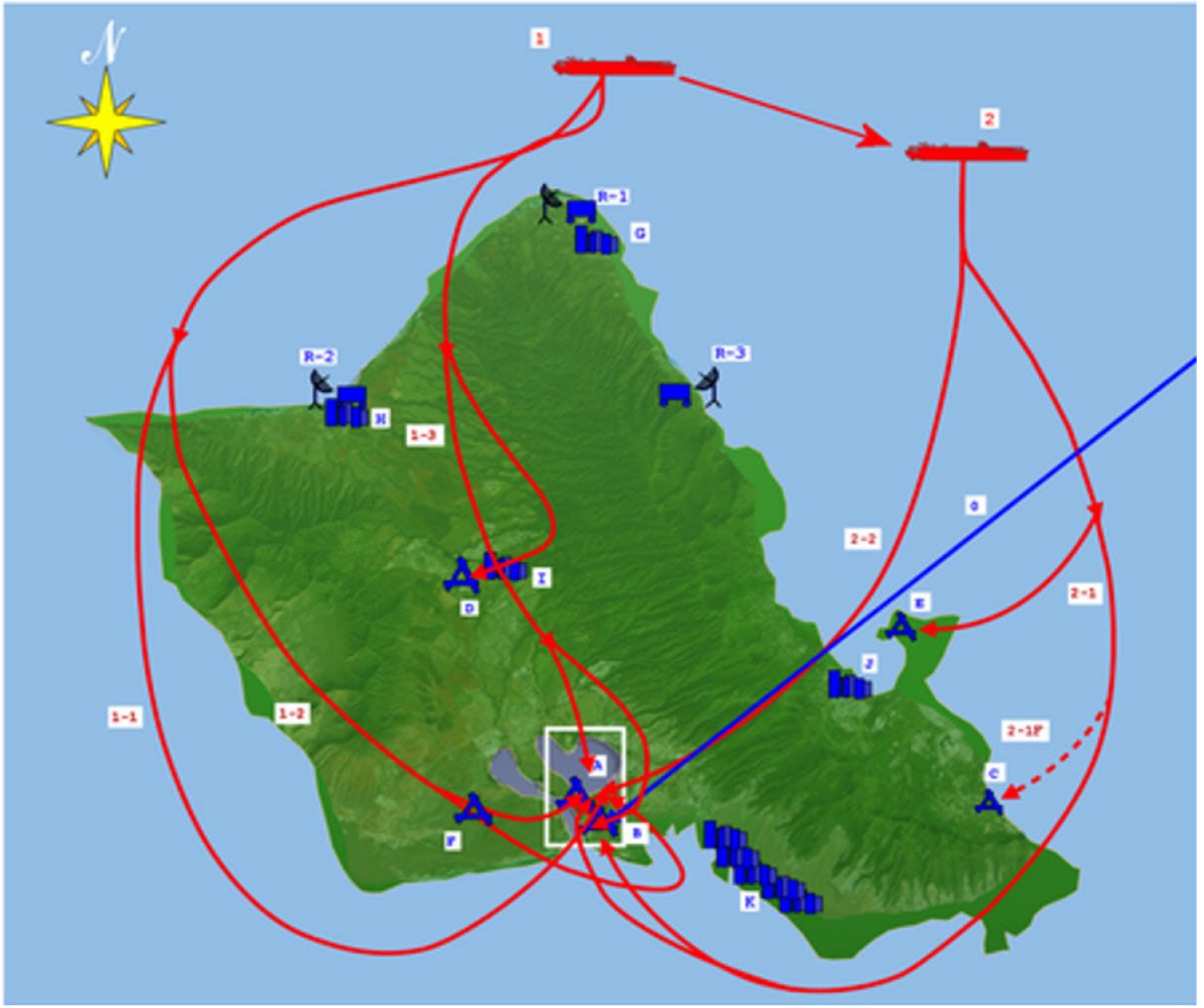
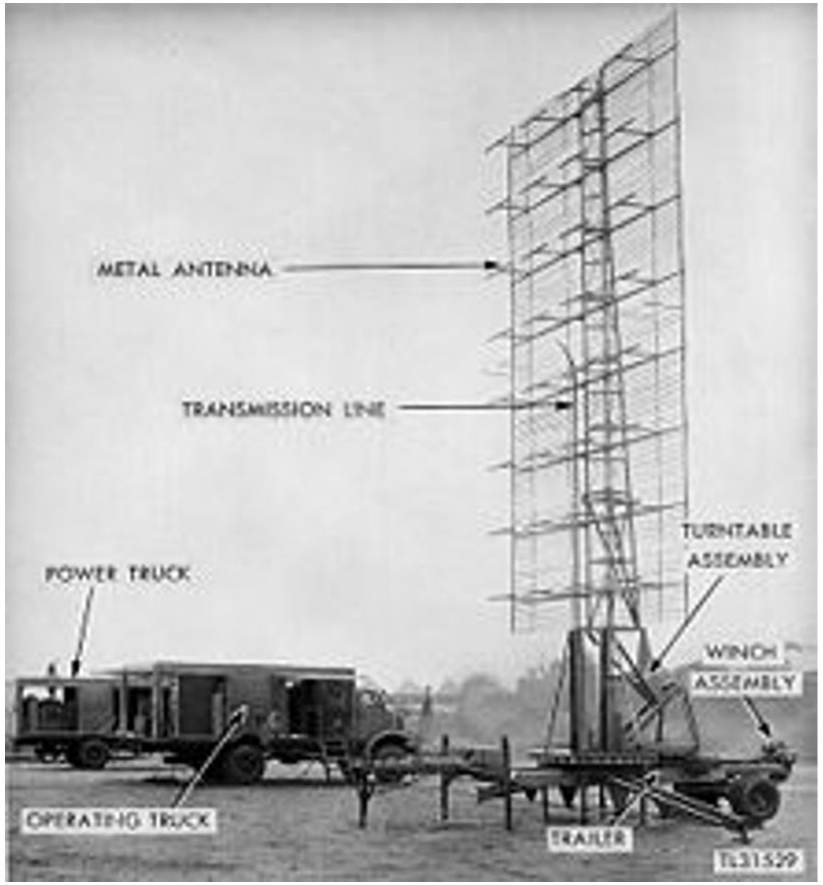
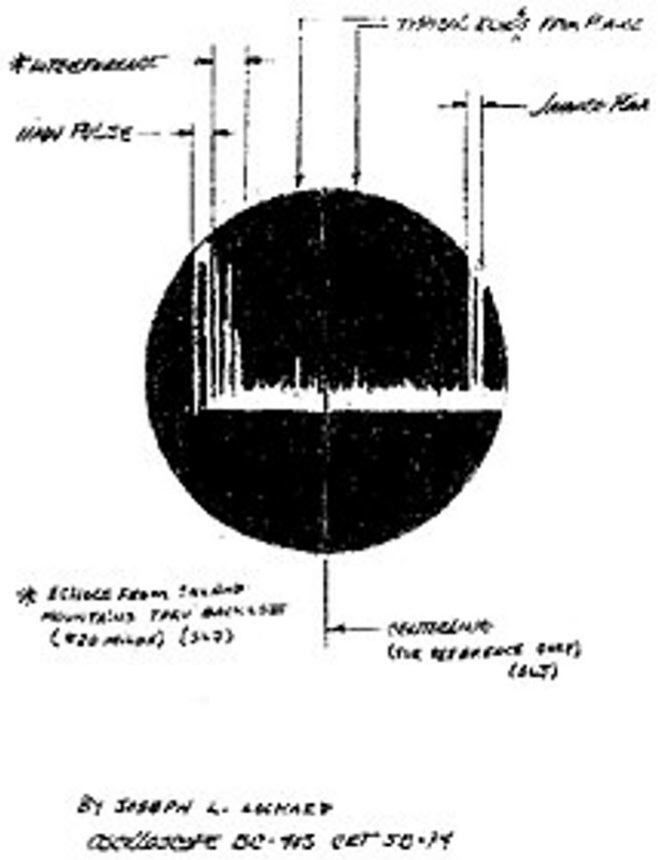
Statistical analysis for NAPPA

