RESEARCH ARTICLE

# Socioeconomic status, mental health, and nutrition are the principal traits for low back pain phenotyping: Data from the osteoarthritis initiative

ZeYu Huang[1,2] | Weihua Guo[3] | John T. Martin[2,4] 

[1]Department of Orthopaedic Surgery, West China Hospital, West China Medical School, SiChuan University, ChengDu, SiChuan Province, People's Republic of China

[2]Department of Orthopaedic Surgery, School of Medicine, Duke University, Durham, North Carolina, USA

[3]Department of Immuno-Oncology, City of Hope Comprehensive Cancer Center, Duarte, California, USA

[4]Department of Orthopedic Surgery, Rush University Medical Center, Chicago, Illinois, USA

**Correspondence**
John T. Martin, Rush University Medical Center, Department of Orthopedic Surgery, Chicago, IL, USA.
Email: john_martin@rush.edu

Weihua Guo, Department of Immuno-Oncology, City of Hope Comprehensive Cancer Center, Duarte, CA, USA.
Email: wguo@coh.org

## Abstract

**Background:** Low back pain (LBP) is a heterogeneous disease with biological, physical, and psychosocial etiologies. Models for predicting LBP severity and chronicity have not made a clinical impact, perhaps due to difficulty deciphering multidimensional phenotypes. In this study, our objective was to develop a computational framework to comprehensively screen metrics related to LBP severity and chronicity and identify the most influential.

**Methods:** We identified individuals from the observational, longitudinal Osteoarthritis Initiative cohort ($N = 4796$) who reported LBP at enrollment ($N = 215$). OAI descriptor variables ($N = 1190$) were used to cluster individuals via unsupervised learning and uncover latent LBP phenotypes. We also developed a dimensionality reduction algorithm to visualize clusters/phenotypes using Uniform Manifold Approximation and Projection (UMAP). Next, to predict chronicity, we identified those with acute LBP ($N = 40$) and persistent LBP over 8 years of follow-up ($N = 66$) and built logistic regression and supervised machine learning models.

**Results:** We identified three LBP phenotypes: a "high socioeconomic status, low pain severity group", a "low socioeconomic status, high pain severity group", and an intermediate group. Mental health and nutrition were also key clustering variables, while traditional biomedical factors (e.g., age, sex, BMI) were not. Those who developed chronic LBP were differentiated by higher pain interference and lower alcohol consumption (a correlate to poor physical fitness and lower soceioeconomic status). All models for predicting chronicity had satisfactory performance (accuracy 76%–78%).

**Conclusions:** We developed a computational pipeline capable of screening hundreds of variables and visualizing LBP cohorts. We found that socioeconomic status, mental health, nutrition, and pain interference were more influential in LBP than traditional biomedical descriptors like age, sex, and BMI.

**KEYWORDS**
alcohol, low back pain, machine learning, mental health, socioeconomic status

# 1 | INTRODUCTION

Low back pain (LBP) is one of the most common complaints from people of all ages,[1,2] affecting approximately 540 million annually (7.3% prevalence).[3] LBP is a heterogeneous disease with biological, physical, and psychosocial etiologies[4] and the underlying combination of factors that drive pain (i.e., the pain phenotype) cannot be identified in 90% of those who seek care.[5] Furthermore, while most LBP will resolve acutely, 30% of LBP patients have symptoms that become chronic.[6] Models to predict which patients transition from acute to chronic LBP have not been adopted into daily clinical practice, perhaps due to difficulty in deciphering multidimensional pain phenotypes.

Individual risk factors for LBP include physical[7] and mental health history,[7] lifestyle factors like diet and exercise,[8] genetics, socioeconomic factors, comorbidities[9] and others. Traditionally, risk factors are assessed in cohort studies where regression is used to determine how specific factors affect LBP prevalence and progression.[9] Cohort sizes necessarily scale with the number of risk factors to achieve the appropriate statistical power, limiting the extent of a potential study. To evaluate a larger variable set in a well-controlled prospective study, researchers have organized cohorts into specific sub-groups based on symptom temporality and severity using clustering,[10] principal components,[11] or latent class analyses.[7] These efforts have identified predictors for severe disease such as poor physical function, depression, and older age.[12] However, for a heterogeneous, multidimensional disease like LBP, a comprehensive screen of personal attributes (mental and physical health history, lifestyle factors, demographics and socioeconomics, anthropometry, comorbidities, genetics, spine health, etc.) is necessary to develop a predictive model. Thus, a clinical tool that simplifies pain phenotypes by identifying the most dominant factors in pain progression remains elusive.

Machine learning is a powerful analytic tool that recognizes patterns in data without explicit programming.[13] Unlike traditional regression, machine learning can identify complex and nonlinear relationships between factors to identify phenotypes in an unbiased manner and generate sophisticated predictions.[14,15] Predictive machine learning models can help providers identify risk in patients whose constellation of factors may have otherwise been unidentified or identify novel and unexpected predictors of disease.[16,17] Machine learning has been applied in many clinical scenarios to determine disease risk, complications, and survival outcomes.[18,19] A computational tool that leverages the benefits of machine learning could be integrated into an electronic health record to make point-of-care predictions of LBP risk.

In this study, our objective was to use machine learning to identify multidimensional LBP phenotypes and predict pain chronicity in the Osteoarthritis Initiative (OAI) dataset. The OAI is a longitudinal, observational cohort study of musculoskeletal health with 4796 enrollees. We chose the OAI for this analysis because it has a comprehensive set of descriptor variables relevant to musculoskeletal health (1000+ metrics, including demographics, anthropometry, diet, physical activity, mental and physical health, medical history, socioeconomics, and medication use) and 8 years of longitudinal LBP data. We identified participants that reported LBP at enrollment and used unsupervised learning (clustering) to establish pain phenotypes with no *a priori* hypothesis as to which variables were most influential. In doing so, we *developed a computational pipeline using* a dimensionality reduction method typical in biomedical sciences (Uniform Manifold Approximation and Projection, UMAP)[20] to visualize clinical datasets. we established a technique for visualizing clinical datasets. We next identified participants with acute or chronic pain and performed supervised learning to develop a model to predict pan chronicity. We compared three common supervised learning models [random forest (RF), support vector machine (SVM), artificial neural network (ANN)] to traditional logistic regression (LR) to determine the most effective algorithm.

# 2 | METHODS

## 2.1 | Unsupervised learning and dimensionality reduction for identifying back pain phenotypes

### 2.1.1 | Cohort identification

The OAI is a longitudinal cohort study initiated in 2004 that recruited men and women aged 45–79 years from four clinical sites in the United States. The OAI dataset includes 4796 subjects and 1190 variables at enrollment with 8 years follow-up (Figure 1A,B). Subjects who responded to "How many days were you limited by back pain in the last 30 days?" with ≥14 days and indicated that pain was located in their low back were included for evaluation ($n = 223$). To reduce noise in the clinical data, an initial data cleaning process was performed to remove: (1) participants with <80% of variables completed ($n = 8$); (2) variables focused on knee health with no analogue in another joint ($n = 455$); (3) variables completed by <80% of participants ($n = 232$); (4) variables with zero variance ($n = 6$). A total of 215 participants and 485 relevant variables were left for further evaluation (Table 1; and Tables S1–S3).

