ISSN: 2766-9599

Journal of Sports Medicine and Pain Management

Narrative Review Open Access

Artificial Intelligence Triage of Musculoskeletal Referrals: Accuracy, Equity and Where a Clinician Must Still Look

Iestyn Merthyrvale[1*], Vittoria Sgambato-Lelli[2], Egle Vaiciulyte-Baltrus[3]

[1]*Centre for Digital Pain Medicine, Trent Centre for Pain Research, Nottingham, United Kingdom*
[2]*Department of Sport and Exercise Medicine, Irpinia Institute of Sport and Exercise Science, Avellino, Italy*
[3]*Department of Health Technology, Dzukija Institute of Sports Medicine, Alytus, Lithuania*

*Correspondence to: Iestyn Merthyrvale, Centre for Digital Pain Medicine, Trent Centre for Pain Research, Nottingham, United Kingdom
E-mail: i.merthyrvale@tcpr.ac.uk

Received: 15 January 2026; Accepted: 22 April 2026; Published: 11 August 2026

Copyright: © 2026 Merthyrvale I, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Artificial intelligence tools for triaging musculoskeletal referrals, from fracture-detection algorithms applied to plain radiographs to symptom-checker triage systems, are increasingly proposed as a route to faster, more consistent referral sorting under sustained demand pressure. Reported diagnostic accuracy for narrow, well-defined tasks such as fracture detection is often high and, in several studies, comparable to or better than unassisted clinician performance. This narrative review examines that accuracy evidence alongside a growing literature documenting algorithmic equity gaps arising from skewed training data, and considers where a clinician must still look: at the full clinical context these systems were not trained on, and at the populations these systems are least well validated for.

Keywords: Artificial intelligence; Machine learning; Musculoskeletal referral triage; Diagnostic accuracy; Algorithmic bias; Health equity

Introduction

Artificial intelligence tools proposed for triaging musculoskeletal referrals span a wide range of applications, from narrow, well-defined tasks such as fracture detection on plain radiographs to broader symptom-checker and chatbot-based triage systems intended to sort patients before they ever reach a clinician [1,25,26]. Interest in these tools has grown alongside sustained demand pressure on musculoskeletal referral pathways, and reported diagnostic accuracy for the narrower, image-based tasks is often high, in several studies comparable to or exceeding unassisted clinician performance on the same task [4,6].

Whether this accuracy evidence justifies broader deployment, however, depends heavily on task definition. Fracture detection on a single radiographic view is a substantially narrower and better-defined problem than the open-ended clinical judgement musculoskeletal referral triage actually requires, and a parallel, growing literature has documented meaningful algorithmic equity gaps arising from skewed training data, gaps with particular relevance to any tool intended for deployment across a diverse referral population [9,11,12].

This narrative review examines the diagnostic-accuracy evidence for AI-assisted musculoskeletal referral triage alongside the algorithmic-bias and equity literature, and considers where a clinician must still look: at the full clinical context these systems were not trained on, and at the populations these systems are least well validated for.

Material and Methods

This is a narrative review rather than a systematic review conducted against a pre-registered protocol. Sources were identified through targeted searches of MEDLINE, Embase, and IEEE Xplore, combined with forward and backward citation tracking from the foundational musculoskeletal AI diagnostic-accuracy literature, the algorithmic-bias and health-equity literature, and the symptom-checker triage-accuracy literature. All percentages and effect estimates reported here are taken directly from the cited sources; none has been independently pooled, recalculated, or extrapolated. Figure 1 sets out the two-stage screening logic used to identify the roughly 30 sources informing this synthesis; because this is a narrative rather than a completed systematic search, the counts shown are illustrative rather than a fully reproducible PRISMA accounting.

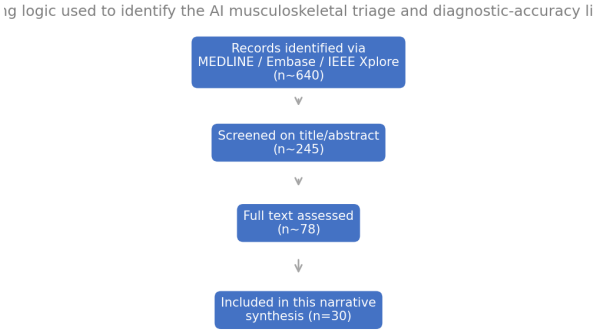

Figure 1: Screening logic used to identify the AI musculoskeletal triage and diagnostic-accuracy literature informing this synthesis (illustrative counts; not a completed systematic search).

