

ECON 145

**ECON 145 – Introductory Data
Analytics in Healthcare**

**Lecture 5: Risk Factor Analysis II
(Logistic Regression)**

Lecture 4:

- Risk Factor Analysis
- Linear Regression
- Framingham Heart Study

Data Analysis (Cont.)

Class Outline

- Logistic Regression
- Case Study: Risk of Disease in Framingham Heart Study Using Logistic Regression

Multivariate Risk Factors: Heart Failure

The routinely measured risk factors used in constructing the heart failure profile include age,

electrocardiographic left ventricular hypertrophy, cardiomegaly on chest x-ray film, heart rate, systolic blood pressure, vital capacity, diabetes mellitus, evidence of myocardial infarction, and valvular disease or hypertension.

The probability of developing heart failure was determined in subjects aged 45 through 94 years who had coronary disease, hypertension, or valvular heart disease, but were free of the condition at baseline. Over the course of the 38-year study period, among those with these predisposing conditions, there were 6354 person-examinations with follow-up in men and 8913 in women.

Based on 486 heart failure cases during 38 years of follow-up, 4-year probabilities of failure were computed using the pooled logistic regression model for each sex...

Table 3. Pooled Logistic Regression Model With Coefficients and Odds Ratios*

Variables	Units	Regression Coefficient	OR (95% CI)	P
Men				
Intercept		-9.2087		
Age	10 y	0.0412	1.51 (1.31-1.74)	<.001
LVH	Yes/no	0.9026	2.47 (1.31-3.77)	<.001
Heart rate	10 bpm	0.0166	1.18 (1.08-1.29)	<.001
Systolic blood pressure	20 mm Hg	0.00804	1.17 (1.04-1.32)	.007
CHD	Yes/no	1.6079	4.99 (3.80-6.55)	<.001
Valve disease	Yes/no	0.9714	2.64 (1.89-3.69)	<.001
Diabetes	Yes/no	0.2244	1.25 (0.89-1.76)	.20
Women				
Intercept		-10.7988		
Age	10 y	0.0503	1.65 (1.42-1.93)	<.001
LVH	Yes/no	1.3402	3.82 (2.50-5.83)	<.001
Heart rate	10 bpm	0.0105	1.11 (1.01-1.23)	.03
Systolic blood pressure	20 mm Hg	0.00337	1.07 (0.96-1.20)	.24
CHD				
Valve disease				
Diabetes				
BMI				
Valve disease and diabetes				

* Excludes vital capacity odds ratio; CI, confidence interval; CHD, congenital heart disease.

ORIGINAL INVESTIGATION

Profile for Estimating Risk of Heart Failure

William B. Kannel, MD, MPH; Ralph B. D'Agostino, PhD; Halit Silberschatz, PhD; Albert J. Belanger, MS; Peter W. F. Wilson, MD; Daniel Levy, MD

Context: We devised a risk appraisal function to assess the hazard of heart failure in persons who are predisposed by coronary disease, hypertension, or valvular heart disease.

Objectives: To provide general practitioners and internists with a cost-effective method to select people at high risk who are likely to develop left ventricular systolic dysfunction. This may then allow for further evaluation and aggressive preventive measures.

Methods: The routinely measured risk factors used in constructing the heart failure profile include age, electrocardiographic left ventricular hypertrophy, cardiomegaly on chest x-ray film, heart rate, systolic blood pressure, vital capacity, diabetes mellitus, evidence of myocardial infarction, and valvular disease or hypertension. Based on 486 heart failure cases during 38 years of follow-up, 4-year probabilities of failure were computed using the pooled logistic regression model for each sex; a single point score system was employed. A multivariate profile was also produced without the vital cap-

acity or chest x-ray film because these may not be readily available in some clinical settings.

Results: Using the risk factors that make up the multivariate risk formulation—derived from ordinary office procedures—the probability of developing heart failure can be estimated and compared with the average risk for persons of the same age and sex. Using this risk profile, for example, one may then be able to identify subjects in the top quintile of multivariate risk.

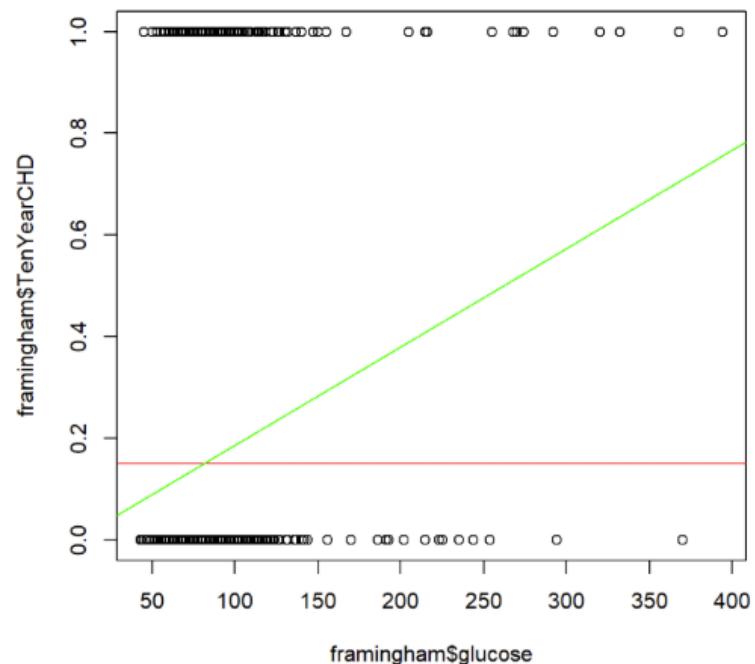
Conclusions: Using this multivariate risk formulation, it is possible to identify high-risk candidates for heart failure who are likely to have a substantial yield of positive findings when tested for heart failure evidence and symptoms of left ventricular dysfunction. The risk profile may also identify candidates who are at high risk for heart failure because of multiple, marginal risk factor abnormalities that might otherwise be overlooked.

Arch Intern Med. 1999;159:1197-1204

Linear Probability Model - FHS

$$TenYearCHD = -0.6529 + 0.0522 \text{ male} + 0.0074 \text{ age} + 0.0023 \text{ cigsPerDay} + 0.2290 \text{ prevalentStroke} + 0.0356 \text{ prevalentHyp} + 0.0003 \text{ totChol} + 0.0015 \text{ sysBP} + 0.0013 \text{ glucose}$$

- Easy to interpret
- However, the predicted values can be negative or larger than one
 - Not interpretable anymore
- All the points will lie on either 0 or 1 for the response variable (*TenYearCHD*)



Lecture 3:

- Risk Factor Analysis
- Linear Regression
- Framingham Heart Study

Analysis (Cont.)

Class Outline

- **Logistic Regression**
- Case Study: Risk of Disease in Framingham Heart Study Using Logistic Regression

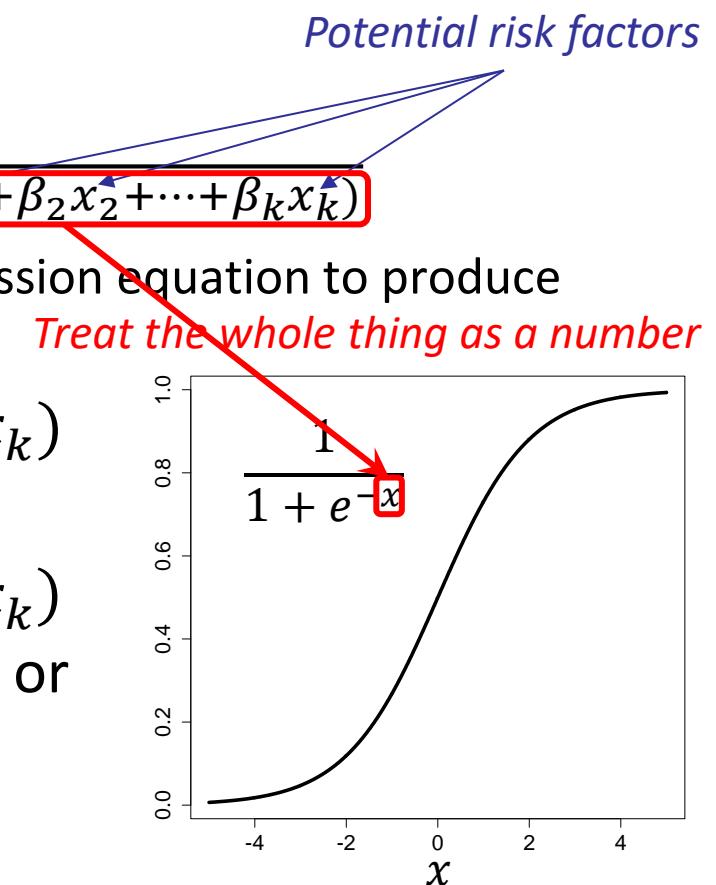
Logistic Regression

- **Logistic (response) function**

TenYearCHD

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}}$$

- Nonlinear transformation of linear regression equation to produce numbers between 0 and 1
- When $(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)$ increases, $P(y = 1)$ gets closer to 1
- When $(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)$ decreases, $P(y = 1)$ gets closer to 0, or $P(y = 0)$ gets closer to 1



Properties of the Logistic Function

- **Logistic (response) function**

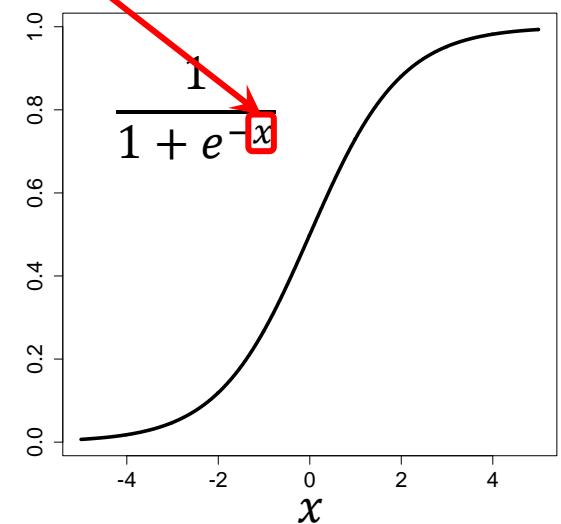
TenYearCHD

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}}$$

- Always between 0 and 1
 - Can be interpreted as probabilities
- Increasing in the input x
 - Allow the coefficients to be easily interpreted
 - Positive β_i
 - An increase in x_i increases $P(y = 1)$
 - Negative β_i
 - An increase in x_i decreases $P(y = 1)$

Potential risk factors

*Treat the whole thing
as a number*



More Interpretation from Logistic Regression

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}}$$

1

- **Odds** of developing CHD in ten years

$$\text{Odds} = \frac{P(y = 1)}{P(y = 0)} = e^{(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)} = e^{\beta_0} e^{\beta_1 x_1} \dots e^{\beta_k x_k}$$

where e is the base of the natural logarithm, and $\beta_0, \beta_1, \dots, \beta_k$ are the coefficients of the logistic regression model.

$$\ln(\text{Odds}) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

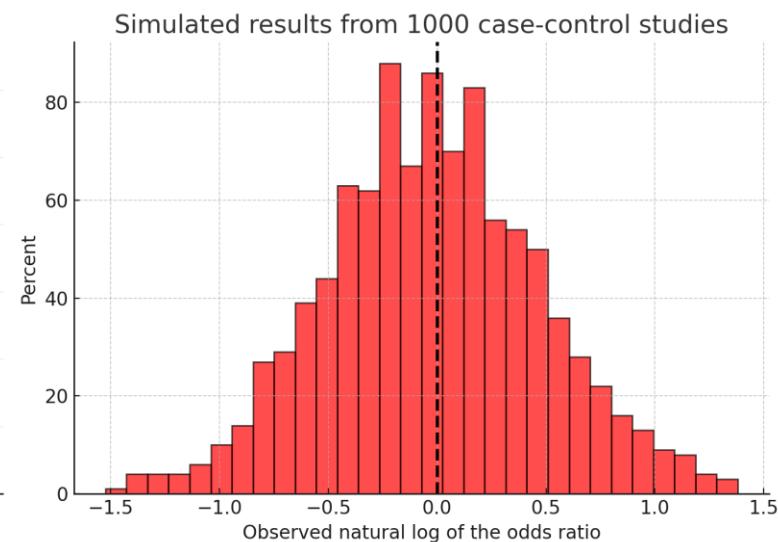
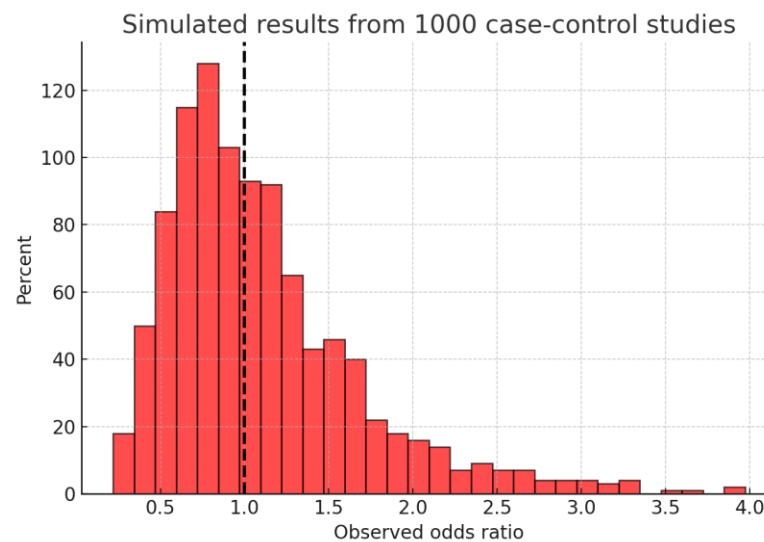
Also known as the **logit** function

Derivation steps:

$$\begin{aligned} P(y = 1) &= \frac{1}{1 + e^{-x}} \\ P(y = 0) &= 1 - \left(\frac{1}{1 + e^{-x}} \right) = \frac{e^{-x}}{1 + e^{-x}} \\ \frac{P(y = 1)}{P(y = 0)} &= \frac{1}{1 + e^{-x}} * \frac{1 + e^{-x}}{e^{-x}} = \frac{1}{e^{-x}} = e^x \end{aligned}$$

Summary of Outcomes Reported

Description	Measures	Min	Max
Probability of Event/Outcome	$P(Y = 1)$	0	1
Odds	$\frac{P(Y=1)}{P(Y=0)}$	0	∞
Log Odds or Logit	$\ln \left(\frac{P(Y=1)}{P(Y=0)} \right)$	$-\infty$	∞

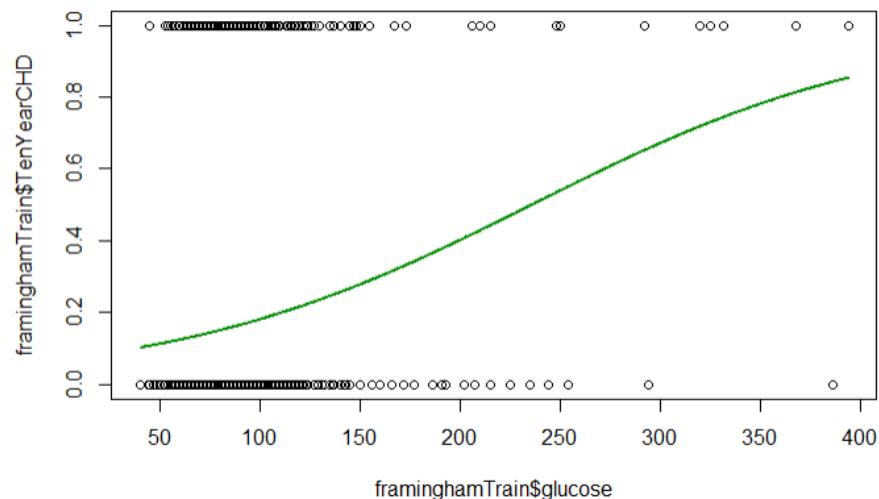
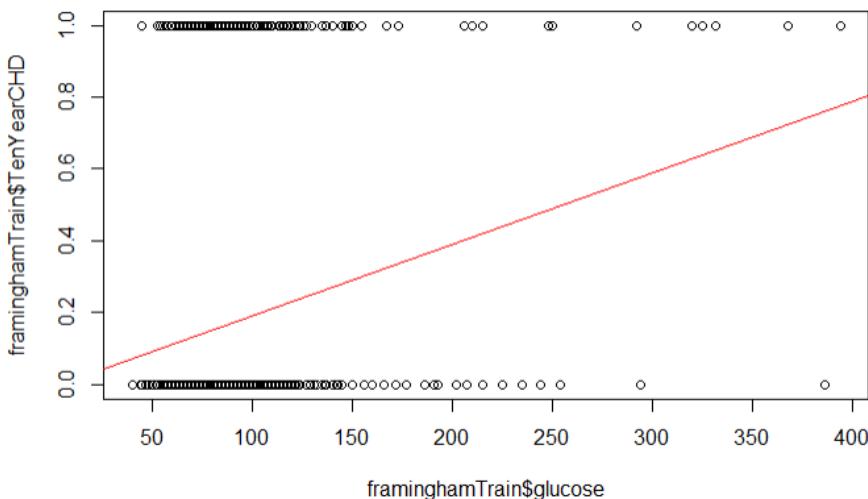


Linear vs. Logistic Regression

TenYearCHD

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}}$$

Potential risk factors



- Note: logistic function is NOT linear in β 's

Training Logistic Regression Models

- **Logistic (response) function**

TenYearCHD

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}}$$

Potential risk factors

$$\text{Odds} = \frac{P(y = 1)}{P(y = 0)} \rightarrow \ln(\text{Odds}) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

Effect sizes ` β 's` for Odds are easier to interpret.

These are represented as: $e^{\beta_0}, e^{\beta_1}, \dots, e^{\beta_k}$

Interpreting R Output

p-value for each independent variable

```
Call:  
lm(formula = TenYearCHD ~ ., data = framinghamTrain)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.71221	-0.18536	-0.10551	-0.01096	1.07341

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.5585219	0.0938940	-5.948	3.08e-09 ***
male1	0.0519583	0.0150741	3.447	0.000576 ***
age	0.0070251	0.0009309	7.546	6.20e-14 ***
education2	-0.0139725	0.0166636	-0.839	0.401827
education3	-0.0254858	0.0200981	-1.268	0.204889
education4	-0.0133774	0.0231004	-0.579	0.562574
currentSmoker1	0.0151090	0.0218574	0.691	0.489470
cigsPerDay	0.0023370	0.0009415	2.482	0.013117 *
BPMeds1	0.0701339	0.0410007	1.711	0.087286 .
prevailingStroke1	0.0484975	0.1091493	0.444	0.656847
prevailingHyp1	0.0190762	0.0208096	0.917	0.359385
diabetes1	0.0350311	0.0526092	0.666	0.505552
totChol	0.0001340	0.0001613	0.831	0.406264
sysBP	0.0028345	0.0005972	4.746	2.19e-06 ***
diaBP	-0.0019303	0.0009734	-1.983	0.047475 *
BMI	0.0002581	0.0018647	0.138	0.889910
heartRate	-0.0004165	0.0005931	-0.702	0.482646
glucose	0.0010791	0.0003444	3.133	0.001748 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
 Residual standard error: 0.3416 on 2543 degrees of freedom
 Multiple R-squared: 0.1027 Adjusted R-squared: 0.09669
 F-statistic: 17.12 on 17 DF, 2543 on 2543 DF, p-value: 5.22e-10

```
call:  
glm(formula = TenYearCHD ~ ., family = "binomial", data = framinghamTrain)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-8.205609	0.848219	-9.674	< 2e-16 ***
male1	0.497280	0.131930	3.769	0.000164 ***
age	0.063948	0.008122	7.874	3.45e-15 ***
education2	-0.112110	0.145016	-0.773	0.439474
education3	-0.232291	0.183097	-1.269	0.204556
education4	-0.111357	0.202087	-0.551	0.581609
currentSmoker1	0.160592	0.187927	0.855	0.392803
cigsPerDay	0.018589	0.007457	2.493	0.012671 *
BPMeds1	0.307977	0.275670	1.117	0.263913
prevailingStroke1	0.265385	0.726877	0.365	0.715034
prevailingHyp1	0.159858	0.166183	0.962	0.336079
diabetes1	0.100131	0.368270	0.272	0.785703
totChol	0.002240	0.001327	1.689	0.091303 .
sysBP	0.017890	0.004585	3.902	9.54e-05 ***
diaBP	-0.009700	0.007648	-1.268	0.204664
BMI	0.009286	0.015385	0.604	0.546138
heartRate	-0.003539	0.005054	-0.700	0.483783
glucose	0.006358	0.002506	2.537	0.011179 *

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2185.3 on 2560 degrees of freedom
 Residual deviance: 1926.9 on 2543 degrees of freedom

Interpretations are different!

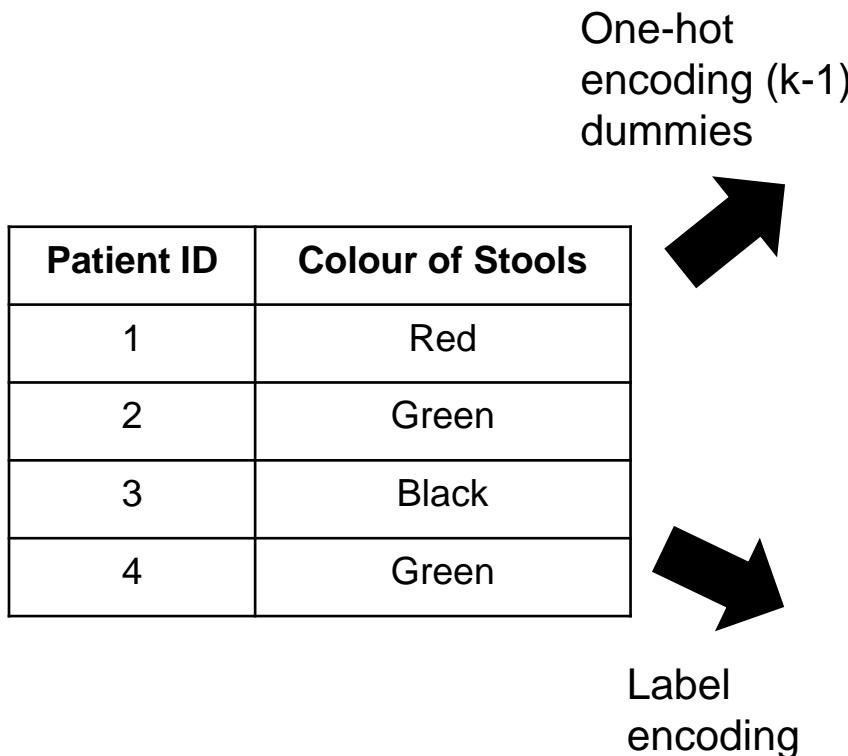
Handling categorical variables

Since regression can only be performed on numerical variables, we need to encode categorical variables to numerical. Types of encoding:

- **Label encoding** – Use unique numerical labels for each category
- **Boolean encoding** - Binary or Two-value variable can be coded as Boolean variables. It is coded as 1 for all observations in that category and 0 for all observations not in that category. E.g., Gender
- **One-hot encoding with k dummy variables**
 - Create additional k new binary variables to represent each possible value. E.g., four races in Singapore needs 4 additional variables to represent.
- **One-hot encoding with k-1 dummy variables encoding** (a.k.a., remove first dummy “drop-first”)
 - Use $(k-1)$ variables dummies where $k = \text{number of categories in the variable}$. E.g., Four races in Singapore needs 3 dummies to represent
 - **To avoid multicollinearity!**

As in many aspects of Analytics, there is no single answer to this problem

Handling categorical variables



Rescaling Data

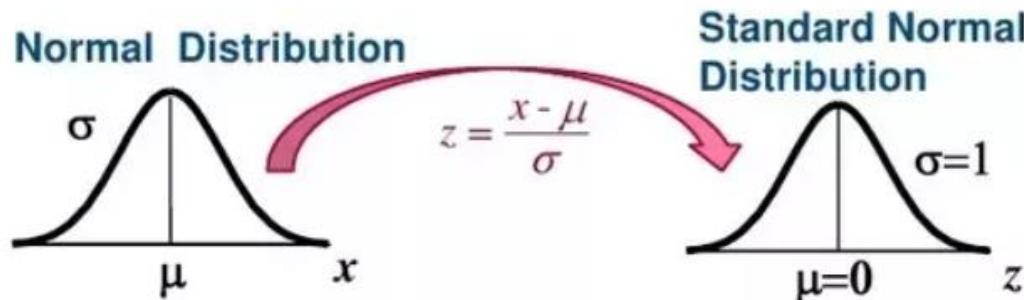
Depending on the nature of the data, **rescaling of data** might be required.

- **Normalization**

- E.g., rescaling by the range of the vector to make all elements lie within a standard range (specified minimum and maximum value)

- **Standardization**

- Most commonly used to transform the variable into a standard normal random variable with mean 0 and standard deviation 1



Rescaling Data

- **Normalization** (Dividing by a norm of the vector)

$$X_{\text{transformed}} = (X - \text{Minimum Value}) / \text{Range of Values} \sim [0, 1]$$

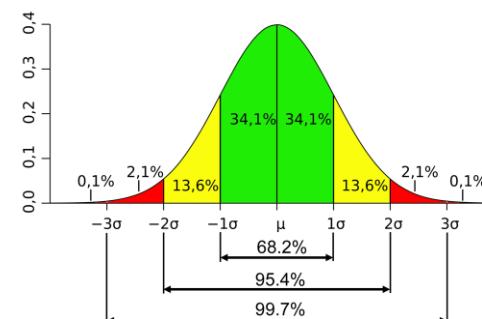
- **Standardization** (or z-score normalization)

$$X_{\text{transformed}} = (X - \text{Mean}) / \text{Standard Deviation}$$

Suppose that the mean and the standard deviation of the values for the hospital bills are \$54,000 and \$16,000 respectively. What does \$73,600 map to?

$$\begin{aligned} v' &= \frac{v - \bar{A}}{\sigma_A} \\ v' &= \frac{73600 - 54000}{16000} \\ v' &= 1.225 \end{aligned}$$

(1.225σ above the mean)



Rescaling Data

When to use? (some guidelines, but use with caution)

Normalization:

- Distribution of the data is unknown or when the distribution is not Gaussian
- Data has varying scales and the algorithm does not make assumptions about the distribution of data (e.g., k-means, neural networks)
- May result in bias if outliers are considered in the scaling (consider other scaling algorithms, e.g., Robust Scaler)

Standardization:

- Distribution of the data appears to be Gaussian for the technique to be more effective (e.g., linear regression, linear discriminant analysis, logistic regression)

Other transformations: Log, Box-Cox, etc. There are other more advanced techniques (e.g., Generalized Estimating Equations)

Note: Transformation of response variable can sometimes help to make the residuals normal (assumption of linear regression models)

Simpson's Paradox

- UC Berkeley almost got sued for sex discrimination in 1973

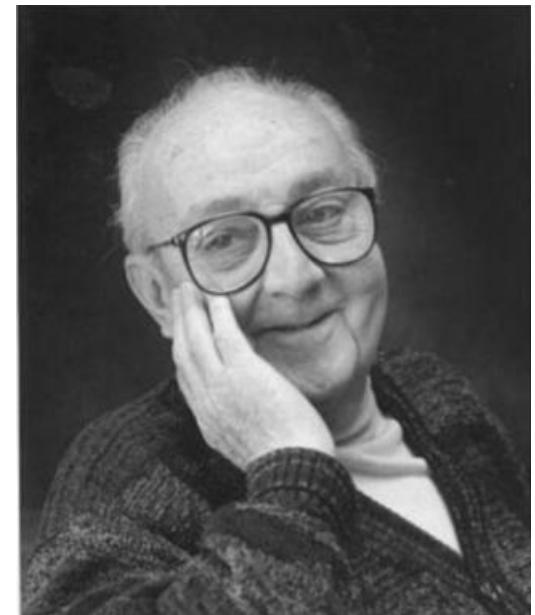
	Applicants	Admitted
Men	8442	44%
Women	4321	35%

- The data from the six largest departments

Department	Men Applicants	Women Applicants	Men Admitted	Women Admitted
A	825	108	62%	82%
B	560	25	63%	68%
C	325	593	37%	34%
D	417	375	33%	35%
E	191	393	28%	24%
F	373	341	6%	7%

Model

“The most that can be expected from any model is that it can supply a **useful** approximation to reality: **All models are wrong; some models are useful.**”



George E. P. Box (1919-2013)
One of the great statistical minds of the 20th century

Break – Next Lecture

Lecture 3:

- Risk Factor Analysis
- Linear Regression
- Framingham Heart Study

Analysis (Cont.)

Class Outline

- Lecture 4 Recap
- Logistic Regression
- **Case Study: Risk of Disease in Framingham Heart Study Using Logistic Regression**

Recall: Coronary Heart Disease (CHD) and Its Risk Factors

- Dependent variable:
Developing CHD in ten years
- Demographic risk factors
 - *male*: sex of patient (**recoded**)
 - *age*: age in years at first examination
 - *education*: Some high school (1), high school/GED (2), some college/vocational school (3), college (4) (**recoded**)

Table 3. Diagnoses among Men Who Were 50 to 59 Years Old at Base Line Who Had Incident Cases of Cardiovascular Disease.*

DIAGNOSIS	COHORT		
	1950	1960	1970
events/1000 (percent of all events)			
Coronary heart disease	124 (65)	125 (71)	103 (67)
Myocardial infarction	54 (28)	63 (36)	43 (28)
Angina pectoris	60 (32)	48 (27)	44 (29)
Sudden death	8 (4)	6 (3)	10 (6)
Other	2 (1)	8 (5)	6 (4)
Cerebrovascular disease	23 (12)	20 (11)	18 (12)
Atherothrombotic brain infarction	15 (8)	7 (4)	10 (6)
Transient ischemic attacks	2 (1)	7 (4)	6 (4)
Other	6 (3)	6 (3)	2 (1)
Other cardiovascular disease	43 (23)	30 (17)	33 (21)†
Intermittent claudication	23 (12)	15 (9)	27 (18)‡
Congestive heart failure	16 (8)	11 (6)	6 (4)
Other	4 (2)	4 (2)	0 (0)
Total events	190	175	154

*The data are from the Framingham Heart Study, 1950 through 1979. Because of rounding, percentages do not always total 100.

†P = 0.77 for the comparison of the three categories across cohorts, by Mantel-Haenszel test.

‡P = 0.78 for the comparison of the diagnoses across cohorts, by Mantel-Haenszel test.

Recall: Coronary Heart Disease (CHD) and Its Risk Factors

- Behavioral risk factors (Smoking behavior)
 - *currentSmoker* (**recoded**), *cigsPerDay*
- Medical history risk factors
 - *BPmeds*: On blood pressure medication at time of first examination(**recoded**)
 - *prevStroke*: Previously had a stroke(**recoded**)
 - *prevHyp*: Currently hypertensive(**recoded**)
 - *diabetes*: Currently has diabetes(**recoded**)

Recall: Coronary Heart Disease (CHD) and Its Risk Factors

- Risk factors from first examination
 - *totChol*: Total cholesterol (mg/dL)
 - *sysBP*: Systolic blood pressure
 - *diaBP*: Diastolic blood pressure
 - *BMI*: Body Mass Index, weight (kg)/height (m)²
 - *heartRate*: Heart rate (beats/minute)
 - *glucose*: Blood glucose level (mg/dL)

Let's get our hands dirty!

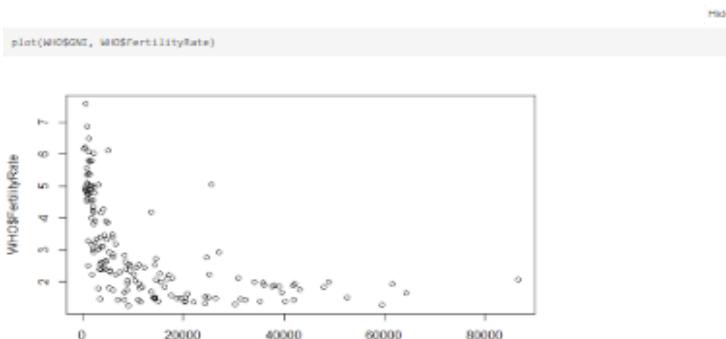
Data Analysis

Let's do some basic data analysis using our WHO data.

```
WHO$Under15
[1] 47.42 21.33 27.42 15.28 47.58 25.96 24.42 28.34 18.95 14.51 22.25 21.62 28.16 30.57 18.99 15.18 16.88 34.4
8 42.95 28.53
[21] 35.23 16.35 33.75 24.56 25.75 13.53 45.66 44.20 31.23 43.08 16.37 38.17 48.87 48.52 21.38 17.95 28.03 42.1
7 42.37 38.61
[41] 23.94 41.48 14.98 16.58 17.16 14.56 21.98 45.11 17.66 33.72 25.96 38.53 38.29 31.25 38.62 38.95 43.18 15.6
9 43.29 28.88
[61] 16.42 18.26 38.49 45.98 17.62 13.17 38.59 14.68 26.96 48.88 42.46 41.55 36.77 35.35 35.72 14.62 28.71 29.4
3 29.27 23.68
[81] 48.51 21.54 27.53 14.04 27.78 13.12 34.13 25.46 42.37 38.18 24.98 38.21 35.61 14.57 21.64 36.75 43.06 29.4
5 15.13 17.46
[101] 42.72 45.44 26.65 29.03 47.14 14.98 38.18 48.22 28.17 29.82 35.81 18.26 27.85 19.81 27.85 45.38 25.28 36.5
9 38.18 35.58
[121] 17.21 20.26 33.37 49.99 44.23 38.61 18.64 24.19 34.31 38.18 28.65 38.37 32.78 29.18 34.53 14.91 14.92 13.2
8 15.25 16.52
[141] 15.85 15.45 43.56 25.96 24.31 25.78 37.88 14.04 41.68 29.69 43.54 16.45 21.95 41.74 16.48 15.08 14.16 40.3
7 47.35 29.53
[161] 42.28 15.28 25.15 41.48 27.83 38.85 16.71 14.79 35.35 35.75 18.47 16.89 46.33 41.89 37.33 28.73 23.22 26.8
0 28.05 38.61
[181] 48.54 14.18 14.41 17.54 44.85 19.63 22.05 28.98 37.37 28.84 22.87 48.72 46.73 40.24
```

```
WHO$Country[which.min(WHO$Under15)]
[1] Japan
194 Levels: Afghanistan Albania Algeria Andorra Angola Antigua and Barbuda Argentina Armenia Australia Austria
... Zimbabwe
```

Let's create some plots for exploratory data analysis (EDA). First, let's create a basic scatterplot of GNI versus FertilityRate.



A Logistic Regression Model

- Used data for 2561 individuals to build the model (70% of the data)

$$P(\text{TenYearCHD} = 1) = \frac{1}{1+e^{(-8.2056 + 0.4973 \cdot \text{male} + \dots + 0.0064 \cdot \text{glucose})}}$$

$$\ln(\text{Odds}) = -8.2056 + 0.4973 \cdot \text{male} + \dots + 0.0064 \cdot \text{glucose}$$

Logistic Regression Output

- Effect Estimation

```

Call:
glm(formula = TenYearCHD ~ ., family = "binomial", data = framinghamTrain)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.205609  0.848219 -9.674 < 2e-16 ***
male1        0.497280  0.131930  3.769 0.000164 ***
age          0.063948  0.008122  7.874 3.45e-15 ***
education2   -0.112110  0.145016 -0.773 0.439474
education3   -0.232291  0.183097 -1.269 0.204556
education4   -0.111357  0.202087 -0.551 0.581609
currentSmoker1 0.160592  0.187927  0.855 0.392803
cigsPerDay    0.018589  0.007457  2.493 0.012671 *
BPMed1        0.307977  0.275670  1.117 0.263913
prevalentStroke1 0.265385  0.726877  0.365 0.715034
prevalentHyp1   0.159858  0.166183  0.962 0.336079
diabetes1      0.100131  0.368270  0.272 0.785703
totChol        0.002240  0.001327  1.689 0.091303 .
sysBP          0.017890  0.004585  3.902 9.54e-05 ***
diaBP          -0.009700  0.007648 -1.268 0.204664
BMI            0.009286  0.015385  0.604 0.546138
heartRate      -0.003539  0.005054 -0.700 0.483783
glucose         0.006358  0.002506  2.537 0.011179 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

```

```

Null deviance: 2185.3 on 2560 degrees of freedom
Residual deviance: 1926.9 on 2543 degrees of freedom
AIC: 1962.9

```

Prediction From Logistic Regression

- The outcome of a logistic regression model is a probability
 - Let's consider Person 6 to Person 10 in the training set from now

	Person 6	Person 7	Person 8	Person 9	Person 10
Prediction	0.123	0.188	0.064	0.236	0.084

- Often, we want to make a class prediction to compare with the actual outcome
 - Will this person develop CHD in ten years?
- We can do this using a threshold value (cut-off), denoted as t
 - If $P(TenYearCHD = 1) \geq t$, predict CHD
 - If $P(TenYearCHD = 1) < t$, predict no CHD
- What value should we pick for t ?

A Note on Threshold Selection for the Logistic Regression Model

- For logistic regression:

$$P(TenYearCHD = 1) = \frac{1}{1 + e^{(-8.2056 + 0.4973 \text{ male} + \dots + 0.0064 \text{ glucose})}}$$

- Applying threshold t (mathematical derivation is not tested)

$$\begin{aligned} P(TenYearCHD = 1) &\geq t \\ \Leftrightarrow \frac{1}{1 + e^{(-8.2056 + 0.4973 \text{ male} + \dots + 0.0064 \text{ glucose})}} &\geq t \\ \Leftrightarrow -8.2056 + 0.4973 \text{ male} + \dots + 0.0064 \text{ glucose} &\geq -\ln \frac{1-t}{t} \end{aligned}$$

- Selecting a threshold for the logistic regression model is equivalent to **choosing a threshold on its logit function**
 - Logit is linear in the independent variables

Threshold Value

- If $P(TenYearCHD = 1) \geq t$, predict CHD ($y = 1$)
- If $P(TenYearCHD = 1) < t$, predict no CHD ($y = 0$)

	Person 6	Person 7	Person 8	Person 9	Person 10	
Prediction	0.123	0.188	0.064	0.236	0.084	
Predicted Outcome	$t = 0.05$	1	1	1	1	1
	$t = 0.15$	0	1	0	1	0
	$t = 0.25$	0	0	0	0	0

- If the threshold t is small, more individuals are classified as 1 (CHD)
- If the threshold t is large, more individuals are classified as 0 (no CHD)

Selecting a Threshold Value

- Let's compare the predictions to the actual outcomes

	Person 6	Person 7	Person 8	Person 9	Person 10
Prediction	0.123	0.188	0.064	0.236	0.084
Predicted Outcome	$t = 0.05$	1	1	1	1
	$t = 0.15$	0	1	0	1
	$t = 0.25$	0	0	0	0
Actual TenYearCHD	0	1	0	0	0

Selecting a Threshold Value

- Let's compare the predictions to the actual outcomes

	Person 6	Person 7	Person 8	Person 9	Person 10
Prediction	0.123	0.188	0.064	0.236	0.084
Predicted Outcome	$t = 0.05$	1	1	1	1
	$t = 0.15$	0	1	0	1
	$t = 0.25$	0	0	0	0
Actual TenYearCHD	0	1	0	0	0

- Confusion matrix (classification matrix)

	Predicted No CHD (0)	Predicted CHD (1)
Actual No CHD (0)	True Negatives (TN) ✓	False Positives (FP) ✗
Actual CHD (1)	False Negatives (FN) ✗	True Positives (TP) ✓

Selecting a Threshold Value

- A different threshold value changes the types of errors

		Person 6	Person 7	Person 8	Person 9	Person 10
Prediction		0.123	0.188	0.064	0.236	0.084
Predicted Outcome	$t = 0.05$	1	1	1	1	1
	$t = 0.15$	0	1	0	1	0
	$t = 0.25$	0	0	0	0	0
Actual TenYearCHD		0	1	0	0	0

$t = 0.05$	Predicted = 0	Predicted = 1	$t = 0.15$	Predicted = 0	Predicted = 1	$t = 0.25$	Predicted = 0	Predicted = 1
Actual = 0	0	4	Actual = 0	3	1	Actual = 0	4	0
Actual = 1	0	1	Actual = 1	0	1	Actual = 1	1	0

Selecting a Threshold Value

- A different threshold value changes the types of errors
 - If t is small, more **non-CHD** cases are classified as CHD
 - More **false positives**
 - If t is large, more **CHD** cases are classified as no-CHD
 - More **false negatives**

$t = 0.05$	Predicted = 0	Predicted = 1	$t = 0.15$	Predicted = 0	Predicted = 1	$t = 0.25$	Predicted = 0	Predicted = 1
Actual = 0	0	4	Actual = 0	3	1	Actual = 0	4	0
Actual = 1	0	1	Actual = 1	0	1	Actual = 1	1	0

Selecting a Threshold Value

- One way to select a threshold is based on accuracy, which measures the overall correct prediction rate
- Accuracy** = $(TN + TP) / (\# \text{ Observations})$
- Choose a threshold that gives high accuracy
 - There are **implicit assumptions when you use accuracy**
 - The **threshold that gives the highest accuracy may result in a useless classification model**



$t = 0.05$	Predicted = 0	Predicted = 1	$t = 0.15$	Predicted = 0	Predicted = 1	$t = 0.25$	Predicted = 0	Predicted = 1	
Actual = 0	0	4	Actual = 0	3	1	Actual = 0	4	0	
Actual = 1	0	1	Actual = 1	0	1	Actual = 1	1	0	
0.2		0.8		0.8		0.8		0.8	

A “Baseline Model”

- When we build classification models, we want to compare our model to a simple baseline model
 - Remember that R^2 does this in linear regression
- A standard **baseline** model is to **predict the most common outcome**
- In this case, 557 participants actually developed CHD in ten years, and 3101 did not
- The baseline model would **predict no-CHD for everyone**, and get an accuracy of $3101/3658 = 84.77\%$
 - The performance is quite good
 - Is the baseline model good?

Issues with Accuracy

- If we want to predict some rare events, e.g., “a man in the airport carrying a bomb”
 - Assume the proportion is 0.0001% (1 out of a million)
- The baseline model will **predict everyone to be innocent** and has an accuracy of 99.9999%
 - Equivalent to set the threshold = 1 in any logistic regression model
 - Many sophisticated models are very likely to perform much worse than the baseline model if a threshold < 1 is used
 - Is the baseline model the best? Is it useful?
- How to set the threshold to trigger an alarm for security check?

When to Use Accuracy

- When accuracy is a good measure to use in threshold selection?
 - Cost of a false positive and cost of a false negative are similar
- $\text{Accuracy} = (\text{TN} + \text{TP}) / (\# \text{ Observations})$
 $= 1 - (\text{FP} + \text{FN}) / (\# \text{ observations})$
 - Implicitly assuming the same weights (costs) on false positive and false negative
- If the costs are different, we need different metrics

Break

Sensitivity and Specificity

	Predicted No CHD (0)	Predicted CHD (1)
Actual No CHD (0)	True Negatives (TN) ✓	False Positives (FP) ✗
Actual CHD (1)	False Negatives (FN) ✗	True Positives (TP) ✓

- **Sensitivity** = $TP/(TP+FN)$
 - Also known as **True Positive Rate** = $1 - \text{False Negative Rate}$
 - Probability that an actual CHD case is predicted to be a CHD case
- **Specificity** = $TN/(TN + FP)$
 - Also known as $1 - \text{False Positive Rate} = \text{True Negative Rate}$
 - Probability that an actual no-CHD case is predicted to be a no-CHD case

Selecting a Threshold Value

Threshold = 0.05

Specificity = $0/(0 + 4) = 0$

Sensitivity = $1/(0 + 1) = 100\%$

	Predicted = 0	Predicted = 1
Actual = 0	0	4
Actual = 1	0	1

Threshold = 0.15

Specificity = $3/(3 + 1) = 75\%$

Sensitivity = $1/(1 + 0) = 100\%$

	Predicted = 0	Predicted = 1
Actual = 0	3	1
Actual = 1	0	1

Threshold = 0.25

Specificity = $4/(4 + 0) = 100\%$

Sensitivity = $0/(1 + 0) = 0$

	Predicted = 0	Predicted = 1
Actual = 0	4	0
Actual = 1	1	0

Threshold Selection and Model Selection

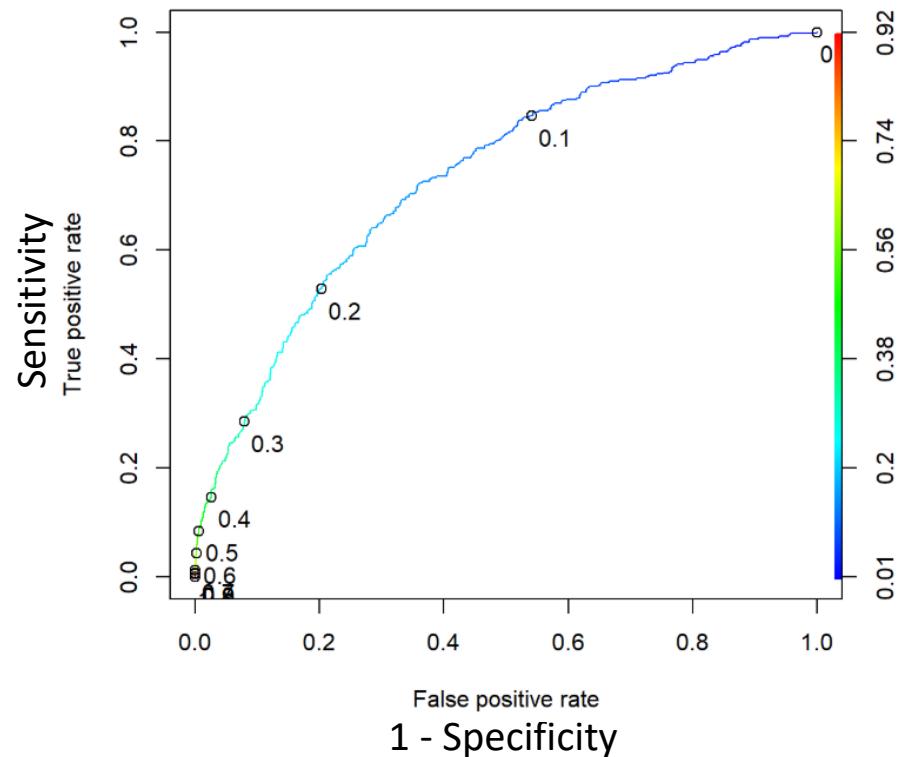
- Choose best threshold for best trade-off between
 - Cost of failing to detect positives
 - Cost of raising false alarms
 - Usually requiring additional information or data
- *Before selecting a threshold, we need to select a model first (variable selection)*

Threshold Selection and Model Selection

- Choose best threshold for best trade-off between
 - Cost of failing to detect positives
 - Cost of raising false alarms
 - Usually requiring additional information or data
- *Before selecting a threshold, we need to select a model first (variable selection)*
- How can we **evaluate the quality of a classification model** like logistic regression?
 - Use *area under the receiver operating characteristic curve*

Model Selection: Receiver Operating Characteristic (ROC) Curve

- True positive rate (sensitivity) on y-axis
 - Proportion of CHD cases caught
- False positive rate (1-specificity) on x-axis
 - Proportion of no-CHD cases labeled as CHD
- Connects all the thresholds



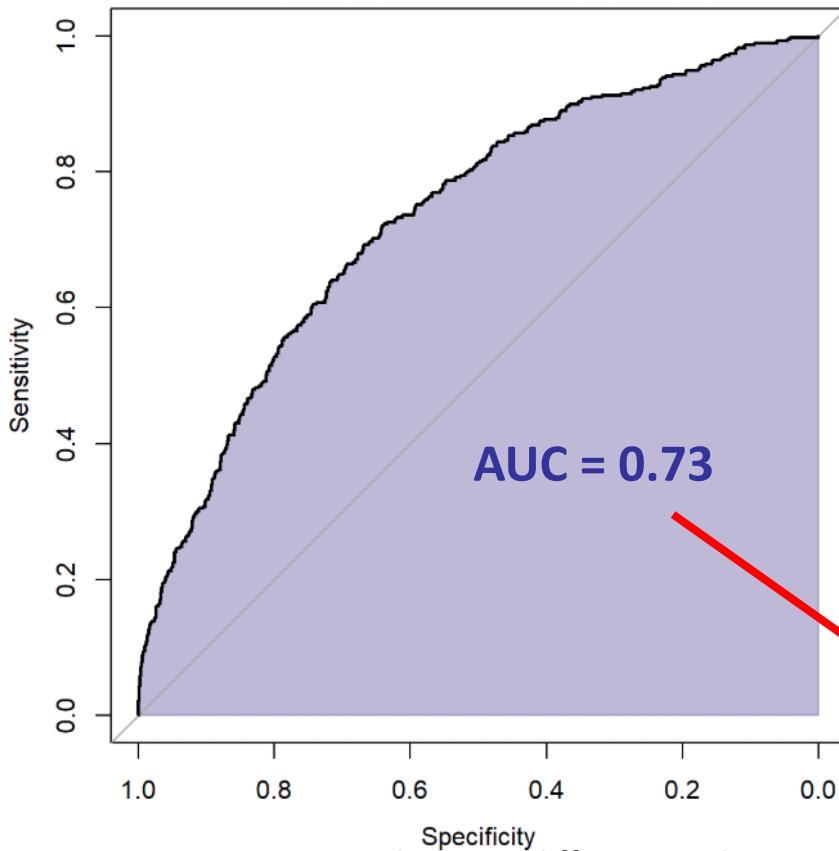
Evaluating the Model

- We can use the area under the ROC Curve (AUROC or simply **AUC**) to evaluate the quality of a classification model like logistic regression
 - Quantify the **discriminative power** of a classification model
 - Similar to R^2 for linear regression in many aspects
- Another advantage
 - **Not** for a specific threshold
- We will use AUC again in classification trees
 - Applicable to **all binary classification models**

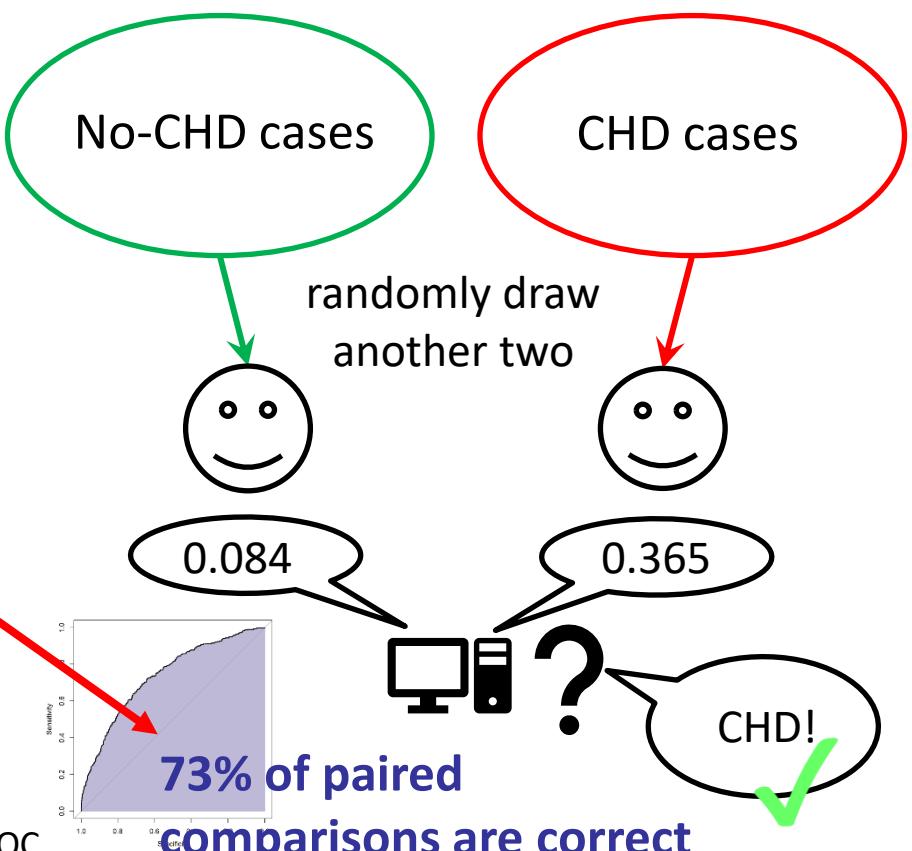
Area Under the ROC Curve (AUC)

- Just take the area under the ROC curve
- Discriminative power
 - Given a random CHD case and a random no-CHD case, proportion of the time the model predicts a higher probability of developing CHD for the CHD case
 - AUC is the Proportion of time the model will assign a higher probability of CHD for actual CHD case (regardless of the threshold) – ability to rank CHD vs no-CHD for actual CHD cases

Area Under the ROC Curve (AUC)

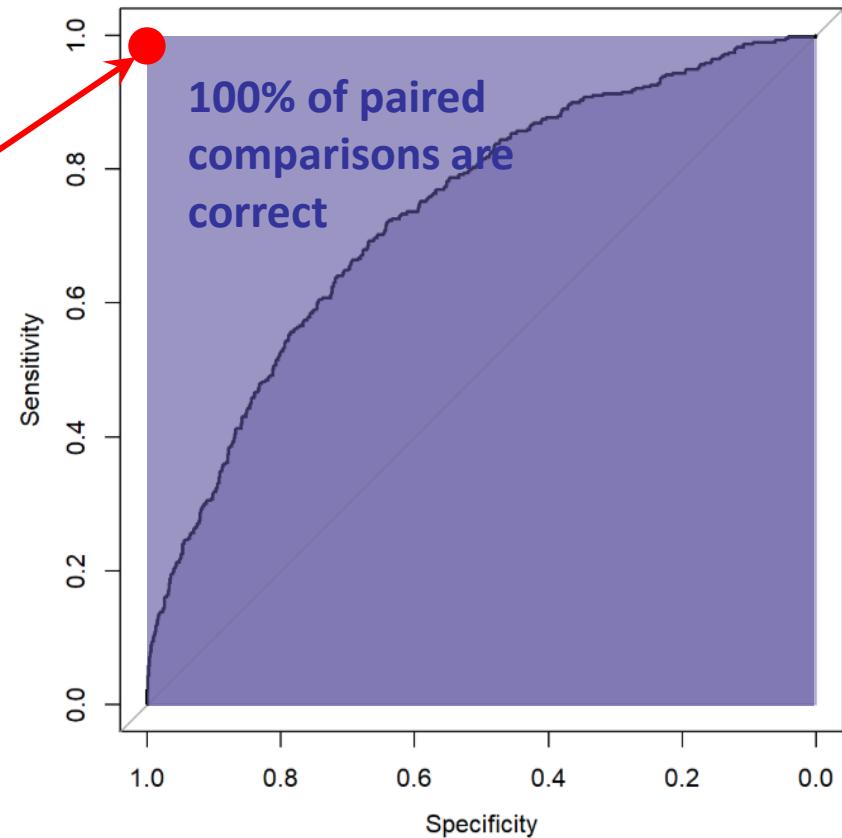


Note: Figure generated using a different package, pROC



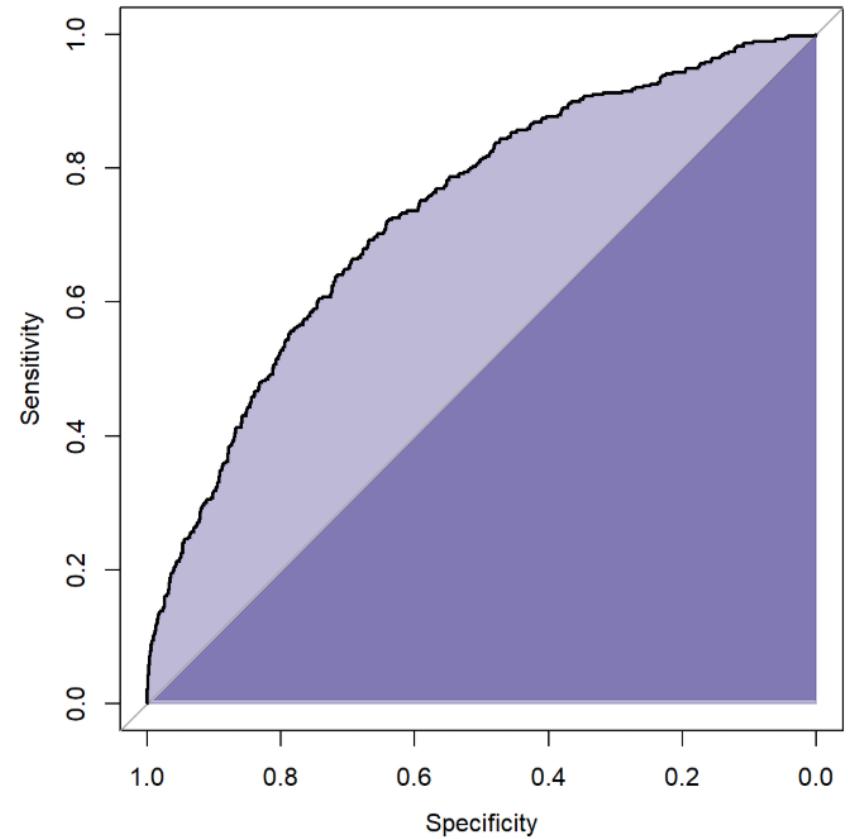
Area Under the ROC Curve (AUC)

- What is a good AUC?
 - Maximum of 1
 - Perfect discrimination
 - There exists a threshold with 100% accuracy



Area Under the ROC Curve (AUC)

- What is a good AUC?
 - Maximum of 1
 - Perfect discrimination
 - There exists a threshold with 100% accuracy
 - Minimum of 0.5
 - Just guessing
 - Lower than 0.5?



