OPEN ACCESS

Journal of Orthopedic Research and Therapy

GAVIN PUBLISHERS

Ahmed G and Rice J. J Orthop Res Ther 5: 1183.

Research Article

DOI: 10.29011/2575-8241.001182

Is Artificial Intelligence Improving the Quality Of Detection in Orthopaedic Imaging? A Systematic Review

G. Ahmed, J. Rice*

Department of Orthopaedics and Trauma, University Hospital Kerry, Tralee, Ireland

***Corresponding author:** John Rice, Department of Orthopaedics and Trauma, University Hospital Kerry, Tralee, Ireland

Citation: Ahmed G and Rice J (2020) Is Artificial Intelligence Improving the Quality Of Detection in Orthopaedic Imaging? A Systematic Review. J Orthop Res Ther 5: 1183. DOI: 10.29011/2575-8241.001183

Received Date: 19 November, 2020; **Accepted Date:** 28 November, 2020; **Published Date:** 02 December, 2020

Abstract

Background: Artificial Intelligence (AI) has heralded huge changes in many facets of our lives. If we were to compare the technological improvements in the automotive industry over the past decades, there has been exponential growth just in the past several years with the development of electric cars, guidance systems, and driverless vehicles. In a similar scale, it is expected that AI will have implications on the future of the practice of orthopaedics. However, there is no easy translation of technology from industrial standards to clinical practice. The most recent example being the attempt to transfer the metal bearing concept from the automotive industry to total hip replacements. [1] our study is a systematic review of the current literature aimed to review the diagnostic studies that have explored the use of AI in areas of orthopaedic imaging. We aimed to look at the benefit of the application of AI in analysing orthopaedic imaging to assess its efficiency in terms of quality of detection in orthopaedic imaging.

Methods: Following a database search for all relevant up to date eligible articles. We carried out a systematic review in accordance with the PRISMA [2] model using PubMed, PubMed Central, Embase and UpToDate databases from start until August 2020 using the terms "Artificial Intelligence", "Orthopaedics", "Fractures", "Deep Learning", "and Imaging". The accuracy range and confidence intervals of the diagnostic studies assessed were recorded. A quality assessment was carried out using the BMJ Diagnostic test studies: assessment and critical appraisal checklist [3].

Results: 1191 records were identified, following the screening process using the PRISMA model a final 14 studies were included in a qualitative synthesis. Given the heterogeneity of the studies included, there was variation between the results of different studies. A total of ten studies applied AI models to detect fractures in plain radiographs of various body parts (X-Ray) with accuracies ranging from 76.9%-99%, 95% Confidence Intervals ranging from 74.2-100%. One study applied an AI model to detect osteoarthritis in hips with an accuracy of 90.2%. Two studies applied AI models to Computed Tomography (CT) to detect fractures in the spine with reported accuracy ranging from 89%-98%. Two further studies applied AI models to Magnetic Resonance Imaging (MRI) to diagnose abnormalities in knee and lumbar spine images, one reported an accuracy of 95.6%, the other reported 95% Confidence Intervals ranging from 78%-99.3%. 10 out of the 14 studies reviewed compared the performances of the AI models to standard references (radiologists, orthopaedic surgeons, clinicians) with accuracy of the standard reference ranging from 77%-99.3%, 95% Confidence Interval range from 76.2%-100%.

Conclusions: Overall, various AI Models applied in diagnostic studies in orthopaedic imaging achieve comparable results to standard references in detecting specific pathologies, mostly fractures, within the limited settings provided in the studies.

Introduction

The term artificial intelligence was introduced in the 1950's [4] where its prospect was explored with great enthusiasm. Since then, the advancements in computational powers and the wide availability of data seems to be turning the initial dream into a current reality in many areas. The definition itself has evolved from the ability of machines to learn without needing to be programmed[4] to encompass a larger concept of machines to be able to think humanly, act humanly, think rationally and act rationally [5]. AI has incorporated itself into many facets in our everyday lives. In this era of Big Data with millions of entries, the sheer quantity makes it difficult for a human or indeed a team of humans to come to meaningful conclusions. There is a lot of enthusiasm (early phase of the Hype Cycle [6]) on the application of AI in healthcare [7]. This has attracted billions in investments and appears to have steered the outlook of current research to further explore and unlock its potential [8]. The field of orthopaedics and trauma relies on objective analysis in the process of input and output. The ability to incorporate technological advancements into practice is part of that process. A tradition that can be dated back to the late Sir Robert Jones, when he introduced the use of the newly developed X-Ray machine for the first time to retrieve a bullet from a man's wrist [9] and later on applied it to routine practice. In the current era, AI is the disruptive technology at hand, and we hope to evaluate its potential value in orthopaedic imaging.