### 2.1.2 | Clustering and dimensionality reduction

Data were then processed by unsupervised clustering to identify LBP phenotypes. First, we generated a correlation matrix and identified highly correlated variables in R v4.0.2 (mixedCor, psych package [v2.1.6]; findCorrelation, caret package [v6.0-88]). After excluding correlated variables with r > 0.7, 297 were left for analysis. Next, a dissimilarity matrix was calculated using the Gower distance (daisy, cluster package [v2.1.2]) and k-medoids clustering was performed (pam, cluster package; distance metric = Euclidean) using the silhouette width to optimize cluster number. To visualize clusters and their phenotypes, UMAP (umap, umap package [v0.2.7.0]; nearest neighbors = 5, minimum distance = 0.7) was used to reduce the dimensions of the dataset from 215 subjects and 297 variables to 215 subjects and 2 UMAP variables. UMAP-1 and UMAP-2 are a nonlinear combination of the 297 variables that preserve information
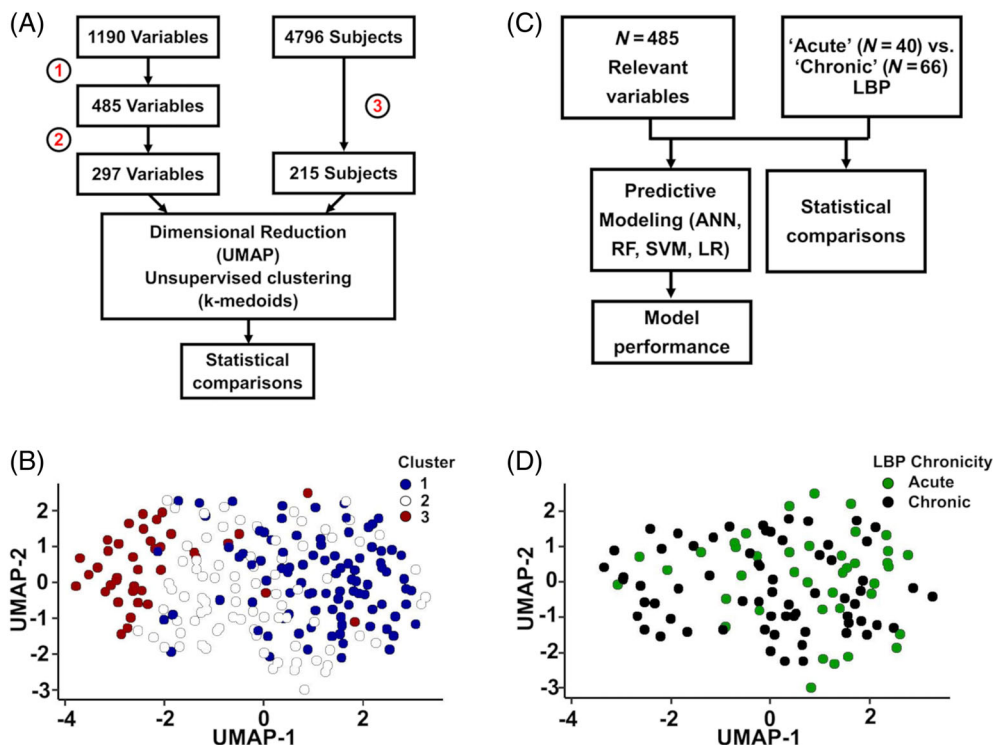
**FIGURE 1** Study design and data visualization. (A) Study I: Data analysis flowchart to identify LBP phenotypes in individuals from the OAI. (1) Knee-specific ($n = 455$), incomplete (completed by <80% participants, $n = 232$), and low variance (variance $= 0$, $n = 6$) variables were eliminated. (2) Highly correlated variables were removed ($r \geq 0.7$, $n = 200$). (3) Individuals with $\geq14$ days of LBP in past month were selected ($n = 223$) and filtered for completeness (>80% of variables completed, $n = 215$). (B) Data visualization (UMAP) with clusters overlaid. Each data point represents an individual, axes represent the UMAP-1 and UMAP-2 variables (a nonlinear combination of hundreds of OAI metrics), and the relative distance between data points represents the dissimilarity between individuals. The color of the data points indicates the cluster an individual was assigned to. (C) Study II: Data analysis flowchart for acute and chronic LBP individuals. (D) Data visualization (UMAP) with LBP chronicity overlaid. Each data point represents an individual, axes represent the UMAP-1 and UMAP-2 variables, and the relative distance between data points represents the dissimilarity between individuals. The color of the data points indicates whether an individual had acute or chronic back pain. ANN, artificial neural network; LBP, low back pain; LR, logistic regression; OAI, osteoarthritis initiative; RF, random forest; SVM, support vector machines; UMAP, uniform manifold approximation and projection.

contained in the 297 variables. UMAP plots were generated (Figure 1) in which axes represent the UMAP-1 and UMAP-2 variables, each data point represents one of the 215 individuals, and the relative distance between data points represents the dissimilarity between individuals. The color of the data points can be manipulated to illustrate a characteristic of an individual (e.g., the cluster the individual belongs to, whether the individual has acute or chronic pain, or a metric that describes the individual like annual income, race, mental health score, etc.).

## 2.1.3 | Statistical analysis

Statistical comparisons were made to identify which variables ($n = 485$) differentiated the clusters. Continuous and ordinal variables were evaluated using the Kruskal–Wallis test (kruskal.test, stats package [v4.2.0]), while categorical variables were evaluated using Fisher's exact test (fisher.test, stats package). Because clustering by its nature maximizes differences between groups, we used a

conservative approach to determine significant variables and then further screened variables based on their effect size. We defined significance with a Bonferroni correction ($p < 0.05/485$) and set effect size cutoffs for continuous, ordinal, and categorical variables. For continuous and ordinal variables, the maximum fold change between groups was calculated and a 1.5-fold increase or decrease was used to define relevant variables. Cramer's V was calculated for categorical variables and effect sizes were defined as: "small" 0.1–0.3, "medium" 0.3–0.5, or "large" > 0.5.

Finally, to develop a clinical tool that provides LBP hazard ratios on an individual basis, we performed a Kaplan–Meier recurrent event survival analysis (survfit, survival package [v3.2-11]) and a Cox proportional hazard ratio analysis (coxph, survival package) to calculate survival curves and hazard ratios for each cluster. We defined an event as any visit at the 1, 2, 3, 4, 6, or 8-year follow-up in which an individual reported $\geq14$ days of LBP in the past 30 days (follow-up time mean [range]: 1.1 years [0.9, 1.6], 2.0 years [1.8, 2.4], 3.1 years [2.8, 3.4], 4.0 years [3.7, 4.4], 6.0 years [5.1, 6.7], 8.0 years [7.7, 8.7]).

