Machine learning in shoulder arthroplasty

a systematic review of predictive analytics applications

From Schulthess Clinic, Zurich, Switzerland

Correspondence should be sent to T. Schneller tim.schneller@kws.ch

Cite this article:
Bone Jt Open 2025;6(2):
126–134.

DOI: 10.1302/2633-1462.62.BJO-2024-0234.R1

T. Schneller,[1] M. Kraus,[1,2] J. Schätz,[1,3] P. Moroder,[1] M. Scheibel,[1,4] A. Lazaridou[1,5]

[1]Department for Shoulder and Elbow Surgery, Schulthess Clinic, Zurich, Switzerland
[2]Department of Traumatology, University Hospital Zurich, Zurich, Switzerland
[3]Institute for Therapies and Rehabilitation, Cantonal Hospital Winterthur, Winterthur, Switzerland
[4]Center for Musculoskeletal Surgery, Charité-Universitaetsmedizin, Berlin, Germany
[5]Department of Anesthesiology, Brigham & Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA

Aims

Machine learning (ML) holds significant promise in optimizing various aspects of total shoulder arthroplasty (TSA), potentially improving patient outcomes and enhancing surgical decision-making. The aim of this systematic review was to identify ML algorithms and evaluate their effectiveness, including those for predicting clinical outcomes and those used in image analysis.

Methods

We searched the PubMed, EMBASE, and Cochrane Central Register of Controlled Trials databases for studies applying ML algorithms in TSA. The analysis focused on dataset characteristics, relevant subspecialties, specific ML algorithms used, and their performance outcomes.

Results

Following the final screening process, 25 articles satisfied the eligibility criteria for our review. Of these, 60% focused on tabular data while the remaining 40% analyzed image data. Among them, 16 studies were dedicated to developing new models and nine used transfer learning to leverage existing pretrained models. Additionally, three of these models underwent external validation to confirm their reliability and effectiveness.

Conclusion

ML algorithms used in TSA demonstrated fair to good performance, as evidenced by the reported metrics. Integrating these models into daily clinical practice could revolutionize TSA, enhancing both surgical precision and patient outcome predictions. Despite their potential, the lack of transparency and generalizability in many current models poses a significant challenge, limiting their clinical utility. Future research should prioritize addressing these limitations to truly propel the field forward and maximize the benefits of ML in enhancing patient care.

Take home message

- Machine learning has the potential to enhance clinical outcomes in total shoulder athroplasty by improving decision-making and tailoring treatment to the patient.
- However, systematic issues regarding transparency, replicability, and external validation have to be resolved until extensive clinical adoption can be achieved.

Introduction

Recent advancements in machine learning (ML) and the increasing availability of big data have opened promising avenues for optimizing orthopaedic treatments. The use of ML techniques has become a cornerstone

in medical research in recent years, offering a wide range of applications that present significant opportunities in orthopaedic surgery.[1]

ML, a subset of artificial intelligence (AI), can be divided into two distinct types – unsupervised and supervised. Unsupervised ML is used to identify patterns or clusters in unlabelled data, such as grouping customers based on their purchasing behaviour to design marketing campaigns. The more popular supervised ML models, on the other hand, are trained on labelled datasets, and the ML model then makes predictions based on the label and the input data. Supervised ML can be further divided into either classification or regression tasks. Classification is where the prediction output can only take a limited number of values (e.g. fraudulent/not fraudulent). In regression tasks, the ML model's training data is labelled with continuous data, thus predictions are on a continuous scale too (e.g. predicting the price of a house). Levin et al[2] recently published an overview on AI in shoulder surgery, summarizing common AI terms and discussing current and future applications; readers unfamiliar with AI or ML are advised to use this article as a reference aid.

A systematic review on studies applying ML to predict clinical outcomes within different orthopaedic disciplines identified 18 studies exploring the efficacy of ML algorithms in predicting clinically significant outcomes (CSOs), with all of them utilizing the minimal clinically important difference as a primary outcome measure.[3] ML algorithms demonstrated favourable performance in predicting CSOs across most studies. This study highlighted the importance of utilizing ML in outcomes-based research due to its capacity to enhance prediction accuracy by discerning complex data relationships through pattern recognition and learning.

Another systematic review focusing on shoulder surgeries assessed the scope and validity of current clinical AI applications.[4] Investigations applying AI to shoulder surgery predominantly centred on two key domains: 1) automated imaging analysis encompassing tasks such as identifying rotator cuff tears and assessing shoulder implants; and 2) risk prediction analyses that included evaluating perioperative complications, functional outcomes, and patient satisfaction. The performance of models varied considerably, with the highest area under the curve (AUC) ranging from 0.681 to a perfect score of 1.00. Remarkably, only two studies reported external validation of their models.

Currently, there is a noticeable gap in the literature regarding systematic reviews that compare multiple ML models and algorithms specifically applied to total shoulder arthroplasty (TSA). This oversight is significant, especially considering the rapid advancements in ML and its increasing application in medical fields. A comprehensive review that synthesizes the most recent findings and methodologies could provide invaluable insights and guidance for future research and clinical applications in TSA. The aim of this systematic review is to identify ML algorithms pertaining to TSA, such as models forecasting clinical outcomes after TSA or models analyzing images. Additionally, this review seeks to evaluate the effectiveness of these models by examining performance metrics.

