Project: Dry eyes

# Goals

2.1 Which are the most important signs or symptoms to represent the disease status of dry eye?

2.2 Given that there are no clear general correlation between signs and symptoms, is this always true or can we see any clear correlation if we choose only the subsets of signs and symptoms?

2.3 Can we identify ideal patient population to predict drug efficacy (Ikervis and Diquas) and to enrich patient population for future clinical development of other new candidates?

# Terminology

Sign, Symptoms and Lab Data

A sign is a condition or parameter that can be subjectively measured by someone other than the patient, for example body temperature. A symptom is something that can only be felt (subjectively) by the patient himself, like pain, dizziness, and itchiness in the eye. We have two types of sign data, one which we call lab data (that which were measured in a pathology lab), and related to chemical composition of blood and urine – for example, proteins, red blood cells etc. The other sign data are measurements of conditions which are directly related to the eye and measured in an eye clinic- for example, staining of cornea, the tear film break-up time, the volume of tears produced in unit time. For clarity, we will use three terms sign, symptom and lab data. Most of the sign and symptom data are integers on a scale of 0-3, where 0 is healthy and 3 is most severe diseased case. Only two signs, TBUT (Tear Breakup time) and Schirmer score (a measure of tear volumes in unit time) are not on this scale but raw measurements.

In case of eyes, the **eye sign data and symptoms data are more important than lab data.** But we want to explore the lab data as to find out if any lab data can show clear correlation to the dry eye disease.

*See Appendix 1 for Table of Sign and Symptoms. This is also contained in brief subset of pcode mapping.*

*Project*

In the context of clinical trial, a Project is the name given to trials for a specific drug along with control drug and Placebo. We have the following projects:

1. **Diquas**
2. **Ikervis**
3. DE-099(Diquas 10%)
4. DE-101
5. DE-110

The table shows the planned treatment undertaken by each project

Table 1 Number of patients under each planned Project subdivided by planned treatment.



*Baseline and Screening*

Typically, when a patient first visits the ophthalmologic clinic, basic measurements are done to see if he should be treated or enrolled in the clinical trial. This first visit is called screening. A few days later, typically two weeks later, he will go for his first treatment. This visit is called Baseline. For checking his status of improvement, the Baseline readings are the reference. Though different projects had different schedules for patients’ visits, they are typically the same within a project. Mostly, week2, week4 and week 6 visits’ data are available.

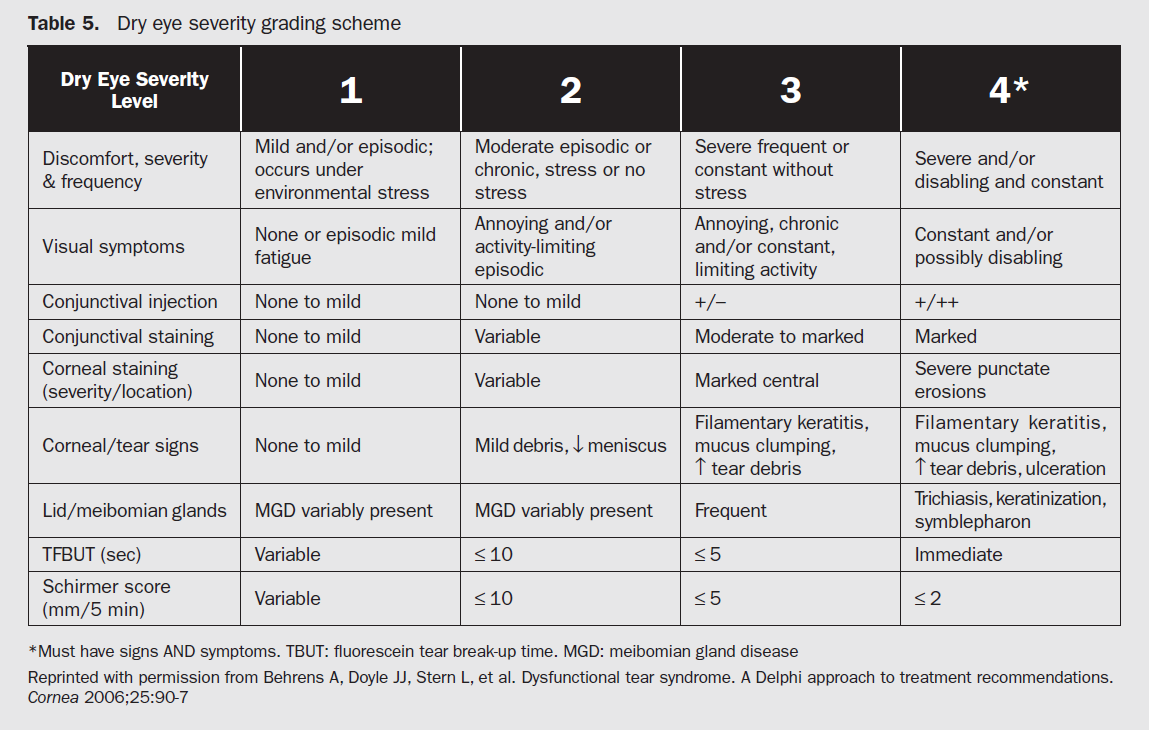
We will identify a record by the *Unique\_Subject\_Identifier* and *Analysis\_Visit\_N* (which is the numeric code of Week number). We have the dates of the Analysis Visit, but these are not important, everything is relative to Baseline, so you will see that Screening will have a negative number for the Analysis\_Visit\_N.

