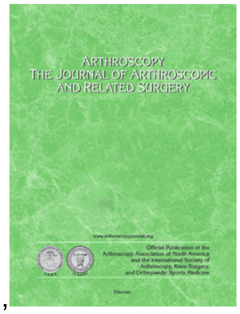
Journal Pre-proof



Diagnostic Performance of Artificial Intelligence for Detection of Anterior Cruciate Ligament and Meniscus Tears: A Systematic Review

Kyle N. Kunze, M.D., David M. Rossi, B.S., Gregory M. White, M.D., Aditya V. Karhade, M.D., M.B.A., Jie Deng, Ph.D., Brady T. Williams, M.D., Jorge Chahla, M.D., Ph.D.

PII: S0749-8063(20)30744-1

DOI: https://doi.org/10.1016/j.arthro.2020.09.012

Reference: YJARS 57125

To appear in: *Arthroscopy: The Journal of Arthroscopic and Related Surgery*

Received Date: 5 April 2020

Revised Date: 2 September 2020

Accepted Date: 9 September 2020

Diagnostic Performance of Artificial Intelligence for Detection of Anterior Cruciate Ligament and Meniscus Tears: A Systematic Review

Kyle N. Kunze, M.D.[1]

David M. Rossi, B.S.[2]

Gregory M. White, M.D.[3]

Aditya V. Karhade, M.D., M.B.A.[4]

Jie Deng, Ph.D.[3]

Brady T. Williams, M.D.[2]

Jorge Chahla, M.D., Ph.D.[2]

1. Department of Orthopaedic Surgery, Hospital for Special Surgery, New York, NY, USA
2. Department of Orthopaedic Surgery, Rush University Medical Center, Chicago, IL, USA
3. Department of Diagnostic Radiology and Nuclear Medicine, Rush University Medical Center, Chicago, IL, USA
4. Department of Orthopedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Correspondence:
Jorge Chahla, M.D., Ph.D.
Rush University Medical Center
1611 W. Harrison St. Suite 300
Chicago, IL, USA 60612
jachahla@msn.com

Diagnostic Performance of Artificial Intelligence for Detection of Anterior Cruciate Ligament and Meniscus Tears: A Systematic Review

ABSTRACT

Purpose: To (1) determine the diagnostic efficacy of artificial intelligence (AI) methods for detecting anterior cruciate ligament (ACL) and meniscus tears and to (2) compare the efficacy to human clinical experts.

Methods: PubMed, OVID/Medline, and Cochrane libraries were queried in November 2019 for research articles pertaining to AI utilization for detection of ACL and meniscus tears. Information regarding AI model, prediction accuracy/area under the curve (AUC), sample sizes of testing/training sets, and imaging modalities were recorded.

Results: A total of 11 AI studies were identified: 5 investigated ACL tears, 5 investigated meniscal tears, and 1 investigated both. The AUC of AI models for detecting ACL tears ranged from 0.895-0.980, and the prediction accuracy ranged from 86.7%-100%. Of these studies, three compared AI models to clinical experts. Two found no significant differences in diagnostic capability, while one found that radiologists had a significantly higher sensitivity for detecting ACL tears (p=0.002) and statistically similar specificity and accuracy. Of the 5 studies investigating the meniscus, the AUC for AI models ranged from 0.847-0.910 and prediction accuracy ranged from 75.0%-90.0%. Of these studies, 2 compared AI models to clinical experts. One found no significant differences in diagnostic accuracy, while one found that the AI model had a significantly lower specificity (p=0.003) and accuracy (p=0.015) than radiologists. Two studies reported that the addition of AI models significantly increased the diagnostic performance of clinicians compared to their efforts without these models.

Conclusion: AI prediction capabilities were excellent and may enhance the diagnosis of ACL and meniscal pathology; however, AI did not outperform clinical experts.

Clinical relevance: AI models promise to improve diagnosing certain pathologies as well as or better than human experts, are excellent for detecting ACL and meniscus tears, and may

enhance the diagnostic capabilities of human experts; however, when compared to these experts, may not offer any significant advantage.

INTRODUCTION

The development and application of deep learning (DL) and machine learning (ML) algorithms to generate prediction models from large datasets is an increasingly utilized statistical tool which relies on pattern recognition and constrained feature selection. By "constraining" feature selection, the algorithms limit the number of variables ultimately chosen and used in subsequent analyses by selecting only those with the greatest predictive value from an initial large pool of potential variables. A clinically relevant application of these artificial intelligence (AI) methods pertains to its ability to diagnose injury and disease on medical imaging studies. AI models learn to recognize disease patterns through repetition and learning, and it is thought that such features may confer the ability to more quickly and accurately identify disease.

