Random Forests and Clustering

Clinical Phenotypes in an Adult Cystic Fibrosis Population

Barbara A. Bailey [1] Douglas J. Conrad [2]

[1]Department of Mathematics and Statistics, SDSU

[2]School of Medicine, UCSD

Outline

- Introduction to Trees and Random Forests
- Making Sense out of a Forest
- Clustering
- CF Clinic Data

Adult Cystic Fibrosis Clinical Phenotyping

The Application of the Statistical Learning Algorithm, Random Forest, to the Generation and Description of Adult Clinical Phenotypes

Douglas Conrad
University of California San Diego

Barbara Bailey
San Diego State University

March 6, 2015

What is Statistical Learning?

- In artificial intelligence, machine learning involves some type of machine that modifies its behavior based on experience.
- In statistics, machine learning uses data to learn.
- Training data: (y, x)'s
  Two types: supervised and unsupervised learning

Some Goals of the Statistical Analysis

- *Classification:* Group data based on predetermined classes, develop criteria for distinguishing between classes (Supervised Method)
- *Clustering:* Discover reasonable groupings within a dataset (Unsupervised Method)
- *Variable Selection:* Reduce the number variables required to perform a classification or clustering task, determine interrelationships between variables (can be Supervised or Unsupervised)

Example: South African Heart Disease Data

- 462 observations on males in South Africa
- Variable of interest is congestive heart disease where a 1 indicates the person has the disease, 0 he does not
- Explanatory variables include measurements on blood pressure, tobacco use, bad cholesterol, adiposity (fat %), family history of disease (absent or present), type A personality, obesity, alcohol usage, and age

- Question: How could you find the best predictors of heart disease?

Trees

- What is a tree?
- Tree-based algorithms (recursive partitioning, etc.)
- How to grow (and prune) a tree?
- Example: South African Heart Disease Data

