

Statistical Evaluation of Diagnostic Tests: ROC curves, sensitivity, and specificity

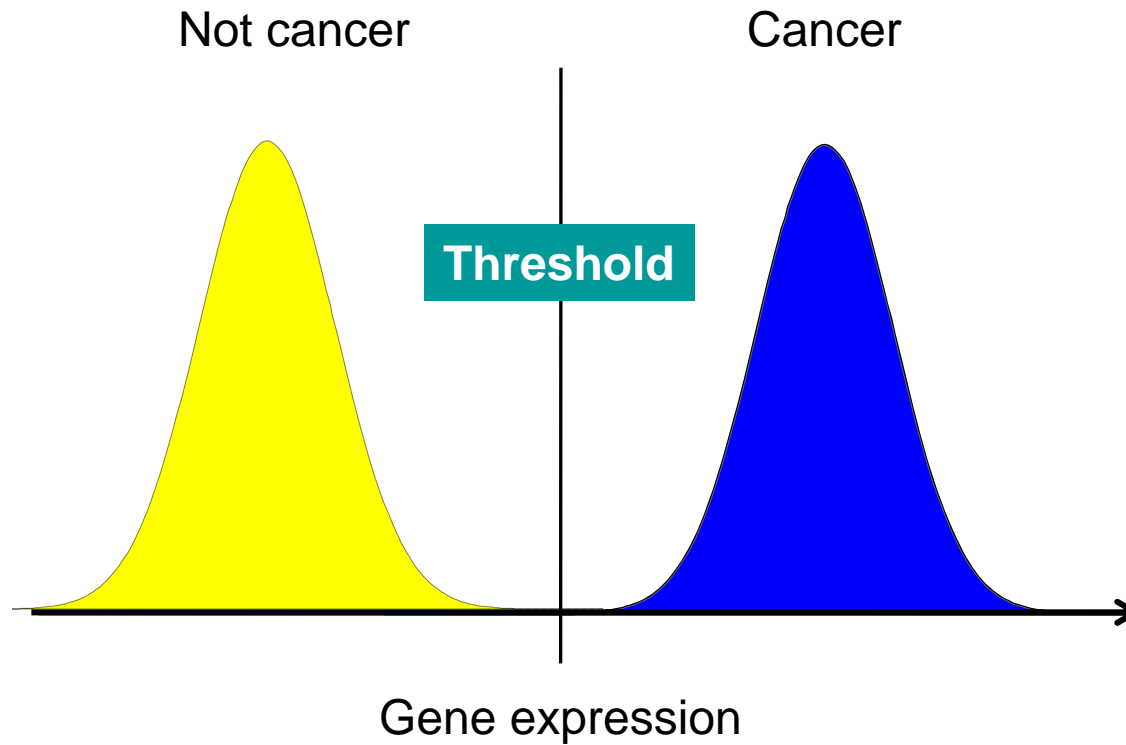
Michael Walker

mwalker@stanfordalumni.org

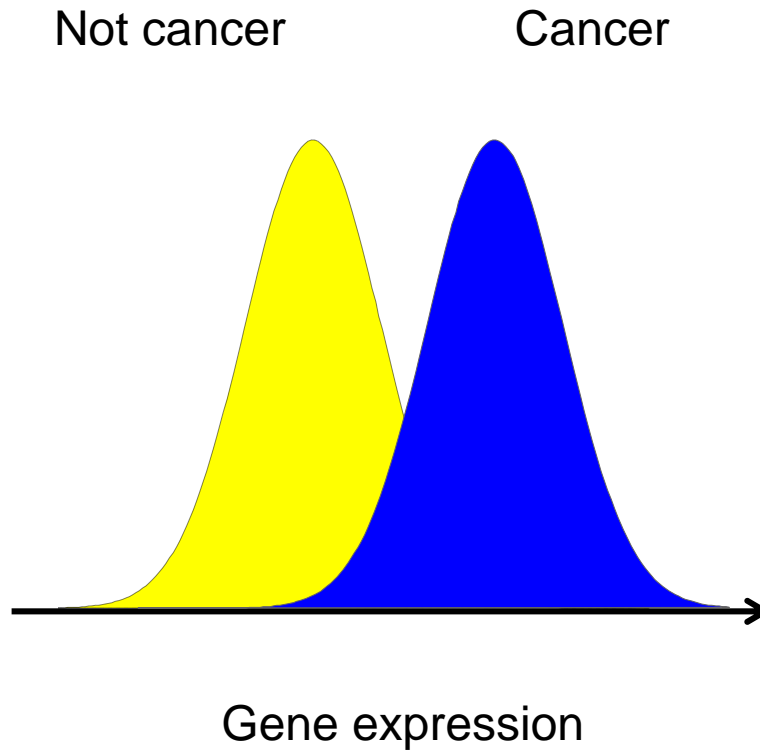
Clinical objectives

- Describe the diagnostic accuracy of a test.
- Compare accuracy of different tests.

- Suppose we wish to know if the expression level of a gene in a tumor can predict if patients will have a recurrence of their cancer.
- High expression => high probability of recurrence



In this case, gene expression separates cancer from non-cancer perfectly.



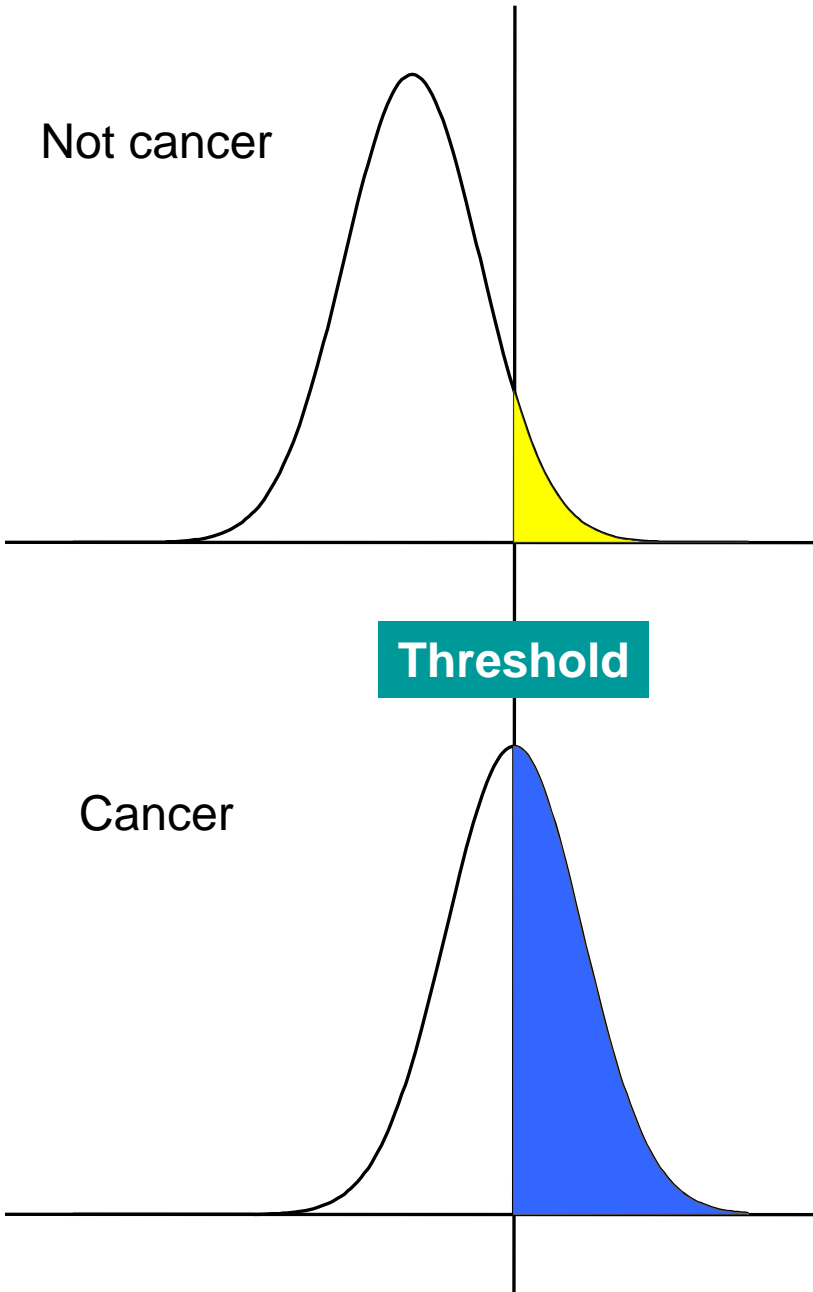
Gene expression levels for cancer and non-cancer overlap.
What threshold should we choose to predict recurrence?