Yunro Chung, Ph.D.
Assistant Professor
Arizona State University

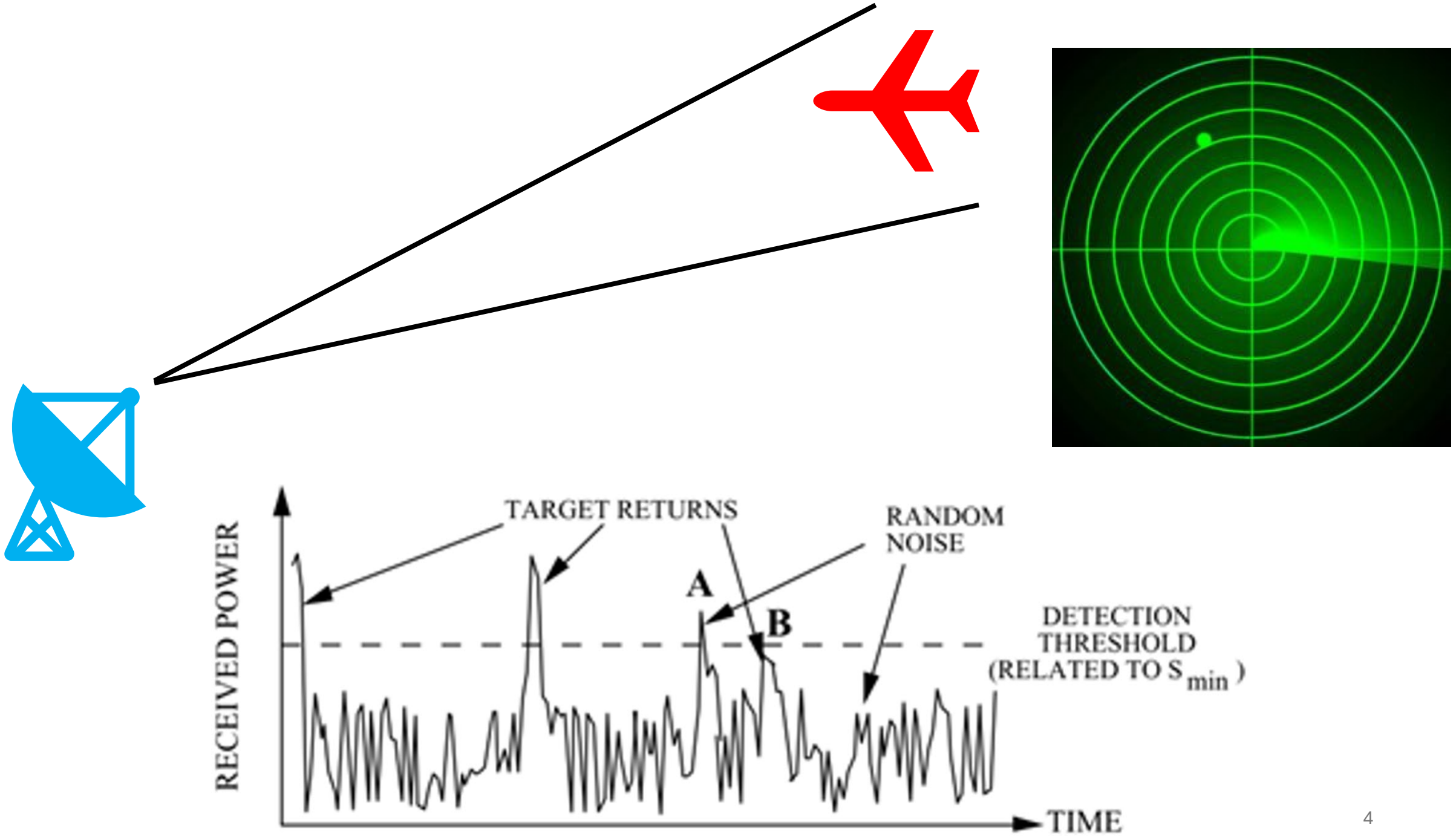


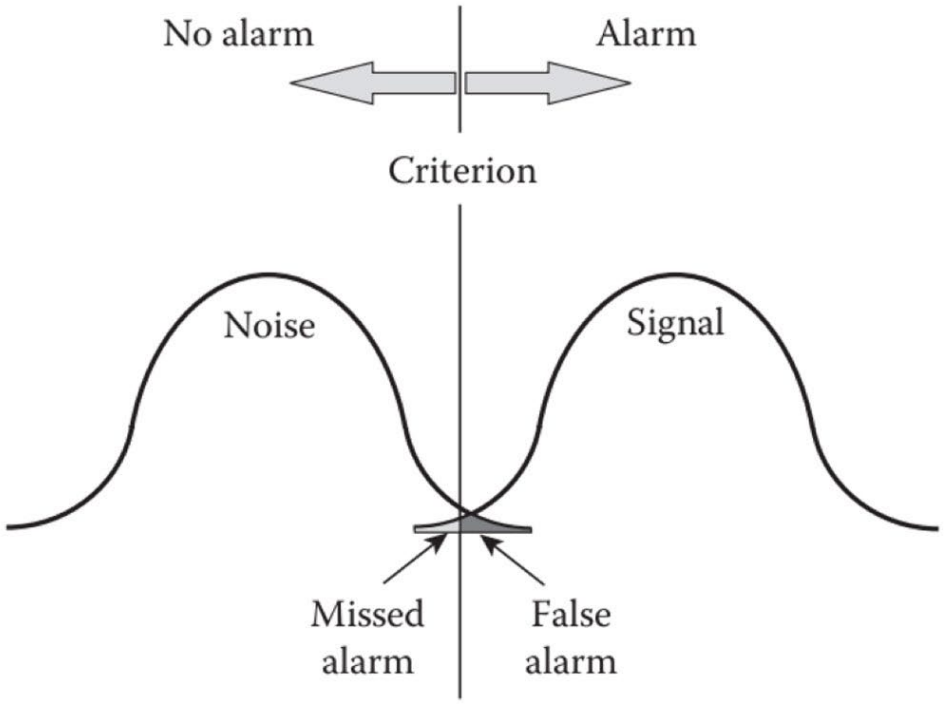
"The Attack on Pearl Harbor was a surprise military strike by the Imperial Japanese Navy Air Service against the United States naval base at Pearl Harbor, Hawaii Territory, on the morning of December 7, 1941." Wikipedia





The SCR-270 (Signal Corps Radio model 270) was the U.S. Army's primary long-distance radar throughout World War II. Its also known as the Pearl Harbor Radar, since it was an SCR-270 set that detected the incoming raid about 45 minutes before the December 7, 1941 attack on Pearl Harbor commenced. Wikipedia


TARGET RETURNS
RANDOM NOISE
A
B
DETECTION THRESHOLD
(RELATED TO S min)
RECEIVED POWER
TIME



Response \ Signal	present	not present
Yes	Hit	False Alarm
No	Miss	Correct Rejection

- Hit = Radar Operator interpreted signal as Enemy Planes and there were Enemy planes (good result)
- Correct Rejection= Radar Operator said no planes and there were none (good result)
- False Alarm = Radar Operator said planes, but there were none (wasted resources)
- Miss = Radar Operator said no plane, but there were planes (very bad outcome, e.g. bombs dropped)

- The receiver operating characteristic (ROC) curve was developed during World War II to detect enemy objects in battle fields.

- For medical diagnostic or screening test:

		Disease	
		Present	Absent
Test	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

TP = test calls the disease and disease is present.
TN= test calls the disease absent and disease is absent.
FP = test calls the disease, but disease is absent.
FN = test calls the disease absent, but disease is present.

Biomarker

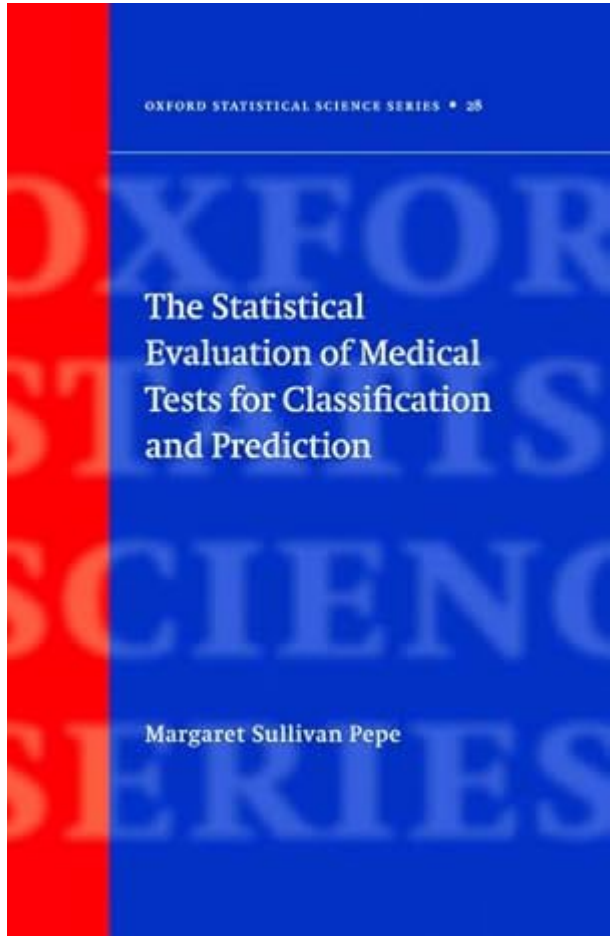
- **Biomarker:** A biological molecule found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease. (NCI)
- **Applications of biomarkers**
  - Screening or diagnostic of disease
  - Treatment selection
  - Surrogate endpoint
  - Disease monitoring
  - …
- *We can also use "test" as biomarker, screening test or diagnostic test*

Biomarker

	Disease: Present	Disease: Absent
Positive	TP	FP
Negative	FN	TN

Criteria for a useful test (Pepe, 2003)

1. Disease should be serious or potentially so
2. Disease should be relative prevalent in the target population
3. Disease should be treatable
4. Treatment should be available to those who test positive
5. The test should not harm the individual
6. The test should accurately classify diseased and non-diseased individuals



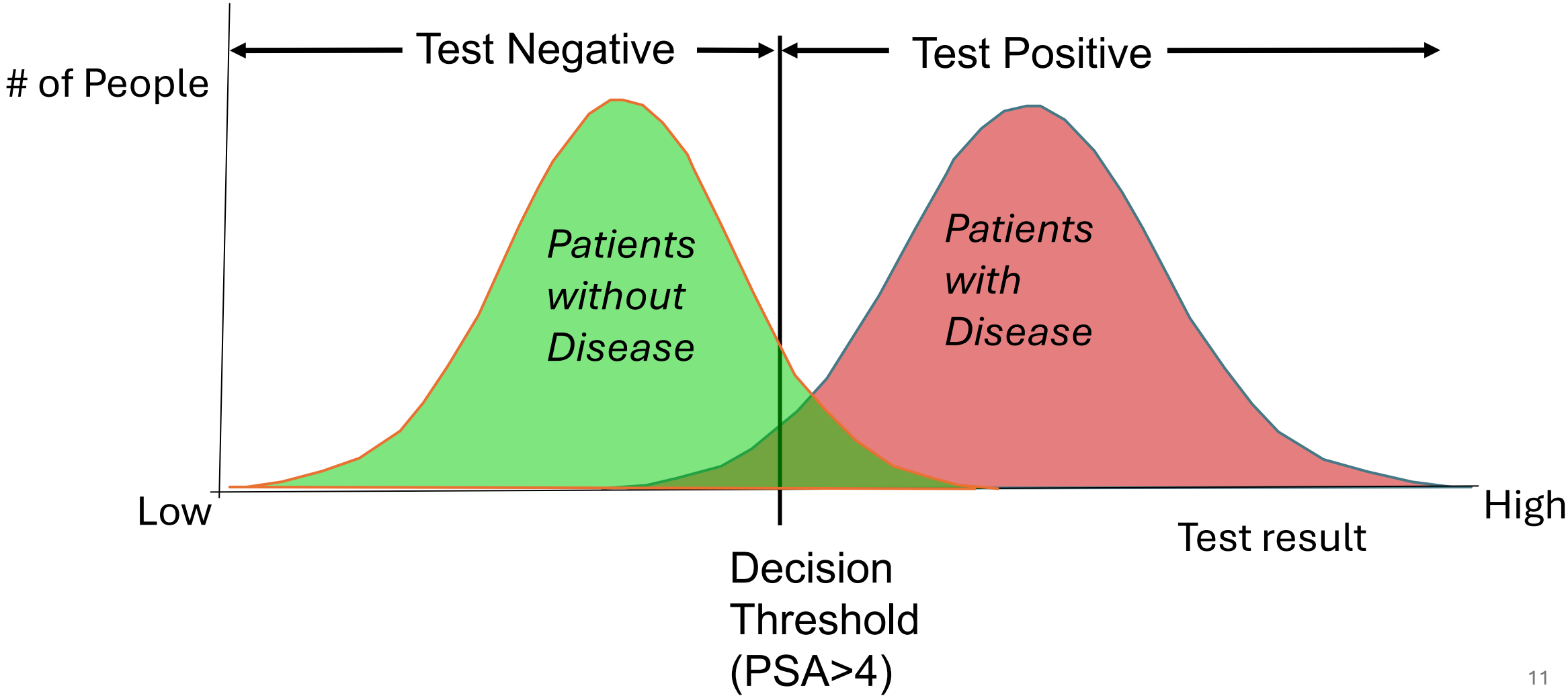
Case-Control Study	Cohort Study
For early biomarker studies typically focusing on people who are **clearly diseased (case)** and **clearly not diseased (control).**	For a late stage for biomarker development.
Small, Inexpensive.	***Large, expensive.***
Exploratory	***Confirmatory***

- Example 1. Case-Control study with 50 case and 50 control.
  - **Prevalence of a disease (or probability of a disease)?**
    50/100=0.5 or 50%? No.
  - *Prevalence is not directly estimable from case-control study*

- Example 2: Control study with n=100,000
  - 100 diseased.
  - Prevalence=100/100,000=0.001 or 0.1%.

Binary test

Example. N=200 with 100 prostate cancers patients and 100 healthy individuals. Prostate-Specific Antigen (PSA) is greater than 4.0 ng/ML as a screening test for PCs.



		Disease	
		Present	Absent
Test	Positive, PSA>4	TP=70	FP=60
	Negative, PSA≤ 4	FN=30	TN=40

- Total Sample size?
  200
- The number of diseased patients?
  100
- The number of healthy patients?
  100

Perfect test:

		Disease	
		Present	Absent
Test	Positive, PSA>4	TP=100	FP=0
	Negative, PSA≤ 4	FN=0	TN=100

In reality:

		Disease	
		Present	Absent
Test	Positive, PSA>4	TP=70	FP=60
	Negative, PSA≤ 4	FN=30	TN=40

Q. How to evaluate the test?

TNF & FNP

		Disease: Present	Disease: Absent
Test	Positive	TP=70	FP=60
	Negative	FN=30	TN=40
	Total	TP+FN=100	FP+TN=100

- **TP Fraction (or Sensitivity)**=P(T+|D+)=$\frac{\text{TP}}{\text{TP+FN}}$= $\frac{70}{70+30}$=0.7
  - no. diseased patients with positive test/no. diseased patients
  - Proportion of those with disease who test positive.
  - Test's ability to correctly classify a person as having a disease (or detect a disease).
- **TN Fraction (or Specificity)**=P(T-|D-)= $\frac{\text{TN}}{\text{FP+TN}}$= $\frac{40}{60+40}$=0.4
  - no. non-diseased patients with negative test/no. non-diseased patients
  - Proportion of those without disease who test negative.
  - Test's ability to correctly classify a person as not having a disease (or detect non-disease).