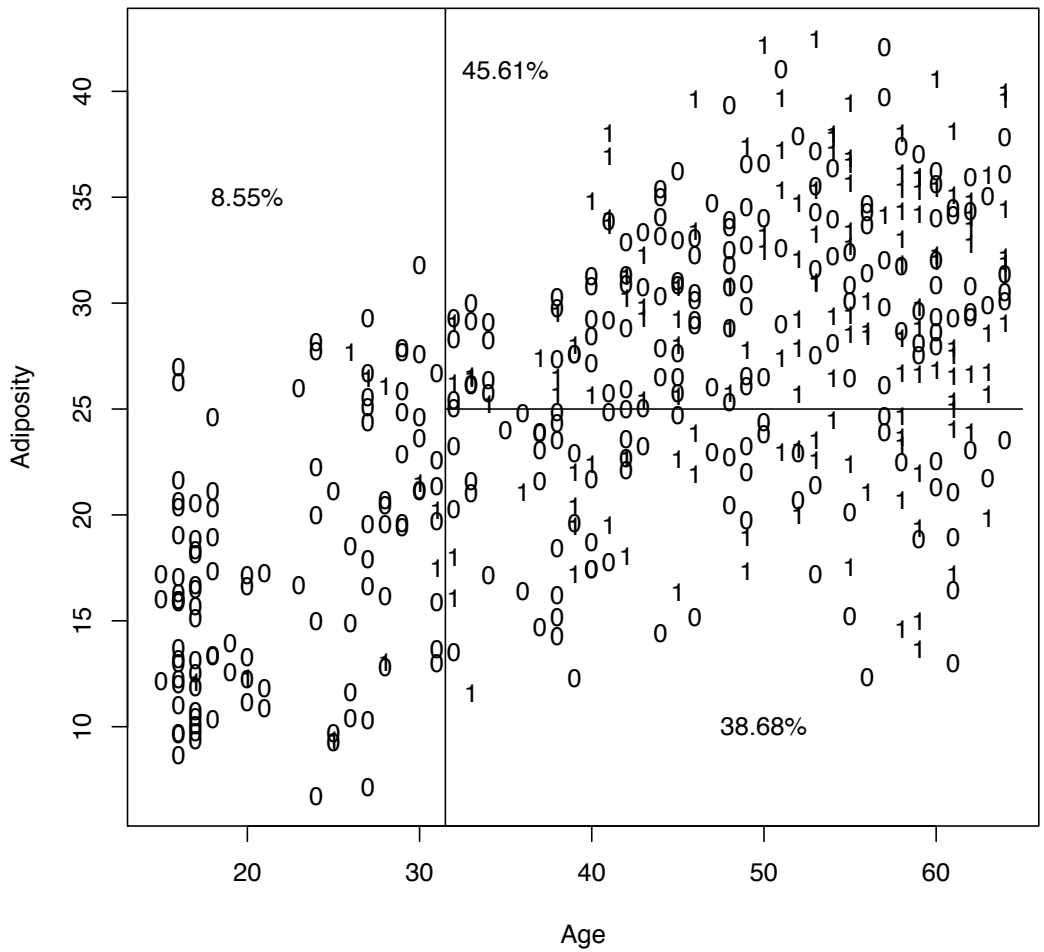


Figure 6.1: Splitting on age and adiposity

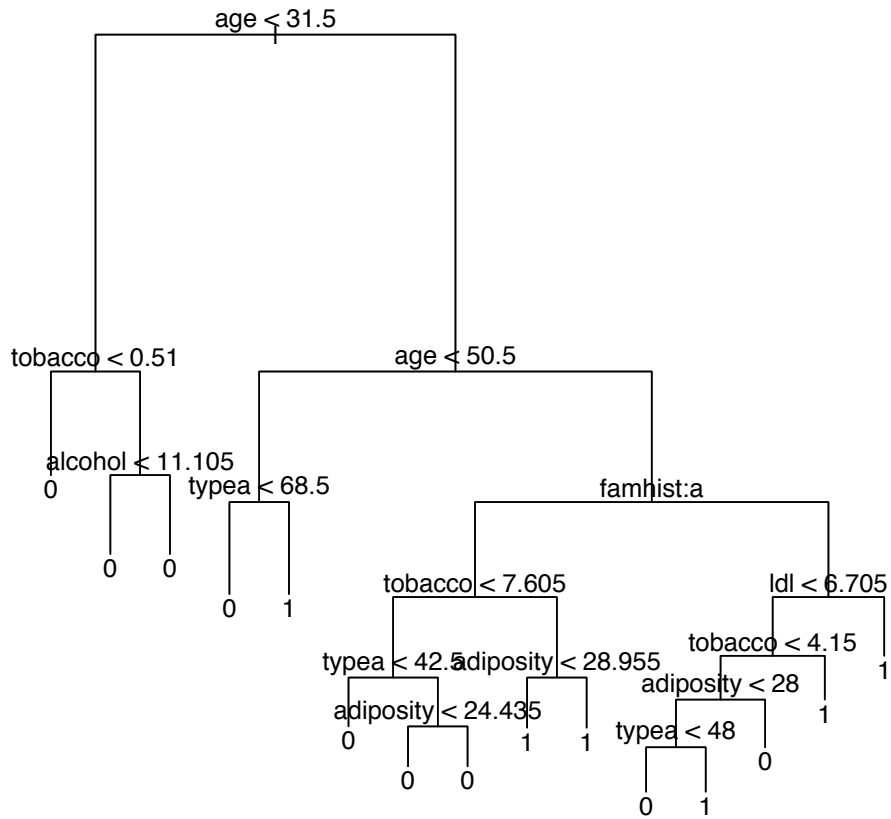


Figure 6.3: A large tree, with classifications at the leaves

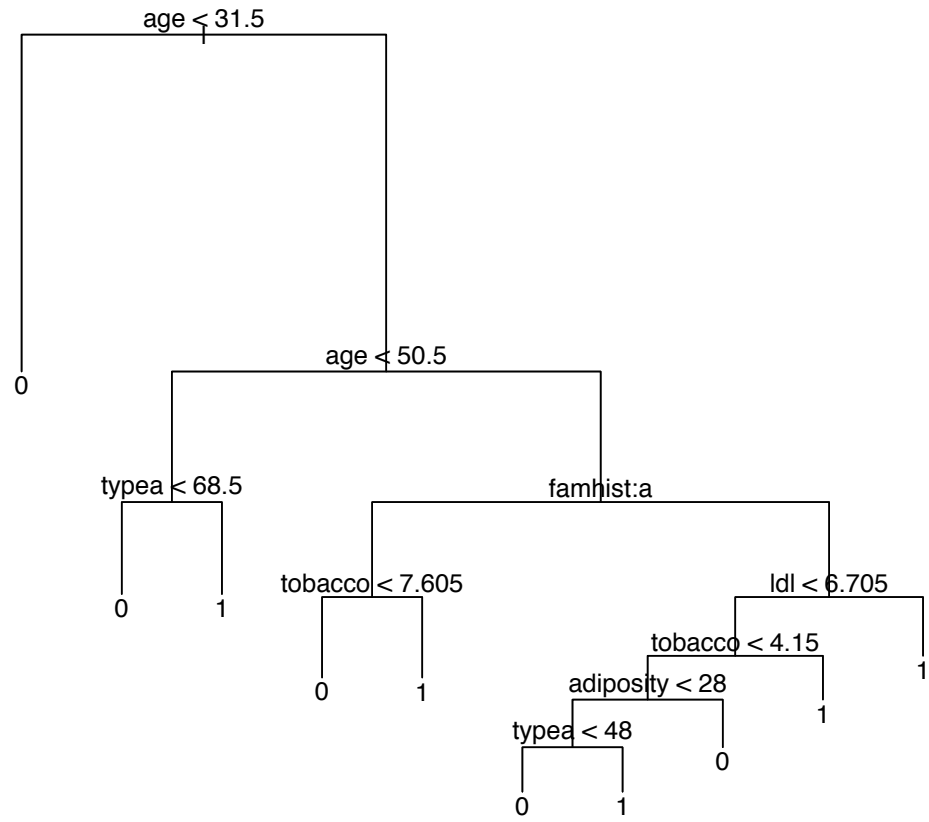
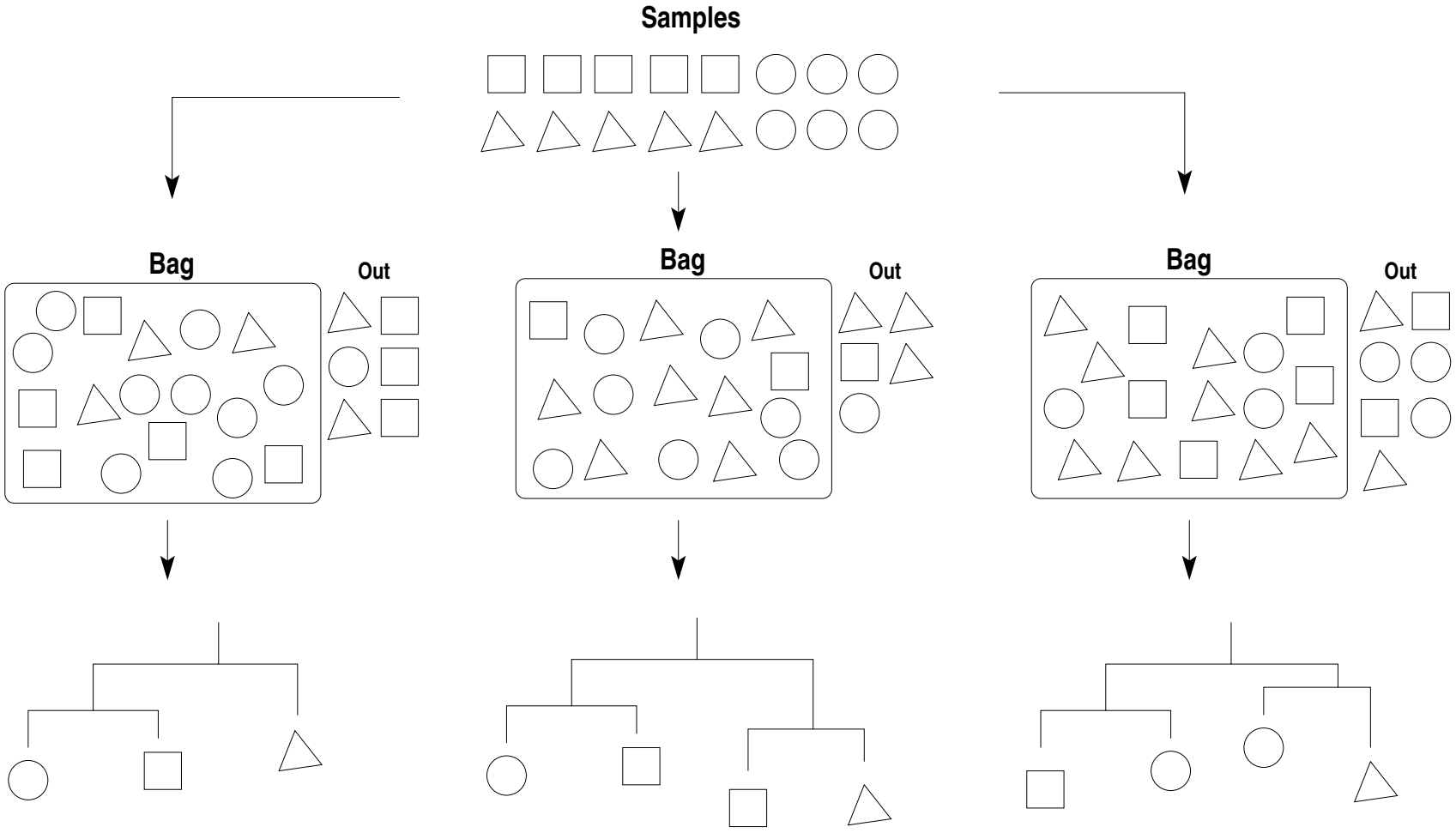


Figure 6.4: The tree, with unnecessary branches snipped

Random Forests

- A Random Forest is a collection or ensemble of trees.
- Each tree in a Random Forest is generated from a different bootstrap sample (sampling with replacement) of the data.
- Each node or split in each tree is determined from a random subset of all the variables.
- Instead of classifying new data by tree branching rules, Random Forest classifies by vote of its component trees.
- RF is an example of bagging or **b**ootstrap **ag**gregation. Unsampled data in each set are called *out-of-bag*.

Random Forest Generation



Why use a Random Forest?

- Performs well compared to many other classifiers.
- User friendly - has 2 parameters (the number of trees in the forest and the number of variables in the random subset at each node).
- Growing trees is fast!
- Easily handles all types of predictor variables.
- Available R package - free!

Supervised and Unsupervised Random Forests

A Random Forest can be supervised or unsupervised.

- Supervised:
  - In a supervised Random Forest, groupings for the training data are input to the algorithm.
  - Estimated classification error is computed using out-of-bag data.

RF: Variable Importance

Random Forest reports which variables are most important during construction. Particular variables are considered more important if:

- The accuracy of prediction of a sample is diminished when that particular variable in the sample is replaced with random noise during error analysis.
- The nodes of the trees become more homogeneous when that particular variable is used.

Unsupervised Random Forests

An unsupervised RF can be used to estimate a proximity matrix for clustering.

