REVIEW ARTICLE

Royal Australasian College of Surgeons

# Supervised machine learning for the prediction of post-operative clinical outcomes of hip and knee replacements: a review

Khashayar Ghadirinejad ,* Roohollah Milimonfared,* Mark Taylor,* Lucian B. Solomon,†‡ Stephen Graves,§ Nicole Pratt,¶ Richard de Steiger§‖ and Reza Hashemi*

*The Medical Device Research Institute, College of Science and Engineering, Flinders University, Clovelly Park, South Australia, Australia
†Department of Orthopaedics and Trauma, Royal Adelaide Hospital, Adelaide, South Australia, Australia
‡Centre for Orthopaedic & Trauma Research, University of Adelaide, Adelaide, South Australia, Australia
¶The Australian Orthopaedic Association National Joint Replacement Registry, South Australian Health and Medical Research Institute, Adelaide, South Australia, Australia
‖Quality Use of Medicines and Pharmacy Research Centre, School of Pharmacy and Medical Sciences, University of South Australia, Adelaide, South Australia, Australia and
§Department of Surgery, Epworth HealthCare, The University of Melbourne, Parkville, Victoria, Australia

**Key words**
data, machine learning, predictive analytics, total hip replacement, total knee replacement.

**Correspondence**
Dr Khashayar Ghadirinejad, PhD, The Medical Device Research Institute, College of Science and Engineering, Flinders University, Clovelly Park, South Australia, Australia.
Email: khashayar.ghadirinejad@flinders.edu.au

**K. Ghadirinejad** PhD; **R. Milimonfared** PhD; **M. Taylor** PhD; **L. B. Solomon** PhD, FRACS; **S. Graves** MBBS, DPhil; **N. Pratt** PhD; **R. de Steiger** MBBS, FRACS; **R. Hashemi** PhD.



Accepted for publication 27 March 2024.

doi: 10.1111/ans.19003

## Abstract

Prediction models are being increasingly used in the medical field to identify risk factors and possible outcomes. Some of these are presently being used to develop guidelines for improving clinical practice. The application of machine learning (ML), comprising a powerful set of computational tools for analysing data, has been clearly expanding in the role of predictive modelling. This paper reviews the latest developments of supervised ML techniques that have been used to analyse data related to post-operative total hip and knee replacements. The aim was to review the most recent findings of relevant published studies by outlining the methodologies employed (most-widely used supervised ML techniques), data sources, domains, limitations of predictive analytics and the quality of predictions.
This paper reviews the latest developments of supervised machine learning (ML) techniques that have been used to analyse data related to post-operative total hip and knee replacements. The aim was to review the most recent findings of relevant published studies by outlining the methodologies employed (most widely used supervised ML techniques), data sources, domains, limitations of predictive analytics and the quality of predictions.

## Introduction

Machine learning (ML) provides systems the ability to produce mathematical models by learning patterns from empirical data. In medical research, ML is mostly used to extract information regarding diagnosis and treatment patterns. Examples include data-driven predictions of drug effects and interactions, the detection of comorbidity groups in autism spectrum disorders,[1] and the identification of type 2 diabetes subgroups.[2]

The most widely used ML approach in medical sciences is supervised learning. This technique estimates the mapping function for new input data in order to predict categorized, real values, or time-to-event outputs. Examples of supervised ML algorithms in orthopaedics include linear regression and similar techniques, decision trees (DTs), random forests (RFs), neural networks (NNs), Naive Bayes, support vector machines (SVMs) and nearest neighbours.[3]

ML technology is relatively new to the field of orthopaedic surgery. Recent applications of ML technology include the development of image-based diagnoses[4–6] and the improvement of value-based care.[7–10] Gait analysis algorithms may be used to notice early warning indications of revision arthroplasty, such as undiagnosed infection or instability.[11] Kotti *et al*.[6] used RF-based modelling to detect osteoarthritis (OA) through gait analysis, reporting a mean accuracy of 72% in 47 patients with this disease. In the area of value-based payment models, Navarro *et al*.[8] evaluated applying a Naïve Bayesian model to assess patient-level factors and forecast value metrics prior to the total knee arthroplasty episode of care. Similarly, Ramkumar *et al*.[7] explored the use of a Naïve Bayesian classifier in primary total hip arthroplasty, and found an excellent predictive capacity with respect to costs and hospital length of stay (LOS).