Machine learning describes statistical processes that exhibit the "learning" associated with human intelligence and leverage this experiential learning to improve and refine programmed algorithms to predict and outcome.[1] The algorithms are applied to a dataset of interest, and self-train based on patterns in the dataset. Once trained, the algorithms can make specific decisions when presented with data that it has not seen before. Each machine learning algorithm makes decisions based on different sets of rules that are out-of-scope of the current study, but allow them to come to decisions in different ways. Algorithms can be modified to optimize their prediction capabilities, and ultimately the predictions made by the algorithm are compared against the true outcome present in the data set to determine how accurate predictions are.[2] This approach has become increasingly popular given the ability of these algorithms to optimize prediction accuracy, whereas traditional statistics may sacrifice accuracy at the cost of favoring interpretability. Machine learning has become of recent interest in orthopaedics given these potential benefits, as evidenced by the recent increase in literature which has applied this methodology.[3-6]

AI technology has been successfully applied in various clinical scenarios. Detection of diabetic retinopathy through analyzing retinal fundus photographs[7, 8] and skin cancer through constructing deep neural networks based on imaging and disease labels[9] have efficacy comparable to, or better than, human experts. Within the field of orthopedic spine and oncologic surgery specifically, AI algorithms are gaining popularity by aiding decision-making and can be used in clinical settings.[5, 10, 11] However, the performance of these models compared to clinical experts in the field remains poorly understood. A recent systematic review by Langerhuizen et al.[12] found that AI algorithms had excellent performance for fracture detection in the orthopedic trauma literature and outperformed human examiners for detecting and classifying hip and proximal humerus fractures. Although it appears that AI methods may may confer diagnostic benefits in other realms of orthopedic surgery, their performance and clinical utility in sports medicine remains poorly defined.

Imaging-based detection of sports medicine injuries of the knee, specifically the use of magnetic resonance imaging (MRI) for anterior cruciate ligament (ACL) and meniscus tears, is the current gold standard for diagnosis. However, the diagnostic accuracy of MRI may be decreased in several circumstances: (1) observer inexperience and bias, (2) small partial or incomplete tears, (3) imaging artifacts, (4) incomplete MRI study, and (5) presence of concomitant injuries. Application of AI methods may address these shortcomings by facilitating clinical decision-making and improving patient management. As such, the purpose of the current study was to (1) determine the diagnostic efficacy of AI methods for detecting ACL and meniscus tears and to (2) compare the efficacy to human clinical experts. The authors hypothesized that AI method performance would be excellent for detection of ACL and meniscus tears and could outperform human examiners.

METHODS

Identification and Selection of Articles

A systematic search in accordance with the 2009 Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) statement[13] was conducted using PubMed, OVID/Medline, and Cochrane libraries. The timeframe for the search was the conception of each online database until November 8, 2019. The following Boolean search syntax was used to conduct the search: (orthopedics OR orthopedic procedures OR ligament OR tear* OR (ligament* AND tear* AND orthop*)) AND (artificial intelligence OR neural network* or deep learning OR machine learning OR machine intelligence) AND (predict* OR predictive value of test OR score OR scores OR scoring system OR scoring systems OR observ* OR observer variation OR detect* or evaluat* OR analy* OR assess* OR measure*). The protocol for the current systematic review was registered on PROSPERO prior to collection and analysis of the data (ID: *blinded for review*).

Articles populated from the above search met inclusion criteria if (1) the study methods and analyses pertained to development or utilization of artificial intelligence or machine learning for detecting or classifying the presence of an ACL or meniscus tear, and (2) was published in the English language. Studies were excluded if (1) data was only published in the form of an abstract, technique paper, cadaveric or animal study, or letter to the editor; or (2) pertained to robotic-assisted surgery. Two observers (blinded for reviewer) independently screened the abstracts and titles of potential articles. Full-text review was only performed during the study selection process if necessary to determine if the articles satisfied inclusion and exclusion criteria. Additionally, all references from the included studies were reviewed and reconciled to verify that no relevant articles were missing from the systematic review. A total of 1,619 records were initially identified, and a total of 11 were ultimately included in the qualitative synthesis (**Figure 1**).

Data Acquisition

All data were recorded into a custom spreadsheet using a modified information extraction table.[14] Categories for data collection for each full article included (1) article information; (2)

input features; (3) imaging plane; (4) size of training and testing samples; (5) ground truth label assignments; (6) output classes; (7) AI models used; (8) use of pretrained Convolutional Neural Networks (CNN); and (9) performance.

Assessment of Heterogeneity and Methodological Quality

A modified MINORS scoring criteria was used to assess quality as has been previously applied in systematic reviews on AI in orthopedics[12] given that studies concerning AI methods are classified as developmental rather than diagnostic. Quality appraisal focused on identification of (1) a clear study aim, (2) description of inclusion and exclusion criteria for input features (all eligible imaging examples included), (3) determination of ground truth (reference standards for AI), (4) report of distribution of data set (training, validation, and testing phases), (5) described how performance of AI model was assessed (area under the curve [AUC]/prediction), and (6) clearly described AI model used. These criteria were applied to and quantified for each study. For reference, the AUC is the quantitative output of a receiver operator curve (ROC) analysis of discrimination. ROC and discrimination analysis is a common performance analysis in diagnostic studies, which assesses the probability that the machine learning model will assign a greater predicted probability to a randomly selected positive case (true positive case, i.e., a patient who actually had an ACL or meniscus tear) relative to a randomly selected negative case (false positive case, i.e., a patient who did not have an ACL or meniscus tear). Each study could score a total of seven points, with a score of zero indicating poor methodological quality, and a score of seven indicating the highest methodological quality. Two independent observers (blinded for review) assessed all included studies. The inter-observer reliability was excellent at 0.97 (95% Confidence interval, 0.93-0.99). Any discrepancies were resolved by consensus.

Statistical Analysis

All data was qualitatively synthesized and reported in both narrative fashion in addition to table format. Extracted data was presented as means and ranges when appropriate with associated p-values given the degree of heterogeneity between studies. All studies considered a p-value

<0.05 to indicate statistical significance. All data extraction and analyses were performed in Microsoft Excel (Microsoft Corporation, Washington, USA).

