

SISCR Module 7
Part I:
Introduction
Basic Concepts for Binary Biomarkers (Classifiers)
and Continuous Biomarkers

Kathleen Kerr, Ph.D.
Associate Professor
Department of Biostatistics
University of Washington

Module Overview

- Part I: Introductory concepts
- Part II: Evaluating Risk Models
- Part III: Evaluating the Incremental Value of New Biomarkers
- Part IV: Some Guidance on Developing Risk Models

- also: R tutorial/demo

Part I Topics

- Motivating and illustrative examples
- True and false positive rates (TPR, FPR)
- Predictive values (PPV, NPV)
- ROC curves and area under the curve (AUC)
- Risk models
- What is “personal risk”?

3

Part 1 Overview

- Some examples
- To start: 1 marker X is binary (a “test”)
- We then move on: 1 marker X is continuous
- Multiple markers X, Y, \dots , and risk model $P(\text{bad outcome} \mid X, Y, \dots)$

4

What is a Marker?

- DEF: a quantitative or qualitative measure that is potentially useful to classify individuals for current or future status
 - current → diagnostic marker
 - future → prognostic marker
- Includes biomarkers measured in biological specimens
- Includes imaging tests, sensory tests, clinical signs and symptoms, risk factors

5

What is the purpose of a classifier or risk prediction tool?

- To inform subjects about risk
- To help make medical decisions
 - Most often: identify individuals with high risk – the assumption is that these individuals have the greatest possibility to benefit from an intervention
 - Sometimes: identify individuals with low risk not likely to benefit from an intervention
- To enrich a clinical trial with “high risk” patients

6

Terminology and Notation

- “case” or “event” is an individual with the (bad) outcome
- “control” or “nonevent” is an individual without the outcome

case	control
D=1	D=0
D	\bar{D}
D	N

7

Terminology and Notation

- X, Y = potential predictors of D
(demographic factors, clinical characteristics, biomarker measurements)
- Often: X is “standard” predictors and Y is a new biomarker under consideration
- $\text{risk}(X) = r(X) = P(D=1 \mid X)$
– $\text{risk}(X,Y) = r(X,Y) = P(D=1 \mid X, Y)$
- prevalence = $P(D=1) = \rho$ (“rho”)

8

What is risk(X)?

- $\text{risk}(x) \equiv P(D=1 | X=x)$ is the frequency of events/disease among the group with $X = x$
- “Personal risk” is not completely personal!
 - Will return to this at the end of Part I

9

Example: Coronary Artery Surgery Study (CASS)

- 1465 men undergoing coronary arteriography for suspected coronary heart disease
- Arteriography is the “gold standard” measure of coronary heart disease
 - Evaluates the number and severity of blockages in arteries that supply blood to the heart
- Simple cohort study
- Possible predictor: Exercise stress test (EST)
- Possible predictor: chest pain history (CPH)

10

Example: EDRN Breast Cancer Biomarkers

- Women with positive mammograms undergo biopsy, the majority turn out to be benign lesions
- Provides motivation to develop serum biomarker to reduce unnecessary biopsies

11

Example: Pancreatic Cancer Biomarkers

- 141 patients with either pancreatitis (n=51) or pancreatic cancer (n=90)
- Serum samples
- Two candidate markers:
 - A cancer antigen CA-125
 - A carbohydrate antigen CA19-9
- Which marker is better at identifying cancer?
- Is either marker good enough to be useful?

Wieand, Gail, James, and James *Biometrika* 1989

Example: Cardiovascular Disease

- Framingham study
- D = CVD event
- Y = high density lipoprotein
- X = demographics, smoking, diabetes, blood pressure, total cholesterol
- $n = 3264$, $n_D = 183$

13

Simulated Data

- Artificial data are useful for exploring/illustrating methodology
- Here I introduce simple but useful models that I will use to illustrate some methods
 - Simulated data on DABS website
 - Simulated data from R packages DecisionCurve and BioPET
 - Normal and MultiNormal biomarker model

14

Example: Simulated data on DABS website

- $n = 10,000$, $n_D = 1017$
- $Y =$ continuous, 1-dimensional
- $X =$ continuous, 1-dimensional
- <http://labs.fhcrc.org/pepe/dabs/> or search “Pepe DABS”

15

Example: Simulated data in R packages

- $n = 500$, $n_D = 60$
- $X =$ sex, smoking status, Marker1
- $Y =$ Marker2
- These simulated data will not appear in lecture notes, but will appear in software demo