Methods

Data sources and search strategy

This study adheres to the reporting standards outlined by the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) reporting guidelines.[5] The study protocol was preregistered on PROSPERO (no. CRD42024524721). An electronic search was performed using PubMed, EMBASE, and the Cochrane Library including Cochrane Reviews and Cochrane Central Register of Controlled Trials databases. A comprehensive search strategy was designed a priori and specifically tailored for each database. The search was restricted to: 1) completed; and 2) human studies. We conducted one literature search. The terms used in our search included the following: ("shoulder arthroplasty" OR "total shoulder arthroplasty" OR "tsa" OR "shoulder prosthesis" OR "shoulder prothesis" OR "shoulder prostheses" OR "reverse shoulder arthroplasty") AND ("neural network" OR "deep learning" OR "deep neural network" OR "machine learning" OR "supervised learning" OR "classification algorithms").

Selection of studies

After identification of the literature, three independent reviewers (TS, MK, JS) each screened one-third of the abstracts yielded to shortlist articles based on predefined exclusion criteria. Studies were excluded if: 1) there were no original, extractable clinical data (e.g. review articles, letters to the editor); or 2) no full-text articles available (e.g. conference proceedings, study protocols); and 3) if the work pertained to surgeries other than TSA; or 4) did not present model performance metrics. Subsequently, full texts were retrieved and screened for exclusion criteria again. Finally, data extraction was performed by the same reviewers manually and independently using a pilot-tested data extraction form. Information extracted included primary author name, year of publication, primary goal of the study, primary algorithm used, use of transfer learning and data augmentation, dataset size, dataset split, type of validation, primary model performance metrics, whether the tool is used in clinical practice, public accessibility of the tool, and code sharing.

Statistical analysis

The methodological quality of all eligible studies was assessed using the Minimum Information about CLinical Artificial Intelligence Modeling (MI-CLAIM) checklist.[6] This tool was specifically developed for assessing ML applications in medicine. While there are alternative quality assessment tools, we elected to use MI-CLAIM as it focuses on critical domains relevant to ML research, including data handling, model evaluation, and reporting standards. The quality assessment, however, was a secondary aim of this manuscript, intended to provide a more structured overview. Any disagreements pertaining to extracted data or assessment of methodological quality between raters were resolved by discussion. As the studies were extremely heterogeneous in terms of the pathology examined and the outcome parameters investigated, we refrained from ranking the clinical relevance. Figure creation was performed using the statistical software R v. 4.4.1 (R Foundation for Statistical Computing, Austria) and the package ggplot2 v. 3.5.1.[7]

Results

The initial screening process yielded a total of 111 studies; 61 articles were excluded based on the abstract and predefined exclusion criteria. Subsequent screening of the potentially relevant 50 full-text studies resulted in a final selection of 25 eligible publications (Figure 1).

This systematic review analyzed 25 articles published between 2016 and 2024 that covered 9,415 medical images and 300,204 patients (Table I). The input for the ML models was divided into image and tabular data derived from 40% and 60% of the articles, respectively.

The MI-CLAIM checklist analysis highlights significant variability in how well imaging and tabular studies adhere to the established reporting standards for ML models (Figure 2). For imaging studies, the most frequent criteria that were unmet included the provision of model details, discussion about data shifts, and code sharing. In contrast, imaging studies consistently reported problem and question statements, cohort characteristics, and sample representativeness.

Tabular studies exhibited a comparable trend, generally adhering well to essential reporting elements like defining the research problem, detailing cohort characteristics, and describing the baseline comparisons. Similarly to the studies involving images, the tabular studies also showed significant deficiencies in providing detailed model information, independence of training and test data, and code sharing.

Implant recognition

Sultan et al[11] developed a model with the goal of classifying implants from different manufacturers, which displayed an average accuracy of 85.92%, an f1-score of 84.69%, a precision of 85.33%, and a recall of 84.11%. In a later study, the same group improved their model with an accuracy of 89.09%, an f1-score of 87.94%, a precision of 89.54%, and a recall of 86.57%.[12]

Geng et al[14] developed a ML model with a similar goal of predicting the manufacturer and specific model of TSA implants. Performance metrics showed an overall accuracy of 93.9% and an average precision, recall, and f1-score of 94%.

Yi et al[10] aimed not only to identify the implant manufacturer, but also to detect the presence of a TSA and determine if the implant was configured anatomical or reverse. The resulting model showed an AUC of 1.00 in predicting the presence of a TSA, an AUC of 0.97 in distinguishing between anatomical TSA (aTSA) and reverse TSA (rTSA), and an AUC of 0.86 to 1.00 in identifying specific TSA models.

Yang et al[17] proposed a model that aimed to execute three separate tasks simultaneously in one process: predicting the side of the shoulder; distinguishing between aTSA, rTSA, and a preoperative shoulder; and assessing the imaging view. Their model achieved an average accuracy, precision, recall, and f1-score of 99.1%.

Urban et al[9] showed superiority of convolutional neural networks (CNNs) over more classic ML approaches like random forest or gradient boosting with regard to classifying the implant manufacturer. The authors proposed a model which was able to predict the manufacturer with an accuracy of 80%, while more classic ML algorithms merely reached accuracies of 51% to 56%.

Kunze et al[13] developed two more CNNs aimed to classify the implant manufacturer and the exact implant type. The first model displayed an overall accuracy of 97.1% and an AUC of 0.99 to 1.00, discriminating between different implant manufacturers. The second model additionally used implant-specific details, achieving an overall accuracy of 99.1% in detecting the implant type. Saliency maps indicate that both proposed CNNs learn from meaningful implant-specific features on the radiographs.

Tendon integrity

Guo et al[15] used a deep CNN to automatically diagnose supraspinatus tears from MRI data in a binary fashion (tear or no tear). The proposed model was both internally and externally validated, achieving respective AUC values of 0.882 and 0.921. The model's performance was benchmarked at the local institution. It demonstrated results comparable to those of senior clinicians and surpassed those of junior clinicians. However, the authors did not offer any explanations for the model's predictions.

