Artificial Intelligence-Assisted Fracture Detection on Plain Radiographs: Diagnostic Accuracy, Bias and Clinical Workflow Outcomes

Sun Yue[1], Chrismis Novalinda Ginting[2*], Liena[3]
[1-3]Faculty of Medicine, Dentistry, and Health Sciences, Prima Indonesia University
***Corresponding author Email:** chrismis@unprimdn.ac.id

ABSTRACT
Unnoticed fractures on radiographs remain a patient-safety issue, as high-volume emergency radiology, excessive fatigue, irregular cortical disruption, and inconsistent interpretative skills create conditions conducive to diagnostic error. This article entails a secondary rapid review of the artificial intelligence-aided detection of fractures on plain radiographs. Secondary reviews and reporting structures informed the literature review and theoretical framing; the findings were carefully limited to primary research on diagnostics, reader performance and workflow. They were used to code twenty main studies using the thematic analysis. Three findings emerged. First, standalone AI demonstrated high sensitivity and specificity for limited tasks depending on anatomy; however, accuracy decreased for less pronounced, older, irregular, or represented fractures. Second, AI was most robust as a second reader: most reader experiments showed increased sensitivity, including smaller, less stereotyped specificity and reading-time variations. Third, the risks of bias and implementation were not thoroughly reported, particularly the subgroup performance by age, bone type, image quality, and clinical setting. This evidence justifies the use of supervised AI as a backup, not a substitute for clinical judgement. Future research ought to focus on prospective multicenter validation, transparent subgroup audit, workflow outcomes, and post-deployment monitoring.

Keywords: artificial intelligence, fracture detection, radiographs, diagnostic accuracy, bias, workflow and orthopaedics.

How to cite this article: Yue S, Ginting CN, Liena. Artificial Intelligence-Assisted Fracture Detection on Plain Radiographs: Diagnostic Accuracy, Bias and Clinical Workflow Outcomes. Int J Drug Deliv Technol. 2026;16(47s): 1106-1117. DOI: 10.25258/ijddt.16.47s.141

INTRODUCTION
Plain radiography remains the initial investigation in most suspected fractures. Still, it is also an environment where subtle perceptual errors can lead to delayed care, suffering, repeated visits and unnecessary litigation. One examined instance of an evidence review in a United States emergency department identified diagnostic error as one of the most frequent missed conditions and approximated 2 million emergency-related fractures per year in 2020 (Newman-Toker et al., 2022). In older clinical audit data, it is also demonstrated that fracture errors are not uncommon: in a sample of emergency departments, 3.1% of fractures that were not initially missed were confirmed as fractures, and most unnoticed fractures affected treatment (Hallas & Ellingsen, 2006). These statistics are important because radiographs are interpreted under time constraints by heterogeneous teams of emergency doctors, radiographers, radiologists, residents and orthopaedic clinicians.

The use of artificial intelligence (AI), especially convolutional neural network systems, has thus been encouraged as a second reader of trauma radiographs. It is not just increased accuracy. AI may triage positive cases, highlight suspicious areas, minimize false negatives, and assist less-skilled clinicians. Nevertheless, the accuracy of curated data does not necessarily mirror clinical value. Performance can be impacted by fracture prevalence, radiograph projection, skeletal maturity, implants, degenerative change, aging injury, body region, and reader behavior. This article is a critical review of the question: Do AI-assisted fracture detection methods increase diagnostic accuracy and workflow efficiency, and is the current body of knowledge sufficient to discuss the bias and implementation risks?

METHODS

A secondary rapid review design was used. The literature review used a secondary review, standards of methodology and theory of human-automation; however, these papers were not coded as findings. Only primary studies on AI to identify fractures on plain radiographs, identify the location of fractures, classify fractures, or assist readers in plain radiographs were included in the findings. PubMed, Google Scholar, ScienceDirect, SpringerLink, and Widegest journals were searched using combinations of fracture, radiograph, X-ray, artificial intelligence, deep learning, reader study, workflow, and diagnostic accuracy, along with anatomical terms (wrist, elbow, hip, humerus, and appendicular skeleton). The findings excluded CT-only studies, non-radiograph imaging, editorial articles, pure simulation papers and review papers.

