Artificial Learning and Machine Learning Decision Guidance
Applications in Total Hip and Knee Arthroplasty: A Systematic Review

Cesar D. Lopez, BS *, Anastasia Gazgalis, BS, Venkat Boddapati, MD, Roshan P. Shah, MD, H. John Cooper, MD, Jeffrey A. Geller, MD

New York Presbyterian/Columbia University Irving Medical Center, New York, NY

ARTICLE INFO

Article history:
Received 23 November 2020
Received in revised form 17 July 2021
Accepted 26 July 2021
Available online 3 September 2021

Keywords:
Machine learning
Artificial intelligence
Deep learning
Artificial neural networks
Orthopedic surgery
Hip and knee arthroplasty

ABSTRACT

Background: Artificial intelligence (AI) and machine learning (ML) modeling in hip and knee arthroplasty (total joint arthroplasty [TJA]) is becoming more commonplace. This systematic review aims to quantify the accuracy of current AI- and ML-based application for cognitive support and decision-making in TJA.

Methods: A comprehensive search of publications was conducted through the EMBASE, Medline, and PubMed databases using relevant keywords to maximize the sensitivity of the search. No limits were placed on level of evidence or timing of the study. Findings were reported according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. Analysis of variance testing with post-hoc Tukey test was applied to compare the area under the curve (AUC) of the models.

Results: After application of inclusion and exclusion criteria, 49 studies were included in this review. The application of AI/ML-based models and average AUC is as follows: cost prediction-0.77, LOS and discharges-0.78, readmissions and reoperations-0.66, preoperative patient selection/planning-0.79, adverse events and other postoperative complications-0.84, postoperative pain-0.83, postoperative discharges-0.78, readmissions and reoperations-0.66, preoperative patient selection/planning-0.79, adverse events and other postoperative complications-0.84, postoperative patient-reported outcomes measures and functional outcome-0.81. Significant variability in model AUC across the different decision support applications was found (P < .001) with the AUC for readmission and reoperation models being significantly lower than that of the other decision support categories.

Conclusions: AI/ML-based applications in TJA continue to expand and have the potential to optimize patient selection and accurately predict postoperative outcomes, complications, and associated costs. On average, the AI/ML models performed best in predicting postoperative complications, pain, and patient-reported outcomes and were less accurate in predicting hospital readmissions and reoperations.

© 2021 The Authors. Published by Elsevier Inc. on behalf of The American Association of Hip and Knee Surgeons. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction and background

Health care has been able to harness artificial intelligence (AI) in an effort to create predictive tools that support clinicians in more complicated decision-making processes. Specifically, a subset of AI known as machine learning (ML) has been used in multiple medical specialties including oncology, neurology, neurosurgery, cardiology, and orthopedic surgery [1-5]. In lower extremity surgery, specifically, ML has been applied in risk assessment and diagnosis, cost analysis, and reimbursement tools [6].

In the simplest form, ML produces useful models from algorithmic analysis of acquired data. One particular strength of ML is that some models are trained on large amounts of information, or “big data”. This optimizes the use of these models in decision-making as the more data the underlying model is trained on, the more accurate its predictions become [7]. Several ML models are used including decision trees, support vector machine, regression analysis, and Bayesian networks [4,5]. In addition, a subset of ML called “deep learning” has been developed, which includes artificial neural network (ANN) models. The advantage of these models is that they do not require the preprocessing of data by humans, but rather can analyze the raw inputs and identify which features are most important for the analysis [8].

Models can be created in several ways; however, supervised learning is the most common one [8,9]. In supervised learning, data sets are labeled so that the model can be built on a “training set” of inputs (variables of interests) with defined outputs (outcomes of interest). Complex patterns and relationships are identified between these inputs and outputs, and the model can then use these...
associations to predict outcomes of interest from novel inputs [10,11]. Once a model is created, it can be tested on a novel data set or a validation data set.

Recently, there has been a substantial increase in the literature describing the use of these models, including the field of hip and knee arthroplasty. As such, it is critical to build a strong understanding of the accuracy and application of these current models to guide their applicability toward further development in the future. The purpose of this review is to assess the accuracy of current applications of AI/ML in hip and knee arthroplasty, namely in (1) administrative/clinical decision support applications (cost, discharge/length of stay (LOS), patient selection and planning, readmission and reoperation risk) and (2) postoperative prediction/management applications (adverse event) other postoperative complication, cardiovascular complication, postoperative pain, postoperative mortality, patient-reported outcomes [PRO], and sustained opioid use.

Material and methods

Search strategy

A comprehensive search of publications, through May 2020, was conducted using the EMBASE, Medline, and PubMed databases in accordance with Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. The search strategy included the following keywords or MeSH-terms: “machine learning”, “artificial intelligence”, “deep learning”, “neural network”, “artificial neural networks”, “support vector machine”, “Bayesian”, “boosting”, “ensemble learning”, “prediction model”, “decision tree”, “random forest”, “total hip arthroplasty”, “total knee arthroplasty”, “total joint arthroplasty”, “THA”, “TKA”, “[TJA]”, “hip replacement”, “knee replacement”, and “joint replacement”. Boolean operators (OR, AND) were used to maximize the sensitivity of the search. Screening of reference lists of retrieved articles also yielded additional studies.

Eligibility criteria

Inclusion criteria comprised original clinical studies, including studies which evaluate AI/ML-based applications in clinical decision-making in hip and knee arthroplasty. Exclusion criteria comprised studies that did not evaluate AI/ML applications in hip and knee arthroplasty, medical imaging analysis studies without explicit reference or application to hip and knee arthroplasty, studies with nonhuman subjects, non-English-language studies, inaccessible articles, conference abstracts, reviews, and editorials. No limits were placed on level of evidence or timing of the study because the majority of the reviewed studies were published within the last 10 years.

