OPINION

Machine-learning-based patient-specific prediction models for knee osteoarthritis

Afshin Jamshidi, Jean-Pierre Pelletier and Johanne Martel-Pelletier

Abstract | Osteoarthritis (OA) is an extremely common musculoskeletal disease. However, current guidelines are not well suited for diagnosing patients in the early stages of disease and do not discriminate patients for whom the disease might progress rapidly. The most important hurdle in OA management is identifying and classifying patients who will benefit most from treatment. Further efforts are needed in patient subgrouping and developing prediction models. Conventional statistical modelling approaches exist; however, these models are limited in the amount of information they can adequately process. Comprehensive patient-specific prediction models need to be developed. Approaches such as data mining and machine learning should aid in the development of such models. Although a challenging task, technology is now available that should enable subgrouping of patients with OA and lead to improved clinical decision-making and precision medicine.

Osteoarthritis (OA) is a leading cause of disability worldwide and one of the most common chronic illnesses, accounting for 40–60% of patients with degenerative diseases of the musculoskeletal system[1]. OA is now recognized as an independent risk factor for increased mortality[2,3]. This disease places a huge burden on health-care services (accounting for 1–2.5% of the gross national product in Western countries)[4], and the cost of OA to these services is expected to double by 2020 and again by 2030 (REF.[5]).

OA can affect many joints but is commonly localized in the weight-bearing joints and most frequently occurs in the knee[6]. Although the main outcome of this disease is cartilage destruction, OA is thought to affect all the tissue structures of the joint[7]. The major risk factors for OA are age (affecting more than half of the world's population aged 65 and older), female sex and obesity[8–10]. OA can be roughly divided into idiopathic (primary) and secondary (has a known cause, such as trauma) OA. Idiopathic OA affects the majority of patients and is difficult to define as its precise aetiology is not known. From a clinical standpoint, idiopathic OA represents a heterogeneous group of disorders with different subgroups that have varying causes and include different clinical and pathological manifestations. For many patients with OA, disease progression can be slow and span many years; however, at least 10% of patients with OA have a rapid disease progression that can lead to the need for total joint replacement[11].

At present, diagnosis of OA occurs mainly during the moderate to severe (late) stage of the disease, by which point the joint tissue often has already become irreversibly damaged. Importantly, no treatment is currently approved by regulatory agencies to cure the disease or prevent disease progression. Current medications are effective only at relieving symptoms and/or pain, and their use could induce major adverse events and even mortality[12]. Progress in developing disease-modifying OA drugs (DMOADs) or therapies that stop or reduce progression of joint tissue deterioration is slow and lagging behind that of other diseases, such as other arthritic diseases. An important hurdle to overcome in OA management is the identification of patients who might benefit the most from such drugs or preventive measures, including individuals at an early stage of the disease or those for whom the disease is not too severe but might progress rapidly. However, existing methods for assessing patients with OA do not provide enough comprehensive information to make robust predictions or prognoses.

In short, current clinical diagnostic procedures do not adequately fulfil the need of clinicians to help patients reduce their risk of disease progression or the need of the health-care industry to develop effective new DMOADs. Prediction models that are capable of analysing large amounts of patient data are needed. The factors involved in OA pathogenesis and, importantly, the interactions between these factors need to be identified and incorporated into prediction models. Although conventional statistical models exist, these methods cannot handle massive amounts of information. Hence, there is a considerable need to develop comprehensive patient-specific risk assessment models using machine-leaning approaches that take into account all the factors and variables and their interactions. The development of computer-based prediction models for analysing large data sets is a promising area in health care in fields such as cancer, genomics and biology[13–16].

In this Perspectives article, we discuss the potential of data mining approaches incorporating statistics and machine learning in the development of prediction models for OA as well as important aspects of these approaches. We also evaluate current prediction models for OA and discuss lessons that can be learnt for future models.

Developing OA prediction models

Predictive modelling in medicine involves the development of models that are capable of analysing data to predict outcomes for an individual. Although historically prediction modelling has involved the use of conventional statistical methods, the use of artificial intelligence approaches such as machine learning (BOX 1) can result in

Box 1 | **Big data analysis**

In health care, artificial intelligence approaches use algorithms and software to approximate human cognition in the analysis of complex medical data. It is now possible to comprehensively apply these approaches to develop osteoarthritis (OA) prediction models that can learn from very large amounts of data. Generally, artificial intelligence approaches outperform conventional statistical approaches on risk prediction tasks owing to their capability and efficiency in finding patterns in data that contain noisy variables and missing and imbalanced data[16]. The use of data mining and machine-learning techniques could lead to the creation of optimized models for real-time decision-making.

Data mining

Data mining refers to an analytical process designed to search a database for consistent patterns and/or systematic relationships between variables, and its ultimate goal is to discover hidden patterns, subtle trends and associations among variables. Interestingly, this approach does not require pre-specification of the outcomes of interest but could identify several important associations. For OA, an immense amount of information is contained but hidden in databases, including information that is potentially important but has not yet been articulated because of the lack of the proper technologies to analyse all the information together.

Machine learning

Machine learning is a method of data analysis. This tool is closely related to computational statistics and automates analytical model building. Machine learning enables computers to find hidden insights without being explicitly programmed using algorithms that iteratively learn from the data. The training step of machine learning involves providing a machine-learning algorithm with a training data set (that, for supervised methods, includes input and outcome variables) to learn from. The learning algorithm finds patterns in the training data set such that the input parameters correspond to the target. The developed model is then used to perform predictions on new data for which the outcome value is not known (for example, to assign a class to a new observation). The accuracy of data mining and machine-learning methods varies depending on the selected variables and the similarity in size between the training and testing data sets.

Machine-learning algorithms include but are not limited to supervised algorithms such as classification and regression algorithms (known as classifiers and regressors, respectively) (TABLE 2) and unsupervised algorithms such as clustering and dimensionality reduction algorithms. Some of the most commonly applied algorithms are support vector machines, k-nearest neighbours, artificial neural networks, decision trees, ensemble methods (such as random decision forests), naive Bayes, clustering and principal component analysis.

models that process complex data quickly and efficiently (FIG. 1). These advanced approaches are based on algorithms that have been specially designed to deal with the uncertainty and imprecision typically found in clinical and biological data sets.

The OA field has been relatively slow in adopting advanced analytical techniques compared with other diseases and fields such as cancer, neuroscience and genomics, to name a few, which could be related to the volume and complexity of the data available in different OA databases as well as large amounts of missing data. Other reasons could be a previous lack of technologies that enable quick and reliable electronic data transfer (such as the Fast and Secure Protocol (FASP))[17], limited computing resources (for example, the need for extremely powerful computers) and a lack of experts in the field of machine learning, as machine learning is a relatively new area in the medical field. Moreover, compared with other diseases (for example, cancer and cardiovascular disease), the big and complete OA databases have become available only in the past decade.

Variables in prediction models

Generally, prediction models are developed using input variables (such as baseline demographic and clinical data) and outcome (or response) variables (such as the presence of knee OA) as assessed by specific methods (for example, Kellgren–Lawrence (KL) grade for classification of knee OA). The resulting model is then used to predict the outcome variables of new data sets (such as patient data). In machine learning, this type of approach is known as supervised learning, as opposed to unsupervised learning, for which only input variables are used to model the data without any corresponding outcome variables (unlabelled data). In addition to supervised and unsupervised learning, there is also semi-supervised learning, in which a small proportion of the data have corresponding outcome variables. Choosing which variables to consider when developing a prediction model is of great importance and can affect how the model performs in longitudinal studies.

Input variables. Knowledge of risk factors can assist in the early detection of knee OA, and these factors are often used as input variables for prediction models (TABLE 1). The risk factors for knee OA can be categorized into six different classes: demographic data, anthropometric features, medical history, biomarkers, imaging assessments and outcomes. During the past two decades, a number of risk factors for the development of knee OA have been identified and confirmed (BOX 2). The risk factors most commonly incorporated into prediction models include age, sex and BMI. In addition, some other factors (such as family history[18,19], ethnicity[19,20], physical activity[18,21,22], knee injury[18,19] and occupational risk[18,19]) have also been considered in different studies. Moreover, a number of models have incorporated miscellaneous risk factors such as soluble vascular cell adhesion protein 1 (VCAM1) expression[23,24] and occupational exposure[19].

