Integration of statistical and machine learning methods to identify factors associated with opioid prescribing in NIH research standards for Low Back Pain survey data: a pilot analysis

Introduction:

The recent development of research standards for low back pain (NIH LBP taskforce reference) creates an opportunity for prospective data standardization. Ultimately, the goal of this standardized data collection is to better understand patterns for treatment response and build predictive care models. While the impact of aggregate data will depend on large-scale integration, the focus of this study is to better understand the relationship and predictive ability of the survey variables, specifically examining predictors of opioid use.

Clinical research is evolving to reflect the need for efficient clinical trial design and data collection, which is reflected by the move toward improved data standardization. A key component of this is adaptive statistical designs and analysis methods. The expert consensus panel developed the NIH task force for research standards questionnaire (LBPTF) to overcome common research barriers while addressing the underlying key clinical questions for low back pain. Specifically, it shifted the focus from anatomic or pathophysiological classification to that of pain interference, functional status, and pain intensity. This focused questionnaire measures these domains using several short forms from PROMIS (Patient-Reported Outcome Measurement Information System).

The novel organizational framework of the LBPTF questionnaire incorporates key clinical self-report measures as well as information about co-morbid conditions, demographic information, and treatment history. Understanding the co-occurence patterns of these data may provide insight into more focused data collection as well as build toward predictive modeling. The inherent limitations of self-reported data are mitigated by the extensive development of the minimum data set variables to incorporate key perceived domains of influence.

Building from this perspective, the objective of this pilot survey was to deconstruct and analyze the inter-relationship of these variables in a way that will provide more meaningful analysis of these data moving forward. Statistical analysis and interpretation can be misleading due to inherent data assumptions, but with the data points selected by expert consensus, this minimum dataset represents the starting point for analyzing these relationships. Recognizing the limitations of a survey snapshot, we planned iterative analyses of a pilot survey obtained during the LBPTF with several statistical and machine learning methods to validate our approach.

Methods

This was a single site study conducted by the Stanford Systems Neuroscience and Pain Lab (PI: Mackey) using the NIH Chronic Low Back Pain Task Force Questionnaire. High visibility ads were placed to recruit subjects in the community with back pain San Francisco Bay Area. After a telephone screen, a website link was generated and provided to the subject. Electronic data capture (EDC) including questionnaire presentation and response collection was performed using RedCap (Vanderbilt).

The study protocol was approved by the Stanford School of Medicine Institute Review Board. From May 2013 to July 2013, a total of 243 subjects underwent testing and contributed data. Of these, 22 records were not valid due to either 1) less than 18 years old, or 2) not having on-going low back pain. There were 221 valid records for the following analysis.

Using Jupyter…. Exploratory analysis was performed to identify and address missing data. Following this, data cleaning and normalization was performed. During statistical and distributional analysis, range and counts were analyzed. Summary statistics were generated. Logistic regression of categorical variables was performed. Statistical significance was determined using chi-square analysis to determine dependency of target variables. Machine learning algorithms were used to identify most important features using: 1) recursive feature selection 2) forward/backward/mixed selection. Decision tree modeling and neural networks were performed to validate modeling approach.

Results

(Exploratory analysis) Initial analysis- some questions are not answered in survey- dealing with missing data and approach.

data cleaning with normalization

Statistical and distributional analysis- ranges and counts

Statistical significance with Chi-square analysis- given a particular question/factor. Based on the distribution does a target variable change- is the target variable dependent (i.e. pain score dependent on back surgery)

Running through a standard dataset

Run machine learning algorithms to identify most important features

1)recursive feature selection

2)forward/backward/mixed selection

Concept is based around finding the feature that is most relevant, retaining that and iteratively adding other features. By retaining features, selecting statistically dependent components

Features X1, X2, X3, X4 to predict why. Which works best, X2, now we use X2 to predict Y, ranked X2, X1, X4, X3…by using factors that are less statistically dependent get better F score

Feature learning potentially more important than the algorithm

Using different machine learning algorithms. Using logistic regression is basically linear model doesn’t work as well as SVM because of the relationship between items that are non linear

Logistic regression- the hunt for the parabola

Decision trees will look at the region and…each split is

Neural networks are nonlinear model

3d separable.

Overfit training data you lose ability to predict.

Different guidelines: n to be 2p or more 22 people with thousands of samples

Principle component analysis

Orthogonal: new coordinate space. Vector going off at right angle

Clustering based on features in genetics

Simplest model

Summary Statistics

Analysis of categorical data

Logistic regression

Model:

Features most predictive