Study selection

Article titles and abstracts were screened initially by 2 reviewers, and full-text articles were subsequently screened based on the selection criteria. The studies were rated by their level of evidence, based on the Oxford Center for Evidence-based Medicine Levels of Evidence [12]. Two authors reviewed each individual article that was included. Discrepancies in inclusion studies were discussed and resolved by consensus.

Data extraction and categorization

A database was generated from all included studies which consisted of the journal of publication, publication year, country of origin, study design, level of evidence, study duration, blinding of the study, number of involved institutions, AI/ML methods and clinical applications, surgical domain, data sources, input variables and output variables, sample size, average patient age, percent female patients, and any additional pertinent findings from the study. The reviewed articles were sorted into different, nonmutually exclusive categories based on the AI/ML clinical application. AI/ML clinical applications were divided into 2 major groups: (1) administrative and clinical decision support and (2) postoperative prediction and management of complications and outcomes. The former group contained the following prediction and optimization subcategories: preoperative planning and cost prediction, hospital discharge and LOS, readmissions, and reoperations. The latter group included postoperative cardiovascular complications, other complications, mortality, and functional and clinical outcomes.

Data analysis

Descriptive statistics were used to summarize important findings and results from the selected articles and to describe trends in AI/ML techniques, clinical applications, and relevant findings associated with its use. Summary data were presented using simple averages, frequencies, and proportions. This study did not evaluate R2 values. AI/ML model performance within the reviewed studies were summarized using various metrics, including the area under the curve (AUC) of receiver operating characteristic curves, accuracy (%), sensitivity (%), and specificity (%). AUC values range from 0.50 to 1 and measure a prediction models’ discriminative ability, with a higher AUC value signifying better predictive ability and overall accuracy of the model correctly placing a patient into an outcome category. A model with an AUC of 1.0 is a perfect discriminator, 0.90 to 0.99 is considered excellent, 0.80 to 0.89 is good, 0.70 to 0.79 is fair, and 0.51 to 0.69 is considered poor [13]. Reported model performance metrics for each AI/ML algorithm type and for each clinical application category were aggregated across the reviewed studies. One-way analysis of variance (ANOVA) with post hoc Tukey tests were performed, with statistical significance set at P < .05. All statistical analyses were performed using Stata (version 16.1; Stata Corporation, College Station, Texas).

Results

Search results and study selection

Predefined search terms resulted in 307 articles, of which 48 duplicate articles were removed. The remaining 259 articles were screened by title and abstract according to inclusion and exclusion criteria. Ultimately, there were 58 articles included for full review, of which 49 met full inclusion and exclusion criteria ([Fig. 1]). Level of evidence of the reviewed studies ranged from II to IV, and over 61% of studies had level of evidence III, 29% of studies had level of evidence II, and 10% had level of evidence IV. The average number of patients included in model testing was 30,624 (standard deviation [SD] 69,069). Although there were no limitations on publication dates in the selection process, the vast majority of studies (42 studies, or 87.5%) were published during the last 3 years (2018 - 2020) ([Fig. 2]). There was variability in the metrics used by authors to report or evaluate AI/ML model performance. AUC was the most frequently reported performance metric, appearing in 39 out of the 49 total reviewed studies (79.6%). In comparison, accuracy was reported less frequently (10 studies, 20.4%), as were sensitivity and specificity (9 studies, or 18.4%).
Administrative and clinical decision support applications

A total of 31 reviewed studies (63.3%) evaluated the use of AI/ML applications in optimizing preoperative patient selection or projecting surgical costs, through prediction of hospital LOS, discharges, readmissions, and other cost-contributing factors (Tables 1 and 2). Sixteen studies (32.7%) evaluated AI/ML applications to accurately predict patient reoperations, operating time, hospital LOS, discharges, readmissions, or surgical and inpatient costs [14-29]. In addition, 16 studies (32.7%) used patients’ preoperative risk factors and other patient-specific variables to optimize the patient selection and surgical planning process through the use of AI/ML-based predictions of surgical outcomes and postoperative complications [30-44]. The majority of the decision support studies evaluated AI/ML model performance using receiver operating characteristic/AUC, accuracy, sensitivity, and specificity. Two studies did not test model performance, but instead used cluster analysis to classify patients based on preoperative risk factors and other variables to predict their association with inpatient costs and functional outcomes [20,45].

AI/ML applications of cost prediction were used in 23 models across 11 studies, which reported an average AUC of 0.77 (SD 0.08) (Table 3). Predictive models of LOS and discharges were used in 6 studies, with an average AUC of 0.78 (SD 0.05) across 11 models. Six studies each evaluated different AI/ML-based predictive models of readmissions and reoperations, with an average AUC of 0.66 (SD 0.04) across 15 models. Applications of preoperative patient selection/planning were used in 62 models across 16 studies, reporting an average AUC of 0.79 (SD 0.11). ANOVA testing found statistically significant variability in model AUC across the different decision support applications (P < .001), and Tukey post-hoc testing confirmed that AI/ML predictive models of readmissions and reoperations reported significantly lower AUC than each of the other administrative and clinical decision support categories (Table 3).

ANOVA testing found statistically significant variability in model accuracy and specificity (P < .001 and P = .026, respectively), and Tukey post hoc testing confirmed statistically significant intergroup differences in those metrics, shown in Table 3. Preoperative planning and patient selection models reported significantly higher average accuracy (95.4%) than each of the other decision support categories, including cost prediction (86.5%; P = .006), discharge/LOS (85.2%; P < .001), and readmissions/reoperations (80.1%; P < .001). Discharge/LOS prediction models had the lowest specificity (72.1%) which was significantly lower than specificity reported for...
preoperative planning and patient selection models (94.6%; \( P < .001 \)) and readmissions/reoperations models (98.3%; \( P < .001 \)). Conversely, there were no significant differences in model sensitivity between the applications (\( P = .866 \)).

