Data-driven identification of predictive risk biomarkers for subgroups of osteoarthritis using an interpretable machine learning framework: a UK biobank study

Ramneek Gupta

rmgp@novonordisk.com

Novo Nordisk Research Centre Oxford https://orcid.org/0000-0001-6841-6676

Rikke Linnemann Nielsen

https://orcid.org/0000-0003-0173-2134

Thomas Monfeuga

Novo Nordisk Research Centre Oxford, UK

Robert Kitchen

Novo Nordisk Research Centre Oxford, UK

Line Egerod

Novo Nordisk Research Centre Oxford, UK

Luis Leal

Novo Nordisk Research Centre Oxford, UK

August Schreyer

Novo Nordisk Research Centre Oxford, UK

Carol Sun

Novo Nordisk Research Centre Oxford, UK

Marianne Helenius

Novo Nordisk Research Centre Oxford, UK https://orcid.org/0000-0003-3613-8338

Lotte Simonsen

Novo Nordisk A/S

Marianne Willert

Novo Nordisk A/S

Abd Tahrani

Novo Nordisk A/S

Zahra McVey

Novo Nordisk Research Centre Oxford, UK

Article

Keywords:

Posted Date: August 10th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3230959/v1

Additional Declarations: Yes there is potential Competing Interest. RLN, TM, RRK, LGL, ATHS, CS, LS, MW, AAT, ZM and RG are employed by Novo Nordisk. LE is currently employed by Nordic Bioscience A/S and was working on this manuscript while being employed by Novo Nordisk A/S.

Version of Record: A version of this preprint was published at Nature Communications on April 1st, 2024. See the published version at https://doi.org/10.1038/s41467-024-46663-4.

Abstract

Osteoarthritis (OA) is increasing in prevalence and has a severe impact on patients' lives. However, our understanding of biomarkers driving OA risk remains limited. We developed a model predicting the five-year risk of OA, integrating clinical, lifestyle and biomarker data from the UK Biobank (19,120 patients with OA, ROC-AUC:0.72 95%CI (0.71 – 0.73)). Higher age, BMI, and prescription of non-steroidal anti-inflammatory drugs contributed most to increased OA risk prediction. 14 sub-groups of OA risk profiles were identified, and validated in an independent set of patients evaluating the 11-year OA risk, with 88% of patients uniquely assigned to one of the sub-groups. Individual OA risk profiles were characterised by personalised biomarkers. Omics integration demonstrated the predictive importance of key OA genes and pathways (e.g. *GDF5* and TGF-β signalling) and identified OA-specific biomarkers (e.g. CRTAC1 and COL9A1). In summary, this work identified opportunities for personalised OA prevention and insights into its underlying pathogenesis.

Introduction

Osteoarthritis (OA) is a chronic degenerative joint disease, that results in decreased quality of life and a high economic health-care burden. Globally, 528 million people are living with OA[1]. The global prevalence has increased by 48% from 1990 to 2019, with further increases expected due to ageing populations and the increasing prevalence of obesity[1,2]. OA has a high impact on health care expenditure and social care cost[3], with the economic impact ranging from 1 to 2.5% of gross national product (GNP) in some countries. The average annual cost of OA for an individual is estimated to be between $700-$15,600 (2019)[1]. Patients are often diagnosed with late-stage disease when irreversible joint damage has already occurred[4]. There are currently no approved treatments that can prevent OA development and progression or cure the disease[4]. The main treatments remain symptomatic, with some patients ultimately needing joint replacement surgery[5]. Hence, there is a need to identify strategies that aid disease prevention and early diagnosis, to facilitate the development of disease modifying therapies.

An improved understanding of OA pathogenesis is necessary to develop new approaches to early diagnosis, prevention and treatment of OA [5,6]. OA is a complex and heterogenous disease, spanning multiple biological mechanisms. These range from cartilage degeneration to systemic and synovial inflammation[5,6]. It has been proposed that OA would benefit from the use of data-driven and machine learning approaches for patient-specific prediction models to dissect the complex relationship between susceptibility biomarkers[4,7–9].

Previous disease predictions of OA have been limited by a focus on knee OA, small sample sizes, restricted sets of input features and more traditional statistical analysis methods[4,8,9]. Additionally, multiple studies focus on disease classification or the prediction of disease progression rather than prediction of disease incidence[4].

Known OA risk biomarkers include age, female sex, BMI/obesity[10], genetic predisposition, as well as joint-specific risk biomarkers such as injuries[11]. However, the complex interplay between risk biomarkers, and to what extent combinations of risk biomarkers can predict individual OA risk, remains to be determined.

Despite efforts to predict individual risk of OA, the majority of these have lacked comprehensive incorporation of genetics, clinical biomarkers and other environmental factors to guide preventative strategies and precision medicine[4,8,12]. Additionally, further work is required to study the impact of changes in modifiable risk biomarkers on the individual risk of OA[9]. Machine learning methods have the unique potential to integrate diverse patient data for the prediction of disease outcomes[13]. In this study, we developed a machine learning model to predict individual five-year risk of OA and identify risk biomarkers. Through the integration of multi-modal patient data, we identified sub-groups of OA with differing risk biomarker profiles. The model captured the broad patient landscape in a UK cohort of ~20,000 people diagnosed with OA, capturing electronic health records (EHR), clinical biomarkers, self-reported questionnaire data, genomics, proteomics, and metabolomics on available subsets of individuals. The model quantified the impact of risk biomarkers on the predicted OA risk at the population and individual level, enabling detailed estimation of the contribution of these biomarkers for OA risk.

Results

OA study population and risk modelling

103,086 patients with OA were identified from primary and secondary EHR data[14] (~21% of all UK biobank study participants, Sup File 1). The UK biobank contains detailed phenotyping of individuals at the recruitment assessment centre between 2006 – 2010. Primary care data enabled the capture of longitudinal OA progression data but was only available for a subset of patients with OA (N = 67,772). An equal number of control participants who never developed OA were identified (N = 67,772). Controls were randomly selected and date-matched with the OA diagnosis dates for case patients. Cases and controls were then filtered for those with an OA diagnosis/matched index date a maximum of five years after the recruitment assessment centre. Controls were required to have observational data and no death registered during the five years prior to the index date (study period: 06/2006 – 09/2015). This resulted in a total of 19,120 patients with OA and 19,252 controls included in the analysis (Fig. 1A left, Sup Fig. 1). The cases and controls were well-matched based on comparison of baseline distribution for known OA risk biomarkers (Fig. 1B). For the patients with OA, the specific joints affected were mapped to the following groups: foot (4%), spine (10%), arm (11%), hip (13%), and knee (28%) (Fig. 1C). In addition to the OA study population, an independent hold-out validation cohort was generated following the same procedure as above. However, for this validation cohort, the time between data collection at the assessment centre and OA diagnosis was more than five years. (Fig. 1A left, 1B, study period: 09/2011 – 09/2017).

Based on the identified cases and controls in the OA study population, an eXtreme Gradient Boosting (XGBoost) machine learning model was trained to predict the five-year risk of developing OA using multi-

model longitudinal data. Performance was evaluated in the test set in a 5*5 cross-validation (Fig. 1A right). The XGBoost model integrated clinical, sociodemographic, diet, physical activity and lifestyle data from the recruitment assessment centre, with clinical data from the five-year longitudinal EHR data (Fig.1A left). The EHR data captured diagnosis of previous post-traumatic OA diagnosis, longitudinal blood and urine biomarkers, clinical measurements, as well as medication data for obesity, OA, and type 2 diabetes. Longitudinal data was captured in the five years prior to OA diagnosis/index date using yearly data bins (Sup Fig. 2). Missingness estimation for each feature included in the machine learning models is presented in Sup File2.

Prediction of OA from five-year multi-modal clinical data

Longitudinal clinical data were integrated in a XGBoost model predicting the five-year risk of OA (Clin model), achieving a cross-validated ROC-AUC performance of 0.72 (95% CI: 0.71 – 0.73, Fig. 2A, 2B and 2C). The Clin model was able to predict 7 in 10 patients who developed OA, and out of all the OA case predictions made by the model, 66% of these were true-positive OA cases (Fig. 2C). Conversely, we were able to predict 6 in 10 patients who did not develop OA, with 67% of the predicted controls being true-negative controls (Fig. 2C). Clin model's predictive performance was robust across random model initialisations and performed significantly better than models trained on permuted OA status labels (Sup Table 1). We assessed if the Clin model had a stronger predictive performance when predicting specific subgroups of OA across different affected joints, specified for some of the OA diagnoses, including arm, foot, hip, knee, or spine. Performance ranged with ROC-AUC: 0.67 – 0.73 (Fig. 2A) and was highest for weight-bearing joints (ROC-AUC: 0.73 for knee OA prediction, and 0.72 for hip OA prediction).

Population risk biomarkers predictive of OA risk

To interpret which risk biomarkers were most important for OA prediction, Shapley additive explanations (SHAP) values were calculated. SHAP values estimate the marginal conditional impact of a feature on the model's prediction in relation to all other features included in the model. The top three predictors of increased OA risk included higher age, prescription of non-steroidal anti-inflammatory drugs (NSAIDs) during the year prior to OA diagnosis, and higher BMI compared to individuals that did not develop OA (Fig. 2D). Following these three risk biomarkers, predictive contributions were made by a variety of features across individuals. Participants who rated their own health as excellent and had a faster walking pace, had a lower risk of OA. Higher levels of vitamin D were predictive of increased OA risk, and individuals that reported taking vitamin D supplements typically had a higher age and higher levels of vitamin D (Sup Fig. 3). Additionally, a higher hand grip strength, and a lower ratio of fat mass to fat free mass were also predictive of lower OA risk. Individuals that had a higher socioeconomic status, as indicated by having a college or university degree and higher income, had a lower risk of OA. In contrast, people working heavy manual or physical jobs, or doing shift work, had increased risk of OA.

Precision sub-groups of OA

The Clin model confirmed that the biological and environmental risk factors underlying OA are heterogenous across individuals. We attempted to capture this heterogeneity and categorise patients into sub-groups with differing risk biomarker profiles. Hence, we clustered the SHAP values, as estimated by the Clin model, for all risk biomarkers across all individuals. The clustering resolution was optimised based on silhouette scores and prediction metrics (Sup Fig. 4) identifying 14 clusters of individuals (Fig. 3A). The clustering allowed us to uncover sub-groups of individuals predicted to have high risk of OA (cf. prediction probabilities shown in Fig. 3B). Furthermore, by using SHAP values, rather than the original input values, we were able to account for the relative importance of features for OA prediction (Fig. 3C). Finally, all identified clusters were described using the average values of the top 6 features in our model, using the original input values to characterise the differences between clusters. This generated an overview of the most defining characteristics of each OA sub-group, capturing predicted archetypes of OA risk with distinct biomarker profiles (Fig. 3D).