Making Predictions

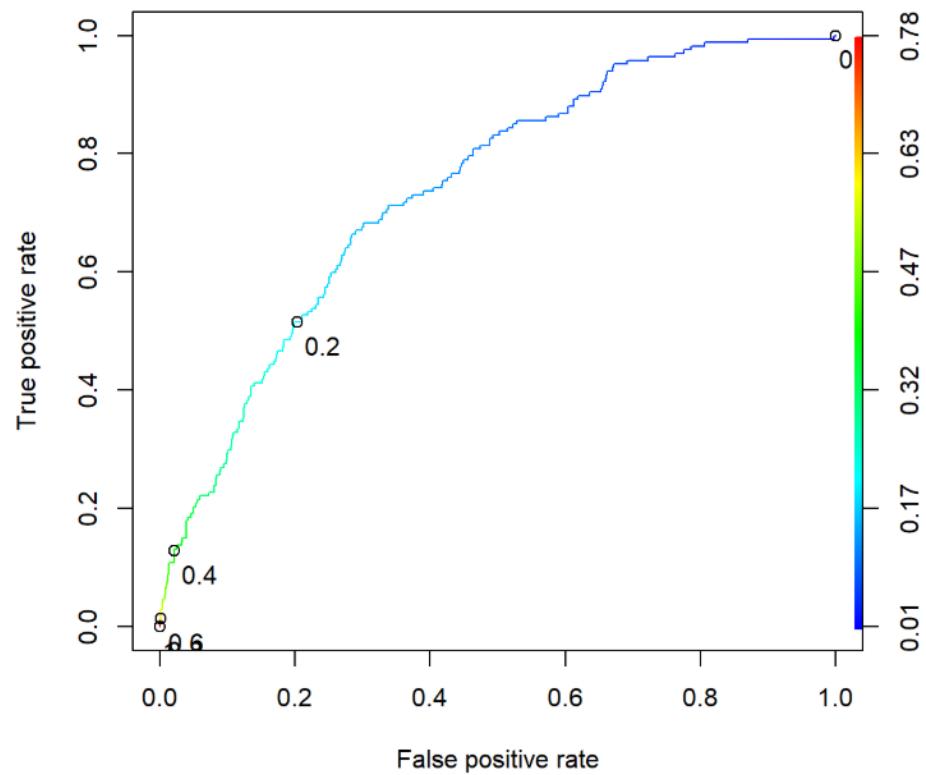
- Just like in linear regression, we want to make predictions on a test set to compute out-of-sample metrics
 - We have 1097 individuals in our test set
- If we use a threshold value of 0.3, we get the following confusion matrix

	Predicted No CHD (0)	Predicted CHD (1)
Actual No CHD (0)	853	77
Actual CHD (1)	126	41

- Out-of-sample accuracy of $894/1097 = 81.49\%$

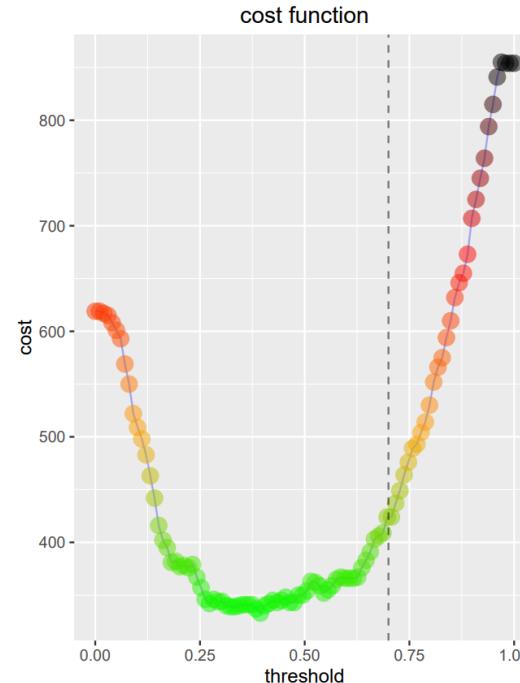
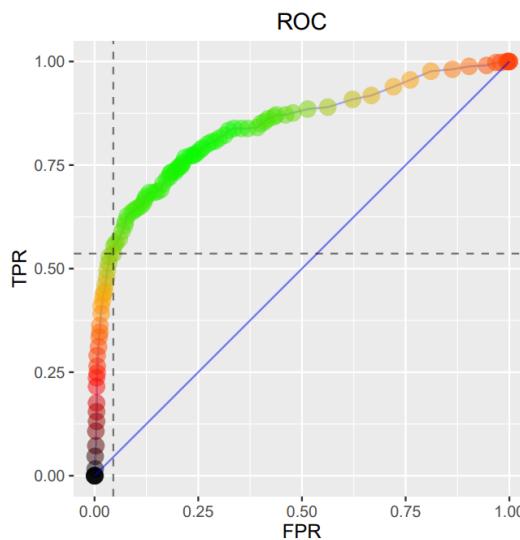
Making Predictions

- Out-of-sample ROC using Test Data
- Out-of-sample AUC = 0.74



ROC and Cost Curve

- AUROC offers a global measure across all different thresholds
- Optimal threshold requires the definition of “Cost” of False Positive vs “Cost” of False Negative
- Context of use is important



Other Quality Measures

- Positive Predictive Value, PPV (Precision):
 - Probability that subjects with a positive screening test truly have the disease
- Negative Predictive Value, NPV:
 - Probability that subjects with a negative screening test truly do not have the disease.
- Both are threshold dependent

Other Quality Measures

		NPV	PPV	Also known as Precision
		Predicted No CHD (0)	Predicted CHD (1)	
Actual No CHD (0)	853	77	Specificity	
	126	41		

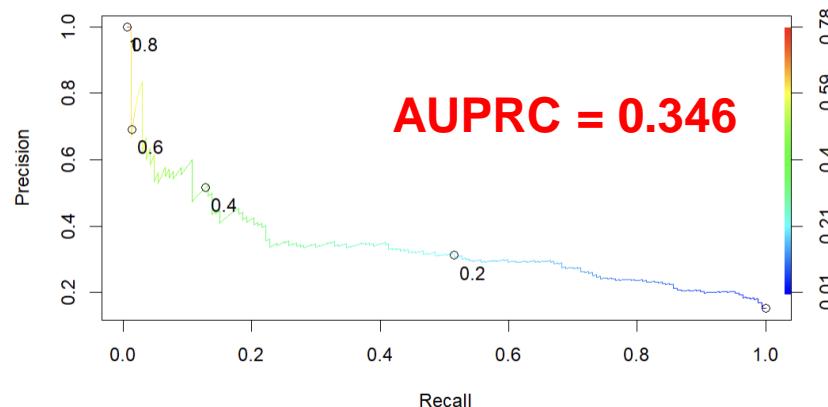
		NPV	PPV	Sensitivity
		Predicted No CHD (0)	Predicted CHD (1)	
Actual No CHD (0)	853	77	Specificity	
	126	41		

Also known as Recall

- Consider the threshold value of 0.3:
 - $PPV = TP / (TP + FP) = 41 / (41 + 77)$
 - $NPV = TN / (TN + FN) = 853 / (853 + 126)$

Area Under the Precision Recall Curve (AUPRC)

- Threshold independent curve similar to AUC
- Precision-Recall Curve plots precision (PPV) against Recall (Sensitivity) for every possible cut-off.
- AUPRC is useful for imbalanced datasets
- AUPRC gives a more accurate picture of performance when positive cases are far outnumbered by negative cases



Let's get our hands dirty!

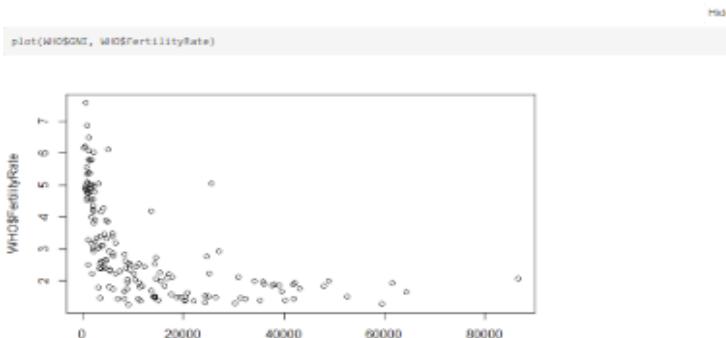
Data Analysis

Let's do some basic data analysis using our WHO data.

```
WHO$Under15
[1] 47.42 21.33 27.42 15.28 47.58 25.96 24.42 28.34 18.95 14.51 22.25 21.62 28.16 30.57 18.99 15.18 16.88 34.4
8 42.95 28.53
[21] 35.23 16.35 33.75 24.56 25.75 13.53 45.66 44.20 31.23 43.08 16.37 38.17 48.87 48.52 21.38 17.95 28.03 42.1
7 42.37 38.61
[41] 23.94 41.48 14.98 16.58 17.16 14.56 21.98 45.11 17.66 33.72 25.96 38.53 38.29 31.25 38.62 38.95 43.18 15.6
9 43.29 28.88
[61] 16.42 18.26 38.49 45.98 17.62 13.17 38.59 14.68 26.96 48.88 42.46 41.55 36.77 35.35 35.72 14.62 28.71 29.4
3 29.27 23.68
[81] 48.51 21.54 27.53 14.04 27.78 13.12 34.13 25.46 42.37 38.18 24.98 38.21 35.61 14.57 21.64 36.75 43.06 29.4
5 15.13 17.46
[101] 42.72 45.64 26.65 29.03 47.14 14.98 38.18 48.22 28.17 29.82 35.81 18.26 27.85 19.81 27.85 45.38 25.28 36.5
9 38.18 35.58
[121] 17.21 20.26 33.37 49.99 44.23 38.61 18.64 24.19 34.31 38.18 28.65 38.37 32.78 29.18 34.53 14.91 14.92 13.2
8 15.25 16.52
[141] 15.85 15.45 43.56 25.96 24.31 25.78 37.88 14.04 41.68 29.69 43.54 16.45 21.95 41.74 16.48 15.08 14.16 40.3
7 47.35 29.53
[161] 42.28 15.28 25.15 41.48 27.83 38.85 16.71 14.79 35.35 35.75 18.47 16.89 46.33 41.89 37.33 28.73 23.22 26.8
0 28.05 38.61
[181] 48.54 14.18 14.41 17.54 44.85 19.63 22.05 28.98 37.37 28.84 22.87 48.72 46.73 40.24
```

```
WHO$Country[which.min(WHO$Under15)]
[1] Japan
194 Levels: Afghanistan Albania Algeria Andorra Angola Antigua and Barbuda Argentina Armenia Australia Austria
... Zimbabwe
```

Let's create some plots for exploratory data analysis (EDA). First, let's create a basic scatterplot of GNI versus FertilityRate.



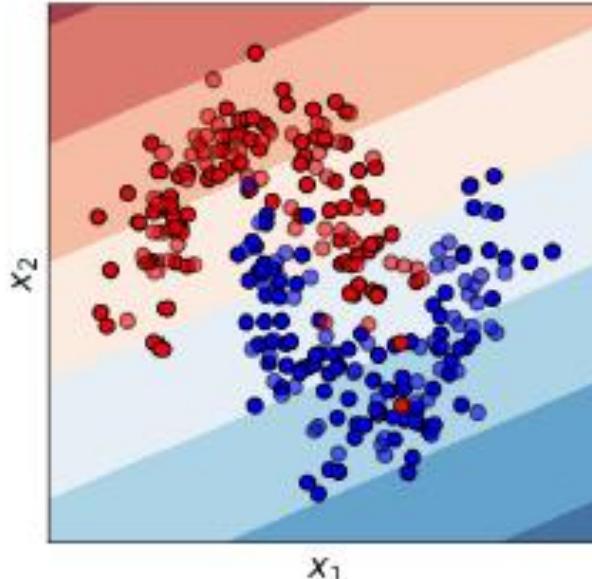
Linear vs. Logistic Regression

- Linear regression can be used to predict general continuous numbers
 - E.g., costs, length of stay, waiting time, etc.
- When predicting a **binary outcome** (i.e., binary classification problem) – **Logistic Regression**
 - Linear and logistic regression models may detect almost the same set of significant independent variables
 - Interpretations differ!
 - Assumptions for linear regression cannot be satisfied

Limitations of Logistic Regression

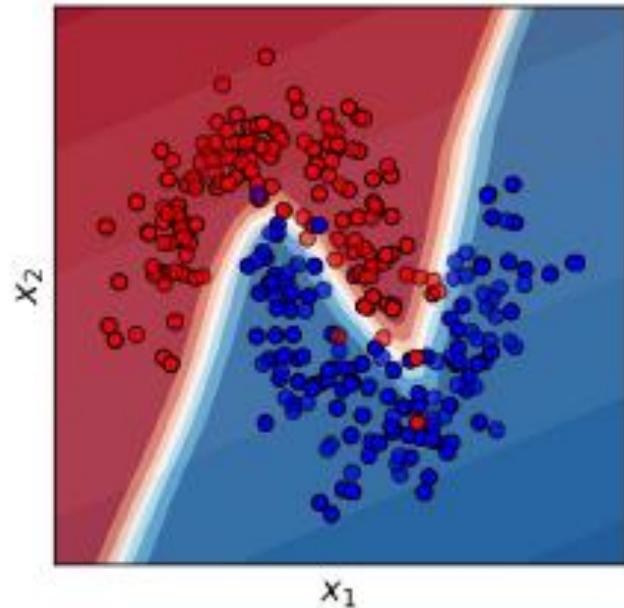
- Logistic Regression is a Linear Classifier
- Can split data into linear trends but can only represent limited relationships
- How to handle non-linearities?

<http://playground.tensorflow.org/>



Sess

tory Data Analy



Other Classification Models

- Decision Tree - Pioneered by ID3 algorithm in Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.
- Advanced Decision Trees and Ensemble Models:
 - Boosting and Bagging
- To be discussed in later lectures ...

Class Outline

- Review: Risk Factor Analysis and Linear Regression
- Logistic Regression
- Case Study: Understanding and Predicting Risk of Disease in Framingham Heart Study Using Logistic Regression

End

ECON 145

ECON 145 – Introductory Data Analytics in Healthcare

Lecture 6: Diagnostic and Causal Analytics

Lecture 5:

- Logistic Regression
- Model Evaluation
- Framingham Heart Study

Class Outline

- **Lecture 5 Recap**
- Diagnostic Analytics and Causal Inference
- Treatment Effect Estimation
- Case Study: Evaluating the Effect of Renal Replacement Therapy Using Propensity Scores

Risk Factor

- “A risk factor is any attribute, characteristic or exposure of an individual that increases the likelihood of developing a disease or injury.”
- The term was coined by William Kannel and Roy Dawber from the Framingham Heart Study



World Health Organization



Multivariate Risk Factors: Heart Failure

The routinely measured risk factors used in constructing the heart failure profile include age,

electrocardiographic left ventricular hypertrophy, cardiomegaly on chest x-ray film, heart rate, systolic blood pressure, vital capacity, diabetes mellitus, evidence of myocardial infarction, and valvular disease or hypertension.

The probability of developing heart failure was determined in subjects aged 45 through 94 years who had coronary disease, hypertension, or valvular heart disease, but were free of the condition at baseline. Over the course of the 38-year study period, among those with these predisposing conditions, there were 6354 person-examinations with follow-up in men and 8913 in women.

Based on 486 heart failure cases during 38 years of follow-up, 4-year probabilities of failure were computed using the pooled logistic regression model for each sex...

Table 3. Pooled Logistic Regression Model With Coefficients and Odds Ratios*

Variables	Units	Regression Coefficient	OR (95% CI)	P
Men				
Intercept		-9.2087		
Age	10 y	0.0412	1.51 (1.31-1.74)	<.001
LVH	Yes/no	0.9026	2.47 (1.31-3.77)	<.001
Heart rate	10 bpm	0.0166	1.18 (1.08-1.29)	<.001
Systolic blood pressure	20 mm Hg	0.00804	1.17 (1.04-1.32)	.007
CHD	Yes/no	1.6079	4.99 (3.80-6.55)	<.001
Valve disease	Yes/no	0.9714	2.64 (1.89-3.69)	<.001
Diabetes	Yes/no	0.2244	1.25 (0.89-1.76)	.20
Women				
Intercept		-10.7988		
Age	10 y	0.0503	1.65 (1.42-1.93)	<.001
LVH	Yes/no	1.3402	3.82 (2.50-5.83)	<.001
Heart rate	10 bpm	0.0105	1.11 (1.01-1.23)	.03
Systolic blood pressure	20 mm Hg	0.00337	1.07 (0.96-1.20)	.24
CHD				
Valve disease				
Diabetes				
BMI				
Valve disease and diabetes				

* Excludes vital capacity odds ratio; CI, confidence interval; CHD, congenital heart disease.

ORIGINAL INVESTIGATION

Profile for Estimating Risk of Heart Failure

William B. Kannel, MD, MPH; Ralph B. D'Agostino, PhD; Halit Silbershatz, PhD; Albert J. Belanger, MS; Peter W. F. Wilson, MD; Daniel Levy, MD

Context: We devised a risk appraisal function to assess the hazard of heart failure in persons who are predisposed by coronary disease, hypertension, or valvular heart disease.

Objectives: To provide general practitioners and internists with a cost-effective method to select people at high risk who are likely to develop left ventricular systolic dysfunction. This may then require further evaluation and aggressive preventive measures.

Methods: The routinely measured risk factors used in constructing the heart failure profile include age, electrocardiographic left ventricular hypertrophy, cardiomegaly on chest x-ray film, heart rate, systolic blood pressure, vital capacity, diabetes mellitus, history of myocardial infarction, and valvular disease or hypertension. Based on 486 heart failure cases during 38 years of follow-up, 4-year probabilities of failure were computed using the pooled logistic regression model for each sex; a single point score was employed. A multivariate profile was also produced without the vital cap-

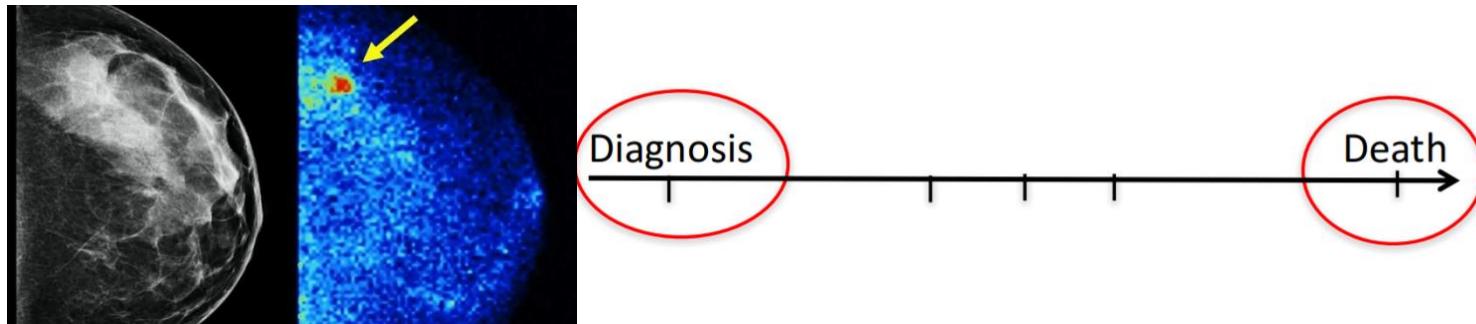
acity or chest x-ray film because these may not be readily available in some clinical settings.

Results: Using the risk factors that make up the multivariate risk formulation—derived from ordinary office procedures—the probability of developing heart failure can be estimated and compared with the average risk for persons of the same age and sex. Using this risk profile, for example, one may then require further evaluation and aggressive preventive measures.

Conclusions: Using this multivariate risk formulation, it is possible to identify high-risk candidates for heart failure who are likely to have a substantial yield of positive findings when tested for heart failure evidence (e.g., cardiomegaly, left ventricular dysfunction). The risk profile may also identify candidates who are at high risk for heart failure because of multiple, marginal risk factor abnormalities that might otherwise be overlooked.

Arch Intern Med. 1999;159:1197-1204

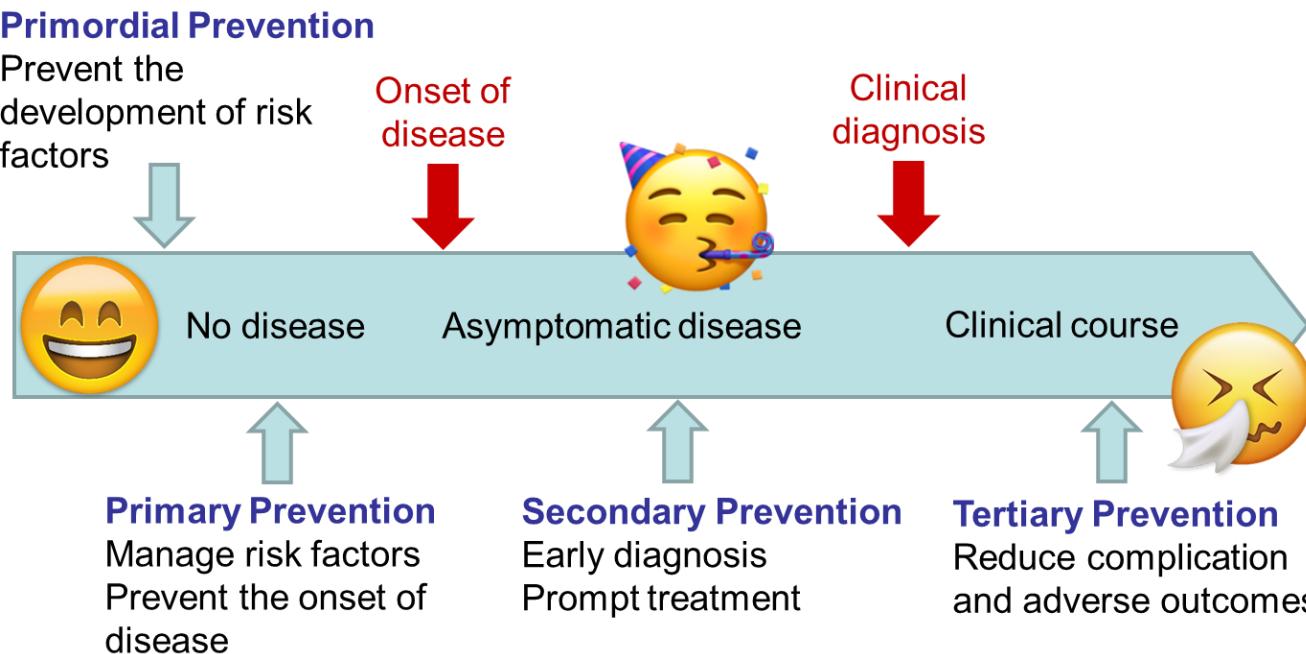
What is the Likelihood This Patient, With Breast Cancer, Will Survive 5 Years?



- Such predictive models are widely used to stage patients
- Should we initiate treatment? How aggressive?
- What could go wrong if we use such models to guide patient care?
 - The average 5-year survival rate for women with invasive breast cancer is 90%, and the average 10-year survival rate is 83% (2019 US Statistics)

Causality and Risk Factor Analysis

- The need to establish causality depends on applications
 - Control risk factors** for effective *prevention* usually **require causality**
 - Identifying signals** for timely *intervention* sometimes only **require correlation**



Recap Popup Quiz



1. Use www.wooclap.com
2. Enter a **nick name** with last 4 digits of your campus ID for class participation points e.g., smartguy7265

This is ungraded.

Lecture 5:

- Logistic Regression
- Model Evaluation
- Framingham Heart Study

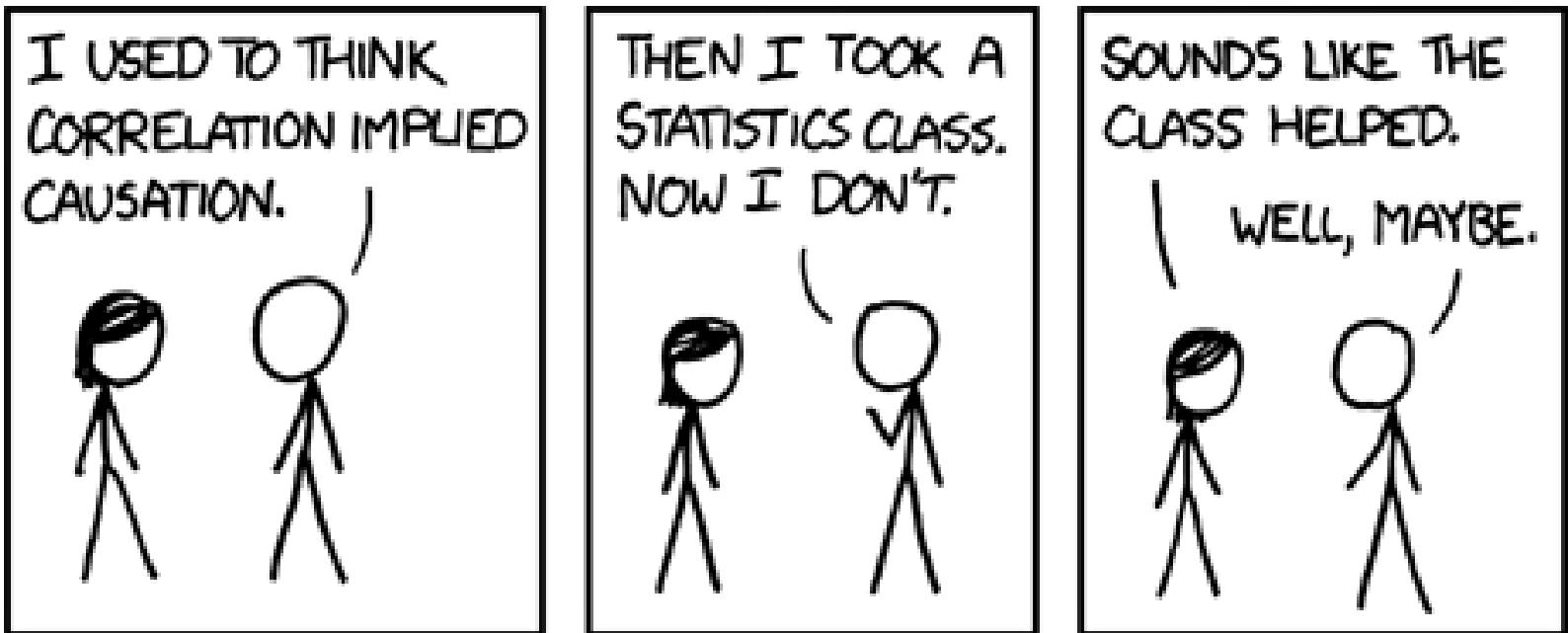
Class Outline

- Lecture 5 Recap
- **Correlation vs Causal Inference**
- Causal Inference Framework
- Case Study: Evaluating the Effect of Renal Replacement Therapy Using Propensity Scores

Diagnostic Analytics

- “Why did it happen?”
 - Identify patterns and anomalies
 - Drill into the analytics (discovery)
 - May require external information or data
 - **Determine causal relationships**
- Intertwined with descriptive and predictive analytics

Correlation vs. Causality



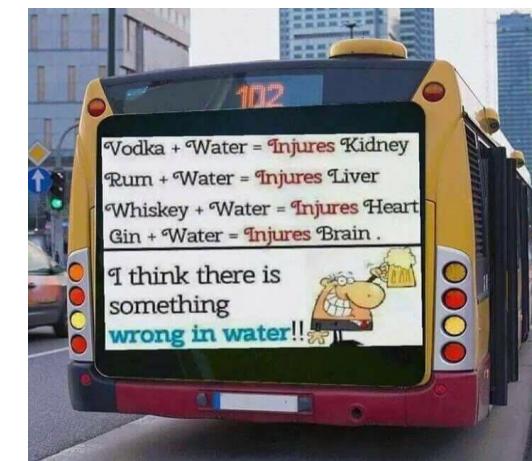
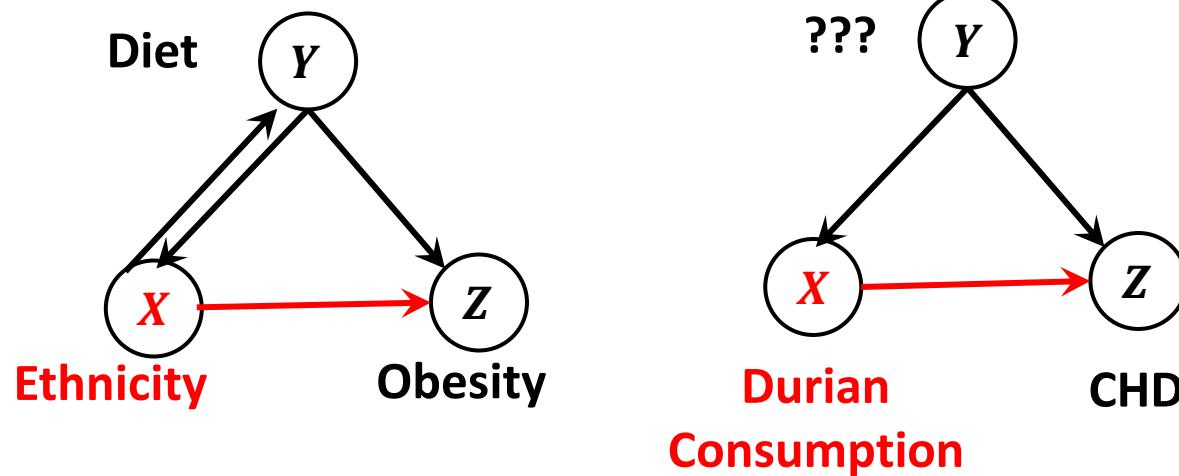
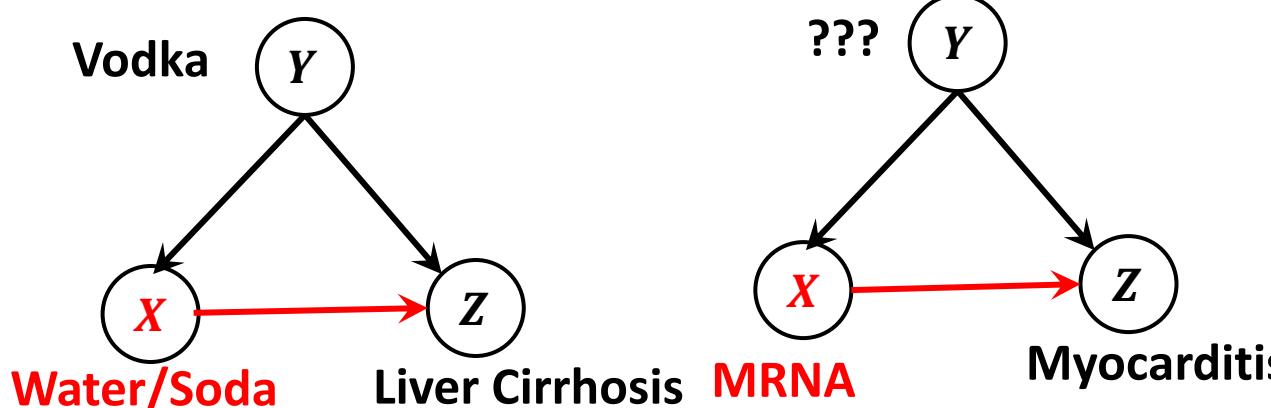
When Does It Matter?

- Not really if you only care about prediction accuracy, but...
- Critical for decision making and policy/treatment recommendation
 - To ensure your recommendation works

The Man Who Would Be Early

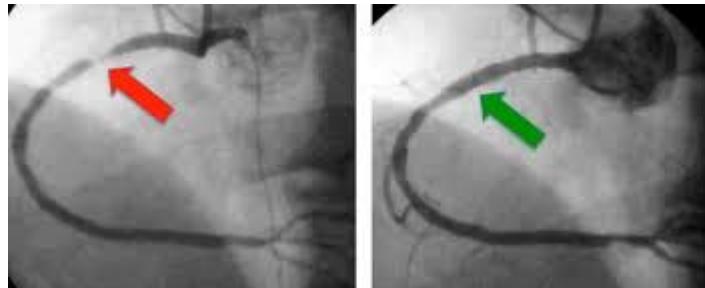
- Observation
 - Students who hurried vigorously on the way to their classes were almost always late
 - Students who walked at a normal pace were almost always on time
- Conclusion: **You must stop hurrying to get to your classes!**

More fun examples...

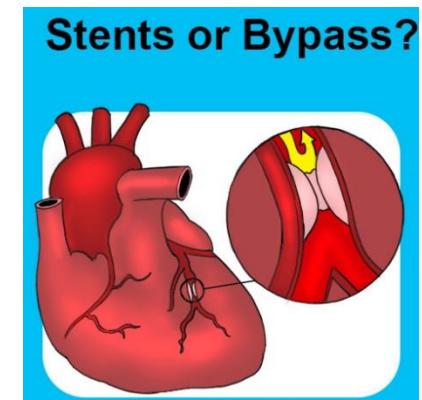




What Treatment Should We Give This Patient?



- People respond differently to treatment
- Use data from other patients and their journeys to guide future treatment decisions
- What could go wrong?



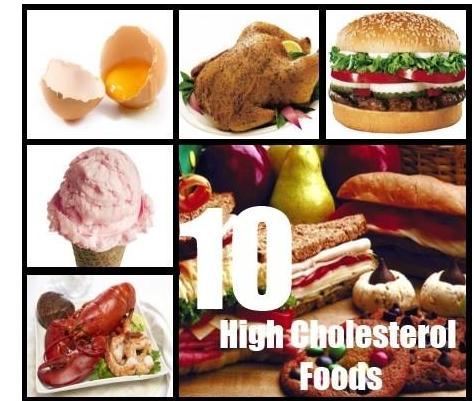
Does Smoking Cause Lung Cancer?

- Doing experiment is unethical
- Could we simply answer this question by comparing the lung cancer incidence rates among smokers and nonsmokers?



Is Cholesterol Intake Bad for Your Health?

- 2015–2020 Dietary Guidelines for Americans removed the recommendations of restricting dietary cholesterol to 300 mg/day
- “It is worth noting that most foods that are rich in cholesterol are also high in saturated fatty acids and thus may increase the risk of CVD due to the saturated fatty acid content.”



Nutrients. 2018 Jun; 10(6): 780.

Published online 2018 Jun 16. doi: [10.3390/nu10060780](https://doi.org/10.3390/nu10060780)

PMCID: PMC6024687

PMID: 29914176

Dietary Cholesterol and the Lack of Evidence in Cardiovascular Disease

Ghada A. Soliman

Is Moderate Drinking Good for You?

Red wine and resveratrol: Good for your heart?

Products and services

The Mayo Clinic Diet

What is your weight-loss goal?

5-10 lbs »

11-25 lbs »

25+ lbs »

Red wine and resveratrol: Good for your heart?

Resveratrol might be a key ingredient that makes red wine heart healthy. Learn the facts — and hype — about red wine and how it affects the heart.

By Mayo Clinic Staff

Red wine, in moderation, has long been thought of as heart healthy. The alcohol and certain substances in red wine called antioxidants may help prevent coronary artery disease, the condition that leads to heart attacks.

Any links between red wine and fewer heart attacks aren't completely understood. But part of the benefit might be that antioxidants in red wine may increase levels of high-density lipoprotein (HDL) cholesterol (the "good" cholesterol) and protect against cholesterol buildup.

NUTRITION

✓ Evidence Based

Can a Glass of Wine Benefit Your Health?

Benefits | Healthiest type | Downsides | Recommendation | Bottom line

People have been drinking wine for thousands of years, and the benefits of doing so have been well documented (1).

Emerging research continues to suggest that drinking wine in moderation — one drink a day — offers several benefits.

This article explains everything you need to know about the health benefits of wine, which type is healthiest, and its potential downsides.

Print

HARVARD HEALTH BLOG

<https://www.health.harvard.edu/blog/is-red-wine-good-actually-for-your-heart-2018021913285>

Is red wine actually good for your heart?

January 29, 2020

By Julie Corliss, Executive Ed.



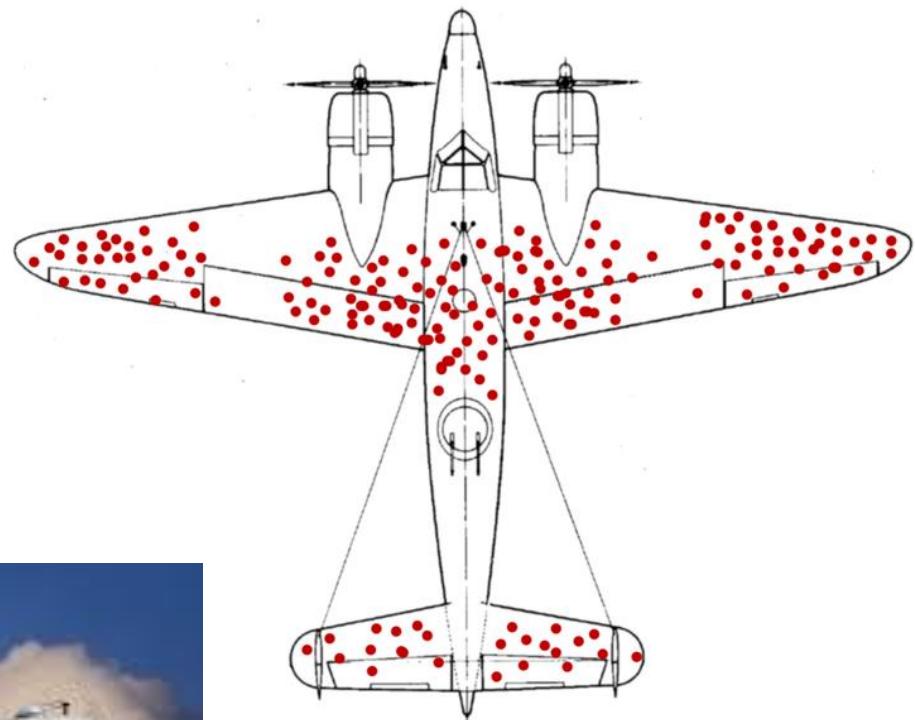
Harvard Health Publishing
HARVARD MEDICAL SCHOOL



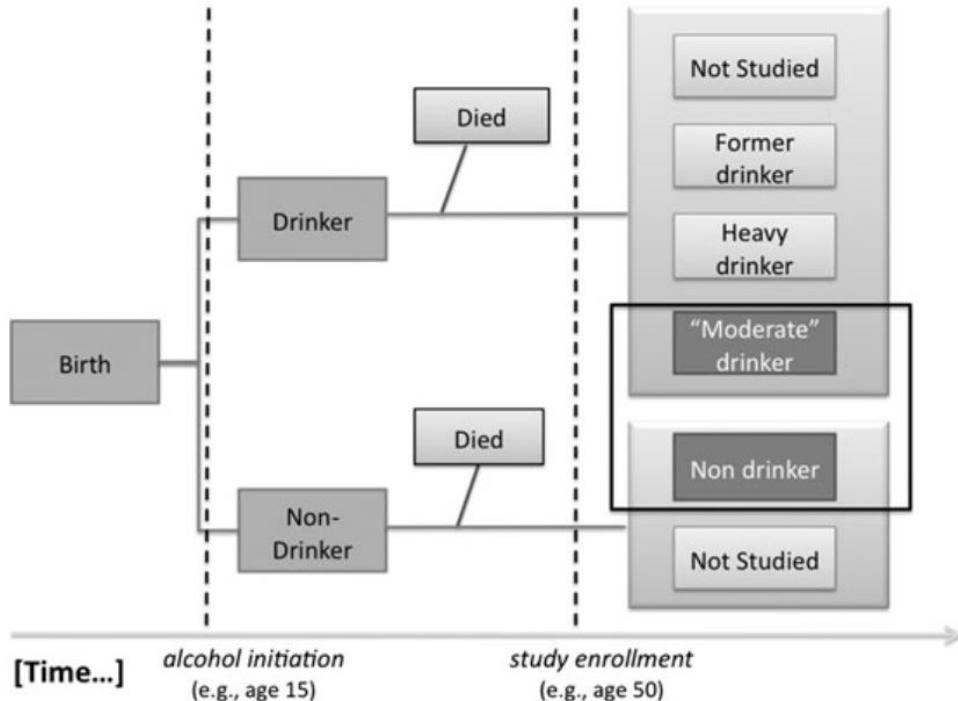
All of the research showing that people who drink moderate amounts of alcohol have lower rates of heart disease is observational. Such studies can't prove cause and effect, only associations.

Moderate drinking — defined as one drink per day for healthy women and two drinks per day for healthy men — is widely considered safe. But to date, the health effects of alcohol have never been tested in a long-term, randomized trial.

What is the story here?



Selection Bias

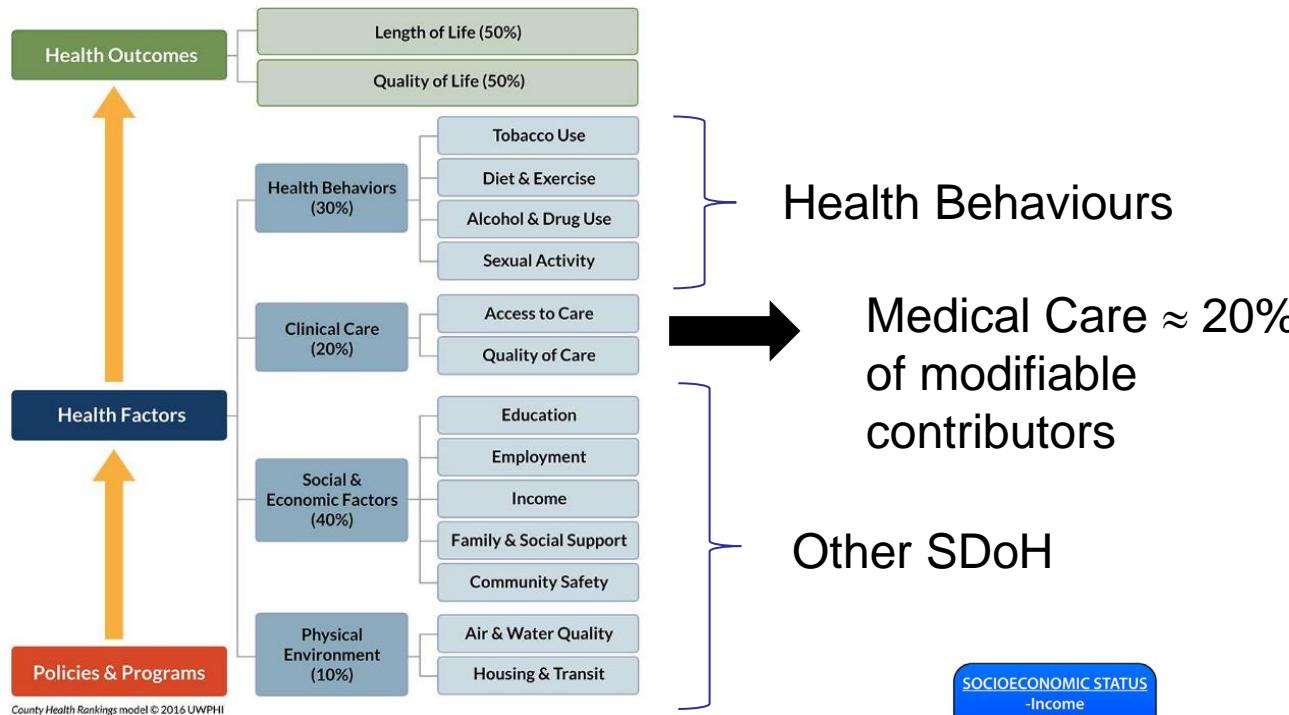


Flow-chart depicting drinking status and possible enrollment in studies in which 'moderate' (i.e. low-volume) drinkers are compared with non-drinkers

Potential Biases:

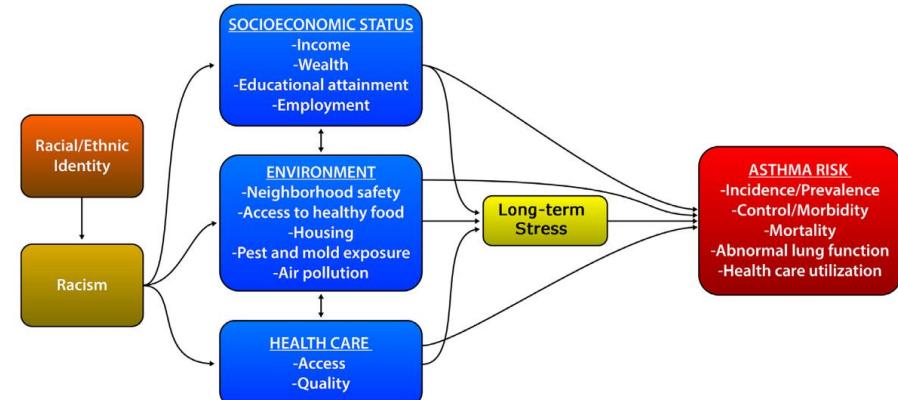
- Self Selection Bias:** A study of young adults found that drinkers had higher incomes and better educational attainment, engaged in more physical activity and were less likely to have illness compared with non-drinkers (Residual confounding)
- Willingness to participate:** Healthier drinkers or less risky drinkers would tend to be over-represented in study cohorts - "healthy survivor bias"
- Data analysis bias:** while alcohol consumption is assessed typically on the basis of average consumption, Most alcohol-related deaths typically from high per-occasion consumption (e.g. binge drinking). Many who drink modest amounts on the basis of average consumption also binge drink.
- There are others...**

SDoH in Health



<https://nam.edu/social-determinants-of-health-101-for-health-care-five-plus-five/>

80-90%: SDoH: **health-related behaviors, socioeconomic factors, and environmental factors**



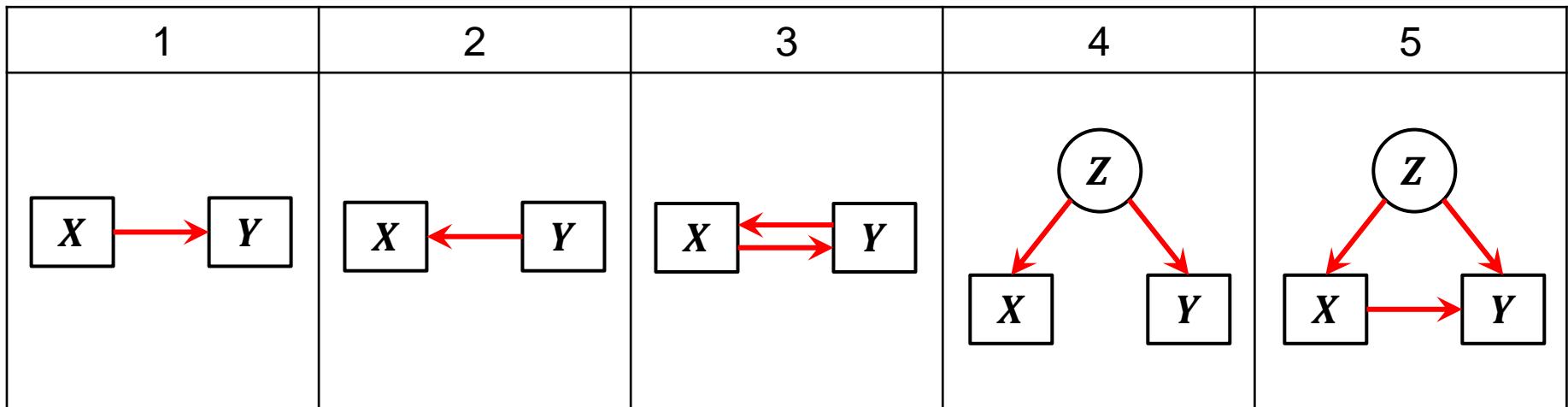
Lecture 5:

- Logistic Regression
- Model Evaluation
- Framingham Heart Study

Class Outline

- Lecture 5 Recap
- Correlation vs Causal Inference
- **Causal Inference Framework**
- Case Study: Evaluating the Effect of Renal Replacement Therapy Using Propensity Scores

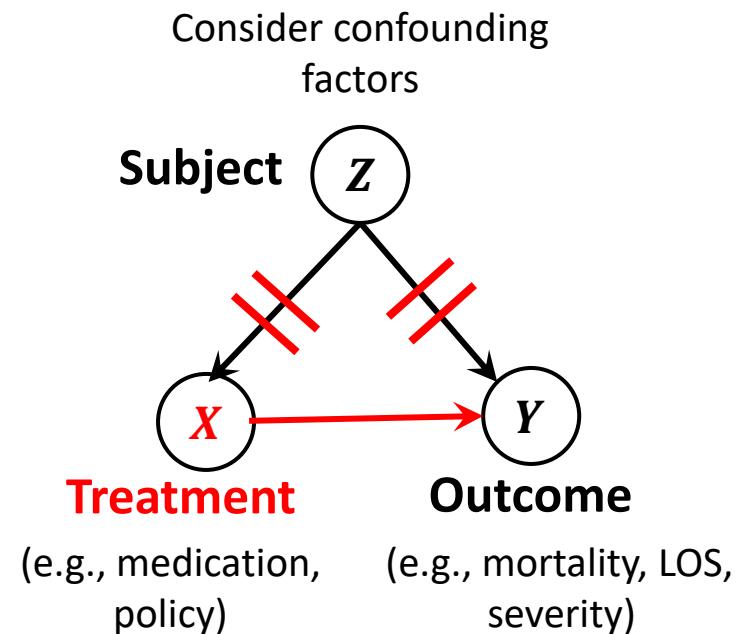
Correlations



- In all cases above, X and Y are correlated
- Is X the reason why Y behaves as such?
- To establish 1, we need to check 2 and 3, and we also need to control for possible Z in 4 and 5

Controlling for Confounders

- Confounders affect variables under study
- **Randomization breaks the links** between observed and unobserved confounders
- **Randomized Controlled Trials:**
 - Subjects participating in the trial are **randomly allocated** to either the group receiving the **treatment** under investigation or to a group receiving standard treatment (or placebo treatment) as the **control**
 - Minimizes **Selection Bias** upfront



Randomized Controlled Trial (RCT)

- Subjects participating in the trial are **randomly** allocated to either the group receiving the **treatment** under investigation or to a group receiving standard treatment (or placebo treatment) as the **control**
- Ideal method for treatment effect assessment
 - **Random selection** into treatment by trial designers
 - Follow standard protocols to perform random selection
 - Need sufficient number of subjects, blinded etc.
 - Well-blinded RCT minimize **selection bias**

These articles are being published Online First to coincide with World Health Organization regarding use of corticosteroids for patients.

