Review

MECP | MEDICAL EDUCATION & CLINICAL PRACTICE

Clinical validation of artificial intelligence in musculoskeletal imaging: a structured narrative review of outcomes, failure modes, and implementation readiness

Validación clínica de la inteligencia artificial en imágenes musculoesqueléticas: una revisión narrativa estructurada de resultados, modos de fallo y preparación para la implementación

Fabriccio J. Visconti-Lopez[1]*, Ivan David Lozada-Martínez[2,3,4]

[1] Universidad Científica del Sur, Lima, Peru.
[2] Center for Meta-Research and Scientometrics in Biomedical Sciences, Barranquilla, Colombia.
[3] Biomedical Scientometrics and Evidence-Based Research Unit, Department of Health Sciences, Universidad de la Costa, Barranquilla, Colombia.
[4] Clínica Iberoamérica, Barranquilla, Colombia.

ABSTRACT

Artificial intelligence (AI) is rapidly entering musculoskeletal (MSK) imaging, yet clinical adoption often outpaces evidence of real-world benefit, widening the gap between diagnostic performance claims and patient-important outcomes. In this structured narrative review, we synthesize published evidence and reporting guidance relevant to clinical validation, failure modes, and implementation readiness of AI in MSK imaging, focusing on where AI may improve care, where it commonly fails to add value, and how validation should be structured before routine adoption. Reported strengths are clustered in narrow, time-sensitive use cases such as fracture triage and workflow support, whereas limitations repeatedly arise from poor generalizability, narrow task framing, bias, reference-standard weaknesses, and workflow mismatch. A central message is that high diagnostic accuracy does not equal clinical value, which must be judged by effects on management, outcomes, safety, efficiency, and equity. We outline minimum validation expectations, including external validation, prospective assessment of clinical impact, and ongoing monitoring for performance drift and bias.

Keywords: Artificial Intelligence; Machine Learning; Deep Learning; Diagnostic Imaging; Musculoskeletal System; Radiography (source: MeSH NLM).

RESUMEN

La inteligencia artificial (IA) está ingresando rápidamente en la imagenología musculoesquelética (MEQ), aunque la adopción clínica a menudo supera la evidencia de beneficio en el mundo real, ampliando la brecha entre las afirmaciones de rendimiento diagnóstico y los desenlaces importantes para los pacientes. En esta revisión narrativa estructurada, sintetizamos la evidencia publicada y las guías de reporte relevantes para la validación clínica, los modos de falla y la preparación para la implementación de la IA en la imagenología MEQ, centrándonos en dónde la IA puede mejorar la atención, dónde comúnmente no logra aportar valor y cómo debe estructurarse la validación antes de la adopción rutinaria. Las fortalezas reportadas se agrupan en casos de uso estrechos y sensibles al tiempo, como el triaje de fracturas y el apoyo al flujo de trabajo, mientras que las limitaciones surgen repetidamente de la pobre generalización, el encuadre estrecho de las tareas, el sesgo, las debilidades del estándar de referencia y el desajuste con el flujo de trabajo. Un mensaje central es que una alta precisión diagnóstica no equivale a valor clínico, el cual debe juzgarse por los efectos sobre el manejo, los desenlaces, la seguridad, la eficiencia y la equidad. Delineamos expectativas mínimas de validación, incluidas la validación externa, la evaluación prospectiva del impacto clínico y el monitoreo continuo de la deriva del rendimiento y el sesgo.

Palabras clave: Inteligencia Artificial; Aprendizaje Automático; Aprendizaje Profundo; Diagnóstico por Imagen; Sistema Musculoesquelético; Radiografía (fuente: DeCS Bireme).

Article history
Received: December 22, 2025
Peer reviewed
Accepted: April 29, 2025
Online: May 20, 2026

How to cite this article
Visconti-Lopez FJ, Lozada-Martínez ID. Clinical validation of artificial intelligence in musculoskeletal imaging: a structured narrative review of outcomes, failure modes, and implementation readiness. Med Educ Clin Pract. 2026;1(2):e017. doi: 10.21142/mecp.2026.1.2.e017.

***Correspondence**
Fabriccio J. Visconti-Lopez
E-mail: fvisconti@cientifica.edu.pe

INTRODUCTION

The rapid proliferation of artificial intelligence (AI) applications in musculoskeletal (MSK) imaging has generated substantial interest, but a critical gap remains between commonly reported diagnostic performance metrics and patient-important outcomes [1]. Although cost-containment pressures and expectations of added value fuel enthusiasm [2], and the AI medical imaging market continues to grow with major financial investment [3], high diagnostic accuracy is often relatively easy to demonstrate for selected tasks in curated datasets [4]. Academic and industry incentives can further emphasize positive, model-centric results while underreporting negative findings, usability limitations, and real-world implementation failures [5]. As a result, "false promises" remain a central risk, in which unsubstantiated outperformance claims persist and many proposed models are never adopted into routine clinical practice despite extensive research [6]. Contributing factors include historical gaps in rigorous statistical and clinical validation [7] and the persistent scarcity of external validation in deep learning studies [8]. Accordingly, despite strong momentum, widespread implementation continues to face practical challenges related to integration with existing systems and demonstration of real-world clinical value [2,9].

In this review, "clinical benefit" extends beyond technical performance or diagnostic accuracy and refers to measurable improvement in patient care and healthcare outcomes [10]. This includes patient-level outcomes such as reduced mortality and morbidity, improved quality of life, and lower healthcare expenditure [10]. At the workflow level, potential benefits may include improved diagnostic accuracy, optimized efficiency, increased productivity, and faster reporting, including better prioritization of urgent cases and reduced misinterpretation [4,11,12]. Clinical benefit may also be reflected in improved decision-making processes, personalized treatment strategies, and increased patient and provider satisfaction [13]. Ultimately, the clinical value of AI should be judged by its effect on real-world outcomes within integrated practice, ideally supported by cost-benefit analyses that include efficiency, expenditure, and patient satisfaction [4,14]. In the present review, AI in MSK imaging is critically appraised by distinguishing diagnostic capability from demonstrated impact on clinical management, patient well-being, and healthcare efficiency [13].

Several recent reviews have addressed AI in MSK imaging from related but distinct perspectives. Debs and Fayad reviewed the promise and limitations of AI across the MSK imaging cycle, including interpretive and non-interpretive tasks [13]. Oettl *et al.* summarized current machine learning models in MSK imaging, including general-purpose multimodal models and specialized foundation models, while highlighting ethical considerations, technical challenges, and the need for validation [15]. Tsang *et al.* focused on emergency MSK imaging, emphasizing workflow triage, modality selection, abnormality detection, fracture classification, pediatric emergency applications, and large language model-assisted reporting [16]. Dubey *et al.* reviewed the role of AI in MSK interventions, with emphasis on procedural planning, image guidance, and clinical integration in interventional practice [17]. Tordjman *et al.* presented a scoping review of practical MSK radiology applications, including accelerated image acquisition, image interpretation, large language models, workflow integration, cost-effectiveness, liability, and education [18]. The present review overlaps with these studies in recognizing the potential of AI to improve diagnostic accuracy and workflow efficiency but differs by centering the analysis on clinical validation: how claims of benefit should be judged, when AI may fail to add value, which evidence levels support adoption, and which validation, reporting, fairness, implementation, and post-deployment monitoring standards should guide clinical use.

METHODS

Review design and scope

This study was conducted as a structured narrative review focused on clinical validation, failure modes, and implementation readiness of AI in MSK imaging. It was not designed as a scoping review, because the aim was not to map the full volume and distribution of the literature, but to critically synthesize evidence and guidance relevant to clinical outcomes, workflow impact, external validation, diagnostic safety, fairness, interpretability, regulatory maturity, and post-deployment monitoring. The objective was interpretive and framework-oriented, and thus, no meta-analysis or quantitative pooling of diagnostic accuracy estimates was performed. Instead, the review used predefined eligibility criteria, explicit evidence prioritization, structured data extraction, and thematic synthesis to evaluate how AI tools should be clinically validated before adoption in MSK imaging practice.