To identify clusters with high predicted OA risk and understand sub-group characteristics, we defined the prediction performance of the XGBoost model within each cluster (F1/PPV/Sensitivity) the cluster case/control ratios and average prediction probability (Fig. 4). The top 3 clusters (12, 11 and 0), representing 23% of all individuals, were the clusters for which individuals were best predicted as OA cases, with F1 > 0.83. Another group of 6 clusters (10, 5, 9, 1, 13 and 8; ~35% of all individuals) had more modest values, but were still predictive for OA (0.73 > F1 > 0.61). The last 5 clusters (6, 7, 3, 4 and 2; ~ 41% of all individuals) were the least predictive for OA (F1 < 0.35).

We used a decision tree-based algorithm (SkopeRules[15]) to define sets of rules, per cluster, based on the input values of the XGBoost model. These rules allowed us to “scope” each predicted OA archetype by identifying the most distinctive variables and values that determined an individual’s cluster allocation with high PPV (Fig. 4, Sup Fig. 5). The rules that defined the top 3 high-risk patient groups included: age, prescription of NSAIDs within the last year, hand grip strength, self-reported walking pace and health rating, Townsend deprivation index and IGF-1 levels (Fig. 4). To validate the potential clinical value of these rules, we applied them to an independent hold-out population (with similar case/control definitions, and with cases being diagnosed more than five years after the assessment center visit; 7,341 cases and 5,999 controls). While 4% of individuals could not be accurately mapped to any sub-group based on these sets of rules, we were able to uniquely assign 88.2% of individuals to a cluster, and 7.8% to two possible clusters. In the latter case, we selected the cluster with the highest case/control ratio in the OA study population in order to minimise the risk of false negatives (Fig. 4 shows cluster attribution). We observed a high correlation of case/control ratios between clusters of the OA study and hold-out populations (R^2 = 0.90), as well as a strong correlation in the proportion of individuals in each cluster (R^2 = 0.68) (Sup Fig. 6).

Personalised risk biomarkers of OA

Interpretation using SHAP values from the XGBoost models enabled us to quantify the impact an individual's patient data had on their predicted risk of OA. We extracted and visualised individual OA risk profiles using waterfall plots, demonstrating the predicted positive and negative impact of personal OA risk biomarkers (Fig. 5). For example, a patient who developed OA from Cluster 1 had a predicted risk of 64% for OA. This predicted risk was predominantly driven by a BMI in the obesity range and age of 65 years (Fig. 5A). However, this patient had not taken NSAIDs 1 year prior to OA diagnosis, and this decreased the predicted risk for OA. Other risk biomarkers had more minor contributions to increasing the predicted OA risk, including a lower muscle strength (indicated by hand grip strength), and lower socioeconomic status (indicated by lower average income and level of education). For this patient, if the BMI was not considered, the predicted OA risk would have decreased to 57%. Our results are suggestive of potential intervention opportunities on high impact modifiable risk biomarkers that may be driving OA risk.

Additional individual OA risk profiles were examined, including an individual from cluster 2 with very low predicted OA risk (Fig. 5B), patient with OA from cluster 12 with multiple signs of poor metabolic health (Fig. 5C), and an individual profile impacted by lifestyle factors such as heavy manual/physical work and shift work from cluster 9 (Fig. 5D).

Multi-omics OA risk biomarkers

To explore molecular risk biomarkers of OA in the context of the clinical prediction model (Clin model), we integrated various types of omics data with the clinical features including OA genetics (ClinSNP, ClinGRS and ClinPath models), metabolomics (ClinMet model) and proteomics data (ClinPro model) for subsets of individuals where this data were available (Fig. 6A). The predictive performance remained unchanged compared to the Clin model (ROC-AUC: 0.70 – 0.72, Fig. 6A, sensitivity analysis of the Clin model on these specific subsets in Sup Table 2). However, the inclusion of OA omics signatures influenced the rankings of OA risk biomarkers in the model (Fig. 6A and 6B).

Firstly, when including a whole-genome polygenic risk score (WGPRS) for OA in our models (ClinWGPRS), it was the sixth most predictive risk feature for OA risk (data not shown). At the population level, the OA WGPRS did not affect the overall performance of the model (ROC-AUC: 0.72) compared to Clin model. However, the predictive performance for the model (F1 and sensitivity) were higher for a population with higher OA WGPRS (Sup Fig. 7). Therefore, although the OA WGPRS provided minimal additional benefit when predicting OA at the population level, it did provide additional value for sub-groups with more extreme genetic risk.

At the gene locus level, the highest ranked gene-level genetic risk scores (gene-GRS) by the OA prediction model were *TGFB1*, *GDF5*, *PTCH1* and *FAM53A*. The relevance of *TGFB1* to OA was further supported by the TGF beta signalling pathway being the top ranked pathway-level polygenic risk score (pathway-PRS). Other pathways identified as being predictive of OA, in the context of clinical features, included glycosphingolipid biosynthesis, adipocytokine signalling and cytokine-cytokine receptor interactions. No

strong predictive signal was identified for previously reported OA-associated single nucleotide variants. Blood plasma metabolites that were identified as being important for the prediction of OA included acetate, valine, 3-hydroxybutyrate, citrate and the percentage of saturated fatty acids to total fatty acids. Strong predictive proteomics signatures were identified where CRTAC1, COL9A1, ACTA2, EDA2R and TACSTD2 were the most predictive of OA risk together with higher age, prescription of NSAIDs during the year prior to OA diagnosis, and higher BMI as seen in the Clin model (Sup Fig. 8). There were significant differences in the normalised expression levels of all five proteins between OA cases and controls. Higher levels of CRTAC1, COL9A1, ACTA2, EDA2R were observed in patients with OA, whereas lower levels of TACSTD2 were observed in patients with OA (Fig. 6B). Interaction dependencies in the XGBoost model between these top five proteins and the most predictive clinical features of OA risk (age, NSAIDs prescriptions one year before diagnosis and BMI) were tested. Significant interactions were identified between age and CRTAC1, COL9A1, EDA2R and TACSTD2 (Sup Fig. 9). Protein levels of COL9A1 were more important for OA prediction in people with an age above 55. EDA2R was generally seen in lower protein levels for individuals under 60 years old, while increased importance of EDA2R as an OA risk biomarker was seen in older individuals.

Discussion

We present a large-scale study of OA, in a UK cohort of ~20,000 patients with OA and ~20,000 controls, encompassing a broad set of OA risk biomarkers. Using interpretable machine learning, we developed a model predictive of an individual's five-year risk of progression to OA. We mapped the complex landscape of risk biomarkers using both multi-modal clinical data and molecular signatures of OA. We identified sub-groups of OA, characterised by distinct profiles of risk biomarkers. Furthermore, our clustering approach enabled us to derive simple clinical association rules for these sub-groups. Finally, we demonstrated how individual patient journeys can be dissected to identify risk biomarkers for OA, which might allow the development of personalised preventative strategies.

OA has significant health consequences with links to frailty and multimorbidity, highlighting the need for OA prevention strategies[2]. However, to develop appropriate preventative strategies, there is a need for reliable predictive models with robust performance. Previous OA risk studies have either focussed on a specific joint (such as knee OA)[2,27], included relatively small sample sizes, used only questionnaire-based data, incorporated a small set of risk biomarkers, or did not dissect clusters of patients with varying OA risk profiles[4,9,12,16,17]. To realise the full potential of machine learning to advance our understanding of OA risk, there is a need for more diverse studies of OA, encompassing multiple OA subtypes and diverse patient cohorts[9]. The OA risk models described in this paper were built using a large sample size of approximately 20,000 patients with OA, utilising a large set of longitudinal clinical, biochemical, and omics-based risk biomarkers, and identified patient clusters which varied in OA risk profiles. We demonstrated the strong and robust predictive performance of our model, with the model able to correctly predict 7 in 10 patients who developed OA, with 66% of these being true-positive OA cases. The model identified and estimated the predictive impact for a range of risk biomarkers. Some of these were

recognised as OA risk biomarkers (e.g. age and BMI)[11], while others were not traditionally considered OA risk biomarkers (e.g. personal health rating, hand grip strength, body composition, and walking pace). This highlights the heterogeneity in OA pathogenesis and the novelty of the OA risk models in our study. Preventative interventions may need to target multiple risk biomarkers to reduce OA incidence. Most of the predictors included in the model are easy to obtain in clinical practice, although some may require specific testing, such as hand grip strength. Risk biomarkers that are easy to obtain clinically may enable preventative strategies that prioritise interventions to those with the highest risk.

Age and BMI were the top and 3rd predictor of OA in the model respectively, both of which are established risk biomarkers for OA[10]. The impact of BMI on the risk of OA was present even in those who did not have BMI $\geq$ 30 kg/m^2, with lower BMI levels reducing OA risk. This is unsurprising since increases in BMI levels, regardless of the BMI category in which the increase occurs, can increase the risk of obesity-related complications (as seen in type 2 diabetes[18]). NSAID prescription one year prior to OA diagnosis was the second most important predictor. This likely reflects the delay or clinical inertia in the diagnosis of OA, reflecting several barriers to OA management in primary care that have been described previously[19,20].

Higher vitamin D levels were also identified as a predictor of increased OA risk in this model. It was observed that individuals taking vitamin D supplements typically had higher vitamin D levels and a higher age. Therefore, vitamin D levels in the Clin model may be capturing an older population who take vitamin supplements. Vitamin D may affect progression and pain associated with OA[21–23], but the levels of vitamin D and use of supplements may vary across different global societies[24]. Additionally, it was seen that a lower socioeconomic status, as indicated by a lower income and level of education, was predictive of an increased OA risk. This agrees with previous work, that demonstrated that social deprivation, including lower education, was associated with increased OA risk[25,26]. Additionally, lower social deprivation and increased OA risk were previously linked to obesity[25].

There have been efforts to identify sub-groups of OA to enable targeted treatment and management of patients. Sub-groups of OA have been proposed based on a shortlist of blood and urine biomarkers; however, these did not incorporate additional types of patient data[6]. Sub-groups based on comorbid symptoms have also been identified, but this was restricted to older individuals with knee or hip OA[27]. Both studies are limited by identifying sub-groups of established disease, which does not contribute to the understanding of the impact of risk biomarkers on developing OA. By modelling OA risk biomarkers ahead of diagnosis, we were able to identify opportunities for early intervention and prevention of disease progression.

Our data-driven approach allowed identification of distinct sub-groups of OA risk profiles. These results suggest that simple sets of rules may be used to assign most individuals to specific sub-groups, which differ in their predicted OA risk and risk biomarkers. This approach may help inform both clinicians and patients to simply assess the potential for belonging to high-risk groups for OA risk, and the need for an OA diagnostic assessment. The interpretable machine learning approach also provided an understanding

of the contributing factors to the predicted OA risk for specific individuals. This explainability generates trust in the predictions and enables identification of high-risk individuals to target for prevention of disease progression. OA Risk C (calculator.oarisk.org) is an online tool that estimates the risk of an individual developing knee OA[28]. However, it is limited by its restricted set of risk biomarkers. We present a more complex OA risk model, encompassing a wider range of OA risk biomarkers. Our study could contribute to the prioritisation of relevant input data for future, more comprehensive OA risk tools. In addition, our incorporation of modifiable risk biomarkers may inform potential interventions targeting the risk biomarkers. Based on machine learning predictions of OA risk, it cannot be fully assessed whether intervention on risk biomarkers would decrease OA risk, and the extent of this decrease. However, it has previously been shown that reduction in modifiable risk biomarkers corresponded with a reduced individual probability of developing knee OA[12].

