Statistics for Genomic Data Analysis

Hypothesis testing and ROC curves; Multiple hypothesis testing

Decision Truth	not rejected	rejected
true H	☺	✗
false H	✗	☺

http://moodle.epfl.ch/course/view.php?id=15271

Hypothesis testing review

- 2 'competing theories' regarding a population parameter:
  - *NULL* hypothesis *H* ('straw man')
  - ALTERNATIVE hypothesis *A* ('claim', or theory you wish to test)
- *H:* NO DIFFERENCE
  - any observed deviation from what we expect to see is due to *chance variability*
- *A:* THE DIFFERENCE IS *REAL*

Test statistic

- Measure how far the observed data are from what is expected *assuming the NULL H* by computing the value of a *test statistic* (TS) from the data
- The particular TS computed depends on the parameter
- For example, to test the population mean μ, the TS is the *sample mean* (or standardized sample mean)

Example

- An experiment is conducted to study the effect of exercise on the reduction of the cholesterol level in slightly obese patients considered to be at risk for heart attack. 80 patients are put on a specified exercise plan while maintaining a normal diet. At the end of 4 weeks the change in cholesterol level will be noted. It is thought that the program will reduce the average cholesterol reading by more than 25 points.
- Data:
  - sample mean = 27
  - sample SD = 18

Steps in hypothesis testing (I)

1. Identify the population parameter being tested

   Here, the parameter being tested is the population mean cholesterol reading μ

2. Formulate the NULL and ALT hypotheses

   H: $\mu = 25$ *(or* $\mu \le 25$*)*

   A: $\mu > 25$

3. Compute the TS

   $t = (27 - 25)/(18/\sqrt{80}) = .99$

Hypothesis Truth vs. Decision: *ONE* hyp test

Decision Truth	not rejected	rejected
true H	☺ specificity	✗ Type I error (False +) α
false H	✗ Type II error (False -) β	☺ Power 1 - β; sensitivity

Some terminology

- The chance of rejecting a NULL which is *true* is α; this type of mistake is called a *Type I error* or *false positive*
- The chance of not rejecting a NULL which is false is β; this type of mistake is called a *Type II error* or a *false negative*
- In medical contexts, these quantities are referred to with other terminology:
  - The *specificity* of a test is the chance that the test result is negative given that the subject is negative; this is just 1 - α
  - The *sensitivity* of a test is the chance that the test result is positive given that the subject is positive; this is just 1 - β, also called *power*

p-value

- Decide on whether or not to *reject* the NULL hypothesis *H* based on the chance of obtaining a TS *as or more extreme* (as far away from what we expected or even farther, in the direction of the ALT) than the one we got, *ASSUMING THE NULL IS TRUE*
- This chance is called the *observed significance level*, or *p-value*
- A TS with a p-value less than some pre-specified false positive *level* (or *size*) α is said to be 'statistically significant' at that level

p-value interpretation

- The interpretation of a p-value is a little tricky
- In particular, it does ***NOT*** tell us the probability that the NULL hypothesis is true
- The p-value represents the chance that we would see a difference as big as we saw (or bigger) *if* there were really nothing happening other than chance variability
- 'a single convenient number giving a measure of the degree of surprise which the experiment should cause a believer of the null hypothesis' (Hodges and Lehmann)

Steps in hypothesis testing (II)

4. Compute the p-value

Here, p = $P(t_{79} > .99) = .16$

5. (Optional) *Decision Rule:* REJECT H if p-value $\leq \alpha$ (This is a type of argument by contradiction)

A typical value of α is .05, but there's no law that it needs to be. If we use .05, the decision here will be DO NOT REJECT H

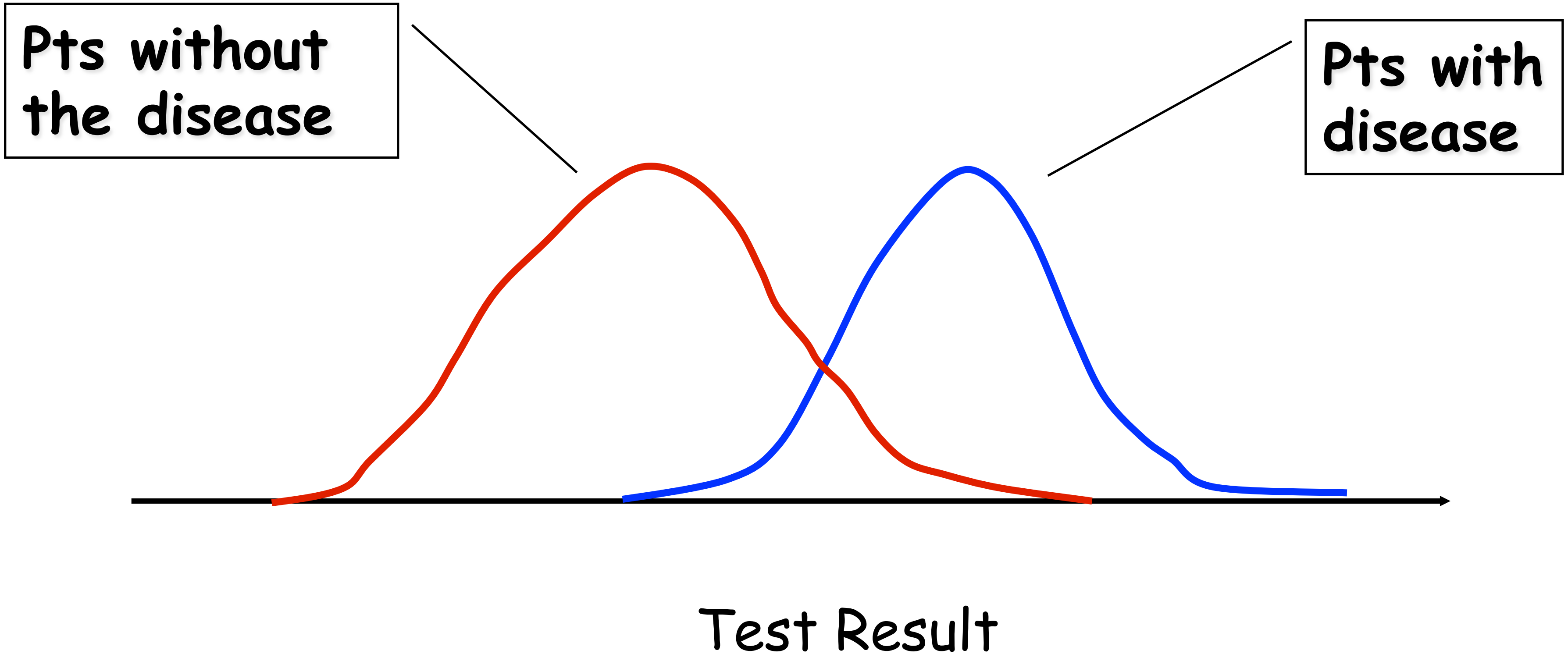
Introduction to ROC curves

- *ROC* = *R*eceiver *O*perating *C*haracteristic
- Started in electronic signal detection theory (1940s - 1950s)
- Has become very popular in biomedical applications, particularly radiology and imaging
- Also used in machine learning applications to assess classifiers
- Can be used to *compare tests*/procedures

ROC curves: simplest case

- Consider diagnostic test for a disease
- Test has 2 possible outcomes:
  - ‘postive’ = suggesting presence of disease
  - ‘negative’
- An individual can test either positive or negative for the disease