One of the main limitations of current models is that mostly conventional risk factors have been included. In order to enhance the predictive accuracy, several other variables and risk factors should be added. These variables include but are not limited to quadriceps strength, oestrogen deficiency, pharmacological treatments, genetic factors, varus or valgus malalignment, genetic predisposition, index-to-ring-finger ratio, yearly income, nutrition, serum biomarkers, MRI markers and other imaging data (such as radiography and ultrasonography data). Identifying the best variables for predicting knee OA outcome is of importance for distinguishing between patients with slow-onset knee OA and those with rapidly progressive knee OA. By identifying the most important variables, cost-effective treatments can be carried out on high-risk patients. In addition, knowing which variables are important to include in prediction models should reduce the number of variables required to measure or assess for each patient, saving considerable time and cost.

Outcome variables. The OA prediction models developed to date included ten different outcomes (TABLE 1). Of these outcomes, KL grade ≥ 2 and the presence of knee OA (including radiographic knee OA), alone or with other outcomes, were most commonly used. However, future prediction studies might elect to use variables that

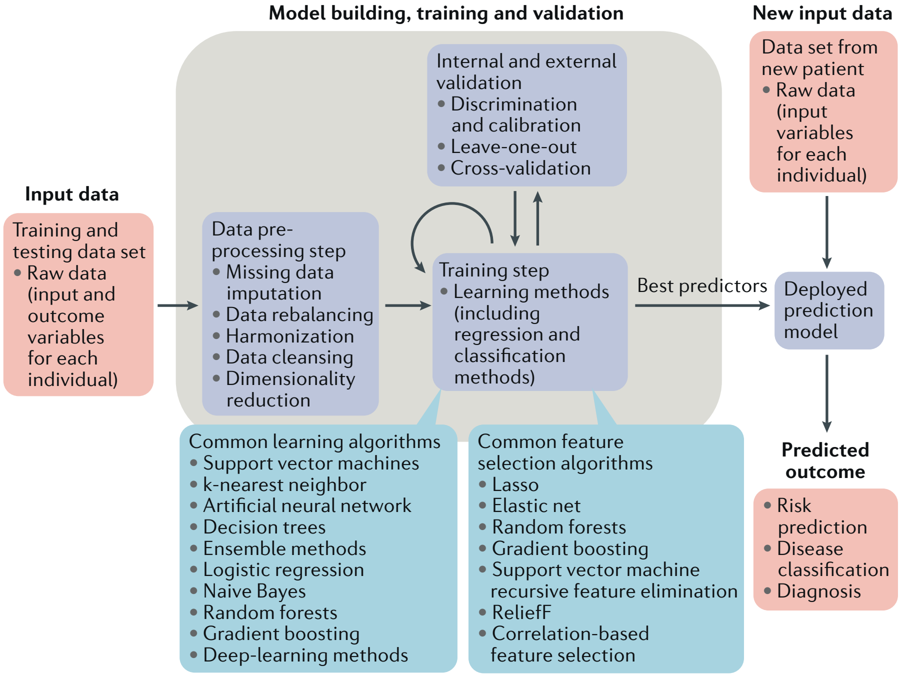

Model building, training and validation
New input data
Input data
Training and testing data set
• Raw data (input and outcome variables for each individual)
Data pre-processing step
• Missing data imputation
• Data rebalancing
• Harmonization
• Data cleansing
• Dimensionality reduction
Internal and external validation
• Discrimination and calibration
• Leave-one-out
• Cross-validation
Training step
• Learning methods (including regression and classification methods)
Best predictors
Data set from new patient
• Raw data (input variables for each individual)
Deployed prediction model
Predicted outcome
• Risk prediction
• Disease classification
• Diagnosis
Common learning algorithms
• Support vector machines
• k-nearest neighbor
• Artificial neural network
• Decision trees
• Ensemble methods
• Logistic regression
• Naive Bayes
• Random forests
• Gradient boosting
• Deep-learning methods
Common feature selection algorithms
• Lasso
• Elastic net
• Random forests
• Gradient boosting
• Support vector machine recursive feature elimination
• ReliefF
• Correlation-based feature selection

Fig. 1 | **A generic scheme for clinical prediction modelling.** Prediction models can be developed using conventional statistical approaches (such as Bayesian belief network). Machine-learning algorithms can also be applied at various stages when developing a prediction model. For both approaches, a training data set (such as raw data from a data set of patients with knee osteoarthritis) is used to build the model. The data are first pre-processed, which can involve data cleansing, data imputation (to account for missing data), dimensionality reduction and the rebalancing and harmonization of the data. Feature selection approaches can be used to select the best variables, which can involve the use of machine-learning algorithms such as Lasso, elastic net and random forest approaches. A training step can be incorporated into the model, such that the model learns from the data to find patterns and match the processed input data with the outcome variables. Machine-learning algorithms commonly used for this step include support vector machines, k-nearest neighbours, artificial neural network, decision trees and ensemble methods. The model can be internally validated or externally validated. The resulting prediction model is then used to predict the outcome (such as the risk, disease classification or diagnosis) of new input data (such as a new patient).

are visible on MRI (such as cartilage loss, meniscal lesions, bone marrow lesions and bone curvature) as outcome variables, as these variables have been strongly associated with progression of OA in clinical trials[25–31]. It should be mentioned that the definition of the outcome variable could relate to the current disease status (cross-sectional), incidence (for example, if in a given time frame the disease status changes from KL grade 0 to KL grade 2 as assessed by radiography, or from healthy cartilage to cartilage degradation as assessed by MRI) or progression (for example, a 0.3 mm per year decrease of joint space width as assessed by radiography or a high level of cartilage degradation as assessed by MRI). Hence, the outcome variable might require data from multiple time points to define. When the outcome variable relates to the incidence or progression of disease, the input data are selected at baseline and the outcome variable is selected from the next time series.

Big data analysis. As with other diseases, the number of parameters included in OA databanks is expanding. With such large amounts of data, the ability to extract useful hidden information is becoming increasingly important. In past OA prediction models, various databases (Supplementary Table 1) were employed as the data source during model development, the most commonly used being the Osteoarthritis Initiative (OAI). As mentioned by Watt et al.[20], the ability of these models to predict disease onset and progression is limited as the variables contributing to OA are both numerous at different time points (follow-up data or time series data) and multifaceted. Furthermore, these models are not well suited to processing the noisy high-dimensional data that are typical of knee OA data sets (for example, MRI data) and new advanced methods are crucially needed. Massive sample sizes and high dimensionality introduce challenges such as overfitting, heterogeneity, noise accumulation, spurious correlation and incidental endogeneity, which make traditional statistical methods inappropriate and unreliable in model development[32]. In order to achieve precise early diagnosis and prognosis, advanced machine-learning and deep-learning approaches are needed to quickly and automatically develop models that can analyse big and complex data and deliver fast and accurate results.

Feature selection. For large and complex data sets, decreasing the number of variables can increase the interpretability of a prediction model. In machine learning and statistics, feature selection (also known as variable selection, attribute selection or variable subset selection) is the process of selecting a subset of relevant variables (dependent and independent variables) for use in prediction model construction. The most important variables are chosen with respect to a given outcome. The selected variables have a vital role in the learning step and for determining the performance of the decision-making step of the model; hence, the variables have to be chosen carefully.

Pre-processing data

The size and quality of the data set, the quality of the selected variables (especially in image analysis) and the choice of outcome(s) have notable effects on the effectiveness of machine-learning-based approaches. Various types of pre-processing can be performed on the data to improve the performance of the developed prediction models.

Class imbalance. The ultimate goal of a prediction model is to assign either a category (known as a class; such as a diagnosis, disease classification or risk group) or a continuous value (such as a risk value) to each item in a collection (for example, the individuals in a data set); these models are known as classification models and regression models, respectively.

Class imbalance refers to a data set in which one or more of the classes occurs more frequently than the others. In such a data set, the most common class is called the majority class, whereas the rarest class is called the minority class[33].