Search strategy and evidence selection

Targeted searches were conducted in PubMed/MEDLINE, Scopus, and Web of Science. Search terms combined AI-related concepts, including "artificial intelligence," "machine learning," "deep learning," "neural network," "radiomics," and "large language model," with MSK imaging concepts, including "musculoskeletal," "orthopaedic," "orthopedic," "radiography," "MRI," "CT," "ultrasound," "fracture," "osteoarthritis," "spine," "ligament," "rotator cuff," "tumor," "osteoporosis," and "sarcopenia." These

terms were combined with validation and implementation concepts, including "external validation," "clinical impact," "workflow," "triage," "prospective," "randomized," "bias," "fairness," "calibration," "PACS," "RIS," "regulatory," and "cost-effectiveness." Reference lists of relevant reviews, reporting guidelines, and implementation frameworks were also screened.

Eligibility criteria

We included review articles, validation studies, reader studies, prospective workflow studies, clinical impact studies, randomized trials, implementation reports, regulatory or governance articles, and reporting guidelines if they addressed AI in MSK imaging or provided directly applicable guidance for clinical validation of imaging AI. We prioritized sources that addressed clinical utility, comparison with usual care, diagnostic safety, external validation, fairness, interpretability, implementation, regulatory maturity, or post-deployment monitoring. We excluded articles not related to imaging, articles focused only on technical model development without clinical validation or implementation relevance, non-medical AI applications, and papers for which the full text or key methodological information was unavailable.

Study selection process

Records identified through database searches and reference-list screening were assessed according to their relevance to the review scope. Titles and abstracts were first reviewed to exclude publications clearly unrelated to MSK imaging, AI, clinical validation, or implementation. Potentially relevant full texts were then assessed using the eligibility criteria described above. Publications were retained when they addressed AI in MSK imaging or provided directly applicable guidance on clinical validation, implementation, reporting, fairness, interpretability, regulatory maturity, or post-deployment monitoring of imaging AI.

Evidence prioritization

Evidence was prioritized according to its relevance to clinical validation. The highest priority was given to randomized or prospective clinical impact studies, followed by prospective workflow or reader studies comparing AI-assisted interpretation with usual care, external validation studies, implementation and post-deployment evaluations, systematic or scoping reviews, and reporting or regulatory guidance. Retrospective technical development studies were used only when they illustrated a specific failure mode, dataset limitation, reference-standard issue, or validation requirement.

Data extraction and synthesis

For each source included, we extracted the clinical use case, imaging modality, AI task, model type when reported, comparator, reference standard, validation design, performance metrics when available, clinical or workflow endpoints, reported failure modes, regulatory or implementation maturity, and relevance to patient care. Extracted information was used to characterize the nature of evidence supporting each clinical use case, including whether the source represented a validation study, reader study, workflow evaluation, implementation report, methodological framework, or review-level synthesis. Synthesis was narrative and framework-based. We did not pool sensitivity, specificity, positive predictive value, negative predictive value, or area under the curve because indications, imaging modalities, reference standards, model types, and clinical workflows were heterogeneous. No numerical cutoff was applied because outcome domains and validation designs varied across studies. The two thematic categories were determined through narrative synthesis rather than by a numerical threshold. No cutoff point for sensitivity, specificity, predictive value, area under the curve, time reduction, or effect size was used to define improvement. A use case was interpreted as having potential clinical or workflow value when the evidence linked AI use to outcomes beyond standalone diagnostic accuracy, including changes in clinical management, reduction in clinically significant errors, triage or reporting efficiency, diagnostic safety, patient-important outcomes, equity, or implementation readiness. Conversely, a use case or evidence domain was interpreted as failing to add value when reported limitations suggested weak generalizability, poor reference standards, limited clinical actionability, workflow mismatch, automation bias, inequitable subgroup performance, lack of prospective validation, or absence of post-deployment monitoring. Studies evaluating efficacy and safety were prioritized when available, but the classification also incorporated validation studies, reader studies, workflow evaluations, implementation reports, methodological frameworks, and review-level evidence.

RESULTS

The narrative synthesis identified two broad themes across the literature reviewed. First, AI in MSK imaging appears to be most clinically relevant when deployed in narrow, well-defined, and workflow-sensitive tasks, including triage support, detection of clinically important imaging findings, structured assessment, and selected opportunistic screening applications. Second, the literature repeatedly indicates that AI may fail to add value when validation is limited to curated datasets, outcome definitions are disconnected from clinical

Table 1. Clinical use cases for artificial intelligence in musculoskeletal imaging and what to measure.