The aim of this study is to examine the current status of published literature on the application of AI in the field of orthopaedic imaging. We pooled data from online publications for diagnostic studies and carried out an objective analysis.

Methods

A systematic review in accordance with the PRISMA model (Figure 1) was conducted using PubMed, PubMed Central, Embase, and UpToDate databases from start until August 2020. The following Keywords were used Artificial Intelligence, Orthopaedics, Fractures, Deep Learning, Trauma, Imaging. The two authors independently screened the titles and abstracts of the records identified using arranged upon measures for inclusion. The inclusion criteria were that articles that used an AI model to test or detect or analyze orthopaedic imaging was to be used. Excluded studies and articles were those not related to orthopaedics, non-diagnostic studies, reviews, conference abstracts, protocol studies, editorials, commentaries, and non-English articles. An additional article was found outside of the database search when one of the authors was exploring the internet for the impact of AI in healthcare. The search resulted in 1191 records identified plus the additional record mentioned. After duplicates were removed, we screened titles and abstracts of 1009 articles, of which 985 did not meet the inclusion criteria and were excluded. That yielded 24 articles which were eligible for full text screening of which 9 were further excluded (3=reviews, 1=duplicate, 1=non-diagnostic, 2=commentary, 2=editorials). The final 15 studies were included in a qualitative synthesis.