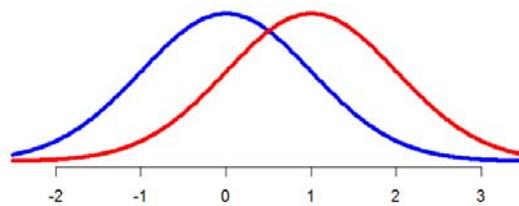
16

Normal Model with 1 Marker

- Biomarker X Normally distributed in **controls** and in **cases**

$$X \sim N(0,1) \text{ in controls}$$

$$X \sim N(\mu,1) \text{ in cases}$$



Distribution of X when $\mu=1$

17

Multivariate Normal Model with 2 Markers (Bivariate Normal)

- Biomarkers (X_1, X_2) are bivariate Normally distributed in controls and in cases

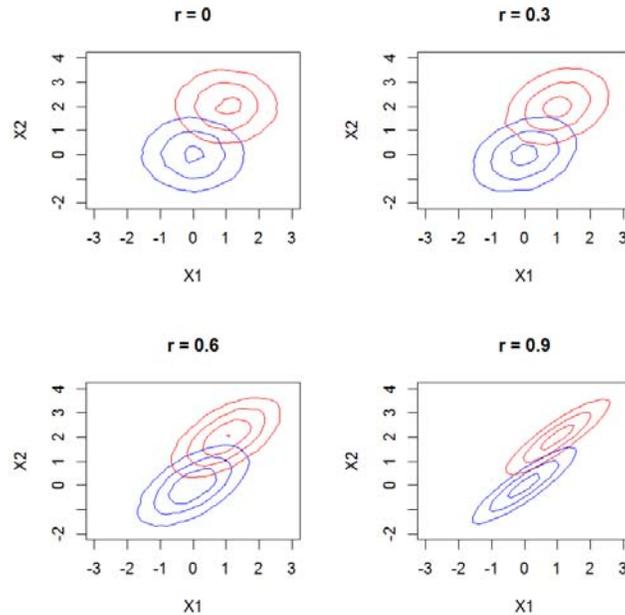
$$\vec{X} \sim MVN(\vec{0}, \Sigma) \text{ in controls}$$

$$\vec{X} \sim MVN(\vec{\mu}, \Sigma) \text{ in cases}$$

$$\Sigma = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

18

In these examples X_1 and X_2 each have mean 0 in **controls** and mean 1 in **cases**. We can picture marker data in 2-dimensional space.



19

- Biomarkers (X_1, X_2) are bivariate Normally distributed in controls and in cases

$$\vec{X} \sim MVN(\vec{0}, \Sigma) \text{ in controls}$$

$$\vec{X} \sim MVN(\vec{\mu}, \Sigma) \text{ in cases}$$

- This data model is useful in research because the logistic regression model holds for each marker **and** for both markers together.

logit $P(D=1 | X_1)$ is linear in X_1

logit $P(D=1 | X_2)$ is linear in X_2

logit $P(D=1 | X_1, X_2)$ is linear in X_1 and X_2

20

Generalization: Multivariate Normal Model

- Biomarkers (X_1, X_2, \dots, X_k) are multivariate Normally distributed in controls and in cases

$$\vec{X} \sim MVN(\vec{0}, \Sigma) \text{ in controls}$$

$$\vec{X} \sim MVN(\vec{\mu}, \Sigma) \text{ in cases}$$

- The linear logistic model holds for every subset of markers

**QUANTIFYING CLASSIFICATION
ACCURACY (BINARY MARKER OR “TEST”)**

Terminology

- D = outcome (disease, event)
- Y = marker (test result)

	D=0	D=1
Y=0	true negative	false negative
Y=1	false positive	true positive

23

Terminology

TPR = true positive rate = $P[Y=1|D=1]$ = sensitivity

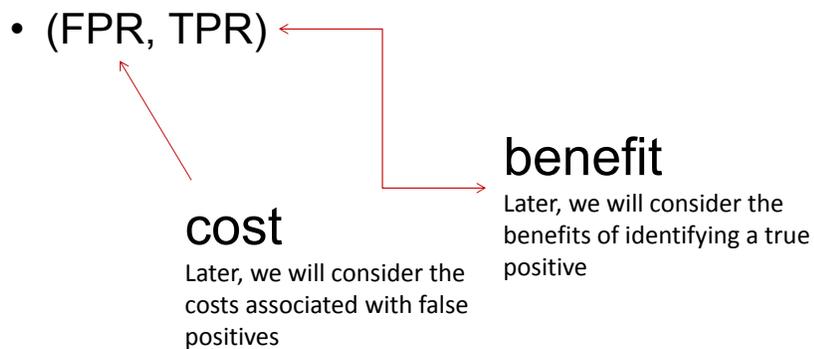
FPR = false positive rate = $P[Y=1|D=0]$ = 1-specificity