Although the integration of omics in our models did not improve the overall performance, it did change the ranking of the risk biomarkers. Some omics biomarkers replaced clinical risk biomarkers, with some being more predictive of OA risk than BMI (CRTAC1 and COL9A1). CRTAC1 was the most predictive protein and has previously been proposed as an OA risk biomarker [29–31]. CRTAC1 is associated with OA diagnosis across multiple joints, and with the severity of OA, including progression to total joint replacement[29–31]. These findings are across multiple cohorts (including the UK Biobank), proteomics technologies, and methods. It has been suggested that CRTAC1 is upregulated in joints in OA by pro-inflammatory cytokines, and there may be a sex specific role for CRTAC1 in OA progression[32]. COL9A1 is both genetically and epigenetically linked to OA with its hypermethylation linked to decreased anabolism in OA[33–37]. Additionally, mutations in COL9A1 are associated with multiple epiphyseal dysplasia, a hereditary condition characterised by early onset OA[36].

The integration of multiple omics biomarkers highlighted relevant biological pathways important for OA disease prediction. An improved understanding of predictive OA molecular risk biomarkers may help guide more OA-specific prevention strategies. In contrast, other clinical risk biomarkers, such as BMI, may reflect a more general risk of obesity-related complications[12].
Proteomics biomarkers were highly important in the prediction of OA risk and other identified proteins included ACTA2 and EDA2R. The smooth muscle cell protein ACTA2 has previously been associated with sub-groups of OA[38], and is associated with clusters of smooth muscle cells in the OA synovium[39]. Lastly, EDA2R has previously been associated with TNF mediated inflammation, in the context of rheumatoid arthritis[40].

It has been demonstrated that multiple genetic risk variants are associated with OA[5]. Although the OA WGPRS had limited predictive value across the general population, we found it may be informative for those with more extreme genetic risk. It has previously been seen that only a small proportion of variance in OA can be explained by genetics[5], and that genetics had a limited ability to predict OA[16,41]. This highlights the importance of including additional clinical and biological data in prediction models, to provide contextual disease information.

Our GRS approach enabled us to estimate the contribution to predicted OA risk from genes previously associated with OA. The GRS for the *TGFB1* locus was the top ranked GRS in the ClinGRS model. TGF-β signalling affects chondrocytes, mesenchymal stem cells and synovial lining cells in OA development and progression[42]. Multiple components of the TGF-β pathway have been genetically associated with OA, including SMAD3, a regulator of *TGFB1* expression that impacts chondrogenic differentiation[42,43]. In phase 3 clinical trials, TGFB1 cell and gene therapy improved function and pain in knee OA[44]. The *GDF5* GRS was predictive for OA and GDF5 has a role in chondrogenesis, a critical process for cartilage and bone development[45]. Higher levels of GDF5 were seen in OA with advanced cartilage damage[46], and GDF5 is currently a target in clinical development for cartilage regeneration indications[43]. Lastly, the *PTCH1* GRS was predictive of OA, and *PTCH1* encodes a transmembrane receptor for Hedgehog ligands. Hedgehog signalling is associated with chondrocyte proliferation, osteogenesis, cartilage degeneration and disease severity in OA[5,47,48]. Loss of *PTCH1* induces OA-like phenotypes[48,49] and *PTCH1* is genetically associated with total hip, knee and joint replacements[5]. Therefore, PTCH1 may be particularly relevant for OA disease progression, and targeting the Hedgehog pathway may be a therapeutic opportunity for OA[47].

Multiple metabolites were relevant for predicting OA and have previous associations with the complex metabolism underlying OA. Acetate was the most predictive metabolite for OA. ACOT12 breaks down acetyl-CoA into acetate and CoA, and is a novel regulator of de novo lipogenesis (DNL) associated cartilage degradation in OA[50]. Valine was also predictive for OA and is an essential branched chain amino acid (BCAA). Valine has previously been associated with the severity of inflammation in synovial tissue[51]. BCAAs may play a role in increased inflammation, reduced autophagy, and increased insulin resistance in OA [52]. The ketone body 3-hydroxybutarate (βHB, β-hydroxybutarate) may have anti-senescence effects which delay OA progression[53]. Finally, citrate levels in synovial fluid have conflicting associations with knee OA[54], and may be linked to altered energy metabolism[55].

There is a complex relationship between fatty acids and OA, that may be impacted by sex, obesity, OA joint and whether measured in the fasted or postprandial state[56,57]. Different fatty acids have distinct effects on OA[58]. Previous work has shown that longer chain saturated fatty acids induced metabolic syndrome and OA-like cartilage degradation[59,60], and induced systemic inflammation, independent of weight-gain in obesity related OA[61].

Currently, omics data may not be easily available in routine clinical practice, although this may change in the future as the utility of omics biomarkers matures. However, even without the inclusion of omics data, our Clin model provided a significant advance in predicting the risk of OA across subgroups and at the individual level, encompassing a broad set of clinical risk biomarkers. The availability of some of the clinical risk biomarkers, such as hand grip strength and regional body composition, might be challenging depending on the health care system and the setting (for example primary vs secondary care). Nonetheless, the top 6 predictors in the Clin model were all easily obtainable. Our study incorporated a

diversity of risk biomarkers, including genetics, proteomics, and metabolomics. Future work would benefit from the incorporation of joint imaging data, WOMAC pain scores, and transcriptomic data, which may contribute to the prediction of OA and a better understanding of biological pathways associated with (progression of) OA[62–64]. It should also be noted that the timing of biological sampling, whether in the UK Biobank or clinical practice, may have an impact on risk biomarker levels and estimated OA risk.

In summary, we present large-scale predictive models for incident OA risk, incorporating a broad range of risk biomarkers. The use of interpretable machine learning for individual patients enabled the identification of personalised modifiable risk biomarkers, opening up the opportunity for tailored OA preventative strategies. Our integration of omics features identified OA-specific risk biomarkers and highlighted the predictive importance of underlying OA disease biology. Taken together, these findings may advance early screening, prevention and treatment of OA, reducing both disease incidence and progression.

Online Methods

UK biobank

The UK Biobank is a human cohort comprising ~ 500,000 individuals, aged between 40 and 69 years at recruitment, from 2006 to 2010[14]. At a recruitment assessment centre, participants contributed to in-depth data collection and banking of biological samples. Biological samples were subsequently used for the generation of various omics data and biomarker measurements. Clinical outcomes for participants can be followed-up through linkage to secondary care data and for ~ 45% of participants primary care data (available data until 31/12 2020). The machine learning models integrated both clinical, metabolomics, proteomics and genetic data from the recruitment assessment centre as well as longitudinal information captured from the EHR data.

Pre-processing of assessment centre data

Clinical assessment centre data

Diverse participant data from the recruitment assessment centre was processed for input into machine learning models to integrate multi-modal signals in the models (Sup File 3). This included data on socio-demographics (including Townsend deprivation index), lifestyle, diet, and physical activity. Additionally, a panel of blood and urine biomarkers were measured at recruitment (Sup File 3) and included as input features. Biomarker data from recruitment was processed to assign outliers outside the 1st and 99th percentiles to the value of the 1st and 99th percentile. Body impedance data reflecting body composition was taken from the first assessment instance and included derived measurement of the fat mass (Kg) to fat-free mass (Kg) ratio across whole body, and specific body areas including trunk area, leg and arms (legs and arms indicated by differences in right and left side of an individual's body composition).

Genetic data

The imputed genetics data was generated and provided by the UK BioBank and processed using PLINK (v2.00a3LM)[65,66]. Single nucleotide polymorphisms (SNPs) were subsequently removed if they had a minor allele frequency (MAF) < 1%, a missingness > 1%, an imputation score (INFO) < 0.8 or with a Hardy-Weinberg Equilibrium exact p-value < 0.00001. This resulted in a set of around 9.4M high quality imputed common genetic variants that were used to generate genetic risk scores (GRS).

Genetic information was integrated into the machine learning models through several representations of the genetic information by i) individual SNPs, ii) whole genome polygenic risk score of OA (WGPRS) as well as genetic risk scores for either iii) specific gene-level genetic risk scores (gene-GRS) or iv) pathway-level polygenic risk scores (pathway-PRS).

Genetic risk scores were generated using the software PRSice (v.2.3.3)[67] and the weights of variants associated with OA from a recent meta-analysis[5] (summary statistics of a sub-analysis excluding the UK Biobank samples).

For gene-GRS, SNPs within 5kb upstream and 1.5kb downstream of a gene were used for scoring after clumping. Furthermore, analyses were run with and without a linkage disequilibrium proxy (LD proxy): including SNPs in LD with the clumped region, with R^2 > 0.8. The gene regions were defined using GENCODE annotations (V43lift37); only protein-coding genes were considered for analyses.

For pathway-PRS, SNPs within genes (as described above) belonging to specific gene-sets were aggregated together into a score (with and without LD proxy). The gene-sets used for the pathway-PRS were generated using KEGG pathways obtained via the Molecular Signature Database (MsigDB; v2022.1).

The default PRSice clumping parameters were used. For WGPRS and gene-GRS/pathway-PRS, no p-value based thresholding was applied to maximise the number of SNPs included in the analyses; however multiple thresholds were tested for WGPRS, with no significant impact on the results (data not shown).

SNPs and genes to be included in the models were guided by using previously identified variants and loci associated with OA risk genes from two recent large GWAS-based studies[5,68]. Indels were removed from individual genetic variants, resulting in a set of 85 SNPs. Specific gene-GRS features to include were prioritised based on the genes annotated to the GWAS significant loci. This included 77 high effector OA genes identified by Boer *et al*[5] and 134 OA genes annotated to genome-wide significant SNPs by McDonald *et al*[68]. In total, 204 unique genes associated to OA risk were identified, where 193 genes were protein-encoding and prioritised for the list of gene-GRS used. All pathways from KEGG were tested for pathway-GRS (n = 186).

Metabolomics data

Circulating metabolite biomarkers from blood EDTA plasma samples were quantified from a subset of UK biobank participants (N = 118,021) using Nightingale's high-throughput proton NMR metabolomics platform[69–71]. A total of 249 metabolites were quantified (168 metabolites given in absolute molar

concentration units (differs per biomarker, see details: https://research.nightingalehealth.com) and 81 ratios of these) spanning amino acids, (apo-)lipoproteins (incl subclasses), cholesterol, cholesteryl esters, fatty acids, biomarkers of fluid balance, glycolysis related metabolites, inflammation biomarkers, ketone bodies, phosphor- and/or other lipids. Metabolites that were measured by absolute concentrations were normalised by square root. Metabolites given as ratio or percentages were not processed further before including in the machine learning models.