Articular margin plane

Only one article included a regression task from medical imaging. Tschannen et al[8] developed a comprehensive automated model based on the random forest regression algorithm to precisely predict coordinates of the articular margin plane (AMP) based on CT images of the upper arm. However, the study lacked explicit documentation regarding the model's performance assessment. Mean absolute errors (MAEs) of 6.51° and 2.4 mm were reported for both the angular and positional attributes of the AMP.

Bone density estimation

Ritter et al[16] developed a model based on the support vector machine algorithm for prediction of bone density. The authors used a combination of clinical data and CT scans of humerus cadavers to develop a model predicting bone density of the proximal humerus. The model achieved an AUC of 0.83 in predicting the surgeons' intraoperative assessment of bone density.

Outcome prediction

Kumar et al[21,28] used a variety of models to accurately predict internal rotation scores at various postoperative intervals for patients undergoing aTSA and rTSA. The same authors additionally demonstrated the effectiveness of the XGBoost model in predicting various postoperative functional outcomes.[23] The model showed robust accuracy across multiple postoperative scores and timepoints. Simmons et al[31] performed an external validation of this specific tool using patients undergoing primary aTSA or rTSA. The validation analysis revealed that the tools' predictions were generally accurate, with MAEs within 10% of the initial internal validation results.

Franceschetti et al[30] aimed to predict postoperative anterior elevation following rTSA. A total of 28 features from 105 patients treated in two clinics were used as input to the model. The support vector regression algorithm demonstrated the best performance with a MAE of 12°.

In a retrospective cohort study of 472 patients with primary glenohumeral osteoarthritis undergoing TSA, McLendon et al[24] employed different ML models to predict postoperative improvements measured with the American

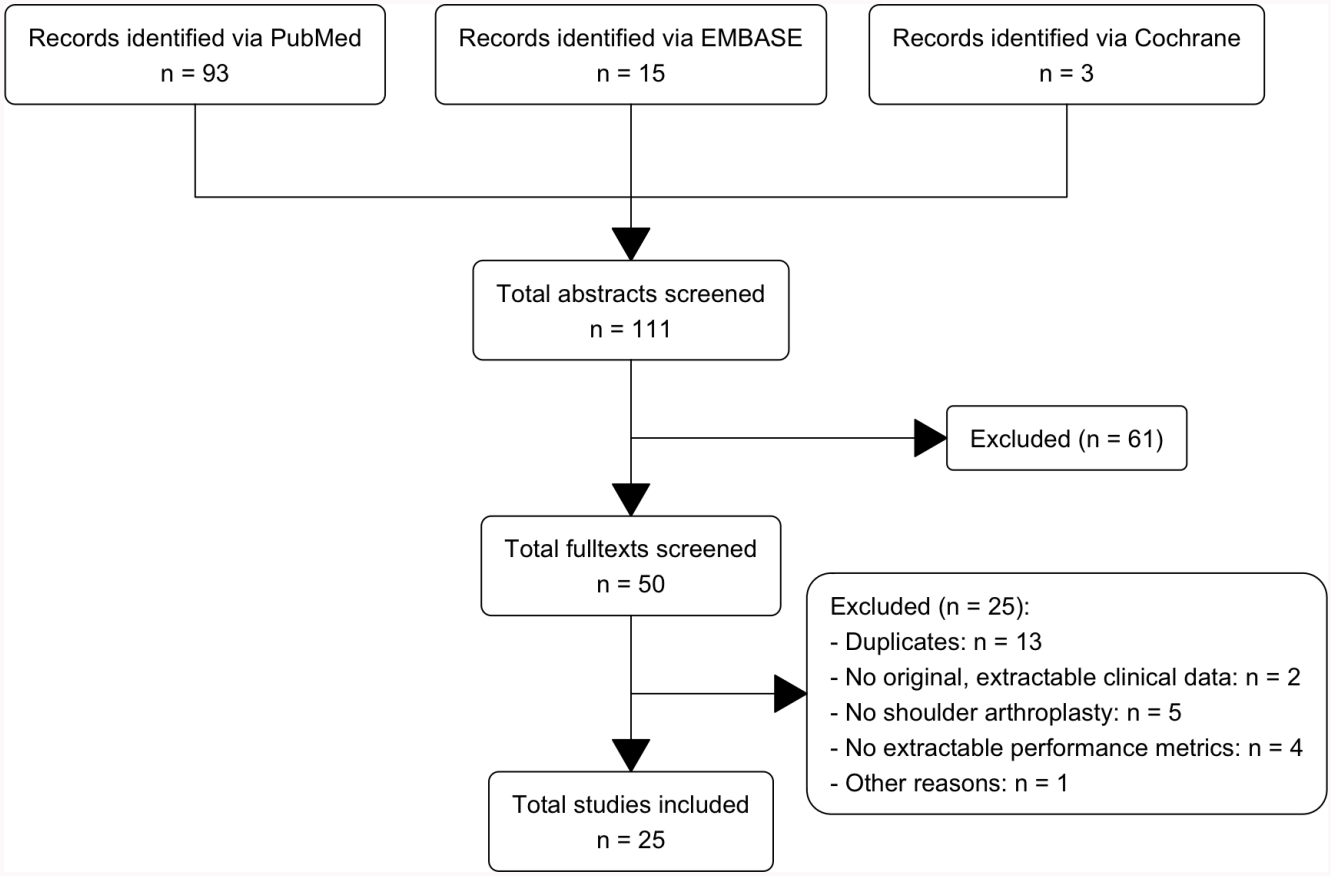

Fig. 1
Of the 25 excluded articles, the article listed under "other reasons" was an unsuitable evaluation of an existing prediction model's fairness based on ethnicity, age, and sex.