Specific Example

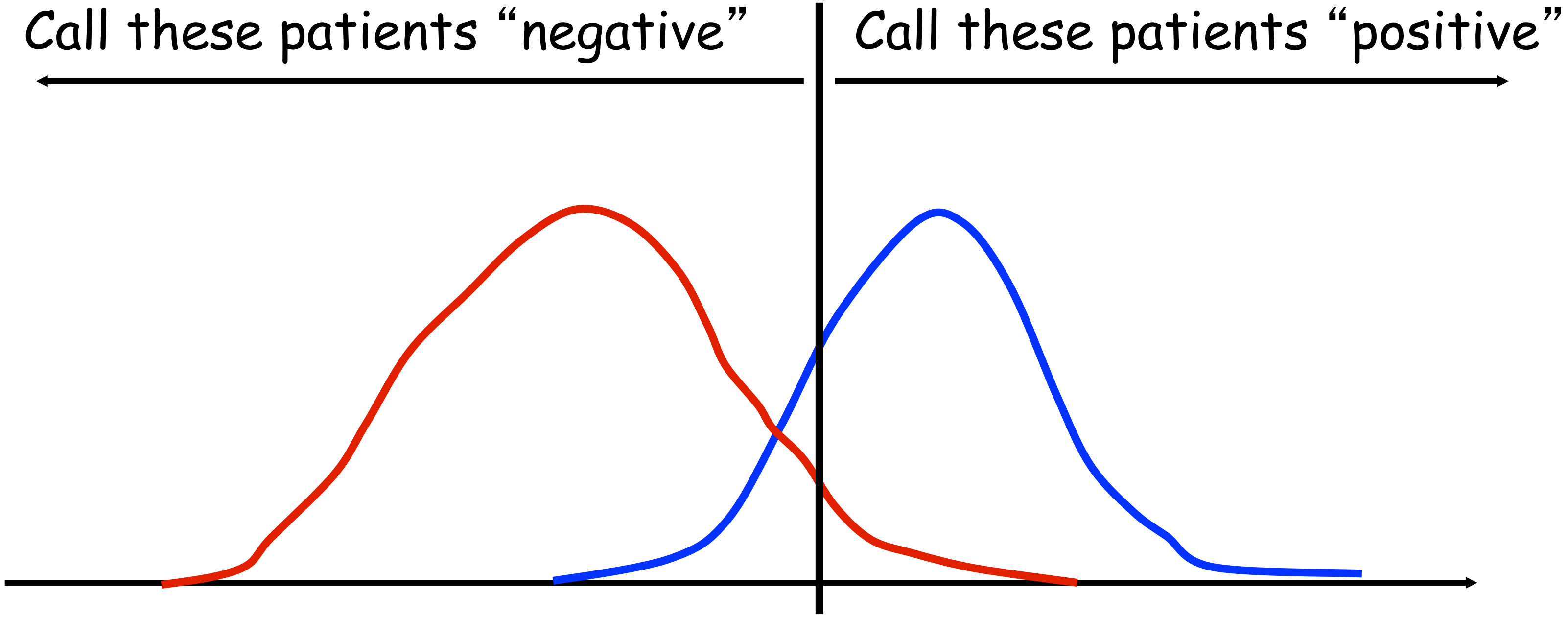


Pts without the disease

Pts with disease

Test Result

Threshold

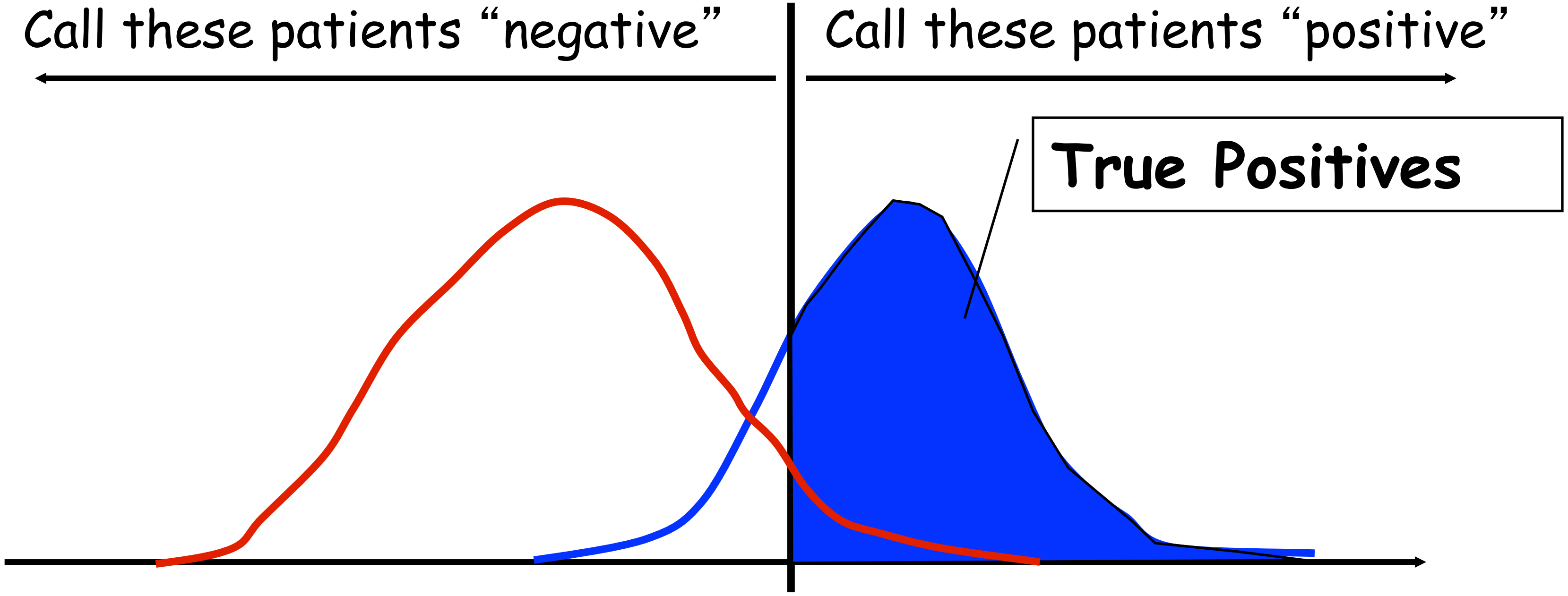


Call these patients “negative”

Call these patients “positive”

Test Result

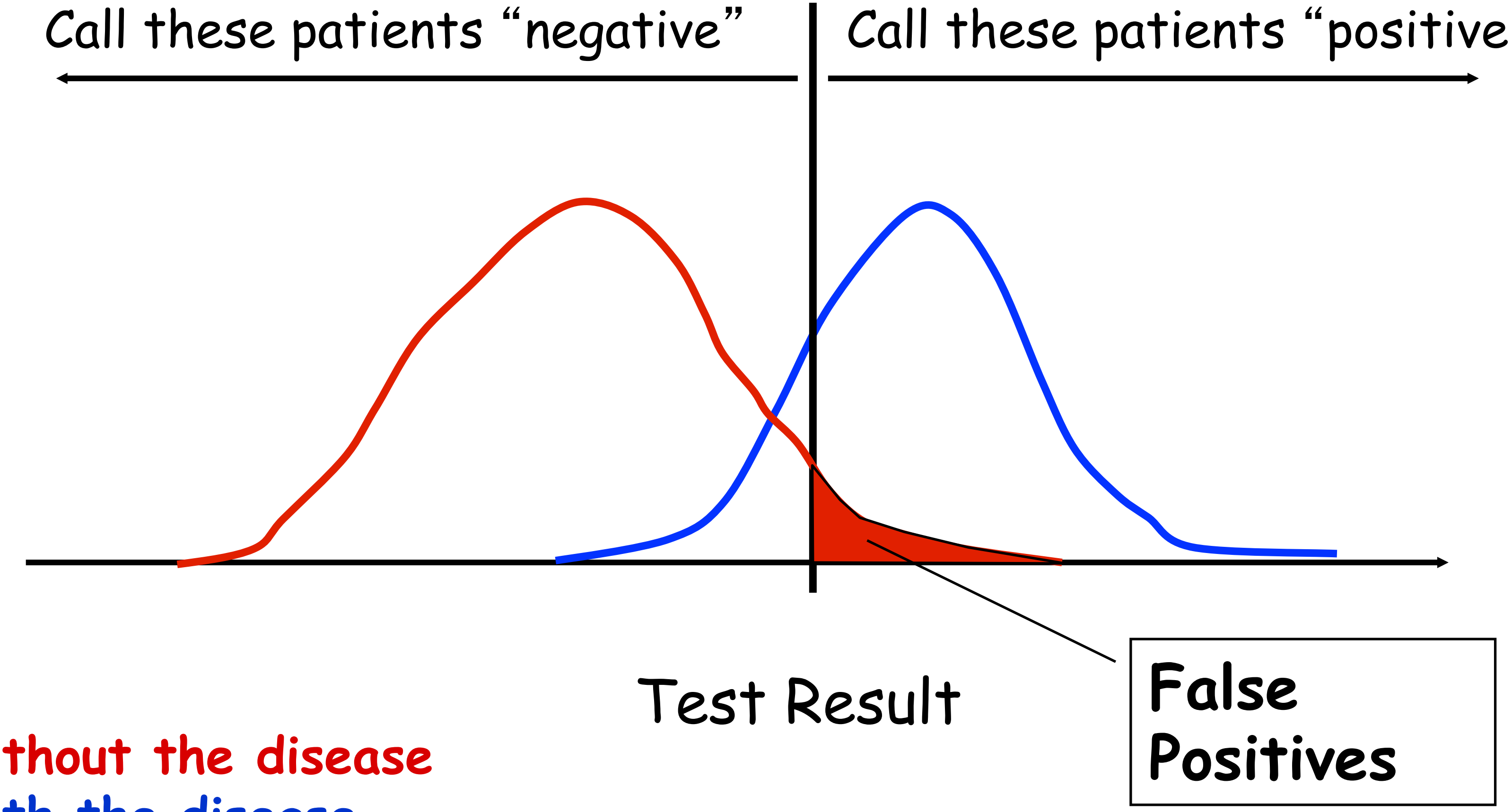
Some definitions ...