# Dry Eye Severity Score

For any supervised learning task, we will have predictors or features which are the independent variables and a response which is the outcome based on these features. This is often called a label. The goal of the machine learning program is to learn from the set of predictors and labels so that it can classify new, previously unseen examples into the correct category(class) based on only the knowledge of the predictors for these examples, and on the previous set of predictor-response on which it was trained. For example, in a Diabetic Retinopathy classification task, we might be given a large number of images of the retina which have been graded by a trained lab technician on an integer scale of 0-4, where 0 indicates a healthy eye and 4 indicates sever diseased status. These labels are assigned not automatically but by a human. Then, we will train our classifier by presenting each image and its label (0, 1, 2, 3, or 4) several times. Finally, we will hope that when a new image is input to our classification program, it can assign a label to it which denotes the disease status. And that over several such images, it will correctly classify with high accuracy, few false positive and false negatives.

In our current study, we do not have a subjective label for each eye. Instead we have a report called DEWS 2007 Report in which Table 5 on page 88 shows how to calculate the severity status of the Dry Eye. This is not an ideal situation but this is all that is feasible in this study.

The Dry Eye severity score is the average of nine category scores. Each category score is itself the average of a few related predictors. Two of the categories are based on symptoms, and the other seven on sign data. The details of this calculation are given below.



**Basic Description of the Data Files**

|  |  |
| --- | --- |
| result\_comb\_lr\_v2.csv | * Data frame with signs and symptoms data, but not lab data. * The rows represent “each patient per visit” data. * The left and right eye readings are averaged. * Some variables with different names but which measure the same condition is combined. |
| result\_comb\_lr\_v3.csv | * Using result\_comb\_lr\_v2, we combine some more variables for dimension reduction. * Not too different from result\_comb\_lr\_v2 |
| result\_comb\_lr\_v4.csv | * Uses result\_comb\_lr\_v3 as its base. * TBUT and Schirmer are converted from continuous scale to integer valued 0-3 scale. * Using k-nearest neighbors, we impute missing values where we can. |
| result\_lab.csv | * Data frame of lab data. |
| result\_comb\_lr\_v2\_lab.csv | * Data frame with sign, symptom and lab data, made from result\_comb\_lr\_v2 and result\_lab. |
| pcode\_mappingdemapping.csv | * Parameter codes and description for all the predictors(sign, symptom and lab data) |
| LabDataMappings.csv | * Parameter codes and description lab data. * This file is now redundant since pcode\_mappings.csv contains this info. |
| dryEyeStatus5\_score\_table.csv | The severity readings based on result\_comb\_lr\_v2.csv |

**Some Guidelines for Modeling**

## Multivariate Statistics:

* 1. Correspondence Analysis in R
  2. Polychoric Correlation in R (this is not multivariate in a strict sense)
  3. Canonical Correlation

## Python Guidelines:

* 1. Use Python 3.5
  2. Create virtualenv called de with the following packages installed: Numpy, Scipy, Pandas, Matplotlib, scikit-learn, readLine.
  3. Use source code version control system like Tortoise SVN or Git.

## Suggested Approaches for Feature Engineering and Dimension Reduction

* 1. PCA
  2. ICA, R-ICA
  3. t-SNE
  4. Gaussian Random Projection
  5. Autoencoder
  6. SVD

The outputs of dimension reduction are used as features, which become inputs for the classifiers.

## Schemes for Classification:

Majority Vote of

1. Random Forest
2. XGBoost
3. Neural Network
4. Softmax (multinomial logistic regression)
5. k-Nearest Neighbours
6. Support Vector Machine

## Examining Model Fitness

* 1. Use k-fold cross validation. Use accuracy, F1 score, Youden’s J, Mathew’s correlation and any other similar scores like precision and recall.
  2. Always plot the confusion matrix.
  3. Show the error rate curve as a function of different parameters.
  4. Use GridSearchCV for parameter tuning.
  5. Examine predictor significance and importance.
  6. Make sure the classes are not too imbalanced.

# Guidelines for Documentation

Presentation of results is next in importance only to a strong and meaningful result. In the absence of a perfect result, presentation occupies first place and the result, second. We should aim to include as many *relevant* charts and graphical elements as possible, taking care that do not look forced or out of place.

While inserting captions for tables and figures, use APA style. Table captions should be placed above the table and figure captions, below the figure.

# Appendix 1: Brief Subset of Parameter Codes and their Dry Eye Category