ORIGINAL INVESTIGATION**Caring for the Critically Ill Patient****Effect of Hydrocortisone on 21-Day Mortality or Respiratory Support Among Critically Ill Patients With COVID-19: A Randomized Clinical Trial**

Pierre-François Dequin, MD, PhD; Nicholas Heming, MD, PhD; Ferhat Mezian, PhD; et al.

Caring for the Critically Ill Patient**Effect of Dexamethasone on Days Alive and Ventilator-Free Days Among Patients With Moderate or Severe Acute Respiratory Distress Syndrome and COVID-19: The CoDEX Randomized Clinical Trial**

Bruno M. Tomazini, MD; Israel S. Maia, MD, MSc; Alexandre B. Cavalcanti, M et al.

Caring for the Critically Ill Patient**Effect of Hydrocortisone on Mortality and Organ Support Among Patients With Severe COVID-19: The REMAP-CAP COVID Corticosteroid Domain Randomized Clinical Trial**

The Writing Committee for the REMAP-CAP Investigators

Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine

L.R. Baden, H.M. El Sahly, B. Essink, K. Kotloff, S. Frey, R. Novak, D. Diemert, S.A. Spector, N. Roush, C.B. Creech, J. McGettigan, S. Khetan, N. Segall, J. Solis, A. Brosz, C. Fierro, H. Schwartz, K. Neuzil, L. Corey, P. Gilbert, H. Janes, D. Follmann, M. Marovich, J. Mascola, L. Polakowski, J. Ledgerwood, B.S. Graham, H. Bennett, R. Pajon, C. Knightly, B. Leav, W. Deng, H. Zhou, S. Han, M. Ivarsson, J. Miller, and T. Zaks, for the COVE Study Group*

ABSTRACT**BACKGROUND**

Vaccines are needed to prevent coronavirus disease 2019 (Covid-19) and to protect persons who are at high risk for complications. The mRNA-1273 vaccine is a lipid nanoparticle-encapsulated mRNA-based vaccine that encodes the prefusion stabilized full-length spike protein of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus that causes Covid-19.

METHODS

This phase 3 randomized, observer-blinded, placebo-controlled trial was conducted at 99 centers across the United States. Persons at high risk for SARS-CoV-2 infection or its complications were randomly assigned in a 1:1 ratio to receive two intramuscular injections of mRNA-1273 (100 µg) or placebo 28 days apart. The primary end point was prevention of Covid-19 illness with onset at least 14 days after the second injection in participants who had not previously been infected with SARS-CoV-2.

RESULTS

The trial enrolled 30,420 volunteers who were randomly assigned in a 1:1 ratio to receive either vaccine or placebo (15,210 participants in each group). More than 96% of participants received both injections, and 2.2% had evidence (serologic, virologic, or both) of SARS-CoV-2 infection at baseline. Symptomatic Covid-19 ill-

The authors' full names, academic degrees, and affiliations are listed in the Appendix. Address reprint requests to Dr. El Sahly at the Departments of Molecular Virology and Microbiology and Medicine, 1 Baylor Plaza, BCM-MS280, Houston, TX 77030, or at hana.elsahly@bcm.edu; or to Dr. Baden at the Division of Infectious Diseases, Brigham and Women's Hospital, 15 Francis St., PBB-A4, Boston, MA 02115, or at lbaden@bwh.harvard.edu.

*A complete list of members of the COVE Study Group is provided in the Supplementary Appendix, available at NEJM.org.

Drs. Baden and El Sahly contributed equally to this article.

This article was published on December 30, 2020, and updated on January 15, 2021, at NEJM.org.

N Engl J Med 2021;384:403-16.

DOI: 10.1056/NEJMoa2035389

Copyright © 2020 Massachusetts Medical Society.

Limitations of RCT

- RCT is not feasible for all causal effect studies
 - Costs, patient recruitment, legal issues, etc.
- Small RCT may still suffer from estimation bias
- Large RCT could be both costly and time-consuming
- **Efficacy ≠ Effectiveness**
- Can we estimate treatment effect from **observational data?**

A Randomized Controlled Trial of a Behavioral Intervention to Reduce High-Risk Sexual Behavior Among Adolescents in STD Clinics

CAROL W. METZLER

ANTHONY BIGLAN

JOHN NOELL

DENNIS V. ARY

LINDA OCHS

Oregon Research Institute

A five-session behavioral intervention to reduce risky sexual behavior was evaluated in a randomized controlled trial, in which 339 adolescents, ages 15 to 19 years, were recruited in public sexually transmitted disease clinics and randomly assigned to receive the intervention or usual care. The intervention targeted (a) decision-making about safer sex goals, (b) social skills for achieving safer sex, and (c) acceptance of negative thoughts and feelings. Compared to the control group at 6-months follow-up, treatment participants reported fewer sexual partners, fewer nonmonogamous partners, and fewer sexual contacts with strangers in the past 3 months, and less use of marijuana before or during sex. Treated adolescents also performed better on a taped situations test of skill in handling difficult sexual situations. Strongest intervention effects were for male and nonminority youth. Further research is needed to develop interventions with strong, durable effects across gender and ethnic groups that can be delivered cost-effectively within existing service systems.

Treatment Program:

1. Prompt decision to reduce risky sexual behaviour; set a safer sex goal
2. Increasing social skills in handling difficult sexual situation
3. Increasing willingness to experience unpleasant thoughts/feelings that accompany changes in sexual behaviours

Metzler CW, Biglan A, Noell J, Ary DV, Ochs L. A randomized controlled trial of a behavioral intervention to reduce high-risk sexual behavior among adolescents in STD clinics. *Behavior Therapy*. 2000;31(1):27-54. doi:[10.1016/S0005-7894\(00\)80003-9](https://doi.org/10.1016/S0005-7894(00)80003-9)

t Estimation

September 21, 2020

Efficacy of Smartphone Applications for Smoking Cessation

A Randomized Clinical Trial

Jonathan B. Bricker, PhD^{1,2}; Noreen L. Watson, PhD¹; Kristin E. Mull, MS¹; et al

» Author Affiliations | Article Information

JAMA Intern Med. Published online September 21, 2020. doi:10.1001/jamainternmed.2020.4055

Key Points

Question Is a smartphone application based on acceptance and commitment therapy (ACT) efficacious for smoking cessation?

Findings In this 2-group stratified, double-blind, individually randomized clinical trial of 2415 adult smokers with a 12-month follow-up and high retention, participants assigned to the smartphone application based on ACT had 1.49 times higher odds of quitting smoking compared with the participants assigned to the smartphone application based on US clinical practice guidelines.

Meaning Compared with a US clinical practice guidelines-based application that teaches avoidance of smoking triggers, an ACT-based application that teaches acceptance of smoking triggers was more efficacious for quitting smoking.

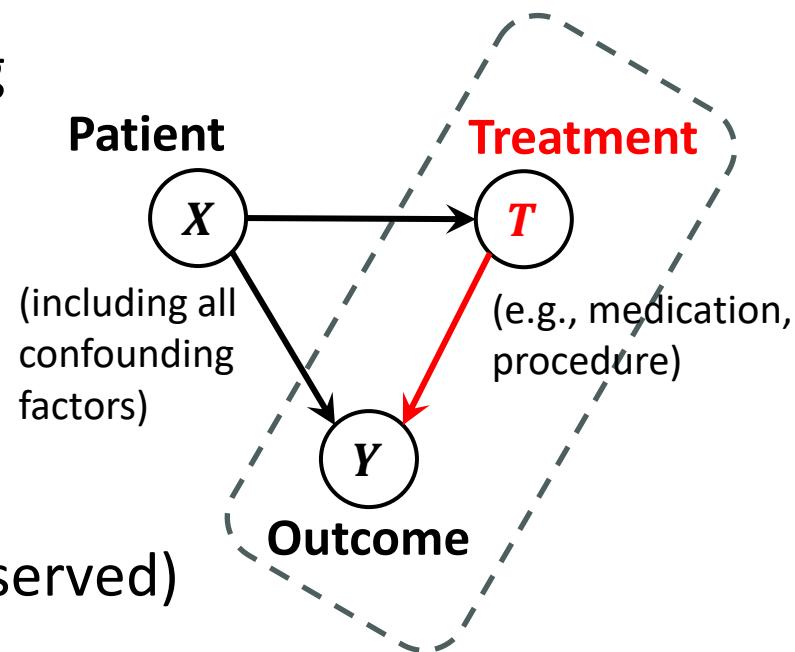
RCT is harder for SDoH!

- Interventions are complex and demanding
- Behaviours and outcomes related to health are influenced by complex social and economic factors – hard to balance our external confounders
- Excluding choice through randomization may introduce bias
- Effectiveness of behavioural interventions can be motivated by choice itself!
- Have to **estimate treatment effect from observational data**

Stephenson J, Imrie J. Why do we need randomised controlled trials to assess behavioural interventions? *BMJ*. 1998;316(7131):611-613. doi:[10.1136/bmj.316.7131.611](https://doi.org/10.1136/bmj.316.7131.611)

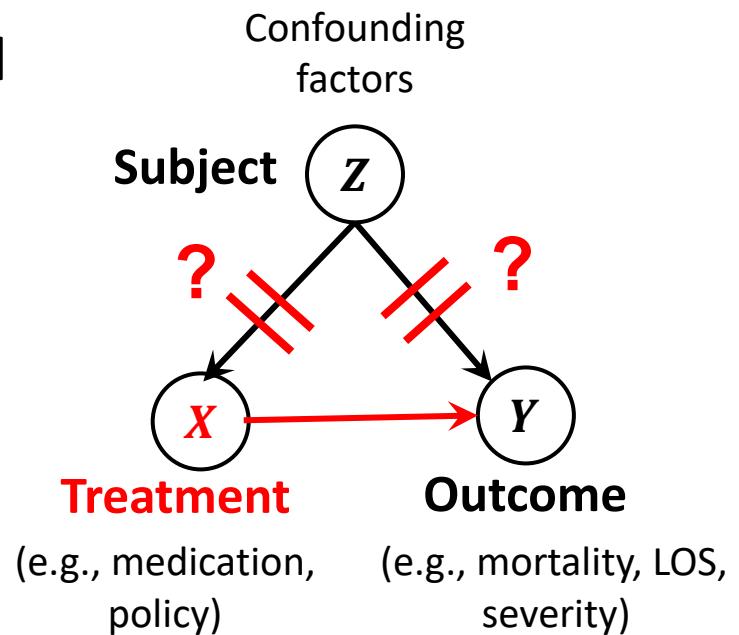
Treatment Effect Estimation

- Investigating the effects of adopting a treatment on one or more outcome variables
- Y : Outcome
- T : Treatment indicator
- X : Other control variables (a.k.a. covariates; observed & unobserved)
- Assessing the ***treatment effect***
 - What would have happened to those who, in fact, received treatment, if they have not received treatment (or vice versa)?



Observational Studies

- Treatment effectiveness could be estimated in a **time- and money-efficient manner with observational data**, but
 - Treatment assignment is not random
- **Observation data are not randomized upfront**
- Traditional statistical analysis, like linear regression, may provide biased results due to self-selection into treatment -> **Selection bias**



Classical Causal Inference

- Bradford Hill criteria
 - Strength
 - Consistency
 - Specificity
 - Temporality
 - Biological gradient
 - Plausibility
 - Coherence
 - Experiment
 - Analogy
- Qualitative and subjective evaluation
 - Weak
 - Moderate
 - Strong

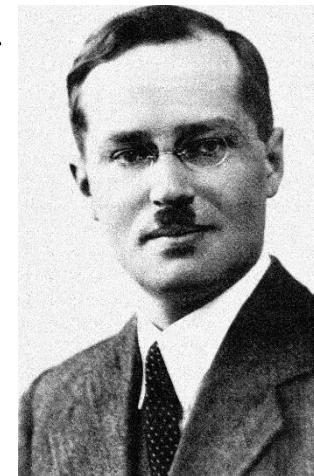


Sir Austin Bradford Hill CBE FRS
(1897 – 1991)

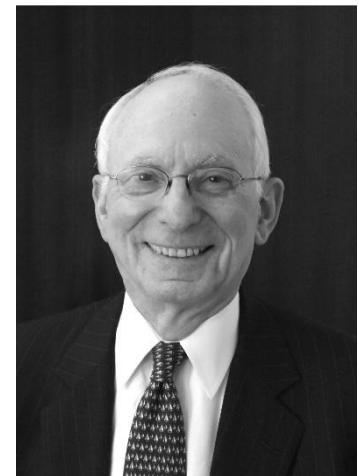
Fedak KM, Bernal A, Capshaw ZA, Gross S. Applying the Bradford Hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. *Emerg Themes Epidemiol.* 2015;12(1):14. doi:[10.1186/s12982-015-0037-4](https://doi.org/10.1186/s12982-015-0037-4)

Potential Outcomes Framework

- Potential outcomes framework
 - Neyman–Rubin causal model
 - Rubin causal model
- Quantitative approach
 - Estimating the treatment effect
 - What would have happened to those who, in fact, received treatment (e.g., medication, policy), if they have not received treatment (or vice versa)?



Jerzy Spława-Neyman
(1894 – 1981)



Donald Bruce Rubin
(1943 –)

Missing Data Problem

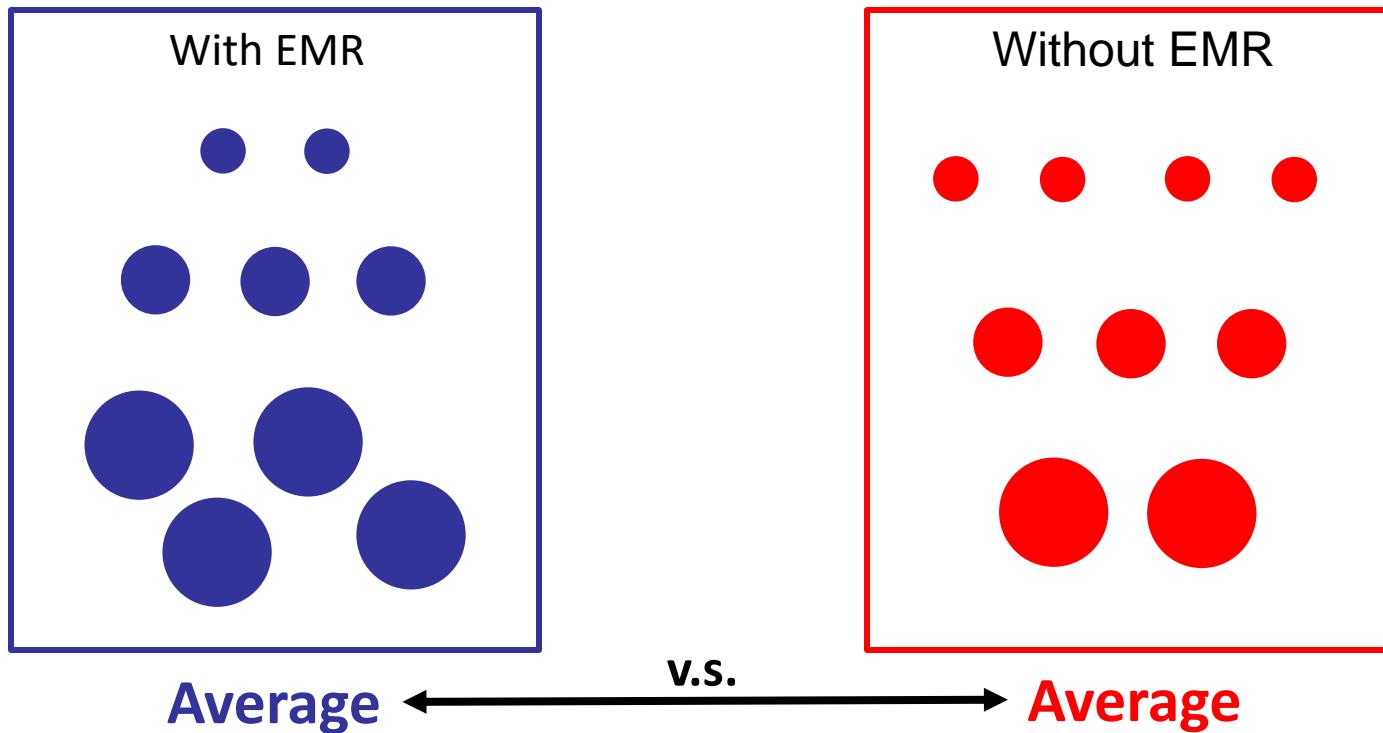
- E.g., Effect of new diabetes management program on reducing HbA1c levels

	HbA1c Level Without Program	HbA1c Level With Program	Reduction in HbA1c
<i>Both actual and potential outcomes are known</i>			
Patient Group A (with program)	8.5%	6.0%	2.5%
Patient Group B (without program)	9.0%	8.0%	1%
<i>Only actual outcomes are known</i>			
Patient Group A (with program)	-	6.0%	?
Patient Group B (without program)	9.0%	-	?

- We only ever observe one of the two outcomes

Estimating Average Treatment Effect (ATE)

- Simple estimation



Selection Bias

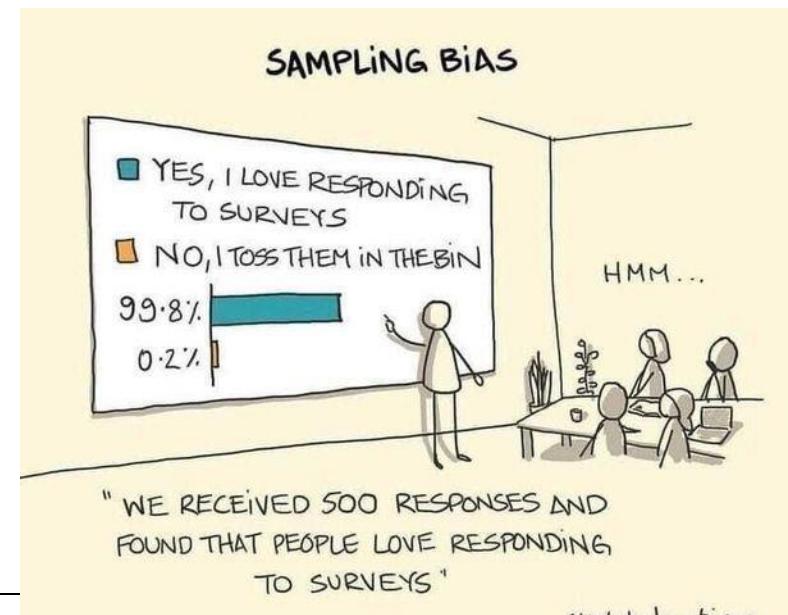
- Selection bias arises when the treatment assignment decisions depend on observed variables or factors that are not observed but also affect the outcome

From the use case:

1st case: Selection on observables

- Patients who are younger in group A vs group B
- More educated and physically more active?

2nd case: Selection on unobservables



Matching

- Find each unit's long-lost counterfactual identical twin, check up on his outcome



*Had Zelensky
become a
President?*

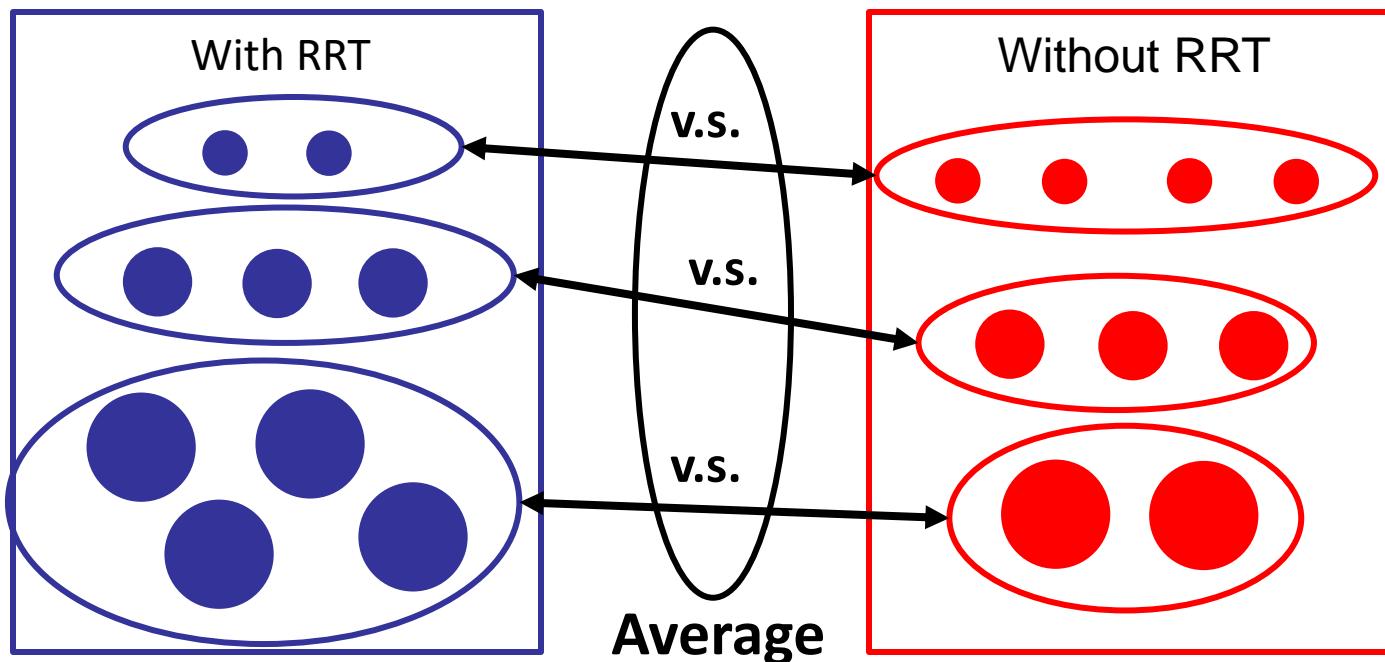


*Had Zelensky stay
as an actor?*

**Identify hidden
“natural
experiments”
from
observational
data**

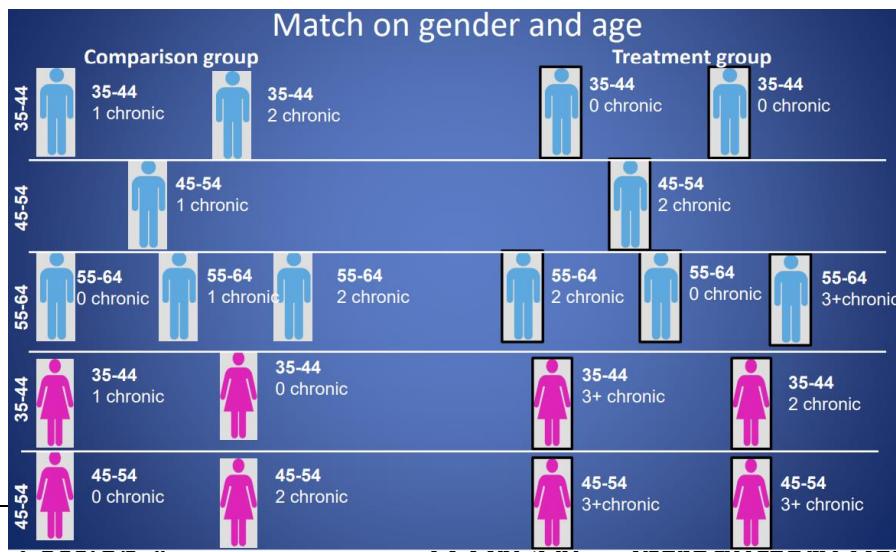
Estimating Average Treatment Effect (ATE)

- Matching: Compare with “similar” subjects



Matching

- To make **Exposed and Unexposed (Control) groups as similar as possible** before proceeding with analysis
- Matching of patients in order to **simulate the random assignments of treatment/exposure** to each group (equal probability of assignment)
- For each treatment group, find at least one untreated patient from the comparison group with similar observable characteristics



What if you have > 10 covariates?

Is there a composite matching score?

Propensity Score Matching (PSM)

- Matching: Match treated and untreated observations on all the observed variables
 - Sometimes can be very challenging
- PSM: Match treated and untreated observations on the estimated probability of being treated (propensity score)
 - Reduce to one dimension, match only a number
 - Need some technical assumptions
 - Rosenbaum and Rubin (1983)

Propensity Score Matching (PSM)

Choose observable variables for the evaluation of the propensity score (PS)

Use logistic regression (commonly used) to obtain a PS for each subject

Choose a matching algorithm

Match exposed and unexposed subjects based on the PS (random assignment)

Check the balance of covariates in the exposed and unexposed groups after matching on PS (Balance Diagnostics)

Proceed with analyses based on sample matched or weighted by propensity score

Propensity Score Matching (PSM)

- Propensity score for a unit i , $e(\mathbf{Y}_i)$, can be estimated from logistic regression of the treatment condition z_i on the covariate vector \mathbf{Y}_i and coefficient vector β (Agresti, 2013):

$$\ln \left(\frac{e(\mathbf{Y}_i)}{1 - e(\mathbf{Y}_i)} \right) = \beta_0 + \boldsymbol{\beta} \mathbf{Y}_i$$

- Propensity score: $\hat{e}(\mathbf{Y}_i)$

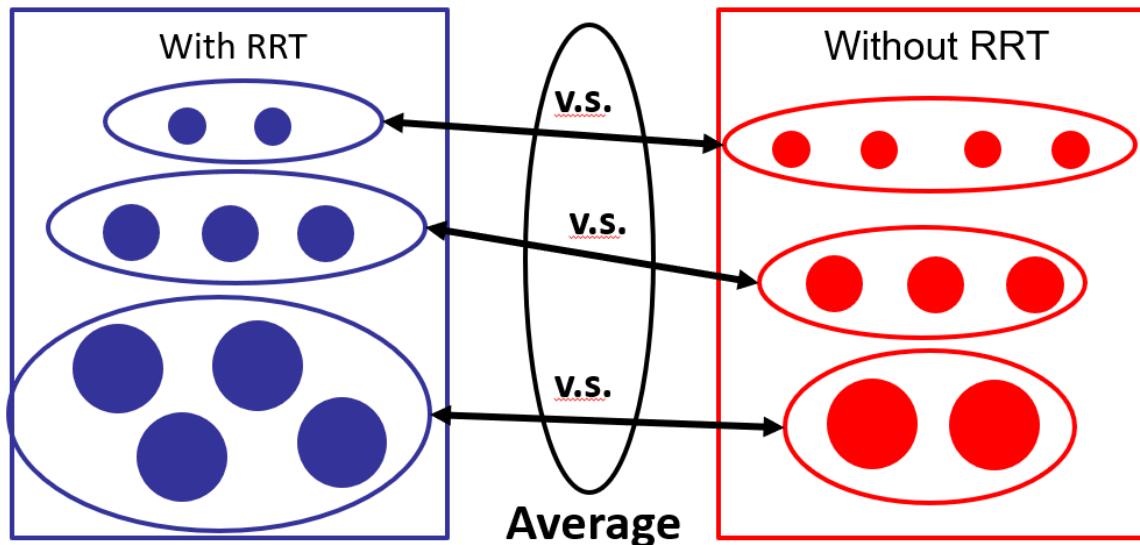
PSM – Simple Example

Patient ID	Age	Severity of Hypertension	Propensity Score	Received Drug?
P1	65	Severe	0.90	Yes
P2	66	Severe	0.89	No
P3	45	Mild	0.20	Yes
P4	44	Mild	0.18	No

Feature	High Propensity Scores	Low Propensity Scores
Likelihood of receiving treatment	High (~80% or more)	Low (~20% or less)
Covariates (age, severity, etc.)	Similar across matched groups	Different from high-propensity patients

After Matching – Average Treatment Effect

- Estimating the causal effects of treatments after dealing with the missing data problem (matching)
- Causal effects are comparisons of responses $r_{1,i}$ and $r_{0,i}$,
 - E.g. $r_{1,i} - r_{0,i}$ or $r_{1,i}/r_{0,i}$
- The usual quantity to be estimated is the Average treatment effect (ATE):
 - $\mathbb{E}(r_1) - \mathbb{E}(r_0)$



Propensity Score Matching (PSM)

- Typically used when **randomization or other quasi experimental options are not possible**
- Propensity score (PS) is the **conditional probability of treatment given the observed covariates (Y)**
- PS attempts to group observations such that the probability of exposure is approximately similar:
$$e(Y) = P(\text{Exposed} = 1|Y) = P(\text{Exposed} = 0|Y)$$
- **Estimated without knowledge of the outcome variable**, only the confounders
- PS helps to **control only for OBSERVABLE confounders**
- What about NON-UNOBSERVABLE confounders?

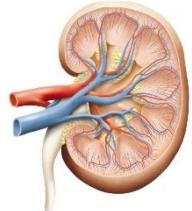
Lecture 5:

- Logistic Regression
- Model Evaluation
- Framingham Heart Study

Class Outline

- Lecture 5 Recap
- Correlation vs Causal Inference
- Causal Inference Framework
- **Case Study: Evaluating the Effect of Renal Replacement Therapy Using Propensity Scores**

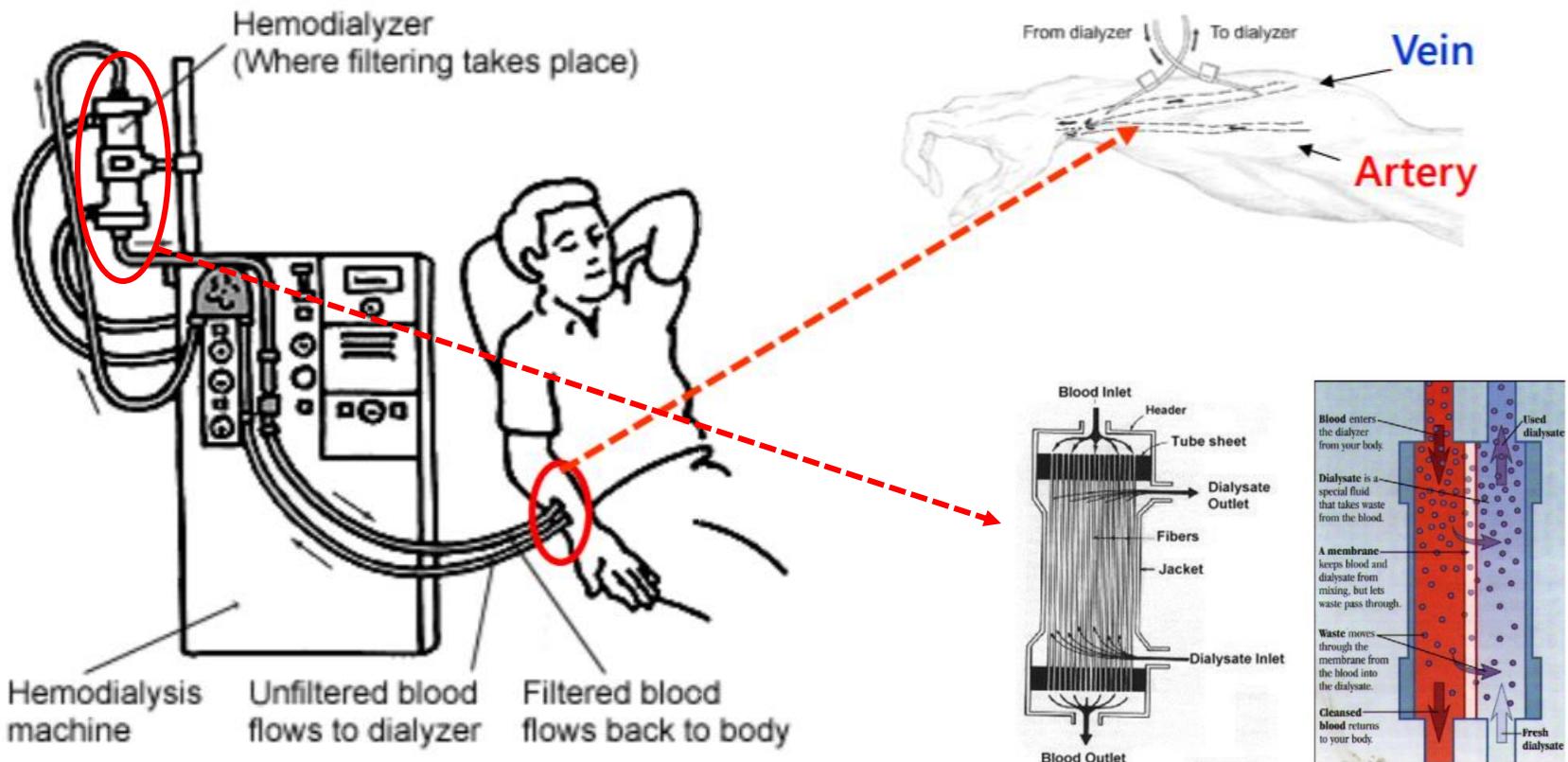
Case Background



- Acute Kidney Injury (AKI)
 - Common complication in critically ill patients: Occurring in 25–65% of the patients in the intensive care units (ICUs)
 - 5–8% of these patients will require RRT
- Renal Replacement Therapy (RRT)
 - Also known as hemodialysis (or simply dialysis)
 - Continuous renal replacement therapy (CRRT) has become increasingly popular
 - No clear guidelines on whether to initiate, when to start/stop, mode of control, dosage, etc.
 - A lot of open debates



Principle of RRT



Pros and Cons of RRT

- Advantages
 - More hemodynamic stability
 - Allowing more adequate fluid removal
 - Better recovery of renal function
 - More efficient removal of small and large metabolites
- Adverse events (side effects)
 - Machine dysfunction
 - Clinical complications
 - Infection, air embolism, hypothermia, anticoagulation, hypotension, electrolyte disturbances, nutritional losses
 - Depression
 - Lifetime dependence

Is RRT Effective in Reducing Mortality?

- Almost unanimous agreement among physicians on the effectiveness of RRT
- Open question in medical research
 - A lot of research effort but no conclusions
- Why?

Is RRT Effective in Reducing Mortality?

- No randomized controlled trials (RCT)
 - Economic, human resources, ethical limitations
- Only uncontrolled observational studies
 - Treatment selection bias
 - Patients' underlying severity are major confounders
 - Need rigorous approach to reduce bias

Effort in Medical Research

- Elseviers et al. (2010)
 - Included individual Stuivenberg Hospital Acute Renal Failure (SHARF) scores in a logistic regression model
 - Increased risk of mortality for RRT
- Clec'h et al. (2012)
 - Used propensity score matching (PSM) to control for treatment selection bias
 - RRT failed to reduce mortality
 - Similar to our conclusion in R exercise

Effort in Medical Research

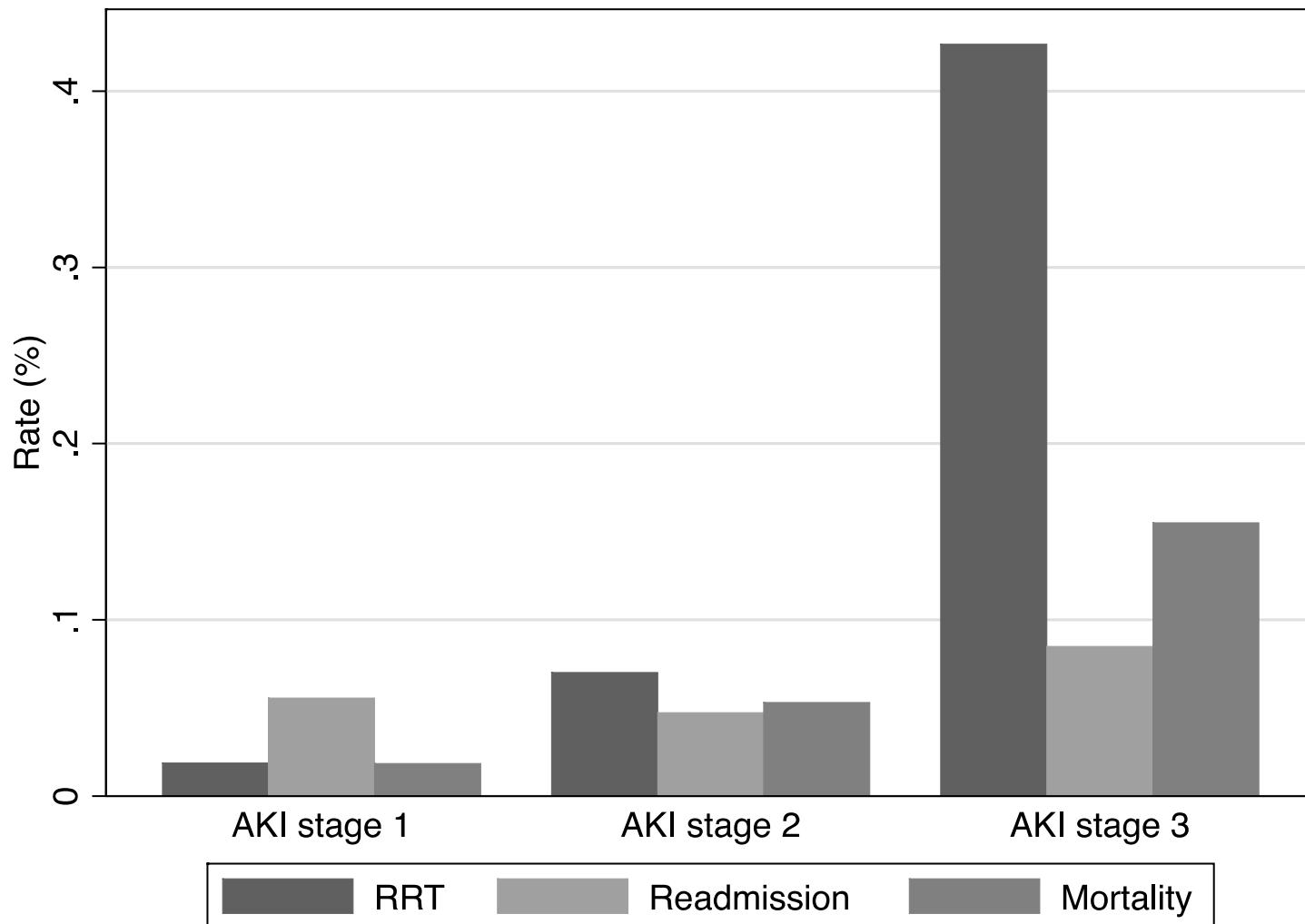
- These results should be cautiously interpreted
- Elseviers et al. (2010)
 - The observed higher mortality “may be due to differences in *severity of disease* in general and renal failure in particular”
- Clec'h et al. (2012)
 - PSM can only account for *observed covariates*, but not for *unobserved covariates*, which may still cause bias

Treatment Effect Estimation

Variable	ALL	RRT	Non-RRT	p-value
Age	60.1 (13.4)	63.2 (12.4)	59.8 (13.5)	0.000
Males	73.9%	69.2%	74.4%	0.010
<i>Race</i>				0.004
Chinese	65.4%	62.8%	65.7%	
Malay	15.1%	19.7%	14.6%	
Indian	8.3%	9.2%	8.2%	
Others	11.2%	8.3%	11.5%	
<i>Admission_type</i>				0.000
Emergency	46.1%	61.3%	44.4%	
Elective	44.3%	24.1%	46.6%	
Other	9.6%	14.6%	9.0%	
<i>AKI_stage</i>				0.000
I	58.2%	11.3%	63.4%	
II	23.6%	13.9%	24.7%	
III	18.2%	74.8%	11.9%	
BaseCr ($\mu\text{mol/L}$)	127.5 (142.7)	297.2 (220.9)	109.4 (118.0)	0.000
SOFA score	5.6 (4.9)	9.4 (4.9)	5.1 (4.7)	0.000
Creatinine ($\mu\text{mol/L}$)	138.9 (152.8)	327.1 (214.6)	116.8 (126.5)	0.000
Potassium (mmol/L)	4.2 (0.6)	4.4 (0.7)	4.2 (0.5)	0.000
pH	7.4 (0.1)	7.4 (0.1)	7.4 (0.1)	0.000
Temperature	36.3 (1.5)	36.0 (1.6)	36.3 (1.5)	0.000
Urea (mmol/L)	7.7 (5.7)	15.0 (8.6)	6.9 (4.6)	0.000
24hr UO (mL)	1601.5 (769.7)	667.1 (744.5)	1694.1 (708.0)	0.000
In-hospital mortality	4.6%	28.4%	1.9%	0.000
30-day readmission	5.7%	9.0%	5.4%	0.001
ICU LOS (days)	4.1 (6.6)	11.2 (15.0)	3.4 (4.3)	0.000
<i>Observation</i>	5,779	584	5,195	

AKI, Acute Kidney Injury; RRT, Renal Replacement Therapy;
 BaseCr, Baseline Creatinine; SOFA, Sequential Organ Failure Assessment;
 UO, Urine Output; LOS, Length of Stay.

Treatment Effect Estimation



Propensity Score Matching

- Compare in-hospital mortality between matched patients with and without RRT, who had a close probability of receiving the treatment

	Treated	Control	ATT ^a	p-value
Before Matching	0.383	0.090	0.293	0.001
After Matching	0.383	0.212	0.171	0.021

^a ATT, Average Treatment Effect

Change in mortality

Let's get our hands dirty!

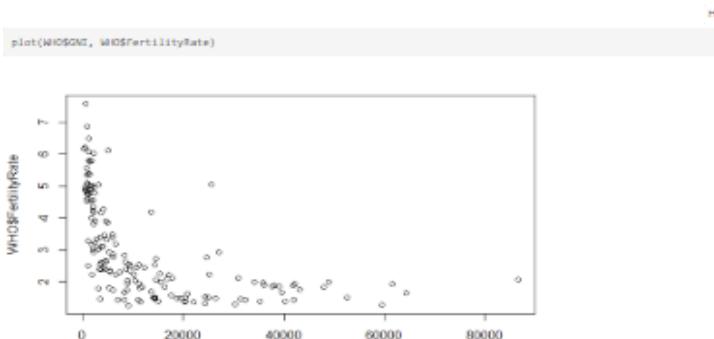
Data Analysis

Let's do some basic data analysis using our WHO data.

```
WHO$Under15
[1] 47.42 21.33 27.42 15.28 47.58 25.96 24.42 28.34 18.95 14.51 22.25 21.62 28.16 30.57 18.99 15.18 16.88 34.4
8 42.95 28.53
[21] 35.23 16.35 33.75 24.56 25.75 13.53 45.66 44.20 31.23 43.08 16.37 38.17 48.87 48.52 21.38 17.95 28.03 42.1
7 42.37 38.61
[41] 23.94 41.48 14.98 16.58 17.16 14.56 21.98 45.11 17.66 33.72 25.96 38.53 38.29 31.25 38.62 38.95 43.18 15.6
9 43.29 28.88
[61] 16.42 18.26 38.49 45.98 17.62 13.17 38.59 14.68 26.96 48.88 42.46 41.55 36.77 35.35 35.72 14.63 28.71 29.4
3 29.27 23.68
[81] 48.51 21.54 27.53 14.04 27.78 13.12 34.13 25.46 42.37 38.18 24.98 38.21 35.61 14.57 21.64 36.75 43.06 29.4
5 15.13 17.46
[101] 42.72 45.64 26.65 29.03 47.14 14.98 38.18 48.22 28.17 29.82 35.81 18.26 27.85 19.81 27.85 45.38 25.28 36.5
9 38.18 35.58
[121] 17.21 20.26 33.37 49.99 44.23 38.61 18.64 24.19 34.31 38.18 28.65 38.37 32.78 29.18 34.53 14.91 14.92 13.2
8 15.25 16.52
[141] 15.85 15.45 43.56 25.96 24.31 25.78 37.88 14.04 41.68 29.69 43.54 16.45 21.95 41.74 16.48 15.08 14.16 40.3
7 47.35 29.53
[161] 42.28 15.28 25.15 41.48 27.83 38.85 16.71 14.79 35.35 35.75 18.47 16.89 46.33 41.89 37.33 28.73 23.22 26.8
0 28.05 38.61
[181] 48.54 14.18 14.41 17.54 44.85 19.63 22.05 28.98 37.37 28.84 22.87 48.72 46.73 40.24
```

```
WHO$Country[which.min(WHO$Under15)]
[1] Japan
194 Levels: Afghanistan Albania Algeria Andorra Angola Antigua and Barbuda Argentina Armenia Australia Austria
... Zimbabwe
```

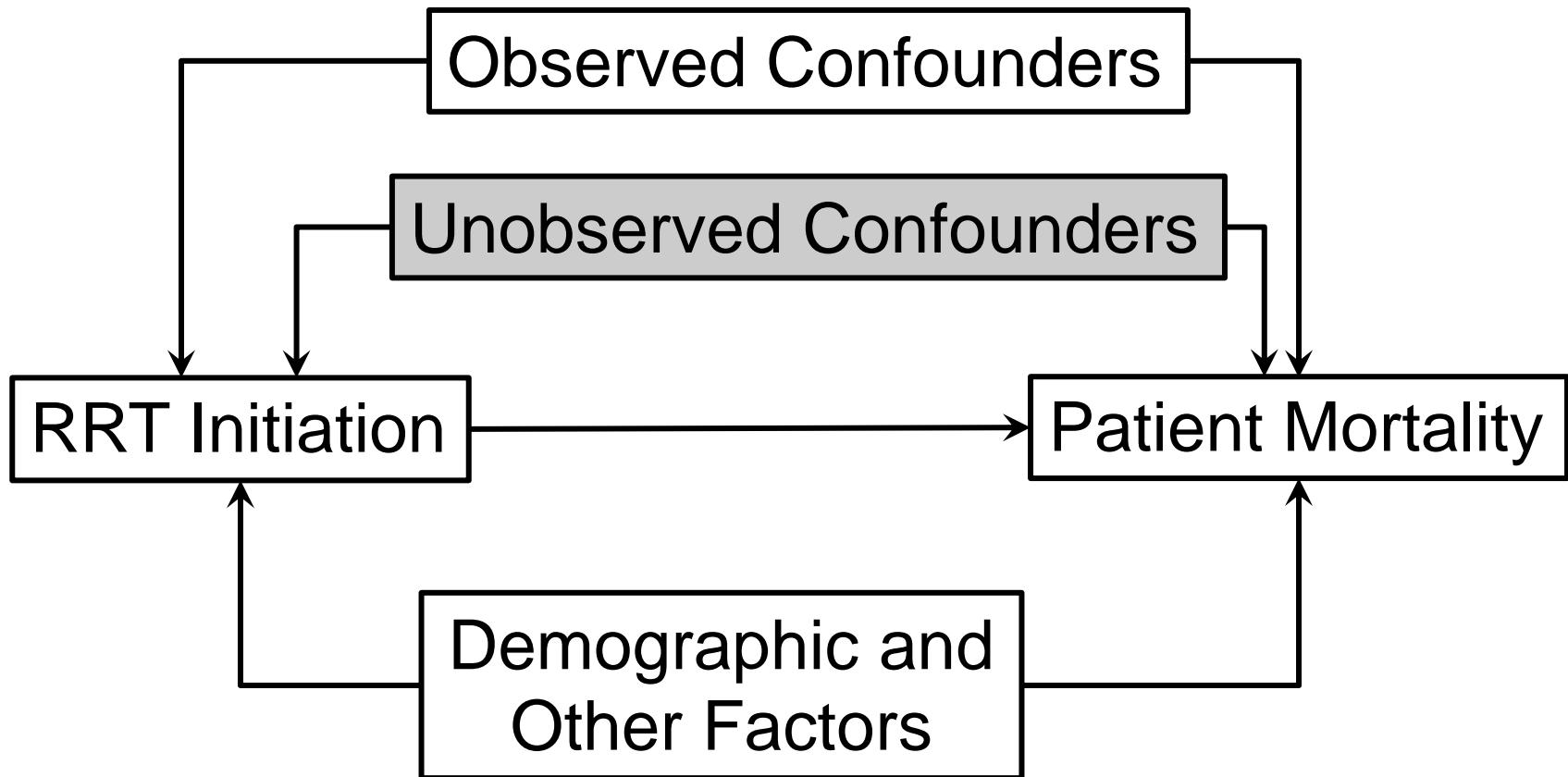
Let's create some plots for exploratory data analysis (EDA). First, let's create a basic scatterplot of GNI versus FertilityRate.



Limitations of PSM

- Matching based on observed variables
- Cannot account for unobserved variables
 - Patient health conditions can only be partially captured by some measurable variables, like AKIN classification, SOFA score, etc.
 - Physicians' treatment choice can depend on patient conditions that are not observed in the data set

Limitations of PSM



Bivariate Probit Model

Different effects!

