Learning Objectives

1. Learn how to calculate basic accuracy statistics such as sensitivity, specificity, likelihood ratios and AUC
2. Understand reasons for differences in diagnostic accuracy: real differences, bias, random variation, cut-offs.
3. Understand the difference between tests conducted under ideal conditions vs real conditions
4. Understand the role of higher-level approaches to performance evaluation

No Disclosures

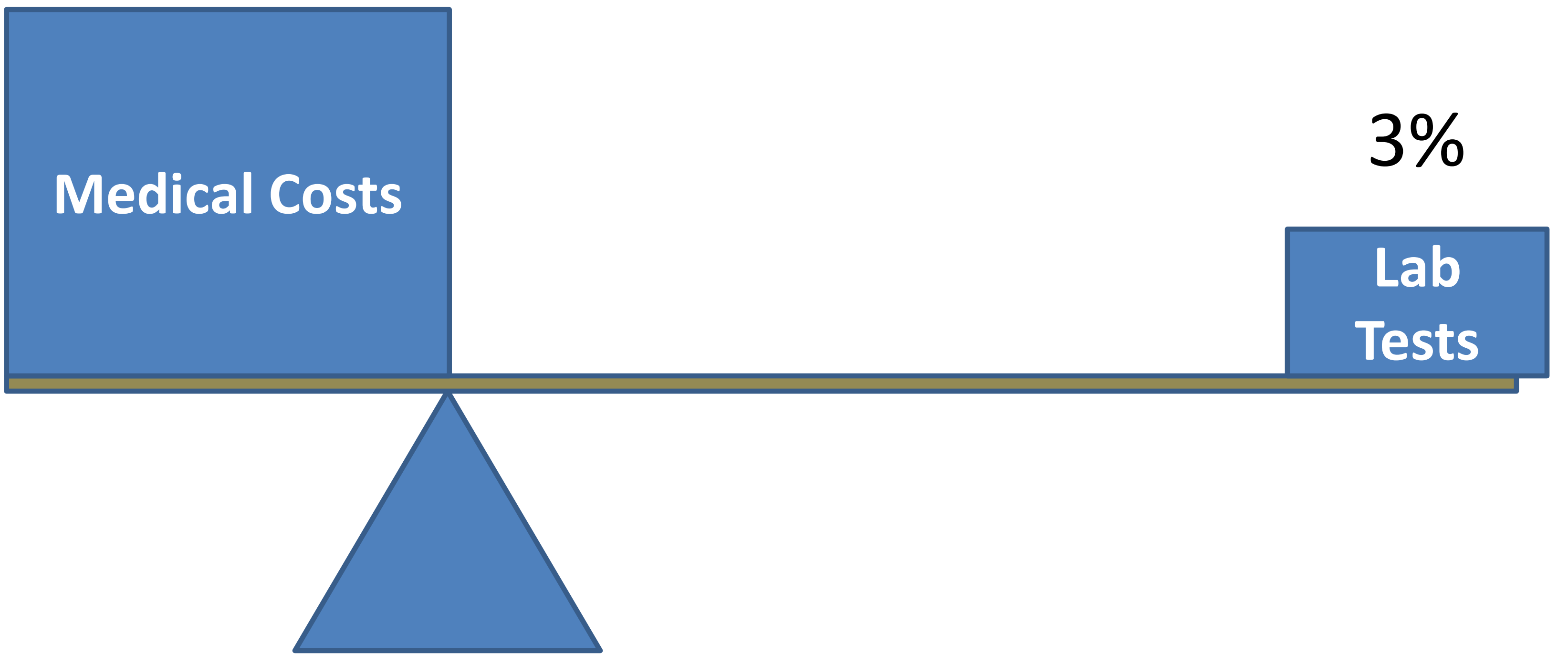
Testing a Test : Beyond Sensitivity and Specificity:

Robert Schmidt MBA MD PhD MMed

Tests are Central to Medicine

- Diagnosis
- Prognosis
- Monitoring
- Management

Tests Exert Great Leverage



Hierarchy of Effectiveness

Societal Impact

Cost effectiveness

Clinical effectiveness

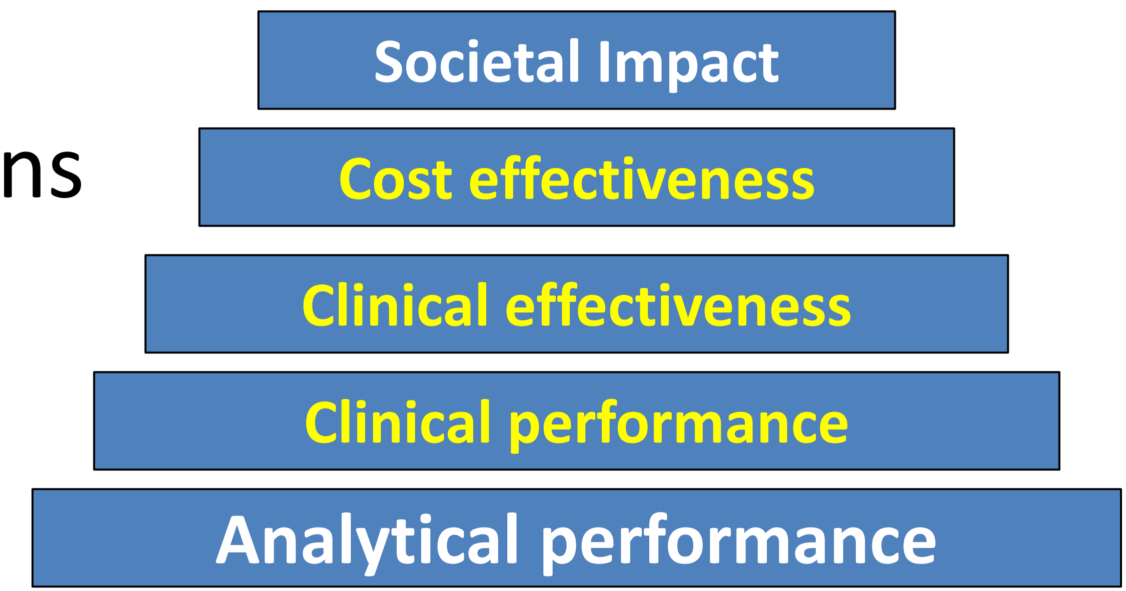
Clinical performance

Analytical performance

What this talk is about

Evaluating Tests:

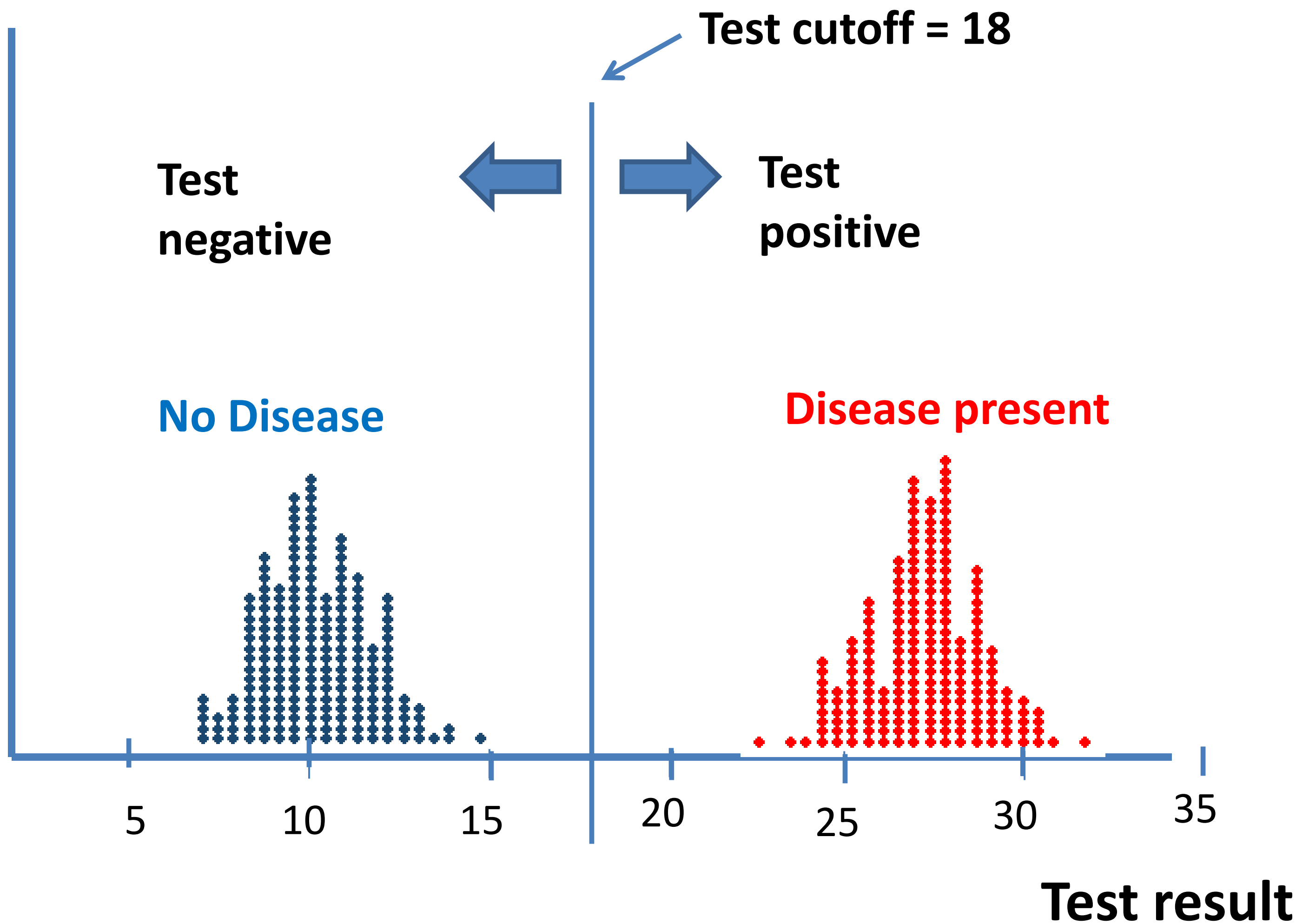
- Accuracy
- Usefulness
- Test Comparisons
- Limitations
- Future Directions



