SISCER 2023 Module 5: Evaluation of Biomarkers and Risk Models

Part I: Basic Concepts for Binary Markers (Classifiers) and Continuous Biomarkers

July 13-14, 2023

8:30am-Noon PT / 11:30am-3:00pm ET

Kathleen Kerr, PhD
Professor of Biostatistics
SISCER Director
University of Washington

Overview

- Part I: Introductory Concepts
- Part II: Evaluating Risk Models
- Part III: Evaluating the Incremental Value of New Biomarkers
- Part IV: Guidance on Developing Risk Models
- Part V: Target Performance for Early Phase Biomarker Research
- Part VI: Prognostic vs. Predictive Biomarkers

Overview

- The focus of this short course is concepts rather than statistical details
  - we will not derive hypothesis tests or distributional results
  - we will examine some mathematical expressions as we explore concepts

Misconceptions about Biomarkers and Risk Models

- A large odds ratio means a biomarker is useful for prediction.
- ROC curves are useful to identify the best biomarker cut-point.
- Decision curves are useful to identify the best risk threshold.
- To assess whether to add new biomarker to a risk model, multiple stages of hypothesis testing are needed.
- The best biomarker to improve a risk model is the one with strongest association with the outcome.
- To improve prediction, a new biomarker should be independent of existing predictors.
- We can often use biomarkers to identify which patients will benefit from treatment.

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”?

Part 1 Overview

- Some examples
- To start: 1 marker X is binary (a “test”)
- Then: 1 marker X is continuous
- Multiple markers X, Y, …, and risk model P(bad outcome | X, Y, …)

What is a Biomarker?

- 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

A brief history of *Biomarker*

- According to Lassere (Stat Methods Medical Research 2008):
  - The term *biomarker* first appeared in a 1973 paper on extraterrestrial biological markers


A search for porphyrin biomarkers in nonesuch shale and extraterrestrial samples

Joon H. Rho, A. J. Bauman, Heinz G. Boettger & Teh Fu Yen

Space life sciences **4**, 69–77 (1973) | Cite this article

  - *biomarker* first appeared in the biomedical literature in 1977
  - Most early biomarker papers were in cancer

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

- To help make medical decisions
  - Often: identify individuals with high risk – individuals at high risk of a clinical event have the greatest potential to benefit from an intervention that could prevent the event
  - Sometimes: identify individuals with low risk who are unlikely to benefit from an intervention
- To enrich a clinical trial with "high risk" patients
- To inform subjects about risk
- Etc.

Guiding principle: evaluate a risk model in a way that relates to how it will be used

Terminology and Notation

- "case" or "event" is an individual with the (bad) outcome
- "control" or "non-event" 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 (biomarkers, demographic factors, clinical characteristics)
- Often: X is "standard" predictor(s) 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/disease among the group with X = x

- Risk is simply a population frequency. "Personal risk" is not completely personal!
  - Will return to this at the end of Part I

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 marker: exercise stress test (EST)
- Possible marker: chest pain history (CPH)

Example: 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 (EDRN – early detection research network)

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
  - Simulated data on DABS website
  - Simulated data from R packages *rmda* (risk model decision analysis) and *BioPET*
  - Normal and MultiNormal biomarker model

Example: Simulated data on DABS website

- n = 10,000, n_D=1017
- X = continuous, 1-dimensional
- Y = continuous, 1-dimensional
- Search “Pepe DABS” or http://research.fredhutch.org/diagnostic-biomarkers-center/
  - “simulated risk reclassification dataset”

Example: Simulated data in R packages

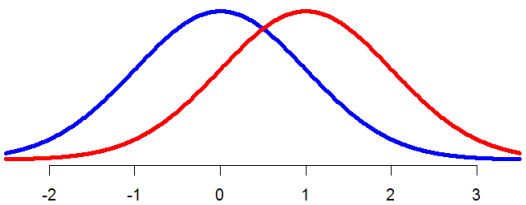
- n = 500, n_D=60
- Four predictors: sex, smoking status, Marker1, Marker2
- Dataset used in software demo (not in course notes)

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)