Call these patients "negative"
Call these patients "positive"
True Positives

Test Result

without the disease

with the disease



Call these patients "negative"

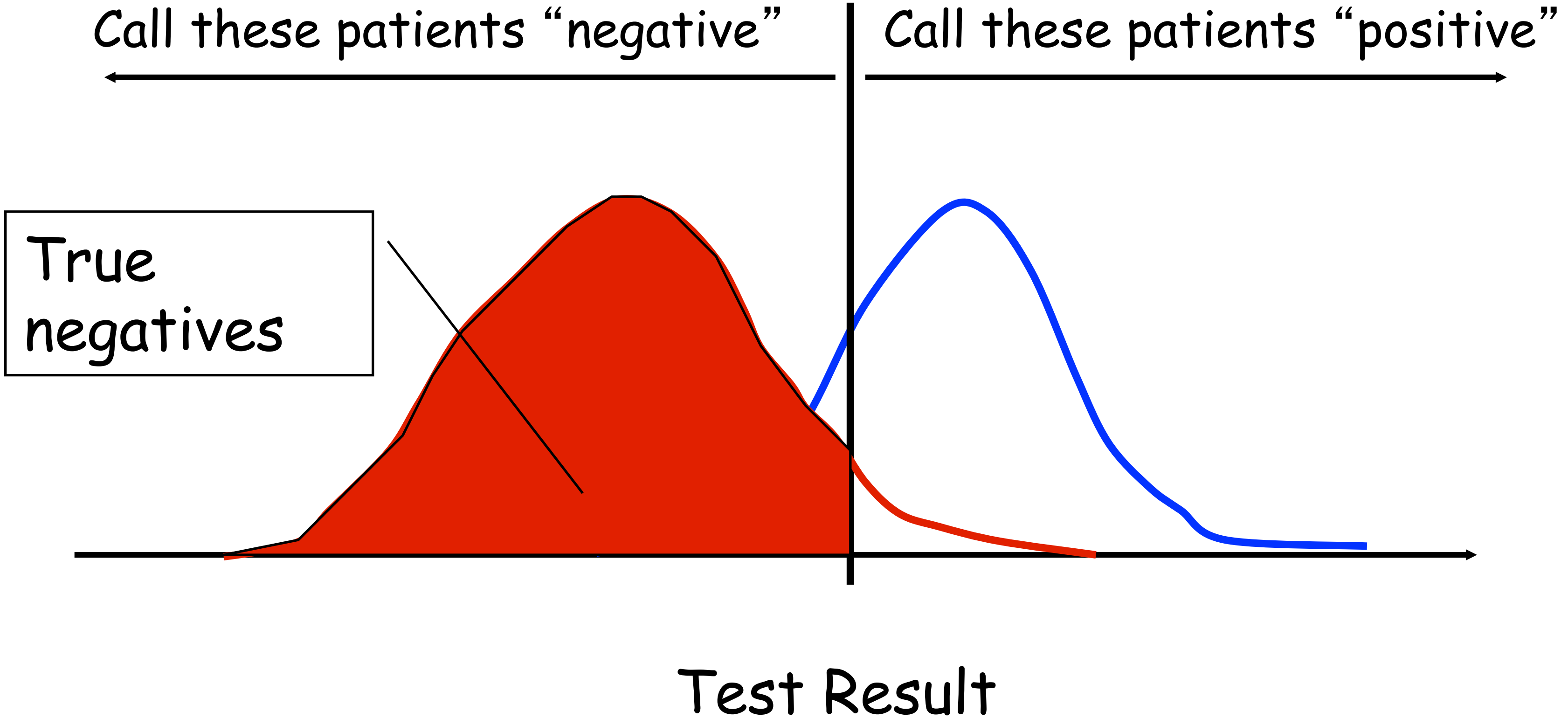
Call these patients "positive"

Test Result

False Positives

without the disease

with the disease



Call these patients “negative”

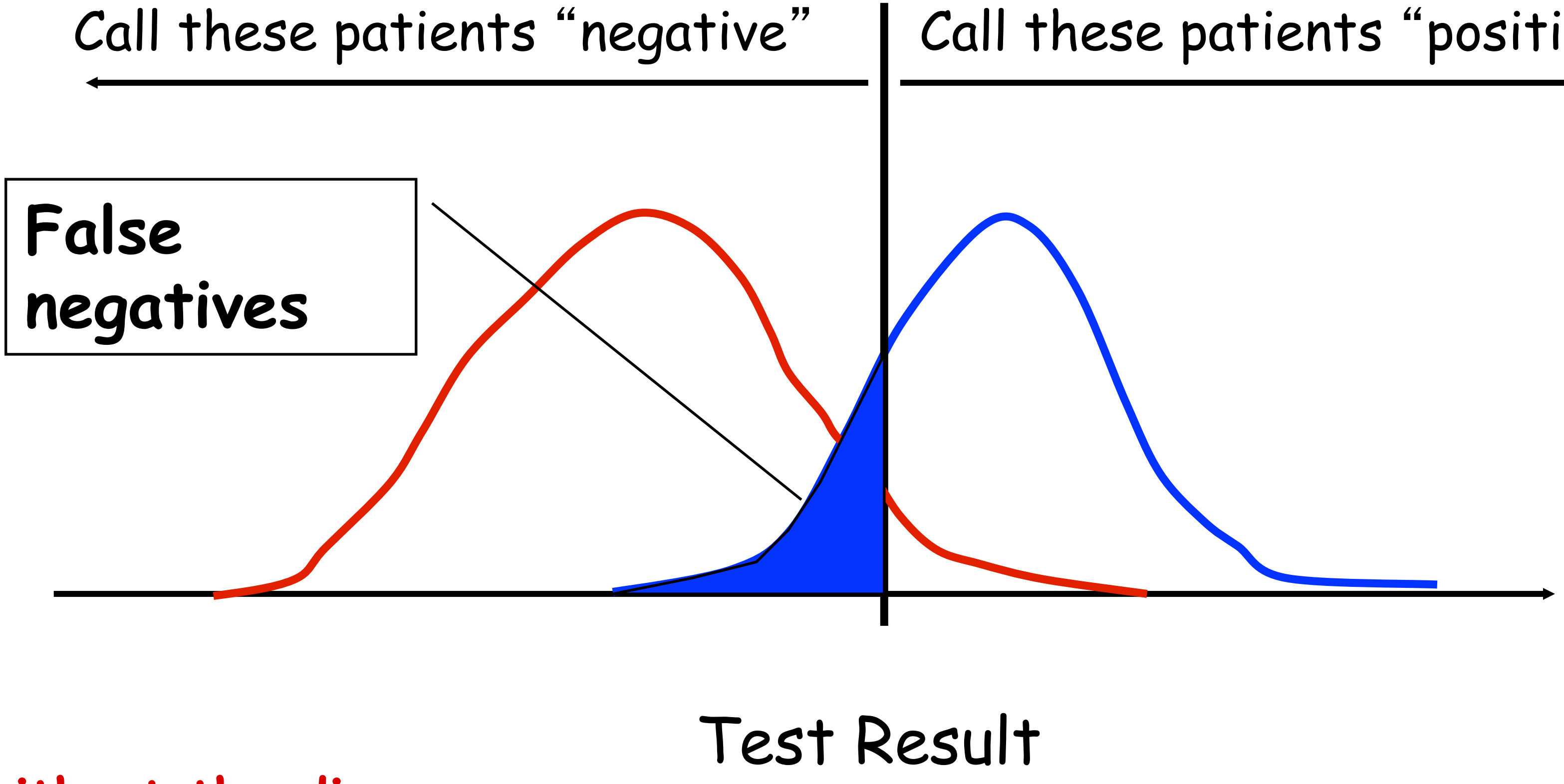
Call these patients “positive”

