Data mining approach to determining gait abnormalities in runners with patellofemoral pain syndrome

Ross J. Brancati[1*], Katherine A. Boyer[1]
[1]Department of Kinesiology, University of Massachusetts Amherst, Amherst, MA
email: *rbrancati@umass.edu

Introduction

Patellofemoral Pain Syndrome (PFPS) is a common cause of anterior knee pain in recreational runners. Alterations in running mechanics may contribute to the onset of PFPS. However, pain may also contribute to additional changes in both gait mechanics and neuromuscular activation patterns. Previous research has found alterations in kinematic variables in runners with PFPS such as greater peak hip adduction angle.[1,2]

To date, most of the research with PFPS has focused on quantifying differences in discrete variables using a traditional hypothesis testing approach, which may miss important characteristics of biomechanical waveforms. With the recent advancements in machine learning, data mining has been applied to gait studies to determine important features for classifying pathological populations, which may have advantages over traditional parametric testing.[3] Shi et al. used machine learning approaches to diagnose[4] and determine biomechanical adaptations of[5] runners with PFPS. They reported hip, knee, and ankle flexion angle as well as electromyography (EMG) from multiple muscles as the most pertinent features for classifying these two groups. However, their approaches lack kinetic data and use the mean of the entire waveforms as features, omitting crucial aspects of waveform variables. This study's aim was to determine the biomechanical characteristics that are most influential on classifying runners with and without PFPS while considering entire waveforms. Identification of these key characteristics may support development of intervention and selection of critical outcomes for future research.

Methods

37 recreational runners - 19 symptomatic PFPS (10f, 22.6±4.3 years, speed: 2.88±0.53 m/s) and 18 who never experienced PFPS (10f, 22±3.5 years, speed: 2.80±0.32 m/s) participated in this study. Participants ran at self-selected pace for 21 minutes on an instrumented treadmill (Treadmetrix, Park City, UT, USA) while eight infrared motion capture cameras (Oqus-5 series, Qualysis, Inc. Gothenburg, Sweeden) and 11 wireless sensors (Delsys Trigno, Delsys, Inc., Natick, MA, USA) collected kinetic, kinematic, and EMG data. Ensemble averages of approximately ten strides at the first minute of the run were exported using Visual 3D (C-Motion Inc., Germantown, MD). Matrices for each kinetic, kinematic, and EMG variable were generated resulting in 27 total matrices (9 kinetic, 9 kinematic, and 11 EMG). A principal component analysis (PCA) was performed on each matrix, and the top two PC scores for each variable generated a data matrix, *[37x54]*, where each row represented the subject's stride, and each column represented the respective PC score.

A support vector machine (SVM) with a linear kernel was then trained on 80% of the data and tested with the remaining 20%. Metrics such as accuracy and F1 score were used to select the model with the best features. After training the model, the weights of each feature were calculated to determine feature importance and each pertinent feature was interpreted by plotting the 5th and 95th percentile of the principal component score, as previously described.[6] All data analyses were performed using Python 3.8.8 and multiple toolboxes such as scikit-learn.

Results and Discussion

An SVM model with eight features was selected to classify healthy and injured runners, resulting in high accuracy (85.7%), precision (90.5%), recall (85.7%), and F1 score (86.4%). The important features, their weights, and interpretations are listed in *Table 1*. The most heavily weighted kinetic features in the classification model were in the transverse plane. For all features, with the exception of the ankle and knee frontal plane moment, the PFPS group has smaller magnitudes and rates of change for the EMG, kinematic, kinetic selected features.

Significance

This analysis technique combining PCA to study entire waveforms with machine learning to mine important features uncovered variables that have not previously been outlined as important characteristics of runners with PFPS. Prior work on runners with PFPS highlights kinematic alterations in the frontal plane, whereas this approach found transverse plane kinetic variables to be essential to the classification procedure. Data mining approaches may be advantageous for determining variables to focus on when designing future studies and interventions to alleviate PFPS.

Acknowledgments

Funding: Deluca Foundation. Carl Jewel, for data collection.

References

[1] Barton et al., 2009. *Gait & Posture*. 30: 405-416
[2] Dierks et al., 2008. *J of Orth. & Sports PT*. 38(8): 448-456
[3] Halilaj et al., 2018. *J of Biomechanics*. 81: 1-11
[4] Shi et al., 2021. *Front. in Pub. Health*. 9: 643191
[5] Shi et al., 2020. *CSRSWTC*. 50769.2020.9372597
[6] Deluzio et al., 2007. *Gait & Posture*. 25: 86-9

Feature	Weight	Interpretation of Feature
Rectus Femoris PC 1	0.561	Healthy group showed greater peak activation during stance and in preparation for loading.
Hip Rotation Moment PC 1	0.283	Healthy group had faster rates of rotational loading change in the stance phase.
Ankle Rotation Moment PC 1	0.180	Healthy group presented a larger internal rotation moment at toe off.
Knee Rotation Moment PC 1	0.135	Healthy group had a larger peak knee flexion moment during stance.
Ankle Inversion Moment PC 1	0.063	Injured group showed a larger peak ankle inversion moment during stance.
Hip Rotation Moment PC 2	0.043	Healthy group had a larger peak internal rotation moment in early stance.
Hip Rotation Angle PC 1	0.034	Healthy group had greater hip external rotation at mid stance and in preparation for loading.
Knee Adduction Moment PC 1	0.001	Injured group presented greater peak knee adduction moment during stance phase.

Table 1: Interpretation of the most relevant features in the SVM-A classification task with the respective feature weights. Waveforms of the 5th and 95th percentiles of each of the features were plotted to interpret their meaning.