FNF & FPP

		Disease	
		Present	Absent
Test	Positive	TP=70	FP=60
	Negative	FN=30	TN=40
	Total	TP+FN=100	FP+TN=100

- **FN Fraction**=P(T-|D+)=$\frac{\text{FN}}{\text{TP+FN}}$= $\frac{30}{70+30}$=0.3
  - *no. diseased patients with negative test/no. diseased patients*
  - *Proportion of those with disease who test negative.*
- **FP Fraction**=P(T+|D-)=$\frac{\text{FP}}{\text{FP+TN}}$= $\frac{60}{60+40}$=0.6
  - *no. non-diseased patients with positive test/no. non-diseased patients*
  - *Proportion of those who without disease who test positive.*
- **FNF=1-TPF; FPF=1-TNF**
- *Test result can be interpreted by any pair of (TPF, TNF); (TPF, FPF); (FNF, TNF); (FNF, FPF).*

Sensitivity & Specificity

- In biomedical research,
  - sensitivity=TPF
  - specificity=TNF
- In statistical hypothesis test,
  - significance level=FPF
  - statistical power=TPF
- In engineering applications and in audiology,
  - hit rate=TPF
  - false alarm rate=FPF

Sensitivity & Specificity

- **Sensitivity** and **specificity** are ranged from 0 to 1 (0% to 100%).
  - A perfect test has sensitivity=1 & specificity=1.
  - A useless test TPF=FPF, i.e. sensitivity=1-specificity.
- *A test with high sensitivity and high specificity is generally preferred.*

Q. Why do we report both sensitivity and specificity, not either sensitivity or specificity?

- Test A with 100% sensitivity.

Test A		Disease: Present	Disease: Absent
	Positive	TP=100	FP=100
	Negative	FN=0	TN=0

- Test B with 100% specificity.

Test B		Disease: Present	Disease: Absent
	Positive	TP=0	FP=0
	Negative	FN=100	TN=100

Q. Why do we report both sensitivity and specificity instead of a single measure?

		Disease	
		Present	Absent
Test	Positive	TP=70	FP=60
	Negative	FN=30	TN=40

- Accuracy=$\frac{TP+TN}{TP+FN+FP+TN}=\frac{110}{200}=0.55$
- Misclassification rate =$\frac{FP+FN}{TP+FN+FP+TN}=\frac{90}{200}=0.45$ (=1-Accuracy)
- OR=$\frac{TP/FP}{FN/TN}=\frac{TP\times TN}{FN\times FP}=\frac{70\times 40}{30\times 60}=1.56$

- These single measures are often used in practice but do not provide enough information.
- Sensitivity and specificity have different meanings, e.g. *miss (bombs dropped) is more serious than false alarm*.
- Lower sensitivity (or higher FN)
  - Missed opportunity for intervention
  - Most common cause for malpractice lawsuits
- Lower specificity (or higher FP)
  - Increased worry/fears
  - Expensive and unnecessary additional testing

Disease A versus B

		Disease	
		A	B
Test	A	TC_A=70	FC_A=60
	B	FC_B=30	TC_B=40
	Total	TC_A+FC_B=100	FC_A+TC_B=100

- True classification fraction for disease A (TCF_A)=$\frac{TC_A}{TC_A+FC_A}=\frac{70}{70+30}=0.7$
  - test's ability to correctly classify a person as having disease A (or detect disease A).
- True classification fraction for disease B (TCF_B)=$\frac{TC_B}{FC_A+TC_B}=\frac{40}{60+40}=0.4$
  - test's ability to correctly classify a person as having disease B (or detect disease B).
- Thus, similar measures can be used to evaluate the test.

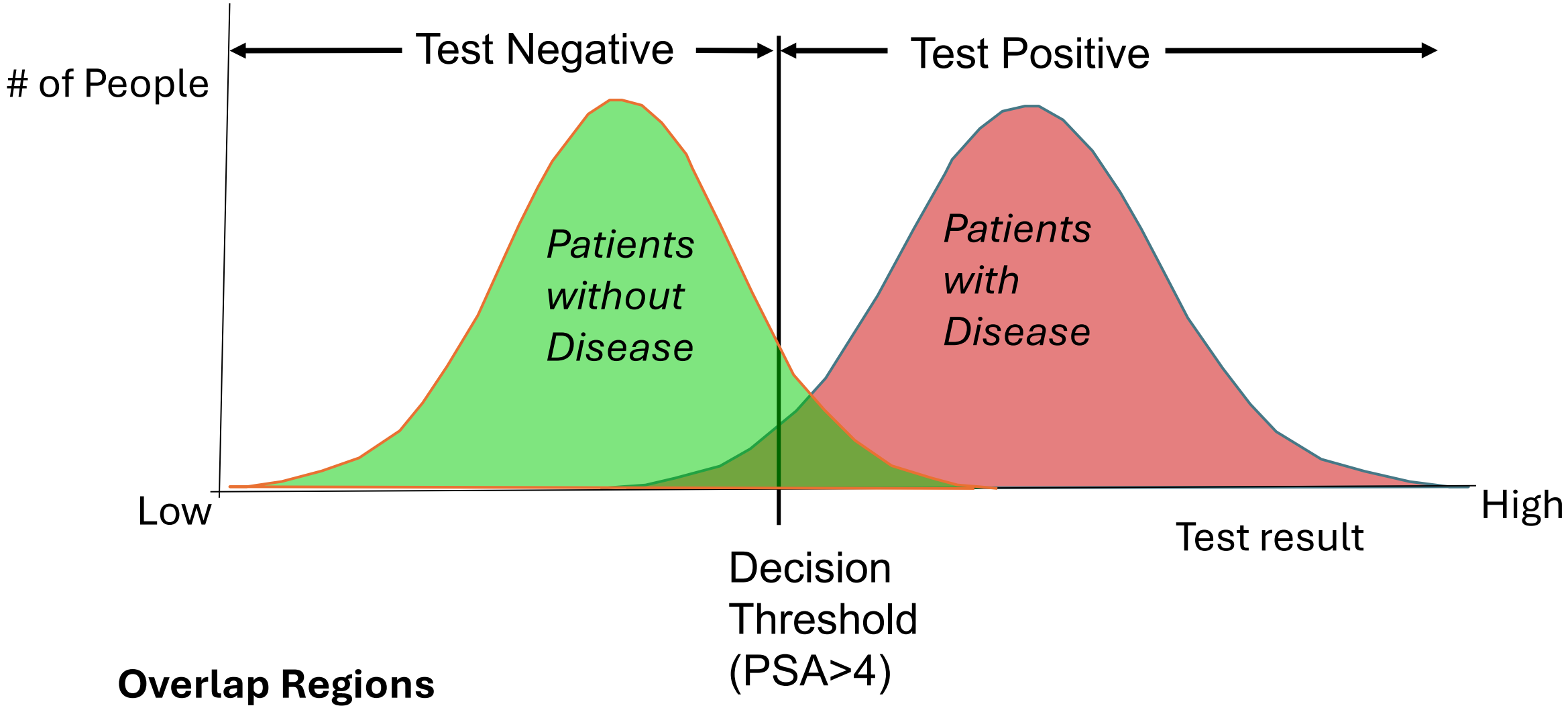
Predictive value

		Disease		
		Present	Absent	Total
Test	Positive	TP=70	FP=60	TP+FP=130
	Negative	FN=30	TN=40	FN+TN=70

- **Positive Predictive Value (PPV)**=P(D+|T+)= $\frac{\text{TP}}{\text{TP+FP}}$= $\frac{70}{70+60}$=0.54
  - *no. diseased patients with positive test/no. diseased with positive test*
  - *The proportion of patients with a positive test who have the disease*
  - *Test's ability to accurately predict disease*
- **Negative Predictive Value (NPV)**=P(D-|T-)=$\frac{\text{TN}}{\text{FN+TN}}$= $\frac{40}{30+40}$=0.57
  - *no. non-diseased patients with negative test/no. patients with negative test*
  - *The proportion of patients with a negative test who do not have the disease*
  - *Test's ability to accurately predict non-disease*

- Sensitivity and Specificity quantify how well the test reflects true disease status.
  - Used to quantify the inherent accuracy of the test.
- PPV and NPV quantify how well the test result predicts true disease status.
  - Used to quantify the clinical value of the test.
- PPV and NPV are also ranged from 0 to 1 (or 0 to 100%).
  - Perfect test: PPV=NPV=1
  - Useless test: PPV=prevalence; NPV=1-prevalence
  - PPV and NPV are not directly estimable from case-control study.