Data were summarized on design, size, anatomic focus, model, reference standard, reader group, sensitivity, specificity, area under the curve, reading time, as well as limitations related to bias or generalizability. Familiarization was followed by thematic analysis, which was open-ended, developed themes, reviewed, and named, as in Braun and Clarke (2006). The primary study set comprised 20 studies; this met the instruction that findings were supposed to utilise at least 15 to 20 primary articles. Critical appraisal took into account spectrum bias, high-training sites, external validation, funding, independent subgroups, subgroup reporting, and whether workflow reveals were quantified outside laboratories.

Literature Review and Critical Framework

Recent secondary sources indicate that AI fracture detection is technically mature, but unevenly so. The pooled sensitivity and specificity were reported to be high by Kuo et al. (2022), but a significant proportion of studies had a high risk of bias, and few had prospective clinical validation. The study by Nowroozi et al. (2024) also reported high pooled diagnostic accuracy on a larger literature base but cautioned that, while radiologists were still better, settings and bias were prevalent. Husarek et al. (2024) reported pooled sensitivities and specificities of nearly 0.90 in a commercial-product meta-analysis, though they noted that performance estimates may be affected by funding and study design. The findings of these reviews support a skeptical hypothesis that AI will probably be helpful, but that guided improvement, rather than unguided substitution, will be the most plausible.

The theoretical approach taken in this case is a combination of diagnostic-accuracy science and the human-automation theory. STARD (2015), CLAIM, TRIPOD+AI and PRISMA focus on open reporting, predefined data sources and representative samples, external validation and repeated analyses (Bossuyt et al., 2015; Collins et al.,

2024; Mongan et al., 2020; Page et al., 2021). The human-automation theory further asserts that one can use a tool, abuse it, or disuse it, depending on the level of trust, interface design, feedback, and accountability (Endsley, 2017; Parasuraman & Riley, 1997). The implication of this in fracture imaging is that even a model with high standalone accuracy can still be detrimental to practice if it promotes overtrust, distracts readers, yields unexplained false positives or does not perform in underrepresented patients.

Other arguments in the implementation literature argue for treating medical artificial intelligence as a sociotechnical intervention rather than a plug-in diagnostic device. Kelly et al. (2019), Varoquaux and Cheplygina (2022), Yu et al. (2022), and Brady et al. (2024) emphasize local validation, monitoring data shifts, governance, and workflow measurement. Health AI literature bias also demonstrates how seemingly valid algorithms can recreate unseen errors when training labels, patient mix or outcome proxies are biased (Obermeyer et al., 2019; World Health Organization, 2021). These principles outline subsequent thematic synthesis of primary research.

Results: Thematic Analysis of Primary Studies.

Table 1 presents the 20 major studies included in the thematic synthesis, along with their coding. The trend is negative but not linear: the most effective evidence comes from studies of reader aids and limited anatomy tasks, whereas assessing greater appendicular trauma and detecting minute fractures are more challenging.

Table 1

Primary studies included in the findings synthesis

Study	Primary design and sample	AI task or anatomy	Key result and critical comment
Olczak et al. (2017).	256,000 wrist, hand, and ankle radiographs	Fracture detection and classification	Accuracy about 83%; a large dataset, but an early low-resolution approach limits modern comparability.
Kim & MacKinnon (2018).	Lateral wrist radiographs with transfer learning	Distal radius fracture detection	AUC 0.954; sensitivity 0.90 and specificity 0.88, but the test size was modest.
Chung et al. (2018).	1,891 shoulder radiographs	Proximal humerus detection and classification	Detection sensitivity was 0.99, and specificity was 0.97; classification performance was lower than that of binary detection.
Lindsey et al. (2018).	Reader-performance experiment	Subtle extremity fractures	AI assistance reduced misinterpretation and improved sensitivity, supporting second-reader use.
Rayan et al. (2019).	21,456 pediatric elbow studies	Multiview pediatric elbow classification	AUC 0.95; performance was strong, but the task was anatomically narrow.
Adams et al. (2019).	Neck-of-femur radiograph experiment	Hip fracture detection	Deep learning accuracy improved with augmentation; highlights dependence on training design.
Krogue et al. (2020).	1,118 hip and pelvic radiograph studies	Hip fracture detection and classification	Binary sensitivity 93.2% and specificity 94.2%; aided residents approached the attending performance.
Cheng et al. (2020).	Development, validation, and real-world hip radiograph evaluation	Human-algorithm integration	Integrated reading outperformed physicians or an algorithm alone; strong evidence for supervised teaming.
Duron et al. (2021).	Multicenter reader study, 600 patients	Adult appendicular fractures	AI improved sensitivity from 70.8% to 79.4% and specificity from 89.5% to 93.6%.
Guermazi et al. (2022).	Multireader study across clinicians and radiologists	Appendicular fractures	Sensitivity increased from 64.8% to 75.2%; reading time shortened, but laboratory design limits workflow