**TABLE 1** Demographics of individuals with LBP in the OAI dataset

| | | LBP clusters | | | | Control | Low back pain type | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 (N = 94) | 2 (N = 81) | 3 (N = 40) | All Clusters (N = 215) | No LBP (N = 4581) | Acute (N = 40) | Chronic (N = 66) | All (N = 106) |
| Age (years) | | 61.6 (9.7) | 60.3 (9.2) | 56.8 (8.2) | 60.2 (9.3) | 61.2 (10.3) | 59.5 (9.4) | 59.2 (8.5) | 59.3 (6.6) |
| Sex | Female | 48 (51.1%) | 52 (64.2%) | 26 (65%) | 126 (58.6%) | 2678 (58.5%) | 21 (52.5%) | 37 (56.1%) | 58 (54.7%) |
| | Male | 46 (48.9%) | 29 (35.8%) | 14 (35%) | 89 (41.4%) | 1903 (41.5%) | 19 (47.5%) | 29 (43.9%) | 48 (45.3%) |
| Race | Other Non-White | 1 (1.1%) | 1 (1.2%) | 1 (2.5%) | 3 (1.4%) | 79 (1.7%) | 1 (2.5%) | 2 (3.0%) | 3 (2.8%) |
| | White or Caucasian | 81 (86.2%) | 44 (54.3%) | 15 (37.5%) | 140 (65.1%) | 3650 (79.7%) | 30 (75%) | 41 (62.2%) | 71 (67.0%) |
| | Black or African American | 10 (10.6%) | 35 (43.3%) | 24 (60%) | 69 (32.1%) | 805 (17.6%) | 8 (20%) | 22 (33.3%) | 30 (28.3%) |
| | Asian | 1 (1.1%) | 0 (0%) | 0 (0%) | 1 (0.5%) | 44 (1.0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| | No response | 1 (1%) | 1 (1.2%) | 0 (0%) | 2 (0.9%) | 3 (0%) | 1 (2.5%) | 1 (1.5%) | 2 (1.9%) |
| BMI (kg/m$^2$) | | 29.4 (4.9) | 31.5 (5.1) | 33.2 (4.2) | 30.8 (5.2) | 28.5 (5.2) | 28.5 (3.6) | 31.6 (5.6) | 30.4 (5.0) |
| Education | Less than high school graduate | 0 (0%) | 7 (8.7%) | 10 (25%) | 17 (7.9%) | 151 (3.3%) | 1 (2.5%) | 4 (6.0%) | 5 (4.7%) |
| | High school graduate | 11 (11.7%) | 11 (13.6%) | 9 (22.5%) | 31 (14.4%) | 576 (12.6%) | 6 (15%) | 8 (12.1%) | 14 (13.2%) |
| | Some college | 23 (24.5%) | 35 (43.2%) | 15 (37.5%) | 73 (34.0%) | 1073 (23.4%) | 8 (20%) | 25 (37.9%) | 33 (31.1%) |
| | College graduate | 19 (20.2%) | 12 (14.8%) | 5 (12.5%) | 36 (16.7%) | 965 (21.1%) | 11 (27.5%) | 8 (12.1%) | 19 (18.0%) |
| | Some graduate school | 5 (5.3%) | 6 (7.4%) | 0 (0%) | 11 (5.1%) | 386 (8.4%) | 2 (5%) | 5 (7.6%) | 7 (6.6%) |
| | Graduate degree | 36 (38.3%) | 10 (12.3%) | 0 (0%) | 46 (21.4%) | 1390 (30.3%) | 12 (30%) | 16 (24.3%) | 28 (26.4%) |
| | No response | 0 (0%) | 0 (0%) | 1 (2.5%) | 1 (0.5%) | 40 (0.9%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Income | <$10 K | 4 (4.3%) | 4 (4.9%) | 14 (35%) | 22 (10.2%) | 138 (3.0%) | 1 (2.5%) | 9 (13.6%) | 10 (9.4%) |
| | $10 K to <$25 K | 4 (4.3%) | 16 (19.8%) | 11 (27.5%) | 31 (14.4%) | 423 (9.2%) | 2 (5%) | 13 (19.7%) | 15 (14.2%) |
| | $25 K to $50 K | 29 (30.9%) | 36 (44.4%) | 7 (17.5%) | 72 (33.5%) | 1063 (23.2%) | 15 (37.5%) | 16 (24.2%) | 31 (29.2%) |
| | $50 K to <$100 K | 32 (34.0%) | 12 (14.8%) | 1 (2.5%) | 45 (20.9%) | 1565 (34.2%) | 13 (32.5%) | 14 (21.2%) | 27 (25.5%) |
| | >$100 K | 13 (13.8%) | 7 (8.7%) | 1 (2.5%) | 21 (9.8%) | 1054 (23.0%) | 5 (12.5%) | 9 (13.6%) | 14 (13.2%) |
| | No response | 12 (12.7%) | 6 (7.4%) | 6 (15%) | 24 (11.2%) | 338 (7.4%) | 4 (10%) | 5 (7.7%) | 9 (8.5%) |
| Employment | Not working other reasons | 26 (27.7%) | 18 (22.2%) | 5 (12.5%) | 49 (22.8%) | 1481 (32.3%) | 8 (20%) | 10 (15.2%) | 18 (17.0%) |
| | Not working in part due to health | 12 (12.8%) | 22 (27.2%) | 24 (60%) | 58 (27.0%) | 187 (4.1%) | 3 (7.5%) | 26 (39.4%) | 29 (27.4%) |
| | Unpaid work for family business | 2 (2.0%) | 0 (0%) | 0 (0%) | 2 (0.9%) | 52 (1.1%) | 0 (0%) | 0 (0%) | 0 (0%) |
| | Works for pay | 53 (56.4%) | 41 (50.6%) | 11 (27.5%) | 105 (48.8%) | 2838 (62.0%) | 29 (72.5%) | 30 (45.4%) | 59 (55.6%) |
| | No response | 1 (1.1%) | 0 (0%) | 0 (0%) | 1 (0.5%) | 23 (0.5%) | 0 (0%) | 0 (0%) | 0 (0%) |
| Symptomatic knee OA | No | 52 (55.3%) | 48 (59.3%) | 15 (37.5%) | 115 (53.5%) | 3289 (71.8%) | 27 (67.5%) | 34 (51.5%) | 61 (57.5%) |
| | Yes | 42 (44.7%) | 33 (40.7%) | 25 (62.5%) | 100 (46.5%) | 1291 (28.2%) | 13 (32.5%) | 32 (48.5%) | 45 (42.5%) |
| Radiographic knee OA | No | 38 (40.4%) | 42 (51.9%) | 15 (37.5%) | 95 (44.2%) | 2021 (44.1%) | 21 (52.5%) | 28 (42.4%) | 49 (46.2%) |
| | Yes | 56 (59.6%) | 39 (48.2%) | 25 (62.5%) | 120 (55.8%) | 2559 (55.9%) | 19 (47.5%) | 38 (57.5%) | 57 (53.8%) |

Abbreviations: BMI, body mass index; LBP, low back pain; OAI, osteoarthritis initiative.

FIGURE 2  Unsupervised learning identifies three LBP phenotypes with varying pain severity. (A) Days with activity-limiting LBP at enrollment (screening variable). There were no significant differences between clusters. (B) LBP severity at enrollment. Cluster 3 has the highest proportion with severe LBP, followed by Cluster 2 and then Cluster 1. (C) Survival analysis over the 8-year follow-up period with an episode of LBP as the survival event. There were no differences between clusters. (D) Survival analysis over the 8-year follow-up period with an episode of severe LBP as the survival event. Cluster 3 had the worst survival outcomes. LBP, low back pain.
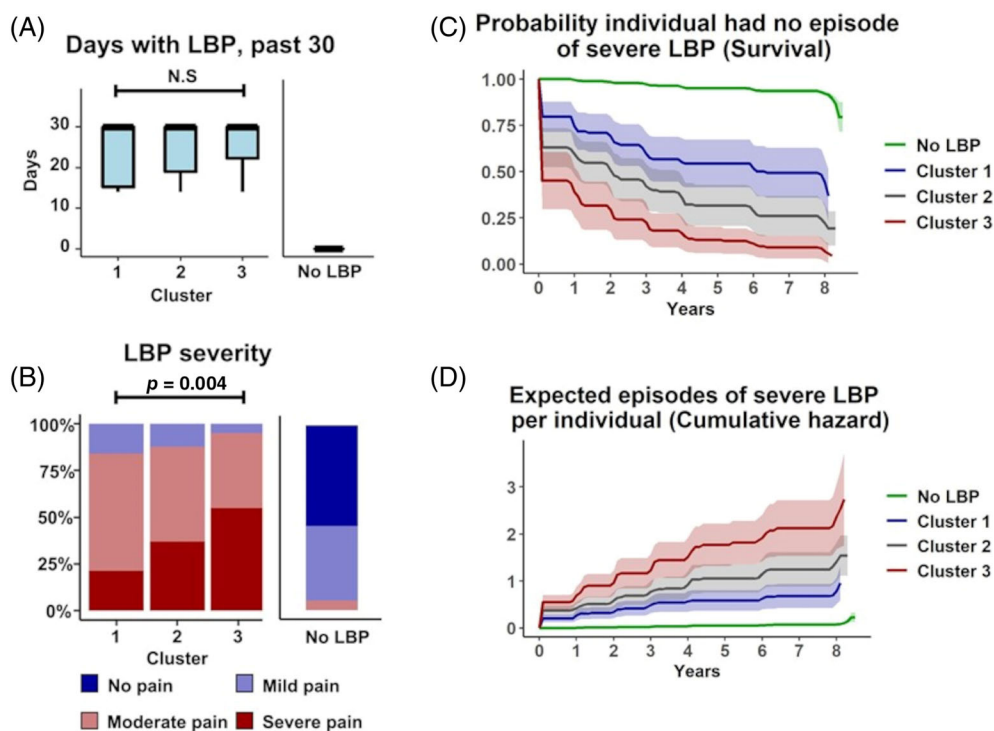


## 2.2 | Supervised learning for predicting back pain chronicity

### 2.2.1 | Cohort identification

Of the 215 individuals with LBP, we identified 40 with acute pain (≥14 days in the past month with activity-limiting LBP at enrollment and <14 days at six subsequent visits) and 66 with chronic pain (≥14 days with activity-limiting LBP at enrollment and in at least three of six visits between enrollment and the 8-year follow-up) (Figure 1C,D, Table 1). The remaining individuals were excluded; 37 could not be classified as either acute or chronic and 72 who were lost to follow-up.