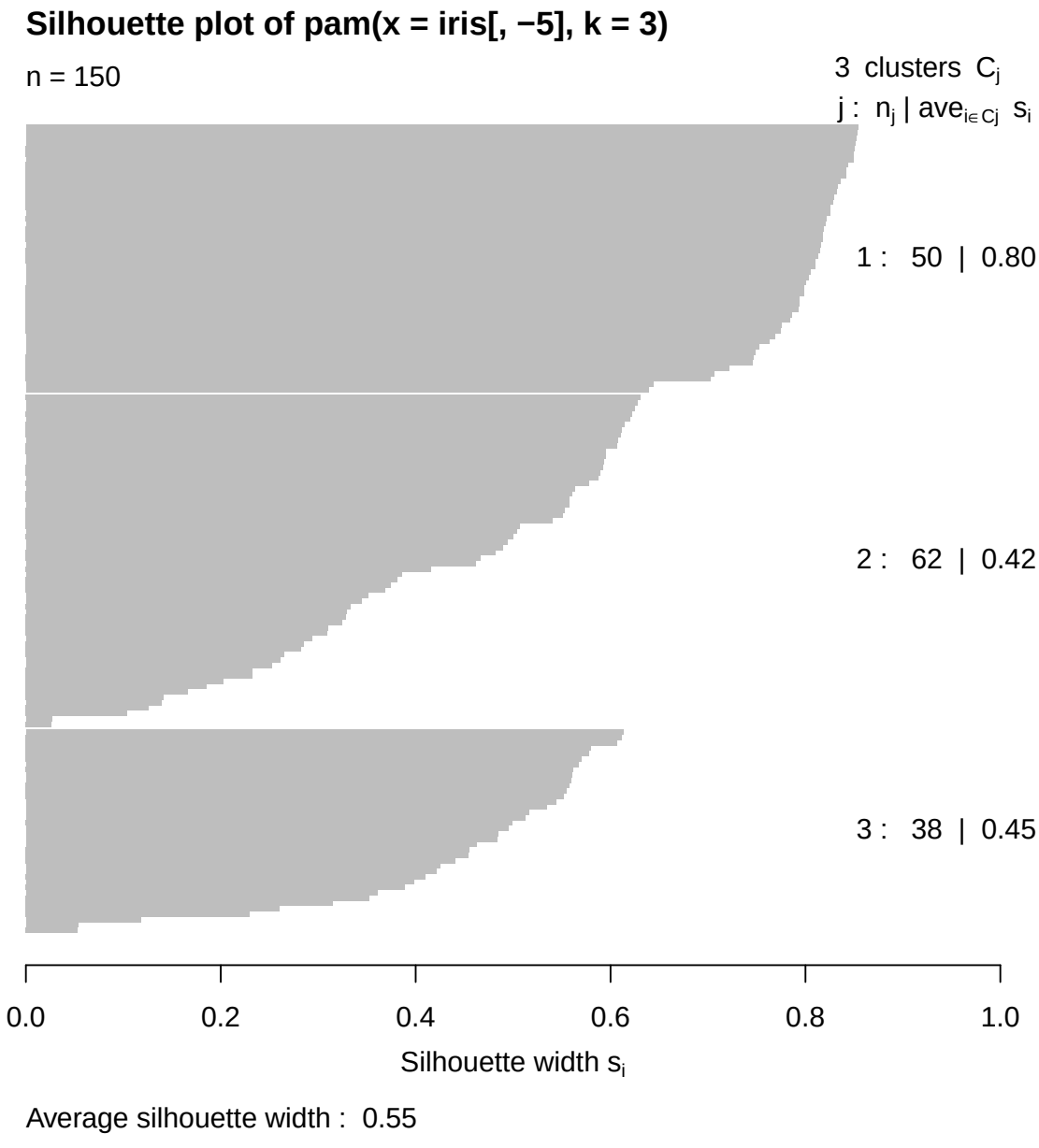
- The (i, j) element of the matrix is the fraction of trees that i and j fall in the same terminal node.
- Trick:
  - Call original data "class 1".
  - Generate synthetic "class 2" data by sampling uniformly within the range of each variables.
  - Use supervised RF on the above 2 classes to estimate the proximity matrix.

Clustering with the Proximity Matrix

- We choose Partioning around Medoids (PAM)
  - Similar to k-means but uses the median.
  - More robust to outliers and noise.
  - Choose the "best" number of classes using silhouettes.

Silhouettes

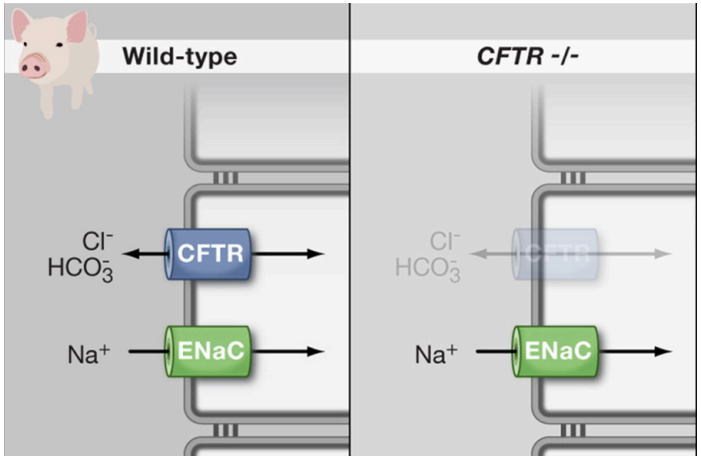
- Can be used with any clustering algorithm.
- Description for each proposed clusters number k:
  - For each data point, first find the average distance between it and all other points in the same cluster.
  - Then find the average distance between the data point and all points in the nearest cluster.
  - The silhouette coefficient for each data point is defined as the difference between the above, divided by the greater of the two.
  - Use the average silhouette coefficient to obtain an "overall" measure.
- Calculates a measure of dissimilarity (so high is good).
- Use average silhouette plot over a range of the number of clusters k to determine best number of groups.


