Use of machine learning in osteoarthritis research: a systematic literature review

Marie Binvignat,1,2,3 Valentina Pedoia,4 Atul J Butte,2 Karine Louati,1 David Klatzmann,3,5 Francis Berenbaum,1 Encarnita Mariotti-Ferrandiz,3 Jérémie Sellam1

ABSTRACT

Objective The aim of this systematic literature review was to provide a comprehensive and exhaustive overview of the use of machine learning (ML) in the clinical care of osteoarthritis (OA).

Methods A systematic literature review was performed in July 2021 using MEDLINE PubMed with key words and MeSH terms. For each selected article, the number of patients, ML algorithms used, type of data analysed, validation methods and data availability were collected.

Results From 1148 screened articles, 46 were selected and analysed; most were published after 2017. Twelve articles were related to diagnosis, 7 to prediction, 4 to severity and 11 to progression. The number of patients included ranged from 18 to 5749. Overall, 35% of the articles described the use of deep learning and 74% imaging analyses. A total of 85% of the articles involved knee OA and 15% hip OA. No study investigated hand OA. Most of the studies involved the same cohort, with data from the OA initiative described in 46% of the articles and the MOST and Cohort Hip and Cohort Knee cohorts in 11% and 7%. Data and source codes were described as publicly available respectively in 54% and 22% of the articles. External validation was provided in only 7% of the articles.

Conclusion This review proposes an up-to-date overview of ML approaches used in clinical OA research and will help to enhance its application in this field.

INTRODUCTION

The development of artificial intelligence (AI), especially machine learning (ML), in healthcare has led to important improvements and discoveries, notably in rheumatology and osteoarthritis (OA).1–3 There are many definitions for AI, but it could be summarised as the ability for a computer system to perform intellectual tasks normally requiring human skills.

AI includes ML,4,5 defined as the ability to ‘learn’ or progressively improving performance from data. ML methods can be supervised or unsupervised. In supervised analysis, outcomes are known, and data are labelled. Conversely, in unsupervised ML, the outcomes and data are unknown and unlabeled. Two additional categories have been further proposed: semisupervised learning and reinforcement learning, with the outcome only partially known.6 Semisupervised learning models consist of a mix of labelled and unlabeled data and are based on weak supervision, with limited labelled data used to provide information and supervision for unlabeled data. However, reinforcement learning is an ML paradigm in which learning occurs iteratively via a series of trial-and-error cycles to maximise the reward received after each trial and therefore improve the learning. Supervised ML methods are the most commonly used in medicine and healthcare.7 Among supervised ML, we can distinguish different algorithms such as random forest, support vector machine, and convolutional neural

Key messages

What is already known about this subject?
► This is the first systemic literature review of machine learning and osteoarthritis.

What does this study add?
► Most (85%) of the machine learning articles focused on knee osteoarthritis, and radiological data investigation predominated clearly over clinical or biological data.
► Almost half of the selected articles described use of the osteoarthritis initiative database, and external validation was poorly used (7% of the articles).

How might this impact on clinical practice or further developments?
► Application of machine learning is needed in other sites of osteoarthritis such as the hand or foot osteoarthritis, and new cohorts need to be established.
► Improving reproducibility and understanding of machine learning in the osteoarthritis field is needed.
networks according to the type of analyses (figure 1). Deep learning (DL) is a subtype of ML based on multiple layers of a neuron-architecture network allowing the model to improve and train itself and leading to high accuracy via high-level feature extraction from data.

In these ML algorithms, we can distinguish explainable ML models (eg, linear models, naïve Bayes, logistic regression) from unexplainable ML models (eg, decision tree models, neural network, support vector machine), also known as interpretable ML or ‘white-box’ models for which the results of the algorithm can be understood by human intelligence. In contrast, unexplainable ML models (or ‘black-box models’) are algorithms for which, theoretically, one cannot possibly explain how and why the algorithm achieved a specific decision. Interpretability tools and methods developed to improve the models explainability include gradient-based methods for convolutional neural networks (eg, gradient class activation map), shapley additive explanation for decision-tree models and local interpretable model-agnostic explanation. These interpretability tools are important because they help determine which features contribute to a specific model decision.

In supervised ML, usually, the data used are separated in two parts with a training dataset to teach the algorithm and a testing dataset to test the performance of the model. Performance in ML is evaluated with different prediction metrics such as accuracy, sensitivity, specificity and precision. Finally, a step of validation is needed to assess the reproducibility of the dataset and avoid overfitting, which can be applied by k-fold cross-validation, bootstrap, leave one-out or splitting dataset, or using external data. Validation plays a key role in ML study because reproducibility remains one of the main critical issues and challenges in ML.

In rheumatology, ML analyses have improved our knowledge of patient trajectories via disease and care modelling as well as response to treatment or disease phenotyping predictions with immunological signatures. ML is also applied in the field of OA, especially with the establishment of large cohorts such as the OA Initiative (OAI), an observational cohort study of knee OA; the Multi-Centre Osteoarthritis Study (MOST), a longitudinal prospective and observational study of knee OA; and the Cohort Hip and Cohort Knee (CHECK), a prospective observational cohort of knee and hip OA. However, despite its trending and increasing applications, ML remains an emerging field with incredible potential but also limitations. Thus, a better delineation and understanding of the ML methods used in OA is needed.

The aim of this systemic literature review was to give a comprehensive overview of ML in clinical OA.