### 2.2.2 | Predictive modeling

To develop a model that predicts pain chronicity, traditional LR was compared to three common supervised learning models: RF, SVM, ANN. In each case, the model output was a classification of whether an individual's pain would be acute or chronic. The model input consisted of an optimal subset of predictor variables identified through a feature elimination approach. Optimized models incorporated three to seven predictor variables (the best predictors of 485 total variables) and one output variable (acute vs. chronic classification).

To construct the LR model, the acute/chronic pain condition served as the output variable and an optimal set of predictor variables was chosen by minimizing the Akaike information criterion (AIC) (stepAIC, MASS package [v7.3-54]). Then, 100-fold cross validation with iterated shuffling (75/25 training/testing split) was used to calculate model performance

metrics: prediction accuracy, area under the receiver operating characteristic curve (AUC), sensitivity, and specificity.

Supervised learning models were constructed via RF (randomForest, randomForest package [v 4.6–14]) and SVM (svm, e1071 package [v 1.7-7]). In both cases, the top 10 variables based on the LR AIC analysis were chosen as candidates for predictive modeling and a grid search was performed to simultaneously optimize the variable set and model hyperparameters (optimal sets: RF, number of trees = 500, variables per node = 1; SVM, kernel = radial, cost = 1.8, gamma = 0.3). The output variable was the acute/chronic pain condition. Model performance metrics were then calculated using the cross-validation procedure as described above.

Next, we constructed an ANN composed of one input layer, two hidden layers, and one output layer with dropout and weight regularization (keras package [v2.4.0]). The model output was the acute/chronic pain condition, and a random search was used to simultaneously optimize the variable set (<10, a subset from the top 10 variables identified in LR AIC analysis) and model hyperparameters (optimal set: nodes/dropout/activation function at layer 1, 40/30%/relu; layer 2, 25/50%/relu; learning rate = 0.003). Model performance metrics were calculated as described above. In all predictive models, missing data was imputed using a RF approach (rfImpute, randomForest package).

## 3 | RESULTS

## 3.1 | Unsupervised learning identifies three LBP phenotypes with varying pain severity and chronicity

We identified 3 clusters (3 LBP phenotypes) through unsupervised learning (Figure 1B, Table 1), where an individual's cluster assignment
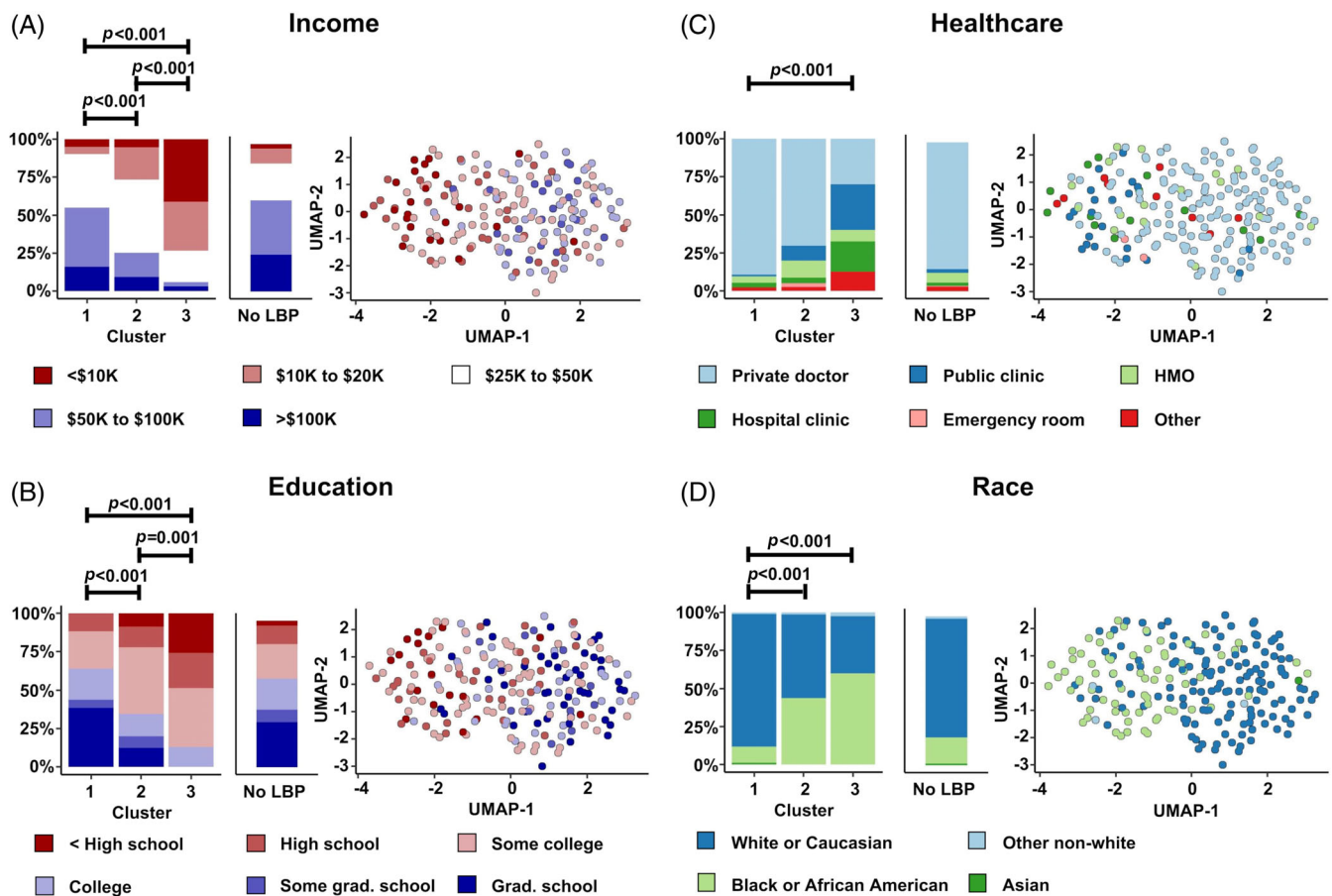
**FIGURE 3** Socioeconomic status and race distinguish clusters. Cluster 3 had the lowest socioeconomic status and were majority Black or African American, while Cluster 1 had the highest socioeconomic status and were majority white or Caucasian. Cluster 2 was an intermediate. (A) Annual income (V00INCOME). (B) Education (V00EDCV). (C) Type of healthcare (V00HLTHCAR). (D) Race (P02RACE). Grad., graduate; HMO, health maintenance organization; LBP, low back pain; UMAP, Uniform manifold approximation and projection.

was consistent in 89% ± 12% of validation runs (Figure S1). UMAP visualization revealed the greatest differences between Clusters 1 and 3, while Cluster 2 was an intermediate/transition phenotype (Figure 1B). Clusters were differentiated in terms of LBP severity and chronicity. Cluster 3 had the highest proportion reporting severe LBP at enrollment. Based on survival and hazard analyses, Cluster 3 was at the greatest risk for episodes of severe LBP over 8 years of follow-up. (Hazard Ratio [95% CI]: C1, 10.7 [7.4, 15.5]; C2, 18.0 [13.2, 24.5]; C3, 30.9 [22.6, 42.2]) (Figure 2, Table S4).