Proteomics data

A subset of UK biobank participants (as previously described[72]) had biomarkers measured from plasma blood samples by Olink Proximity Extension assays across 58,634 samples for 54,308 individuals. The data was filtered to only include individuals with samples obtained from the first assessment centre, samples were processed and no withdrawn consent as per reported by UK Biobank (48,040 samples and 46,673 individuals). In total, 1,472 proteomics markers were measured across four Olink panels (cardiometabolic panel: 369, inflammation: 368, neurology: 367, oncology: 368). The quality of the assay and sample quality was checked by Olink's built-in quality control. Sample with < 85% of assays not completely passing Olink's built-in quality control were excluded for analysis. The protein expression was normalised into Normalized Protein eXpression (NPX), which is Olink's arbitrary unit (Log2 scale). NPX values were compared across individuals and extreme outliers were removed if the mean NPX for an individual was < 0.2% or > 99.8% percentile of the overall cohort mean NPX values. Proteins that were identified by non-normal / bimodal residuals from age, sex, BMI were removed from the analysis. CXCL8, IL6, and TNF were measured across all four Olink assays and an average NPX value was calculated. Furthermore, some samples were assayed twice for replication and for machine learning analyses, an average NPX value was calculated.

Pre-processing of longitudinal primary and secondary care data (biomarker & medication) during the 5 years pre-index date

Longitudinal data was captured 5-years prior to OA diagnosis/matched index date from the primary and secondary health care data (Fig. 1 and Sup Fig. 1–2*)*.

Primary care data was used as a source of longitudinal biomarkers, clinical measurements, and prescription data. Biomarkers and clinical measurements included haematological measurements, liver biomarkers, anthropometric measurements, cardiovascular biomarkers, renal biomarkers, bone & joint biomarkers, lifestyle measurements, hormonal measurements, and diabetes biomarkers (Sup File 4). Where appropriate, data was processed to align units and outliers. Medication for type 2 diabetes, obesity and OA was extracted from prescription data (Sup File 5).

Longitudinal data was captured in yearly snapshots 5 years prior to OA diagnosis. For continuous data, this was represented as the median value across an 11-month period in the year. For medication data, prescriptions were grouped into medication classes and a binary value indicated whether the patient had a prescription for the medication class during the year.

Finally, primary and secondary health care data was used to indicate if OA patients had a diagnosis of posttraumatic OA before onset of their OA diagnosis, as this has been shown to increase individual OA risk[73]. Diagnoses of post-traumatic OA were identified and annotated if any codes were given ahead of the OA date / control proxy day.

OA study and validation populations

OA diagnoses were identified from primary and secondary health care data linked to individual participants (> 18 years old) using read2, read3, ICD-9 and ICD-10 clinical codes (Sup File 1). Further inclusion criteria for the study included filtering of patients with available primary care data in the UK Biobank. An equal number of control participants, who never developed OA and had observational data during the entire study period (e.g., due to death), were identified. Controls were date-matched with the OA diagnosis dates for case patients to enable retrospective capture of longitudinal data five year before OA diagnosis or OA proxy date. Cases and controls were then filtered for those with an OA diagnosis/matched index date after the date of the recruitment assessment centre, and maximum five years between the OA diagnosis and assessment centre, allowing the use of this recruitment data in predictive models to estimate the five-year risk (Fig. 1A).

Machine learning models

Extreme Gradient Boosting (XGBoost) models with decision-tree based learners were implemented by R version 4.2.0 using the publicly available R libraries; *xgboost (v 1.6.0.1), pROC (v 1.18)* and *caret (v6.0-93)*. XGboost models were trained using a nested two-level five-fold case-control stratified cross-validation. Missing values were handled by imputation within each cross-validation fold based on the median value available across cases and controls in the training dataset. Features with > 50% missing data were excluded from analysis. Additionally, near zero variance features were excluded using 'near0var' from the *caret* library. Hyperparameters for the cross-validated models were optimised via a cross-validation grid search, where optimal parameters were selected based on highest cross-validated ROC-AUC in the inner cross-validation fold. The grid optimised for the following parameters *eta*: (0.05, 0.10, 0.15, 0.20, 0.25,0.30) and *nrounds*: (50, 100, 200, 300, 500, 700, 1000) with the following other parameters fixed in the XGBoost model; *booster: gbtree, max depth: 10, eval_metric: logloss, objective: binary:logistic, subsample = 0.8, sample:method = uniform* and *min_child_weight = 50*. For models where omics data (genomics, metabolomics or proteomics) were integrated the XGBoost model also had *colsample_bytree = 0.8*. The short depth of each decision tree was set to help guide feature selection and reduce risk of overfitting. The selected hyperparameters were selected for re-training of one model in the outer cross-validation level. The model performance was validated on the outer test set that has been held out from any model training. Model performance was estimated by evaluating classification of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) classifications done by the machine learning model. To account for variance between the five cross-validation sets, both the average and 95% confidence interval (CI) across the five validation datasets were reported. The classification performance is reported by the area under the receiver operating characteristic curve (ROC-AUC),

sensitivity (recall, ($\frac{TP}{TP+FN}$)), positive predictive value (PPV, precision, ($\frac{TP}{TP+FP}$)), specificity ($\frac{TN}{TN+FP}$) and negative predictive value (NPV, $\frac{TN}{TN+FP}$). ROC and precision-recall curves were generated using the R library *yardstick (v 1.1.0)*. The robustness and stability of the machine learning models was evaluated across 100 random model initialisation and against 100 random model initialisations with a permuted prediction outcome of OA cases and controls that both accessed impact on predictive performance. Transparent reporting of a multivariate prediction model for individual prognosis or diagnosis (TRIPOD) checklist for prediction model development were followed to ensure clarity and reproducibility of the multivariate prediction model of individual OA risk (Sup File 6).

Model interpretation

To explore how each individual features contribute to the risk prediction of OA, Shapley Additive explanation (SHAP) values were calculated from the XGBoost models using Tree SHAP[74]. The SHAP values were given as the log-odds of the individual contributions and were visualised using the R libraries *SHAPforxgboost (v 0.1.1)* and *shapviz (v 0.4.1)*. For global estimation of feature importance across the five outer test datasets, the mean absolute SHAP values were calculated, which quantifies, on average, the magnitude of a feature's ability to predict individual risk of osteoarthritis.

Cluster analyses

To explore differences in feature importance in the prediction model between subgroups of individuals, clustering was performed on the SHAP values. For this, the *Seurat*[75] (v.4.3.0) implementation of the Louvain clustering algorithm[76] was used after reducing the data to 10 dimensions by principal component analysis (PCA). A three-steps approach was used to determine the number of clusters. First, the package *chooseR*[77] was used to perform a subsampling-based approach to generate silhouette scores (as a measure of cluster robustness) for various resolutions (number of clusters). Secondly, the PPV, sensitivity and F1 (a balanced metric evaluating accuracy of case classification, $\frac{2}{sensitivity^{-1}+PPV^{-1}}$) values of each cluster were calculated, based on the prediction results of the XGBoost model. Finally, these two steps were integrated by looking at the cluster resolutions leading to the best combination of these metrics with the formula: $clusterscore = \frac{weightedmean(silhouettescores)}{1-weightedmedian(predictionvalues)}$.

After manually selecting the optimal resolution parameter to maximise cluster robustness, number of clusters and per-cluster F1 values (resolution = 0.5, n = 14 clusters), each cluster was first characterised by averaging the values of the top 6 predictive features to plot as a heatmap using the *ComplexHeatmap*[78]R package(v.2.13.1). Secondly, the *SkopeRules* Python package (v. 1.0.1)[15] was employed to generate interpretable rules to define each cluster using the original input values used in the XGBoost model (*n_estimators: int = 10, recall_min: float = 0.2, max_depth: int = 5, max_depth_duplication: int = 7*). When multiple set of rules were given for a cluster, the rules appearing in the highest numbers of decision trees were chosen, and the rules with highest out-of-bag PPV and sensitivity in case of equality.

Declarations

Data Availability

UK biobank is available to researchers following application to the UK biobank database (https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access). All field ID and clinical codes used for extraction of data has been provided in Supplementary files. The use of UK biobank for this study was performed under research application numbers 53639 and 65851.

The OA GWAS summary statistics files used to generate the genetic risk scores were created as part of a published study by Boer et al, Cell 2021.

Code availability

All analyses were performed on publicly available software, and parameters provided wherever relevant. The code used for this study was tailored to the data and the fields of the UK biobank data, and is thus not provided since it is of no use as a standalone without access to the data. However, the authors welcome being contacted to provide more information to reproduce the results presented in this paper if needed.

Acknowledgements

The work was supported by research grants from the Innovation Foundation Denmark and the Danish Diabetes Academy, which is funded by the Novo Nordisk Foundation, grant number NNF17SA0031406.

We also thank Eleftheria Zeggini (Helmholtz Munich / Wellcome Sanger Institute) for providing summary level results from the genetic analyses published in Boer *et al*, 2021 for the analysis subset excluding the UK Biobank samples.

We also thank the UK Biobank participants and researchers for providing this research resource.

Author information

Novo Nordisk Research Centre Oxford, UK.

Rikke Linnemann Nielsen, Thomas Monfeuga, Robert R. Kitchen, Line Egerod, Luis G. Leal, August Thomas Hjortshøj Schreyer, Carol Sun, Marianne Helenius,Zahra McVey, Ramneek Gupta.

University of Oxford, UK.

Carol Sun.

Technical University of Denmark, Denmark.

Marianne Helenius.

Novo Nordisk A/S, Denmark.

Lotte Simonsen, Marianne Willert, Abd A. Tahrani.

Contributions

Contributions to the conception or design of the work: RLN, TM, MW, AAT, ZM, RG.

Acquisition, analysis, or interpretation of data: All authors.

Creation of new software used in the work: RLN, TM.

Drafted the manuscript: RLN, TM, ZM.

Substantively revision and approval of the submitted manuscript: All authors.

Competing interests

RLN, TM, RRK, LGL, ATHS, CS, LS, MW, AAT, ZM and RG are employed by Novo Nordisk. LE is currently employed by Nordic Bioscience A/S and was working on this manuscript while being employed by Novo Nordisk A/S. MH declares no competing interests.

