Krish Rai

BIOSTAT202

Final Project

08/28/23

Title: Using Supervised Learning to Predict Osteoarthritis Pain with non-MRI data

Background: Osteoarthritis, which involves the wearing down of joints, affects older adults at an alarming rate. 10% of men and 13% of women in the United States above 60 years of age have osteoarthritis of the knee [1]. Predicting worsening osteoarthritis pain can help limit potential long term mobility issues. Much of this relies on MRI tests of the knee to assess cartilage, but the cost of an MRI remains high averaging around $600 in a particular study[2]. This type of testing is cost prohibitive and is therefore not the best approach for predicting progression of the disease in the larger population. Another study showed the charge per procedure to be $410 per X-ray compared to $2048 for an MRI [3]. Therefore, there is a need for a better approach to predict onset and progression of the disease without relying on expensive MRI test data. Using X-ray data and other epidemiological information to predict progression would be useful in improving care.

Research Question: Can X-ray data of joint space and cartilage thickness along with other survey/epidemiological measures like race accurately predict progressing osteoarthritis pain (measured by jspainprg- increasing pain and decreasing joint space width) for osteoarthritis patients without numerical MRI data?

Variable/Predictors: The variables used as predictors of jspainprg were as follows:

For **Categorical:** Hisp (Hispanic or not), sex, race, cohort, xr_jsm (Xray assessment of degree of joint space narrowing in medial compartment), xr_jsl (Xray assessment of degree of joint space narrowing in lateral compartment), and xr_kl (Kellgren-Lawrence grading of severity of osteoarthritis scale of 1-5) were chosen. For **Numerical**: xr_min_thk_med (minimal cartilage thickness in the weight-bearing portion of the medial compartment via Xray) was chosen. All other variables were removed during methods.

Outcome/Objective: Outcome variable is jspainprg, which notes increasing pain and decreased joint space width for osteoarthritis patients. This is listed as a 1 for pain progression or 0 for no progression pain. The objective is to use supervised learning to predict a risk of progressed osteoarthritis pain from jspainprg without MRI data.

Methods: The osteoarthritis dataset (Osteoarthritis.xlsx) was used for this research question. This data is part of the OAI (Osteoarthritis Initiative) and was used to learn more about the disease.

The dataset was added to Orange upon which the variables were checked to ensure they were correctly sorted into categorical or numerical. After this, jspainprg was assigned as the target variable and the MRI variables were ignored. Using the Rank widget, all variables with correlation to the target (using Information gain > 0.00) were included. Age, BMI, as well as categorical variables fkp (frequency of knee pain), and inj (history of knee injury) were ignored due to Information gain or gain ratio being low, 0.00/0.01. Patient ID was removed as well.

The data was normalized to an interval of [-1,1]. Using the Data Sampler Widget, 75% of the data was used for training/cross validation and 25% for testing. The cross-validation method was used to select the best model, and then the best model was picked for testing. This method was used to minimize overfitting. 4 models were then selected: Random Forest, k Nearest Neighbor (kNN), Neural Network, and Support Vector Machines (SVM).