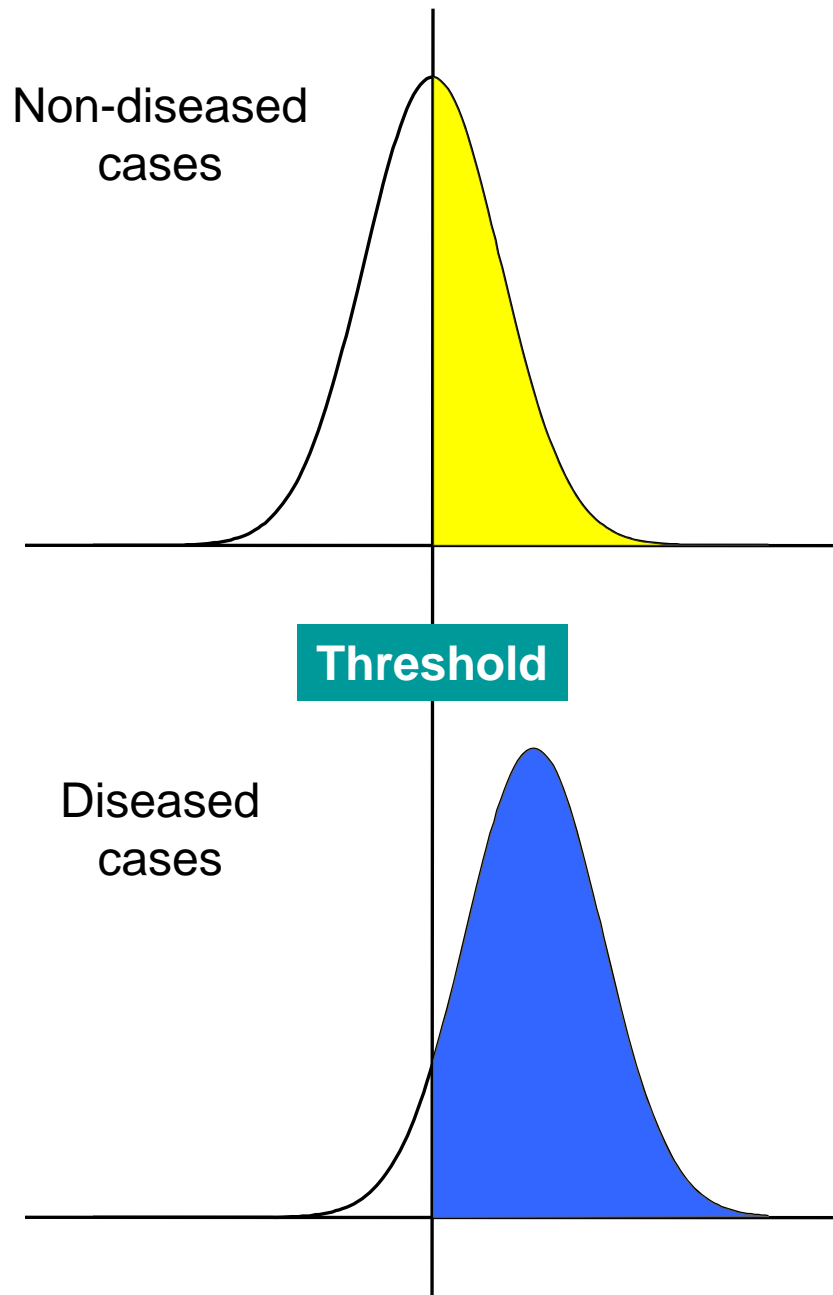
Not cancer

Choose a high threshold:
Few false positive predictions,
but
lots of false negatives

Threshold

Cancer





Choose a low threshold:
Many false positive predictions,
but
few false negatives

Notation for conditional probability

- Suppose that the patient has the disease (according to gold standard)
- Notation to specify the probability that the test for the patient is positive, given that the patient has the disease:

$P(\text{Test positive} \mid \text{patient has disease})$

$P(T+ \mid D+)$

- This notation describes the conditional probability.

Sensitivity and Specificity

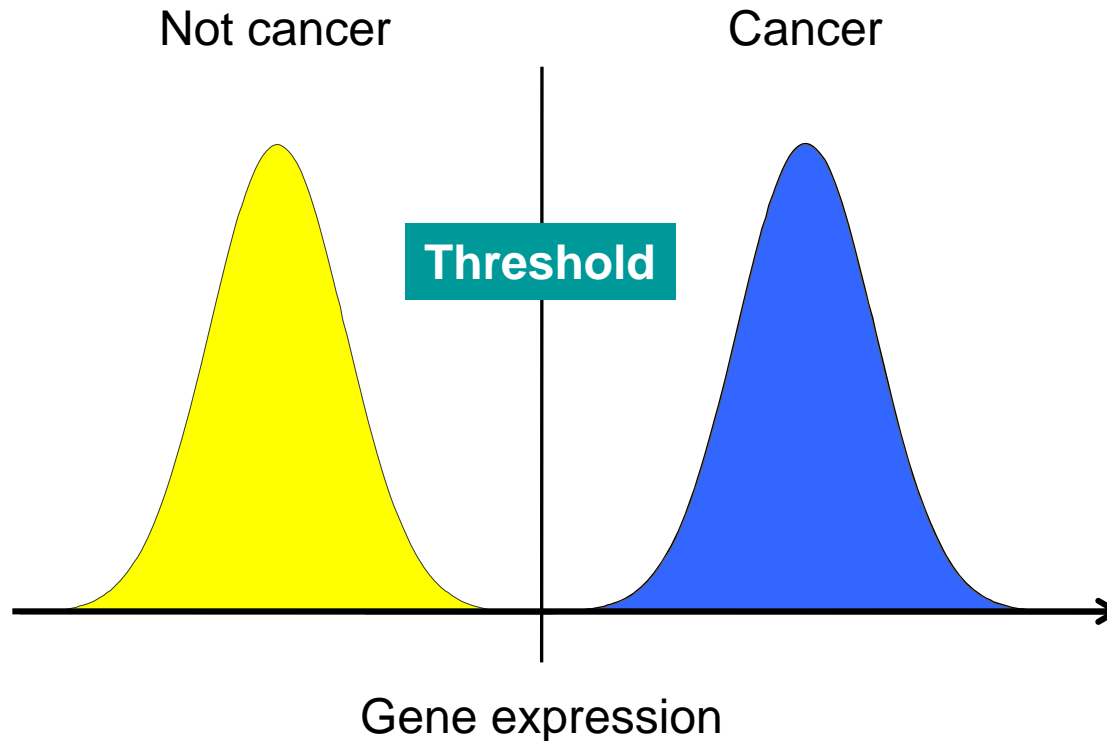
- We want the test to be positive when the patient has the disease:
- Sensitivity
= $P(\text{Test positive} \mid \text{patient has disease})$
- We want the test to be negative when the patient does not have the disease:
- Specificity
= $P(\text{Test negative} \mid \text{patient does not have disease})$

Sensitivity and Specificity

- **Sensitive** => find **ALL** disease
- Sensitivity
= $P(\text{Test positive} \mid \text{patient has disease})$
- **Specific** => find **ONLY** disease
- Specificity
= $P(\text{Test negative} \mid \text{patient does not have disease})$

- Sensitivity and specificity depend on
 - How well the test separates the two groups
 - What threshold we choose

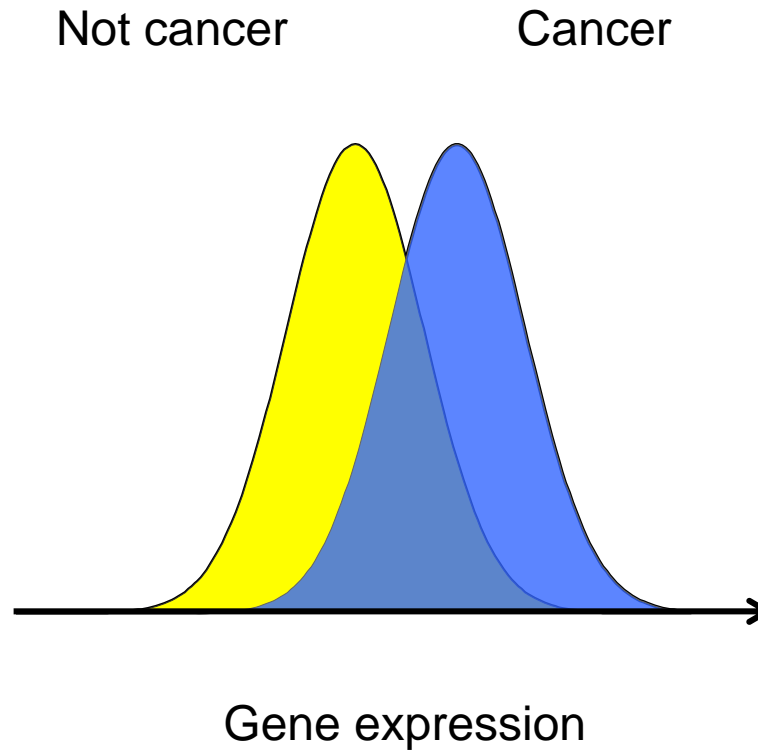
If gene expression separates cancer from non-cancer perfectly:



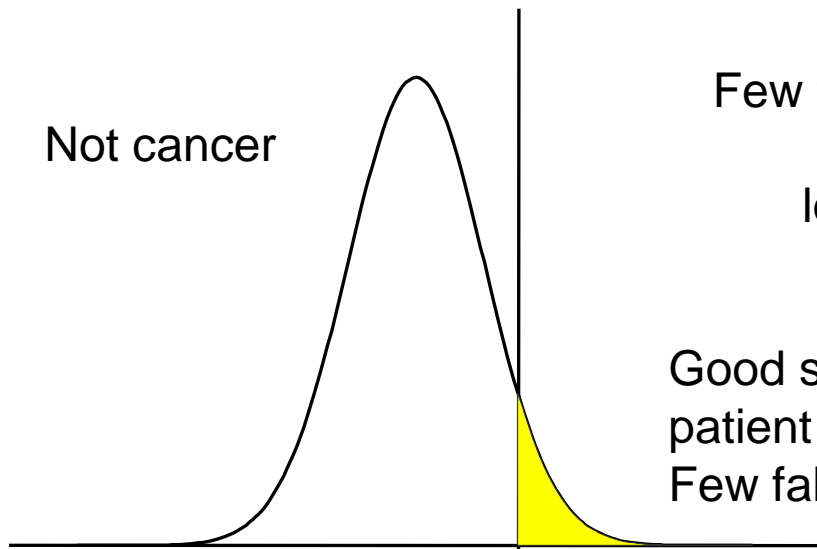
Good sensitivity = $P(\text{Test positive} \mid \text{patient has disease})$

Good specificity = $P(\text{Test negative} \mid \text{patient does not have disease})$

If gene expression levels for cancer and non-cancer overlap:



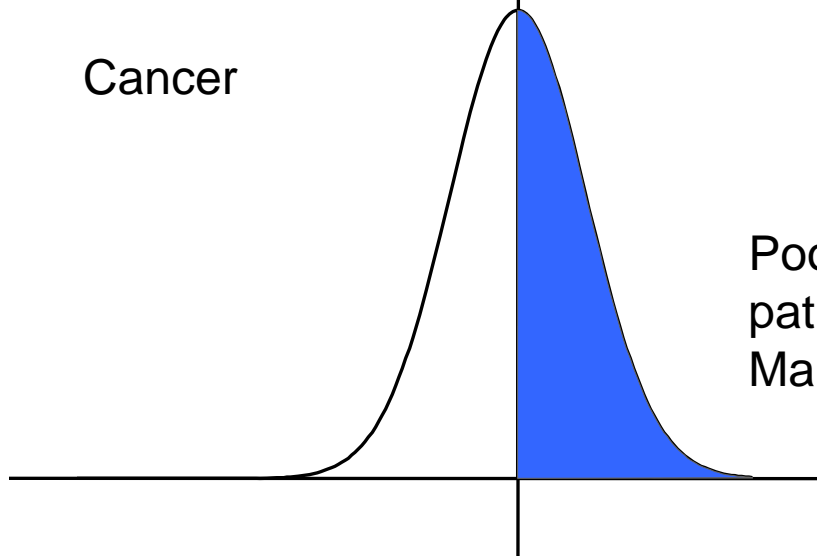
Sensitivity and specificity depend on what threshold we choose



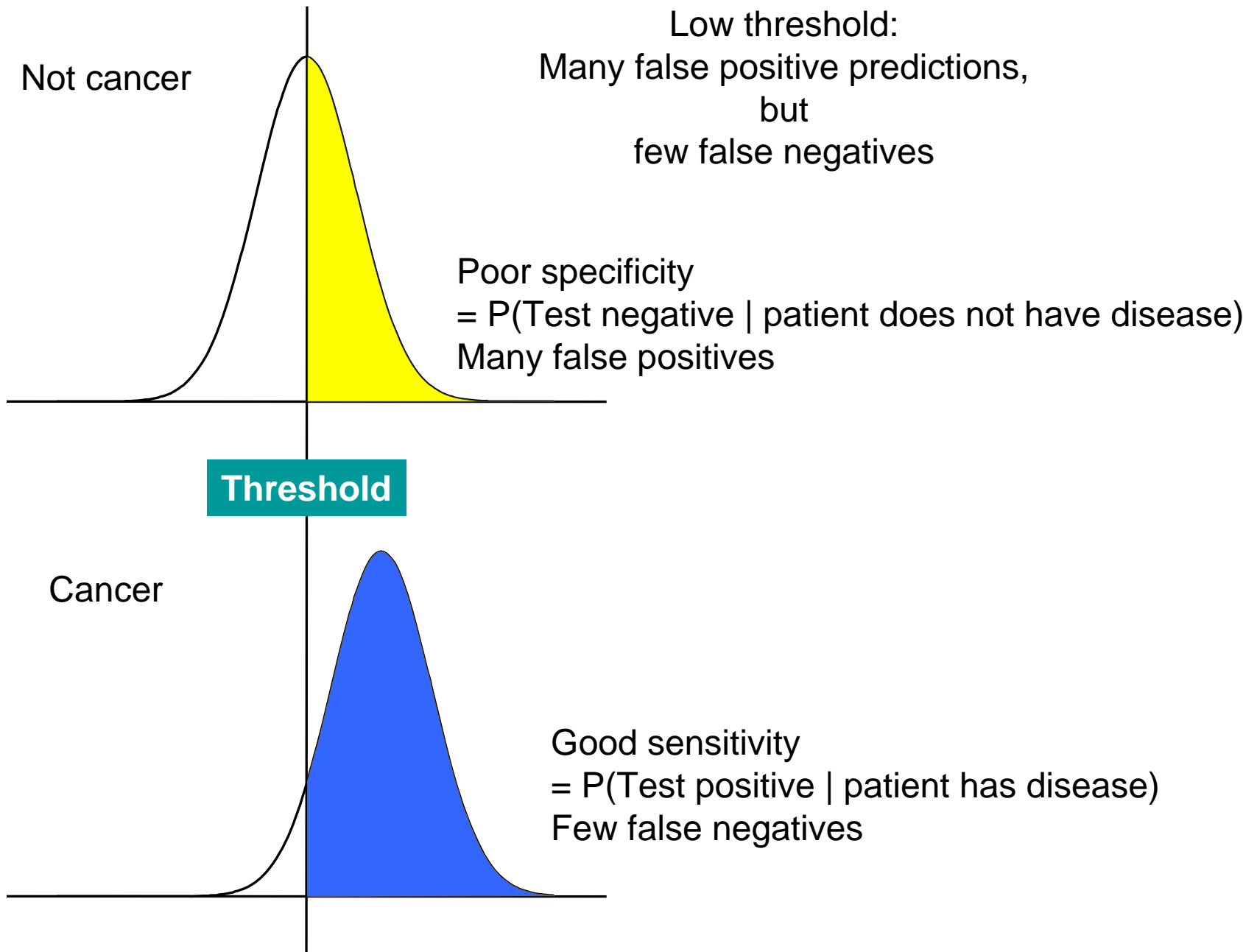
High threshold:
Few false positive predictions,
but
lots of false negatives

Good specificity = $P(\text{Test negative} \mid \text{patient does not have disease})$
Few false positives

Threshold



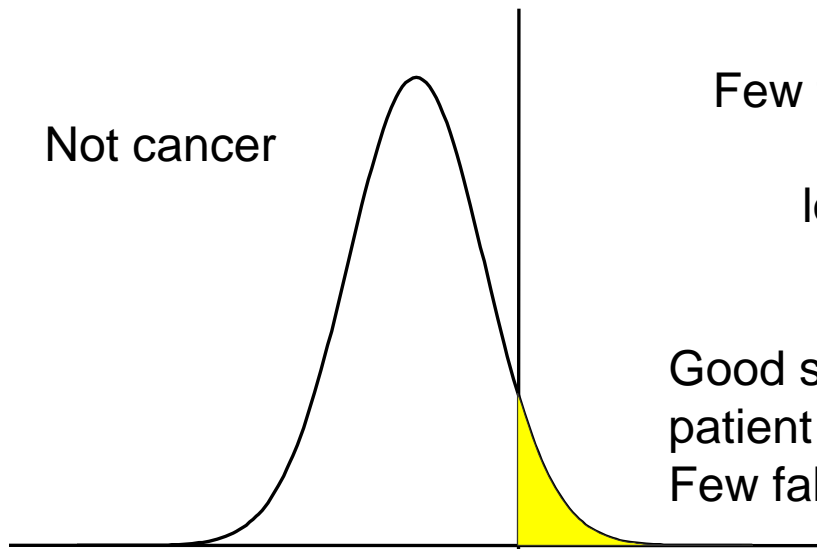
Poor sensitivity = $P(\text{Test positive} \mid \text{patient has disease})$
Many false negatives



Sensitivity and Specificity

- Sensitivity
 - = $P(\text{Test positive} \mid \text{patient has disease})$
 - = $P(T+ \mid D+)$
 - = True positive rate
- Specificity
 - = $P(\text{Test negative} \mid \text{patient does not have disease})$
 - = $P(T- \mid D-)$
 - = True negative rate

- False positive rate (FPR) = $1 - \text{specificity}$
- False negative rate (FNR) = $1 - \text{sensitivity}$



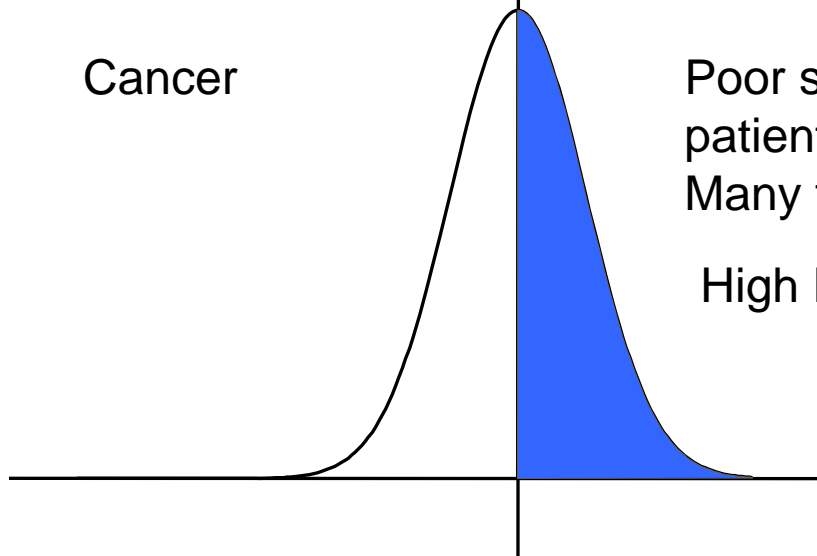
Not cancer

High threshold:
Few false positive predictions,
but
lots of false negatives

Good specificity = $P(\text{Test negative} \mid \text{patient does not have disease})$
Few false positives

Threshold

Low False positive rate (FPR) = $1 - \text{specificity}$



Cancer

Poor sensitivity = $P(\text{Test positive} \mid \text{patient has disease})$
Many false negatives

High False negative rate (FNR) = $1 - \text{sensitivity}$

- Sensitivity and specificity tell us about the test result, given that we know if the patient has the disease or not.
- In clinic, we don't know if the patient has the disease; that's what we want the test to tell us.

PPV and NPV

- Positive predictive value (PPV)
= $P(\text{patient has disease} \mid \text{Test positive})$
= $P(D+ \mid T+)$
- Negative predictive value (NPV)
= $P(\text{patient does not have disease} \mid \text{Test negative})$
= $P(D- \mid T-)$

Sensitivity and Specificity

- Sensitivity = $P(T+ \mid D+)$
- Specificity = $P(T- \mid D-)$
- PPV = $P(D+ \mid T+)$
- NPV = $P(D- \mid T-)$

- PPV and NPV are a function of the prevalence (the proportion of the population that has the disease), as we will see shortly
- Sensitivity and specificity do not depend on the prevalence. They are conditional on the patient either having or not having the disease.

