Data-Sharing and Usage Agreement

Spanish Back Pain Research Network

This agreement establishes the terms and conditions under which the Spanish Back Pain Research Network, henceforth referred to as the Data Provider, will release data to Dr Bernard Liew (University of Essex, UK), and Dr David Rugamer (Ludwig-Maximilians-Universität München, Germany), henceforth referred to as the Data Recipients.

1. Confidentiality of data pertaining to individuals will be protected as follows: The Data Provider will provide pseudonymized data, with a detachment of fields/identifiers from the personal data record
2. The Data Recipients will not release data to a third party or deposit data in a public data repository without prior written approval from the Data Provider. Failure to maintain confidentiality may result in legal action and termination of authorization to access the data.
3. The Data Recipients will not share, publish, or otherwise release any findings or conclusions derived from analysis of data obtained from the Data Provider without prior written approval from the Data Provider.
4. Data transferred pursuant to the terms of this Agreement shall be utilized solely for the purposes detailed below in the "Data summary" section. Any modifications or changes in the project purpose, scope, or design will require an updated version of the "Data summary".
5. All data transferred to the Data Recipients shall remain the property of the Data Provider. Upon termination of this agreement, or completion of the project, the Data Recipient would fully delete and/or destroy any data covered by this agreement.
6. The Data Recipients agree to share any codes and derivative works with the Data Provider, upon request from the latter.
7. This agreement will be effective upon signature, and will be valid for a period of four years, unless terminated earlier subject a written notice issued by either party thirty days in advance. Upon termination, both parties can also agree to renew this agreement, yearly.

Key roles in the collaboration include:

- Development of research planned analyses (BL, Spanish Back Pain Research Network)
- Statistical modelling analyses (BL, DR).
- Review of results and refinement of the analyses (BL, Spanish Back Pain Research Network)
- Write up of methods and results (BL, DR)
- Write up of introduction and discussion (BL, Spanish Back Pain Research Network, by negotiation)
- Completion of first draft, responsibility for paper submission and 1st authorship (Spanish Back Pain Research Network, BL, by negotiation)

Data Summary

Provision of data in the form of an excel spreadsheet which contains the data analysed in the study "Predicting the evolution of neck pain episodes in routine clinical practice (BMC MusculoskeletDisord 2019. 2019;20:620https://doi.org/10.1186/s12891-019-2962-9), namely, the following information for 3,001 patients who requested care for neck pain:

- Age
- Gender
- Employment status
- Duration of the pain episode
- Duration of the pain episode categorised
- Time since first episode
- Baseline intensity of neck pain
- Baseline intensity of arm pain
- Baseline disability
- Improvement of neck pain
- Improvement of arm pain
- Improvement of disability
- Diagnostic procedures during the episode (X-Ray, MRI, CT scan, EMG, Other)
- Imaging findings (Disc degeneration, Facet joint degeneration, Scoliosis, Spinal stenosis, Disc protrusion, Disc herniation (extrusion), Other findings, No findings);
- Clinical diagnosis (Spinal stenosis, Disc protrusion/Herniation, Nonspecific syndrome)
- Pharmacological treatments (Analgesics, NSAIDs, Steroids, Muscle relaxants, Opioids, Other)
- Non pharmacological treatments (Physical therapy, Rehabilitation, Neuro-reflexotherapy, Surgery)
- Other treatments (Rhizolysis, Epidural injections, Referral to pain unit, Other treatments)

Additional databases from other studies may be subsequently added as deemed necessary by the persons named below.

The intent is to publish the results of any analyses in an academic journal. Data will be provided in a deidentified manner. Additional variables may subsequently be added as deemed necessary by those named below.

Proposal 1

The first proposal is to compare different state-of-art machine learning methods in their predictive performance over the standard method of logistic regression. The methods vary from simple such as penalized logistic regression (LASSO), to methods with greater flexibility such as support vector machines, extreme gradient boosting, random forest, and neural networks. Depending on preliminary data exploration, the data may be split into a training (80%), and testing set (20%). The training set is for hyperparameter tuning for each method and internal validation using resampling methods, whilst model performance (area under receiver operating curve [AUC]) will be determined on the testing set.

Proposal 2

Variable selection is an important pre-processing step regardless of any statistical methods used. The predominant approach is using step-wise regression techniques. Other methods exist within the machine learning framework, namely: filtering methods, wrapper methods, and embedded methods. **Filtering**: An external algorithm computes a rank of the variables. Then, variables are subsetted by a certain threshold criteria. The selected variables will then be used to fit a model. **Wrapper Methods:** Here, no ranking of variables is done. Instead, an optimization algorithm selects a subset of the variables, evaluates the set by calculating the resampled predictive performance, and then proposes a new set of variables (or terminates). A simple example is the sequential forward selection. This method is usually computationally very intensive as a lot of models are fitted. After undertaken all of these steps, the final set of selected variables is again fitted. **Embedded Methods:** Many learners internally select a subset of the variables which they find helpful for prediction. This proposal aims to compare different variable selection methods and determing which method provides the most parsimonious model with the best accuracy.

Proposal 3

Many state of the art machine learning methods, e.g. random forest, are “black box” models, making them unappealing to be used in a clinical context. The aim of this proposal is to illustrate how certain post-hoc methods can be used on black-box models to understand how each predictor influences the outcome prediction. Such methods may include, partial dependence plot, global surrogate, Shapley, can be used on a wide range of black-box machine learning methods.

Additional collaborators, as described below, could be involved depending on resources required for each analysis.
- Ana Royuela, PhD; Biostatistics Unit; Puerta de Hierro Biomedical Research Institute (IDIPHISA); Madrid, Spain

Data released to: Dr Bernard Liew (University of Essex, UK), and Dr David Rugamer (Ludwig-Maximilians-Universität München, Germany)

Agreement Completed by: Dr. Francisco Kovacs for the Spanish Back Pain Research Network.

Signed by Bernard Liew, on behalf of David

(Signature)

Bernard Liew

University of Essex, UK

4th March 2021

Signed by Francisco Kovacs in representation of the Spanish Back Pain Research Network

F-M. Kovacs

Francisco Kovacs, MD, PhD

Spanish Back Pain Research Network, Spain

12th March 2021