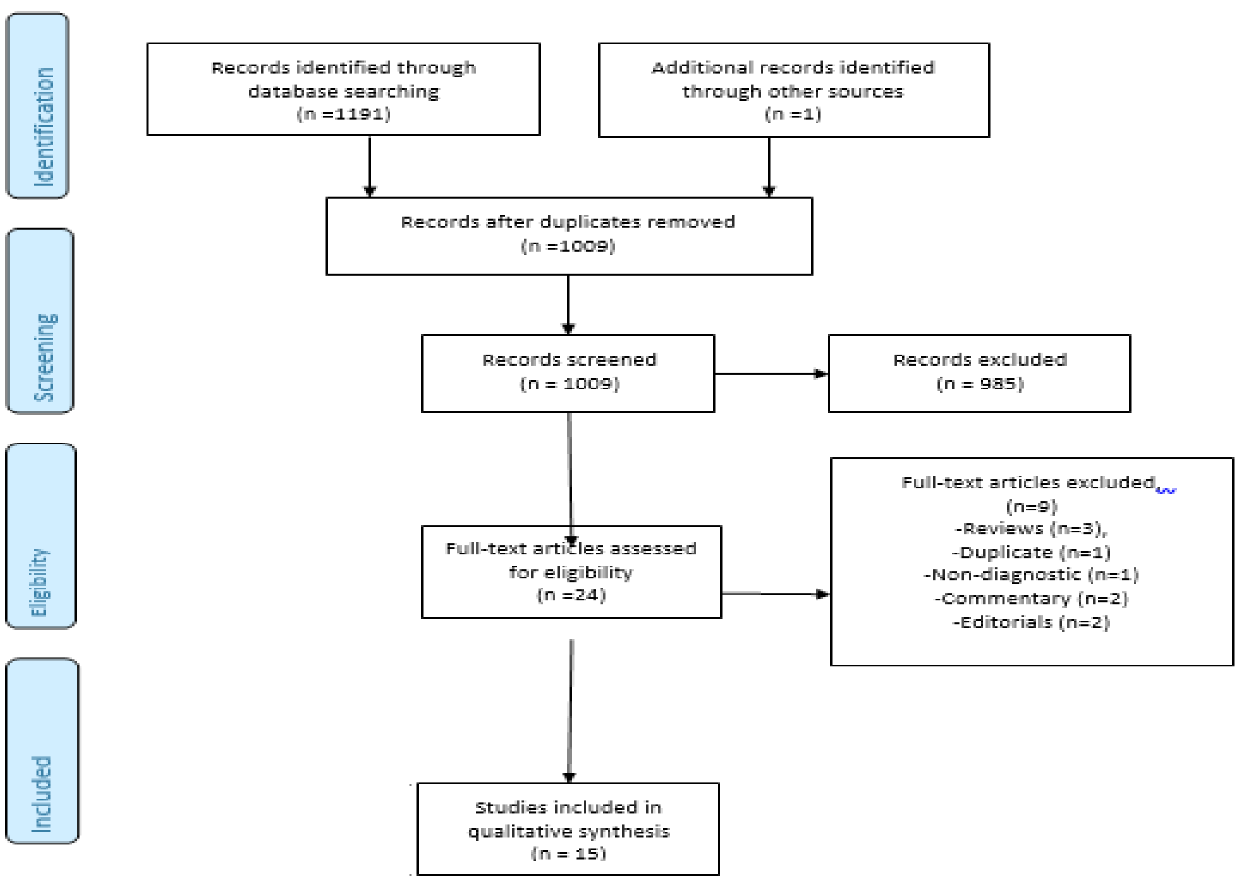

Figure 1: PRISMA Flow Diagram.

The two authors assessed and appraised the quality of the final included studies. This was conducted separately. We applied the principles of quality appraisal as recommended from the BMJ Diagnostic test studies: assessment and critical appraisal tool [3].

The data extraction was performed using a standard data extraction (Table 1). The author, year and country of each study was recorded. The specific diagnostic aim, image modality, body part imaged, and sample size of each study was recorded. For the AI models used in each study, we extracted the type of AI model used, the ground truth labelling, the comparison group, and the results of the AI model performance in terms of accuracy and confidence intervals of each study. The primary outcome measure of the studies was to establish how well an AI model performs in detecting relevant specifics pathologies in each image modality.

Data Author, Year, Country	AIM	Image Modality	Sample Size	Body Part	Ground Truth Labelling	AI Version used	Comparison Group	AI Model Results (Accuracy/ 95% CI)	Decision reasoning/ validation
Olczak et al, 2017, Sweden	Fracture detection	X-Ray	256,458	Wrist, Hand, Ankle	Radiology Report	CNN VGG-16, VGG-19, VGG S, BVLC, Network-in Network	Orthopaedic Surgeons	83%/ 79-87 %	
Chung et al, 2018, S Korea	Fracture detection	X -Ray	1,891	Proximal Humerus	Shoulder Specialists +Radiologist	CNN Microsoft ResNet-152	Gen. Physicians, Gen. Orthopaedics, Specialized Orthopaedics	96%/ 94-97%	
Chi-Tung Cheng et al, 2019, Taiwan	Fracture detection	X-Ray	3,605	Hip/Pelvis	Trauma Surgeon	DCNN DenseNet-121	Primary Physicians, Emergency Physicians, Orthopaedic Surgeons, Radiologists	91%/ 84-96%	Grad-CAM
Urakawa et al, 2018, Japan	Fracture detection	X-Ray	3,346	Proximal Femur	Orthopaedic Surgeons	CNN VGG-16	Orthopaedic Surgeons	95.5 %/ 93.1-97.6	
Daniel Pinto dos Santos et al, Germany 2019	Fracture detection	X-Ray	157	Ankle	Radiologist	CNN Inception-V3	N/A	76.9% / 74.2-79.6 %	

Kaifeng Gan et al, 2019, China	Fracture detection	X-Ray	2,340	Wrist	Orthopaedic Surgeons	Faster R-CNN Inception- V4	Orthopaedics, Radiologists	93%/ 90-96%	
Kemal Üreten et al, 2020, Turkey	Osteoarthritis detection	X-Ray	868	Pelvis	Physiotherapist, Rheumatologist, Radiologist	CNN VGG-16	N/A	90.2%/ N/A	
Gale at al, 2017, Australia	Fracture detection	X-RAY	53,278	Pelvis	Surgical records, AI model, Radiology reports, Radiologist	CNN DenseNet	Radiological reports	99%/ 99-100%	
Ozkaya et al, 2020, Turkey	Fracture detection	X-RAY	390	Wrist	Radiologist	CNN ResNet50	ED Physicians, Orthopaedics	N/A/ 75.3-90.6%	
Yoshi Sato et al, 2020, Japan	Fracture detection	X-RAY	10,484	Pelvis	Orthopaedic Surgeons	CNN EfficientNet-B4	N/A	96.1% / 94.9-97.3%	Grad-CAM
Adams et al, 2018, Australia	Fracture detection	X-RAY	805	Neck of Femur	Surgical records	DCNN AlexNet, GoogLeNet	Medically Naïve Individuals, Board certified Radiologists	88.1%, 94.4%/ 86-97%, 88-98%	
Al-Helo et al, 2012, Jordan	Fracture detection	CT	50	Lumbar Spine	N/A	K-means, Neural Networks	N/A	98%,93.2%/ N/A	
Tomita et al, 2018, USA	Osteoporotic Fracture detection	CT	1432	Spine	Reports	CNN ResNet34	Radiologist Report	89%/ N/A	