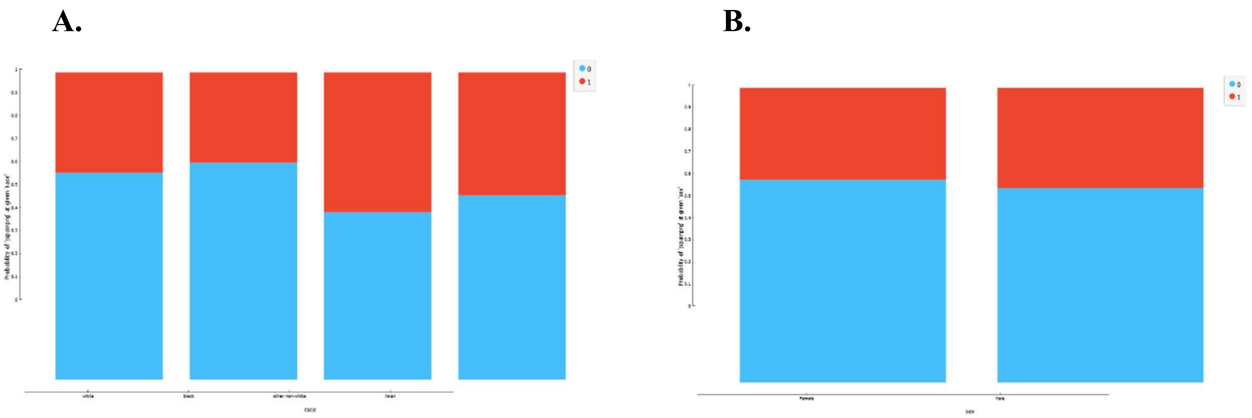
The variables that were included (after using Rank to see which variables correlated with target) were: Hisp, xr_min_thk_med, xr_kl, cohort, xr_jsm, sex, race, and xr_jsl.

The models were then applied to Test and Score in Orange to measure AUC. Parameter tuning was also done to optimize AUC scores/limit overfitting. For k nearest neighbors, this meant changing the number of neighbors. For Random Forest, it meant changing the number of trees. Random Forest number of attributes/trees was reduced to 3 to prevent overfitting. The Cost C value was changed for SVM and regularization was changed for Neural Network as well.

Table 1. Breakdown of Progression of Osteoarthritis Pain (jspainprg = 0 or 1) by Race

Race	N (%) for jspainprg = 0	N (%) for jspainprg = 1
White	320 (78.8)	155 (79.9)
Black	77 (18.9)	32 (16.5)
Asian	3 (0.74)	2 (1.03)
Other non White	6 (1.47)	5 (2.57)

Figure 1.



C.

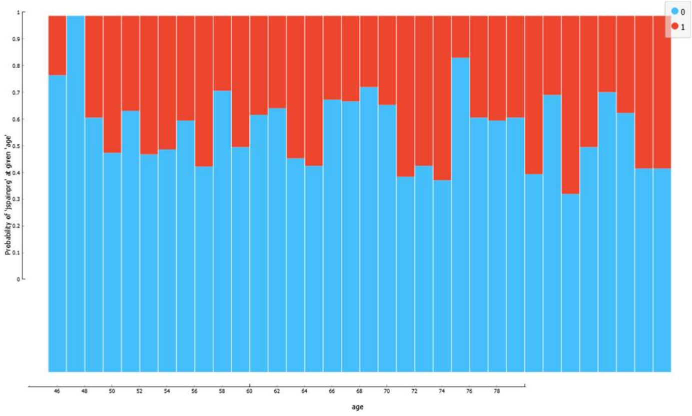


Figure 1. Breakdown of Participants by A. Race, B. Sex. And C. Age for jspainprg. **A.** 32.63% of white patients, 29.36% of black patients, 45.45% of other non white patients, and 40% of Asian patients (low sample size) suffered the progression of osteoarthritis pain (jspainprg =1). **B.** 31.16% of females and 34.01% of males were also under jspainprg=1. **C.** Age does not show clear trend in relation to where jspainprg=1.

Results:

78.8% of patients were white, 18.9% of patients were black, 0.74% of patients were Asian, and 1.47% patients were other race out of total who answered no to increased pain from osteoarthritis (jspainprg = 0), while 79.9% of patients were white, 16.5% of patients were black, 1.03% of patients were Asian, and 2.57% patients were other race who answered yes (jspainprg = 1). The percentage of patients by race was therefore relatively consistent between jspainprg = 0 and 1 (Table 1).

While different races and sexes had different percentages for jspainprg=1, age had no clear trend with jspainprg (Figure 1). Age (along with BMI and others) was therefore removed as per methods.

The data was as follows: Precision and Recall were calculated as well for each Test and Score as well as CA (accuracy), but AUC remained primary performance metric.