Health-care data sets are often imbalanced (for example, in the general population, many more individuals are likely to be healthy than have a risk of OA); this imbalance can result in erroneous prediction (heavily biased towards the majority class). Sampling-based approaches can be used for rebalancing the data[33–37]. These approaches can be categorized

Table 1 | **Common risk factors incorporated into current models**

Study	Input variables									Outcome variable(s)
	Age	Sex	Family history	Ethnicity	Physical activity	BMI	Knee injury	Occupational risk	Miscellaneous factors	
Watt et al.[20]	x	x	–	x	–	x	–	–	20 medical history variables, 9 physical exam variables, markers (including knee replacement, JSN, baseline symptoms and osteophyte score) and 7 joint variables	• Knee OA • WOMAC score • JSN • Knee pain • Osteophyte score • OA • Baseline symptoms
Schett et al.[23]	x	x	–	–	–	x	–	–	Soluble VCAM1	Incidence of knee or hip replacement surgery for severe OA
Schett et al.[24]	x	x	–	–	–	x	–	–	Soluble VCAM1	Incidence of knee or hip replacement surgery for severe OA
Takahashi et al.[66]	x	x	–	–	–	x	–	–	Risk alleles of three susceptibility genes (*ASPN*, *GDF5* and *DVWA* (also known as *COL6A4P1*))	Knee OA
Zhang et al.[18]	x	x	x	–	x	x	x	x	None	• Incidence of radiographic knee OA (KL grade ≥ 2) • Incidence of symptomatic knee OA (KL grade ≥ 2 plus knee pain) • Progression of knee OA (≥ 1 KL grade increase from baseline)
Kinds et al.[68]	x	x	–	–	–	x	–	–	ESR, pain intensity, WOMAC pain, WOMAC function, KL grade and quantitative radiographic variables	• Radiographic OA: KL grade ≥ 2 • Clinical OA: WOMAC pain score • Clinical OA: WOMAC function score • Clinical OA: knee pain
Kerkhof et al.[67]	x	x	–	–	–	x	–	–	Knee pain, disability index, general health, smoking, educational level, heavy work, genetic score, hand OA, hip OA and KL grade	• OA group: KL grade ≥ 2 • Non-OA group: KL grade < 2
Losina et al.[19]	x	x	x	x	–	x	x	x	Occupational exposure	Individual risk of knee OA and total knee replacement
Yoo et al.[21]	x	x	–	–	x	x	–	–	Educational status (graduated from college), hypertension and knee pain	• Radiographic knee OA • KL grade 2 • KL grade 3 • KL grade 4 • Symptomatic knee OA
Long et al.[71]	x	x	–	–	–	x	x	–	Biomechanical gait parameters	Gait abnormalities characteristic of knee OA in injured populations
Swan et al.[69]	–	–	–	–	–	–	–	–	Proteomic and transcriptomic data	NA

Table 1 (cont.) | **Common risk factors incorporated into current models**

Study	Input variables									Outcome variable(s)
	Age	Sex	Family history	Ethnicity	Physical activity	BMI	Knee injury	Occupational risk	Miscellaneous factors	
Lazzarini et al.[22]	x	–	–	–	x	x	–	–	Imaging variables (active shape modes) extracted from radiographs, soluble biomarkers, pain questionnaires, food questionnaires and physical examinations	• Incidence of 'combined radiographic and clinical ACR criteria' • Incidence of frequent knee pain • Lateral JSN of 1.0 mm • Medial JSN of 1.0 mm • Incidence of KL grade 2
Ashinsky et al.[70]	–	–	–	–	–	–	–	–	Multislice T2-weighted knee images	• Baseline WOMAC score ≤ 10 with a baseline KL grade < 2 (non-progression) • WOMAC score ≤ 10 with a KL grade < 2 and 36-month change in WOMAC > 10 (symptomatic progression of OA)
Minciullo et al.[72]	–	–	–	–	–	–	–	–	Shape, texture and appearance parameters extracted from lateral knee radiographic images	• Non-OA: KL grade 0 or 1 • OA group: KL grade 2–4

ESR, erythrocyte sedimentation rate; JSN, joint space narrowing; KL, Kellgren–Lawrence; NA, not applicable; OA, osteoarthritis; VCAM1, vascular cell adhesion protein 1; WOMAC, Western Ontario and McMaster Universities Arthritis Index; x, input variable included; –, input variable not included.

into three categories: undersampling approaches (in which the minority class data are replicated), oversampling approaches (in which some of the majority class data are removed) and hybrid approaches (a mixture of undersampling and oversampling)[33]. In health care, resampling-based ensemble methods (for example, random forests)[38] are also widely employed for imbalanced data. Ensemble methods create multiple classification models and combine them to produce more accurate results than a single model.

Missing data. Missing data occur when some of the values in a data set are not stored or are invalid. Missing data are a common issue and important to consider as their occurrence can affect the predicted outcome. Imputation methods such as multiple imputation (for example, multivariate imputation by chained equations (MICE) in the R package)[39] and imputation using machine-learning algorithms such as k-nearest neighbour (k-NN) algorithms, support vector machines (SVMs) and regression methods (such as sequential regression multiple imputation (SRMI) or stochastic regression) are among the best solutions for handling missing data. In addition, a variety of software such as SPSS[40], SAS (MI procedure)[41], STATA[42] and WEKA[43] include imputation methods to tackle this problem. However, the right method to use is still controversial and is an area of active research[44].

Data cleansing. Data cleansing involves detecting and correcting (or removing) incorrect, inaccurate or inconsistent parts of the data in order to improve the data quality. This process provides a clean, uniform and consistent data set to enable data harmonization. Several data cleansing tools are available, such as OpenRefine (open source)[45], Wrangler Pro (Trifacta)[46], the Paxata adaptive information platform (Paxata)[47], the data cleansing tool from Alteryx[48], Pandas (an open-source Python data analysis library)[49] and Dplyr (an open-source R package)[50]. Currently, the fastest open-source tool that enables data cleansing in a distributed and robust manner is the software Optimus[51]. This tool is very easy to install, use and understand.

Data harmonization. Data harmonization is the process of combining data from different sources and file formats, both of which are rarely uniform; for example, data sets can have different naming conventions or different numbers of variables[52]. This process can be difficult and should strike a good balance between practicality (harmonizing information that is similar and works together) and purity (harmonizing information that corresponds exactly). The harmonization process involves a number of steps including collecting the data from eligible sources (if eligibility criteria exist for inclusion in the harmonized data set); assessing whether the data and sources are suitable for harmonization; processing the source-specific data into a common (harmonized) format (for example, using open-source software such as Opal[53], Mica[53] or DataSHIELD[54]); and analysing the harmonized data set[55].

The above-mentioned issues (missing data, imbalanced data, data cleansing and data harmonization) are often not considered in the field of knee OA predictive modelling and deserve further attention. Data harmonization methods are beyond

Box 2 | Risk factors in osteoarthritis

Demographics

- Age
- Sex
- Education
- Occupation
- Yearly income category
- Lifestyle
- Living environment
- Nutrition
- Knee symptoms
- Physical activity
- Weight loss
- History of knee injury and/or surgery
- Family history (for example, of knee replacement)

Anthropometric features

- BMI
- Waist circumference

Medical history

- Concomitant joint-affected diseases
- Knee pain
- Knee stiffness
- Pharmacological treatment

Biospecimens

- Serum, plasma and/or urine biomarkers
- Genetic data

Imaging

- Radiographic data
- Quantitative MRI structure assessments

Outcome

- Knee and hip replacements

the scope of this Perspectives article but are discussed in detail elsewhere[52,54,55].

Model evaluation

The performance of a model can be evaluated using data from the same source (internal validation) or an independent source (external validation).

Cross-validation approaches[24] are an accurate method of internal validation. The simplest form of cross validation (the holdout method) is one in which the data set is partitioned into a training data set (which is used for developing the model) and a test data set (which is used to assess the performance of the model). There is no optimum partition between the training and testing set, but researchers usually apply heuristic approaches, such as assigning 80% of the study population to the training set and 20% to the testing set. However, this approach can bias the results and the findings might not be generalizable. *N*-fold cross validation is an alternative approach; in this approach, the data set is partitioned into a number of equal-sized partitions (n), multiple rounds of cross validation are performed (that is, each round uses a different partition as the testing set) and the average result is used to estimate the model's performance.

External validation of a model can improve its generalizability and support the general applicability of a prediction model. In the literature, most researchers have applied calibration and discrimination measures[56] as external validation methods[57].