**Prediction and management of postoperative outcomes and complications**

A total of 25 reviewed studies (51.0%) (Table 4) used various AI/ML models to predict outcomes, complications, and adverse events, including postoperative risk of cardiac complications, pulmonary complications, renal complications, venous thromboembolism, blood transfusion, periprosthetic and surgical site infections, vomiting, sustained opioid use, and mortality (ranging from 30 to 90 days) (Table 2). Postoperative prediction categories were sorted into 5 groups based on application: adverse events and other complications, cardiovascular complications, postoperative pain, postoperative mortality, PROs and other functional outcomes, and sustained opioid use (Table 2).

Model performance significantly varied across postoperative management applications, which was confirmed by ANOVA testing for average AUC and sensitivity values (\( P = .002 \) and \( P = .042 \), respectively) (Table 5). Models predicting adverse events and other postoperative complications averaged 0.84 (SD 0.10, 14 models), models predicting postoperative cardiovascular complications averaged an AUC of 0.77 (SD 0.08, 8 models), models predicting postoperative pain models averaged an AUC of 0.83 (SD 0.05, 10 models), postoperative mortality models averaged an AUC of 0.81 (SD 0.07, 3 models), and postoperative PRO and functional outcome models averaged an AUC of 0.81 (SD 0.08, 56 models). Tukey post hoc testing found statistically significant differences between postoperative sustained opioid use models (average AUC of 0.71) and models predicting adverse events/other complications (AUC 0.84; \( P = .002 \)), postoperative pain (AUC 0.83; \( P = .003 \)), and PROs/functional outcomes (AUC 0.81; \( P = .011 \)) (Table 5). Average sensitivity was also found to be significantly different between adverse event/other complication models (97.7%) and postoperative pain (78.8%; \( P < .001 \)) and PROs/functional outcome models (76.9%; \( P < .001 \)) (Table 5). There was no significant variation in reported accuracy or specificity values (\( P = .279 \) and \( P = .167 \), respectively) (Table 5).

**Comparison of AI/ML algorithms**

Various AI/ML algorithms were used in the reviewed studies, including ANNs, decision trees (including random forest), logistic regressions, gradient boosting/ensemble learning, and Bayesian networks (Tables 2 and 6). The most commonly applied AI/ML algorithms in the reviewed studies were logistic regression (24 studies, 60.0%) and neural networks (23 studies, 46.9%). Decision trees were used in 16 studies (32.7%), boosting/ensemble learning models were used in 11 studies (22.4%), support vector machines were used in 7 studies (14.3%), Bayesian networks were used in 5 studies (10.2%), and cluster analysis was only included in 2 studies (4.1%).

Across all ML types, ANNs and Bayesian networks each had the highest average AUC (0.81; SD 0.11 across 56 models and SD 0.07 across 8 models, respectively) Boosting/ensemble learning models had average AUC of 0.79 (SD 0.07, 19 models), followed by decision tree models (AUC 0.78, SD 0.10, 41 models), regression models (AUC 0.77, SD 0.07, 62 models), and support vector machines (AUC 0.77, SD 0.11, 26 models) (Table 6). When comparing AI/ML model performance across various algorithm types, one-way ANOVA testing did not find statistically significant variation, except for specificity (\( P = .019 \)). No significant intergroup differences were found for any performance metrics on Tukey post-hoc testing (Table 6).