**MATERIALS AND METHODS**

**Information sources and search strategy**

The systemic literature review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-analyses guidelines and was registered in the International Prospective Register of Systematic Review PROSPERO (CRD42021272975). Articles in MEDLINE PubMed were searched beginning on 9 July 2021, by using the following MeSH and standard terms ((human [MeSH Terms]) AND (osteoarthritis [MeSH Terms]) AND ((algorithms [MeSH Terms]) OR (machine learning) OR (“information systems”[MeSH Terms]) OR (“artificial intelligence”[MeSH Terms]) OR (artificial intelligence)). The choice of these terms was motivated by the complexity of the definition of ML and by a willingness to be as exhaustive as possible. The definition of ML was based on classification and algorithms listed by scikit-machine learning module documentation.

**Eligibility criteria**

We searched for and included only original articles using AI and ML algorithms with clinical application in human OA. We excluded articles in a language other than English and articles related to surgery (especially those related to robotics and outcomes after total knee replacement); articles focused on locomotor metrics related to physical therapy outcomes; articles related to therapeutics, spine OA and temporo-mandibular OA; basic
research articles and/or studies using murine models; basic cellular or molecular biology articles; and articles related to basic and fundamental imaging as well as theoretical ML (table 1).

**Article selection and data extraction**

Two article selection steps based on eligible criteria have been used. A first selection was based on abstracts and a second selection on full-text articles. The final choice of articles was independently validated by the three coauthors. Data have been extracted data by using a csv file extraction from the National Center for Biotechnology Information (NCBI) database.

For each article we collected the following:

- The domain of application of the article: diagnosis, prediction, phenotyping, severity, and progression of OA.
- Number of patients, year of publication, localisation of OA (knee, hand, foot and/or hip), main ML method of analysis and the notion of supervised and unsupervised analysis, use of DL, explainable ML and interpretability tools, type of data analysed (clinical, biological and imaging data), name of the cohort, presence of testing and training dataset, validation method when used and data and source code availability.

**Statistical analysis**

Descriptive analyses, tabulation, visual display of the results and subgroup analysis were performed with R V.4.1.1 (2021-08-10). Graphical visualisation involved using the software BioRender and Affinity Designer.

**RESULTS**

**Selection flow chart**

We retrieved 1148 articles from the search. The flow chart is in figure 2. In the first selection step based on the abstracts for 1148 articles, 956 articles were excluded, including 196 reviews, 240 articles related to surgery, 134 on fundamental and theoretical imaging, 43 on reeducation, 57 related to treatment, 42 on basic research and 60 related to other diseases, and 192 articles remained. After reading complete articles, we excluded 43 articles related to theoretical radiology, 9 to surgery outcomes and robotics, 10 to reeducation outcomes, 45 to molecular biology and 10 to other diseases. We finally selected

---

**Table 1 Inclusion and exclusion criteria of the systemic literature review**

| Inclusion criteria | Exclusion criteria |
|--------------------|-------------------|
| OA                 | Review and meta-analysis |
| Human              | Non-clinical OA articles |
| Machine learning algorithms | - Surgery. |
|                    | - Non-applied radiology. |
|                    | - Physical therapy. |
|                    | - Treatments. |
|                    | Experimental OA |
|                    | - Molecular biology. |
|                    | - Murine model. |
|                    | - Cell biology. |
|                    | Temporo-mandibular OA |
|                    | Spine OA |
|                    | Non-available articles |
|                    | - Full text not available. |
|                    | - Non-English articles. |

OA, osteoarthritis.
and analysed 46 articles (online supplemental table S1).

**Systematic literature overview**

Among the 46 identified articles published between 2007 and 2021, 74% were published after 2017 (figure 3); 12 were devoted to the diagnosis of OA at an early stage, 7 to the prediction of developing OA in healthy volunteers, 4 to the identification of OA phenotypes, 12 an automated estimation of OA structural severity classification and 11 to the identification of the progression of OA, notably patients with rapid disease progression. A complete descriptive analysis of the review is summarised in table 2 and online supplemental table S2.

**Number of patients**

The number of patients in each study was heterogeneous. Indeed, the mean number of patients was 1359 and median 525 (range 18–5749). Studies related to OA severity had a high number of patients, with mean 1803 and median 942 (range 18–4504); and studies related to phenotypes had the smallest number of patients, with mean 518 and median 559 (range 102–52). The first studies of ML in the field of OA were mainly related to diagnosis and OA prediction (median publication year 2017 (range 2008–2020)), whereas recent articles mainly focused on phenotype identification (median year 2018 (range 2015–2019)), and disease progression (median year 2019 (range 2012–2020)).

**Source of the data**

Most of the analyses involved knee OA, 85% (N=39) of the articles, whereas only 15% (N=7) involved hip OA. No study investigated hand or foot OA. Overall, 64% of the articles described use of the OAI, the MOST and CHECK cohorts. The OAI database was described in 46% (N=21) of articles, the MOST database in 11% (N=5) and the CHECK database in 7% (N=3).

**ML methods**

ML methods included the use of supervised algorithms, in 87% of articles (N=40), and unsupervised and semi-supervised algorithms, in 9% (N=4) and 4% (N=2). The most frequently used algorithms were convolutional neural network approaches. In total, 80% of the supervised algorithms were linked to classifications (N=32) and 20% (N=8) to regression analysis. Mixed ML algorithms were described in 28% (N=13) of articles. The ML methods differed according to the field of interest.

- In diagnosis articles, methods with convolutional neural network and random forest were predominant.
- Studies related to OA prediction involved methods such as elastic net regularisation and the multipurpose image classifier method: weighted neighbor distance using compound hierarchy of algorithms representing morphology (WND-CHARM).
- Studies related to phenotypes involved unsupervised methods such as latent cluster analysis.
- Studies related to estimation of OA severity involved methods such as convolutional neural network and densely connected convolutional neural network.
- In progression-related articles, 45% (N=5) described methods derived from logistic regression.