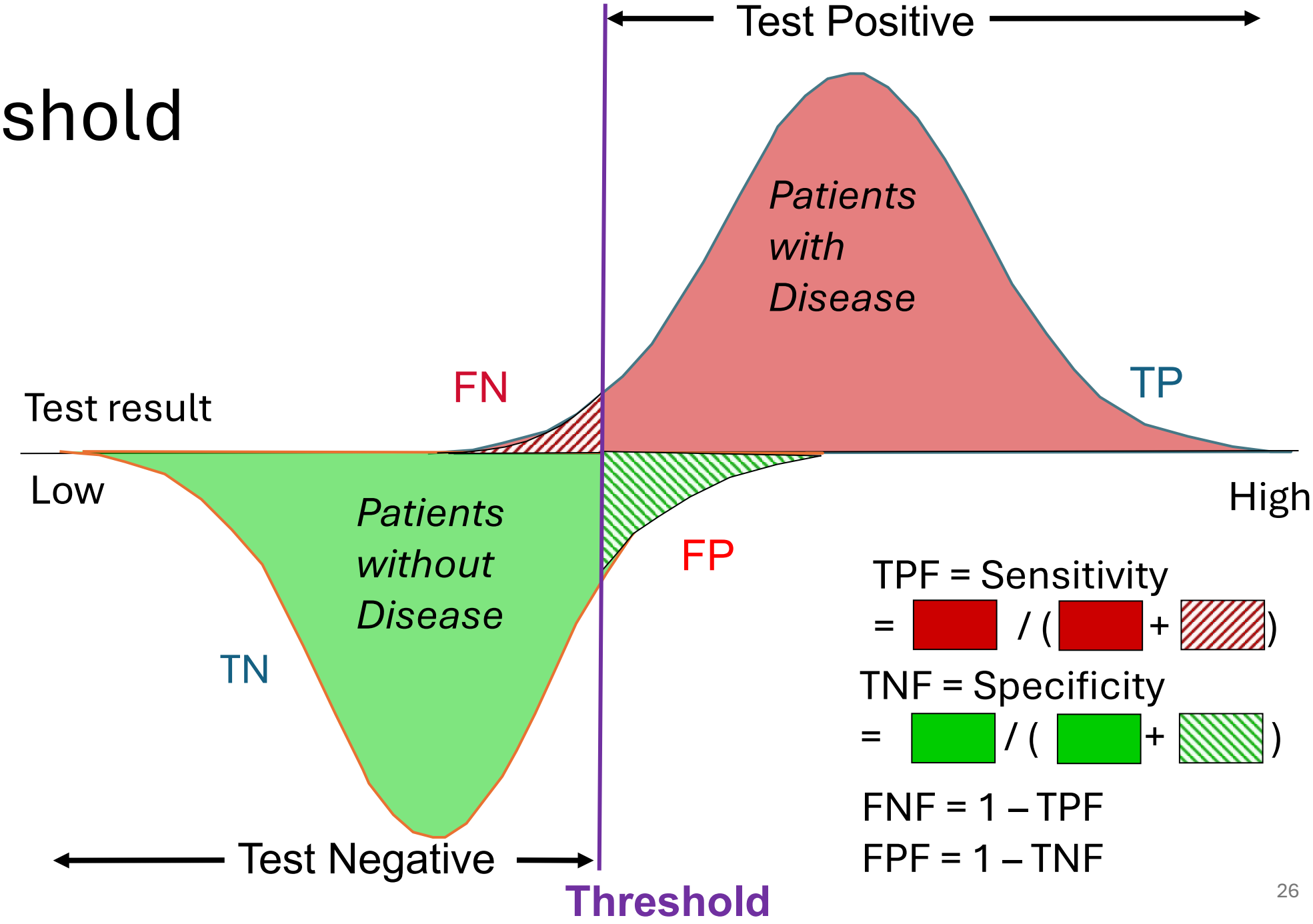
ROC Curve



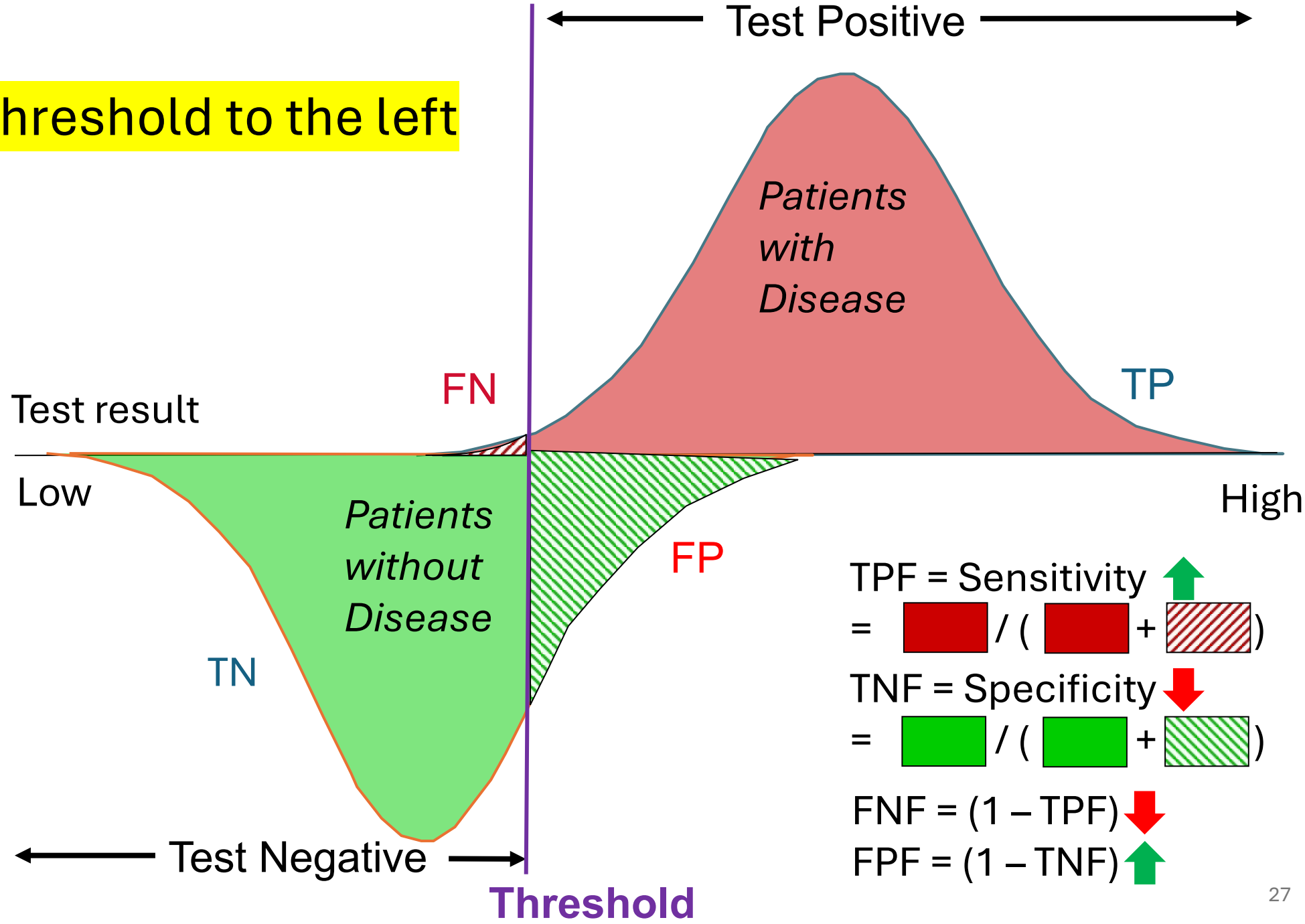
Overlap Regions

- Threshold (or cutoff value) defines a positive or negative result
  - Divides overlap region in patients with target disease and patients without the target disease into FN results and FP results

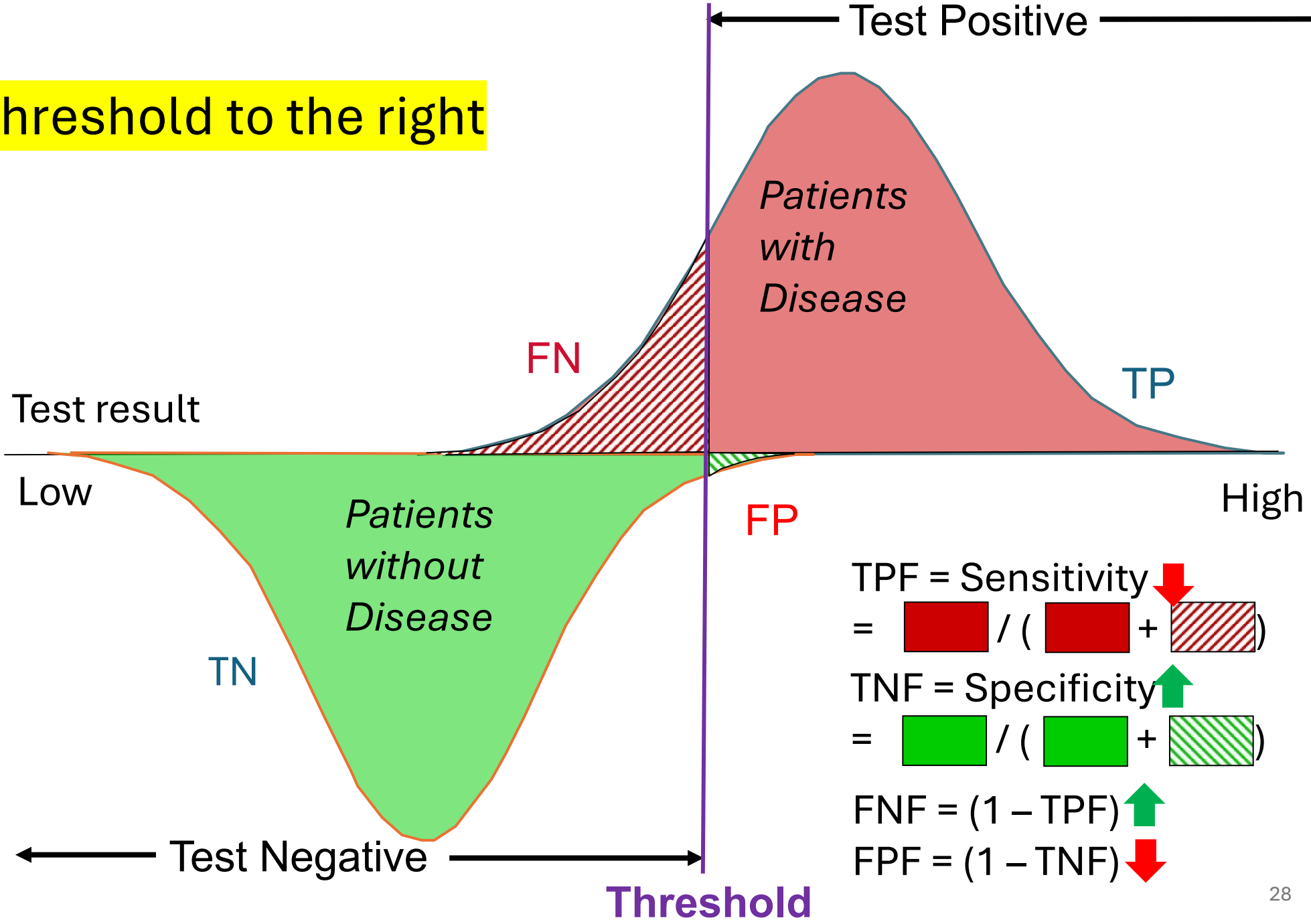
Threshold


Test Positive
Patients with Disease
FN
TP
Test result
Low
High
Patients without Disease
FP
TN
Test Negative
Threshold
TPF = Sensitivity
= / (+)
TNF = Specificity
= / (+)
FNF = 1 – TPF
FPF = 1 – TNF

Move threshold to the left



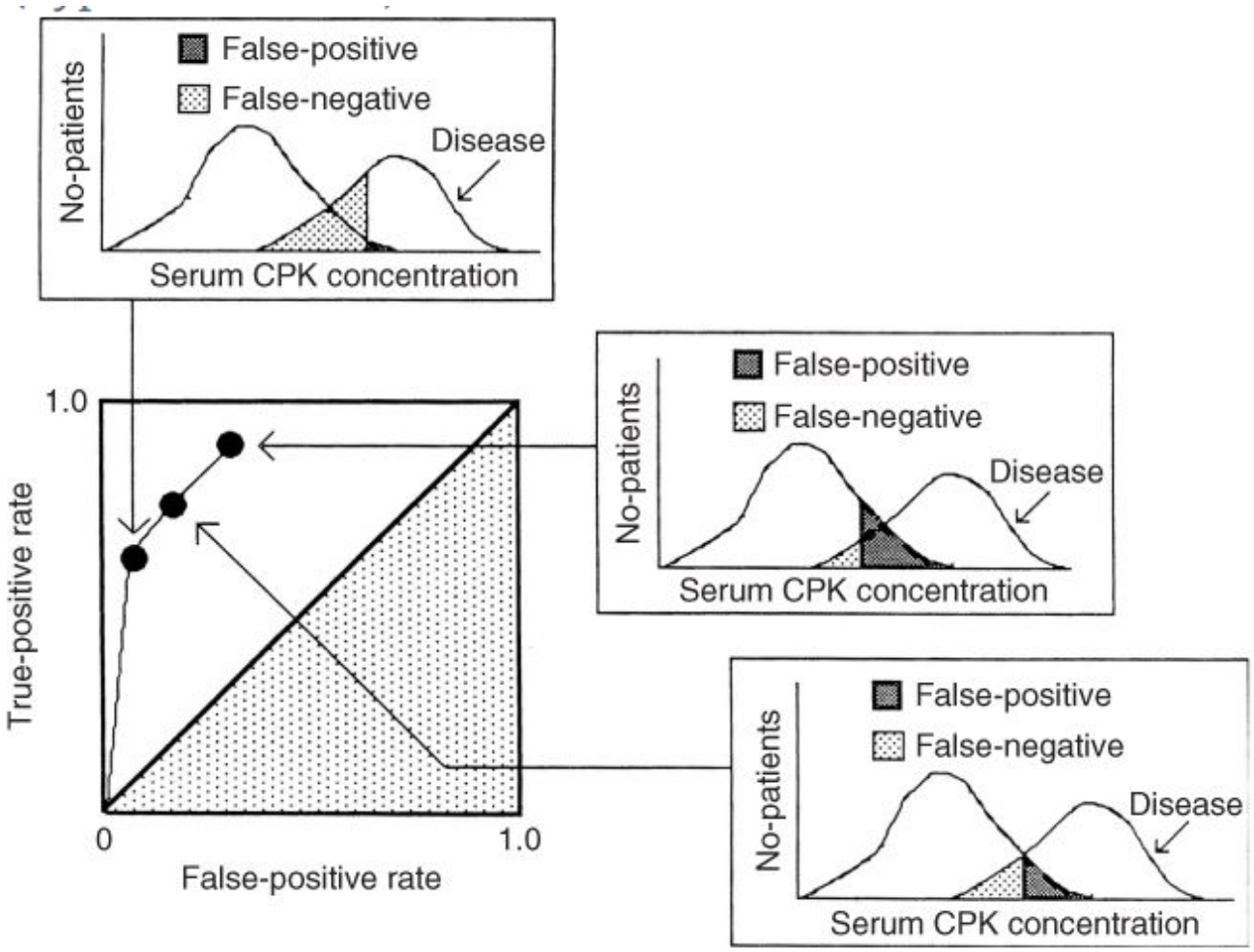
Move threshold to the right



- There is a tradeoff between sensitivity and specificity.
  - If threshold increases, sensitivity increases but specificity decreases.
  - If threshold decreases, specificity increases but sensitivity decreases.
  - Any increase in sensitivity will be accompanied by a decrease in specificity, or vice versa.

