The VADA Program
Visual and Automated Disease Analytics
Graduate Training Program

Natural Sciences and Engineering Research Council of Canada
Conseil de recherches en sciences naturelles et en génie du Canada
Canada

Multi-SNP prediction model for lung function decline

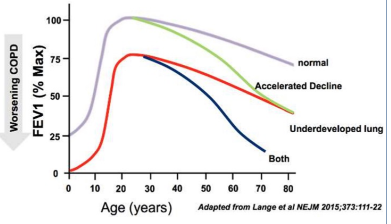
Yushan Hu 13/06/2022

University of Victoria

Background

- Chronic obstructive pulmonary disease (COPD) is a progressive lung disease.
- Forced expiratory volume in 1 second (FEV1) and COPD exacerbation rate are widely used measurements of the progression of COPD.
- Understanding individual level longitudinal progression of disease is to understand disease pathogenesis.

Figure 1. Different Pathways to COPD



Adapted from Lange et al NEJM 2015;373:111-22

Introduction

- **Our Problem:**

  Predict COPD clinical outcome using genetic information and baseline characteristics

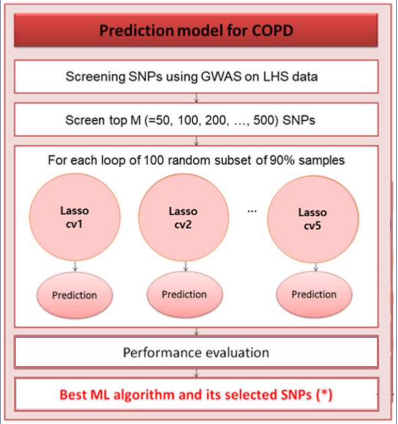
- **Our Goals:**
  - ➢ Build the scoring system from fitted model (called ensemble model)
  - ➢ Use score system to help experts to make decisions if a given patient should receive treatment
- **Our Approach:**
  - ➢ Use the machine learning approach to discover important interactions of treatment.
  - ➢ Discover multi-SNP marker that can predict patterns of FEV1 trajectories.
  - ➢ Validate with ECLIPSE data by trying to find the association of this score with other clinical outcomes.

Dataset

LHS (Lung Health Study)

- Usual Care / Smoking Intervention and inhaled bronchodilator/ Smoking Intervention and inhaled Placebo
- 5,887 participants with early stages of COPD
- Variables: age, sex, pack-years of cigarette smoking, smoking status, genetic information, etc.
- The level of FEV1 was measured for each patient for 5 years after their first visit

Methods



Prediction model for COPD

Screening SNPs using GWAS on LHS data

Screen top M (=50, 100, 200, ..., 500) SNPs

For each loop of 100 random subset of 90% samples

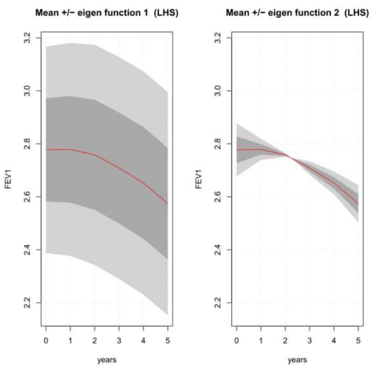
Lasso cv1 → Prediction
Lasso cv2 → Prediction
...
Lasso cv5 → Prediction

Performance evaluation

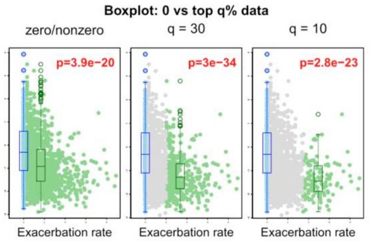
Best ML algorithm and its selected SNPs (*)

- Derive outcome variables for each patient
- Fit linear model in GWAS
- Select M important SNP candidates by the significance of their effects on the outcome
- Fit multi-SNP prediction models of M SNPs with lasso regression
- Prediction performance evaluation by MSE
- Get the best LASSO models
- Build the ensemble model by averaging the coefficients of selected SNPs from the best LASSO model
- Compare the MSE of ensemble model and single-SNP models

Results



- FEV1 trajectory over the year 0 to 5 (red line) and the effect of the fPC scores (shaded area) for the first two fPC scores (sc5.1, sc5.2) on LHS (without UC）
- **sc5.1** —> overall FEV1 level from year 1 to year 5
- **sc5.2** —> FEV1 decline speed from year 1 to year 5



- The blue points represent when the exacerbation rate is 0, and the green points represent the q% quantile of the exacerbation rate.
- The rest of the points are colored as grey. The numbers in the plot are the Wilcoxon p-value between the two boxplots.

Conclusion

- **The effect of the ensemble model**
  - ➢ The scores calculated from the ensemble model with M=300 SNPs can estimate the outcome variables better than the single lasso model.
- **Prediction of COPD exacerbation rate**
  - ➢ Prediction of exacerbation rate by scores: highest |correlation| = 0.31.
  - ➢ Classification of exacerbation rate by scores: clear separation of scores between high/low exacerbation rates.
    - ✓ can use our ensemble model to classify patients at row risk or high risk of COPD exacerbation.
- **The effect of genetic information**
  - ➢ When comparing the model with and without the genetic information, there were some increases in the correlation when including the genetic information.

Literature Cited

Nicholas R Anthonisen, John E Connett, James P Kiley, Murray D Altose, William C Bailey, A Sonia Buist,William A Conway, Paul L Enright, Richard E Kanner, Peggy O', et al. Effects of smoking intervention and the use of an inhaled anticholinergic bronchodilator on the rate of decline of fev1: the lung health study. Jama, 272(19):1497–1505, 1994