**Type of data**

The main type of data studied was imaging, in 74% (N=34) of articles; 61% (N=28) of the articles described analysis of X-ray data and 22% (N=10) MRI. Overall, 41% (N=19) of articles described analysis of clinical data, mainly demographic data and OA evaluation scores. Only 15% of the articles (N=7) described analysis of biological data in ML models, including 13% (N=6) patient serum and 4% (N=2) synovial fluid. Most of these studies focused on only one type of data, and only 24% (N=11) of articles described considering multiple data in their model (≥2 types of data), such as clinical and imaging data. Data analysis differed according to the field of interest:

- Articles related to OA prediction, representing 86% of articles, described use of imaging data (N=6), as compared with only 50% of articles related to OA early diagnosis.
- Studies related to OA severity estimation did not use biological data in their model.
- Studies related to phenotypes always used multiple types of data in their models, whereas those related to early diagnosis used a single type of data.

**Reproducibility**

A separate training and testing set were described in 63% (N=29) of articles: internal validation in 80% (N=37), cross validation as the main internal validation method in 43% (N=20), leave one-out in 9% (N=4) and bootstrap in 4% (N=2). Only 26% (N=12) of the articles described splitting the cohort to validate the data. No validation was described in articles related to phenotypes, using...
Table 2  Descriptive analysis of 46 selected articles

| Overall (N=46) | Diagnosis (N=12) | Prediction (N=7) | Phenotypes (N=4) | Severity (N=12) | Progression (N=11) |
|---------------|------------------|-----------------|-----------------|----------------|-------------------|
| No of patients |                  |                 |                 |                |                   |
| Mean          | 1 359            | 978             | 1 254           | 518            | 1 803             | 1 662             |
| Median (range)| 525 (18–5 749)   | 263 (60–5 749)  | 601 (68–4 796)  | 559 (102–852)  | 942 (18–4 504)    | 728 (100–4 796)   |
| Year of publication |      |                 |                 |                |                   |
| Mean          | 2017             | 2017            | 2015            | 2018           | 2018              | 2018              |
| Median (range)| 2019 (2007–2021) | 2018 (2012–2020) | 2018 (2008–2020) | 2018 (2015–2019)| 2020 (2007–2021) | 2019 (2012–2020) |
| Type of method |                  |                 |                 |                |                   |
| Supervised    | 40 (87%)         | 11 (92%)        | 7 (100%)        | 1 (25%)        | 11 (92%)          | 10 (91%)          |
| Unsupervised  | 4 (9%)           | 0 (0%)          | 0 (0%)          | 3 (75%)        | 0 (0%)            | 1 (9%)            |
| Semi-supervised | 2 (4%)         | 1 (8%)          | 0 (0%)          | 0 (0%)         | 1 (8%)            | 0 (0%)            |
| Method subtype |                  |                 |                 |                |                   |
| Most frequently used | Convolutional neural network | Convolutional neural network random forest | Elastic net WND-CHARM | Latent class analysis topological data analysis | Convolutional neural network densely connected convolutional network | Logistic regression convolutional neural network |
| Mixed algorithms |                  |                 |                 |                |                   |
| Yes          | 13 (28%)         | 4 (33%)         | 0 (0%)          | 0 (0%)         | 4 (33%)           | 5 (45%)           |
| No           | 33 (72%)         | 8 (67%)         | 7 (100%)        | 4 (100%)       | 8 (67%)           | 6 (55%)           |
| Deep learning |                  |                 |                 |                |                   |
| Yes          | 16 (35%)         | 4 (33%)         | 0 (0%)          | 0 (0%)         | 9 (75%)           | 3 (27%)           |
| No           | 30 (65%)         | 8 (67%)         | 7 (100%)        | 4 (100%)       | 3 (25%)           | 8 (73%)           |
| Explainable model |                  |                 |                 |                |                   |
| Yes          | 15 (33%)         | 3 (25%)         | 3 (43%)         | 4 (100%)       | 0 (0%)            | 5 (45%)           |
| No           | 31 (67%)         | 9 (75%)         | 4 (57%)         | 0 (0%)         | 12 (100%)         | 6 (55%)           |
| Unexplainable model |                  |                 |                 |                |                   |
| Yes          | 10 (22%)         | 2 (17%)         | 1 (14%)         | 0 (0%)         | 4 (33%)           | 3 (27%)           |
| No           |                  |                 |                 |                |                   |                   |
| OA localisation |                  |                 |                 |                |                   |
| Knee         | 39 (85%)         | 10 (83%)        | 5 (71%)         | 3 (75%)        | 10 (83%)          | 12 (100%)         |
| Hip          | 7 (15%)          | 1 (8%)          | 2 (29%)         | 1 (25%)        | 2 (17%)           | 1 (9%)            |
| Hand         | 0 (0%)           | 0 (0%)          | 0 (0%)          | 0 (0%)         | 0 (0%)            | 0 (0%)            |
| Foot         | 0 (0%)           | 0 (0%)          | 0 (0%)          | 0 (0%)         | 0 (0%)            | 0 (0%)            |
| Clinical data | 19 (41%)         | 2 (17%)         | 3 (43%)         | 4 (100%)       | 3 (25%)           | 7 (64%)           |
| Biological data | 7 (15%)         | 4 (33%)         | 1 (14%)         | 1 (25%)        | 0 (0%)            | 1 (9%)            |
| Serum        | 6 (13%)          | 3 (25%)         | 1 (14%)         | 1 (25%)        | 0 (0%)            | 1 (9%)            |