## 3.2 | Socioeconomic status, mental health, nutrition, and analgesic use are signature cluster markers

There were 58 variables that differentiated the LBP clusters by meeting both p-threshold and effect size criteria (Figure S2, Tables S1–S3). These variables reflected differences in socioeconomic status (SES), mental and physical health, nutrition, and analgesic use. We defined clusters as: *Cluster 1*, "high SES, low severity"; *Cluster 2*, "intermediate SES, intermediate severity;" *Cluster 3*, "low SES, high severity". Cluster 1 ("high SES, low

severity" group) had the least severe LBP phenotype and had higher income, higher education level, were more likely to access to private health care, were proportionally more white/Caucasian (Figure 3), and had better mental and physical health than the other clusters (Figure 4). Subjects from Cluster 2 ("intermediate SES, intermediate severity" group) had an intermediate LBP phenotype and had intermediate income, education level, healthcare access, and were racially mixed (Figure 3), with intermediate mental and physical health (Figure 4). They also consumed more healthy foods like fruits and vegetables and more vitamin supplements than the other clusters (Figure 5). Subjects from Cluster 3 ("low SES, high severity" group) had the most severe LBP phenotype and had low income, worse employment status, lower education level, were primarily Black/African American, and had worse mental and physical health. Additionally, Cluster 3 individuals had a diet with comparatively low nutritional value, with less vitamin supplements (Figure 5A–C), and along with those in Cluster 2, were more likely to use prescription analgesics (Figure 4B) and less likely to drink alcoholic beverages, including wine/wine coolers (Figure 5D). There were no differences in clusters by sex, height, weight, BMI, age, diagnoses of spine arthritis, family history of musculoskeletal health (knee/hip replacement), cardiovascular health, and comorbidities (Figure S3) among other metrics.
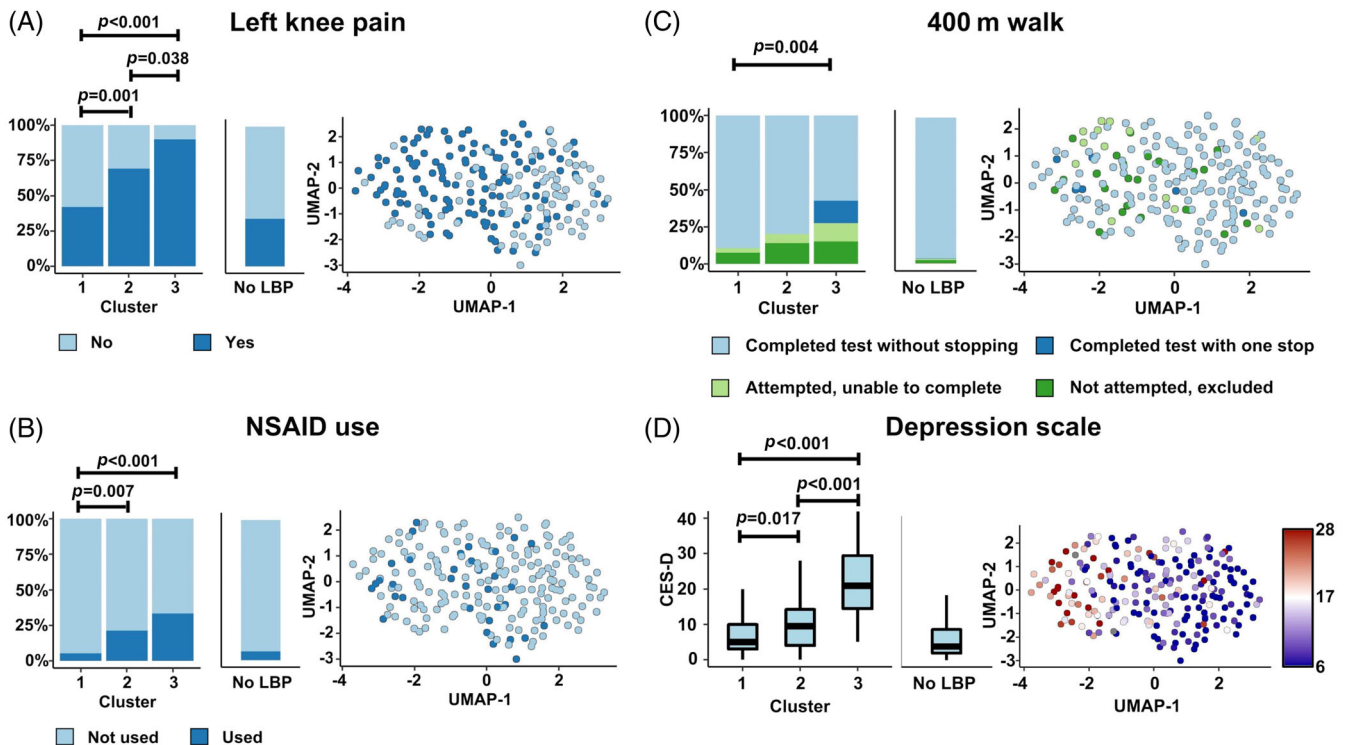
**FIGURE 4** Physical and mental health variables distinguish clusters. Cluster 3 had the worst physical and mental health, with Cluster 1 as the best and Cluster 2 as an intermediate. (A) Left knee pain, aching, or stiffness (P01KPL12CV). (B) NSAID use (V00NSAIDRX). (C) Completion of 400-meter walk (V00400MCMP). (D) Depression scale (Center for Epidemiology Studies) (V00CESD). LBP, low back pain; NSAID, nonsteroidal anti-inflammatory drug; UMAP, uniform manifold approximation and projection.
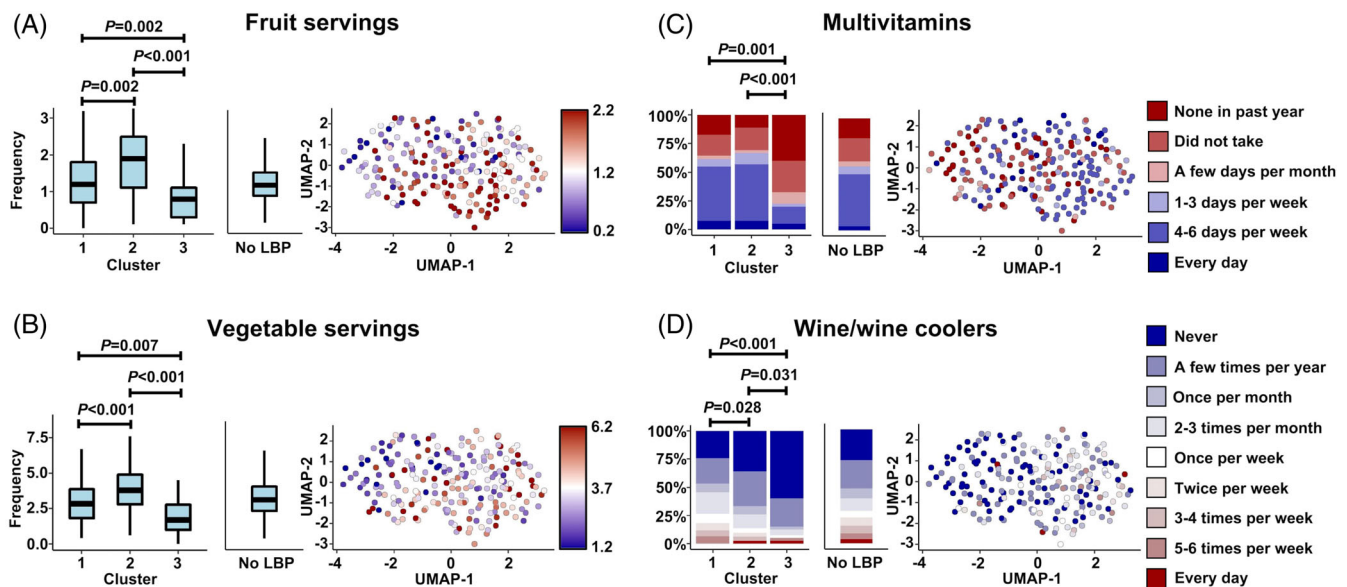


**FIGURE 5** Diet and nutrition variables distinguish clusters. Cluster 3 had the worst diet and nutrition compared to the other clusters but consumed the least alcohol (e.g., wine/wine coolers). (A) Daily servings of fruit (V00SRVFRT); (B) Daily servings of vegetables (V00SRVVEG); (C) Multivitamin use (V00MVIT1CV); (D) Wine/wine cooler consumption (V00FFQ71). LBP, low back pain; UMAP, Uniform manifold approximation and projection.