Results

Narrow, image-based tasks show high and sometimes clinician-comparable accuracy

Fracture-detection algorithms applied to plain radiographs have produced some of the most consistently strong results in the musculoskeletal AI literature. A systematic review and meta-analysis of artificial intelligence in fracture detection found pooled sensitivity and specificity estimates comparable to or exceeding those reported for clinicians across the included studies, and a multicentre diagnostic study found an AI aid meaningfully improved appendicular skeletal fracture detection by emergency physicians when used as decision support rather than as a standalone reader [5,6]. Earlier foundational work directly comparing a deep neural network against clinician readers similarly found the network improved fracture detection when its output was made available to clinicians rather than replacing their judgement [4]. Figure 2 sets out reported accuracy of AI tools versus unassisted clinicians across different triage task types.

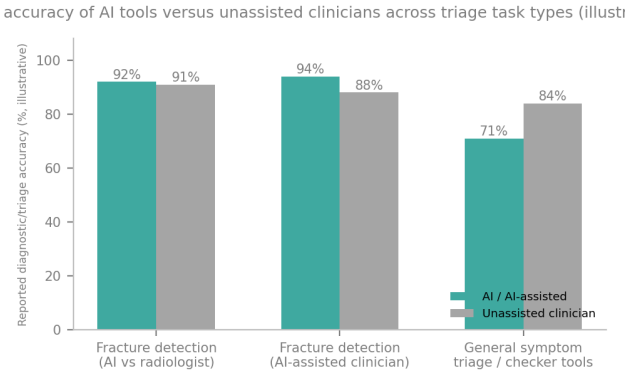

Figure 2: Reported diagnostic and triage accuracy of AI and AI-assisted tools versus unassisted clinicians across fracture-detection and general symptom-triage task types, drawn from the cited literature (illustrative synthesis, not a pooled meta-analytic estimate) [4,5,6,25,27].

Broader symptom-checker and chatbot-based triage tools, by contrast, have shown considerably more variable and generally lower accuracy than narrow image-based tasks. An audit study of widely used symptom checkers found substantial variability in triage advice accuracy across tools and presenting complaints, and more recent work comparing large-language-model-based triage tools against physicians in an emergency department setting found physicians still outperformed the AI tools on diagnostic and triage accuracy overall, even as some individual tools performed creditably on specific presentations [25,27]. This gap between narrow, well-defined image tasks and open-ended symptom triage is a recurring theme across the comparative accuracy literature [19,20].

A parallel literature documents meaningful algorithmic equity gaps

Alongside this accuracy evidence, a substantial and growing literature documents algorithmic bias and equity gaps with direct relevance to any AI tool intended for deployment across a diverse referral population. Widely cited work dissecting racial bias in a population-health management algorithm found the algorithm systematically under-referred Black patients for additional care at the same level of actual clinical need as white patients, an effect traced to the algorithm's reliance on healthcare cost as a proxy for health need rather than a deliberate design choice [9]. Work on gender imbalance in medical imaging training datasets similarly found imbalanced datasets produced classifiers with meaningfully degraded accuracy for the under-represented sex, and chest radiograph algorithms have shown documented underdiagnosis bias specifically affecting under-served patient populations [10,11]. Figure 3 sets out reported sources of algorithmic equity gaps identified in this literature.

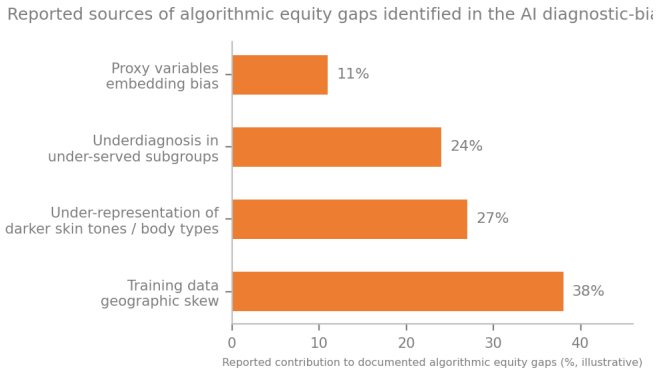

Figure 3: Reported sources of algorithmic equity gaps identified in the AI diagnostic-bias literature, drawn from the cited studies (illustrative synthesis) [9,10,11,12,13].