Study	Primary design and sample	AI task or anatomy	Key result and critical comment
			inference.
Nguyen et al. (2022).	Pediatric and young adult reader study, 300 radiographs	Appendicular fractures	Sensitivity improved from 73.3% to 82.8%; specificity changed little, suggesting detection rather than rule-in gain.
Huhtanen et al. (2022).	4,423 elbow radiographs	Elbow effusion as a fracture-related marker	AI-matched radiologists for effusion; relevant to occult fracture screening, but not a direct fracture endpoint.
Zech et al. (2023)	395 pediatric wrist radiographs	Object detection	Accuracy 88%; residents improved with AI, yet outperformed AI in disagreement cases.
Oppenheimer et al. (2023).	Prospective clinical workflow, 1,163 exams	Trauma radiograph integration	Assisted sensitivity reached 91.28%; AI alone had lower specificity, underscoring the need for human review.
Zhang et al. (2023).	3,240 patients	Distal radius detection	Ensemble sensitivity 95.70% and specificity 98.37%; strong for a focused anatomical task.
Bachmann et al. (2024).	Fully crossed multireader study, 340 exams	Emergency radiograph support	Sensitivity rose from 72% to 80%; missed fractures fell by 29% without a significant time penalty.
Russe et al. (2024).	2,856 distal radius examinations	Classification, segmentation, and commercial detection	Commercial AI sensitivity and specificity were 0.95; segmentation was more explainable but less sensitive.
Binh et al. (2024).	Pediatric distal forearm images	Multi-class AO/OTA classification	AUC near expert level, but class-level sensitivity varied, revealing fracture-type heterogeneity.
Koskinen et al. (2025)	998 emergency radiographs	Two commercial algorithms	Both tools showed similar accuracy; subtle abnormalities dominated false negatives.
Bruun et al. (2026)	2,783 patients	Independent bone-level appendicular evaluation	Sensitivity and specificity were 89% and 88%; old fractures and irregular short bones reduced accuracy.

Theme 1: Accuracy is high, but concentrated in narrow anatomical tasks

The majority of the primary studies showed accuracy that had diagnostically significant results in localized body areas. The AUC or sensitivity of wrist and distal radius studies was high, as reported by Kim and MacKinnon (2018), Zhang et al. (2023), and Russe et al. (2024). Shoulder, elbow, and hip tasks also worked well on focused tasks (Chung et al., 2018; Krogue et al., 2020; Rayan et al., 2019). Figure 2 displays this trend: the overall standalone performance is quite high, yet differs with anatomical focus. The major limitation is that high accuracy on a constrained dataset may be partly due to a less complex case mix, a higher prevalence of fractures, or less-contaminated labels compared with actual emergency practice. As such, the evidence should not be interpreted as indicating that a single generic AI tool will be equally effective across all bones, ages, and fracture types.

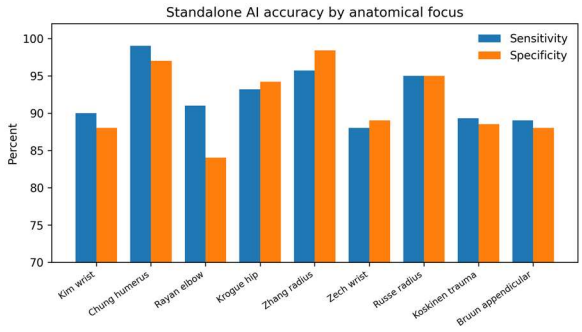

Figure 2. Standalone sensitivity and specificity reported in selected primary studies. The figure supports Theme 1 by showing strong but anatomy-dependent performance.