True negatives

Test Result

without the disease

with the disease



Call these patients "negative"

Call these patients "positive"

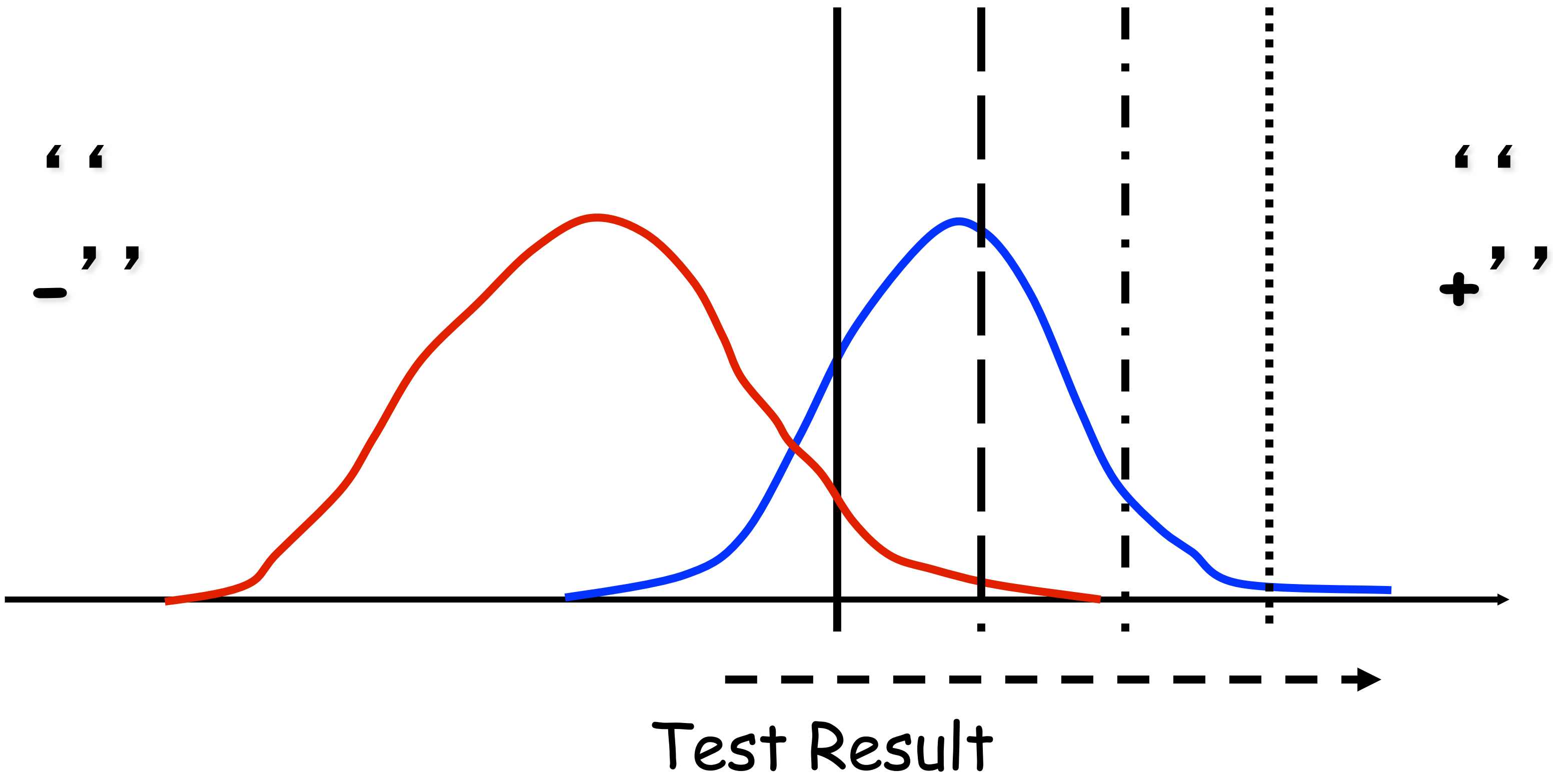
False negatives

Test Result

without the disease

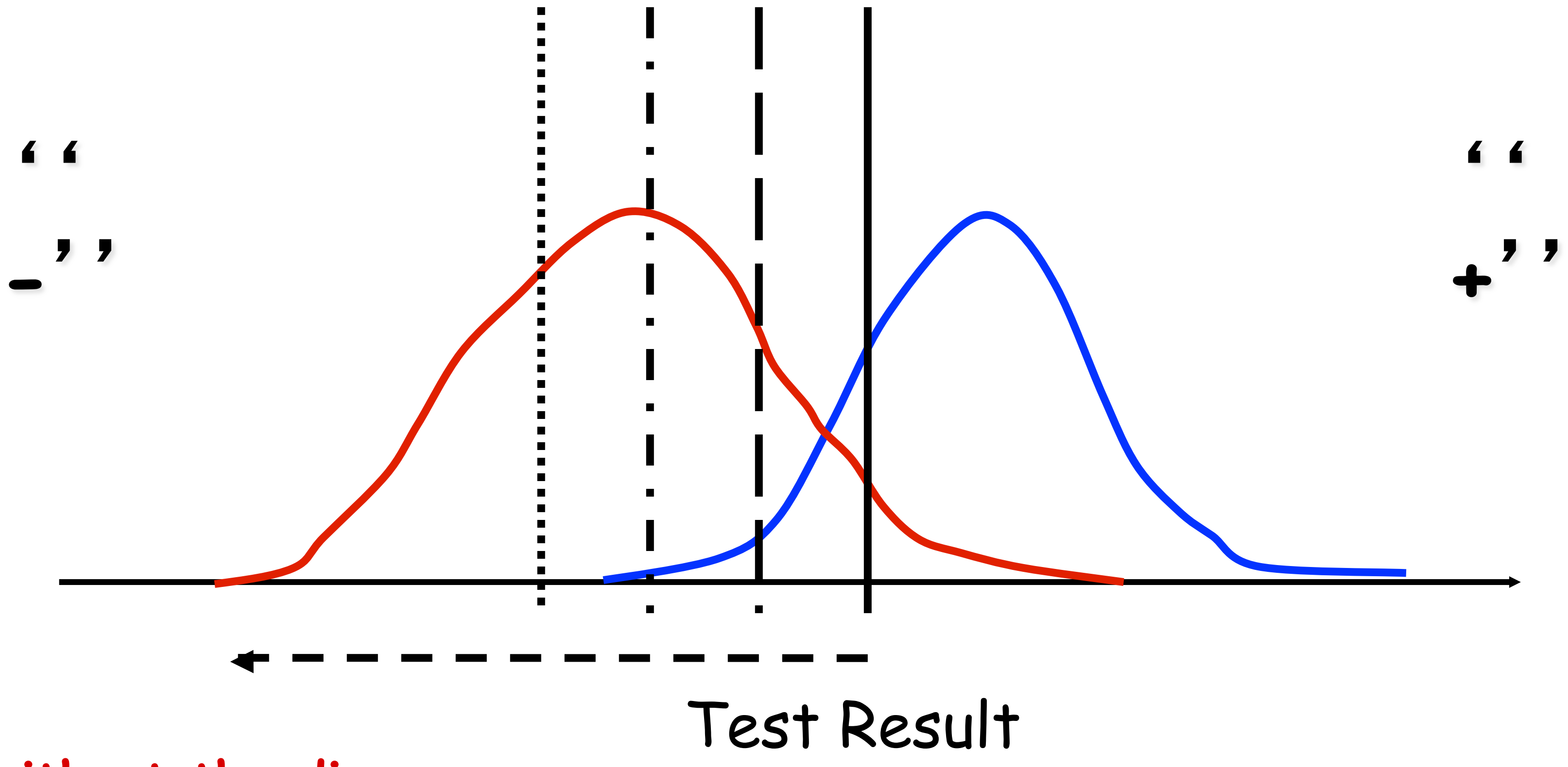
with the disease

Moving the Threshold: right