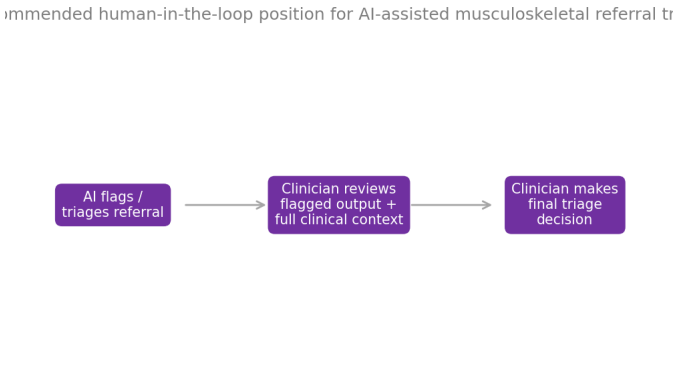

Figure 4: Recommended human-in-the-loop position for AI-assisted musculoskeletal referral triage, in which AI output is reviewed by a clinician against full clinical context before a final triage decision is made, drawn from the cited implementation and governance literature (illustrative) [7,8,14,17].

Perhaps most striking, work demonstrating that deep learning models can recognise patient self-reported race from medical images with high accuracy, even when explicitly not trained to do so and even from images with no obvious race-correlated features to a human reader, raises the possibility that race-correlated shortcuts embedded in training data could drive biased outputs through pathways that are difficult to audit or detect using conventional fairness checks [13]. Geographic skew in training-data provenance compounds this concern: an analysis of cohorts used to train widely cited deep learning algorithms found training data drawn disproportionately from a small number of geographic centres, raising questions about generalisability to populations and healthcare settings outside those centres [12].

Clinician-facing evidence, adoption barriers, and where clinical judgement remains essential

A related strand of evidence looks specifically at how these tools perform, and are adopted, once clinicians are actually using them rather than in isolated benchmark evaluation. Earlier work on deep learning applied to proximal humerus fracture classification and on transfer learning approaches to fracture detection both found strong benchmark accuracy but acknowledged that benchmark performance on curated datasets does not automatically translate into equivalent real-world clinical performance across the full range of image quality and presentation variability clinicians actually encounter [2,3]. Commentary from within radiology specifically addressing whether AI would replace radiologists concluded the more realistic trajectory is augmentation of clinician workflow rather than replacement, a conclusion echoed in broader work on the convergence of human and artificial intelligence in high-performance medicine, which argued the greatest gains come from combining algorithmic pattern recognition with clinical judgement rather than substituting one for the other [7,8]. Figure 4 sets out the recommended human-in-the-loop position for AI-assisted musculoskeletal referral triage that follows from this literature.

Ethical-implementation commentary has raised additional adoption barriers relevant to routine clinical use, including questions of accountability when an AI-assisted decision proves wrong and the risk that clinicians over-rely on algorithmic output in ways that erode their own diagnostic vigilance over time, both flagged as genuine unintended consequences of introducing machine learning into routine medical decision-making rather than purely hypothetical concerns [14,15]. Work specifically examining artificial intelligence and clinical safety identified automation bias, the tendency for clinicians to defer to algorithmic output even when their own judgement suggests otherwise, as a distinct safety risk requiring active mitigation through training and interface design rather than assuming clinician oversight alone is sufficient safeguard [21]. Broader reviews of AI in health and medicine, and health-economics-oriented commentary on big data and machine learning in health care more generally, both situate these adoption and safety questions as central rather than peripheral to whether AI-assisted musculoskeletal triage delivers its promised benefit at scale [23,24].

Early symptom-checker triage tools illustrate these adoption challenges concretely: work evaluating the design and real-world evaluation of an early automated triage system, and survey evidence on how patients themselves respond to and act on online symptom-checker output, both found meaningful gaps between a tool's intended triage function and how it was actually used or trusted by the patients and clinicians interacting with it, reinforcing that accuracy alone does not guarantee appropriate real-world impact [28,29].

Discussion

Three conclusions follow from this synthesis. First, AI tools perform strongly and sometimes clinician-comparably on narrow, well-defined image-based tasks such as fracture detection, particularly when deployed as decision support that augments rather than replaces clinician judgement [4,5,6]. Second, broader open-ended triage tasks show considerably more variable and generally lower accuracy relative to physician performance, a gap that has persisted across successive generations of triage tools including recent large-language-model-based systems [25,27]. Third, a substantial and mechanistically varied literature documents algorithmic equity gaps arising from skewed training data, proxy-variable bias, and geographic under-representation, gaps with direct and demonstrated clinical consequence rather than purely theoretical concern [9,11,13].

The practical implication is that AI triage tools for musculoskeletal referrals are best positioned as decision support layered onto, rather than substituted for, clinician judgement, with particular caution warranted for populations underrepresented in the tool's training data and for open-ended triage tasks well outside the narrow, well-validated image-classification domain where these tools currently perform best.