FNR = false negative rate = $P[Y=0|D=1]$ = 1-TPR

TNR = true negative rate = $P[Y=0|D=0]$ = 1-FPR

Ideal test: FPR=0 and TPR=1

24



25

Coronary Artery Surgery Study (CASS)

Coronary Artery Disease

		D=0	D=1
Exercise Test	Y=0	327	208
	Y=1	115	815
		442	1023

$$\text{FPR} = 115/442 = 26\%$$

$$\text{TPR} = 815/1023 = 80\%$$

26

What about Odds Ratios?

- Odds ratios are very popular:
 - Because logistic regression is popular
 - Odds Ratio estimable from case-control study
 - $OR \approx$ relative risk for rare outcome
- $OR = \frac{TPR(1-FPR)}{FPR(1-TPR)}$
- Good classification (high TPR and low FPR)
→ large odds ratio
- However, large odds ratio does NOT imply good classification!

27

Good classification → large odds ratio

E.g., TPR=0.8, FPR=0.10

$$OR = \frac{0.8 \times 0.9}{0.1 \times 0.2} = 36$$

28

Coronary Artery Surgery Study (CASS)

		Coronary Artery Disease		
		D=0	D=1	
Exercise Test	Y=0	327	208	FPR=115/442=26%
	Y=1	115	815	TPR=815/1023=80%
		442	1023	OR ≈ 11.1

OR is large but classification performance is not exceptional.

29

large odds ratio does NOT imply good classification!

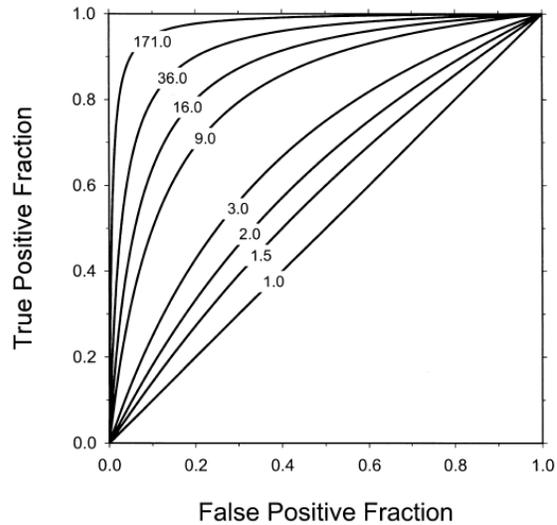


FIGURE 1. Correspondence between the true-positive fraction (TPF) and the false-positive fraction (FPF) of a binary marker and the odds ratio. Values of (TPF, FPF) that yield the same odds ratio are connected.

Pepe et al, American Journal of Epidemiology 2004;
159:882-890.

- Need to report *both* FPR and TPR
- Collapsing into one number (e.g., OR) is not sufficient
 - important information is lost

31

Misclassification Rate

$$\begin{aligned} \text{MR} &= \text{error rate} = P(Y \neq D) \\ &= P(Y=0, D=1) + P(Y=1, D=0) \\ &= \rho(1-\text{TPR}) + (1-\rho)\text{FPR} \end{aligned}$$

- ρ is the prevalence $P(D=1)$
- only appropriate if the cost of false positives equals the cost of false negatives
- seldom useful or appropriate

32

Misclassification Rate

- There are **two kinds of wrong decisions** and the MR equates these. In order to be clinically relevant we must consider the **cost of each kind of error**
 - ... later today

33

- FPR, TPR condition on true status (D)
- they address the question: “to what extent does the biomarker reflect true status?”