ML is increasingly used in the medical sciences because it offers alternative approaches to address the probability of confounding, particularly in high-dimensional datasets. For instance, Kaplan–Meier is a common statistical method that uses lifetime data to estimate the survival function of primary total hip replacement (THR) and total knee replacement (TKR) implants.[12–20] However, an ideal method uses a time-to-event endpoint while limiting the confounding effect of other variables.

There are a number of important questions that need to be addressed given the recent advances and increasing use of supervised learning methods in various medical areas, including orthopaedics. These comprise: What are the main justifications for using supervised ML methods and their effectiveness in assisting with the THR and TKR procedures? Are the ML results affected by data volume and data quality?

The aim of this paper was to address these questions by reviewing the use of supervised ML techniques in regression, classification and survival problems associated with the post-operative outcomes of THR and TKR. The different types of ML techniques (including RF, SVM, Naive Bayes, and deep learning) were reviewed, focusing on the data source, domains, limitations, and quality of outcomes reported in the literature.

## Method (literature search and selection criteria)

For this review paper, the English-based literature was searched online, including PubMed search engine and Scopus Elsevier databases using various key terms: supervised learning, machine learning, hip replacement, knee replacement, predictive, and data. A comprehensive search was conducted across these databases for the period of each database inception to the end of 2019. Only articles and their corresponding references reporting studies that utilized supervised ML techniques were reviewed for inclusion. Non-peer-reviewed studies, non-English language studies, unpublished manuscripts were excluded. The studies using unsupervised (or semi-supervised) ML learning approaches, or employing learning methods to train unreal data, or with a focus on pre-operative outcomes of THR and TKR were not considered. Titles and abstracts of the remaining articles were then carefully screened.

### Random forest (RF)

The RF is a tree-based ensemble learning method widely used to predict an outcome or rank and select the most significant variables. Cafri *et al*.[21] defined the time to first revision in elective primary THR as the outcome in order to compare two ML techniques (elastic-net Cox *vs*. random survival forest) with the principal aim of assessing their performance in identifying recalled components. The concept of training an ML model to identify significant features differs from predicting the survival probability of components. The authors used 348 unique components as indicator variables in addition to patient covariates, which were all categorized and treated as potential confounders to detect the components based on the statistical significance. Two of the six recalled components (ASR shell/head and Rejuvenate) with $P < 0.001$ and minimal depth rank of 1 and 2 in the RF model, were identified in both approaches. However, one more component (Durom shell/Metasul femoral head) was also picked by the regularized Cox model, even while maintaining the false discovery rate at 0.05. The results suggested that the ML methods can be effective for detection, although the Cox technique with a more traditional way to address confounding performed more effectively.[21]

Gabriel *et al*.[22] trained predictive models using RF, ridge and lasso regression, and multivariable logistic regression to determine those patients who would not need prolonged hospital LOS after THR. The discriminatory ability was reported as 0.735 (95% confidence interval, 0.675–0.787) using the area under the receiver operating characteristic curve (AUC) for multivariable logistic regression (the best-fit algorithm). Also, '$P = 0.37 > 0.05$' was obtained as fitting goodness by the Hosmer–Lemeshow test. Nine variables—age, sex, anaemia, opioid use, obesity, metabolic equivalents score, chronic obstructive pulmonary disease, primary anaesthesia type, and hypertension—were included in the proposed calculator. The authors stated that this model might assist clinicians in the strategic planning of bed availability to reduce both overcrowding and underutilisation.[22] However, this sort of single-institution studies needs to add external validations and use larger sample sizes before reporting big statements.