RESULTS

A total of 11 studies were identified in the search.[15-25] All 11 studies were included in the qualitative data analysis. Of these 11 studies, five investigated the use of AI to detect ACL tears,[15, 17, 20, 21, 24] five investigated the use of AI to detect tears of the meniscus,[18, 19, 22, 23, 25] and one investigated both.[16]

AI Model Performance: ACL Tear Detection

All six studies that investigated the performance of AI on ACL tear detection utilized knee MRI acquired in standard imaging planes. Four (66.6%) of the studies analyzed sagittal-plane images only,[15, 20, 21, 24] one study analyzed coronal images only,[17] and one study analyzed sagittal, coronal, and axial images.[16] A total of four (66.6%) studies reported AUC data for complete ACL tear detection. The AUC for these AI models ranged from 0.895-0.980 (**Table 1**). Additionally, four studies report AI model prediction accuracy for specifically detecting complete ACL tear (range 86.7%-100%).

Štadjuhar et al.[24] utilized two different feature extraction techniques: Histogram of Oriented Gradient (HOG) and Generalized Search Tree (GIST). These feature extraction techniques were subsequently paired with two commonly used machine learning models: Support Vector Machine (SVM) and Random Forest (RF). They found that their best performing machine learning model that combined HOG with linear-kernel SVM (HOG+lin-SVM) performed the best, producing an AUC of 0.894 for differentiating between an injured ACL and healthy ACL and an AUC of 0.943 for detecting completely ruptured ACL cases only.

Abdullah et al.[15] described a diagnostic system consisting of image pre-processing, feature extraction, and finally classification. The authors utilized k-Nearest Neighbor (K-NN) and Back Propagation Artificial Neural Network (BP-ANN) classifiers to determine the best accuracy for ACL tear classification. They found that BP-ANN produced a higher classification accuracy of 94.44%, compared to 87.33% for k-NN.

Chang et al.[17] evaluated multiple customized CNN models with variations in the input fields-of-view (i.e. full slice, cropped slice, dynamic patch-based sampling) and dimensionality (single slice, three slices, five slices) for detection of complete ACL tears. They determined that the model created to dynamically sample random cropped patches of images of the ACL performed the best in terms of detecting ACL tears when compared to a similar model that utilized the entire uncropped MRI slices. The model that utilized dynamic sampling had an accuracy of 96.7% and AUC of 0.971.

Bien et al.[16] used a fully automated deep learning CNN model with logistic regression for predicting the presence or absence of ACL tears on MRI after image pre-processing (intact=normal, mucoid degeneration, ganglion cyst, sprain; tear=low-grade partial tear with <50% fibers torn, high-grade partial tear with >50% of fibers torn, or complete tear). They reported that their best performing model produced an AUC of 0.965 (95% CI 0.938-0.993) for ACL tear detection. The specificity, sensitivity, and accuracy of the model were also reported as 0.968 (95% CI 0.890-0.991), 0.759 (95% CI 0.635-0.850), and 0.867 (95% CI 0.794-0.916), respectively.

AI Model Performance Compared with Human Observers: : ACL Tear Detection

Three (50.0%) studies compared the performance of an AI model for ACL tear detection with human medical experts.[16, 20, 21]

Liu et al.[20] trained multiple CNNs and applied them to a test set of 50 MRI images of full thickness ACL tears and 50 MR images with intact ACLs. They found that their model with the best overall diagnostic performance for detecting the presence or absence of a full thickness ACL tear produced an AUC of 0.98 (95% CI: 0.93-1.00, p <0.001) However, there was no statistically significant difference in diagnostic performance found between the AI model and clinical radiologist performance (Radiologist 0.90 (95% CI: 0.95-1.00); Fellow 0.90 (95% CI: 0.95-1.00); Resident1 0.93 (95%CI 0.88-0.98); Resident2 0.97 (95%Ci 0.94-1.00); Resident3 0.98 (95% CI 0.95-1.00).

Mazlan et al.[21] tested an SVM algorithm on 60 samples from MR images of 100 non-injured ACLs, 100 partially-torn ACLs, and 100 completely-torn ACLs that underwent pre-processing. They reported that the SVM model had an accuracy of 100% for classifying ACL MRI samples as normal, partial-tear, or complete-tear. The authors also sought to compare the diagnostic capability of their AI model to that of two medical experts. No statistically significant differences between the AI model and radiologists were found in terms of diagnostic capabilities, as the SVM and both medical experts correctly identified all 10 samples with 100% accuracy.

Bien et al.[16] compared their MRNet model's performance for detecting ACL tears to three musculoskeletal (MSK) radiologists on a testing set of 120 knee MR images, with the majority vote of 3 musculoskeletal (MSK) radiologists serving as the reference standard. They also evaluated changes in the diagnostic performance of clinical experts when the AI model predictions were provided to the radiologists during interpretation. Their model for detecting complete ACL tear produced an AUC of 0.968 (95% CI 0.890-0.991) compared to radiologist specificity of 0.933 (95% CI 0.906-0.953). However, results were not statistically significant (p-value=0.441). Radiologists achieved significantly higher sensitivities for tear diagnosis than the AI model (AUC 0.906 vs. 0.759, p-value=0.002). The AI model accuracy for ACL tear detection was 0.867 (95% CI 0.794-0.916), which was lower than the MSK radiologist accuracy of 0.920 (95% CI 0.900-0.937), which was not statistically significant (p-value=0.075). When provided with model assistance, there was a statistically significant increase (4.8%, p<0.001) in the clinical experts' specificity in identifying ACL tears. They reported that because the testing set consisted of 62 exams that were negative for ACL tear, the represented increase in specificity in the optimal clinical setting would potentially translate to the avoidance of three unnecessary surgeries for suspected ACL tears.