**Training data sets**

For several studies, data sets used for training were extracted from large national and multicenter databases (Tables 1 and 4). The most commonly used were the Medicare databases (5 studies). Other administrative and private insurance databases were also
| Author, year | Pathology | ML algorithms | Prediction outputs | Patients in testing set (n) | Avg. age | %Female | Data source |
|-------------|-----------|---------------|-------------------|-----------------------------|---------|---------|-------------|
| Alam et al., 2019 [14] | THA | ANN, regression | Costs | 10,000 | — | — | Multicenter |
| Aram et al., 2018 [15] | TKA | ANN, decision tree | Readmissions/reoperation | 6137 | 70.2 | 57.1 | National Joint Registry (UK) |
| Bonakdari et al., 2020 [16] | TKA, THA | ANN | Readmissions/reoperation | 5251 | 60.3 | 66.3 | UK National Institute for Health and Care Excellence (NICE) |
| Borjali et al., 2020 [30] | THA | ANN | Preop patient selection/planning | 25 | 61.3 | 47 | Single institution |
| Cafri et al., 2019 [31] | TKA | Decision tree, regression | Preop patient selection/planning | 74,520 | 65.5 | 57.4 | Single institution |
| Fontana et al., 2019 [32] | TJA | Regression, SVM, decision tree | Preop patient selection/planning | 3430 | 63 | — | Patient database |
| Gabriel et al., 2019 [17] | THA | Regression, decision tree | Discharge/LOS | 240 | — | 50.5 | Single institution |
| Hirvasnemi et al., 2019 [33] | THA | Regression | Preop patient selection/planning | 197 | 55.7 | 83.8 | CHECK cohort |
| Hyer et al., 2019 [18] | TJA | Regression, decision tree | Preop patient selection/planning | 262,290 | 73 | 55.8 | Medicare |
| Hyer et al., 2020 [19] | All | Decision tree | Costs, readmissions/reoperation | 262,290 | 73 | 55.8 | Medicare |
| Hyer et al., 2020 [20] | THA, TKA | Cluster analysis | Preop patient selection/planning | 19,522 | — | — | Medicare |
| Jafarzadeh et al., 2020 [44] | TKA | ANN, regression | Preop patient selection/planning | 2357 | 61.6 | 62 | Multicenter |
| Jodeiri et al., 2020 [34] | THA | ANN | Preop patient selection/planning | 95 | — | — | Single institution |
| Jones et al., 2019 [21] | THA, TKA | Regression, boosting | Readmissions/reoperation | — | — | — | Medicare |
| Kang et al., 2020 [35] | THA | ANN | Preop patient selection/planning | 1202 | — | — | Multicenter |
| Karnuta et al., 2019 [22] | Hip fracture | Bayesian | Costs, discharge/LOS | 98,562 | — | 73.5 | New York Statewide Planning and Research Cooperative System database |
| Karnuta et al., 2019 [46] | TJA | ANN | Costs | 73,901 | — | — | New York State inpatient administrative database |
| Lee et al., 2017 [26] | TJA | Regression | Readmissions/reoperation | 26 | — | — | Analysis of patient records to provide risk prediction for readmissions. |
| Lee et al., 2019 [27] | TJA | Boosting, regression Bayesian | Costs | 131 | — | — | Single institution |
| Navarro et al., 2018 [28] | TKA | Bayesian | Costs, discharge/LOS | 35,362 | — | — | Administrative database |
| Pareek et al., 2019 [37] | Knee fracture | ANN, decision tree, regression, boosting, Bayesian | Preop patient selection/planning | 62 | 64.6 | 68 | Single institution |
| Ramkumar et al., 2019 [23] | TKA | ANN | Costs, discharge/LOS | 175,042 | 73.5 | 64 | National inpatient sample |
| Ramkumar et al., 2019 [24] | THA | Bayesian | Costs, discharge/LOS | 30,584 | — | — | Patient database |
| Ramkumar et al., 2019 [25] | THA | ANN | Costs, discharge/LOS | 78,335 | 75.3 | 63.6 | National inpatient sample |
| Sheralatif et al., 2020 [38] | THA | ANN | Preop patient selection/planning | 78 | 63.1 | 47 | Single institution |
| Tiulpin et al., 2019 [39] | TKA | ANN, regression, boosting | Preop patient selection/planning | 3918 | 61.16 | 62.50 | Osteoarthritis Initiative (OAI) and MOST data sets |
| Tolpadi et al., 2020 [40] | TKA, THA | ANN | Preop patient selection/planning | 719 | 61 | 58 | OAI database |
| Twiggs et al., 2019 [41] | TKA | Bayesian | Preop patient selection/planning | 150 | 65.7 | 53 | Single institution |
| Van et al., 2019 [29] | THA | ANN | Costs, preop patient selection/planning | 100 | — | — | Single institution |
| Yi et al., 2019 [42] | TKA | ANN | Preop patient selection/planning | 154 | — | — | Single institution |
| Yoo et al., 2013 [43] | TKA | SVM | Preop patient selection/planning | — | 0 | 0 | Single institution |

ANN, artificial neural network; LOS, length of stay; ML, machine learning; SVM, support vector machine; THA, total hip arthroplasty; TJA, total joint arthroplasty; TKA, total knee arthroplasty.
used, including the Kaiser Permanente Total Joint Replacement Registry, the American College of Surgeons National Surgical Quality Improvement Program, National Inpatient Sample, and New York Statewide Planning and Research Cooperative System databases. Finally, some studies used training data sets comprising patients from multicenter or single-center cohorts.

**Discussion**

To our knowledge this systematic review is the first of its kind, evaluating the accuracy and reliability of AI/ML applications in hip and knee arthroplasties across 49 studies. The included studies investigated the role of AI/ML in clinical decision-making and surgical planning by optimizing patient selection and predicting cost and complication risks. AI/ML models performed best, average AUC > 0.8, when predicting postoperative adverse events and mortality, as well as postoperative pain and PROs. Deep learning/ANN models resulted in the highest average AUC and accuracy of all the model types and were presented in 47% of the studies.

There are multiple benefits of AI/ML-based predictive modeling in hip and knee arthroplasties. First, AI/ML-based model capable of predicting the need for surgery remains an important tool for surgeons, given increased cost-consciousness with health-care expenditures. Hip and knee arthroplasty typically involve an older and highly comorbid patient population, and these tools can be especially helpful in identifying patient-specific needs and risks within this population. Examples of how these models can enable providers to create and optimize personalized treatment plans include accurate identification of an implant from a previous surgery for revision procedures and classifying total knee arthroplasty (TKA) surgical candidates based on patient-specific risk factors [29-31,33-35,37-42,44,62]. Hyer et al. demonstrated an AI/ML model

### Table 2
Characteristics of AI/ML applications, including applied ML algorithms and prediction outputs.

| Administrative/clinical decision support applications | Applied ML algorithms | Prediction outputs |
|-------------------------------------------------------|-----------------------|--------------------|
| Costs                                                 | ANN, Bayesian, boosting, decision tree, regression, cluster analysis | Hospital charges, procedural costs, cost-effective interventions, payment, postoperative resource utilization |
| Discharge/LOS                                          | ANN, Bayesian, decision tree, regression | Discharge disposition, LOS |
| Preop patient selection/planning                       | ANN, Bayesian, boosting, decision tree, regression, SVM | Preop OA progression/prognosis, preop THA/TKA |
| Readmissions/reoperation                              | ANN, boosting, decision tree, regression | 30-d readmission, unplanned readmission, revision |