Use case, modality	Where AI fits in the workflow	Evidence characteristics and supporting sources	Technological maturity and regulatory status	Primary endpoints to judge clinical value	Secondary endpoints	Key risks and failure modes	Minimum validation expectations before adoption
ED fracture detection and triage, radiography	Queue prioritization, decision support for initial read	Commonly evaluated in retrospective validation studies, reader studies, emergency workflow studies, and reviews of MSK imaging AI. Evidence usually focuses on diagnostic performance and turnaround time rather than patient-important outcomes [1,13,16,20,21]	Relatively mature compared with other MSK AI use cases, with some commercial deployment in radiology workflows. Regulatory approval is product-specific and jurisdiction-dependent. Approval status was not inferred unless reported in the original source or product documentation	Time to immobilization or operative management, missed clinically significant fractures leading to harm, unplanned return visits	Sensitivity at triage threshold, false negative rate for "can't miss" fractures, turnaround time	Spectrum bias, automation bias, alert fatigue, weak reference standard	External validation across sites, prospective workflow evaluation, predefined escalation pathway for positive flags
Occult fracture support, CT or MRI in selected scenarios	Second reader for subtle findings	Evidence is usually based on modality-specific validation studies, reader studies, and broader reviews of MSK AI. Data are heterogeneous across scanners, protocols, anatomical regions, and reference standards [1,13,20,21]	Mostly validation-stage or workflow-evaluation stage, with maturity varying by anatomical site and modality. Regulatory status is not consistently reported across studies	Reduced delayed diagnosis, fewer complications from missed injuries, time to definitive treatment	Miss rate versus standard care, inter-reader agreement, reporting time	Label noise, confounding by severity, domain shift across scanners	External validation on heterogeneous scanners, prospective impact on miss rate and treatment timing
Joint dislocation and alignment abnormalities, radiography	ED support and triage	Evidence is generally extrapolated from emergency MSK imaging AI, radiographic triage, and abnormality detection literature. These high-risk conditions require safety-focused validation because prevalence and presentation vary across emergency settings [13,16,20,21]	Less mature than general fracture detection because these are less frequent and higher-risk presentations. Regulatory status is variable and not consistently reported	Time to reduction, fewer missed dislocations, reduced neurovascular complications	Sensitivity and specificity, time to diagnosis	Rare but high-risk errors, prevalence shift, automation bias	External validation in real ED prevalence, prospective safety monitoring for misses
Osteoarthritis severity grading, radiography or MRI	Standardized grading for clinic decision-making	Evidence includes imaging AI reviews, ML studies in rheumatic and MSK diseases, and model validation studies using radiographic or MRI-based grading systems. Comparability is limited by different grading scales, reference standards, and clinical decision thresholds [1,13,21,24]	Moderate technical maturity for grading and quantification tasks, but clinical adoption depends on whether outputs change management. Regulatory status is product-specific and often not reported in validation studies	Shared decision quality, appropriate timing of nonoperative versus operative management, function and pain outcomes if measured clinically	Agreement with reference grading, reproducibility, calibration	Shortcut learning, site artifacts, limited clinical utility if grading does not change management	External validation, assessment of decision impact, subgroup performance checks
Spine degeneration and stenosis assessment, MRI	Structured reporting support, triage for specialist review	Evidence is usually derived from MRI-based model development, validation studies, and MSK AI reviews. Clinical relevance depends on avoiding overinterpretation of degenerative findings that may not explain symptoms [1,13,20,21]	Mainly validation-stage or structured-reporting support, with variable clinical maturity. Regulatory approval is not consistently reported and should be verified for each product	Reduced delays to appropriate care, fewer unnecessary referrals or imaging cascades	Report completeness, agreement with expert reads, triage precision	Overcalling degenerative findings, incidentalomas, downstream overuse	External validation, prospective evaluation of referral patterns and downstream testing
Meniscus, ACL, and ligament injury support, MRI	Decision support for reporting and triage	Evidence is commonly based on retrospective MRI validation studies, reader studies, and MSK imaging AI reviews. Reported performance varies by injury type, prior surgery, image protocol, and reference standard [1,13,20,21]	Mostly retrospective validation or reader-study stage, with limited prospective clinical impact evidence. Regulatory status is not consistently reported across models	Time to definitive management, reduced unnecessary arthroscopy, improved patient selection for surgery	Sensitivity and specificity, calibration, reading time	Spectrum bias, label noise, confounding by prior injury or surgery	External validation including postoperative cases, prospective evaluation of management changes
Rotator cuff tear assessment, MRI or ultrasound	Reporting support, triage	Evidence includes imaging AI reviews and modality-specific validation studies. Ultrasound applications require special attention to operator dependence, acquisition variability, and device heterogeneity [1,13,20,21]	Mostly validation-stage, with ultrasound maturity limited by operator and acquisition variability. Regulatory status is product-specific and not consistently reported	Time to treatment, appropriate surgical referral, function outcomes when collected	Agreement with reference reads, reading time, confidence calibration	Operator dependence, variable protocols, partial tear ambiguity	External validation across operators and devices, prospective monitoring of referral and treatment patterns
Tumor detection and characterization, MRI	Triage for urgent specialist review, decision support	Evidence is limited by rarity, small datasets, and limited external validation. Scoping review evidence suggests that MSK malignancy AI remains constrained by data scarcity and the need for multicenter validation [19,23]	Early-stage or validation-stage because rarity limits large multicenter evaluation. Regulatory approval should not be assumed without product-specific evidence	Time to diagnosis and referral, avoidance of delayed malignant diagnosis, reduced unnecessary biopsies	Sensitivity at high-recall threshold, false negative audit rate, time to report	Rare disease scarcity, high-consequence false negatives, dataset shift	Multicenter external validation, prospective safety-focused evaluation, mandatory discordance review process
Opportunistic osteoporosis screening, CT	Automated flagging for follow-up evaluation	Evidence commonly comes from opportunistic screening, body composition, and imaging AI literature. Clinical value depends on whether automated detection leads to follow-up, treatment initiation, and fracture prevention strategies [1,13,21]	More mature than many opportunistic imaging applications when linked to automated measurement and follow-up pathways. Regulatory status remains product-specific and jurisdiction-dependent	Treatment initiation rates, fracture prevention strategies implemented, downstream fracture events when available	Screening yield, calibration, referral completion	Overdiagnosis, inequitable follow-up access, false reassurance	External validation, workflow plan for follow-up, monitoring of downstream actions and equity
Opportunistic sarcopenia or frailty markers, CT	Risk stratification and care planning support	Evidence is mainly based on opportunistic imaging, segmentation, and risk stratification studies. The main challenge is actionability, because prognostic markers only add value if linked to effective care pathways [1,13,21]	Mostly exploratory, segmentation, or risk-stratification stage. Regulatory status is usually not reported, and clinical maturity depends on whether outputs trigger effective interventions.	Postoperative complications, recovery trajectories, functional outcomes if linked to care pathways	Feasibility, calibration, referral or intervention uptake	Confounding, unclear actionability, drift across protocols	External validation, prospective evaluation of whether flags trigger beneficial interventions

ACL: anterior cruciate ligament, AI: artificial intelligence, CT: computed tomography, ED: emergency department, MRI: magnetic resonance imaging, MSK: musculoskeletal, ML: machine learning.

Med Educ Clin Pract. 2026;1(2):e017. doi: 10.21142/mecp.2026.1.2.e017.

decisions, reference standards are weak, subgroup performance is untested, or workflow integration increases cognitive burden. The results below present these themes and summarize the selected literature and use-case implications in Tables 1 and 2.

Theme 1: Clinical contexts in which AI may improve outcomes or workflow

Evidence of clinical impact in MSK imaging is most credible when AI is evaluated as decision support inside real workflows and when outcomes extend beyond standalone accuracy. We did not classify improvement using a fixed numerical cutoff because validation designs, imaging modalities, reference standards, and outcome domains varied across studies. Instead, clinical relevance was assessed narratively according to whether AI was linked to clinically meaningful endpoints, including management changes, reduction in clinically significant errors, triage or reporting efficiency, diagnostic safety, patient-important outcomes, equity, and implementation readiness. In practice, most published work still focuses on diagnostic performance, while fewer studies quantify downstream management changes, patient-important outcomes, or unintended consequences. Accordingly, this section summarizes use cases where impact outcomes are most often assessed and explicitly highlights where evidence remains limited to intermediate endpoints (Table 1). To complement this use-case framework, Table 2 summarizes selected studies and guidance sources that reported on the main themes of the narrative synthesis, including clinical value, validation hierarchy, reporting transparency, implementation barriers, data governance, and post-deployment monitoring.

Comparative evidence between conventional clinical practice and AI-supported practice remains uneven across MSK imaging. The comparative evaluations available most often assessed AI as a second reader, triage tool, or workflow support system against unaided clinician interpretation or usual radiology workflow. These comparisons are most informative when they measure clinically relevant endpoints such as missed clinically significant findings, time to report, time to escalation, change in management, downstream testing, and safety monitoring. However, much of the evidence available still emphasizes diagnostic performance or workflow efficiency rather than patient-important outcomes. Therefore, AI should be interpreted as a potential adjunct to conventional practice rather than as a replacement for clinician judgment, particularly when evidence is limited to retrospective datasets, reader studies, or intermediate process outcomes [4,6,10,13,19].

Theme 2: Clinical contexts in which AI may fail to add value

Despite impressive diagnostic accuracy in controlled settings, AI applications often fail to translate gains in MSK imaging into tangible clinical benefit or widespread adoption [4,6]. A recurring pattern is that high diagnostic accuracy does not reliably translate into improved patient outcomes or effective clinical integration [4,6]. Claims of outperformance over human experts can be unsubstantiated, and many models ultimately remain unadopted in clinical practice [6]. In addition, many studies reporting strong performance are neither prospective nor randomized and carry a high risk of bias; lack of adherence to reporting standards further limits interpretability and clinical relevance [19]. Accordingly, the emphasis should shift away from algorithm-centric accuracy metrics toward evidence that AI improves real-world outcomes when used in combination with clinicians [4].

Narrow task framing and clinical complexity. Many AI models are optimized for single tasks and may struggle with the complex, multifaceted presentations typical of clinical practice [13]. This narrow focus can also lead to incidental findings without sufficient clinical context and may contribute to overdiagnosis by identifying patterns that are statistically recognizable but not clinically relevant in the scenario presented [20].

Generalizability, bias, and shortcut learning. A key weakness of MSK imaging AI is limited generalizability to different populations and clinical environments [21]. Datasets may be unrepresentative and can propagate pre-existing biases and inequities present in training data [13]. Reliance on highly preprocessed public datasets further limits performance in unseen settings, and shortcut learning, in which models exploit nonmorphological patterns, can exacerbate this problem [21,22]. When external validation is lacking, performance outside the original development environment remains difficult to discern [21,22].

Rare conditions and limited clinical scope. Development for rare MSK conditions is constrained by insufficient availability of large, diverse datasets [21,23]. Consequently, many algorithms focus on common pathologies, limiting utility across the broader spectrum of orthopaedic disease [12]. Studies in rare diseases often face small sample sizes and limited external validation, restricting the ability of AI to support diagnosis for less common conditions [24]. Table 3 summarizes the main contexts in which AI may fail to add value in MSK imaging, the type of evidence informing each limitation, the likely clinical consequence, and the expected technological maturity or regulatory status.

