SISCR Module 7

Part I:
Introduction
Basic Concepts for Binary Classification Tools 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 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, …, and risk model P(bad outcome | X, Y, …)

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

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

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	$\overline{D}$
D	N

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
- risk(X) = r(X) = P(D=1 | X)
  - risk(X,Y) = r(X,Y) = P(D=1 | X, Y)
- prevalence = P(D=1) = ρ   (“rho”)

What is risk(X)?

- risk(x) ≡ P(D=1 | X=x) is the frequency of events among the group with X = x
- “Personal risk” is not completely personal!
  - Will return to this at the end of Section 1

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)

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

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

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

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"

Example: Simulated data in R packages

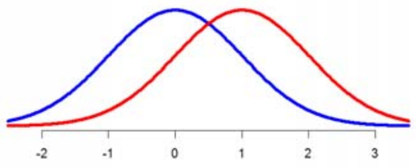
- n = 500, n_D=60
- X = sex, smoking status, Marker1
- Y = Marker2
- These data will not appear in lecture notes, but will appears in software demo

Normal Model with 1 Marker

- Biomarker X Normally distributed in controls and in cases

X ~ N(0,1) in controls

X ~ N(μ,1) in cases



Distribution of X when μ=1

Multivariate Normal Model with 2 Markers (Bivariate Normal)

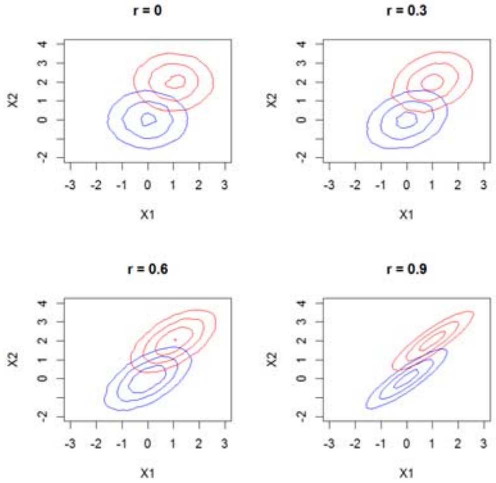
- Biomarkers (X_1, X_2) are bivariate Normally distributed in controls and in cases

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

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

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

In these examples X1 and X2 each have mean 0 in controls and mean 1 in cases. We can picture marker data in 2-dimensional space.



- 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_1, X_2) is linear in X_1 and X_2

Generalization: Multivariate Normal Model

- Biomarkers (X_1, X_2, …, 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

Terminology

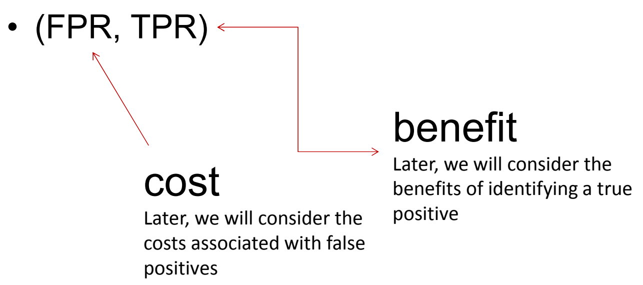
TPR = true positive rate = P[Y=1|D=1] = sensitivity

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



- (FPR, TPR)

cost

Later, we will consider the costs associated with false positives

benefit

Later, we will consider the benefits of identifying a true positive

Coronary Artery Surgery Study (CASS)

Coronary Artery Disease

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

FPR=115/442=26%

TPR=815/1023=80%

What about Odds Ratios?

- Odds ratios are very popular:
  - – Because logistic regression is popular
  - – Odds Ratio estimable from case-control study
  - – OR≈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!

Good classification → large odds ratio

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

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

Coronary Artery Surgery Study (CASS)

Coronary Artery Disease

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

FPR=115/442=26%

TPR=815/1023=80%

OR ≈ 11.1

large odds ratio does NOT imply good classification!

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

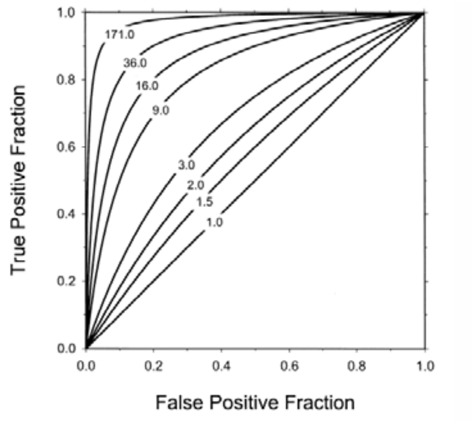


FIGURE 1. Correspondence between the true-positive fraction (TPF) and the false-positive fraction (FPF) of a binary marker and the odds ratio. Values of (TPF, FPF) that yield the same odds ratio are connected.