**Postoperative prediction/management applications**

| Adverse event/other complication                      | ANN, boosting, decision tree, regression, SVM | 90-d postoperative complications, any complication, periprosthetic joint infection, postoperative complications, postoperative vomiting, pulmonary complication, renal complication, surgical site infection |
| Cardiovascular complication                           | Decision tree, regression | Cardiac complication, risk of allogenic blood transfusion (ALBT) in primary lower limb, VTE |
| Postoperative pain                                    | ANN, boosting, decision tree, regression, SVM | Improvement in SF-36 pain score, VAS score, severe pain |
| Postoperative mortality                               | Decision tree, regression | 30-d mortality, 90-d mortality, death |
| PROMs/Outcomes                                        | ANN, boosting, decision tree, regression, SVM, cluster analysis | Hip OA at 8 y postoperatively, HOOS JR, Hip OA at 10 y postoperatively, KOOS JR, patient satisfaction, postoperative Q-score, postoperative functional outcomes, clinically meaningful improvement for the patient-reported health state, postoperative walking limitation, SF-36 MCS, SF-36 PCS, unfavorable outcomes |
| Sustained opioid use                                  | ANN, boosting, decision tree, regression, SVM | 90-d postoperative outcome-opioid use, postoperative sustained opioid use |

### Table 3
Statistical comparisons of reported model performance metrics, by administrative/clinical decision support application.

| Administrative/clinical decision support applications | Performance metrics: mean (SD, n) |
|-------------------------------------------------------|----------------------------------|
|                                                       | AUC | Accuracy | Sensitivity | Specificity |
| 1. Costs                                              | 0.77 (0.08, 23) | 86.5 (4.7, 4) | — | — |
| 2. Discharge/LOS                                       | 0.78 (0.05, 11) | 85.2 (3.2, 2) | 64.5 (—, 1) | 72.1 (—, 1) |
| 3. Preoperative patient selection/planning            | 0.79 (0.11, 62) | 95.4 (5.4, 10) | 70.1 (32.7, 9) | 94.6 (7.1, 9) |
| 4. Readmissions/reoperation                            | 0.66 (0.04, 15) | 80.1 (3.1, 3) | 81.8 (2.4, 2) | 98.3 (0.2, 2) |
| ANOVA                                                 | P < .001 | P < .001 | P < .001 | 2 vs 3 (P < .001) |
| Tukey Post Hoc Tests (stat. significant results)      | 4 vs 1 (P < .003) | 3 vs 1 (P < .006) | — | — |
|                                                       | 4 vs 2 (P < .005) | 3 vs 2 (P < .001) | — | — |
|                                                       | 4 vs 3 (P < .001) | 3 vs 4 (P < .001) | — | — |

ANOVA, analysis of variance; AUC, area under curve; LOS, length of stay; SD, standard deviation.
which classified TKA and total hip arthroplasty patients based on surgical complexity scores [19].

AI/ML models may aid in predicting postoperative complications and creating personalized postoperative management protocols to avoid or manage those obstacles and maximize outcomes. Several of the reviewed studies used AI/ML models to accurately predict the risk of a range of postoperative complications and adverse events [14,19,29,47,50,51,60]. TKA and total hip arthroplasty revisions and reoperations are also modeled with AI/ML algorithms in some studies, [15,16,21,64] as well as hospital readmissions [20,21,26,27]. In the postoperative period, AI/ML tools offer surgeons the ability to predict patients’ outcomes after surgery, including functional outcomes and PRO scores [14,32,33,43,45,48,53,54,57–59,61]. Postoperative pain has also been shown to be predicted with AI/ML [43,53,56,55] including identification of patients at high risk for prolonged postoperative opioid prescriptions. These tools may better inform analgesic and pain management protocols, especially for opioid prescriptions. They may also enable surgeons to better

Table 4
Reviewed studies of postoperative outcome prediction in hip and knee arthroplasty.

| Author, year       | Pathology | ML algorithms | Prediction outputs | Patients in testing set (n) | Avg. age | %Female | Data source                  |
|--------------------|-----------|---------------|--------------------|-----------------------------|----------|----------|------------------------------|
| Alam et al., 2019  | THA       | ANN, regression | PROs/outcomes      | 10,000                      | —        | —        | Multicenter                  |
| Bini et al., 2019  | TJA       | Cluster analysis | PROs/outcomes      | —                           | 63       | 68       | Single institution           |
| Fontana et al., 2019| TJA     | Regression, SVM, decision tree | PROs/outcomes | 2744                       | 63       | —        | Patient database             |
| Galivanche et al., 2019 | THA    | Boosting | Adverse event/other complication | 34,982                    | —        | —        | ACS-NSQIP database           |
| Gielis et al., 2020 | TJA     | Regression | PROs/outcomes | 1044                       | 55.9     | 87.3     | CHECK cohort                 |
| Gong et al., 2014  | TJA       | ANN, regression | Adverse event/other complication | —                         | 69.6     | 53.3     | Single institution           |
| Harris et al., 2019 | TJA     | Regression | Adverse event/other complication, cardiovascular complication, postoperative mortality | —                         | 65.7     | 59.4     | ACS-NSQIP database           |
| Hirvasnemi et al., 2019 | THA | Regression  | PROs/outcomes      | 197                         | 55.7     | 83.8     | CHECK cohort                 |
| Huang et al., 2018 | TJA, TKA  | Decision tree, regression | Cardiovascular complication | 3797                      | 62       | 66       | Multicenter                  |
| Huang et al., 2018 | TKA       | Decision tree  | Postoperative pain | —                          | —        | —        | Administrative database      |
| Huber et al., 2019 | THA       | Boosting, ANN, regression | Postoperative pain, PROs/outcomes | 31,905          | —        | 59.7     | NHS PRO data                |
| Hyer et al., 2019  | TKA, TKA  | Regression | Adverse event/other complication | 1,049,160        | —        | —        | Medicare                     |
| Hyer et al., 2020  | Ali       | Decision tree  | Adverse event/other complication, postoperative mortality | 524,580          | 73       | 55.8     | Medicare                     |
| Jacobs et al., 2016 | TKA     | Decision tree  | PROs/outcomes      | 325                         | 263      | 59       | 38.7                         | Single institution Multicenter |
| Kharhade et al., 2019 | THA | Boosting, decision tree, SVM, ANN, regression | Sustained opioid use | 2508                       | 67       | 60.3     | Single institution           |
| Katakam et al., 2020 | TKA | Decision tree, ANN, decision tree, SVM, regression, boosting | Sustained opioid use | 263                       | 59       | 38.7     | Single institution           |
| Kluge et al., 2018 | TKA       | Decision tree  | PROs/outcomes      | —                           | 64       | 66.666667 | Single institution           |
| Kunze et al., 2020 | THA       | ANN, decision tree, SVM, regression, boosting | PROs/outcomes | 183                         | 62       | 57.3     | Single institution           |
| Onsem et al., 2016 | TKA       | Regression | PROs/outcomes      | 113                         | 65.2     | 56       | Single institution Multicenter |
| Parvizi et al., 2018 | TKA  | Decision tree  | Adverse event/other complication | 422                        | 65.4     | 52.3     | Single institution           |
| Pua et al., 2019   | TKA       | Decision tree  | PROs/outcomes      | 1208                        | 67.8     | 75       | Single institution           |
| Schwartz et al., 1997 | TKA | Decision tree, regression, boosting, ANN, regression | Postoperative pain | 221                         | 63       | 57       | THR outcomes database at Center for Clinical Effectiveness of the Henry Ford Health System |
| Van et al., 2019   | THA       | ANN         | Adverse event/other complication | 100                         | —        | —        | Single institution           |
| Wu et al., 2016    | TJA       | Regression, SVM | Adverse event/other complication | —                          | 69.6     | 53.3     | Single institution           |
| Yoo et al., 2013   | TKA       | SVM         | PROs/outcomes      | —                           | 0        | 0        | Single institution           |