Shoulder and Elbow Surgeons score. The best performing model used baseline functional scores and morphological variables, achieving the highest accuracy, with probability values of 0.94 for minimal, 0.95 for moderate, and 0.94 for significant improvement.

Devana et al[26] developed several models with the goal to identify patients at risk of complications or readmissions within 30 days post surgery. Five different ML algorithms were compared. The XGBoost model exhibited superior efficacy, with an AUC of 0.681 in predicting the occurrence of major postoperative complications or readmissions within 30 days.

Karnuta et al[19] compared the effectiveness of artificial neural networks (ANNs) to conventional statistical approaches in predicting complications after aTSA and rTSA. The authors focused on predicting prolonged duration of stay, discharge disposition, and inpatient expenses using data from over 100,000 patients. The ability of ANNs to provide accurate forecasts for these outcomes following TSA was demonstrated, with accuracies of 91.8%, 73.1%, and 76.5%, respectively. Results demonstrated that ANNs have the capacity to handle complicated data interactions, providing more predictive power than traditional statistical models.

Oeding et al[29] explored the risk of early dislocation leading to revision surgery within three months following rTSA. The proposed model, employing the XGBoost algorithm, was able to accurately identify rare dislocation events with a recall of 84%.

Gowd et al[18] used data from the American College of Surgeons–National Surgical Quality Improvement Program to develop predictive models for the occurrence of adverse events following TSA. The study analyzed data from 17,119 cases. Preoperative and intraoperative factors, including patient demographics and health status, were considered. Random forest achieved the highest accuracy in predicting any adverse event, while logistic regression provided the highest AUC for this outcome.

Patient selection

Lopez et al[22] developed models to assist healthcare providers in preparing adequate postoperative care and support arrangements. They applied boosted decision trees and ANNs to develop predictive models from a large national surgical registry with 21,544 patients predicting non-home discharge following TSA. The best performing models were ANNs reaching an overall accuracy of 92.5%.

Biron et al[20] developed a clinical decision support tool that helps to select patients for outpatient TSA based on length of stay. They applied demographic and comorbidity data from a multicentre study to a random forest model to predict the length of stay in a binary fashion, achieving an AUC of 0.77.

Polce et al[25] developed a tool to predict patient satisfaction two years after TSA based on demographic and patient-specific factors, as well as whether an aTSA or rTSA is implanted. The tool was initially developed based on

Table I. Overview of the most important characteristics of all 25 eligible studies selected for systematic review.

First author	Primary goal	Primary algorithm	Transfer learning	Data augmentation	Dataset size	Dataset split	Validation	Primary metric	Tool used in clinical practice	Public accessibility of tool	Code policy available
Imaging*											
Tschannen et al[8]	Articular margin plane prediction	RF regression	No	Yes	72 images	N/A	Internal	MAE 2.4 mm, 6.5°	No	No	No
Urban et al[9]	Implant recognition	CNN	Yes	Yes	597 images	N/A	Internal	ACC 80.4%	No	In part	N/A
Yi et al[10]	Implant recognition	CNN	Yes	Yes	482 images	70/10/20	Internal	AUC 0.86 to 1.0	No	N/A	N/A
Sultan et al[11]	Implant recognition	CNN	Yes	Yes	538 images	90/10	Internal	ACC 85.92%	No	On request	On request
Sultan et al[12]	Implant recognition	CNN	Yes	Yes	597 images	90/2/8	Internal	ACC 89.09%	No	Yes	No
Kunze et al[13]	Implant recognition	CNN	Yes	Yes	3,060 images	80/20	Internal	ACC 97.1%	No	No	Yes
Geng et al[14]	Implant recognition	CNN	Yes	No	696 images	70/30	Internal	ACC 93.9%	No	No	No
Guo et al[15]	Tendon integrity recognition	CNN	No	Yes	770 images	72/19/9	External	ACC 82% to 87%	No	On request	On request
Ritter et al[16]	Bone density estimation	SVM	No	No	300 images	N/A	External	ACC 87%	No	No	No
Yang et al[17]	Implant recognition	CNN	Yes	Yes	2,303 images	80/20	Internal	ACC 95%	No	No	No
Tabular*											
Gowd et al[18]	Prediction of postop complications	RF	No	No	17,119 ptns	80/20	Internal	ACC 95.4%	No	No	No
Karnuta et al[19]	Cost prediction	ANN	No	No	111,147 ptns	70/10/20	Internal	ACC 76.5%	No	No	Yes
Biron et al[20]	Identification of ptns suited for outpatient treatment	RF	No	No	4,500 ptns	70/30	Internal	AUC 0.77	No	No	No
Kumar et al[21]	Prediction of postop outcome	WD	Yes	No	4,782 ptns	66.7/33.3	Internal	ACC > 85%	Yes	Yes	No
Lopez et al[22]	Prediction of non-home discharge	ANN	No	No	21,544 ptns	80/20	Internal	AUC 0.85	No	N/A	N/A
Kumar et al[23]	Prediction of postop outcome	XGB	No	No	5,774 ptns	66.7/33.3	Internal	ACC > 82%	No	N/A	N/A
McLendon et al[24]	Prediction of postop outcome	N/A	No	No	300 ptns	N/A	Neither	Recall 84% to 95%	N/A	N/A	N/A
Polce et al[25]	Prediction of postop outcome	SVM	No	No	413 ptns	80/20	Internal	AUC 0.8	No	Yes	No
Devana et al[26]	Prediction of postop complications	XGB	No	No	2,799 ptns	80/20	Internal	AUC 0.68	No	No	No
Gowd et al[27]	Cost prediction	Gradient boosting trees	No	No	49,354 ptns	80/20	Internal	AUC 0.87	No	N/A	N/A
Kumar et al[28]	Prediction of postop outcome	WD	Yes	No	6,468 ptns	66.7/33.3	Internal	Mean MAE 1.09	Yes	Yes	N/A
Oeding et al[29]	Prediction of postop complications	XGB	No	No	74,697 ptns	80/20	Internal	AUC 0.71	No	No	No