34

Predictive Values

Positive predictive value $PPV = P(D=1|Y=1)$

Negative predictive value $NPV = P(D=0|Y=0)$

- condition on biomarker results (Y)
- address the question: “Given my biomarker value is Y, what is the chance that I have the disease?” This is the question of interest for patients and clinicians in interpreting the result of a biomarker test

35

Predictive Values

PPV and NPV are functions of TPR and FPR *and* the prevalence ρ

$$PPV = \frac{\rho TPR}{\rho TPR + (1 - \rho)FPR}$$
$$NPV = \frac{(1 - \rho)(1 - FPR)}{(1 - \rho)(1 - FPR) + \rho(1 - TPR)}$$

- TPR, FPR are **properties of a test**, but PPV, NPV are **properties of a test in a population**
- For low prevalence conditions, PPV tends to be low, even with very sensitive tests

36

False Discovery Rate

$$\begin{aligned}\text{False Discovery Rate } \text{FDR} &= P(D=0|Y=1) \\ &= 1 - \text{PPV}\end{aligned}$$

“False Discovery Rate” and “False Positive Rate” sound similar, but they are not the same!

- FPR: among all those who are not diseased, how many were called positive
- FDR: among all those you called positive, how many were not actually diseased.
- We will not use or further discuss FDR further today.

Motivation

- Most biomarkers are continuous

Convention

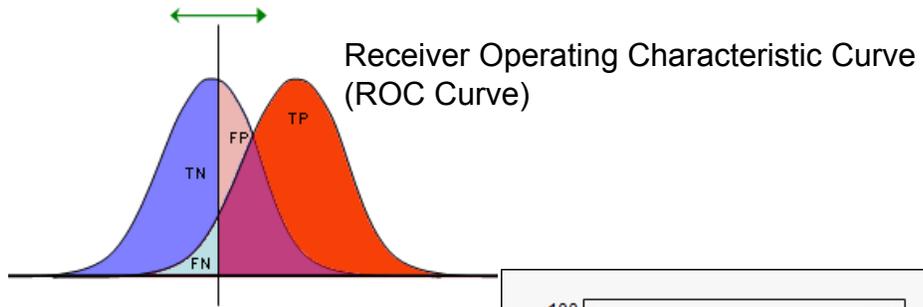
- Assume larger Y more indicative of disease
 - otherwise replace Y with $-Y$
- Formally: $P(D=1 | Y)$ increasing in Y

39

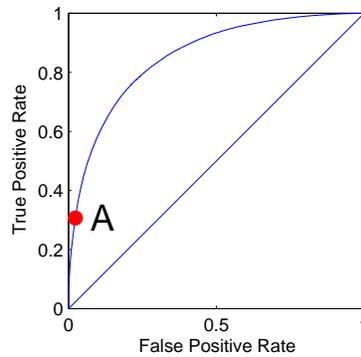
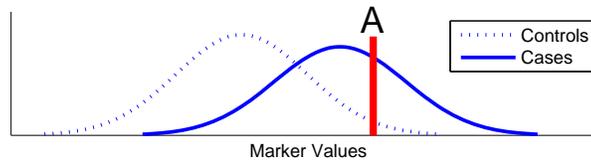
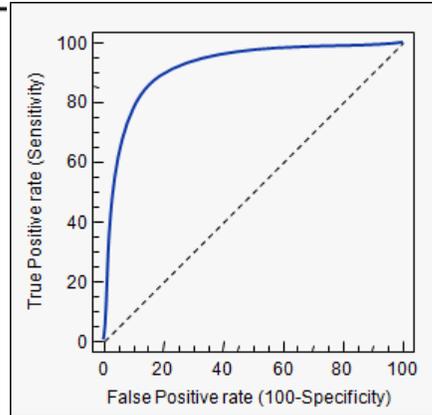
Receiver Operating Characteristic (ROC) Curve

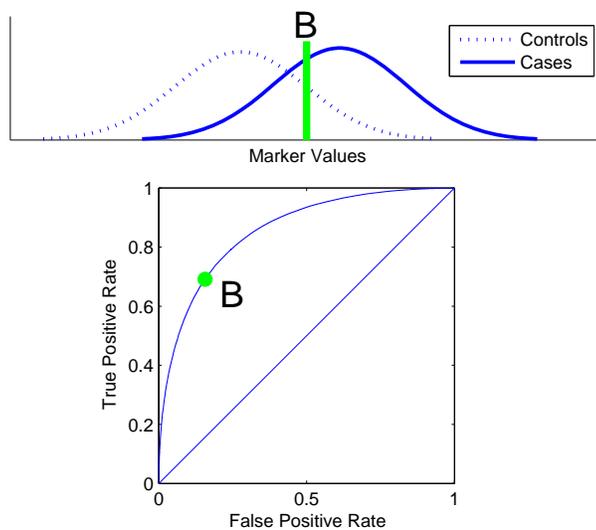
- generalizes (FPR, TPR) to continuous markers
- considers rules based on thresholds " $Y \geq c$ "
 - makes sense if $P(D=1|Y)$ increasing in Y
- $TPR(c) = P(Y \geq c | D=1)$
- $FPR(c) = P(Y \geq c | D=0)$
- $ROC(\cdot) = \{FPR(c), TPR(c) ; c \text{ in } (-\infty, \infty)\}$