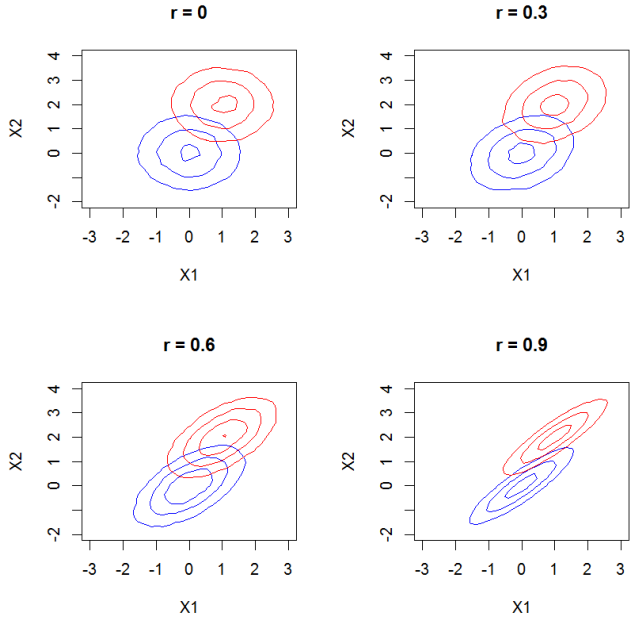
- Biomarkers (X_1, X_2) are bivariate Normally distributed in controls and in cases

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

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

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

In these examples, (X_1, X_2) has mean (0,0) in controls and mean (1,2) in cases. We can visualize 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_2) is linear in X_2

  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")

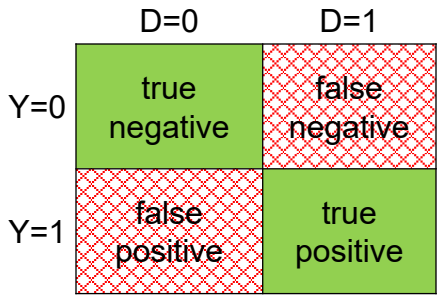
Terminology

- 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

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



27

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

28

- (FPR, TPR)

benefit

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

cost

Later, we will consider the costs associated with false positives

Coronary Artery Surgery Study (CASS)

Coronary Artery Disease

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

FPR=115/442=26%

TPR=815/1023=80%

Odds Ratios do not summarize predictive performance

- 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.
- Need to report *both* FPR and TPR. Collapsing into one number (the OR) is not adequate.

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

OR is large but classification performance is mediocre.

large odds ratio does NOT imply good classification!

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

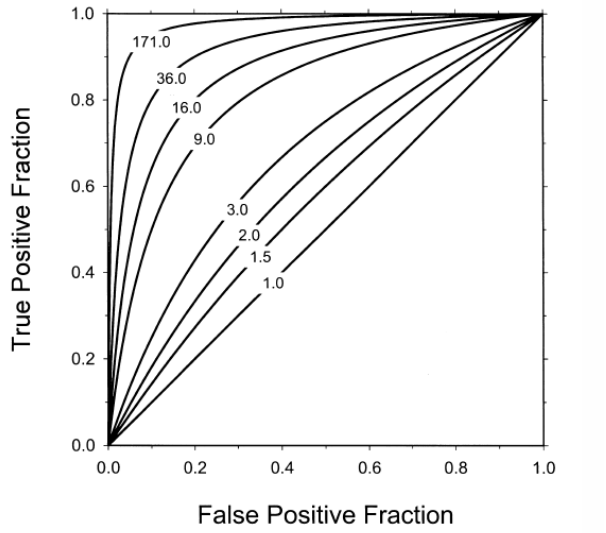


FIGURE 1. Correspondence between the true-positive fraction (TPF) and the false-positive fraction (FPF) of a binary marker and the odds ratio. Values of (TPF, FPF) that yield the same odds ratio are connected.

Classification Accuracy

$$\text{Accuracy} = P(Y = D)$$
$$= P(Y=0, D=0) + P(Y=1, D=1)$$
$$= \rho(TPR)+(1- \rho)(1 - FPR)$$

ρ is the prevalence, P(D=1)

- There are two kinds of errors (false positives and false negatives); accuracy treats them as equally bad.
- In order to be clinically relevant we must consider the harms of each kind of error.
- Accuracy is seldom an appropriate metric in any biomedical application

- 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)
- “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 when interpreting the result of a biomarker or medical 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

Predictive Values - Example

A serious disease affects 1 in 10,000 in a population.