(*Continued*)

(*Continued*)

First author	Primary goal	Primary algorithm	Transfer learning	Data augmentation	Dataset size	Dataset split	Validation	Primary metric	Tool used in clinical practice	Public accessibility of tool	Code policy available
Franceschetti et al[30]	Prediction of postop outcome	SVR	No	No	105 ptns	70/30	Internal	MAE 11.6°	No	No	No
Simmons et al[31]	External validation of CDST	N/A	No	No	243 ptns	N/A	External	10% worse to 31.6% better	Yes	Yes	No
Eghbali et al[32]	Prediction of glenohumeral joint forces	CNN	No	No	959 synthetic ptns	85/15	Internal	MAE 11.1 N	No	No	No

*Studies were categorized based on source data stemming from radiological images (i.e. imaging study) or patient (ptn) data (i.e. tabular study).
ACC, accuracy; ANN, artificial neural network; AUC, area under curve; CDST, clinical decision support tools; CNN, convolutional neural network; MAE, mean absolute error; N/A, not available; postop, postoperative; RF, random forest; SVM, support vector machine; SVR, support vector regression; WD, wide and deep; XGB, extreme gradient boosting.

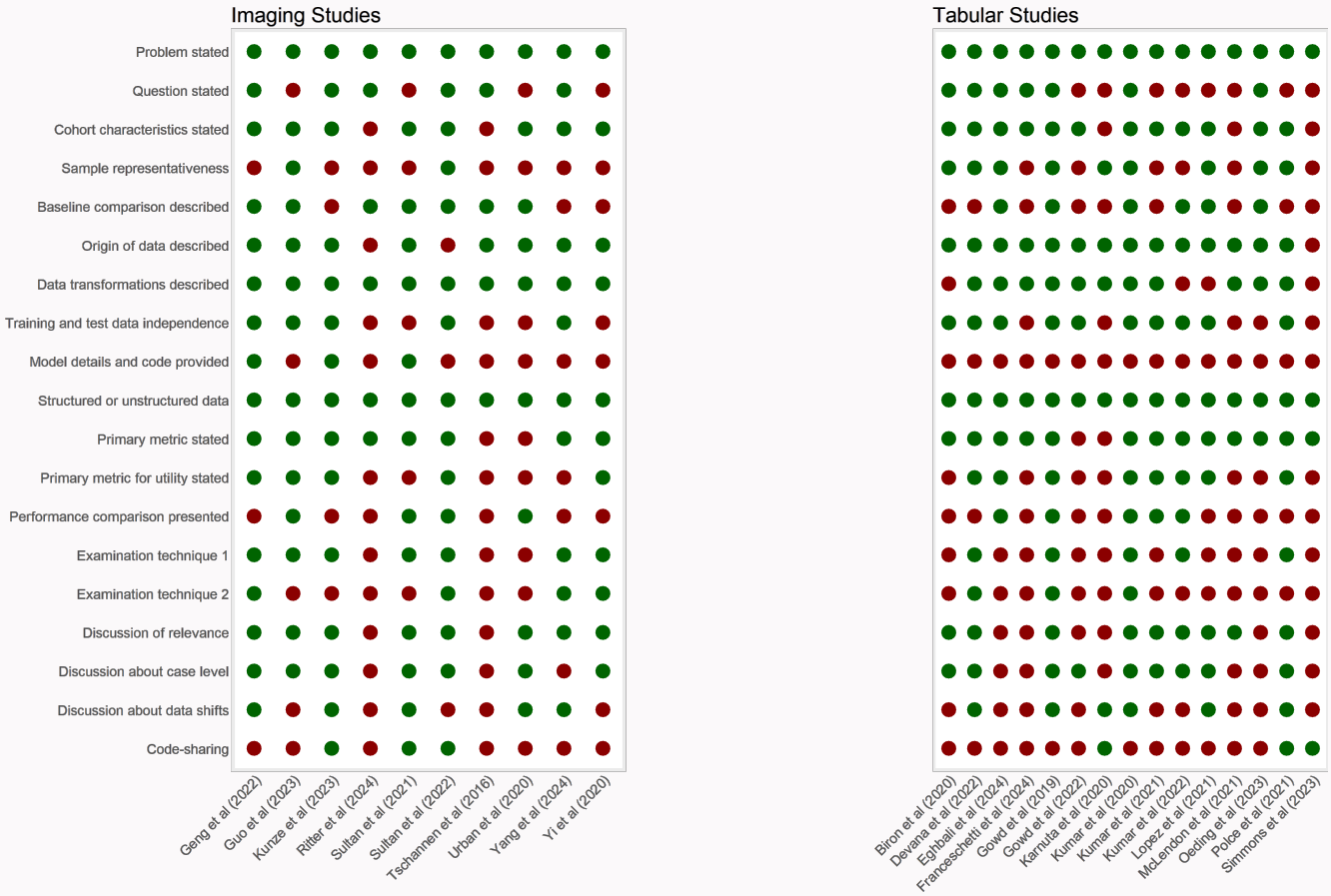

Fig. 2
Traffic light plots highlighting the methodological quality of all 25 eligible studies based on the Minimum Information about CLinical Artificial Intelligence Modeling (MI-CLAIM) checklist. All studies were categorized as either imaging (left) or tabular (right) studies based on data sourced from radiological images and patient data, respectively. Each row lists a specific criterion of the checklist, while each column corresponds to an individual study. Green dots indicate that the study met the specified criterion, while red dots signify that the study did not adhere to the indicated criterion.

