaging analysis are rapidly entering clinical practice, with many systems demonstrating impressive technical performance in detecting vertebral fractures, quantifying degenerative changes, and measuring alignment parameters. However, conservative spine care practitioners-including chiropractors, physiotherapists, and osteopaths-face unique challenges in evaluating whether these tools genuinely support patient care. Technical accuracy alone does not establish clinical utility, particularly when imaging findings correlate poorly with symptoms and functional outcomes. Furthermore, concerns persist regarding algorithmic bias, equity of access, deskilling risks, inadequate post-market surveillance, and potential for technological paternalism that undermines shared decision-making.

Methods: We synthesized evidence from recent systematic reviews of AI performance in spine imaging, implementation science frameworks, health equity literature, professional competency standards, and patient-centered care principles. Drawing on Kumar, et al.'s [1] comprehensive technical review and extending it with multidimensional considerations relevant to conservative practice, we developed a six-step clinical decision algorithm. This framework integrates critical appraisal domains spanning technical validity, clinical utility for functional outcomes, equity and bias, clinical reasoning support, safety governance, and patient-centeredness.

Results: The resulting algorithm guides practitioners through systematic evaluation of AI tools before adoption. Step 1 assesses technical validity and generalizability across relevant populations. Step 2 examines clinical utility specifically for conservative care contexts, requiring evidence linking AI outputs to functional outcomes rather than merely anatomical classifications. Step 3 evaluates equity dimensions, including dataset representativeness and performance across demographic subgroups. Step 4 considers whether the tool enhances clinical reasoning without promoting automation bias. Step 5 reviews governance mechanisms for continuous safety monitoring. Step 6 ensures patient-centered implementation supporting shared decision-making. Each step includes specific evidence requirements and decision criteria, with tools assigned to one of four adoption categories: Ready, Conditional, Research-Only, or Not Recommended.

Conclusion: This practice-based decision framework provides conservative practitioners with structured criteria for evaluating AI imaging tools before clinical deployment. By extending beyond technical performance metrics to encompass clinical utility for functional outcomes, equity considerations, clinical reasoning support, safety governance, and patient-centeredness, the algorithm addresses critical dimensions often overlooked in AI validation studies. Conservative practitioners can use this tool to make evidence-informed decisions that prioritize patient welfare and professional integrity over technological enthusiasm, ensuring AI adoption genuinely enhances rather than compromises the quality and equity of spine care delivery.

Keywords: artificial intelligence, spine imaging, clinical decision-making, manual therapy, chiropractic, physiotherapy, osteopathy, health equity, patient-centered care, diagnostic imaging

ClinMed International Library

Citation: Wenban AB, Denis FL, Sandino SM (2026) A Clinical Decision Framework for Evaluating Artificial Intelligence Tools in Spine Imaging: Ensuring Safe and Effective Conservative Care. J Musculoskelet Disord Treat 12:135. doi.org/10.23937/2572-3243.1510135

Received: January 19, 2026 : **Accepted:** February 09, 2026 : **Published:** February 12, 2026

Introduction

The integration of artificial intelligence (AI) into spine imaging interpretation represents a watershed moment for conservative musculoskeletal care. Deep learning algorithms now achieve diagnostic accuracy approaching or exceeding expert radiologists for vertebral fracture detection, disk degeneration classification, and spinal stenosis quantification [1]. Beyond imaging, these models show promise in enhancing clinical decision-making, providing treatment recommendations, and facilitating patient triage [2,3]. These advances have been accompanied by the rapid deployment of commercial AI systems despite concerning validation gaps [4,5].

Additionally, conservative spine and healthcare practitioners face a critical challenge: technical sophistication does not guarantee clinical value. Most AI systems are trained to detect structural abnormalities that frequently occur in asymptomatic populations and correlate poorly with patient-reported pain and functional limitation [6,7]. An algorithm achieving 95% accuracy in Pfirrmann grading may offer negligible clinical utility - or cause harm through medicalization - if such classifications fail to inform conservative treatment decisions or predict outcomes.

Furthermore, several concerning gaps persist in AI validation paradigms [6]. Many systems demonstrate performance degradation when deployed outside their training environments, with particular concerns about underrepresentation of racial minorities, older adults, and patients from community practice settings [8,9]. Post-market surveillance remains inconsistent, and mechanisms for detecting algorithmic bias or performance drift are underdeveloped [10]. Professional competency frameworks risk not keeping pace with technology, leaving practitioners uncertain about requisite skills for safe AI integration. Recently, the World Federation of Chiropractic emphasized the need for educational institutions to adapt curricula for technology integration, promote digital solutions for patient-centered care, and incorporate augmented reality into clinical training [11].