ACS-NSQIP, American College of Surgeons National Surgical Quality Improvement Program; ANN, artificial neural network; CHECK, Cohort Hip and Cohort Knee; ML, machine learning; NHS, National Health Service; PRO, patient-reported outcome; SVM, support vector machine; THA, total hip arthroplasty; THR, total hip reconstruction; TJA, total joint arthroplasty; TKA, total knee arthroplasty.

which classified TKA and total hip arthroplasty patients based on surgical complexity scores [19].

AI/ML models may aid in predicting postoperative complications and creating personalized postoperative management protocols to avoid or manage those obstacles and maximize outcomes. Several of the reviewed studies used AI/ML models to accurately predict the risk of a range of postoperative complications and adverse events [14,19,29,47,50,51,60]. TKA and total hip arthroplasty revisions and reoperations are also modeled with AI/ML algorithms in some studies, [15,16,21,64] as well as hospital readmissions [20,21,26,27]. In the postoperative period, AI/ML tools offer surgeons the ability to predict patients’ outcomes after surgery, including functional outcomes and PRO scores [14,32,33,43,45,48,53,54,57–59,61]. Postoperative pain has also been shown to be predicted with AI/ML [43,53,56,55] including identification of patients at high risk for prolonged postoperative opioid prescriptions. These tools may better inform analgesic and pain management protocols, especially for opioid prescriptions. They may also enable surgeons to better
tailor treatment for their patients and perhaps offer nonoperative management, especially if there is a high predicted risk of revision surgery or potentially serious postoperative complications.

A powerful predictive tool that aids in clinical decision-making by being able to integrate a large amount of information and identify complex patterns, ML is still vulnerable to the biases faced in other forms of clinical research [65–68]. Among these biases are those related to nonrandom missing data, limited sample size and underestimation to avoid overfitting by the models, and misclassification of disease or discrepancies in measurement between providers [65,67]. These result in specific biases as there can be a limited number of inputs based on researchers’ belief of which variables are important or the ability to collect all possible variables, thus limiting the accuracy and generalizability of the model. Models created on single-institution data sets may not be generalizable because of variation in measurement or reporting of the variables or outputs [65]. Some national data sets that are often used to create these models do not provide granular data which can lead to errors. Other databases can be subject to several biases including selection bias and misclassification of diagnosis as some of these sets have been created for purposes other than research, such as billing [67,68]. Sample size and the population from which the training set is sourced from are also important when considering generalizability; these models may be better at making predictions for those individuals with high access to care as they are built of their data [67]. In both smaller single and large multicenter databases, there is often a lack of information related to social determinants of health which may contribute to disparities seen within our current system [65,68]. Controversy surrounding the use of gender and race information in AI/ML models raises ethical concerns regarding potential introduction of bias into prediction models which are designed to optimize outcomes based on historically inequitable health-care data [66,69,70]. These challenges must be taken into account when implementing the use of AI/ML in clinical settings, especially given the well-studied systemic racial and socioeconomic disparities that exist in US health care [70].

The application of AI/ML to clinical decision-making in hip and knee arthroplasties may result in optimized outcomes by aiding in accurate patient selection and surgical planning during the perioperative period. Several studies in our review demonstrated the use of AI/ML models to predict hospital LOS and readmissions and associated inpatient costs after total joint arthroplasty [14,22-25,27-29,46]. Other studies have demonstrated AI/ML potential to reduce unnecessary expenditures and create risk-adjusted reimbursement models [24,27,28,46]. AI/ML may even enable insurers to more accurately account for individual patient risk and case complexity, especially for bundled payment models. However, the shift away from fee-for-service models and toward models which reward cost-efficiency and incentivize treating a low-acuity patient population may indirectly exclude certain patients from accessing care, especially when based on ethnic or socioeconomic background and associated comorbidities and risk factors [71,72]. AI/ML is a powerful tool that can be broadly adopted in health care and, more specifically, within the field of hip and knee arthroplasty, to optimize patient outcomes, but the data sets upon which these models are trained on must be carefully constructed not only to be of a sufficient size but also to adequately represent the complexities of our patient population. This study has several potential limitations, as we did not use criteria to evaluate the quality of the various data sets used in the reviewed studies, and additional studies using more standardized data sets would be required before making conclusions about clinical efficacy.

