OPEN ACCESS

TRANSLATIONAL SCIENCE

Predicting rapid progression in knee osteoarthritis: a novel and interpretable automated machine learning approach, with specific focus on young patients and early disease

Simone Castagno ,[1] Mark Birch,[1] Mihaela van der Schaar,[2] Andrew McCaskie[1]

Handling editor Josef S Smolen

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/ard-2024-225872).

[1]Department of Surgery, University of Cambridge, Cambridge, UK
[2]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK

Correspondence to
Dr Simone Castagno; sc2257@cam.ac.uk

Received 26 March 2024
Accepted 13 August 2024

ABSTRACT

Objectives To facilitate the stratification of patients with osteoarthritis (OA) for new treatment development and clinical trial recruitment, we created an automated machine learning (autoML) tool predicting the rapid progression of knee OA over a 2-year period.

Methods We developed autoML models integrating clinical, biochemical, X-ray and MRI data. Using two data sets within the OA Initiative—the Foundation for the National Institutes of Health OA Biomarker Consortium for training and hold-out validation, and the Pivotal Osteoarthritis Initiative MRI Analyses study for external validation—we employed two distinct definitions of clinical outcomes: Multiclass (categorising OA progression into pain and/or radiographic) and binary. Key predictors of progression were identified through advanced interpretability techniques, and subgroup analyses were conducted by age, sex and ethnicity with a focus on early-stage disease.

Results Although the most reliable models incorporated all available features, simpler models including only clinical variables achieved robust external validation performance, with area under the precision-recall curve (AUC-PRC) 0.727 (95% CI: 0.726 to 0.728) for multiclass predictions; and AUC-PRC 0.764 (95% CI: 0.762 to 0.766) for binary predictions. Multiclass models performed best in patients with early-stage OA (AUC-PRC 0.724–0.806) whereas binary models were more reliable in patients younger than 60 (AUC-PRC 0.617–0.693). Patient-reported outcomes and MRI features emerged as key predictors of progression, though subgroup differences were noted. Finally, we developed web-based applications to visualise our personalised predictions.

Conclusions Our novel tool's transparency and reliability in predicting rapid knee OA progression distinguish it from conventional 'black-box' methods and are more likely to facilitate its acceptance by clinicians and patients, enabling effective implementation in clinical practice.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Osteoarthritis (OA) is a common degenerative joint disease which causes pain, stiffness and reduced joint motion. It affects over 500 million people globally, leading to significant healthcare costs.
⇒ The heterogeneity of OA makes the development of effective clinical therapies challenging with current treatments focusing on symptomatic relief and, in advanced cases, joint replacement.
⇒ Identifying patients at risk for rapid OA progression is a fundamental aspect of accurate patient stratification, allowing for the development of new treatments and more strategic clinical trial recruitment (especially in younger patients and those with early-stage disease).
⇒ Machine learning (ML) is recognised as having significant potential in early-stage disease prediction.

WHAT THIS STUDY ADDS

⇒ We developed an autoML tool for predicting rapid knee OA progression, using clinical, X-ray, MRI and biochemical data.
⇒ We demonstrated robust performance of models which included only clinical or 'core' variables, facilitating their practical implementation in clinical settings where extensive data collection is not always feasible.
⇒ We identified key predictors of OA progression, such as patient-reported outcome measures (PROMs) and MRI features, enhancing model transparency and potential for clinical adoption.
⇒ To the best of our knowledge, we are the first to apply these predictive models and assess feature importance in multiple subgroups of patients with OA.
⇒ We also developed web-based applications called clinical demonstrators to facilitate understanding and visualisation of our models' personalised predictions.

INTRODUCTION

Osteoarthritis (OA) is a degenerative joint disease whose primary symptoms are pain, stiffness and reduced joint motion.[1 2] OA affects over 500 million people worldwide[3] and its direct healthcare costs are estimated to be 1–2.5% of national gross domestic product.[4] The heterogeneity of the disease presents a significant challenge in developing effective clinical therapies.[5–7] As a result, there is a clear global unmet clinical need with no approved treatments to halt or reverse disease progression.[8] The primary treatment options remain focused on providing symptomatic relief and, in advanced cases, resorting to prosthetic joint replacement.[9]

Identifying patients at risk for rapid OA progression is crucial for accurate patient stratification, particularly in the early stages of the disease and

To cite: Castagno S, Birch M, van der Schaar M, *et al*. *Ann Rheum Dis* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/ard-2024-225872

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Our predictive models for knee OA progression could significantly impact clinical practice by enabling earlier and more accurate identification of high-risk patients (particularly younger individuals and those with early-stage OA), thereby guiding timely and targeted interventions.

⇒ By integrating PROMs, clinical, biochemical, and imaging data into our models, our approach has the potential to be extrapolated to other complex degenerative diseases, which often rely on this information for patient monitoring and management.

among younger patients, a demographic increasingly affected by OA.[10 11] This stratification is key to successful patient selection in clinical trials to develop and evaluate new treatments.[12] Younger individuals frequently face a 'treatment gap',[13] where conservative management often falls short in managing symptoms and arthroplasty, though potentially beneficial, may not suit their active lifestyles and carries a higher risk of aseptic loosening and future revision surgery. To optimise non-surgical and surgical approaches ahead of joint replacement (including regenerative therapies aimed at joint preservation), a stratified approach is necessary. We hypothesise that machine learning (ML), a branch of artificial intelligence (AI) that uses algorithms to learn from data and make predictions without being explicitly programmed to do so,[14 15] can be leveraged to identify individuals with OA who are at risk of rapid progression, especially in early stages of disease.

In this study, we introduce and validate an innovative, interpretable automated ML (autoML) tool to predict the rapid progression of knee OA (the most common form of OA[4 16]), focusing on early-stage disease and a younger patient demographic.

METHODS

Data sets

Data used in this study are from the Osteoarthritis Initiative (OAI),[17] a multicentre, longitudinal, prospective observational study of 4796 men and women aged 45–79, designed to identify biomarkers and risk factors for the development and progression of OA. OAI data are publicly available and can be accessed at https://nda.nih.gov/oai/.

ML models were developed and trained using data from the Foundation for the National Institutes of Health (FNIH) OA Biomarkers Consortium Project,[18 19] a nested case-controlled study of 600 patients (one index knee per participant) selected from the OAI. Patients in this study were followed up for a total of 4 years at 1-year intervals, and inclusion criteria included the presence of at least one knee with a Kellgren and Lawrence grade (KLG) of 1–3 at baseline.[12] Data collected included clinical, radiographic (X-ray), MRI, blood and urine biospecimen data[17] in a tabular format.

Additionally, patients from the Pivotal Osteoarthritis Initiative MRI Analyses (POMA) study[20–22] were used to further validate our models. POMA is a nested case-controlled study within the OAI, aimed at understanding the progression of OA using MRI.

Preprocessing and class definition

We followed a similar methodology to that used by Widera *et al*.[23] For each patient, we used all available periods that were 2 years in length (ie, baseline to year 2, year 1 to year 3 and year 2 to year 4): Therefore, each instance represented a period, rather than a patient. For each period, we defined four outcome classes:

- **Class 0**: No disease progression.
- **Class 1**: Pain-only progression—based on Western Ontario and McMaster (WOMAC) pain scores (ranging 0–20).
- **Class 2**: Radiographic-only progression—based on minimum medial joint space width (JSW) and KLG
- **Class 3**: Both pain and radiographic progression.

The exact definitions of pain and radiographic progression are as follows:

Pain progression (Eq. 1):

An increase of at least 2 points in the WOMAC pain scale over a two-year period ($\Delta p \geq 2$) AND substantial pain at the end of the period ($p_{end} \geq 8$)

OR

A rapid increase in pain ($\Delta p \geq 4$) AND a lower end pain ($p_{end} \geq 7$)

OR

Sustained substantial pain throughout the period ($p_{start} \geq 8$ AND $p_{end} \geq 8$).

$$(\Delta p \geq 2 \cap p_{end} \geq 8) \cup (\Delta p \geq 4 \cap p_{end} \geq 7) \cup (p_{start} \geq 8 \cap p_{end} \geq 8) \quad \text{(Eq.1)}$$

Radiographic progression (Eq. 2):

A decrease in minimum medial JSW of at least 0.6 mm over a 2-year period

A KLG of 4 at the end of the period ($KLG_{end} = 4$)—this condition was introduced to identify patients with radiographic 'end-stage' OA at the end of the period, independently of medial JWS narrowing.

$$(\Delta width \leq -0.6\ mm) \cup (KLG_{end} = 4) \quad \text{(Eq.2)}$$

Periods were excluded if the outcome class could not be assigned due to missing values, resulting in a total of 1691 instances. Variables with more than 85% missing values and those not relevant to our analysis, such as patient ID, visit number, dates and barcodes were also removed. This resulted in a total of 304 features for analysis. Online supplemental table 1 shows all variables with their definitions.

The above process was then repeated using binary class labels only, with Class 0 representing 'non-progressors' and Class 1 'progressors'.

We performed an 80–20 training-testing split on the data set, ensuring that instances with the same patient ID were consistently placed in either the training or testing set. This resulted in a training set with 1353 instances and a hold-out (or testing) set with 338. Model development and training were exclusively conducted on the training set while the testing set was held out for further validation (figure 1 shows a schematic overview of our study methodology).

Model development using AutoPrognosis V.2.0

AutoPrognosis V.2.0 was used to develop models predicting accelerated knee OA progression. The framework, which is an updated and enhanced version of the original AutoPrognosis,[24] uses advanced optimisation techniques to automatically create a weighted ensemble of ML pipelines, tailored to the specific variables and outcomes of the study population.[25] These pipelines include choices for data imputation, feature processing and classification algorithms, along with their respective hyperparameters. AutoPrognosis V.2.0 design space encompasses 7 feature scaling algorithms, 7 feature selection algorithms, 12 imputation algorithms and 23 classification algorithms (full list in online supplemental table 2). In this study, to enhance computational efficiency, we used the default classification algorithms of

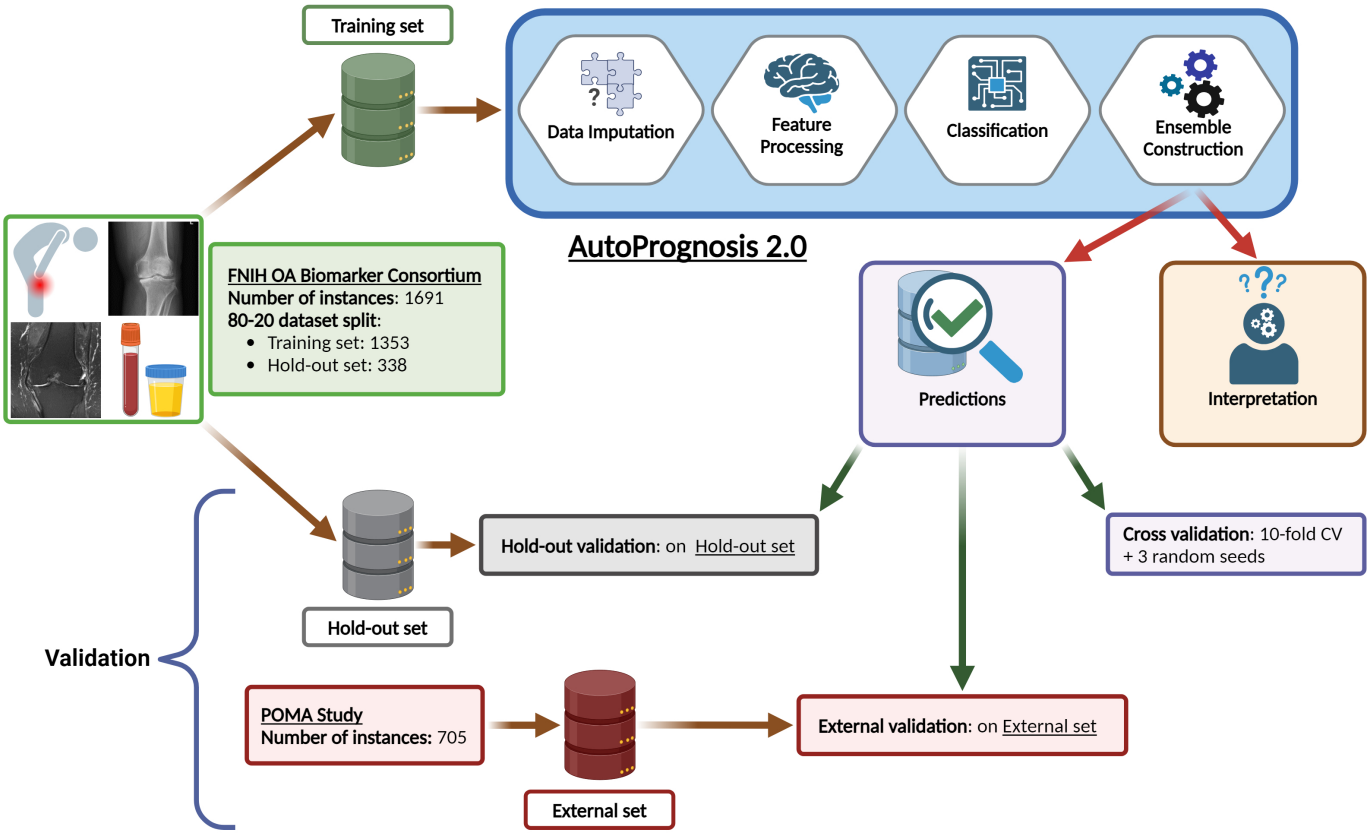


Figure 1 Methodology overview. This figure delineates our methodical approach towards model development and validation. Initially, our data set underwent a random partitioning: 80% allocated to the training set ($N_{training}$=1353) and 20% to the hold out (or testing) set ($N_{hold\text{-}out}$=338). The training phase was strictly confined to the training set preserving the testing set for subsequent validation. Predictive models for rapid knee OA progression were built using AutoPrognosis V.2.0, and key predictors of progression were identified through post-hoc interpretability analysis. Model reliability was rigorously evaluated via internal validation on the training set and hold-out validation on the testing set. Additionally, further validation was conducted by testing our clinical and streamlined models (incorporating only the top five predictors) on patients from the POMA study. (Created with BioRender.com). FNIH, Foundation for the National Institutes of Health; OA, osteoarthritis; POMA, Pivotal Osteoarthritis Initiative MRI Analyses.

AutoPrognosis V.2.0 (highlighted in bold in online supplemental table 2), selected for their speed and efficiency.

Our model was trained by conducting 100 iterations of Bayesian optimisation.[24] At each iteration, the algorithm searched for a new ML pipeline and optimised its hyperparameters. Area under the precision-recall curve (AUC-PRC) was used to evaluate the performance of each pipeline and three weighted ML pipelines were combined to produce the final model. AUC-PRC was chosen as a pipeline evaluation metric because it can be applied to both binary and multi-label classification tasks, effectively addresses the class imbalance in our data set and enables performance comparison independent of classification thresholds. Additionally, AUC-PRC allows for the detection and differentiation of both positive and negative cases, providing a more comprehensive evaluation of model performance.[26]

Model development and training were performed for various data subsets: (1) Clinical data including demographic information, patient-reported outcome measures (PROMs) and simple X-ray features such as KLG, joint space narrowing and medial minimum JSW (detailed in online supplemental table 1); (2) clinical and X-ray data with advanced X-ray features such as fractal bone trabecular integrity; (3) biochemical markers; (4) MRI data; and (5) the entire data set. The whole analysis was performed for both multiclass and binary predictions.

Additionally, streamlined models were built using only five 'core' variables, identified in our post-hoc interpretability analysis as pivotal in influencing model predictions.

Model interpretation

Another benefit of AutoPrognosis V.2.0 is its integration of advanced model interpretability tools that enable the evaluation of variables' contributions to model predictions.

A post-hoc interpretability tool called 'KernelSHAP' was employed to agnostically assess the relative importance of features used to build our models. 'KernelSHAP' uses a weighted linear regression model to compute the importance of each feature.[27] The five most highly ranked attributes were selected as 'core' variables and used for the development of new prediction models.

Validation of model performance

Stratified 10-fold cross-validation with three random seeds was conducted on our training set. The models, optimised for AUC-PRC during the development phase, were evaluated using multiple metrics: AUC-PRC, area under the receiver operating characteristic curve (AUC-ROC), weighted precision (or positive predictive value), weighted recall (also known as sensitivity or true-positive rate) and weighted F1-score (which is the harmonic mean of precision and recall). By 'weighted' we intend the average metric for all labels, weighted by the number of true instances for each label. For simplicity, in the rest of this paper we have omitted the word 'weighted' when discussing these metrics. For each metric, AutoPrognosis V.2.0 also allowed the calculation of 95% CIs.

Further cross-validation was conducted on the hold-out set (representing unseen data excluded from model development and training) and the external data set containing baseline data from the POMA study (figure 1). For this data set, knee OA outcomes were assessed at the 2-year follow-up time point. From the 1170 patients in the POMA study, 183 were also part of the FNIH OA Biomarkers Consortium and were therefore excluded from our validation set. Consequently, the validation cohort consisted of 987 patients encompassing 601 right and 502 left knees (1103 instances in total). Knees lacking sufficient data for outcome class assignment due to missing values were omitted. When data for both knees were available for a patient, only one knee was randomly selected, resulting in a total of 705 patients (383 right, 322 left knees).

Subgroup analysis

Subgroup analyses by age (age<60 vs age≥60), sex and ethnicity were conducted using the hold-out set. Evaluations were then performed on three distinct subgroups within the external data set: Patients under 60 years, patients without initial X-ray signs of OA (KLG 0, a demographic not included in our training set) and patients displaying early-stage OA (KLG 0–1). Online supplemental figure 1 illustrates our subgroup analysis schematically.

Clinical demonstrators

A working prototype of a web-based application or 'clinical demonstrator' was also developed to illustrate the practical application of our tool to predict rapid knee OA progression (although it is not currently intended for use on any individual, including in any clinical or medical setting). This demonstrator was built and deployed using 'Streamlit' (https://streamlit.io/).

RESULTS

Study population

The complete data set included 1691 instances, of which 41% were men and 59% were women, with ages ranging between 45 and 81 (online supplemental table 3). 60.6% of instances were OA non-progressors (Class 0), 7.7% pain-only progressors (Class 1), 25.9% radiographic-only progressors (Class 2) and 5.7% both pain and radiographic progressors (Class 3).

Model performance

Table 1 shows the predictive performance of all our models developed with AutoPrognosis V.2.0 while the final ML pipeline ensembles of each model are illustrated in online supplemental table 4.

The highest performance was achieved when all 304 variables were included (models AP5_mu and AP5_bi in table 1) with AUC-PRC 0.678 (95% CI: 0.676 to 0.680) for multiclass predictions; and AUC-PRC 0.635 (95% CI: 0.629 to 0.641) for binary predictions. The lowest performance was observed in models AP3_mu and AP3_bi, trained solely on biochemical marker data, with AUC-PRC 0.600 (95% CI: 0.597 to 0.603) and 0.523 (95% CI: 0.509 to 0.537), respectively. AUC-PRC and AUC-ROC were higher in multiclass models, whereas F1-score, precision and recall were higher in binary models.

Additionally, models AP5_top5_mu and AP5_top5_bi created using only five 'core' variables (identified by post-hoc interpretability analysis as the strongest contributors to the models' predictions as outlined in figure 2 and online supplemental table 5), achieved performance scores similar to those of the larger models (AUC-PRC 0.648 (95% CI: 0.647 to 0.649) and 0.618 (95% CI: 0.613 to 0.623), respectively).

Notably, models AP1_mu and AP1_bi (including clinical data and simple X-ray features like KLG) also demonstrated robust performance: AP1_mu achieved AUC-PRC 0.648 (95% CI: 0.646 to 0.650); whereas AP1_bi yielded AUC-PRC 0.613 (95% CI: 0.605 to 0.621).