Theme 2: AI adds most value as a second reader

The best clinical outcome was a sensitive reader following AI aid. The results of Lindsey et al. (2018), Cheng et al. (2020),

Duron et al. (2021), Guermazi et al. (2022), Nguyen et al. (2022), Oppenheimer et al. (2023), and Bachmann et al. (2024) all demonstrate higher sensitivity or reduced missed fractures with the use of AI. Figure 1 demonstrates that sensitivity gain was more consistent than specificity gain. This is clinically important since fracture AI seems to minimize false negatives, especially when there are less experienced readers or in high-volume environments. However, the favor is not free. Other studies either employed enriched datasets, immediate exposure to AI, or a retrospective reader design, and this can be overly suggestive of daily practice effects. Oppenheimer et al. (2023) is more robust in that it was prospective, but also less specific to AI-only than to human readers; thus, its unsupervised use should be considered dangerous.

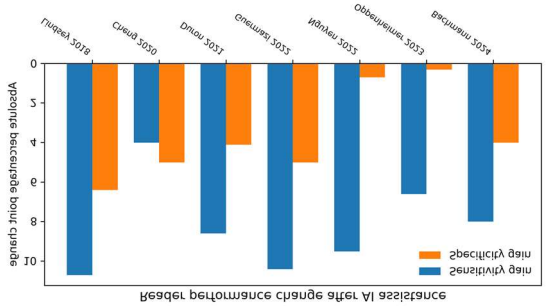

Figure 1. Absolute percentage point changes in reader sensitivity and specificity after AI assistance. The figure indicates that AI mainly improves detection sensitivity.

Theme 3: Bias and failure modes are visible but incompletely reported

The main evidence unravels many modes of failure. In pediatrics, a single case of bacterial infection was reported as worthy of an AI application, but performance is worse depending on the type of fracture and radiographic appearance (Binh et al., 2024; Nguyen et al., 2022; Rayan et al., 2019; Zech et al., 2023). In wider emergency radiographs, commercial algorithms failed to detect numerous minor abnormalities (Koskinen et al., 2025). Weaker outcomes were observed in old fractures and irregular short bones in independent tests of fracture bone at a bone level (Bruun et al., 2026). These results are clinically significant because emergency radiographs are not a balanced test set. They consist of old trauma, growth plates, casts, hardware, osteopenia, degenerative change, and positioning issues. Not many reported performance by age, sex, ethnicity, scanner, radiographic projection, fracture chronicity, or socioeconomic group. Therefore, bias cannot be eliminated due to the high headline accuracy.

Theme 4: Workflow benefit depends on governance, interface, and accountability

There was less reporting of workflow outcomes than of accuracy. Guermazi et al. (2022) and Duron et al. (2021) proposed that it might save reading time or, at least, prevent interpretation delays, and Bachmann et al. (2024) did not report a significant time penalty. Oppenheimer et al. (2023) demonstrated that prospective integration can alter clinical interpretations and lead to the detection of more fractures. However, there is still scant evidence in the workflow: only a handful of studies have measured patient outcomes, downstream imaging, return visits, reporting delays, alert fatigue, and clinician trust. Figure 3 summarizes the thematic coding and shows that accuracy was mentioned relatively often, whereas workflow and bias were less frequently mentioned.

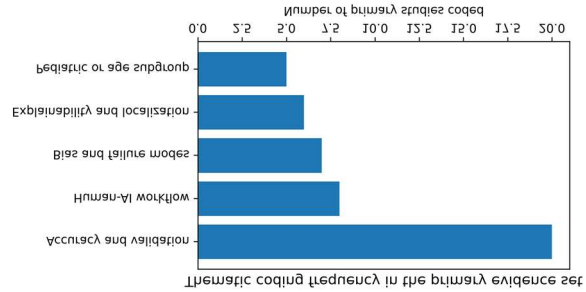

Figure 3. Thematic coding frequency across the 20 primary studies. Evidence of accuracy was abundant,

while evidence of workflow and bias was less developed.