16 predictive variables and comparing different ML algorithms. The best performing algorithm was the support vector machine, achieving an AUC of 0.8 in detecting patient satisfaction in a binary fashion.

Financial implications

Gowd et al's[27] study applied a combination of ML techniques to predict total healthcare costs during the 90-day perioperative period. Findings based on the dataset of 49,354 cases showed an average initial surgery cost of USD $19,364. Cases were considered high-cost if they exceeded USD $32,883,

which represents more than one SD above the average cost. Gradient boosting trees were the most effective predictive model, showing an AUC of 0.85.

Karnuta et al[19] used ANNs to predict the length of stay, discharge dispositions, and inpatient charges after TSA, with AUCs ranging from 0.71 to 0.89, using a publicly available dataset with sociodemographic data and geographical location comprising 111,147 patients.

Biomechanics

Eghbali et al[32] developed a ML model to predict glenohumeral joint forces in TSA, using a dataset of 959 virtual subjects derived from clinical registries and musculoskeletal models. The fully connected deep-learning model exhibited high accuracy in predicting joint forces across various abduction angles. The model achieved coefficients of determination between 0.97 and 0.98.

Discussion

The aim of this systematic review was to evaluate the effectiveness of ML algorithms in TSA, specifically those used for predicting clinical outcomes and image analysis.

Among the studies included in our review that utilize imaging data, the majority employ deep CNNs (Table I). This preference is understandable, as these algorithms do not rely on manually engineered features, instead performing feature extraction autonomously. However, achieving optimal performance with deep-learning algorithms such as CNNs typically requires large datasets, unless a pre-trained model is used. In this case, the pre-trained model can be adapted to the available training data and still perform well, even if the available training data are sparse. Another advantage of transfer learning is faster training, as most of the learning has already been done before.[33] Most of the studies included in this review that used images take advantage of transfer learning, especially studies with CNNs. In addition, it is easier to perform data augmentation with images than with tabular data, which can easily be done by any combination of rotating and cropping the images or adjusting brightness and contrast, greatly increasing the size of the dataset used to train the model, resulting in better performance. On the other hand, deep learning can be computationally intensive and may not be possible without the right hardware. In the studies included in this review that utilize tabular data, deep learning is notably less prevalent. This is expected, as shallow ML algorithms such as random forests and XGBoost have demonstrated superior performance compared to deep-learning algorithms when applied to the same tabular datasets.[34–36] These models do not require large datasets, a long training time, or high computational power either, which makes them more accessible to use. However, data augmentation proved to be more complicated in tabular data. Unlike augmenting images, where rotations or other image transformations conserve the underlying meaning, new data points in a tabular dataset cannot simply be generated without potentially changing the underlying relationships of the features in the dataset or creating unrealistic combinations. This might explain why none of the studies in our review that used tabular datasets employed data augmentation.

Our review demonstrated a significant gap in the transparency and replicability of studies employing ML techniques in the field of TSA. The lack of detailed model reporting, data, and code availability in both imaging and tabular studies inherits a risk of compromising scientific rigour in limiting the ability to reproduce and verify results. Further, the lack of external validation in many studies is a notable issue, raising doubts about the applicability of these models in different clinical environments. The widespread absence of code sharing highlights a significant challenge in the field of orthopaedic surgery, where concerns over proprietary rights and issues surrounding open science can impede collaborative advancement and innovation in ML applications.

ML models possess the potential for numerous clinical applications in the field. For instance, they might become important in determining whether to do inpatient or outpatient TSA.[20,22] Models recommending non-home discharge (e.g. rehabilitation facilities) play an important role in ensuring appropriate postoperative care. They can be particularly beneficial for patients at risk for complications or requiring more extensive postoperative care. This approach not only protects patient health but also optimizes the use of healthcare resources, reducing the number of preventable hospital readmissions. The ability to predict patient satisfaction based on preoperative characteristics could also significantly impact decisions regarding the suitability of an outpatient surgery, ultimately improving patient satisfaction and results.[25]

Moreover, predicting potential complications and unplanned readmissions can mitigate unforeseen costs, leading to more financially stable operations within healthcare institutions and potentially facilitating better negotiations with insurance providers.

A critical endpoint that most studies have overlooked is pain. In fact, only two publications focused on building a ML model for predicting pain following TSA,[21,23] which most likely is explained by the fact that pain is multifactorial and highly subjective. Future initiatives should focus on filling this gap, perhaps by allowing healthcare practitioners to conduct a comprehensive assessment of outcomes, including postoperative pain.

Recent research utilizing ML techniques has heavily focused on image analysis for implant detection using deep learning. While these models are technically sophisticated, their practical utility in clinical settings is limited due to the narrow variety of implants used. For instance, the variability in the types of implants used makes these models applicable only to providers who use the exact same selection of implants. Identifying conditions such as fractures, dislocations, and infections would be a more clinically valuable application. In addition, broadening ML to cover the prediction of a wider spectrum of postoperative outcomes could be of extra value. These outcomes can include prediction of complications, postoperative pain, patient satisfaction, and, to a lesser extent, postoperative function.

Several limitations should be considered before interpreting the results of this systematic review. One limitation is that only one critical appraisal tool was used to quantify the quality of studies. Furthermore, screening, data extraction, and critical appraisal were only conducted in a single fashion without control by another reviewer to offset potential mistakes; areas of inconsistency were, however, discussed among the authors.

For clinicians, effectively integrating ML can transform the field of TSA by streamlining decision-making processes, crafting personalized treatment plans, managing patient expectations, and enhancing the accuracy of predicting patient outcomes. However, to fully realize these benefits it is important to establish a robust validation framework. This approach is foundational to advancing personalized medicine, which aims to tailor healthcare strategies to individual patient profiles, optimizing treatment efficacy and patient satisfaction.