without the disease
with the disease

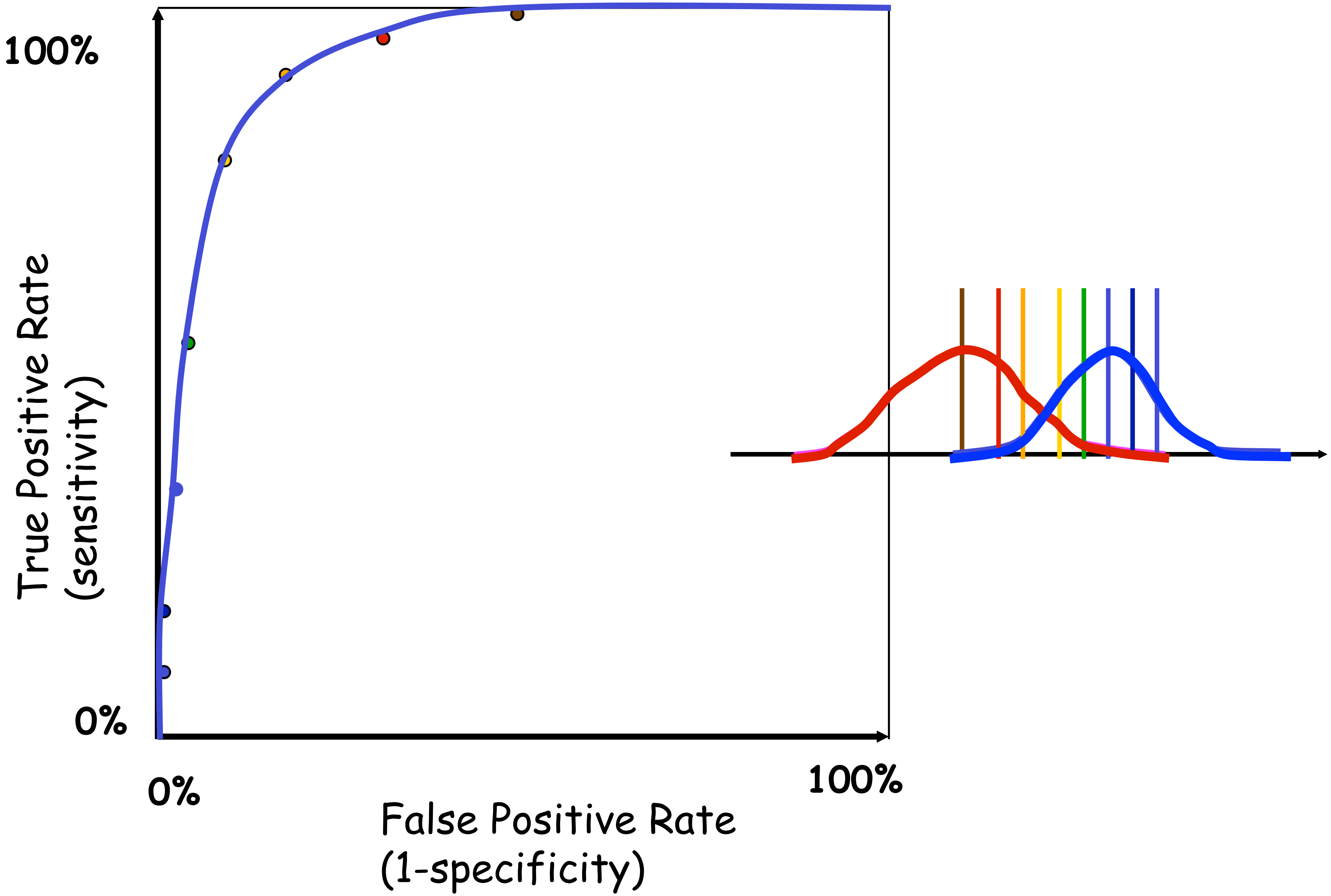
Moving the Threshold: left



without the disease

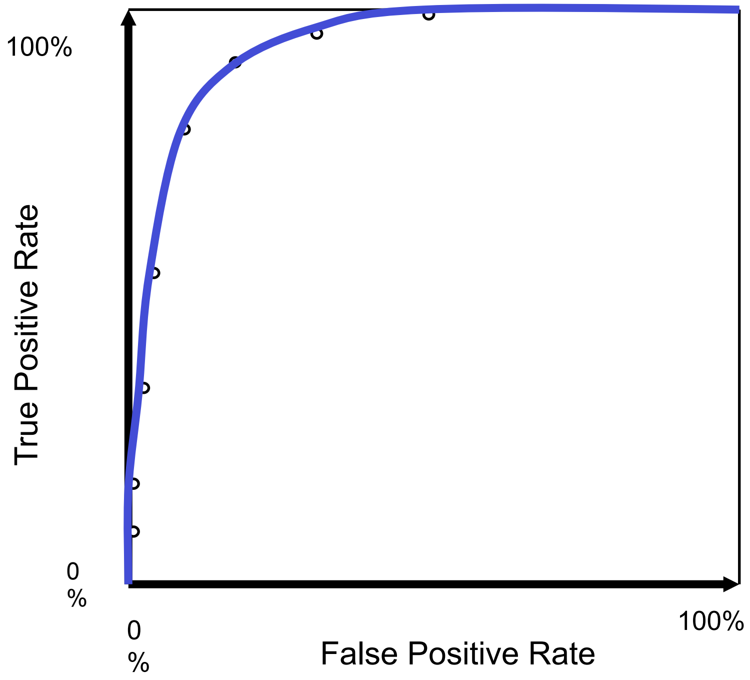
with the disease

ROC curve

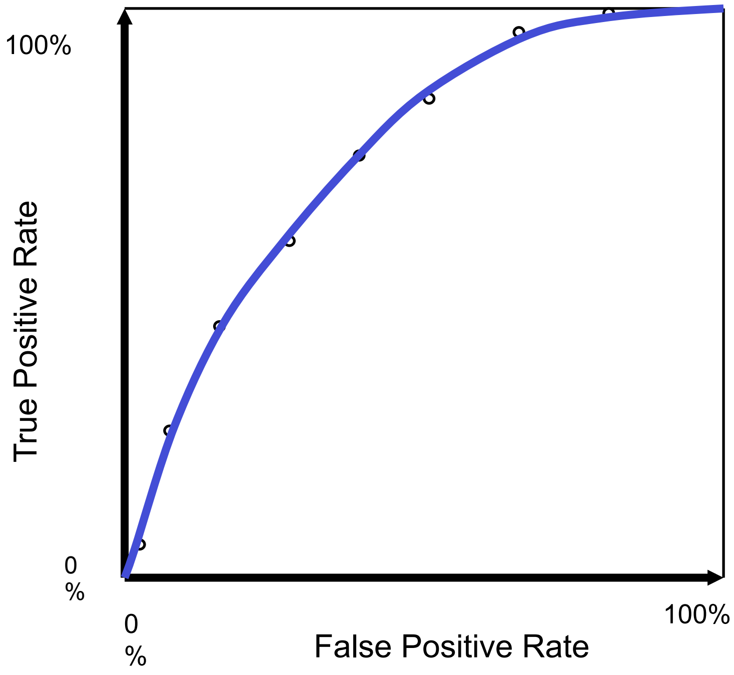

100%
0%
True Positive Rate
(sensitivity)
0%
100%
False Positive Rate
(1-specificity)