Table 2
Thematic synthesis matrix

Theme	Supporting primary evidence	Interpretation	Critical gap
Accuracy and validation	Kim; Chung; Rayan; Krogue; Zhang; Russe	Strong performance in focused tasks supports technical feasibility.	Generalizability to broader emergency practice is uncertain.
Human-AI workflow	Lindsey; Cheng; Duron; Guermazi; Nguyen; Oppenheimer; Bachmann	AI behaves most credibly as a second reader, reducing misses.	Patient-level outcomes, alert fatigue, and clinical accountability are rarely measured.
Bias and failure modes	Binh; Zech; Koskinen; Bruun; Nguyen	Accuracy varies by age, anatomy, chronicity, and subtlety.	Subgroup audit is incomplete, so inequitable performance remains possible.
Explainability	Russe; Zech; Lindsey; commercial-reader studies	Localization may help build trust and correct misses.	Black-box classification can be accurate but difficult to verify at the point of care.

DISCUSSION

This review concurs with the secondary meta-analyses that AI detection of fractures has high average diagnostic accuracy, but it introduces an additional, more conservative interpretation. Kuo et al. (2022) and Nowroozi et al. (2024) have reported pooled sensitivities and specificities of 0.90 or greater, but the primary studies illustrate why pooled values are misleading. AI performs best when anatomy, projection and fracture types are managed. An unbalanced case mix, such as subtle abnormalities, old fractures, irregular short bones, or pediatric forms, adds to the fragility of performance (Bruun et al., 2026; Koskinen et al., 2025; Zech et al., 2023). Thus, it is not whether AI is capable of detecting fractures, but rather where, to whom, and under what supervision it can safely decrease diagnostic error.

The results also assert the human-automation theory. According to Parasuraman and Riley (1997) and Endsley (2017), automation may enhance performance but also falsifies attention and trust. The largest gains were recorded in the primary evidence, where AI served as a visible second reader rather than an independent decision-maker (Cheng et al., 2020; Duron et al., 2021; Guermazi et al., 2022). This is in line with the review results, which indicate that sensitivity tends to improve rather than specificity. That makes AI clinically applicable as a safety net for missed fractures, but not a replacement for a radiologist or orthopaedic opinion. The decreased standalone specificity indicated in certain prospective and commercial environments diminishes the probability of false alarms, redundant follow-up, or overreliance in the case of poorly regulated implementation (Husarek et al., 2024; Oppenheimer et al., 2023).

The primary literature is not very advanced compared to the methodological guidance. CLAIM and STARD 2015 promote representative populations, open reporting, external validation, and bias evaluation (Bossuyt et al., 2015; Collins et al., 2024; Mongan et al., 2020; Wolff et al., 2019). PROBAST and TRIPOD+AI promote similar measures (Bossuyt et al., 2015; Collins et al., 2024; Mongan et al., 2020; Wolff et al., 2019). However, numerous AI research projects still tend to focus on retrospective accuracy instead of prospective workflow and equality of outcomes. This is a fundamental limitation, since emergency radiography is a living service, not merely an image-classification bench test. Local validation, reader training, threshold tuning, false-positive and false-negative monitoring, and subgroup reporting must be considered minimum governance conditions, in line with general suggestions on clinical AI use

(Brady et al., 2024; Kelly et al., 2019; Yu et al., 2022).

Lastly, the bias question remains unanswered. More general literature on AI ethics cautions that algorithms can seem objective, even though they replicate concealed injustices in data, labels, or deployment contexts (Obermeyer et al., 2019; World Health Organization, 2021). In fracture identification, bias can be caused by age-related anatomy, childhood growth plates, osteoporosis, hardware effect, image quality, or poor/underperforming patterns of trauma. The primary studies included seldom looked at these issues systematically. Therefore, the potential of AI-enhanced fracture detection is not an empty threat, and the evidence base is solid for technical augmentation rather than outcome-enhancing transformation in the form of equity.

CONCLUSION

Detection of fractures on plain radiographs with the aid of AI has a promising clinical potential, particularly as a second reader, which is less prone to missing fractures and can support shortcuts or high-volume interpretation. In 20 major studies, most systems demonstrated useful accuracy, and reader studies generally indicated better sensitivity. Nonetheless, the scientific grounds are not robust enough to warrant the independent replacement of clinicians. Anatomy dependence: Accuracy is incompletely reported; subgroup bias: Subgroup bias is dependent on anatomy; workflow evidence usually stops at reading time, not patient outcomes. Only with local validation, clear accountability, prospective audit, and ongoing monitoring of subgroup performance, false negatives, false positives, and subgroup performance, should AI come to orthopaedic and radiology services. This article concludes that AI has the potential to enhance fracture diagnosis when used as a guided clinical aid, but not as a black-box replacement for experience.

