MATH411 | Fall 2018 | Exam II

Your Name Here

Monday in class, 11/16/2018

Background

The present study in this exam involves a pulmonary function parameter called **DLCO** or Lung diffusing capacity for carbon monoxide (CO). This is a clinical parameter for evaluating the gas exchange function of lungs. **DLCO** is measured by making the subject inspires a gas mixture that contains Helium and CO then measuring the partial pressure difference between inspired and expired CO after 10 seconds of breath-hold. **DLCO** is defined as volume of CO diffused into lung capillary blood during 1 minute for each pressure gradient unit (ml/min/mmHg or mmol/min/kPa). **DLCO** could help detecting the respiratory diseases such as COPD, lung fibrosis, pulmonary hypertension...

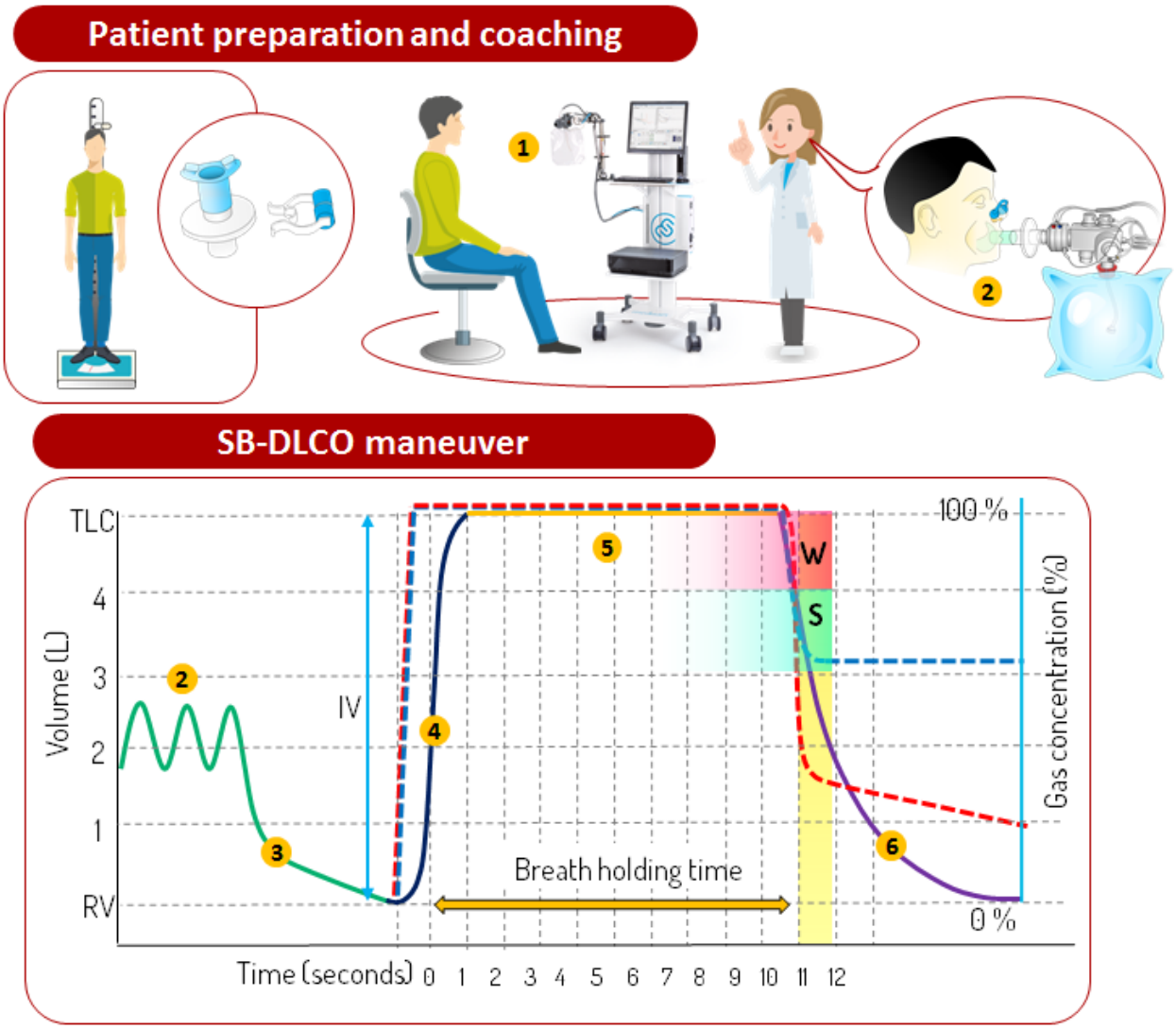

Figure 1: Background of DLCO

This study implies a real dataset of DLCO measured in 487 healthy Caucasians. **Our goal is to develop a predictive model that allows to estimate `DLCO` values by `Gender` (Male or Female), `Height` (cm) and `Age` (year)**. It could also be considered as a prediction of the mean DLCO value (mean predicted) of a virtual population of many peoples who are characterized by the same gender, age and height values.

Your Task

As described above, your goal is to find the best model to predict the response variable, i.e., `DLCO` value. As a measure of the prediction ability of the model, let's use **RMSE**, which is defined as

$$RMSE = \sqrt{\frac{\sum_i^n \hat{y}_i - y_i}{n}}$$

. To achieve this task, you need to:

1. Split the data into two sets, i.e., train set and test set. Let's do so by using 80% of the data as the **train set** and the remaining 20% as the **test set** (When you are splitting the data, please use `set.seed(2018411)`). Then the **RMSE** can be calculated based on the **test set**.

2. The you'll use the train data set to find the best model. There are multiple things you need to consider, for instance, if the response variable needs to be transformed; whether or not a predictor needs to be included into the model and in what form, etc...

3. To submit:

3.1 You only need to provide your final model, i.e., your best model along with the **RMSE** value. **But you need to describe your strategy of how you achieve this model**.

3.2 Save and submit the **test set** as a `csv` file. Within the file add a new column that contains your predicts.