ROC curve comparison

A good test:

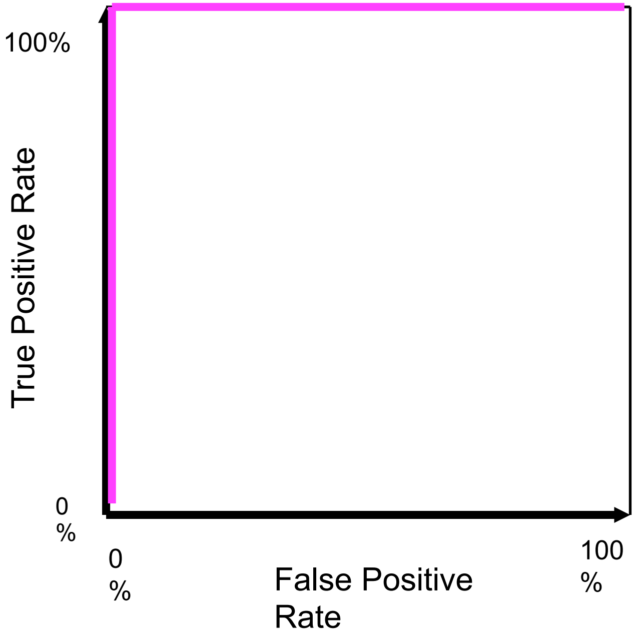


A poor test:



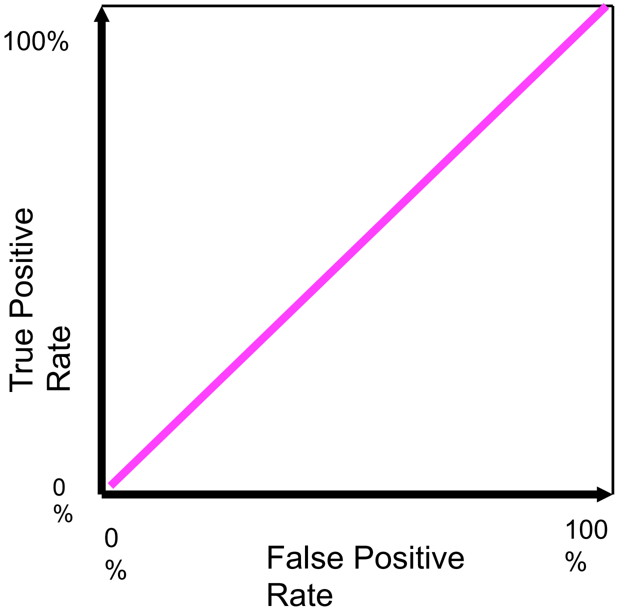
ROC curve extremes

Best Test:



The distributions don't overlap at all

Worst test:

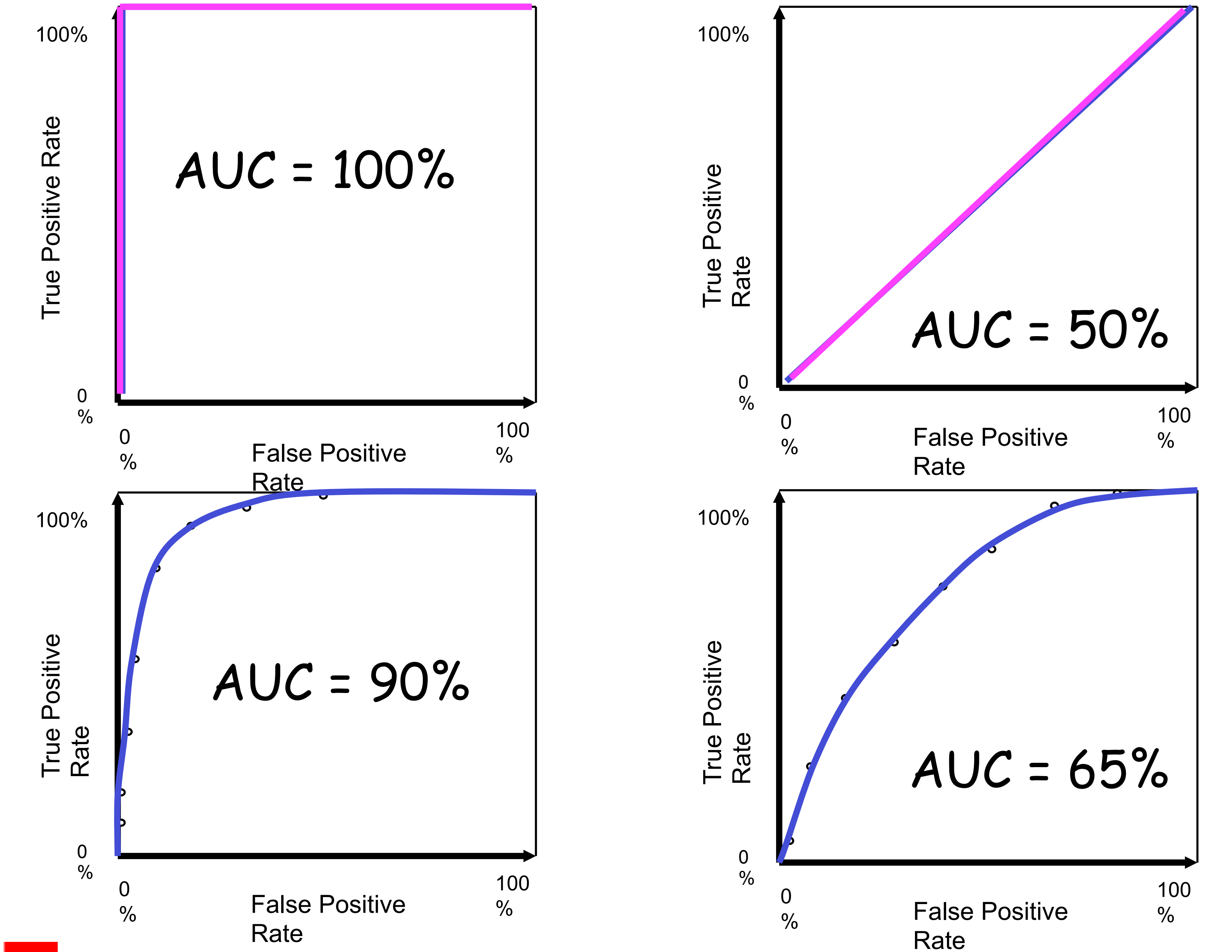


The distributions overlap completely

Area under ROC curve (AUC)

- *Overall measure* of test performance
- *Comparisons* between two tests based on differences between (estimated) AUC
- For continuous data, AUC equivalent to *Mann-Whitney U-statistic* (nonparametric test of difference in location between two populations)

AUC for ROC curves



Interpretation of AUC

- AUC can be interpreted as the probability that the test result from a randomly chosen diseased individual is more indicative of disease than that from a randomly chosen nondiseased individual: $P(X_i \geq X_j \mid D_i = 1, D_j = 0)$
- So can think of this as a nonparametric distance between disease/nondisease test results