References

1. Leifer, V. P., Katz, J. N. & Losina, E. The burden of OA-health services and economics. Osteoarthritis Cartilage 30, 10–16 (2022).
2. Roos, E. M. & Arden, N. K. Strategies for the prevention of knee osteoarthritis. Nat Rev Rheumatol 12, 92–101 (2016).
3. Cook, M. J., Verstappen, S. M. M., Lunt, M. & O'Neill, T. W. Increased Frailty in Individuals With Osteoarthritis and Rheumatoid Arthritis and the Influence of Comorbidity: An Analysis of the UK Biobank Cohort. Arthritis Care Res (Hoboken) 74, 1989–1996 (2022).
4. Jamshidi, A., Pelletier, J. P. & Martel-Pelletier, J. Machine-learning-based patient-specific prediction models for knee osteoarthritis. *Nature Reviews Rheumatology* vol. 15 49–60 Preprint at https://doi.org/10.1038/s41584-018-0130-5 (2019).
5. Boer, C. G. *et al.* Deciphering osteoarthritis genetics across 826,690 individuals from 9 populations. Cell 184, 4784–4818.e17 (2021).
6. Angelini, F. *et al.* Osteoarthritis endotype discovery via clustering of biochemical marker data. Ann Rheum Dis 81, 666–675 (2022).
7. Stiglic, G. *et al.* Interpretability of machine learning-based prediction models in healthcare. Wiley Interdiscip Rev Data Min Knowl Discov 10, 1–13 (2020).
8. Appleyard, T., Antcliff, D., Thomas, M. & Peat, G. Prediction Models To Estimate Future Individual Risk of Osteoarthritis in the General Population: a Systematic Review. Osteoarthritis Cartilage 30, S22

(2022).

9. Binvignat, M. *et al.* Use of machine learning in osteoarthritis research: A systematic literature review. RMD Open 8, 1–10 (2022).
10. O'Neill, T. W., McCabe, P. S. & McBeth, J. Update on the epidemiology, risk factors and disease outcomes of osteoarthritis. Best Pract Res Clin Rheumatol 32, 312–326 (2018).
11. Palazzo, C., Nguyen, C., Lefevre-Colau, M.-M., Rannou, F. & Poiraudeau, S. Risk factors and burden of osteoarthritis. Ann Phys Rehabil Med 59, 134–138 (2016).
12. Zhang, W. *et al.* Nottingham knee osteoarthritis risk prediction models. Ann Rheum Dis 70, 1599–1604 (2011).
13. Pettit, R. W., Fullem, R., Cheng, C. & Amos, C. I. Artificial intelligence, machine learning, and deep learning for clinical outcome prediction. Emerg Top Life Sci 5, 729–745 (2021).
14. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med 12, 1–10 (2015).
15. Goix, N. *et al.* scikit-learn-contrib/skope-rules v1.0.1. Preprint at https://doi.org/10.5281/zenodo.4316671 (2020).
16. Kerkhof, H. J. M. *et al.* Prediction model for knee osteoarthritis incidence, including clinical, genetic and biochemical risk factors. Ann Rheum Dis 73, 2116–2121 (2014).
17. Black, J. E., Terry, A. L. & Lizotte, D. J. Development and evaluation of an osteoarthritis risk model for integration into primary care health information technology. Int J Med Inform 141, 104160 (2020).
18. Carey, V. J. *et al.* Body fat distribution and risk of non-insulin-dependent diabetes mellitus in women: The nurses' health study. Am J Epidemiol 145, 614–619 (1997).
19. Egerton, T., Diamond, L. E., Buchbinder, R., Bennell, K. L. & Slade, S. C. A systematic review and evidence synthesis of qualitative studies to identify primary care clinicians' barriers and enablers to the management of osteoarthritis. *Osteoarthritis and Cartilage* vol. 25 625–638 Preprint at https://doi.org/10.1016/j.joca.2016.12.002 (2017).
20. Martel-Pelletier, J. *et al.* A new decision tree for diagnosis of osteoarthritis in primary care: international consensus of experts. Aging Clin Exp Res 31, 19–30 (2019).
21. Laslett, L. L. *et al.* Moderate Vitamin D deficiency is associated with changes in knee and hip pain in older adults: A 5-year Longitudinal study. Ann Rheum Dis 73, 697–703 (2014).
22. Joseph, G. B. *et al.* Associations Between Vitamins C and D Intake and Cartilage Composition and Knee Joint Morphology Over 4 Years: Data From the Osteoarthritis Initiative. Arthritis Care Res (Hoboken) 72, 1239–1247 (2020).
23. Park, C. Y. Vitamin D in the prevention and treatment of osteoarthritis: From clinical interventions to cellular evidence. *Nutrients* vol. 11 Preprint at https://doi.org/10.3390/nu11020243 (2019).
24. Amrein, K. *et al.* Vitamin D deficiency 2.0: an update on the current status worldwide. Eur J Clin Nutr 74, 1498–1513 (2020).

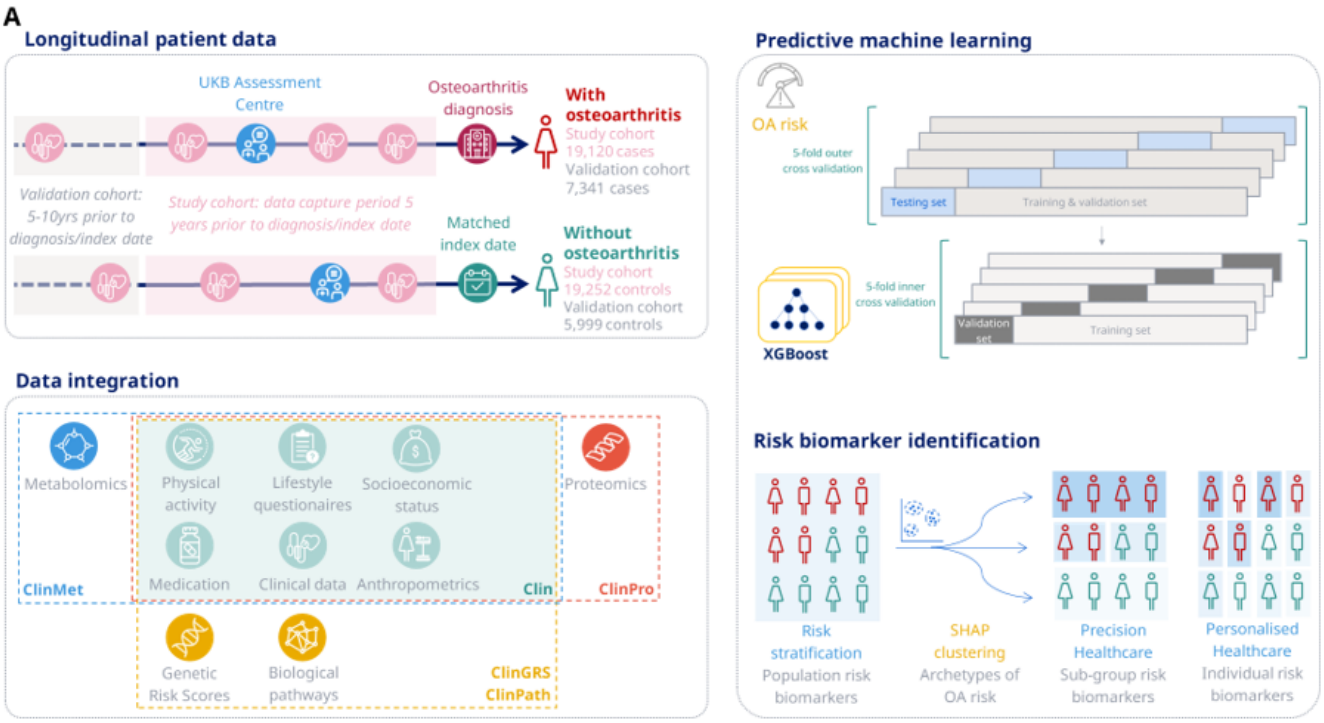
25. Reyes, C. *et al.* Socio-economic status and the risk of developing hand, hip or knee osteoarthritis: A region-wide ecological study. Osteoarthritis Cartilage 23, 1323–1329 (2015).
26. Lee, J. Y., Han, K., Park, Y. G. & Park, S. H. Effects of education, income, and occupation on prevalence and symptoms of knee osteoarthritis. Sci Rep 11, 1–8 (2021).
27. Murphy, S. L., Lyden, A. K., Phillips, K., Clauw, D. J. & Williams, D. A. Subgroups of older adults with osteoarthritis based upon differing comorbid symptom presentations and potential underlying pain mechanisms. Arthritis Res Ther 13, (2011).
28. 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, 1–12 (2015).
29. Styrkarsdottir, U. *et al.* The CRTAC1 Protein in Plasma Is Associated With Osteoarthritis and Predicts Progression to Joint Replacement: A Large-Scale Proteomics Scan in Iceland. Arthritis and Rheumatology 73, 2025–2034 (2021).
30. Styrkarsdottir, U. *et al.* Cartilage Acidic Protein 1 in Plasma Associates With Prevalent Osteoarthritis and Predicts Future Risk as Well as Progression to Joint Replacements: Results From the UK Biobank Resource. Arthritis and Rheumatology 0, 1–9 (2022).
31. Szilagyi, I. *et al.* Plasma Proteomics Identifies Crtac1 As Biomarker for Osteoarthritis Severity and Progression. Ann Rheum Dis 80, 61.1–62 (2021).
32. Xianpeng, G. *et al.* Sex-specific protection of osteoarthritis by deleting cartilage acid protein 1. PLoS One 11, 1–17 (2016).
33. Imagawa, K. *et al.* Association of reduced type IX collagen gene expression in human osteoarthritic chondrocytes with epigenetic silencing by DNA hypermethylation. Arthritis and Rheumatology 66, 3040–3051 (2014).
34. Mustafa, Z. *et al.* Linkage analysis of candidate genes as susceptibility loci for osteoarthritis - Suggestive linkage of COL9A1 to female hip osteoarthritis. Rheumatology 39, 299–306 (2000).
35. Hu, K. *et al.* Pathogenesis of osteoarthritis-like changes in the joints of mice deficient in type IX collagen. Arthritis Rheum 54, 2891–2900 (2006).
36. Li, Y., Xu, L. & Olsen, B. R. Lessons from genetic forms of osteoarthritis for the pathogenesis of the disease. Osteoarthritis Cartilage 15, 1101–1105 (2007).
37. Alizadeh, B. Z. *et al.* Evidence for a role of the genomic region of the gene encoding for the α1 chain of type IX collagen (COL9A1) in hip osteoarthritis: A population-based study. Arthritis Rheum 52, 1437–1442 (2005).
38. Soul, J. *et al.* Stratification of knee osteoarthritis: Two major patient subgroups identified by genome-wide expression analysis of articular cartilage. Ann Rheum Dis 77, 423–430 (2018).
39. Chou, C. H. *et al.* Synovial cell cross-talk with cartilage plays a major role in the pathogenesis of osteoarthritis. Sci Rep 10, 1–14 (2020).