Model interpretation

Figure 2 illustrates the overall impact of features in models AP5_mu and AP5_bi (encompassing all 304 variables) ranked according to their contributions to predictive outcomes. WOMAC pain and disability scores as well as MRI features such as MRI Osteoarthritis Knee Score (MOAKS) and percentage area of subchondral bone denuded of cartilage, emerged as the

Table 1 Models' performance on 10-fold cross-validation. Cross-validation performance of autoML models created using AutoPrognosis V.2.0 for multiclass and binary predictions of rapid knee OA progression

Multiclass predictions							
Model*	**Features**	**N features**	**AUC-PRC (95% CI)**	**AUC-ROC (95% CI)**	**F1 score (95% CI)**	**Precision (95% CI)**	**Recall (95% CI)**
AP1_mu	Clinical	11	0.648 (0.646 to 0.650)	0.851 (0.851 to 0.851)	0.534 (0.533 to 0.535)	0.581 (0.575 to 0.587)	0.630 (0.629 to 0.631)
AP2_mu	Clinical+X-ray	17	0.654 (0.652 to 0.656)	0.851 (0.850 to 0.852)	0.476 (0.474 to 0.478)	0.489 (0.477 to 0.501)	0.612 (0.611 to 0.613)
AP3_mu	Biomarkers	26	0.600 (0.597 to 0.603)	0.820 (0.818 to 0.822)	0.457 (0.457 to 0.457)	0.367 (0.367 to 0.367)	0.606 (0.606 to 0.606)
AP4_mu	MRI	261	0.667 (0.665 to 0.669)	0.851 (0.850 to 0.852)	0.465 (0.464 to 0.466)	0.459 (0.454 to 0.464)	0.606 (0.606 to 0.606)
AP5_mu	All features	304	0.678 (0.676 to 0.680)	0.858 (0.857 to 0.859)	0.512 (0.511 to 0.513)	0.522 (0.517 to 0.527)	0.620 (0.618 to 0.622)
AP5_top5_mu	Top 5†	5	0.648 (0.647 to 0.649)	0.852 (0.852 to 0.852)	0.564 (0.563 to 0.565)	0.580 (0.574 to 0.586)	0.636 (0.635 to 0.637)
Binary predictions							
Model	**Features**	**N features**	**AUC-PRC (95% CI)**	**AUC-ROC (95% CI)**	**F1 score (95% CI)**	**Precision (95% CI)**	**Recall (95% CI)**
AP1_bi	Clinical	11	0.613 (0.605 to 0.621)	0.681 (0.674 to 0.688)	0.637 (0.628 to 0.646)	0.681 (0.673 to 0.689)	0.675 (0.668 to 0.682)
AP2_bi	Clinical+X-ray	17	0.617 (0.614 to 0.620)	0.692 (0.691 to 0.693)	0.644 (0.642 to 0.646)	0.653 (0.651 to 0.655)	0.662 (0.660 to 0.664)
AP3_bi	Biomarkers	26	0.523 (0.509 to 0.537)	0.624 (0.617 to 0.631)	0.597 (0.592 to 0.602)	0.598 (0.591 to 0.605)	0.600 (0.596 to 0.604)
AP4_bi	MRI	261	0.613 (0.604 to 0.622)	0.718 (0.713 to 0.723)	0.636 (0.635 to 0.637)	0.666 (0.664 to 0.668)	0.668 (0.666 to 0.670)
AP5_bi	All features	304	0.635 (0.629 to 0.641)	0.730 (0.723 to 0.737)	0.653 (0.649 to 0.657)	0.682 (0.679 to 0.685)	0.681 (0.677 to 0.685)
AP5_top5_bi	Top 5	5	0.618 (0.613 to 0.623)	0.693 (0.689 to 0.697)	0.660 (0.656 to 0.664)	0.675 (0.670 to 0.680)	0.680 (0.676 to 0.684)

*AP = AutoPrognosis; mu = multi-class; bi = binary.
†"Core" features identified through *post-hoc* interpretability analysis.
AUC-PRC, area under the precision-recall curve; AUC-ROC, area under the receiver operating characteristic curve; OA, osteoarthritis.

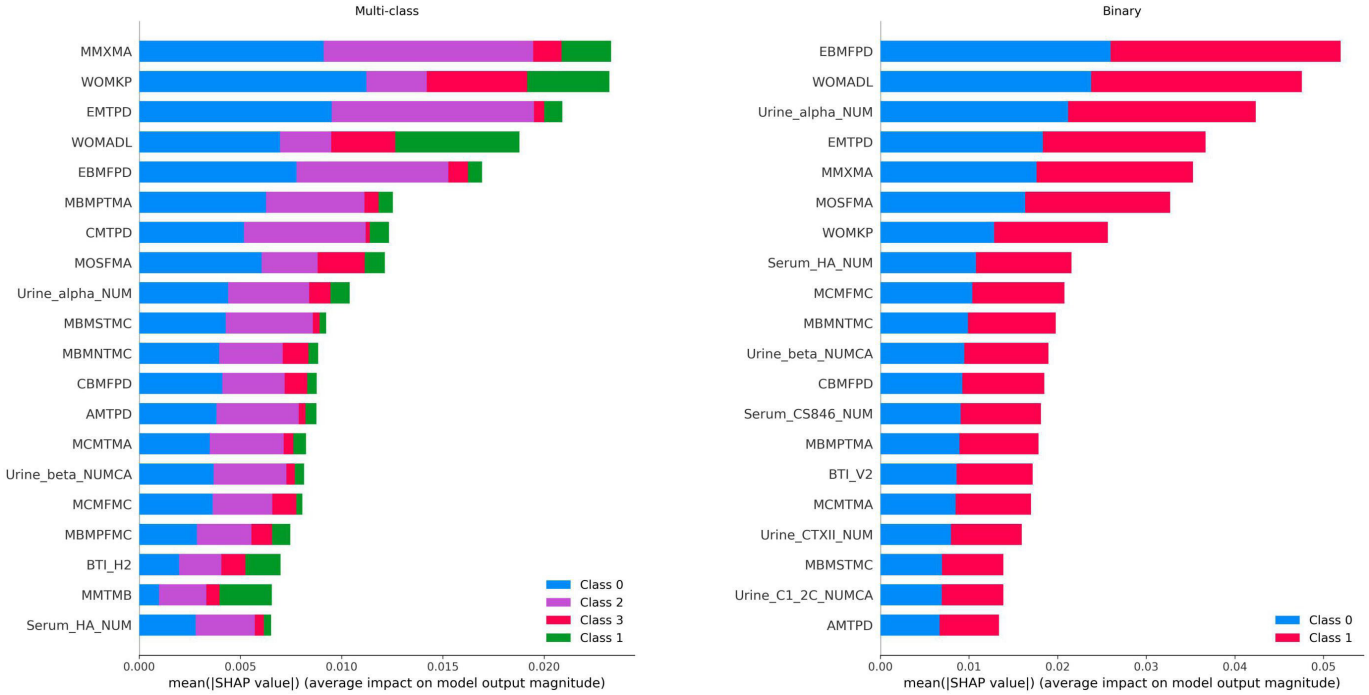


Figure 2 Overall feature importance. This figure illustrates the overall importance of features in models AP5_mu (left) and AP5_bi (right). All 304 variables were included in the analysis. A full description of each feature is outlined in online supplemental table 1.

strongest predictors. Detailed descriptions of the five 'core' variables used to create our streamlined models are presented in online supplemental table 5, while descriptions of all other variables are shown in online supplemental table 1. Feature MOSFMA (a component of MOAKS used to assess the size of osteophytes at the femur's medial anterior trochlear region) was used as a 'core' feature in model AP5_top5_bi in place of biochemical marker Urine_alpha_NUM (urine CTX-1a) as the latter was not available in the external data set.

The impact distribution and average impact magnitude for the most important features across each outcome class in these models are illustrated in figures 3 and 4.

For multiclass predictions, MRI features and WOMAC scores were the most significant contributors across all outcome classes (figure 3). Urine CTX-1a (Urine_alpha_NUM) emerged as the most important biochemical marker significantly affecting the prediction of all classes. Pain-only progression (Class 1) was also influenced by BTI_H2 (an imaging biomarker used to assess the microstructural integrity of trabecular bone in the horizontal plane), the use of medication for knee pain, aching or stiffness (P01KPMEDCV) and age (figure 3B).

Similar results were observed for binary predictions except for a stronger contribution from urine CTX-1a and serum hyaluronic acid (Serum_HA_NUM) (figure 4).

Validation of model performance

Hold-out validation

Models AP1_mu and AP1_bi (only clinical features), AP5_mu and AP5_bi (all available features) and AP5_top5_mu and AP5_top5_bi (five 'core' features) were validated on the hold-out set. All models obtained similar performance scores to those from internal cross-validation, as shown in table 2. Again, multiclass models yielded higher AUC-PRC and AUC-ROC scores while binary models had greater F1-score, precision and recall.

Interestingly, clinical models AP1_mu and AP1_bi, and streamlined models AP5_top5_mu and AP5_top5_bi achieved similar or better performance than the most comprehensive models.

Precision-recall curves (PRCs) and confusion matrices for each model are displayed in online supplemental figure 2 and online supplemental figure 3.

External validation

Due to the absence of several features in the POMA data set (including biochemical markers and complex X-ray features), only clinical models AP1_mu and AP1_bi and streamlined models AP5_top5_mu and AP5_top5_bi were further validated on this external set.

The POMA data set exhibited comparable proportions across outcome classes to our training set; however, it included a much greater fraction of patients with KLG 1 (32.9% vs 11.0%), and, notably, a substantial number of patients with KLG 0 (23.4%), a group absent from our training set (online supplemental table 6).

Highest performance was achieved by models AP1_mu and AP1_bi with AUC-PRC 0.727 (95% CI: 0.726 to 0.728) and 0.764 (95% CI: 0.762 to 0.766), respectively. All external validation results are presented in table 2 while PRCs and confusion matrices for each model are displayed in online supplemental figure 4 and online supplemental figure 5.

Subgroup analysis

Hold-out subgroups

The demographic profiles of the hold-out subpopulations studied are presented in online supplemental table 7. Only White and Black ethnicities were analysed due to the small number of patients belonging to the other groups.

The performance scores achieved by models AP5_mu and AP5_bi (our most comprehensive models) for each subgroup are presented in table 3. Model AP5_mu demonstrated improved performance in patients aged≥60 with AUC-PRC 0.685

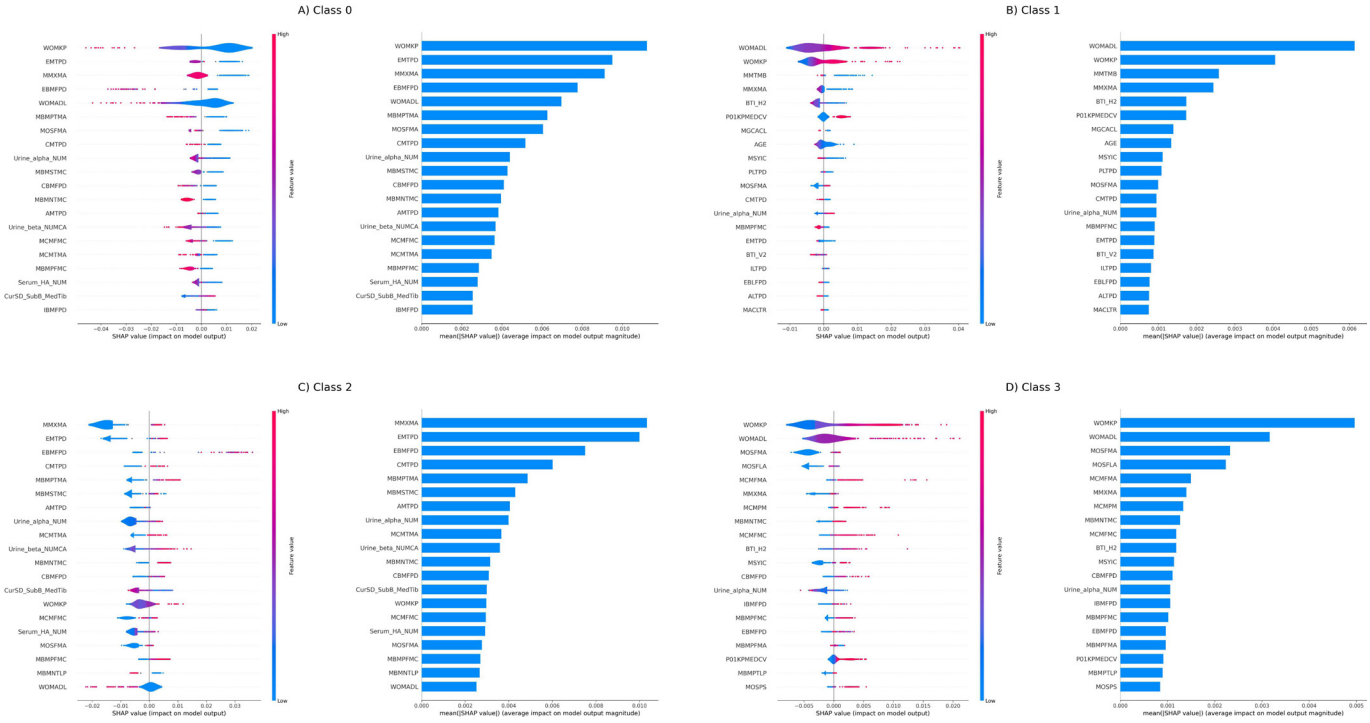


Figure 3 Assessment of feature impact on multiclass predictions. Assessment of feature impact on non-progression (Class 0, panel A), pain-only progression (Class 1, panel B), radiographic progression (Class 2, panel C) and both pain and radiographic progression (Class 3, panel D), using 'Kernel-SHAP' for multiclass predictions with model AP5_mu. Left—impact distribution of the most important features. The colour represents the feature value (red=high, blue=low). A positive SHAP value represents a positive impact on class prediction. Right—average impact magnitude of the most important features on class prediction.

compared with 0.644 in those aged<60. Sex differences were also evident with female patients achieving an AUC-PRC of 0.707, significantly higher than the 0.619 observed in male patients. Ethnicity-related performance varied with patients of White ethnicity showing an AUC-PRC of 0.706, markedly higher than the 0.558 observed in Black and African-American patients. Similarly, model AP5_bi showed higher AUC-PRC scores for female patients (0.702) compared with men (0.522) and better performance in younger patients (AUC-PRC 0.676 for age<60) compared with older ones (AUC-PRC 0.564 for age≥60). Ethnicity disparities persisted in binary models with White patients achieving an AUC-PRC of 0.563 versus 0.632 for Black patients.

The results of our post-hoc interpretability analyses of each subgroup are illustrated in figure 5. For multiclass predictions, WOMAC pain and disability scores were particularly significant for all subgroups, especially for young, women and Black patients. MRI features, including MOAKS, cartilage thickness and the percentage area of subchondral bone denuded of cartilage also consistently ranked highly across all subgroups. Urine CTX-1a emerged once again as the most important biochemical marker, especially for patients of Black ethnicity.

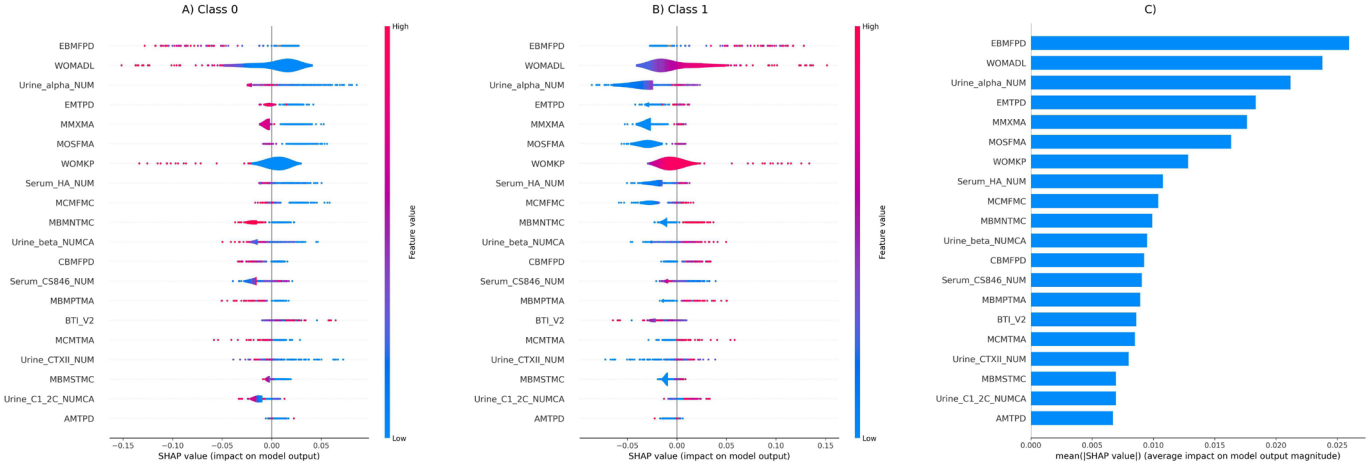


Figure 4 Assessment of feature impact on binary predictions. Assessment of feature impact on non-progression (Class 0) and progression (Class 1), using "Kernel-SHAP" for binary predictions with model AP5_bi. Left and middle—impact distribution of the most important features for Class 0 (left) and Class 1 (middle). The colour represents the feature value (red=high, blue=low). A positive SHAP value represents a positive impact on class prediction. Right—average impact magnitude of the most important features on class prediction.

Table 2 Validation of models' performance. Hold-out and external validation performance scores for multiclass and binary predictions. External validation was conducted on patients from the POMA study

Multiclass predictions							
Validation	**Model**	**N features**	**AUC-PRC (95% CI)**	**AUC-ROC (95% CI)**	**F1 score (95% CI)**	**Precision (95% CI)**	**Recall (95% CI)**
Hold-out	AP1_mu	11	0.649 (0.646 to 0.652)	0.847 (0.847 to 0.847)	0.529 (0.528 to 0.530)	0.559 (0.532 to 0.586)	0.633 (0.633 to 0.633)
	AP5_mu	304	0.665 (0.661 to 0.669)	0.842 (0.841 to 0.843)	0.488 (0.485 to 0.491)	0.500 (0.469 to 0.531)	0.614 (0.612 to 0.616)
	AP5_top5_mu	5	0.650 (0.647 to 0.653)	0.846 (0.845 to 0.847)	0.552 (0.549 to 0.555)	0.558 (0.537 to 0.579)	0.629 (0.627 to 0.631)
External	AP1_mu	11	0.727 (0.726 to 0.728)	0.881 (0.881 to 0.881)	0.593 (0.592 to 0.594)	0.637 (0.631 to 0.643)	0.670 (0.670 to 0.670)
	AP5_top5_mu	5	0.589 (0.583 to 0.595)	0.838 (0.836 to 0.840)	0.563 (0.555 to 0.571)	0.566 (0.552 to 0.580)	0.596 (0.585 to 0.607)
Binary predictions							
Validation	**Model**	**N features**	**AUC-PRC (95% CI)**	**AUC-ROC (95% CI)**	**F1 score (95% CI)**	**Precision (95% CI)**	**Recall (95% CI)**
Hold-out	AP1_bi	11	0.637 (0.636 to 0.638)	0.700 (0.699 to 0.701)	0.631 (0.630 to 0.632)	0.669 (0.666 to 0.672)	0.666 (0.666 to 0.666)
	AP5_bi	304	0.598 (0.590 to 0.606)	0.675 (0.674 to 0.676)	0.622 (0.619 to 0.625)	0.636 (0.633 to 0.639)	0.648 (0.645 to 0.651)
	AP5_top5_bi	5	0.618 (0.613 to 0.623)	0.693 (0.689 to 0.697)	0.660 (0.656 to 0.664)	0.675 (0.670 to 0.680)	0.680 (0.676 to 0.684)
External	AP1_bi	11	0.764 (0.762 to 0.766)	0.780 (0.780 to 0.780)	0.714 (0.713 to 0.715)	0.733 (0.732 to 0.734)	0.726 (0.726 to 0.726)
	AP5_top5_bi	5	0.688 (0.683 to 0.693)	0.702 (0.691 to 0.713)	0.653 (0.615 to 0.691)	0.669 (0.641 to 0.697)	0.663 (0.626 to 0.700)

AP, AutoPrognosis; AUC-PRC, area under the precision-recall curve; AUC-ROC, area under the receiver operating characteristic curve; POMA, Pivotal Osteoarthritis Initiative MRI Analyses.

For binary predictions, WOMAC disability score and MRI features remained important predictors across all subgroups. Urine CTX-1a also demonstrated a very strong contribution while serum hyaluronic acid emerged as an additional important predictor, especially in young patients. WOMAC pain, on the other hand, was significantly less influential in binary models compared with multiclass models.

Online supplemental figures 6–17 illustrate the impact distribution and average impact magnitude of the most important features across each outcome class for all subgroups.