A company markets a test as “98% accurate” because both sensitivity and specificity are estimated to be 98%.

Those who test positive are recommended to undergo an invasive procedure for definitive diagnosis.

Should there be general screening with this test in this population?

Predictive Values - Example

Disease affects 1 in 10,000 in a the population.

Test has sensitivity=specificity=98%.

A person from the population tests negative. What is the probability that person is truly not diseased?

A person from the population tests positive. What is the probability that person has the disease?

Predictive Values - Example

Disease affects 1 in 10,000 in a the population.

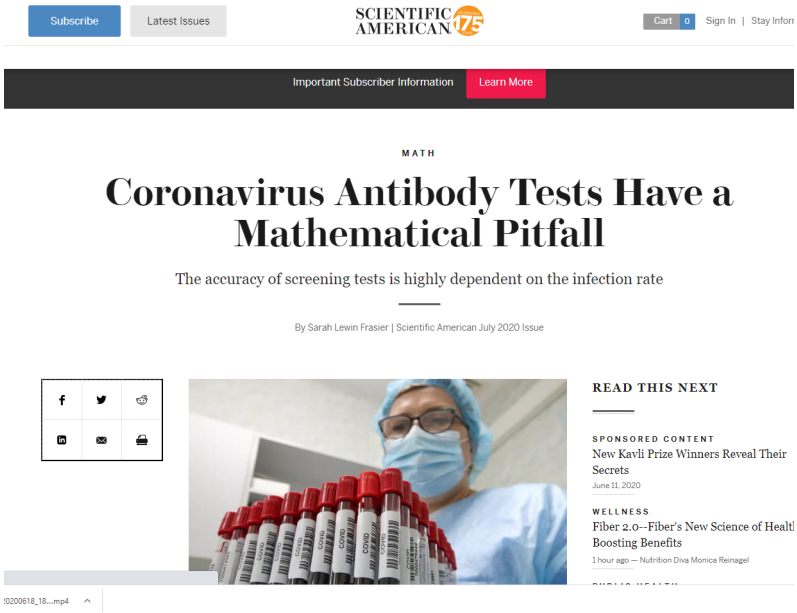
Test has sensitivity=specificity=98%.

What is the probability that person who tests negative is truly not diseased?

What is the probability that person who tests positive truly has the disease?

NPV = ?

PPV = ?



False Discovery Rate

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

=1 – PPV

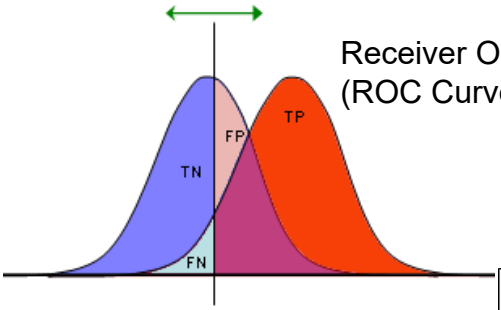
"False Positive Rate" and "False Discovery Rate": sound very similar, but are very different

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

CONTINUOUS MARKERS: ROC CURVES

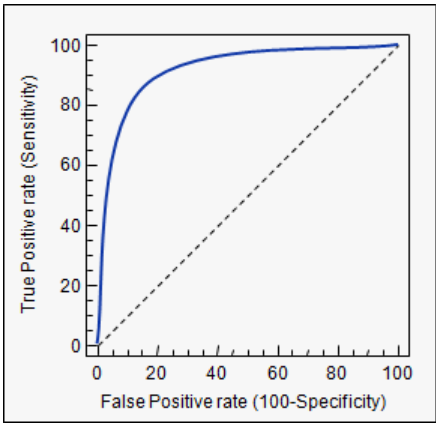
Receiver Operating Characteristic (ROC) Curve

- generalizes (FPR, TPR) to continuous markers
- considers rules based on thresholds "Y≥c"
  - makes sense when 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.”

