Louie Alexander

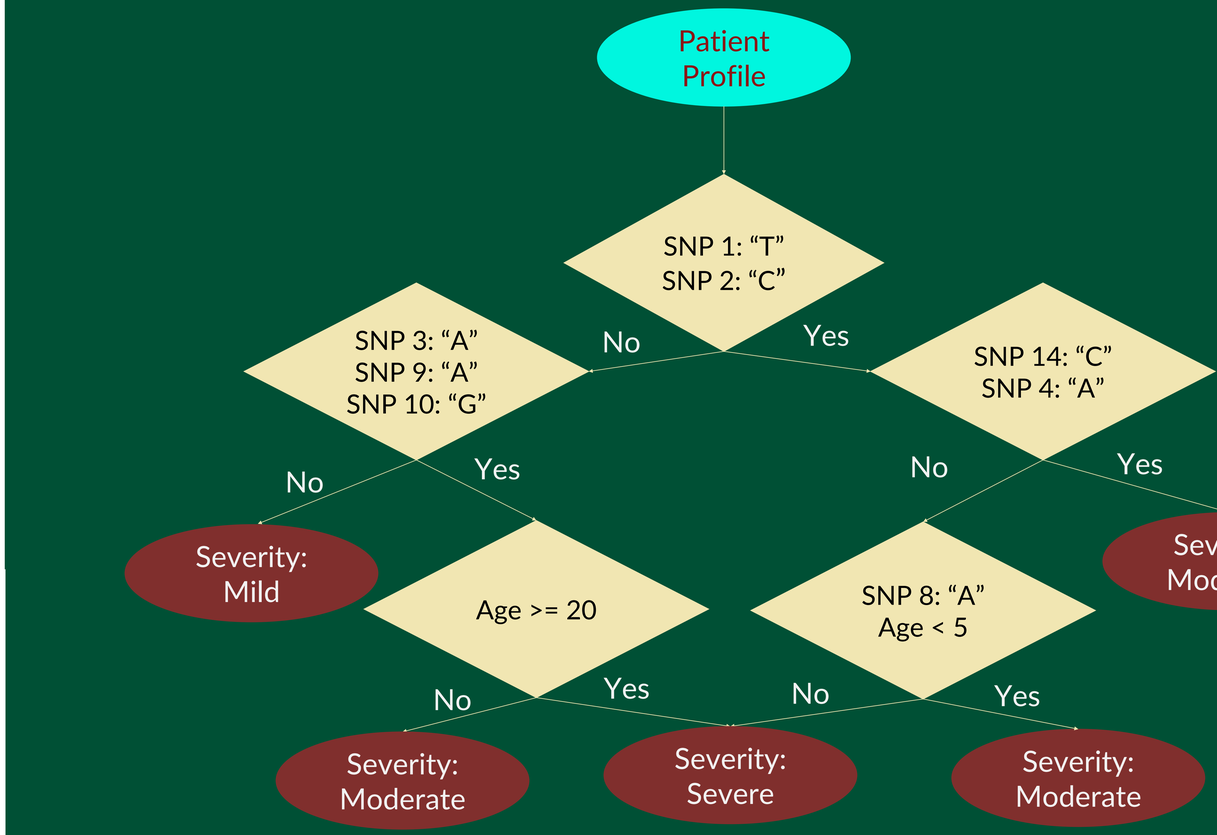
Random Forest can Identify Variables that Affect Cystic Fibrosis Severity

BACKGROUND:

Cystic fibrosis is a disease characterized by poor lung function due to mutations in the cystic fibrosis transmembrane conductance regulator (CFTR.) The effects of mutations in CFTR have been described; however, individuals with the same CFTR genotype can have differing symptoms and severity.

Other regulatory and functional genetic elements likely contribute to these differing phenotypes. However, it can be hard to identify all genes or SNPs that contribute to the disease and *in-vitro* research on one gene or SNP is time-intensive.

Random forest can direct *in-vitro* research by identifying features that contribute strongly to distinguishing between phenotypes of cystic fibrosis severity.



Database of SNP data from cystic fibrosis patients

Train Random Forest to classify cystic fibrosis severity

Identify demographic and SNP variables that best distinguish CF severity

Validate SNPs associated with disease and find mechanism

Tutorial of randomForest in R:

Can also use:

- **randomForest (R)**
- **cForest (R)**
- **Caret (R)**
- **RandomForest (python)**
- **scikit-learn (python)**

Other Applications:

- Classification or Regression
- Diagnostic aid
- Improve sedative use for veterinary surgery
- ID drug responders vs non-responders
- ID variables that differ markedly between pheno/genotypes
- Direct *in-vitro* research to validate "important" biological variables

Results:

- Presence/Absence of SNP would be binary
- Can use VIM to determine impactful SNPs
- Need to consider whether SNPs are correlated
  - ❖ Linkage disequilibrium
  - ❖ Could select 1 SNP per block
  - ❖ Could use permutation VIM
- Permutation VIM can better separate correlated predictors
- If traditional VIM are used, correlated SNPs may still show high importance and point to functional significance or a functional block

VIP: Heart Disease