The rapidly advancing integration of AI into clinical decision making has been accompanied by a consensus emphasizing that responsibility for the judicious use of AI in patient care is a shared model involving several key actors, including AI developers, healthcare institutions and individual clinicians [12]. In response to these rapid developments, regulatory bodies have developed benchmarks for assessing AI models and monitoring their performance over time [13]. However, despite these efforts, a tangible gap exists between such guidelines and their practical implementation by clinicians upon whom professional bodies continue to place the ultimate responsibility for their judicious use.

With these concerns, evidence and statements in mind we have developed six simple but critical questions that we believe conservative healthcare practitioners must address before adopting AI imaging tools:

1) Has the tool been validated for conservative care contexts where functional outcomes matter more than anatomical classifications?

2) Will deployment exacerbate existing health disparities?

3) Does the tool enhance rather than replace clinical reasoning?

4) Are adequate governance mechanisms in place for continuous safety monitoring?

5) How does integration affect professional identity and scope of practice?

6) Does implementation support shared decision-making rather than technological paternalism?

This article presents a practical, stepwise algorithm integrating these considerations into a coherent decision framework. Healthcare practitioners can use this tool to systematically evaluate AI imaging systems before adoption, ensuring deployment genuinely enhances patient care rather than merely automating tasks of uncertain clinical value.

Methods

This clinical decision framework was developed through a narrative synthesis methodology combining evidence from multiple knowledge domains relevant to AI adoption in conservative spine care. The development process followed established guidelines for creating clinical decision tools and implementation frameworks [14,15]. We systematically identified and synthesized evidence across six core domains and we integrated the three-tier competency framework proposed by Cao, et al. [16], which ensured alignment between our evaluation framework and emerging professional standards. Drawing on Kumar et al.'s comprehensive technical review [1] as the foundation for imaging AI capabilities, we constructed a sequential six-step decision algorithm with a structure reflecting hierarchical decision-making. Finally, we developed four mutually exclusive adoption recommendations (Supplement I - Method Section (Details) - Available upon request).

Results

The six-step decision algorithm

Figure 1 presents the complete decision algorithm, which practitioners apply sequentially. Each step requires specific evidence before proceeding; deficiencies at any stage may warrant rejection or conditional adoption with safeguards. Detailed criteria for each step follow.

Step	Critical Question	Evidence Required to Proceed
1	Is the tool technically valid for my population?	External validation in ≥ 1 independent dataset; performance metrics stratified by age, sex, severity; evaluation in community practice settings; imaging protocol compatibility documented
2	Does it improve patient outcomes relevant to conservative care?	Evidence linking AI outputs to functional outcomes (pain, disability); impact on treatment decisions documented; evaluation in non-surgical populations; relationship between detected findings and clinical presentation established
3	Will it reduce or exacerbate health disparities?	Training data demographic composition reported; performance analyzed across race/ethnicity, socioeconomic status; validation in safety-net/community settings; accessibility barriers identified and addressed
4	Does it support or replace clinical reasoning?	Explainability features present; positioned as decision support not autonomous diagnosis; practitioner override mechanisms clear; training resources available; automation bias risks assessed
5	Are safety governance mechanisms adequate?	Regulatory clearance appropriate for intended use; post-market surveillance plans documented; adverse event reporting mechanisms exist; version control and update procedures clear; liability framework defined
6	Does it support patient-centered care?	Patient communication supports available; informed consent processes defined; integration with shared decision-making demonstrated; psychological impacts (anxiety, medicalization) considered

Figure 1: Six-step decision algorithm for evaluating AI tools in spine imaging for conservative care. Practitioners apply steps sequentially; deficiencies at any stage warrant caution or rejection.

Step-by-step application guide

Step 1: Technical validity and generalizability: Before considering any AI tool, practitioners must verify technical performance in populations resembling their own patients. Kumar, et al. [1] document impressive accuracies (sensitivity > 90%, specificity > 95%) for vertebral fracture detection across multiple CNNs, but note critical variability when algorithms encounter imaging protocols, equipment, or patient characteristics differing from training conditions [1]. For example, one FDA-cleared system showed 68% sensitivity in Danish populations versus expected 90+%, highlighting generalizability failures [17].