AI Model Performance: Meniscus Tear Detection

All six studies that investigated the performance of AI models on meniscus tear detection utilized MRI. Five (83.3%) of the studies analyzed sagittal-plane images only,[18, 19, 22, 23, 25] and one study analyzed sagittal, coronal, and axial images.[16] Four (66.67%) studies reported AUC data for meniscus tear detection, ranging from 0.847-0.910 (**Table 2**).

Fu et al.[19] compared the performance of two SVM models to detect meniscus tears. One model was created to select relevant meniscus MR features, while the other model implemented the SVM model without feature selection. The SVM model without feature selection produced an AUC of 0.727 for meniscus tear detection, while their model with feature selection yielded an AUC value of 0.912 for meniscus tear detection.

Couteax et al.[18] used an R-CNN model for tear detection (tear in any meniscus) and localization (anterior or posterior). The anterior meniscus was classified as torn when at least one network had detected a torn anterior meniscus and the posterior meniscus was classified as torn when the strict majority of the networks had detected a torn posterior meniscus. When they applied their model to a test set of 700 MRIs, the authors found that the model produced a weighted AUC score of 0.906.

Roblot et al.[23] used a three-step AI model where an image was transferred into a R-CNN trained to detect menisci as torn or normal, meniscus tear location, and whether the tear was horizontal or vertical. The model was tested on a dataset of 700 MRI images to perform detection of meniscus tear presence, position, and orientation. The AI model produced an AUC of 0.94 for presence of a meniscal tear, 0.92 for detection of the position of the two meniscal horns, and 0.83 for orientation of the tear. The overall combined AUC was 0.90.

Pedoia et al.[22] created a deep-learning model that combined meniscus segmentation and a 3D CNN for the detection and severity staging of meniscus lesions. The model was first built to discriminate between the presence of a lesion versus no lesion (including no lesion and intrasubstance abnormalities), and subsequently lesion severity (severe lesion=maceration of the meniscus; mild-moderate lesion=non-displaced tears and displaced and complex tears without deformity; no lesion= lesion absence and intrasubstance abnormalities). This model produced a lesion versus no lesion AUC of 0.89 on the test dataset and accuracies of 80.74%, 78.02%, and 75.00% for determining severe lesion versus mild-moderate lesion versus no lesion, respectively.

Bien et al.[16] also investigated the ability of their CNN models to detect meniscus tears following their investigations of diagnostic capabilities for ACL tears. For meniscus tear diagnosis, this group reported an accuracy of 0.725 (95% CI 0.639-0.797) and an AUC of 0.847 (95% CI 0.780-0.914).

Fazel-Zarandi et al.[25] used AI for MR image segmentation followed by the application of a Perceptron Neural Network (PNN) for classification of meniscal tears. A testing dataset of 50 MRIs were fed into the PNN and resulted in meniscus tear versus no meniscus tear accuracy of 90%. Classification rate (precision %) was also reported for five different settings of meniscus tear including: (1) medial anterior horn and posterior horn normal (88.82%), (2) lateral anterior horn and posterior horn normal (92.13%), (3) medial anterior horn normal and posterior horn torn (84.24%), (4) lateral anterior horn normal and posterior horn torn (91.96%) and (5) lateral anterior horn torn and posterior horn normal (87.64%).

AI Model Performance Compared with Humans: Meniscus

Two (33.3%) studies compared the performance of using an AI model for meniscus tear detection with human medical experts.[16, 22]

Bien et al.[16] compared the performance of their AI model with unassisted MSK radiologists for detecting meniscus tear (intact=normal, degenerative changes without tear, postsurgical changes without tear; tear=increased signal reaching the articular surface on at least two slices or morphologic deformity). When compared to the MSK radiologists in the study, the AI model had a statistically significant lower specificity (AUC 0.882, 95% CI 0.847-0.910 versus AUC 0.741, 95% CI 0.616-0.837; p-value=0.003) and accuracy (accuracy (0.849, 95% CI 0.823-0.871 versus 0.725, 95% CI 0.639-0.797, p=0.015). The sensitivity (0.820, 95% CI 0.781-0.853 versus 0.710, 95% CI 0.587-0.808; p=0.504) was also shown to be lower for the AI model compared to MSK radiologists, although this was not statistically significant.

Pedoia et al.[22] analyzed 1,478 MRI studies and utilized automatic segmentation of cartilage and meniscus using 2D U-Net. Detection and severity staging of meniscus and cartilage lesion was performed with a 3D CNN. Comparisons were made between their model and experts, where they sought to determine the inter-rater variability between three MSK radiologists (expert 1: >20 years of experience, expert 2: 10 years of experience, <1 year of experience) for determining meniscus lesion severity on selected cases. They found an average agreement between the three experts of 86.27% for no meniscus lesion, 66.48% for mild-moderate lesion, and 74.66% for severe lesion, while the best AI model obtained accuracies of 80.74% for no meniscus lesion, 78.02% for mild-moderate lesion, and 75.00% for severe lesion.