- We calculate PPV and NPV using Bayes' rule

Positive predictive value

$$P(D+|T+) = \frac{P(T+|D+)P(D+)}{P(T+|D+)P(D+) + P(T+|D-)P(D-)}$$

- $P(D+|T+)$ = Probability of disease given test +ve
- $P(T+|D+)$ = sensitivity
- $P(D+)$ = prevalence
- $P(T+|D-)$ = 1 – specificity
- $P(D-)$ = 1 - prevalence

- Pre-test probability = estimated probability that the patient has the disease before getting the diagnostic test result.
 - The pre-test probability = prior probability
- Post-test probability = estimated probability that the patient has the disease after getting the diagnostic test result.
 - The post-test probability = posterior probability

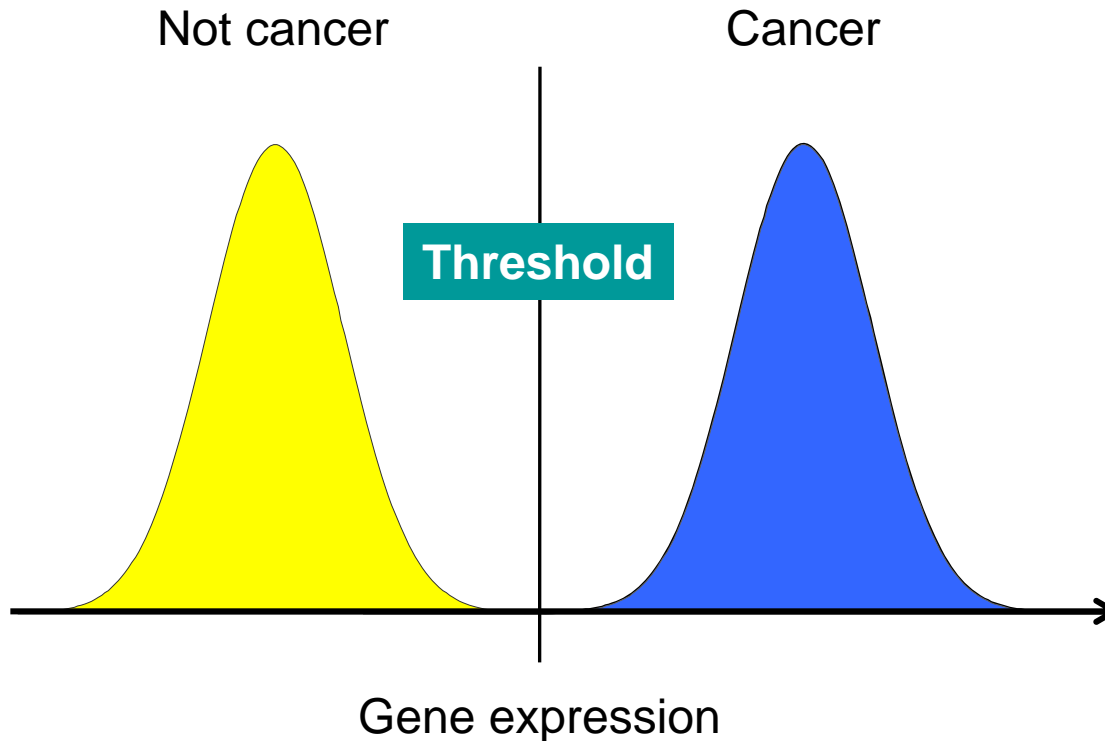
ROC curves

- We want to be able to compare the accuracy of diagnostic tests.
- Sensitivity and specificity are candidate measures for accuracy, but have some problems, as we'll see.
- ROC curves are an alternative measure

ROC curves

- We plot sensitivity against $1 - \text{specificity}$ to create the ROC curve for a test

A test that perfectly separates the two groups

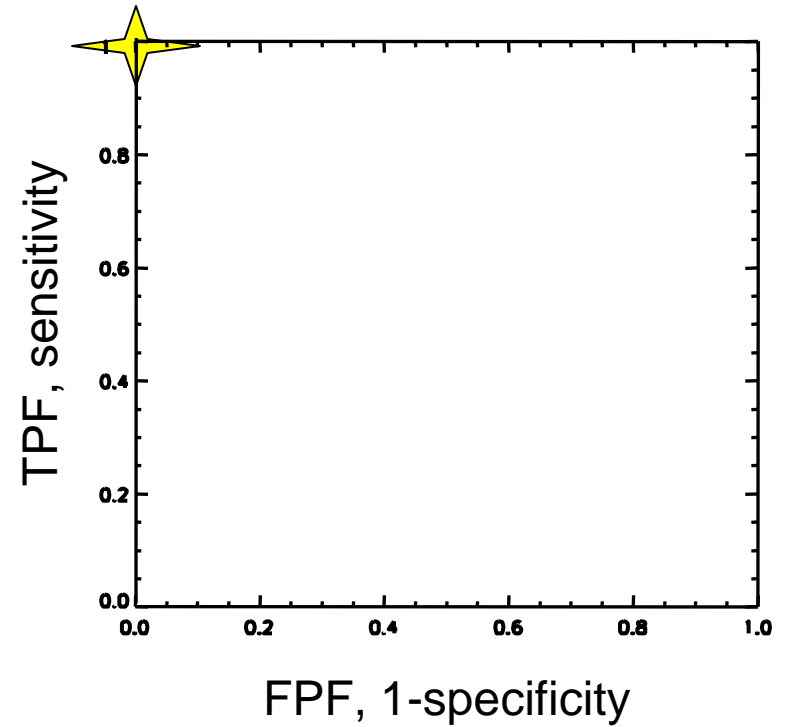


Sensitivity = $P(\text{Test positive} \mid \text{patient has disease}) = 1.0$

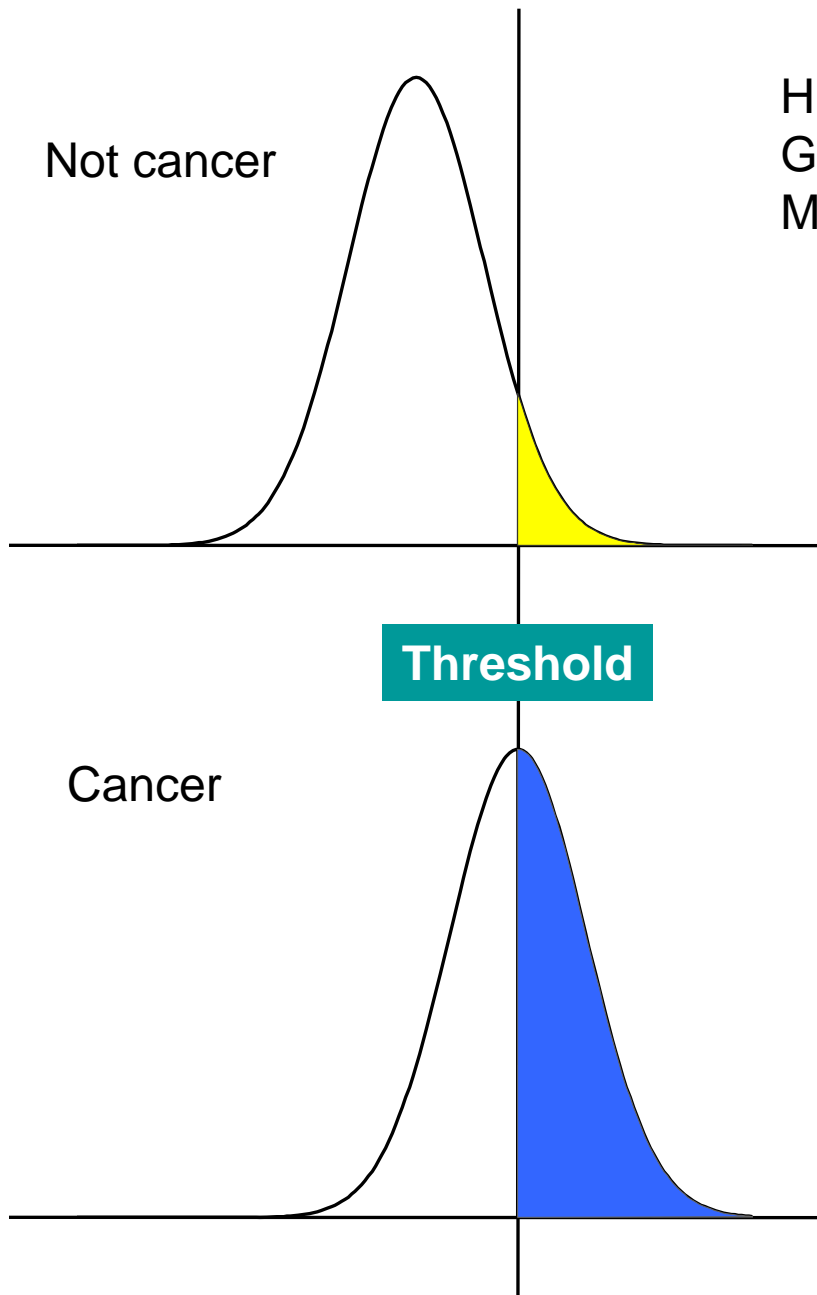
Specificity = $P(\text{Test negative} \mid \text{patient does not have disease}) = 1.0$