Continued
|                               | Overall    | Diagnosis                          | Prediction                          | Phenotypes | Severity | Progression |
|-------------------------------|------------|------------------------------------|-------------------------------------|------------|----------|-------------|
| (N=48)                        | (N=12)     | (N=7)                              | (N=4)                               | (N=12)     | (N=11)   |
| Synovium                      | 2 (4%)     | 2 (16%)                            | 0 (0%)                              | 0 (0%)     | 0 (0%)   |
| Imaging data                  | 34 (74%)   | 6 (50%)                            | 6 (86%)                             | 3 (75%)    | 10 (83.3%) | 9 (82%)     |
| X-ray                         | 28 (51%)   | 5 (42%)                            | 5 (70%)                             | 3 (75%)    | 9 (75%)  |
| MRI                           | 10 (22%)   | 1 (8%)                             | 2 (17%)                             | 2 (50%)    | 1 (8%)  |
| Multiple data                 |            |                                    |                                     |            |          |
| Yes                           | 11 (24%)   | 0 (0%)                             | 2 (28.6%)                           | 3 (75%)    | 1 (8.3%) |
| No                            | 35 (76%)   | 12 (100%)                          | 5 (71%)                             | 1 (25%)    | 11 (91.7%) | 6 (55%)     |
| Training and testing sets     |            |                                    |                                     |            |          |
| Yes                           | 29 (63%)   | 9 (75%)                            | 6 (86%)                             | 0 (0%)     | 10 (83%) |
| No                            | 17 (37%)   | 3 (25%)                            | 1 (14%)                             | 4 (100%)   | 2 (17%) |
| Internal validation           |            |                                    |                                     |            |          |
| Yes                           | 37 (80%)   | 12 (100%)                          | 6 (85.7%)                           | 0 (0%)     | 11 (92%) |
| No                            | 9 (20%)    | 0 (0%)                             | 1 (14.3%)                           | 4 (100%)   | 1 (8%)  |
| Type of validation            |            |                                    |                                     |            |          |
| Cross                         | 20 (43%)   | 6 (50%)                            | 4 (67%)                             | 0 (0%)     | 3 (27%) |
| Split                         | 12 (25%)   | 4 (33%)                            | 0 (0%)                              | 0 (0%)     | 7 (58%) |
| Leave one out                 | 4 (8.7%)   | 2 (17%)                            | 1 (17%)                             | 0 (0%)     | 1 (9%)  |
| Bootstrap                      | 2 (4.3%)   | 0 (0%)                             | 1 (17%)                             | 0 (0%)     | 0 (0%)  |
| External validation           |            |                                    |                                     |            |          |
| Yes                           | 3 (7%)     | 1 (8.3%)                           | 1 (14%)                             | 0 (0%)     | 1 (8%)  |
| No                            | 43 (93%)   | 11 (91.7%)                         | 6 (86%)                             | 4 (100%)   | 11 (92%) |
| Cohort                        |            |                                    |                                     |            |          |
| OAI                           | 21 (46%)   | 3 (25%)                            | 2 (29%)                             | 1 (25%)    | 6 (50%) |
| MOST                          | 5 (11%)    | 1 (8.3%)                           | 0 (0%)                              | 1 (25%)    | 2 (16.7%) |
| CHECK                         | 3 (7%)     | 0 (0%)                             | 2 (29%)                             | 0 (0%)     | 0 (0%)  |
| Publicly available dataset    |            |                                    |                                     |            |          |
| Yes                           | 25 (54%)   | 3 (25%)                            | 4 (57%)                             | 3 (75%)    | 6 (50%) |
| No                            | 21 (46%)   | 9 (75%)                            | 3 (43%)                             | 1 (25%)    | 6 (50%) |
| Source code available         |            |                                    |                                     |            |          |
| Yes                           | 10 (22%)   | 2 (17%)                            | 1 (14%)                             | 0 (0%)     | 4 (33%) |
| No                            | 36 (78%)   | 10 (83%)                           | 6 (86%)                             | 4 (100%)   | 8 (67%) |

CHECK, Cohort Hip and Cohort Knee; MOST, Multicentre Osteoarthritis Study; OA, osteoarthritis; OAI, osteoarthritis initiative; WND-CHARM, Weighted Neighbor Distance using Compound Hierarchy of Algorithms Representing Morphology.
unsupervised algorithms. External validation with an independent dataset was described only in 7% (N=3) of the articles. Datasets were described as publicly available in 54% (N=25) of the articles; however, source codes were available in only 22% (N=10).

DISCUSSION

This systematic review gives an exhaustive overview of ML approaches in OA research, currently a very dynamic field. Indeed, most of the articles related to ML in OA were published in the last 5 years. Our study highlights that (1) 85% of the ML articles focused on knee OA, mainly using the OAI database, with only 15% focusing on hip OA and none focusing on other sites such as hand or foot OA; (2) radiological data investigations predominated over clinical or biological data; (3) DL is increasingly being used and was described in 35% of the articles and (4) external validation was poorly used (7% of the articles) (figure 4).

Importantly, one major strength of applying ML in OA research is the wide range of clinical applications, covering the current scientific questions and main challenges in OA such as diagnosis of OA at an early stage, predicting the development of OA in the population, identifying OA phenotypes, estimating OA structural severity and identifying patients with slow and rapid disease progression. However, we found few articles on OA symptoms such as pain, function, or physical activity, and articles related to phenotypes were few (N=4), representing only 9% of the selected articles.