REFERENCES

Adams, M., Chen, W., Holcdorf, D., McCusker, M. W., Howe, P. D. L., & Gaillard, F. (2019). Computer vs human: Deep learning versus perceptual training for the detection of neck of femur fractures. Journal of Medical Imaging and Radiation Oncology, 63(1), 27–32. https://doi.org/10.1111/1754-9485.12828

Bachmann, R., Gunes, G., Hangaard, S., Nexmann, A., Lisouski, P., Boesen, M., Lundemann, M., & Baginski, S. G. (2024). Improving traumatic fracture detection on radiographs with artificial intelligence support: A multi-reader study. BJR Open, 6(1), tzae011. https://doi.org/10.1093/bjro/tzae011

Benjamens, S., Dhunnoo, P., & Mesko, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms. NPJ Digital Medicine, 3, 118. https://doi.org/10.1038/s41746-020-00324-0

Binh, L. N., Nhu, N. T., Vy, V. P. T., Son, D. L. H., Hung, T. N. K., Bach, N., Huy, H. Q., Tuan, L. V., Le, N. Q. K., & Kang, J. H. (2024). Multi-class deep learning model for detecting pediatric distal forearm fractures based on the AO/OTA classification. Journal of Imaging Informatics in Medicine, 37(2), 725-733. https://doi.org/10.1007/s10278-024-00968-4

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., Moher, D., Rennie, D., de Vet, H. C. W., & Kressel, H. Y.

(2015). STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. BMJ, 351, h5527. https://doi.org/10.1136/bmj.h5527

Brady, A. P., Allen, B., Chong, J., Kotter, E., Kottler, N., Mongan, J., Oakden-Rayner, L., Prabhakar, A. M., & Tang, A. (2024). Developing, purchasing, implementing and monitoring AI tools in radiology: Practical considerations. Radiology: Artificial Intelligence, 6, e230513. https://doi.org/10.1148/ryai.230513

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Bruun, F. J., Muller, F. C., Nybing, J. U., Hansen, P., Gosvig, K. K., Boesen, M. P., & colleagues. (2026). Independent bone-level diagnostic accuracy study of an AI tool for detecting appendicular skeletal fractures on radiographs. European Radiology. https://doi.org/10.1007/s00330-026-12489-5

Cheng, C. T., Chen, C. C., Cheng, F. J., Chen, H. W., Su, Y. S., Yeh, C. N., Chung, I. F., & Liao, C. H. (2020). A human-algorithm integration system for hip fracture detection on plain radiography: System development and validation study. JMIR Medical Informatics, 8(11), e19416. https://doi.org/10.2196/19416

Chung, S. W., Han, S. S., Lee, J. W., Oh, K. S., Kim, N. R., Yoon, J. P., Kim, J. Y., Moon, S. H., Kwon, J., Lee, H. J., Noh, Y. M., & Kim, Y. (2018). Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthopaedica, 89(4), 468-473. https://doi.org/10.1080/17453674.2018.1453714

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., & colleagues. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378. https://doi.org/10.1136/bmj-2023-078378

Duron, L., Ducarouge, A., Gillibert, A., Laine, J., Allouche, C., Cherel, N., Zhang, Z., Nitche, N., Lacave, E., Pourchot, A., Felter, A., Lassalle, L., Regnard, N. E., & Feydy, A. (2021). 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(1), 120-129. https://doi.org/10.1148/radiol.2021203886

Endsley, M. R. (2017). From here to autonomy: Lessons learned from human-automation research. Human Factors, 59(1), 5-27. https://doi.org/10.1177/0018720816681350

Guermazi, A., Tannoury, C., Kompel, A. J., Murakami, A. M., Ducarouge, A., Gillibert, A., Li, X., Tournier, A., Lahoud, Y., Jarraya, M., Lacave, E., Rahimi, H., Pourchot, A., Parisien, R. L., Merritt, A. C., Comeau, D., Regnard, N. E., & Hayashi, D. (2022). Improving radiographic fracture recognition performance and

efficiency using artificial intelligence. Radiology, 302(3), 627-636. https://doi.org/10.1148/radiol.210937