Of note, individuals in these clusters had a higher rate of symptomatic knee OA (46.5%) compared to individuals who did not meet the criteria for LBP (28.2%) (Table 1), while the prevalence of radiographic knee OA was comparable between clusters (55.8%) and those who did not meet the criteria for LBP (55.9%) (Table 1). Further, we found that subjects from Cluster
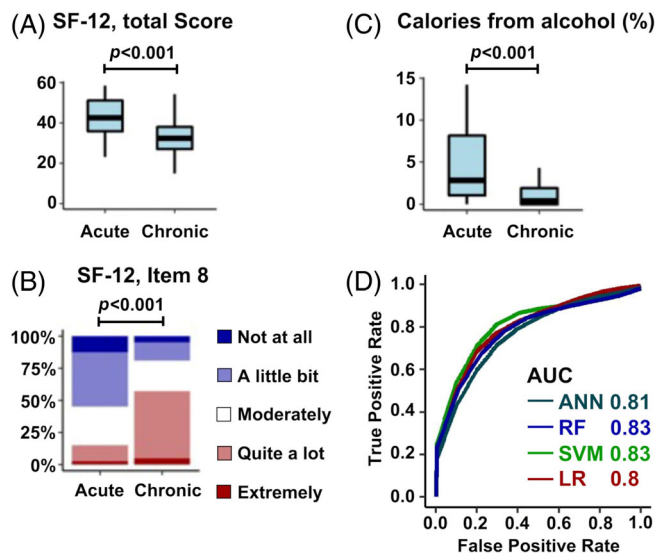
**FIGURE 6** Key variables that distinguished acute and chronic LBP groups. Three variables significantly different the groups (A) SF-12 total score (V00HSPSS). (B) Item 8 of SF-12: "During the past 4 weeks, how much did pain interfere with your normal work?" (V00SF8) (C) Daily calories from alcohol (% total daily calories) (V00PCTALCH). (D) ROC curve comparisons of ANN, RF, SVM and LR models for predicting pain chronicity. There was a mild improvement in performance with machine learning. ANN, artificial neural network; AUC, area under curve; LBP, low back pain; LR, logistic regression; RF, random forest; ROC, receiver operating characteristic curve; SF-12, short form-12 questionnaire; SVM, support vector machine.

3 had the highest percentage lost to follow-up (C1: 30%, C2: 32%, C3: 45%).

## 3.3 | Supervised learning predicts LBP chronicity

We identified three variables that distinguished the acute and chronic LBP groups (Figure 6, Tables S5–S7). One variable (daily calories from alcoholic beverages) met both significance and effect size criteria, while two additional variables (Physical Summary Scale from the SF-12 questionnaire; Item 8 from SF-12 questionnaire "How much did pain interfere with your normal work [including work outside home and housework]") met significance criteria but not effect size criteria.

After optimizing each model's hyperparameters and input variables, prediction accuracy was relatively similar across models (76%–78%) (Table 2, Figure 6D). For each model, the optimized variable set contained Item 8 from the SF-12 questionnaire. In addition, 3 out of 4 models contained daily alcohol consumption, Item 3 from the SF-12 questionnaire ("How much did health limit climbing several flights of stairs?"), and whether the individual used prescription medication (e.g., narcotics) for pain in half of the days in the past month. Training and testing accuracies were calculated to diagnose overfitting, where LR and RF had the smallest differential, followed by ANN and SVM (Table 2).

## 4 | DISCUSSION

Current methods for evaluating LBP patients cannot consistently identify the underlying drivers of pain or predict whether acute pain will become chronic. In this study, we probed the OAI database to determine latent pain phenotypes and develop models to determine chronicity. Our analysis incorporated elements of traditional statistics and machine learning to screen hundreds of variables and identify the most impactful in LBP, despite a limited LBP cohort size. We identified three phenotypes in which SES, mental health, nutrition, and analgesic use had strong associations with LBP. Specifically, individuals with lower or intermediate SES were more likely to have severe pain at enrollment and develop long-term severe pain. Toward clinical implementation, long-term LBP symptoms were predicted by pain interference and alcohol consumption using three machine learning models and traditional regression.

Social determinants of health influence the occurrence or progression of many diseases, such as hypertension,[21] joint degeneration,[22] breast cancer,[23] and kidney disease.[24] In this study, those with lower income and education level were more likely to have severe back pain at enrollment and over 8 year. In agreement, Chen et al.[7] reported that lower social class and unemployment were the strongest predictors for fluctuating and persistent severe LBP in 5 years. In addition, a large network analysis showed that individuals living in rural areas with worse health insurance were more likely to develop LBP.[8] Furthermore, we identified race as an indicator for LBP phenotyping, where Cluster 3 ("low SES, high severity" cluster) had a higher proportion of Black or African American individuals and was more likely to have severe chronic LBP. Disparities are apparent in other health conditions as well,[25–27] including stark disparities in COVID-19 outcomes.[28] Thus, the social, cultural, and biological consequences of racial discrimination, as in other chronic diseases, likely drive LBP. Interestingly, the diet of individuals in Cluster 3 ("low SES, high severity" cluster) was lower in nutritional content than the other clusters, consistent with previous reports of geographic disparities like food deserts.[21,29] Patients with lower SES are also prescribed different pain management strategies. Data from the National Ambulatory Medical Care Survey (NAMCS) suggests thatpatients that are non-white, are from rural areas, or have public health insurance are more likely to be prescribed opioids compared to white patients.[30,31] In this study, prescription analgesic use was increased in the low SES/high pain severity cluster and prescription narcotic use was a predictor of chronic LBP. It may be that those with severe pain and low SES are more likely to use prescription analgeiscs, however this did not prevent chronic LBP in this population. Taken together, these data suggest that US healthcare disparities are impacting LBP, warranting urgent attention from healthcare policy makers.

The associations between mental health, especially depression, and LBP has long been established.[32–34] A synchrony of depression and pain (i.e., when pain severity changes, levels of depression symptoms change in the same direction) has been observed in both short-term[35,36] and long-term follow-up studies.[37] For example, Stevans et al.[6] reported that patients with depression/anxiety were more likely

**TABLE 2** Predicting pain chronicity: Optimized input variable sets and performance metrics

| Model | Optimized variable set | Variable description | Accuracy (95% CI) | AUC (95% CI) | Sensitivity (95% CI) | Specificity (95% CI) |
|---|---|---|---|---|---|---|
| Artificial neural network (ANN) | V00PCTALCH | Block Brief 2000: daily percent of calories from alcoholic beverages (kcal) (calc) | Testing: 78% (77%, 80%) Training: 84% (83%, 84%) | 0.81 (0.79, 0.82) | 81% (79%, 83%) | 73% (70%, 76%) |
| | V00SF8 | SF-12: how much did pain interfere with normal work (include work outside home and housework), past 4 weeks | | | | |
| | V00SF3 | SF-12: how much health limit climbing several flights of stairs | | | | |
| | V00NARCOT | Q50e.Used strong prescription pain medications (e.g., narcotics) for joint pain or arthritis more than half the days of the month, past 30 days | | | | |
| Support vector machine (SVM) | V00PCTALCH | Block Brief 2000: daily percent of calories from alcoholic beverages (kcal) (calc) | Testing: 78% (77%, 79%) Training: 94% (94%, 94%) | 0.83 (0.81, 0.84) | 82% (80%, 84%) | 72% (69%, 74%) |
| | V00BONEFX | Doctor ever said you broke or fractured bone after age 45 | | | | |
| | V00HSPSS | SF-12: physical summary scale for the MOS 12-item short-form health survey (SF-12) v2 (calc) | | | | |
| | V00RXNSAID | MIF: Rx NSAID use indicator (calc) | | | | |
| | V00SF8 | SF-12: how much did pain interfere with normal work (include work outside home and housework), past 4 weeks | | | | |
| | V00SF3 | SF-12: how much health limit climbing several flights of stairs | | | | |
| | V00NARCOT | Q50e.Used strong prescription pain medications (e.g., narcotics) for joint pain or arthritis more than half the days of the month, past 30 days | | | | |
| Random forest (RF) | V00PCTALCH | Block Brief 2000: daily percent of calories from alcoholic beverages (kcal) (calc) | Testing: 76% (75%, 78%) Training: 76% (75%, 77%) | 0.83 (0.81, 0.84) | 85% (83%, 86%) | 70% (67%, 73%) |
| | V00SF8 | SF-12: how much did pain interfere with normal work (include work outside home and housework), past 4 weeks | | | | |
| | V00SF3 | SF-12: how much health limit climbing several flights of stairs | | | | |
| | V00NARCOT | Q50e.Used strong prescription pain medications (e.g., narcotics) for joint pain or arthritis more than half the days of the month, past 30 days | | | | |
| Logistic regression (LR) | V00BONEFX | Doctor ever said you broke or fractured bone after age 45 | Testing: 76% (75%, 78%) Training: 77% (76%, 77%) | 0.80 (0.78, 0.82) | 83% (80%, 85%) | 66% (63%, 69%) |
| | V00SF8 | SF-12: how much did pain interfere with normal work (include work outside home and housework), past 4 weeks | | | | |
| | P01BPTOT | Total days in bed and/or limited activity due to back pain, past 30 days (calc) | | | | |