Quality Assessment

The average modified MINORS score among all studies was 4.9±1.0 (**Table 3**), indicating moderate to high methodological quality of the included studies on average. The most common reasons for loss of quality points was failure to describe both the inclusion and exclusion criteria of input features including patient and imaging selection (n=4, 36.4%) and failure to describe ground truth assignment (n=4, 36.4%). In the absence of clearly defined inclusion/exclusion criteria, selection bias cannot be excluded for the four studies. Failure to clearly describe the ground truth assignment risks publishing data from poorly trained AI models that may be inaccurate. The remaining limitations to quality was failure to describe the distribution of data, which also potentiates selection bias and misinterpretation of conclusions.

DISCUSSION

The main finding of the current study was that the AUC and prediction accuracy of AI models for detecting ACL tears ranged from 0.895-0.980 and 86.7%-100%, while the AUC and prediction accuracy for detecting meniscus tears ranged from 0.847-0.910 and 75.0%-90.0%. Additionally, in two studies that compared AI models to clinical experts, one found no significant differences in diagnostic accuracy, while one found that the AI model had a significantly lower specificity and accuracy than radiologists. Two studies reported that the addition of AI models significantly increased the diagnostic performance of clinicians compared to their efforts without these models. However, the heterogeneity of the studies and methodology identified in this systematic review suggests several areas for improvement and makes interpretation across studies challenging.

AI models are mathematical computing algorithms trained to integrate big data and autonomously assign labels to unseen data. Through multiple statistical iterations and pattern recognition, these models can apply learned features from training data sets and apply them to test sets to detect or classify lesions on many imaging modalities. Discrimination is also employed in conjunction with these AI models through generating a receiver operator curve (ROC) and generating a c-statistic (area under the curve, AUC). An AUC of 1.0 indicates perfect discrimination, while an AUC of 0.5 indicates discrimination similar to chance.[26] The current study found that the AUC for detecting ACL tears was near perfect ranging from 0.90-0.98 and that for detecting meniscus tears was excellent at 0.85-0.91.

Perhaps most importantly, the current study found that a combination of AI and human experts outperformed human experts or AI alone for diagnosis of ACL and meniscal tears, similar to the results of a prior systematic review of natural and artificial intelligence in neurosurgery.[27] AI methods have been previously applied to achieve or exceed human-level performance for tasks ranging from detection of distal radius fractures and hip fractures, to malignant pulmonary nodules.[28-32] The clinical relevance of AI applications for ACL and meniscal tears may be divided into the following categories: (1) human-level performance or better on routine tasks and (2) human-level performance or better on difficult tasks. During algorithm development, cohort curation, and study design, the studies included in this systematic review did not distinguish between these ultimate goals. If the intended purpose of AI algorithms is to diagnose lesions that are difficult for humans (partial tears, poor imaging), the AI algorithms should be specifically trained for this purpose. On the other hand, if the purpose of AI algorithms is to diagnose simple lesions, the output of these algorithms should be designed into clinically relevant categories ("definitely normal", "definitely abnormal", "not sure") to improve the efficiency and workflow of diagnosticians.

Interestingly, none of the included studies compared the use of AI to the gold standard of confirming ACL and meniscus lesions, which is arthroscopy. This is likely a function of the designs of the included studies, in which those that made a comparison to human experts were intended to do so. In the two studies that compared AI models to clinical experts, one found no significant differences in diagnostic accuracy and one found that AI was inferior in diagnostic accuracy. If it is assumed that clinical radiologist experts with only images at their disposal are less accurate at diagnosing these lesions than the gold standard of arthroscopy, then by transitive property, AI may be less accurate than arthroscopy as well. Future studies are warranted to determine the accuracy of AI for diagnosis of these lesions in comparison to arthroscopy as a ground truth label.

Another area for improvement is the requirement for adherence to peer-reviewed AI-specific guidelines. Efforts are underway to update the TRIPOD guidelines and develop a standardized system for AI applications in healthcare.[33] Several of the studies included in this analysis did not report measures of model performance such as precision-recall curves and Brier score that are key to interpreting diagnostic studies, particularly when outcomes are not balanced.[34, 35] AI is often criticized for the "black-box" nature of transformations required to take input data and produce meaningful outcomes. This block-box limits our ability to understand the specific imaging features an AI method utilized to produce its probability outputs. However, prior studies have provided explainable algorithms where output not only predicted probabilities, but also explanations in the form of augmented input images with heat maps highlighting the regions of interest for the specific diagnosis task.[29] Success by Lindsey et al. in applying these explainable AI techniques for distal radius fracture detection should also become the norm for sports medicine.[29]

AI has significant implications for the future of diagnosis in orthopaedic sports medicine, but clinicians must be informed and critical consumers of this rapidly evolving field.[36] The focus of this review was on diagnostic applications of AI in orthopaedic sports medicine and a similar analysis for prognostic applications of AI remains to be undertaken. Interestingly, AI did not outperform human experts, which is potentially a result of the early applications of AI in this field and the need to further refine and modify algorithms. It is possible that as these algorithms continue to be trained with more data, that prediction accuracies and image recognition improves and may eventually outperform these experts. As described above, the drawbacks of current AI applications for diagnosis of ACL and meniscal tears should inspire future studies to follow standardized guidelines to allow for reliable and reproducible research. However, advantages of AI include the potential for the rapid and accurate identification and diagnosis of pathology, such as ligamentous and meniscal tears, which may initially be missed by the human eye. Eventually, AI and related technology may progress to the point where fewer working personnel are required to perform these tasks (i.e., the development of pathology-specific AI algorithms, where only one attending radiologist is required to double-check the finding made by the algorithm, as opposed to the current use of teams of multiple radiologists who are burdened with large numbers of images with multiple views to read). This, in turn, may increase timeliness of reads and decrease healthcare costs. Ultimately, ensuring clinical relevance at every step in algorithm conception, design, and development will lead to true progress for AI in orthopaedic sports medicine. Applications of AI in orthopaedic surgery are rapidly growing and periodic updates will be required to appropriately represent the state of the literature in the years to come. At present, inadequate reference standards to train and test AI is the biggest hurdle to overcome prior to integration into clinical workflows.