40. Xiao, H., Bartoszek, K. & Lio', P. Multi-omic analysis of signalling factors in inflammatory comorbidities. BMC Bioinformatics 19, 1–18 (2018).
41. Takahashi, H. *et al.* Prediction model for knee osteoarthritis based on genetic and clinical information. Arthritis Res Ther 12, 1–6 (2010).
42. Shen, J., Li, S. & Chen, D. TGF-β signaling and the development of osteoarthritis. Bone Res 2, (2014).
43. Tachmazidou, I. *et al.* Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data. Nat Genet 51, 230–236 (2019).
44. Kim, M. K. *et al.* A Multicenter, Double-Blind, Phase III Clinical Trial to Evaluate the Efficacy and Safety of a Cell and Gene Therapy in Knee Osteoarthritis Patients. Hum Gene Ther Clin Dev 29, 48–59 (2018).
45. Sun, K., Guo, J., Yao, X., Guo, Z. & Guo, F. Growth differentiation factor 5 in cartilage and osteoarthritis: A possible therapeutic candidate. Cell Prolif 54, 1–13 (2021).
46. Kania, K. *et al.* Regulation of Gdf5 expression in joint remodelling, repair and osteoarthritis. Sci Rep 10, 1–11 (2020).
47. Xiao, W. feng, Li, Y. sheng, Deng, A., Yang, Y. tao & He, M. Functional role of hedgehog pathway in osteoarthritis. *Cell Biochem Funct* 38, 122–129 (2020).
48. Lin, A. C. *et al.* Modulating hedgehog signaling can attenuate the severity of osteoarthritis. Nat Med 15, 1421–1425 (2009).
49. Deng, Q. *et al.* Activation of hedgehog signaling in mesenchymal stem cells induces cartilage and bone tumor formation via wnt/β - catenin. Elife 8, 1–24 (2019).
50. Park, S., Baek, I. J., Ryu, J. H., Chun, C. H. & Jin, E. J. PPARα – ACOT12 axis is responsible for maintaining cartilage homeostasis through modulating de novo lipogenesis. Nat Commun 13, 1–12 (2022).
51. Murillo-Saich, J. D. *et al.* Synovial tissue metabolomic profiling reveal biomarkers of synovial inflammation in patients with osteoarthritis. Osteoarthr Cartil Open 4, 100295 (2022).
52. Rockel, J. S. & Kapoor, M. The metabolome and osteoarthritis: Possible contributions to symptoms and pathology. Metabolites 8, (2018).
53. Xia, G. *et al.* β-Hydroxybutyrate alleviates cartilage senescence through hnRNP A1-mediated up-regulation of PTEN. Exp Gerontol 175, 112140 (2023).
54. Akhbari, P. *et al.* Differences in the composition of hip and knee synovial fluid in osteoarthritis: a nuclear magnetic resonance (NMR) spectroscopy study of metabolic profiles. Osteoarthritis Cartilage 27, 1768–1777 (2019).
55. Mickiewicz, B. *et al.* Metabolic analysis of knee synovial fluid as a potential diagnostic approach for osteoarthritis. Journal of Orthopaedic Research 33, 1631–1638 (2015).
56. Felson, D. T. *et al.* Fatty acids and osteoarthritis: the MOST study. Osteoarthritis Cartilage 29, 973–978 (2021).

57. Loef, M. *et al.* The association of plasma fatty acids with hand and knee osteoarthritis: the NEO study. Osteoarthritis Cartilage 28, 223–230 (2020).
58. Loef, M., Schoones, J. W., Kloppenburg, M. & Ioan-Facsinay, A. Fatty acids and osteoarthritis: different types, different effects. *Joint Bone Spine* vol. 86 451–458 Preprint at https://doi.org/10.1016/j.jbspin.2018.07.005 (2019).
59. Sekar, S. *et al.* Saturated fatty acids induce development of both metabolic syndrome and osteoarthritis in rats. Sci Rep 7, 1–11 (2017).
60. Prasadam, I., Sr., Y. & Xiao, V. Effects of dietary saturated fatty acid consumption on cartilage health and trauma-induced osteoarthritis in rats. Osteoarthritis Cartilage 26, S12 (2018).
61. Tan, L., Harper, L. R., Armstrong, A., Carlson, C. S. & Yammani, R. R. Dietary saturated fatty acid palmitate promotes cartilage lesions and activates the unfolded protein response pathway in mouse knee joints. PLoS One 16, 1–13 (2021).
62. 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).
63. Swan, A. L. *et al.* A machine learning heuristic to identify biologically relevant and minimal biomarker panels from omics data. BMC Genomics 16, 1–12 (2015).
64. Gandhi, R., Tsvetkov, D., Dhottar, H., Davey, J. R. & Mahomed, N. N. Quantifying the pain experience in hip and knee osteoarthritis. Pain Res Manag 15, 224–228 (2010).
65. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81, 559–575 (2007).
66. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7 (2015).
67. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience 8, giz082 (2019).
68. McDonald, M.-L. N. *et al.* Novel genetic loci associated with osteoarthritis in multi-ancestry analyses in the Million Veteran Program and UK Biobank. Nat Genet 54, 1816–1826 (2022).
69. Soininen, P., Kangas, A. J., Würtz, P., Suna, T. & Ala-Korpela, M. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. Circ Cardiovasc Genet 8, 192–206 (2015).
70. Würtz, P. *et al.* Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Large-Scale Epidemiology: A Primer on -Omic Technologies. Am J Epidemiol 186, 1084–1096 (2017).
71. Julkunen, H. *et al.* Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank. Nat Commun 14, 604 (2023).
72. Sun, B. B. *et al.* Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. *bioRxiv* 2022.06.17.496443 (2022) doi:10.1101/2022.06.17.496443.

73. Lotz, M. K. New developments in osteoarthritis: Posttraumatic osteoarthritis: pathogenesis and pharmacological treatment options. Arthritis Res Ther 12, 211 (2010).
74. Lundberg, S. M. & Lee, S.-I. Consistent feature attribution for tree ensembles. CoRR abs/1706.0, (2017).
75. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol 33, 495–502 (2015).
76. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008, P10008 (2008).
77. Patterson-Cross, R. B., Levine, A. J. & Menon, V. Selecting single cell clustering parameter values using subsampling-based robustness metrics. BMC Bioinformatics 22, 39 (2021).
78. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32, 2847–2849 (2016).

Figures



B

	OA cohort		
	No OA (N=19252)	OA (N=19120)	P-value
Sex			
F	10411 (54.1%)	11282 (59.0%)	<0.001
M	8841 (45.9%)	7838 (41.0%)	
Age			
Mean (SD)	57.8 (8.22)	61.8 (7.17)	<0.001
Median [Min, Max]	58.0 [40.0, 75.0]	63.0 [41.0, 75.0]	
BMI			
Mean (SD)	26.9 (4.49)	28.6 (5.16)	<0.001
Median [Min, Max]	26.3 [14.7, 62.8]	27.8 [14.3, 62.7]	
Missing	104 (0.5%)	113 (0.6%)	

	Hold-out set		
	No OA (N=5999)	OA (N=7341)	P-value
Sex			
F	3268 (54.5%)	4345 (59.2%)	<0.001
M	2731 (45.5%)	2996 (40.8%)	
Age			
Mean (SD)	55.8 (8.18)	58.6 (7.45)	<0.001
Median [Min, Max]	56.0 [39.0, 71.0]	60.0 [40.0, 71.0]	
BMI			
Mean (SD)	27.2 (4.48)	28.4 (4.96)	<0.001
Median [Min, Max]	26.6 [16.9, 55.0]	27.7 [16.9, 67.4]	

C

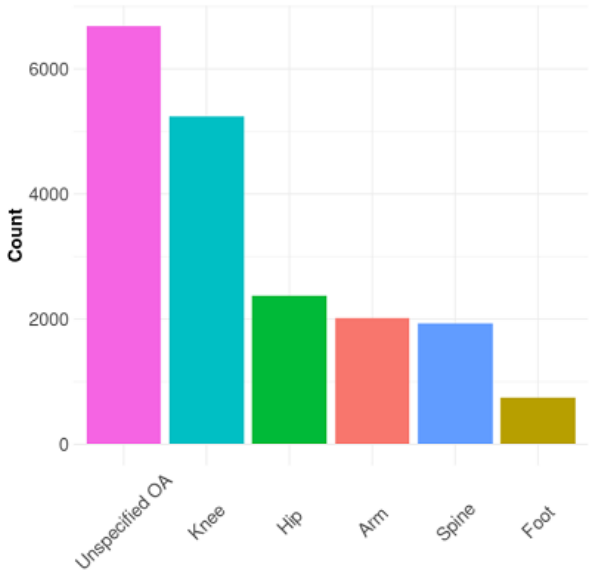
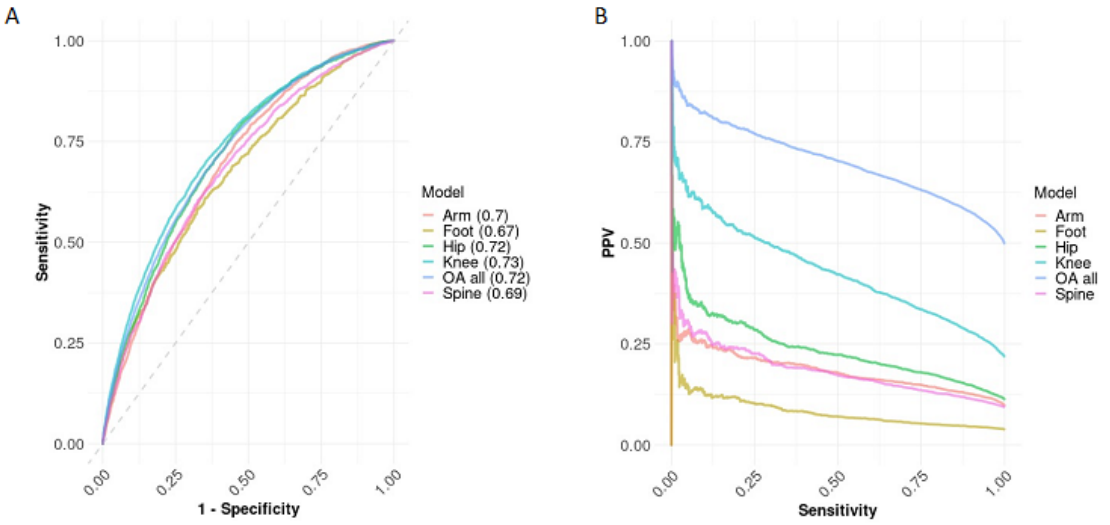


Figure 1

Study design and population characteristics. 1A) Overview of study design including example of date matching cases and controls (longitudinal patient data). For cases with OA, the OA diagnosis date was identified and a data capture period of 5 years prior to diagnosis created. For controls without OA, a matched index date, equivalent to the OA diagnosis date for the case used for matching, was identified. For controls, a data capture period of 5 years prior to the index date was created. For both cases and

controls, longitudinal EHR data and data from the UK biobank assessment centre were captured in the 5-year data capture period. Data integration shows the different models and data integrated for prediction of OA risk. Predictive machine learning shows the machine learning modelling setup for training and model validation as well as interpretation for precision and personalised OA healthcare.1B) OA risk factors summarised across OA cases and matched non-OA controls used for modelling in the study. P-values generated by two-sided t.test for continuous features and chi-squared test of independence for sex. 1C) Count of joints affected that could be extracted from the OA diagnosis codes. These groups were used for stratification in the prediction models. When a patient with OA had multiple joints affected, both diagnoses were included in this joint mapping. It was not possible to map all OA diagnoses to a specific joint.