Bien et al, 2018, Croatia	Diagnosis of Knee abnormalities	MRI	1,370	Knee	Radiologists	CNN MRNet	Radiologists, Orthopaedics	85%/ 77.5%-90.3%	
Jamaludin et al, 2017, United Kingdom	Automation of radiological features of lumbar spine	MRI	12,018	Lumbar Spine	Radiologist	CNN	Radiologist	95.6%/ N/A	

Table 1: The data extraction.

Ten studies demonstrated the use of AI models in detecting fractures on anteroposterior (AP) views of plain radiographs of wrists, ankles, proximal humerus, pelvis/hip, proximal femur and neck of femur [10-20]. All studies opted to use a supervised deep learning Convolutional Neural Network (CNN) model of AI, which consists of an algorithm which is run across multiple layers of neural networks aimed to mimic the way the human brain works [7]. The algorithm is supervised meaning the input and output are both known to the model, it is then trained with various combination of inputs until it can conjure the desired output. Following the training process comes a validation test and finally the algorithm is tested on a completely different set of input variables. Two of the ten studies used similar CNN, (VGG 16) [10,13]. Eight of the ten studies compared the performance of the AI model to that of a standard reference (radiologist, reports, orthopaedic surgeon, clinician) [10-15,17-19]. Further two studies used a gradient-weighted class activation mapping (Grad-CAM) to confirm the validity of the AI models applied [12,20]. One article conducted a reverse study to evaluate the accuracy of perceptual training in medically-naïve individuals to that of Deep Convolutional Neural Networks (DCNN) for detecting neck of femur fractures [19].

A single study applied an AI model, CNN VGG-16 to detect hip osteoarthritis on pelvis AP X-Rays [16]. The performance of the model was not compared to a standard reference. Two studies applied AI models on CT's of the spine to detect fractures [21,22]. One used two types of machine learning algorithms [22], a supervised neural network and an unsupervised (k-means) algorithm. The other used a CNN (ResNet34) to detect osteoporotic fractures. The final two studies [23,24] applied CNNs to knee (MRNet) and lumbar spine MRIs, respectively.

Quality Appraisal

The study aim was clear in all fifteen included studies. The population sample size represented by the number of images used to train, validate, and test the AI model was available in all studies. The methodology and testing specifications was described in all included studies. Standard references in the form of comparison groups ranging from clinicians, radiologists to orthopaedic surgeons of varying experience levels was described in eleven studies (73.3%), leaving four (26%) studies not suitable to be considered as valid diagnostic studies. Thirteen studies (86.6%) validated their AI models by testing it on a second independent group, while two (13.3%) did not. This allows us to reserve judgement with regards to accuracy of those results. There was clear blinding in four (26.6%) of studies to comparative groups, leaving eleven (73.4%) with no clear blinding methods mentioned. This leads to the accuracy of interpretation open to subjective bias. The accuracy of the tests was shown in thirteen (86.6%) studies, leaving two studies (13.3%) with no accuracies reported. 95% Confidence Intervals were reported in eleven (73.4%) studies and four were without confidence intervals (26.6%), hence the findings in those four studies cannot be considered generalizable. Sensitivities and Specificities were reported in 10 studies (66.6%), while five did not mention it (33.3%). Predictive values were present in six studies (40%) rendering nine (60%) not reporting predictive values.

An additional validation checklist item was included using the term decision reasoning, this was to examine which studies attempted to "uncover" the decision-making process of the AI model applied. The decision reasoning was found to be reported in two (13.3%) studies, rendering thirteen (86.6%) with no method of examining the process of decision-making within the model.