Table 3 Models' performance in subgroup analysis. Models' performance in various subgroups from the hold-out and external validation sets

Multiclass predictions							
Validation	**Model**	**Subgroup**	**AUC-PRC (95% CI)**	**AUC-ROC (95% CI)**	**F1 score (95% CI)**	**Precision (95% CI)**	**Recall (95% CI)**
Hold-out	AP5_mu	Age<60	0.644 (0.642 to 0.646)	0.825 (0.824 to 0.826)	0.448 (0.448 to 0.448)	0.414 (0.410 to 0.418)	0.593 (0.593 to 0.593)
		Age≥60	0.685 (0.683 to 0.687)	0.852 (0.852 to 0.852)	0.511 (0.506 to 0.516)	0.503 (0.470 to 0.536)	0.627 (0.624 to 0.630)
		Male	0.619 (0.616 to 0.622)	0.845 (0.844 to 0.846)	0.462 (0.460 to 0.464)	0.495 (0.457 to 0.533)	0.584 (0.584 to 0.584)
		Female	0.707 (0.704 to 0.710)	0.833 (0.833 to 0.833)	0.515 (0.514 to 0.516)	0.497 (0.475 to 0.519)	0.641 (0.641 to 0.641)
		White ethnicity	0.706 (0.704 to 0.708)	0.874 (0.874 to 0.874)	0.518 (0.513 to 0.523)	0.526 (0.496 to 0.556)	0.637 (0.635 to 0.639)
		Black ethnicity	0.558 (0.552 to 0.564)	0.710 (0.709 to 0.711)	0.387 (0.387 to 0.387)	0.299 (0.299 to 0.299)	0.547 (0.547 to 0.547)
External	AP1_mu	Age<60	0.751 (0.749 to 0.753)	0.896 (0.896 to 0.896)	0.604 (0.603 to 0.605)	0.655 (0.651 to 0.659)	0.691 (0.691 to 0.691)
		KLG 0–1	0.797 (0.795 to 0.799)	0.924 (0.924 to 0.924)	0.644 (0.644 to 0.644)	0.564 (0.564 to 0.564)	0.749 (0.749 to 0.749)
		KLG 0	0.806 (0.805 to 0.807)	0.928 (0.928 to 0.928)	0.637 (0.637 to 0.637)	0.557 (0.557 to 0.557)	0.744 (0.744 to 0.744)
	AP5_top5_mu	Age<60	0.633 (0.611 to 0.655)	0.854 (0.844 to 0.864)	0.543 (0.511 to 0.575)	0.539 (0.523 to 0.555)	0.579 (0.550 to 0.608)
		KLG 0–1	0.731 (0.729 to 0.733)	0.908 (0.908 to 0.908)	0.652 (0.652 to 0.652)	0.583 (0.582 to 0.584)	0.754 (0.754 to 0.754)
		KLG 0	0.724 (0.721 to 0.727)	0.907 (0.907 to 0.907)	0.644 (0.643 to 0.645)	0.565 (0.562 to 0.568)	0.750 (0.750 to 0.750)
Binary predictions							
Validation	**Model**	**Subgroup**	**AUC-PRC (95% CI)**	**AUC-ROC (95% CI)**	**F1 score (95% CI)**	**Precision (95% CI)**	**Recall (95% CI)**
Hold-out	AP5_bi	Age<60	0.676 (0.666 to 0.686)	0.690 (0.681 to 0.699)	0.625 (0.619 to 0.631)	0.658 (0.651 to 0.665)	0.654 (0.650 to 0.658)
		Age≥60	0.564 (0.559 to 0.569)	0.666 (0.661 to 0.671)	0.619 (0.617 to 0.621)	0.629 (0.625 to 0.633)	0.645 (0.645 to 0.645)
		Male	0.522 (0.518 to 0.526)	0.575 (0.571 to 0.579)	0.571 (0.570 to 0.572)	0.575 (0.573 to 0.577)	0.591 (0.591 to 0.591)
		Female	0.702 (0.691 to 0.713)	0.759 (0.756 to 0.762)	0.667 (0.662 to 0.672)	0.711 (0.703 to 0.719)	0.699 (0.696 to 0.702)
		White ethnicity	0.563 (0.558 to 0.568)	0.674 (0.671 to 0.677)	0.638 (0.636 to 0.640)	0.648 (0.642 to 0.654)	0.672 (0.672 to 0.672)
		Black ethnicity	0.632 (0.629 to 0.635)	0.610 (0.601 to 0.619)	0.514 (0.505 to 0.523)	0.535 (0.527 to 0.543)	0.542 (0.535 to 0.549)
External	AP1_bi	Age<60	0.693 (0.689 to 0.697)	0.759 (0.757 to 0.761)	0.709 (0.707 to 0.711)	0.725 (0.724 to 0.726)	0.729 (0.729 to 0.729)
		KLG 0–1	0.415 (0.413 to 0.417)	0.665 (0.664 to 0.666)	0.713 (0.713 to 0.713)	0.707 (0.705 to 0.709)	0.742 (0.742 to 0.742)
		KLG 0	0.468 (0.458 to 0.478)	0.665 (0.663 to 0.667)	0.712 (0.711 to 0.713)	0.705 (0.698 to 0.712)	0.738 (0.738 to 0.738)
	AP5_top5_bi	Age<60	0.617 (0.606 to 0.628)	0.693 (0.686 to 0.700)	0.669 (0.660 to 0.678)	0.678 (0.664 to 0.692)	0.686 (0.671 to 0.701)
		KLG 0–1	0.379 (0.377 to 0.381)	0.581 (0.577 to 0.585)	0.675 (0.675 to 0.675)	0.741 (0.727 to 0.755)	0.757 (0.757 to 0.757)
		KLG 0	0.385 (0.376 to 0.394)	0.575 (0.570 to 0.580)	0.670 (0.669 to 0.671)	0.682 (0.655 to 0.709)	0.756 (0.756 to 0.756)

AP, AutoPrognosis; AUC-PRC, area under the precision-recall curve; AUC-ROC, area under the receiver operating characteristic curve; KLG, Kellgren and Lawrence grade .

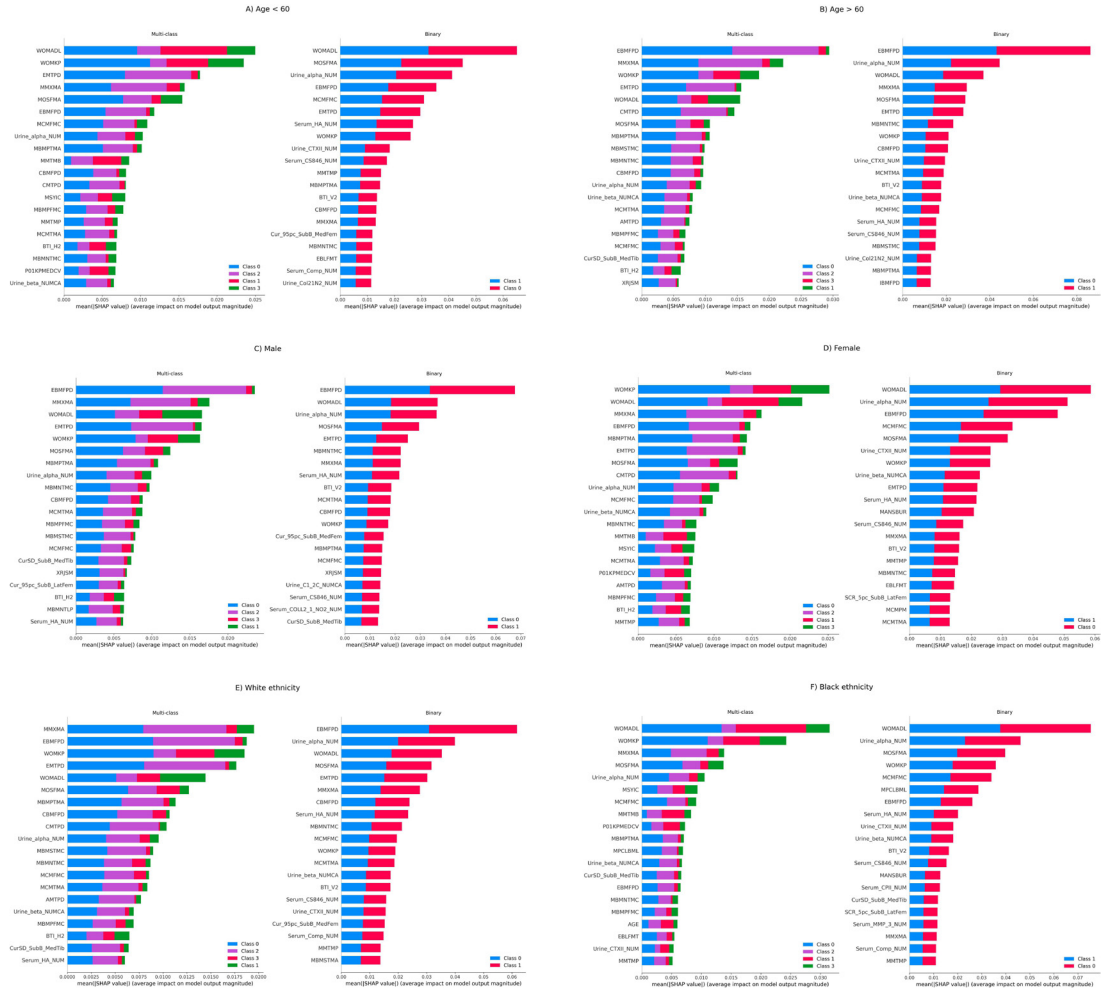


Figure 5 Overall feature importance in the hold-out subgroup analysis. This figure illustrates the overall importance of features in models AP5_mu (left) and AP5_bi (right) for the following subgroups in the hold-out data set: (A) Age<60; (B) Age≥60; (C) Male; (D) Female; (E) White ethnicity; and (F) Black ethnicity.

External subgroups

Online supplemental file 8 shows the demographic characteristics of the subpopulations in the external validation set. Notably, the young cohort exhibited significantly higher proportions of knees classified as KLG 0 or 1 (27.8% and 41.3%, respectively), in comparison to our training data set (0% and 11.0%). Additionally, subgroups with early-stage OA (KLG 0–1) and no initial radiographic signs of OA (KLG 0) demonstrated substantially greater rates of non-progression (74.9% and 74.4%) than observed in our training set (60.6%).

Performances of models AP1_mu, AP1_bi, AP5_top5_mu and AP5_top5_bi on these subgroups are presented in table 3. Both multiclass models achieved high predictive performance, particularly in the KLG 0–1 and KLG 0 subgroups (AUC-PRC 0.724–0.806).

In contrast, binary models exhibited comparatively lower AUC-PRC and AUC-ROC scores, but higher F1-score, precision and recall. Although performance remained robust in the young subgroup (AUC-PRC 0.617–0.693), both binary models showed mixed results in the KLG 0–1 and KLG 0 cohorts with low AUC-PRC and AUC-ROC values, but high F1-score, precision and recall.

Clinical demonstrators

Clinical demonstrators were built using our clinical models and can be accessed via these links:

- https://autop-knee-oa-multiclass.streamlit.app/?embed_options=light_theme for multiclass predictions;
- https://autop-knee-oa-binary.streamlit.app/?embed_options=light_theme for binary predictions.

WOMAC pain and disability scores were not included as variables in these prototypes to prevent any possible copyright infringement.

These clinical demonstrators allow intuitive and streamlined visualisation of our models' predictions for individual patients along with the relative impact of each feature on these personalised predictions, as elucidated by Kernel SHAP.

DISCUSSION

We developed autoML models to predict rapid knee OA progression over 2 years. Our most reliable models incorporated clinical, X-ray, MRI and biochemical features resulting in an 'information gain' compared with models using only a subset of these data. Additionally, AutoPrognosis V.2.0 introduced a 'modelling gain', by selecting the most suitable algorithms in a fully data-driven manner, without prior assumptions. In light of this 'modelling gain', model performance was not significantly affected when only 'core' variables were used. This is important as it facilitates the translation of our models to clinical practice where it may not be feasible, nor logical, to measure over 300 variables for each patient.

Despite several studies previously attempting to predict knee OA progression using baseline biomarker data (with or without ML),

direct comparison with our research is impossible, primarily due to inconsistent definitions of OA progression and validation methods. For instance, Hunter *et al*[19] employed logistic regression models with imaging and biochemical markers achieving AUC-ROC 0.716–0.732 for radiographic progression and AUC-ROC 0.668–0.694 for both pain and radiographic progression over 4 years. Widera *et al*,[23] in contrast, constructed random forest models to predict progression over 2 years, using similar class definitions to ours but relying solely on clinical and X-ray data, resulting in F1-scores of 0.560–0.698.

Unlike earlier research, we took a completely data-driven approach to model development by employing AutoPrognosis V.2.0. We incorporated a wider variety of data types, including clinical data, PROMs, X-rays, MRIs and biochemical markers which enriched our predictive models. We also made significant efforts to enhance the transparency of our models through post-hoc interpretability analysis and the development of clinical demonstrators.

We agnostically identified key predictors of rapid knee OA progression, particularly PROMs like WOMAC pain and disability scores, MRI features such as MOAKS and area of subchondral bone denuded of cartilage and biochemical markers such as urine CTX-1a. We believe this transparency will help build trust among clinicians and patients, potentially accelerating healthcare adoption. Furthermore, our analysis highlights the importance of PROMs in prognostic modelling of a complex condition like knee OA, reflecting a critical step towards the humanisation of AI in healthcare.[28] By incorporating PROMs, our tool assimilates the patients' own perceptions of their symptoms, empowering collaborative, informed healthcare decisions.

To the best of our knowledge, this study is the first to apply these predictive models and assess feature importance in multiple OA patient subgroups, including patients under 60 who constitute a significant proportion of the knee OA population and may particularly benefit from early intervention.[10 11 13 29] Subgroup analysis is essential to identify and address potential biases ensuring the models' accuracy and applicability across diverse populations.[30 31]

A critical component of the study was the thorough validation of our models using multiple performance metrics alongside techniques such as stratified 10-fold cross-validation, hold-out validation and external validation with the POMA cohort, representing a separate, nested study within the broader OAI framework. Although this methodology confirmed our models' reliability, future research in diverse clinical settings and new cohorts is essential to assess their clinical utility and generalisability across diverse patient demographics.[32]

Even though our training cohort included only patients with radiographic evidence of knee OA (KLG 1–4), our models demonstrated robust performance when validated on the POMA data set which has a high proportion of patients with KLG 0–1. Interestingly, models using only clinical variables showed the strongest external validation performance (despite missing features in the external data set preventing validation of the most comprehensive models). Relying on clinical features is advantageous in clinical practice as they are inexpensive and easily collected. This is particularly relevant in resource-constrained environments where comprehensive data collection might be challenging.

Our multiclass models demonstrated high predictive performance in younger patients and those with early-stage OA, offering the dual advantage of reliability in high-risk groups and patient phenotyping based on progression type. This aligns with our aim to predict early disease progression, providing a potential 'window of opportunity' for interventions (ranging from lifestyle modifications and rehabilitation, to reparative and regenerative therapies) to arrest or slow down disease progression.[33] In contrast, our binary models, while performing well on the entire POMA study cohort, showed mixed performance across metrics when applied to early-OA subgroups. This underscores the need to refine these models by incorporating data specifically from patients in the early stages of OA.

Our study has other limitations that should be addressed in future work. The use of data sets from the same overall study (OAI) for both training and validation may restrict generalisability despite employing cross-validation techniques and conducting validation on multiple data sets and subgroups. Future research should validate these models on completely independent data sets from diverse geographic and demographic backgrounds to ensure broader applicability. Additionally, although WOMAC scores are commonly used in research, their copyright protection may limit their use in clinical practice. Finally, when validating our models, confusion matrices revealed that classes with the smallest sample sizes were less accurately predicted, especially in the multiclass models. However, the accuracy of these minority classes can be significantly improved by adjusting the probability threshold during class assignment (as demonstrated by the PRCs) and the same models achieved high AUC-PRC and AUC-ROC indicating strong overall performance independent of classification thresholds.

While these limitations highlight the need for model refinement and further training prior to clinical implementation, this study demonstrates the significant potential of a fully data-driven autoML approach and the utility of biomarker identification and subgroup analysis in predicting knee OA progression.

We believe our approach is not only applicable to OA but could also serve as a model for other complex degenerative conditions (such as multiple sclerosis and Parkinson's disease) which share common challenges, including chronicity, unmet clinical needs and difficulties in early diagnosis.[34 35] By tailoring data inputs and fine-tuning models to these diseases, our method holds significant potential for the prediction and monitoring of such conditions. This ML application represents a step towards a more tailored and precise approach to healthcare, addressing on the one hand the personalised needs of the individual patient while on the other delivering impact on a societal scale.

Acknowledgements We extend our gratitude to the participants of the Osteoarthritis Initiative for their invaluable contributions to this research. Their willingness to share data and experiences has been instrumental in advancing our understanding of osteoarthritis. Additionally, we acknowledge the dedicated members of the patient and public involvement panel at Addenbrooke's Hospital, Cambridge, UK, for their insights and guidance which have greatly enriched the scope and relevance of our study. A previous version of our work was presented at the 2023 European Orthopaedic Research Society and British Orthopaedic Research Society conferences.

Contributors All authors contributed to the conceptualisation and design of the study. SC contributed to the curation and analysis of the data. MB, MvdS and AM supervised the study. All authors contributed to the interpretation of the data, the drafting of the article and final approval of the version to be submitted. AM is the guarantor of the study. ChatGPT, an AI language model developed by OpenAI, was used exclusively to assist in improving the clarity and legibility of few sentences in the initial drafting of the manuscript, though these sections have been substantially revised by the authors to generate the final version. It did not contribute to the creation of content or the analysis of data.

Funding SC is supported by the Louis and Valerie Freedman Studentship in Medical Sciences from Trinity College Cambridge, the ORUK/Versus Arthritis: AI in MSK Research Fellowship (G124606) and the Addenbrooke's Charitable Trust (ACT) Research Advisory Committee grant (G123290). At the start of the study, SC was also supported by the National Institute for Health and Care Research (NIHR) (ACF-2021-14-003). AM and MB are supported by the NIHR Cambridge Biomedical Research Centre (NIHR203312) and receive funding from Versus Arthritis (grant 21156). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. The funders of the study were not involved in the design, data collection, analysis, interpretation or writing of this study.

Competing interests None declared. We confirm that we have read the journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

Patient and public involvement Patients and the public were involved early in our research, contributing to the development of our research questions and outcome measures. Their input, gathered through a focus group with the Patient and Public Involvement team at Addenbrooke's Hospital, Cambridge, UK, informed the design of our study and our clinical demonstrators. While direct involvement in recruitment and study conduct was not applicable due to the nature of our data, their perspectives on the usability and implications of our research were integral. Our dissemination strategy includes regular interactions with this group, collaborations with patient groups and relevant charities (such as Osteoarthritis Research UK (ORUK) and Versus Arthritis), and public-friendly summaries of our findings to ensure ongoing, reciprocal communication and feedback.

Patient consent for publication Not applicable.

Ethics approval This study was performed retrospectively using data from human subjects which was openly accessible through the Osteoarthritis Initiative (OAI) (https://nda.nih.gov/oai). Since the OAI had already secured ethical approval and obtained informed consent from the participants and the data was released under an open access permission group, there was no need for additional ethical approval for our study. All individuals had provided informed consent prior to their inclusion in the OAI study.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. Data and/or research tools used in the preparation of this manuscript were obtained and analysed from the controlled access data sets distributed from the Osteoarthritis Initiative (OAI), a data repository housed within the National Institute of Mental Health (NIMH) Data Archive. OAI is a collaborative informatics system created by the NIMH and the National Institute of Arthritis, Musculoskeletal and Skin Diseases to provide a worldwide resource to quicken the pace of biomarker identification, scientific investigation and OA drug development. (DOI: 10.15154/1vhq-h028).Data provided from the FNIH OA Biomarkers Consortium Project (available at https://nda.nih.gov/oai/) made possible through grants and direct or in-kind contributions by: AbbVie; Amgen; Arthritis Foundation; Artialis; Bioiberica; BioVendor; DePuy; Flexion Therapeutics; GSK; IBEX; IDS; Merck Serono; Quidel; Rottapharm | Madaus; Sanofi; Stryker; the Pivotal OAI MRI Analyses study, NIH HHSN2682010000 21C; and the Osteoarthritis Research Society International. The OAI is a public-private partnership comprised of five contracts (N01-AR-2-2258; N01-AR-2-2259; N01AR-2-2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health. Funding partners include Merck Research Laboratories; Novartis Pharmaceuticals, GlaxoSmithKline; and Pfizer. Private sector funding for the consortium and OAI is managed by the Foundation for the National Institutes of Health. Code availability. The AutoPrognosis V.2.0 open-source software package is available at https://www.autoprognosis.vanderschaar-lab.com/.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD
Simone Castagno http://orcid.org/0000-0002-3411-5880

REFERENCES

1 Glyn-Jones S, Palmer AJR, Agricola R, *et al*. Osteoarthritis. *Lancet* 2015;386:376–87.
2 Sharma L. Osteoarthritis of the Knee. *N Engl J Med* 2021;384:51–9.
3 Hunter DJ, March L, Chew M. Osteoarthritis in 2020 and beyond: a Lancet Commission. *Lancet* 2020;396:1711–2.
4 Hunter DJ, Bierma-Zeinstra S. Osteoarthritis. *Lancet* 2019;393:1745–59.
5 Deveza LA, Nelson AE, Loeser RF. Phenotypes of osteoarthritis: current state and future implications. *Clin Exp Rheumatol* 2019;37 Suppl 120:64–72.
6 Deveza LA, Loeser RF. Is osteoarthritis one disease or a collection of many? *Rheumatol (Oxford)* 2018;57:iv34–42.
7 Ghouri A, Conaghan PG. Prospects for Therapies in Osteoarthritis. *Calcif Tissue Int* 2021;109:339–50.
8 Bijlsma JW, Berenbaum F, Lafeber FP. Osteoarthritis: an update with relevance for clinical practice. *Lancet* 2011;377:2115–26.
9 National Institute for Health and Care Excellence (NICE). Osteoarthritis in over 16s: diagnosis and management (NG226). 2022. Available: https://www.nice.org.uk/guidance/ng226
10 Losina E, Weinstein AM, Reichmann WM, *et al*. Lifetime risk and age at diagnosis of symptomatic knee osteoarthritis in the US. *Arthritis Care Res (Hoboken)* 2013;65:703–11.
11 Ackerman IN, Kemp JL, Crossley KM, *et al*. Hip and Knee Osteoarthritis Affects Younger People, Too. *J Orthop Sports Phys Ther* 2017;47:67–79.
12 Felson D, Niu J, Sack B, *et al*. Progression of osteoarthritis as a state of inertia. *Ann Rheum Dis* 2013;72:924–9.
13 Khan M, Adili A, Winemaker M, *et al*. Management of osteoarthritis of the knee in younger patients. *CMAJ* 2018;190:E72–9.
14 Bhowmik RT, Jung YS, Aguilera JA, *et al*. A multi-modal wildfire prediction and early-warning system based on a novel machine learning framework. *J Environ Manage* 2023;341:117908.
15 Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput Sci* 2021;2:160.
16 Michael JWP, Schlüter-Brust KU, Eysel P. The epidemiology, etiology, diagnosis, and treatment of osteoarthritis of the knee. *Dtsch Arztebl Int* 2010;107:152–62.
17 Wirth W, Hunter DJ, Nevitt MC, *et al*. Predictive and concurrent validity of cartilage thickness change as a marker of knee osteoarthritis progression: data from the Osteoarthritis Initiative. *Osteoarthr Cartil* 2017;25:2063–71.
18 Hunter DJ, Nevitt M, Losina E, *et al*. Biomarkers for osteoarthritis: current position and steps towards further validation. *Best Pract Res Clin Rheumatol* 2014;28:61–71.
19 Hunter DJ, Deveza LA, Collins JE, *et al*. Multivariable Modeling of Biomarker Data From the Phase I Foundation for the National Institutes of Health Osteoarthritis Biomarkers Consortium. *Arthritis Care Res (Hoboken)* 2022;74:1142–53.
20 Roemer FW, Guermazi A, Hannon MJ, *et al*. Presence of Magnetic Resonance Imaging-Defined Inflammation Particularly in Overweight and Obese Women Increases Risk of Radiographic Knee Osteoarthritis: The POMA Study. *Arthritis Care Res (Hoboken)* 2022;74:1391–8.
21 Roemer FW, Kwoh CK, Hannon MJ, *et al*. What comes first? Multitissue involvement leading to radiographic osteoarthritis: magnetic resonance imaging-based trajectory analysis over four years in the osteoarthritis initiative. *Arthritis Rheumatol* 2015;67:2085–96.
22 Roemer FW, Kwoh CK, Hannon MJ, *et al*. Can structural joint damage measured with MR imaging be used to predict knee replacement in the following year? *Radiology* 2015;274:810–20.
23 Widera P, Welsing PMJ, Ladel C, *et al*. Multi-classifier prediction of knee osteoarthritis progression from incomplete imbalanced longitudinal data. *Sci Rep* 2020;10:8427.
24 Alaa AM, van der Schaar M. 10. AutoPrognosis: automated clinical prognostic modeling via bayesian optimization with structured kernel learning.
25 Alaa AM, Bolton T, Di Angelantonio E, *et al*. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. *PLoS One* 2019;14:e0213653.
26 Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;10:e0118432.
27 Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. Curran Associates, Inc; 2017 Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
28 Cruz Rivera S, Liu X, Hughes SE, *et al*. Embedding patient-reported outcomes at the heart of artificial intelligence health-care technologies. *Lancet Dig Health* 2023;5:e168–73.
29 Losina E, Klara K, Michl GL, *et al*. Development and feasibility of a personalized, interactive risk calculator for knee osteoarthritis. *BMC Musculoskelet Disord* 2015;16:312.
30 Kapur S. Reducing racial bias in AI models for clinical use requires a top-down intervention. *Nat Mach Intell* 2021;3:460.
31 Chan SCC, Neves AL, Majeed A, *et al*. Bridging the equity gap towards inclusive artificial intelligence in healthcare diagnostics. *BMJ* 2024;384:q490.
32 Youssef A, Pencina M, Thakur A, *et al*. External validation of AI models in health should be replaced with recurring local validation. *N Med* 2023;29:2686–7.
33 Mahmoudian A, Lohmander LS, Mobasheri A, *et al*. Early-stage symptomatic osteoarthritis of the knee - time for action. *Nat Rev Rheumatol* 2021;17:621–32.
34 Solomon AJ, Ascherio A. Early Diagnosis of Multiple Sclerosis: Further Evidence for Missed Opportunity. *Neurol (ECronicon)* 2021;96:1111–2.
35 Soman K, Nelson CA, Cerono G, *et al*. Early detection of Parkinson's disease through enriching the electronic health record using a biomedical knowledge graph. *Front Med (Lausanne)* 2023;10:1081087.