40

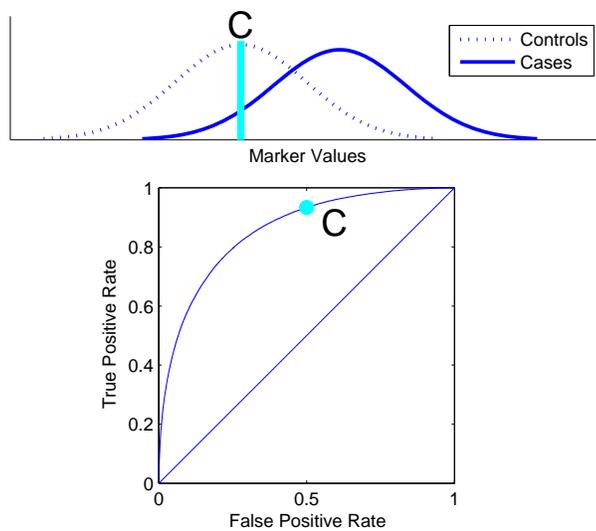


Each point on the ROC curve corresponds to a threshold for declaring “marker-positive.”

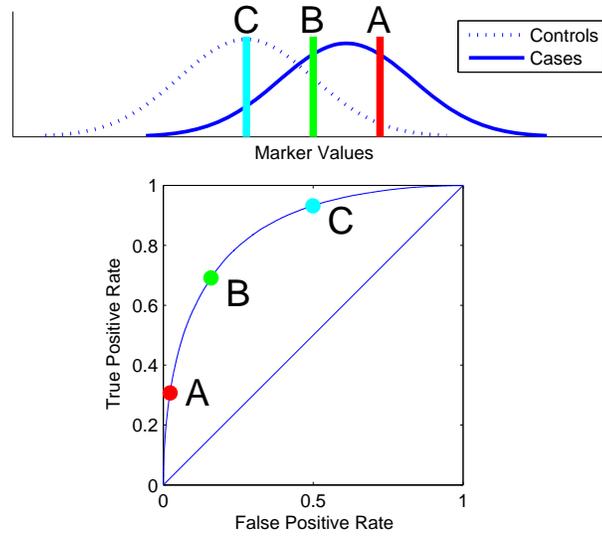




43

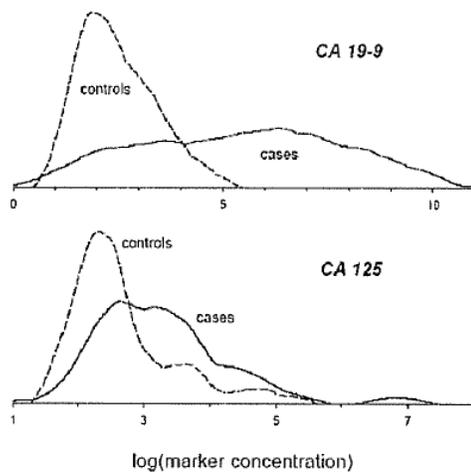


44

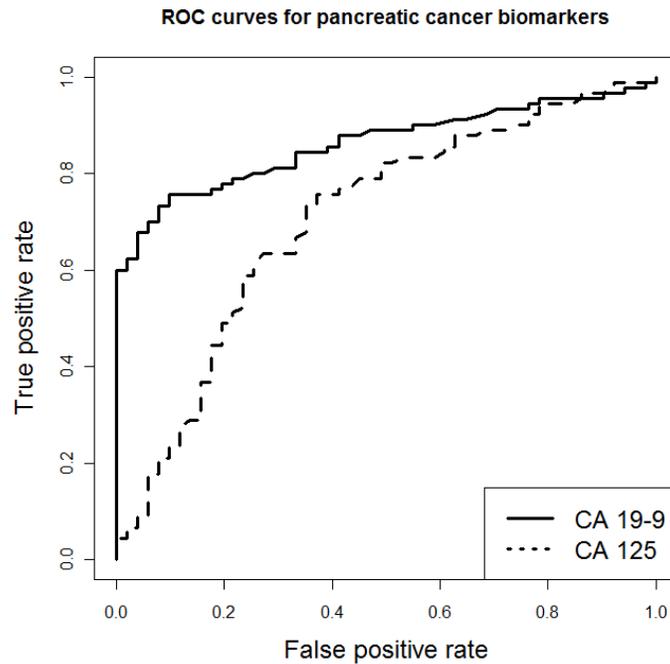


45

Pancreatic cancer biomarkers (Wieand et al 1989)



46



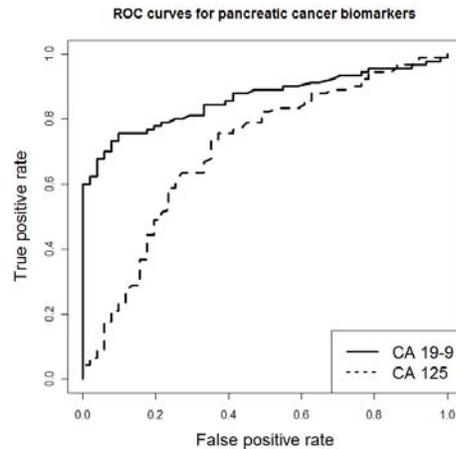
47

Properties of ROC curves

- non-decreasing from (0,0) to (1,1) as threshold decreases from $c=\infty$ to $c=-\infty$
- *ideal* marker has control distribution completely disjoint from case distribution; ROC through (0,1)
- *useless* marker has ROC equal to 45 degree line
- doesn't depend on scale of Y: invariant to monotone increasing transformations of Y
- puts different markers on a common relevant scale
- shows entire range of possible performance

48

CA-19-9 appears to be the more accurate diagnostic biomarker for pancreatic cancer



- for most fixed FPR, CA-19-9 has the better corresponding TPR
- for most fixed TPR, CA-19-9 has the better corresponding FPR

49

Comparing ROC Curves: AUC

- **AUC** is Area under ROC curve
- $AUC = \int_0^1 ROC(t) dt = \text{average}(TPR)$
 - average is uniform over (0,1)
- commonly used summary of an ROC curve
 - also called the c-index or c-statistic
- ideal test: $AUC=1.0$
- useless test: $AUC=0.5$
- A single number summary of a curve is necessarily a crude summary

50

AUC: probabilistic interpretation