Case:

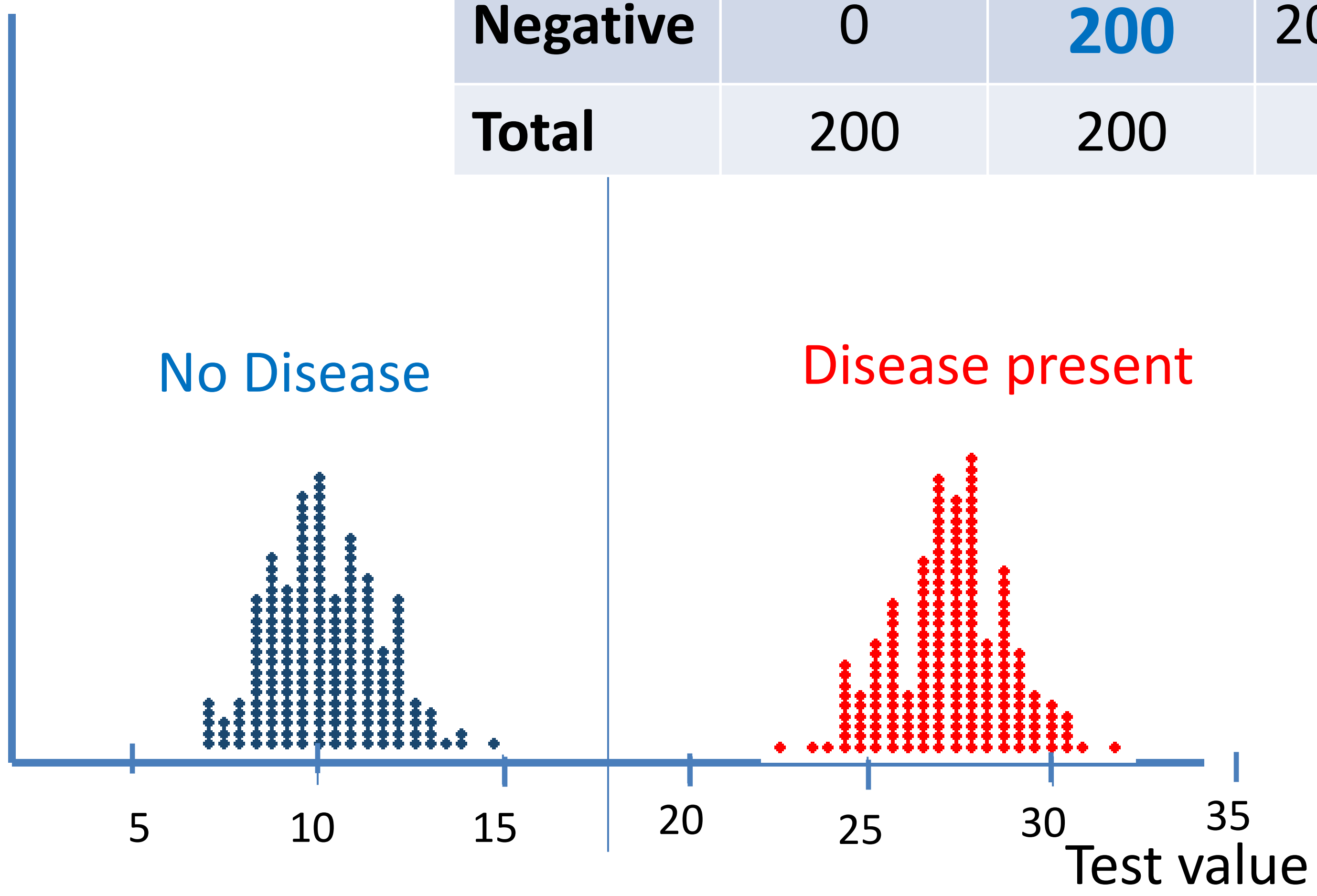
Your father has just returned from his annual physical. His doctor suggested that he consider a prostatic specific antigen (PSA) test to screen for prostate cancer. He is unsure what to do and asks your advice. Should he take the test?

The basic task: classification



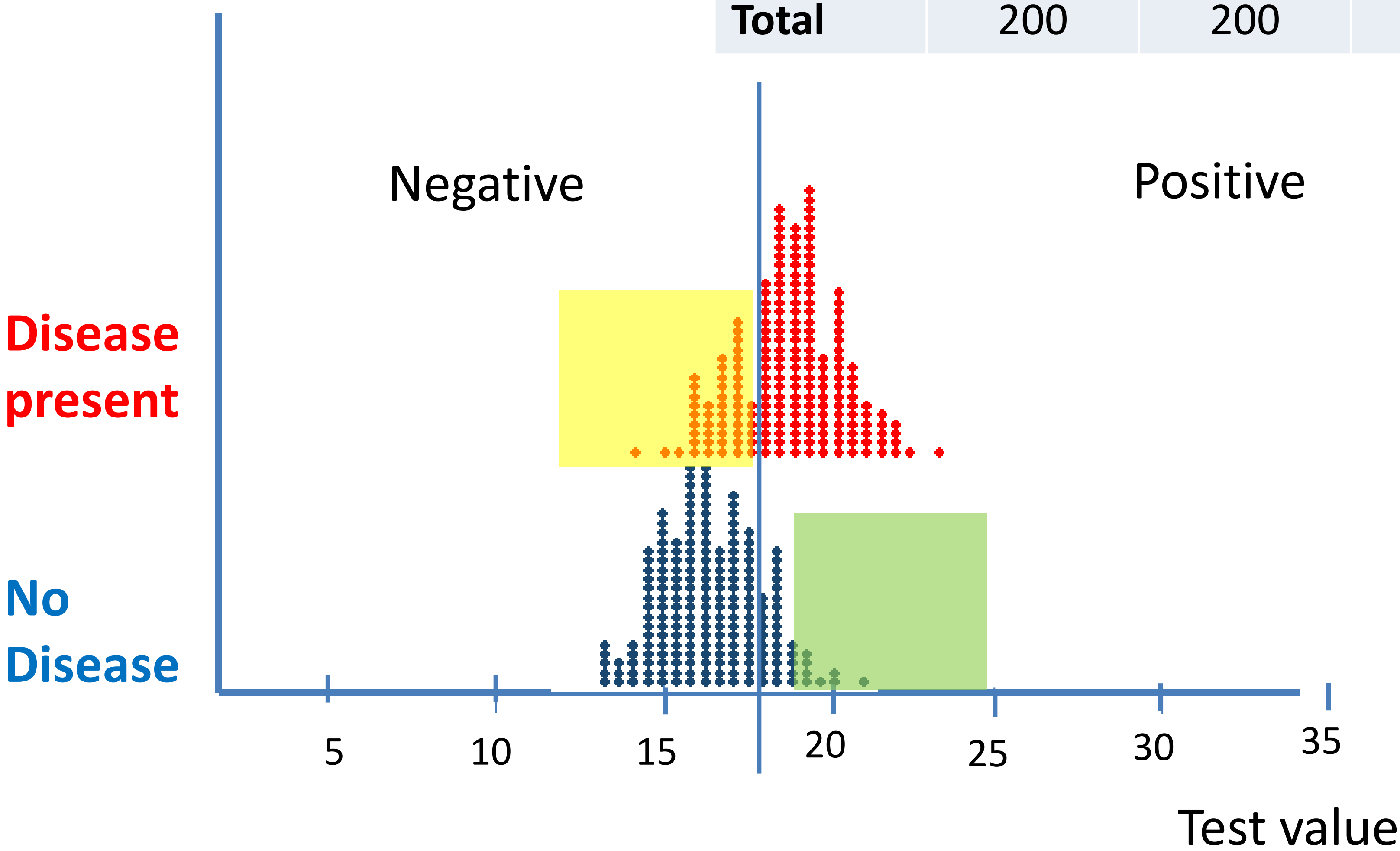
Perfect Test:

	Disease		
Test	**Present**	**Absent**	**Total**
Positive	**200**	0	200
Negative	0	**200**	200
Total	200	200	

