Get real: A synthetic dataset illustrating clinical and genetic covariates

Ted Laderas[1], Nicole Vasilevsky[1,2], Melissa Haendel[1,2] , Shannon McWeeney[1], David A. Dorr[1,3]

[1]Department of Medical Informatics and Clinical Epidemiology, [2]Library, [3]Department of Medicine, Oregon Health & Science University, Portland, OR

Goals

- Develop a script to generate realistic synthetic datasets for hands-on learning in BD2K workshops

Learning Objectives

- Learn the difficulties and challenges of working with both clinical data and genetic data
- Learn the strengths/weaknesses of machine learning algorithms for classification in a “safe context”
- Highlight known issues with integrating clinical and genetic data

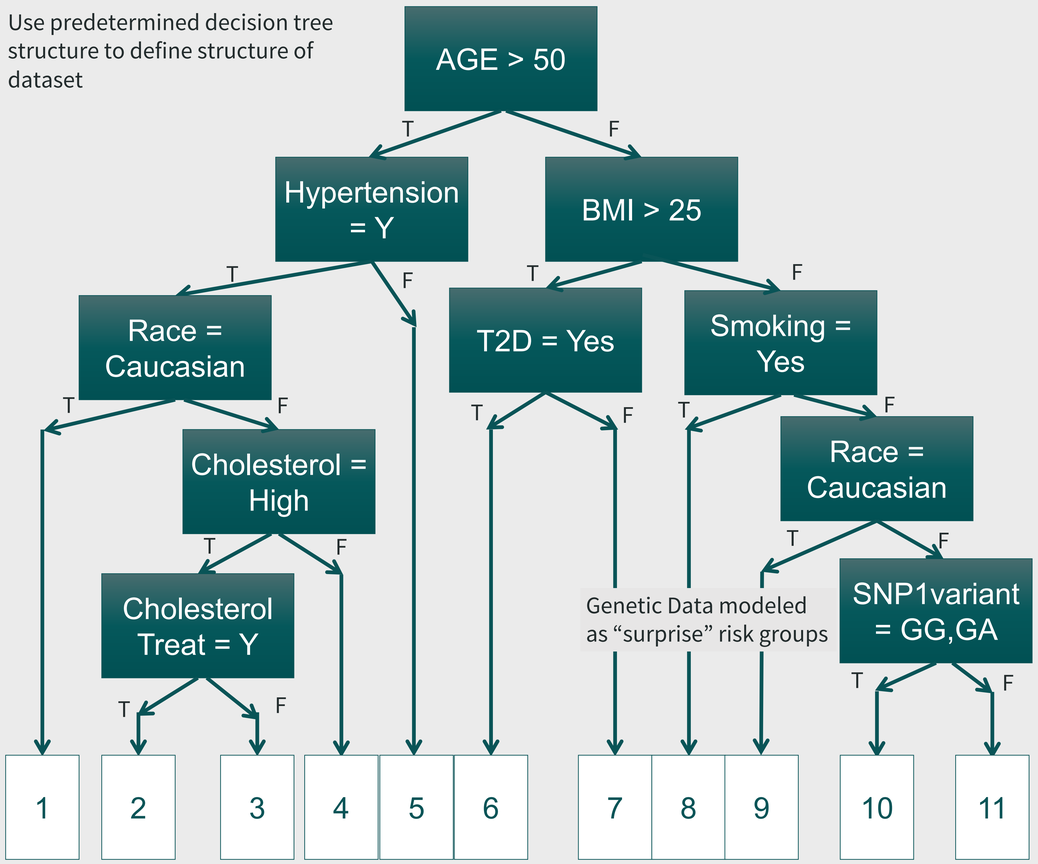
Scenario

- Task is a classification problem, where students predict (low or high) cardiovascular risk (< or =>7.5%/10 years), given the following data:
- ***Clinical Data*** derived from Electronic Health Record systems (EHRs)
  - Lab Values (LOINC)
  - Medications (RxNorm)
- ***Genetic data*** modeled as SNPs
- Task structure is to train a model using initial set, then test algorithm on (hidden) set
- To synthesize data, we use variable ‘importance’, defined as population affected and predictive impact determines level in decision tree

Variable	Importance/level
Age	1
BMI	2
Hypertension	2
Type 2 Diabetes	3
Race	4
Cholesterol	5
Gender	6
SNPs	6

Decision Trees for Generating Realistic Data

Use predetermined decision tree structure to define structure of dataset



Decision Tree branches define restrictions on sampling space for covariates

Decision Tree Structure defines what risk groups exist in dataset

Risk Group	Group Definition	Frequency	P(CVD=High)
1	Age >50,Hyp=Y,Race=C	0.310	0.81
2	Age >50,Hyp=Y,Race=NC, Chol=H, CholTreat=Y	0.001	0.10
3	Age >50,Hyp=Y,Race=NC, Chol=H, CholTreat =N	0.001	0.20
4	Age>50,Hyp=Y,Race=NC, Chol=Low	0.010	0.70
5	Age>50, Hyp=N	0.280	0.08
6	Age<50, BMI>25, T2D=Y	0.166	0.78
7	Age<50, BMI>25, T2D=N	0.220	0.20
8	Age<50, BMI<25, Smoking = Y	0.010	0.60
9	Age<50, BMI<25,Race=C	0.010	0.01
10	Age < 50, BMI<25, Race=NC, SNP1=GG,GA	0.001	0.20
11	Age<50, BMI<25, SNP1=AA	0.001	0.05

Frequencies of risk groups derived from real patient data

Difficulty of problem can be tuned by adjusting conditional probabilities of CVD risk association for each risk group

Risk group frequencies determine sampling probability of that risk group in synthetic dataset

Discussion Questions

- Decision Trees provide framework for integrating clinical and genetic datatypes, but they are overly simplistic.
- How difficult do we make the problem?
  - Tradeoffs between learning and difficulty
- How can we incorporate other techniques such as natural language processing (NLP) into this task?
- What are other scenarios we can model?
- What extra covariates should we include for clinical and genetic data?
  - Should they be extraneous or collinear with other variables?

Tuning Difficulty Level

- Difficulty of problem can be refined using iterative testing framework
- Can refine probabilities, estimate difficulty from ROC curves, and recalculate dataset until desired difficulty reached

Reference Parameters

The Look AHEAD Research Group. Prospective association of a genetic risk score and lifestyle intervention with cardiovascular morbidity and mortality among individuals with type 2 diabetes: the Look AHEAD randomised controlled trial. *Diabetologia.* 2015;58(8):1803-1813. doi:10.1007/s00125-015-3610-z.

Next Steps

Sample Datasets and the generation script will be made open-access and open-source on GitHub. We encourage collaboration to improve this idea.

For More Information

www.ohsu.edu/bd2k

ACKNOWLEDGEMENTS:
This work is supported by NIH Grant 1R25EB020379-01

OHSU