Required evidence includes: (a) external validation in ≥ 1 independent dataset; (b) performance stratified by age, sex, and disease severity; (c) evaluation in community practice or safety-net settings if relevant to your context; (d) imaging protocol compatibility documented. If validation evidence derives exclusively from academic medical centers using standardized protocols, exercise caution when deploying in community practices with heterogeneous equipment and scanning techniques.

Step 2: Clinical utility for functional outcomes: Technical accuracy detecting anatomical abnormalities does not establish clinical value for conservative care, where functional recovery often matters more than structural restoration. Degenerative disk changes, moderate stenosis, and disk bulges occur in 30–80% of asymptomatic adults depending on age [6,7]. An AI tool precisely identifying such findings may increase medicalization without improving outcomes.

Required evidence includes: (a) Demonstrated relationship between AI-detected findings and patient-reported outcomes (pain, function, health status, quality of life); (b) Impact on clinical decision-making documented (treatment selection, referral pathways); (c) Evaluation in conservative care populations, not only surgical candidates; (d) Analysis showing AI outputs inform prognosis or guide management. Absent such evidence, tools detecting findings of uncertain clinical significance should be approached skeptically.

Step 3: Equity and algorithmic fairness: Training datasets overwhelmingly represent patients from high-resource settings, with documented underrepresentation of racial and ethnic minorities [8,9]. Algorithms that perform well in academic centers may fail in community practices serving diverse populations. Moreover, dataset bias can perpetuate health inequities if certain demographic groups experience higher false-negative or false-positive rates.

Required evidence includes: (a) Demographic composition of training data reported and representative; (b) Performance analyzed across race/ethnicity and socioeconomic subgroups; (c) Validation in safety-net hospitals or community practices; (d) Consideration of accessibility barriers (cost, infrastructure requirements, and digital literacy). If subgroup performance data are absent or concerning, deployment risks exacerbating disparities.

Step 4: Clinical Reasoning and Human-AI Interaction: AI integration should enhance clinical reasoning without promoting 'automation bias' - uncritical acceptance of algorithmic outputs despite conflicting clinical information [18]. Conservative practitioners develop expertise integrating imaging findings with history, examination, and psychosocial factors; AI tools must support rather than undermine this synthesis.

Required evidence includes: (a) Explainability features enabling practitioners to understand AI reasoning; (b) Clear positioning as decision support, not autonomous diagnosis; (c) Straightforward override mechanisms preserving clinical judgment; (d) Training resources demonstrating achievable competence; (e) Assessment of automation bias risks. Systems functioning as 'black boxes' without interpretable outputs fail to support professional reasoning and should be avoided.

Step 5: Safety governance and post-market surveillance: Many AI systems secured regulatory clearance through pathways not requiring prospective clinical trials [10]. Post-market surveillance mechanisms often remain underdeveloped, creating risks of undetected performance degradation ('model drift') or accumulating evidence of harm.

Required evidence includes: (a) Regulatory clearance appropriate for intended use (not off-label deployment); (b) Post-market surveillance plans documenting ongoing performance monitoring; (c) Adverse event reporting systems accessible to end-users; (d) Version control and update procedures transparent; (e) Liability framework defining responsibilities. Absent robust governance, practitioners bear disproportionate medico-legal risk, including lifecycle management competencies for drift detection and version control [14].

Step 6: Patient-centeredness and shared decision-making: AI-generated imaging findings require thoughtful communication that contextualizes structural changes, manages expectations, and supports informed decision-making. Presenting algorithmic outputs as objective, unchallengeable facts risks both technological paternalism and automation bias and may increase patient anxiety or catastrophizing about findings of uncertain significance [19,20].

Required evidence includes: (a) Patient communication supports translating AI outputs into understandable language; (b) Informed consent processes addressing AI use; (c) Integration with shared decision-making tools; (d) Consideration of psychological impacts (anxiety, medicalization). Tools lacking patient-facing communication supports may compromise therapeutic alliance and informed consent.

Making the adoption decision

After systematic evaluation across all six steps, practitioners assign tools to adoption categories based on evidence adequacy:

Ready for adoption: Adequate evidence across all domains; demonstrated clinical utility improving functional outcomes; robust equity and safety data; clear implementation pathway. Proceed with confidence but maintain ongoing monitoring.