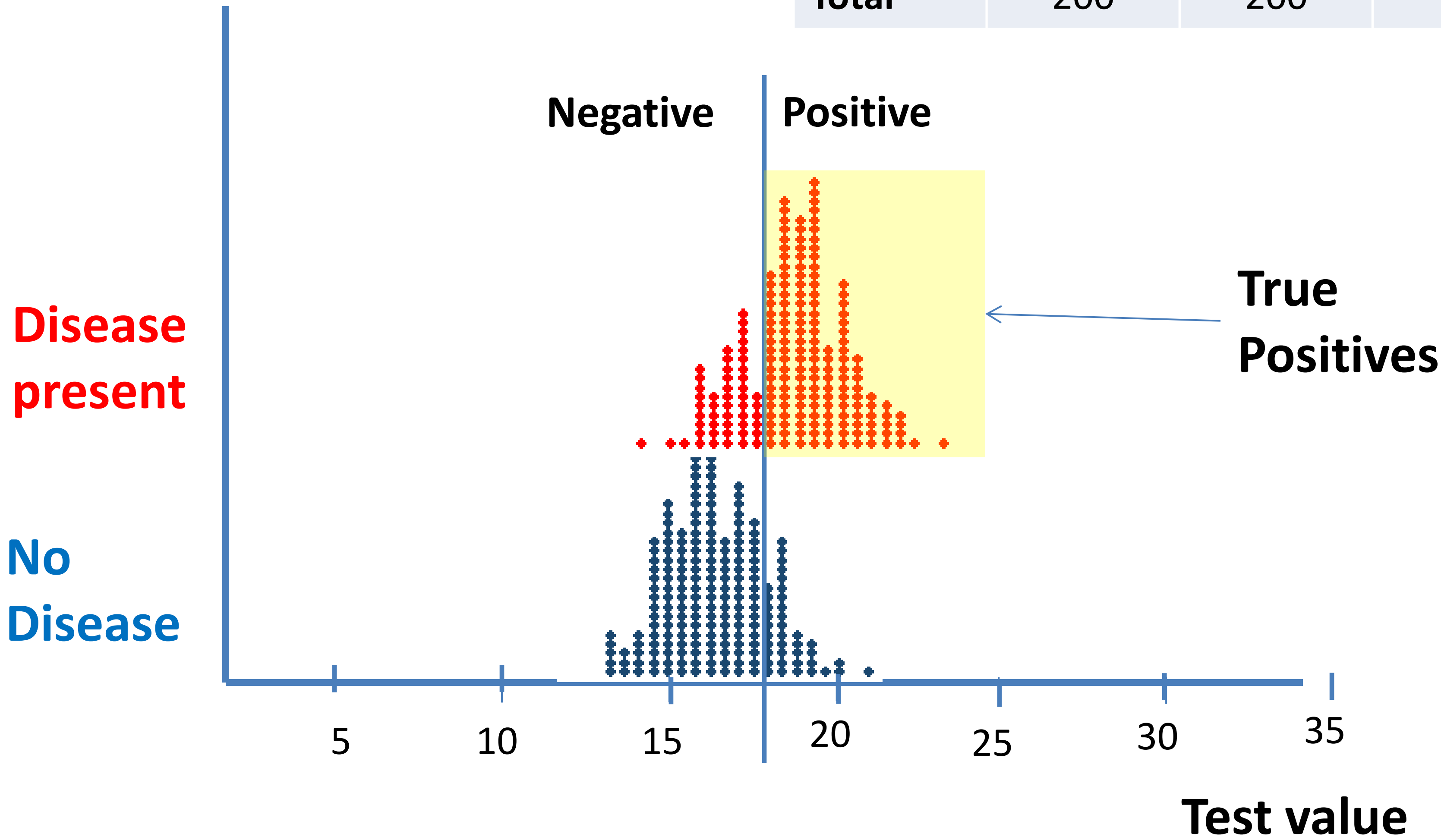


Not so Perfect Test:

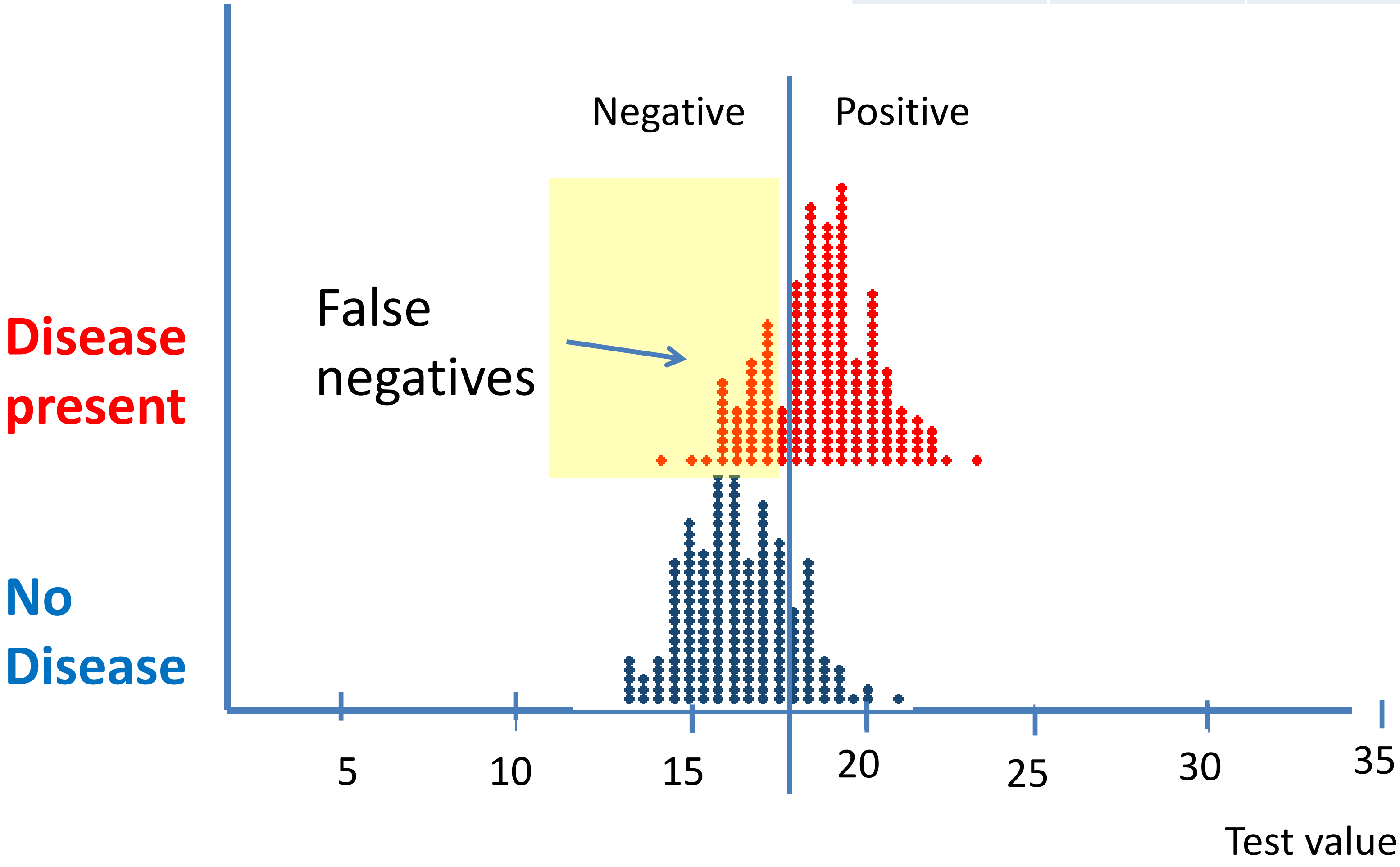
	Disease		
Test	**Present**	**Absent**	**Total**
Positive	148	**40**	188
Negative	**52**	160	212
Total	200	200	



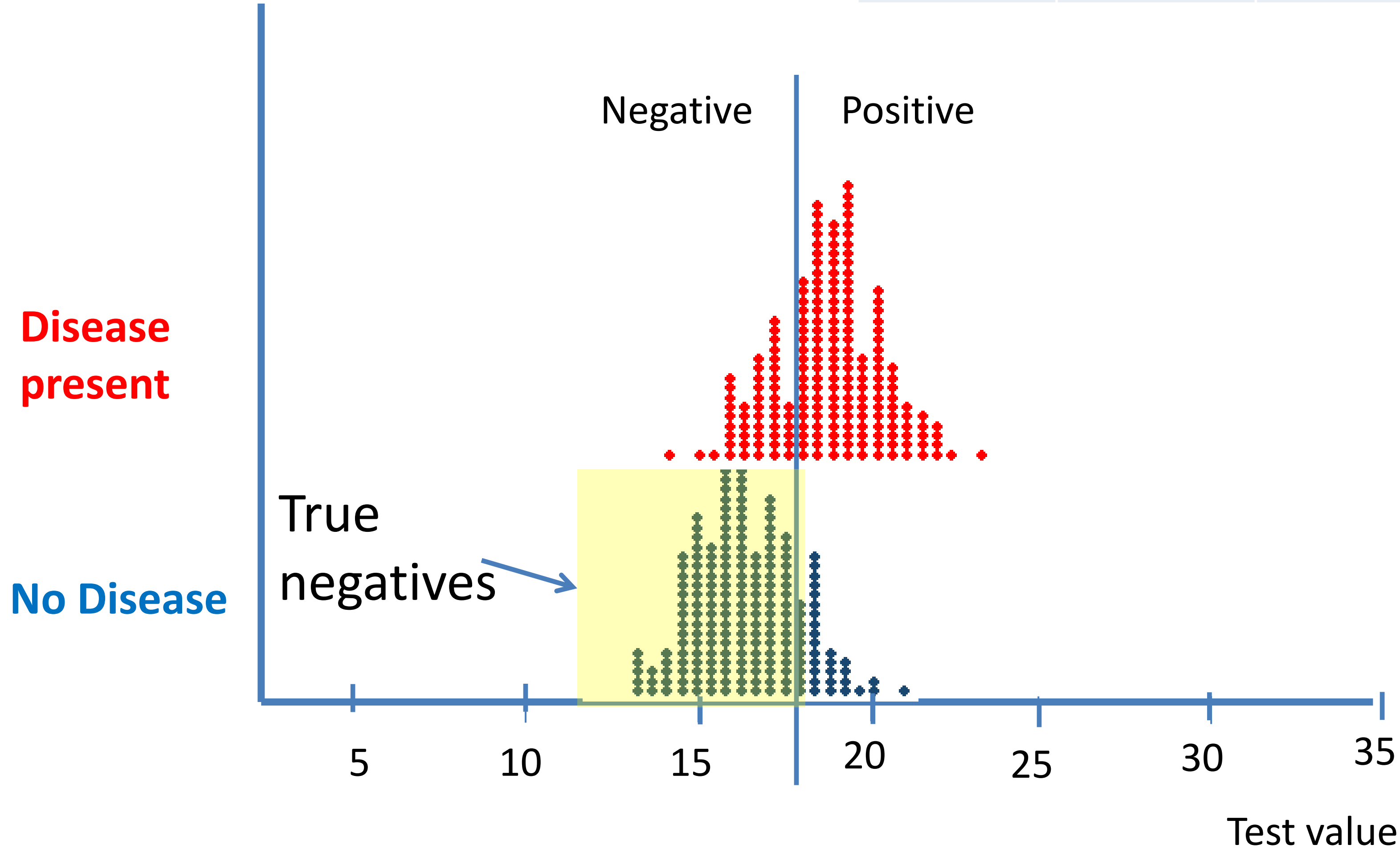
	Disease		
Test	**Present**	**Absent**	**Total**
Positive	**148**	40	188
Negative	52	160	212
Total	200	200	