A major limitation of current prediction models, including those based on machine-learning approaches, relates to the validation approach used, including the size of the populations considered for evaluating the performance of the algorithms. Hence, these models require further validation in order to be applicable in practice. To develop comprehensive prediction models that are useable in other similar populations, external validation (to assess generalizability) is needed[57–60]. So far, only one of the knee OA prediction models developed using machine-learning approaches (Supplementary Table 1) has been externally validated with other data sets[21]. This model performed less well for the external data sets than for the internal data sets[21], highlighting the importance of validation with other cohorts.

When developing a model, a variety of methods can also be used for increasing the generalizability of a model during external validation. For example, one proposed internal–external cross-validation approach[53,61] could be useful for models that use data from a large number of small trials. For this approach, a preliminary model is developed that excludes the data from one of the trials, and the data from the excluded trial are then used to validate the model. This process is then repeated for another preliminary model by excluding data from a different trial. Meta-analysis is then performed on the summarized data of the different preliminary models (for example, to help identify sources of heterogeneity[57], which in meta-analysis refers to the variation in study outcomes between studies measured using statistical tests such as Cochran's Q test and Higgins's I^2 test) before a final model is developed.

Researchers have also proposed a framework for quantifying how the populations used to develop and validate a model relate to each other (such as the distribution and variety of patient characteristics)[62], which might enhance the interpretability of results from validation studies of developed prediction models.

The input and/or outcome variable definitions and the scoring systems used often vary in different populations, which limits the generalizability of developed clinical prediction models[63]. In order to overcome this shortcoming, one could verify whether extensions of the prediction models or modifications of data in external cohorts are possible for the developed prediction model. For example, the prediction model could be adjusted to take into account local and/or contemporary circumstances. A variety of approaches such as calibration-in-the-large or re-calibration methods could be applied for this issue. Strategies for upcoming clinical prediction models have been reviewed in more detail elsewhere[63].

Lessons learnt from previous models

A wide range of prediction models have been developed for diseases other than OA using artificial intelligence tools that have had varying degrees of clinical implementation and utility[15,16,64,65]; however, a paucity of literature exists on models that estimate the risk of developing knee OA. The existing prediction models in OA can be divided into two categories: models developed using conventional statistical approaches[18–20,23,24,66–68] and models developed using machine-learning approaches[21,22,69–72] (Supplementary Table 1). Although in the past decade, a number of prediction models for knee OA that use conventional scoring systems have been developed[18–24,66–72], little progress has been achieved so far with respect to developing prediction models of knee OA using machine-learning methods. Furthermore, only a few of the developed models deal with the prediction of knee OA in patients in the early stages of disease and other studies selected only variables (for use in prediction models) or developed models that were not designed for use in the general population (TABLE 1).

Conventional statistical approaches

One of the first developed knee OA prediction models was based on a Bayesian belief network (BBN)[20], a graphical model that depicts the probabilistic relationships between variables (such as between diseases and symptoms). This model could consistently predict the presence of knee pain and knee OA in the study cohort. However, developing a BBN-based model is extremely demanding and requires a high level of expertise, even if the network structure is already in place. BBN-based models are also limited in their ability to analyse high-dimensional data, and the results of these models can be difficult to interpret. The use of fuzzy cognitive maps (FCMs)[64,65,73], another graphical framework that represents the relationship between variables, is an alternative approach to using a BBN and is widely used in biomedicine. Evolutionary-based FCMs can be applied to predict disease using multivariate time series data[65] (which is not possible using a BBN) and can generate models from raw input data without any expert intervention.

Various other models relating to knee OA have since been developed, the results of which should inform the development of future models. For example, the level of soluble VCAM1 in severe OA is a predictor of joint replacement[23]; hence, two models have incorporated VCAM1 measurements for predicting the risk of joint replacement[23]. In one of the models, prediction was most accurate for patients who required bilateral joint replacement (that is, patients with the most severe form of OA) and was unaffected by the concomitance of other diseases (such as cardiovascular disease or other autoimmune diseases).

The purpose of most knee OA prediction models developed to date has been to predict the risk or presence of knee OA or knee OA progression. In one study, the incorporation of both genetic and clinical information improved the ability of a model to predict the presence of OA compared with the incorporation of genetic information alone[66]. Modifiable risk factors are often incorporated into prediction models. In the three prediction models developed by Zhang et al.[18] (which were among the first models developed to predict the presence and progression of knee OA), the investigators demonstrated that by modifying certain risk factors (for example, obesity), the risk predicted by the models was reduced by a much greater extent than by modifying other factors.

As well as conventional risk factors, the addition of imaging variables has also been tested in various models. For example, Kinds et al.[68] assessed whether and which separate quantitative variables on knee radiographs of individuals with recent-onset knee pain are associated with the presence of radiographic OA and persistence and/or progression of clinical OA during a 5-year follow-up. Incorporating measurements of osteophyte area and minimum joint space width to demographic and clinical characteristics improved the prediction of incident radiographic OA 5 years later. The evaluation of separate quantitative variables performed slightly better than KL grading in predicting clinical OA, whereas radiographic characteristics hardly added to the prediction of clinical OA. In another study, the addition of radiographic variables greatly improved the ability of a model[67] to predict the presence of knee OA in an elderly population whereas the incorporation of easily obtainable questionnaire variables, genetic markers, OA at other joint sites and biochemical markers only modestly improved this model.

The development of online risk calculators might enable the public use of developed prediction models, and online risk calculators are already available for other diseases such as cancer, heart disease and diabetes. Losina et al.[19] developed and assessed the feasibility of a computer-based interactive risk calculator for knee OA (OA Risk C). However, this study had several major shortcomings (Supplementary Table 1). The average lifetime risk estimated by the study participants was 38%, whereas the average risk estimated by the OA Risk C was 25%.

Machine-learning-based approaches

Incorporating machine-learning algorithms. In the past 3 years, researchers have started applying artificial intelligence tools to predict early knee OA. The first model of knee OA to use machine-learning techniques was developed by Yoo et al.[21]; in this study, the investigators developed and validated a self-assessment scoring system and showed that the performance of this model improved considerably by incorporating an artificial neural network (ANN) (TABLE 2).

In 2017, Long et al.[71] developed a prediction model for knee OA that used a k-NN algorithm (TABLE 2). A combination of hip and knee kinetic variables (such as ground reaction force, maximum vertical loading rate, first peak rotation angle and maximum adduction moment) and the quality of life outcome score produced the strongest performing prediction model with the lowest error rate for predicting the risk of knee OA in injured individuals. In addition, the investigators found that individuals with lower limb injury and knee OA had lower Knee Injury and Osteoarthritis Outcome Scores (KOOSs) than asymptomatic individuals. As highlighted by the investigators, this finding gives credence to the idea that KOOSs related to peak knee adduction moment during gait, which is a valid proxy for medial joint loading, are a sensitive measure for predicting those at risk of developing poor knee function over time and could be used in a clinical setting. Additionally, these findings strengthen the idea that alternative diagnostic techniques might be effective, compared with using only self-reported questionnaires or clinical symptom assessment, and provide another option over costly MRI. Moreover, the findings of this study support previous evidence suggesting that worsening knee-related quality of life and knee functionality are linked to peak knee adduction moment, particularly in individuals who have undergone knee surgery. The findings from both Long et al.[71] and Yoo et al.[21] suggest that the incorporation of a machine-learning algorithm such as k-NN or ANN could be a viable cost-effective method if used in conjunction with biomechanical gait analysis for the diagnosis of early knee OA.

Finally, in another study, Minciullo et al.[72] developed two OA-related prediction models using decision trees (TABLE 2). In these two models, the use of a modified version of a random forest outperformed the use of a standard random forest. In a standard random forest, each tree in the forest contains binary decision nodes that decide whether a sample should be passed to one of two nodes (known as leaves). Minciullo et al.[72] demonstrated that making this step less decisive (that is, introducing a soft decision at each node, where some samples might go to multiple leaves) can improve the performance of a prediction model.