Table 2.

Model	Cross Validation Data (AUC/CA/Precision/Recall)	Testing Data (AUC/CA/Precision/Recall)
SVM - C = 0.90 - C = 1.00	0.548/0.648/0.621/0.648 **0.592**/0.643/0.622/0.643	**0.602**/0.628/0.602/0.628
K Nearest Neighbors	0.534/0.698/0.487/0.698	

-50 neighbors -100 neighbors	0.562/0.698/0.487/0.698	
Random Forest -20 trees -30 trees	0.560/0.704/0.637/0.704 0.560/0.707/0.650/0.707	
Neural Network -200 iterations -175 iterations	0.454/0.698/0.487/0.698 0.460/0.698/0.487/0.698	

The AUC (Area under the ROC) is an indicator of performance, especially useful for binary target variables such as jspainprg. The closer to 1.0, the better the model. Upon significant parameter tuning while ensuring not to lead to overfitting (too many trees, neighbors, etc.), the following models' AUC scores were computed in Test and Score (Table 2).

Random Forest had a cross validation AUC score of 0.560. Neural network scores remained under 0.500 for cross validation AUC score, and kNN had a score of 0.562 for 100 neighbors. The highest cross validation AUC was SVM with C=1.00, which had a score of 0.592 (Table 2). The SVM model was therefore selected for the testing data which had an AUC score of 0.602.

Discussion:

While SVM (C=1.00) performed the best of the models in the validation step, the validation AUC score of 0.592 and the testing score of 0.602 were not as high as expected.

Therefore, this data without MRI data was not a strong predictor of jspainprg. The variables in the Rank widget that ranked highest to the target such as identifying as Hispanic, race, and sex seem like poor predictors for pain. Even adding MRI data back did not improve AUC.

Additionally, the survey data like frequency of knee pain and prior injury which seem like better pain predictors than Hisp or race showed no correlation to the target variable. It is likely that the small size of the data (600 participants) and high number needed for adequately training the data impaired the ability to predict and is the largest limitation of this study. In conclusion, the question was not adequately answered.

Another study that measured knee pain of osteoarthritis had a higher external validation AUC score around 0.7 but had more participants and more quantitative variables such as physical activity and hypertension [4]. Since the age and BMI data in this study showed no correlation with jspainprg, it is possible that the data was simply too small and limited. Prediction without MRI data is likely still possible, but more participants with more quantitative measurements would be a next step. Additionally, Osteoarthritis Data was searched online on UCSF MOST.

This data had almost 1400 participants and seemed like a good test to compare and understand scores, but it is only available to researchers. Requesting this may be a next step in understanding the issues in this data.

Citations:

[1] Y. Zhang and J. M. Jordan, “Epidemiology of osteoarthritis,” *Clinics in Geriatric Medicine*, vol. 26, no. 3. 2010. doi: 10.1016/j.cger.2010.03.001.

[2] K. E. Rudisill ; P.P. Ratnasamy; P.Y. Joo; L.E. Rubin; J.N. Grauer; (n.d.). *Magnetic Resonance Imaging in the year prior to total knee arthroplasty: A potential overutilization of healthcare resources*. Journal of the American Academy of Orthopaedic Surgeons. Global research & reviews. https://pubmed.ncbi.nlm.nih.gov/37205731/

[3] C. L. Sistrom and N. L. McKay, “Costs, charges, and revenues for hospital diagnostic imaging procedures: Differences by modality and hospital characteristics,” *Journal of the American College of Radiology*, vol. 2, no. 6, 2005, doi: 10.1016/j.jacr.2004.09.013.

[4] T. K. Yoo, D. W. Kim, S. B. Choi, E. Oh, and J. S. Park, “Simple scoring system and artificial neural network for knee osteoarthritis risk prediction: A cross-sectional study,” *PLoS One*, vol. 11, no. 2, 2016, doi: 10.1371/journal.pone.0148724.