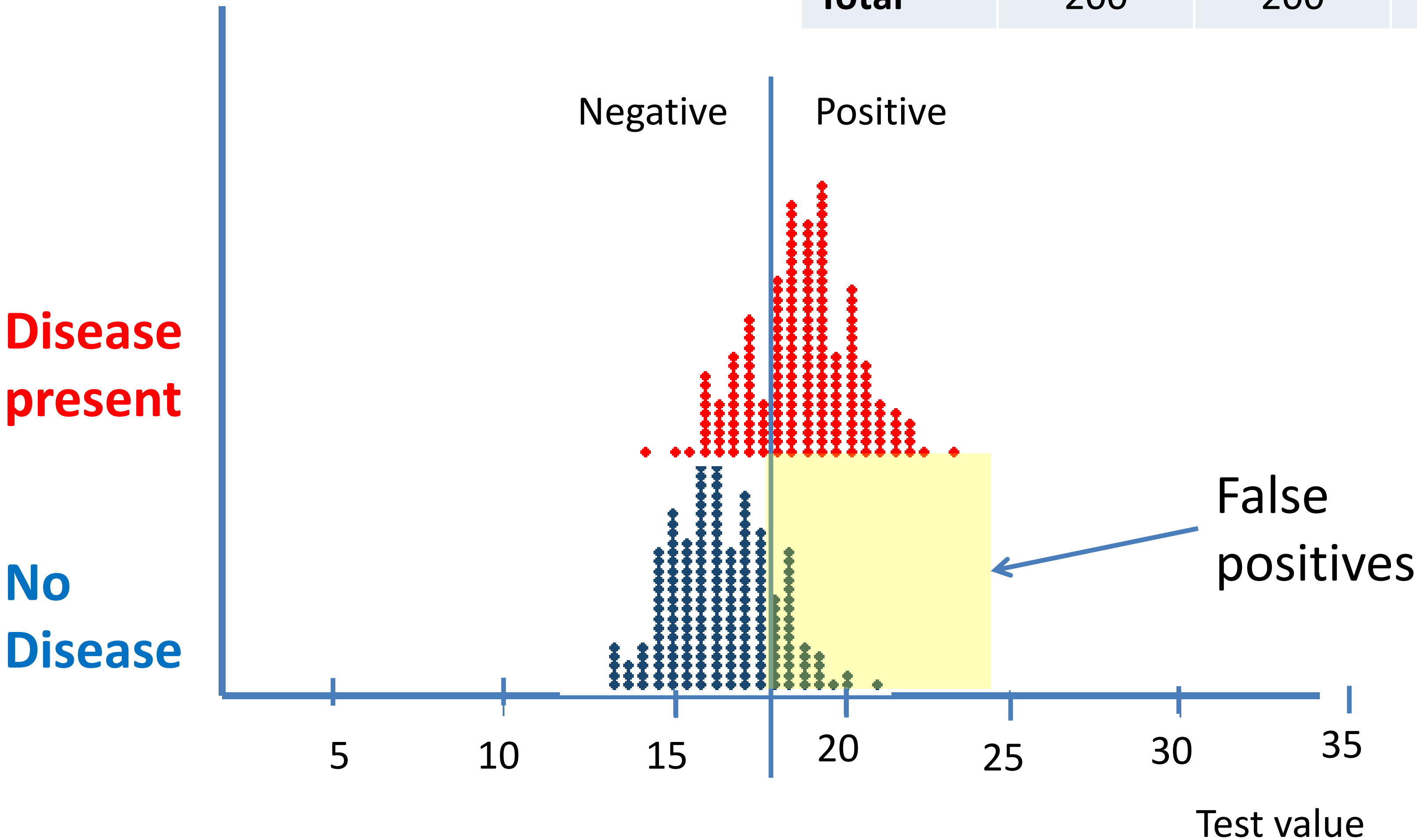
	Disease		
Test	**Present**	**Absent**	**Total**
Positive	148	40	188
Negative	**52**	160	212
Total	200	200	



	Disease		
Test	**Present**	**Absent**	**Total**
Positive	148	40	188
Negative	52	**160**	212
Total	200	200	



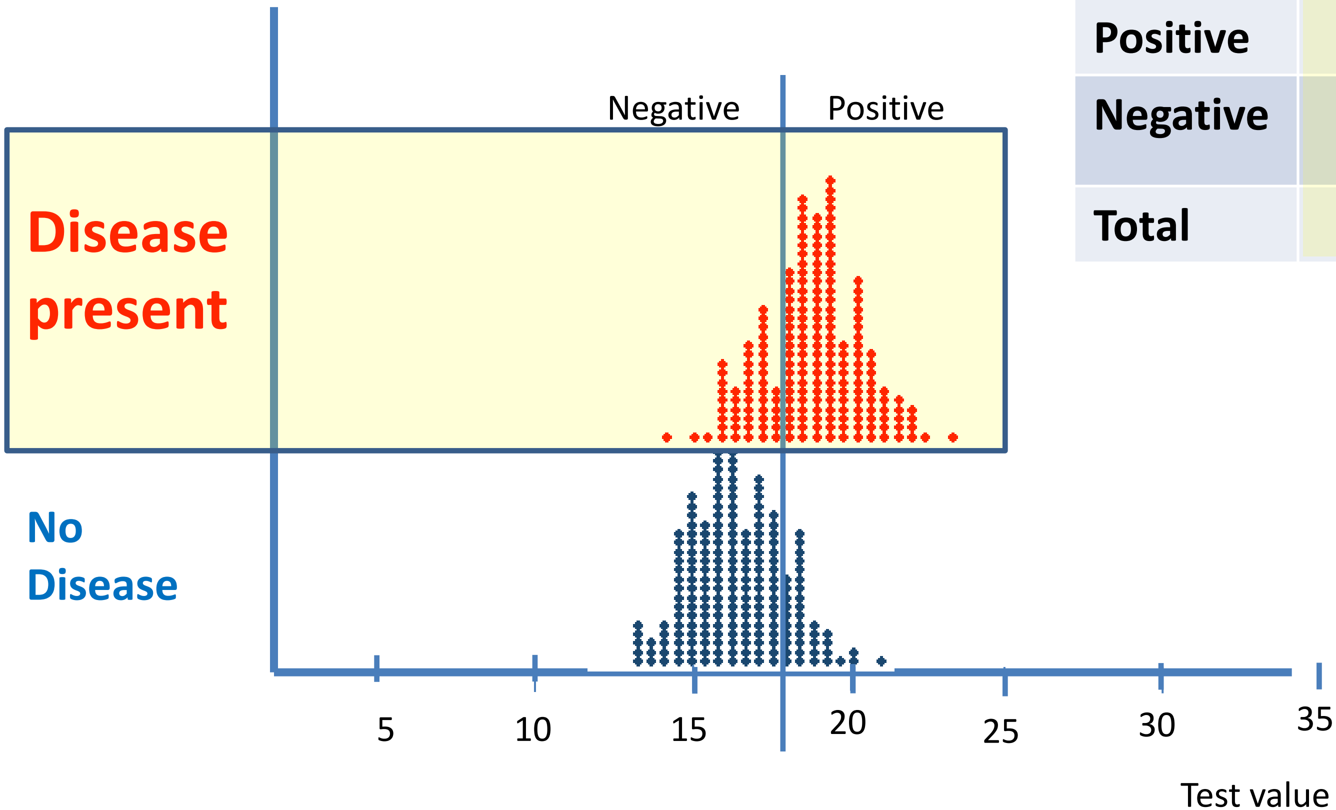
	Disease		
Test	**Present**	**Absent**	**Total**
Positive	148	**40**	188
Negative	52	160	212
Total	200	200	



How well did we classify those with disease?