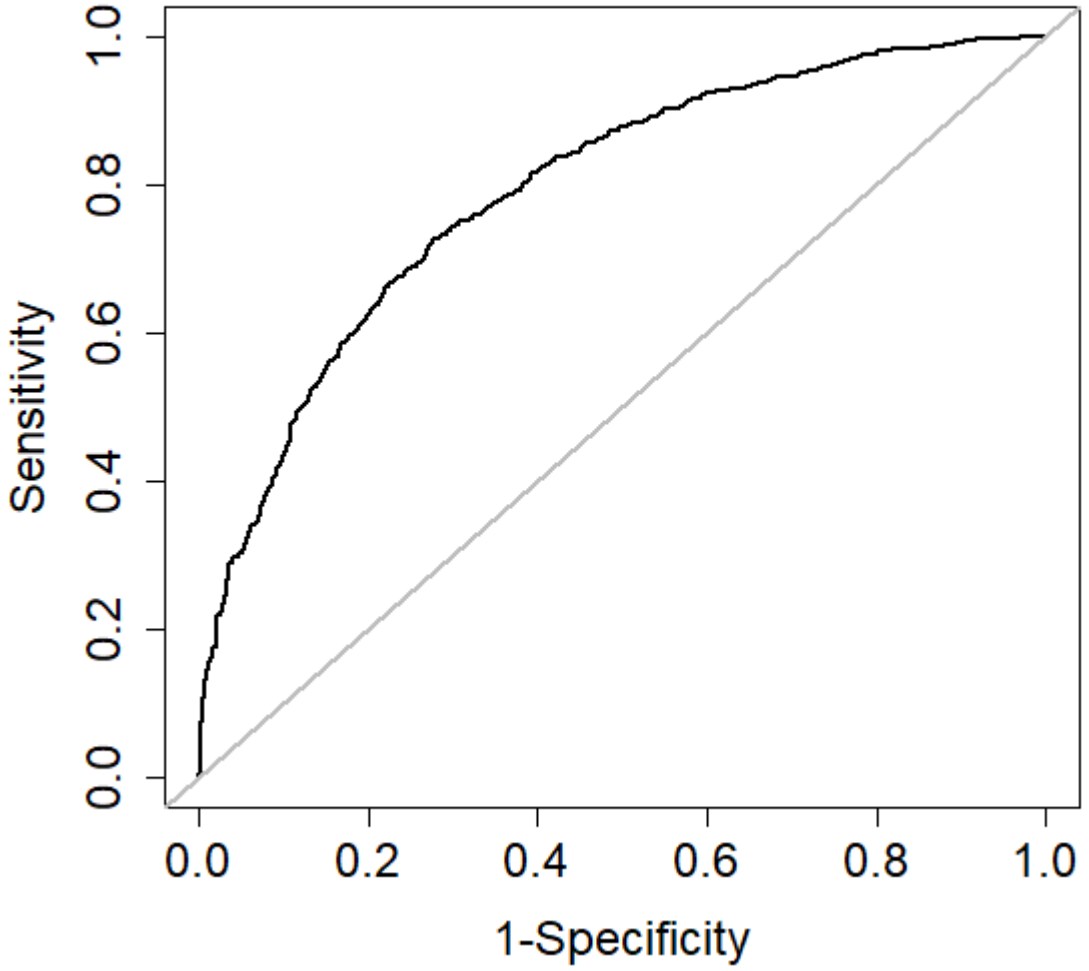
- **The receiver operating characteristics (ROC) curve** plots **sensitivity (or TPF)** versus **1-specificity (or FPF)** for all possible thresholds of the test result.

ROC Curve



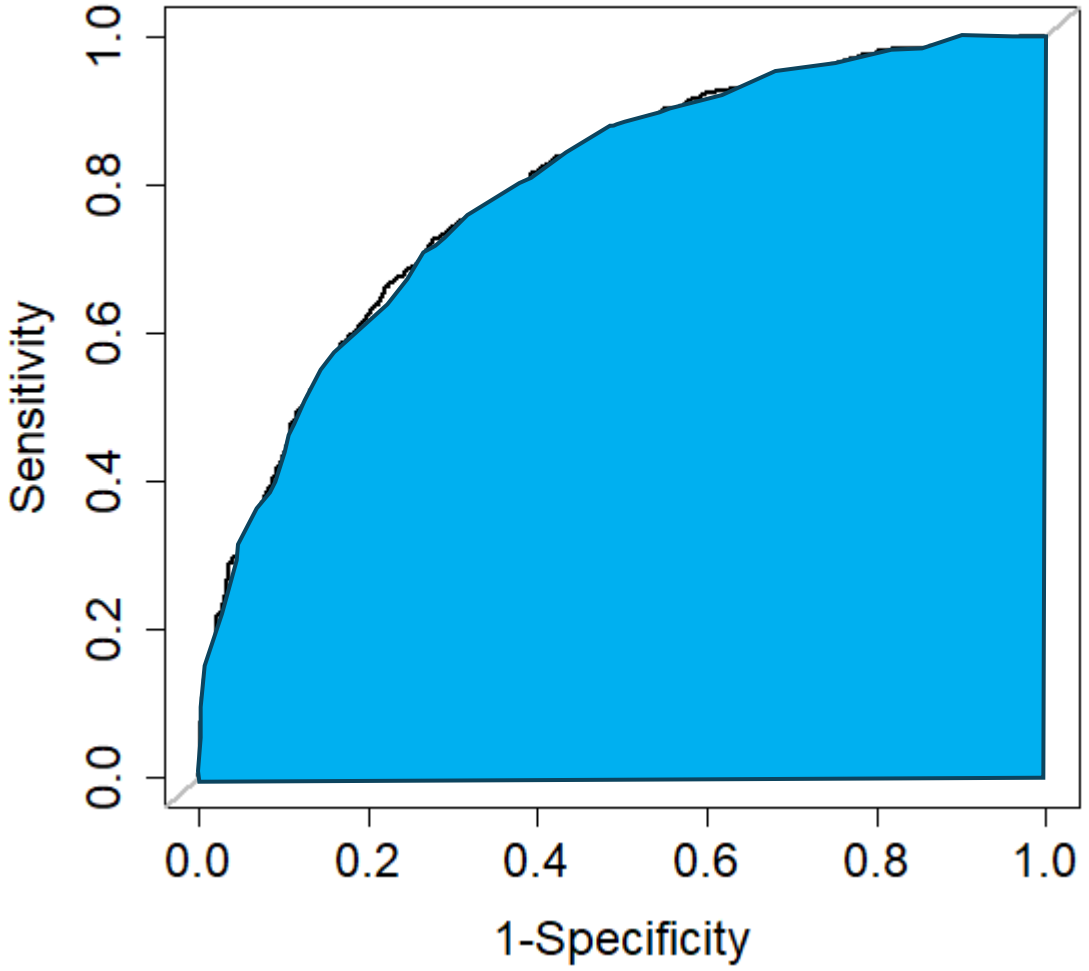
Harold C., et al. *Medical decision making*. ACP Press, 2007.

ROC Curve

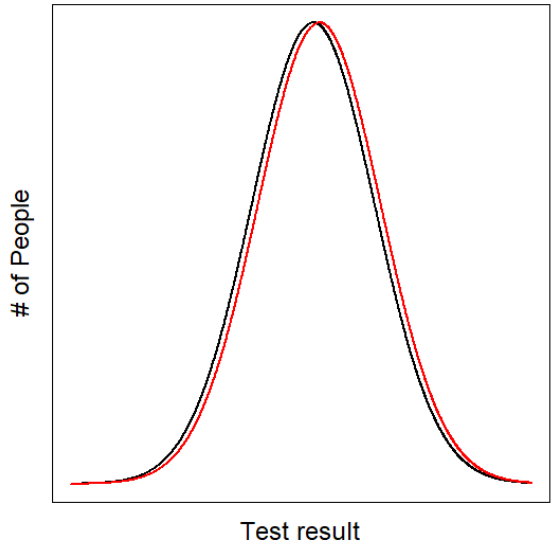


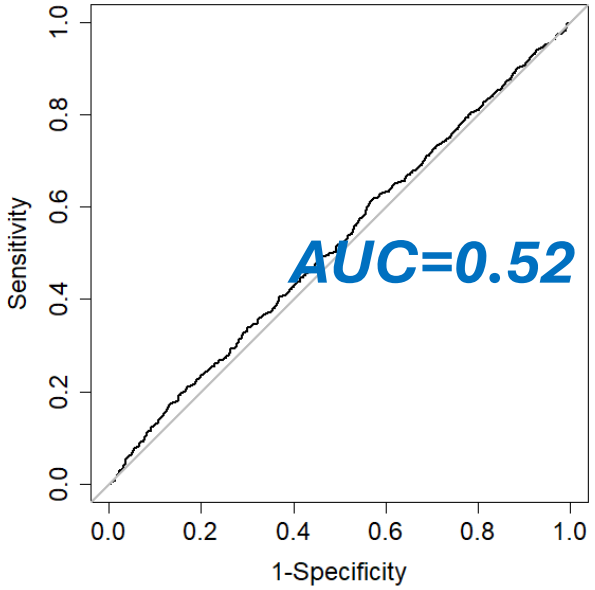
- **45 degree line**: TPF=FPF (*uninformative test*).
- **Perfect test**: the lower left corner, straight up to the upper left corner, and then to the upper right corner.

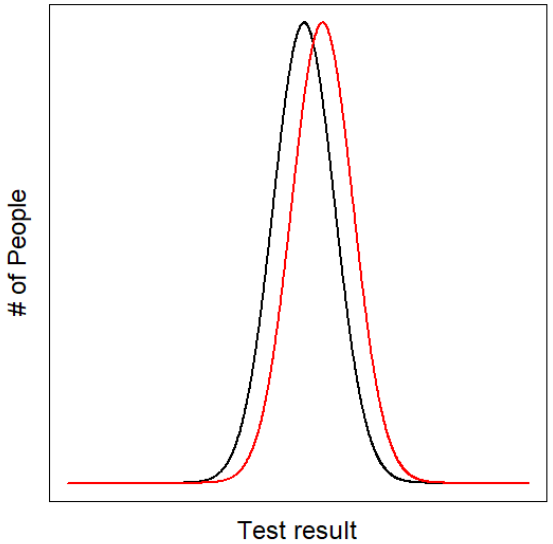
AUC

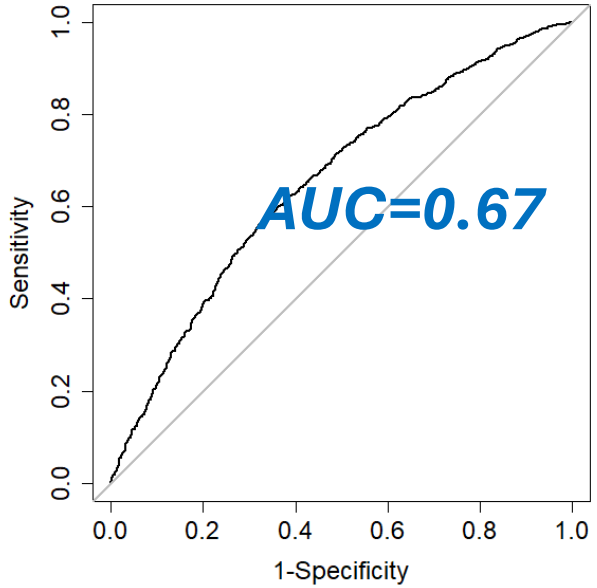


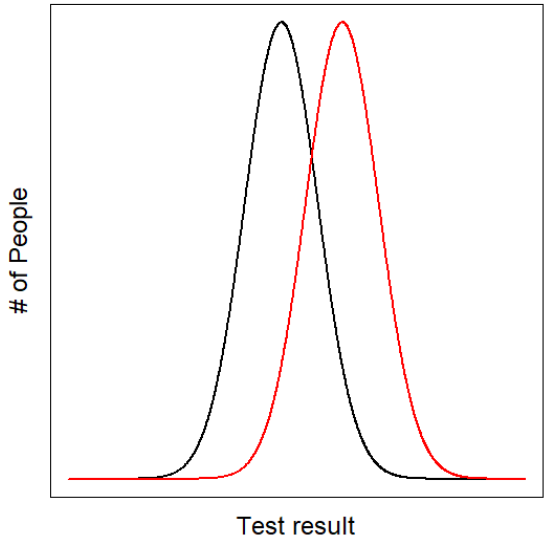
- Area under the ROC curve (AUC) is a measure of how well a test can distinguish between two groups.
  - AUC is ranged from 0 to 1.
  - AUC=1 for perfect test
  - AUC=0.5 for useless test