Prediction of patient-reported outcomes (130 945 observations) by eight supervised binary classifiers (logistic regression, extreme gradient boosting, multistep adaptive elastic-net, RF, neural net, Naïve Bayes, k-nearest neighbours and boosted logistic regression) was the aim of a study on THR and TKR.[23] The generic and disease-specific improvement was considered as the dependent outcome based on the Oxford Hip and Knee Score (Q score) and the EQ-5D-3L visual analogue scale (VAS). Results showed that RF, extreme gradient boosting, linear model, and multistep elastic net had the highest overall J-statistic (as a statistic that shows diagnostic tests' performance). The AUC of the best-fit models was reported as around 0.86 (VAS) and 0.70 (Q score) for knee replacements, and 0.87 (VAS) and 0.78 (Q score) for hip replacements. All these models were used to depict the most contributive variables but some methods, such as RF with random permutations, can introduce bias and artificial variable selection under specific circumstances.[24,25] If several significant variables were correlated, they share the importance, suggesting that the variable importance may be shown lower than the reality.[26]

## Support-vector machine (SVM)

The SVM is a supervised ML algorithm, suitable for creating subtle patterns from complex datasets in both classification and regression problems. The SVM classifier was examined through an image-based approach for its usefulness in rating the corrosion damage on the THR prostheses (at the head–neck taper junction).[27] The classifier was applied to capture local and textural information (as two approaches of object recognition); then, Goldberg's scores were given to rank the images. The hyperparameters were tuned to minimize the cross-validation error by Bayesian optimisation; the features with greater discriminatory power were selected after analysis of the Neighbourhood components as a supervised learning method to classify the multivariable dataset into separate groups. An accuracy level of 85% was obtained using five-fold cross-validation, whereas a limited pool of available prostheses made a significant limitation in terms the validity. Fontana *et al.*[28] investigated whether ML algorithms are able to predict the patients who will attain Minimal Clinically Important Difference in THR and TKR post-operatively. Based on patient-reported outcome measures (PROMs), 6480 TKRs and 7239 THRs were selected from only a single hospital. Linear SVM, logistic LASSO, and RF were trained on 80% of the dataset to predict two-year minimal clinically important differences. The AUCs of the three ML methods varied from 0.60 to 0.89 with the best result for the LASSO but Theoretically these values cover a range of poor to acceptable prediction but the presence of high multicollinearity breaks one of the assumptions promising that logistic regression can produce unbiased coefficients. Although the authors noted that ML holds much guarantee for assisting as a clinical decision-making support system, it should be considered that most similar studies were only limited to small number of observations.

## Naïve Bayes

Ramkumar *et al.*[7] aimed to develop and validate a Naïve Bayesian model using pre-operative primary THR data to predict LOS and patient-specific inpatient payments, and then recommended the use of a risk-adjusted patient-specific payment model that reflects patient comorbidity. The data of 122 334 primary THRs, including race, age, gender, and comorbidity scores, was used to train and evaluate the model using AUC and training accuracy. Inpatient payments were categorized as the output variable, and the AUC showed the validity of 0.71 and 0.87 for payment and LOS, respectively. Naïve Bayesian methods assume conditional independence which, however, fail to identify confounding variables. The validation of the developed model required that, first, an initial viability be established before proceeding with the resource-intensive task of developing other available models. This may mean that the other ML methods such as deep learning can create a more accurate model.[29,30] SVM and NNs algorithms can take into account confounding relationships among the variables and may create better machine automation.[30]

## Deep learning

There are several major differences between deep learning and other ML methods.[31] Deep learning is a subset of machine learning that reproduces the mechanisms of the human brain in learning from big data and generating patterns for decision makings. Deep learning techniques have become popular in research studies since they can automatically perform the raw data engineering by finding the optimal inner representation, which is necessary for the discriminative (mapping) task. Deep learning methods are often mysterious because of their black-box nature, which is often the main source of concern in medical applications.[32] However, they can analyse data efficiently and can capture the more complex structure of big datasets for THR and TKR despite computational complexity.[33] For instance, in a recent study, Qiu *et al.*[34] used a large commercial claims dataset to identify patients with a strong likelihood of requiring TKR and THR surgeries. Supervised ML methods (RF, LASSO, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN)) were investigated using 540 000 observations of patient data and 2000 variables. The deep learning methods showed much better performance than LASSO, RF, and RNN (a common type of ANN in the medical area[35,36]) with a pooling mechanism that recorded the best accuracy using different metrics: $0.8339 \pm 0.0024$ as AUC and $0.0662 \pm 0.0008$ for precision with recall set to 0.9. As a function to reduce the number of parameters and computation in the network, the applied pooling mechanism positively influenced the performance by detecting the additional signals from the hidden intermediate states.