Sensitivity = 148/200 = 74%

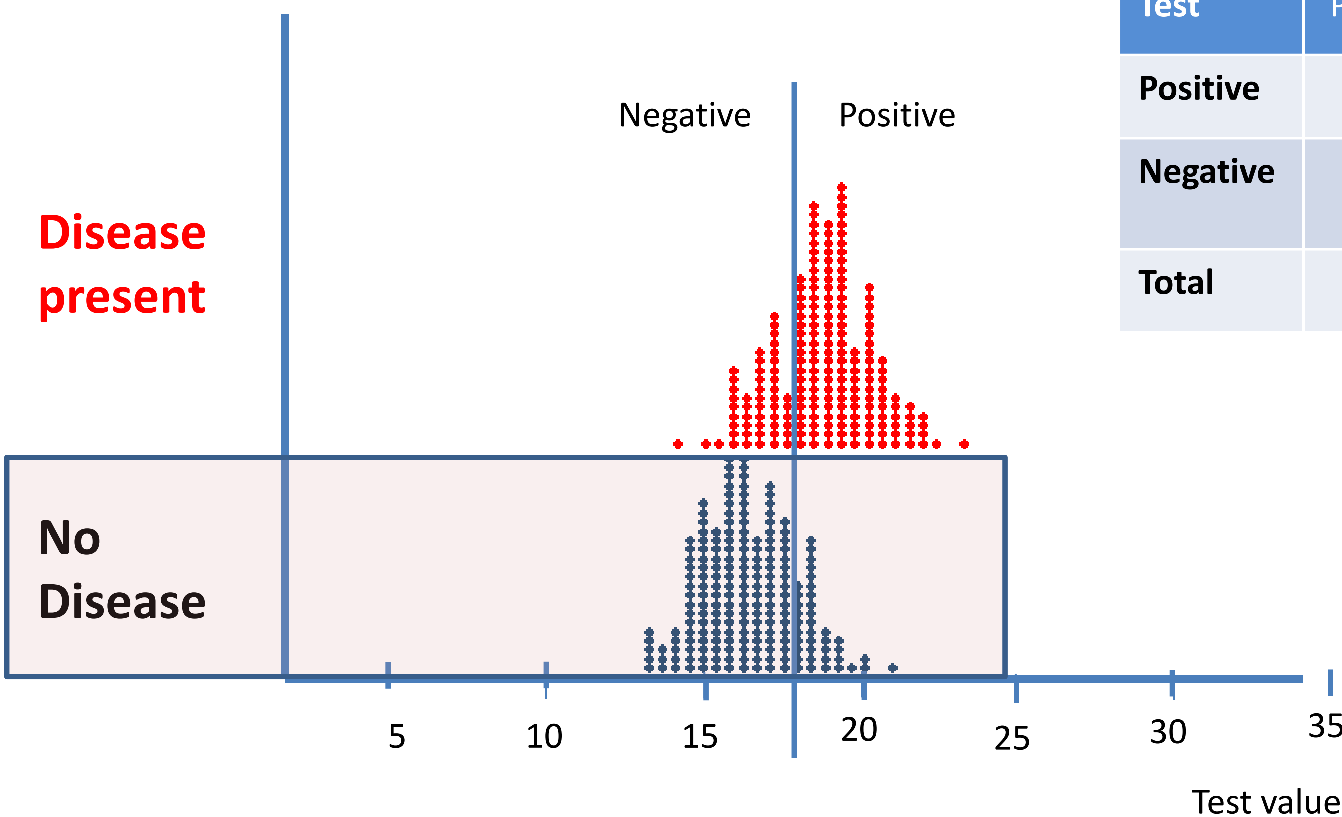
Test	Disease Present	Absent	Total
Positive	**148**	40	188
Negative	**52**	160	212
Total	**200**	200	



How well did we classify those without disease?

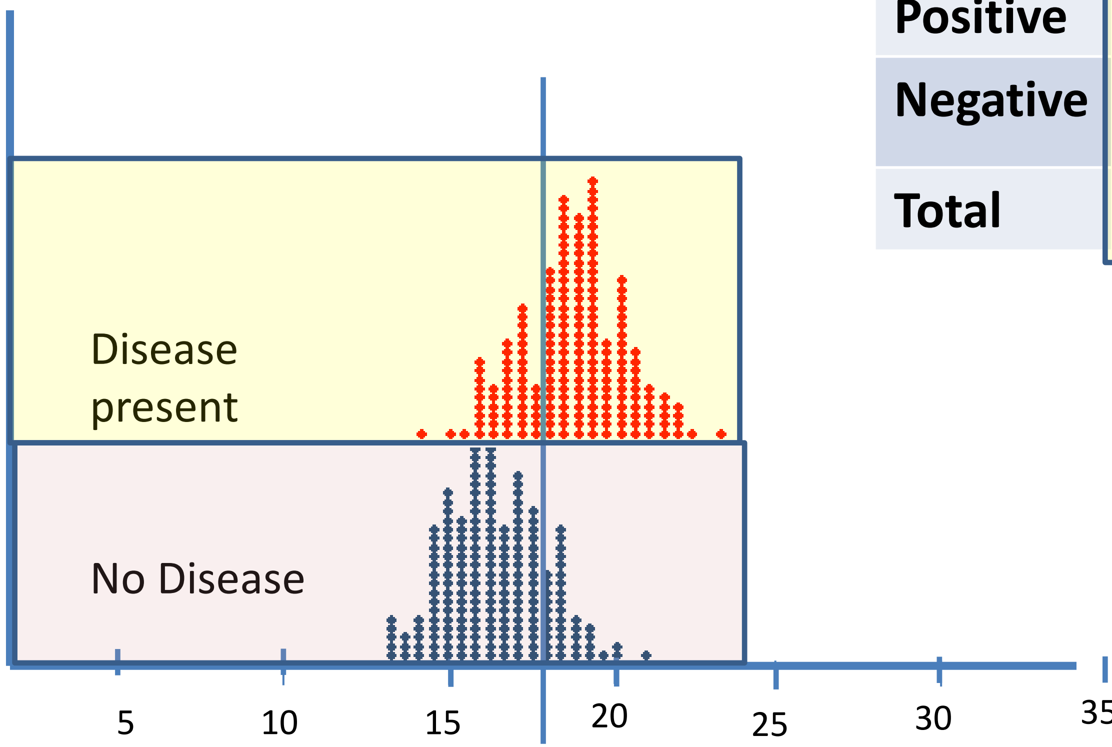
Specificity = 160/200 = 80%

	Disease		
Test	Present	Absent	Total
Positive	148	40	188
Negative	52	160	212
Total	200	200	



Sensitivity = accuracy in the **diseased** group
Specificity = accuracy in the **nondiseased** group

	Disease		
Test	**Present**	**Absent**	**Total**
Positive	148	40	188
Negative	52	160	212
Total	200	200	



Two useful mnemonics*

SnNout:

High ***Sen***sitivity Test with a ***N***egative result rules ***out***

SpPin:

High ***Sp***ecificity Test with a ***p***ositive result rules ***in***

*Guyatt & Sackett, Guide to EBM, 2002

PSA test

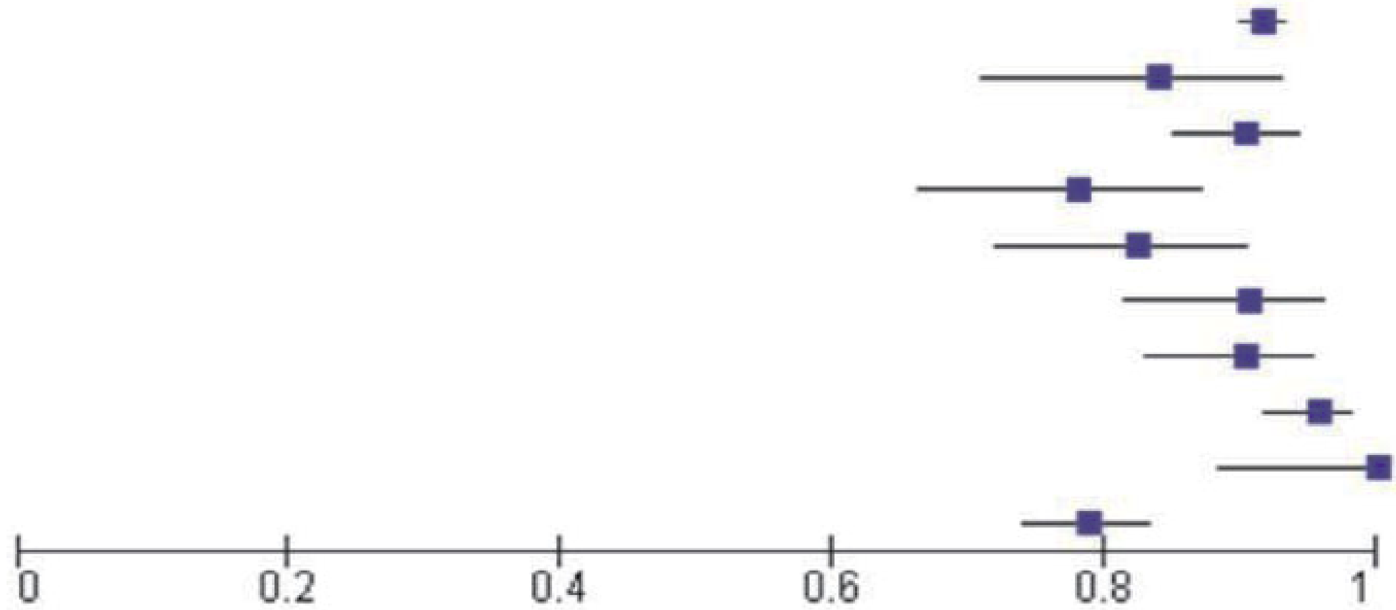
- Sensitivity 90%
- Specificity 20%

How might this test be useful?