The quality appraisal checklist is shown in Figure 2.

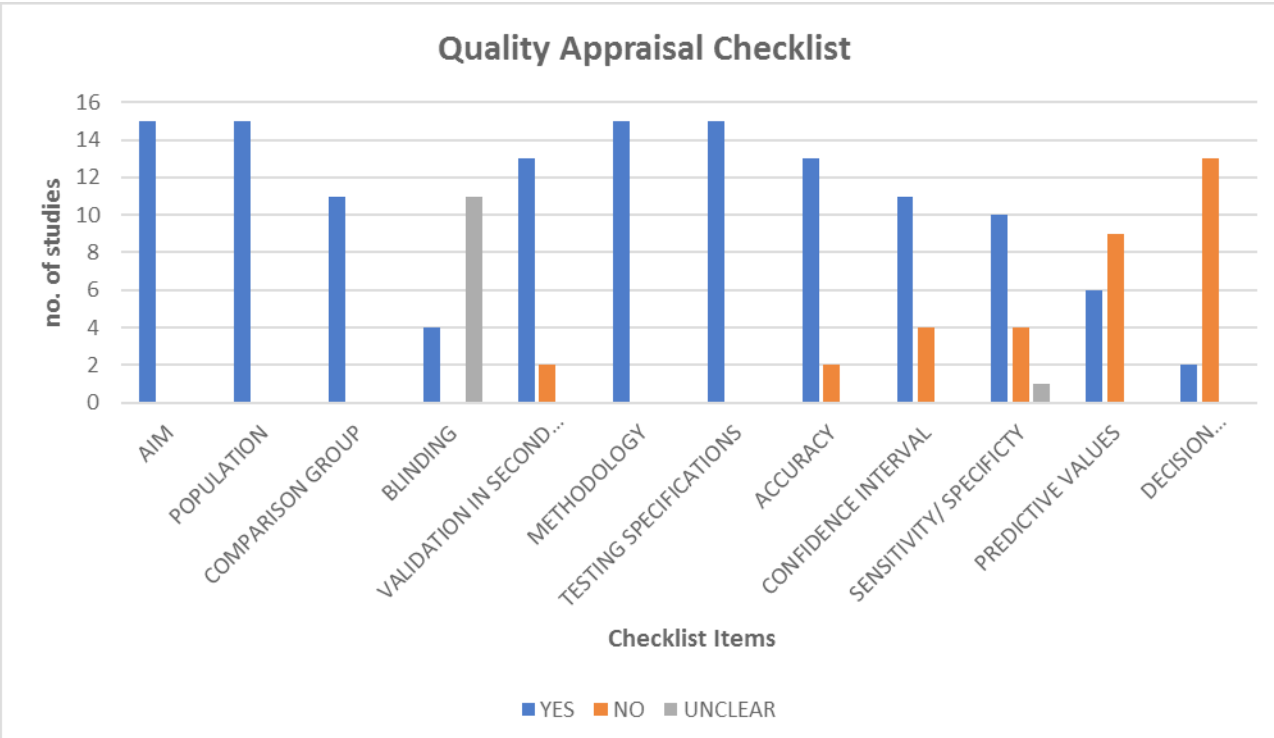

Figure 2: Quality Appraisal Checklist.

Results

A total of ten studies [10-15,17-19] applied AI models to detect fractures in X-Rays of various body parts, with accuracies ranging from 76.9%-99%, 95% Confidence Intervals ranging from 74.2-100%. Chung et al [11] also showed a CNN with 0.99/0.97 sensitivity/specificity and 0.97 Youden index for detecting proximal humerus fractures. In addition their model also showed a 0.88/0.83-0.97/0.94 sensitivity/specificity and 0.71-0.90 Youden index for classifying fracture type. Chi-Tung Cheng et al [12] and Yochi Sato et al [20] are the only two studies to assign the visualisation algorithm Grad-CAM to confirm validity of the AI model used to detect hip fractures. They achieved an accuracy of 91% and 96.1%, respectively. Sensitivity of 98% and 95.2% respectively. Specificity of 84% and 96.9 %, respectively. The Grad-CAM had an accuracy of 95.9% and 96.1%, respectively. Chi-Tung Cheng et al had a false negative rate of 2%. Yochi Sato et al had a F-value of 0.961. One study [16] applied an AI model to detect osteoarthritis in hips with an accuracy of 90.2%, sensitivity 97.6%, specificity 83.0%, and precision of 84.7%. An evaluation of scaphoid fractures [18] yielded a 76% sensitivity, 92% specificity, 0.680 Youden index and 0.826 F score value.