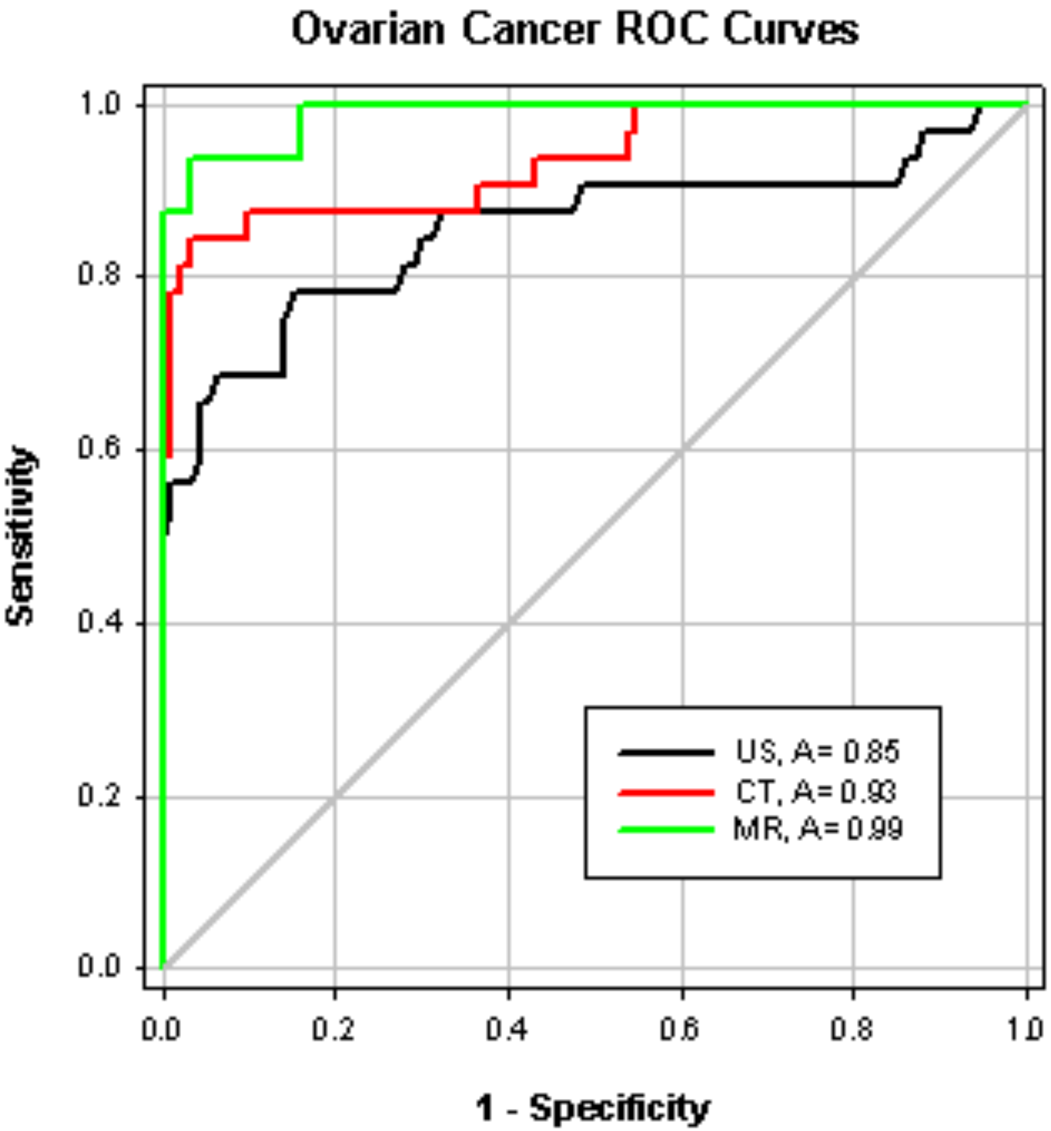
Problems with AUC

- *No clinically relevant meaning*
- A lot of the area is coming from the range of *large false positive* values, no one cares what's going on in that region (need to examine restricted regions)
- The curves might *cross*, so that there might be a meaningful difference in performance that is not picked up by AUC

Examples using ROC analysis

- Threshold selection for ‘tuning’ an already trained classifier (e.g. neural nets)
- Defining signal thresholds in DNA microarrays (Bilban *et al.*)
- Comparing test statistics for identifying differentially expressed genes in replicated microarray data (Lönnstedt and Speed)
- Assessing performance of different protein prediction algorithms (Tang *et al.*)
- Inferring protein homology (Karwath and King)

Ovarian cancer ROC curves



(BREAK)

Multiple testing problem

- Simultaneously test *m* null hypotheses, one for each gene *j*

  H_j: *no association* between expression measure of gene *j* and the covariate

- Because microarray experiments simultaneously monitor expression levels of thousands of genes, there is a *large multiplicity issue*

- Increased chance of *at least one false positive result*

- Would like some sense of how 'surprising' the observed results are

Hypothesis Truth vs. Decision: *m* tests

Decision Truth	# not rejected	# rejected	Totals
# true H	U	V (F +) Type I error	m_0
# non-true H	T Type II error	S	m_1
totals	m - R	R	m

Hypothesis Truth vs. Decision: *m* tests

Decision Truth	# not rejected	# rejected	Totals
# true H	U	V (F +) Type I error	m_0
# non-true H	T Type II error	S	m_1
totals	W (= m - R)	R	m

Random Variables

constants

Hypothesis Truth vs. Decision: *m* tests

Unobservable

Decision Truth	# not rejected	# rejected	Totals
# true H	U	V (F +) Type I error	m_0
# non-true H	T Type II error	S	m_1
totals	W (= m - R)	R	m

Observable

Type I (false positive) error rates

- *Per-family Error Rate*

  $$\mathrm{PFER} = E(V)$$

- *Per-comparison Error Rate*

  $$\mathrm{PCER} = E(V)/m$$

- *Family-wise Error Rate*

  $$\mathrm{FWER} = p(V \geq 1)$$

- *False Discovery Rate*

  $\mathrm{FDR} = E(Q)$, where

  $Q = V/R$ if $R > 0$; $Q = 0$ if $R = 0$

Strong vs. weak control

- All probabilities are *conditional* on which hypotheses are true
- *Weak control* refers to control of the Type I error rate only under the *complete null hypothesis* (i.e. *all* nulls true)
- *Strong control* refers to control of the Type I error rate under *any combination* of true and false nulls
- In general, *weak control* without other safeguards is *unsatisfactory*

Adjusted p-values (p*)

- *Test level* (*e.g.* 0.05) does not need to be determined in advance
- Some procedures *most easily described* in terms of their adjusted *p*-values
- Usually *easily estimated using resampling*
- Procedures can be *readily compared* based on the corresponding adjusted *p*-values

A Little Notation

- For hypothesis H_j, $j = 1, \ldots, m$

  observed test statistic: t_j

  observed *unadjusted* (nominal) p-value: p_j

- Ordering of observed (absolute) t_j: $\{r_j\}$ such that $|t_{r_1}| \geq |t_{r_2}| \geq \ldots \geq |t_{r_m}|$

- Ordering of observed p_j: $\{r_j\}$ such that $|p_{r_1}| \leq |p_{r_2}| \leq \ldots \leq |p_{r_m}|$

- Denote corresponding RVs by upper case letters (T, P)

Methods for obtaining p*

- *Single-step* adjustment
  - *p*-values compared to a *predetermined value*
  - *same adjustment* for every *p*-value
- *Step-down* adjustment
  - p-values adjusted from smallest to largest
  - when find 'large' *p*-value, that null and all nulls with larger *p*-values are not rejected
- *Step-up* adjustment
  - p-values adjusted from largest to smallest
  - when find 'small' p-value, that null and all nulls with smaller p-values are rejected

- One-step Bonferroni: p#(i)=n*p(i)
- One-Step Sidak: p#(i)=1-(1-p(i))^n

- - Step-down Holm: p#(i)=(n-i+1)*p(i)
- - Step-down Sidak: p#(i)=1-(1-p(i))^(n-i+1)

