Journal of Statistics on Health Decision

EXTENDED ABSTRACT

Applying machine learning methods to predict the Parkinson´s Disease Questionnaire-39 Summary Index

Daniel Magano[1,2], Joana B. Moreira[1,3], Pedro P. Rodrigues[4]

[1] Ph.D. Program in Health Data Science, Faculty of Medicine, University of Porto, Porto, Portugal;
[2] BIAL – Portela & Ca S.A., Coronado, Portugal
[3] Anaesthesiology, Emergency and Intensive Care Department, Centro Hospitalar Universitário do Porto, Largo do Prof. Abel Salazar, Porto, Portugal
[4] CINTESIS@RISE, MEDCIDS - Faculty of Medicine of the University of Porto, Porto, Portugal

Introduction:

Health-related quality of life (HRQoL) gained importance in the last decades and is considered to be an important outcome measure in chronic diseases studies. HRQoL disease specific instruments reflect the consequences of that disease to a particular person and are sensitive to change in perceived HRQoL. In Parkinson’s disease (PD) several disease specific HRQoL instruments have become available in the past few years. A 2002 systematic review of HRQoL instruments in PD concluded that in many situations the Parkinson´s Disease Questionnaire -39 (PDQ-39) is the most appropriate HRQoL instrument, because it is the scale that has been tested most thoroughly, has adequate clinimetric characteristics, has been used in the largest number of studies, and is available in many languages [1].

The PQD-39 is a 39-item self-report questionnaire, which assesses PD specific health related quality over the last month. It assesses how often patients experience difficulties across the 8 quality of life dimensions and the impact of PD on specific dimensions of functioning and well-being [2]. Its summary index (PDQ-39SI) is a widely used patient-reported clinical trial endpoint. Peter Hagell and Maria H. Nilsson assessed the unidimensionality of the PDQ-39 and PDQ-39SI using Rasch and confirmatory factor analysis and concluded its multidimensional nature [3].

The objective of this study is to do an exploratory analysis on the multidimensional nature of the PDQ-39SI using machine learning methods to improve the interpretability of the score and obtain a model to predict the perceived quality of life of people with PD.

Methods:

Data analysis was completed with R version 4.1.1. Data used in the preparation of this article were obtained from the Parkinson’s Real-World Impact assesSMent (PRISM) database. PRISM study and database was funded by BIAL – Portela & C^{a}, S.A., designed in collaboration with The Cure Parkinson’s Trust, an advocacy group based in the United Kingdom (UK), and reviewed by the PRISM steering committee.

The PRISM database contains data from 861 people from 5 European countries with PD collected in the context of an observational study with cross-sectional design [4].

PDQ-39SI multivariable nature was first assessed using the Expectation-Maximization (EM) algorithm (Figure 1). Based on the latent classes, a transformation of PDQ-39SI from continuous to binomial outcome was performed: mild vs severe symptoms impacting their quality of life.

Clinically relevant variables for modeling the PDQ-39SI were initially selected from the PRISM database. Univariate logistic regressions were created with the PDQ-39SI as a function of each of the previously selected variables. The variables presenting a p-value below 0.05 were ordered according to their impact on the odds ratio and included in the machine learning models by forward selection based on the resulting AUC values. Incomplete observations were then removed resulting in a dataset with 615 patients

Before training the models, 20% of the dataset was saved to be used as independent validation data. With the remaining 80% of the dataset (training set), several models were trained to predict the classes obtained through EM. A repeated 10-fold cross validation was performed on this training set to estimate the respective ROC curves (Figure 2). Additionally, the performance of each model obtained with the 20% independent validation set was also assessed through the plotting of ROC curves (Figure 3).