References

1. **Padash S**, **Mickley JP**, **Vera Garcia DV**, **et al**. An overview of machine learning in orthopedic surgery: an educational paper. *J Arthroplasty*. 2023;38(10):1938–1942.
2. **Levin JM**, **Lorentz SG**, **Hurley ET**, **et al**. Artificial intelligence in shoulder and elbow surgery: overview of current and future applications. *J Shoulder Elbow Surg*. 2024;33(7):1633–1641.
3. **Kunze KN**, **Krivicich LM**, **Clapp IM**, **et al**. Machine Learning algorithms predict achievement of clinically significant outcomes after orthopaedic surgery: a systematic review. *Arthroscopy*. 2022;38(6):2090–2105.
4. **Gupta P**, **Haeberle HS**, **Zimmer ZR**, **Levine WN**, **Williams RJ**, **Ramkumar PN**. Artificial intelligence-based applications in shoulder surgery leaves much to be desired: a systematic review. *JSES Rev Rep Tech*. 2023;3(2):189–200.
5. **Page MJ**, **McKenzie JE**, **Bossuyt PM**, **et al**. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021; 372:n71.
6. **Norgeot B**, **Quer G**, **Beaulieu-Jones BK**, **et al**. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26(9):1320–1324.
7. **Wickham H**. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.
8. **Tschannen M**, **Vlachopoulos L**, **Gerber C**, **Székely G**, **Fürnstahl P**. Regression forest-based automatic estimation of the articular margin plane for shoulder prosthesis planning. *Med Image Anal*. 2016;31:88–97.
9. **Urban G**, **Porhemmat S**, **Stark M**, **Feeley B**, **Okada K**, **Baldi P**. Classifying shoulder implants in X-ray images using deep learning. *Comput Struct Biotechnol J*. 2020;18:967–972.
10. **Yi PH**, **Kim TK**, **Wei J**, **et al**. Automated detection and classification of shoulder arthroplasty models using deep learning. *Skeletal Radiol*. 2020;49(10):1623–1632.
11. **Sultan H**, **Owais M**, **Park C**, **Mahmood T**, **Haider A**, **Park KR**. Artificial intelligence-based recognition of different types of shoulder implants in X-ray scans based on dense residual ensemble-network for personalized medicine. *J Pers Med*. 2021;11(6):482.
12. **Sultan H**, **Owais M**, **Choi J**, **et al**. Artificial intelligence-based solution in personalized computer-aided arthroscopy of shoulder prostheses. *J Pers Med*. 2022;12(1):109.
13. **Kunze KN**, **Jang SJ**, **Li TY**, **et al**. Artificial intelligence for automated identification of total shoulder arthroplasty implants. *J Shoulder Elbow Surg*. 2023;32(10):2115–2122.
14. **Geng EA**, **Cho BH**, **Valliani AA**, **et al**. Development of a machine learning algorithm to identify total and reverse shoulder arthroplasty implants from X-ray images. *J Orthop*. 2023;35:74–78.
15. **Guo D**, **Liu X**, **Wang D**, **Tang X**, **Qin Y**. Development and clinical validation of deep learning for auto-diagnosis of supraspinatus tears. *J Orthop Surg Res*. 2023;18(1):426.
16. **Ritter D**, **Denard PJ**, **Raiss P**, **Wijdicks CA**, **Bachmaier S**. Preoperative 3-dimensional computed tomography bone density measures provide objective bone quality classifications for stemless anatomic total shoulder arthroplasty. *J Shoulder Elbow Surg*. 2024;33(7):1503–1511.
17. **Yang L**, **Oeding JF**, **de Marinis R**, **Marigi E**, **Sanchez-Sotelo J**. Deep learning to automatically classify very large sets of preoperative and postoperative shoulder arthroplasty radiographs. *J Shoulder Elbow Surg*. 2024;33(4):773–780.
18. **Gowd AK**, **Agarwalla A**, **Amin NH**, **et al**. Construct validation of machine learning in the prediction of short-term postoperative complications following total shoulder arthroplasty. *J Shoulder Elbow Surg*. 2019;28(12):e410–e421.
19. **Karnuta JM**, **Churchill JL**, **Haeberle HS**, **et al**. The value of artificial neural networks for predicting length of stay, discharge disposition, and inpatient costs after anatomic and reverse shoulder arthroplasty. *J Shoulder Elbow Surg*. 2020;29(11):2385–2394.
20. **Biron DR**, **Sinha I**, **Kleiner JE**, **et al**. A novel machine learning model developed to assist in patient selection for outpatient total shoulder arthroplasty. *J Am Acad Orthop Surg*. 2020;28(13):e580–e585.
21. **Kumar V**, **Roche C**, **Overman S**, **et al**. What is the accuracy of three different machine learning techniques to predict clinical outcomes after shoulder arthroplasty? *Clin Orthop Relat Res*. 2020;478(10):2351-2363.
22. **Lopez CD**, **Constant M**, **Anderson MJJ**, **Confino JE**, **Heffernan JT**, **Jobin CM**. Using machine learning methods to predict nonhome discharge after elective total shoulder arthroplasty. *JSES Int*. 2021;5(4): 692–698.
23. **Kumar V**, **Roche C**, **Overman S**, **et al**. Using machine learning to predict clinical outcomes after shoulder arthroplasty with a minimal feature set. *J Shoulder Elbow Surg*. 2021;30(5):e225–e236.
24. **McLendon PB**, **Christmas KN**, **Simon P**, **et al**. Machine learning can predict level of improvement in shoulder arthroplasty. *JB JS Open Access*. 2021;6(1):e20.00128.
25. **Polce EM**, **Kunze KN**, **Fu MC**, **et al**. Development of supervised machine learning algorithms for prediction of satisfaction at 2 years following total shoulder arthroplasty. *J Shoulder Elbow Surg*. 2021;30(6): e290–e299.
26. **Devana SK**, **Shah AA**, **Lee C**, **et al**. Development of a machine learning algorithm for prediction of complications and unplanned readmission following primary anatomic total shoulder replacements. *J Shoulder Elb Arthroplast*. 2022;6:24715492221075444.
27. **Gowd AK**, **Agarwalla A**, **Beck EC**, **et al**. Prediction of total healthcare cost following total shoulder arthroplasty utilizing machine learning. *J Shoulder Elbow Surg*. 2022;31(12):2449–2456.
28. **Kumar V**, **Schoch BS**, **Allen C**, **et al**. Using machine learning to predict internal rotation after anatomic and reverse total shoulder arthroplasty. *J Shoulder Elbow Surg*. 2022;31(5):e234–e245.
29. **Oeding JF**, **Lu Y**, **Pareek A**, **et al**. Understanding risk for early dislocation resulting in reoperation within 90 days of reverse total shoulder arthroplasty: extreme rare event detection through cost-sensitive machine learning. *J Shoulder Elbow Surg*. 2023;32(9):e437–e450.
30. **Franceschetti E**, **Gregori P**, **De Giorgi S**, **et al**. Machine learning can predict anterior elevation after reverse total shoulder arthroplasty: a new tool for daily outpatient clinic? *Musculoskelet Surg*. 2024;108(2):163–171.
31. **Simmons C**, **DeGrasse J**, **Polakovic S**, **et al**. Initial clinical experience with a predictive clinical decision support tool for anatomic and reverse total shoulder arthroplasty. *Eur J Orthop Surg Traumatol*. 2024;34(3):1307–1318.
32. **Eghbali P**, **Becce F**, **Goetti P**, **Büchler P**, **Pioletti DP**, **Terrier A**. Glenohumeral joint force prediction with deep learning. *J Biomech*. 2024;163:111952.
33. **Prinzi F**, **Currieri T**, **Gaglio S**, **Vitabile S**. Shallow and deep learning classifiers in medical image analysis. *Eur Radiol Exp*. 2024;8(1):26.
34. **Ram Kumar RP**, **Polepaka S**. Performance comparison of random forest classifier and convolution neural network in predicting heart diseases. International Conference on Computational Intelligence and Informatics; 2018, Hyderabad, India
35. **Shwartz-Ziv R**, **Armon A**. Tabular data: deep learning is not all you need. *Inf Fus*. 2022;81:84–90.
36. **Grinsztajn L**, **Varoquaux G. OE**. Why do tree-based models still outperform deep learning on typical tabular data?. *Adv Neural Inf Process Syst*. 2022;35:507–520.