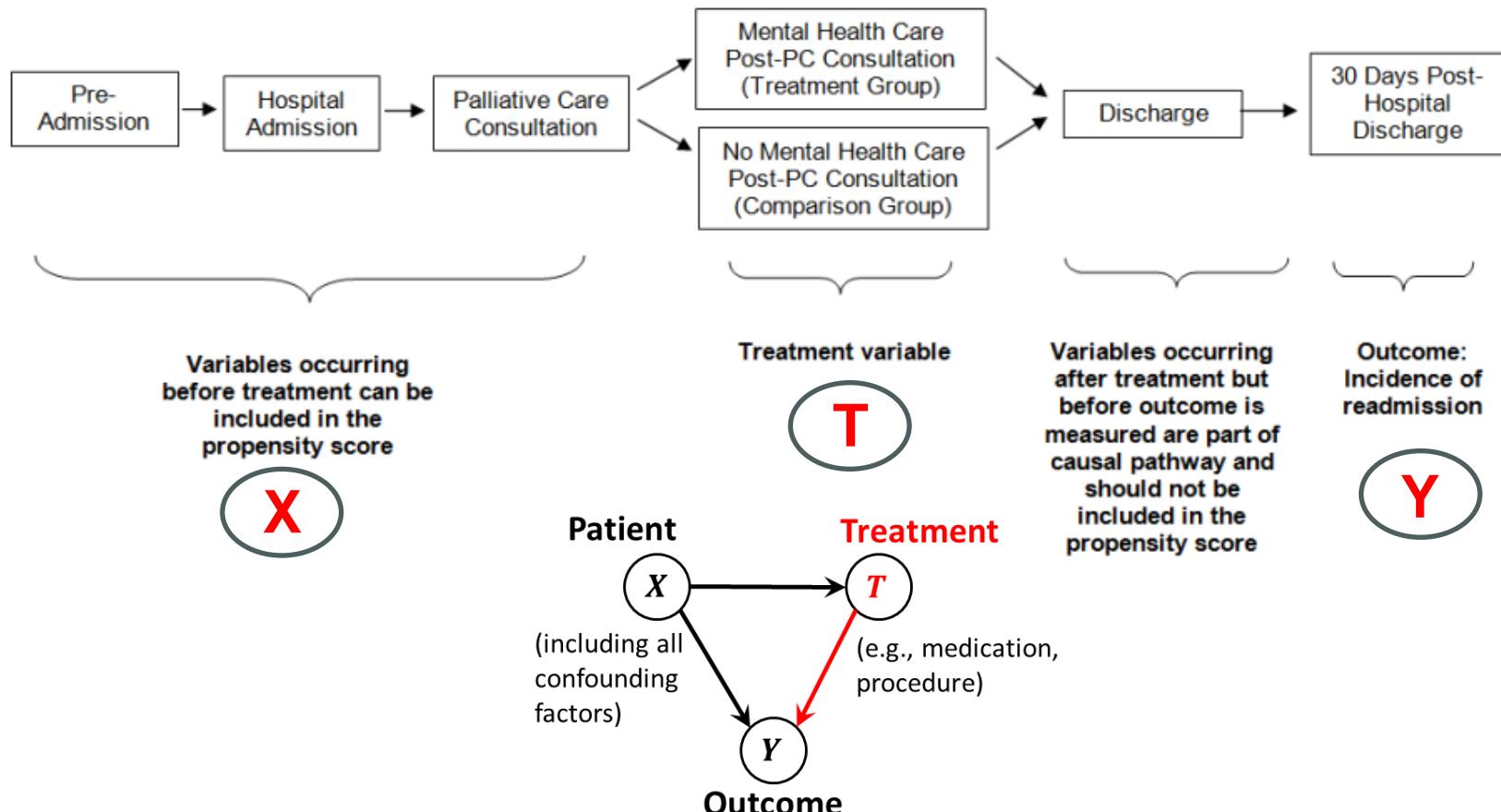
	Marg. Effect	95% CI	p-value
Overall			
Univariate probit Model	0.16	[0.13, 0.19]	0.000
Bivariate probit Model	-0.03	[-0.06, 0.00]	0.066
Subsample			
BaseCr<300	-0.02	[-0.05, 0.01]	0.184
BaseCr \geq 300	-0.05	[-0.08, -0.01]	0.006

CI, Confidence Interval; BaseCr, Baseline Creatinine;

The creatinine cutoff $300 \mu\text{mol/L}$ was chosen in parallel to the assessment of renal system in SOFA score.

- By-product: a statistical test confirms that some unobserved factors increase the likelihood of RRT initiation and mortality simultaneously

Choosing Observables (Confounders)



More Advanced Techniques to Establish Causality

- Instrumental variables
- Regression discontinuity design (RDD)
- Difference in differences (DID)
- Causal tree/forest
- Near-far matching
- Etc.

Class Outline

- Lecture 5 Recap
- Correlation vs Causal Inference
- Causal Inference Framework
- Case Study: Evaluating the Effect of Renal Replacement Therapy Using Propensity Scores

End

ECON 145

ECON 145 – Introductory Data Analytics in Healthcare

Lecture 7: Patient Segmentation

Class Outline

- Assignment 2 Brief
- Review: Diagnostic Analytics
- Patient Segmentation
 - Supervised vs Unsupervised Learning
 - Case Study: Segmentation of Heart Patients using Clustering

Assignment 2 Brief

- Assignment 2 has two parts:
- Part 1: Structured Questions (20 marks)
- Part 2: Healthcare Analytics Case Study (30 marks)

Healthcare Analytics Case Study

SMU-20-0026 :

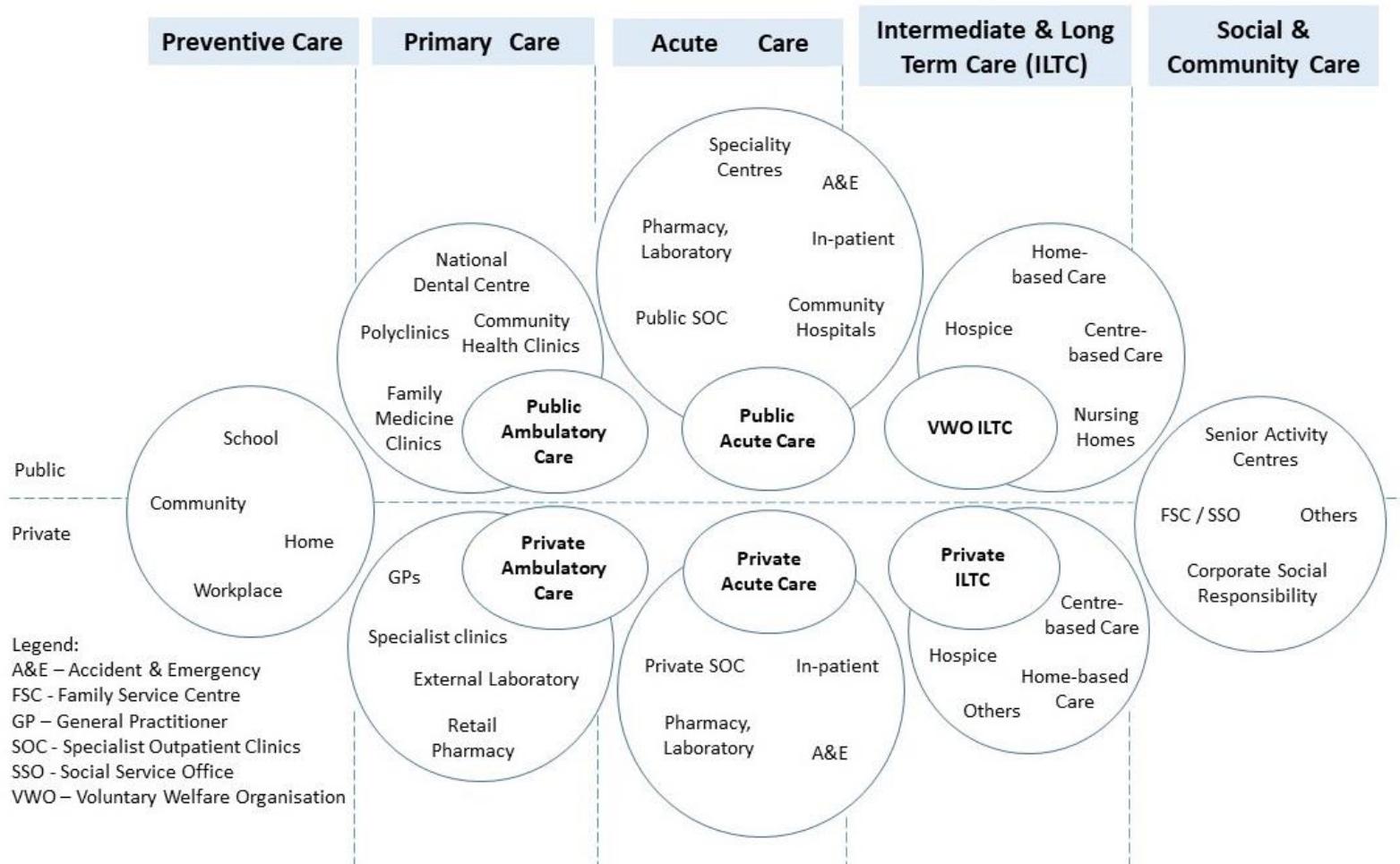
THE HEALTHCARE ANALYTICS LANDSCAPE IN SINGAPORE: EVOLVING TO DELIVER BETTER CARE

LEARNING OUTCOMES:

1. Broad awareness of the healthcare landscape and stakeholders in Singapore
2. Appreciate the advantages and disadvantages of centralised and decentralised management of data in Singapore's healthcare landscape
3. Understand the role of open data to facilitate co-creation of healthcare solutions between the government and the public
4. Gain an overview of the real-world applications using data analytics in healthcare settings
5. Appreciate the trends in smart healthcare and the role of technology in healthcare in an urban context

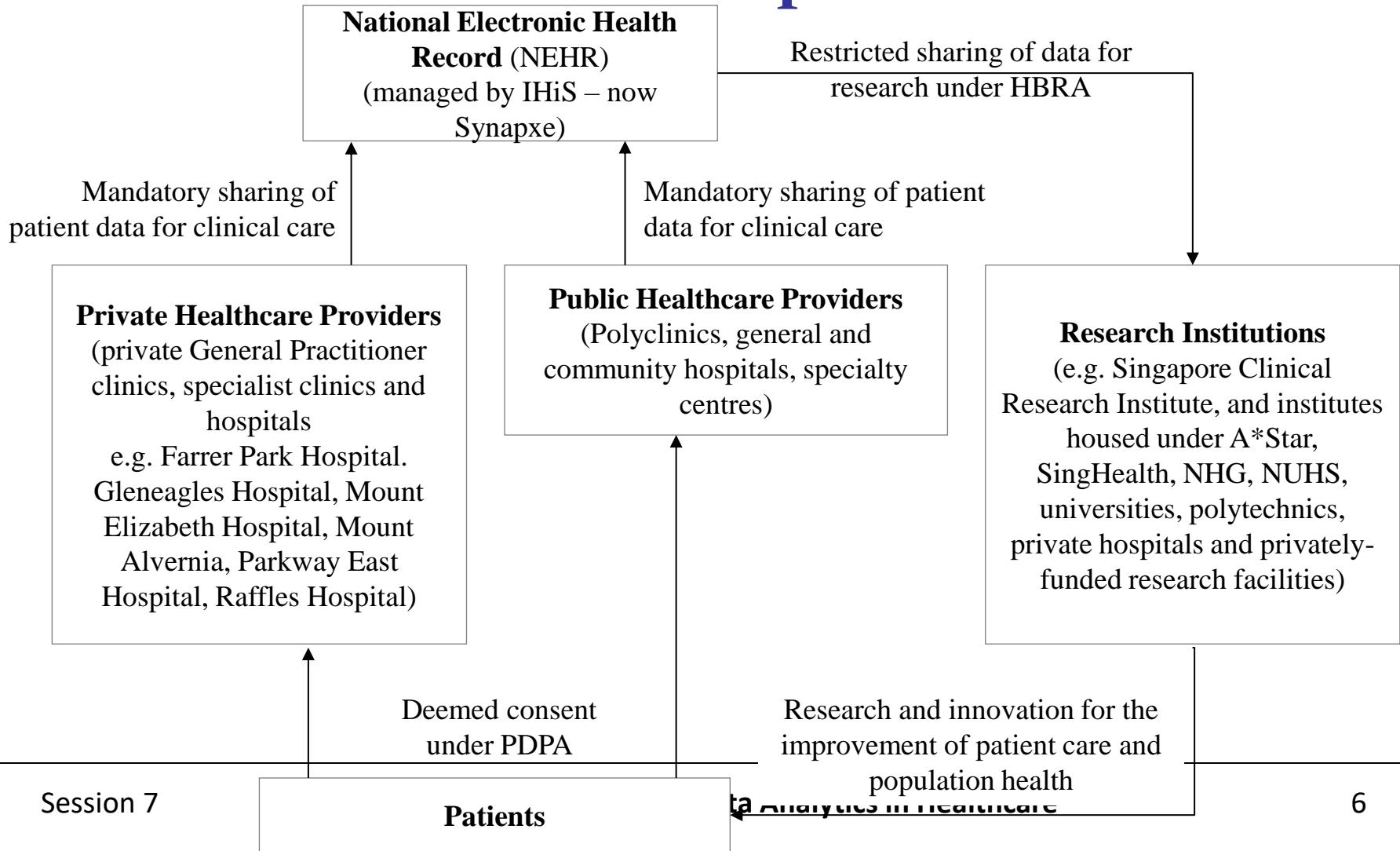
Patient Segmentation

Singapore Healthcare Ecosystem



Patient Segmentation

Singapore Healthcare Analytics Landscape



Case Questions

Part 1: Healthcare in Singapore

- a. How has healthcare in Singapore progress over the years (from the 1960s to 2010s)?
- b. How did Singapore ensure that “Singaporeans always have access to good quality and affordable health care” through its Healthcare 2020 Masterplan?

Part 2: Healthcare Analytics in Singapore

- a. Looking at Singapore’s healthcare ecosystem, do you think healthcare data is centralised or decentralised? In what ways do you think the data is (i) centralised and (ii) decentralised?
- b. What may be some of the arguments supporting centralised (e.g., NEHR, data in eHints, etc.) and decentralised data management?
- c. “Open data facilitates evidence-based healthcare and resident co-creation of solutions especially in the urbanisation context”. Do you agree or disagree with this statement?
- d. With the advent of Artificial Intelligence (AI), and the growing awareness of AI in healthcare, do you agree that open data could play a role in achieving effective data-driven healthcare solutions?
- e. How will trends in smart healthcare and the role of data analytics impact healthcare in the Singapore context?

Marking Rubric for Case Study

TOTAL: 60 Marks



Content Analysis (20)

- Does the student identify key issues and discuss them critically?
- Explore beyond content discussed in case study
- Insightful connections made in context



Critical Reflection and Content Synthesis (20)

- Well developed insights, depth in perceptions and understanding of key challenges/ opportunities
- Challenges or opportunities deep and well-considered?
- Synthesizes current experience to future implications?



Writing Quality (20)

- Include relevant references where necessary
- Good organisation and clear flow of arguments
- Are there any grammatical errors or typos, etc?

Report Guidelines:

Maximum 2000 words
(Exclude References
and all Figures and
Tables)

Guidelines

- Address all the questions listed for the case study
- Scrutinize your **broader experience and public information** to build up your arguments
- Write clearly and concisely to put forward your arguments. Include any figures and tables that can help to articulate your arguments.
- State your arguments clearly, explain with examples and provide the reference where necessary to reinforce your points
- Give **tangible examples** in all your arguments; **list references** where necessary
- No right or wrong answers!**

Pedagogical Knowledge

Content Reflection

Sense-making

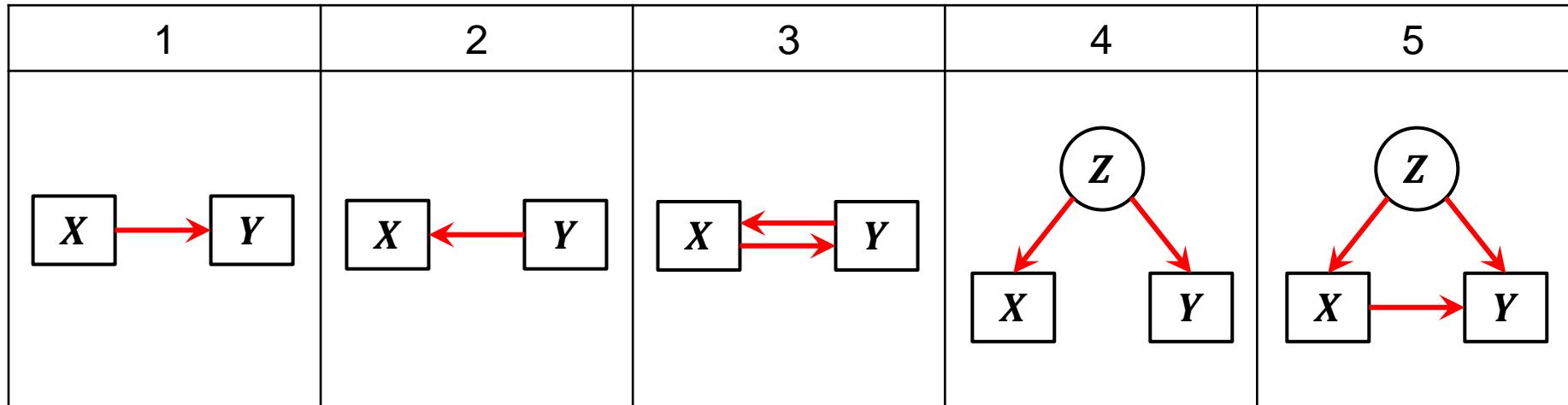
Futures-planning

Class Outline

- Review: Diagnostic Analytics
- Patient Segmentation
 - Supervised vs Unsupervised Learning
 - Case Study: Segmentation of Heart Patients using Clustering

Patient Segmentation

Correlation \neq Causality



- In all cases above, X and Y are correlated
- Is X the reason why Y behaves as such?
- To establish 1, we need to check 2 and 3, and we also need to control for possible Z in 4 and 5

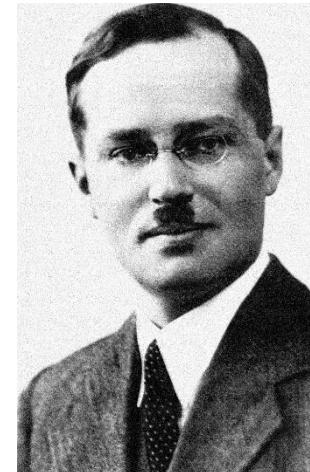
Causal Inference Framework

- Classical Framework
 - Hill's criteria
- More qualitative and subjective evaluation
 - Strength
 - Consistency
 - Specificity
 - Temporality
 - Biological gradient
 - Plausibility
 - Coherence
 - Experiment
 - Analogy
- Potential outcomes framework
 - Neyman–Rubin causal model
 - Rubin causal model
- More quantitative approach
 - Estimating the treatment effect
 - What would have happened to those who, in fact, received treatment (e.g., medication, policy), if they have not received treatment (or vice versa)?

Both are important and widely used.

Potential Outcomes Framework

- Potential outcomes framework
 - Neyman–Rubin causal model
 - Rubin causal model
- We only ever observe one of the two potential outcomes



Jerzy Spława-Neyman
(1894 – 1981)



Donald Bruce Rubin
(1943 –)

Selection Bias in Observational Studies

- Randomized controlled trial (RCT) is not always feasible
 - Unethical, legal issues, economic constraints, etc.
- Have to estimate treatment effect from **observational data**
- Selection bias arises when the treatment assignment decisions depend on observed variables or factors that are not observed but also affect the outcome



Had Zelensky become a President? [Observed]



Had Zelensky stay as an actor? [Counterfactual]

Patient Segmentation

Challenge

- E.g., estimating the reduction in medical errors of hospitals adopting electronic medical record (EMR) systems

	Error Rate With EMR	Error Rate Without EMR	Reduction in Error Rate
<i>Both actual and potential outcomes are known</i>			
Hospital A (with EMR)	1%	3%	2%
Hospital B (without EMR)	8%	10%	2%
<i>Only actual outcomes are known</i>			
Hospital A (with EMR)	1%	—	?
Hospital B (without EMR)	—	10%	?

- We only ever observe one of the two outcomes

Propensity Score Matching (PSM)

- Matching: Match treated and untreated observations on all the observed variables
 - Sometimes can be very challenging
- PSM: Match treated and untreated observations on the estimated probability of being treated (propensity score)
 - Reduce to one dimension, match only a number
 - Need some technical assumptions

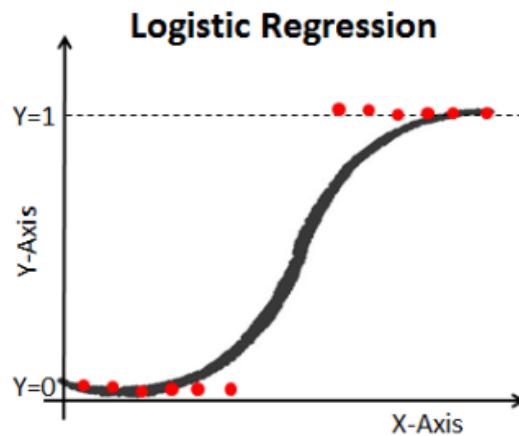
Hands-on Propensity Score Matching

Class Outline

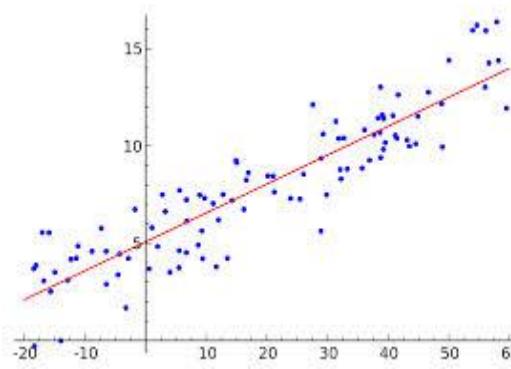
- Patient Segmentation
 - Supervised vs Unsupervised Learning
 - Machine Learning Workflow
 - Case Study: Segmentation of Heart Patients using Clustering

Patient Segmentation

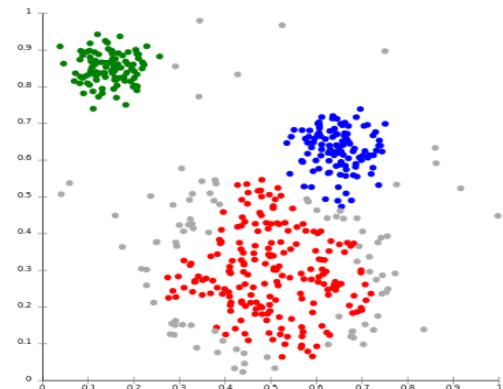
Supervised vs Unsupervised Learning



Classification
Logistics Regression



Linear Regression



Clustering

Supervised Learning

Unsupervised Learning

Patient Segmentation

Supervised vs Unsupervised Learning

Supervised Learning

	X					Y
	SBP	temperature	Outcome
Pt 1						1
Pt 2						0
Pt 3						0
...						1
Pt N						0

Unsupervised Learning

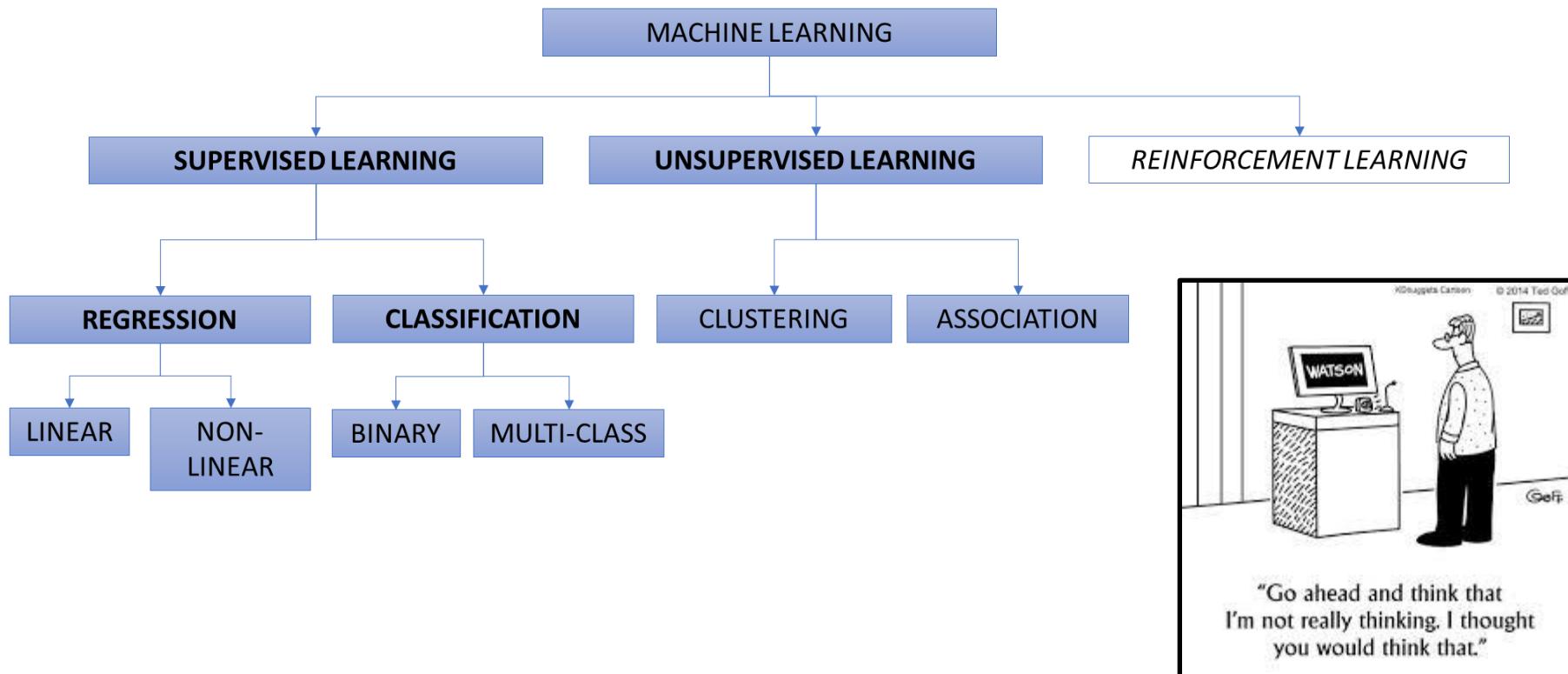
	X						?
	SBP	temperature	Outcome
Pt 1							?
Pt 2							?
Pt 3							?
...							?
Pt N							?

Supervised vs Unsupervised

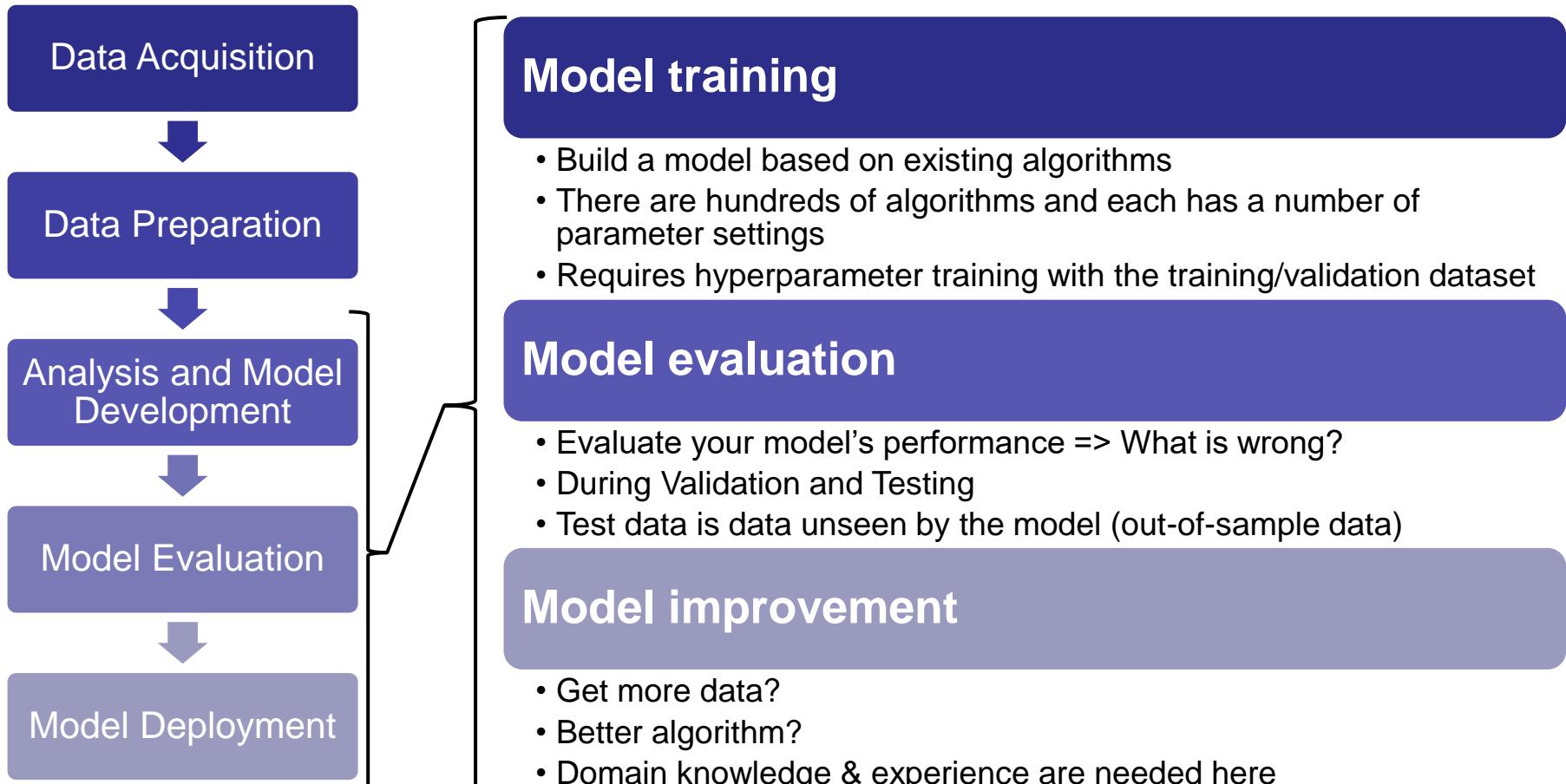
- **Supervised Learning** -- where a **training set of actual target/outcome is available** e.g., Accept / Reject, Group 1 to Group N, numerical values
- The algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples.
- E.g., in regression – data on the target must be available
- **Unsupervised Learning** -- grouping data into categories based on some measure of inherent similarity to **understand pattern without specifying purpose or target**
- Clustering is one of the approaches to unsupervised learning
- No guarantee that clusters formed will be useful

Patient Segmentation

Machine Learning

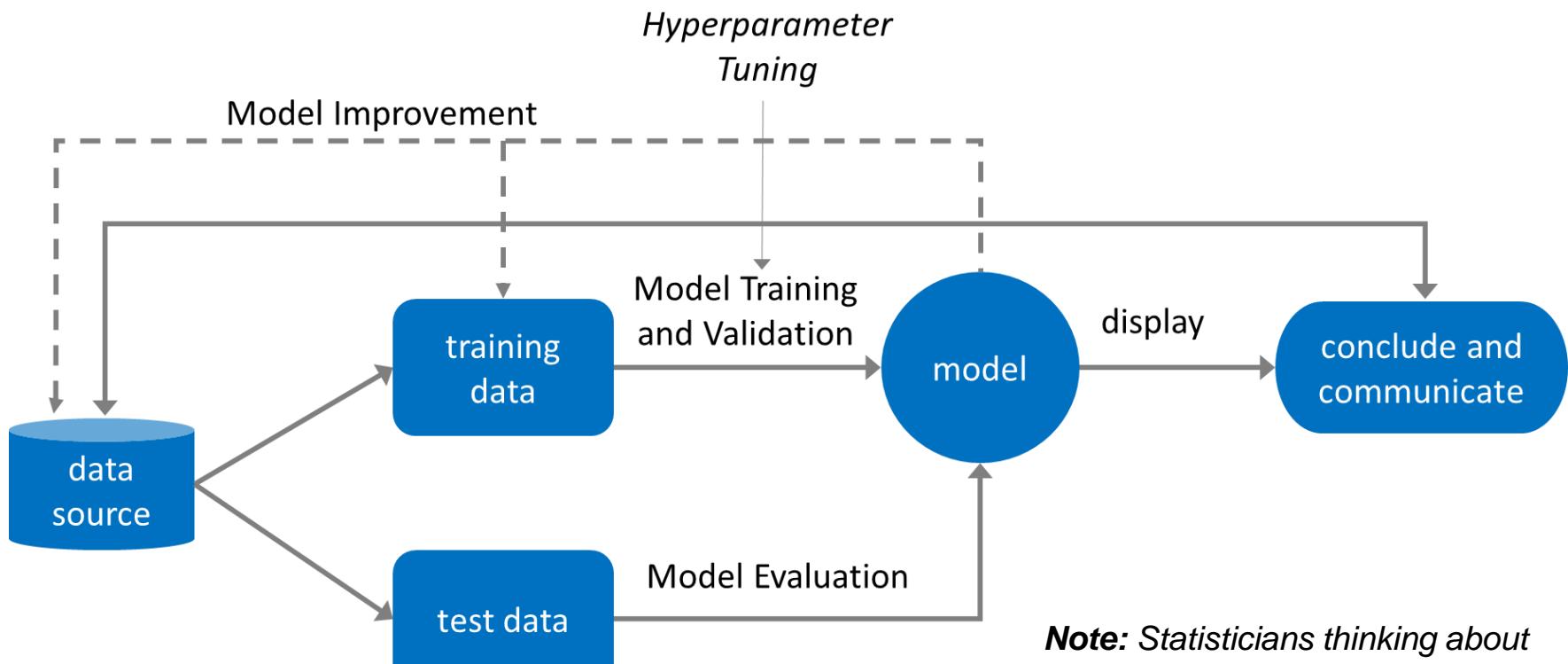


Generic ML Workflow



Patient Segmentation

Generic ML Workflow



Note: Statisticians thinking about Training/Validation/Test may be different from Machine Learning folks.

Simple Class Experiment

- Task: Split the people in the picture into 3 groups. How would you do it?

Note: You can use multiple attributes.

Image credit: IBM



Clustering

- Clustering is the process of examining a collection of “points,” and grouping the points into “clusters” by their natural characteristics.
- Similarity measure is based on some distance measure
- The goal is that
 - points in the **same cluster** have a **small distance** from one another,
 - while points in **different clusters** are at a **large distance** from one another.

Patient Segmentation

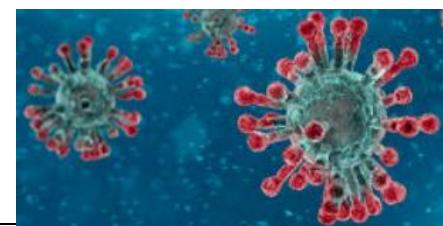
Ex. 1: Segmentation of Symptoms

Covid-19: Study reveals six clusters of symptoms that could be used as a clinical prediction tool

An analysis of data obtained from a symptom tracker app has shown that there are six distinct “types” of covid-19, each distinguished by a cluster of symptoms. Train dataset: 1653 users, tested on 1000 users.

Clusters

1. “Flu-like” with no fever—headache, loss of smell, muscle pains, cough, sore throat, chest pain, no fever
2. “Flu-like” with fever—headache, loss of smell, cough, sore throat, hoarseness, fever, loss of appetite
3. Gastrointestinal—headache, loss of smell, loss of appetite, diarrhoea, sore throat, chest pain, no cough
4. Severe level one, fatigue—headache, loss of smell, cough, fever, hoarseness, chest pain, fatigue
5. Severe level two, confusion—headache, loss of smell, loss of appetite, cough, fever, hoarseness, sore throat, chest pain, fatigue, confusion, muscle pain
6. Severe level three, abdominal and respiratory—headache, loss of smell, loss of appetite, cough, fever, hoarseness, sore throat, chest pain, fatigue, confusion, muscle pain, shortness of breath, diarrhoea, abdominal pain.



Reference: <https://www.bmjjournals.org/content/370/bmjm2911>

Patient Segmentation

Ex. 2: Segmentation of Medical Cost

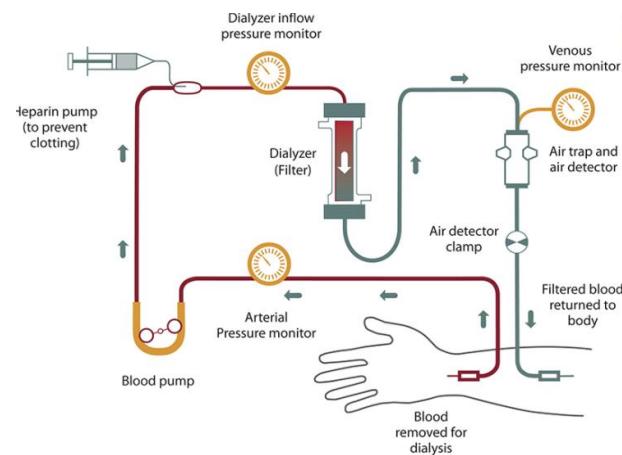
End-stage Renal Disease (ESRD) Cost Analysis: Identify cost change patterns of patients with ESRD

Purpose: To provide useful information to healthcare decision-makers in relation to the cost burden of hemodialysis (HD) therapy by the characteristics of the patients receiving HD.

Demographic and clinical characteristics of patients grouped into 4 proposed clusters using K-means CA

	Cluster 1: Average to High (n = 113)	Cluster 2: Very High to High (n = 89)	Cluster 3: Average to Average (n = 16,624)	Cluster 4: Increasing Costs, High at Both Points (n = 1554)
Age (y), mean (SD)	57.6 (11.6)	55.5 (14.8)	63.9 (14.0)	56.2 (12.8)
Age (y), n (%)				
18-24	0 (0.0)	4 (4.5)	121 (0.7)	33 (2.1)
25-34	2 (1.8)	7 (7.9)	355 (2.1)	54 (3.5)
35-44	15 (13.3)	6 (6.7)	1026 (6.2)	156 (10.0)
45-54	24 (21.2)	19 (21.3)	2401 (14.4)	375 (24.1)
55-64	50 (44.2)	33 (37.1)	4652 (28.0)	609 (39.2)
65+	22 (19.5)	20 (22.5)	8069 (48.5)	327 (21.0)
Sex, n (%)				
Male	66 (58.4)	48 (53.9)	9599 (57.7)	924 (59.5)
Female	47 (41.6)	41 (46.1)	7025 (42.3)	630 (40.5)
Region in the United States, n (%)				
Northeast	12 (10.6)	12 (13.5)	1843 (11.1)	192 (12.4)
North central	32 (28.3)	18 (20.2)	6084 (36.6)	444 (28.6)
South	38 (33.6)	39 (43.8)	6354 (38.2)	625 (40.2)
West	30 (26.5)	19 (21.3)	2235 (13.4)	286 (18.4)
Unknown	1 (0.9)	1 (1.1)	108 (0.6)	7 (0.5)
Health insurance type, n (%)				
FFS	87 (77.0)	71 (79.8)	13,967 (84.0)	1237 (79.6)
HMO and POS capitation	20 (17.7)	17 (19.1)	2304 (13.9)	270 (17.4)
Missing	6 (5.3)	1 (1.1)	353 (2.1)	4 (3.0)
Comorbidity Score Indices ^a	Pre-HD Period	Post-HD Period	Pre-HD Period	Post-HD Period
ECI, mean (SD)	6.9 (3.4)	10.8 (3.5)	9.0 (4.1)	9.3 (3.7)
CCI, mean (SD)	5.0 (2.7)	7.1 (2.4)	5.6 (3.2)	6.2 (2.9)
	Pre-HD Period	Post-HD Period	Pre-HD Period	Post-HD Period
	5.7 (2.5)	6.8 (2.8)	6.5 (3.0)	9.5 (3.2)
	5.1 (2.3)	5.0 (2.5)		6.5 (2.7)

Reference: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4776444/>



Reference:

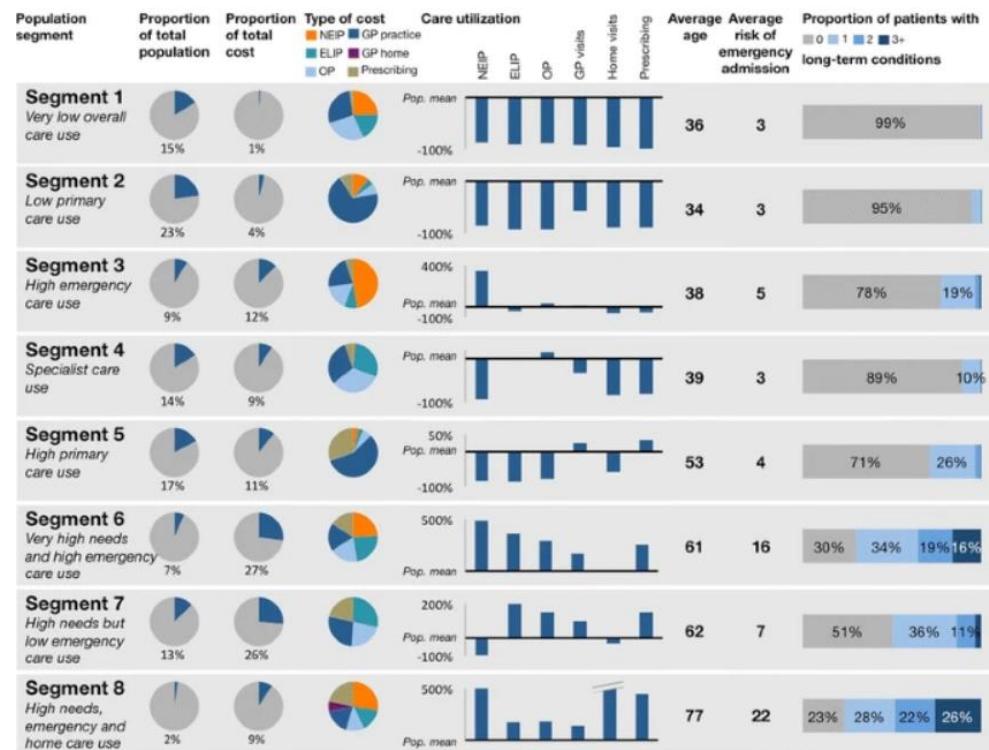
<https://www.niddk.nih.gov/health-information/kidney-disease/kidney-failure/hemodialysis>

Patient Segmentation

Ex. 3: Patient Segmentation

A quantitative evidence base for population health: applying utilization-based cluster analysis to segment a patient population

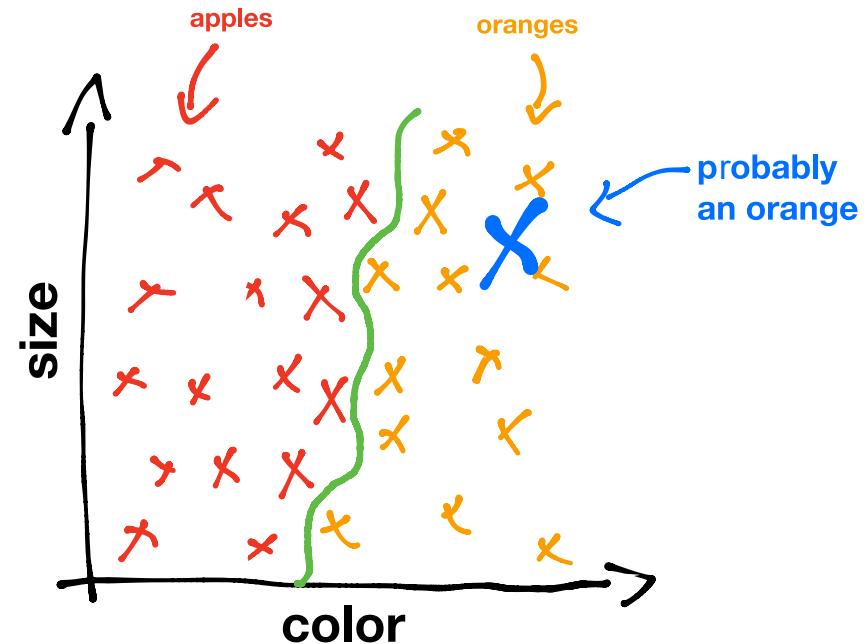
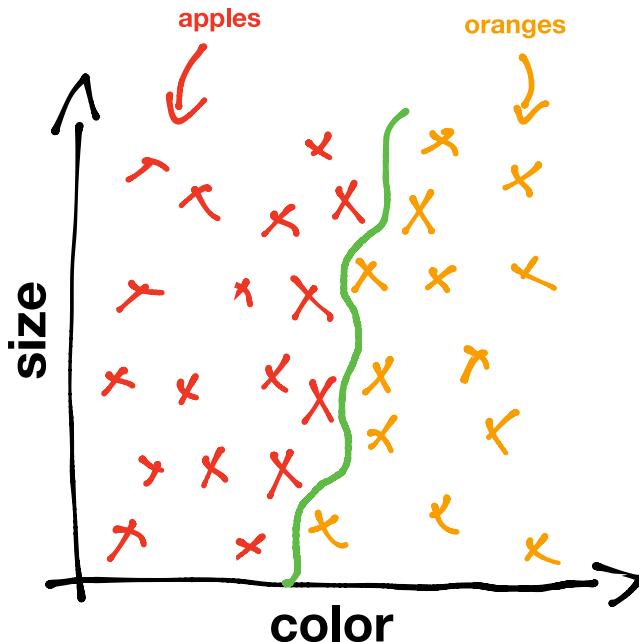
- Utilization-based cluster analysis segments a patient population into distinct groups
- Identify unique care priorities
- Provide quantitative evidence to improve population health
- Segments lower-needs populations → preventive intervention.
- Identify different user types to provide appropriate care needs



Reference: <https://pophealthmetrics.biomedcentral.com/articles/10.1186/s12963-016-0115-z>

Patient Segmentation

Supervised Learning - Classification



Use classification if your data can be tagged, categorized, or separated into specific groups or classes (labels are available)

[Machine Learning Crash Course: Part 1](#), KD Nuggets

Patient Segmentation

Unsupervised Learning - Clustering



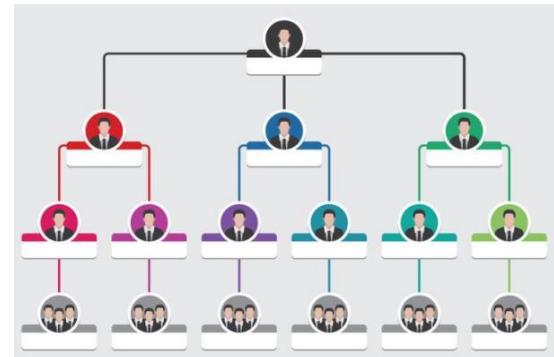
In clustering, labels are inferred from the data

[What Is Machine Learning? 3 things you need to know](#), MathWorks

Clustering Methods

1. Hierarchical:

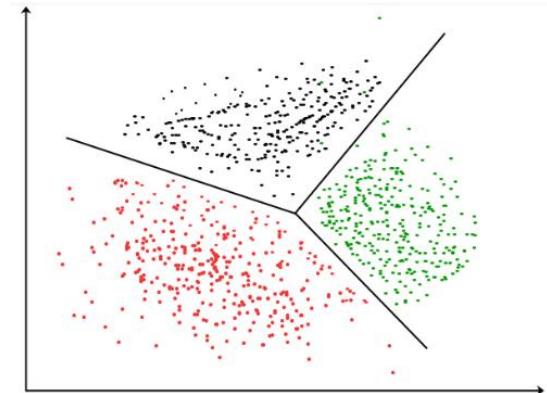
- Initially, each point in cluster by itself.
- Repeatedly combine the two “closest” clusters into one.



Self study: Hierarchical clustering using video
<https://www.youtube.com/watch?v=EUQY3hL38cw>

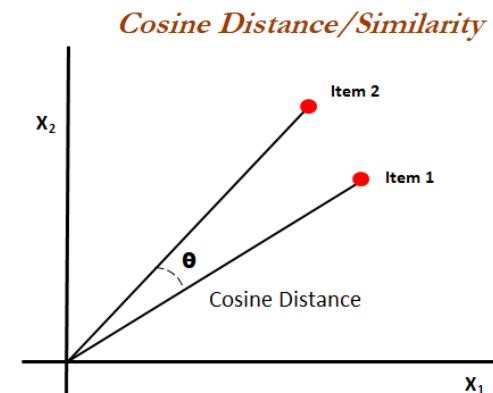
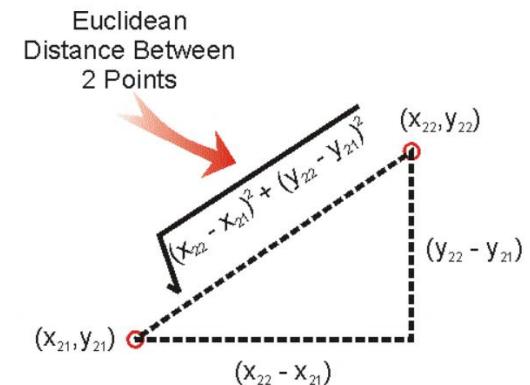
2. Partitioning Method:

- Maintain a set of clusters.
- Place points into “closest” cluster.
- Example: K-Means Clustering



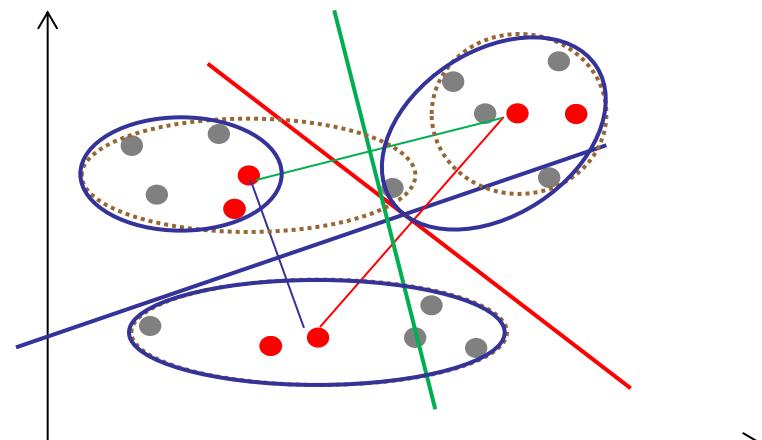
Distance Measures

- Clustering problems are based on some notion of a **distance** between two points.
- Distances can be divided into two classes: E.g.,
 - **Euclidean**
 - You can take an **average** of two points.
 - A Euclidean distance is based on the **locations** of points in space.
 - **Cosine Similarity**



K-Means Clustering

1. Value of **K** is **decided** and **K seeds** are **assigned**
2. Each **observation** (point or row) is **allocated** to the **closest seed** to get K clusters
3. **Compute** the **centroid** of each **cluster** as the new seed
4. **Reassign** each observation to the **new** clusters, based on the new distance from the new seeds
5. Iterate step 2 to 4 till until a **stopping criteria** is met (in practice, five to 25 iterations is likely to arrive at a stable solution)



K-Means Clustering

Possible stopping criteria

1. no data points change clusters (convergence)
2. sum of the square distances to the centroid (a.k.a. total within-cluster sum of square) has no or very small change (convergence)
3. maximum number of iterations is reached

Handling different types of variables

- Discussion:
 - What type of variables (e.g., numerical, categorical) can we compute distances?
 - What if we have categorical variables?

Handling categorical data

- K-Means algorithm is not robust to handle categorical data
 - Categorical variables are discrete and do not have any natural origin.
 - Computing Euclidean distance for such variables is not meaningful.
- Some variations to K-Means for Categorical Data:
 - K-Modes which is suitable for data with categorical features.
 - K-Prototypes is an extension of the k-Modes algorithm that works for mixed categorical and numerical features.
 - Partitioning Around Medoids (PAM) is a popular method.

K-prototypes:

<https://cran.r-project.org/web/packages/clustMixType/clustMixType.pdf>

K-modes:

<https://cran.r-project.org/web/packages/klaR/klaR.pdf>

PAM:

[RPubs - Using K-means and PAM clustering for Customer Segmentation](#)

Rescaling Data

Guidelines:

- Rescaling should be performed for **all** variables
- For highly skewed variables, additional transformations like log or square root can reduce the skewness for the **affected** variables.
- Always check for outliers in the variables
 - *Outliers may result in bias if outliers are considered in the rescaling (consider other scaling algorithms, e.g., Robust Scaler)*
 - *Note: In some cases, we can also choose to remove outliers, then transformation may not be required.*

Other Transformations:

Log, Box-Cox, Square-root, etc. There are other more advanced techniques (e.g., Generalized Estimating Equations)

Evaluation: Total Sum-of-Squares vs Total Within-Cluster Sum-of-Squares

Total Sum-of-Squares TSS: Total variance in the data (without clustering)

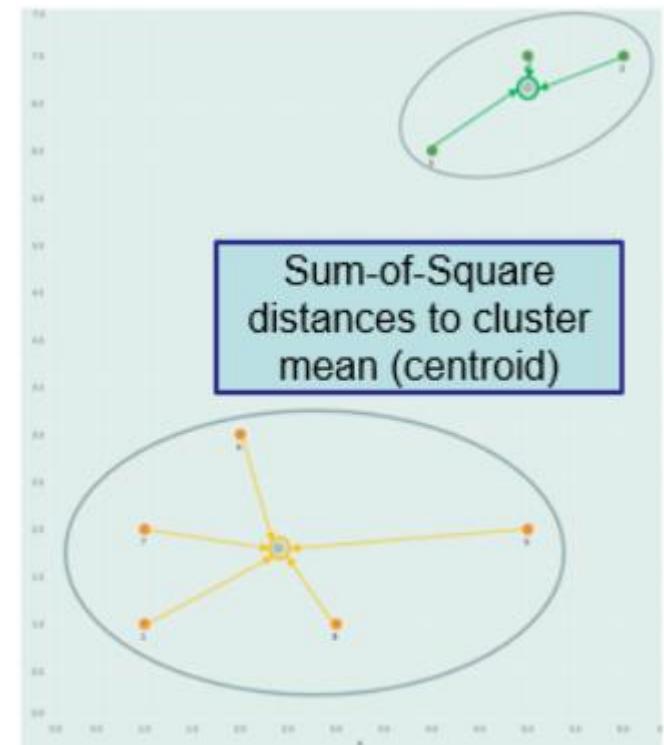
$$TSS = \sum_{i=1}^n (x_i - \bar{x})^2$$

Total Within-Cluster Sum-of-Squares WSS:

Each observation is assigned to one of the k clusters,

1. **Cluster mean (or centroid):** The coordinates representing the mean of the observations within the cluster for each dimension.
2. **Compute Sum of Squared Errors (distances) $d_{i,j}^2$** for each observation i in cluster j , to the **respective cluster centroid**. For K clusters,

$$WSS = \sum_{j=1}^K \sum_{i=1}^{n_j} d_{i,j}^2$$

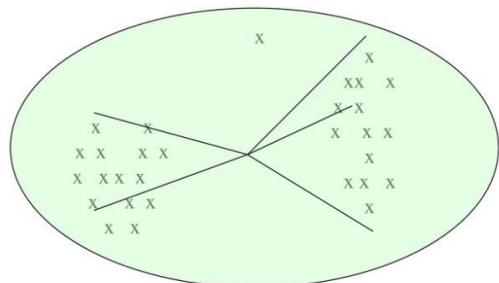


Between-cluster sum of squares and variance explained by model

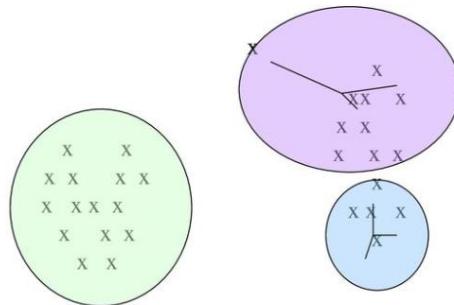
- Another result **between-cluster sum of squares** (i.e., between_SS) is the difference between TSS and WSS
$$\text{between_SS} = \text{TSS} - \text{WSS}$$
- **(Ratio between_SS / TSS) percentage:** A measure of the **total variance in your data set that is explained by the clustering**. k-means algorithm **minimizes the within group dispersion and maximizes the between-group dispersion**. By assigning the samples to k clusters, the model achieves a reduction in sums of squares of by (between_SS/ TSS)%.