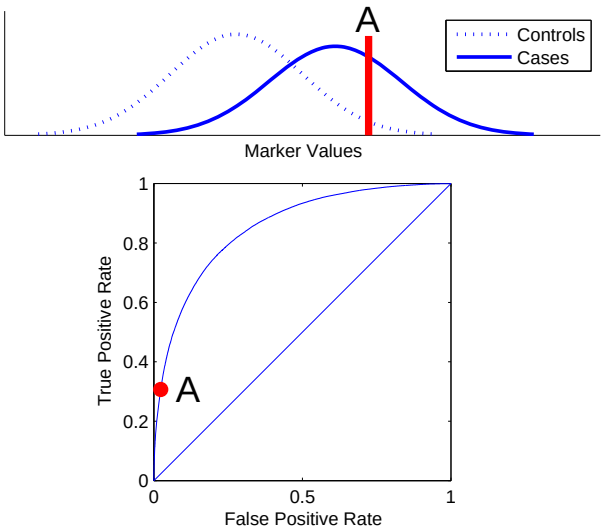


Motivation

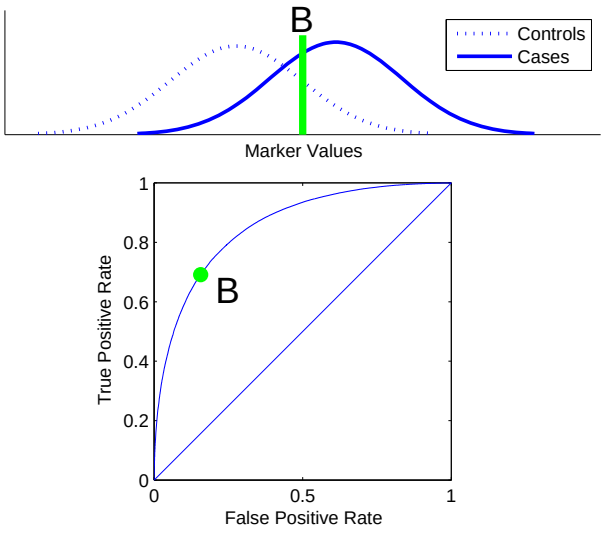
- Most biomarkers are continuous

Convention

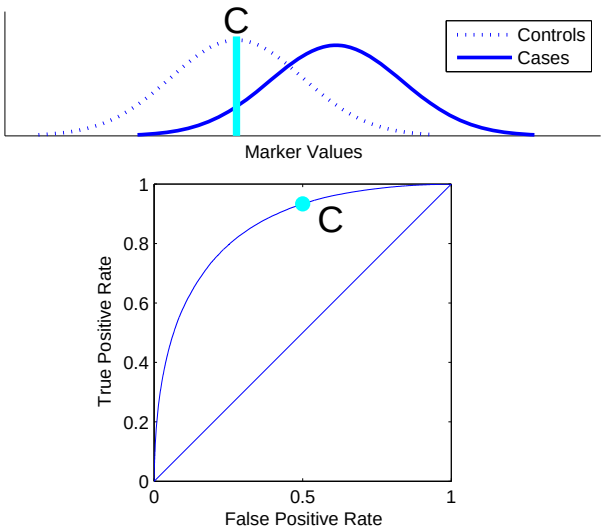
- Assume larger Y more indicative of disease
  - otherwise replace Y with -Y
- Formally: P(D=1 | Y) increasing in Y