C

	ROC-AUC	Sensitivity	PPV	Specificity	NPV
Clin	0.72 (0.71 – 0.73)	0.67 (0.66 – 0.68)	0.66 (0.65 – 0.67)	0.65 (0.64 – 0.66)	0.67 (0.66 – 0.67)

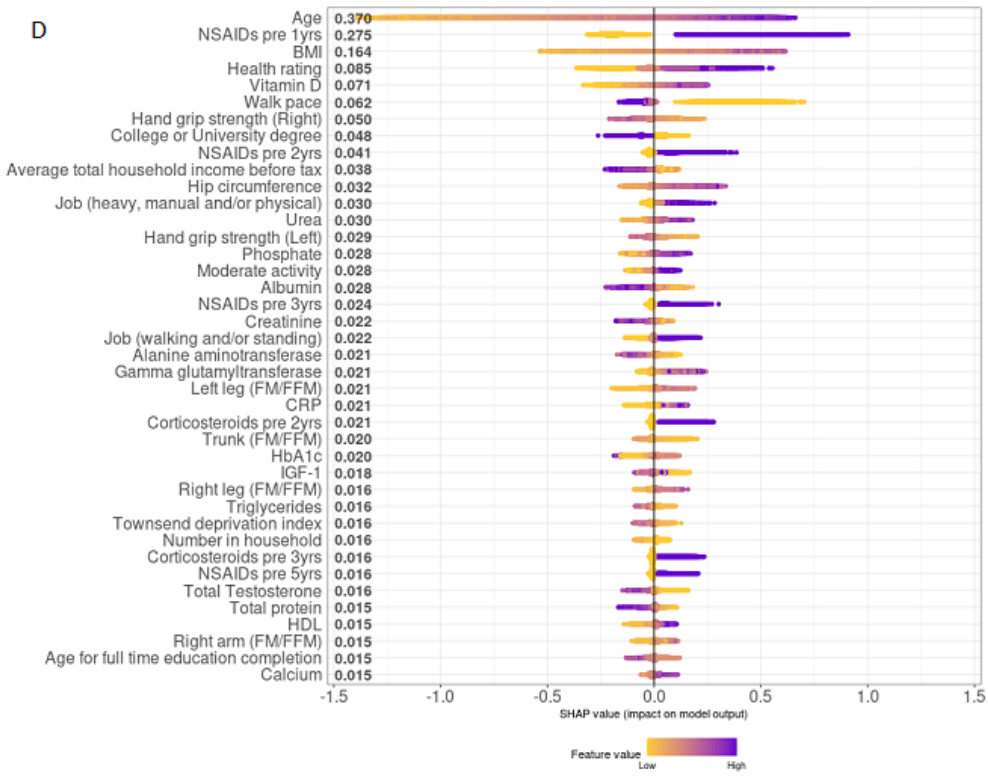


Figure 2

Clinical prediction model of OA risk (Clin model). 2A) ROC curves of OA prediction models 2B) Precision-recall curves of OA prediction models. 'OA all' incudes all cases of OA independent of specific joint subsets (Arm, Foot, Hip, Knee, Spine). An OA case can have multiple OA joints affected and a case is included per joint affected (meaning these can be repeated). 2C) Performance metrics of clinical model on independent five-fold cross-validation test dataset. PPV = positive predictive value. NPV = Negative

predictive value. 2D) Ranked feature importance of OA model by SHAP additive explanations. NSAIDs = non-steroidal anti-inflammatory steoroid drugs. FM/FFM = Fat mass / Fat-free mass.

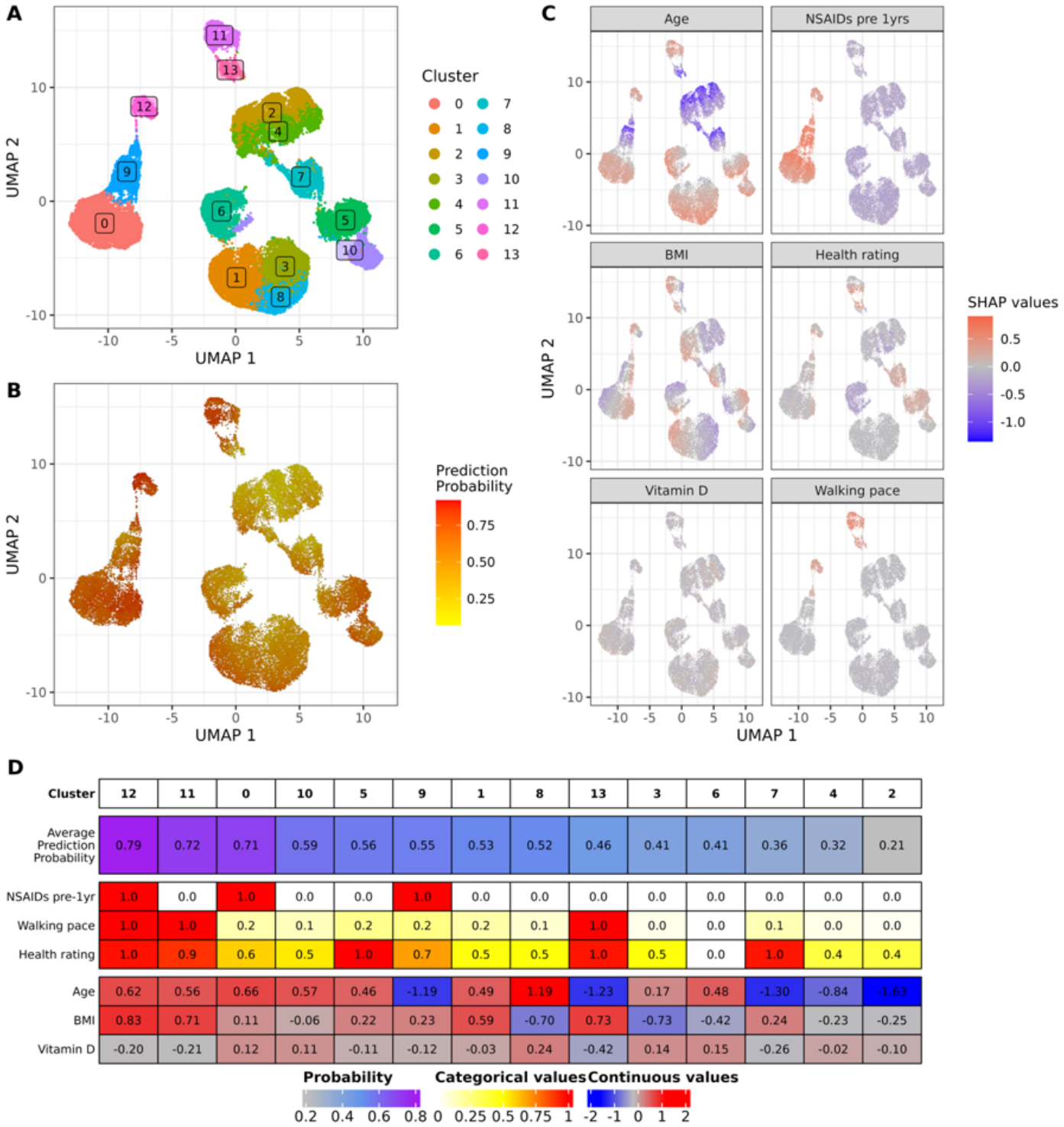


Figure 3

OA patient clustering. 3A) Clusters obtained based on the SHAP values (Louvain clustering algorithm). 3B) Prediction probability per individuals. 3C) SHAP values used for clustering; the colour scheme allows

visualisation of the importance of each feature in the prediction model as well as their impact on clustering. For (3B) and (3C), points were binned to increase readability; the average values within each of the bins are plotted. Clusters were obtained after dimensionality reduction of the SHAP data with a principal component analysis (10 PCs) and visualised after further dimensionality reduction with UMAP. 3D) Average values of prediction probabilities and the top 6 most predictive features. Categorical values encoding before averaging: NSAIDs (pre-1years): 1 and 0 represent patient taking or not taking the drug, respectively; walking paces were rescaled from 0 to 1 corresponding to a range from fastest to slowest; health ratings were rescaled from 0 to 1 corresponding to a range, from healthiest to least healthy. Continuous values encoding before averaging values were transformed into Z-scores for visualisation purposes.

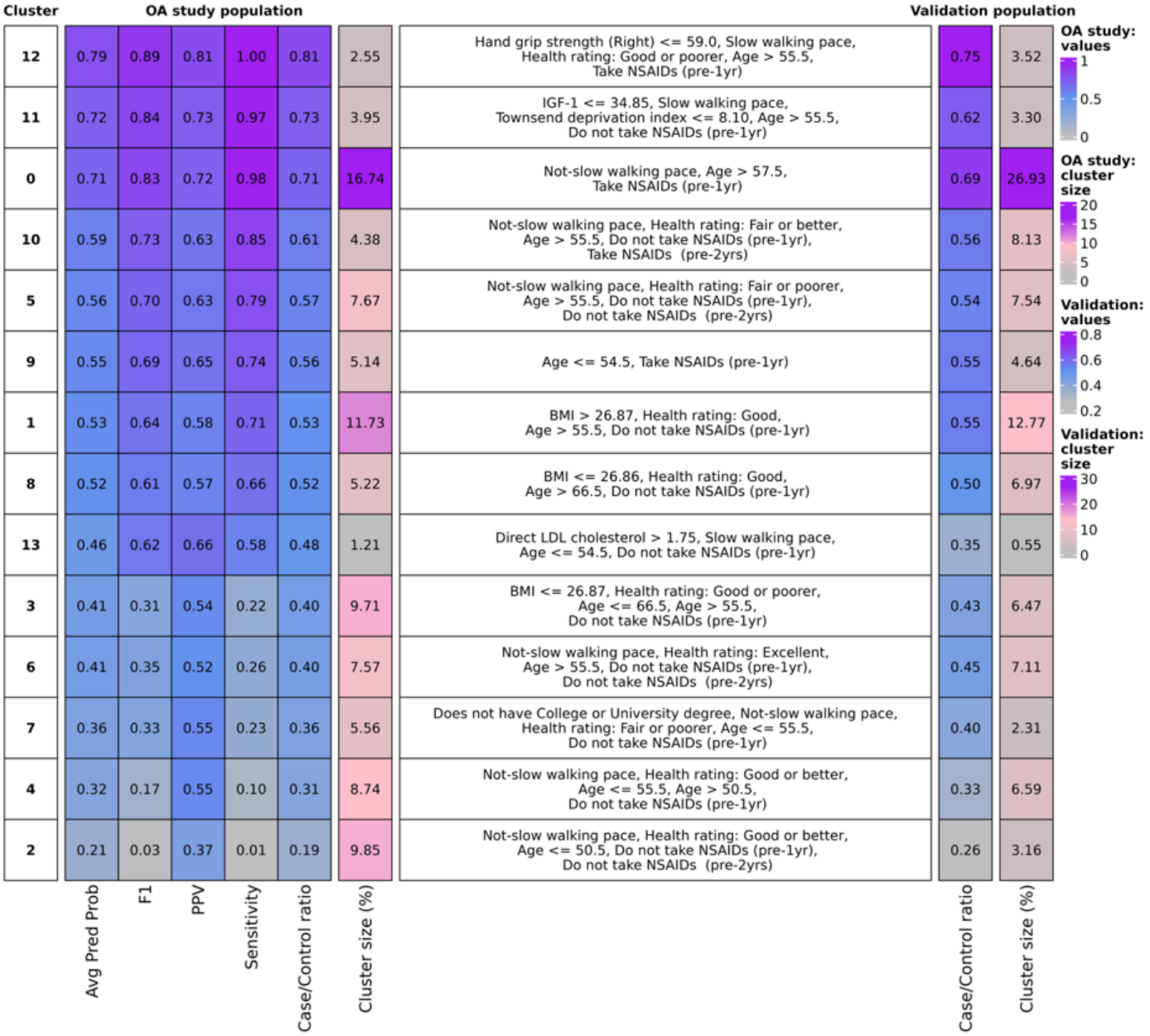


Figure 4

Cluster prediction metrics, defining rules and validation in an independent population. **Left**(heatmap): each cluster is defined by prediction metrics, case/control ratio, cluster size (%). **Middle** (text): set of rules best defining each cluster, based on the model input values and generated by a decision tree model. **Right**(heatmap): Case/control ratio and cluster size (%) in an independent population in which individuals were attributed to clusters according to their corresponding rules.