Hallas, P., & Ellingsen, T. (2006). Errors in fracture diagnoses in the emergency department: Characteristics of patients and diurnal variation. BMC Emergency Medicine, 6, 4. https://doi.org/10.1186/1471-227X-6-4

Huhtanen, J. T. J., Nyman, M., Mohanty, M., Kuitunen, I., Sopanen, M., Stogiannos, N., & Hakkarinen, A. (2022). Automated detection of elbow fractures by deep learning. Scientific Reports, 12, 11803. https://doi.org/10.1038/s41598-022-16154-x

Husarek, J., Hess, S., Razaeian, S., Ruder, T. D., Sehmisch, S., Muller, M., & Liodakis, E. (2024). Artificial intelligence in commercial fracture detection products: A systematic review and meta-analysis of diagnostic test accuracy. Scientific Reports, 14, 23053. https://doi.org/10.1038/s41598-024-73058-8

Kalmet, P. H. S., Sanduleanu, S., Primakov, S., Wu, G., Jochems, A., Refaee, T., Ibrahim, A., Hulst, L. V., Lambin, P., & Poeze, M. (2020). Deep learning in fracture detection: A narrative review. Acta Orthopaedica, 91(3), 362-369. https://doi.org/10.1080/17453674.2019.1711323

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17, 195. https://doi.org/10.1186/s12916-019-1426-2

Kim, D. H., & MacKinnon, T. (2018). Artificial intelligence in fracture detection: Transfer learning from deep convolutional neural networks. Clinical Radiology, 73(5), 439-445. https://doi.org/10.1016/j.crad.2017.11.015

Kitamura, G., Chung, C. Y., & Moore, B. E. (2019). Ankle fracture detection utilizing a convolutional neural network ensemble implemented with a small sample, de novo training, and multiview incorporation. Journal of Digital Imaging, 32, 672-677. https://doi.org/10.1007/s10278-018-0167-7

Koskinen, S. K., Pudas, T. K., Kajander, S., Niemi, P., Aronen, H. J., & Hirvonen, J. (2025). Comparative accuracy of two commercial AI algorithms for musculoskeletal trauma detection in emergency radiographs. Emergency Radiology. https://doi.org/10.1007/s10140-025-02353-2

Krogue, J. D., Cheng, K. V., Hwang, K. M., Toogood, P., Meinberg, E. G., Geiger, E. J., Zaid, M., McGill, K. C., Patel, R., Sohn, J. H., & Feeley, B. T. (2020). Automatic hip fracture identification and functional subtype classification with deep learning. Radiology: Artificial Intelligence, 2(2), e190023. https://doi.org/10.1148/ryai.2020190023

Kuo, R. Y. L., Harrison, C., Curran, T. A., Jones, B., Freethy, A., Cussons, D., Stewart, M., Collins, G. S., & Furniss, D. (2022). Artificial intelligence in fracture detection: A systematic review and meta-analysis. Radiology, 304(1), 50-62.

https://doi.org/10.1148/radiol.211785

Lindsey, R., Daluiski, A., Chopra, S., Lachapelle, A., Mozer, M., Sicular, S., Hanel, D., Gardner, M., Gupta, A., Hotchkiss, R., & Potter, H. (2018). Deep neural network improves fracture detection by clinicians. Proceedings of the National Academy of Sciences, 115(45), 11591-11596. https://doi.org/10.1073/pnas.1806905115

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J. R., Schmid, M. K., Balaskas, K., Topol, E. J., Bachmann, L. M., Keane, P. A., & Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging. The Lancet Digital Health, 1(6), e271-e297. https://doi.org/10.1016/S2589-7500(19)30123-2

London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. Hastings Center Report, 49(1), 15-21. https://doi.org/10.1002/hast.973

Mongan, J., Moy, L., & Kahn, C. E. (2020). Checklist for artificial intelligence in medical imaging. Radiology: Artificial Intelligence, 2(2), e200029. https://doi.org/10.1148/ryai.2020200029

Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., Topol, E. J., Ioannidis, J. P. A., Collins, G. S., & Maruthappu, M. (2020). Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. BMJ, 368, m689. https://doi.org/10.1136/bmj.m689