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

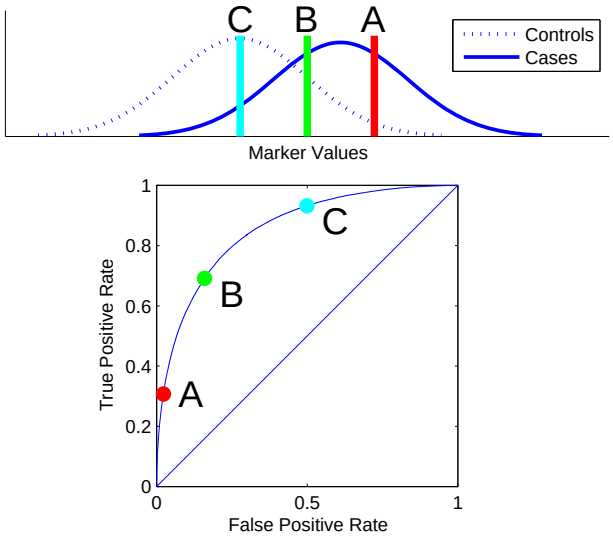
47


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

48


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

49

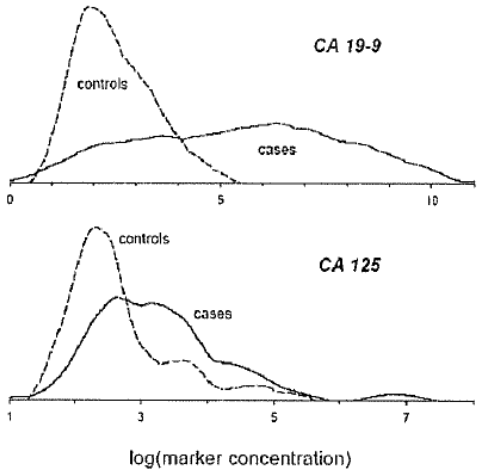

C
B
A
Controls
Cases
Marker Values
C
B
A
True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1
0
0.5
1

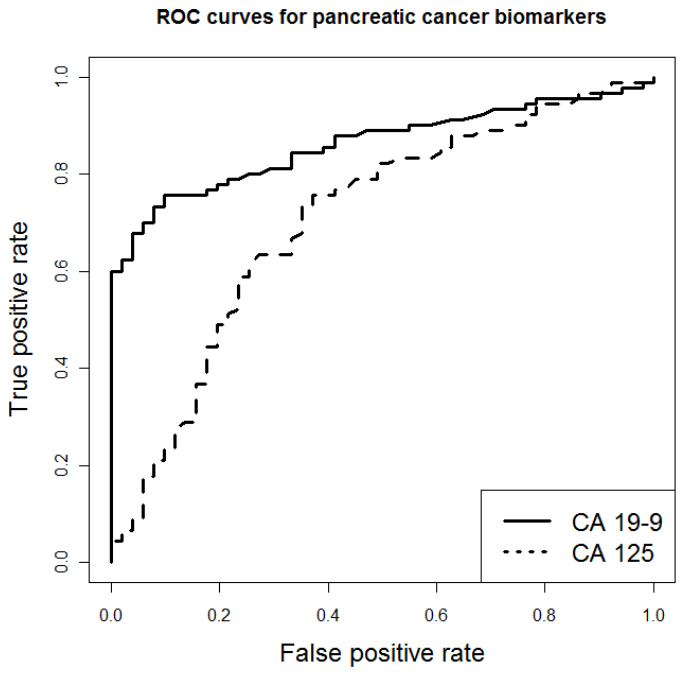
50

Properties of ROC curves

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

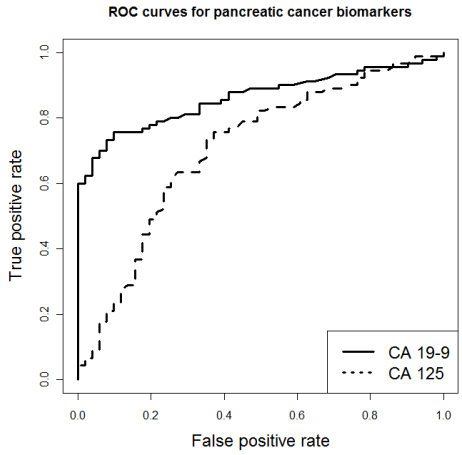
Pancreatic cancer biomarkers (Wieand et al 1989)





53

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



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

54

ROC limitations

- ROC curve summarizes (FPR, TPR) across all possible cut-points for the continuous marker
  - Alternatively, (specificity, sensitivity)
  - Aids in assessing: How well can the marker discriminate between controls and cases ?
- While useful, ROC curves do not contain crucial information
  - Prevalence
  - Value of TP, Cost of FP
- →There is no way to determine an optimal cut-point from an ROC curve

Summarizing ROC Curves: AUC

- AUC is Area under ROC curve
  - AUC sometimes called the c-index or c-statistic
- AUC = $\int_0^1 ROC(t)\, dt$ = average(TPR)
  - average is uniform over (0,1)
- ideal marker: AUC=1.0
- useless marker: AUC=0.5
- A single number summary of a curve is necessarily a crude summary
- Commonly used to compare biomarkers

AUC: probabilistic interpretation

- For a randomly selected case D and a randomly selected control N,

$$AUC = P(Y_D > Y_N)$$

- AUC is interpretable, but its interpretations (as an area; as a probability) show that AUC is not clinically meaningful