Incorporating imaging-based information. Similar to demographic and clinical data, the incorporation of imaging-based information can improve machine-learning-based prediction models[22]. Some of the data used to develop prediction models, such as radiographic data, are not very sensitive (for example, changes in joint space measurements over time and the presence and size of osteophytes) or rely on patient assessments, which could be very subjective and dependent on the population and/or ethnicity (for example, pain). The incorporation of more objective quantitative metrics such as direct and quantitative measurements of imaging variables by MRI will improve the modelling procedure. Using MRI, numerous tissues of the joint can be quantitatively assessed for the classification of individuals[74], as highlighted in the study by Lazzarini et al.[22]. The findings of this study suggest that imaging data could be used in primary care settings[22].

In another study by Ashinsky et al.[70], a machine-learning algorithm (WND-CHRM) was used to select variables of articular cartilage visible by MRI (performed in vivo) that were indicative

Table 2 | **Examples of supervised machine-learning algorithms for disease prediction and/or classification models**

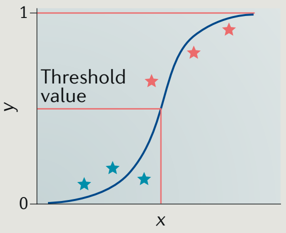
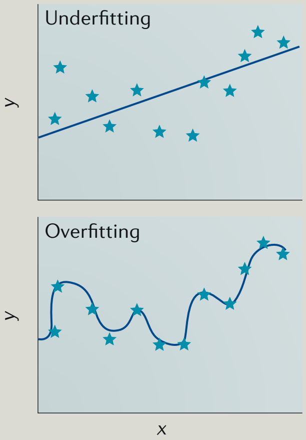
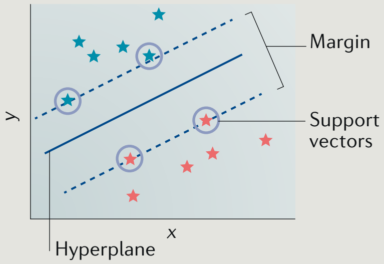
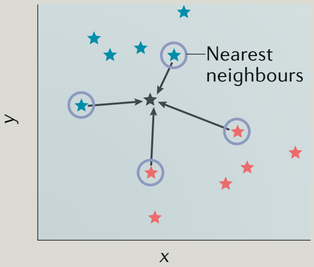
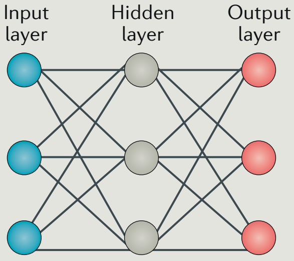
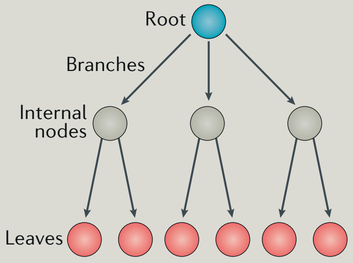
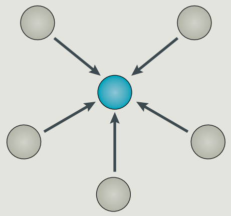
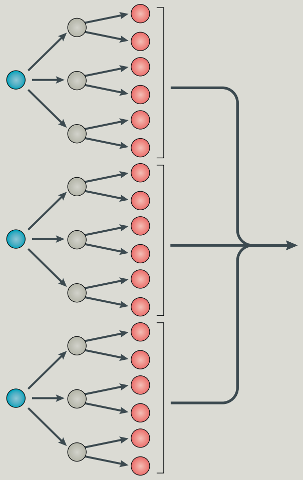
Approach	Type of supervised learning	Description	Graphical depiction
Logistic regression	Classification or regression	• Logistic regression is a conventional statistical technique that can be applied in machine learning. It is typically used for two-class classification problems (that is, data for which the outcome variable is binary) • The goal of this approach is to represent the relationship between a binary outcome variable (*y*) and one or more input variables (*x*) in an equation (which can also be graphically represented) • In the equation, the input variables are combined with coefficients that weight each input variable (*x*), and a logistic function is used to transform the value into a probability between 0 and 1 (*y*). A threshold value is used to assign the binary outcome (0 or 1)	
Lasso regression[a]	Classification or regression	• Lasso regression is a type of regularized linear regression method. Regularized linear regression methods use a linear equation to represent the relationship between the outcome variable (*y*) and input variables (*x*) and impose a penalty to regularize the coefficients (to reduce the complexity of the model and strike a balance between underfitting and overfitting the data) • In Lasso regression, this penalty is equal to the absolute value of the magnitude of coefficients. To reduce this penalty, Lasso regression tends to shrink a set of regression coefficients to zero. As variables with a coefficient of zero are effectively omitted from the model, this method can be used for feature selection (that is, any variables with non-zero regression coefficients are selected)	
Support vector machine	Classification	• The goal of a support vector machine is to identify a hyperplane that best divides the data into the classes • This hyperplane could be a line (for separating 2D data), a plane (for separating 3D data) or a hyperplane (for separating 4D data) • The support vector machine finds the coefficients that result in the best separation of the classes by trying to maximize the margin between the hyperplane and the closest points to the hyperplane	
k-NN	Classification or regression	The k-NN algorithm is a nonparametric method (that is, it makes no assumptions on the underlying data distribution). The algorithm is based on feature similarity (that is, how closely a new item resembles each item in the training set). The item is classified by a majority vote of its neighbours (that is, the new item is assigned to the class most common among its neighbours)	
Artificial neural network	Classification	An artificial neural network consists of units (neurons) arranged in layers, with the aim of converting an input vector into some output. The layers between the input and output layers are often hidden. Each unit takes an input, applies a (often nonlinear) function to it and passes it onto the next layer. Weights are applied to the signals passing from one unit to another, which are modified during the training phase	

Table 2 (cont.) | **Examples of supervised machine-learning algorithms for disease prediction and/or classification models**

Approach	**Type of supervised learning**	**Description**	**Graphical depiction**
Decision tree	Classification	A decision tree builds a model in the form of a tree-like structure to describe the data, where the root (the starting node; representing the input data) connects to leaves (the terminal nodes; representing the class labels) via branches that divide at internal nodes. Each internal node performs a test on the data to decide which branch to move onto	
Naive Bayes	Classification	• Naive Bayes is a classification technique that predicts a class value given a set of variables on the basis of applying Bayes's theorem with the assumption of independence between the variables • In simple terms, the probability of each class for different input values for each variable is calculated during training. For new sets of data, all the variables contribute independently to predicting the probability of the class	
Random forest[a]	Classification or regression	Random forest builds multiple decision trees and merges them together to get a more accurate prediction. The output is the mode of the predicted class (classification) or the mean of the predicted class (regression) of the individual trees	

k-NN, k-nearest neighbours. [a]Can be used for feature selection.

of OA progression, which were then used to develop a disease classification model; the selected variables correlated with the Western Ontario and McMaster Universities Arthritis Index (WOMAC) score, and the developed model had a 75% accuracy in predicting what patients would progress to having symptomatic OA. The model had several limitations (Supplementary Table 1), and the investigators concluded that future versions of the model should incorporate additional cartilage slices in the MRI analysis and combine T2 maps with additional MRI contrast modalities.

Selecting the best variables. Gathered data from numerous variables of knee OA such as patient history, biomarkers and image assessments and, more importantly, the complex interactions between the different variables, are needed to achieve accurate risk prediction and early detection. However, selecting what variables to incorporate into prediction models can be a complex task owing to the great number of variables (both relevant and irrelevant) to choose from.

To address this issue, Swan et al.[69] proposed a heuristic method for feature selection called ranked guided iterative feature elimination (RGIFE) to identify biomarkers of OA, articular cartilage degradation and joint inflammation. RGIFE is an iterative machine-learning-based feature elimination approach in which (in each iteration) the variables are ranked on the basis of their importance in the model and blocks of attributes are removed. This dynamic approach differs from other proposed methods for feature selection that use a static (fixed) approach. The authors obtained 100% classification accuracy by combining the RGIFE feature selection and BioHEL (a rule-based machine-learning method for classifying samples). Swan et al.[69] identified biologically relevant proteins using this feature selection method. This algorithm was used in a later study by Lazzarini et al.[22] to identify biomarkers for use in various models of knee OA. Although no new biomarkers were identified in this later study (owing to a lack in overlap between biomarkers identified for each model), their findings suggest that the evaluation of body fluids of structural degradation products from the extracellular matrix might provide valuable information on the development and prediction of OA.