ROC curve for a perfect test

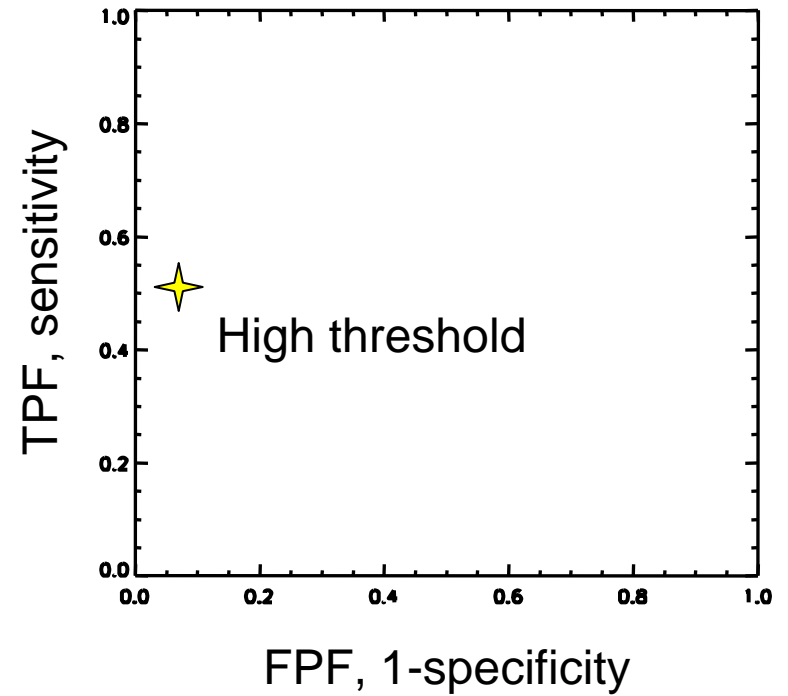
Sensitivity = 1
Specificity = 1
 $1 - \text{specificity} = 0$

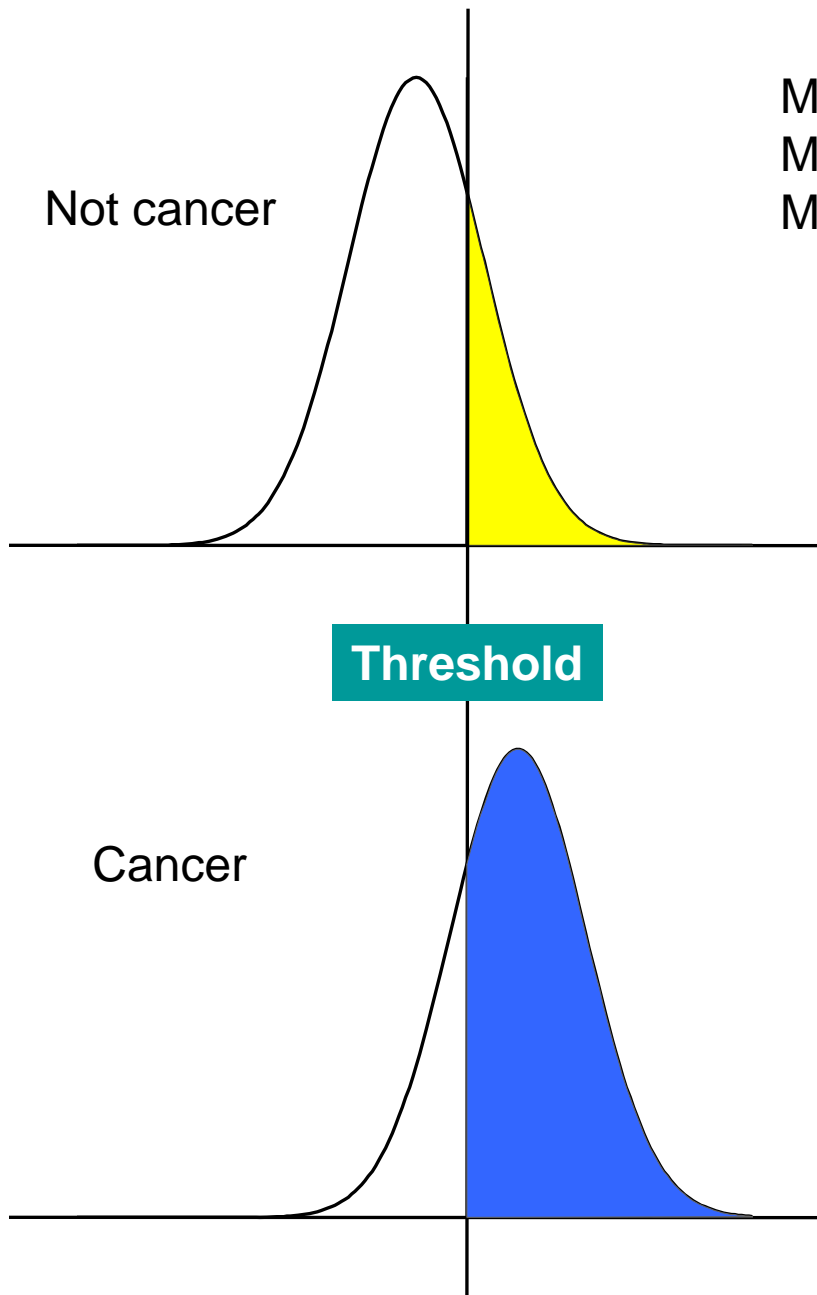


- For a single diagnostic test, sensitivity and specificity vary with the threshold we use.



High threshold:
Good specificity = $P(T- | D-)$
Medium sensitivity = $P(T+ | D+)$

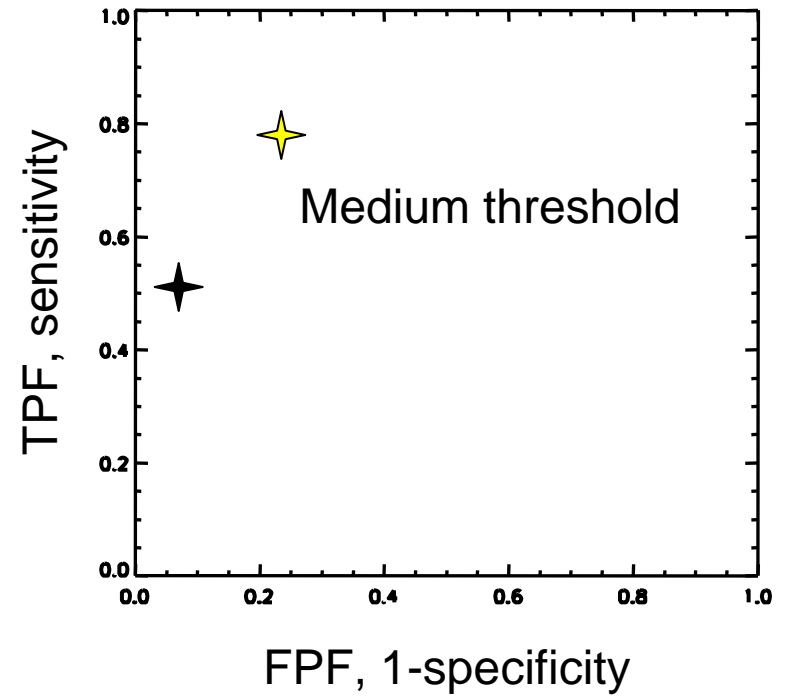


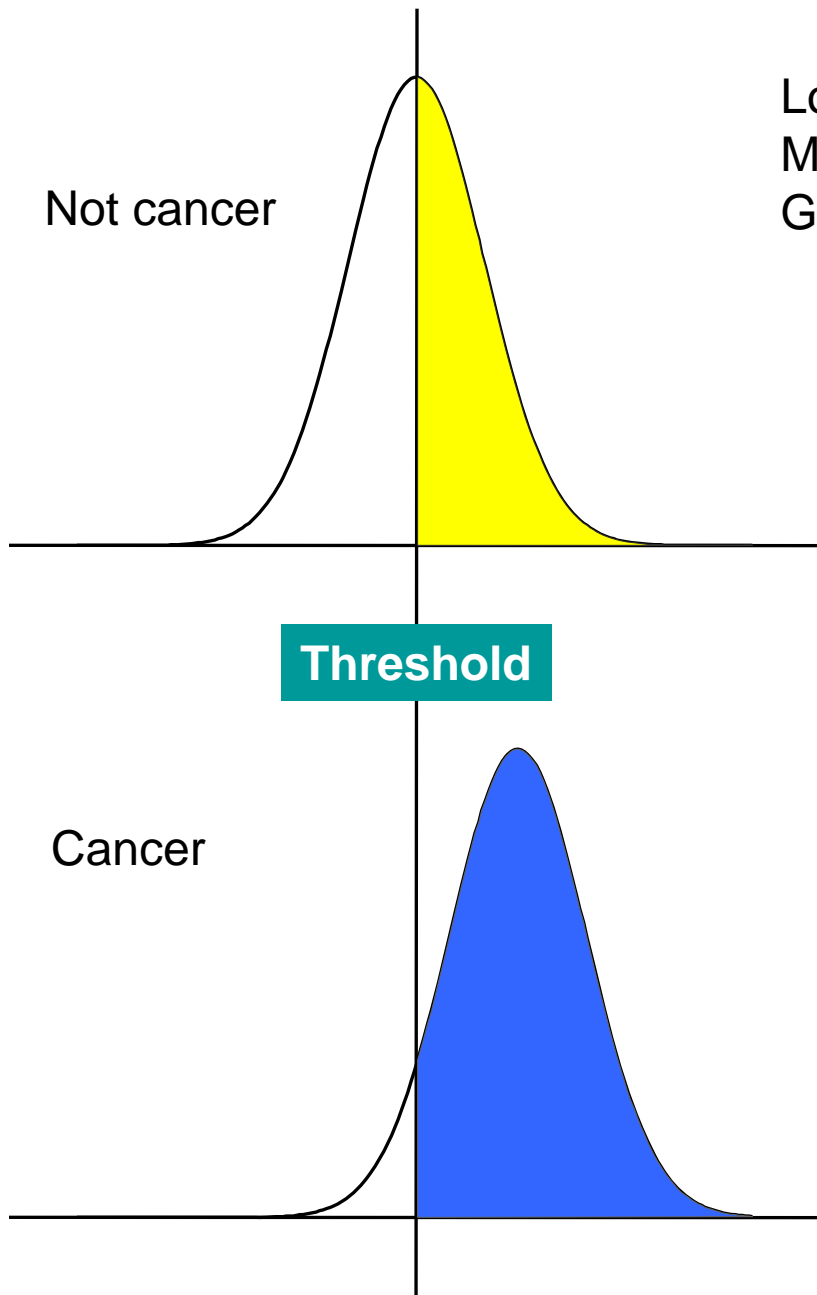


Medium threshold:

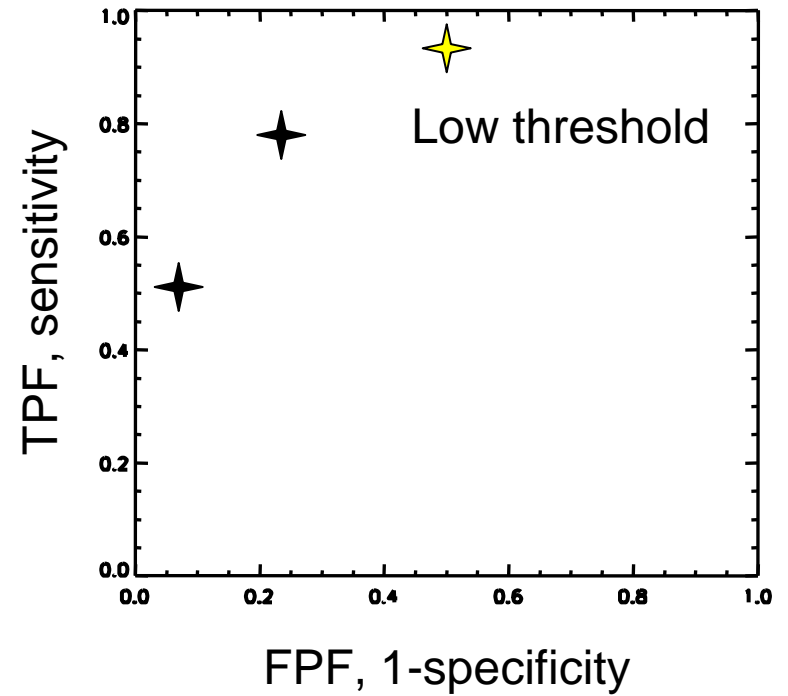
Medium specificity = $P(T- | D-)$

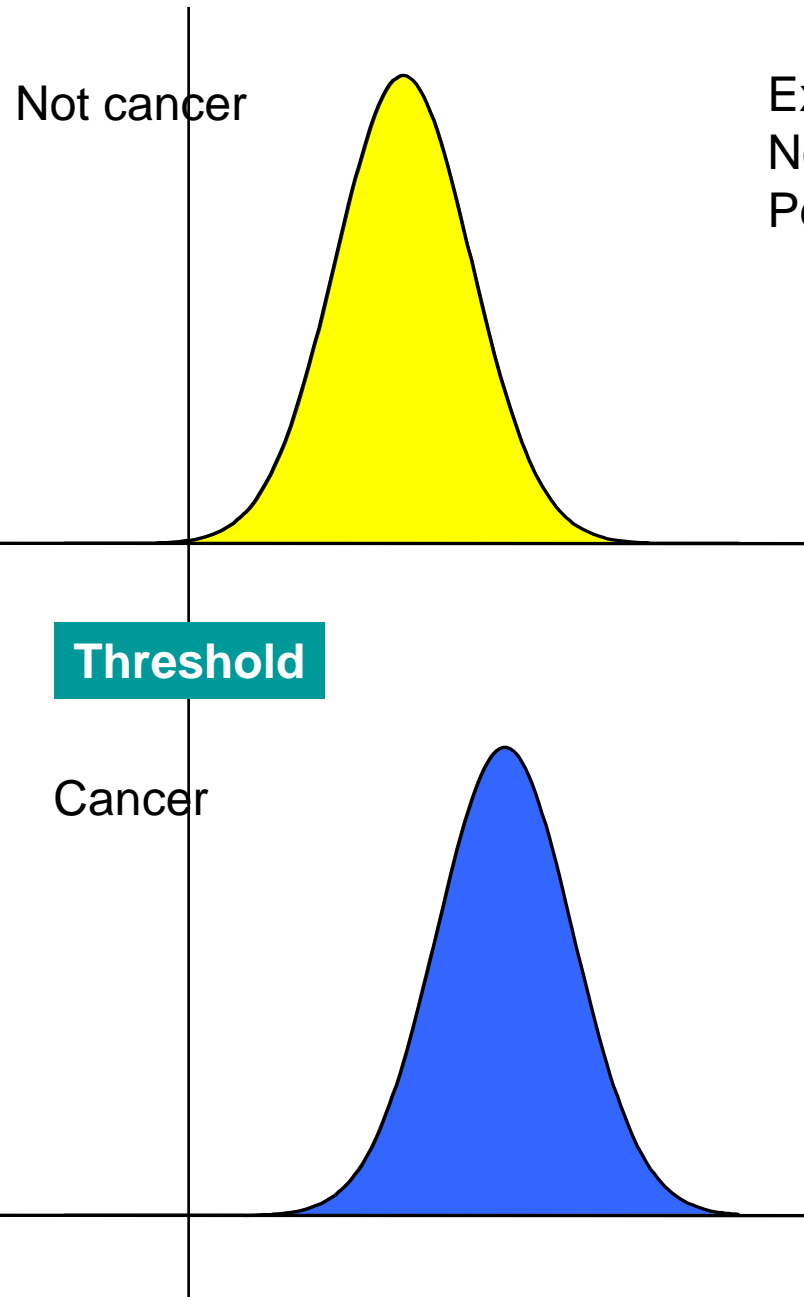
Medium sensitivity = $P(T+ | D+)$



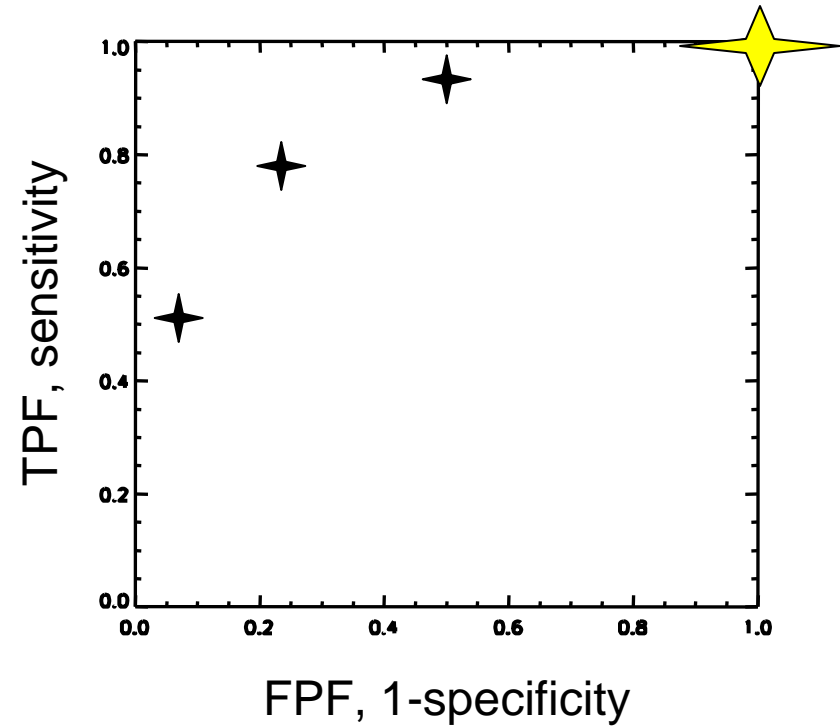


Low threshold:
Medium specificity = $P(T-|D-)$
Good sensitivity = $P(T+|D+)$



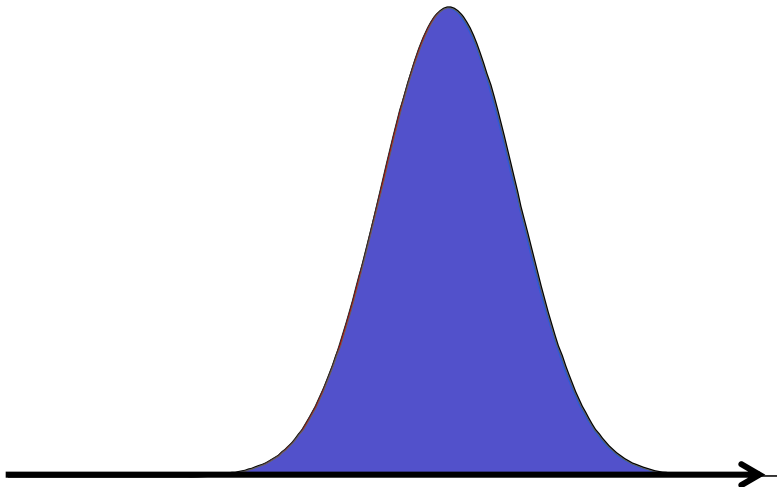
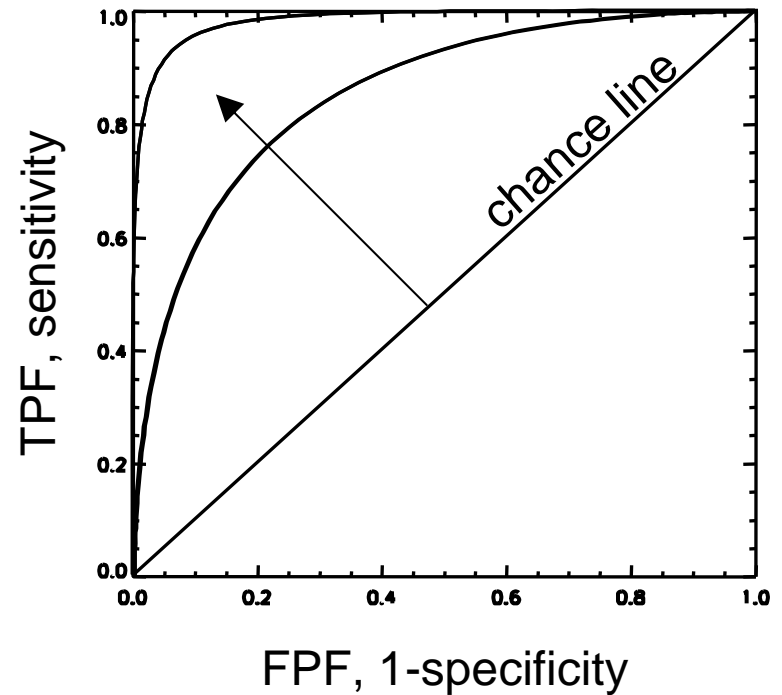