RISK PREDICTION

Risk Model: Huntington's Disease

- Huntington's Disease is caused by the *HTT* gene 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 develops* Huntington's Disease in middle-age. The genetic abnormality is dominant ─ one abnormal gene causes disease.
  - *40+ times: always develop HD
  - *36-39 times: might not develop HD (ignoring this small possibility)

Risk Model: Huntington's Disease

- Relevant Population: Individuals with a biological parent who has Huntington's Disease
- Within this population, an individual has a 50% chance of developing HD depending on whether he or she inherited the abnormal or normal *HTT* gene from the affected parent.
- P(D) = ½ = ρ in this population.

Risk Model: Huntington's Disease

- An individual can choose to have their *HTT* gene genotyped. Say HTT=0 means 0 copies of abnormal gene; HTT=1 means 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.

Risk model

- risk prediction model – gives a risk based on 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, we might not think of it as a risk model

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, blood pressure

Risk model examples

- E.g. STS risk score for dialysis following cardiac surgery is formed via:
  - STS risk score = $f(\alpha + \beta_1$ 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)

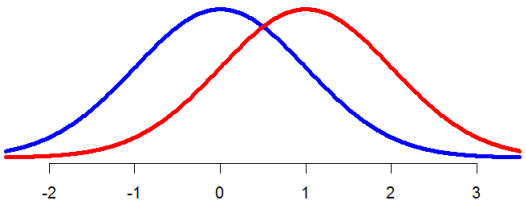
What is “personal risk”?

- 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”?

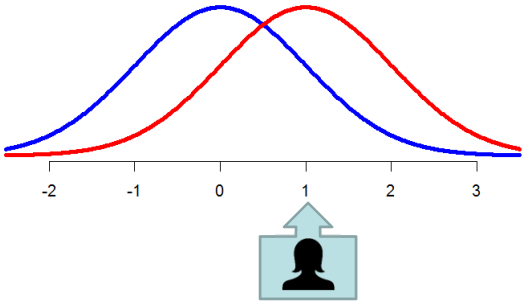
- Suppose the prevalence of D in “Population A” is 1%
  – Without any additional information, the only valid risk prediction instrument is to assign everyone in the population risk=1%
- We have a marker X that tends to be higher in cases than controls



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

What is “personal risk”?

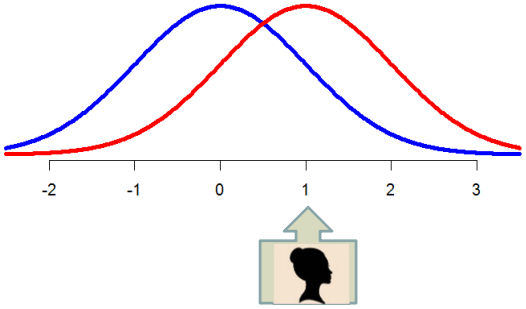
- Alice is an individual in Population A. Alice has X=1.
- We can calculate Alice’s risk(X=1)≈1.6%
  - calculation uses Bayes’ rule



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%.
- Betty, an individual in Population B, has X=1. Betty’s 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 in the sense that it is calculated from personal characteristics
- However, personal risk is not completely divorced from population characteristics. The previous example shows that the population (specifically, the population prevalence) affects “personal” risk.

What is “personal risk”?

- Occasionally one hears 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 from now.
- The best I can objectively 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.”

Summary of Part I

- Example datasets
- FPR (1 – specificity), TPR (sensitivity)
- PPV, NPV
  - function of FPR, TPR and disease prevalence
- ROC curves
- AUC
  - geometric interpretation as area under curve
  - probability interpretation
- A risk model gives population frequencies:
  risk(X)=P(D=1|X)

Misconceptions about Biomarkers and Risk Models

- A large odds ratio means a biomarker is useful for prediction. ✖
- ROC curves are useful to identify the best biomarker cut-point. ✖
- Decision curves are useful to identify the best risk threshold.
- To assess whether to add new biomarker to a risk model, multiple stages of hypothesis testing are needed.
- The best biomarker to improve a risk model is the one with strongest association with the outcome.
- To improve prediction, a new biomarker should be independent of existing predictors.
- We can often use biomarkers to identify which patients will benefit from treatment.