Keywords:
Health-Related Quality of Life, Machine Learning, Parkinson Disease, Parkinson´s Disease Questionnaire -39, Parkinson´s Disease Questionnaire -39 Summary Index

Corresponding author:
Daniel Magano
danielmagano@outlook.com

Conflict of interest:
the authors declare no conflict of interests

First published: 20JUL2022

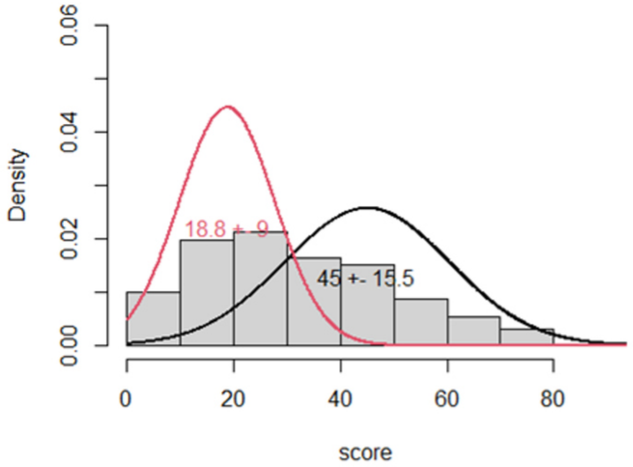

Figure 1 - Latent classes on Parkinson´s Disease Summary Index. ***Legend:*** With the use of Expectation Maximization algorithm (2 kernels) 2 latent classes are described: **(i)** red distribution with mean of 18.8 and standard deviation of ± 9, and **(ii)** black distribution with mean of 45 and standard deviation of 15.5. Interception of the two distributions at the 31 value.

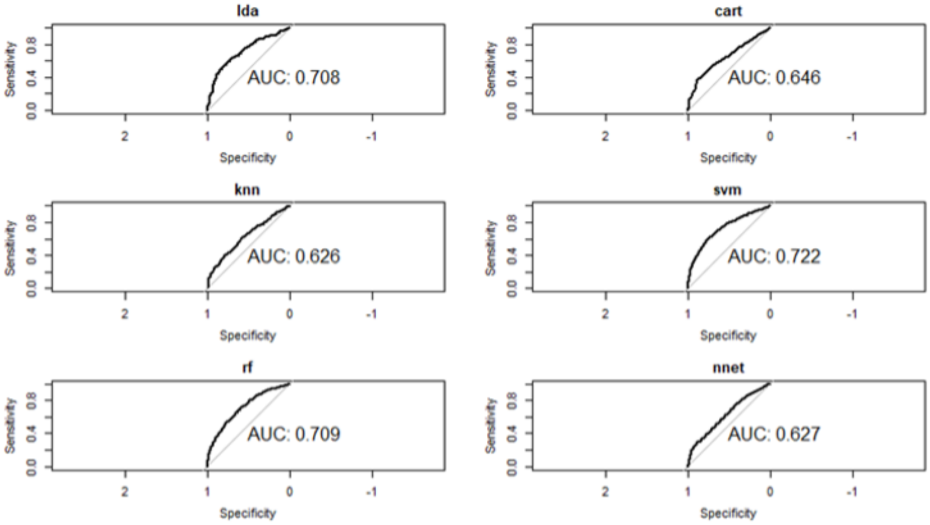

Figure 2 - Performance estimation of different machine learning methods for PDQ-39SI modeling on the training set using repeated 10-fold cross validation. ***Legend:*** ROC curves (receiver operating characteristic curves) of the repeated 10-fold cross validation training set (80% of the dataset) with **(i)** lda (Linear Discriminant Analysis), **(ii)** cart (Classification and Regression Tree, splitting index: gini), **(iii)** knn (K-nearest Neighbors, K=3), **(iv)** svm (Support Vector Machine, kernel: radial basis function), **(v)** rf (Random Forest, number of trees = 500), **(vi)** nnet (Feedforward Neural Network, 1 hidden layer with 5 neurons). AUC (area under the curve)