Supplemental Table 1: Features overview

This table outlines each feature, accompanied by its respective description, across the Clinical, Biochemical, X-ray, and MRI datasets as part of the FNIH OA Biomarker Consortium.

Feature	Description
Clinical Data	
P01KPMEDCV	Either knee, used med for pain, aching or stiffness more than half the days of a month, past 12mo (calc, used for study eligibility)
P01BMI	Body mass index (calc)
P02SEX	Gender, male or female
P02RACE	Racial background, self-reported (calc)
XRJSM	(BU): joint space narrowing (OARSI grades 0-3) medial compartment
XRKL	(BU): Kellgren and Lawrence (grades 0-4)
XRJSL	(BU): joint space narrowing (OARSI grades 0-3) lateral compartment
MCMJSW	reading (JD): medial minimum JSW [mm]
WOMKP	WOMAC Pain Score
WOMADL	WOMAC Disability Score
AGE	Age
Xray data	
BTI_H0	Fractal Bone Trabecular Integrity Horizontal - constant
BTI_H1	Fractal Bone Trabecular Integrity Horizontal - slope
BTI_H2	Fractal Bone Trabecular Integrity Horizontal - quadratic
BTI_V0	Fractal Bone Trabecular Integrity Vertical - constant

BTI_V1	Fractal Bone Trabecular Integrity Vertical - slope
BTI_V2	Fractal Bone Trabecular Integrity Vertical - quadratic
Biochemical Data	
Serum_C1_2C_NUM	Serum C1,2C (numeric, interpolated research value if below lower limit)
Serum_C2C_NUM	Serum C2C (numeric, interpolated research value if below lower limit)
Serum_CPII_NUM	Serum CPII (numeric, interpolated research value if below lower limit)
Serum_PIIANP_NUM	Serum PIIANP (numeric, interpolated research value if below lower limit)
Serum_COLL2_1_NO2_NUM	Serum COLL2-1 NO2 (numeric, interpolated research value if below lower limit)
Serum_CS846_NUM	Serum CS846 (numeric, interpolated research value if below lower limit)
Serum_CTXI_NUM	Serum CTXI (numeric, interpolated research value if below lower limit)
Serum_Comp_NUM	Serum cartilage oligomeric matrix protein (COMP) (numeric, interpolated research value if below lower limit)
Serum_HA_NUM	Serum HA (numeric, interpolated research value if below lower limit)
Serum_MMP_3_NUM	Serum MMP-3 (numeric, interpolated research value if below lower limit)
Serum_NTXI_NUM	Serum NTXI (numeric, interpolated research value if below lower limit)
Urine_CTXII_NUM	Urinary CTXII (numeric, interpolated research value if below lower limit)
Urine_C1_2C_NUM	Urine C1,2C (numeric, interpolated research value if below lower limit)
Urine_C2C_NUM	Urine C2C (numeric, interpolated research value if below lower limit)
Urine_Creatinine_NUM	Urine Creatinine (numeric, interpolated research value if below lower limit)
Urine_NTXI_NUM	Urine NTXI (numeric, interpolated research value if below lower limit)
Urine_alpha_NUM	Urine CTX-1a (Urine_alpha) (numeric, interpolated research value if below lower limit)
Urine_beta_NUM	Urine CTX-1b (Urine_beta) (numeric, interpolated research value if below lower limit)

Urine_Col21N2_NUM	Urine Coll21NO2 (numeric, interpolated research value if below lower limit)
Urine_CTXII_NUMCA	Urinary CTXII creatinine adjusted (numeric, interpolated research value if below lower limit)
Urine_C1_2C_NUMCA	Urine C1,2C creatinine adjusted (numeric, interpolated research value if below lower limit)
Urine_C2C_NUMCA	Urine C2C creatinine adjusted (numeric, interpolated research value if below lower limit)
Urine_NTXI_NUMCA	Urine NTXI creatinine adjusted (numeric, interpolated research value if below lower limit)
Urine_alpha_NUMCA	Urine CTX-1a (Urine_alpha) creatinine adjusted (numeric, interpolated research value if below lower limit)
Urine_beta_NUMCA	Urine CTX-1b (Urine_beta) creatinine adjusted (numeric, interpolated research value if below lower limit)
Urine_Col21N2_NUMCA	Urine Coll21NO2 creatinine adjusted (numeric, interpolated research value if below lower limit)
MRI data	
MF_tAB	(iMorphics): total area of subchondral bone - femur medial (MF.tAB) [mm^2]
LF_tAB	(iMorphics): total area of subchondral bone - femur lateral (LF.tAB) [mm^2]
MT_tAB	(iMorphics): total area of subchondral bone - tibia medial (MT.tAB) [mm^2]
LT_tAB	(iMorphics): total area of subchondral bone - tibia lateral (LT.tAB) [mm^2]
MP_tAB	(iMorphics): total area of subchondral bone - patella medial (mP.tAB) [mm^2]
LP_tAB	(iMorphics): total area of subchondral bone - patella lateral (lP.tAB) [mm^2]
notch	(iMorphics): total area of subchondral bone - femoral notch (cTrF.tAB) [mm^2]
TrFLat	(iMorphics): total area of subchondral bone - femoral trochlea lateral (lTrF.tAB) [mm^2]
TrFMed	(iMorphics): total area of subchondral bone - femoral trochlea medial (mTrF.tAB) [mm^2]
nFemurOAVector	(iMorphics): vector of 3D shape - femur [normalized units +1 = mean OA shape, -1 = mean non-OA shape]
nTibiaOAVector	(iMorphics): vector of 3D shape - tibia [normalized units +1 = mean OA shape, -1 = mean non-OA shape]
nPatellaOAVector	(iMorphics): vector of 3D shape - patella [normalized units +1 = mean OA shape, -1 = mean non-OA shape]

MedialTibialCartilage	Cartilage Volume (mm^3) - Medial Tibia
LateralTibialCartilage	Cartilage Volume (mm^3) - Lateral Tibia
MedialFemoralCartilage	Cartilage Volume (mm^3) - Medial Femur
LateralFemoralCartilage	Cartilage Volume (mm^3) - Lateral Femur
PatellarCartilage	Cartilage Volume (mm^3) - Patella
MedialMeniscus	Meniscal Volume (mm^3) - Medial
LateralMeniscus	Meniscal Volume (mm^3) - Lateral
WMTVCL	(FE): volume of cartilage - medial tibia [mm^3]
WMTSBA	(FE): total area of subchondral bone - medial tibia [cm^2]
WMTVCN	(FE): normalized cartilage volume - medial tibia
WMTMTH	(FE): mean cartilage thickness - medial tibia [mm]
WMTACS	(FE): area of cartilage surface - medial tibia [cm^2]
WMTPD	(FE): % area of subchondral bone denuded of cartilage - medial tibia [%]
WMTCAAB	(FE): area of subchondral bone covered by cartilage - medial tibia
WMTMTC	(FE): mean cartilage thickness - medial tibia [mm]
WMTMAV	(FE): maximum cartilage thickness - medial tibia (MT.ThCtAB) [mm]
WMTCTS	(FE): SD of cartilage thickness - medial tibia
WMTACV	(FE): CV of cartilage thickness - medial tibia
CMTMAT	(FE): minimum cartilage thickness - medial tibia (center) [mm]
CMTMTH	(FE): mean cartilage thickness - medial tibia (center) [mm]
EMTMTH	(FE): mean cartilage thickness - medial tibia (external) [mm]

IMTMTH	(FE): mean cartilage thickness - medial tibia (internal) [mm]
AMTMTH	(FE): mean cartilage thickness - medial tibia (anterior) [mm]
PMTMTH	(FE): mean cartilage thickness - medial tibia (posterior) [mm]
CMTPD	(FE): % area of subchondral bone denuded of cartilage - medial tibia (center) [%]
EMTPD	(FE): % area of subchondral bone denuded of cartilage - medial tibia (external) [%]
IMTPD	(FE): % area of subchondral bone denuded of cartilage - medial tibia (internal) [%]
AMTPD	(FE): % area of subchondral bone denuded of cartilage - medial tibia (anterior) [%]
PMTPD	(FE): % area of subchondral bone denuded of cartilage - medial tibia (posterior) [%]
BMFVCL	(FE): volume of cartilage - central medial femur [mm^3]
BMFSBA	(FE): total area of subchondral bone - central medial femur [cm^2]
BMFVCN	(FE): normalized cartilage volume - central medial femur
BMFMTH	(FE): mean cartilage thickness - central medial femur (internal) [mm]
BMFACS	(FE): area of cartilage surface - central medial femur
BMFPD	(FE): % area of subchondral bone denuded of cartilage - central medial femur (internal) [%]
BMFCAAB	(FE): area of subchondral bone covered by cartilage - central medial femur [cm^2]
BMFMTC	(FE): mean cartilage thickness (excldAB) - central medial femur [mm]
BMFMAV	(FE): maximum cartilage thickness - central medial femur [mm]
BMFCTS	(FE): SD of cartilage thickness - central medial femur
BMFACV	(FE): CV of cartilage thickness - central medial femur [%]
CBMFMAT	(FE): minimum cartilage thickness - central medial femur (center) [mm]
CBMFMTH	(FE): mean cartilage thickness - central medial femur (center) [mm]

EBMFMTH	(FE): mean cartilage thickness - central medial femur (external) [mm]
IBMFMTH	(FE): mean cartilage thickness - central medial femur (internal) [mm]
CBMFPD	(FE): % area of subchondral bone denuded of cartilage - central medial femur (center) [%]
EBMFPD	(FE): % area of subchondral bone denuded of cartilage - central medial femur (external) [%]
IBMFPD	(FE): % area of subchondral bone denuded of cartilage - central medial femur (internal) [%]
WMTFVCL	(FE): cartilage volume - medial tib-fem compartment [mm^3]
WMTFVCN	(FE): normalized cartilage volume - medial tib-fem compartment [cm^2]
WMTFMTH	(FE): mean cartilage thickness - medial tib-fem compartment [mm]
WMTFMAV	(FE): maximum cartilage thickness - medial tib-fem compartment [mm]
BMTFMAT	(FE): minimum cartilage thickness - central medial tib-fem compartment (weight bearing) [mm]
BMTFMTH	(FE): mean cartilage thickness - central medial tib-fem compartment (weight bearing) [mm]
WLTVCL	(FE): volume of cartilage - lateral tibia (LT.VC) [mm^3]
WLTSBA	(FE): total area of subchondral bone - lateral tibia [cm^2]
WLTVCN	(FE): normalized cartilage volume - lateral tibia
WLTMTH	(FE): mean cartilage thickness - lateral tibia [mm]
WLTACS	(FE): area of cartilage surface - lateral tibia
WLTPD	(FE): % area of subchondral bone denuded of cartilage - lateral tibia [%]
WLTCAAB	(FE): area of subchondral bone covered by cartilage - lateral tibia [cm^2]
WLTMTC	(FE): mean cartilage thickness (excl dAB) - lateral tibia [mm]
WLTMAV	(FE): maximum cartilage thickness - lateral tibia [mm]
WLTCTS	(FE): SD of cartilage thickness - lateral tibia

WLTACV	(FE): CV of cartilage thickness - lateral tibia
CLTMAT	(FE): minimum cartilage thickness - lateral tibia (center) [mm]
CLTMTH	(FE): mean cartilage thickness - lateral tibia (center) [mm]
ELTMTH	(FE): mean cartilage thickness - lateral tibia (external) [mm]
ILTMTH	(FE): mean cartilage thickness - lateral tibia (internal) [mm]
ALTMTH	(FE): mean cartilage thickness - lateral tibia (anterior) [mm]
PLTMTH	(FE): mean cartilage thickness - lateral tibia (posterior) [mm]
CLTPD	(FE): % area of subchondral bone denuded of cartilage - lateral tibia (center) [%]
ELTPD	(FE): % area of subchondral bone denuded of cartilage - lateral tibia (external) [%]
ILTPD	(FE): % area of subchondral bone denuded of cartilage - lateral tibia (internal) [%]
ALTPD	(FE): % area of subchondral bone denuded of cartilage - lateral tibia (anterior) [%]
PLTPD	(FE): % area of subchondral bone denuded of cartilage - lateral tibia (posterior) [%]
BLFVCL	(FE): volume of cartilage - central lateral femur [mm^3]
BLFSBA	(FE): total area of subchondral bone - central lateral femur [cm^2]
BLFVCN	(FE): normalized cartilage volume - central lateral femur
BLFMTH	(FE): mean cartilage thickness - central lateral femur
BLFACS	(FE): % area of subchondral bone denuded of cartilage - central lateral femur (external) [%]
BLFPD	(FE): % area of subchondral bone denuded of cartilage - central lateral femur [%]
BLFCAAB	(FE): area of subchondral bone covered by cartilage - central lateral femur [cm^2]
BLFMTC	(FE): mean cartilage thickness (excl dAB) - central lateral femur
BLFMAV	(FE): maximum cartilage thickness - central lateral femur

BLFCTS	(FE): SD of cartilage thickness - central lateral femur
BLFACV	(FE): CV of cartilage thickness - central lateral femur
CBLFMAT	(FE): minimum cartilage thickness - central lateral femur (center) [mm]
CBLFMT	(FE): mean cartilage thickness - central lateral femur (center) [mm]
EBLFMT	(FE): mean cartilage thickness - central lateral femur (external) [mm]
IBLFMT	(FE): mean cartilage thickness - central lateral femur (internal) [mm]
CBLFPD	(FE): % area of subchondral bone denuded of cartilage - central lateral femur (center) [%]
EBLFPD	(FE): % area of subchondral bone denuded of cartilage - central lateral femur (external) [%]
IBLFPD	(FE): % area of subchondral bone denuded of cartilage - central lateral femur (internal) [%]
WLTFVCL	(FE): cartilage volume - lateral tib-fem compartment [mm^3]
WLTFVCN	(FE): normalized cartilage volume - lateral tib-fem compartment
WLTFMTH	(FE): mean cartilage thickness - lateral tib-fem compartment [mm]
WLTFMAV	(FE): maximum cartilage thickness - lateral tib-fem compartment [mm]
BLTFMAT	(FE): minimum cartilage thickness - central lateral tib-fem compartment (weight bearing)) [mm]
BLTFMTH	(FE): mean cartilage thickness - central lateral tib-fem compartment (weight bearing) [mm]
SubBArea_MedFem	(Qmetrics): central medial femur subchondral bone plate (SBP) area
CurAverage_SubB_MedFem	(Qmetrics): central medial femur average SBP mean curvature
CurSD_SubB_MedFem	(Qmetrics): central medial femur standard deviation SBP mean curvature
Cur_5pc_SubB_MedFem	(Qmetrics): central medial femur 5% percentile SBP mean curvature
Cur_95pc_SubB_MedFem	(Qmetrics): central medial femur 95% percentile SBP mean curvature
SCRAverage_SubB_MedFem	(Qmetrics): central medial femur average SBPMRI signal contrast ratio (MRI signal contrast ratio: DESS signal contrast

	between cartilage tissue and bone tissue at the subchondral bone plate)
SCRSD_SubB_MedFem	(Qmetrics): central medial femur standard deviation SBP MRI signal contrast ratio
SCR_5pc_SubB_MedFem	(Qmetrics): central medial femur 5% percentile SBP MRI signal contrast ratio
SCR_95pc_SubB_MedFem	(Qmetrics): central medial femur 95% percentile SBP MRI signal contrast ratio
SubBArea_LatFem	(Qmetrics): central lateral femur subchondral bone plate (SBP) area
CurAverage_SubB_LatFem	(Qmetrics): central lateral femur average SBP mean curvature
CurSD_SubB_LatFem	(Qmetrics): central lateral femur standard deviation SBP mean curvature
Cur_5pc_SubB_LatFem	(Qmetrics): central lateral femur 5% percentile SBP mean curvature
Cur_95pc_SubB_LatFem	(Qmetrics): central lateral femur 95% percentile SBP mean curvature
SCRAverage_SubB_LatFem	(Qmetrics): central lateral femur average SBPMRI signal contrast ratio (MRI signal contrast ratio: DESS signal contrast between cartilage tissue and bone tissue at the subchondral bone plate)
SCRSD_SubB_LatFem	(Qmetrics): central lateral femur standard deviation SBP MRI signal contrast ratio
SCR_5pc_SubB_LatFem	(Qmetrics): central lateral femur 5% percentile SBP MRI signal contrast ratio
SCR_95pc_SubB_LatFem	(Qmetrics): central lateral femur 95% percentile SBP MRI signal contrast ratio
SubBArea_MedTib	(Qmetrics): medial tibia subchondral bone plate (SBP) area
CurAverage_SubB_MedTib	(Qmetrics): medial tibia average SBP mean curvature
CurSD_SubB_MedTib	(Qmetrics): medial tibia standard deviation SBP mean curvature
Cur_5pc_SubB_MedTib	(Qmetrics): medial tibia 5% percentile SBP mean curvature
Cur_95pc_SubB_MedTib	(Qmetrics): medial tibia 95% percentile SBP mean curvature
SCRAverage_SubB_MedTib	(Qmetrics): medial tibia average SBP MRI signal contrast ratio (MRI signal contrast ratio: DESS signal contrast between cartilage tissue and bone tissue at the subchondral bone plate)
SCRSD_SubB_MedTib	(Qmetrics): medial tibia standard deviation SBP MRI signal contrast ratio

SCR_5pc_SubB_MedTib	(Qmetrics): medial tibia 5% percentile SBP MRI signal contrast ratio
SCR_95pc_SubB_MedTib	(Qmetrics): medial tibia 95% percentile SBP MRI signal contrast ratio
SubBArea_LatTib	(Qmetrics): lateral tibia subchondral bone plate (SBP) area
CurAverage_SubB_LatTib	(Qmetrics): lateral tibia average SBP mean curvature
CurSD_SubB_LatTib	(Qmetrics): lateral tibia standard deviation SBP mean curvature
Cur_5pc_SubB_LatTib	(Qmetrics): lateral tibia 5% percentile SBP mean curvature
Cur_95pc_SubB_LatTib	(Qmetrics): lateral tibia 95% percentile SBP mean curvature
SCRAverage_SubB_LatTib	(Qmetrics): lateral tibia average SBP MRI signal contrast ratio (MRI signal contrast ratio: DESS signal contrast between cartilage tissue and bone tissue at the subchondral bone plate)
SCRSD_SubB_LatTib	(Qmetrics): lateral tibia standard deviation SBP MRI signal contrast ratio
SCR_5pc_SubB_LatTib	(Qmetrics): lateral tibia 5% percentile SBP MRI signal contrast ratio
SCR_95pc_SubB_LatTib	(Qmetrics): lateral tibia 95% percentile SBP MRI signal contrast ratio
MCMPM	(BI): MOAKS: cartilage morphology - patella medial
MCMPL	(BI): MOAKS: cartilage morphology - patella lateral
MCMFMA	(BI): MOAKS: cartilage morphology - femur medial anterior (trochlear)
MCMFLA	(BI): MOAKS: cartilage morphology - femur lateral anterior (trochlear)
MCMFMP	(BI): MOAKS: cartilage morphology - femur medial posterior
MCMFLP	(BI): MOAKS: cartilage morphology - femur lateral posterior
MCMFMC	(BI): MOAKS: cartilage morphology - femur medial central
MCMFLC	(BI): MOAKS: cartilage morphology - femur lateral central
MCMTMA	(BI): MOAKS: cartilage morphology - tibia medial anterior