Silhouette plot of pam(x = iris[, −5], k = 3)
n = 150
3 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 50 | 0.80
2 : 62 | 0.42
3 : 38 | 0.45
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.55

References for Random Forest (and more)

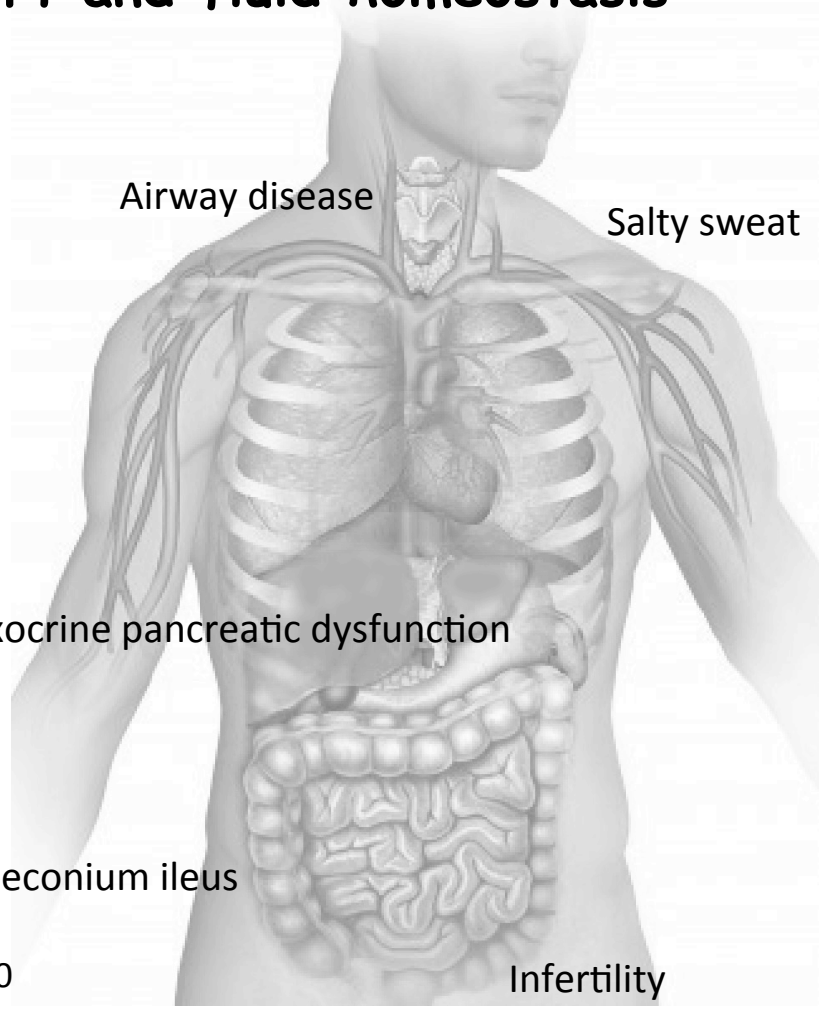
- Leo Breiman's webpage: www.stat.berkeley.edu/ breiman/RandomForests/
- R package: randomForest
- Notes on Statistical Learning by John Marden

Defective Cystic Fibrosis Transmembrane Regulator affects anions transport and fluid homeostasis



Defective HCO_3^- transport

Loss of Cl^- conductance



Airway disease

Salty sweat

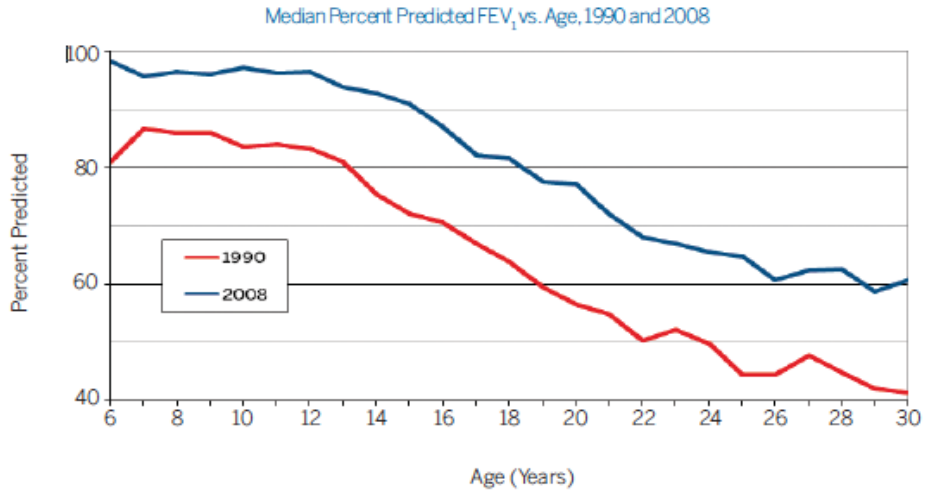
Exocrine pancreatic dysfunction

Meconium ileus

Infertility

Chen *et al.* 2010; Kopelman *et al.* 1988; Quinton 2010

Why Classify?