Table 2. Summary of selected studies and guidance sources included in the narrative synthesis.

Reference(s)	Clinical or methodological focus	Key findings relevant to this review	Main limitation or relevance to interpretation
Chang and Link [1]	AI in MSK radiology	Highlights the rapid growth of AI applications in MSK radiology and supports the need to distinguish technical promise from clinical value	Provides contextual framing rather than primary validation evidence
Strohm *et al.* [2], Kim *et al.* [9]	AI implementation in radiology	Identify workflow integration, organizational readiness, usability, and stakeholder engagement as central determinants of successful AI adoption	Not specific to all MSK use cases, but directly relevant to radiology implementation
Farhadi *et al.* [4], Ounasser *et al.* [14]	Clinical and surgical applications of AI	Support the view that AI may improve diagnostic and workflow processes, but clinical value depends on demonstrated effects on decision-making, outcomes, and healthcare efficiency	Most of the evidence included remained heterogeneous and often emphasized technical performance
Christodoulou *et al.* [6], Nagendran *et al.* [19]	Validity of AI outperformance claims and reporting quality	Show that claims of AI superiority over clinicians may be overstated when study design, reporting quality, risk of bias, and clinical applicability are critically assessed	Primarily addressed medical imaging or deep learning broadly, rather than only MSK imaging
Van Leeuwen *et al.* [8]	Scientific evidence behind commercial AI	Reports that many commercial radiology AI products have limited published scientific evidence and scarce external validation	Commercial evidence may evolve rapidly and may not capture unpublished implementation data
Guermazi *et al.* [11], Debs and Fayad [13]	MSK imaging AI	Describe promising MSK imaging applications, including fracture detection, workflow support, osteoarthritis grading, and MRI-based lesion assessment, while emphasizing limitations in generalizability, validation, and clinical integration	Mostly synthesis and expert interpretation rather than interventional outcome evidence
Hirschmann *et al.* [20], Diao *et al.* [21]	Current challenges and trends in MSK AI	Identify narrow task framing, dataset limitations, limited generalizability, and lack of real-world validation as recurring barriers to adoption	Do not provide pooled clinical outcome estimates
Hinterwimmer *et al.* [23]	Machine learning for imaging-driven diagnosis of MSK malignancies	Shows that AI applications for MSK malignancy imaging remain limited by rarity, dataset scarcity, and the need for multicenter validation	Evidence is constrained by small datasets and limited external validation
Nelson and Arbeeva [24]	Machine learning in rheumatic and MSK diseases	Emphasizes bias, target definition, clinical relevance, and the need to align model development with meaningful clinical goals	Broader rheumatic and MSK focus, not limited to imaging AI
Pham *et al.* [10], Park *et al.* [26], You *et al.* [27]	Levels of evidence, clinical validation, and AI deployment	Support a stepwise validation model progressing from retrospective testing to external validation, prospective impact studies, randomized trials, and post-deployment monitoring	Framework-based sources that guide interpretation rather than provide MSK-specific outcome estimates
Boverhof *et al.* [25], Omoumi *et al.* [30]	Value-based AI deployment and evaluation of commercial AI	Provide practical criteria for assessing whether radiology AI adds value in clinical workflows and before purchasing or implementing commercial solutions	Not limited to MSK imaging, but highly relevant to adoption decisions
Duggan *et al.* [32], Duncan *et al.* [31]	Reference standards for AI validation	Highlight the importance of defensible reference standards, adjudication, and bias control when validating AI imaging tools	Examples derive partly from radiography outside MSK, but the reference-standard principles are applicable
Reinke *et al.* [33], Subbaswamy *et al.* [34]	Metric selection and subgroup performance	Support the need for appropriate performance metrics, calibration, operating thresholds, and subgroup evaluation to avoid misleading conclusions and inequitable performance	Methodological focus, not specific to one MSK pathology
CONSORT-AI, SPIRIT-AI, TRIPOD+AI, CLAIM, and DECIDE-AI [37,41-47]	Reporting of AI prediction models, AI trials, imaging AI, and early clinical evaluation	Provide structured reporting standards for development, validation, early evaluation, and clinical trials of AI-enabled tools	Applicable according to study design and claim. They do not replace clinical validation
Blezek *et al.* [49], Chatterjee *et al.* [50], Wiggins *et al.* [52]	Clinical integration of AI into radiology systems	Support the importance of PACS/RIS integration, interoperability, reporting pathways, and operational deployment for real-world AI use	Implementation evidence may be context-specific and dependent on local infrastructure
Cestonaro *et al.* [53], Contaldo *et al.* [54], Saw and Ng [55]	Medical responsibility and governance of AI	Emphasize that AI deployment requires governance, human oversight, liability clarity, and regulatory alignment	Regulatory requirements differ by jurisdiction and may evolve
Kondylakis *et al.* [56], Araújo *et al.* [57], Jiménez-Sánchez *et al.* [58]	Medical imaging data infrastructure and dataset quality	Support the review's emphasis on dataset completeness, harmonization, governance, privacy, robustness, and fairness	Mostly infrastructure-focused, with indirect links to patient outcomes
Van Kooten *et al.* [59], Soleimantabar *et al.* [60], Ng and Tan [61]	AI education, adoption barriers, and trust among radiologists	Support the need for clinician education and trust-building as part of implementation readiness	Focuses on training and perceptions rather than diagnostic effectiveness

AI: artificial intelligence, MRI: magnetic resonance imaging, MSK: musculoskeletal, PACS: Picture Archiving and Communication Systems, RIS: Radiology Information Systems.

Med Educ Clin Pract. 2026;1(2):e017. doi: 10.21142/mecp.2026.1.2.e017.

DISCUSSION

The findings of this narrative synthesis indicate that the central challenge in MSK imaging AI is not whether models can achieve high standalone diagnostic performance, but whether their use improves care in a safe, equitable, and operationally sustainable manner. This distinction is clinically important because MSK imaging decisions may influence emergency triage, surgical referral, follow-up imaging, rehabilitation pathways, and resource allocation. Therefore, implementation should depend on evidence linking AI outputs to management decisions, a reduction in clinically significant errors, workflow improvement, and post-deployment safety monitoring, rather than on retrospective accuracy alone [4,6,10,13,25-27].

Evidence Hierarchy for Clinical Validation

AI in MSK imaging can be interpreted through a stepwise clinical validation hierarchy, with which each level provides a different degree of evidence regarding technical performance, workflow integration, clinical utility, and long-term safety. This hierarchy was adapted narratively from existing AI validation and deployment frameworks rather than developed as a new validated scoring tool [10,25-27]. It ranges from initial retrospective technical validation to post-deployment monitoring and clarifies what type of claim can be legitimately supported at each step. To make the framework more transparent, each level below includes examples of the type of evidence that would correspond to that level.

Level 1: Retrospective internal validation (development setting). AI models are tested in existing historical datasets from the same institution in which the model was developed. This level supports claims of early technical efficiency and diagnostic performance and supports assessment of dataset quality and model development [10]. Models are typically trained, validated, and tested using different splits of the same retrospective dataset, often in ratios such as 80:10:10 [13]. However, results are confined by characteristics and biases of the development dataset [10] and cannot support claims of generalizability to new populations or settings, real-world clinical impact, or improved patient outcomes [10]. For example, this level includes studies in which an algorithm is trained and tested in historical radiographs, computed tomography (CT) images, or magnetic resonance imaging (MRI) examinations from the same institution using internal data splits.

Level 2: External validation (new institutions or populations). The model is evaluated in datasets from institutions or populations other than where it was developed, supporting claims of reproducibility, generalizability, and mitigation of bias, including performance on out-of-distribution data [10]. However, external validation alone cannot support claims about workflow impact, patient management, or patient outcomes in real time, because it tests performance with unseen data rather than integration into practice. For example, this level includes testing a fracture detection, osteoarthritis grading, or MRI classification model in a different hospital, scanner environment, geographic population, or time period from the original development dataset.