Conditional adoption: Strong evidence in some domains but gaps in others. May deploy with specific safeguards: enhanced monitoring, restricted use cases, systematic outcome tracking, commitment to discontinuation if harms emerge. Document decision-making rationale.

Research/Pilot Only: Insufficient evidence for routine clinical use. Confine deployment to controlled research settings with IRB oversight, informed consent emphasizing experimental nature, and rigorous outcome assessment. Generate evidence to inform future decisions.

Not recommended: Evidence of harm, unacceptable equity implications, or fundamental misalignment with conservative care principles. Reject deployment regardless of technical performance or commercial pressures.

Discussion

This decision algorithm addresses a critical gap in AI implementation guidance for healthcare providers who provide care related to the spine. Existing frameworks like the recently proposed three-tier competency framework by Cao, et al. [16] emphasize technical validation metrics (sensitivity, specificity, AUC) while overlooking dimensions central to conservative practice: clinical utility for functional outcomes, equity across diverse populations, support for clinical reasoning, safety governance, and patient-centered communication [16]. More specifically, the Cao, et al. [16] framework proposed a three-tier competency framework spanning foundational skills (safe use including output verification and privacy awareness), intermediate skills (bias detection and clinical integration), and advanced skills (governance and lifecycle management) while overlooking dimensions central to conservative practice: clinical utility for functional outcomes, equity across diverse populations, support for clinical reasoning, safety governance, and patient-centered communication. The complementary nature of our framework and that of Cao, et al. [16], suggests that healthcare providers evaluating the spine must not only critically evaluate AI tools using the six dimensions proposed here but also develop progressive competencies to deploy them responsibly. Educational institutions and professional organizations should consider integrating both evaluation literacy and usage competencies into undergraduate and graduate curricula and continuing education programs.

Our framework's sequential structure reflects hierarchical priorities. Technical validity (Step 1) constitutes necessary but insufficient grounds for adoption; tools failing external validation should not advance. Clinical utility for functional outcomes (Step 2) distinguishes conservative care contexts from radiology-centric applications - detecting structural changes matters only if such findings inform treatment decisions or predict outcome. Equity assessment (Step 3) protects vulnerable populations from algorithmic bias, recognizing that impressive overall performance may mask unacceptable disparities across demographic subgroups.

Steps 4-6 address implementation dimensions often neglected until post-deployment. Clinician competencies spanning safe use, evaluative proficiency, and governance [16] complement this evaluation framework. Clinical reasoning evaluation (Step 4) prevents deskilling and automation bias by ensuring tools genuinely support professional judgment. Safety governance assessment (Step 5) establishes accountability frameworks before adverse events occur. Patient-centeredness evaluation (Step 6) maintains therapeutic alliance and informed consent.

Our framework complements institutional governance approaches like those of Saenz, et al. [13], who developed comprehensive AI integration guidelines at Mass General Brigham with nine principles operationalized through multidisciplinary committees. Where Saenz, et al. address organizational-level governance, our framework enables individual practitioners to evaluate AI tools before adoption, particularly when institutional governance structures are absent.

Several limitations warrant acknowledgment. The algorithm cannot compensate for absent evidence; many AI tools lack adequate validation across multiple domains, leaving practitioners balancing potential benefits against uncertainty. The framework emphasizes conservative care priorities, which may differ from surgical or emergency contexts. Beyond initial validation, ongoing model lifecycle management remains critical as AI systems adapt over time [16]. Despite these limitations, the algorithm provides structured decision-making guidance previously unavailable to conservative practitioners.

Conclusion

AI technologies hold genuine promise for enhancing spine care by improving advanced imaging evaluation and by the detection of serious pathology, reduced radiation exposure, structural/functional analysis and support for less-experienced practitioners in under-resourced settings. Realizing this promise requires that technical development proceed alongside rigorous validation in relevant clinical contexts, proactive attention to equity, thoughtful integration preserving clinical reasoning, robust safety governance, and unwavering commitment to patient-centered care - dimensions often overlooked in AI validation.

This six-step algorithm provides conservative practitioners a practical tool for evaluating AI imaging systems, beyond technical metrics, before adoption. Implementation requires complementary development of student and clinician competencies spanning safe use, evaluation, and governance [21]. This integrated approach ensures AI adoption genuinely enhances spine care quality, equity and patient outcomes.