Our review revealed several limitations in how ML is applied in OA studies. First, most of the ML algorithms applied were based on supervised approaches, which may limit the power of identifying novel phenotypes based on data. In addition, we found high heterogeneity in terms of algorithms used depending on the study. Explainable ML was described in 33% (N=15) articles, and among the articles based on unexplainable ML models, only one third described interpretability tools (N=10/31 (32%)). These results highlight the need for increasing awareness of the need to develop explainable AI and ML models. Second, 74% of the selected articles were related to imaging, and few articles described the use of clinical and biological data, which limits the discovery of new phenotypes, biomarkers of severity progression or diagnosis. Third, the major focus on knee OA in studies using ML is questionable because of high heterogeneity among OA subtypes or localisations, which remains unclear and calls for diversifying the studies to better understand these diseases. With the several available hand-OA cohorts such as the Digital Cohort Design cohort, the Hand Osteoarthritis in Secondary care and the Nor-hand study, studies using ML tools in hand OA research are expected to better understand the characteristics, specificities and course of this OA localisation in the future. Finally, data and source codes for analyses were not available in 46% and 78% of the articles, but they are critical to ensure reliability and reproducibility of the ML analyses. Furthermore, external validation was described in only 7% of the selected articles, which is also a crucial point because ‘reproducibility crisis’ is one of the main challenges in science, particularly the ML field.

One major bias that could be highlighted in the studies applying ML to OA is that they mostly involved using the OAI, MOST and CHECK cohorts, with 46% of the articles involving the OAI database. Importantly, most of the cohorts currently available predate the interest of ML analyses in the field. Therefore, the study design and consent form may not have included broad data-sharing, which limits the use of the data for ancillary studies. This situation reinforces the need for purpose-built cohorts for ML analyses. As an example, the consortium for Applied Public–Private Research enabling OsteoArthritis Clinical Headway (APPROACH) aims at creating a broad
The use of ML checklists has been promoted in other fields and in an interventionual clinical trial using ML according to the Consolidated Standards of Reporting Trials-Artificial Intelligence guidelines. The application of these checklists could improve the quality, standardisation, and reproducibility of ML studies in OA research.

In conclusion, ML is a fast-growing field providing better knowledge of human OA disease (diagnosis assistive tool especially for early OA, prediction for progression or severity of OA, characterisation of new therapeutic targets). This systematic review provides a comprehensive overview of ML applications in OA and delineates some methodological caveats that can and should be resolved to improve the quality of ML studies in OA research.

Author affiliations
1Department of Rheumatology, Hôpital Saint-Antoine, Assistance Publique – Hôpitaux de Paris (AP-HP), Centre de Recherche Saint-Antoine, Inserm UMR 938, Assistance Publique – Hôpitaux de Paris (AP-HP), Sorbonne Université, Paris, France
2Bakar Computational Health Science Institute, University of California, San Francisco, California, USA
Immunology Immunopathology Immunotherapy UMR 959, Sorbonne Université, Paris, France
3Center for Intelligent Imaging (C2I), Department of Radiology and Biomedical Imaging, University of California, San Francisco, California, USA
4Biotherapy (CIC-BTI) and Inflammation Immunopathology-Biotherapy Department (12B), Hôpital Pitié-Salpêtrière, AP-HP, Paris, France

Twitter Marie Binvignat @m_binvignat
Contributors MB has contributed to the conception and design of the study, acquisition, collection and assembly of the data, analysis and interpretation of data, drafting the article. VP has contributed to analysis and interpretation of data, revising it critically for important intellectual content and final approval of the article. AJB has contributed to analysis and interpretation of data, revising it critically for important intellectual content and final approval of the article. KL has contributed to analysis and interpretation of data, revising it critically for important intellectual content and final approval of the article. EM-F has contributed to the conception and design of the study, analysis and interpretation of the data, drafting the article, revising it critically for important intellectual content and final approval of the article. JS has contributed to the conception and design of the study, analysis and interpretation of the data, drafting the article, revising it critically for important intellectual content and final approval of the article. MB (marie.binvignat@sorbonne-universite.fr), JS (jeremie.seliam@aphp.fr) and EM-F (enancita.mariotti@sorbonne-universite.fr) accept full responsibility for the work and/or the conduct of the study, had access to the data, and controlled the decision to publish.

Funding MB is funded by the Sorbonne Université doctoral award fellowship and by the French Society of Rheumatology mobility fellowship.

Competing interests AJB is a cofounder and consultant to Personalis and NuMedici; consultant to Samsung, Mango Tree Corporation, and in the recent past, 10x Genomics, Helix, Pathway Genomics, and Verinata (Illumina); has served on paid advisory panels or boards for Geisinger Health, Regenstrief Institute, Gerson Lehman Group, AlphaSights, Covance, Novartis, Genentech, and Merck, and Roche; is a shareholder in Personalis and NuMedici; is a minor shareholder in Apple, Facebook, Alphabet (Google), Microsoft, Amazon, Snap, 10 x Genomics, Illumina, CVS, Nuna Health, Assay Depot, Vetz24even, Regeneron, Sanofi, Royalty Pharma, AstraZeneca, Moderna, Bioeas, Paraxel, and Soturo; and several other health-related companies and mutual funds; and has received honoraria and travel reimbursement for invited talks from Johnson and Johnson, Roche, Genentech, Pfizer, Merck, Lilly, Takeda, Varian, Mars, Siemens, Optum, Abbott, Celgene, AstraZeneca, AbbVie, Westat and many academic institutions, medical or disease specific foundations and associations, and health systems. AJB receives royalty payments through Stanford University, for several patents and other disclosures licensed to NuMedici and Personalis. AJB’s research has been funded by NIH, Northrup Grumman (as the prime on an NIH contract), Genentech, Johnson and
Osteoarthritis