Level 3: Prospective workflow evaluation (observational "in practice" use). The AI model is introduced into a real clinical workflow and clinician interaction is observed in a live environment without direct intervention based on AI output alone. This level supports claims about real-world technical performance, early workflow integration, and preliminary signals related to diagnostic process or efficiency when clinicians use the tool [25]. However, because this evidence remains observational, confounding limits causal claims and it cannot support definitive outcome improvement or cost-effectiveness. For example, this level includes studies in which AI output is integrated into a live reporting or triage environment, and investigators observe usability, reporting workflow, alert burden, or clinician interaction without random allocation.

Level 4: Prospective impact studies (practice and intermediate outcomes). These studies assess the direct effect of AI integration on clinical practice, workflow, and intermediate outcomes, supporting claims such as reduced reading times, lower misinterpretation rates, or changes in clinical decision-making [10]. They can also support evidence of added value in specific use cases [25]. However, they cannot support claims about long-term outcomes, societal cost-effectiveness, or the absence of unintended consequences without more rigorous designs. For example, this level includes prospective studies measuring whether AI implementation changes reporting time, escalation time, diagnostic error rates, follow-up actions, referral patterns, or other intermediate clinical outcomes.

Level 5: Randomized controlled trials. Randomized controlled trials compare groups with and without AI (or different care standards) and support claims of causal links between AI integration and improvements in patient outcomes [10]. They provide the strongest evidence for clinical efficacy and health outcomes in real-world care [26], but they may not fully capture long-term generalizability across settings or continuous performance as models evolve. For example, this level includes randomized comparisons of AI-assisted interpretation versus usual care, with outcomes such as clinically significant diagnostic errors, time to management, downstream testing, patient outcomes, or service-level efficiency.

Table 3. Evidence contexts in which artificial intelligence may fail to add value in musculoskeletal imaging.

Limitation or failure mode	Evidence context	Clinical consequence	Technological maturity and regulatory status	Validation or implementation implication
High diagnostic accuracy without clinical impact	Retrospective validation studies, reader studies, and broad AI performance reviews	Accuracy gains may not translate into improved management, safety, outcomes, or efficiency	Many tools remain at validation or early implementation stage. Regulatory approval, when present, does not necessarily prove patient benefit	Prospective workflow or clinical impact evaluation needed before claiming clinical value
Narrow task framing	Model development studies and MSK imaging reviews focused on single-label or single-task outputs	AI may detect one finding but fail to support complex diagnostic reasoning, multimorbidity, or clinically relevant differential diagnosis	Common in technically mature model development but less mature for integrated clinical decision support. Regulatory status is usually task-specific	Define the clinical decision point and test whether the output changes management
Weak or inconsistent reference standards	Validation studies using radiology reports, single-reader labels, incomplete follow-up, or non-adjudicated ground truth	Label noise may overestimate or underestimate model performance and reduce transportability	Present across early and advanced model development. Regulatory status does not remove the need for defensible reference standards	Use adjudication, follow-up, operative findings when appropriate, and blinded assessment
Limited generalizability and dataset shift	External validation studies, reviews of commercial AI evidence, and methodological guidance	Model performance may decrease across scanners, protocols, institutions, populations, or disease prevalence	Tools may appear mature in the development site but remain immature across external settings. Regulatory approval may not cover all local populations or workflows	Multicenter external validation and local performance monitoring needed
Bias and uneven subgroup performance	Fairness studies, subgroup performance frameworks, and reviews of AI bias	AI may worsen inequity if performance is lower in underrepresented groups or specific clinical subgroups	Fairness evaluation remains inconsistently reported across AI maturity levels. Regulatory status rarely guarantees equitable subgroup performance	Report subgroup performance by clinically relevant variables, including site, age, sex, device, and postoperative status when applicable
Shortcut learning and spurious correlations	Technical validation studies, dataset-quality studies, and methodological reviews	Models may rely on non-anatomical artifacts, site markers, acquisition patterns, or preprocessing features rather than disease morphology	Often detected after external validation or stress testing. Regulatory approval does not exclude local shortcut learning	Use external validation, stress testing, explainability checks, and dataset audits
Workflow mismatch and alert fatigue	Implementation reports, workflow studies, and radiology AI deployment frameworks	Poor integration may increase cognitive burden, unnecessary alerts, delayed reporting, or clinician distrust	More relevant at implementation stage. Some commercial tools may be available, but maturity depends on local PACS/RIS integration	Evaluate usability, alert burden, override rates, turnaround time, and clinician interaction
Automation bias and overreliance	Human factors studies, reader studies, and implementation guidance	Clinicians may accept incorrect AI outputs or underweight discordant clinical information	Can occur even with technically mature tools. Regulatory approval does not eliminate human factors risk	Test AI-assisted performance, unaided performance, discordance handling, and escalation pathways
Lack of actionability	Opportunistic screening, grading, segmentation, and risk-stratification studies	AI outputs may not improve care if no follow-up pathway, referral process, or treatment decision is linked to the result	Common in exploratory or validation-stage tools, but may also occur after deployment	Define the downstream action, responsible clinician, follow-up pathway, and audit process
Absence of post-deployment monitoring	Deployment frameworks, regulatory discussions, and safety monitoring guidance	Drift, new bias, scanner changes, or workflow changes may reduce safety after implementation	Relevant to all deployed tools. Regulatory approval is not a substitute for ongoing surveillance	Establish monitoring metrics, audit frequency, model-update governance, and revalidation triggers

AI: artificial intelligence, MSK: musculoskeletal, PACS: Picture Archiving and Communication Systems, RIS: Radiology Information Systems.

Level 6: Post-deployment monitoring and surveillance. Once implemented in routine practice, continued monitoring supports claims about real-world effectiveness and safety, detection of performance drift, emergence of new biases, and long-term system impact, including evolving regulatory compliance needs [27]. However, this phase does not establish initial causal impact; it primarily confirms and refines understanding of performance and impact across diverse uncontrolled settings. For example, this level includes ongoing audits after deployment that track model drift, override rates, false negative cases, subgroup performance, scanner or protocol changes, and safety events requiring revalidation or rollback.

Illustrative Application of the Validation Framework to the Reviewed Evidence

When applied to the literature reviewed, the validation hierarchy shows that most AI applications in MSK imaging remain concentrated on retrospective validation, external validation, and early workflow evaluation. Use cases categorized under "Theme 1: Clinical contexts in which AI may improve outcomes or workflow", such as fracture detection and triage, selected structured grading tasks, and opportunistic screening, appear to be more clinically promising when studies move beyond standalone diagnostic performance and assess workflow integration, reader support, escalation time, or downstream management. However, even in these areas, evidence for patient-important outcomes, cost-effectiveness, and long-term safety remains limited.

In contrast, use cases categorized under "Theme 2: Clinical contexts in which AI may fail to add value" commonly reflect lower or incomplete levels of validation. These include applications supported mainly by retrospective internal datasets, narrow task definitions, weak or inconsistent reference standards, limited subgroup assessment, poor external validation, or unclear clinical actionability. In such cases, high diagnostic performance alone should not be interpreted as evidence of clinical benefit. The framework, therefore helps distinguish technical feasibility from clinical readiness by asking whether an AI tool has progressed from internal performance testing to external validation, prospective workflow evaluation, clinical impact assessment, and post-deployment monitoring.

This illustrative application also clarifies that the classification was not based on a fixed cutoff for sensitivity, specificity, predictive value, or area under the curve. Instead, the contribution of AI was assessed narratively according to clinical significance, including whether the tool could plausibly improve management decisions, reduce clinically significant errors, support timely triage, improve workflow, preserve safety, and maintain equitable performance across settings. This approach is consistent with the purpose of the review, which is to guide interpretation of clinical validation evidence rather than to produce a formal meta-analysis of diagnostic accuracy.