In these two studies, the researchers used data sets with high numbers of parameters, including proteomics and transcriptomic data with thousands of parameters[69], and the PROOF data set[22]. However, although the study by Lazzarini et al.[22] using the PROOF data set was comprehensive, only 186 variables and 365 individuals were included. Additional comprehensive prediction models for early knee OA should be developed, inspired by this study[22], and should include complete clinical data in addition to income, educational status and quantitative MRI structure assessment variables. Furthermore, future models should also contain much larger numbers of individuals and consider linear and nonlinear interactions between variables. By creating a visual map of the selected variables and performing advanced analysis and modelling, the relationship among selected variables can be identified. Swan et al.[69] applied the cytoscape[75] software platform to visualize and analyse their data sets, but the FCM method could also be applied for this task. Lastly, more advanced machine-learning algorithms are needed to develop a prediction model for such a complex disease as OA.

The identification of new biomarkers and further validation of existing biomarkers have the potential to facilitate DMOAD development and improve treatments. Methods for feature selection will probably become increasingly useful in the field of rheumatology as the use of large 'omics' data sets increases. Swan et al.[69] showed that the RGIFE method in combination with the BioHEL rule-based machine-learning method is more suitable for the analysis of transcriptomic and proteomic OA data than five other machine-learning-based methods (that is, correlating-based feature selection, SVM-recursive feature elimination, random forest feature selection, naive Bayes feature selection and χ^2 feature selection). At present, regularized linear regression machine-learning algorithms such as least absolute shrinkage and selection operator (Lasso) and elastic net[76–80] (TABLE 2) are among the most effective and efficient solutions for feature selection. However, deep-learning-based feature selection is also becoming increasingly used[81–83].

Unlike linear models, deep-learning-based models can take into account the nonlinearity of features and can be used in situations that involve more than two classes (such as multiclass classification as opposed to binary classification)[83]. For example, in one study[81] a feedforward-network-based deep-learning approach enabled the accurate selection of features that could best predict the oestrogen receptor status of patients with breast cancer using metabolomics data. Among six other machine-learning methods (recursive partitioning and regression trees, linear discriminant analysis, SVM, deep learning, random forest, generalized boosted models and prediction analysis for microarrays), the deep-learning approach had the highest accuracy in classifying patients as oestrogen receptor positive or negative based on its selected features. Furthermore, another deep feature selection approach, called SAFS, could outperform other feature selection methods (such as Lasso and random forest methods) in the evaluation and prioritization of risk factors for hypertension in a high-risk demographic subgroup (African-American patients)[82]. Hence, deep-learning-based feature selection is an area of active research.

In addition to the above methods, new feature selection methods that are able to handle multivariate time series data sets such as the Python package tsfresh[84] or software package hctsa[85] could increase the prediction accuracy and interpretability of prediction models.

Encouraging interdisciplinary research collaborations. Experts in the field of knee OA could benefit from looking at other fields that have already switched from using conventional methods for disease classification or prediction to using more advanced methods such as models that incorporate machine learning[64,65,73]. To move towards personalized medicine, interdisciplinary research teams are necessary; OA researchers could benefit

Glossary

Artificial intelligence
The process of creating systems that can learn from experience and adjust to new inputs in order to perform human-like tasks. Machine-learning is a fundamental concept of artificial intelligence.

Calibration
Calibration measurements represent the level of accuracy of a model in estimating the absolute risk (that is, the agreement between the observed and predicted risk). Poorly calibrated models will underestimate or overestimate the outcome of interest.

Classification models
In statistics and machine-learning, classification is the process of identifying the category of a new observation on the basis of a training set of data containing observations for which the category (outcome value) is known. In the field of osteoarthritis, an example could be classification of patients into slow progressors and fast progressors on the basis of several input variables.

Deep-learning
A subfield of machine-learning that is based on advanced artificial neural networks; this field has enabled doctors in different fields of medicine to obtain a precise 3D understanding of 2D images.

Discrimination
Discrimination measurements identify to what extent a model discriminates items of different classes (for example, individuals with disease and without disease). For binary outcomes, the receiver operating characteristic curve or C-statistic could be applied for discrimination measurement.

Feature selection
Feature selection refers to the process of obtaining a subset of variables from an original set of variables according to certain feature selection criteria. The feature selection step precedes the learning step of a prediction model and good feature selection results can improve the learning accuracy, reduce learning time and simplify learning results.

Generalizability
Refers to the accuracy with which a prediction model developed from one study population can be used for the population at large.

Imputation
In machine-learning and statistics, imputation is the process of replacing missing data with substituted values to avoid bias or inaccuracies in the results.

Interpretability
Model interpretability describes the ability of the user to understand the model, which includes understanding the relationships between the input and outcome variables (for example, knowing how the selected input variables contribute to the outcome variable).

Regression models
Regression is the process of identifying the value of a new observation on the basis of a training set of data containing observations for which the category (outcome value) is known. In the field of osteoarthritis, an example could be predicting the probability of disease.

Semi-supervised learning
Semi-supervised learning is typically when only a small amount of data are labelled (that is, have both input and output variables) and a large amount are unlabelled (that is, have only input data); this method falls between unsupervised learning and supervised learning.

Supervised learning
Supervised learning is where you have input variables (x) and an output variable (y) and use an algorithm to learn the mapping function from the input to the output $y = f(x)$.

Training
The training for machine learning involves providing a machine-learning algorithm with training data (input and outcome variables) to learn from. The learning algorithm finds patterns in the training data such that the input parameters correspond to the target. Machine-learning models are applied to do predictions on new data for which the outcome value is not known (for example, to determine to which class the new observation belongs).

Unsupervised learning
In unsupervised learning, only input data (x) exist and there are no corresponding output variables. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

from collaborating with others in the field of computer and data science with expertise, for example, in data mining, machine learning and image processing. In addition, research funding agencies, universities and governments are encouraged to connect researchers from diverse scientific backgrounds (including bioinformatics, biomathematics and biostatistics) on topics related to OA and personalized medicine. For example, a large European interdisciplinary research team completed the D-BOARD project in 2017. In this 5-year project, European researchers identified novel diagnostics, genes and biomarkers (by analysing proteomic, metabolomic, genomic and transcriptomic profiles) that could help diagnose OA at an early stage. In the ongoing APPROACH project, another European interdisciplinary team is working together to combine biomedical data from more than 10,000 individuals with and without OA into a unified bioinformatics platform with the aim of identifying different OA phenotypes.

Conclusions

Early diagnosis of knee OA and the ability to track disease progression is challenging. Progress is needed to help knee OA physicians and scientists make decisions on the basis of massive data sets within a short time. Accurate predictive modelling for OA progression might be difficult to achieve without sensitive imaging techniques that can detect early changes before morphological alterations are detectable. In addition to the quantitative MRI assessment of articular tissues, compositional MRI techniques for cartilage (for example, T2 and T1ρ relaxometry) could in principle uncover such early changes in this tissue. However, these techniques have limitations that need to be overcome before they can be commonly used; for example, these techniques are sensitive to regional variation in the tissue, T2 is prone to an artefact named magic angle effect (which is a potential source of diagnostic error) and T1ρ could be difficult to implement in the clinical setting as this technique requires dedicated hardware and software. Moreover, alterations in the cartilage might not be the earliest changes that occur in OA, and bone curvature changes assessed by quantitative MRI seem to precede cartilage loss[30,31].

Efficient and reliable screening of patients with early OA and patients who will progress rapidly using prediction models is important, not only from a medical and patient standpoint but also for the pharmaceutical industry, scientific community and society in general. Such screening could be used as a tool to guide clinical decision-making, representing a major advance towards attaining precision medicine, which in turn will also help to distinguish the responders from the non-responders of a given therapy. For the scientific community, effective prediction models should boost research into drugs and drug targets and the design and development of effective and specific personalized therapeutic interventions for these patients. The models should also enable substantial savings in medical resources and societal costs by reducing patient morbidity and improving the quality of life of patients. Moreover, research in other complex, slow and unpredictable diseases could benefit from fine-tuning such developed prediction models.