Abbreviation: AUC, area under the receiver operator characteristic curve.

to transition from acute to chronic LBP. Similarly, in our study, individuals from Cluster 3 ("low SES, high severity" cluster) had more severe depression, as assessed by the CES-D scale, and were more likely to develop severe chronic LBP. In a meta-analysis including 37 studies, Caruso et al.[38] found that antidepressants could improve both quality of life and pain symptoms, suggesting that targeting mental health could have synergistic treatment effects on pain. Interestingly, Quiton et al.[39] found that lower SES was also associated with depression, suggesting sociodemographic factors create unique social identities that impact pain. Our clustering analysis identified unique social identities in each cluster and could be useful in a primary care setting to direct multimodal physical and mental health treatment plans. Furthermore, we observed the highest attrition rate in Cluster 3. Previously, a meta-analysis of 54 studies (5852 subjects) showed that individuals with higher levels of depression were more likely to drop-out of RCTs.[40] These results should be taken into consideration when designing longitudinal studies that target depression and LBP.

Predicting and preventing long-term LBP remains an active area of research. Several studies have utilized trajectory analysis and other conventional techniques to identify risk factors associated with the transition to chronic LBP.[6,7] However, the number of identified risk factors is prohibitive for adoption into daily clinical practice. Here, we optimized predictive models with critical predictors, decreasing model complexity while still achieving satisfactory accuracy, sensitivity, and specificity. Of note, we identified that an individual's perception of how pain interferes with their daily work (SF-12, Item 8) was among the strongest predictors of long-term pain, perhaps, because this item incorporates elements of mental, physical, and socioeconomic health. Furthermore, there is a complicated relationship between LBP and alcohol consumption. In our study, daily alcohol consumption was one of the strongest predictors of long-term LBP, where *increased* alcohol intake was *protective* from LBP. In contrast, previous work associates increased alcohol consumption with psychosocial risk factors in LBP patients.[41,42] Here, we attribute the positive impact of alcohol to its positive correlation to physical fitness,[43,44] as, in our study, the strongest correlates to alcohol intake were dominated by variables related to physical health (20 meter walk time, pace, and step count; 400 meter walk completion status; SF12-Physical Summary Scale, SF-12 items 3 and 8 related to pain interference during activity; ability to perform household activities; BMI). Furthermore, moderate alcohol consumption (wine in particular) is an indicator of higher SES,[45–48] and we detected a positive association between alcohol intake and income. Thus, in our study, chronic pain was avoided by those whose pain did not interfere with their daily activities and, perhaps, were socioeconomically advantaged at study enrollment.

The strength of the OAI dataset is that it provides hundreds of metrics relevant to MSK disease and 8 years of follow-up data, however caution should be used before extrapolating these results to other populations. First, only a small subset of the OAI enrollment cohort reported LBP, limiting broader generalizations. In addition, findings related to race are likely due to the specific racial demographics and social implications of race in the United States, as nearly all individuals in this study identified as Black/African American and

White/Caucasian. Similarly, alcohol consumption has different cultural implications outside the United States as well. Next, individuals enrolled in the OAI study are predisposed to musculoskeletal disease by design; individuals included in the current study had a higher rate of symptomatic and radiographic knee OA than those in the global population (LBP clusters vs. global population: radiographic OA: 55.8% vs. 28.7%, symptomatic OA: 46.5% vs. 12.4%),[49] though the prevalence of knee OA is typically higher in a population with LBP (50%–70%).[50,51] A future prospective LBP-specific cohort studies is warranted to compare social and biological drivers of LBP in a cohort representative of the LBP population. Such a study should consider other factors like genetics and spine pathologies (e.g., disc degeneration) as well, as these are linked to back pain[52,53] but not covered by the OAI.

## SUMMARY

Clinical strategies to diagnose and treat LBP have failed and research efforts to improve outcomes have been incremental. It may be that reasonable preconceptions for what drives LBP have biased the approach to diagnosing and treating back pain and the research strategies for uncovering the underlying causes. In this work, we drop these preconceptions and explore a musculoskeletal database with no *a priori* hypothesis, determining that SES, mental health, nutrition, and analgesic use are the strongest correlates to LBP severity and pain interference was predictive of chronic pain. Our analysis framework was able to screen hundreds of variables to identify the most impactful in LBP, despite a relatively small cohort, and is applicable to any disease type. We opine that focusing on the traditional biomedical descriptors of LBP obscures the need for social programs, exercise programs, access to healthy foods, and education that are necessary to treat LBP.

### CONFLICT OF INTEREST
John T. Martin is a consultant for DiscGenics Inc. ZeYu Huang is a consultant for DePuy Synthes. Neither these companies nor the funding sources for this work contributed to the study design, data

collection, data analysis, manuscript preparation, or decision to submit this manuscript.