How to Validate AI Clinically: A Practical Framework

Careful clinical validation is required to determine whether an AI tool improves MSK care rather than simply improving test-set performance. The framework below summarizes minimum validation actions across the life cycle of an AI tool and aligns evaluation with outcomes that matter to patients and clinical services.

The clinical validation of AI in MSK imaging should follow a stepwise evaluation pathway. This pathway begins by defining the intended clinical use and decision target, then proceeds through the definition of the reference standard, internal and external validation, subgroup and explainability assessment, workflow integration, prospective impact evaluation, cost and scalability assessment, and post-deployment monitoring. The steps below translate this pathway into a practical framework for evaluating whether an AI tool is ready for clinical adoption.

Step 1: Define the clinical question, use-case, and comparator. Specify the target population, setting, modality, decision point, and what clinical action the AI is intended to change. Predefine the comparator (usual care) and the minimum clinically important effect on workflow or outcomes [28].

Step 2: Assemble representative datasets. Use datasets that reflect real-world prevalence, device and protocol heterogeneity, and clinically relevant edge cases. Clearly separate development data from external validation data and justify site selection [29,30].

Step 3: Specify a defensible reference standard. Define how ground truth is established (adjudication, follow-up, operative findings when appropriate) and protect against incorporation bias. Report blinding procedures for adjudicators in prospective evaluations [31,32].

Step 4: Pre-specify endpoints, metrics, and operating thresholds. Choose endpoints aligned with the clinical decision (for example, sensitivity at a triage threshold) and include calibration when probabilistic outputs are used. Lock thresholds before evaluation and report decision-analytic implications when relevant [33].

Step 5: Evaluate subgroup performance and fairness. Report performance across clinically relevant subgroups (age, sex, device, site, postoperative status when applicable) and identify any underperformance that could worsen inequity [34].

Step 6: Test workflow integration and human factors. Evaluate usability, interpretability, and operational fit within Picture Archiving and Communication Systems (PACS)/Radiology Information Systems (RIS). Measure clinician interaction effects, including automation bias and alert fatigue risk, using predefined process measures [35,36].

Step 7: Conduct prospective impact evaluation. Use prospective designs to quantify changes in clinician behavior, turnaround time, and clinically relevant errors, and evaluate downstream management effects when feasible. Use protocols consistent with AI trial guidance when interventional evaluation is planned [37].

Step 8: Assess cost and scalability. Evaluate resource utilization, downstream testing, throughput, and staffing implications to determine whether observed effects translate into sustainable service value [38].

Step 9: Implement post-deployment monitoring and governance. Define monitoring metrics, drift detection, audit frequency, and governance for model updates, including triggers for revalidation or rollback [39].

Reporting and Transparency Standards

Robust reporting and transparency are essential for reliability, reproducibility, and clinical translation of AI research in MSK imaging. Reporting guidelines support standardized description of study design and methods and help reviewers evaluate submissions [40,41].

Key guidelines relevant to AI in medical imaging include TRIPOD+AI for prediction model development and validation studies [42], CONSORT-AI for clinical trial reports evaluating interventions with an AI component [43,44], and SPIRIT-AI for clinical trial protocols involving AI interventions and robust prospective evaluation [39,45]. DECIDE-AI focuses on early-stage clinical evaluation of AI-driven decision support systems and specifies minimum reporting items in this phase [45]. CLAIM promotes transparent and reproducible reporting for medical imaging AI and provides a checklist for authors and a structured framework for reviewers [46,47].

For interpretation, guideline selection should match the study's primary claim. When the claim concerns model performance and validation, reviewers should expect TRIPOD+AI-aligned reporting [42]. When the claim concerns the clinical impact of an AI-enabled intervention, trial protocols and reports should align with SPIRIT-AI and CONSORT-AI [36,45]. When the claim concerns early-stage workflow evaluation or the feasibility of AI-driven decision support, DECIDE-AI is particularly relevant [45]. Across designs, reviewers should seek sufficient detail to enable independent replication [40] and should critically evaluate sources of bias and limitations while maintaining focus on human oversight of clinical decisions [48]. Studies lacking sufficient detail in design, methods, or results remain difficult to assess and may inhibit translation [41]; therefore, adherence to appropriate reporting guidance is an indicator of quality and clinical readiness [46,47].

Implementation Realities

Successful translation of AI into clinical use in MSK imaging depends on both technical validation and practical implementation. Integration with PACS and RIS is central because many commercial AI solutions still face difficulties integrating seamlessly into established radiology information technology systems and complex workflows [49,50]. Embedding AI tools across the radiology workflow requires standards and interoperability and is central to realizing potential diagnostic improvements in routine practice [51,52].

Regulatory considerations and liability also shape implementation. Attribution of responsibility in cases of error remains under active discussion, and regulatory frameworks for AI in medicine continue to evolve [53]. Although AI may augment decision-making, primary responsibility typically remains with the physician, with potential shared liability involving developers under specific circumstances within medical device regulations [54]. Organizations such as the World Health Organization and the US Food and Drug Administration have begun issuing ethical and regulatory frameworks to support safe AI use [55].

Data governance is foundational for development and deployment. Effective governance requires large, complete, harmonized datasets [56], and challenges include balancing privacy with innovation and developing common data models that capture relevant information [56,57]. Dataset quality and governance directly influence algorithm accuracy, robustness, and fairness [58]. Finally, clinician education is critical for implementation. Integrating AI into training programs can prepare clinicians to use and oversee AI tools appropriately and can address technological, workflow, and organizational gaps noted in the literature [9,59-61].

Ethical and Economic Considerations in Implementation

Ethical considerations are relevant across the full life cycle of AI in MSK imaging, from model development and validation to implementation and post-deployment monitoring. In clinical practice, AI should support rather than replace clinician judgment, because responsibility for patient care, interpretation of uncertain findings,

and communication of clinically relevant results remain embedded in professional decision-making. Key ethical issues include transparency of model outputs, human oversight, accountability for errors, privacy protection, data governance, bias, fairness, and the risk of automation bias. These concerns are particularly relevant in MSK imaging because AI outputs may influence emergency triage, surgical referral, follow-up imaging, opportunistic screening, and prioritization of specialist review. Therefore, implementation should include predefined escalation pathways, audit mechanisms, subgroup performance monitoring, and procedures for managing discordance between AI outputs and clinician interpretation [34,39,52-57].

Cost-effectiveness should also be assessed before routine adoption. AI tools may generate value if they reduce clinically significant errors, shorten reporting or escalation time, improve workflow efficiency, or increase appropriate follow-up. However, these potential benefits must be weighed against acquisition costs, software maintenance, integration with PACS/RIS, cybersecurity requirements, staff training, monitoring infrastructure, false positive burden, downstream testing, and possible inequities in access. For MSK imaging services, economic evaluation should therefore extend beyond diagnostic accuracy and include resource utilization, throughput, clinician workload, downstream referrals, patient follow-up, and sustainability in the local health system. Without such assessment, technically promising AI tools may fail to provide value in routine practice despite favorable validation metrics [26,28,39,49-52].

Conclusions

AI in MSK imaging is most likely to deliver clinical value when it targets narrow, well-defined decisions within real workflows, such as triage support and selected efficiency functions. However, high diagnostic performance alone should not be interpreted as proof of patient benefit. Tools can fail to add value, or cause harm, when generalizability is weak, reference standards are unreliable, subgroup performance is uneven, or workflow integration increases cognitive burden and automation bias. For orthopedic and MSK imaging services, adoption decisions should be anchored to clinically meaningful endpoints: changes in management, reduction in clinically significant errors, validated workflow improvements, and equity and safety impact. Minimum validation should include external validation across institutions and populations, prospective workflow evaluation, and explicit monitoring plans for drift and emerging bias after deployment.

Author contributions

FJVL: Conceptualization, Methodology, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing. **IDLM:** Writing – Original Draft, Writing – Review & Editing. All authors assume full responsibility for the content of the manuscript.

Funding

This study did not receive any funding.

Conflict of interest statement

The authors declare no conflicts of interest.