One article [19] conducted a reverse study to evaluate the accuracy of perceptual training in medically-naïve individuals to that of deep convolutional neural networks (DCNN) for detecting neck of femur fractures. The pretrained DCNNs, AlexNet and GoogleNet, showed accuracies of 88.1% and 94.4%, respectively. Accuracy for perceptual training for medically-naïve individuals was at 90.5%. Two studies [21,22] applied AI models to CT to detect fractures in the spine with reported accuracy ranging from 89%-98%. Two further studies [23,24] applied AI models to MRI to diagnose abnormalities, anterior cruciate ligament (ACL) tears, and meniscal teras in the knee. Results showed an accuracy of 85%, sensitivity of 0.879 and a specificity of 0.71. Jamuludin et al [24] developed a model to automate and grade lumbar spine (degenerative changes) images, they reported an accuracy of 95.6%. Eleven [10-17,19,21,23,25] out of the fifteen studies reviewed compared the performances of the AI models to standard references (radiologists, orthopaedic surgeons, clinicians) with accuracy of the standard reference ranging from 77%-99.3%, 95% Confidence Interval range from 76.2%-100%.

Discussion

There is an expectation that AI/Machine Learning is going to be a new departure in providing orthopaedic/radiological services. The use of AI models in assessing decision in image interpretation in orthopaedic surgery currently exists and has been examined in a scientific way. Most studies included in this review have compared the outcomes of performances of trained deep neural network AI models to trained clinicians as the current clinical standard reference. This showed varying degrees of success. Overall results were comparable in terms of accuracy in detection of specific pathologies mostly fractures, within the limited settings applied to the diagnostic test. Taking into consideration that the current overall day-to-day radiologist error rate has an estimated average of 3-5% [26], the current review demonstrates AI models tested within limited settings are comparable and in some instances more accurate than current standard references. The single detection ability demonstrated in nearly all the studies is the major limitation as machines cannot be expected to appreciate unanticipated findings such as the incidental finding of an asymptomatic tumour on an x-ray carried out to assess for a fracture. The ability of AI models to detect relevant incidental concomitant pathologies/ findings on given set images would require specific training / programming of the machine to detect each possible eventuality. In that context, AI models do not outperform humans as would be the requirement of a human during routine clinical practice to be able to identify incidental findings or concomitant requirements. That is not to say that such algorithmic models do not exist, but such ability has not been exposed while conducting this review. For AI models to be able to be applied at an industrial scale in the health service, this limitation has to be addressed and would appear to require an inordinate amount of work.

Another concept to be examined is the ability to understand how and why an AI model comes to decision making. This is currently a mystery of machine learning referred to as the "black box" [25] of the algorithmic decision process. In that sense we have no reasonable idea of understanding how the AI is analysing the image, what it is basing its predictions upon, and how it is arriving at the final output. Such decisions cannot be trusted without proper validation of the process. AI models used in a diagnostic setting should be scrutinized to uncover the "black box" of algorithmic decision processes. In our series only two studies [12,20] attempted to validate their models by including a gradient-weighted class activation mapping software (Grad-Cam). This allowed for visual validation for where the AI model is looking, verifying that it is indeed looking at the correct patterns in the image and activating around those patterns. At this stage training humans specifically might still be rewarding. The study conducted by Adams et al [19] where medically-naïve individuals (medical students) through perceptual training achieved high accuracy rates in detecting neck of femur fractures, reminds us of a fundamental concept - that the human mind is still very capable of improving quality detection with respect to imaging when provided with appropriate training. Machine learning may represent as the Holy Grail in crossing over from individual variability and subjectivity towards achieving objectivity in assessing orthopaedic imaging. However experience in the field is still quite limited. The current literature is mainly composed of retrospective diagnostic tools. Further research with prospective studies and randomized controlled trails should be conducted to deliver higher quality evidence that AI can be considered as an independent diagnostic tool. Review of the current literature would suggest that machine learning systems in diagnostic imaging in orthopaedics are not yet at a stage where they can exist as an independent clinical tool.