- - Step-up Hommel: p#(i)=n*Cn*p(i)/i , with Cn=1+1/2+ ... +1/n
- - Step-up Hochberg: p#(i)=(n-i+1)*p(i)
- - Step-up Simes: p#(i)=n*p(i)/i

Control of the FWER

- *Bonferroni single-step* adjusted p-values

  $$p_j^* = \min(mp_j, 1)$$

- *Sidak single-step (SS)* adjusted p-values

  $$p_j^* = 1 - (1 - p_j)^m$$

- *Sidak free step-down (SD)* adjusted p-values

  $$p_{(j)}^* = 1 - (1 - p_{(j)})^{(m - j + 1)}$$

Control of the FWER

- *Holm (1979) step-down* adjusted p-values

  $$p_{r_j}^* = \max_{k = 1\ldots j} \{\min((m-k+1)p_{r_k}, 1)\}$$

  - *Intuitive explanation*: once $H_{(1)}$ rejected by Bonferroni, there are only m-1 remaining hyps that might still be true (then another Bonferroni, *etc.*)

- *Hochberg (1988) step-up* adjusted p-values (Simes inequality)

  $$p_{r_j}^* = \min_{k = j\ldots m} \{\min((m-k+1)p_{r_k}, 1)\}$$

Control of the FWER

- Westfall & Young (1993) step-down minP adjusted p-values

  $p_{r_j}^* = \max_{k = 1 \ldots j} \{ p(\min_{l \in \{k,\ldots,m\}} P_{r_l} \leq p_{r_k} \mid H_0^C) \}$

- Westfall & Young (1993) step-down maxT adjusted p-values

  $p_{r_j}^* = \max_{k = 1 \ldots j} \{ p(\max_{l \in \{k,\ldots,m\}} |T_{r_l}| \geq |t_{r_k}| \mid H_0^C) \}$

Westfall & Young (1993) Adjusted *p*-values

- Step-down procedures: successively *smaller adjustments* at each step
- Take into account *joint distribution* of test stats.
- Takes dependence structure between genes into account, which gives in many cases (positive dependence between genes) *higher power*
- *Less conservative* than Bonferroni, Sidak, Holm, or Hochberg adjusted *p*-values
- Estimated by *resampling* - computer-intensive (especially for minP)

maxT vs. minP

- The maxT and minP adjusted p-values are the *same* when the test statistics are identically distributed (id)
- When the test statistics are not id, maxT adjustments may be *unbalanced* (not all tests contribute equally to the adjustment)
- maxT *more computationally tractable* than minP
- maxT can be *more powerful* in ‘small n, large m’ situations

Control of the FDR

- *Benjamini & Hochberg (1995): step-up* procedure which controls the FDR under some dependency structures

$$p_{r_j}^* = \min_{k = j \ldots m} \{ \min ([m/k]\, p_{r_k}, 1) \}$$

- *Benjamini & Yuketieli (2001): conservative step-up* procedure which controls the FDR under general dependency structures

$$p_{r_j}^* = \min_{k = j \ldots m} \{ \min (m\Sigma_{j=1}^{m}[1/j]/k]\, p_{r_k}, 1) \}$$

- *Yuketieli & Benjamini (1999):* resampling based adjusted p-values for controlling the FDR under certain types of dependency structures

q-value

- *p-value:* the smallest *false positive rate* for which the test can be called 'significant'

- *q-value:* the smallest *false discovery rate* for which the test can be called 'significant' (Storey *et al.*)

R: multiple testing

- The BioConductor package **multtest** has implemented a number of adjustments for multiple hypotheses
- The package vignette reviews the functionality
- **limma** also adjusts for multiplicity
- The BioConductor package **qvalue** computes q-values and various plots
- http://www.bioconductor.org

What about pre-screening?

- To get around the problem of loss of power when adjusting, some have recommended *'pre-screening'* expression levels and only testing those showing sufficient variation
- This is an example of *'data snooping'* : looking at the data before deciding what to test
- Unless the screening statistic is *independent of the test statistic under the null*, the Type I error rate ***will not be correct***
- In addition, any *p*-value for the test may be *difficult to interpret*

Controversies

- *Whether* multiple testing methods (adjustments) should be applied at all
- *Which tests* should be included in the *family* (e.g. all tests performed within a single experiment; define 'experiment')
- Alternatives
  - Bayesian approach
  - Meta-analysis

Situations where inflated error rates are a concern

- It is plausible that *all nulls may be true*
- A *serious claim* will be made whenever any p < .05 (say) is found
- *Much data manipulation* may be performed to find a 'significant' result
- The analysis is planned to be *exploratory* but wish to claim 'sig' results are real
- Experiment *unlikely to be followed up* before serious actions are taken

Steps in hypothesis testing

1. Identify the *population parameter* being tested (*e.g.* population mean)
2. Formulate the *NULL* (H) and *ALT* (A) hypotheses
3. Compute an appropriate *TS (Test Statistic)*
4. Compute the *p-value*
5. (Optional) *Decision Rule:* REJECT H if the p-value ≤ α

Computing the *p-value*

- For a *parametric* hypothesis test, we start with some *assumptions* to derive the *sampling distribution of the TS* assuming that the NULL hypothesis is true

- *Example:* If our samples come from a normal (Gaussian) distribution with a known SD, the sampling distribution of the (standardized) sample mean is also normal

- Use this sampling distribution to get the *p-value:* chance of obtaining a TS *as or more extreme* than the one we got, *ASSUMING THE NULL IS TRUE*

Permutation test

- A type of *nonparametric hypothesis test*
- Also called *randomization test*, *rerandomization test*, *exact test*
- Very widely applicable class of tests
- Introduced in the 1930s (that's right, before everyone had a desktop/laptop computer!)
- Usually require only a few weak assumptions
- These tests often have good power
- 'I lost the labels' story

5 Steps to a permutation test

1. Analyze the problem: identify the NULL and ALT hypotheses
2. Choose a test statistic (TS)
3. Compute the TS for the original labeling of the observations
4. *** *Rearrange (permute) the labels and recompute the TS for the rearranged labels* (do for all possible permutations) ***
5. Decide whether to reject NULL based on this *permutation distribution*

Permutations

- A *permutation* is a reordering of the numbers 1, ..., *n*
- *Example:* What are some permutations of the numbers 1, 2, 3, 4*??*
- The *NULL* specifies that the permutations are *all equally likely*
- The sampling distribution of the TS under the NULL is computed by forming all permutations, calculating the TS for each and considering these values all equally likely

Example

- How could we carry out a permutation test to test the NULL hypothesis of no difference between treated and untreated ??

Treated	121	118	110	90
Untreated	95	34	22	12

Step 4: reference NULL

- This is an example of an *unpaired 2-sample test*
- Here, we have to find all of the combinations (since order within each group doesn't matter)
- These are …
- *Handout: What's a p-value?*

Advantages

- Can get a permutation test for any TS, even if it’s sampling distribution is unknown
- This gives more freedom in choosing a TS
- Can use on unbalanced designs
- Can combine dependent tests on mixtures of different data types (*e.g.* with numerical and categorical data)

Limitations

- Assumption that the observations are *exchangeable under the NULL*
- This assumption is what allows us to randomly move observations between the groups
- For example, when testing for a difference in 2 group means you would need to assume that the distributions in both groups have *the same shape and spread*
- *Cannot use* for testing hypotheses in a single population, or to compare groups that are different under the NULL

Sampling permutations

- What if the total number of possible permutations is too large for complete enumeration?
- Use *Monte Carlo sampling:* that is, randomly select some of the permutations
- Monte Carlo methods are based on the use of random numbers and probability to investigate problems
- The number of permutations to sample depends on desired accuracy

Pitfalls in hypothesis testing

- Even if a result is ‘statistically significant’, *it can still be due to chance*
- Statistical significance is not the same as *practically meaningful*
- A test of significance does not say *how meaningful* the difference is, or *what caused it*
- A test does not check the study *design*
- If the test is applied to a *nonrandom sample* (or the whole population), the resulting *p*-value is generally *meaningless*
- *Data-snooping* makes *p*-values hard to interpret