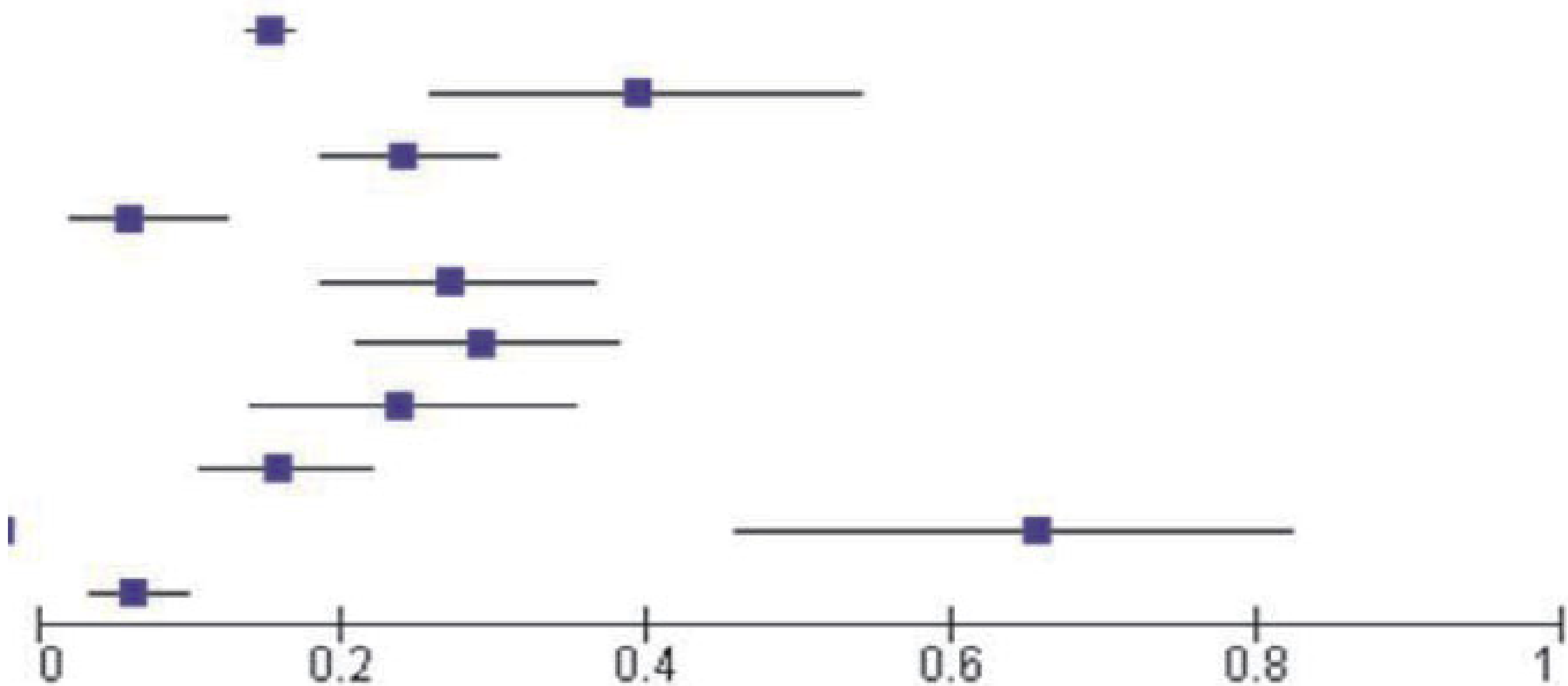
- **SnNout?**
- **SpPin?**

Results May Vary……

Sensitivity of PSA Tests



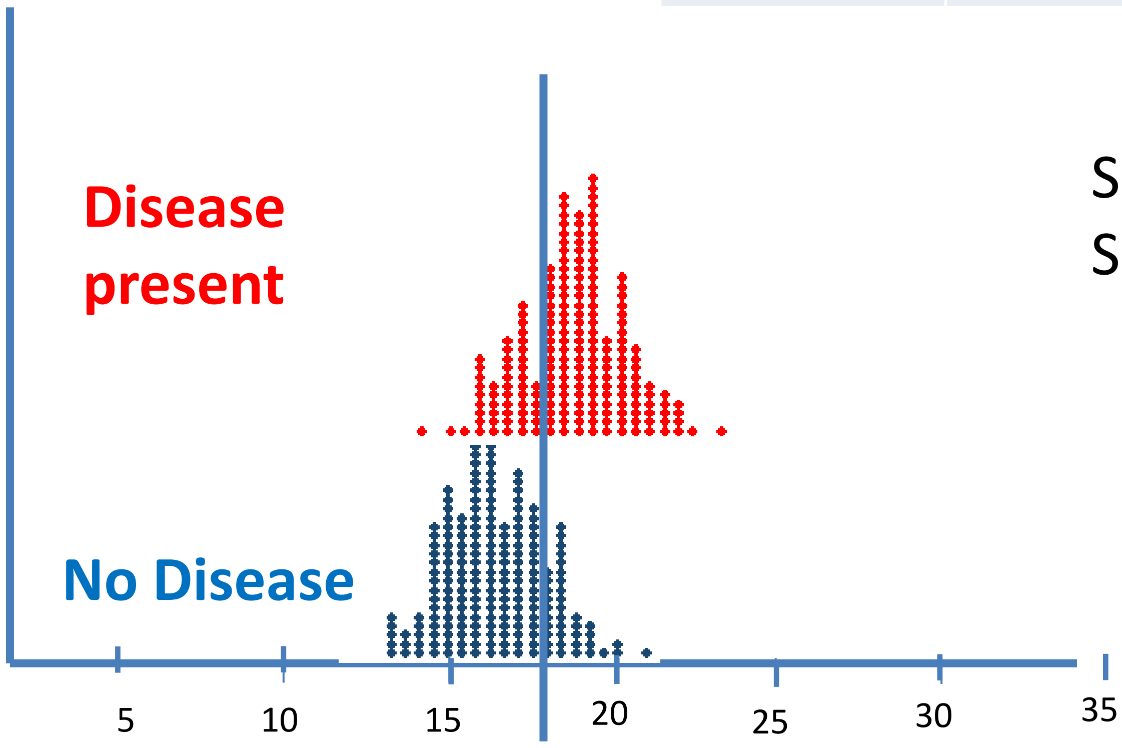
Specificity of PSA Test



Sensitivity and Specificity Depend on Cutoff Values

	Cutoff A	Cutoff B
Sensitivity	74%	22%
Specificity	80%	98%

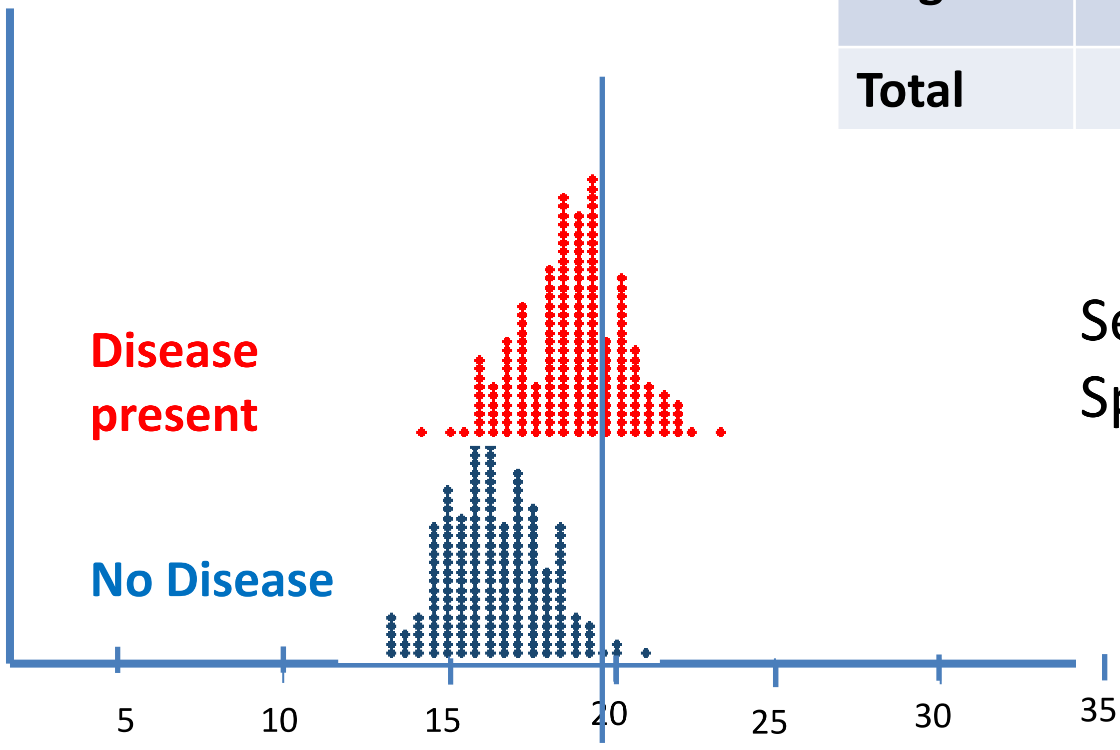
	Disease		
Test	**Present**	**Absent**	**Total**
Positive	148	40	188
Negative	52	160	212
Total	200	200	



Sensitivity = 74%
Specificity = 80%

Threshold Effects on Test Performance

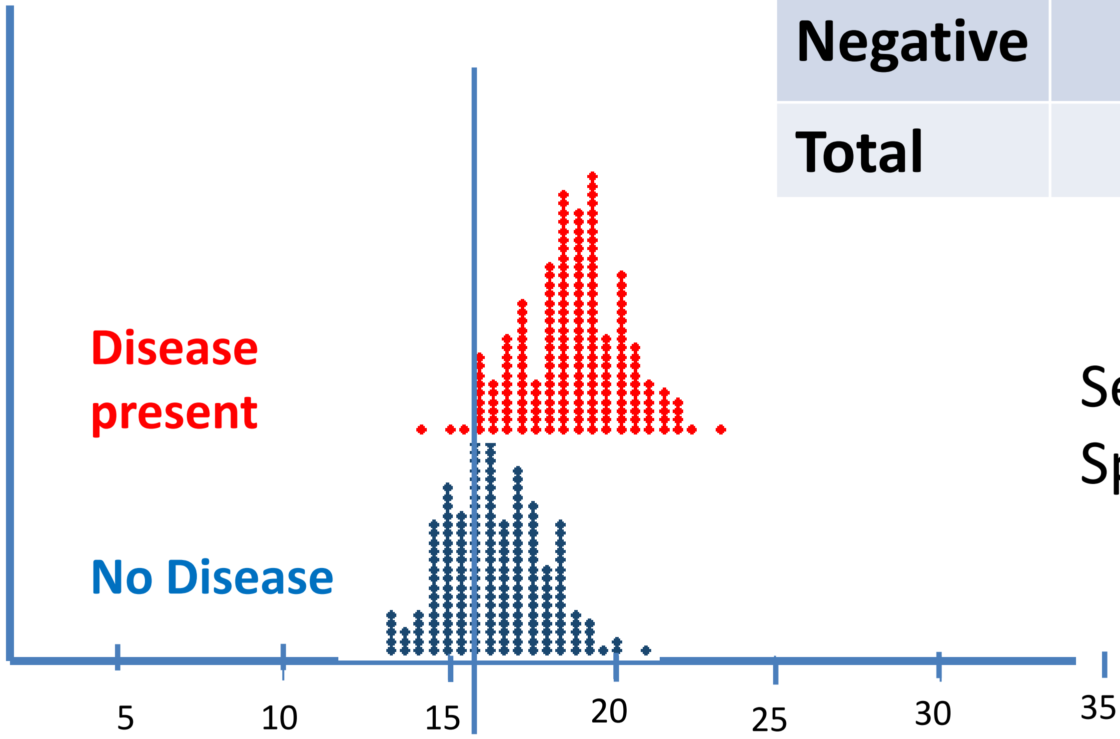
	Disease		
Test	Present	Absent	Total
Positive	44	4	88
Negative	156	196	312
Total	200	200	