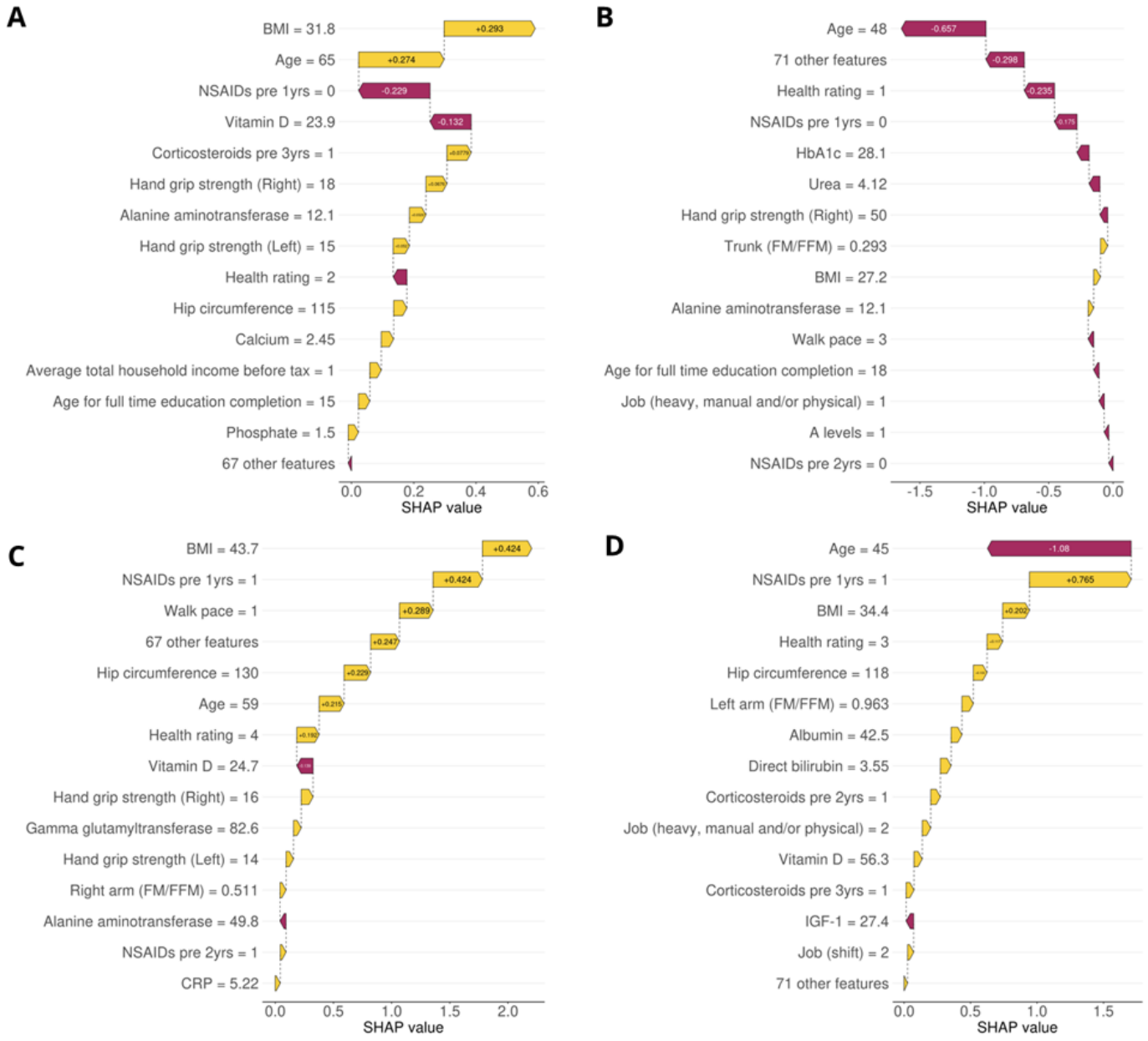


Figure 5

Individual risk profiles for example patients from 5A) cluster 1, 5B) cluster 2, 5C) cluster 12 and 5D) cluster 9. Waterfall plots show the top 15 most important features for estimating the OA risk at the

individual level. Yellow bars (positive SHAP value) indicate features that increased predicted OA risk; red bars (negative SHAP value) indicate features that decreased predicted OA risk. Numbers within bars represent the SHAP value for the feature; numbers on the y-axis represent the value for this feature, both are specific to the individual shown and represented the magnitude of the effect of the risk biomarker on predicted OA risk.

A

Model name	ClinSNP	ClinGRS	ClinPath	ClinMet	ClinPro
Omics data	Confident OA SNPs	Gene OA Risk Score	KEGG pathway OA PRS	NMR metabolomics	Olink proteomics
Number of omics features	85	193	186	249	1461
Number of Cases/Controls	18,625 / 18,779	18,625 / 18,779	18,625 / 18,779	4,502 / 4,519	1,723 / 1,816
ROC-AUC	0.72	0.72 [1,2]	0.72 [1,2]	0.72	0.70
Top 1 feature		*TGFB1* [1,2]*	TGF beta signaling pathway [1,2]*	Acetate	CRTAC1
Top 2 feature		*GDF5* [1,2]*	Glycosphingolipid biosynthesis globo series [2]	Valine	COL9A1
Top 3 feature		*PTCH1* [1,2]	Adipocytokine signaling pathway [1]*	3-Hydroxybutyrate	ACTA2
Top 4 feature		*FAM53A* [1]	Cytokine cytokine receptor interaction [1]*	Citrate	EDA2R
Top 5 feature				Saturated fatty acids to total fatty acids percentage	TACSTD2

(1): GRS obtained with proxy
(2): GRS obtained without proxy
* Identified for models with genetic features corrected for population stratification

B

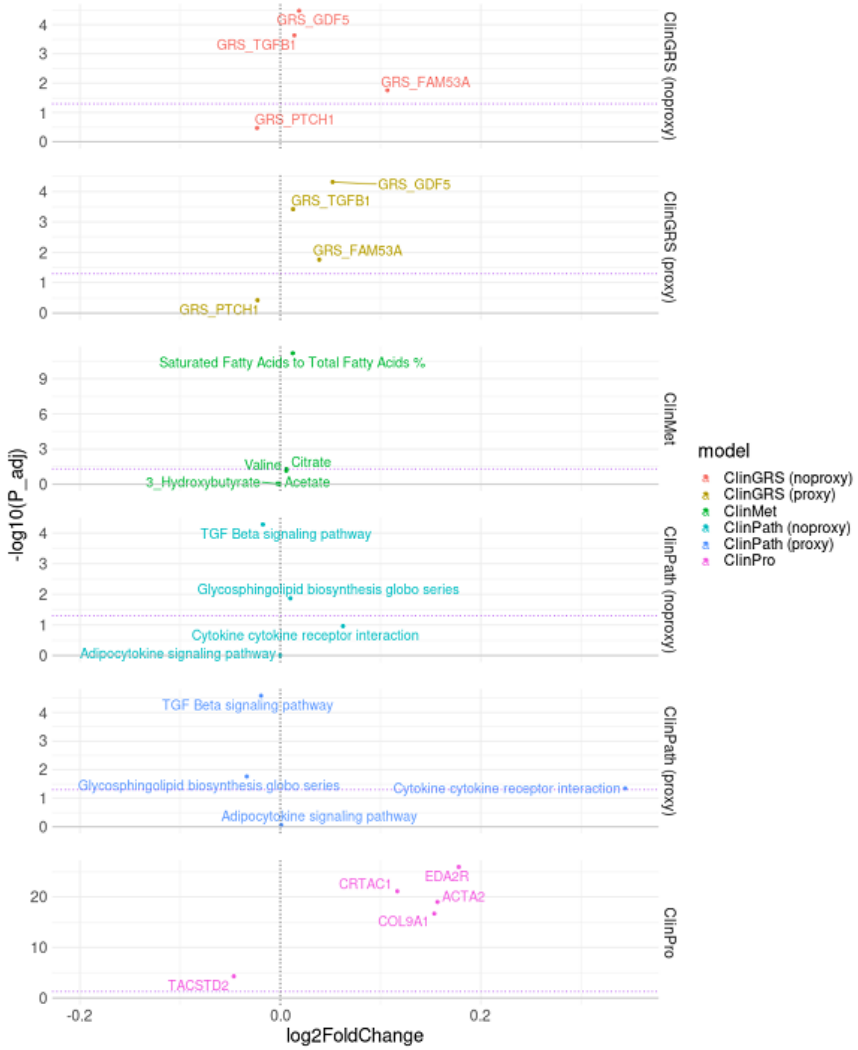


Figure 6

6A) Top ranked features from omics models in the context of multi-modal clinical features. The top 5 omics features that appeared important for prediction of OA based on the average marginal SHAP value ranking amongst top 40 predictive features. For ClinSNP, ClinGRS and ClinPath, several sensitivity checks were done for the genetic features including results marked with: (1): GRS obtained with proxy, (2): GRS obtained without proxy and * Identified for models with genetic features corrected for population stratification (Sup Table 3 for details). 6B) Sparse volcano plot of log2 fold changes of identified 'omics features between OA cases in relation to controls across the six molecular OA models. P-values generated by two-sided t.test and FDR-adjusted (P_adj). The horisontal dotted purple line corresponds to P_adj=0.05.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplFile1OAclinicalcodes.xlsx
- SupplFile2missingclinicaldata.xlsx
- SupplFile3InputFeaturesAC.xlsx
- SupplFile4BiomarkerReadCodes.xlsx
- SupplFile5PrimaryMedicationCodes.xlsx
- SupplFile6TripodCheclistPredictionModelDevelopment.pdf
- OAsupplv27.docx