MCMTLA	(BI): MOAKS: cartilage morphology - tibia lateral anterior
MCMTMC	(BI): MOAKS: cartilage morphology - tibia medial central
MCMTLC	(BI): MOAKS: cartilage morphology - tibia lateral central
MCMTMP	(BI): MOAKS: cartilage morphology - tibia medial posterior
MCMTLP	(BI): MOAKS: cartilage morphology - tibia lateral posterior
MBMSFMA	(BI): MOAKS: BML size - femur medial anterior (trochlear)
MBMPFMA	(BI): MOAKS: BML (% lesion that is edema) - femur medial anterior (trochlear)
MBMNFMA	(BI): MOAKS: number of BML lesions - femur medial anterior (trochlear)
MBMSFLA	(BI): MOAKS: BML size - femur lateral anterior (trochlear)
MBMPFLA	(BI): MOAKS: BML (% lesion that is edema) - femur lateral anterior (trochlear)
MBMNFLA	(BI): MOAKS: number of BML lesions - femur lateral anterior (trochlear)
MBMSFMC	(BI): MOAKS: BML size - femur medial central
MBMPFMC	(BI): MOAKS: BML (% lesion that is edema) - femur medial central
MBMNFMC	(BI): MOAKS: number of BML lesions - femur medial central
MBMSFLC	(BI): MOAKS: BML size - femur lateral central
MBMPFLC	(BI): MOAKS: BML (% lesion that is edema) - femur lateral central
MBMNFLC	(BI): MOAKS: number of BML lesions - femur lateral central
MBMSFMP	(BI): MOAKS: BML size - femur medial posterior
MBMPFMP	(BI): MOAKS: BML (% lesion that is edema) - femur medial posterior
MBMNFMP	(BI): MOAKS: number of BML lesions - femur medial posterior
MBMSFLP	(BI): MOAKS: BML size - femur lateral posterior

MBMPFLP	(BI): MOAKS: BML (% lesion that is edema) - femur lateral posterior
MBMNFLP	(BI): MOAKS: number of BML lesions - femur lateral posterior
MBMSSS	(BI): MOAKS: BML size - tibia sub-spinous
MBMPSS	(BI): MOAKS: BML (% lesion that is edema) - tibia sub-spinous
MBMNSS	(BI): MOAKS: number of BML lesions - tibia sub-spinous
MBMSTMA	(BI): MOAKS: BML size - tibia medial anterior
MBMPTMA	(BI): MOAKS: BML (% lesion that is edema) - tibia medial anterior
MBMNTMA	(BI): MOAKS: number of BML lesions - tibia medial anterior
MBMSTLA	(BI): MOAKS: BML size - tibia lateral anterior
MBMPTLA	(BI): MOAKS: BML (% lesion that is edema) - tibia lateral anterior
MBMNTLA	(BI): MOAKS: number of BML lesions - tibia lateral anterior
MBMSTMC	(BI): MOAKS: BML size - tibia medial central
MBMPTMC	(BI): MOAKS: BML (% lesion that is edema) - tibia medial central
MBMNTMC	(BI): MOAKS: number of BML lesions - tibia medial central
MBMSTLC	(BI): MOAKS: BML size - tibia lateral central
MBMPTLC	(BI): MOAKS: BML (% lesion that is edema) - tibia lateral central
MBMNTLC	(BI): MOAKS: number of BML lesions - tibia lateral central
MBMSTMP	(BI): MOAKS: BML size - tibia medial posterior
MBMPTMP	(BI): MOAKS: BML (% lesion that is edema) - tibia medial posterior
MBMNTMP	(BI): MOAKS: number of BML lesions - tibia medial posterior
MBMSTLP	(BI): MOAKS: BML size - tibia lateral posterior

MBMPTLP	(BI): MOAKS: BML (% lesion that is edema) - tibia lateral posterior
MBMNTLP	(BI): MOAKS: number of BML lesions - tibia lateral posterior
MBMSPM	(BI): MOAKS: BML size - patella medial
MBMPPM	(BI): MOAKS: BML (% lesion hat is edema) - patella medial
MBMNPM	(BI): MOAKS: number of BML lesions - patella medial
MBMSPL	(BI): MOAKS: BML size - patella lateral
MBMPPL	(BI): MOAKS: BML (% lesion that is edema) - patella lateral
MBMNPL	(BI): MOAKS: number of BML lesions - patella lateral
MMTMA	(BI): MOAKS: medial meniscal morphology - anterior horn
MMTLA	(BI): MOAKS: lateral meniscal morphology - anterior horn
MMTMB	(BI): MOAKS: medial meniscal morphology - body
MMTLB	(BI): MOAKS: lateral meniscal morphology - body
MMTMP	(BI): MOAKS: medial meniscal morphology - posterior horn
MMTLP	(BI): MOAKS: lateral meniscal morphology - posterior horn
MMHMA	(BI): MOAKS: medial meniscal hypertrophy - anterior horn
MMHLA	(BI): MOAKS: lateral meniscal hypertrophy - anterior horn
MMHMB	(BI): MOAKS: medial meniscal hypertrophy - body
MMHLB	(BI): MOAKS: lateral meniscal hypertrophy - body
MMHMP	(BI): MOAKS: medial meniscal hypertrophy - posterior horn
MMHLP	(BI): MOAKS: lateral meniscal hypertrophy - posterior horn
MMSMA	(BI): MOAKS: medial meniscal signal abnormality - anterior horn

MMSLA	(BI): MOAKS: lateral meniscal signal abnormality - anterior horn
MMSMB	(BI): MOAKS: medial meniscal signal abnormality - body
MMSLB	(BI): MOAKS: lateral meniscal signal abnormality - body
MMSMP	(BI): MOAKS: medial meniscal signal abnormality - posterior horn
MMSLP	(BI): MOAKS: lateral meniscal signal abnormality - posterior horn
MMXMM	(BI): MOAKS: medial meniscal extrusion - medially
MMXMA	(BI): MOAKS: medial meniscal extrusion - anteriorly
MMXLA	(BI): MOAKS: lateral meniscal extrusion - anteriorly
MMXLL	(BI): MOAKS: lateral meniscal extrusion - laterally
MMRTM	(BI): MOAKS: medial meniscal morphology - posterior root tear
MMRTL	(BI): MOAKS: lateral meniscal morphology - posterior root tear
MOSPS	(BI): MOAKS: osteophyte size - patella superior
MOSPI	(BI): MOAKS: osteophyte size - patella inferior
MOSPM	(BI): MOAKS: osteophyte size - patella medial
MOSPL	(BI): MOAKS: osteophyte size - patella lateral
MOSFMA	(BI): MOAKS: osteophyte size - femur medial anterior (trochlear)
MOSFLA	(BI): MOAKS: osteophyte size - femur lateral anterior (trochlear)
MOSFMP	(BI): MOAKS: osteophyte size - femur medial posterior
MOSFLP	(BI): MOAKS: osteophyte size - femur lateral posterior
MOSFMC	(BI): MOAKS: osteophyte size - femur medial central
MOSFLC	(BI): MOAKS: osteophyte size - femur lateral central

MOSTM	(BI): MOAKS: osteophyte size - tibia medial
MOSTL	(BI): MOAKS: osteophyte size - tibial lateral
MACLTR	(BI): MOAKS: ACL tear
MACLBML	(BI): MOAKS: BML associated with ACL insertion
MACLRP	(BI): MOAKS: ACL repair
MPCLTR	(BI): MOAKS: PCL tear
MPCLBML	(BI): MOAKS: BML associated with PCL insertion
MPCLRP	(BI): MOAKS: PCL repair
MPTSIG	(BI): MOAKS: patella tendon signal abnormality
MGCACL	(BI): MOAKS: ACL ganglion cyst
MGCPCL	(BI): MOAKS: PCL ganglion cyst
MGCTIB	(BI): MOAKS: tibial ganglion cyst
MGCSM	(BI): MOAKS: semimembranosus ganglion cyst
MGSST	(BI): MOAKS: semitendinosus ganglion cyst
MGCOTH	(BI): MOAKS: other ganglion cyst
MANSBUR	(BI): MOAKS: pes anserine bursa
MIPBUR	(BI): MOAKS: infrapatellar bursa
MPPBUR	(BI): MOAKS: prepatella bursa
MSYIC	(BI): MOAKS: inter-condylar synovitis
MEFFWK	(BI): MOAKS: whole knee effusion
MPOPCYS	(BI): MOAKS: popliteal cyst

Supplemental Table 2: ***AutoPrognosis 2.0*** **algorithm inventory**

List of algorithms currently included in *AutoPrognosis 2.0*, grouped by pipeline stage. Algorithms highlighted in bold represent the default selections, selected for their speed and efficiency (detailed at AutoPrognosis 2.0 website: https://www.autoprognosis.vanderschaar-lab.com/).

Pipeline Stage	Algorithms				
Feature Scaling	MaxAbs	Standard Scaler	L2 Normalizer	Normal Transform	Uniform Transform
	MinMax	None			
Feature Selection	Feature Agglomeration	Fast ICA	Variance Threshold	Gaussian Projection	PCA
	Data Cleanup	None			
Imputation	Sinkhorn	EM	(M)ICE	ICE	HyperImpute
	Most-Frequent	Median	MissForest	SoftImpute	None
	Mean	Gain			
Classifiers	**XGBoost**	**CatBoost**	KNN	ADABoost	Bernoulli Naïve Bayes
	Neural Nets	Linear SVM	QDA	Decision Trees	**Logistic Regression**
	Hist Gradient Boosting	ExtraTree	Bagging	Gradient Boosting	Ridge Classifier
	Gaussian Process	Perceptron	**LGBM**	**Random Forest**	TabNet
	Multinomial Naïve Bayes	LDA	Gaussian Naïve Bayes		

Supplemental Table 3*:* Demographic characteristics of study population

This table provides a breakdown of the demographic characteristics among the study population.

Characteristics	Training Set	Testing Set
Number of instances	1353	338
Sex:		
Male	540 (39.9%)	154 (45.6%)
Female	813 (60.1%)	184 (54.4%)
Age (years):		
Range	45 – 81	45 – 81
Median	62	63
Ethnicity:		
Asian	12 (0.9%)	2 (0.6%)
Black or African American	221 (16.3%)	75 (22.2%)
White or Caucasian	1102 (81.4%)	247 (73.1%)
Other Non-White	18 (1.3%)	14 (4.1%)
Baseline KLG:		
KLG 1	149 (11.0%)	33 (9.8%)
KLG 2	681 (50.3%)	166 (49.1%)
KLG 3	504 (37.3%)	139 (41.1%)
KLG 4	19 (1.4%)	0 (0.0%)
Clinical Outcomes:		
Class 0	820 (60.6%)	205 (60.7%)
Class 1	97 (7.2%)	34 (10.1%)
Class 2	361 (26.7%)	77 (22.8%)
Class 3	75 (5.5%)	22 (6.5%)

Supplemental Table 4: Pipeline ensembles

Pipeline ensemble for each model developed using *AutoPrognosis 2.0*, for both multi-class and binary predictions.

Multi-class predictions	
Model	**Pipeline Ensemble**
AP1_mu	gain->data_cleanup->random_forest + softimpute->variance_threshold->feature_normalizer->data_cleanup->logistic_regression + ice->feature_agglomeration->maxabs_scaler->data_cleanup->lgbm --> gain->data_cleanup->random_forest
AP2_mu	0.624999999609375 * (mice->pca->minmax_scaler->data_cleanup->random_forest) + 0.1874999998828125 * (median->variance_threshold->feature_normalizer->data_cleanup->logistic_regression) + 0.1874999998828125 * (missforest->variance_threshold->scaler->data_cleanup->lgbm)
AP3_mu	0.3124999998046875 * (gain->data_cleanup->random_forest) + 0.24999999984375 * (gain->data_cleanup->lgbm) + 0.4374999997265625 * (gain->data_cleanup->xgboost)
AP4_mu	0.3157894735180055 * (median->nop->nop->data_cleanup->random_forest) + 0.3684210524376731 * (gain->data_cleanup->lgbm) + 0.3157894735180055 * (median->feature_agglomeration->minmax_scaler->data_cleanup->logistic_regression)
AP5_mu	average(median->nop->nop->data_cleanup->random_forest + gain->data_cleanup->lgbm + gain->data_cleanup->xgboost)
AP5_top5_mu	gain->data_cleanup->random_forest + missforest->data_cleanup->scaler->data_cleanup->logistic_regression + gain->data_cleanup->lgbm --> gain->data_cleanup->random_forest

Binary predictions	
Model	**Pipeline Ensemble**
AP1_bi	ice->nop->scaler->data_cleanup->random_forest + median->feature_agglomeration->scaler->data_cleanup->catboost + nop->data_cleanup->uniform_transform->data_cleanup->lgbm --> ice->nop->scaler->data_cleanup->random_forest
AP2_bi	0.5882352937716263 * (mice->pca->minmax_scaler->data_cleanup->random_forest) + 0.2352941175086505 * (median->feature_agglomeration->scaler->data_cleanup->catboost) + 0.1764705881314879 * (ice->nop->uniform_transform->data_cleanup->xgboost)
AP3_bi	0.5555555552469136 * (ice->data_cleanup->nop->data_cleanup->catboost) + 0.2777777776234568 * (softimpute->variance_threshold->maxabs_scaler->data_cleanup->xgboost) + 0.16666666657407406 * (gain->data_cleanup->random_forest)
AP4_bi	median->nop->nop->data_cleanup->random_forest + gain->data_cleanup->lgbm + softimpute->nop->minmax_scaler->data_cleanup->catboost --> median->nop->nop->data_cleanup->random_forest
AP5_bi	median->nop->nop->data_cleanup->random_forest + gain->data_cleanup->lgbm + softimpute->nop->minmax_scaler->data_cleanup->catboost --> median->nop->nop->data_cleanup->random_forest
AP5_top5_bi	average(median->pca->minmax_scaler->data_cleanup->random_forest + gain->nop->uniform_transform->data_cleanup->catboost + ice->pca->maxabs_scaler->data_cleanup->logistic_regression)

Supplemental Table 5: Overview of the core features for multi-class and binary predictions

This table ranks and describes the five "core" features for multi-class and binary predictions, identified by *post-hoc* interpretability analysis, that were used to create our streamlined models AP5_top5_mu and AP5_top5_bi.

Multi-class predictions		
Rank	**Feature**	**Description**
1	**MMXMA**	MOAKS: medial meniscal extrusion - anteriorly
2	**WOMKP**	WOMAC pain score
3	**EMTPD**	% area of subchondral bone denuded of cartilage - medial tibia (external) [%]
4	**WOMADL**	WOMAC disability score
5	**EBMFPD**	% area of subchondral bone denuded of cartilage - central medial femur (external) [%]

Binary predictions		
Rank	**Feature**	**Description**
1	**EBMFPD**	% area of subchondral bone denuded of cartilage - central medial femur (external) [%]
2	**WOMADL**	WOMAC disability score
3	**EMTPD**	% area of subchondral bone denuded of cartilage - medial tibia (external) [%]
4	**MMXMA**	MOAKS: medial meniscal extrusion - anteriorly
5	**MOSFMA**[i]	MOAKS: osteophyte size - femur medial anterior (trochlear)

[i] MOSFMA was utilised as a core feature in our streamlined model AP5_top5_bi in place of Urine_alpha_NUM (urine CTX-1a), as the latter was not available in the external dataset.

Supplemental Table 6*:* Demographic characteristics of the POMA dataset used for validation
This table provides a breakdown of the demographic characteristics among the POMA population, on which external validation of our streamlined models was conducted.

Characteristics	External validation set
Number of instances	705
Sex:	
Male	281 (39.9%)
Female	424 (60.1%)
Age (years):	
Range	45 – 79
Median	61
Ethnicity:	
Asian	6 (0.9%)
Black or African American	106 (15.0%)
White or Caucasian	581 (82.4%)
Other Non-White	12 (1.7%)
Baseline KLG:	
KLG 0	165 (23.4%)
KLG 1	232 (32.9%)
KLG 2	87 (12.3%)
KLG 3	114 (16.2%)
KLG 4	107 (15.2%)
Clinical Outcomes:	
Class 0	414 (58.7%)
Class 1	40 (5.7%)
Class 2	207 (29.4%)
Class 3	44 (6.2%)

Supplemental Table 7*:* Demographic characteristics of the subpopulations in the hold-out set
This table provides a breakdown of the demographic characteristics among the subpopulations in the hold-out validation set used in our subgroup analysis.

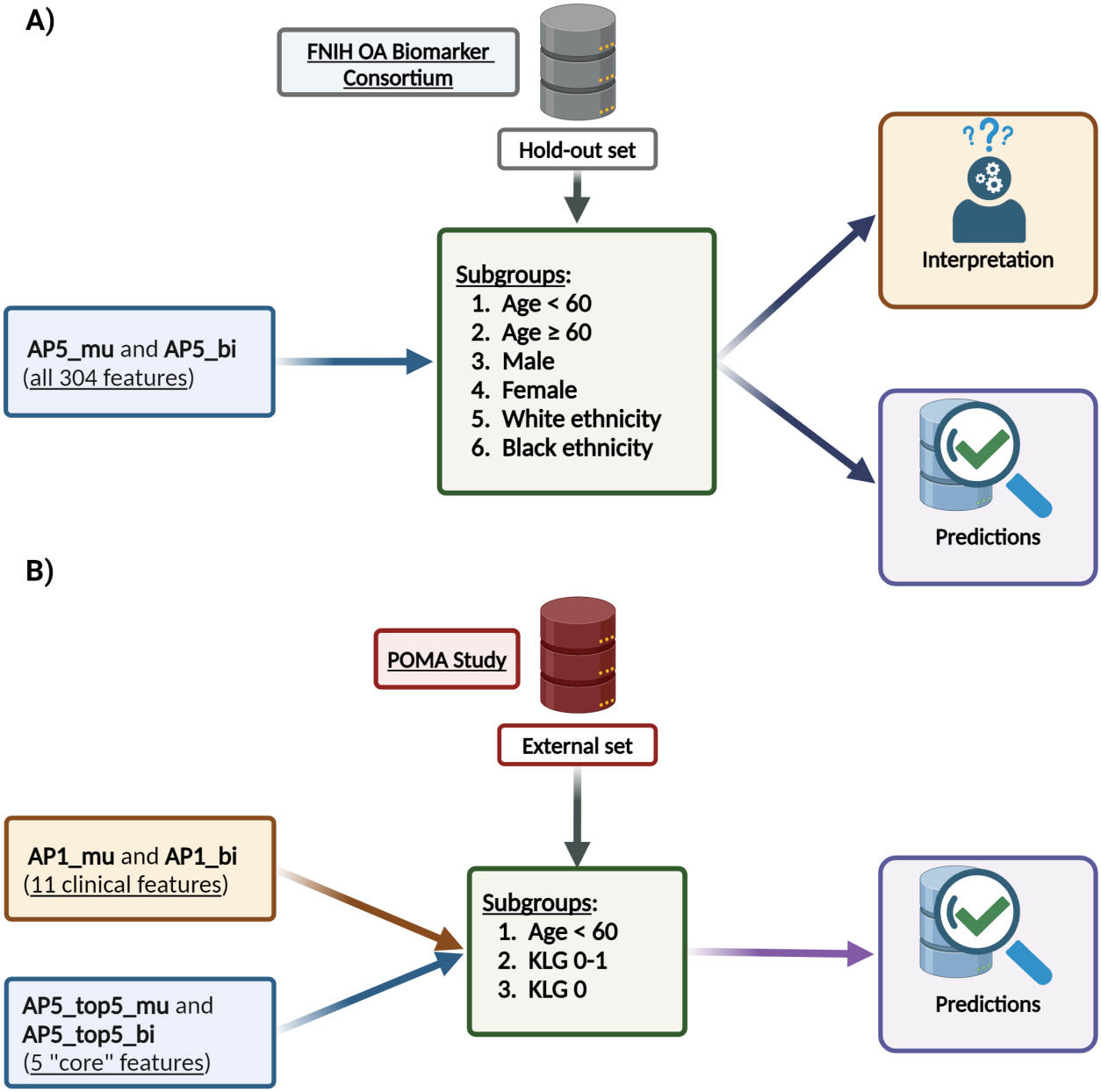
Characteristics	Age <60	Age >60	Male	Female	White ethnicity	Black ethnicity
Number of instances	135	203	154	184	247	75
Sex:						
Male	66 (48.9%)	88 (43.3%)	154 (100.0%)	0 (0.0%)	126 (51.0%)	20 (26.7%)
Female	69 (51.1%)	115 (56.7%)	0 (0.0%)	184 (100.0%)	121 (49.0%)	55 (73.3%)
Age (years):						
Range	45 – 59	60 – 81	45 – 80	46 – 81	45 – 81	48 – 80
Median	53.0	69.0	63.0	64.0	64.0	59.0
Ethnicity:						
Asian	2 (1.5%)	0 (0.0%)	0 (0.0%)	2 (1.1%)	0 (0.0%)	0 (0.0%)
Black or African American	40 (29.6%)	35 (17.2%)	20 (13.0%)	55 (29.9%)	0 (0.0%)	75 (100.0%)
White or Caucasian	88 (65.2%)	159 (78.3%)	126 (81.8%)	121 (65.8%)	247 (100.0%)	0 (0.0%)
Other Non-White	5 (3.7%)	9 (4.4%)	8 (5.2%)	6 (3.3%)'}	0 (0.0%)	0 (0.0%)
Baseline KLG:						
KLG 0	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
KLG 1	12 (8.9%)	21 (10.3%)	22 (14.3%)	11 (6.0%)	30 (12.1%)	3 (4.0%)
KLG 2	89 (65.9%)	77 (37.9%)	55 (35.7%)	111 (60.3%)	107 (43.3%)	56 (74.7%)
KLG 3	34 (25.2%)	105 (51.7%)	77 (50.0%)	62 (33.7%)	110 (44.5%)	16 (21.3%)
KLG 4	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Clinical Outcomes:						
Class 0	79 (58.5%)	126 (62.1%)	90 (58.4%)	115 (62.5%)	157 (63.6%)	41 (54.7%)
Class 1	18 (13.3%)	16 (7.9%)	8 (5.2%)	26 (14.1%)	15 (6.1%)	19 (25.3%)
Class 2	28 (20.7%)	49 (24.1%)	49 (31.8%)	28 (15.2%)	65 (26.3%)	4 (5.3%)
Class 3	10 (7.4%)	12 (5.9%)	7 (4.5%)	15 (8.2%)	10 (4.0%)	11 (14.7%)

Supplemental Table 8*:* Demographic characteristics of the subpopulations in the external validation set

This table provides a breakdown of the demographic characteristics among the subpopulations in the external validation set used in our subgroup analysis.