- $P(Y_D > Y_N)$ for a randomly selected case D and a randomly selected control N
 - Provides an interpretation for AUC beyond “area under ROC curve”
- The AUC is a summary of an ROC curve that is commonly used to compare ROC curves – it is interpretable, but the interpretation also shows that AUC is not clinically meaningful

Risk Model: Huntington's Disease

- Huntington's Disease is caused by the gene *HTT* on human chromosome 4. There is a CAG segment that is repeated 10-35 times in non-diseased individuals. If the segment is repeated 36-120+ times, a person always* develops Huntington's Disease in middle-age. The genetic abnormality is dominant, meaning one abnormal gene causes disease.
 - *>40 times: always develop HD
 - *36-39 times: might not develop HD (ignoring this small possibility for this example)

53

Risk Model: Huntington's Disease

- Relevant Population: Individuals with a biological parent who have Huntington's Disease
- These individuals have a 50% chance of developing HD depending on whether they inherited the abnormal or normal version of the gene from their affected parents.
- $P(D) = 0.5 = p$ in this population.

54

Risk Model: Huntington's Disease

- These individuals can choose to have their *HTT* gene genotyped. Say $HTT=0$ means 0 copies of abnormal gene; $HTT=1$ mean 1 copy of abnormal gene.
- $P(D|HTT=0)=0\%$; $P(D|HTT=1)=100\%$.
- The marker *HTT stratifies* the patient population (risk=50%) into the subgroup with 0% risk and the subgroup with 100% risk.