Limitations

There are a number of limitations that must be acknowledged to appropriately interpret the results of this study. This was a systematic review that followed the PRISMA guidelines, but did not include a more formal quantitative meta-analysis due to study heterogeneity. Despite the comprehensive search, the total number of studies included in this analysis was relatively small. Another limitation is that no studies included diagnostic arthroscopy as the gold standard reference to diagnose ACL or meniscus lesions, which may limit the applicability of the findings to clinical practice. Finally, studies were inherently heterogeneous given the AI models used, inclusion/exclusion criteria, ground truth label assignments, and imaging protocols (**Tables 1 and 2**).

CONCLUSION

AI prediction capabilities were excellent and may enhance the diagnosis of ACL and meniscus pathology; however, AI did not outperform clinical experts.

REFERENCES

1. Bini SA. Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? *J Arthroplasty.* 2018;33:2358-2361.

2. Helm JM, Swiergosz AM, Haeberle HS, et al. Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions. *Curr Rev Musculoskelet Med.* 2020;13:69-76.

3. Kunze KN, Karhade AV, Sadauskas AJ, Schwab JH, Levine BR. Development of Machine Learning Algorithms to Predict Clinically Meaningful Improvement for the Patient-Reported Health State After Total Hip Arthroplasty. *J Arthroplasty.* 2020.

4. Karhade AV, Ahmed AK, Pennington Z, et al. External validation of the SORG 90-day and 1-year machine learning algorithms for survival in spinal metastatic disease. *Spine J.* 2020;20:14-21.

5. Karhade AV, Schwab JH, Bedair HS. Development of Machine Learning Algorithms for Prediction of Sustained Postoperative Opioid Prescriptions After Total Hip Arthroplasty. *J Arthroplasty.* 2019;34:2272-2277 e2271.

6. Karhade AV, Shah AA, Bono CM, et al. Development of machine learning algorithms for prediction of mortality in spinal epidural abscess. *Spine J.* 2019;19:1950-1959.

7. Ogunyemi O, Kermah D. Machine Learning Approaches for Detecting Diabetic Retinopathy from Clinical and Public Health Records. *AMIA Annu Symp Proc.* 2015;2015:983-990.

8. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA.* 2016;316:2402-2410.

9. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542:115-118.

10. Thio Q, Karhade AV, Ogink PT, et al. Development and Internal Validation of Machine Learning Algorithms for Preoperative Survival Prediction of Extremity Metastatic Disease. *Clin Orthop Relat Res.* 2019.

11. Thio Q, Karhade AV, Ogink PT, et al. Can Machine-learning Techniques Be Used for 5-year Survival Prediction of Patients With Chondrosarcoma? *Clin Orthop Relat Res.* 2018;476:2040-2048.

12. Langerhuizen DWG, Janssen SJ, Mallee WH, et al. What Are the Applications and Limitations of Artificial Intelligence for Fracture Detection and Classification in Orthopaedic Trauma Imaging? A Systematic Review. *Clin Orthop Relat Res.* 2019;477:2482-2491.

13. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ.* 2009;339:b2535.

14. Harris JD, Quatman CE, Manring MM, Siston RA, Flanigan DC. How to write a systematic review. *Am J Sports Med.* 2014;42:2761-2768.

15. Abdullah AA A-ZN. Design of an Intelligent Diagnostic System for Detection of Knee Injuries. *Applied Mechanics and Materials.* 2013;339:219-224.

16. Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Med.* 2018;15:e1002699.
17. Chang PD, Wong TT, Rasiej MJ. Deep Learning for Detection of Complete Anterior Cruciate Ligament Tear. *J Digit Imaging.* 2019;32:980-986.
18. Couteaux V, Si-Mohamed S, Nempont O, et al. Automatic knee meniscus tear detection and orientation classification with Mask-RCNN. *Diagn Interv Imaging.* 2019;100:235-242.
19. Fu J LC, Wang C, Ou Y. Computer-aided diagnosis for knee meniscus tears in magnetic resonance imaging. *Journal of Industrial and Production Engineering.* 2013;30:67-77.
20. Liu F GB, Zhou Z, Samsonov A, Rosas H, Lian K, Sharma R, Kanarek A, Kim J, Guermazi A, Kijowski R. Fully Automated Diagnosis of Anterior Cruciate Ligament Tears on Knee MR Images by Using Deep Learning. *Radiology: Artifical Intelligence.* 2019;1:1-10.
21. Mazlan SS AM, Kadir Bakti ZA. Anterior Cruciate Ligament (ACL) Injury Classification System Using Support Vector Machine (SVM). *Proc. Int. Engin and Tech. .* 2017:1-5.
22. Pedoia V, Norman B, Mehany SN, Bucknor MD, Link TM, Majumdar S. 3D convolutional neural networks for detection and severity staging of meniscus and PFJ cartilage morphological degenerative changes in osteoarthritis and anterior cruciate ligament subjects. *J Magn Reson Imaging.* 2019;49:400-410.
23. Roblot V, Giret Y, Bou Antoun M, et al. Artificial intelligence to diagnose meniscus tears on MRI. *Diagn Interv Imaging.* 2019;100:243-249.
24. Stajduhar I, Mamula M, Miletic D, Unal G. Semi-automated detection of anterior cruciate ligament injury from MRI. *Comput Methods Programs Biomed.* 2017;140:151-164.
25. Zarandi MH, Khadangi A, Karimi F, Turksen IB. A Computer-Aided Type-II Fuzzy Image Processing for Diagnosis of Meniscus Tear. *J Digit Imaging.* 2016;29:677-695.
26. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation.* 2007;115:928-935.
27. Senders JT, Arnaout O, Karhade AV, et al. Natural and Artificial Intelligence in Neurosurgery: A Systematic Review. *Neurosurgery.* 2018;83:181-192.
28. Nam JG, Park S, Hwang EJ, et al. Development and Validation of Deep Learning-based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. *Radiology.* 2019;290:218-228.
29. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. *Proceedings of the National Academy of Sciences of the United States of America.* 2018;115:11591-11596.
30. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine.* 2019;25:44-56.
31. Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting hip fractures with radiologist-level performance using deep neural networks. *arXiv preprint arXiv:1711.06504.* 2017.
32. Lee H, Yune S, Mansouri M, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nature biomedical engineering.* 2019;3:173-182.

33. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet (London, England).* 2019;393:1577-1579.
34. Ozenne B, Subtil F, Maucort-Boulch D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of clinical epidemiology.* 2015;68:855-859.
35. Brier GW. Verification of forecasts expressed in terms of probability. *Monthly weather review.* 1950;78:1-3.
36. Ramkumar PN, Kunze KN, Haeberle HS, et al. Clinical and Research Medical Applications of Artificial Intelligence: Fundamentals for the Orthopaedic Surgeon. *Arthroscopy.* 2020.

Figure legends:

Figure 1: Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) flowchart for included studies.

Tables

Table 1. Artificial Intelligence and Methodology for Anterior Cruciate Ligament Studies

Study	Input Features	Imaging plane	Dataset size	Anatomic structure	Ground/truth label assignment	Output classes	AI models used	Pretrained CNN	Size training set	Size validation set/validation method	Size test set	Performance (accuracy/AUC)
Štadjuhar et al. 2017	MR	Sagittal	917	ACL	Two Radiologists	2	HOG+linSVM	NA	NA	10- fold cross-validation	NA	NA/0.894 (linear-kernel SVM+HOG: injury-detection problem) NA/0.943 (linear-kernel SVM +HOG: complete rupture)
							HOG+RF					NA/0.884 (injury-detection) NA/0.937 (complete rupture)
							GIST+rbfSVM					NA/0.889 (injury-detection) NA/0.913 (complete rupture)
							GIST+RF					NA/0.880 (injury-detection) NA/0.895 (complete rupture)
Mazlan et al. 2017	MR	Sagittal	300	ACL	Two Medical Experts with >7 years of experience	3	SVM	NA	210 (70%)	30 (10%)	60 (20%)	100%/NA
Chang et al. 2019	MR	Coronal	260	ACL	Visual inspection by a board-certified subspecialist	2	CNN	ResNet-Derived, U-net	160	40/5-fold cross-validation	60	0.967/0.971

					MSK radiologist							
Bien et al. 2018	MR	Sagittal, Coronal, Axial	1,370	ACL	Three MSK radiologists' majority vote, average 12 years in practice	3	CNN	AlexNet, MRNet	1,130	120	120	Model ACL tear: 0.867 [95%CI 0.794, 0.916]/ 0.965 [95% CI 0.938, 0.993] Model abnormality detection: NA/0.937 [95%CI 0.895,0.980]
Abdullah et al. 2013	MR	Sagittal	90	ACL	NA	3	BP ANN, k-NN	NA	72	NA/5-fold and 6-fold	18	BP ANN: 0.9444/NA k-NN: 0.878333/NA
Liu et al. 2019	MR	Sagittal	350	ACL	Orthopedic Surgeon + Fellowship trained MSK radiologist with >15 years of clinical experience	2	CNN	LeNet-5, YOLO, DenseNet, VGG16, AlexNet	200 (57%)	50 (14%)	100 (29%)	NA/0.98 (DenseNet 95% CI (0.93-1.0) p<0.001

MR, magnetic resonance; NA, not available; ACL, anterior cruciate ligament; CNN, convolutional neural network; ANN, artificial neural network; BP, back-propagation; k-NN, K-nearest neighbors; RF, random forest; HOG, histogram of oriented gradients; GIST, generalized search tree; rbf, radial basis function; MSK, musculoskeletal; YOLO, you only look once; VGG, visual graphics group (type of neural network architecture); AUC, area under the curve; AI, artificial intelligence.