Patient Segmentation

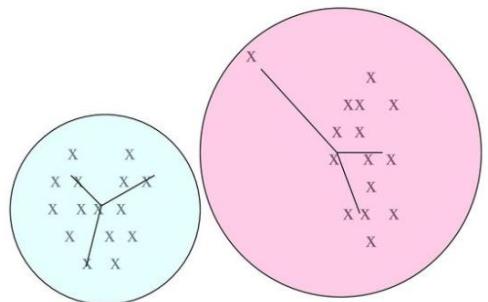
Modeling – Choosing value of K



Too few clusters, long distances to centroid



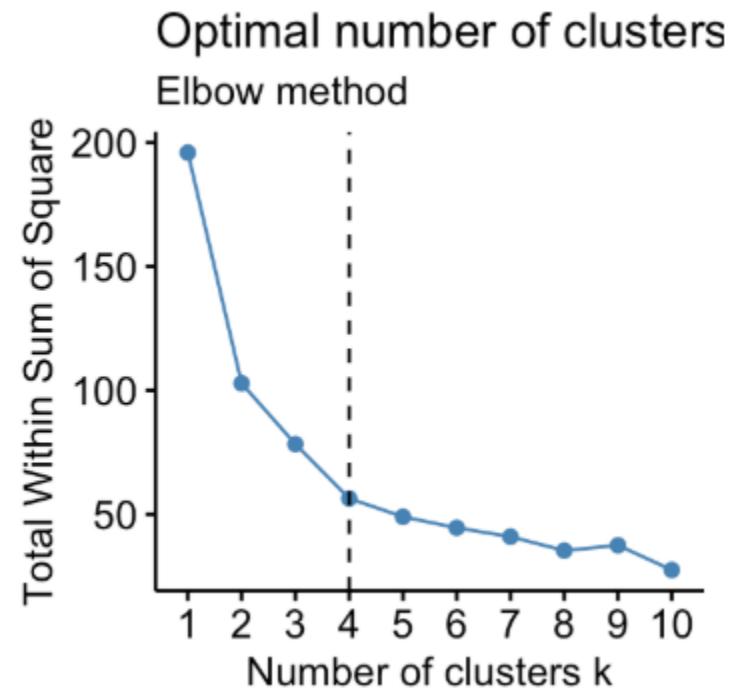
Too many clusters, little improvement in distances to centroid



Reasonable distances to centroid

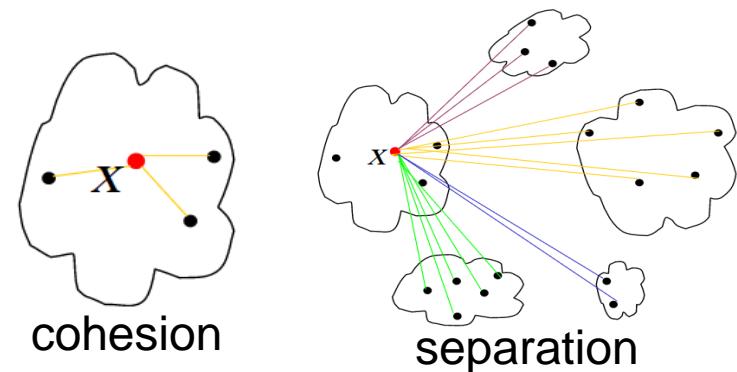
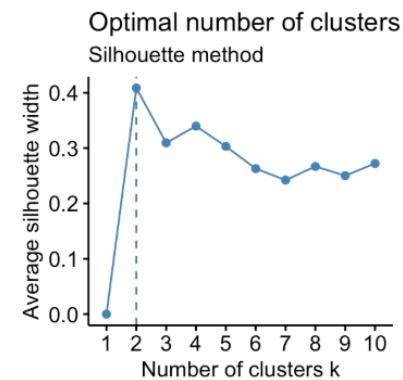
Choosing K using Within-Cluster Sum-of-Squares (Elbow method)

- Try different k , looking at the change in the average distance to centroid as k increases
- Plot k vs Within-Cluster Sum-of-Squares
- Look for an “elbow” where WSS falls rapidly until the changes are small with increasing k .



Choosing K using Silhouette Method

- Silhouette Coefficient combines ideas of both cohesion and separation
- For an individual point, i
 - a = average distance of i to the points in the same cluster
 - b = min (average distance of i to points in another cluster)
 - silhouette coefficient of i :
 $s = 1 - a/b$ if $a < b$
 - Typically between 0 and 1.
 - The closer to 1 the better.



Class Outline

- Review: Treatment Effect Estimation
- Patient Segmentation
- Case Study: Segmentation of Heart Patients using Clustering

Patient Segmentation using Heart Disease Data

- Source: University of California, Irvine (UCI) Machine Learning [[Link](#)]
- The dataset includes healthy subjects and heart disease patients, aged 34-77.
- The entire dataset contains 76 attributes, but all published experiments refer to using a subset of 14 of the parameters.
- The overall dataset contains 4 databases concerning heart disease diagnosis from four locations:
 - Cleveland Clinic Foundation (Cleveland data);
 - Hungarian Institute of Cardiology, Budapest (Hungarian data);
 - V.A. Medical Center, Long Beach, CA (Long Beach VA data),
 - University Hospital, Zurich, Switzerland (Switzerland data)

Patient Segmentation

Data Dictionary (1)

Variable	Type	Description
age	int	age in years
sex	factor	gender (1 = male; 0 = female)
cp	factor	chest pain type -- Value 1: typical angina -- Value 2: atypical angina -- Value 3: non-anginal pain -- Value 4: asymptomatic
trestbps	dbl	resting blood pressure (in mm Hg on admission to the hospital)
chol	dbl	serum cholestorol in mg/dl
fbs	factor	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

Patient Segmentation

Data Dictionary (2)

Variable	Type	Description
restecg	factor	<p>resting electrocardiographic results</p> <ul style="list-style-type: none"> -- Value 0: normal -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
thalach	dbl	maximum heart rate achieved
exang	factor	exercise induced angina (1 = yes; 0 = no)
oldpeak	dbl	ST depression induced by exercise relative to rest

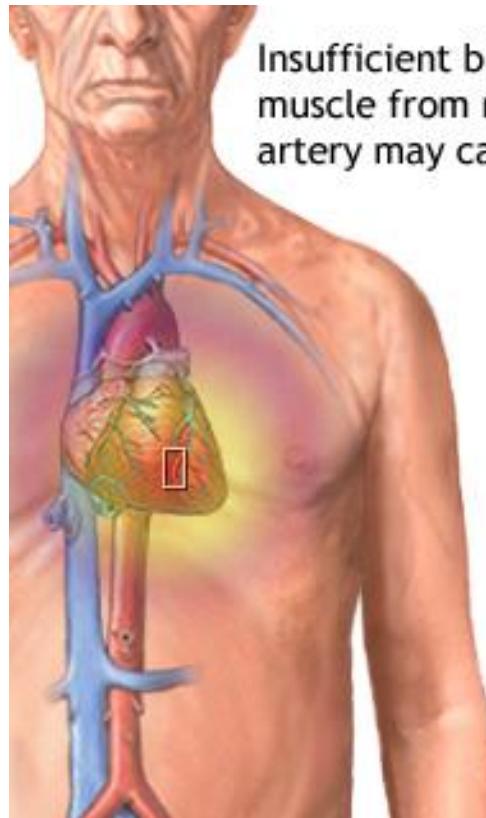
Recall: Coronary Heart Disease (CHD)

- CHD is a disease of the blood vessels supplying the heart
- CHD occurs when the flow of oxygen-rich blood to the heart is blocked or reduced by build-up of fatty material in the coronary arteries.

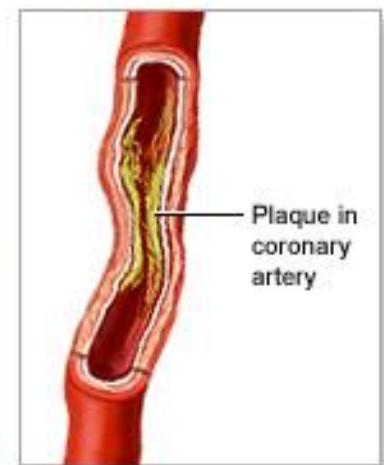
Patient Segmentation

Angina

- As the blood vessels narrowing, blood supply to the heart will be restricted which will cause angina (Chest Pain)
- Angina is a pain or discomfort often felt in your chest area. The most common cause of angina is coronary heart disease
- <https://www.youtube.com/watch?v=k5VjGgk7Wqc>



Insufficient blood flow to the heart muscle from narrowing of coronary artery may cause angina (chest pain)

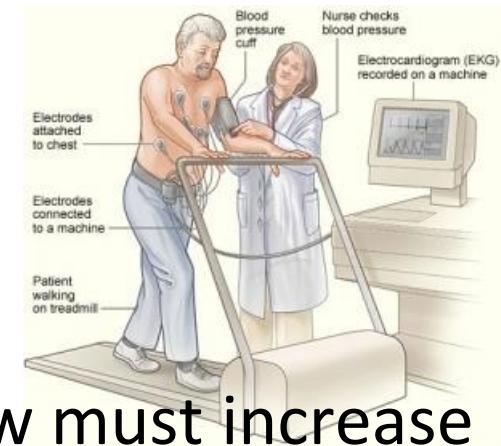


ADAM.

<https://medlineplus.gov/ency/article/000198.htm>

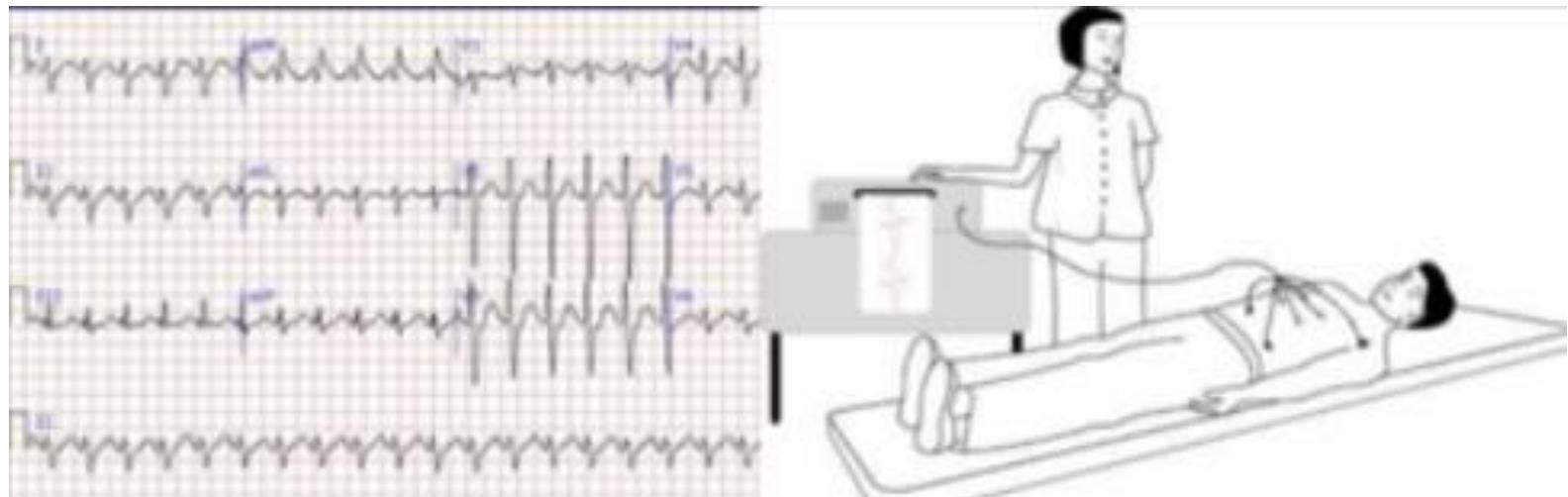
Exercise tolerance testing

- Exercise tolerance testing (also known as stress testing) is used routinely in evaluating patients
 - who present with chest pain
 - have chest pain on exertion,
 - with known heart disease.
- During the exercise, coronary blood flow must increase to meet the higher metabolic demands of the myocardium.
- Limiting the coronary blood flow may result in electrocardiographic changes.



Electrocardiography (ECG)

- Detects heart abnormalities, disease and damage by measuring the heart's rhythms and electric impulses.



Risk Factors of Angina

There are many risk factors, e.g.,

- Fasting Sugar Level / Diabetes
- High LDL cholesterol / Low HDL cholesterol
- Smoking
- Advancing age
- Male gender
- Exercise
- Emotional stress
- Abnormal heart rhythms (your heart beats very quickly or your heart rhythm is not regular)

What are the characteristics of people at risk?



Patient Segmentation

Let's get our hands dirty!

Data Analysis

Let's do some basic data analysis using our WHO data.

```
WHO$Under15
[1] 47.42 21.33 27.42 15.28 47.58 25.96 24.42 28.34 18.95 14.51 22.25 21.62 28.16 30.57 18.99 15.18 16.88 34.4
8 42.95 28.53
[21] 35.23 16.35 33.75 24.56 25.75 13.53 45.66 44.20 31.23 43.08 16.37 38.17 48.87 48.52 21.38 17.95 28.03 42.1
7 42.37 38.61
[41] 23.94 41.48 14.98 16.58 17.16 14.56 21.98 45.11 17.66 33.72 25.96 38.53 38.29 31.25 38.62 38.95 43.18 15.6
9 43.29 28.88
[61] 16.42 18.26 38.49 45.98 17.62 13.17 38.59 14.68 26.96 48.88 42.46 41.55 36.77 35.35 35.72 14.62 28.71 29.4
3 29.27 23.68
[81] 48.51 21.54 27.53 14.04 27.78 13.12 34.13 25.46 42.37 38.18 24.98 38.21 35.61 14.57 21.64 36.75 43.06 29.4
5 15.13 17.46
[101] 42.72 45.64 26.65 29.03 47.14 14.98 38.18 48.22 28.17 29.82 35.81 18.26 27.85 19.81 27.85 45.38 25.28 36.5
9 38.18 35.58
[121] 17.21 20.26 33.37 49.99 44.23 38.61 18.64 24.19 34.31 38.18 28.65 38.37 32.78 29.18 34.53 14.91 14.92 13.2
8 15.25 16.52
[141] 15.85 15.45 43.56 25.96 24.31 25.78 37.88 14.04 41.68 29.69 43.54 16.45 21.95 41.74 16.48 15.08 14.16 40.3
7 47.35 29.53
[161] 42.28 15.28 25.15 41.48 27.83 38.85 16.71 14.79 35.35 35.75 18.47 16.89 46.33 41.89 37.33 28.73 23.22 26.8
0 28.05 38.61
[181] 48.54 14.18 14.41 17.54 44.85 19.63 22.05 28.98 37.37 28.84 22.87 48.72 46.73 40.24
```



```
WHO$Country[which.min(WHO$Under15)]
[1] Japan
194 Levels: Afghanistan Albania Algeria Andorra Angola Antigua and Barbuda Argentina Armenia Australia Austria
... Zimbabwe
```


Let's create some plots for exploratory data analysis (EDA). First, let's create a basic scatterplot of GNI versus FertilityRate.

```
plot(WHO$GNI, WHO$FertilityRate)
```



Patient Segmentation

Dataset: Heart Disease

Var_num	Variable	Role	Type	Description	Missing Values
1	sex	Feature	Categorical	gender	
2	cp	Feature	Categorical	chest pain type -- Value 1: typical angina -- Value 2: atypical angina -- Value 3: non-anginal pain -- Value 4: asymptomatic	
3	fbs	Feature	Categorical	fasting blood sugar > 120 mg/dl	
4	restecg	Feature	Categorical	resting electrocardiographic results -- Value 0: normal -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria	
5	exang	Feature	Categorical	exercise induced angina	no
6	age	Feature	Integer	years	
7	trestbps	Feature	Integer	resting blood pressure (on admission to the hospital)	no
8	chol	Feature	Integer	serum cholesterol	no
9	thalach	Feature	Integer	maximum heart rate achieved	no
10	Oldpeak	Feature	Integer	ST depression induced by exercise relative to rest	no

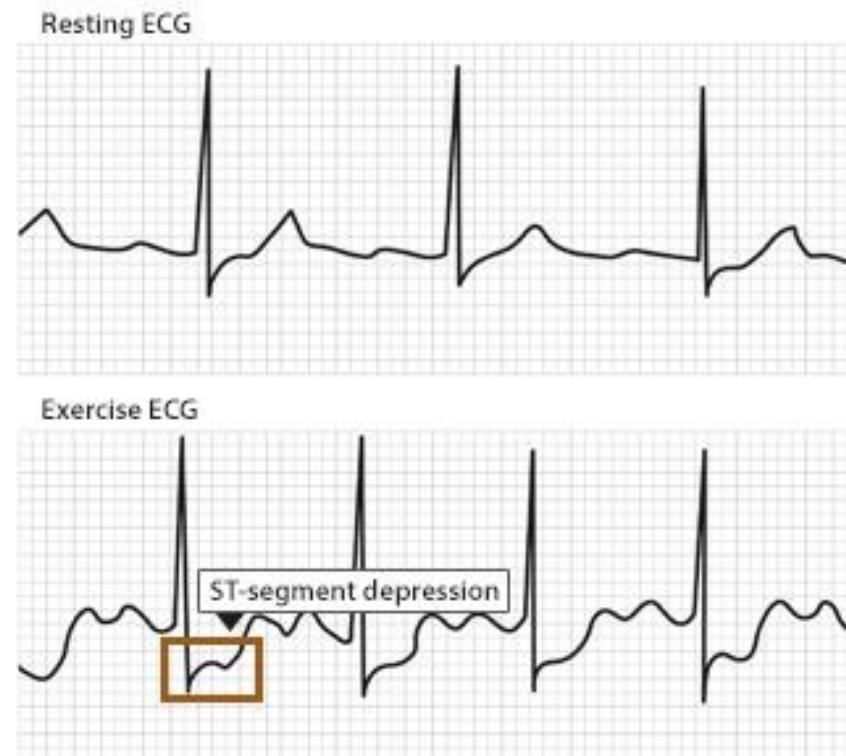
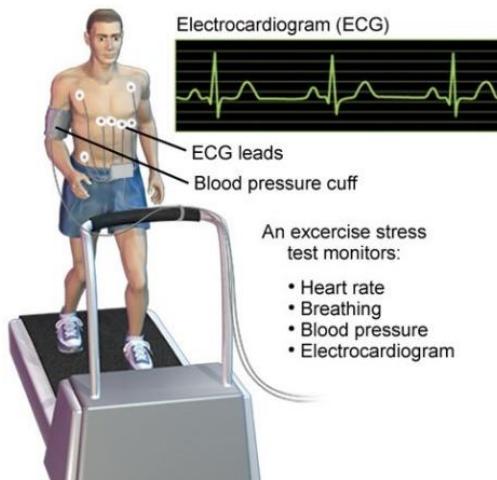
<https://archive.ics.uci.edu/dataset/45/heart+disease>

Note: No outcomes labels!

Domain Knowledge

- ST depression exercise vs rest
- "ST segment depression suggesting "ischemia" or lack of oxygen getting to the heart."*

<https://www.acpjournals.org/doi/10.7326/0003-4819-106-6-793>



Interpreting Clustering Output

K-means clustering with 2 clusters of sizes 172, 130

cluster means:

	age	trestbps	chol	thalach	oldpeak
1	-0.5378132	-0.3386598	-0.2362374	0.5210886	-0.4515563
2	0.7115682	0.4480729	0.3125602	-0.6894403	0.5974437

within cluster sum of squares by cluster:

[1] 527.6549 603.6768
(between_SS / total_SS = 24.8 %)

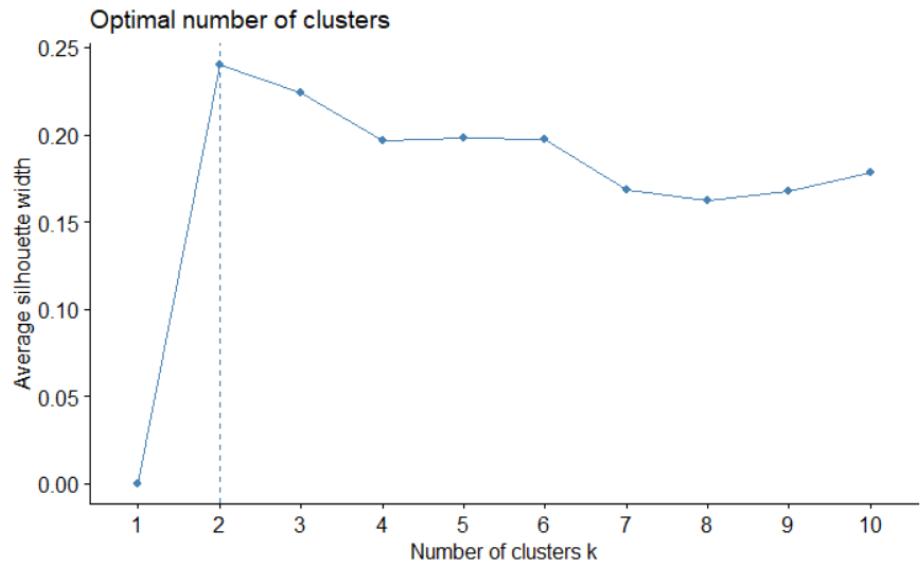
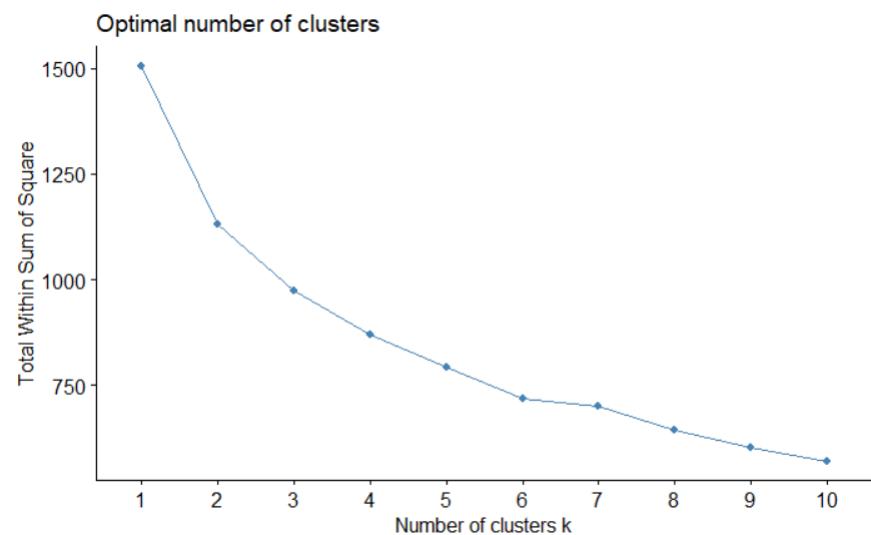
- Are the clusters well distributed?
- Which cluster is more homogeneous?
- How much variance can the model explain?
- What are the profiles of the clusters?

Cluster 1:

- Younger (Age)
- Lower resting blood pressure (on admission to the hospital)
- Lower serum cholesterol
- Higher max heart rate (thalach)
- Lower ST depression exercise vs rest (oldpeak)

Patient Segmentation

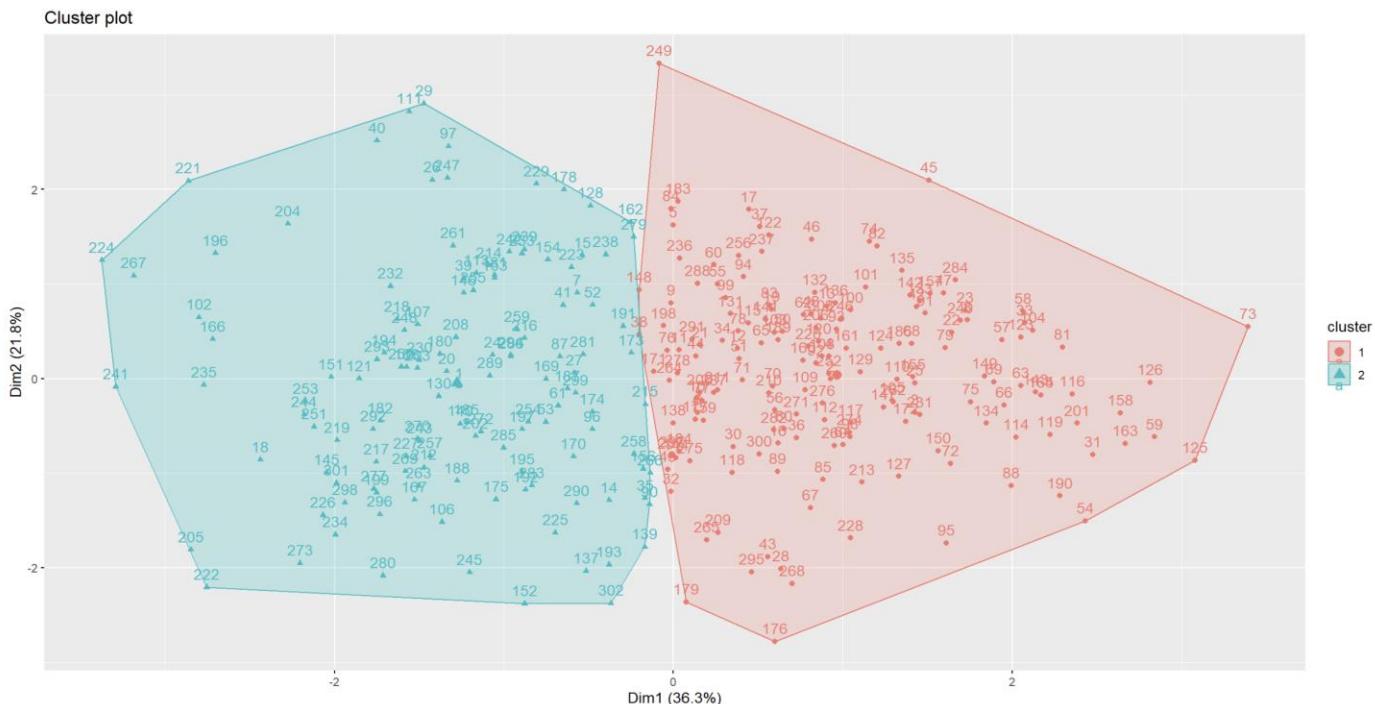
Picking K



Patient Segmentation

Descriptive Insights

	age	trestbps	chol	thalach	oldpeak
1	-0.5378132	-0.3386598	-0.2362374	0.5210886	-0.4515563
2	0.7115682	0.4480729	0.3125602	-0.6894403	0.5974437



Finding the Principal Components

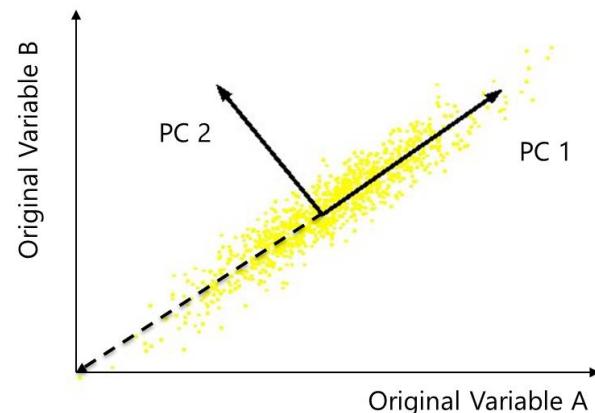
- Allows for better visualization of clusters across 2-3 Principal Components
- Too many observations and factors – how to visualize?
- Reduce to smaller set of factors
- Better representation without losing too much information
- Combination of observed variables may be more effective for insights even if physical meaning is obscured
- **Factors are a combination of Observed Variables in PCA**

Healthcare data are usually high dimensional!

PCA – Key Principles

- Linear projection to reduce the number of parameters
- Map data into a space of lower dimensionality
- Transfer set of correlated variables to a new set of uncorrelated variables (new axes are orthogonal)

1. Projections of PC1 discriminate the most along one axis
2. PC1 – direction of greatest variability in the data
3. After removing the variability along PC1, centre the axis at the centroid of all data points
4. Next PC's are orthogonal to the earlier PCs



Patient Segmentation

PCA Scores

- Variance explained is computed by the **eigendecomposition** of the covariance matrix (eigenvectors and eigenvalues)
- Eigenvectors:** directions of dataset with the most variance (**Principal Components**)
- PCA scores computed by the **projection of the original standardized data onto the eigenvectors** using matrix multiplication

Covariance matrix:

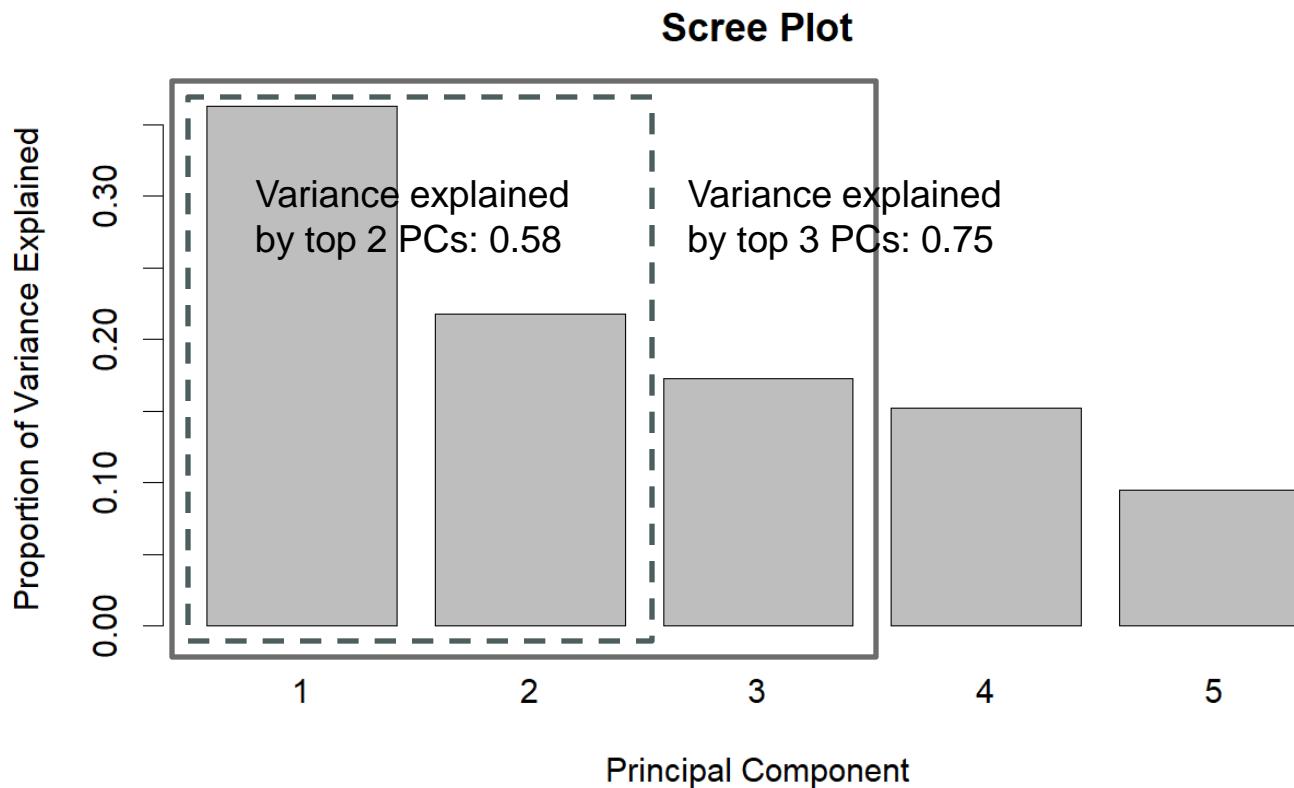
	age	trestbps	chol	thalach	oldpeak
age	1.0000000	0.28507352	0.19876790	-0.40204119	0.20853386
trestbps	0.2850735	1.00000000	0.15253060	-0.04535749	0.19510314
chol	0.1987679	0.15253060	1.00000000	-0.02047239	0.04718279
thalach	-0.4020412	-0.04535749	-0.02047239	1.00000000	-0.34516242
oldpeak	0.2085339	0.19510314	0.04718279	-0.34516242	1.00000000

* Covariances between pairs of variables in a dataset.

	PC1	PC2	PC3	PC4	PC5
1	-1.2662167122	-0.079995175	-0.80049816	-0.006271851	0.69563158
2	0.9268814773	-0.004699616	-1.42995002	-2.748721200	0.59392883
3	1.4254967428	-0.356538322	-1.09019242	-0.768311787	-0.01638062
4	0.9185311739	0.237751166	0.06953513	0.043495979	1.08637380
5	-0.0003562269	1.623320711	1.64834504	-0.734490093	0.37199793
6	0.1446112007	0.350200022	0.66487002	1.105101725	0.071441010

Patient Segmentation

Variance Explained



Patient Segmentation

Cluster Centroid across PC₁ and PC₂

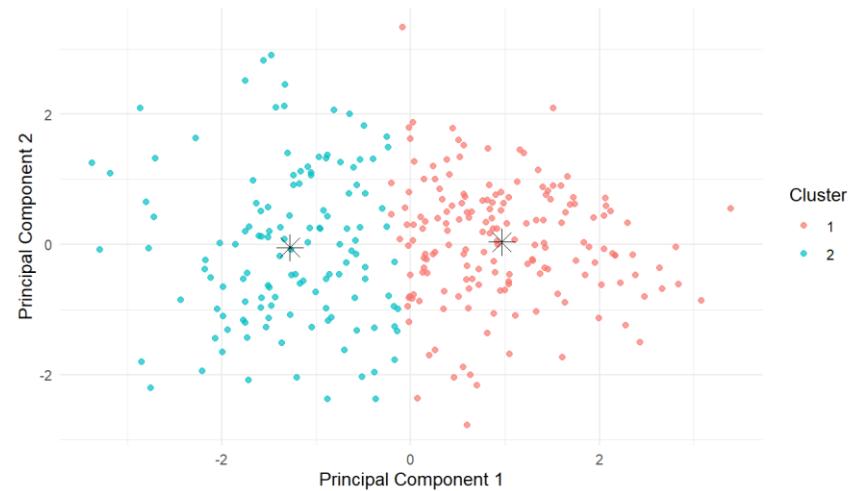
- Loading across the 2 dominant PC:
 - PC1 : age (-), thalach (+), oldpeak (-)
 - PC2 : trestbps (+), chol (+), thalach (+)



- *Plausible* interpretation of the PC:
 - PC1: **Age-related heart function**
(younger, higher heart rate vs. older, higher ST depression).
 - PC2: **Cardiovascular risk factors** (high BP, cholesterol, and heart rate).

Variable	Description
age	years
trestbps	resting blood pressure (on admission to the hospital)
chol	serum cholesterol
thalach	maximum heart rate achieved
oldpeak	ST depression induced by exercise relative to rest

PCA Plot with Cluster Centroids (kmeans)



```
K-means clustering with 2 clusters of sizes 172, 130
```

```
Cluster means:
```

	age	trestbps	chol	thalach	oldpeak
1	-0.5378132	-0.3386598	-0.2362374	0.5210886	-0.4515563
2	0.7115682	0.4480729	0.3125602	-0.6894403	0.5974437

Class Outline

- Review: Treatment Effect Estimation
- Patient Segmentation
- Case Study: Segmentation of Heart Patients using Clustering

End

ECON 145

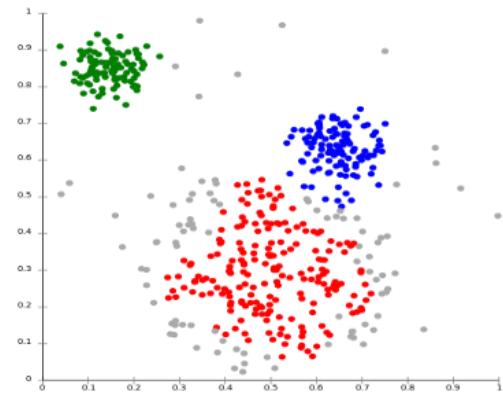
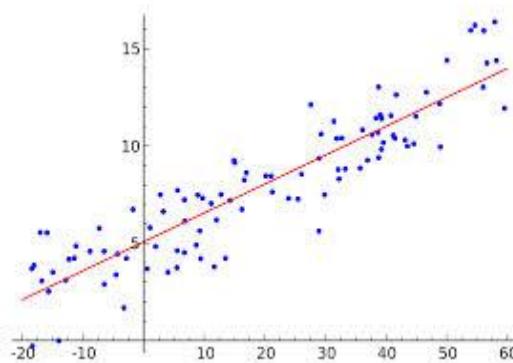
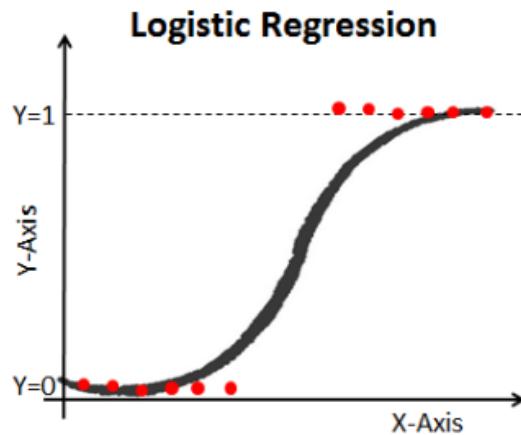
ECON 145 – Introductory Data Analytics in Healthcare

Lecture 8: Quality of Care I

Class Outline

- Review: Patient Segmentation (Clustering)
- Quality of Care
- Classification and Regression Trees
- Case Study: Predicting Quality of Diabetes Care Using Trees and Random Forests

Supervised vs. Unsupervised Learning



Classification
Logistics Regression

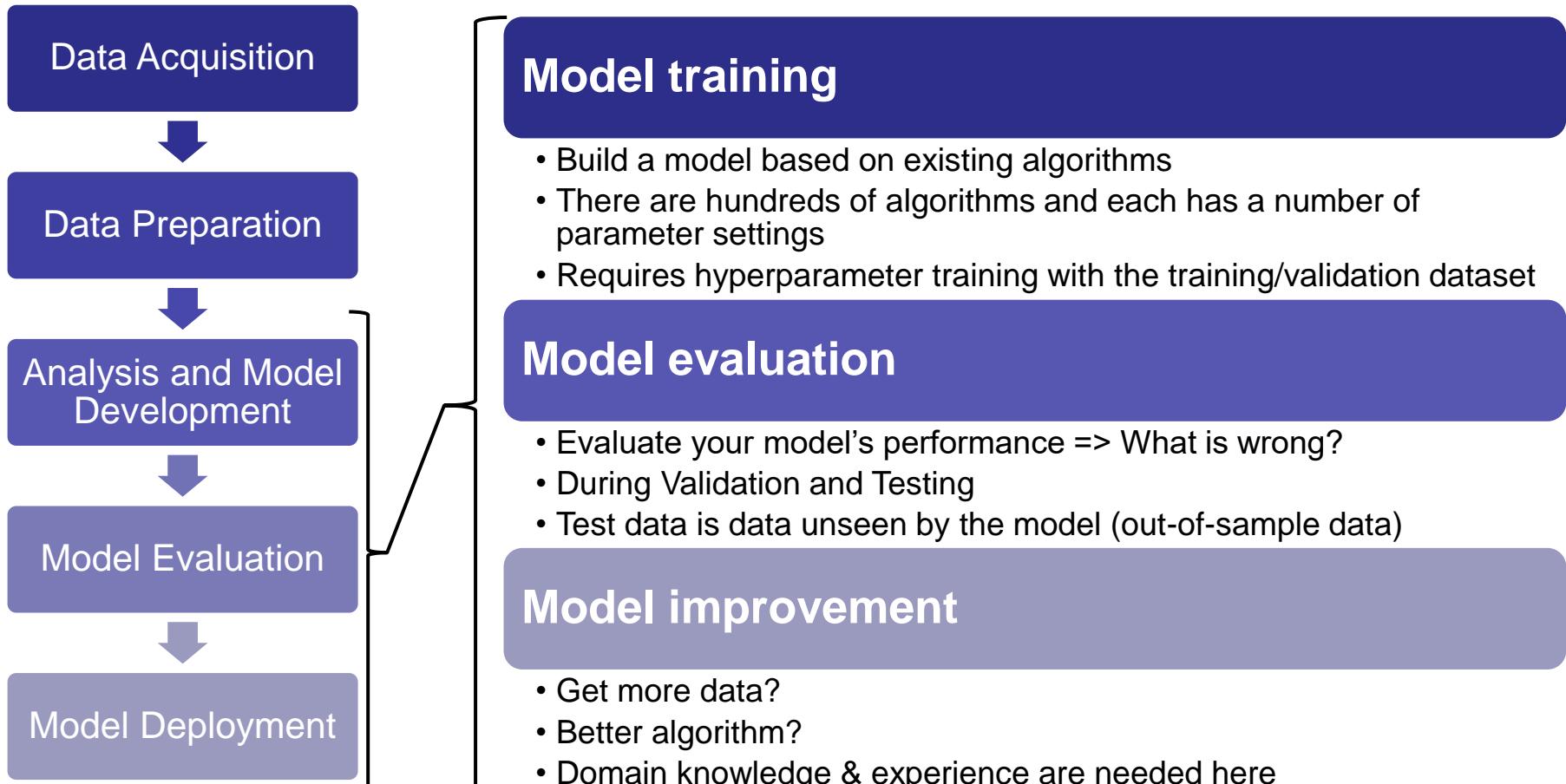
Linear Regression

Clustering

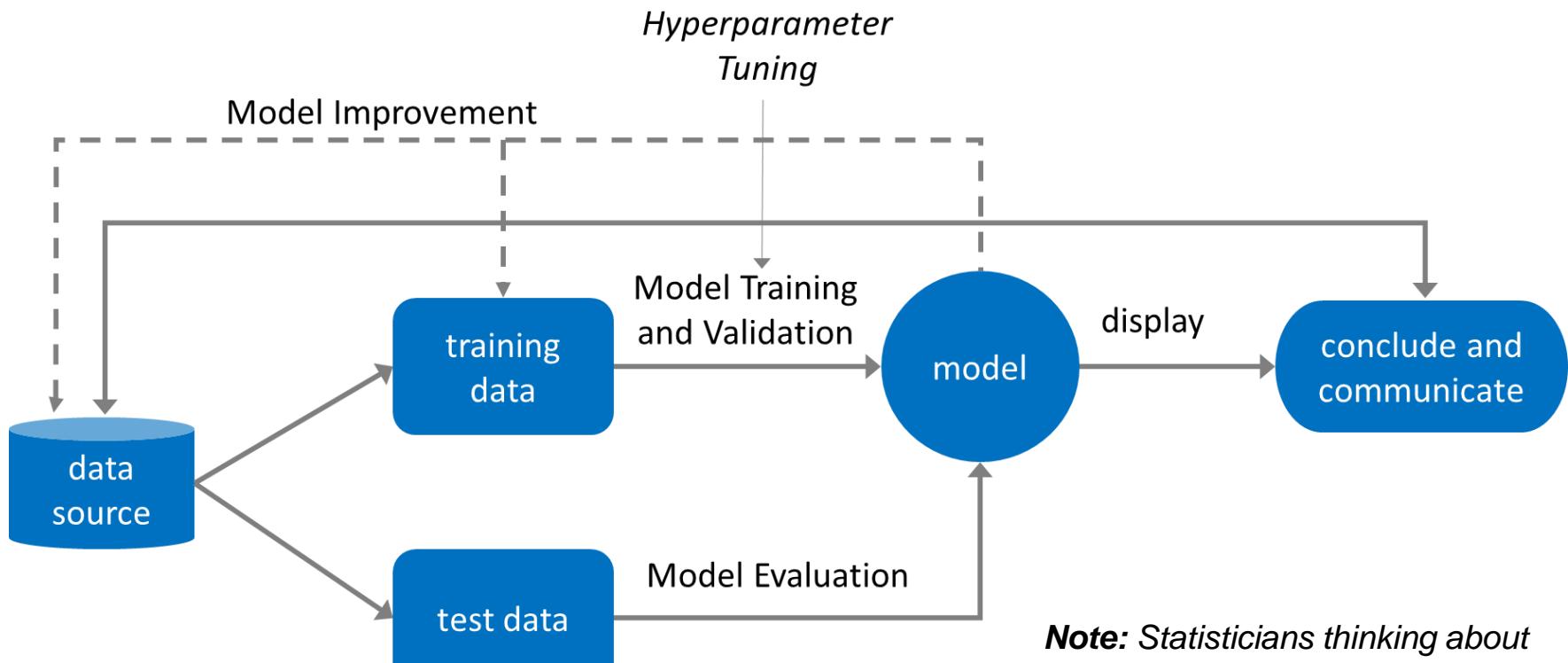
Supervised Learning

Unsupervised Learning

Generic ML Workflow



Generic ML Workflow



Note: Statisticians thinking about Training/Validation/Test may be different from Machine Learning folks.

Unsupervised Learning - Clustering

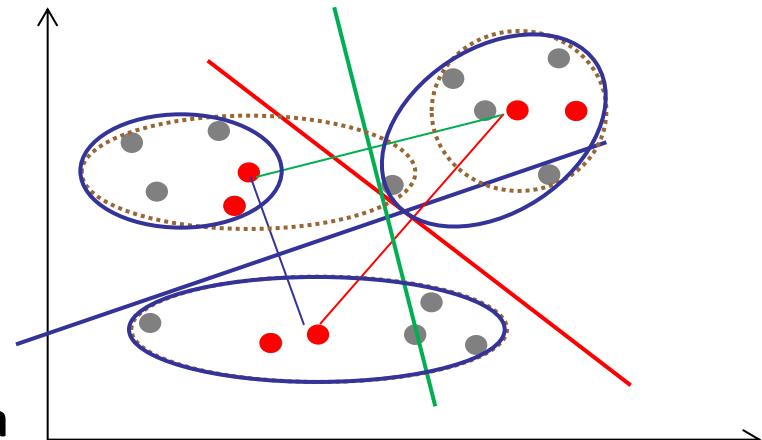


In clustering, labels are inferred from the data

[What Is Machine Learning? 3 things you need to know](#), MathWorks

K-Means Clustering

1. Value of **K** is **decided** and **K seeds** are **assigned**
2. Each **observation** (point or row) is **allocated** to the **closest seed** to get **K clusters**
3. **Compute** the **centroid** of each **cluster** as the **new seed**
4. **Reassign** each observation to the **new** clusters, based on the new distance from the **new seeds**
5. Iterate step 2 to 4 till until a **stopping criteria** is met (in practice, five to 25 iterations is likely to arrive at a stable solution)



Pre-Processing Data Before Clustering

Rescaling with Normalization and Standardization

Guidelines:

- Rescaling should be performed for **all** variables
- For highly skewed variables, additional transformations like log or square root can reduce the skewness for the **affected** variables.
- Always check for outliers in the variables

Total Sum-of-Squares vs. Total Within-Cluster Sum-of-Squares

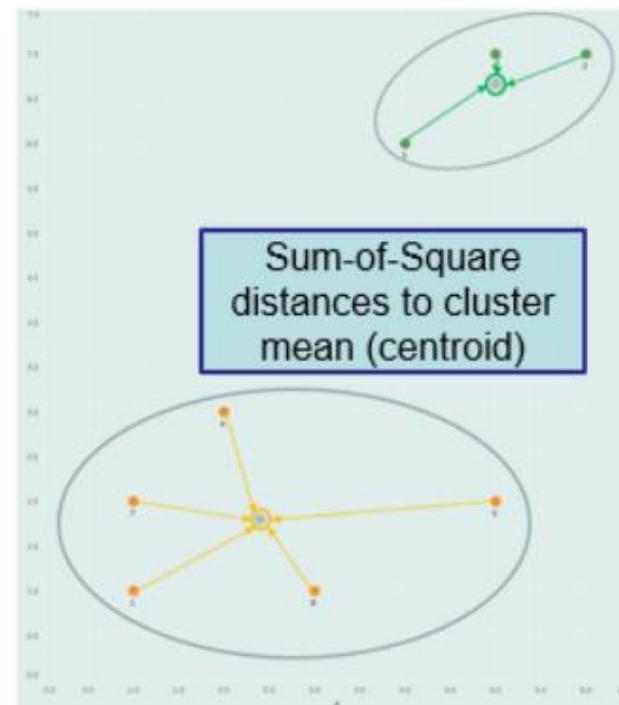
Total Sum-of-Squares (TSS): Total variance in the data (without clustering)

Total Within-Cluster Sum-of-Squares (WSS):

Each observation is assigned to one of the k clusters,

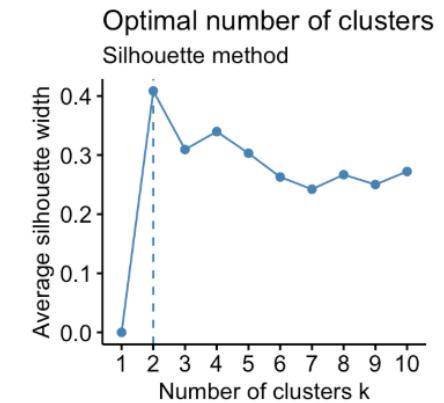
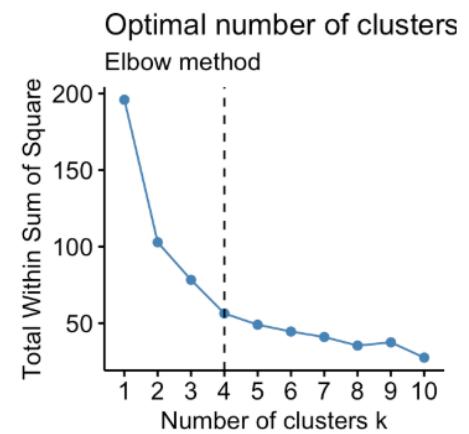
1. **Cluster mean (or centroid):** The coordinates representing the mean of the observations within the cluster for each dimension.
2. **Compute Sum of Squared Errors (distances) d^2** for each observation i , to the **respective cluster centroid**.

$$\text{Total WSS} = \sum_{i=1}^n d_i^2$$



Finding a Suitable K (Number of Cluster)

- Elbow / Within-Cluster Sum-of-Squares Method
 - Look for an “elbow” in the graph of k vs WSS. Pick the point where WSS falls rapidly until the changes are small with increasing k.
- Silhouette Method
 - Silhouette Coefficient combines ideas of both **(a)** cohesion and **(b)** separation
 - Plot k vs silhouette coefficient
 - Silhouette coefficient $s = 1 - a/b$ (typically $a < b$)
 - The closer to 1 the better



Assignment 2

- Deadline: **6 April 2025**
 - Part 1: Structured Questions (40 marks)
 - Part 2: Healthcare Analytics Case Study (60 marks)

Assignment Reminder

- Assignment: 15%
 - Individual Assignment
 - Deadline for submission is given below

	Posted	Deadline	Weightage
Assignment 1	2 Feb	2 Mar	6%
Assignment 2	8 Mar	6 Apr	9%

Class Outline

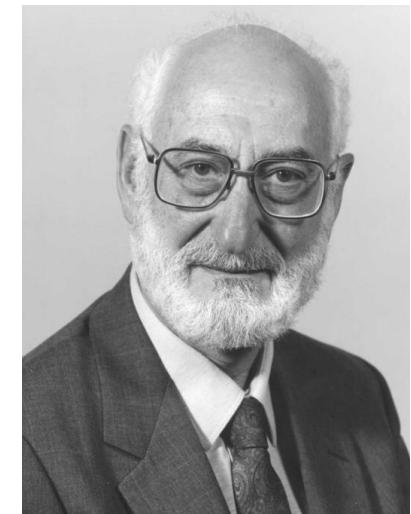
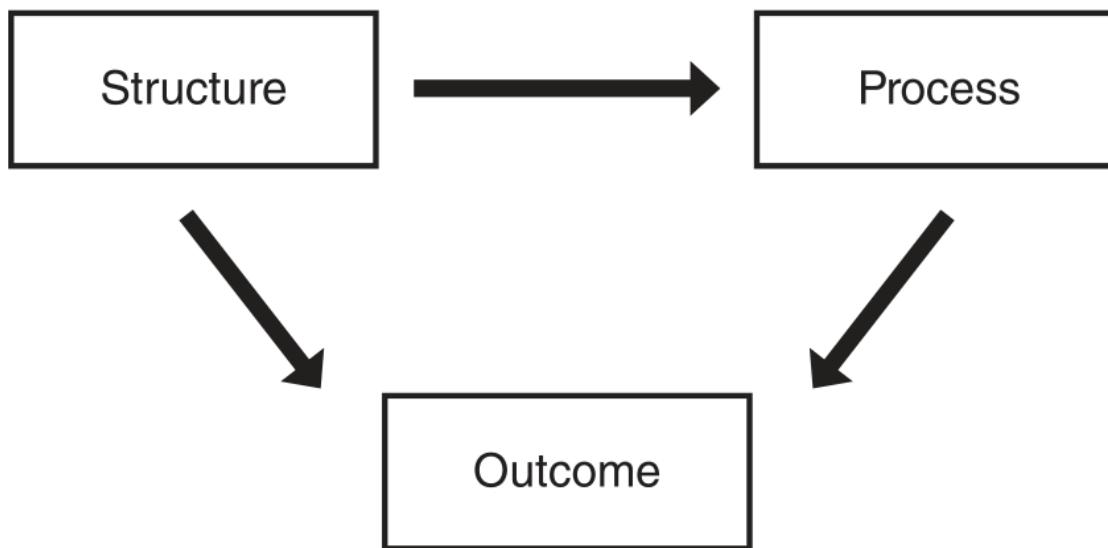
- Review: Patient Segmentation
- Quality of Care
- Classification and Regression Trees
- Case Study: Predicting Quality of Diabetes Care Using Trees and Random Forests

Healthcare Quality Assessment

- **Importance:** Critical in improving care quality and efficiency in healthcare delivery
 - Timely intervention to revert poor-quality care
 - Improve operational efficiency without sacrificing quality
 - Growing demand due to aging population and lifestyle problems
 - Capacity shortage in healthcare providers
- **Challenge**
 - How should we assess the quality of care?
 - Can we quantify the quality of care?