Generative Artificial Intelligence use statement

During the preparation of this work, the authors used Chat-GPT5 for English-language editing to improve clarity and grammar. After using this tool, the authors reviewed and carefully edited the content as needed and take full responsibility for the content of the publication.

Data availability statement

Not aplicable.

Acknowledgments

None.

ORCID

Fabriccio J. Visconti-Lopez: https://orcid.org/0000-0002-8056-2112

Ivan David Lozada-Martínez: https://orcid.org/0000-0002-1960-7334

REFERENCES

1. Chang CY, Link TM. Introduction to the special issue on artificial intelligence in musculoskeletal radiology. Skeletal Radiol. 2022;51(2):233. doi: 10.1007/s00256-021-03922-5.
2. Strohm L, Hehakaya C, Ranschaert ER, Boon WPC, Moors EHM. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. Eur Radiol. 2020;30(10):5525-5532. doi: 10.1007/s00330-020-06946-y.
3. York TJ, Raj S, Ashdown T, Jones G. Clinician and computer: a study on doctors' perceptions of artificial intelligence in skeletal radiography. BMC Med Educ. 2023;23(1):16. doi: 10.1186/s12909-022-03976-6.

4. Farhadi F, Barnes MR, Sugito HR, Sin JM, Henderson ER, Levy JJ. Applications of artificial intelligence in orthopaedic surgery. Front Med Technol. 2022;4:995526. doi: 10.3389/fmedt.2022.995526.
5. Fayaz-Bakhsh A, Tania J, Lutfi SL, Jha AK, Rahmim A. What Is Implementation Science: And Why It Matters for Bridging the Artificial Intelligence Innovation-to-Application Gap in Medical Imaging. PET Clin. 2026;21(1):1-16. doi: 10.1016/j.cpet.2025.09.002.
6. Christodoulou E, Reinke A, André P, Godau P, Kalinowski P, Houhou R, *et al.* False Promises in Medical Imaging AI? Assessing Validity of Outperformance Claims. arXiv (Cornell University) [Internet]. 2025 [cited 2025 Nov]. Available from: http://arxiv.org/abs/2505.04720
7. Mazurowski MA. Do We Expect More from Radiology AI than from Radiologists? Radiol Artif Intell. 2021;3(4):e200221. doi: 10.1148/ryai.2021200221.
8. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol. 2021;31(6):3797-3804. doi: 10.1007/s00330-021-07892-z.
9. Kim B, Romeijn S, van Buchem M, Mehrizi MHR, Grootjans W. A holistic approach to implementing artificial intelligence in radiology. Insights Imaging. 2024;15(1):22. doi: 10.1186/s13244-023-01586-4.
10. Pham N, Hill V, Rauschecker A, Lui Y, Niogi S, Fillipi CG, *et al.* Critical Appraisal of Artificial Intelligence-Enabled Imaging Tools Using the Levels of Evidence System. AJNR Am J Neuroradiol. 2023;44(5):E21-E28. doi: 10.3174/ajnr.A7850.
11. Guermazi A, Omoumi P, Tordjman M, Fritz J, Kijowski R, Regnard NE, *et al.* How AI May Transform Musculoskeletal Imaging. Radiology. 2024;310(1):e230764. doi: 10.1148/radiol.230764.
12. Román-Belmonte JM, De la Corte-Rodríguez H, Rodríguez-Damiani BA, Rodríguez-Merchán EC. Artificial Intelligence in Musculoskeletal Conditions. Artificial Intelligence. IntechOpen; 2023. doi: 10.5772/intechopen.110696.
13. Debs P, Fayad LM. The promise and limitations of artificial intelligence in musculoskeletal imaging. Front Radiol. 2023;3:1242902. doi: 10.3389/fradi.2023.1242902.
14. Ounasser N, Rhanoui M, Mikram M, Asri BE. A Systematic Review on Artificial Intelligence in Orthopedic Surgery. Revue d intelligence artificielle. International Information and Engineering Technology Association. 2024;38(4):1143. doi: 10.18280/ria.380409.
15. Oettl FC, Zsidai B, Oeding JF, Hirschmann MT, Feldt R, Fendrich D, *et al.* Artificial intelligence-assisted analysis of musculoskeletal imaging: A narrative review of the current state of machine learning models. Knee Surg Sports Traumatol Arthrosc. 2025;33:3032-3038. doi:10.1002/ksa.12702.
16. Tsang B, Gong B, Probyn L, Patlas MN. Artificial intelligence in emergency musculoskeletal imaging: A critical review of current applications. Diagn Interv Imaging. 2026:S2211-5684(26)00028-8. doi: 10.1016/j.diii.2026.02.003.
17. Dubey A, Uldin H, Khan Z, Panchal H, Iyengar KP, Botchu R. Role of artificial intelligence in musculoskeletal interventions. Cancers. 2025;17(10):1615. doi: 10.3390/cancers17101615.
18. Tordjman M, Fritz J, Regnard NE, Kijowski R, Mihoubi F, Taouli B, *et al.* Artificial intelligence in musculoskeletal radiology: Practical aspects and latest perspectives. BJR Open. 2025;7(1):tzaf029. doi:10.1093/bjro/tzaf029.
19. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, *et al.* Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368:m689. doi: 10.1136/bmj.m689.
20. Hirschmann A, Cyriac J, Stieltjes B, Kober T, Richiardi J, Omoumi P. Artificial Intelligence in Musculoskeletal Imaging: Review of Current Literature, Challenges, and Trends. Semin Musculoskelet Radiol. 2019;23(3):304-311. doi: 10.1055/s-0039-1684024.
21. Diao X, Wang X, Qin J, Wu Q, He Z, Fan X. A Review of the Application of Artificial Intelligence in Orthopedic Diseases. Computers, materials & continua. 2024;78(2):2617. doi: 10.32604/cmc.2024.047377.
22. Paluh JL, Sunil S, Rajeev RS, Chatterjee A, Pilitsis JG, Mukherjee A. SIENNA: A Generalizable Parameter-Efficient Machine Learning Diagnostic for Clinical Magnetic Resonance Imaging. Research Square [Internet]. 2024 [cited 2025 Aug]. Available from: https://doi.org/10.21203/rs.3.rs-4087784/v1
23. Hinterwimmer F, Consalvo S, Neumann J, Rueckert D, von Eisenhart-Rothe R, Burgkart R. Applications of machine learning for imaging-driven diagnosis of musculoskeletal malignancies-a scoping review. Eur Radiol. 2022;32(10):7173-7184. doi: 10.1007/s00330-022-08981-3.
24. Nelson AE, Arbeeva L. Narrative Review of Machine Learning in Rheumatic and Musculoskeletal Diseases for Clinicians and Researchers: Biases, Goals, and Future Directions. J Rheumatol. 2022;49(11):1191-1200. doi: 10.3899/jrheum.220326.
25. Boverhof BJ, Redekop WK, Bos D, Starmans MPA, Birch J, Rockall A, *et al.* Radiology AI Deployment and Assessment Rubric (RADAR) to bring value-based AI into radiological practice. Insights Imaging. 2024;15(1):34. doi: 10.1186/s13244-023-01599-z.
26. Park SH, Choi J, Byeon JS. Key Principles of Clinical Validation, Device Approval, and Insurance Coverage Decisions of Artificial Intelligence. Korean J Radiol. 2021;22(3):442-453.
27. You JG, Hernandez-Boussard T, Pfeffer MA, Landman A, Mishuris RG. Clinical trials informed framework for real world clinical implementation and deployment of artificial intelligence applications. NPJ Digit Med. 2025;8(1):107. doi: 10.1038/s41746-025-01506-4.
28. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, *et al.* Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature. 2022;28(5):924. doi: 10.1038/s41591-022-01772-9.
29. Benjamin M, Engelhard G, Aisen AM, Aradi Y, Benjamin E. A Multisite, Report-Based, Centralized Infrastructure for Feedback and Monitoring of Radiology AI/ML Development and Clinical Deployment. arXiv (Cornell University) [Internet]. 2022 [cited 2025 Nov]. Available from: http://arxiv.org/abs/2008.13781
30. Omoumi P, Ducarouge A, Tournier A, Harvey H, Kahn CE Jr, Louvet-de Verchère F, *et al.* To buy or not to buy-evaluating commercial AI solutions in radiology (the ECLAIR guidelines). Eur Radiol. 2021;31(6):3786-3796. doi: 10.1007/s00330-020-07684-x.
31. Duncan SF, Kidd AC, Lampignano JP, Cannon P, Hall M, Stobo DB, *et al.* Reference standard methodology in the clinical evaluation of AI chest X-ray algorithms for lung cancer detection: A systematic review. Eur J Radiol. 2025;192:112409. doi: 10.1016/j.ejrad.2025.112409.
32. Duggan GE, Reicher JJ, Liu Y, Tse D, Shetty S. Improving reference standards for validation of AI-based radiography. Br J Radiol. 2021;94(1123):20210435. doi: 10.1259/bjr.20210435.
33. Reinke A, Tizabi MD, Baumgartner M, Eisenmann M, Heckmann-Nötzel D, Kavur AE, *et al.* Understanding metric-related pitfalls in image analysis validation. Nat Methods. 2024;21(2):182-194. doi: 10.1038/s41592-023-02150-0.