## Other ML methods used in THR and TKR

In one of the first studies to develop a pre-operative algorithm for predicting post-operative opioid use after THR,[37] five ML algorithms (stochastic gradient boosting, RF, SVM, NNs, elastic-net penalized logistic regression) were developed. The elastic-net penalized logistic regression attained the best performing method for calibration, discrimination (C-statistic = 0.77), and decision curve analysis; whereas, the NNs and stochastic gradient boosting models recorded the same AUC (0.77) with elastic-net penalized logistic regression.

Predictive risk models were developed and validated in another study[38] to forecast the risk of death and major complications after THR and TKR. This involved 70 569 observations of OA patients who received primary THR and TKR and the highest C-statistics and bootstrapped confidence intervals (CIs) were reported for 30-day mortality (0.73; 0.66–0.79) and cardiac complications (0.75; 0.71–0.79) based on the cross-validation of the boosted regression models. The lowest values were also reported for returns to the operating room (0.60; 0.57–0.63), and for deep vein thrombosis (0.59; 0.55–0.64).

A similar study to predict the risk of complications and 30-day mortality after TKR and THR trained LASSO regression models on 107 792 data, including clinical inputs and pre-operative demographic variables. The results demonstrated an acceptable prediction accuracy for death (0.73; 95% CI, 0.70–0.76), a renal complication (C-statistic, 0.78; 95% CI, 0.76–0.80), and a cardiac complication (0.73; 95% CI, 0.71–0.75) within 30 days of arthroplasty, although poor accuracy was reported for venous thromboembolism (C-statistic, 0.61; 95% CI, 0.60–0.62). Importantly, it was suggested that these are the most accurate and validated prediction models; however, the models performed poorly in terms of external validation (prediction of outcomes from another dataset).[39]

Given the effect of THR on health-related quality of life (HRQoL), Nemes *et al*.[40] suggested a clinical decision support system (DSS) using Swedish joint registry data to help clinicians assessing the future benefits of THR by offering predictions of 1-year post-operative HRQoL. Three groups of supervised ML algorithms were used: (1) linear regression and its variants, (2) nonlinear regression algorithms, and (3) classification trees and rule-based models. The multivariate adaptive regression splines ($R^2 = 0.158$) were shown to have the best predictive capability. However, it was not significantly better than the developed linear regression model ($R^2 = 0.157$). Although 11 patient-related predictors were considered, more variables need to be analysed as predictors to construct a comprehensive and successful DSS. There is no set criterion on a good $R^2$ value as it may increase by adding even non-important predictors in a multivariate model; it is often preferred to compare the performance of models with the same variables.[41]

To predict patients' pain and function levels after undergoing TKR, 1649 patient-reported data in the UK were studied with the aim of training and validating a supervised ML model. Clinical factors and patient characteristics were used as pre-operative inputs to predict the Oxford Knee Score (OKS) after 12-months of TKR. This prediction model provided an individualized estimation of post-operative OKS, and also changes in OKS. The bootstrap backward linear regression showed predictive validity with $R^2$ of 0.175 (internal validation) and 0.211 (external validation).[42] These low values explained 17.5% and 21.1% variability in the outcome, suggesting that the models' generalizability is dependant on considering more potential predictive factors.