Johnson, FDA, Robert Wood Johnson Foundation, Leon Lowenstein Foundation, Intervalen Foundation, Priscilla Chan and Mark Zuckerberg, the Barbara and Gerson Bakar Foundation, and in the recent past, the March of Dimes, Juvenile Diabetes Research Foundation, California Governor’s Office of Planning and Research, California Institute for Regenerative Medicine, L’Oréal, and Progenity. JG reports personal fees from MSD, Pfizer, Abivio, Fresenius Kabi, BMS, Roche Chugai, Sandoz, Lilly, Gilead, Novartis, Janssen and grant research from Pfizer, MSD, Schwa Medico, BMS.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in this article or uploaded as online supplemental information. All data relevant to the study are included in the article or uploaded as online supplemental information. Complementary data are available on request.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: https://creativecommons.org/licenses/by/4.0/.

ORCID iDs
Marie Binvgigant http://orcid.org/0000-0001-7473-7636
Francis Berenbaum http://orcid.org/0000-0001-8252-7815

REFERENCES

1. Bohr A, Mernarzadeh K. The rise of artificial intelligence in healthcare applications. In: Artificial intelligence in healthcare. Elsevier, 2020: 25–60.
2. Pandit A, Radstake TRDJ. Machine learning in rheumatology approaches the clinic. Nat Rev Rheumatol 2020;16:69–70.
3. Kokkotis C, Moustakidis S, Papaageorgiou E, et al. Machine learning in knee osteoarthritis: a review. Osteoarthritis Cartil 2020;28:S10069.
4. Helm JM, Swiergosz AM, Haebeler HS, et al. Machine learning and artificial intelligence: definitions, applications, and future directions. Curr Rev Musculoskelet Med 2020;13:69–76.
5. Cao L. Data science: a comprehensive overview. ACM Comput Surv 2017;50.
6. Sarkar IH. Machine learning: algorithms, real-world applications and research directions. SN Comput Sci 2021;2:160.
7. Lo Vercio L, Amador K, Bannister JJ, et al. Supervised machine learning tools: a tutorial for clinicians. J Neural Eng 2020;17:062001.
8. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol 2019;19:64.
9. Pedroia V. Machine learning and artificial intelligence. Osteoarthritis Cartilage 2020;28:S16.
10. Dubey SR, Chakraborty S, Roy SK, et al. diffGRad: an optimization method for Convolutional neural networks. IEEE Trans Neural Netw Learn Syst 2020;31:4500–11.
11. Nohara Y, Matsumoto K, Soejima H, et al. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. Comput Methods Programs Biomed 2022;214:106584.
12. Ribeiro M, Singh S, Guestrin C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. San Diego, California: Association for Computational Linguistics, 2016: 97–101.
13. Linardatos P, Papastefanopoulou V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy 2020;22:18.
14. McDermott MBA, Wang S, Marinsek N, et al. Reproducibility in machine learning for health research: still a ways to go. Sci Transl Med 2021;13:eabf1655.
15. Norgeot B, Glicksberg BS, Trupin L, et al. Assessment of a deep learning model based on electronic health record data to forecast clinical outcomes in patients with rheumatoid arthritis. JAMA Netw Open 2019;2:e190606.
16. Orange DE, Agius P, DiCarlo EF, et al. Identification of three rheumatoid arthritis disease subtypes by machine learning integration of synovial histologic features and RNA sequencing data. Arthritis Rheumatol 2018;70:690–701.
17. Eng SWM, Aeschlimann FA, van Veenendaal M, et al. Patterns of joint involvement in juvenile idiopathic arthritis and prediction of disease course: a prospective study with multilayer non-negative matrix factorization. PLoS Med 2019;16:e1002750.
18. Igle M, Omouri F, van Laar JM, et al. Applied machine learning and artificial intelligence in rheumatology. Rheumatol Adv Pract 2020;4:kaa005.
19. Gosses L, Kedra J, Servy H, et al. EULAR points to consider for the use of big data in rheumatic and musculoskeletal diseases. Ann Rheum Dis 2020;79:69–76.
20. Peterly CG, Schneider E, Nevitt M. The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee. Osteoarthritis Cartilage 2008;16:1433–41.
21. Segal NA, Nevitt MC, Gross KD, et al. The multicenter osteoarthritis study: opportunities for rehabilitation research. PM R 2013;5:647–54.
22. Wesseling J, Boers M, Viergever MA, et al. Cohort profile: cohort hip and cohort knee (check) study. Int J Epidemiol 2016;45:36–44.
23. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6:e100060.
24. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009;339:b2700.
25. Booth A, Clarke M, Ghersi D, et al. An international registry of systematic-review protocols. Lancet 2011;377:108–9.
26. Pedregosa F, Varoquaux G, Gramfort A. Scikit-learn: machine learning in python. J Mach Learn Res 2012;12.
27. Kundu S, Ashinsky BG, Boughrara M, et al. Enabling early detection of osteoarthritis from presymptomatic cartilage texture maps via transport-based learning. Proc Natl Acad Sci U S A 2020;117:24709–19.
28. Tiulpin A, Klein S, Bierma-Zeinstra SMA, et al. Multimodal machine Learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. Sci Rep 2019;9:20038.
29. Nelson AE, Fang F, Arbeeva L, et al. A machine learning approach to knee osteoarthritis phenotyping: data from the FNHI biomarkers Consortium. Osteoarthritis Cartilage 2019;27:994–1001.
30. Hu T, Oksanen K, Zhang W, et al. An evolutionary learning and network approach to identifying key metabolites for osteoarthritis. PLoS Comput Biol 2018;14:e1005988.
31. Lim J, Kim J, Cheon S. A deep neural network-based method for early detection of osteoarthritis using statistical data. Int J Environ Res Public Health 2019;16:1281.
32. Abraham A, Jannane R, Riad R, et al. A decision support tool for early detection of knee osteoarthritis using X-ray imaging and machine learning: data from the osteoarthritis initiative. Comput Med Imaging Graph 2019;73:11–18.
33. Kotti M, Duffell LD, Faisal AA, et al. Detecting knee osteoarthritis and its discriminating parameters using random forests. Med Eng Phys 2017;43:19–29.
34. Ahmed U, Anwar A, Savage RS, et al. Protein oxidation, nitration and glycation biomarkers for early-stage diagnosis of osteoarthritis of the knee and typing and progression of arthritic disease. Arthritis Res Ther 2016;18:250.
35. Han MY, Dai JJ, Zhang Y, et al. Identification of osteoarthritis biomarkers by proteomic analysis of synovial fluid. J Int Med Res 2012;40:2243–50.
36. Marques J, Gerant HK, Lillhom M. Diagnosis of osteoarthritis and prognosis of tibial cartilage loss by quantification of tibia trabecular bone from MRI: diagnosis of osteoarthritis and prognosis of cartilage loss. Magn Reson Med 2013;70:568–75.
37. Xue Y, Zhang R, Deng Y, et al. A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. PLoS One 2017;12:e0178992.
38. Heard BJ, Rosvold JM, Fritzier MJ, et al. A computational method to differentiate normal individuals, osteoarthritis and rheumatoid arthritis patients using serum biomarkers. J R Soc Interface 2014;11:20140428.
39. Shamir L, Ling SM, Scott W, et al. Early detection of radiographic knee osteoarthritis using computer-aided analysis. Osteoarthritis Cartilage 2009;17:1307–12.
40. Örten K, Arlan T, Gültekin KE, et al. Detection of hip osteoarthritis by using plain pelvic radiographs with deep learning methods. Skeletal Radiol 2020;49:1369–74.
41. Lazzarini N, Runhaar J, Bay-Jensen AC, et al. A machine learning approach for the identification of new biomarkers for knee osteoarthritis development in overweight and obese women. Osteoarthritis Cartilage 2017;25:2014–21.
42. Ashinsky BG, Boughrara M, Coletta CE, 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 2017;35:2243–50.