### Table 5
Statistical comparison of reported model performance metrics, by postoperative predictions/management applications.

| Postoperative prediction/management applications | Performance metrics: mean (SD, n) | AUC | Accuracy | Sensitivity | Specificity |
|-------------------------------------------------|-----------------------------------|-----|----------|-------------|-------------|
| 1. Adverse event/other complication              | 0.84 (0.1, 14)                   | —   | 97.7 (—, 1) | 99.5 (—, 1) |
| 2. Cardiovascular complication                  | 0.77 (0.08, 8)                   | —   | —         | —           |
| 3. Postoperative pain                           | 0.83 (0.05, 10)                  | 78.8 (2.2, 7) | 78.7 (7.5, 7) | 78.8 (4,9, 7) |
| 4. Postoperative mortality                      | 0.81 (0.07, 3)                   | —   | —         | —           |
| 5. PROs/outcomes                                | 0.81 (0.08, 56)                  | 75.1 (8.4, 12) | 76.9 (7.1, 13) | 64.9 (24, 13) |
| 6. Sustained opioid use                         | 0.71 (0.09, 10)                  | —   | —         | —           |
| ANOVA                                           | P = .002                         | P = .279 | P = .042 | P = .167 |
| Tukey Post Hoc Tests (stat. significant results) | 6 vs 1 (P = .002)                | -    | 1 vs 3 (P < .001) | - |
|                                                 | 6 vs 3 (P = .003)                | -    | 1 vs 5 (P < .001) | - |
|                                                 | 6 vs 5 (P = .011)                | -    | -         | -           |

ANOVA, analysis of variance; AUC, area under curve; PRO, patient-reported outcome; SD, standard deviation.

### Table 6
Statistical comparisons of reported model performance metrics, by AI/ML algorithm.

| AI/ML algorithm | Performance metrics: Mean (SD, n) | AUC | Accuracy | Sensitivity | Specificity |
|-----------------|-----------------------------------|-----|----------|-------------|-------------|
| ANN             | 0.81 (0.11, 56)                   | 87.6 (11.7, 14) | 70.69 (24.18, 15) | 88.4 (12.9, 15) |
| Bayesian         | 0.81 (0.07, 8)                    | 84.1 (2.6, 4) | —         | —           |
| Boosting         | 0.79 (0.07, 19)                   | 77.3 (7.1, 7) | 77.8 (5.36, 5) | 72.8 (11.7, 5) |
| Decision tree    | 0.78 (0.1, 41)                    | 89 (—, 1) | 86.35 (16.05, 2) | 99.8 (0.4, 2) |
| Regression       | 0.77 (0.07, 62)                   | 79 (8.7, 7) | 75.75 (11.28, 6) | 70.4 (14.6, 6) |
| SVM              | 0.77 (0.11, 26)                   | 83.2 (10, 5) | 86.1 (7.34, 5) | 80.5 (16.1, 5) |
| ANOVA           | P = .252                         | P = .228 | P = .497 | P = .019 |
| Tukey Post Hoc Tests (stat. significant results) | —    | —    | —         | —           |

AI/ML, artificial intelligence/machine learning; ANN, artificial neural network; ANOVA, analysis of variance; AUC, area under curve; SD, standard deviation; SVM, support vector machine.
Conclusions

The body of literature on AI/ML-based applications in hip and knee arthroplasties is growing rapidly. Currently, these models are doing well in predicting some postoperative complications but are still limited in predicting postoperative opioid use and need for readmission or reoperation. The accuracy of these predictive tools has the potential to increase with technological advancements and larger data sets, but these models also require external validation. Future work in AI/ML-based applications should aim at creating accurate commercially ready tools that can be integrated into existing systems and to fulfill their role as an aid to physicians and patients in clinical decision-making.

Conflicts of interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: R.P.S. is a paid consultant for KCI USA, Inc and LinkBio Corporation, H.J.C. is a paid consultant for KCI USA, Inc and Zimmer Biomet. J.A.G. is a paid consultant for Smith & Nephew Inc.