Sensitivity = 22%
Specificity = 98%

Threshold Effects on Test Performance

	Disease		
Test	**Present**	**Absent**	**Total**
Positive	196	156	312
Negative	4	44	88
Total	200	200	



Sensitivity = 98%
Specificity = 22%

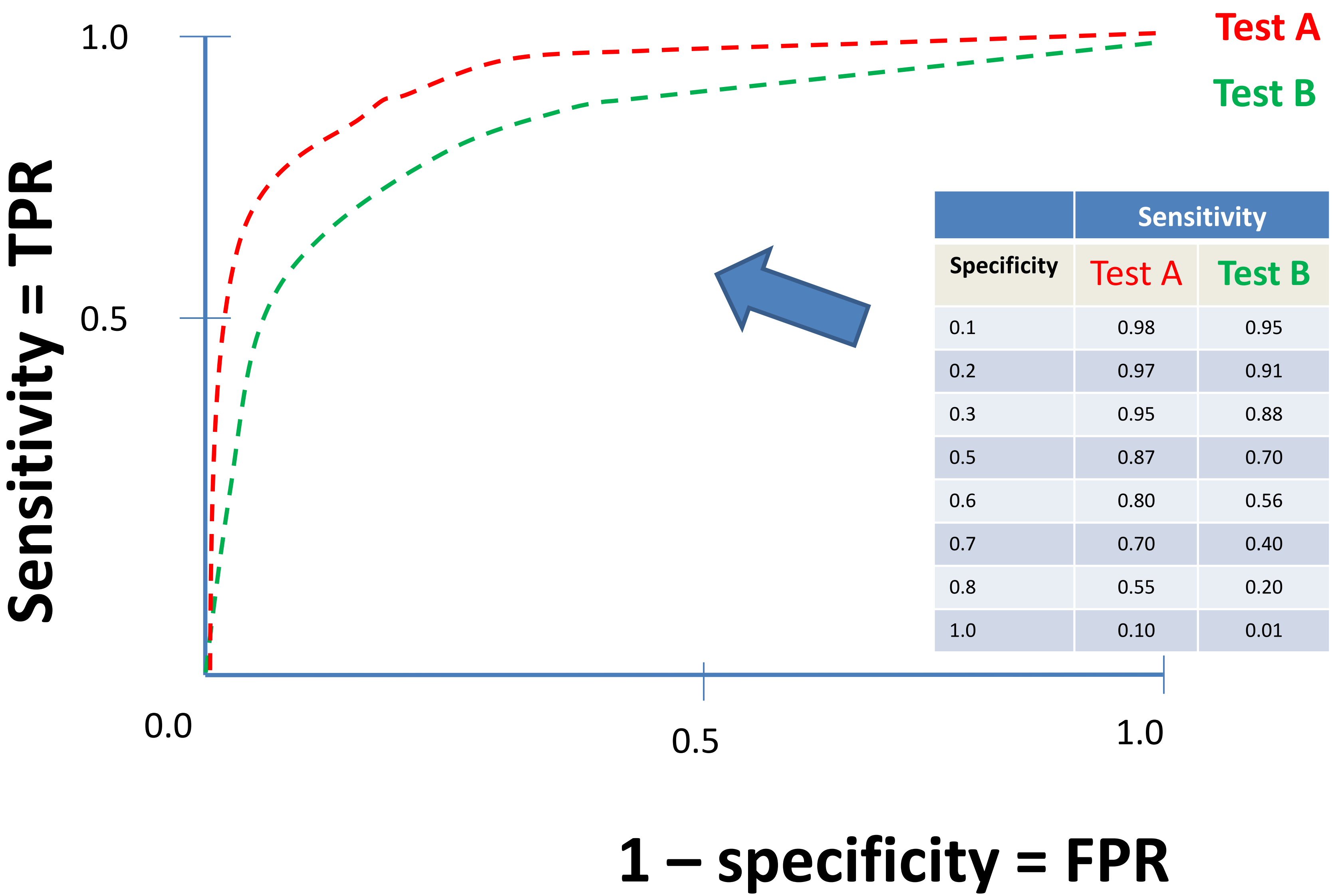
Tradeoff: Specificity vs Sensitivity

Threshold	Sensitivity	Specificity
15	98	22
18	74	80
20	22	98

How to Compare Tests

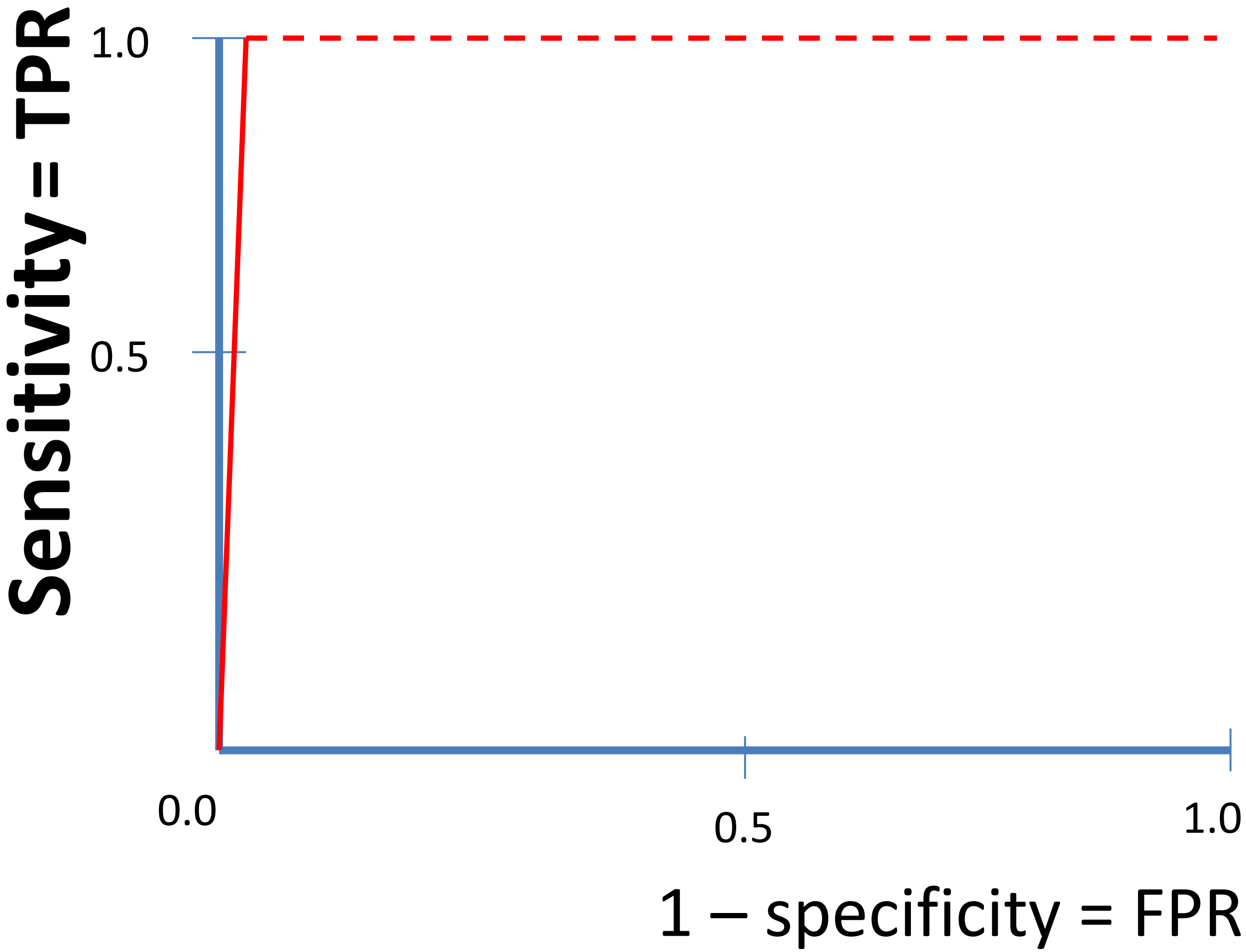
	Sensitivity	
Specificity	**Test A**	**Test B**
0.1	0.98	0.95
0.2	0.97	0.91
0.3	0.95	0.88
0.4	0.91	0.78
0.5	0.87	0.70
0.6	0.80	0.56
0.7	0.70	0.40
0.8	0.55	0.20
0.9	0.38	0.10
1.0	0.10	0.01