References

1. Cohen D (2011) Out of joint: The story of the ASR. British Medical Journal 342: 1-7.
2. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009) Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6.
3. British Medical Journal (2020) Diagnostic test studies: assessment and critical appraisal.
4. Amisha Malik P, Pathania M, Rathaur VK (2019) Overview of artificial intelligence in medicine. Journal of family medicine and primary care 8: 2328-2331.
5. Russell S and Bohannon J (2015) Artificial intelligence. Fears of an AI pioneer. Science 349.
6. Car J, Sheikh A, Wicks P, Williams MS (2019) Beyond the hype of big data and artificial intelligence: building foundations for knowledge and wisdom. BMC Medicine 143.
7. Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25: 44-56.
8. Johnson K (2020) CB Insights: AI startup funding hit new high of $26.6 billion in 2019.
9. Tham W, Sng S, Lum YM, Chee YH (2014) A Look Back in Time: Sir Robert Jones, 'Father of Modern Orthopaedics'. Malaysian orthopaedic journal 8: 37-41.
10. Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, et al. (2017) Artificial intelligence for analyzing orthopaedic trauma radiographs. Acta Orthopaedica 88: 581-586.
11. Chung SW, Han SS, Lee JW, Oh KS, Kim NR, et al. (2018) Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthopaedica 89: 468-473.
12. Cheng CT, Ho TY, Lee TY, Chang CC, Chou CC, et al. (2019) Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. European Radiology 29: 5469-5477.
13. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, et al. (2019) Detecting intertrochanteric hip fractures with orthopaedist-level accuracy using a deep convolutional neural network. Skeletal Radiology, 48: 239-244.
14. Pinto dos Santos D, Brodehl S, Baeßler B, Arnhold G, Dratsch T, et al. (2019) Structured report data can be used to develop deep learning algorithms: a proof of concept in ankle radiographs. Insights into Imaging, 10: 1-8.
15. Gan K, Xu D, Lin Y, Shen Y, Zhang T, et al. (2019) Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments. Acta Orthopaedica 90: 394-400.

16. Üreten K, Arslan T, Gültekin KE, Demir AN, özer HF, et al. (2020) Detection of hip osteoarthritis by using plain pelvic radiographs with deep learning methods. Skeletal Radiology 49: 1369-1374.

17. Gale W, Carneiro G, Oakden-Rayner L, Palmer LJ, Bradley AP, et al. (2017) Detecting hip fractures with radiologist-level performance using deep neural networks. ArXiv, 1-6.

18. Ozkaya E, Topal FE, Bulut T, Gursoy M, Ozuysal M, et al. (2020) Evaluation of an artificial intelligence system for diagnosing scaphoid fracture on direct radiography. European Journal of Trauma and Emergency Surgery 2020: 1-8.

19. Adams M, Chen W, Holcdorf D, McCusker MW, Howe PDL, et al. (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: 27-32.

20. Sato Y, Takegami Y, Asamoto T, Ono Y, Tsugeno H, et al. (2020) A Computer-Aided Diagnosis System Using Artificial Intelligence for Hip Fractures -Multi-Institutional Joint Development Research. Physics Medical Physics 2020: 1-9.

21. Tomita N, Cheung YY, Hassanpour S (2018) Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. Computers in Biology and Medicine 2018: 8-15.

22. Al-Helo S, Alomari RS, Ghosh S, Chaudhary V, Dhillon G, et al. (2013) Compression fracture diagnosis in lumbar: a clinical CAD system. International Journal of Computer Assisted Radiology and Surgery 8: 461-469.

23. Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, et al. (2018) Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. PLoS Medicine 15: 1-10.

24. Jamaludin A, Lootus M, Kadir T, Zisserman A, Urban J, et al. (2017) Automation of reading of radiological features from magnetic resonance images (MRIs) of the lumbar spine without human intervention is comparable with an expert radiologist. European Spine Journal 26: 1374-1383.

25. Bathee Y (2018) The Artificial Intelligence Black Box and the Failure of Intent and Causation. Harvard Journal of Law and Technology 31: 890-938.

26. Berlin L (2007) Radiologic errors and malpractice: a blurry distinction. AJR Am J Roentgenol 189: 517-522.