## Discussion

Both ML and conventional statistical methods with similarities and differences have made great strides in the support they can offer clinicians, although conventional methods have been the main statistical approach in the domain of THR and TKR to date. For example, the generation of risk-predictive models is a common approach taken for estimating the risk of an event of interest occurring in post-operative THR and TKR. In previous studies, most of these models have been developed using the conventional methods (e.g., logistic regression, Cox proportional hazards regression)[43,44] rather than the more modern ML strategies. These strategies are becoming the main approach for addressing prediction problems across a wide range of industry and science domains. Although to date there has been very limited adoption of these strategies for the purpose of THR and TKR predictions, it is anticipated that more studies will be published on ML predictive models for THR and TKR.

One misconception is that conventional statistical methods rely on predetermined assumptions and mathematical equations to formalize relations between the variables, whereas ML techniques use the data to recognize these relationships.[45] The capability to link a large amount of data and variables together and capture complex non-linear relationships, is the key benefit of ML methods over the conventional statistical methods. ML, as a useful and powerful set of computational tools, is now a common choice for the development of predictive models in the medical community.[46–48] The successful adoption of several electronic medical record (EMR) systems developed for various purposes (prognosis, diagnosis, or treatment) have been noted in several studies.[49,50] Greatly improved subsets of ML models, known as ANNs, have been notable in total joint replacement contexts because of a great potential for processing 'big data'.[51]

ML has proved its undeniable capability, although it is not free of issues. The accuracy of predictive models is dependent on the quality of the data sources, and predictions may be significantly affected by the amount of data and the number of variables included. Therefore, care should be taken when dealing with limited data, as it is not advisable to report that these models are reliable with acceptable accuracy levels. Furthermore, ML models should be assessed and evaluated using a randomized cohort of studies and controlled trials in real-world settings. Hence, more improvements are needed in ML orthopaedic applications to translate the research aims into clinical practices.

It is essential to understand the difference between two different types of studies with a focus on the impact of variables on the outcome or predicting outcomes for a separate data. While ML has the potential to offer more accurate predictions, this can cost a poorer understanding of the relationships among the variables. The output of ML models needs to be interpreted carefully, and the expectations of predictive analytics can be raised with a consciousness of the matters associated with misinterpretation and over-fitting in clinical settings.

To date, multivariable predictive models have been developed for THR and TKR based mainly on patient-reported factors and imaging variables. The literature shows that ML adoption for post-operative THR and TKR is still in the basic phase with only a few studies confirming that the models are entirely available for a THR or TKR practice. This suggests future research opportunities for studies on the post-operative clinical outcomes of THR and TKR. There is still a need for models that can predict various outcomes such as the early identification of prostheses outliers based on the available big data from the national joint registries around the world. Perhaps, this indicates that now is the time to enter a new era of THR and TKR by developing decision-making support systems comprising effective predictors based on big data. A future global direction of ML in the domain of joint arthroplasty could be to enable surgeons to determine what is the best for their patients.

## Author contributions

**Khashayar Ghadirinejad:** Writing – original draft. **Roohollah Milimonfared:** Writing – review and editing. **Mark Taylor:** Writing – review and editing. **Lucian B. Solomon:** Writing – review and editing. **Stephen Graves:** Supervision; writing – review and editing. **Nicole Pratt:** Writing – review and editing. **Richard de

**Steiger:** Writing – review and editing. **Reza Hashemi:** Supervision; writing – review and editing.

## Acknowledgement

Open access publishing facilitated by Flinders University, as part of the Wiley - Flinders University agreement via the Council of Australian University Librarians.

## Conflicts of interest

The authors declare that there is no conflict of interest.