Donabedian Model

- A framework for quality measurement in health services research proposed in 1966



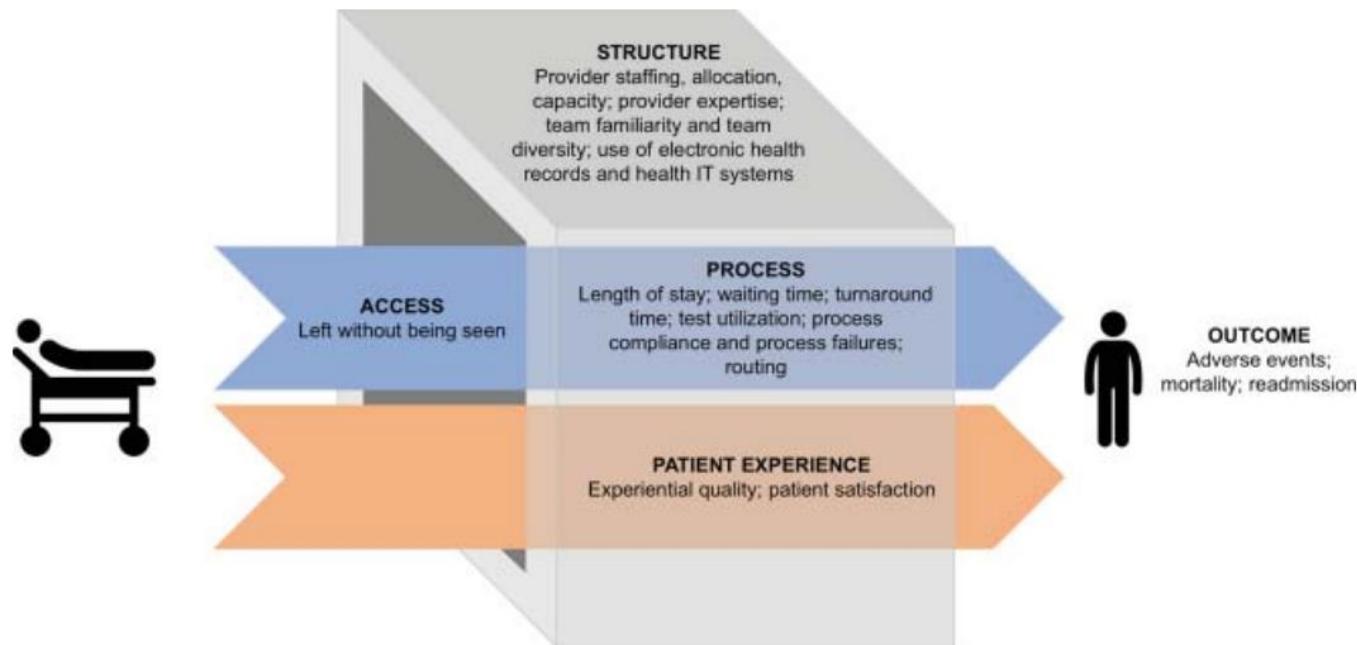
Avedis Donabedian
(1919 – 2000)

Donabedian Model

Domain	Definition	Examples
Structure	Attributes of the setting in which care is delivered, including material resources, human resources, and organizational structure	Nurse-to-patient ratios, levels of clinician expertise or skill, use of an electronic medical record system, reimbursement method
Process	What is actually done in giving and receiving care; what healthcare-related activities are performed	Percent of diabetic patients being checked for hemoglobin A1C levels, percent of patients receiving surgery for whom a surgical safety checklist was followed
Outcome	Effects of care on the health status of patients and populations resulting from healthcare services	Blood pressure control rate, 30-day mortality, functional status, changes in patient's health-related behaviors

AHRQ Framework

- Agency for Healthcare Research and Quality (AHRQ)
- New dimensions: *access* and *patient experience*



Poor Quality of Care

- Patient safety
- Medical errors
- Overuse, underuse, and misuse of care
- Etc.
- Can we timely detect the poor quality of care of every individual patient so that proper intervention can take place?

Experts Assessment

- Health professionals are experts in quality-of-care assessment
 - Expert physicians can evaluate quality by examining a patient's records
 - This process is time consuming and inefficient
 - Experts are limited by memory and time
 - They cannot assess quality for millions of patients
- Similar practice in other industries
 - Accreditation in education, manufacturing, etc.

Replicating Expert Assessment

- Can we develop analytical tools that replicate expert assessment?
 - Learn from expert human judgment
 - Develop a model, interpret results, and adjust the model
- Make predictions/evaluations on a large scale

Quality of Care and Claims Data

- In 2013, Professor Dimitris Bertsimas from MIT tried to predict the quality of care for diabetes patients using patients' health insurance claims and expert's rating on the quality of care
- Identify potential candidates for case management



Claims Data

Medical Claims

Diagnosis, Procedures,
Doctor/Hospital, Cost

Pharmacy Claims

Drug, Quantity, Doctor,
Medication Cost

- Electronically available
- Standardized and well-regarded as data have financial implication
- But they are not 100% accurate
- Claims for hospital visits can be vague

Creating the Dataset – Claims Samples

Claims Sample

- Large health insurance claims database
- Randomly selected 131 diabetes patients
- Ages range from 35 to 55
- Costs \$10,000 – \$20,000
- September 1, 2003 – August 31, 2005

Creating the Dataset – Expert Review

Claims Sample

Expert Review

- Expert physician reviewed claims and wrote descriptive notes:
 - “Ongoing use of narcotics”
 - “Only on Avandia, not a good first choice drug”
 - “Had regular visits, mammogram, and immunizations”
 - “Was given home testing supplies”

Creating the Dataset – Expert Assessment

Claims Sample

- Rated quality on a two-point scale (poor/good)

Expert Review

“I’d say **care was poor** – poorly treated diabetes”

Expert Assessment

“No eye care, but overall I’d say **high quality**”

Creating the Dataset – Variable Extraction

Claims Sample

Expert Review

Expert Assessment

Variable Extraction

- Dependent Variable
 - Quality of care
- Independent Variables
 - Ongoing use of **narcotics**
 - **Only on Avandia**, not a good first choice drug
 - Had **regular visits, mammogram, and immunizations**
 - Was given **home testing supplies**

Creating the Dataset – Variable Extraction

Claims Sample

Expert Review

Expert Assessment

Variable Extraction

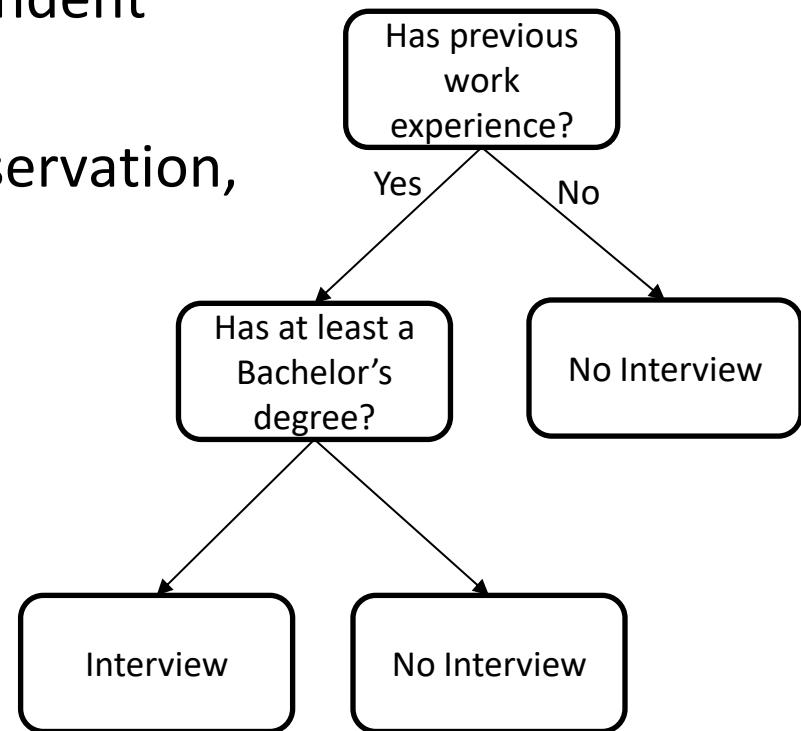
- Dependent Variable
 - Quality of care
- Independent Variables
 - Diabetes treatment
 - Patient demographics
 - Healthcare utilization
 - Providers
 - Claims
 - Prescriptions

Class Outline

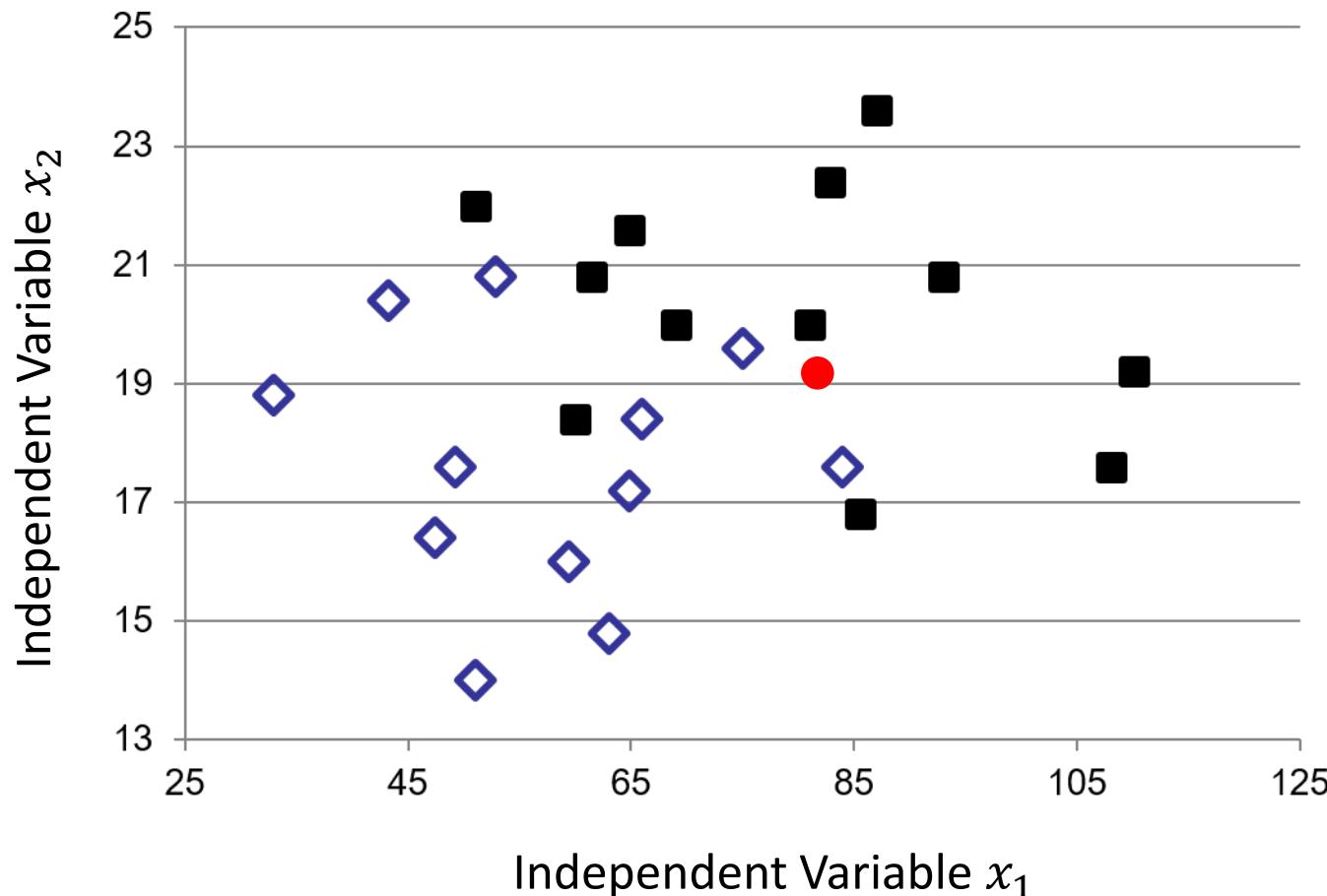
- Review: Patient Segmentation
- Quality of Care
- Classification and Regression Trees
- Case Study: Predicting Quality of Diabetes Care Using Trees and Random Forests

Classification and Regression Trees (CART)

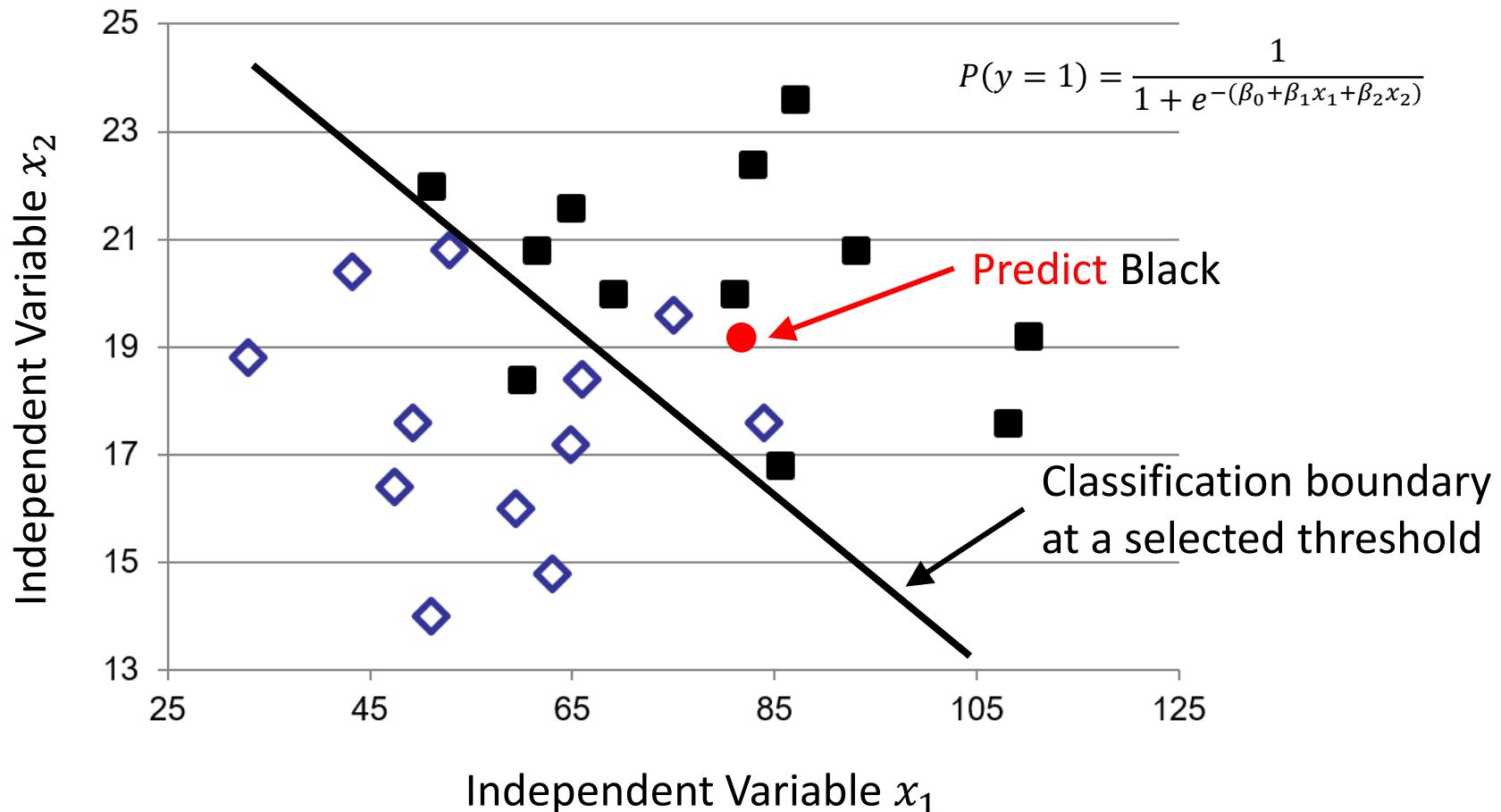
- Build a tree by splitting on independent variables
- To predict the outcome for an observation, follow the splits and at the end...
- Advantages
 - Does not assume a linear model
 - More Interpretable
 - Easier variable selection
 - Handle multicollinearity problem in a better way
 - Control tree growth via parameters



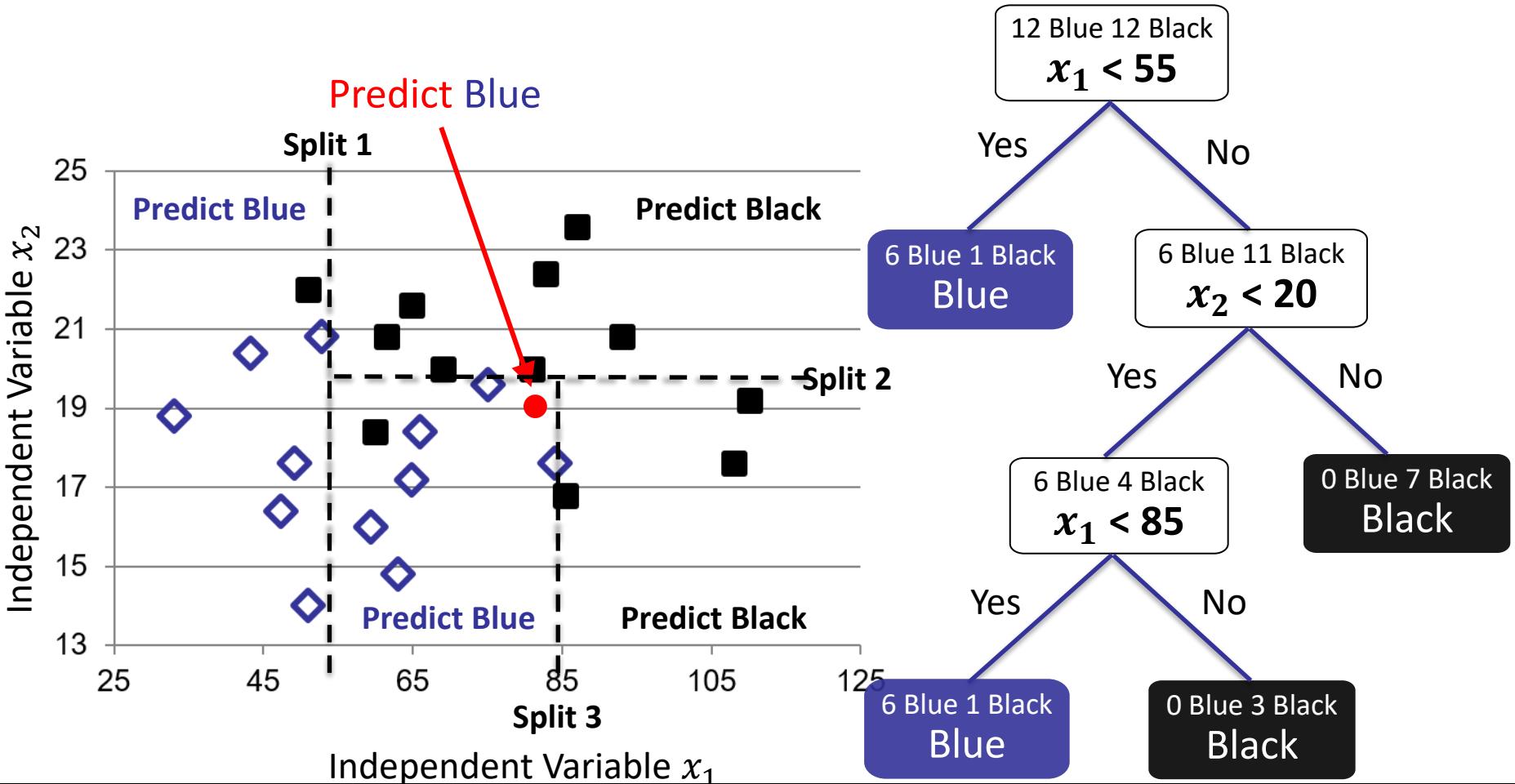
Example: Predict Black or Blue?



Predict Black or Blue Using Logistic Regression



Predict Black or Blue Using CART



Predictions from CART

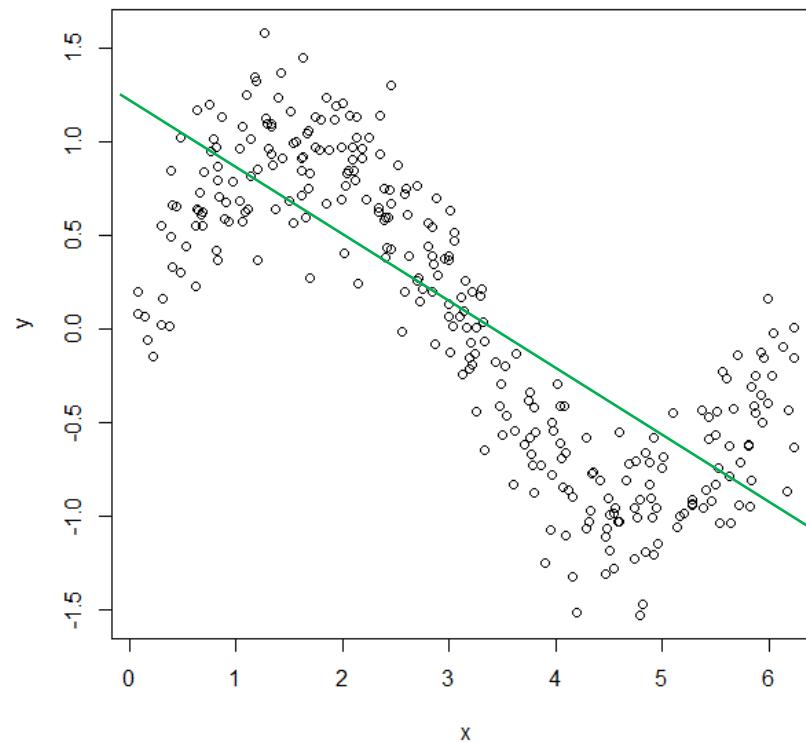
- At each ***terminal (leaf) node*** of the tree, we have a bucket of observations, which may contain both outcomes (i.e., black or blue)
- For each leaf node, we can compute percentage of points in one group (“purity”)
 - For example, 6 blue, 1 black $\Rightarrow P(\text{blue}) = 6/(6 + 1) = 0.857$
- Just like in logistic regression, we can use a threshold to obtain a prediction
 - Default threshold of 0.5 for CART
 - Predicting the most frequent outcome

Regression Trees

- *Classification trees* are used to predict categorical variables
- *Regression trees* predict numerical variables
- Regression trees are constructed similar to classification trees
- To obtain predictions from regression trees, we simply report the average of the values at the leaf node

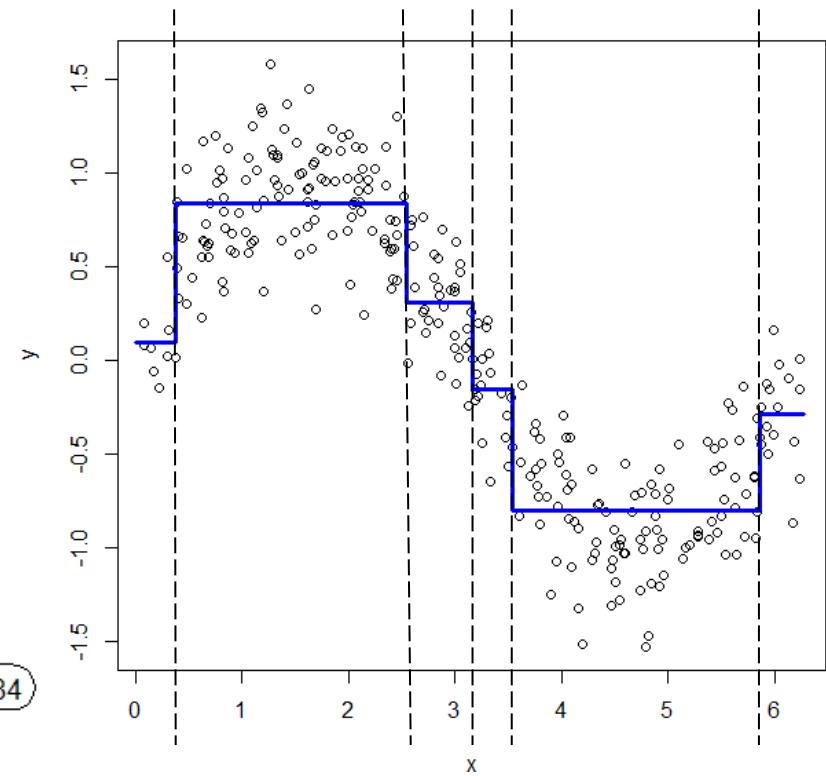
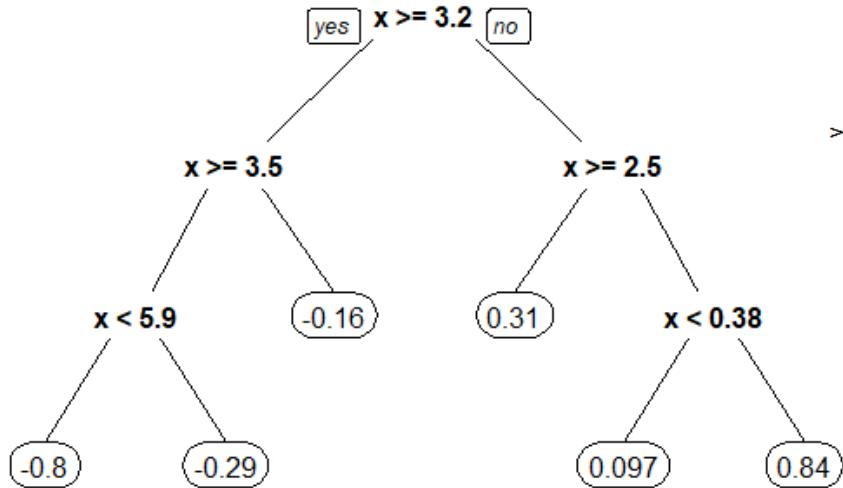
Example – Regression Tree

- Fit a linear regression model using x to predict y



Example – Regression Tree

- Fit a regression tree model using x to predict y
- 5 splits are used



Example – Classification Tree

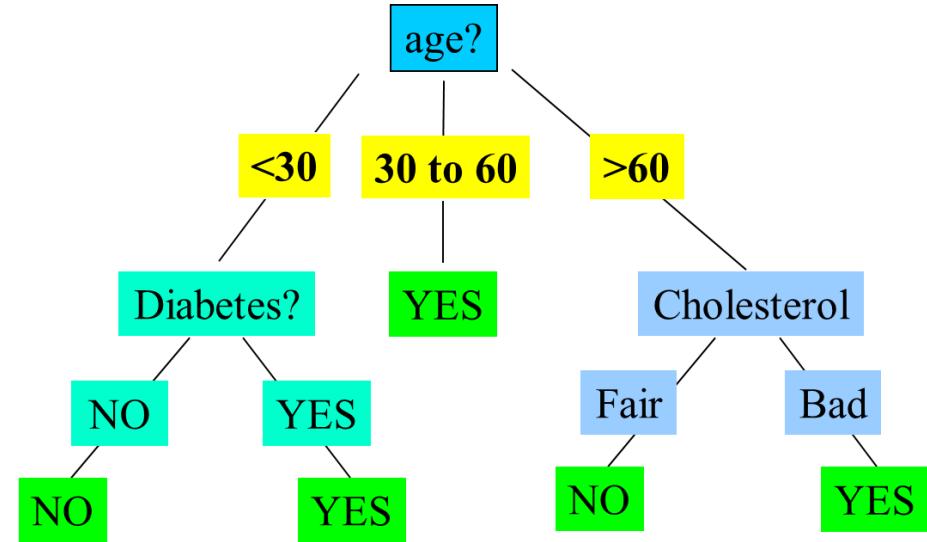
Predict ED Admission (within 1 year)

- Consider the following independent variables:
 - Age
 - Income
 - Diabetes Flag
 - Cholesterol Flag
- Outcome variable:
 - ED Admission (within 1 year)

Age	Income	Diabetes	Cholesterol	ED Admission
<30	High	NO	BAD	NO
<30	High	NO	FAIR	NO
30-60	High	NO	BAD	YES
>60	Medium	NO	BAD	YES
>60	Low	YES	BAD	YES
>60	Low	YES	FAIR	NO
30-60	Low	YES	FAIR	YES
<30	Medium	NO	BAD	NO
<30	Low	YES	BAD	YES
>60	Medium	YES	BAD	YES
<30	Medium	YES	FAIR	YES
30-60	Medium	NO	FAIR	YES
30-60	High	YES	BAD	YES
>60	Medium	NO	FAIR	NO

Example – Classification Tree

Age	Income	Diabetes	Cholesterol	ED Admission
<30	High	NO	BAD	NO
<30	High	NO	FAIR	NO
30-60	High	NO	BAD	YES
>60	Medium	NO	BAD	YES
>60	Low	YES	BAD	YES
>60	Low	YES	FAIR	NO
30-60	Low	YES	FAIR	YES
<30	Medium	NO	BAD	NO
<30	Low	YES	BAD	YES
>60	Medium	YES	BAD	YES
<30	Medium	YES	FAIR	YES
30-60	Medium	NO	FAIR	YES
30-60	High	YES	BAD	YES
>60	Medium	NO	FAIR	NO



Adapted from: Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1), 81-106.

How does CART Branch?

How Does CART Branch Out?

- All input variables and all possible split points are evaluated and chosen in a **greedy** manner
 - Minimize the **Gini index** (also known as **Gini impurity**) for classification trees (predicting categorical variables)
 - Gini index measures how “pure” the leaf nodes are
 - Gini index = 0 if each leaf node has only one class
 - Minimize the **Entropy** associated with the random variable
 - Minimize the **SSE** for regression trees (predicting numerical variables)

Concept: Gini Impurity

- Measure of the “**impurity**” in the classification
- Given an rv Y with n classes, Gini impurity is defined as:

$$Gini(Y) = \sum_{j=1}^n p_j(1 - p_j) = 1 - \sum_{j=1}^n p_j^2$$

where p_j : Relative frequency of class j in Y ; n : number of classes

- Given Y is split into two subsets (D_1 and D_2) by attribute A , the **revised Gini impurity** is (**weighted by the subgroup size**):

$$Gini_A(Y) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

- The **change in Gini impurity (reduction in impurity)** is given by:

$$\Delta Gini(A) = Gini(Y) - Gini_A(Y)$$

- The attribute which provide the **largest reduction in impurity** is chosen to split the node (greediness)

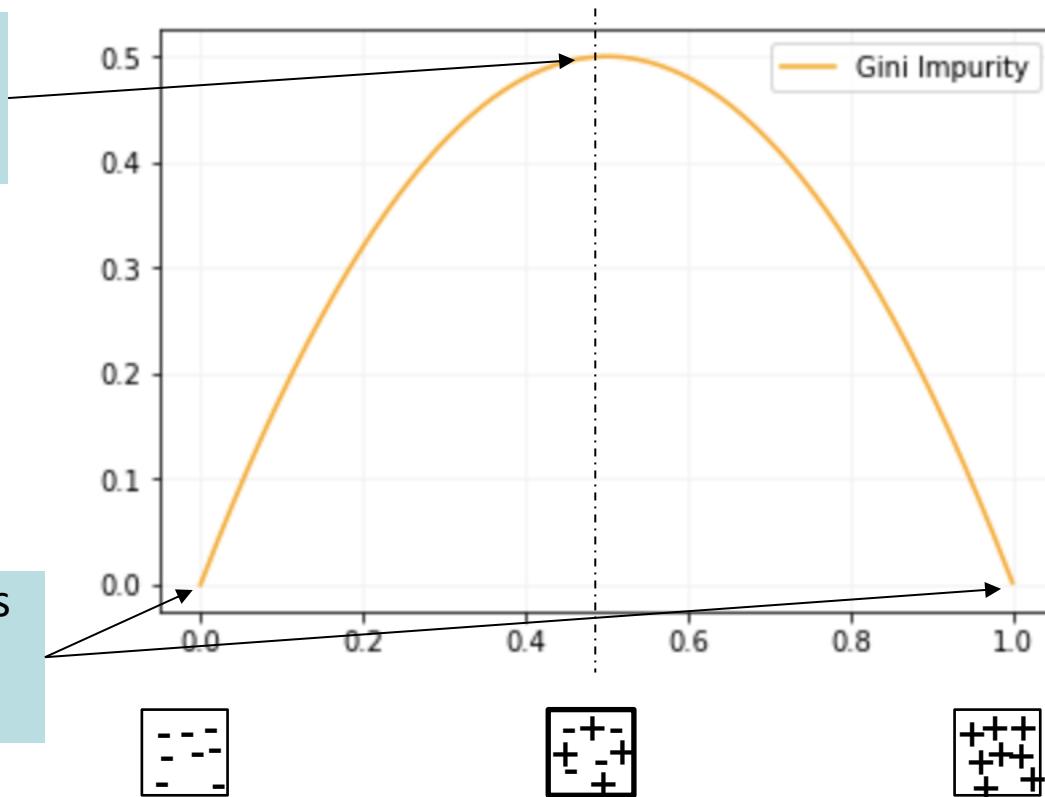
Reference: Provost, Foster; Fawcett, Tom. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking

Concept: Gini Impurity

Gini Impurity curve

If we have mix of +
and - samples:
High uncertainty

If we have only + samples
or only – samples:
Low uncertainty

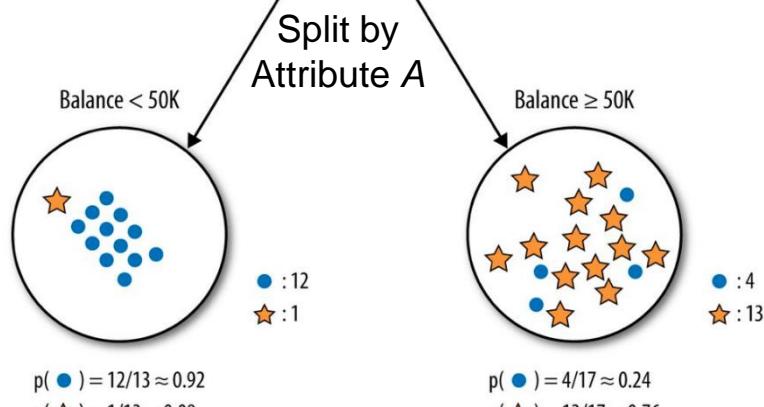
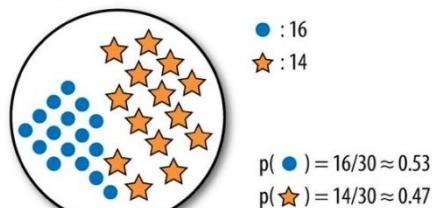


Concept: Gini Impurity

$$Gini(Y) = 1 - \left(\frac{16}{30}\right)^2 - \left(\frac{14}{30}\right)^2 = 0.459$$

Gini: 0.49778

Entire population (30 instances)



Gini: 0.142012

Gini: 0.35986

These are well-implemented in packages

- Do **not** have to worry about it and do **not** have control over it
- We can and have to control when CART should stop splitting

$$Gini_A(Y) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

$$Gini_A(Y) = 0.265$$

$$\Delta Gini(A) = 0.232$$

Concept: Gini impurity

- Predict ED Admission (within 1 year)

Age	Income	Diabetes	Cholesterol	ED Admission
<30	High	NO	BAD	NO
<30	High	NO	FAIR	NO
30-60	High	NO	BAD	YES
>60	Medium	NO	BAD	YES
>60	Low	YES	BAD	YES
>60	Low	YES	FAIR	NO
30-60	Low	YES	FAIR	YES
<30	Medium	NO	BAD	NO
<30	Low	YES	BAD	YES
>60	Medium	YES	BAD	YES
<30	Medium	YES	FAIR	YES
30-60	Medium	NO	FAIR	YES
30-60	High	YES	BAD	YES
>60	Medium	NO	FAIR	NO

We have 14 examples at the root for “ED Admission” (9 are “YES” and 5 are “NO”):

$$Gini(Y) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459$$

Suppose the Attribute “Income” partitions Y into 10 patients with D_1 : {low, medium} and 4 patients with D_2 {High},

$$Gini_{D}(Y)$$

$$= \frac{10}{14} \left(1 - \left(\frac{7}{10} \right)^2 - \left(\frac{3}{10} \right)^2 \right) + \frac{4}{14} \left(1 - \left(\frac{2}{4} \right)^2 - \left(\frac{2}{4} \right)^2 \right)$$

$$= 0.443$$

7 ED and 3 no ED for
low and medium
income.

2 ED and 2 no ED
among 4 high income
customers

The gain after split on **Income** is:
 $0.459 - 0.443 = 0.016$

Concept: Entropy

Entropy: **Measure of uncertainty** associated with a random variable (rv). For a discrete rv (Y) taking n distinct values (y_1, \dots, y_n) ,

Entropy is: $H(Y) = -\sum_{i=1}^n p_i \log_2(p_i)$ where $p_i = P(Y = y_i)$

- Interpretation:
 - Higher Entropy \rightarrow Higher Uncertainty in the rv
 - Lower Entropy \rightarrow Lower Uncertainty in the rv
- In Estimation problems,

Conditional Entropy is given by:

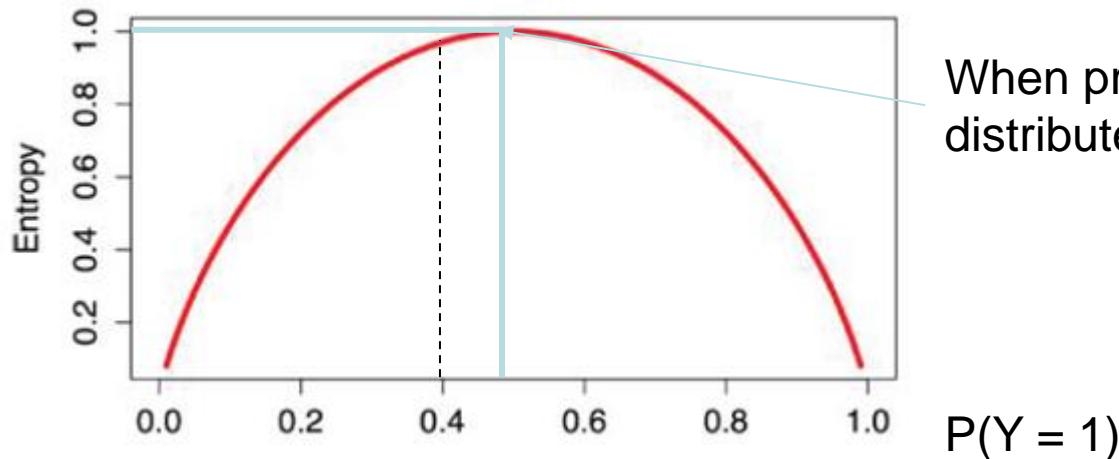
$$H(Y|X) = \sum_x p(x)H(Y|X=x)$$

Concept: Entropy

Entropy:

- If there are 40% “1” and 60% “0” in rv Y , then the entropy is 0.9709506 calculated as:

$$\begin{aligned} H(Y) &= - \sum_{i=1}^n p_i \log_2(p_i) \\ &= -[0.6 \times \log_2(0.6) + 0.4 \times \log_2(0.4)] = 0.9709506 \end{aligned}$$

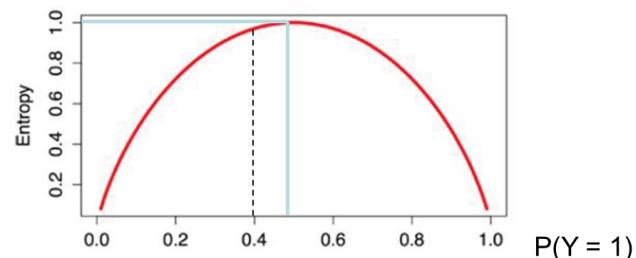


When proportions are equally distributed, Entropy = 1

Log base 2 Calculations

$$\log_2 x = \frac{\log_{10} x}{\log_{10} 2}$$

$$\begin{aligned}\text{Entropy } H(Y) &= -\sum_{i=1}^n p_i \log_2(p_i) \\&= -[0.6 \times \log_2(0.6) + 0.4 \times \log_2(0.4)] \\&= -[0.6 \times \frac{\log_{10} 0.6}{\log_{10} 2} + 0.4 \times \frac{\log_{10} 0.4}{\log_{10} 2}] \\&= 0.9709506\end{aligned}$$



Concept: Entropy

- Before any split, **Expected Entropy** of Y with n classes is:

$$Info(Y) = - \sum_{i=1}^n p_i \log_2(p_i)$$

- With a new feature (attribute A) to **split into m partitions**, the **expected entropy** is:

$$Info_A(Y) = \sum_{j=1}^m \frac{|Y_j|}{|Y|} \times Info(Y_j)$$

- **Information gain** by splitting according to A is defined as:

$$Gain(A) = Info(Y) - Info_A(Y)$$

Concept: Entropy

- Calculating Entropy

Age	Income	Diabetes	Cholesterol	ED Admission
<30	High	NO	BAD	NO
<30	High	NO	FAIR	NO
30-60	High	NO	BAD	YES
>60	Medium	NO	BAD	YES
>60	Low	YES	BAD	YES
>60	Low	YES	FAIR	NO
30-60	Low	YES	FAIR	YES
<30	Medium	NO	BAD	NO
<30	Low	YES	BAD	YES
>60	Medium	YES	BAD	YES
<30	Medium	YES	FAIR	YES
30-60	Medium	NO	FAIR	YES
30-60	High	YES	BAD	YES
>60	Medium	NO	FAIR	NO

Predict ED Admission

- We have 14 examples at the root for “ED Admission” (9 are “YES” and 5 are “NO”)
- The Expected Entropy for this outcome variable is thus:

$$\begin{aligned}
 Info(Y) &= I(9,5) \\
 &= -\frac{9}{14} \log_2\left(\frac{9}{14}\right) - \frac{5}{14} \log_2\left(\frac{5}{14}\right) \\
 &= 0.940
 \end{aligned}$$

Information Gain

- In each iteration of the tree, we select the attribute with the highest **information gain**.
- The algorithm **calculates the information gain of each feature** and pick the best one (highest)
- For “Age”, upon splitting by this attribute, expected entropy is:

$$Info_{age}(Y) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0.694$$

- Information Gain:

$$Gain(Age) = Info(Y) - Info_A(Y) = 0.246$$

Information Gain – Computing $Info_{age}$

$$Info_{age}(Y) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694$$

“Age<30” has 5 out of 14 samples, with 2 YES and 3 NO

“Age>60” has 5 out of 14 samples, with 3 YES and 2 NO

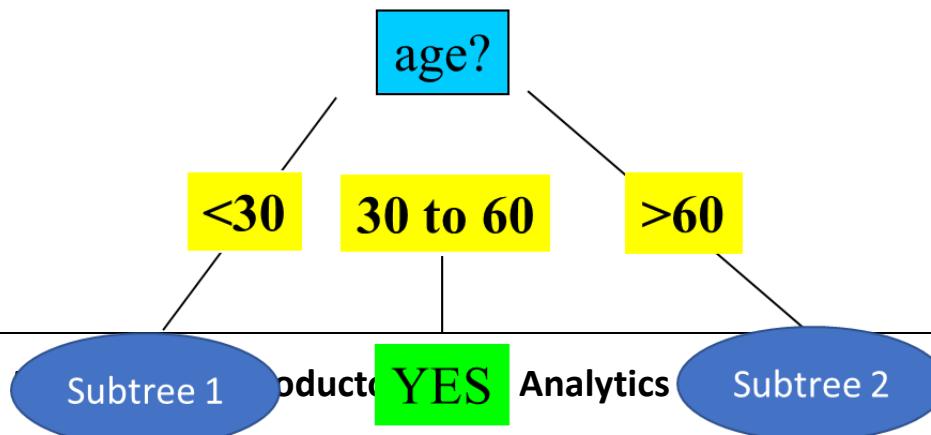
Age	YES	NO	I(YES, NO)
<30	2	3	0.971
30...60	4	0	0
>60	3	2	0.971

$$I(2,3) = I(3,2) = -(2/5)*\log_2(2/5) - (3/5)*\log_2(3/5) = 0.970951$$

Age	Income	Diabetes	Cholesterol	ED Admit
<30	High	NO	BAD	NO
<30	High	NO	FAIR	NO
30-60	High	NO	BAD	YES
>60	Medium	NO	BAD	YES
>60	Low	YES	BAD	YES
>60	Low	YES	FAIR	NO
30-60	Low	YES	FAIR	YES
<30	Medium	NO	BAD	NO
<30	Low	YES	BAD	YES
>60	Medium	YES	BAD	YES
<30	Medium	YES	FAIR	YES
30-60	Medium	NO	FAIR	YES
30-60	High	YES	BAD	YES
>60	Medium	NO	FAIR	NO

Information Gain

- We can calculate the information gain of the other 3 variables from the root split:
 - Gain(Income) = 0.029, Gain(Diabetes) = 0.151, and Gain(Cholesterol) = 0.048.
- Gain(Age) gives the best split for the first level as it reduces entropy the most
- We repeat the process for the sub-trees below “Age”
 - We need to do so for only Age<30 and Age>60 sub-tree
 - Age between 30-60 is completed because all examples are “yes” already.



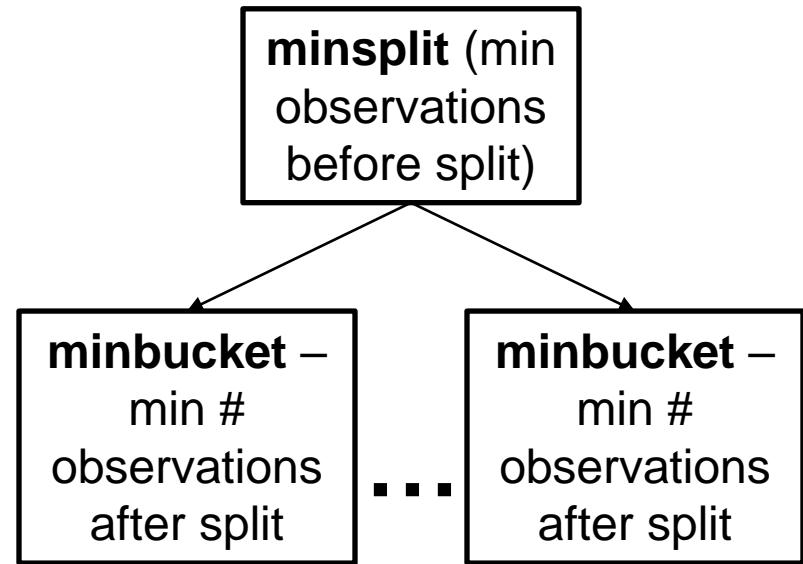
How does CART Stop?

When Does CART Stop Splitting?

- There are different ways to control how many splits are generated
- One way is by setting a lower bound for the **number of observations in any leaf node**
 - To avoid leaf nodes that have only a few observations
- In R, a parameter that controls this is ***minbucket***
 - The smaller it is, the more splits will be generated
 - If it is **too small, overfitting** will occur
 - If it is **too large, model will be too simple** and accuracy will be poor
 - Its default value is 7, which is 1/3 of another parameter...