In developing prediction models, interdisciplinary research teams are needed. Furthermore, to increase the prediction accuracy and interpretability of OA knee prediction models, data mining approaches should include advanced machine-learning algorithms, multivariate time series data sets, nonlinear feature selection methods and other imaging variables such as MRI.

*Afshin Jamshidi, Jean-Pierre Pelletier and Johanne Martel-Pelletier**

Osteoarthritis Research Unit, University of Montreal Hospital Research Centre (CRCHUM), Montreal, Quebec, Canada.

**e-mail: jm@martelpelletier.ca*

https://doi.org/10.1038/s41584-018-0130-5
Published online 6 December 2018

1. Arden, N. & Cooper, C.in *Osteoarthritis Handbook* (Taylor & Francis, London, 2006).
2. McGuire, D. A., Carter, T. R. & Shelton, W. R. Complex knee reconstruction: osteotomies, ligament reconstruction, transplants, and cartilage treatment options. *Arthroscopy* **18**, 90–103 (2002).
3. Cooper, C. & Arden, N. K. Excess mortality in osteoarthritis. *BMJ* **342**, d1407 (2011).
4. Hochberg, M. C. Mortality in osteoarthritis. *Clin. Exp. Rheumatol* **26**, S120–S124 (2008).
5. Bitton, R. The economic burden of osteoarthritis. *Am. J. Manag. Care* **15**, S230–S235 (2009).
6. Prieto-Alhambra, D. et al. Incidence and risk factors for clinically diagnosed knee, hip and hand osteoarthritis: influences of age, gender and osteoarthritis affecting other joints. *Ann. Rheum. Dis.* **73**, 1659–1664 (2014).
7. Martel-Pelletier, J. et al. Osteoarthritis. *Nat. Rev. Dis. Primers* **2**, 16072 (2016).
8. Blagojevic, M., Jinks, C., Jeffery, A. & Jordan, K. P. Risk factors for onset of osteoarthritis of the knee in older adults: a systematic review and meta-analysis. *Osteoarthritis Cartilage* **18**, 24–33 (2010).
9. Zhang, W. Risk factors of knee osteoarthritis — excellent evidence but little has been done. *Osteoarthritis Cartilage* **18**, 1–2 (2010).
10. McWilliams, D. F., Leeb, B. F., Muthuri, S. G., Doherty, M. & Zhang, W. Occupational risk factors for osteoarthritis of the knee: a meta-analysis. *Osteoarthritis Cartilage* **19**, 829–839 (2011).
11. Raynauld, J. P. et al. Long term evaluation of disease progression through the quantitative magnetic resonance imaging of symptomatic knee osteoarthritis patients: correlation with clinical symptoms and radiographic changes. *Arthritis Res. Ther.* **8**, R21 (2006).
12. Solomon, D. H. et al. The comparative safety of analgesics in older adults with arthritis. *Arch. Intern. Med.* **170**, 1968–1978 (2010).
13. Marx, V. Biology: the big challenges of big data. *Nature* **498**, 255–260 (2013).
14. Dolinski, K. & Troyanskaya, O. G. Implications of big data for cell biology. *Mol. Biol. Cell* **26**, 2575–2578 (2015).
15. Cintolo-Gonzalez, J. A. et al. Breast cancer risk models: a comprehensive overview of existing models, validation, and clinical applications. *Breast Cancer Res. Treat.* **164**, 263–284 (2017).
16. Cosma, G., Brown, D., Archer, M., Khan, M. & Pockley, A. G. A survey on computational intelligence approaches for predictive modeling in prostate cancer. *Expert Syst. Appl.* **70**, 1–19 (2017).
17. Fast and Secure protocol — FASP (Aspera, Inc., Emeryville, CA, USA).
18. Zhang, W. et al. Nottingham knee osteoarthritis risk prediction models. *Ann. Rheum. Dis.* **70**, 1599–1604 (2011).
19. Losina, E., Klara, K., Michl, G. L., Collins, J. E. & Katz, J. N. Development and feasibility of a personalized, interactive risk calculator for knee osteoarthritis. *BMC Musculoskelet. Disord.* **16**, 312 (2015).
20. Watt, E. W. & Bui, A. A. Evaluation of a dynamic Bayesian belief network to predict osteoarthritic knee pain using data from the osteoarthritis initiative. *AMIA Annu. Symp. Proc.* **2008**, 788–792 (2008).
21. Yoo, T. K., Kim, D. W., Choi, S. B., Oh, E. & Park, J. S. Simple scoring system and artificial neural network for knee osteoarthritis risk prediction: a cross-sectional study. *PLoS ONE* **11**, e0148724 (2016).
22. Lazzarini, N. et al. A machine learning approach for the identification of new biomarkers for knee osteoarthritis development in overweight and obese women. *Osteoarthritis Cartilage* **25**, 2014–2021 (2017).
23. Schett, G. et al. Vascular cell adhesion molecule 1 as a predictor of severe osteoarthritis of the hip and knee joints. *Arthritis Rheum.* **60**, 2381–2389 (2009).
24. Schett, G., Zwerina, J., Axmann, R., Willeit, J. & Stefan, K. Risk prediction for severe osteoarthritis. *Ann. Rheum. Dis.* **69**, 1573–1574 (2010).
25. Berthiaume, M. J. et al. Meniscal tear and extrusion are strongly associated with the progression of knee osteoarthritis as assessed by quantitative magnetic resonance imaging. *Ann. Rheum. Dis.* **64**, 556–563 (2005).
26. Raynauld, J. P. et al. Correlation between bone lesion changes and cartilage volume loss in patients with osteoarthritis of the knee as assessed by quantitative magnetic resonance imaging over a 24-month period. *Ann. Rheum. Dis.* **67**, 683–688 (2008).
27. Tanamas, S. K. et al. Bone marrow lesions in people with knee osteoarthritis predict progression of disease and joint replacement: a longitudinal study. *Rheumatology* **49**, 2413–2419 (2010).
28. Raynauld, J. P. et al. Risk factors predictive of joint replacement in a 2-year multicentre clinical trial in knee osteoarthritis using MRI: results from over 6 years of observation. *Ann. Rheum. Dis.* **70**, 1382–1388 (2011).
29. Pelletier, J. P. et al. What is the predictive value of MRI for the occurrence of knee replacement surgery in knee osteoarthritis? *Ann. Rheum. Dis.* **72**, 1594–1604 (2013).
30. Neogi, T. et al. Magnetic resonance imaging-based three-dimensional bone shape of the knee predicts onset of knee osteoarthritis: data from the osteoarthritis initiative. *Arthritis Rheum.* **65**, 2048–2058 (2013).
31. Raynauld, J. P. et al. Bone curvature changes can predict the impact of treatment on cartilage volume loss in knee osteoarthritis: data from a 2-year clinical trial. *Rheumatology* **56**, 989–998 (2017).
32. Fan, J., Han, F. & Liu, H. Challenges of big data analysis. *Natl Sci. Rev.* **1**, 293–314 (2014).
33. Haixiang, G. et al. Learning from class-imbalanced data: review of methods and applications. *Expert Syst. Appl.* **73**, 220–239 (2017).
34. Fu, X., Wang, L., Chua, K. S. & Chu, F. Training RBF Neural Networks on Unbalanced Data. *Proc. 9th Int. Conf. Neural Inform. Processing (ICONIP'02)* **2**, 1016–1020 (2002).
35. Wasikowski, M. & Chen, X. W. Combating the small sample class imbalance problem using feature

selection. *IEEE Trans. Knowl. Data Eng.* **22**, 1388–1400 (2010).