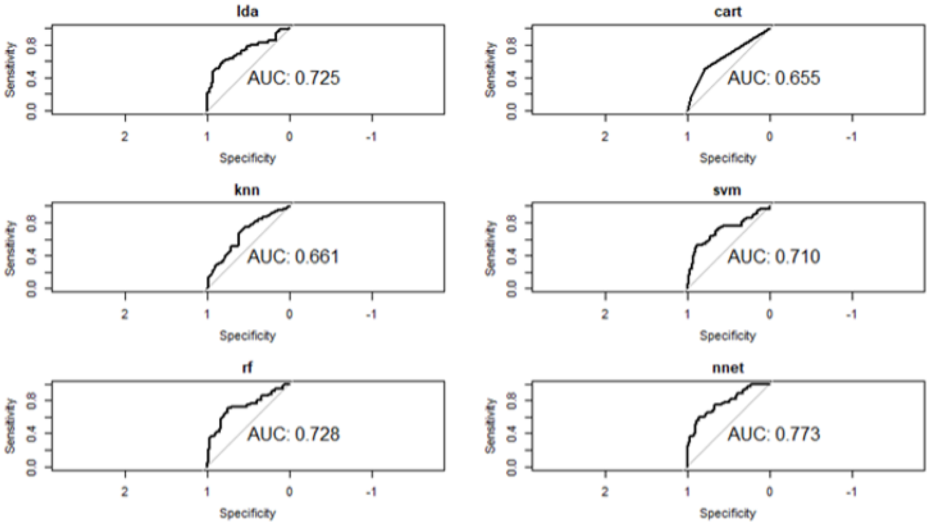

Figure 3 - Performance of different machine learning methods for PDQ-39SI modeling on the independent validation set. ***Legend:*** ROC curves (receiver operating characteristic curves) of the independent validation set (20% of the dataset) with **(i)** lda (Linear Discriminant Analysis), **(ii)** cart (Classification and Regression Tree, splitting index: gini), **(iii)** knn (K-nearest Neighbors, K=3), **(iv)** svm (Support Vector Machine, kernel: radial basis function), **(v)** rf (Random Forest, number of trees = 500), **(vi)** nnet (Feedforward Neural Network, 1 hidden layer with 5 neurons). AUC (area under the curve)

Results:

The PDQ-39SI was converted from a continuous outcome into two categories based on the impact on the quality of life of the person with PD: mild (with a PDQ-39SI below 31) and severe (PDQ-39SI above 31).

The ROC curves obtained with repeated cross-validation for each of the trained models are presented in figure 2 while the ROC curves obtained with the independent validation data set are presented in figure 3.

Discussion:

The obtention of a binomial classification for the PDQ-39SI can facilitate its interpretation and harmonize its utilization by different health care providers (HCP). Additionally, machine learning models applied to clinical variables can be used to predict to which of these classes a person with Parkinson’s disease would belong. This information would allow HCPs to predict which of the patients are expected to have a lower quality of life.

The PDQ-39SI is based on the patient’s interpretation about several dimensions measuring his/her perceived quality of life. The subjectivity of the information collected through a survey is considered a limitation of this study.

The results obtained demonstrate a good performance of Linear Discriminant Analysis, Support Vector Machine and Random Forest methods. Future work will focus on hyperparameters optimization and on studying a wider range of variables to accommodate possible confounding factors.

References:

1. Marinus J, Ramaker C, van Hilten JJ, Stiggelbout AM. Health related quality of life in Parkinson´s disease: a systematic review of disease specific instruments. J Neurol Neurosurg Psychiatry. 2002 Feb 1;72(2):241 LP – 248. https://doi.org/10.1136/jnnp.72.2.241
2. parkinsons-disease-questionnaire-pdq-39-pdq-8. Available from: https://innovation.ox.ac.uk/outcome-measures/parkinsons-disease-questionnaire-pdq-39-pdq-8/
3. Hagell P, Nilsson MH. The 39-item Parkinson’s Disease Questionnaire (PDQ-39): is it a unidimensional construct? Ther Adv Neurol Disord. 2009 May 28;2(4):205–14. https://doi.org/10.1177/1756285609103726
4. Parkinson’s Real World Impact assesSMent (PRISM). Available from: www.prism.bial.com