34. Subbaswamy A, Sahiner B, Petrick N, Pai V, Adams R, Diamond MC, *et al.* A data-driven framework for identifying patient subgroups on which an AI/machine learning model may underperform. NPJ Digit Med. 2024;7(1):334. doi: 10.1038/s41746-024-01275-6.
35. de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, *et al.* Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med. 2022;5(1):2. doi: 10.1038/s41746-021-00549-7.
36. Juluru K, Shih HH, Keshava Murthy KN, Elnajjar P, El-Rowmeim A, Roth C, *et al.* Integrating AI Algorithms into the Clinical Workflow. Radiol Artif Intell. 2021;3(6):e210013. doi: 10.1148/ryai.2021210013.
37. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI, *et al.* Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med. 2020;26(9):1351-1363. doi: 10.1038/s41591-020-1037-7.
38. Alnasser B. A Review of Literature on the Economic Implications of Implementing Artificial Intelligence in Healthcare. E-Health Telecommunication Systems and Networks. 2023;12(3):35. doi: 10.4236/etsn.2023.123003.
39. Abulibdeh R, Celi LA, Sejdić E. The illusion of safety: A report to the FDA on AI healthcare product approvals. PLOS Digit Health. 2025;4(6):e0000866. doi: 10.1371/journal.pdig.0000866.
40. Klontzas ME, Gatti AA, Tejani AS, Kahn CE Jr. AI Reporting Guidelines: How to Select the Best One for Your Research. Radiol Artif Intell. 2023;5(3):e230055. doi: 10.1148/ryai.230055.
41. Park SH, Suh CH. Reporting Guidelines for Artificial Intelligence Studies in Healthcare (for Both Conventional and Large Language Models): What's New in 2024. Korean J Radiol. 2024;25(8):687-690. doi: 10.3348/kjr.2024.0598.
42. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi: 10.1136/bmj-2023-078378.
43. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI, *et al.* Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26(9):1364-1374. doi: 10.1038/s41591-020-1034-x.
44. Liu X, Cruz Rivera S, Moher D. Calvert MJ, Denniston AK; The SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26:1364-1374. doi: 10.1038/s41591-020-1034-x.
45. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, *et al.* Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022;377:e070904. doi: 10.1136/bmj-2022-070904.
46. Tejani AS, Klontzas ME, Gatti AA, Mongan J, Moy L, Park SH, *et al.* Updating the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) for reporting AI research. Nat Mach Intell. 2023;5(9):950. doi: 10.1038/s42256-023-00717-2.
47. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, *et al.* Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. Radiol Artif Intell. 2024;6(4):e240300. doi: 10.1148/ryai.240300.
48. Tong MW, Zhou J, Akkaya Z, Majumdar S, Bhattacharjee R. Artificial intelligence in musculoskeletal applications: a primer for radiologists. Diagn Interv Radiol. 2025;31(2):89-101. doi: 10.4274/dir.2024.242830.
49. Blezek DJ, Olson-Williams L, Missert A, Korfiatis P. AI Integration in the Clinical Workflow. J Digit Imaging. 2021;34(6):1435-1446. doi: 10.1007/s10278-021-00525-3.
50. Chatterjee N, Duda J, Gee J, Elahi A, Martin K, Doan V, *et al.* A Cloud-Based System for Automated AI Image Analysis and Reporting. J Imaging Inform Med. 2025;38(1):368-379. doi: 10.1007/s10278-024-01200-z.
51. Pérez-Sanpablo AI, Quinzaños-Fresnedo J, Gutiérrez-Martínez J, Lozano-Rodríguez IG, Roldan-Valadez E. Transforming Medical Imaging: The Role of Artificial Intelligence Integration in PACS for Enhanced Diagnostic Accuracy and Workflow Efficiency. Curr Med Imaging. 2025;21:e15734056370620. doi: 10.2174/0115734056370620250403030638.
52. Wiggins WF, Magudia K, Schmidt TMS, O'Connor SD, Carr CD, Kohli MD, *et al.* Imaging AI in Practice: A Demonstration of Future Workflow Using Integration Standards. Radiol Artif Intell. 2021;3(6):e210152. doi: 10.1148/ryai.2021210152.
53. Cestonaro C, Delicati A, Marcante B, Caenazzo L, Tozzo P. Defining medical liability when artificial intelligence is applied on diagnostic algorithms: a systematic review. Front Med (Lausanne). 2023;10:1305756. doi: 10.3389/fmed.2023.1305756.
54. Contaldo MT, Pasceri G, Vignati G, Bracchi L, Triggiani S, Carrafiello G. AI in Radiology: Navigating Medical Responsibility. Diagnostics (Basel). 2024;14(14):1506. doi: 10.3390/diagnostics14141506.
55. Saw SN, Ng KH. Current challenges of implementing artificial intelligence in medical imaging. Phys Med. 2022;100:12-17. doi: 10.1016/j.ejmp.2022.06.003.
56. Kondylakis H, Kalokyri V, Sfakianakis S, Marias K, Tsiknakis M, Jimenez-Pastor A, *et al.* Data infrastructures for AI in medical imaging: a report on the experiences of five EU projects. Eur Radiol Exp. 2023;7(1):20. doi: 10.1186/s41747-023-00336-x.
57. Lutz de Araujo A, Wu J, Harvey H, Lungren MP, Graham M, Leiner T, *et al.* Medical Imaging Data Calls for a Thoughtful and Collaborative Approach to Data Governance. PLOS Digit Health. 2025;4(10):e0001046. doi: 10.1371/journal.pdig.0001046.
58. Jiménez-Sánchez A, Avlona NR, Juodelyte D, Sourget T, Vang-Larsen C, Zając HD, *et al.* Copycats: the many lives of a publicly available medical imaging dataset. arXiv (Cornell University) [Internet]. 2024 [cited 2025 Nov]. Available from: http://arxiv.org/abs/2402.06353
59. van Kooten MJ, Tan CO, Hofmeijer EIS, van Ooijen PMA, Noordzij W, Lamers MJ, *et al.* A framework to integrate artificial intelligence training into radiology residency programs: preparing the future radiologist. Insights Imaging. 2024;15(1):15. doi: 10.1186/s13244-023-01595-3.
60. Soleimantabar H, Mahdavi A, Qorbani M, Madanipour M, Tasharrofi MA, Okhovvat B. Artificial Intelligence in Radiology: Perceptions, Adoption Barriers, and Trust Among Iranian Radiologists in a Global Context. InfoScience Trends. 2025;2(3):1. doi: 10.61186/ist.202502.03.01.
61. Ng KH, Tan CH. It is Time to Incorporate Artificial Intelligence in Radiology Residency Programs. Korean J Radiol. 2023;24(3):177-179. doi: 10.3348/kjr.2022.1023.