- Define disease risk
  - Prognosis/Risk Assessment
  - Optimization of Therapy.
    - Aggressiveness of therapeutic approach
    - Application of specific therapies (do all patients equally benefit from all therapies)
    - Combinations of Existing Therapies

Why Classify?

- Correlation of metagenomic/ metabolomic patterns with specific clinical classifications.
- Clarify genomic risk
  - Begin to define gene by gene
- Clarify genetic risk.
  - Better genotype-phenotype correlation

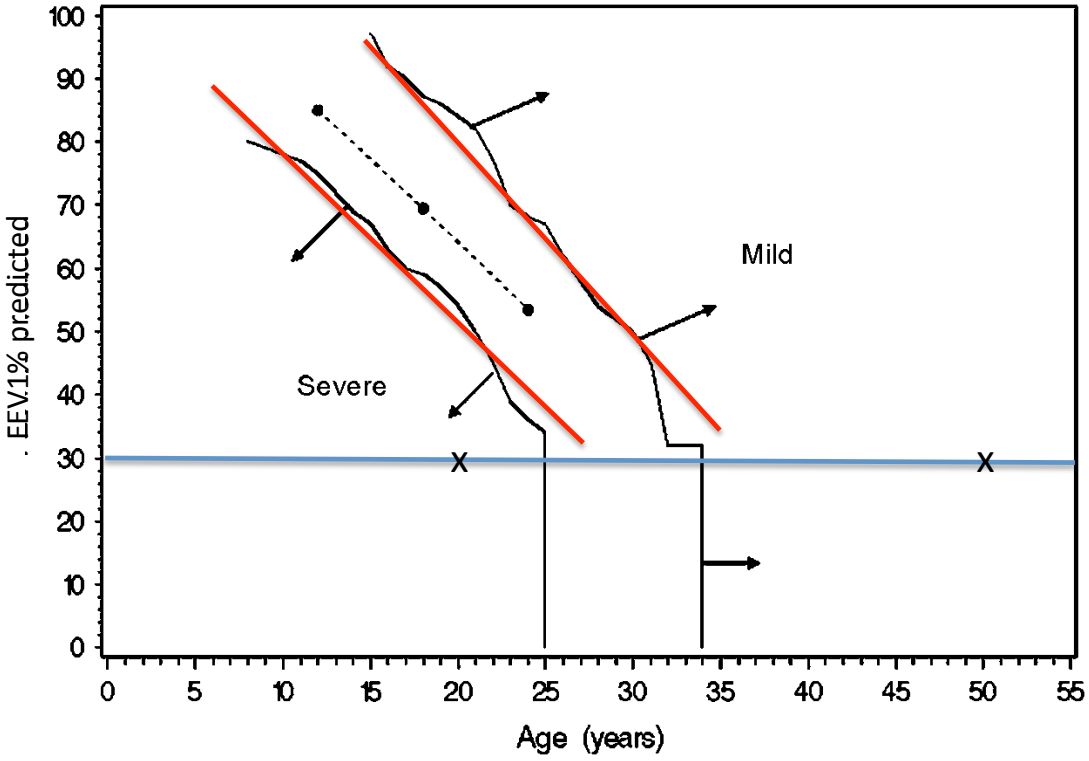
Why Multi-dimensional Phenotyping?

- CF is a multi-systemic disease
- Although CF mortality is driven by pulmonary disease, there are likely interacting phenotypes
- Nutrition/GI phenotype
- Gender Differences
- CFTR Genotype
- Microbiology

Available Data (Cross Sectional)

- Demographic
  - CFTR mutations/Class
  - Gender/Age
- Physiology
  - Best FEV1 and FVC in the past year
  - Age x %FEV1 predicted product
- Radiographic score (Brasfield)
  - Individual components and final score
- Microbiology standard micro studies
  - Simple taxonomy PA/MRSA/MSSA/Achromo/Cepacia/ Fungi, NTM
- Nutritional
  - Height/Weight/BMI/PI.PS status

Age*FEV1% product



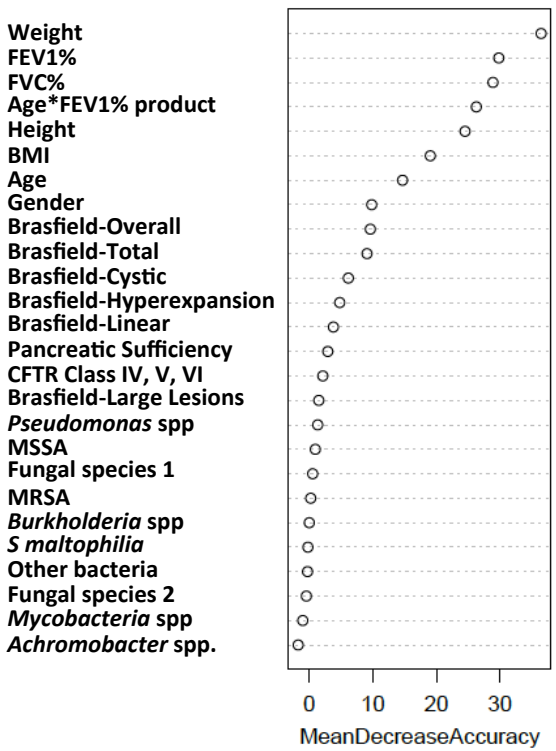
Schluchter et. el. Am J Respir Crit Care Med Vol 174. pp 780–786, 2006