Table 2. Artificial Intelligence and Methodology for Meniscus Studies

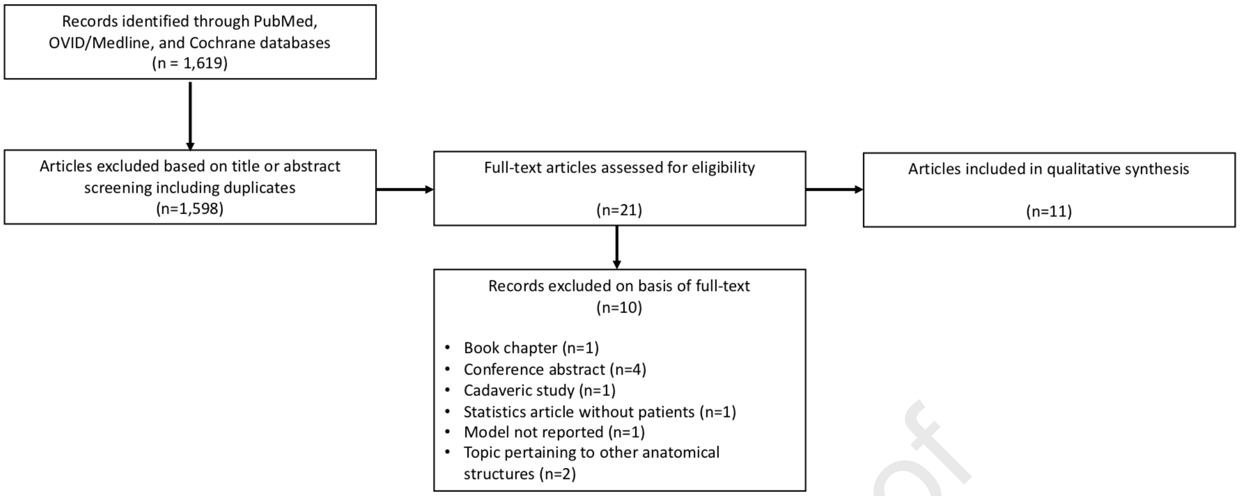
Study	Input Features	Imaging plane	Dataset size	Anatomic location	Ground truth label assignment	Output classes	AI models	Pretrained CNN	Size training set	Size validation set/validation method	Size test set	Performance (accuracy/AUC)
Pedoia et al. 2018	MR	Sagittal	1,478	Meniscus	Expert 1: >20 years experience, Expert 2: >10 years experience, Expert 3: <1 year training as a radiologist	2	2D U-Net, 3D CNN	NA	960 (65%)	295 (20%)	221 (15%)	Binary mode (lesion vs non-lesion): NA/0.89
	MR	Sagittal	1,478	Meniscus	Expert 1: >20 years experience, Expert 2: >10 years experience, Expert 3: <1 year training as a radiologist	3	2D U-Net, 3D CNN, RF	NA	960 (65%)	295 (20%)	221 (15%)	(no lesion, mild-moderate lesion, severe lesion): 80.74%/NA, 78.02%/NA, 75.00%/NA
Fazel-Zarandi et al. 2016	MR	Sagittal	248	Meniscus	NA	2	IT2FCM, IT2PCM, PNN	NA	198 (80%)	NA	50 (20%)	0 and 1 mode: 90%/NA Binary mode: 78%/NA
Couteaux et al. 2019	MR	Sagittal	1,128	Meniscus	NA	6	R-CNN, ConvNet	ResNet-101, ConvNet, R-CNN	246	54	700	NA/0.906
Fu et al. 2013	MR	Sagittal	166	Meniscus	NA	2	SVM	NA	NA	5-FCA	NA	SVM model: NA/0.727 SFFS+SVM: NA/0.912
Roblot et al. 2019	MR	Sagittal	2,246	Meniscus	CSV file	6	Fast RCNN, Faster RCNN	NA	1,123	NA	700	NA/0.90
Bien et al. 2018	MR	Sagittal, Coronal, Axial	1,370	Meniscus	Three MSK radiologists' majority vote, average 12 years in practice on an internal validation set of 120 exams	3	CNN	AlexNet, MRNet	1,130	120	120	Model Meniscal tear: 0.725 [95%CI 0.639, 0.797]/0.847 (95% CI 0.780-0.914);

MR, magnetic resonance; NA, not available; 2D, two-dimensional; 3D, three-dimensional; CNN, convolutional neural network; R-CNN, regions with CNN; SVM, support vector machine; IT2FCM, Interval type-2 fuzzy c-means; PNN, probabilistic neural network;

SFSS, sequential floating forward selection; AI, artificial intelligence; AUC, area under the curve; MSK, musculoskeletal; CSV, comma-separated values.

Table 3. Quality appraisal of included studies.

Study	Quality Appraisal Score
Abdullah et al. 2013	4 – Failure to describe both inclusion/exclusion criteria of input features and ground truth assignment
Fu et al. 2013	3 – Failure to describe inclusion/exclusion criteria of input features, ground truth assignment, and distribution of data
Fazel-Zarandi et al. 2016	4 – Failure to describe ground truth assignment and distribution of data
Mazlan et al. 2017	6
Štadjuhar et al. 2017	5 – Failure to describe distribution of data
Bien et al. 2018	5 – Failure to describe inclusion/exclusion criteria of input features
Pedoia et al. 2018	6
Chang et al. 2019	6
Couteaux et al. 2019	4 – Failure to describe inclusion/exclusion criteria of input features and ground truth assignment
Liu et al. 2019	6
Roblot et al. 2019	5 – Failure to describe inclusion/exclusion criteria of input features


Records identified through PubMed, OVID/Medline, and Cochrane databases
(n = 1,619)
Articles excluded based on title or abstract screening including duplicates
(n=1,598)
Full-text articles assessed for eligibility
(n=21)
Articles included in qualitative synthesis
(n=11)
Records excluded on basis of full-text
(n=10)
• Book chapter (n=1)
• Conference abstract (n=4)
• Cadaveric study (n=1)
• Statistics article without patients (n=1)
• Model not reported (n=1)
• Topic pertaining to other anatomical structures (n=2)