Newman-Toker, D. E., Peterson, S. M., Badihian, S., Hassoon, A., Nassery, N., Parizadeh, D., Wilson, L. M., Jia, Y., Omron, R., Tharmarajah, S., Guerin, L., Bastani, P. B., Fracica, E. A., Kotwal, S., Robinson, K. A., Segal, J. B., & Kansagara, D. (2022). Diagnostic errors in the emergency department: A systematic review. Agency for Healthcare Research and Quality. https://www.ncbi.nlm.nih.gov/books/NBK588120/

Nguyen, T., Maarek, R., Hermann, A. L., Kammoun, A., Marchi, A., Khelifi-Touhami, M. R., Collin, M., Jaillard, A., Kompel, A. J., Hayashi, D., Guermazi, A., & Ducou Le Pointe, H. (2022). Assessment of an artificial intelligence aid for the detection of appendicular skeletal fractures in children and young adults by senior and junior radiologists. Pediatric Radiology, 52(11), 2215-2226. https://doi.org/10.1007/s00247-022-05496-3

Nowroozi, A., Salehi, M. A., Shobeiri, P., Agahi, S., Momtazmanesh, S., Kaviani, P., & Kalra, M. K. (2024). Artificial intelligence diagnostic accuracy in fracture detection from plain radiographs and comparing it with clinicians: A systematic review and meta-analysis. Clinical Radiology, 79(8), 579-588. https://doi.org/10.1016/j.crad.2024.04.009

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations.

Science, 366(6464), 447-453. https://doi.org/10.1126/science.aax2342

Olczak, J., Fahlberg, N., Maki, A., Razavian, A. S., Jilert, A., Stark, A., Skoldenberg, O., & Gordon, M. (2017). Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthopaedica, 88(6), 581-586. https://doi.org/10.1080/17453674.2017.1344459

Oppenheimer, J., Luken, S., Hamm, B., & Niehues, S. M. (2023). A prospective approach to integration of AI fracture detection software in radiographs into clinical workflow. Life, 13(1), 223. https://doi.org/10.3390/life13010223

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hrobjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230-253. https://doi.org/10.1518/001872097778543886

Rayan, J. C., Reddy, N., Kan, J. H., Zhang, W., & Annapragada, A. (2019). Binomial classification of pediatric elbow fractures using a deep learning multiview approach emulating radiologist decision making. Radiology: Artificial Intelligence, 1(1), e180015. https://doi.org/10.1148/ryai.2019180015

Russe, M. F., Rebmann, P., Tran, P. H., Kellner, E., Reisert, M., Bamberg, F., Kotter, E., & Kim, S. (2024). AI-based X-ray fracture analysis of the distal radius: Accuracy between representative classification, detection and segmentation deep learning models for clinical practice. BMJ Open, 14(1), e076954. https://doi.org/10.1136/bmjopen-2023-076954

van Leeuwen, K. G., Schalekamp, S., Rutten, M., van Ginneken, B., & de Rooij, M. (2021). Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. European Radiology, 31, 3797-3804. https://doi.org/10.1007/s00330-021-07892-z

Varoquaux, G., & Cheplygina, V. (2022). Machine learning for medical imaging: Methodological failures and recommendations for the future. NPJ Digital Medicine, 5, 48. https://doi.org/10.1038/s41746-022-00592-y

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Annals of Internal Medicine, 170(1), 51-58. https://doi.org/10.7326/M18-1376

World Health Organization. (2021). Ethics and governance of artificial intelligence for health. World Health Organization.

https://www.who.int/publications/i/item/9789240029200

Yu, A. C., Mohajer, B., & Eng, J. (2022). External validation of deep learning algorithms for radiologic diagnosis: A systematic review. Radiology: Artificial Intelligence, 4(3), e210064. https://doi.org/10.1148/ryai.210064

Zech, J. R., Carotenuto, G., Igbinoba, Z., Tran, C. V., Insley, E., Baccarella, A., & Wong, T. T. (2023). Detecting pediatric wrist fractures using deep-learning-based object detection. Pediatric Radiology, 53(6), 1125-1134. https://doi.org/10.1007/s00247-023-05588-8

Zhang, J., He, X., Li, W., Wang, C., Wang, J., & colleagues. (2023). Deep learning assisted diagnosis system: Improving the diagnostic accuracy of distal radius fractures. Frontiers in Medicine, 10, 1224489. https://doi.org/10.3389/fmed.2023.1224489