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

Misclassification Rate

MR = error rate = P(Y ≠ D)

= P(Y=0, D=1) + P(Y=1, D=0)

= ρ(1-TPR)+(1- ρ)FPR

- 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

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

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

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

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

False Discovery Rate

False Discovery Rate FDR=P(D=0|Y=1)

=1 – PPV

“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.

CONTINUOUS MARKERS: ROC CURVES

Motivation

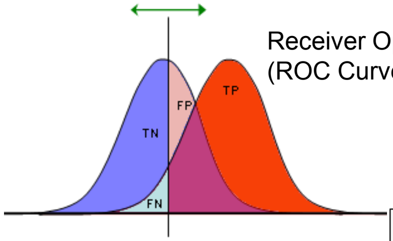
- Most biomarkers are continuous

Convention

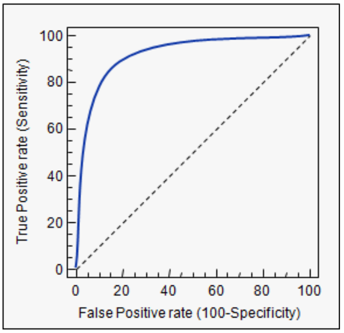
- Assume larger Y more indicative of disease
  - otherwise replace Y with -Y
- Formally: P(D=1 | Y) increasing in Y

Receiver Operating Characteristic (ROC) Curve

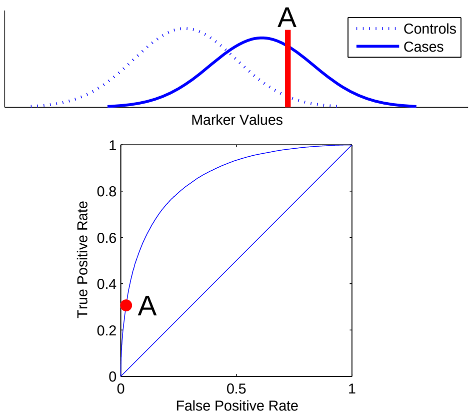
- generalizes (FPR, TPR) to continuous markers
- considers rules based on thresholds “Y≥c”
  - makes sense if P(D=1|Y) increasing in Y
- TPR(c)=P(Y ≥ c | D=1)
- FPR(c)=P(Y ≥ c | D=0)
- ROC(·)={FPR(c), TPR(c) ; c in (-∞,∞)}

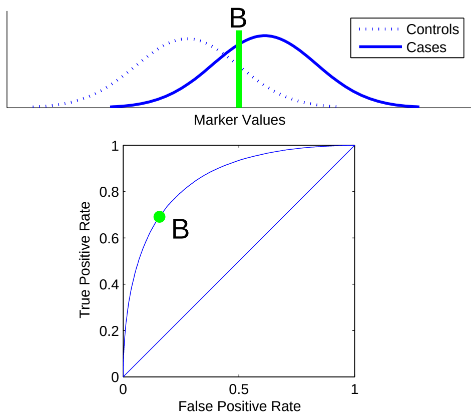


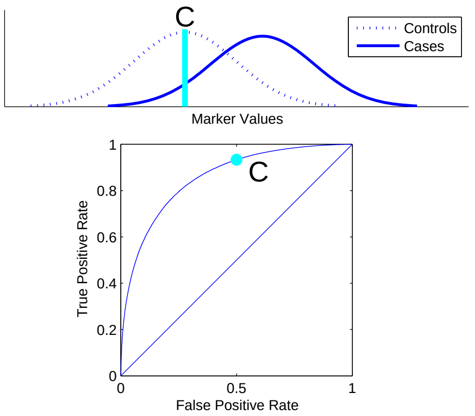
Receiver Operating Characteristic Curve (ROC Curve)

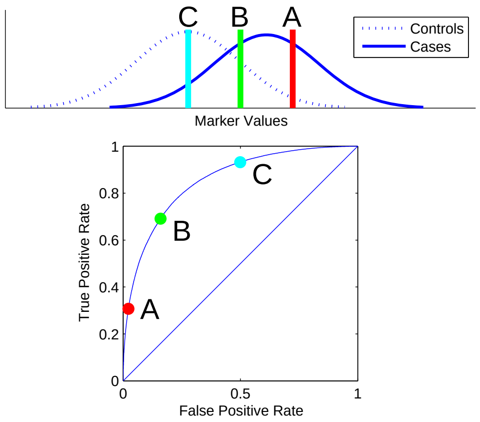


Each point on the ROC curve corresponds to a threshold for declaring “marker-positive.”