## References

1. Doshi-Velez F, Ge Y, Kohane I. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics* 2014; **133**: e54–63.
2. Schulze MB, Hoffmann K, Boeing H *et al*. An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. *Diabetes Care* 2007; **30**: 510–5.
3. Cabitza F, Locoro A, Banfi G. Machine learning in orthopedics: a literature review. *Front. Bioeng. Biotechnol.* 2018; **27**: 75.
4. Kozic N, Weber S, Büchler P *et al*. Optimisation of orthopaedic implant design using statistical shape space analysis based on level sets. *Med. Image Anal.* 2010; **14**: 265–75.
5. Eskinazi I, Fregly BJ. Surrogate modeling of deformable joint contact using artificial neural networks. *Med. Eng. Phys.* 2015; **37**: 885–91.
6. Kotti M, Duffell LD, Faisal AA, McGregor AH. Detecting knee osteoarthritis and its discriminating parameters using random forests. *Med. Eng. Phys.* 2017; **43**: 19–29.
7. Ramkumar PN, Navarro SM, Haeberle HS *et al*. Development and validation of a machine learning algorithm after primary total hip arthroplasty: applications to length of stay and payment models. *J. Arthroplasty* 2019; **34**: 632–7.
8. Navarro SM, Wang EY, Haeberle HS *et al*. Machine learning and primary total knee arthroplasty: patient forecasting for a patient-specific payment model. *J. Arthroplasty* 2018; **33**: 3617–23.
9. Ramkumar PN, Karnuta JM, Navarro SM *et al*. Preoperative prediction of value metrics and a patient-specific payment model for primary total hip arthroplasty: development and validation of a deep learning model. *J. Arthroplasty* 2019; **34**: 2228–34.
10. Karnuta JM, Navarro SM, Haeberle HS, Billow DG, Krebs VE, Ramkumar PN. Bundled care for hip fractures: a machine-learning approach to an untenable patient-specific payment model. *J. Orthop. Trauma* 2019; **33**: 324–30.
11. Haeberle HS, Helm JM, Navarro SM *et al*. Artificial intelligence and machine learning in lower extremity arthroplasty: a review. *J. Arthroplasty* 2019; **34**: 2201–3.
12. Paxton EW, Ake CF, Inacio MC, Khatod M, Marinac-Dabic D, Sedrakyan A. Evaluation of total hip arthroplasty devices using a total joint replacement registry. *Pharmacoepidemiol. Drug Saf.* 2012; **21**: 53–9.
13. Robertsson O. Knee arthroplasty registers. *J. Bone Joint Surg. Br.* 2007; **89**: 1–4.
14. Inacio MC, Lorimer M, Davidson DC, De Steiger RN, Lewis PL, Graves SE. What is the risk of revision surgery in hydroxyapatite-coated femoral hip stems? Findings from a large national registry. *Clin. Orthop. Relat. Res.* 2018; **476**: 2353–66.
15. Southern DA, Faris PD, Brant R *et al*. Kaplan–Meier methods yielded misleading results in competing risk scenarios. *J. Clin. Epidemiol.* 2006; **59**: 1110–4.
16. Paxton EW, Cafri G, Nemes S *et al*. An international comparison of THA patients, implants, techniques, and survivorship in Sweden, Australia, and the United States. *Acta Orthop.* 2019; **90**: 148–52.
17. de Steiger R, Lorimer M, Graves SE. Cross-linked polyethylene for total hip arthroplasty markedly reduces revision surgery at 16 years. *J. Bone Joint Surg. Am.* 2018; **100**: 1281–8.
18. de Steiger R, Peng A, Lewis P, Graves S. What is the long-term survival for primary THA with small-head metal-on-metal bearings? *Clin. Orthop. Relat. Res.* 2018; **476**: 1231–7.
19. Lie SA, Pratt N, Ryan P *et al*. Duration of the increase in early postoperative mortality after elective hip and knee replacement. *J. Bone Joint Surg. Am.* 2010; **92**: 58–63.
20. Lewis PL, Graves SE, de Steiger RN *et al*. Does knee prosthesis survivorship improve when implant designs change? Findings from the Australian Orthopaedic Association National Joint Replacement Registry. *Clin. Orthop. Relat. Res.* 2020; **478**: 1156–72.
21. Cafri G, Graves SE, Sedrakyan A *et al*. Postmarket surveillance of arthroplasty device components using machine learning methods. *Pharmacoepidemiol. Drug Saf.* 2019; **28**: 1440–7.
22. Gabriel RA, Sharma BS, Doan CN, Jiang X, Schmidt UH, Vaida F. A predictive model for determining patients not requiring prolonged hospital length of stay after elective primary total hip arthroplasty. *Anesth. Analg.* 2019; **129**: 43–50.
23. 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.
24. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* 2007; **8**: 1–21.
25. Liaw A, Wiener M. Classification and regression by randomForest. *R News* 2002; **2**: 18–22.
26. Ishwaran H, Kogalur UB, Chen X, Minn AJ. Random survival forests for high-dimensional data. *Stat. Anal. Data Min. ASA Data Sci. J.* 2011; **4**: 115–32.
27. Milimonfared R, Oskouei RH, Taylor M, Solomon LB. An intelligent system for image-based rating of corrosion severity at stem taper of retrieved hip replacement implants. *Med. Eng. Phys.* 2018; **1**: 13–24.
28. Fontana MA, Lyman S, Sarker GK, Padgett DE, MacLean CH. Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? *Clin. Orthop. Relat. Res.* 2019; **477**: 1267–79.
29. Gulshan V, Peng L, Coram M *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; **316**: 2402–10.
30. Ravì D, Wong C, Deligianni F *et al*. Deep learning for health informatics. *IEEE J. Biomed. Health Inform.* 2016; **21**: 4–21.
31. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521**: 436–44.
32. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA* 2017; **318**: 517–8.
33. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining 2015 Aug 10 (pp. 1721–1730).
34. Qiu R, Jia Y, Wang F *et al*. Predictive modeling of the total joint replacement surgery risk: a deep learning based approach with claims data. *AMIA Jt. Summits Transl. Sci. Proc.* 2019; **2019**: 562.

35. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. *J. Am. Med. Inform. Assoc.* 2017; **24**: 361–70.
36. Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks. arXiv preprint arXiv:1511.03677. 2015 Nov 11.
37. Karhade AV, Schwab JH, Bedair HS. Development of machine learning algorithms for prediction of sustained postoperative opioid prescriptions after total hip arthroplasty. *J. Arthroplasty* 2019; **34**: 2272–7.
38. Harris AH, Kuo AC, Bowe T, Gupta S, Nordin D, Giori NJ. Prediction models for 30-day mortality and complications after total knee and hip arthroplasties for veteran health administration patients with osteoarthritis. *J. Arthroplasty* 2018; **33**: 1539–45.
39. Harris AH, 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**: 452–60.
40. Nemes S, Rolfson O, Garellick G. Development and validation of a shared decision-making instrument for health-related quality of life one year after total hip replacement based on quality registries data. *J. Eval. Clin. Pract.* 2018; **24**: 13–21.
41. Saunders LJ, Russell RA, Crabb DP. The coefficient of determination: what determines a useful R2 statistic? *Invest. Ophthalmol. Vis. Sci.* 2012; **53**: 6830–2.
42. Sanchez-Santos MT, Garriga C, Judge A *et al*. Development and validation of a clinical prediction model for patient-reported pain and function after primary total knee replacement surgery. *Sci. Rep.* 2018; **8**: 1–9.
43. Mezhov V, Laslett LL, Ahedi H *et al*. Predictors of total hip replacement in community based older adults: a cohort study. *Osteoarthr. Cartil.* 2019; **27**: S79–80.
44. Inacio MC, Pratt NL, Roughead EE, Graves SE. Evaluation of three co-morbidity measures to predict mortality in patients undergoing total joint arthroplasty. *Osteoarthr. Cartil.* 2016; **24**: 1718–26.
45. Bayliss L, Jones LD. The role of artificial intelligence and machine learning in predicting orthopaedic outcomes. *Bone Joint J.* 2019; **101**: 1476–8.
46. Miller DD, Brown EW. Artificial intelligence in medical practice: the question to the answer? *Am. J. Med.* 2018; **131**: 129–33.
47. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N. Engl. J. Med.* 2016; **375**: 1216–9.
48. Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *N. Engl. J. Med.* 2017; **376**: 2507–9.
49. Iwashyna TJ, Liu V. What's so different about big data? A primer for clinicians trained to think epidemiologically. *Ann. Am. Thorac. Soc.* 2014; **11**: 1130–5.
50. Khoury MJ, Ioannidis JP. Big data meets public health. *Science* 2014; **346**: 1054–5.
51. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA* 2018; **319**: 1317–8.