References

[1] Wong D, Yip S. Machine learning classifies cancer. Nature 2018;555(7697): 58-60.
[2] Johnson KW, Torres Soto J, Glicksberg BS, et al. Artificial intelligence in cardiology. J Am Coll Cardiol 2018;71(23):2668.
[3] Saber H, Somai M, Rajah GR, Scalzo F, Lieshendk DS. Predictive analytics and machine learning in stroke and neurovascular medicine. Neurol Res 2019;41(8):681.
[4] Shamer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. Machine learning in cardiovascular medicine: are we there yet? Heart 2018;104(14): 1156.
[5] Cabitza F, Locoro A, Banfi G. Machine learning in orthopedics: a literature review. Front Bioeng Biotechnol 2018;6:75.
[6] Haebeler HS, Helm JM, Navarro SM, et al. Artificial intelligence and machine learning in lower extremity arthroplasty: a review. J Arthroplasty 2019;34(10):2201.
[7] 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 2020;35(2):1034.
[8] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436.
[9] Rajkomar A, Dean J, Kohane IS. Machine learning in medicine. N Engl J Med 2019;380(14):1347.
[10] Deo RC. Machine learning in medicine. Circulation 2015;132(20):1920.
[11] Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med 2019;380(13):1216.
[12] Medicine CE-B. OCEBM levels of evidence. In: Oxford Centre for evidence-based medicine levels of evidence. Oxford, UK: University of Oxford; 2020.
[13] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143(1):29.
[14] Alam MF, Briggs A. Artificial neural network metamodel for sensitivity analysis in a total hip replacement health economic model. Expert Rev Pharmacoecon Outcomes Res 2019;1,1.
[15] Aram P, Trela-Larsen L, Sayers A, et al. Estimating an individual’s probability of revision surgery after knee replacement: a comparison of modeling approaches using a national data set. Am J Epidemiol 2018;187(10):2252.
[16] Bonaldhari H, Pelletier JP, Martel-Pelletier J. A reliable time-series method for predictive early prediction of value metrics and a patient-specific prediction model for primary total hip arthroplasty: development and validation of a deep learning model. J Arthroplasty 2019;34(10):2228.
[17] Ramkumar PN, Navarro SM, Haebeler HS, et al. Development and validation of a machine learning algorithm after primary total hip arthroplasty: application to length of stay and payment models. J Arthroplasty 2019;34(4):632.
[18] Ramkumar PN, Karna JM, Navarro SM, et al. Preoperative predictive value of revision surgery: a comparative study using deep learning algorithms. J Arthroplasty 2019;34(10):2228.
[19] Lee HK, Jin R, Feng Y, Bain PA, Coffinett J, Baker C, LiJ. Modeling and analysis of postoperative interventions for total knee replacement super-utilizer patients using simulations. IEEE J Biomed Health Inform 2017;568.
[20] Lee HK, Jin R, Feng Y, et al. An analytical framework for TRJ readmission prediction and cost-effective intervention. IEEE J Biomed Health Inform 2015;23(4):1760.
[21] Navarro SM, Wang EY, Haebeler HS, et al. Machine learning and primary total knee arthroplasty: patient forecasting for a patient-specific payment model. J Arthroplasty 2018;33(12):3617.
[22] Karna JM, Navarro SM, Haebeler HS, et al. Predicting inpatient payments prior to lower extremity arthroplasty using deep learning: which model architecture is best? J Arthroplasty 2019;34(10):2235.
Gong CS, Yu L, Ting CK, et al. Predicting postoperative vomiting for orthopedic patients receiving patient-controlled epidural analgesia with the application of an artificial neural network. Biomed Res Int 2014;2014:786418.

Harris AHS, Kuo AC, Weng Y, Trickey AW, Bowe T, Giori NJ. Can machine learning methods produce accurate and easy-to-use prediction models of 30-day complications and mortality after knee or hip arthroplasty? Clin Orthop Relat Res 2019;477(2):452.

Huang Z, Huang C, Xie J, et al. Analysis of a large data set to identify predictors of blood transfusion in primary total hip and knee arthroplasty. Transfusion 2018;58(8):1655.

Huang Z, Huang C, Xie J, et al. Pre-operative predictors of post-operative pain relief: pain score modeling following total knee arthroplasty. Transfusion 2018;22.

Huber M, Kurz C, Leidl R. Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning. BMC Med Inform Decis Mak 2019;19(1):3.

Jacobs CA, Christensen CP, Karthikeyan T. Using machine learning to identify factors associated with satisfaction after TKA for different patient populations. J Orthop Res 2016;34.

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(10):2272.

Katakam A, Karhade AV, Schwab JH, Chen AF, Bedair HS. Development and validation of machine learning algorithms for postoperative opioid prescriptions after TKA. J Orthop 2020;22:95.

Kluge F, Hännikä J, Paluusto C, et al. Pre-operative sensor-based gait parameters predict functional outcome after total knee arthroplasty. Gait Posture 2018;66:194.

Knunz 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;

Van Osem S, Van Der Straeten C, Arnout N, Deprez P, Van Damme G, Victor J. A New prediction model for patient satisfaction after total knee arthroplasty. J Arthroplasty 2016;31(12):2660.

Parvizi J, Tan TL, Goswami K, et al. The 2018 definition of periprosthetic hip and knee infection: an evidence-based and validated criteria. J Arthroplasty 2018;33(5):1309.

Pua YH, Kang H, Thumboo J, et al. Machine learning methods are comparable to logistic regression techniques in predicting severe walking limitation following total knee arthroplasty. Knee Surg Sports Traumatol Arthrosc 2019.

Schwartz MH, Ward RE, Macwilliam C, Verner J. Using neural networks to identify patients unlikely to achieve a reduction in bodily pain after total hip replacement surgery. Med Care 1997;35(10):1020.

Wu HY, Gong CA, Lin SP, Chang KY, Tsou MY, Ting CK. Predicting post-operative vomiting among orthopedic patients receiving patient-controlled epidural analgesia using SVM and LR. Sci Rep 2016;6:27041.

Zhang T, Liu P, Zhang Y, et al. Combining information from multiple bone turnover markers as diagnostic indices for osteoporosis using support vector machines. Biomarkers 2019;24(2):120.

Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018;178(11):1544.

Chai DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med 2018;378(11):981.

Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019;25(9):1337.

Chen JH, Asch SM. Machine learning and prediction in medicine - beyond the peak of inflated expectations. N Engl J Med 2017;376(26):2507.

Gibney E. The battle for ethical AI at the world’s biggest machine-learning conference. Nature 2020;577(7792):609.

Peagin J, Bennefield Z. Systemic racism and U.S. health care. Soc Sci Med 2014;103(7).

Sullivan R, Jarvis LD, O’Cara T, Langfitt M, Emory C. Bundled payments in total joint arthroplasty and spine surgery. Curr Rev Musculoskelet Med 2017;10(2):218.

Dunnit LA, Kahvecioglu D, Marrufo G, et al. Association between hospital participation in a Medicare bundled payment initiative and payments and quality outcomes for lower extremity joint replacement episodes. JAMA 2016;316(12):1267.