Extreme low threshold:
No specificity = $P(T- | D-)$
Perfect sensitivity = $P(T+ | D+)$

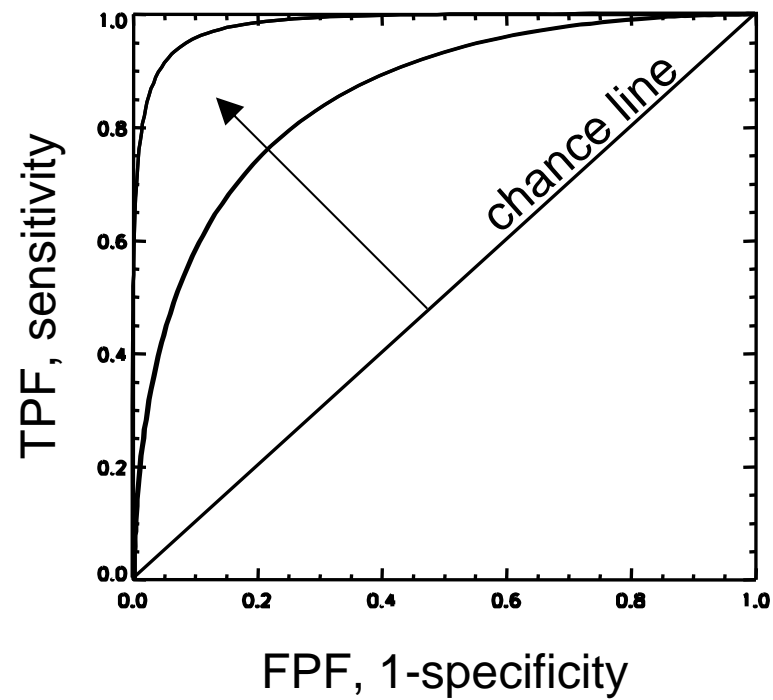
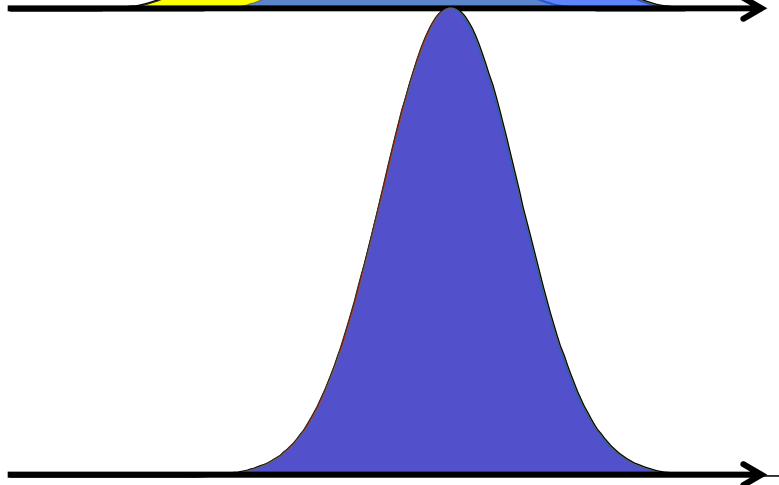
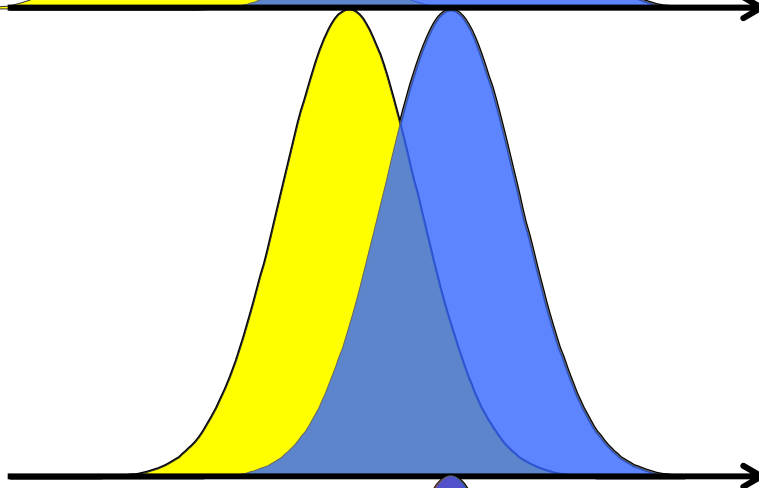
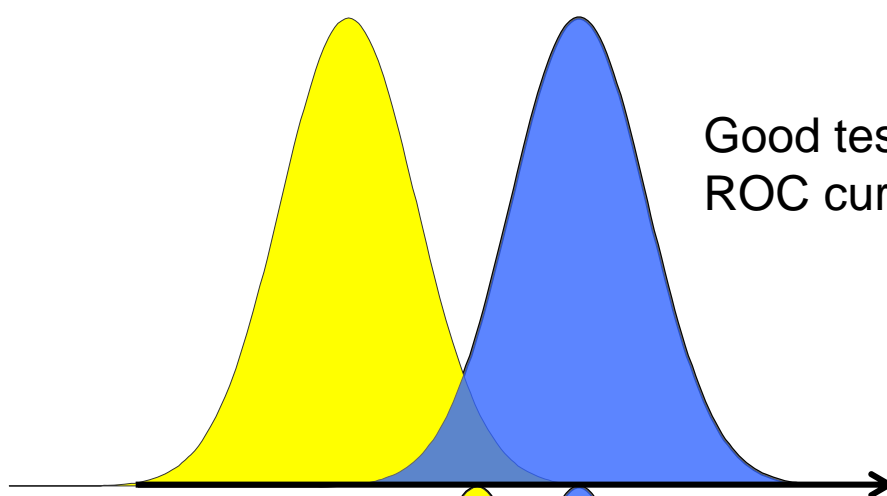


For a test that cannot separate the two classes, the ROC curve is a straight 45 degree line.

Good tests approach the top left corner of the ROC curve.



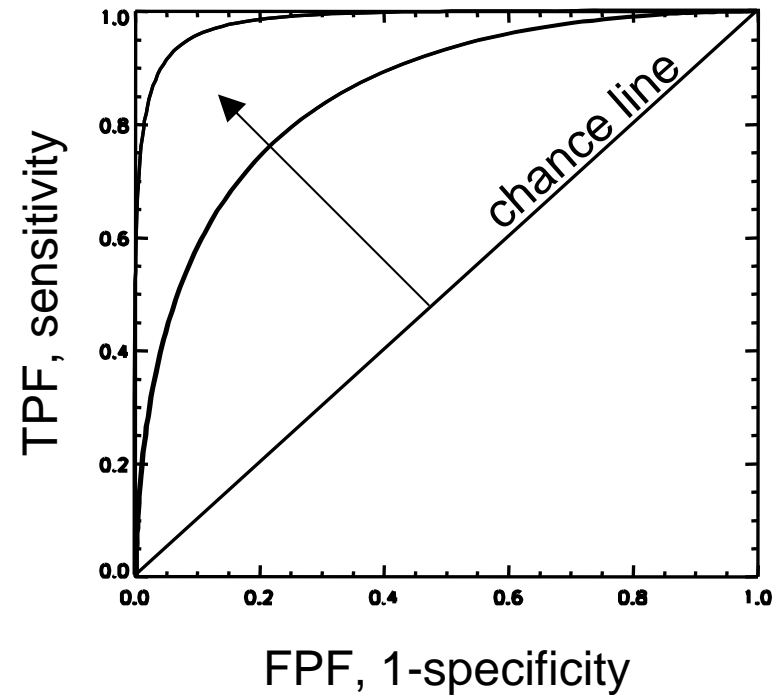
Good tests approach the top left corner of the ROC curve.