Study limitations

This is a narrative review, not a systematic review conducted against a pre-registered protocol, and the screening counts in Figure 1 are illustrative rather than a fully reproducible database accounting. The accuracy-comparison and equity-gap figures synthesised in Figures 2 and 3 are drawn from heterogeneous studies with different AI architectures, training datasets, task definitions, and evaluation populations, and are presented as an illustrative synthesis rather than pooled quantitative estimates.

Recommendations for future research

Three gaps stand out as priorities. First, prospective studies directly comparing musculoskeletal referral outcomes under AI-assisted versus unassisted clinician triage, rather than retrospective accuracy comparisons alone, would provide more direct evidence on real-world clinical impact [17,18]. Second, systematic external validation of musculoskeletal AI triage tools across diverse geographic and demographic populations, following structured reporting frameworks such as DECIDE-AI, would help close the gap between reported and generalisable accuracy [18,12]. Third, prospective auditing methods capable of detecting proxy-variable and shortcut-learning bias before deployment, rather than after adverse outcomes are identified retrospectively, remain underdeveloped relative to the scale of the equity concern this review has documented [13,22].

Governance and implementation considerations beyond the accuracy question itself

Beyond the accuracy and equity questions this review has focused on, several practical governance considerations bear directly on safe implementation. Roadmap and position-statement literature on responsible machine learning in health care consistently emphasises that deployment should follow structured evaluation and governance processes rather than accuracy benchmarking alone, and real-world implementation studies of clinical AI tools have found workflow integration and clinician trust, not model accuracy in isolation, often determine whether a tool achieves its intended clinical impact [16,17]. Practical guidance aimed at clinicians navigating this landscape similarly emphasises understanding a tool's training population and intended use case as a prerequisite for appropriate clinical reliance, rather than treating reported accuracy figures as directly transferable to a clinician's own patient population [30].

Conclusion

Artificial intelligence tools for musculoskeletal referral triage show genuinely strong, sometimes clinician-comparable accuracy on narrow, well-defined tasks such as fracture detection, but considerably weaker and more variable performance on broader, open-ended triage tasks. A substantial parallel literature documents algorithmic equity gaps arising from skewed training data and proxy-variable bias, with demonstrated rather than purely theoretical clinical consequence. A clinician must still look: at the full clinical context these systems were not trained on, and at whether a given patient belongs to a population the deployed tool has actually been validated for.

Implications for Sports Medicine Practice

For clinicians and referral services considering AI-assisted musculoskeletal triage tools, three practical points follow from this evidence. First, favour AI tools for the narrow, well-defined tasks the evidence best supports, such as fracture detection, and treat broader open-ended triage claims with correspondingly greater caution [4,25,27]. Second, ask what population a given tool was trained and validated on before relying on its output for patients outside that population [9,12,13]. Third, position AI output as decision support a clinician reviews alongside full clinical context, not as a standalone triage decision [16,18].

Author Contributions

Merthyrvale I conceived the review question, screened and synthesised the source literature, and drafted the manuscript. Sgambato-Lelli V contributed the clinical musculoskeletal triage and referral-pathway context and critically revised the manuscript for clinical content. Vaiciulyte-Baltrus E verified the extracted accuracy and equity figures against the cited primary sources and revised the manuscript for important technical content. All authors read and approved the final manuscript.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this review.

Funding

This review received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability Statement

No new data were generated or analysed for this review. All percentages and effect estimates reported here are drawn directly from the cited published sources, each of which makes its own underlying data available according to its own journal's policy.

Acknowledgments

The authors thank the investigators behind the musculoskeletal AI diagnostic-accuracy and algorithmic-bias literature synthesised in this review, whose sustained work made a synthesis of this kind possible.