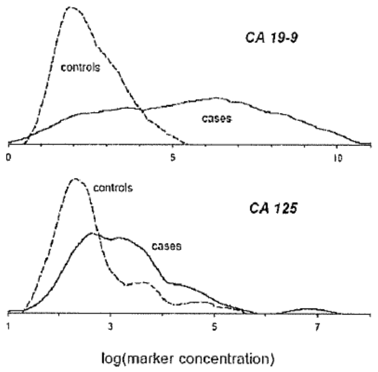

A
Controls
Cases
Marker Values
1
0.8
0.6
0.4
0.2
0
True Positive Rate
A
0
0.5
1
False Positive Rate

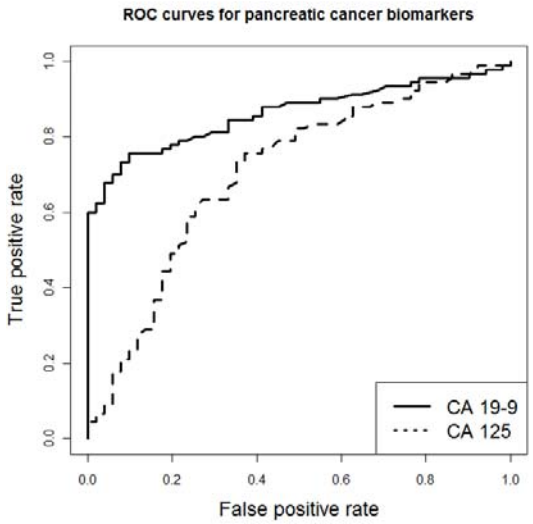

B
Controls
Cases
Marker Values
1
0.8
0.6
0.4
0.2
0
True Positive Rate
B
0
0.5
1
False Positive Rate


C
Controls
Cases
Marker Values
1
0.8
0.6
0.4
0.2
0
True Positive Rate
C
0
0.5
1
False Positive Rate


C
B
A
Controls
Cases
Marker Values
1
0.8
0.6
0.4
0.2
0
True Positive Rate
C
B
A
0
0.5
1
False Positive Rate

Pancreatic cancer biomarkers (Wieand et al 1989)

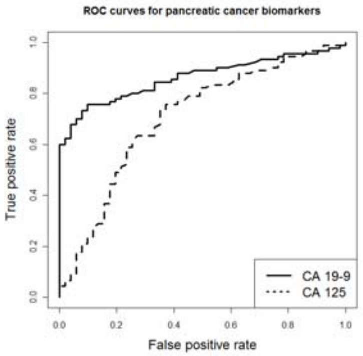




Properties of ROC curves

- non-decreasing from (0,0) to (1,1) as threshold decreases from c=∞ to c= −∞
- *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

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

Comparing ROC Curves: AUC

- AUC is Area under ROC curve
- AUC = $_0\int^1$ ROC(t) dt = 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

AUC: another 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 shows that AUC is not clinically meaningful

RISK PREDICTION

Risk model

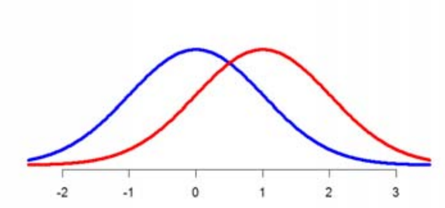
- risk prediction model – gives a risk for a marker value or a combination of markers
- Predicted risks are in the interval [0,1] and interpreted as probabilities
- E.g. STS risk score for dialysis following cardiac surgery is formed via:
  - STS risk score = f(α + β_1 Age + β_2 Surgery Type + β_3 Diabetes + β_4 MI Recent + β_5 Race + β_6Chronic Lung Disease + β_7 Reoperation + β_8NYHA Class + β_9 Cardiogenic Shock+ β_{10}Last Serum Creatinine)

What is “personal risk”?

- Recall: risk(x) ≡ 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)

What is “personal risk”?

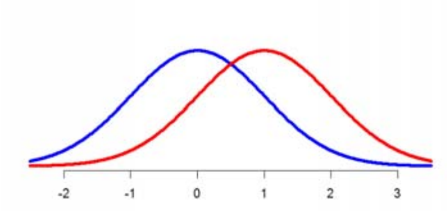
- Suppose the prevalence of D in “Population A” is 1%
  - Without any additional information, the only valid risk prediction instrument is to assign everyone in the population risk=1%
- Suppose we have a marker X that tends to be higher in the cases than controls



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

What is “personal risk”?

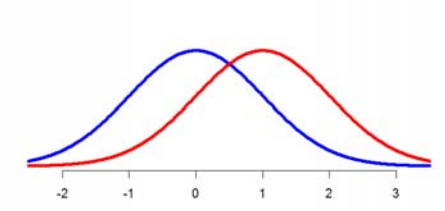
- Suppose an individual in Population A has X measured as 1.
- We can calculate his risk(X=1)≈1.6%
  - We can calculate the risk using Bayes’ rule



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

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 ≈15.5%



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

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.

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: risk(X)=P(D=1|X)