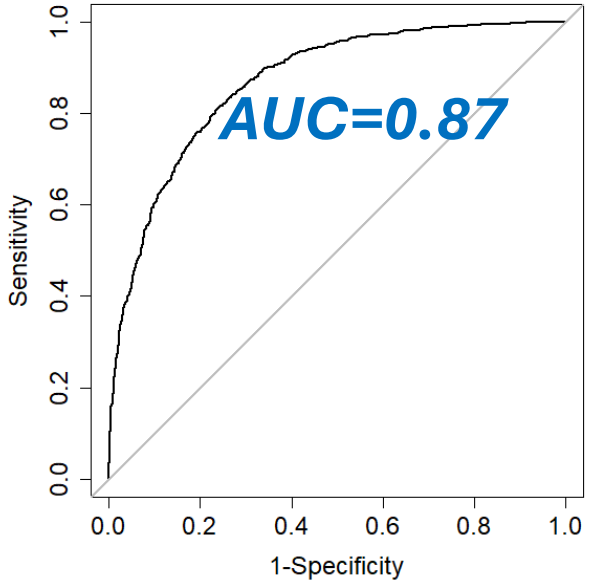

# of People
Test result

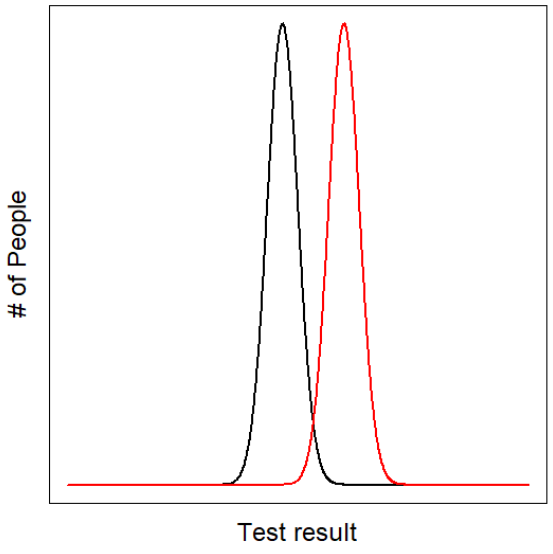

Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0
AUC=0.52

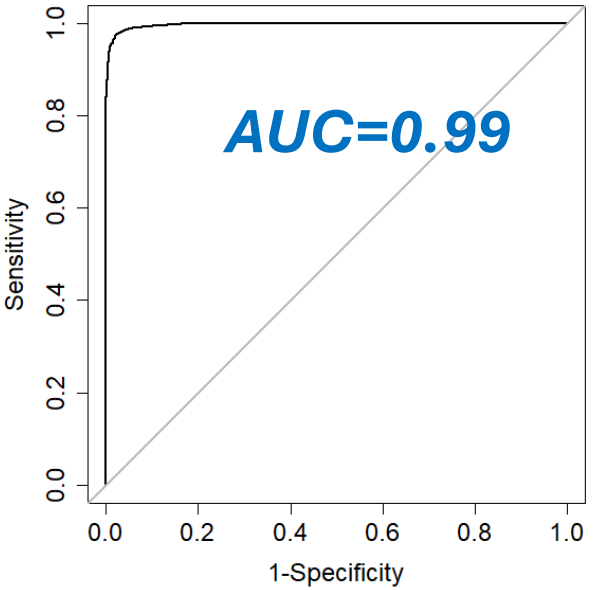

# of People
Test result


Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0
AUC=0.67


# of People
Test result


Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0
AUC=0.87

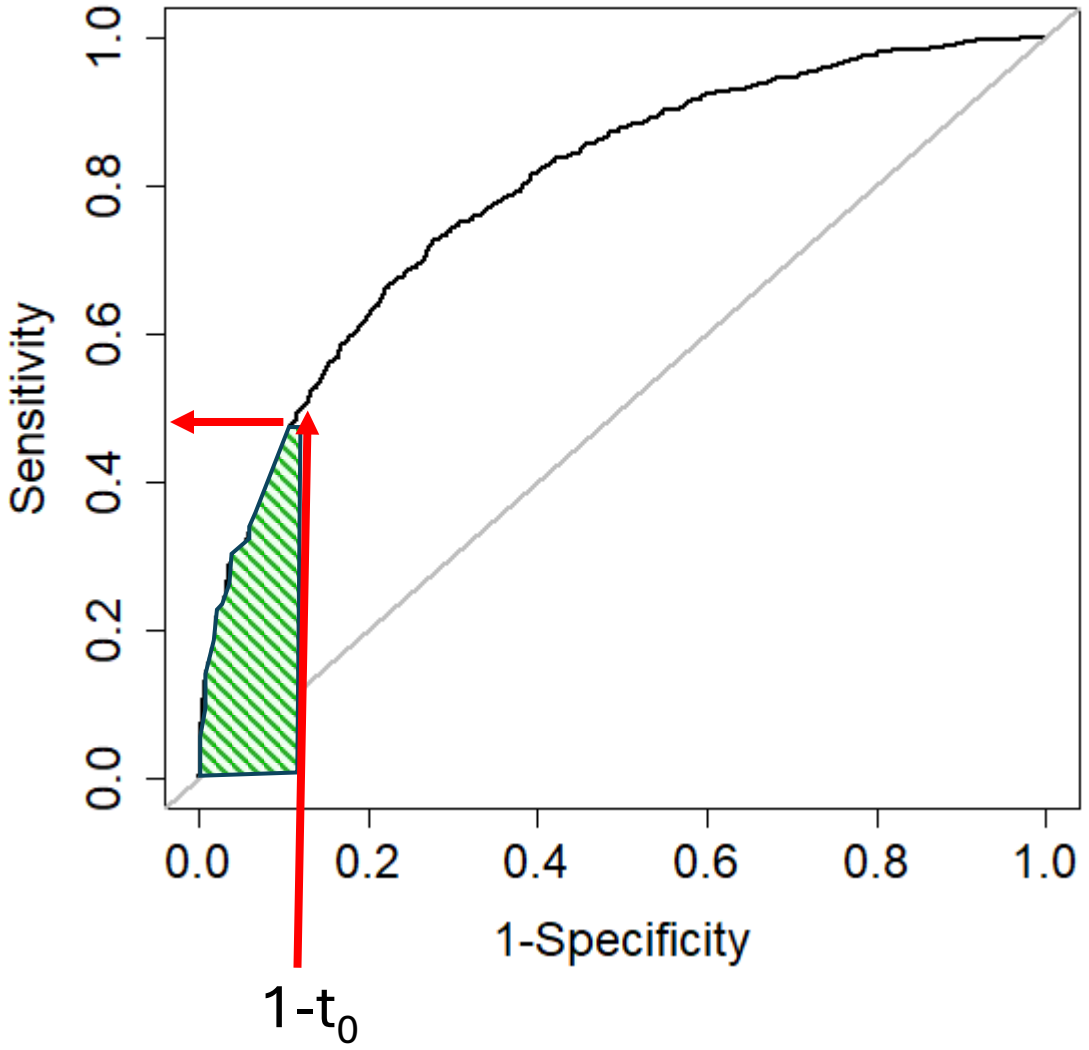

# of People
Test result


Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0
AUC=0.99

AUC

- AUC is a global measure based on “all thresholds”.
  - Higher AUC indicates a test can separate disease and non-disease group more accurately.
- AUC is a clinically relevant measure?
  - **No**. Not all thresholds are of interested.
  - For example, low specificity may not be useful for disease screening or diagnosis.
- Alternatives:
  - Interprets AUC on some thresholds that are of interest only.

Sensitivity at a fixed specificity



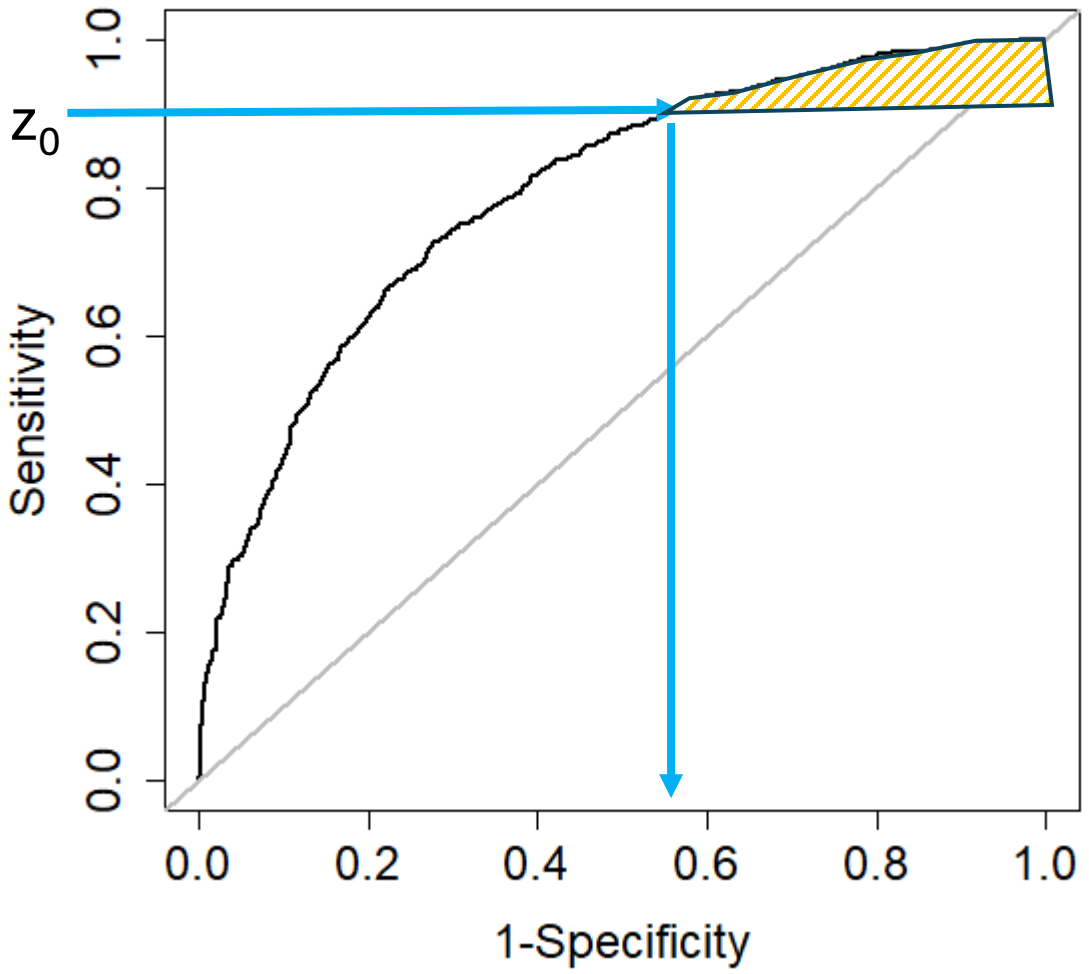
Sensitivity at t_0 specificity.