Methods

- 1) The unsupervised Random Forest algorithm was used to generate a proximity matrix using all listed clinical variables.
- 2) Initial proximity matrix with PAM clustering generated the initial classes.
- 3) A supervised Random Forest analysis of the initial classes
  - Assess classification error rate
  - Variable Importance Plot
- 4) A dimension reduction strategy
  - minimize classification error rates,
  - capture the complexity of the multisystemic disease
  - generate clinically meaningful phenotypes
- 5) PAM clustering was repeated using the second proximity matrix.
- 6) Supervised Random Forest analysis of the new classes demonstrated the lower out of bag error rates and generated a confusion matrix

Variable Importance and Confusion Matrix Assessments

Variable Importance Plot (k=5)

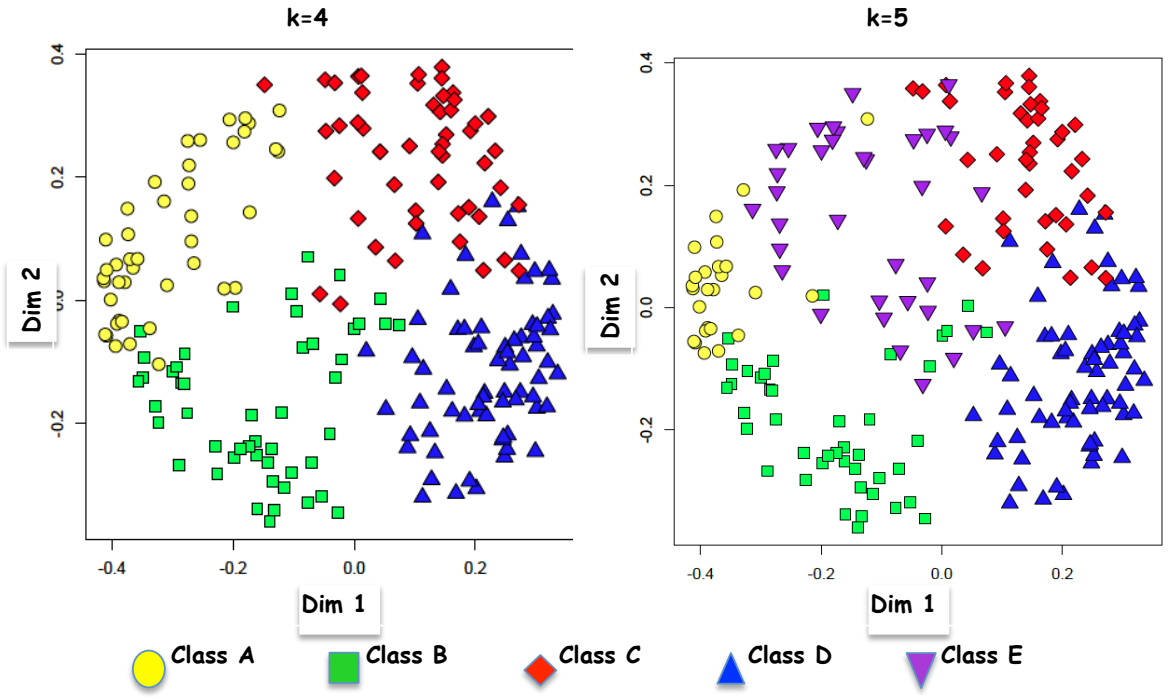


Dimension Reduction

Confusion Matrix

	Class	1	2	3	4		Class Error
k=4	1	40	2	1	0		0.07
	2	2	44	1	2		0.1
	3	1	2	43	4		0.14
	4	0	0	2	67		0.03
k=5		1	2	3	4	5	
	1	21	1	1	0	2	0.16
	2	2	38	0	1	2	0.12
	3	0	0	34	4	2	0.15
	4	0	0	2	65	0	0.03
	5	0	3	2	2	29	0.19

MDS Visualization of Clinical Phenotypes

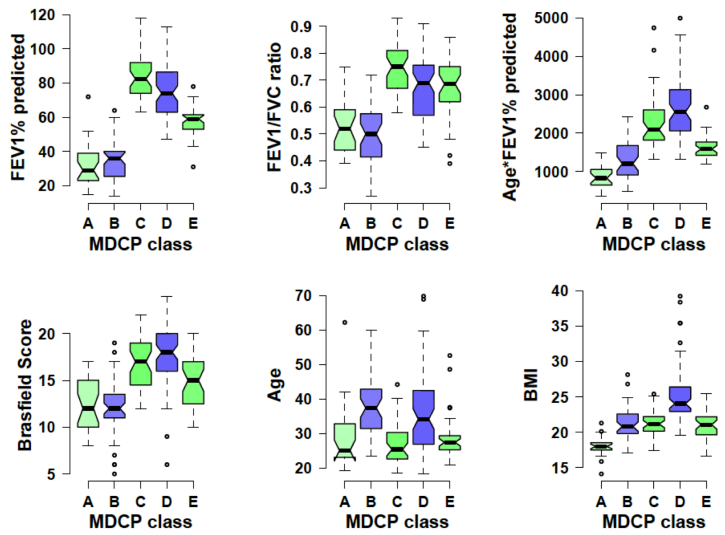


Phenotype Descriptions

A Clinical Phenotype Classes (k=5)