The area under the ROC curve describes test accuracy

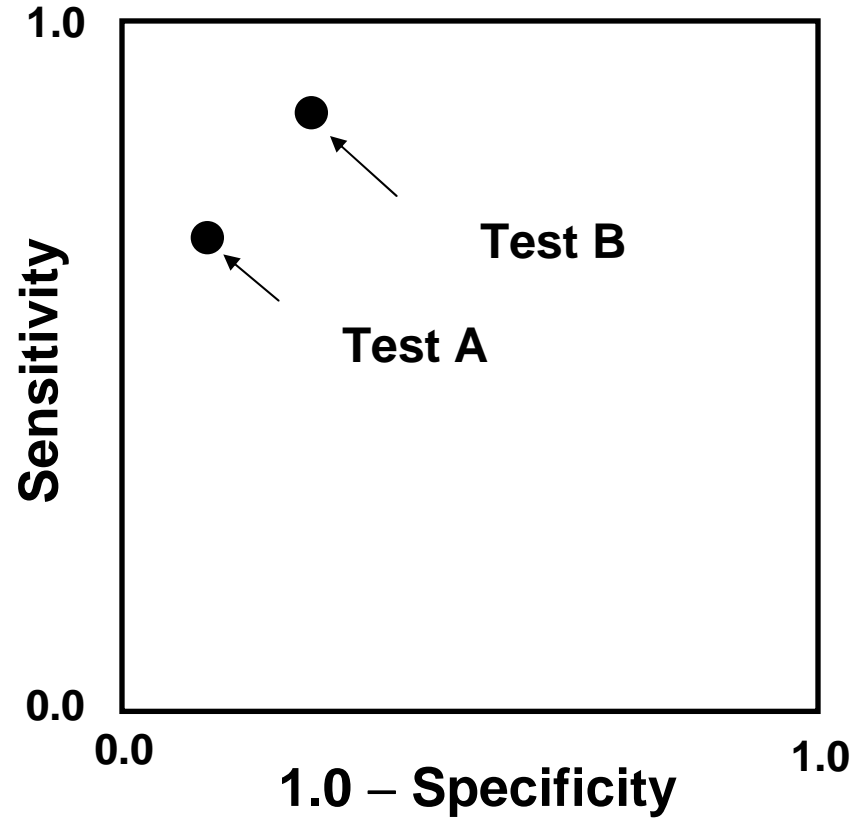
Poor test: ROC area near 0.5

Good test: ROC area near 1.0

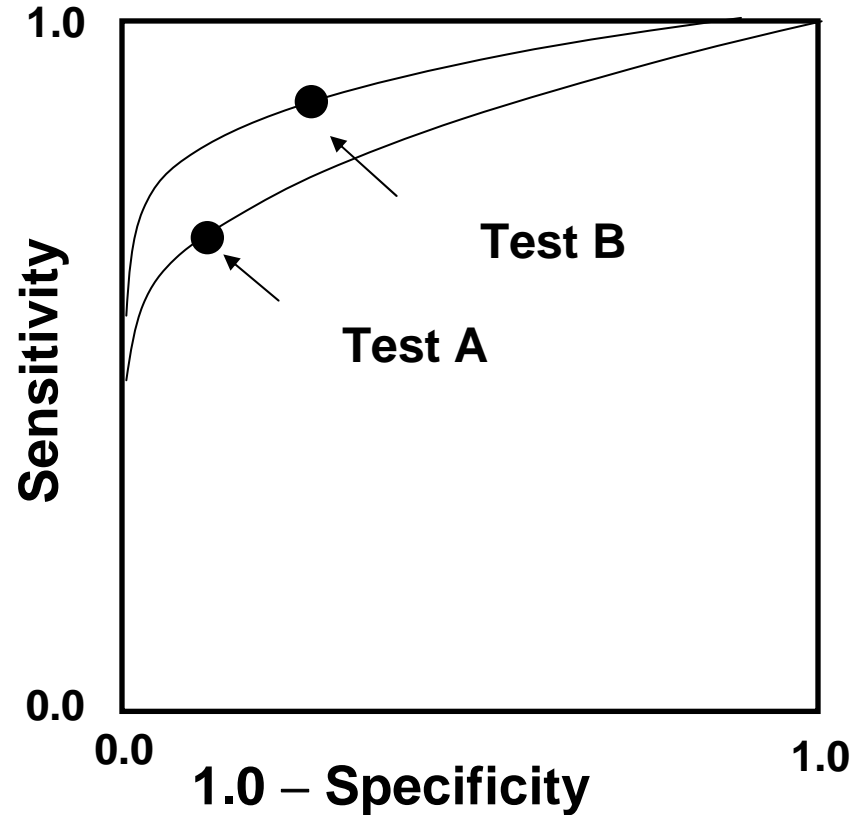


- Sensitivity and specificity don't always make it clear which of two diagnostic tests is better

Which test is better?

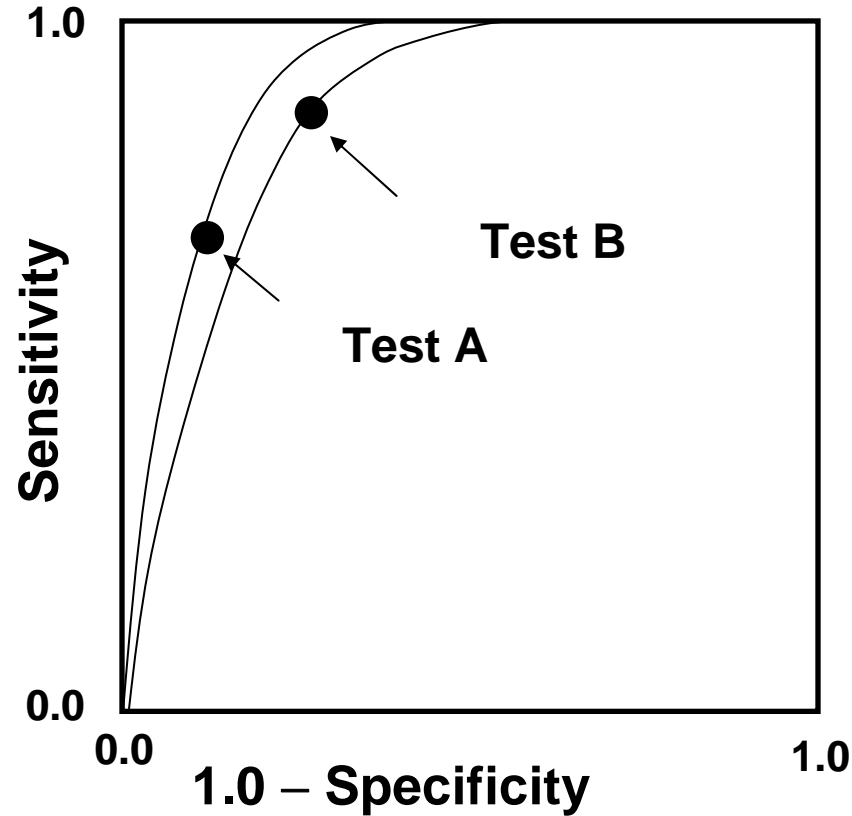


Different ROC for each test



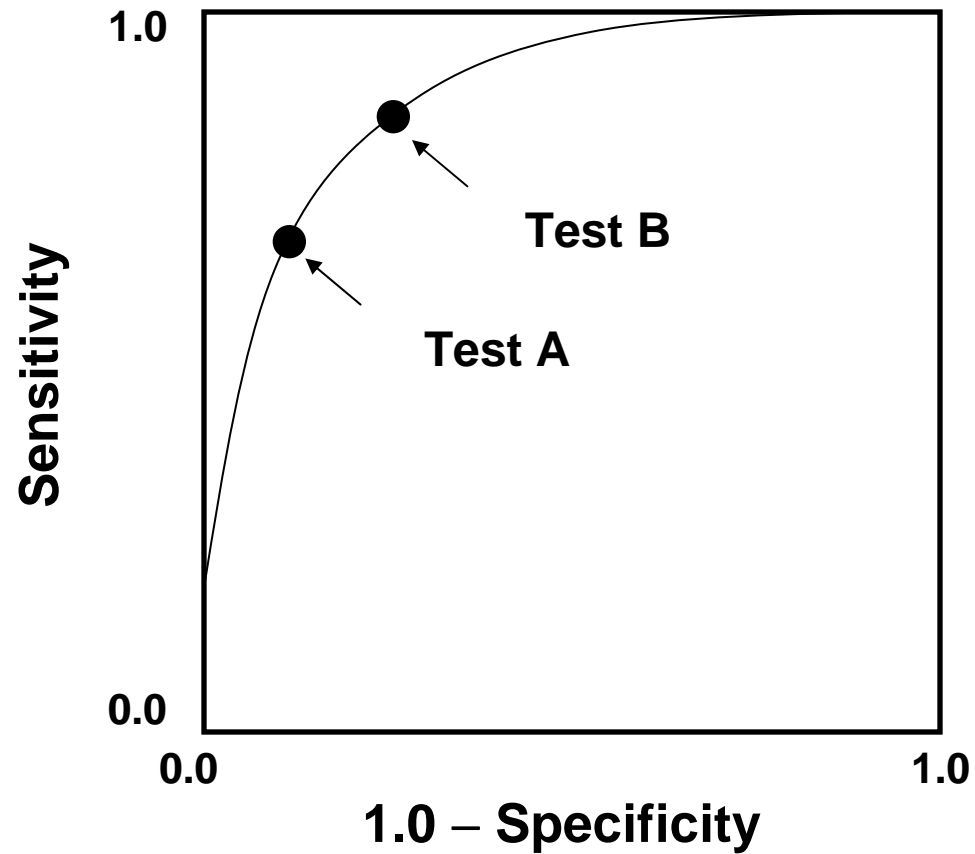
**Test B is better
Greater ROC area**

Different ROC for each test



**Test A is better
Greater ROC area**

Both tests are on the same ROC



**Tests have same
area under ROC**

ROC curves may cross.

In this case, total area under the ROC curve may not be a good measure for comparing tests

Potential ROC issues

- Lack of gold standard for diagnosis
- Lack of reproducibility
 - E.g., disagreement among pathologists
- Bias in sample selection, spectrum of disease used in evaluating test
 - Choose sickest patients, healthy controls
- Problems in ascertainment
 - Genetic disease may not be manifest
- Can't always reliably measure ROC area

Bayes' rule

- How to use Bayes' rule to determine the posterior probability

Bayes' rule

- $P(D+)$ (*prior probability*), the prior probability that the patient has the disease in the absence of any test data (prevalence)
- $P(T+)$: probability of a positive test result (including both true positive and false positive)
- $P(D|T)$ (*posteriori probability*), the probability of disease given the test result

Bayes' rule

$$P(D | T) = \frac{P(T | D)P(D)}{P(T)}$$

$$P(D+ | T+) = \frac{P(T+ | D+)P(D+)}{P(T+ | D+)P(D+) + P(T+ | D-)P(D-)}$$

Positive predictive value

$$P(D+|T+) = \frac{P(T+|D+)P(D+)}{P(T+|D+)P(D+) + P(T+|D-)P(D-)}$$

- $P(D+|T+)$ = Probability of disease given test +ve
- $P(T+|D+)$ = sensitivity
- $P(D+)$ = prevalence
- $P(T+|D-)$ = 1 – specificity
- $P(D-)$ = 1 - prevalence

- Example: sensitivity = 0.9, specificity = 0.8, prevalence=0.5
- $P(T+|D+) = \text{Sensitivity} = 0.9$
- $P(D+) = 0.5$
- $P(T+|D-) = 1 - \text{specificity} = 1 - 0.8 = 0.2$
- $P(D-) = 1 - \text{prevalence} = 1 - 0.5 = 0.5$

$$P(D+|T+) = \frac{P(T+|D+)P(D+)}{P(T+|D+)P(D+) + P(T+|D-)P(D-)}$$

$$P(D+|T+) = \frac{0.9 * 0.5}{0.9 * 0.5 + 0.2 * 0.5} = 0.82$$