Characteristics	Age <60	KLG 0-1	KLG 0
Number of instances	317	399	164
Sex:			
Male	124 (39.1%)	150 (37.6%)	70 (42.7%)
Female	193 (60.9%)	249 (62.4%)	94 (57.3%)
Age (years):			
Range	45 – 59	45 – 78	45 – 78
Median	54.0	58.0	58.5
Ethnicity:			
Asian	6 (1.9%)	5 (1.3%)	2 (1.2%)
Black or African American	63 (19.9%)	58 (14.5%)	22 (13.4%)
White or Caucasian	242 (76.3%)	332 (83.2%)	136 (82.9%)
Other Non-White	6 (1.9%)	4 (1.0%)	4 (2.4%)
Baseline KLG:			
KLG 0	88 (27.8%)	164 (41.1%)	164 (100.0%)
KLG 1	131 (41.3%)	235 (58.9%)	0 (0.0%)
KLG 2	38 (12.0%)	0 (0.0%)	0 (0.0%)'
KLG 3	31 (9.8%)	0 (0.0%)	0 (0.0%)
KLG 4	29 (9.1%)	0 (0.0%)	0 (0.0%)
Clinical Outcomes:			
Class 0	202 (63.7%)	299 (74.9%)	122 (74.4%)
Class 1	17 (5.4%)	14 (3.5%)	4 (2.4%)
Class 2	81 (25.6%)	78 (19.5%)	35 (21.3%)
Class 3	17 (5.4%)	8 (2.0%)	3 (1.8%)

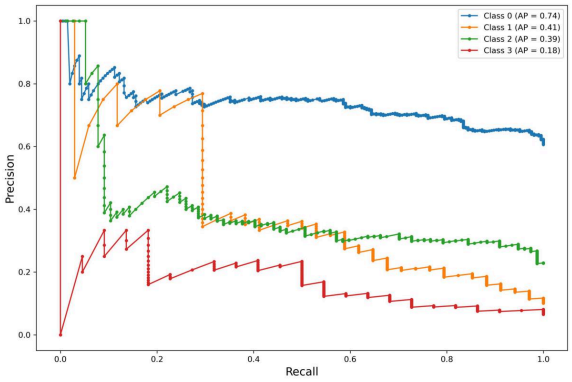
Supplemental Figure 1: Overview of subgroup analysis. This figure schematically presents our approach to subgroup analysis. **A)** Initially, model AP5_mu and AP5_bi (including all 304 variables), were assessed on various subgroups from the hold-out validation set. A *post-hoc* interpretability analysis was performed for these subgroups using the comprehensive AP5_mu and AP5_bi models. **B)** This was followed by the evaluation of models AP1_mu and AP1_bi (including clinical data) and our streamlined models AP5_top5_mu and AP5_top5_bi, on three distinct subgroups within the external validation set: 1)under-60 participants from the POMA study; 2) patients with early-stage OA (KLG 0-1); 3) patients without initial radiographic OA signs (KLG 0). (Created with BioRender.com).



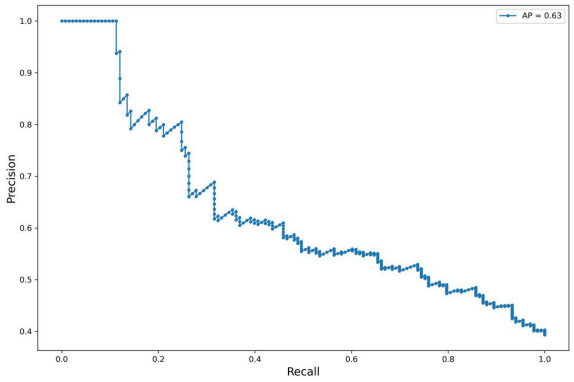
Supplemental Figure 2: Precision-Recall curves for hold-out validation. Precision-Recall curves for models AP1_mu, AP1_bi, AP5_mu, AP5_bi, AP5_top5_mu and AP5_top5_bi, validated on the hold-out set. The predicted classes were determined by assigning the class with the highest predicted probability as the outcome, rather than applying a specific probability threshold.

A)

AP1_mu

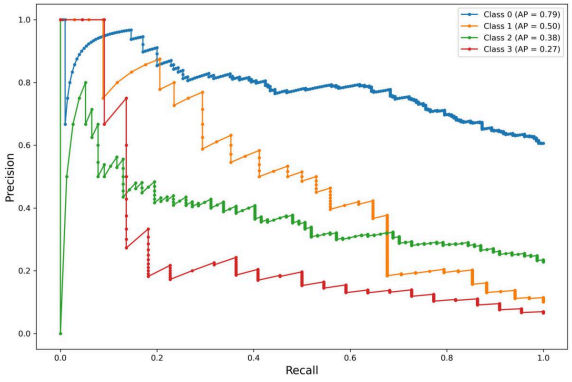


AP1_bi

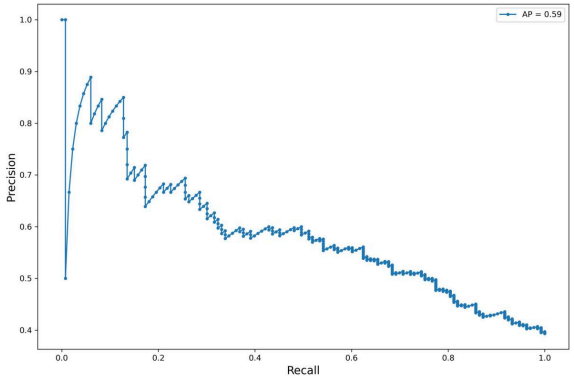


B)

AP5_mu

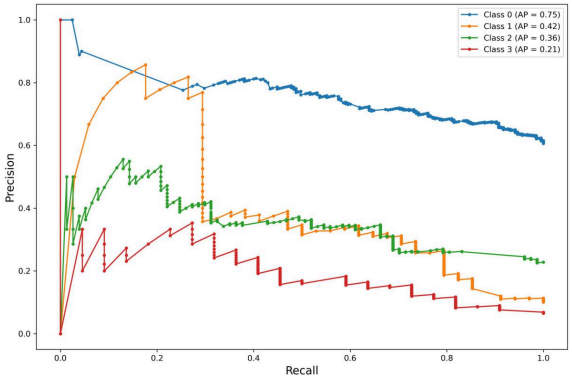


AP5_bi

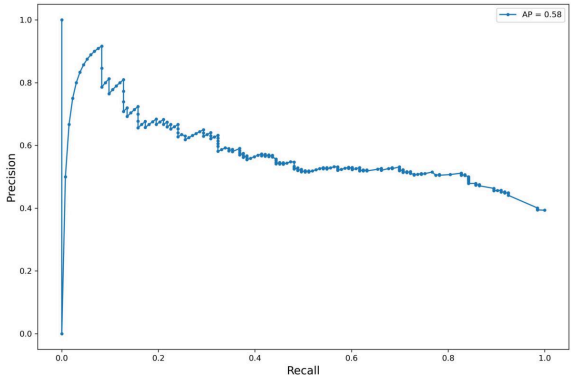


C)

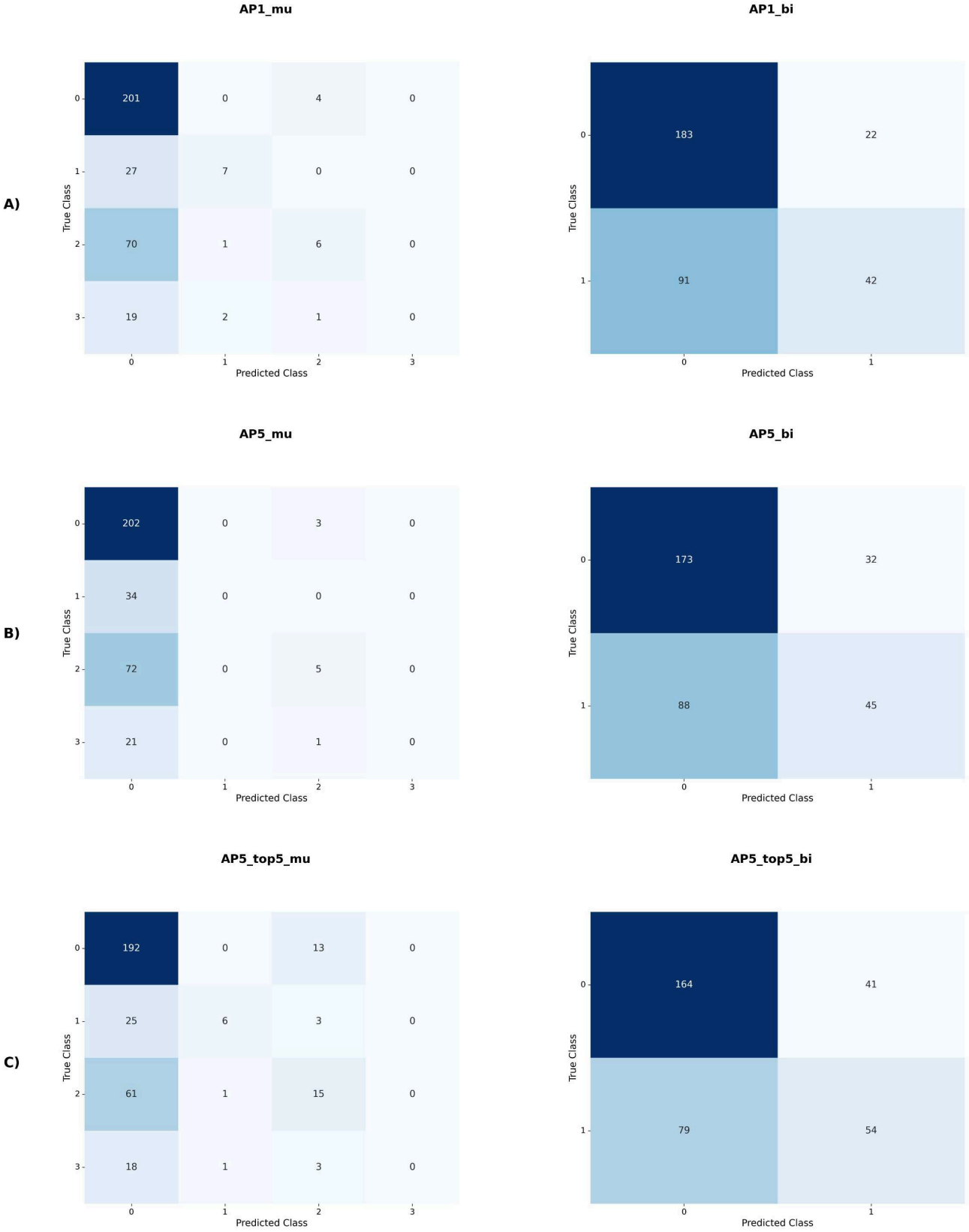
AP5_top5_mu



AP5_top5_bi

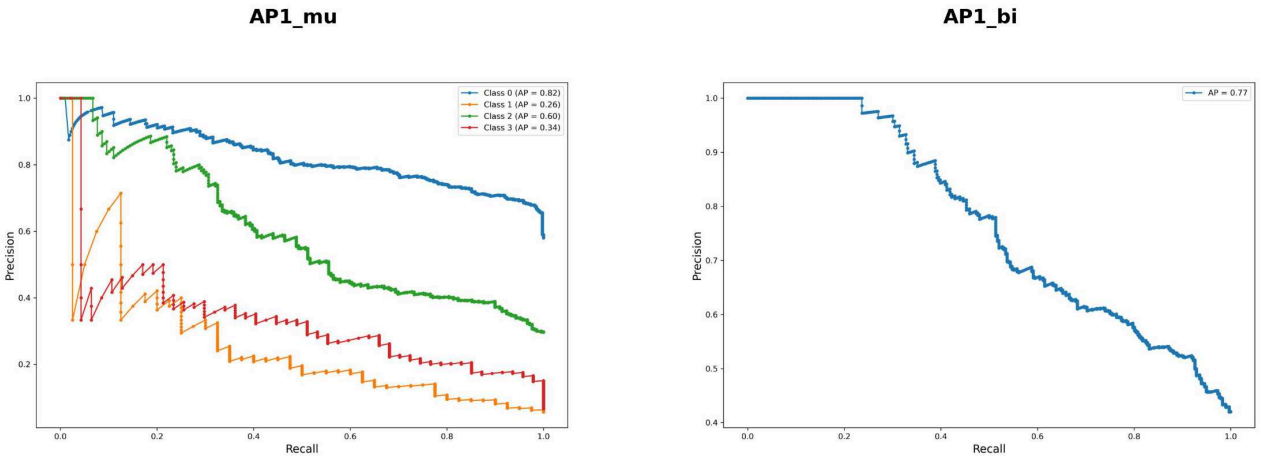


Supplemental Figure 3: Confusion matrices for hold-out validation. Confusion matrices for models AP1_mu, AP1_bi, AP5_mu, AP5_bi, AP5_top5_mu and AP5_top5_bi, validated on the hold-out set. The predicted classes were determined by assigning the class with the highest predicted probability as the outcome, rather than applying a specific probability threshold.

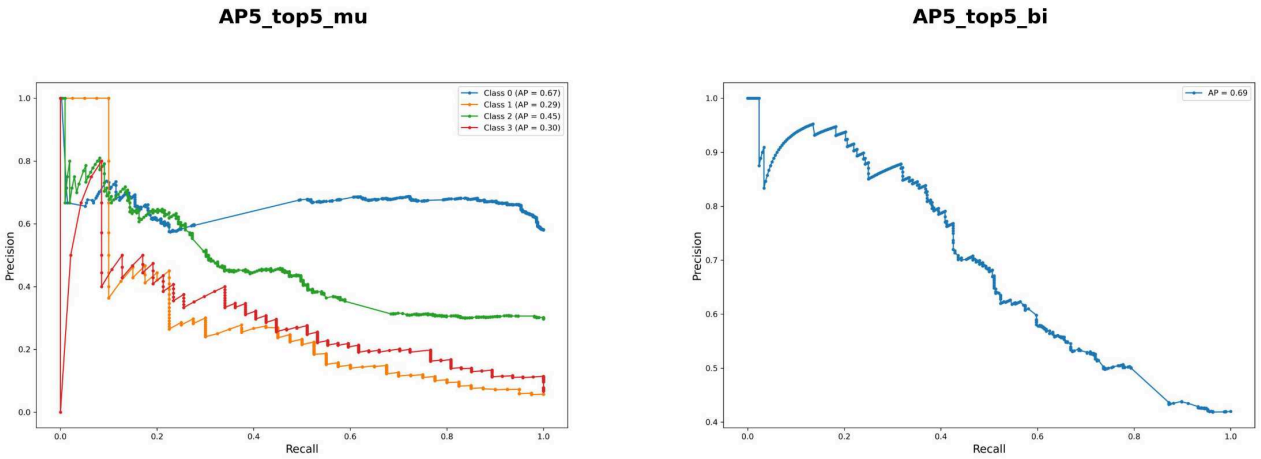


Supplemental Figure 4: Precision-Recall curves for external validation. Precision-Recall curves for models AP1_mu, AP1_bi, AP5_top5_mu and AP5_top5_bi, validated using patients from the POMA study. The predicted classes were determined by assigning the class with the highest predicted probability as the outcome, rather than applying a specific probability threshold.

A)

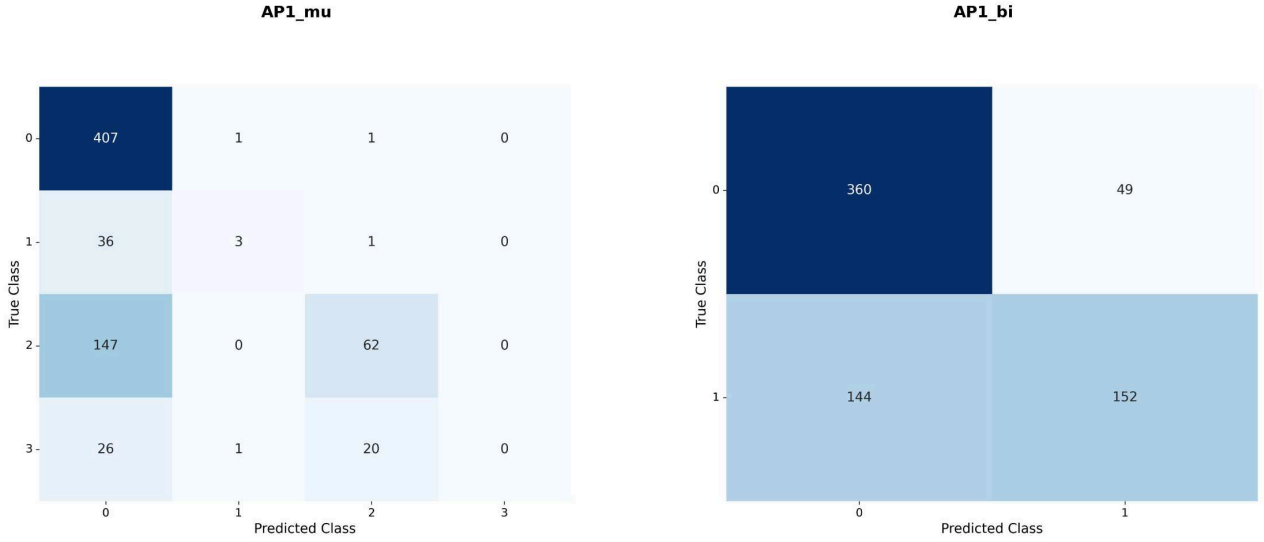


B)

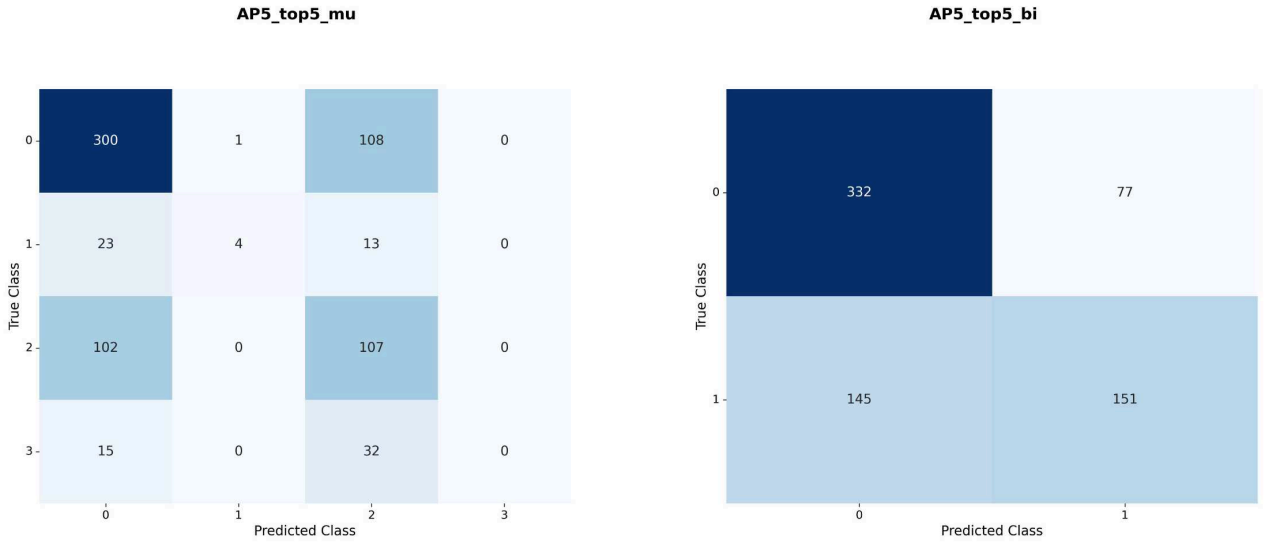


Supplemental Figure 5: Confusion matrices for external validation. Confusion matrices for models AP1_mu, AP1_bi, AP5_top5_mu and AP5_top5_bi, validated using patients from the POMA study. The predicted classes were determined by assigning the class with the highest predicted probability as the outcome, rather than applying a specific probability threshold.