Partial AUC between 1 and t_0 specificities.

For example, sensitivity at 90% specificity (or TPF at 10% FPF.)

- For a useless test,
  - TPF at 10% FPF=0.1 (or 10%)
  - PAUC=$0.1^2/2$
- For a perfect test,
  - TPF at 10% FPF=1 or (100%)
  - PAUC=0.1

Specificity at a fixed sensitivity



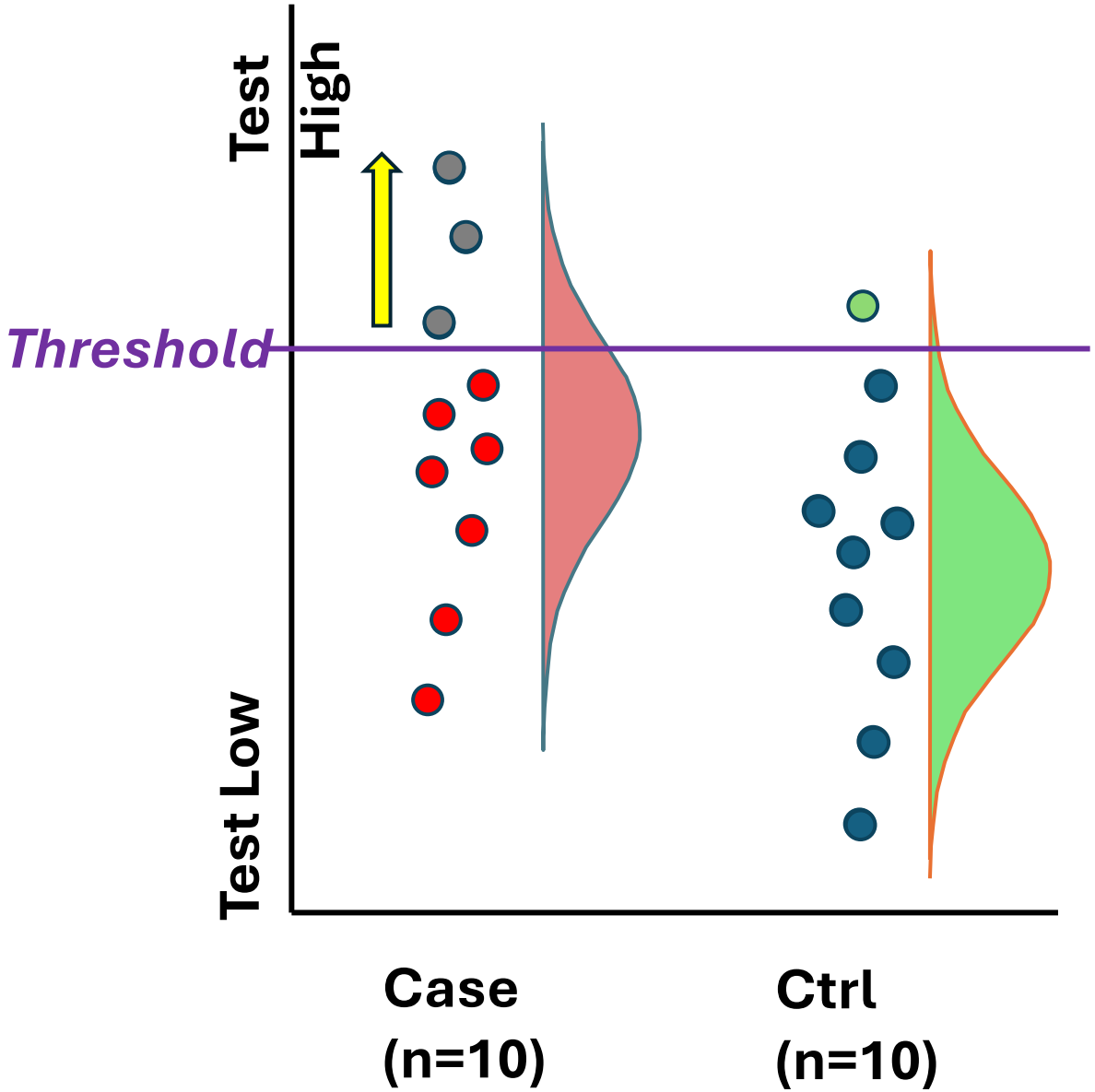
Specificity at z_0 sensitivity.

PAUC between 1 and z_0 sensitivity.

For example, specificity at 90% sensitivity (or FPF at 90% TPF)

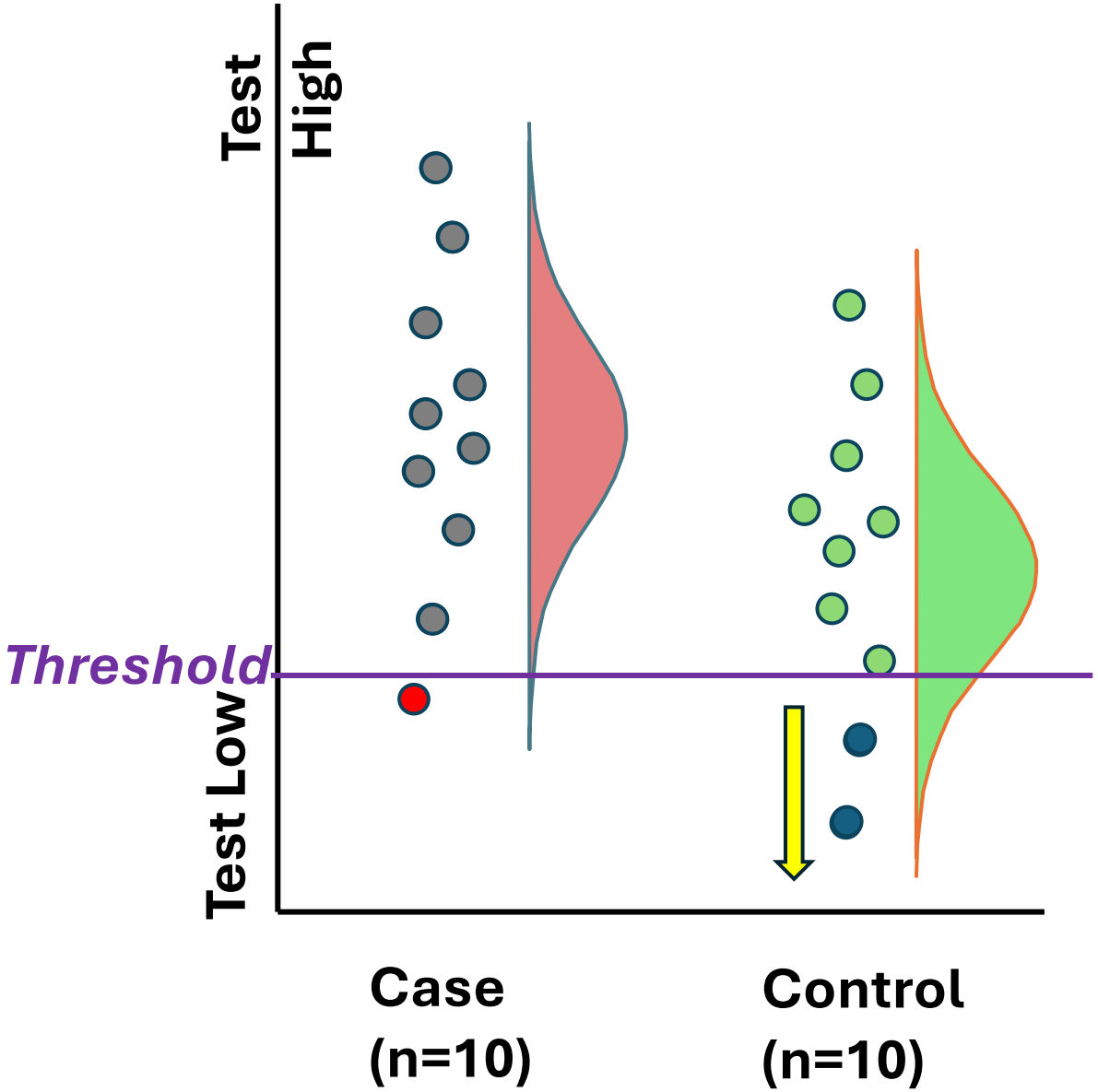
- For a useless test,
  - FPF at 90% TPF=0.9 (or 90%)
  - PAUC= $0.1^2/2$
- For a perfect test,
  - FPF at 90% TPF=0 (or 0%)
  - PAUC=0.1

Sensitivity at 90% specificity



- St at 90% sp=30%
- Easy to understand to non-statisticians

Specificity at 90% sensitivity



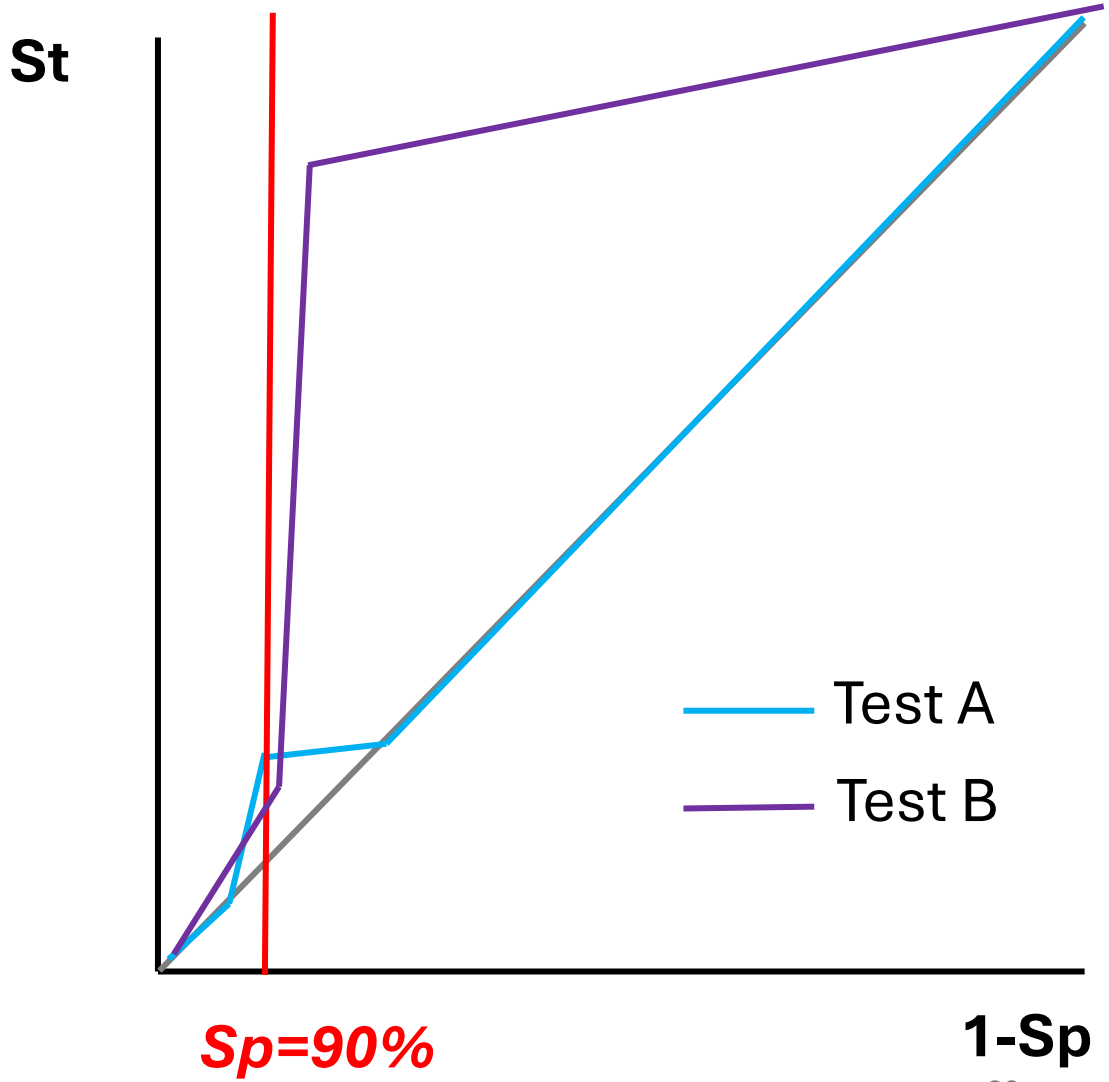
- Sp at 90% st=20%
- This is often less of interest.