55

Risk model

- risk prediction model – gives a risk for a marker value or a combination of markers
- Predicted risks are in the interval $[0,1]$ and interpreted as probabilities
- It is rare that a risk model is definitive like the HD example
 - In fact, because the genetic test for Huntington's Disease is definitive, most people do not even think of it as a risk model

56

Risk model examples

- Most risk models combine information from multiple risk factors
- E.g., Gail model for breast cancer risk
 - for use in women with no history of breast cancer
 - Estimates 5-year risk of breast cancer based on current age, age at menarche, age at first birth, family history, race.
- E.g., Framingham CHD risk score
 - Estimates risk of CHD based on age, sex, smoking status, total and HDL cholesterol, BP

57

Risk model examples

- E.g. STS risk score for dialysis following cardiac surgery is formed via:
 - STS risk score = $f(\alpha + \beta_1 \text{ Age} + \beta_2 \text{ Surgery Type} + \beta_3 \text{ Diabetes} + \beta_4 \text{ MI Recent} + \beta_5 \text{ Race} + \beta_6 \text{ Chronic Lung Disease} + \beta_7 \text{ Reoperation} + \beta_8 \text{ NYHA Class} + \beta_9 \text{ Cardiogenic Shock} + \beta_{10} \text{ Last Serum Creatinine})$

58

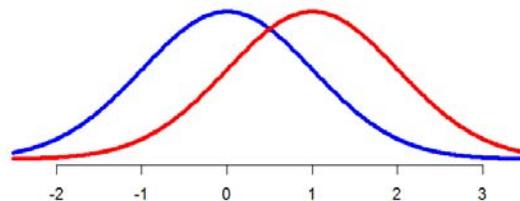
What is “personal risk”?

- Recall: $\text{risk}(x) \equiv P(D=1 | X=x)$ is the frequency of events among the group with marker values x
- “Personal risk” is not completely personal!
 - (next example)

59

What is “personal risk”?

- Suppose the prevalence of D in “Population A” is 1%
 - Without any additional information, the only valid risk prediction instrument is to assign everyone in the population $\text{risk}=1\%$
- Suppose we have a marker X that tends to be higher in the cases than controls

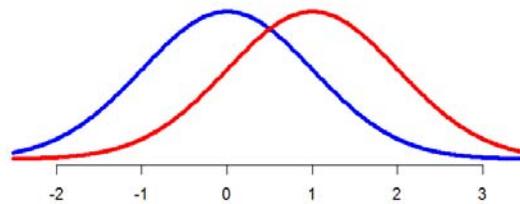


Distribution of marker X in controls (blue) and cases (red)

60

What is “personal risk”?

- Suppose an individual in Population A has X measured as 1.
- We can calculate his risk($X=1$) \approx 1.6%
 - We can calculate the risk using Bayes’ rule

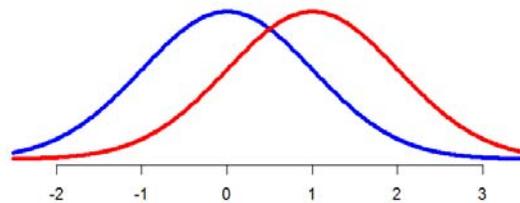


Distribution of marker X in controls (blue) and cases (red)

61

What is “personal risk”?

- Suppose the marker acts exactly the same in Population B. The only difference between Populations A and B is that B has prevalence=10%.
- An individual in Population B has $X=1$. For that individual, his risk is \approx 15.5%



Distribution of marker X in controls (blue) and cases (red)

62

What is “personal risk”?

- “Personal risk” is a term that is prone to be misconstrued
- Risk is personal when calculated based on personal characteristics
- However, personal risk is not completely divorced from population characteristics. For example, the previous example shows that the population (specifically, the population prevalence) affects “personal” risk.

63

What is “personal risk”?

- Occasionally one will hear mention of estimating a person’s “individual risk” or “true personal risk.”
- Frequentist statisticians cannot really claim to do so.
- One might claim John’s “true risk” of a heart attack in the next 5 years is 7%. But we can only observe John having or not having a heart attack in the next 5 years. I cannot observe John having a heart attack in 7% of 5-year periods.
- The best I can claim is that “among people with John’s characteristics, 7% will have a heart attack in the next 5 years.”
 - More than one way to define “people like John.”

64

Summary

- Some example datasets
- FPR, TPR
- PPV, NPV
 - function of FPR, TPR and disease prevalence
- ROC curves
- AUC
 - geometric interpretation as area under curve
 - probability interpretation
- risk model: $\text{risk}(X) = P(D=1|X)$