43 Hirvasniemi J, Giels WP, Arbabi S, et al. Bone texture analysis for prediction of incident radiographic hip osteoarthritis using machine learning: data from the cohort hip and cohort knee (check) study. Osteoarthr Cartilage 2019;27:906–14.

44 Yoo TK, Kim DW, Choi SB, et al. Simple scoring system and artificial neural network for knee osteoarthritis risk prediction: a cross-sectional study. PLoS One 2016;11:e0148724.

45 Giels WP, Weins H, Welsing PMJ, et al. An automated workflow based on hip shape improves personalized risk prediction for hip osteoarthritis in the check study. Osteoarthr Cartilage 2020;28:62–70.

46 Waarsing JH, Biemstra-Zeinstra SMA, Weins H. Distinct subtypes of knee osteoarthritis: data from the osteoarthritis initiative. Rheumatology 2015;54:1650–8.

47 Carlesso LC, Segal NA, Frey-Law L, et al. Pain susceptibility phenotypes in those free of knee pain with or at risk of knee osteoarthritis: the multicenter osteoarthritis study. Arthritis Rheumatol 2019;71:542–9.

48 Pedoia V, Norman B, Mehary SN, et al. 3D convolutional neural networks for detection and severity staging of meniscus and PFJ cartilage morphological degenerative changes in osteoarthritis and anterior cruciate ligament subjects. J Magn Reson Imaging 2019;49:400–10.

49 von Schacky CE, Sohn JH, Liu F, et al. Development and validation of a Multitask deep learning model for severity grading of hip osteoarthritis features on radiographs. Radiology 2020;295:136–45.

50 Abedin J, Antony J, McGuinness K, et al. Predicting knee osteoarthritis severity: comparative modeling based on patient’s data and plain X-ray images. Sci Rep 2019;9:5761.

51 Norman B, Pedoia V, Noworolski A, et al. Applying densely connected Convolutional neural networks for staging osteoarthritis severity from plain radiographs. J Digit Imaging 2019;32:471–7.

52 Chen P, Gao L, Shi X, et al. Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. Comput Med Imaging Graph 2019;75:84–92.

53 Boniati L, Costanioud L, Cavouris D, et al. Assessing hip osteoarthritis severity with a probabilistic neural network based classification scheme. Med Eng Phys 2007;29:227–37.

54 Liu B, Luo J, Huang H, Toward automatic quantification of knee osteoarthritis severity using improved faster R-CNN, Int J Comput Assist Radiol Surg 2020;15:457–66.

55 Moustakidis SP, Theocharis JB, Glakas G. A fuzzy decision tree-based SVM classifier for assessing osteoarthritis severity using ground reaction force measurements. Med Eng Phys 2010;32:1145–60.

56 Giels WP, Ku Y, Han H-S, uk-soo HH, et al. A machine learning-based diagnostic model associated with knee osteoarthritis severity, Sci Rep 2020;10:15743.

57 Nguyen HH, Saarakalaka S, Blaschko MB, et al. Semixup: in- and Out-of-Manifold regularization for deep Semi-Supervised knee osteoarthritis severity grading from plain radiographs. IEEE Trans Med Imaging 2020;39:4346–56.