Q. Why ROC first, instead of st and sp at a particular threshold?

Test A: St at 90% Sp is 20%

vs

Test B: St at 90% Sp is 15%



ROC Curve – Reverse Direction

Reversed Direction

- So far, we assume higher test result indicates more disease.
- What if higher test result indicates less disease?

		Disease	
		Present	Absent
Test	T+ (Test>th)	TP	FP
	T- (Test≤ th)	FN	TN

		Present	Absent
Test	T+ (Test≤ th)	TP	FP
	T- (Test>th)	FN	TN

- *Definitions of sensitivity & specificity (as well as the others) do not change.*

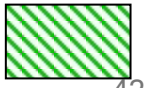
Case high direction

St =  / ( + )

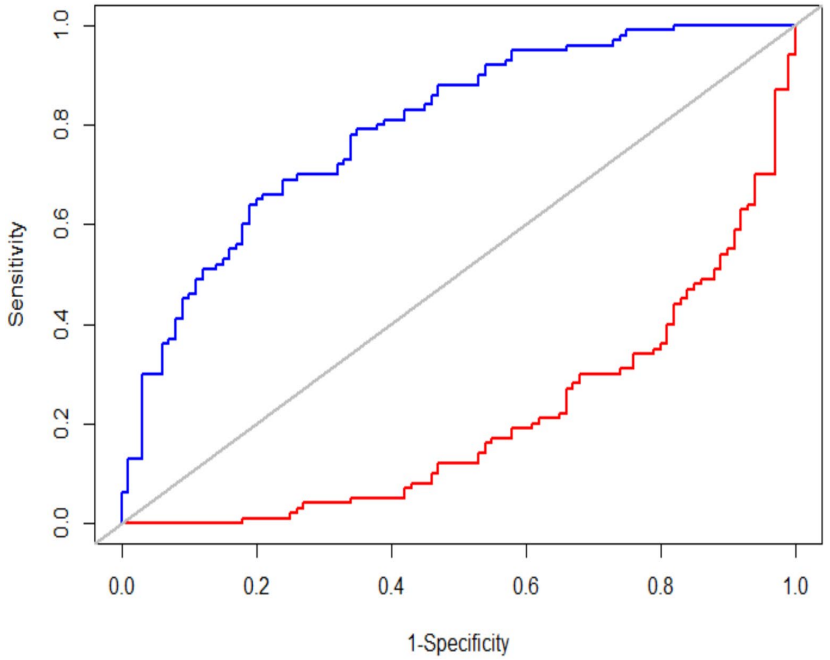
Sp =  / ( + )

Control high direction

St = / (+ )

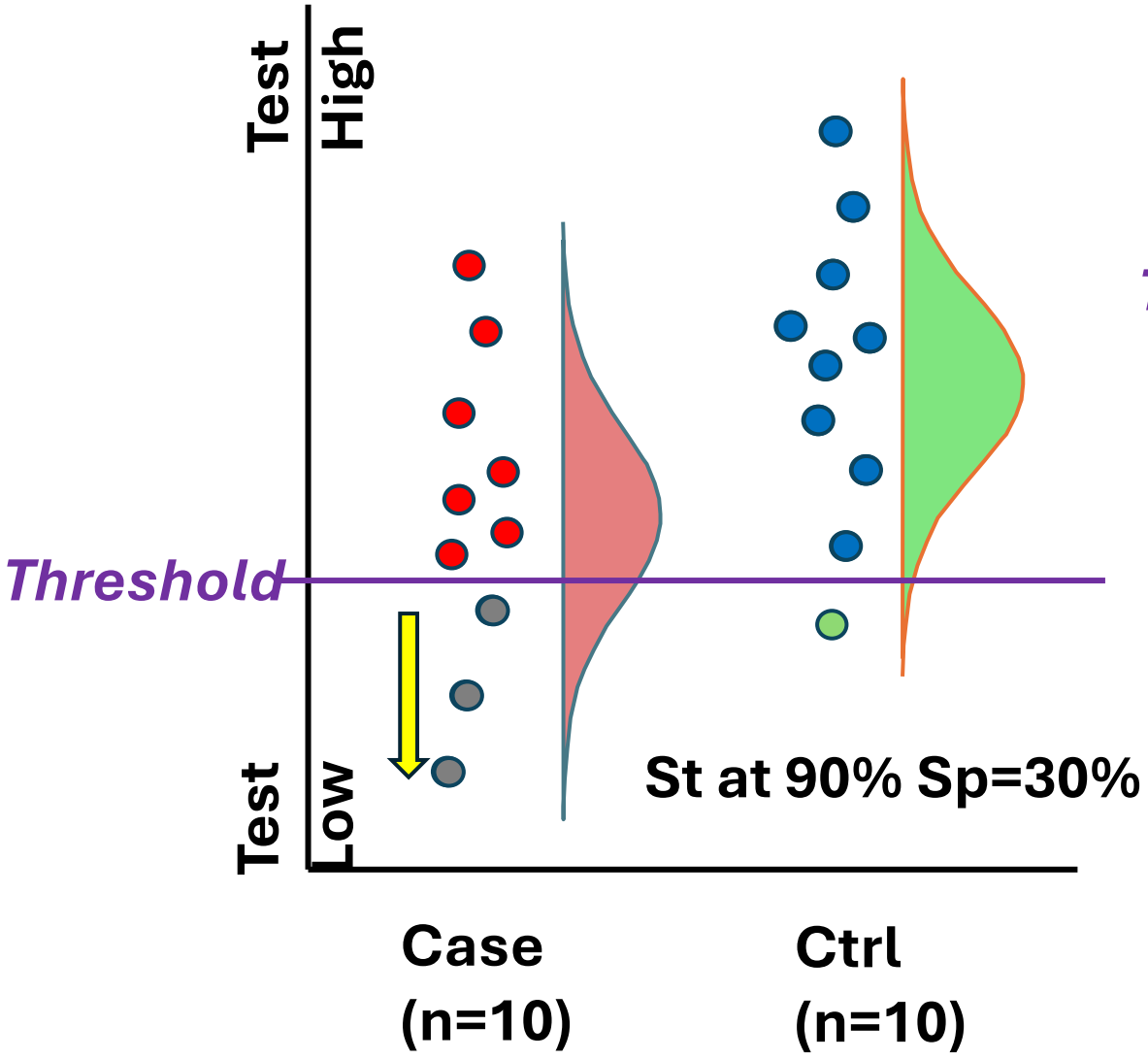
Sp = / (+ )

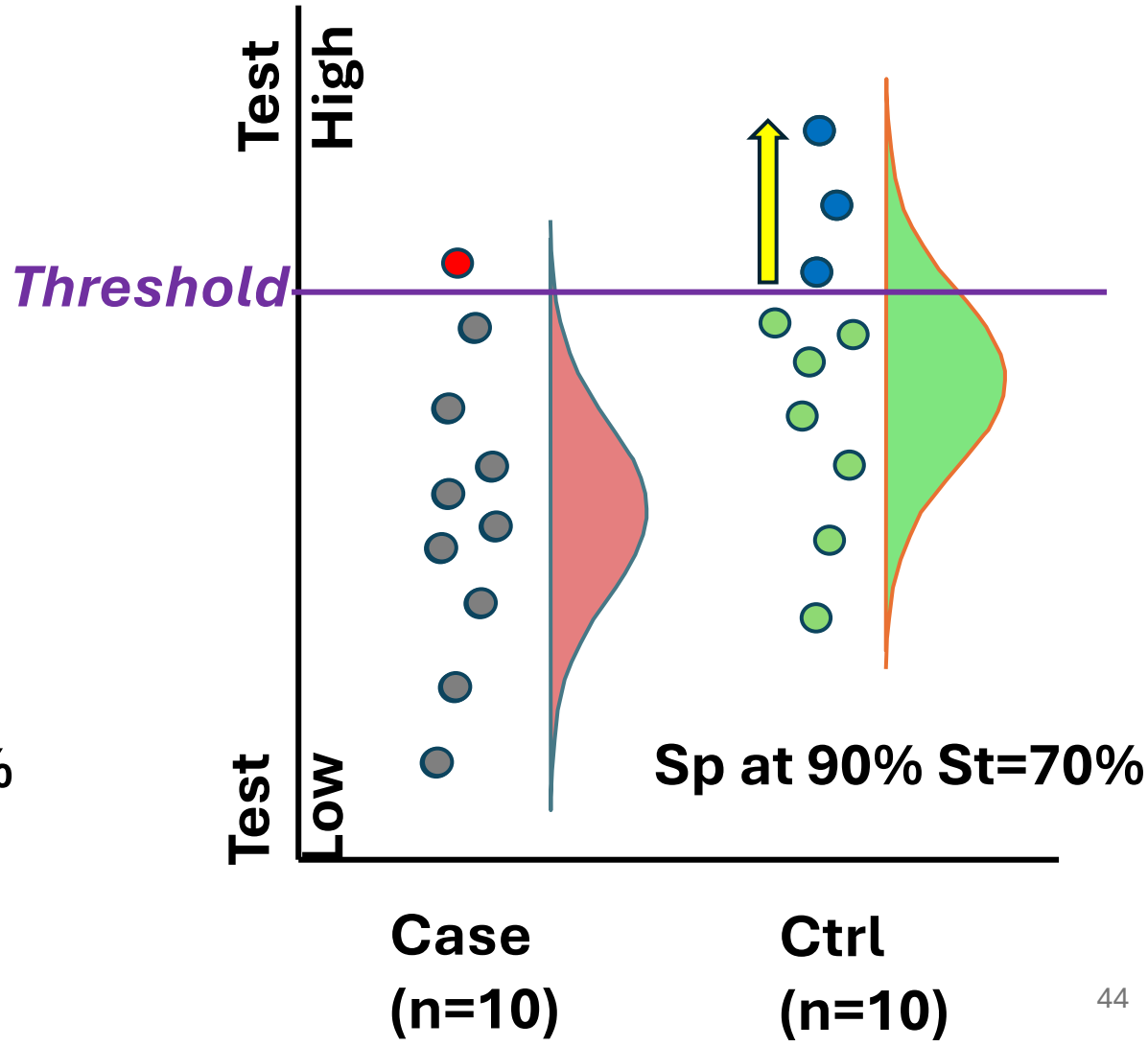
- There are two ROC curves for case high and control high directions:



- How to decide the direction?
  - Knowledge based.
  - Data-driven based, e.g. compare means, medians, or AUCs between case and control high directions.

Control high direction





Summary

- Univariate analysis:
  - Decide evaluation metric(s), e.g., AUC, PAUC, Sensitivity at t% specificity.
  - Method 1: Select biomarkers based on the chosen metric, e.g., AUC>0.8
  - Method 2: Select biomarkers based on p-values.
- How to choose t% ?
  - It depends on the situations.
  - In general, a higher t_0 is preferred, but there is no data point if t_0 is set too higher.