A Related Parameter

- Another parameter related to ***minbucket*** is ***minsplit***
 - The **minimum number of observations that must exist in a node in order for a split to be attempted**
 - Its default value is 20
 - If only one of *minbucket* or *minsplit* is specified, *minbucket* = *minsplit*/3, or *minsplit* = 3**minbucket*



Controlling the Depth of the Final Tree

- We can also set the **maximum depth of any node of the final tree**
- In R, this parameter is called *maxdepth*
 - The root node counted as depth 0
 - Its default value is 30
 - 30 is a very big number

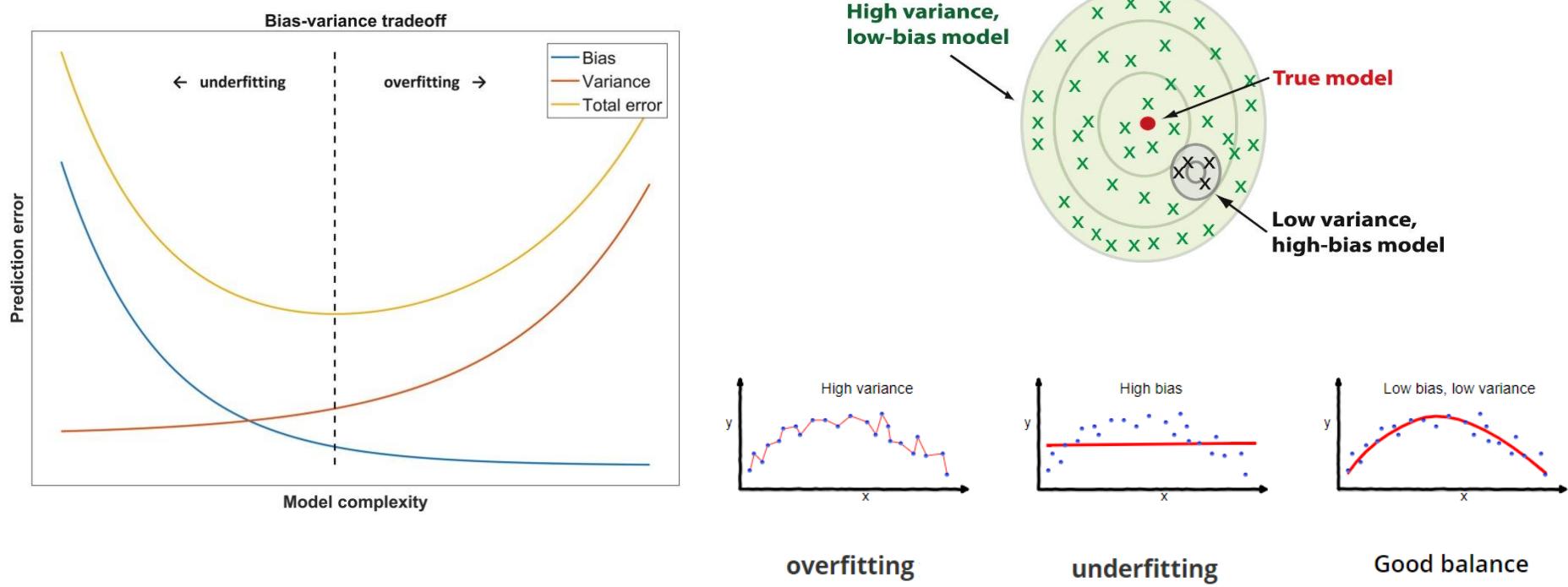
The *cp* Parameter

- There is another parameter called *cp*
 - Complexity parameter with value between 0 and 1
 - Control tree growth by penalizing each split
 - The algorithm has to “pay a cost” to generate any additional split
- Larger *cp* (higher penalty on each split) leads to a smaller tree
 - *cp = 1* will result in a tree with no splits, i.e., the smallest tree; equivalent to set *minbucket* = data size
- Smaller *cp* (lower penalty on each split) leads to a bigger tree
 - *cp = 0* will result in a tree with each data point as a leaf node, i.e., the biggest tree; equivalent to set *minbucket* = 1

How Penalty Works

- Common idea used in all predictive models
 - Trade-off between ***training error*** (also known as ***training loss***) and ***model complexity***
 - The second term is usually called a ***regularization*** term
- Regularization:
 - Model parameters are shrunk to zero to achieve ***Model Parsimony***
 - How to evaluate the best set of parameters?
 - **Hyperparameter** tuning
 - **Cross-validation**

Bias Variance Tradeoffs



Elaborate more in the next lecture

“minbucket” vs. “cp”

	<i>Minbucket</i>	<i>cp</i>
Larger Values	<ul style="list-style-type: none">Allow larger leaf nodesEncourage smaller (simpler) treesMay predict poorly	<ul style="list-style-type: none">Penalize more for each splitEncourage smaller (simpler) treesMay predict poorly
Smaller Values	<ul style="list-style-type: none">Allow smaller leaf nodesEncourage larger (more complicated) treesMay overfit	<ul style="list-style-type: none">Penalize less for each splitEncourage larger (more complicated) treesMay overfit

Class Outline

- Review: Patient Segmentation
- Quality of Care
- Classification and Regression Trees
- Case Study: Predicting Quality of Diabetes Care Using Trees and Random Forests

Let's get our hands dirty!

Data Analysis

Let's do some basic data analysis using our WHO data.

[WHO\\$Under15](#) Hide

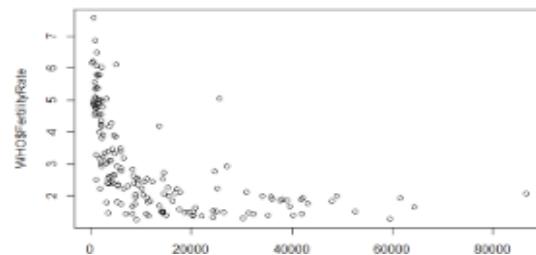
```
[1] 47.42 21.33 27.42 15.28 47.58 25.96 24.42 28.34 18.95 14.51 22.25 21.62 28.16 30.57 18.99 15.18 16.88 34.4
8 42.95 28.53
[21] 35.23 16.35 33.75 24.56 25.75 13.53 45.66 44.20 31.23 43.08 16.37 38.17 48.87 48.52 21.38 17.95 28.03 42.1
7 42.37 38.61
[41] 23.94 41.48 14.98 16.58 17.16 14.56 21.98 45.11 17.66 33.72 25.96 38.53 38.29 31.25 38.62 38.95 43.18 15.6
9 43.29 28.88
[61] 16.42 18.26 38.49 45.98 17.62 13.17 38.59 14.68 26.96 48.88 42.46 41.55 36.77 35.35 35.72 14.63 28.71 29.4
3 29.27 23.68
[81] 48.51 21.54 27.53 14.04 27.78 13.12 34.13 25.46 42.37 38.18 24.98 38.21 35.61 14.57 21.64 36.75 43.06 29.4
5 15.13 17.46
[101] 42.72 45.64 26.65 29.03 47.14 14.98 38.18 48.22 28.17 29.82 35.81 18.26 27.85 19.81 27.85 45.38 25.28 36.5
9 38.18 35.58
[121] 17.21 20.26 33.37 49.99 44.23 38.61 18.64 24.19 34.31 38.18 28.65 38.37 32.78 29.18 34.53 14.91 14.92 13.2
8 15.25 16.52
[141] 15.85 15.45 43.56 25.96 24.31 25.78 37.88 14.04 41.68 29.69 43.54 16.45 21.95 41.74 16.48 15.08 14.16 40.3
7 47.35 29.53
[161] 42.28 15.28 25.15 41.48 27.83 38.85 16.71 14.79 35.35 35.75 18.47 16.89 46.33 41.89 37.33 28.73 23.22 26.8
0 28.05 38.61
[181] 48.54 14.18 14.41 17.54 44.85 19.63 22.05 28.98 37.37 28.84 22.87 48.72 46.73 40.24
```

[WHO\\$Country\[which.min\(WHO\\$Under15\)\]](#) Hide

```
[1] Japan
194 Levels: Afghanistan Albania Algeria Andorra Angola Antigua and Barbuda Argentina Armenia Australia Austria
... Zimbabwe
```

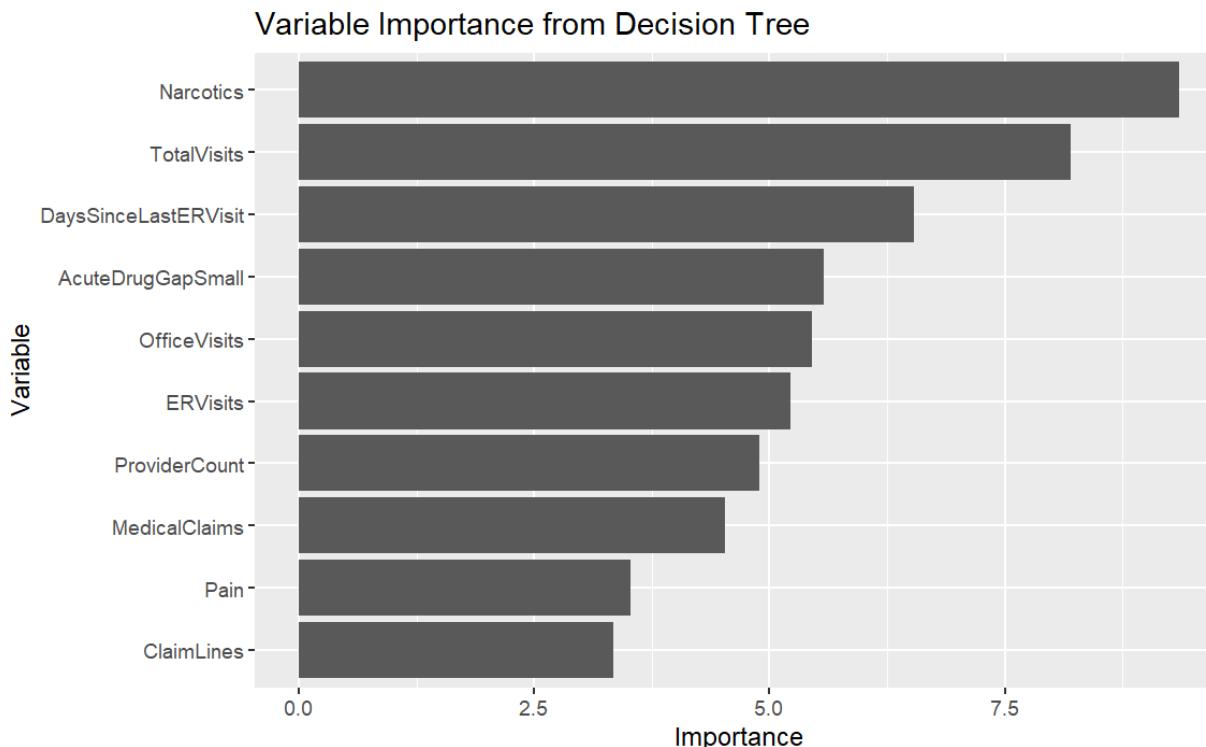
Let's create some plots for exploratory data analysis (EDA). First, let's create a basic scatterplot of GNI versus FertilityRate.

[plot\(WHO\\$GNI, WHO\\$FertilityRate\)](#) Hide



Variable Importance

- Relative importance of each variable on the model's prediction



Parameter Selection

- We can use either “*minbucket*” or “*cp*” to control the tree growth
- **Parameter selection** is **model selection** in CART
- For comparison, in linear and logistic regressions, candidate models are represented by different sets of independent variables
 - That is why we also call the process **variable selection**
- How should we set the parameter’s value?
 - Quality of Care II (Cross Validation)

Class Outline

- Review: Patient Segmentation
- Quality of Care
- Case Study: Predicting Quality of Diabetes Care Using Trees and Random Forests

End

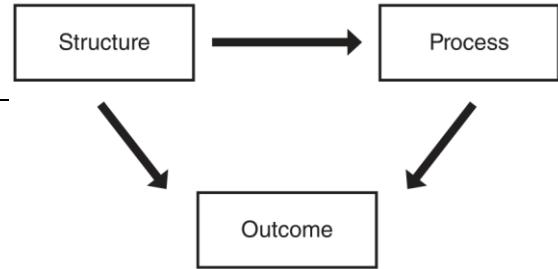
ECON 145

ECON 145 – Introductory Data Analytics in Healthcare

Lecture 8: Quality of Care II

Class Outline

- Review: Quality of Care and CART
- Model Selection: Cross-Validation
- Advanced Decision Trees
- Case Study Continued: Predicting Quality of Diabetes Care Using Trees and Random Forests

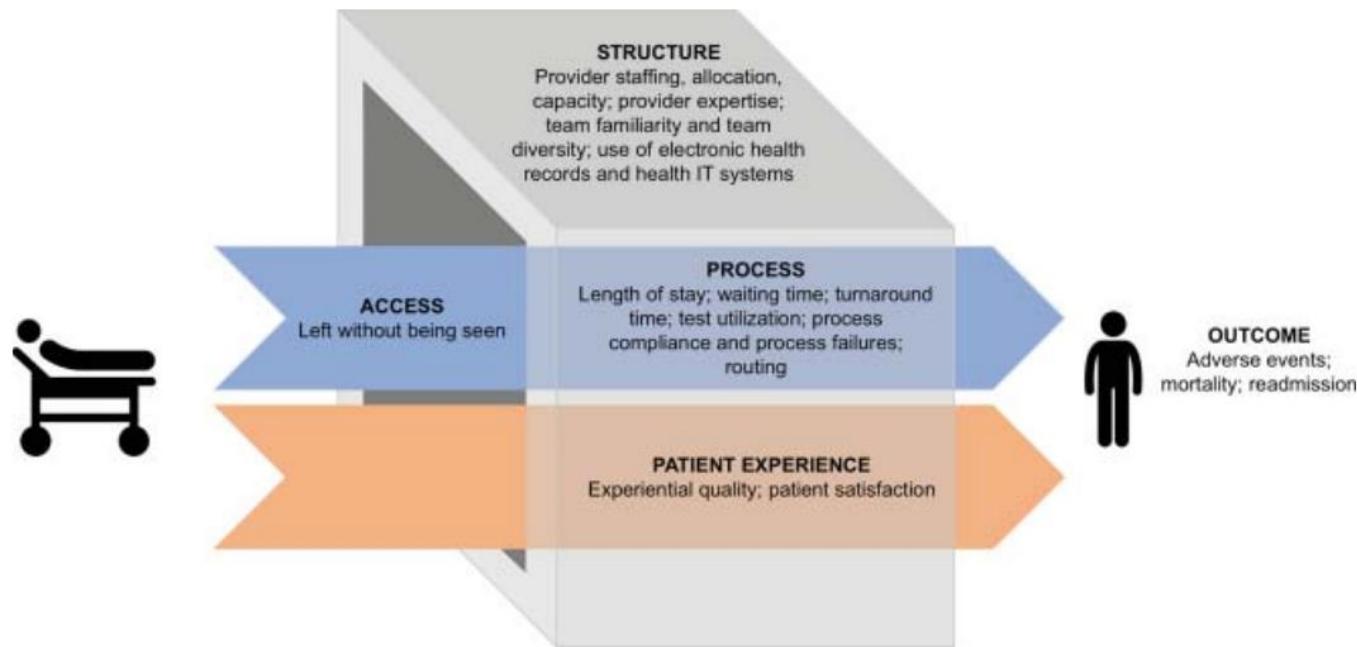


Donabedian Model

Domain	Definition	Examples
Structure	Attributes of the setting in which care is delivered, including material resources, human resources, and organizational structure	Nurse-to-patient ratios, levels of clinician expertise or skill, use of an electronic medical record system, reimbursement method
Process	What is actually done in giving and receiving care; what healthcare-related activities are performed	Percent of diabetic patients being checked for hemoglobin A1C levels, percent of patients receiving surgery for whom a surgical safety checklist was followed
Outcome	Effects of care on the health status of patients and populations resulting from healthcare services	Blood pressure control rate, 30-day mortality, functional status, changes in patient's health-related behaviors

AHRQ Framework

- Agency for Healthcare Research and Quality (AHRQ)
- New dimensions: *access* and *patient experience*

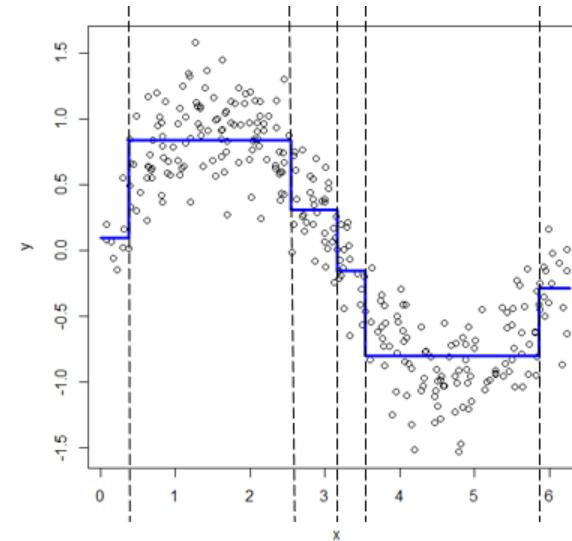
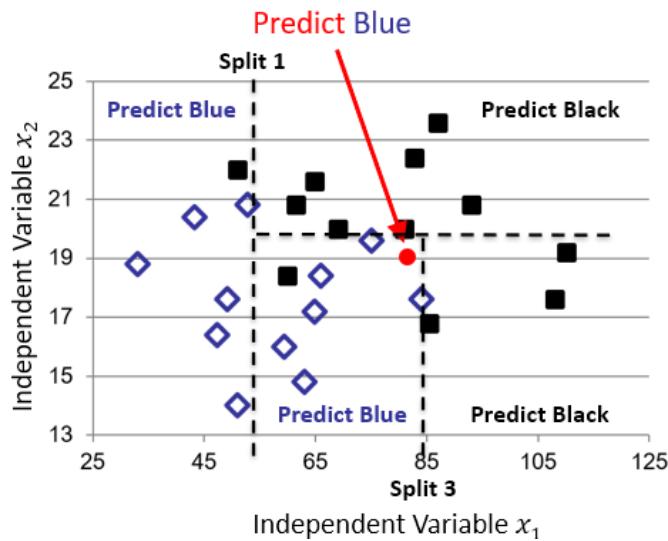


Experts Assessment

- Health professionals are experts in quality of care assessment
 - Expert physicians can evaluate quality by examining a patient's records
 - This process is time consuming and inefficient
 - Experts are limited by memory and time
 - They cannot assess quality for millions of patients
- Can we develop analytical tools that replicate expert assessment?
 - Learn from expert human judgment
 - Develop a model, interpret results, and adjust the model
 - Make predictions/evaluations on a large scale

Classification and Regression Trees (CART)

- Build a tree by splitting on independent variables
- To predict the outcome for an observation, follow the splits and at the end (leaf node) report the average outcome
- Advantages: Does not assume a linear model; more interpretable; easier variable selection



How Does CART Branch Out?

- All input variables and all possible split points are evaluated and chosen in a **greedy** manner at each split
 - Minimize the **Gini index** (also known as **Gini impurity**) for classification trees (predicting categorical variables)
 - Gini index measures how “pure” the leaf nodes are
 - Gini index = 0 if the leaf node has only one class, e.g., all blue points or all black points
 - Minimize the **SSE** for regression trees (predicting numerical variables)
- Well-implemented in packages
 - Do **not** have to worry about it and do **not** have control over it
 - We can and have to control when CART should stop splitting

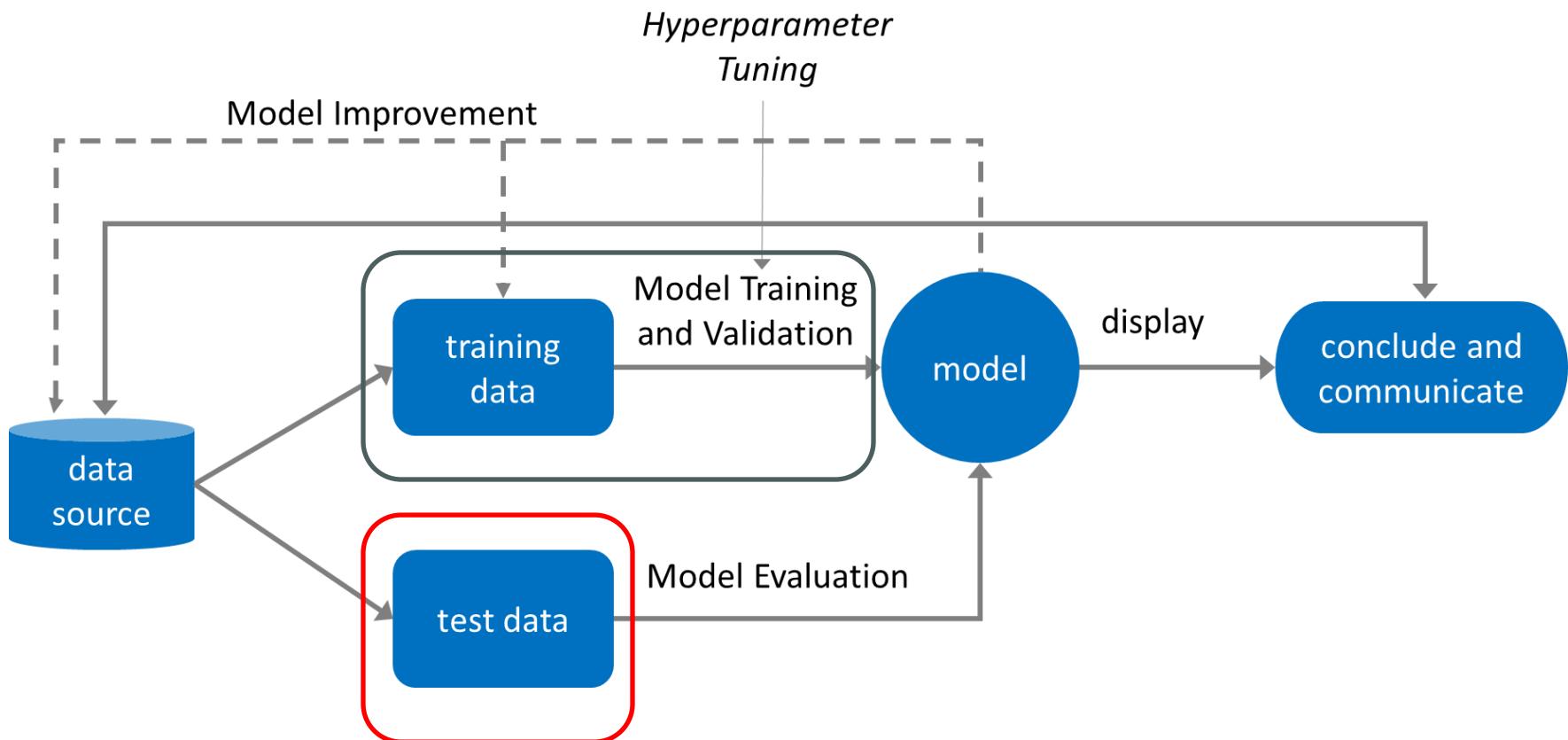
When Does CART Stop Splitting?

- We can use various parameters to control how many splits are generated, e.g.,
 - ***minbucket***: the minimum number of observations in any terminal node
 - ***minsplit***: the minimum number of observations that must exist in a node in order for a split to be attempted
 - ***maxdepth***: the maximum depth of any node of the final tree
 - ***cp***: the penalty for each additional split

Class Outline

- Review: Quality of Care and CART
- Model Selection: Cross-Validation
- Advanced Decision Trees
- Case Study Continued: Predicting Quality of Diabetes Care Using Trees and Random Forests

Generic ML Workflow



Model (Variable) Selection

- How can we choose the best model?
 - For linear and logistic regressions, candidate models are represented by different **sets of independent variables**
 - That is why we also call the process ***variable selection***
- *Can we choose the model with the largest R^2 or AUC in the training set?*
 - **NO**, because of overfitting
- *Can we choose the model with the largest R^2 or AUC in the test set? **NO!***

Model Selection

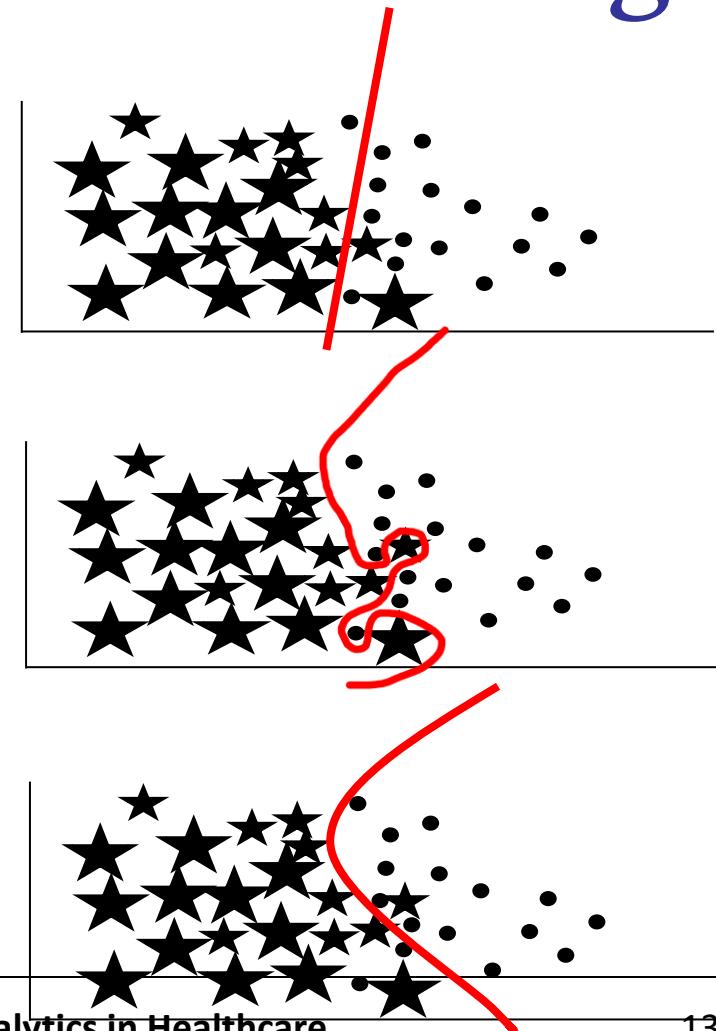
- Choosing the model with the largest R^2 or AUC in the **training set**
 - **Overfitting**
- Choosing the model with the largest R^2 or AUC in the **test set**
 - Implicitly using test set to **build** the model
 - The test dataset **should never be explicitly or implicitly used in building or finetuning the model**
 - Test set can only be used to evaluate the model **after** model is finalized
 - Test data set is the “**future**”
 - Final **reporting**
- Then what can we do with the training set?

Overfitting and Underfitting

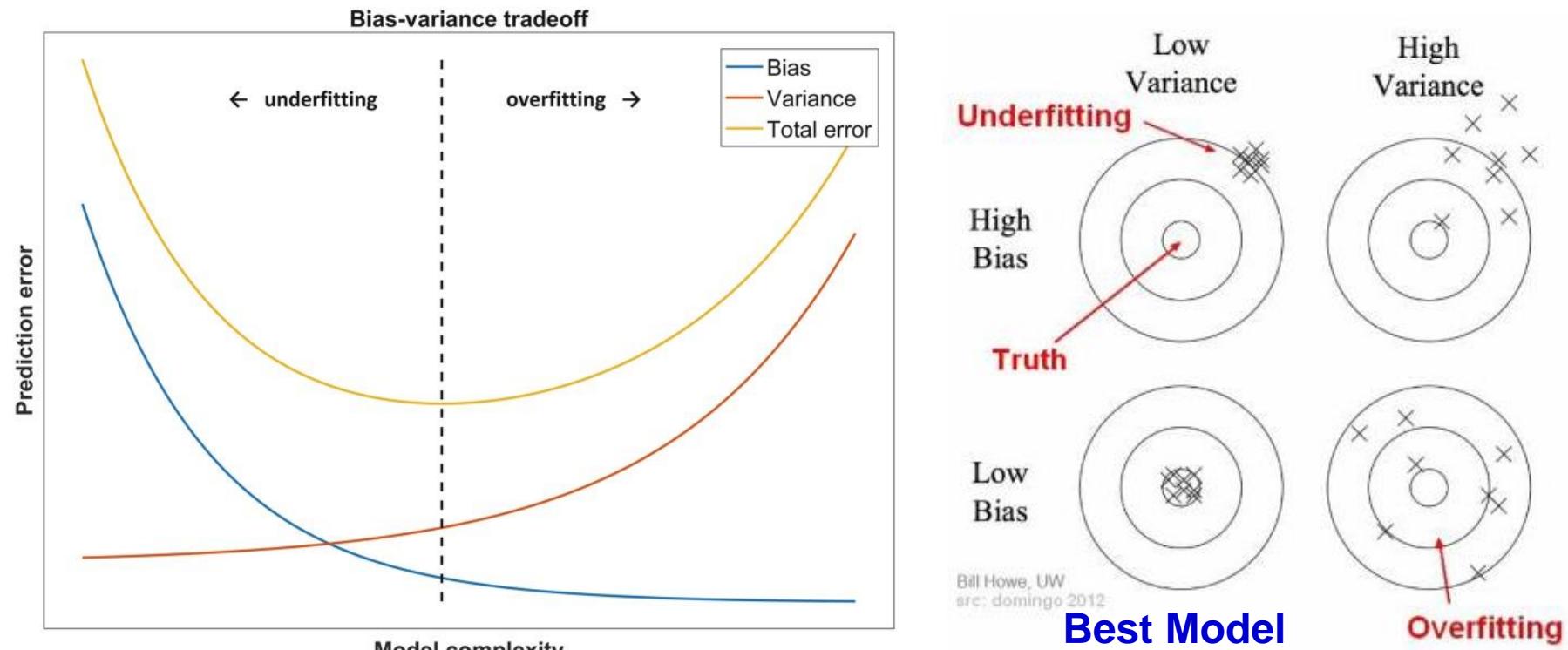
Underfitting: Model too simple, **fewer features, smaller weights,** weak learning.

Overfitting: Model too complex, **too many features, larger weights,** weak generalization.

'Good Fit' Model: Compromise between **fit** and **complexity**



Concept: Bias–Variance Tradeoffs



Mathematical Representation (not tested)

$$Err(x) = \left(E[\hat{f}(x)] - f(x) \right)^2 + E \left[\left(\hat{f}(x) - E[\hat{f}(x)] \right)^2 \right] + \sigma_e^2$$

Training, Validation, and Test

- We have seen examples of building models with the ***training*** set and then checking the performance on the ***test*** dataset

Training

Test

- To unbiasedly evaluate a model fit on the training dataset while **tuning the model's parameters**, we can divide the original training dataset into ***training*** and ***validation*** datasets
- The validation datasets provide the out-of-sample performance to help finetune the model's parameters

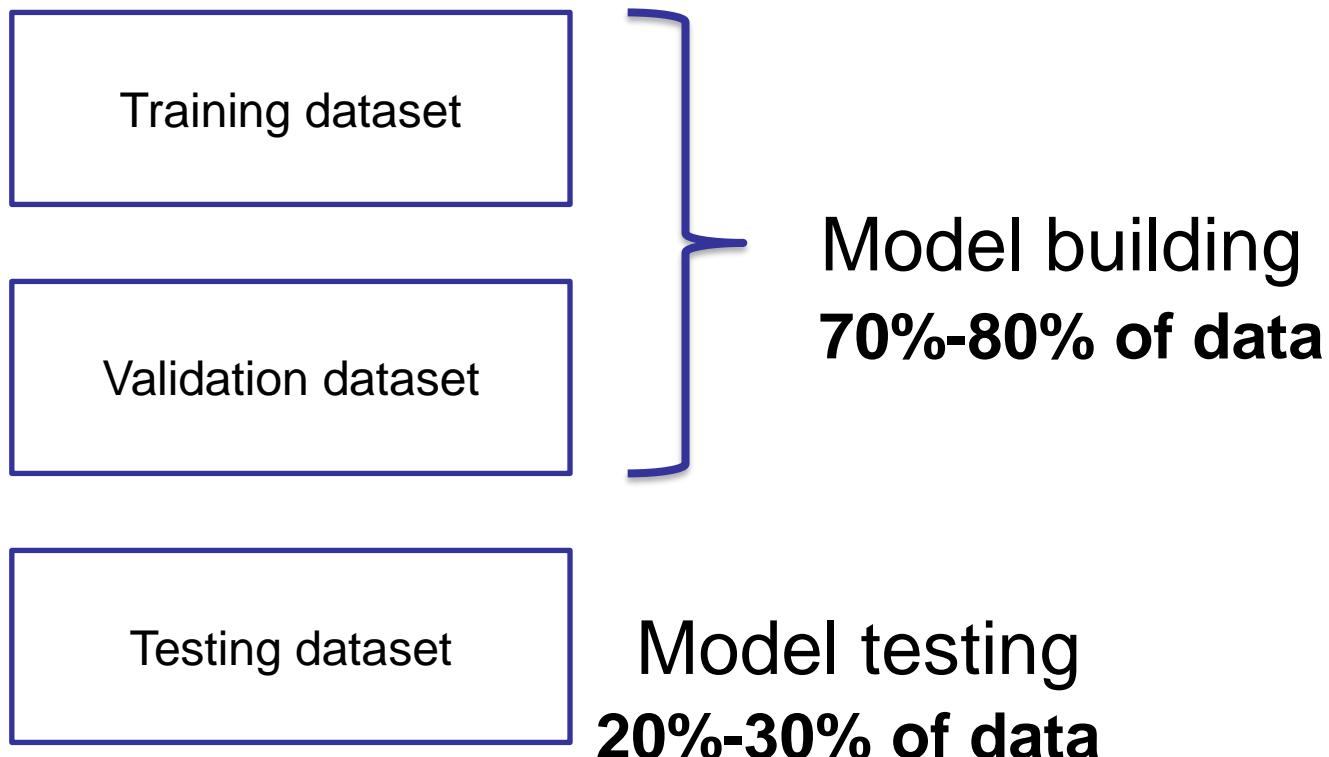
Training

Validation

Test

- This is called ***cross-validation***

Training, Validation and Test Data

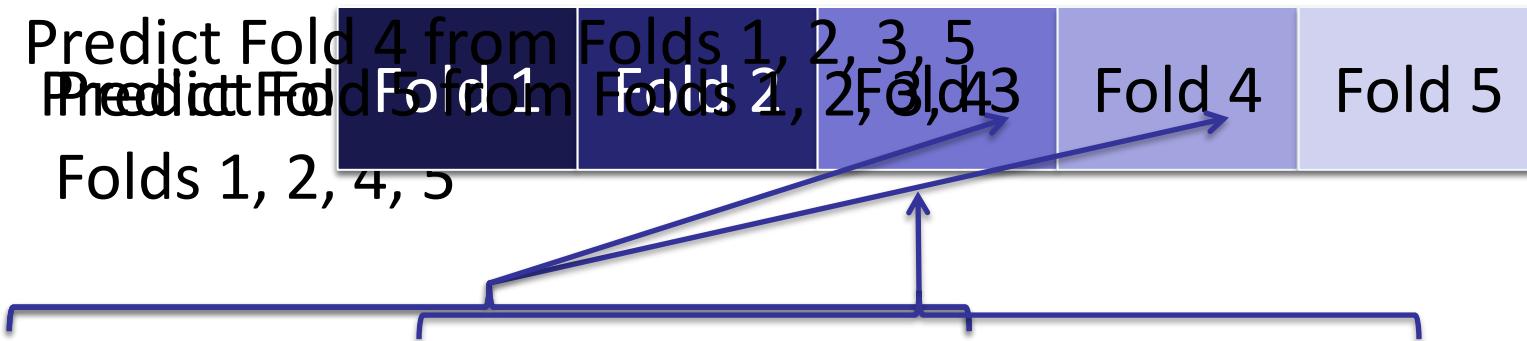


k-fold Cross-Validation

- One commonly used cross-validation procedures is *k-fold cross-validation*
 - Given the **training** set, split into k pieces (“folds”)
 - Use $(k - 1)$ folds to estimate a model, and test model on remaining one fold (which acts as a validation set) for each candidate model
 - Repeat for each of the k folds
 - For each candidate model, average performance over the k folds, or validation sets

k-fold Cross-Validation Graphically

- Assume five folds ($k = 5$)



- Continue to predict Fold 2 and Fold 1...

Output of k -fold Cross-Validation

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Model 1	0.72	0.61	0.68	0.59	0.75	0.67
Model 2
Model 3
...

- List the candidate models that you want to consider
- Use k -fold cross-validation to obtain their performance (e.g., AUC) and fill up the above table
 - Most packages will directly report the last column
- Choose the model with the largest average AUC

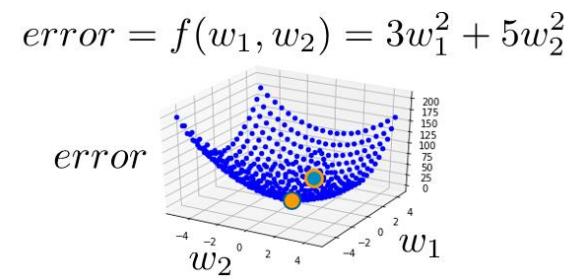
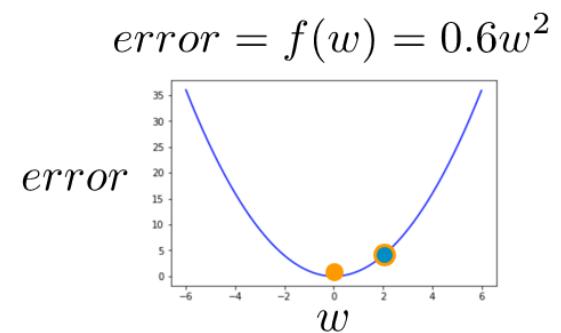
k-fold Cross-Validation for CART

Parameter	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
1	0.72	0.61	0.68	0.59	0.75	0.67
2
3
...

- For linear or logistic regression, candidate models are represented by different sets of independent variables
- For CART, they are indexed (ordered) by parameter values

ML Optimization

- In ML, use **optimization** to **minimize** an **error function** of the **ML model**
 - **Loss function:** $error = f(w)$
 - **Optimizing the error function:**
 - **Minimizing** means **finding** the hyperparameters w that results in the **lowest error**
 - **Maximizing** means finding w that gives the **largest error**



Reference: AWS ML University

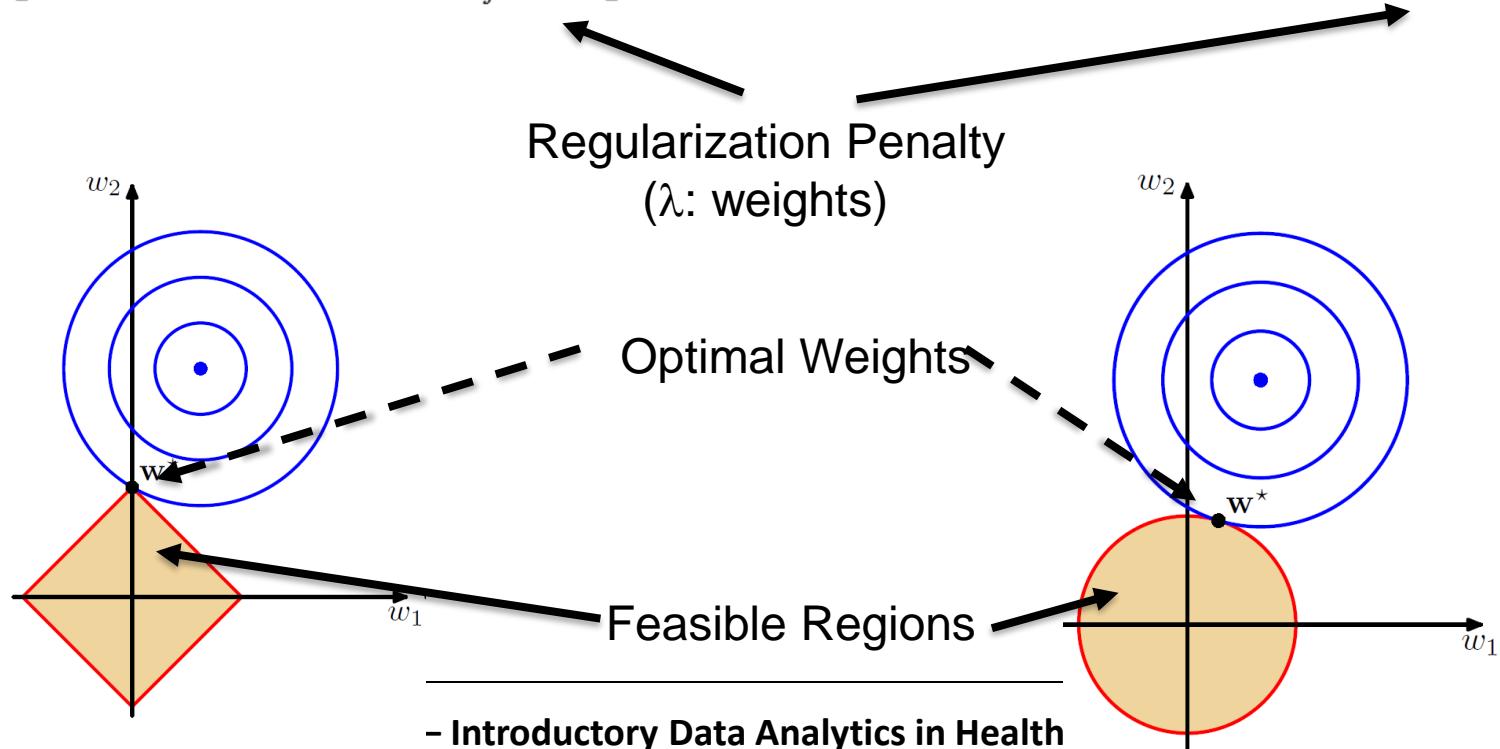
Regularization \rightarrow Parsimony

- Lasso Regularization [L1]

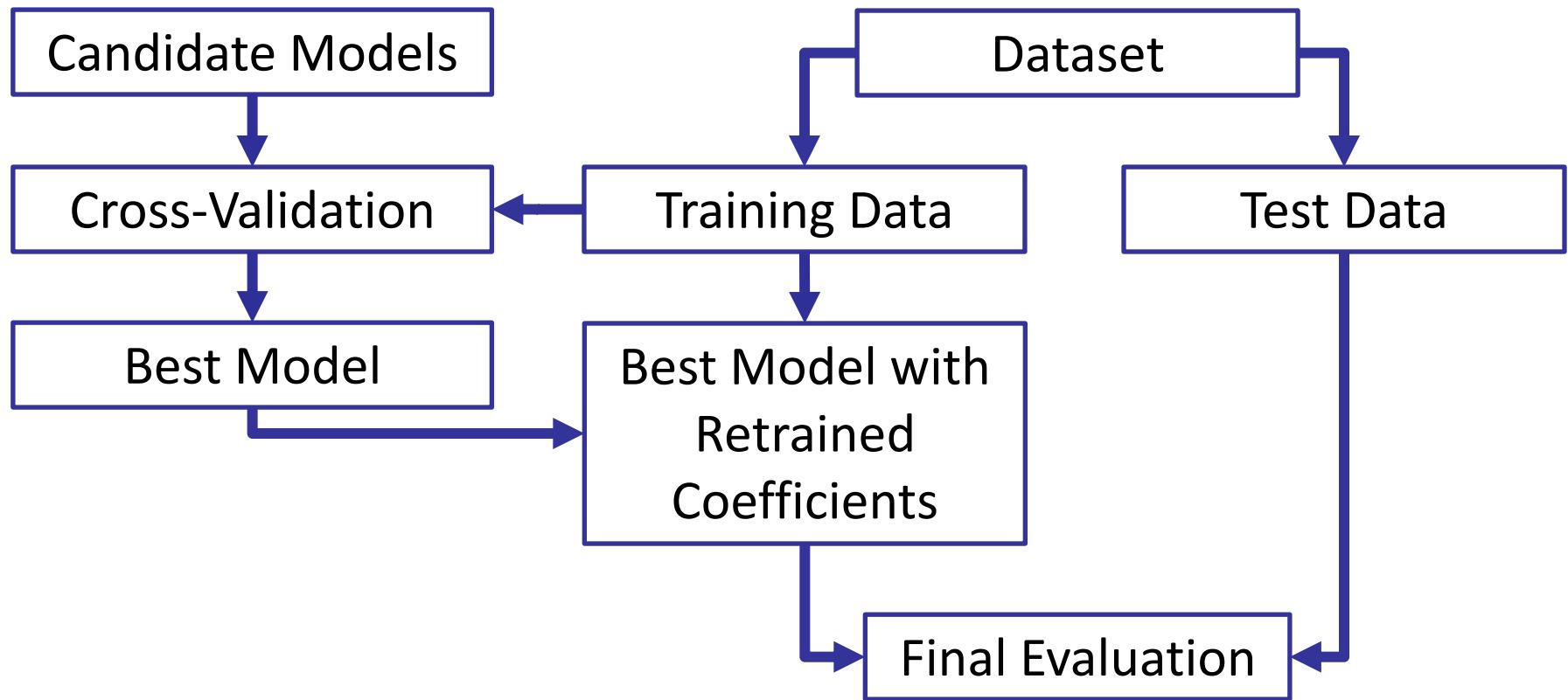
$$\min_{\mathbf{w}} \left[\frac{1}{2} \sum_{n=1}^N (t_n - \mathbf{w}^T \phi(\mathbf{x}_n))^2 + \frac{\lambda}{2} \sum_{j=1}^M |w_j| \right]$$

- Ridge Regularization [L2]

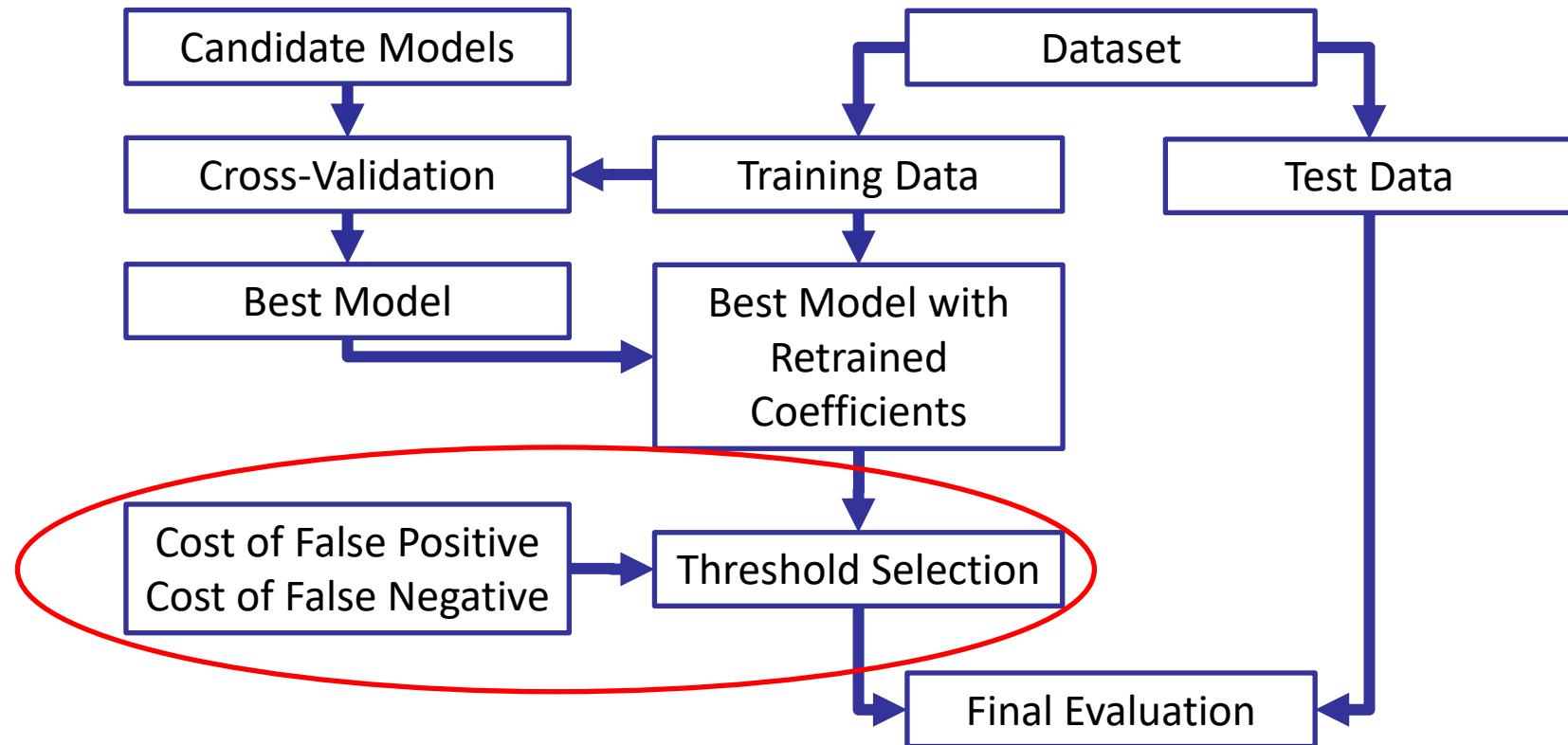
$$\min_{\mathbf{w}} \left[\frac{1}{2} \sum_{n=1}^N (t_n - \mathbf{w}^T \phi(\mathbf{x}_n))^2 + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} \right]$$



k-fold Cross-Validation to determine weights

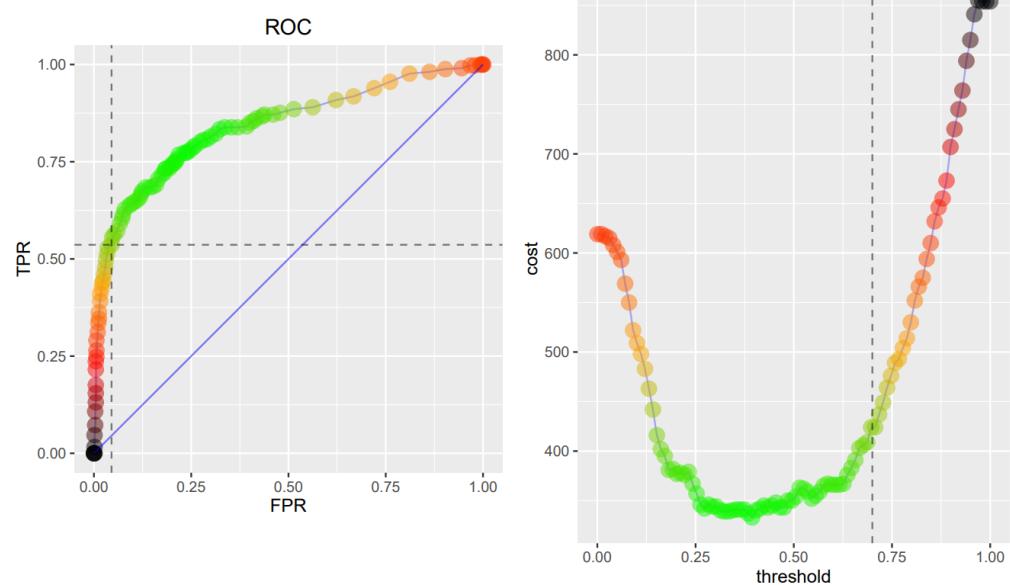


Threshold Selection for Classification Problems



Threshold Selection

- AUROC offers a global measure across all different thresholds
- Optimal threshold requires the definition of Cost of False Positive vs Cost of False Negative
- Requires a “Cost Function”



More About k -fold Cross-Validation

- (k -fold) cross-validation can be applied to any predictive (machine learning) models for model selection
- How can we set the value for k ?
 - “The choice of k is usually 5 or 10, but there is **no formal rule**. As k gets larger, the difference in size between the training set and the resampling subsets gets smaller. As this difference decreases, the bias of the technique becomes smaller” — Max Kuhn, Kjell Johnson (2013) *Applied Predictive Modeling*
 - Computational effort is usually a big concern
 - Leave-one-out cross-validation (LOOCV)
 - Extreme case when k is chosen to be the number of data points in the training set

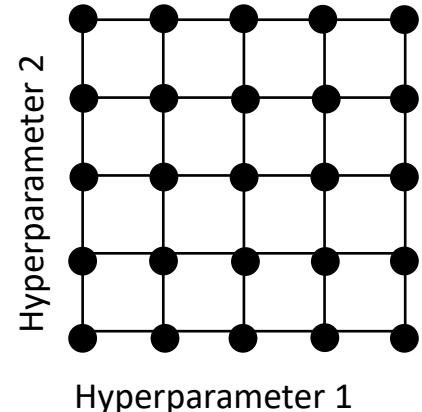
Cross-Validation for CART in R

- When we carry out the cross-validation for CART in R, we often use “*cp*” instead of “*minbucket*”
- Normally, we do not have to test the full range of *cp* parameter from 0 to 1
- Beware that the training function in R uses average accuracy (assuming a threshold of 0.5) by default to select the best *cp* parameter
 - Sometimes not recommended

Hyperparameter Tuning

Grid Search

- **Basic hyperparameter tuning method**, finds the optimum **combination of hyperparameters** by **exhaustive search** over **specified parameter values**