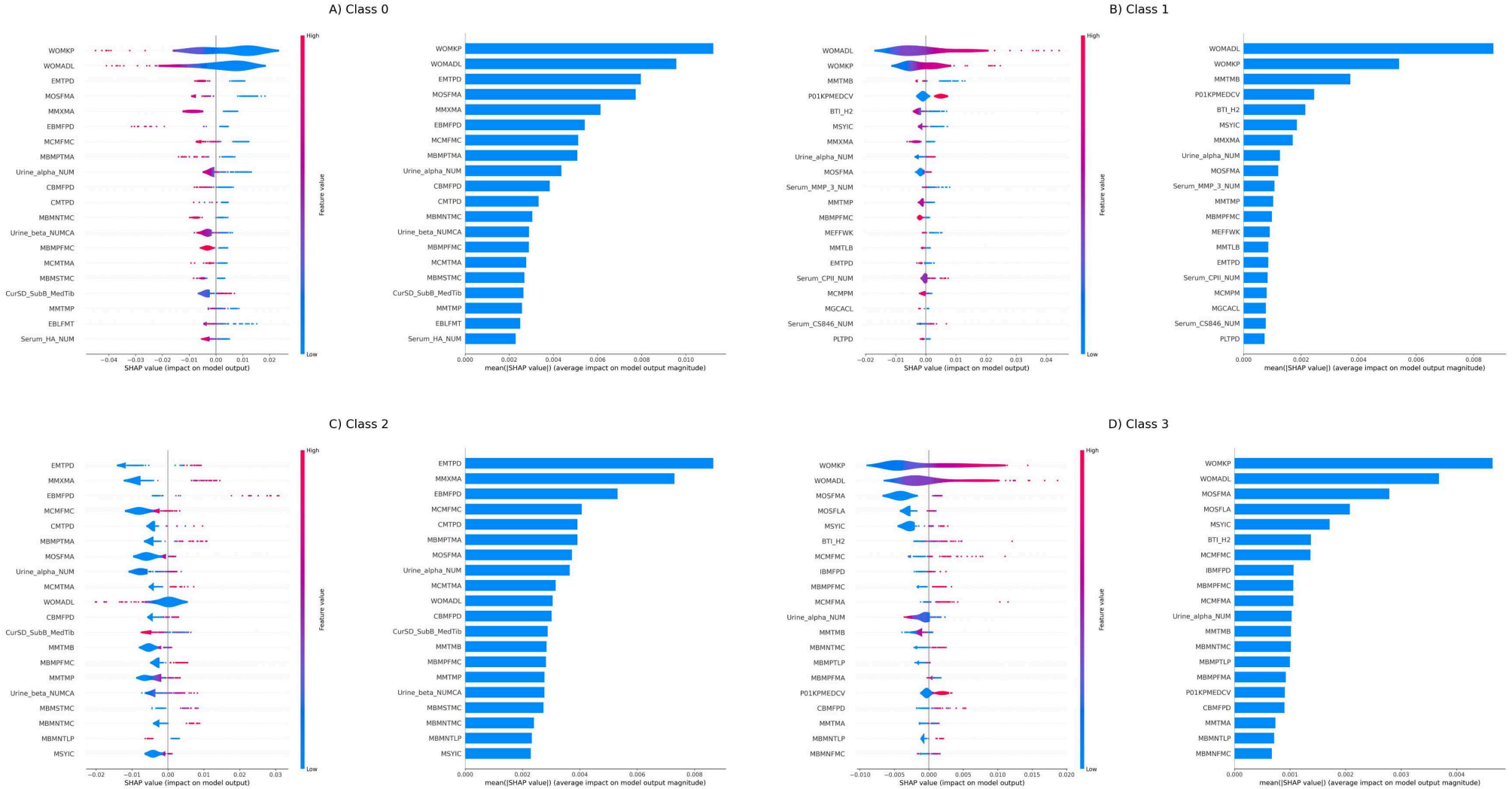
A)



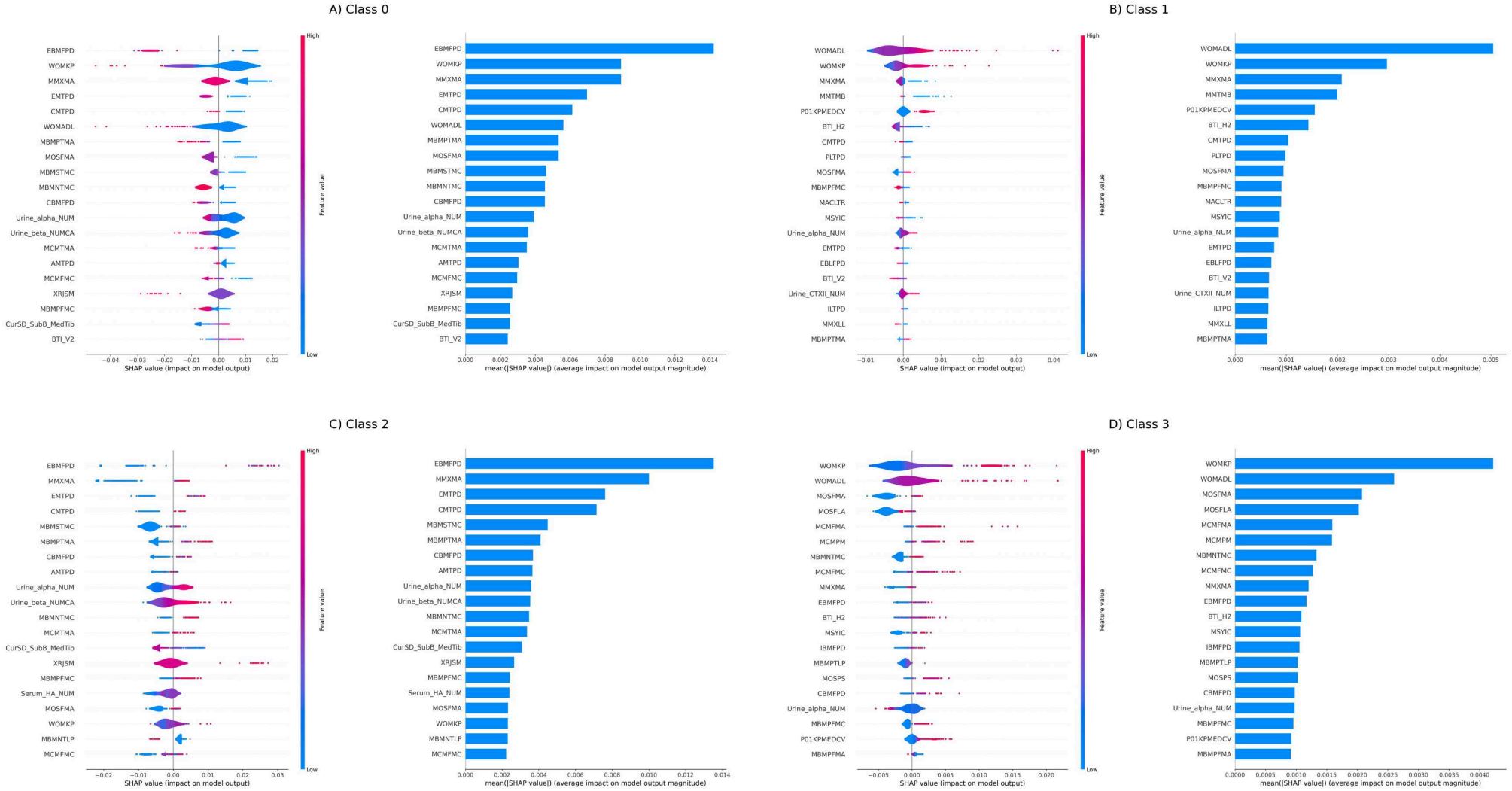
B)



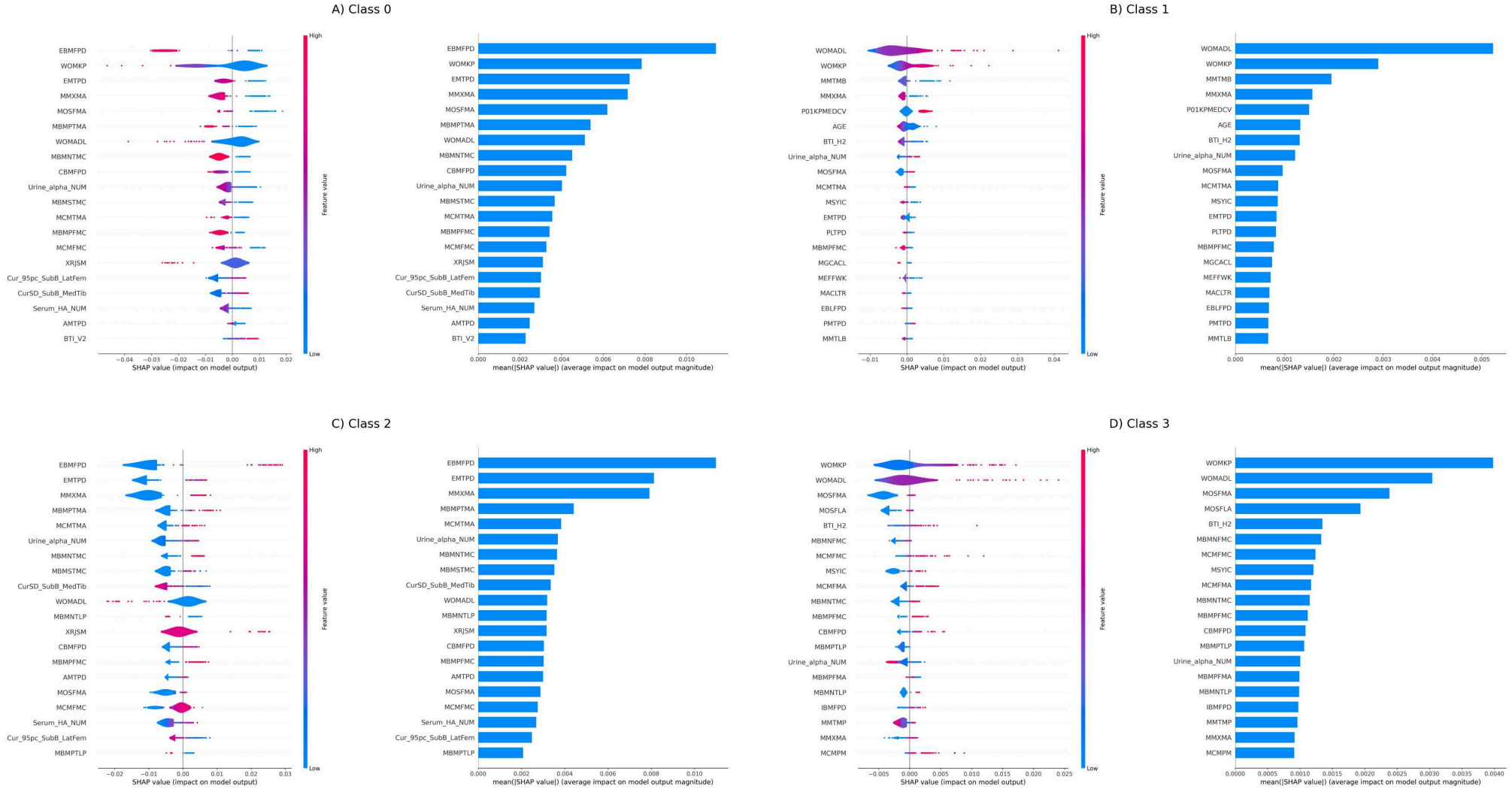
Supplemental Figure 6: Assessment of feature impact on MULTI-CLASS predictions for patients with AGE < 60 in the hold-out set. Assessment of feature impact on non-progression (Class 0, **panel A**), pain-only progression (Class 1, **panel B**), radiographic progression (Class 2, **panel C**) and both pain and radiographic progression (Class 3, **panel D**), using "Kernel-SHAP" for multi-class predictions with model AP5_mu. **Left** – impact distribution of the most important features. The colour represents the feature value (red = high, blue = low). A positive SHAP value represents a positive impact on class prediction. **Right** – average impact magnitude of the most important features on class prediction.



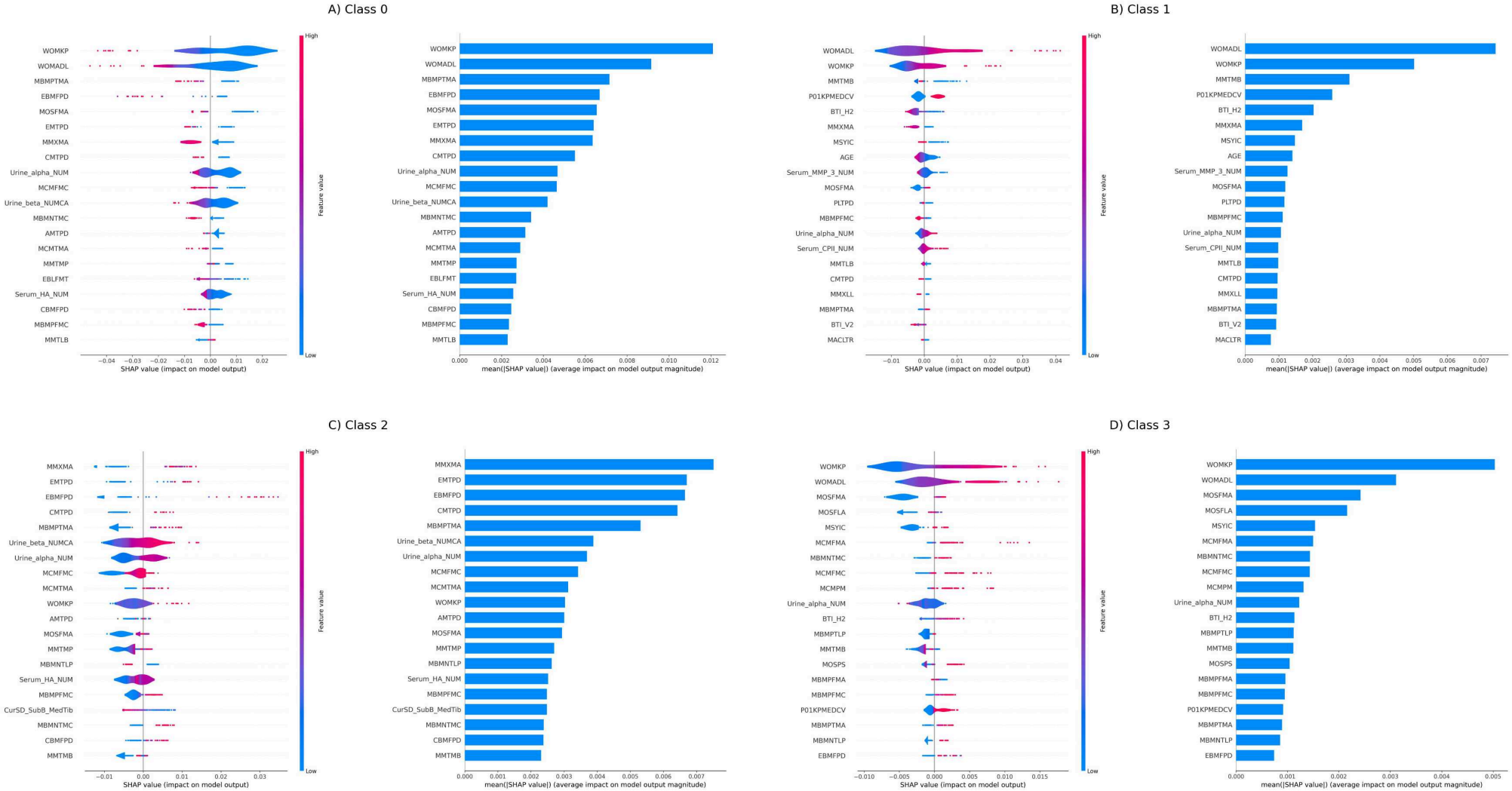
Supplemental Figure 7: Assessment of feature impact on MULTI-CLASS predictions for patients with AGE ≥ 60 in the hold-out set. Assessment of feature impact on non-progression (Class 0, **panel A**), pain-only progression (Class 1, **panel B**), radiographic progression (Class 2, **panel C**) and both pain and radiographic progression (Class 3, **panel D**), using "Kernel-SHAP" for multi-class predictions with model AP5_mu. **Left** – impact distribution of the most important features. The colour represents the feature value (red = high, blue = low). A positive SHAP value represents a positive impact on class prediction. **Right** – average impact magnitude of the most important features on class prediction.



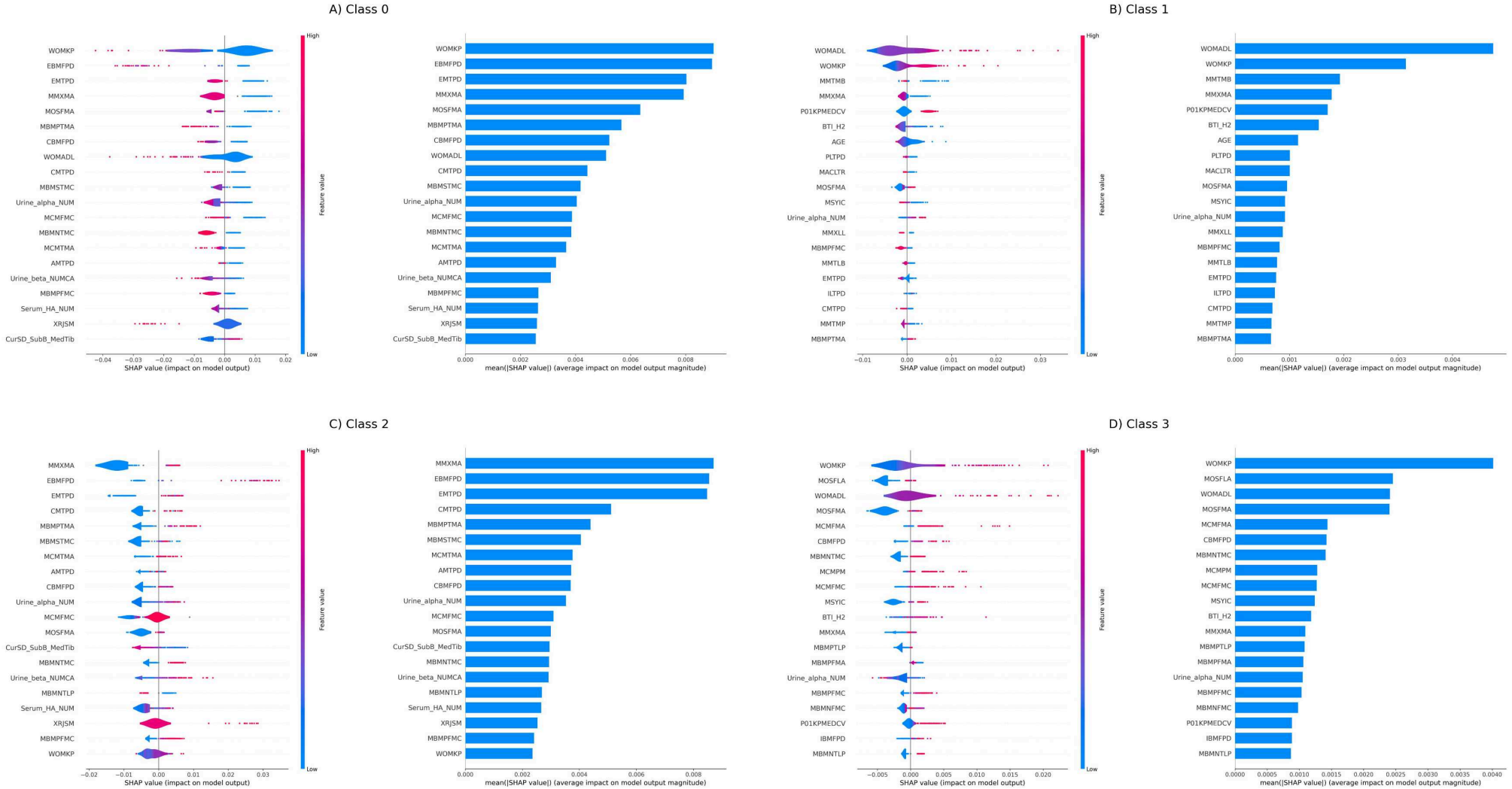
Supplemental Figure 8: Assessment of feature impact on MULTI-CLASS predictions for MALE patients in the hold-out set. Assessment of feature impact on non-progression (Class 0, **panel A**), pain-only progression (Class 1, **panel B**), radiographic progression (Class 2, **panel C**) and both pain and radiographic progression (Class 3, **panel D**), using "Kernel-SHAP" for multi-class predictions with model AP5_mu. **Left** – impact distribution of the most important features. The colour represents the feature value (red = high, blue = low). A positive SHAP value represents a positive impact on class prediction. **Right** – average impact magnitude of the most important features on class prediction.