36. Khalilia, M., Chakraborty, S. & Popescu, M. Predicting disease risks from highly imbalanced data using random forest. *BMC Med. Inform. Decis. Mak.* **11**, 51 (2011).
37. Wang, K. J., Makond, B. & Wang, K. M. An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data. *BMC Med. Inform. Decis. Mak.* **13**, 124 (2013).
38. Ozcift, A. Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. *Comput. Biol. Med.* **41**, 265–271 (2011).
39. van Buuren, S. & Groothuis-Oudshoorn, K. MICE: multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**, 1–68 (2011).
40. IBM SPSS Statistics for Windows, version 25.0, released 2017 (IBM Corp., Armonk, NY, USA).
41. SAS/STAT® version 14.1 (SAS Institute Inc., Cary, NC, USA).
42. STATA Statistical Software, release 15, 2017 (StataCorp LLC, College Station, TX, USA).
43. Frank, E., Hall, M. A. & Witten, I. H. The WEKA workbench: online appendix for data mining: practical machine learning tools and techniques. *UoW* https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf (2016).
44. Zhang, Z. Missing data imputation: focusing on single imputation. *Ann. Transl Med.* **4**, 9 (2016).
45. Verborgh, R. & De Wilde, M. *Using OpenRefine* (Packt Publishing Ltd., Burmingham, UK, 2013).
46. Trifacta. Data wrangling tools & software. *Trifacta* https://www.trifacta.com (2018).
47. Paxata, Inc. Self-service data preparation for data analytics. *Paxata* https://www.paxata.com (2018).
48. Baruti, R. *(ed.) Learning Alteryx: A Beginner's Guide to Using Alteryx for Self-Service Analytics and Business Intelligence* (Packt Publishing Ltd., Birmingham, UK, 2017).
49. McKinney, W. pandas: a foundational python library for data analysis and statistics. *DLR* http://www.dlr.de/sc/Portaldata/15/Resources/dokumente/pyhpc2011/submissions/pyhpc2011_submission_9.pdf (2011).
50. OBiBa. Open source software for epidemiology. *OBiBa* http://www.obiba.org (2018).
51. Optimus Company. Data cleansing and exploration made simple. *Optimus* https://hioptimus.com (2018).
52. Griffith, L. E. et al. Statistical approaches to harmonize data on cognitive measures in systematic reviews are rarely reported. *J. Clin. Epidemiol.* **68**, 154–162 (2015).
53. Royston, P., Parmar, M. K. & Sylvester, R. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. *Stat. Med.* **23**, 907–926 (2004).
54. Doiron, D. et al. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. *Emerg. Themes Epidemiol.* **10**, 12 (2013).
55. Doiron, D., Raina, P., Ferretti, V., L'Heureux, F. & Fortier, I. Facilitating collaborative research: implementing a platform supporting data harmonization and pooling. *Nor. Epidemiol.* **21**, 221–224 (2012).
56. Alba, A. C. et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *JAMA* **318**, 1377–1384 (2017).
57. Steyerberg, E. W. & Harrell, F. E. Jr. Prediction models need appropriate internal, internal-external, and external validation. *J. Clin. Epidemiol.* **69**, 245–247 (2016).
58. Siontis, G. C., Tzoulaki, I., Castaldi, P. J. & Ioannidis, J. P. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J. Clin. Epidemiol.* **68**, 25–34 (2015).
59. Tugwell, P. & Knottnerus, J. A. Clinical prediction models are not being validated. *J. Clin. Epidemiol.* **68**, 1–2 (2015).
60. Tugwell, P. & Knottnerus, J. A. Transferability/generalizability deserves more attention in 'retest' studies in diagnosis and prognosis. *J. Clin. Epidemiol.* **68**, 235–236 (2015).
61. Debray, T. P., Moons, K. G., Ahmed, I., Koffijberg, H. & Riley, R. D. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat. Med.* **32**, 3158–3180 (2013).
62. Debray, T. P. et al. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J. Clin. Epidemiol.* **68**, 279–289 (2015).
63. Steyerberg, E. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating* (Springer New York, 2010).
64. Papageorgiou, E. I., Subramanian, J., Karmegam, A. & Papandrianos, N. A risk management model for familial breast cancer: a new application using Fuzzy Cognitive Map method. *Comput. Methods Programs Biomed.* **122**, 123–135 (2015).
65. Froelich, W., Papageorgiou, E. I., Samarinas, M. & Skriapas, K. Application of evolutionary fuzzy cognitive maps to the long-term prediction of prostate cancer. *Appl. Soft Comput.* **12**, 3810–3817 (2012).
66. Takahashi, H. et al. Prediction model for knee osteoarthritis based on genetic and clinical information. *Arthritis Res. Ther.* **12**, R187 (2010).
67. Kerkhof, H. J. et al. Prediction model for knee osteoarthritis incidence, including clinical, genetic and biochemical risk factors. *Ann. Rheum. Dis.* **73**, 2116–2121 (2014).
68. Kinds, M. B. et al. Evaluation of separate quantitative radiographic features adds to the prediction of incident radiographic osteoarthritis in individuals with recent onset of knee pain: 5-year follow-up in the CHECK cohort. *Osteoarthritis Cartilage* **20**, 548–556 (2012).
69. Swan, A. L. et al. A machine learning heuristic to identify biologically relevant and minimal biomarker panels from omics data. *BMC Genomics* **16**, S2 (2015).
70. Ashinsky, B. G. et al. Predicting early symptomatic osteoarthritis in the human knee using machine learning classification of magnetic resonance images from the osteoarthritis initiative. *J. Orthop. Res.* **35**, 2243–2250 (2017).
71. Long, M. J., Papi, E., Duffell, L. D. & McGregor, A. H. Predicting knee osteoarthritis risk in injured populations. *Clin. Biomech.* **47**, 87–95 (2017).
72. Minciullo, L., Bromiley, P. A., Felson, D. T. & Cootes, T. F. Indecisive trees for classification and prediction of knee osteoarthritis. *8th Int. Workshop MLMI 2017 MICCAI 2017 Proc.* **10541**, 283–290 (2017).
73. Jamshidi, A., Ait-kadi, D., Ruiz, A. & Rebaiaia, M. L. Dynamic risk assessment of complex systems using FCM. *Int. J. Prod. Res.* **56**, 1070–1088 (2017).
74. Meher, S. K. & Pal, S. K. Rough-wavelet granular space and classification of multispectral remote sensing image. *Appl. Soft Comput.* **11**, 5662–5673 (2011).
75. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
76. Hastie, T., Tibshirani, R. & Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations* (Chapman and Hall/CRC, 2015).
77. Meinshausen, N. & Bühlmann, P. High-dimensional graphs and variable selection with the lasso. *Ann. Stat.* **34**, 1436–1462 (2006).
78. Huang, J., Ma, S. & Zhang, C. H. Adaptive Lasso for sparse high-dimensional regression models. *Stat. Sin.* **18**, 1603–1618 (2008).
79. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. A. Sparse-group lasso. *J. Comput. Graph. Stat.* **22**, 231–245 (2013).
80. Friedman, J. et al. Package 'glmnet'. *The Comprehensive R Archive Network* https://cran.r-project.org/web/packages/glmnet/glmnet.pdf (2018).
81. Alakwaa, F. M., Chaudhary, K. & Garmire, L. X. Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data. *J. Proteome Res.* **17**, 337–347 (2018).
82. Nezhad, M. Z., Zhu, D., Li, X., Yang, K. & Levy, P. SAFS: a deep feature selection approach for precision medicine. Preprint at arXiv https://arxiv.org/abs/1704.05960 (2017).
83. Li, Y., Chen, C. Y. & Wasserman, W. W. Deep feature selection: theory and application to identify enhancers and promoters. *J. Comput. Biol.* **23**, 322–336 (2016).
84. Christ, M., Braun, N., Neuffer, J. & Kempa-Liehr, W. A. Time series feature extraction on basis of scalable hypothesis tests (tsfresh — a Python package). *Neurocomputing* **307**, 72–77 (2018).
85. Fulcher, B. D. & Jones, N. S. *hctsa*: a computational framework for automated time-series phenotyping using massive feature extraction. *Cell Syst.* **5**, 527–531 (2017).

Acknowledgements

The authors thank V. Wallis for her assistance with manuscript preparation.

Author contributions

J.M.-P. and A.J. researched data for the article. All authors wrote the article, made substantial contribution to discussions of the content and reviewed and/or edited the manuscript before submission.

Competing interests

The authors declare no competing interests.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information

Supplementary information is available for this paper at https://doi.org/10.1038/s41584-018-0130-5.

RELATED LINKS

D-BOARD: https://cordis.europa.eu/project/rcn/105314_en.html
APPROACH: https://approachproject.eu
DataSHIELD: http://www.datashield.ac.uk/
Automatic Image Registration: http://air.bmap.ucla.edu/AIR5
tsfresh: http://tsfresh.readthedocs.io/en/latest
hctsa: https://github.com/benfulcher/hctsa