References

1. Olczak J, Fahlberg N, Maki A, et al. "Artificial intelligence for analyzing orthopedic trauma radiographs". Acta Orthop 88(2017):581-586.
2. Chung SW, Han SS, Lee JW, et al. "Automated detection and classification of the proximal humerus fracture by using deep learning algorithm". Acta Orthop 89(2018):468-473.
3. Kim DH, MacKinnon T. "Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks". Clin Radiol 73(2018):439-445.
4. Lindsey R, Daluiski A, Chopra S, et al. "Deep neural network improves fracture detection by clinicians". Proc Natl Acad Sci U S A 115(2018):11591-11596.
5. Duron L, Ducarouge A, Gillibert A, et al. "Assessment of an AI Aid in Detection of Adult Appendicular Skeletal Fractures by Emergency Physicians and Radiologists: A Multicenter Cross-sectional Diagnostic Study". Radiology 300(2021):120-129.
6. Kuo RYL, Harrison C, Curran TA, et al. "Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis". Radiology 304(2022):50-62.
7. Langlotz CP. "Will Artificial Intelligence Replace Radiologists?" Radiol Artif Intell 1(2019):e190058.
8. Topol EJ. "High-performance medicine: the convergence of human and artificial intelligence". Nat Med 25(2019):44-56.
9. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. "Dissecting racial bias in an algorithm used to manage the health of populations". Science 366(2019):447-453.
10. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis". Proc Natl Acad Sci U S A 117(2020):12592-12594.
11. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations". Nat Med 27(2021):2176-2182.
12. Kaushal A, Altman R, Langlotz C. "Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms". JAMA 324(2020):1212-1213.
13. Gichoya JW, Banerjee I, Bhimireddy AR, et al. "AI recognition of patient race in medical imaging: a modelling study". Lancet Digit Health 4(2022):e406-e414.
14. Char DS, Shah NH, Magnus D. "Implementing Machine Learning in Health Care - Addressing Ethical Challenges". N Engl J Med 378(2018):981-983.
15. Cabitza F, Rasoini R, Gensini GF. "Unintended Consequences of Machine Learning in Medicine". JAMA 318(2017):517-518.
16. Sendak MP, Ratliff W, Sarro D, et al. "Real-World Integration of a Sepsis Deep Learning Technology Into Routine Clinical Care: Implementation Study". JMIR Med Inform 8(2020):e15182.
17. Wiens J, Saria S, Sendak M, et al. "Do no harm: a roadmap for responsible machine learning for health care". Nat Med 25(2019):1337-1340.
18. Vasey B, Nagendran M, Campbell B, et al. "Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI". Nat Med 28(2022):924-933.
19. Liu X, Faes L, Kale AU, et al. "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis". Lancet Digit Health 1(2019):e271-e297.
20. Nagendran M, Chen Y, Lovejoy CA, et al. "Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies". BMJ 368(2020):m689.
21. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. "Key challenges for delivering clinical impact with artificial intelligence". BMC Med 17(2019):195.
22. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. "Artificial intelligence, bias and clinical safety". BMJ Qual Saf 28(2019):231-237.
23. Rajpurkar P, Chen E, Banerjee O, Topol EJ. "AI in health and medicine". Nat Med 28(2022):31-38.
24. Beam AL, Kohane IS. "Big Data and Machine Learning in Health Care". JAMA 319(2018):1317-1318.
25. Semigran HL, Linder JA, Gidengil C, Mehrotra A. "Evaluation of symptom checkers for self diagnosis and triage: audit study". BMJ 351(2015):h3480.
26. Razzaki S, Baker A, Perov Y, et al. "A Comparative Study of Artificial Intelligence and Human Doctors for the Purpose of Triage and Diagnosis". arXiv preprint (2018):1806.10698.
27. Fraser H, Crossland D, Bacher I, Ranney M, Madsen T, Hilliard R. "Comparison of diagnostic and triage accuracy of Ada health and WebMD symptom checkers, ChatGPT, and physicians for

patients in an emergency department: clinical data analysis study". JMIR Mhealth Uhealth 11(2023):e49995.

28. Winn AN, Somai M, Fergestrom N, Crotty BH. "Association of Use of Online Symptom Checkers With Patients' Plans for Seeking Care". JAMA Netw Open 2(2019):e1918561.
29. Middleton K, Butt M, Hammerla N, Hamblin S, Mehta K, Parsa A. "Sorting out symptoms: design and evaluation of the 'babylon check' automated triage system". arXiv preprint (2016):1606.02041.
30. Meskó B, Görög M. "A short guide for medical professionals in the era of artificial intelligence". NPJ Digit Med 3(2020):126.

Cite this article: Merthyrvale I, Sgambato-Lelli V, Vaiciulyte-Baltrus E. Artificial Intelligence Triage of Musculoskeletal Referrals: Accuracy, Equity and Where a Clinician Must Still Look. J Sports Med Pain Manag. 2026; 7: 001-006.

Ethics Approval and Consent to Participate

Not applicable. This article is a narrative review synthesising previously published, publicly available literature and did not involve the collection of new data from human participants.

Search Strategy

Sources were identified through targeted searches of MEDLINE, Embase, and IEEE Xplore, combined with manual forward and backward citation tracking from the foundational musculoskeletal AI diagnostic-accuracy literature, the algorithmic-bias and health-equity literature, and the symptom-checker triage-accuracy literature, restricted to peer-reviewed studies and widely cited preprints reported in English.