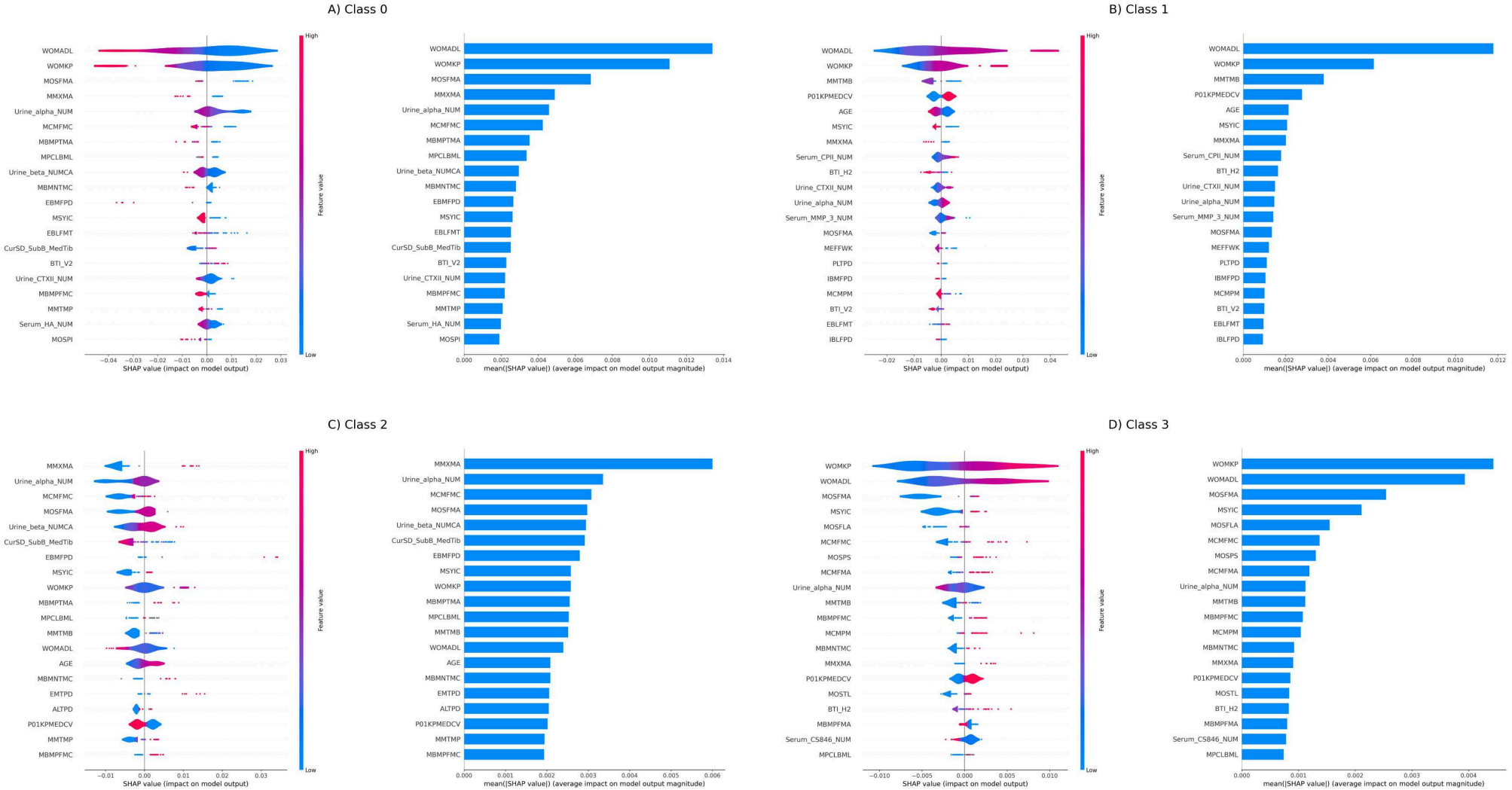
Supplemental Figure 9: Assessment of feature impact on MULTI-CLASS predictions for FEMALE patients in the hold-out set. Assessment of feature impact on non-progression (Class 0, **panel A**), pain-only progression (Class 1, **panel B**), radiographic progression (Class 2, **panel C**) and both pain and radiographic progression (Class 3, **panel D**), using "Kernel-SHAP" for multi-class predictions with model AP5_mu. **Left** – impact distribution of the most important features. The colour represents the feature value (red = high, blue = low). A positive SHAP value represents a positive impact on class prediction. **Right** – average impact magnitude of the most important features on class prediction.



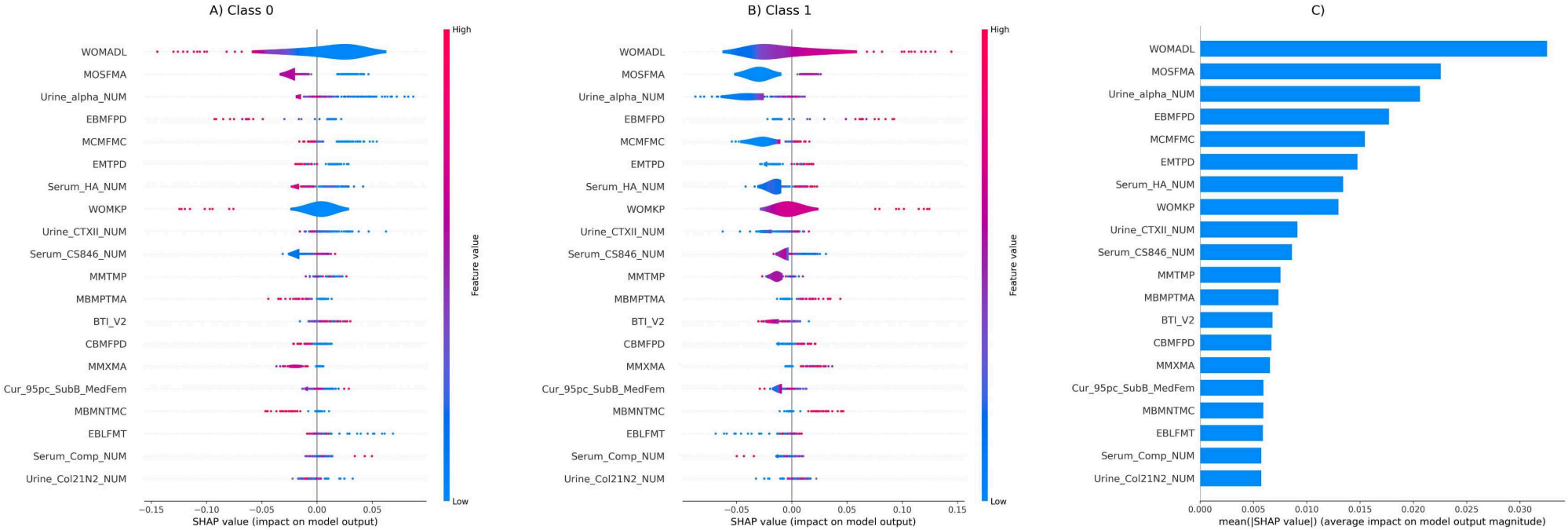
Supplemental Figure 10: Assessment of feature impact on MULTI-CLASS predictions for patients of WHITE ethnicity in the hold-out set. Assessment of feature impact on non-progression (Class 0, **panel A**), pain-only progression (Class 1, **panel B**), radiographic progression (Class 2, **panel C**) and both pain and radiographic progression (Class 3, **panel D**), using "Kernel-SHAP" for multi-class predictions with model AP5_mu. **Left** – impact distribution of the most important features. The colour represents the feature value (red = high, blue = low). A positive SHAP value represents a positive impact on class prediction. **Right** – average impact magnitude of the most important features on class prediction.



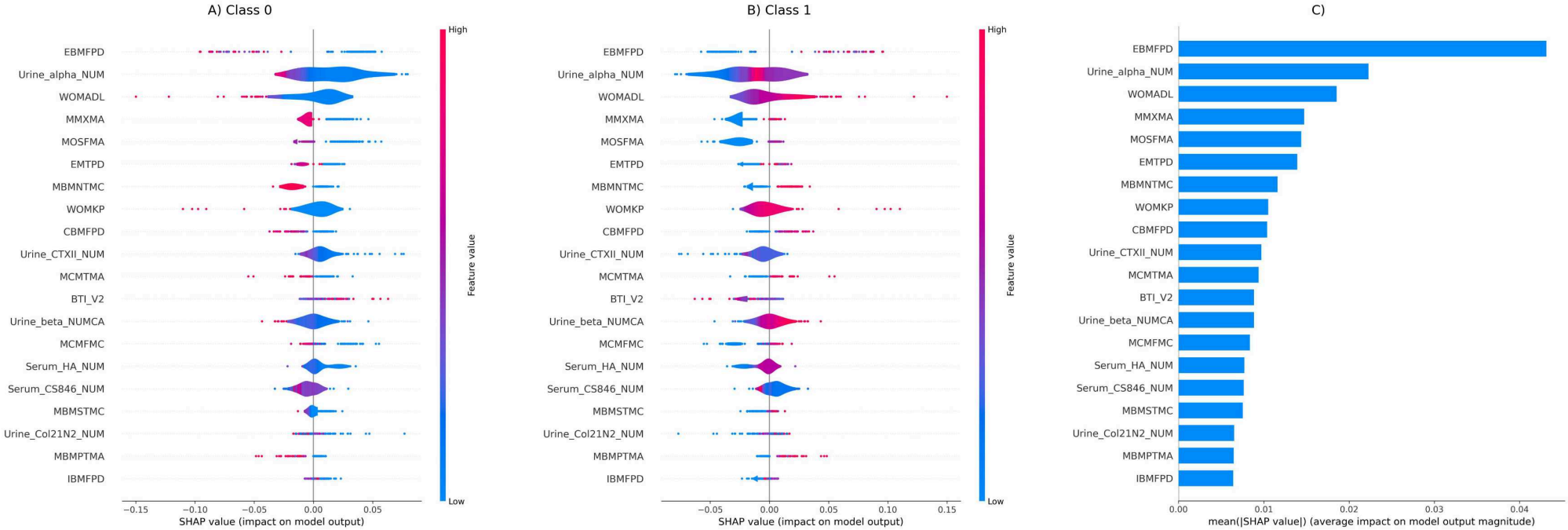
Supplemental Figure 11: Assessment of feature impact on MULTI-CLASS predictions for patients of BLACK ethnicity in the hold-out set. Assessment of feature impact on non-progression (Class 0, **panel A**), pain-only progression (Class 1, **panel B**), radiographic progression (Class 2, **panel C**) and both pain and radiographic progression (Class 3, **panel D**), using "Kernel-SHAP" for multi-class predictions with model AP5_mu. **Left** – impact distribution of the most important features. The colour represents the feature value (red = high, blue = low). A positive SHAP value represents a positive impact on class prediction. **Right** – average impact magnitude of the most important features on class prediction.



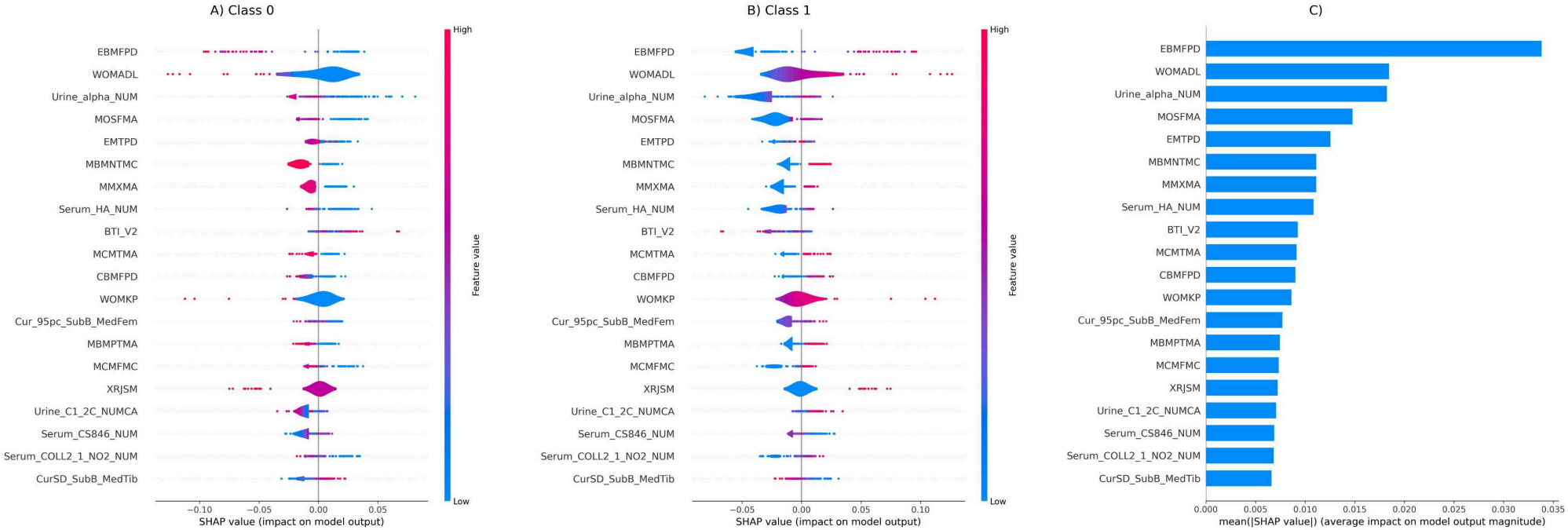
Supplemental Figure 12: Assessment of feature impact on BINARY predictions for patients with AGE < 60 in the hold-out set. Assessment of feature impact on non-progression (Class 0) and progression (Class 1), using "Kernel-SHAP" for binary predictions with model AP5_bi. **Left and Middle** – impact distribution of the most important features for Class 0 (Left) and Class 1 (Middle). The colour represents the feature value (red = high, blue = low). A positive SHAP value represents a positive impact on class prediction. **Right** – average impact magnitude of the most important features on class prediction.



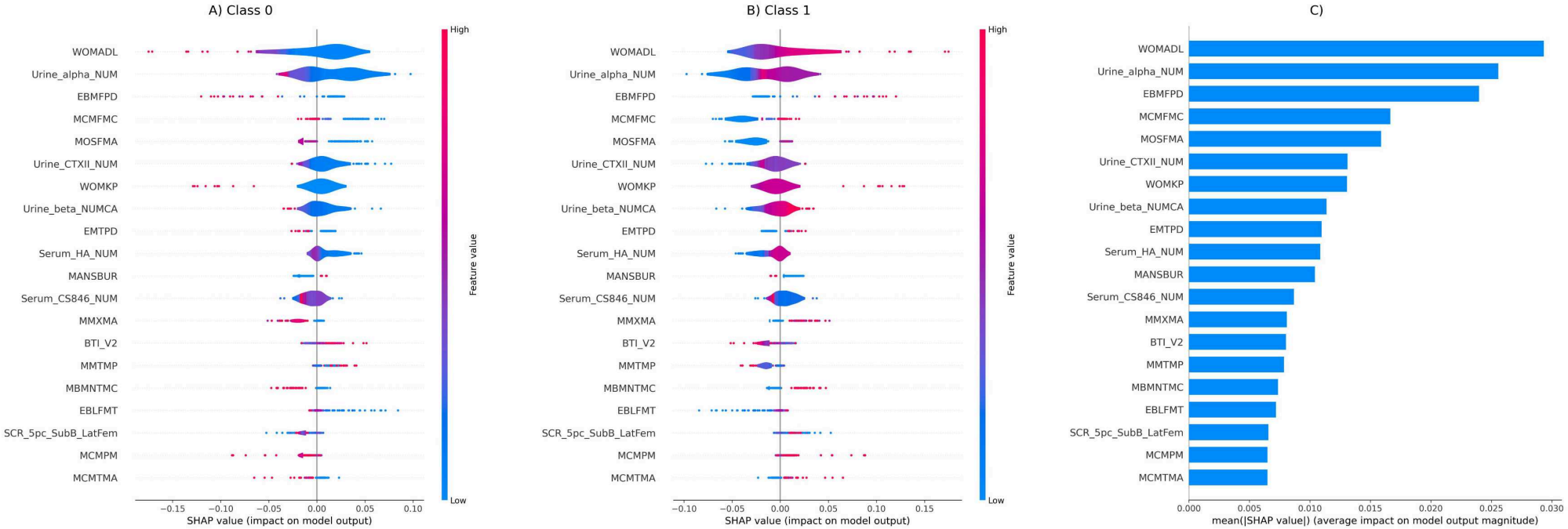
Supplemental Figure 13: Assessment of feature impact on BINARY predictions for patients with AGE ≥ 60 in the hold-out set. Assessment of feature impact on non-progression (Class 0) and progression (Class 1), using "Kernel-SHAP" for binary predictions with model AP5_bi. **Left and Middle** – impact distribution of the most important features for Class 0 (Left) and Class 1 (Middle). The colour represents the feature value (red = high, blue = low). A positive SHAP value represents a positive impact on class prediction. **Right** – average impact magnitude of the most important features on class prediction.



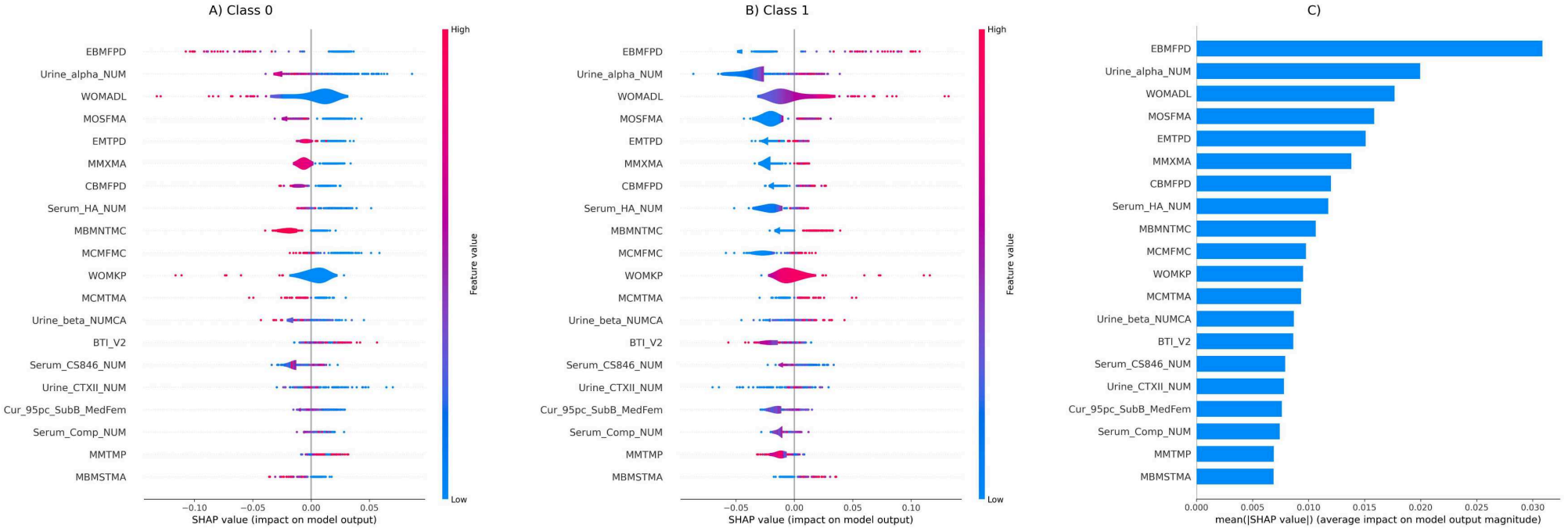
Supplemental Figure 14: Assessment of feature impact on BINARY predictions for MALE patients in the hold-out set. Assessment of feature impact on non-progression (Class 0) and progression (Class 1), using "Kernel-SHAP" for binary predictions with model AP5_bi. **Left and Middle** – impact distribution of the most important features for Class 0 (Left) and Class 1 (Middle). The colour represents the feature value (red = high, blue = low). A positive SHAP value represents a positive impact on class prediction. **Right** – average impact magnitude of the most important features on class prediction.



Supplemental Figure 15: Assessment of feature impact on BINARY predictions for FEMALE patients in the hold-out set. Assessment of feature impact on non-progression (Class 0) and progression (Class 1), using "Kernel-SHAP" for binary predictions with model AP5_bi. **Left and Middle** – impact distribution of the most important features for Class 0 (Left) and Class 1 (Middle). The colour represents the feature value (red = high, blue = low). A positive SHAP value represents a positive impact on class prediction. **Right** – average impact magnitude of the most important features on class prediction.



Supplemental Figure 16: Assessment of feature impact on BINARY predictions for patients of WHITE ethnicity in the hold-out set. Assessment of feature impact on non-progression (Class 0) and progression (Class 1), using "Kernel-SHAP" for binary predictions with model AP5_bi. **Left and Middle** – impact distribution of the most important features for Class 0 (Left) and Class 1 (Middle). The colour represents the feature value (red = high, blue = low). A positive SHAP value represents a positive impact on class prediction. **Right** – average impact magnitude of the most important features on class prediction.



Supplemental Figure 17: Assessment of feature impact on BINARY predictions for patients of BLACK ethnicity in the hold-out set. Assessment of feature impact on non-progression (Class 0) and progression (Class 1), using "Kernel-SHAP" for binary predictions with model AP5_bi. **Left and Middle** – impact distribution of the most important features for Class 0 (Left) and Class 1 (Middle). The colour represents the feature value (red = high, blue = low). A positive SHAP value represents a positive impact on class prediction. **Right** – average impact magnitude of the most important features on class prediction.