58 Schwartz AJ, Clarke HD, Spangehl MJ, et al. Can a Convolutional neural network classify knee osteoarthritis on plain radiographs as accurately as Fellowship-Trained knee arthroplasty surgeons? J Arthroplasty 2020;35:2432–8.

59 Swiezicki A, Li N, O’Donnell J, et al. Deep learning-based algorithm for assessment of knee osteoarthritis severity in radiographs matches performance of radiologists. Comput Biol Med 2021;133:104334.

60 Tormařekho S, Aarnio E, Mononen ME, et al. Eight-Year trajectories of changes in health-related quality of life in knee osteoarthritis: data from the osteoarthritis initiative (OAI), PLoS One 2019;14:e0219902.

61 Du Y, Almajali R, Shan J, et al. A novel method to predict knee osteoarthritis progression on MRI using machine learning methods. IEEE Trans Nanobioscience 2018;17:228–36.

62 Woloszynski T, Potsdadi P, Stachowiak G, et al. A dissimilarity-based multiple classifier system for trabecular bone texture in detection and prediction of progression of knee osteoarthritis. Proc Inst Mech Eng H 2012;226:887–94.

63 Passey C, Kimko H, Nandy P, et al. Osteoarthritis disease progression model using six year follow-up data from the osteoarthritis initiative. J Clin Pharmacol 2015;55:269–78.

64 LaValley MP, Lo GH, Price LL, et al. Development of a clinical prediction algorithm for knee osteoarthritis structural progression in a cohort study: value of adding measurement of subchondral bone density. Arthritis Res Ther 2017;19:95.

65 Leung K, Zhang B, Tan T, et al. Prediction of total knee replacement and diagnosis of osteoarthritis by using deep learning on knee radiographs: data from the osteoarthritis initiative. Radiology 2020;296:584–93.

66 Widera P, Welsing PMJ, Ladel C, et al. Multi-classifier prediction of knee osteoarthritis progression from complete imbalanced longitudinal data. Sci Rep 2020;10:8427.

67 Tolpadi AA, Lee JJ, Pedoia V, et al. Deep learning predicts total knee replacement from magnetic resonance images. Sci Rep 2020;10:6371.

68 Bonakkadi H, Tardif G, Abram F, et al. Serum adipokines/related inflammatory factors and ratios as predictors of infrapatellar fat pad volume in osteoarthritis: applying comprehensive machine learning approaches. Sci Rep 2020;10:9993.

69 Rossi-deviophies V, Pedoia V, Samaan MA, et al. Using multidimensional topological data analysis to identify traits of hip osteoarthritis. J Magn Reson Imaging 2018;48:1048–55.

70 Tiupalin A, Thevenot J, Rahtu E. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep Learning-Based approach. Sci Rep 2018;8.

71 Watt EW, Watt E, Bui AAT, et al. Evaluation of a dynamic Bayesian belief network to predict osteoarthritic knee pain using data from the osteoarthritis initiative. AMIA Annu Symp Proc 2008;2008:788–92.

72 Pedoia V, Haefeli F, Morioka K. MRI and biomechanics multidimensional data analysis reveals R 2 -R 1p as an early predictor of cartilage lesion progression in knee osteoarthritis: multidimensional data analysis to study oa. J Magn Reson Imaging 2018;47:78–90.

73 Orlov N, Shamir L, Macura T, et al. WND-CHARM: multi-purpose image classification using compound image transforms. Pattern Recognit Lett 2008;29:1684–93.

74 Sellam J, Maheu E, Crema MD, et al. The DIGICOD cohort: A hospital-based observational prospective cohort of patients with hand osteoarthritis - methodology and baseline characteristics of the population. Joint Bone Spine 2021;88:105171.

75 Damman W, Liu R, Kroon FPB, et al. Do comorbidities play a role in hand osteoarthritis disease burden? data from the hand osteoarthritis in secondary care cohort. J Rheumatol 2017;44:1659–66.

76 Gloersen M, Mulrooney E, Mathiessen A, et al. A hospital-based observational cohort study exploring pain and biomarkers in patients with hand osteoarthritis in Norway: the Nor-Hand protocol. BMJ Open 2017;7:e016938.

77 Research, reuse, repeat. Nat Mach Intell 2020;2:729.

78 The Lancet Respiratory Medicine. Opening the black box of machine learning. Lancet Respir Med 2018;6:801.

79 Beam AL, Mannai AK, Ghassemi M. Challenges to the reproducibility of machine learning models in health care. JAMA 2021;323:305.

80 van Helvoort EM, van Spil WE, Jansen MP, et al. Cohort profile: the applied public-private research enabling osteoarthritis clinical Headway (IMI-APPROACH) study: a 2-year, European, cohort study to describe, validate and predict phenotypes of osteoarthritis using clinical, imaging and biochemical markers. BMJ Open 2020;10:e035101.

81 Widera P. A machine learning “APPROACH” to recruitment in OA. Osteoarthr Cartilage 2019;27:S15.

82 Mobasheri A, Kapoor M, Ali SA. The future of deep phenotyping in osteoarthritis: how can high throughput omics technologies advance our understanding of the cellular and molecular taxonomy of the disease? Osteoarthr Cartil Open 2021.

83 Scott I, Carter S, Coiera E. Clinician checklist for assessing suitability of machine learning applications in healthcare. BMJ Health Care Inform 2021;28:e100251.

84 Artrith N, Butler KT, Coudert F-X, et al. Best practices in machine learning for chemistry. Nat Chem 2021;13:503–8.

85 Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 2020;26:1364–74.