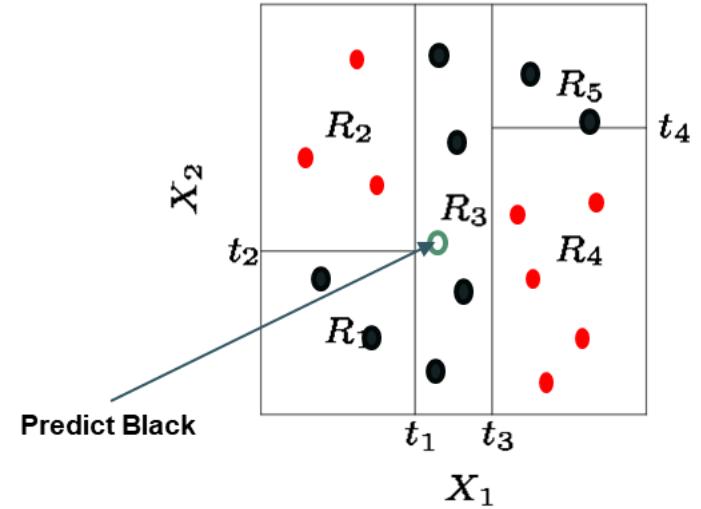
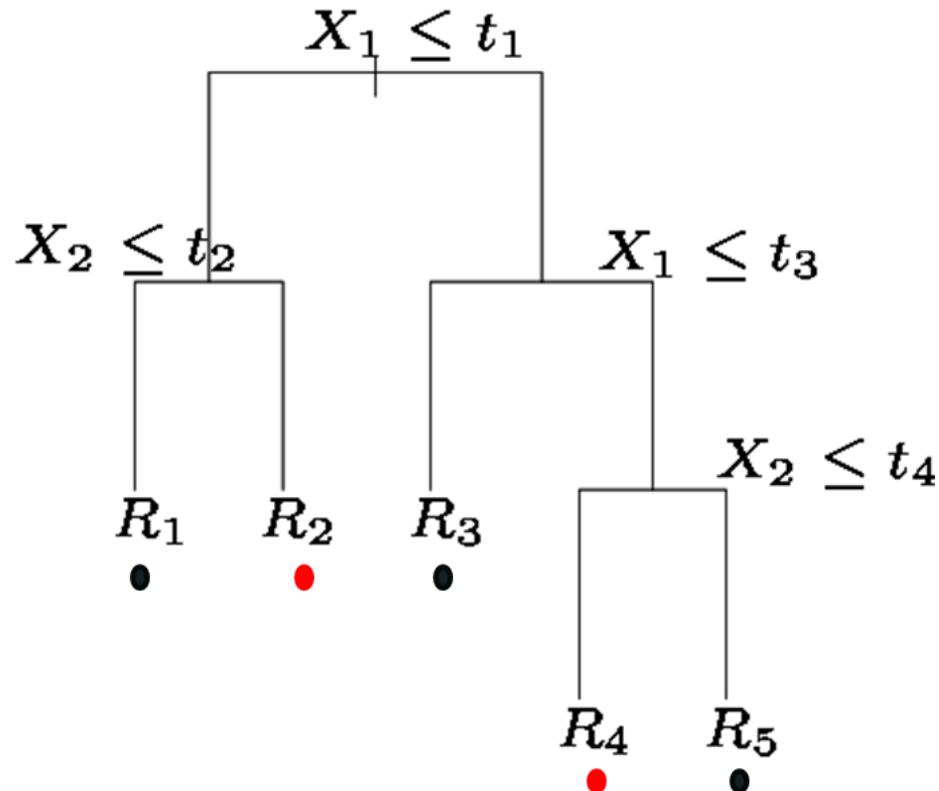


```
• fitControlAcc <- trainControl(method = "cv", number = 5)  
• cpGridAcc <- expand.grid(.cp = (1:50)*0.005)  
• cvResultsAcc <- train(PoorCare ~ . - MemberID, data = train, method =  
  "rpart", trControl = fitControlAcc, tuneGrid = cpGridAcc)
```

Class Outline

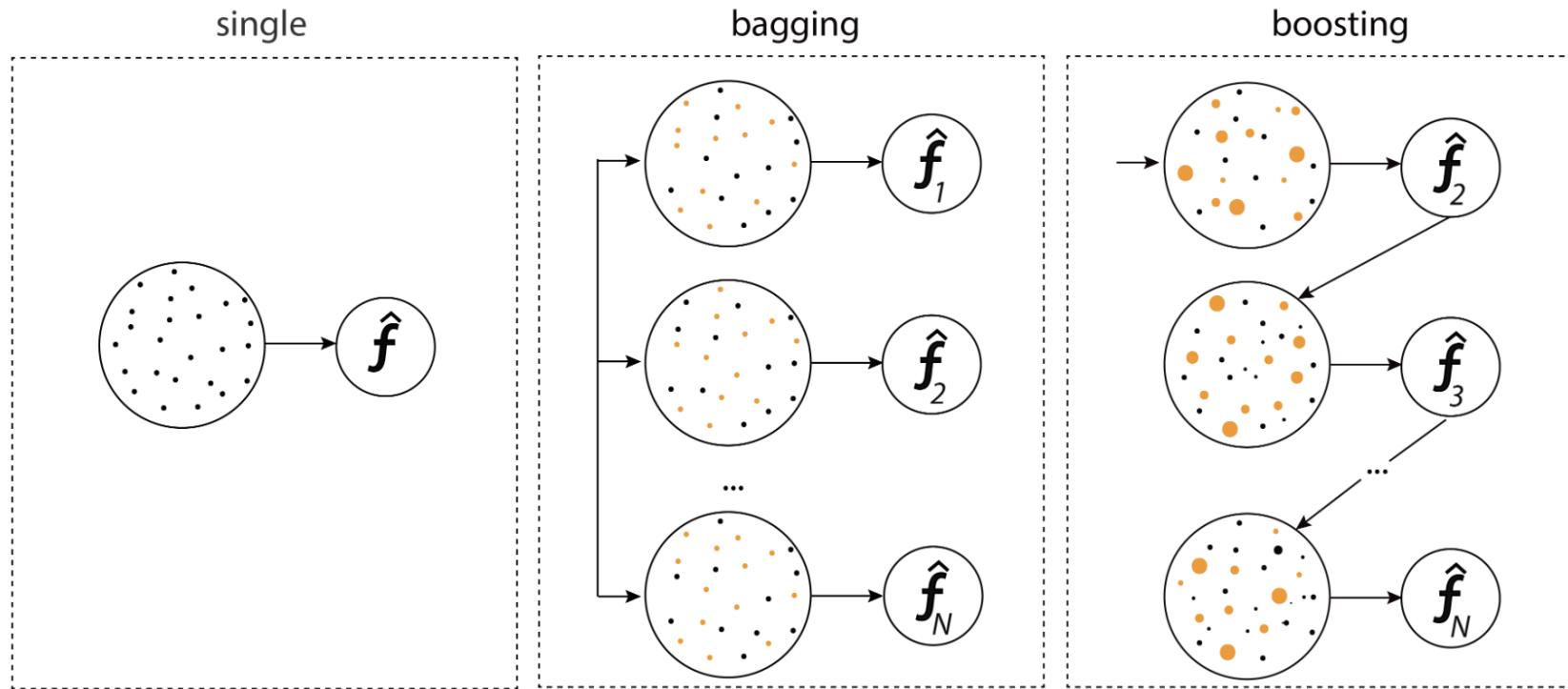
- Review: Quality of Care and CART
- Cross-Validation
- Advanced Decision Trees
- Case Study Continued: Predicting Quality of Diabetes Care Using Trees and Random Forests

Decision Tree (DT)



- With DTs, we can create partitions using a tree structure.
- Non-linear boundaries can reduce the errors in predictions
- Provides a very simple way to explain the model to a non-expert

Advanced Decision Trees



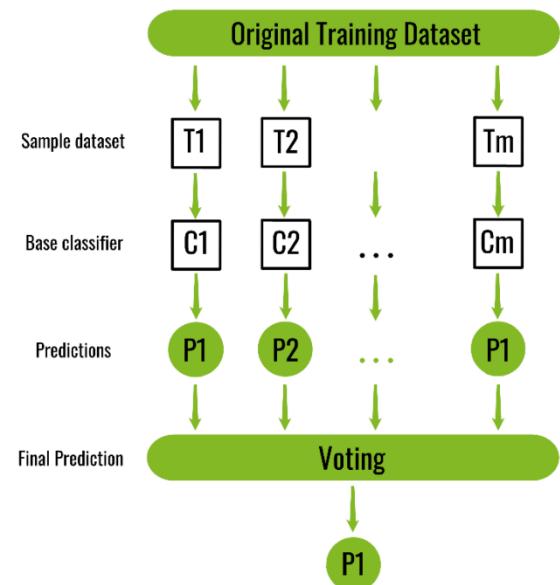
<https://incredible.ai/machine-learning/2015/10/25/Bias-Variance-Tradeoff-Ensemble/>

Building Many Trees - Bagging

- A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on **random subsets** of the original dataset and then **aggregate their individual predictions – bootstrap technique**
 - Select observations randomly with replacement
 - Example – Original data: 1 2 3 4 5 6 7 8 9
 - New “data” with samples of 5 are:
 - 2 3 1 2 5 → 1st tree
 - 3 1 4 5 1 → 2nd tree
 - 4 4 2 1 5 → 3rd tree
 - ⋮ ⋮
- 
- Build **independent** estimators of same type on each subset
 - Majority vote or average the predictions from all estimators

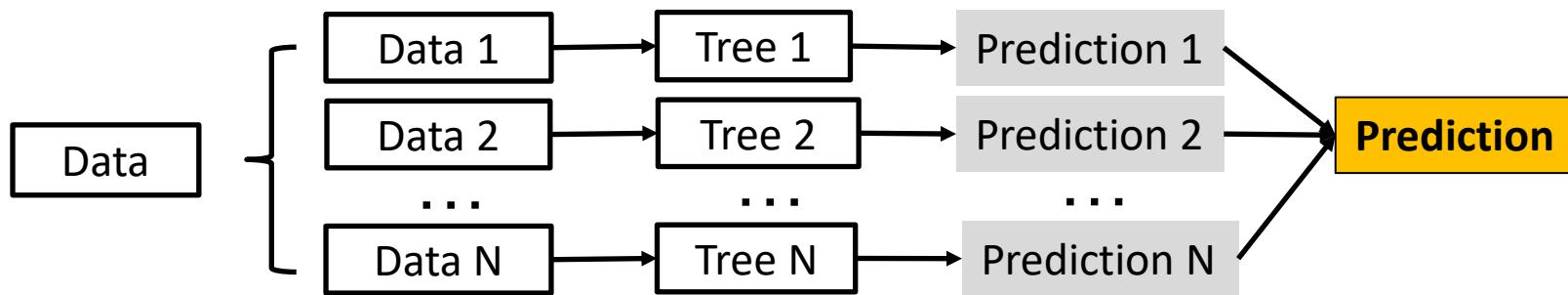
Building Many Trees - Bagging

- A way to reduce the variance of a black-box estimator (e.g., a decision tree), by **introducing randomization into its construction procedure**
- Each time a split is going to be generated in any tree, not all variables are considered
- A **random set of independent variables** is chosen as split candidates
 - Models are trained in parallel
 - Aggregation (either by voting or by averaging) to form a final prediction
- **Trees can be different from each other** in a random forest since they are built on **different data sets** and the splits are based on different candidate independent variables



Random Forests (Intelligent Bagging)

- Random forest is an extension of bagging that also randomly selects **subsets of features** used in each data sample.
- Designed to improve prediction accuracy of CART
- Works by building a large number of CART trees
 - Cons: Less interpretable
- To make a prediction for a new observation, each tree makes its own prediction, and we aggregate all the predictions by majority voting or averaging

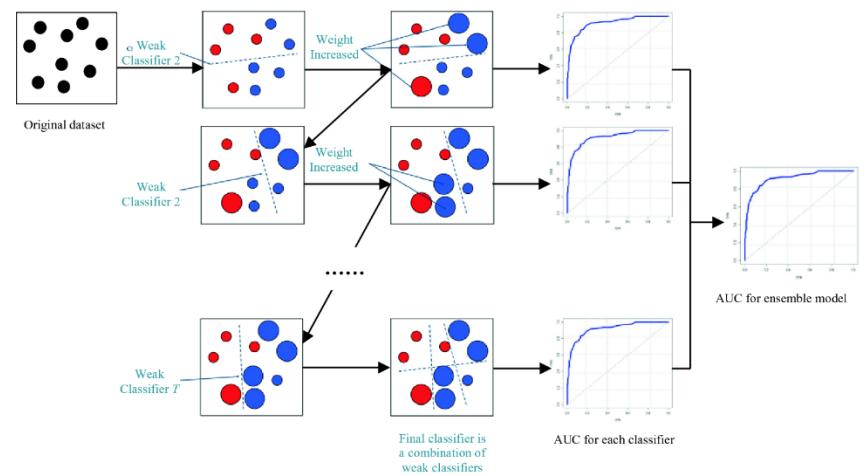
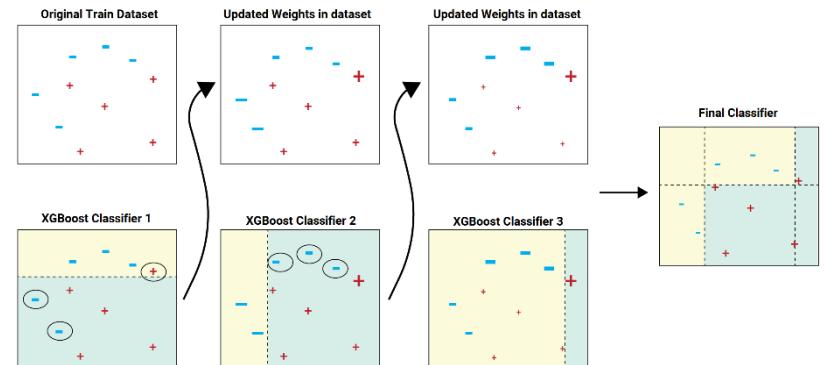


Random Forest Parameters

- Minimum number of observations in a leaf node
 - In R, this is controlled by the “*nodesize*” parameter
- Number of trees
 - In R, this is the “*ntree*” parameter
 - Should not be too small, because bagging procedure may miss observations
 - More trees take longer to build
- Number of variables available for splitting at each tree node
 - In R, this is the “*mtry*” parameter
- You can also use cross-validation to select the parameters for random forest models

Another Method - Boosting

- Boosting is a general learning paradigm for **putting together a Strong Learner, given a collection (possibly infinite) of Weak Learners.**
- Bagging control for high variance whereas **Boosting controls both bias and variance**
- Common types of Boosting algorithm:
 - Adaptive Boosting (AdaBoost)
 - Gradient Boosting
 - Extreme Gradient Boosting (XG Boost)
 - CatBoost
 - Light GBM



Boosting vs Bagging

Bagging	Boosting
	<ul style="list-style-type: none">• Both are ensemble techniques• Both combine models of the same Classifier type
Reduces variance	Reduces bias and variance
Models are built in parallel	Each new model is influenced by the performance of previous models
Parallel ensemble	Sequential ensemble
Equal weight is given to all models	Weight models contribution by performance

David Opitz and Richard Maclin. 1999. Popular ensemble methods: an empirical study. *J. Artif. Int. Res.* 11, 1 (July 1999), 169–198.
Odeguia, Rising. (2019). An Empirical Study of Ensemble Techniques (Bagging, Boosting and Stacking).

Prediction of CKD

Ilyas et al. BMC Nephrology (2021) 22:273
<https://doi.org/10.1186/s12882-021-02474-z>

RESEARCH ARTICLE

Open Access

Chronic kidney disease diagnosis using decision tree algorithms

Hamida Ilyas^{1,2}, Sajid Ali^{1,2,3}, Mahvish Ponum^{1*}, Osman Hasan¹, Muhammad Tahir Mahmood^{1,4}, Mehwish Iftikhar^{1,5} and Mubasher Hussain Malik^{1,2}



Early detection and cure of CKD is extremely desirable as it can lead to the prevention of unwanted consequences.

... this study is significant, as not a single previous research is conducted to detect the stages of CKD using age, sex, race and Serum Creatinine attributes.

In this study, we **focus on using two machine learning algorithms i.e. J48 and Random Forest, to predict the stages of CKD.**

Our study reveals more accurate results than most of the existing studies, i.e., we **achieved 85.5% accuracy using the J48 algorithm within 0.03 s and 78.25% accuracy using the random forest algorithm within 0.28 s.**

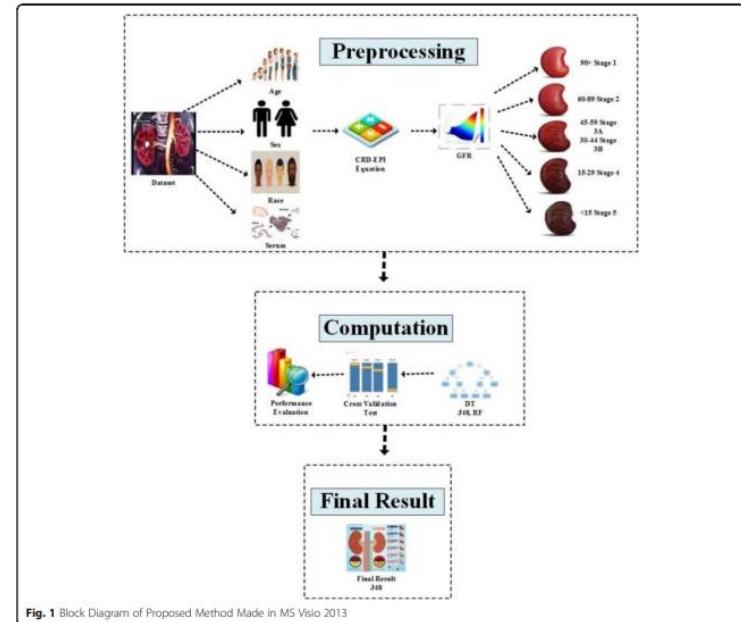


Fig. 1 Block Diagram of Proposed Method Made in MS Visio 2013

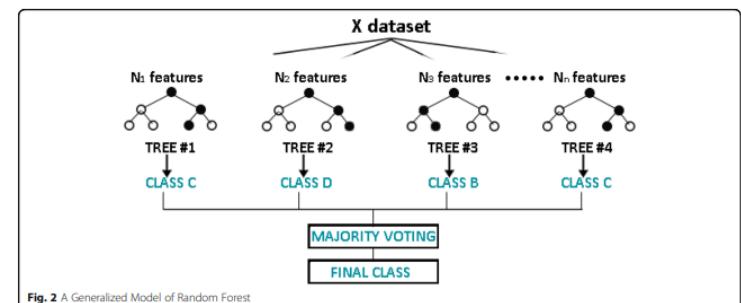


Fig. 2 A Generalized Model of Random Forest

Prediction of Hospitalization

JMIR MEDICAL INFORMATICS

Luo et al

Original Paper

Developing a Model to Predict Hospital Encounters for Asthma in Asthmatic Patients: Secondary Analysis

Gang Luo¹, DPhil; Shan He², DPhil; Bryan L Stone³, MSc, MD; Flory L Nkoy³, MPH, MSc, MD; Michael D Johnson³, MD

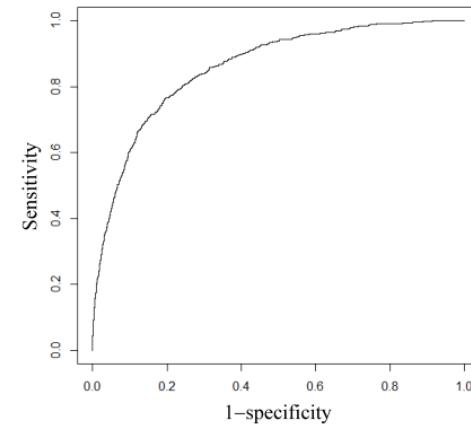
¹Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, United States

²Care Transformation, Intermountain Healthcare, Salt Lake City, UT, United States

³Department of Pediatrics, University of Utah, Salt Lake City, UT, United States

... we considered **235 candidate features** derived from the structured attributes in our dataset. These features came from 4 sources: the >100 potential risk factors for **asthma exacerbations reported in the literature** [9,22,27-34]; features used in the **existing models for predicting asthma exacerbations** [9-22]; factors impacting patients' **general health status mentioned in the literature** [31,35,36]; and **features suggested by the clinical experts in our team**.

ROC Curve from the model



XGBoost compared with existing models

Table 6. A comparison of our final model and multiple prior models for predicting inpatient stays and emergency department visits in asthmatic patients.

Model	Prediction target	Classification algorithm	Features used in the model, n	Data instances, n	Area under the receiver operating characteristic curve	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)
Our final model	Hospital encounters for asthma	Extreme gradient boosting	142	334,564	0.859	53.69	91.93	22.65	97.83
Loymans et al [10]	Asthma exacerbation	Logistic regression	7	611	0.8	— ^a	—	—	—
Schatz et al [11]	Inpatient stay for asthma in children	Logistic regression	5	4197	0.781	43.9	89.8	5.6	99.1
Schatz et al [11]	Inpatient stay for asthma in adults	Logistic regression	3	6904	0.712	44.9	87.0	3.9	99.3
Eisner et al [12]	Inpatient stay for asthma	Logistic regression	1	2858	0.689	—	—	—	—
Eisner et al [12]	ED ^b visit for asthma	Logistic regression	3	2415	0.751	—	—	—	—
Sato et al [13]	Severe asthma exacerbation	Classification and regression tree	3	78	0.625	—	—	—	—
Miller et al [15]	Hospital encounters for asthma	Logistic regression	17	2821	0.81	—	—	—	—
Yerk et al [15]	Hospital encounters for asthma	Logistic regression	11	4888	0.78	77	63	82	56

Class Outline

- Review: Quality of Care and CART
- Cross-Validation
- Advanced Decision Trees
- Case Study Continued: Predicting Quality of Diabetes Care Using Trees and Random Forests

Let's get our hands dirty!

Data Analysis

Let's do some basic data analysis using our WHO data.

[WHO\\$Under15](#) Hide

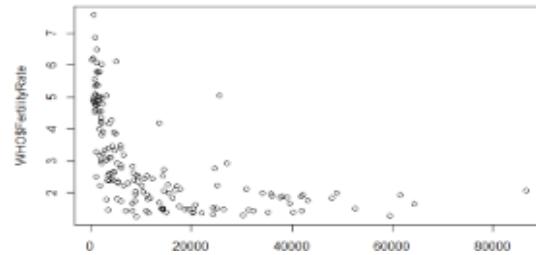
```
[1] 47.42 21.33 27.42 15.28 47.58 25.96 24.42 28.34 18.95 14.51 22.25 21.62 28.16 30.57 18.99 15.18 16.88 34.4
8 42.95 28.53
[21] 35.23 16.35 33.75 24.56 25.75 13.53 45.66 44.20 31.23 43.08 16.37 38.17 48.87 48.52 21.38 17.95 28.03 42.1
7 42.37 38.61
[41] 23.94 41.48 14.98 16.58 17.16 14.56 21.98 45.11 17.66 33.72 25.96 38.53 38.29 31.25 38.62 38.95 43.18 15.6
9 43.29 28.88
[61] 16.42 18.26 38.49 45.98 17.62 13.17 38.59 14.68 26.96 48.88 42.46 41.55 36.77 35.35 35.72 14.63 28.71 29.4
3 29.27 23.68
[81] 48.51 21.54 27.53 14.04 27.78 13.12 34.13 25.46 42.37 38.18 24.98 38.21 35.61 14.57 21.64 36.75 43.06 29.4
5 15.13 17.46
[101] 42.72 45.64 26.65 29.03 47.14 14.98 38.18 48.22 28.17 29.82 35.81 18.26 27.85 19.81 27.85 45.38 25.28 36.5
9 38.18 35.58
[121] 17.21 20.26 33.37 49.99 44.23 38.61 18.64 24.19 34.31 38.18 28.65 38.37 32.78 29.18 34.53 14.91 14.92 13.2
8 15.25 16.52
[141] 15.85 15.45 43.56 25.96 24.31 25.78 37.88 14.04 41.68 29.69 43.54 16.45 21.95 41.74 16.48 15.08 14.16 40.3
7 47.35 29.53
[161] 42.28 15.28 25.15 41.48 27.83 38.85 16.71 14.79 35.35 35.75 18.47 16.89 46.33 41.89 37.33 28.73 23.22 26.8
0 28.05 38.61
[181] 48.54 14.18 14.41 17.54 44.85 19.63 22.05 28.98 37.37 28.84 22.87 48.72 46.73 40.24
```

[WHO\\$Country\[which.min\(WHO\\$Under15\)\]](#) Hide

```
[1] Japan
194 Levels: Afghanistan Albania Algeria Andorra Angola Antigua and Barbuda Argentina Armenia Australia Austria
... Zimbabwe
```

Let's create some plots for exploratory data analysis (EDA). First, let's create a basic scatterplot of GNI versus FertilityRate.

[plot\(WHO\\$GNI, WHO\\$FertilityRate\)](#) Hide



Troubleshooting

You may need to additionally install and load the following packages for cross-validation to work on your computer: "class" and "ggplot2". If you receive an error message after trying to load “caret” and “e1071”, please try installing and loading these two additional packages.

Troubleshooting

When you try to install the package “randomForest”, if you get an installation warning that says:

"Warning: cannot remove prior installation of packages 'randomForest'"

please try quitting and re-starting R.

Expert vs. Analytics

- Expert judgement is the gold standard
- Models do not replace expert judgment
 - Experts can improve and refine models
- Advantages of models
 - Monitor large patient population in real time
 - Reveal important insights by learning from experts
 - Can integrate assessments of many experts into one final unbiased and unemotional prediction

Model-based Thinking

“The most that can be expected from any model is that it can supply a **useful** approximation to reality:

All models are wrong; some models are useful.”



George E. P. Box (1919-2013)
One of the great statistical minds of the 20th century

Class Outline

- Review: Quality of Care and CART
- Cross-Validation
- Advanced Decision Trees
- Case Study Continued: Predicting Quality of Diabetes Care Using Trees and Random Forests

End

ECON 145

ECON 145 – Introductory Data Analytics in Healthcare

Lecture 11: Healthcare Operational Analytics

Class Outline

- Healthcare Operations Management
- Reflections on Healthcare Operations Management
- Examination Briefing

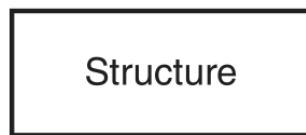
What is Healthcare Operations Management (HOM)?

- Usually refer to administrative, financial and legal activities that responsible for delivering healthcare *services*
- Basically one “example” of OM
 - Other services like education, banking, entertainment, hospitality, etc.
- Inputs include capital, materials, equipment, facilities, suppliers, labor, knowledge, time

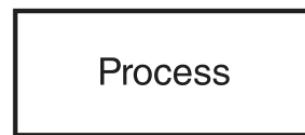
Framework for Healthcare

- Donabedian Framework

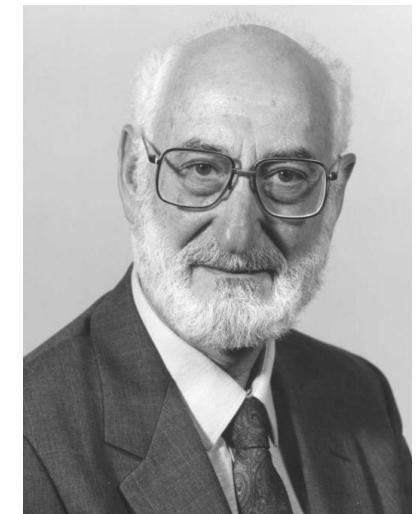
Attributes of the setting in which care is delivered.



What healthcare-related activities are performed.



Effects of care on the health status of patients and populations



Avedis Donabedian
(1919 – 2000)

What is Operations Management (OM)?

- **Production** is the creation of *goods* and *services*; however, has been used primarily to refer to creation of tangible goods
- **Operations** is the term introduced to re-emphasize creation of both goods and services
- **Operations management** is planning, organizing, staffing, directing, and controlling of activities that creates goods and services through the transformation of *inputs* into *outputs*
- **OM** refers to both manufacturing and services

Processes

- Production of goods or services in a sequence of activities performed by a set of specialized resources

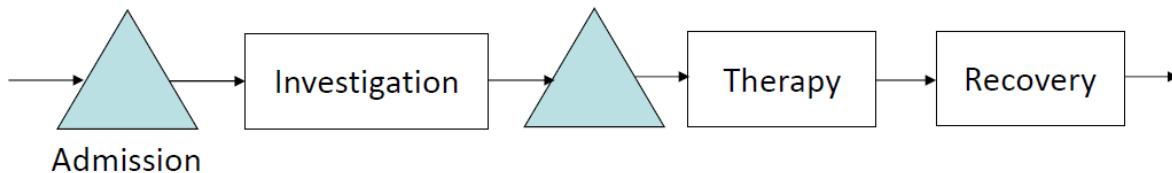


Figure 4. Steps taken to discharge a patient

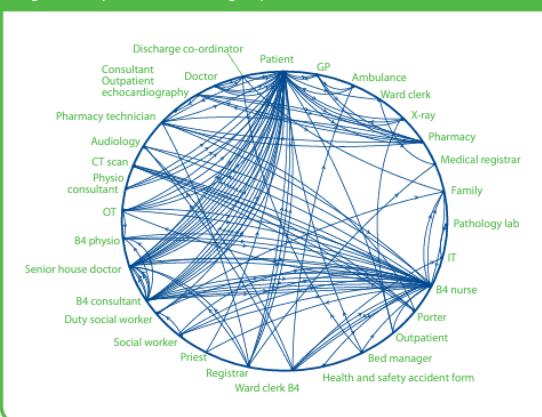
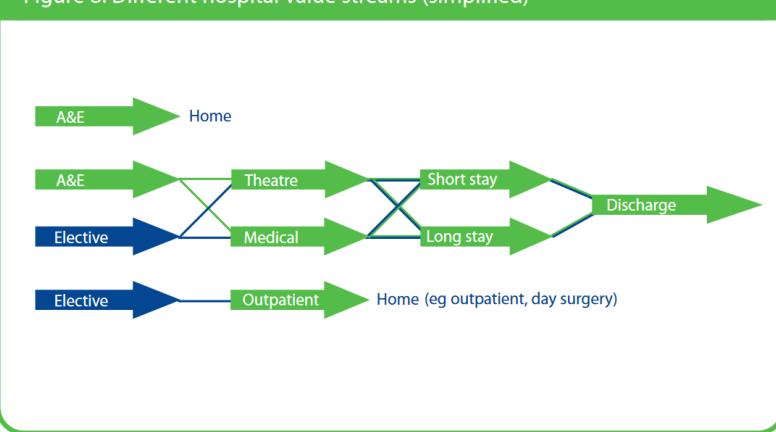


Figure 8. Different hospital value streams (simplified)



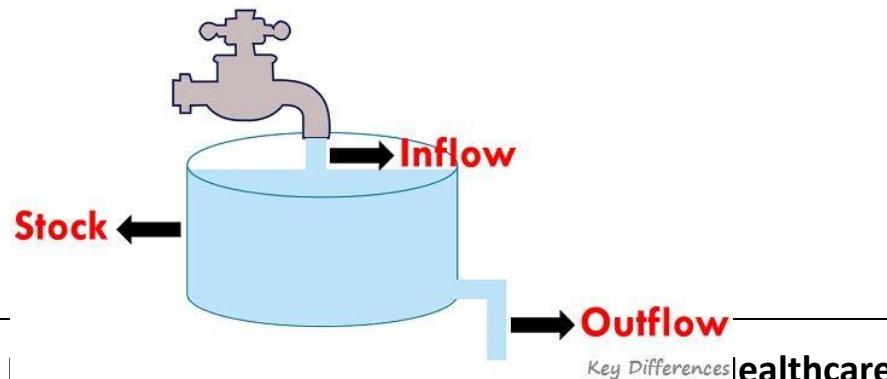
Process Analysis

In any STABLE system,

- *Average Arrival Rate = Average Departure Rate*

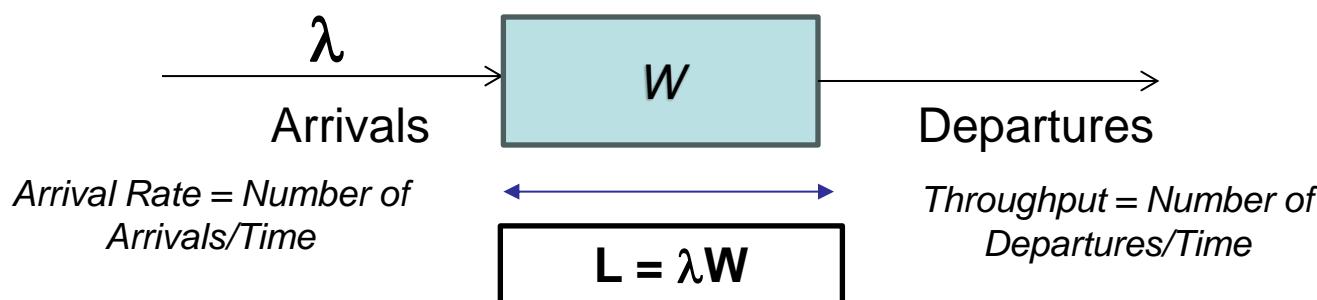
Flow balance assumption in a STABLE system:

- *Items cannot depart faster than they arrive*
- *Items cannot arrive faster than they depart*



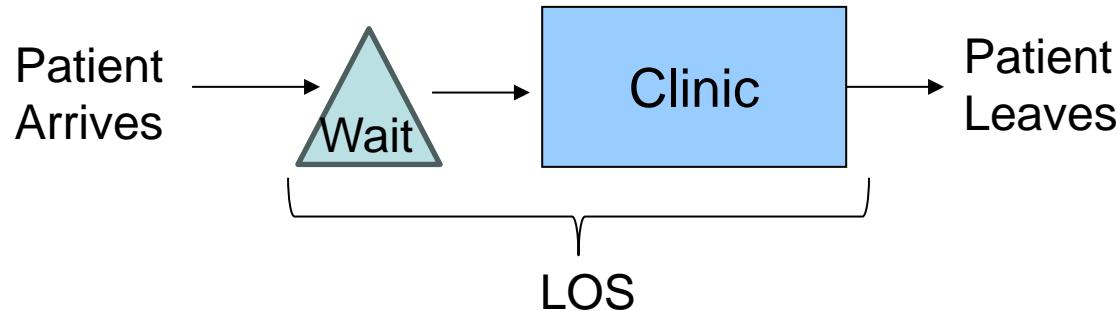
Little's Law

- $L = \text{Inventory}$ = how many flow units are in the process
- $\lambda = \text{Flow Rate}$ = rate at which flow units enter or leave the process
- $W = \text{Flow Time}$ = total time a flow unit is in the process



Little's Law

- For example:



- Little's Law:

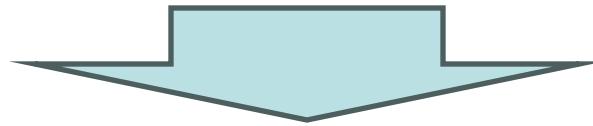
Avg Patients in System (L) =

Arrival/Discharge Rate (λ) X Avg Length of Stay (W)

Little's Law

Flow Rate (λ) = On average 11 patients arrive/discharge per day

LOS (W) = On average a patient spends 2 HOURS in the clinic (or 2/24 day)



Average number of patients in the clinic (L)
 $= \lambda \times W$
 $= 11 \times 2/24 = 0.917$ patients

Another Example

- SGH Statistics:

Key Figures

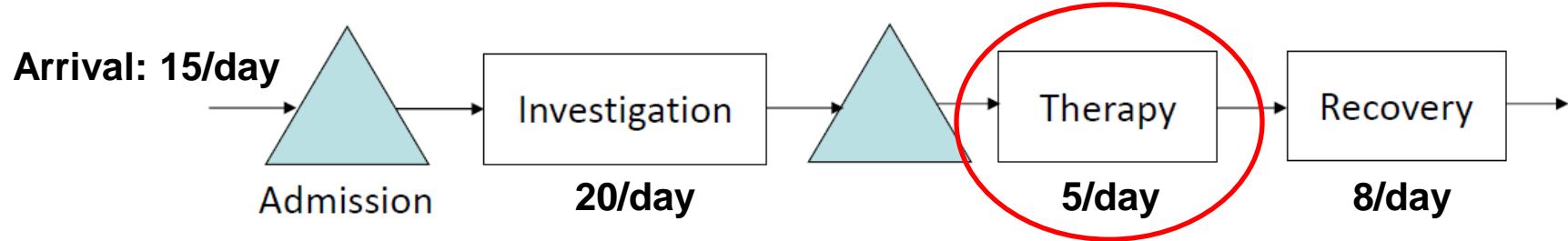
- Bed Distribution - 1,785
- Staff Strength - 9,888
- Patients Discharged - 80,817
- Specialist Outpatient Attendances - 724,480
- A&E Attendances - 128,660
- Inpatient & Electives Surgical Operations - 92,228

<https://www.sgh.com.sg/about-us/corporate-profile/hospital-overview-singapore-general-hospital>

- Beds: 1,785
- Bed Occupancy Rate: 80% (Assumed)
- Discharges: 80,817 per year

$$\begin{aligned} \text{LOS} &= \text{Patients in System/ Discharge Rate} \\ &= (1,785 \times 0.8) / (80,817/365) = 6.45 \text{ days/patient} \end{aligned}$$

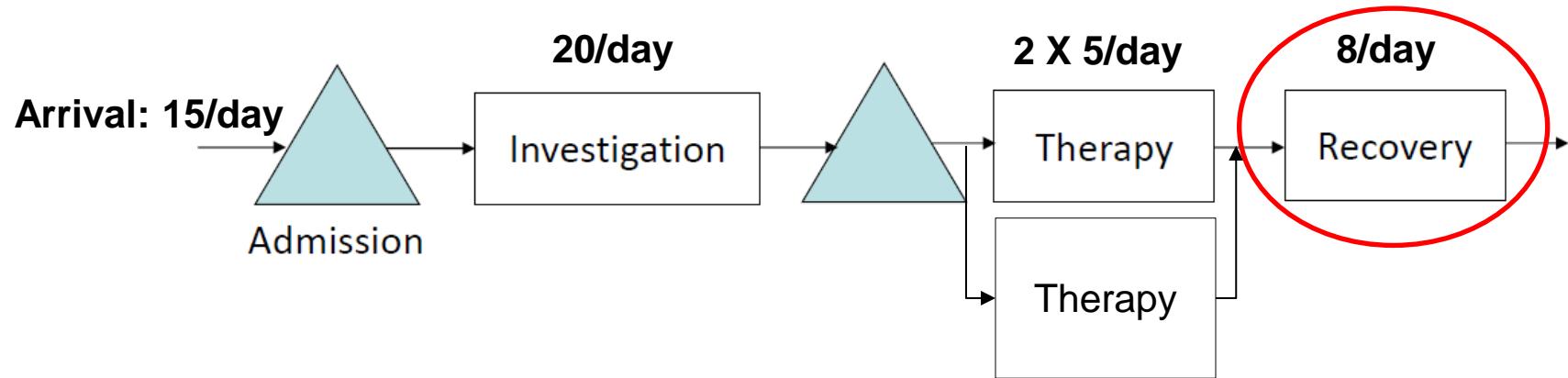
Bottleneck, Process Capacity and Flow Rate



- Process capacity = capacity of the bottleneck
- Flow Rate : Minimum {Process Capacity, Demand}
 - Process can be demand or capacity constrained
- Process utilization: Flow Rate/ Process Capacity

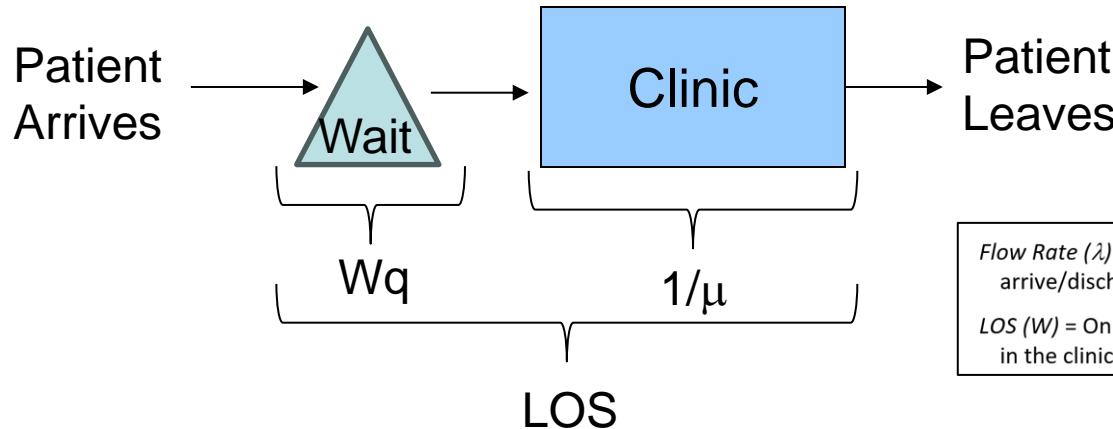
Bottleneck: resource with the lowest capacity

Shifting Bottleneck



- Bottleneck has shifted to “Recovery”
- Capacity = 8/day

Waiting Times



Flow Rate (λ) = On average 11 patients arrive/discharge per day

LOS (W) = On average a patient spends 2 HOURS in the clinic (or 2/24 day)

Service rate, μ = On average 15 patients can be served per day

Utilization rate = $\lambda / \mu = 11/15 = 73.3\%$

Average Waiting time in queue =

$$LOS - 1/\mu = 2/24 - 1/15 = 0.0167 \text{ day}$$

MANPOWER

					TOTAL 29,894
3,778	165	10,832	5,826	9,293	Year ended 31 March 2020

KEY FIGURES

Year ended 31 Mar
2020

ACUTE CARE AND POLYCLINICS

Size

Beds (As at end March) **4,699**

Workload per annum

Bed Occupancy Rate **81.9 %**

Inpatients **255,348**

Total Patient Days **1,217,413**

Average Length of Stay (days) **4.8**

Day Surgeries **182,612**

Inpatient Surgeries **105,533**

Specialist Outpatient Clinic Attendances **2,527,834**

Accident & Emergency Attendances **522,910**

Dental Attendances **244,396**

Dental Procedures **325,005**

Polyclinic Attendances **1,837,506**

COMMUNITY HOSPITALS

Size

Beds (As at end March) **608**

Workload per annum

Bed Occupancy Rate **76.1 %**

Inpatients **4,722**

Total Patient Days **144,860**

rations Management

You can try!



Singapore
General Hospital

Outram Road, Singapore 169608
Tel: 6222 3332 www.sgh.com.sg

					TOTAL 8,993
1,161	0	3,682	1,755	2,395	Year ended 31 March 2020

Size

Beds (As at end March) **1,796**

Workload per annum

Bed Occupancy Rate **82.7 %**

Inpatients **79,557**

Total Patient Days **480,971**

Average Length of Stay (days) **6.0**

Day Surgeries **58,170**

Inpatient Surgeries **45,147**

Specialist Outpatient Clinic Attendances **720,520**

Accident & Emergency Attendances **122,744**

https://www.singhealth.com.sg/about-singhealth/newsroom/Documents/SingHealth%20Duke-NUS%20AR1920-OVERVIEW_final.pdf

Class Outline

- Healthcare Operations Management
- Reflections on Healthcare Operations Management
- Examination Briefing

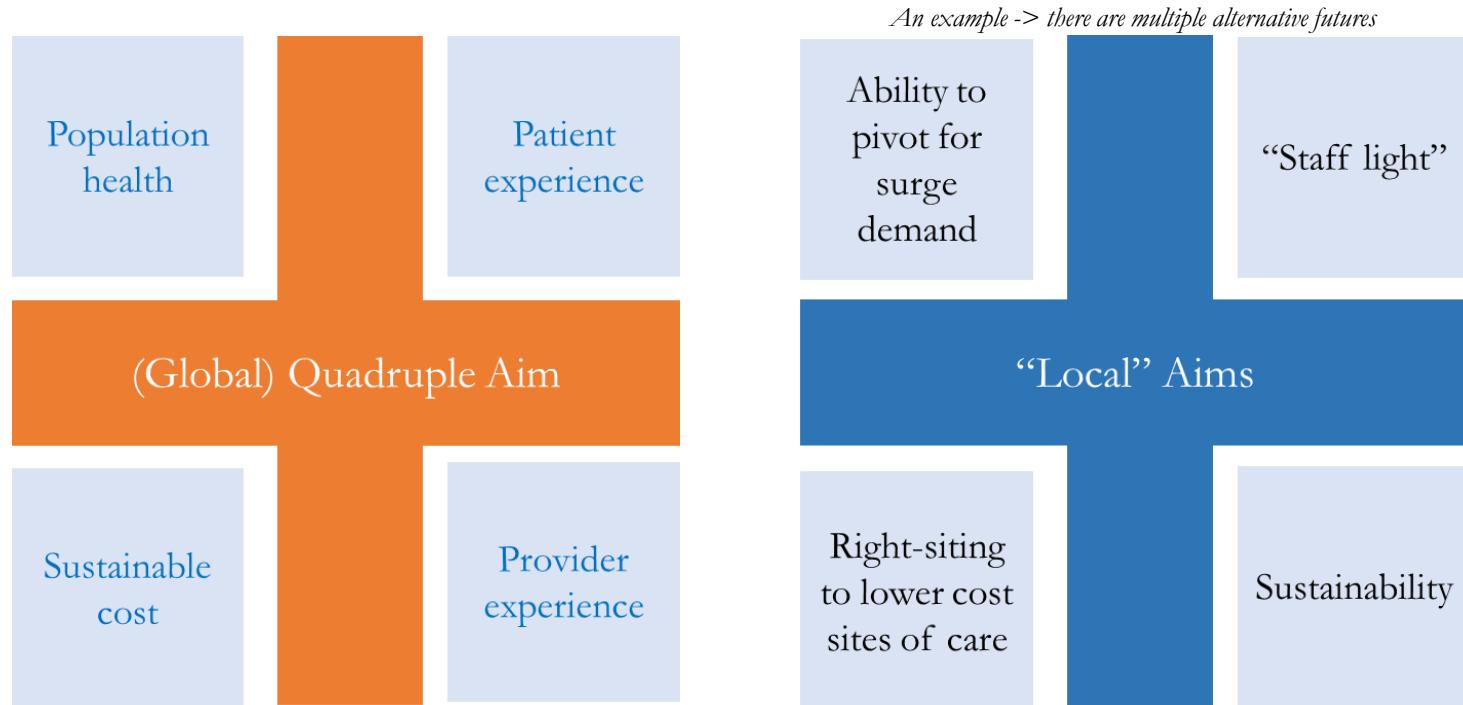


If you had a magic wand and could instantly fix one major challenge in healthcare today, what would it be?

Join at
slido.com
#3793 888



Decoding Challenges Multifactorial and Multiple Objectives



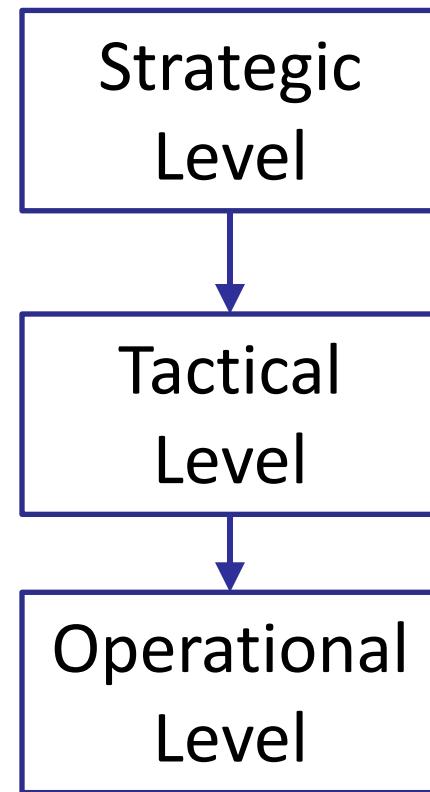
Based on: Quadruple Aim – Bodenheimer et al. (2014); “Local” aims – Typical for Singapore RHS

Objectives in OM and HOM

- Four major operations objectives
 - Cost
 - Quality
 - Dependability/speed (meeting schedules/deadlines),
 - Flexibility (responsiveness to changes/uncertainties)
- Also referred to as operations capabilities
- Can all four objectives be achieved simultaneously?
- Can we provide affordable and inclusive access to quality healthcare, in a timely manner?

Some Topics in OM Related to HOM

- Strategic Operations
- Process management
- Supply chain management
- Operations scheduling
- Quality management
 - Lean manufacturing
 - Theory of constraints



Solution Techniques

- Mathematical programming
- Stochastic programming
- (Distributionally) robust optimization
- Markov decision processes
- (Stochastic) dynamic programming
- Simulation (optimization)
- ...

Class Outline

- Healthcare Operations Management
- Reflections on Healthcare Operations Management
- Examination Briefing

Course Debrief

Problem Formulation and Healthcare Data

Introduction to Analytics

Exploratory Data Analytics

Healthcare Data Management

Statistical Methodologies

Risk Factor Analysis 1: Correlation Analysis, Linear Regression, etc

Risk Factor Analysis 2: Logistic Regression, Model Validation, etc

Causal Analysis (Propensity Score Matching, Bradford Hill Criteria, etc)

Machine Learning

Unsupervised and Supervised Machine Learning

Ops Mgmt

Healthcare Operations Management

LECTURE 1-4

LECTURE 4-6

LECT 6-7

LECT 8-10

LECT 10

Grade Components

Class Participation: 15%

Peer Evaluation: 5%



Class Participation: 20%

Assignments 1 and 2: 15%

Presentation (Case Study) - Group: 15%

Mid-term Checkpoint Quiz: 15%

Final Exam: 35%

Total: 100%



: Grades will be uploaded and released during Study Week;
Please check and confirm the grades are correct

Exam Briefing

- Date: 30 April 2025
- Time: 1PM – 3PM
- Venue: TBA
- **Study Week Consultation:**
 - Date: 15 April 2025 (8AM-2PM)
 - Venue: TBA
 - Not compulsory, **book 30 mins slots as a group with the TA if needed**

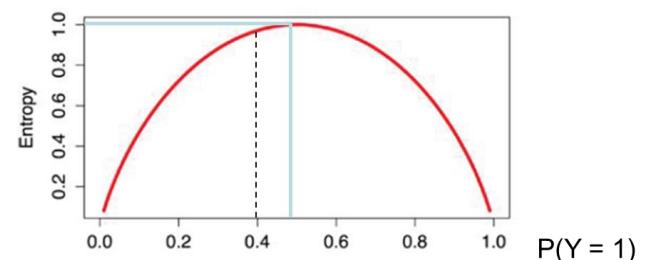
Exam Briefing

- Format:
 - Closed Book with formula sheet provided
 - **Bring your Calculators!**
 - **Section A: 20 MCQs** (5 alternatives each) [40 marks]
 - Like Quiz
 - **Section B: 3 Structured Questions** [60 marks]
 - Short answers – like Assignment 2
 - Multiple parts
 - Both qualitative and quantitative questions

Log base 2 Calculations

$$\log_2 x = \frac{\log_{10} x}{\log_{10} 2}$$

$$\begin{aligned}\text{Entropy } H(Y) &= -\sum_{i=1}^n p_i \log_2(p_i) \\&= -[0.6 \times \log_2(0.6) + 0.4 \times \log_2(0.4)] \\&= -[0.6 \times \frac{\log_{10} 0.6}{\log_{10} 2} + 0.4 \times \frac{\log_{10} 0.4}{\log_{10} 2}] \\&= 0.9709506\end{aligned}$$



Coverage

- Questions will focus on testing understanding of concepts:
 - Introduction to Analytics
 - Exploratory Data Analytics
 - Healthcare Data Management
 - Risk Factor Analysis 1: Correlation Analysis, Linear Regression, etc
 - Risk Factor Analysis 2: Logistic Regression, Model Validation, etc
 - Causal Analysis (Propensity Score Matching, Bradford Hill Criteria, etc)
 - Unsupervised and Supervised Machine Learning
 - Donabedian and AHRQ Framework for Quality of Care
 - Clustering Analysis; CART Decision Trees; Advanced Decision Trees
 - Cross Validation; ML Workflow; Gini Impurity, Entropy, Etc
 - Healthcare Operations Management (Process Analysis, Little's Law)

Coverage

- You will not be asked to write R codes on the spot but understanding of R outputs will be required.
 - For example: interpretation of R regression output and the feature importance chart from CART (Quality.Rmd)
 - Outputs that are covered in class
- Case Studies are excluded
- Assignment 2 (Part 2) Case Study is excluded

```

Call:
lm(formula = TenYearCHD ~ ., data = framinghamTrain)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.80807 -0.18579 -0.10575 -0.00967  1.04095 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.776e-01  9.446e-02 -7.174 9.54e-13 ***
male          5.230e-02  1.481e-02  3.532 0.000419 ***
age           7.455e-03  9.070e-04  8.220 3.21e-16 ***
education    -8.190e-03  6.770e-03 -1.210 0.226498  
currentSmoker 1.002e-02  2.178e-02  0.460 0.645564  
cigsPerDay    2.235e-03  9.329e-04  2.396 0.016665 *  
BPMedS        4.503e-02  4.187e-02  1.075 0.282334  
prevalentStroke 1.846e-01  9.160e-02  2.015 0.043987 *  
prevalentHyp   1.141e-02  2.088e-02  0.547 0.584725  
diabetes       -4.679e-02  5.102e-02 -0.917 0.359218  
totchol         7.731e-05  1.624e-04  0.476 0.634109  
sysBP          2.723e-03  5.932e-04  4.591 4.62e-06 ***  
diaBP          -1.225e-03  9.690e-04 -1.264 0.206259  
BMI            2.088e-03  1.853e-03  1.127 0.260004  
heartRate      -6.279e-04  5.826e-04 -1.078 0.281172  
glucose         1.689e-03  3.596e-04  4.698 2.77e-06 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3399 on 2545 degrees of freedom
Multiple R-squared:  0.1105,    Adjusted R-squared:  0.1053 
F-statistic: 21.08 on 15 and 2545 DF,  p-value: < 2.2e-16

```