Author information

T. Schneller, MSc, Senior Data Manager
P. Moroder, MD, Senior Physician

Department for Shoulder and Elbow Surgery, Schulthess Clinic, Zurich, Switzerland.

M. Kraus, MD, Resident, External Cooperator, Department for Shoulder and Elbow Surgery, Schulthess Clinic, Zurich, Switzerland; Department of Traumatology, University Hospital Zurich, Zurich, Switzerland.

J. Schätz, MSc, Physical Therapist, Research Assistant, Department for Shoulder and Elbow Surgery, Schulthess Clinic, Zurich, Switzerland; Institute for Therapies and Rehabilitation, Cantonal Hospital Winterthur, Winterthur, Switzerland.

M. Scheibel, MD, Chief Physician, Visiting Professor, Department for Shoulder and Elbow Surgery, Schulthess Clinic, Zurich, Switzerland; Center for Musculoskeletal Surgery, Charité-Universitaetsmedizin, Berlin, Germany.

A. Lazaridou, PhD, Research Group Head (Upper Extremity), Assistant Professor, Department for Shoulder and Elbow Surgery, Schulthess Clinic, Zurich, Switzerland; Department of Anesthesiology, Brigham & Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA.

Author contributions

T. Schneller: Data curation, Formal analysis, Investigation, Project administration, Writing – original draft, Methodology.
M. Kraus: Formal analysis, Investigation, Writing – original draft, Methodology.
J. Schätz: Investigation, Visualization, Writing – review & editing.
P. Moroder: Writing – review & editing.
M. Scheibel: Writing – review & editing.
A. Lazaridou: Conceptualization, Methodology, Supervision, Writing – review & editing.

M. Scheibel and A. Lazaridou contributed equally to this work.

T. Schneller and M. Kraus contributed equally to this work.

T. Schneller and M. Kraus are joint first authors.

M. Scheibel and A. Lazaridou are joint senior authors.

Funding statement

The author(s) disclose receipt of the following financial or material support for the research, authorship, and/or publication of this article: The Schulthess Clinic funded the open access fee for this article.

ICMJE COI statement

P. Moroder reports royalties or licenses and consulting fees from Arthrex and Medacta, which are unrelated to this work. M. Scheibel declares grants or contracts from Stryker and Smith & Nephew; royalties or licenses, consulting fees, and payment or honoraria for lectures, presentations, speakers bureaus, manuscript writing, or educational events from Smith & Nephew; and being general secretary of the DACH Shoulder Elbow Society, all of which are also unrelated.

Data sharing

All data generated or analyzed during this study are included in the published article.

Acknowledgements

The authors would like to thank Melissa Wilhelmi, PhD (medical writer at Schulthess Clinic, Zurich, Switzerland) for manuscript copyediting and proofreading.

Open access funding

The authors report that the Schulthess Clinic, Zurich, Switzerland, funded the open access fee for this article.