Author Contributions

ABW and MNW conceived the project,

ABW, SMS and FD contributed to the design or the framework,

All authors produced and provided feedback regarding drafts of the manuscript,

All authors reviewed and approved AI suggestions and the final manuscript.

AI Statement

The authors acknowledge that AI was used in the writing of this paper and that the assist in writing went beyond a check of grammar and spelling. Specifically, AI (Claude.ai) was used to review and critique the paper after it had been developed, researched, written and spell checked by all authors. AI suggestions, if authors agreed to integrate into the original paper, were then checked and validated by all authors.

Funding

None to declare.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Kumar R, Dougherty C, Sporn K, Khanna A, Ravi P, et al. (2025) Intelligence architectures and machine learning applications in contemporary spine care. Bioengineering 12: 967.

2. Clark K, Vendt B, Smith K, Freymann J, Kirby J, et al. (2013) The cancer imaging archive (TCIA): Maintaining and operating a public information repository. J Digit Imaging 26: 1045-1057.

3. Adams R, Henry KE, Sridharan A, Soleimani H, Zhan A, et al. (2022) Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med 28: 1455-1460.

4. Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, et al. (2021) Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Med 13: 152.

5. Hallinan JTPD, Zhu L, Yang K, Makmur A, Algazwi DAR, et al. (2021) Deep learning model for automated detection and classification of central canal, lateral recess, and neural foraminal stenosis at lumbar spine MRI. Radiology 300: 130-138.

6. Brinjikji W, Luetmer PH, Comstock B, Bresnahan BW, Chen LE, et al. (2015) Systematic literature review of imaging features of spinal degeneration in asymptomatic populations. AJNR Am J Neuroradiol 36: 811-816.

7. Nakashima H, Yukawa Y, Suda K, Yamagata M, Ueta T, et al. (2015) Abnormal findings on magnetic resonance images of the cervical spines in 1211 asymptomatic subjects. Spine 40: 392-398.

8. Abulibdeh R, Celi LA, Sejdić E (2025) The illusion of safety: A report to the FDA on AI healthcare product approvals. PLOS Digit Health 4: e0000866.

9. Haider SA, Borna S, Gomez-Cabello CA, Pressman SM, Haider CR, et al. (2026) The algorithmic divide: a systematic review on AI-driven racial disparities in healthcare. J Racial Ethn Health Disparities 13: 188-217.

10. Davis SE, Dorn C, Park DJ, Matheny ME (2025) Emerging algorithmic bias: Fairness drift as the next dimension of model maintenance and sustainability. J Am Medical Informatics Assoc 32: 845-854.

11. (2024) 12th WFC global education conference. World Federation of Chiropractic.

12. As'ad M, Faran N, Joharji H (2025) AI-supported shared decision-making (AI-SDM): Conceptual framework. JMIR AI 4: e75866.

13. Saenz AD, Centi A, Ting D, You JG, Landman A, et al. (2024) Establishing responsible use of AI guidelines: A comprehensive case study for healthcare institutions. NPJ Digit Med 7: 348.

14. Wensing M, Grol R, Grimshaw JM (2020) Improving patient care: The implementation of change in health care. (3rd edn), Wiley-Blackwell.

15. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, et al. (2011) Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 38: 65-76.

16. Cao W, Zhang Q, Liu J, Liu S (2026) From agents to governance: Essential AI skills for clinicians in the large language model era. J Med Internet Res 28: e86550.

17. Nicolaes J, Raeymaeckers S, Robben D, Wilms G, Vandermeulen D, et al. (2019) Detection of vertebral fractures in CT using 3D convolutional neural networks.

18. Goddard K, Roudsari A, Wyatt JC (2014) Automation bias: Empirical results assessing influencing factors. Int J Med Inform 83: 368-375.

19. Flynn TW, Smith B, Chou R (2011) Appropriate use of diagnostic imaging in low back pain: A reminder that unnecessary imaging may do as much harm as good. J Orthop Sports Phys Ther 41: 838-846.

20. Regev GJ, Treister R, Brill S, Ofir D, Salame K, et al. (2023) Low back pain patients' perceptions regarding their own radiology reports: Pre-Intervention survey. J Pain Res 16: 933-941.

21. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, et al. (2020) Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Nat Med 26: 1364-1374.