## ORCID

*John T. Martin* https://orcid.org/0000-0002-5772-3926

## REFERENCES

1. Hartvigsen J, Christensen K, Frederiksen H. Back pain remains a common symptom in old age. A population-based study of 4486 Danish twins aged 70-102. *Eur Spine J*. 2003;12:528-534.
2. Koumtouzoua S, Higgins S. Evaluating and managing the patient with back pain. *Med Clin North Am*. 2021;105:1-17.
3. Global Burden of Disease Study C. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden Of Disease Study 2013. *Lancet*. 2015;386:743-800.
4. Hartvigsen J, Hancock MJ, Kongsted A, et al. What low back pain is and why we need to pay attention. *Lancet*. 2018;391:2356-2367.
5. Koes BW, van Tulder MW, Thomas S. Diagnosis and treatment of low back pain. *BMJ*. 2006;332:1430-1434.
6. Stevans JM, Delitto A, Khoja SS, et al. Risk factors associated with transition from acute to chronic low back pain in US patients seeking primary care. *JAMA Netw Open*. 2021;4:e2037371.
7. Chen Y, Campbell P, Strauss VY, Foster NE, Jordan KP, Dunn KM. Trajectories and predictors of the long-term course of low back pain: cohort study with 5-year follow-up. *Pain*. 2018;159:252-260.
8. Strozzi AG, Pelaez-Ballestas I, Granados Y, et al. Syndemic and syndemogenesis of low back pain in latin-American population: a network and cluster analysis. *Clin Rheumatol*. 2020;39:2715-2726.
9. Yoshimoto T, Ochiai H, Shirasawa T, et al. Clustering of lifestyle factors and its association with low Back pain: a cross-sectional study of over 400,000 Japanese adults. *J Pain Res*. 2020;13:1411-1419.
10. Yadollahpour N, Zahednejad S, Yazdi MJS, Esfandiarpour F. Clustering of patients with chronic low back pain in terms of physical and psychological factors: a cross-sectional study based on the STarT Back Screening Tool. *J Back Musculoskelet Rehabil*. 2020;33:581-587.
11. Molgaard Nielsen A, Binding A, Ahlbrandt-Rains C, Boeker M, Feuerriegel S, Vach W. Exploring conceptual preprocessing for developing prognostic models: a case study in low back pain patients. *J Clin Epidemiol*. 2020;122:27-34.
12. Smart KM, Blake C, Staines A, Thacker M, Doody C. Mechanisms-based classifications of musculoskeletal pain: part 1 of 3: symptoms and signs of central sensitisation in patients with low back ($+/-$ leg) pain. *Man Ther*. 2012;17:336-344.
13. Deo RC. Machine learning in medicine. *Circulation*. 2015;132:1920-1930.
14. Bibault JE, Giraud P, Burgun A. Big data and machine learning in radiation oncology: state of the art and future prospects. *Cancer Lett*. 2016;382:110-117.
15. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375:1216-1219.
16. Waljee AK, Joyce JC, Wang S, et al. Algorithms outperform metabolite tests in predicting response of patients with inflammatory bowel disease to thiopurines. *Clin Gastroenterol Hepatol*. 2010;8:143-150.
17. Waljee AK, Higgins PD, Singal AG. A primer on predictive models. *Clin Transl Gastroenterol*. 2014;5:e44.
18. Gaskin GL, Pershing S, Cole TS, Shah NH. Predictive modeling of risk factors and complications of cataract surgery. *Eur J Ophthalmol*. 2016;26:328-337.
19. Huang Z, Huang C, Xie J, et al. Analysis of a large data set to identify predictors of blood transfusion in primary total hip and knee arthroplasty. *Transfusion*. 2018;58:1855-1862.
20. McInnes L, Healy J and Melville J. Umap: uniform manifold approximation and projection for dimension reduction. *arXiv Preprint arXiv: 180203426* 2018.
21. Brahma VL, Snow J, Tam V, et al. Socioeconomic and geographic disparities in idiopathic intracranial hypertension. *Neurology*. 2021;96:e2854-e2860.
22. Delanois RE, Tarazi JM, Wilkie WA, et al. Social determinants of health in total knee arthroplasty: are social factors associated with increased 30-day post-discharge cost of care and length of stay? *Bone Joint J*. 2021;103-B:113-118.
23. Fong AJ, Lafaro K, Ituarte PHG, Fong Y. Association of living in urban food deserts with mortality from breast and colorectal cancer. *Ann Surg Oncol*. 2021;28:1311-1319.
24. Banerjee T, Crews DC, Wesson DE, et al. Powe NR and Team CCSFood insecurity, CKD, and subsequent ESRD in US Adults. *Am J Kidney Dis*. 2017;70:38-47.
25. Havranek EP, Mujahid MS, Barr DA, et al. Social determinants of risk and outcomes for cardiovascular disease: a scientific statement from the American Heart Association. *Circulation*. 2015;132:873-898.
26. Johnson DA, Lewis TT, Guo N, et al. Associations between everyday discrimination and sleep quality and duration among African Americans over time in the Jackson Heart Study. *Sleep*. 2021;44:zsab162.
27. Ashana DC, D'Arcangelo N, Gazarian PK, et al. "Don't talk to them about goals of care": understanding disparities in advance care planning. *J Gerontol A Biol Sci Med Sci*. 2021;77:339-346.
28. Yancy CW. COVID-19 and African Americans. *JAMA*. 2020;323:1891-1892.
29. Potluri VS, Sawinski D, Tam V, et al. Effect of neighborhood food environment and socioeconomic status on serum phosphorus level for patients on chronic dialysis. *J Am Soc Nephrol*. 2020;31:2622-2630.
30. Prunuske JP, St Hill CA, Hager KD, et al. Opioid prescribing patterns for non-malignant chronic pain for rural versus non-rural US adults: a population-based study using 2010 NAMCS data. *BMC Health Serv Res*. 2014;14:563.
31. Lin HC, Wang Z, Boyd C, Simoni-Wastila L, Buu A. Associations between statewide prescription drug monitoring program (PDMP) requirement and physician patterns of prescribing opioid analgesics for patients with non-cancer chronic pain. *Addict Behav*. 2018;76:348-354.
32. Carroll LJ, Cassidy JD, Cote P. Depression as a risk factor for onset of an episode of troublesome neck and low back pain. *Pain*. 2004;107:134-139.
33. Leino P, Magni G. Depressive and distress symptoms as predictors of low back pain, neck-shoulder pain, and other musculoskeletal morbidity: a 10-year follow-up of metal industry employees. *Pain*. 1993;53:89-94.
34. Rayner L, Hotopf M, Petkova H, Matcham F, Simpson A, McCracken LM. Depression in patients with chronic pain attending a specialised pain treatment centre: prevalence and impact on health care costs. *Pain*. 2016;157:1472-1479.
35. Hawker GA, Gignac MA, Badley E, et al. A longitudinal study to explain the pain-depression link in older adults with osteoarthritis. *Arthritis Care Res (Hoboken)*. 2011;63:1382-1390.
36. Gerrits MM, van Marwijk HW, van Oppen P, van der Horst H, Penninx BW. Longitudinal association between pain, and depression and anxiety over four years. *J Psychosom Res*. 2015;78:64-70.
37. Glette M, Stiles TC, Jensen MP, Nilsen TIL, Borchgrevink PC, Landmark T. Impact of pain and catastrophizing on the long-term course of depression in the general population: the HUNT pain study. *Pain*. 2021;162:1650-1658.
38. Caruso R, Ostuzzi G, Turrini G, et al. Beyond pain: can antidepressants improve depressive symptoms and quality of life in patients with neuropathic pain? A systematic review and meta-analysis. *Pain*. 2019;160:2186-2198.
39. Quiton RL, Leibel DK, Boyd EL, Waldstein SR, Evans MK, Zonderman AB. Sociodemographic patterns of pain in an urban

community sample: an examination of intersectional effects of sex, race, age, and poverty status. *Pain*. 2020;*161*:1044-1051.

40. Cooper AA, Conklin LR. Dropout from individual psychotherapy for major depression: a meta-analysis of randomized clinical trials. *Clin Psychol Rev*. 2015;*40*:57-65.

41. Hurwitz EL, Randhawa K, Torres P, et al. The Global Spine Care Initiative: a systematic review of individual and community-based burden of spinal disorders in rural populations in low- and middle-income communities. *Eur Spine J*. 2018;*27*:802-815.

42. Booker EA, Haig AJ, Geisser ME, Yamakawa K. Alcohol use self report in chronic back pain--relationships to psychosocial factors, function performance, and medication use. *Disabil Rehabil*. 2003;*25*:1271-1277.

43. Leasure JL, Neighbors C, Henderson CE, Young CM. Exercise and alcohol consumption: what we know, what we need to know, and why it is important. *Front Psych*. 2015;*6*:156.

44. Piazza-Gardner AK, Barry AE. Examining physical activity levels and alcohol consumption: are people who drink more active? *Am J Health Promot*. 2012;*26*:e95-e104.

45. Mortensen EL, Jensen HH, Sanders SA, Reinisch JM. Better psychological functioning and higher social status may largely explain the apparent health benefits of wine: a study of wine and beer drinking in young Danish adults. *Arch Intern Med*. 2001;*161*:1844-1848.

46. Dawson DA, Grant BF, Chou SP, Pickering RP. Subgroup variation in U.S. drinking patterns: results of the 1992 national longitudinal alcohol epidemiologic study. *J Subst Abuse*. 1995;*7*:331-344.

47. Casswell S, Pledger M, Hooper R. Socioeconomic status and drinking patterns in young adults. *Addiction*. 2003;*98*:601-610.

48. Moore AA, Gould R, Reuben DB, et al. Longitudinal patterns and predictors of alcohol consumption in the United States. *Am J Public Health*. 2005;*95*:458-465.

49. Cui A, Li H, Wang D, Zhong J, Chen Y, Lu H. Global, regional prevalence, incidence and risk factors of knee osteoarthritis in population-based studies. *EClinicalMedicine*. 2020;*29-30*:100587.

50. Rundell SD, Karmarkar A, Nash M, Patel KV. Associations of multiple chronic conditions with physical performance and falls among older adults with back pain: a longitudinal population-based study. *Arch Phys Med Rehabil*. 2021;*102*:1708-1716.

51. Stupar M, Cote P, French MR, Hawker GA. The association between low back pain and osteoarthritis of the hip and knee: a population-based cohort study. *J Manipulative Physiol Ther*. 2010;*33*:349-354.

52. Battie MC, Videman T, Levalahti E, Gill K, Kaprio J. Heritability of low back pain and the role of disc degeneration. *Pain*. 2007;*131*:272-280.

53. Battie MC, Ortega-Alonso A, Niemelainen R, et al. Lumbar spinal stenosis is a highly genetic condition partly mediated by disc degeneration. *Arthritis Rheumatol*. 2014;*66*:3505-3510.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.