MDCP	FEV1%	Brasfield	Age	Age*FEV1% product	BMI	f Male
A (n=25)	32.0	12.4	28.7	861	18.1	0.36
B (n=43)	34.5	12.0	38.5	1312	21.2	0.74
C (n=40)	83.5	17.0	27.4	2299	21.3	0.23
D (n=67)	75.2	17.7	36.7	2682	25.3	0.79
E (n=36)	57.0	14.9	28.8	1610	21.0	0.25

B



Description of Classes

- Class A: low FEV1 % predicted, low Brasfield scores and not surprisingly lower BMI values.
- Class D: good nutrition, high age*FEV1 % product and dominated by male subjects.
- Class B: class of older, predominantly male patients (mean age of 38 were). This oldest class had intermediate lung disease health with age*FEV1 % products in the 1300-1500 range, severe airway obstruction, and median BMI values.
- Class C and E: The remaining classes were female dominated classes that varied mostly in terms of lung function and lung disease risk.

Description of Classes (cont.)

- We identified a much higher incidence of Pancreatic sufficiency (PS) in Class D and much lower rates of PS in the phenotype with the lowest age*FEV1 % predicted product (Class A, k = 4)
- We found enrichment of CFTR (Class IV, V, and VI) mutation in Class D (k = 4 and k = 5), consistent with their good lung function and well nourished phenotype, while lower frequencies of these mutations were identified in higher risk Class A phenotypes.

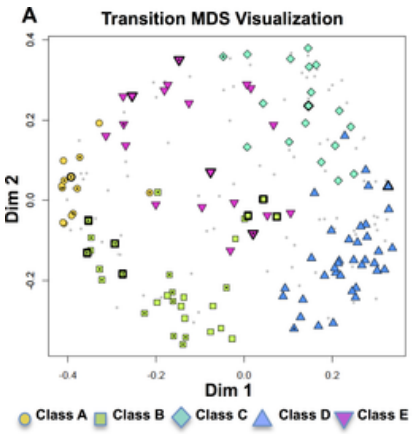
- Chronic pseudomonas (PA) infections have long been associated with accelerated loss of lung function in cystic fibrosis patients. Consistent with these findings, phenotypes with low lung function (Class A and B, k = 5 and Class B, k = 4) had much higher frequencies of PA airway infections while lower rates of PA in sputum were found in the medium risk, female dominated classes. (Class C, k = 5 and k = 4).
- Ecological interactions between fungal and bacterial populations are well studied but the clinical implications are not fully appreciated. We found two strong correlations between a) PA and Candida spp b) and Methicillin Sensitive Staphylococcus aureus (MSSA) and Aspergillus spp. with specific phenotypic classes that warrant further metagenomic and metabolomics studies.

MDCP correlation with clinical traits.

	MDCP Class	PS	CFTR Class IV, V VI	PA	Candida	PA/Candida	MSSA/ASP	Class Interpretation
k=4	A	<.05 (p=.002)						Low A*FEV1
	B			2.7 (p=.04)				Low A*FEV1, older, male
	C		.27 (p=.05)	.37 (p=.05)			4.2 (p=.03)	Median A*FEV1, female
	D	5.0 (p=.0007)	3.2 (p=.01)					High A*FEV1, older, male
k=5	A		<.05 (p=.04)	4.9 (p=.03)	4.4 (p=.005)	4.4 (p=.01)		Low A*FEV1
	B			3.5 (p=.02)				Low A*FEV1, older, male
	C			.25 (p=.001)			4.5 (p=.03)	Median A*FEV1, female
	D	5.2 (p=.0003)	3.4 (p=.01)		.17 (p=.005)	.20 (p=.01)		High A*FEV1, older, male
	E							Low A*FEV1, female

Two-tailed Fisher Exact Test: OR (p-value)

Fig 5. Clinical Phenotype Transitions.



B Class Transitions

Training Set(2014) MDCP Class

Initial (2011) MDCP Class	A	B	C	D	E	n	p Transition
A	11	2	0	0	1	14	0.21
B	1	25	0	0	1	27	0.07
C	0	2	17	2	1	22	0.23
D	0	1	1	35	0	37	0.05
E	1	1	0	0	17	19	0.11
f Transition	.14	.43	.07	.14	.21		

C Deaths/Transplants

MDCP Class Training Set (2014)	k=4	%	k=5	%
A	9	0.30	7	0.23
B	17	0.57	18	0.60
C	2	0.07	1	0.03
D	2	0.07	1	0.03
E	na	na	3	0.10

Conrad DJ, Bailey BA (2015) Multidimensional Clinical Phenotyping of an Adult Cystic Fibrosis Patient Population. PLoS ONE 10(3): e0122705. doi:10.1371/journal.pone.0122705
http://127.0.0.1:8081/plosone/article?id=info:doi/10.1371/journal.pone.0122705