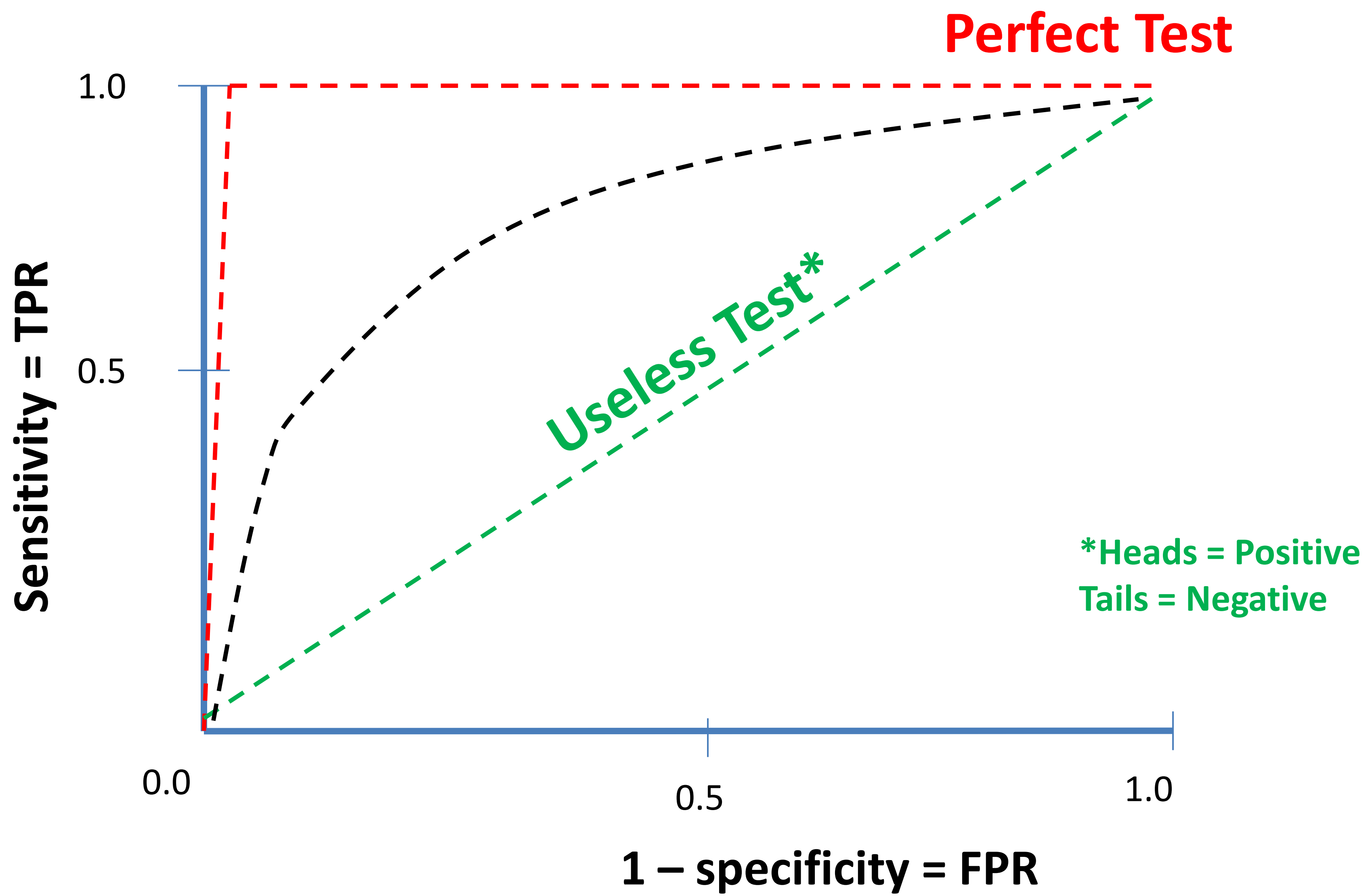
Receiver Operating Characteristic (ROC) Curve



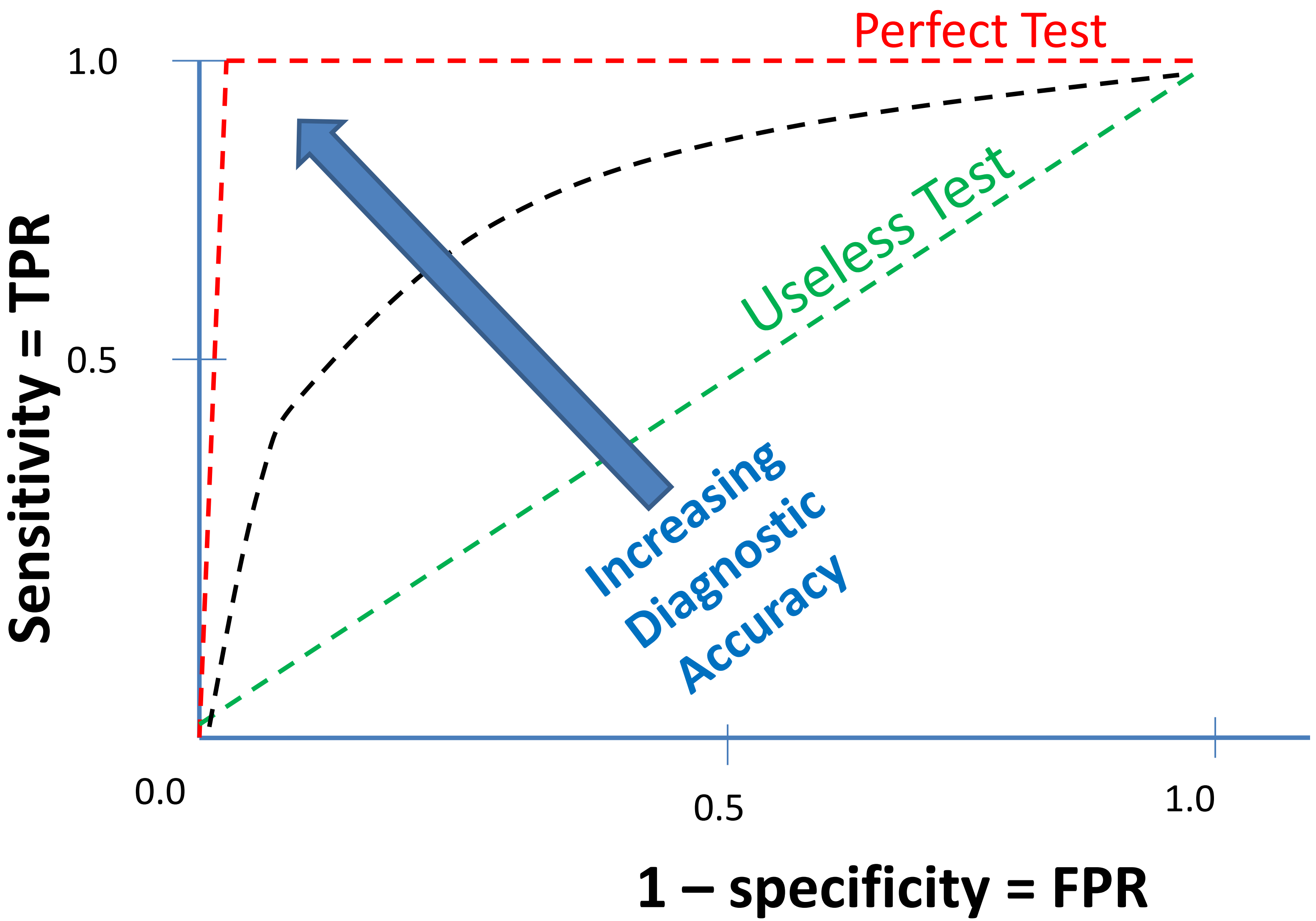
	Sensitivity	
Specificity	Test A	Test B
0.1	0.98	0.95
0.2	0.97	0.91
0.3	0.95	0.88
0.5	0.87	0.70
0.6	0.80	0.56
0.7	0.70	0.40
0.8	0.55	0.20
1.0	0.10	0.01

ROC Curve for The Perfect Test

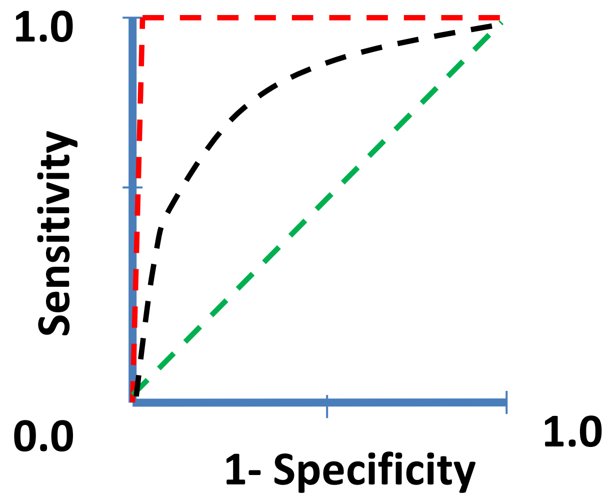


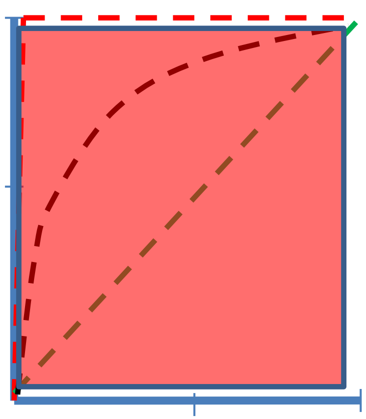

Perfect Test
1.0
0.5
Sensitivity = TPR
Useless Test*
*Heads = Positive
Tails = Negative
0.0
0.5
1.0
1 – specificity = FPR

Test Performance is Related to Area Under the Curve (AUC)

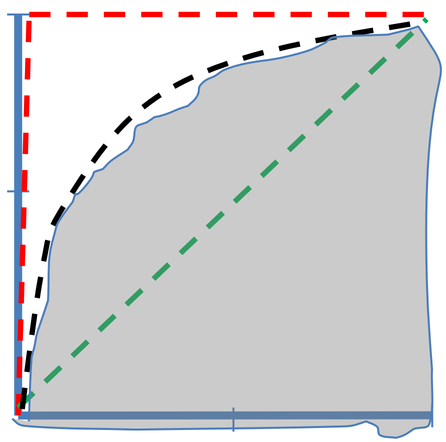


Area Under the Curve (AUC)

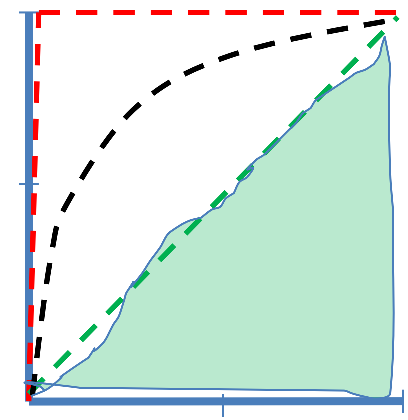




Perfect Test
AUC = 1.0



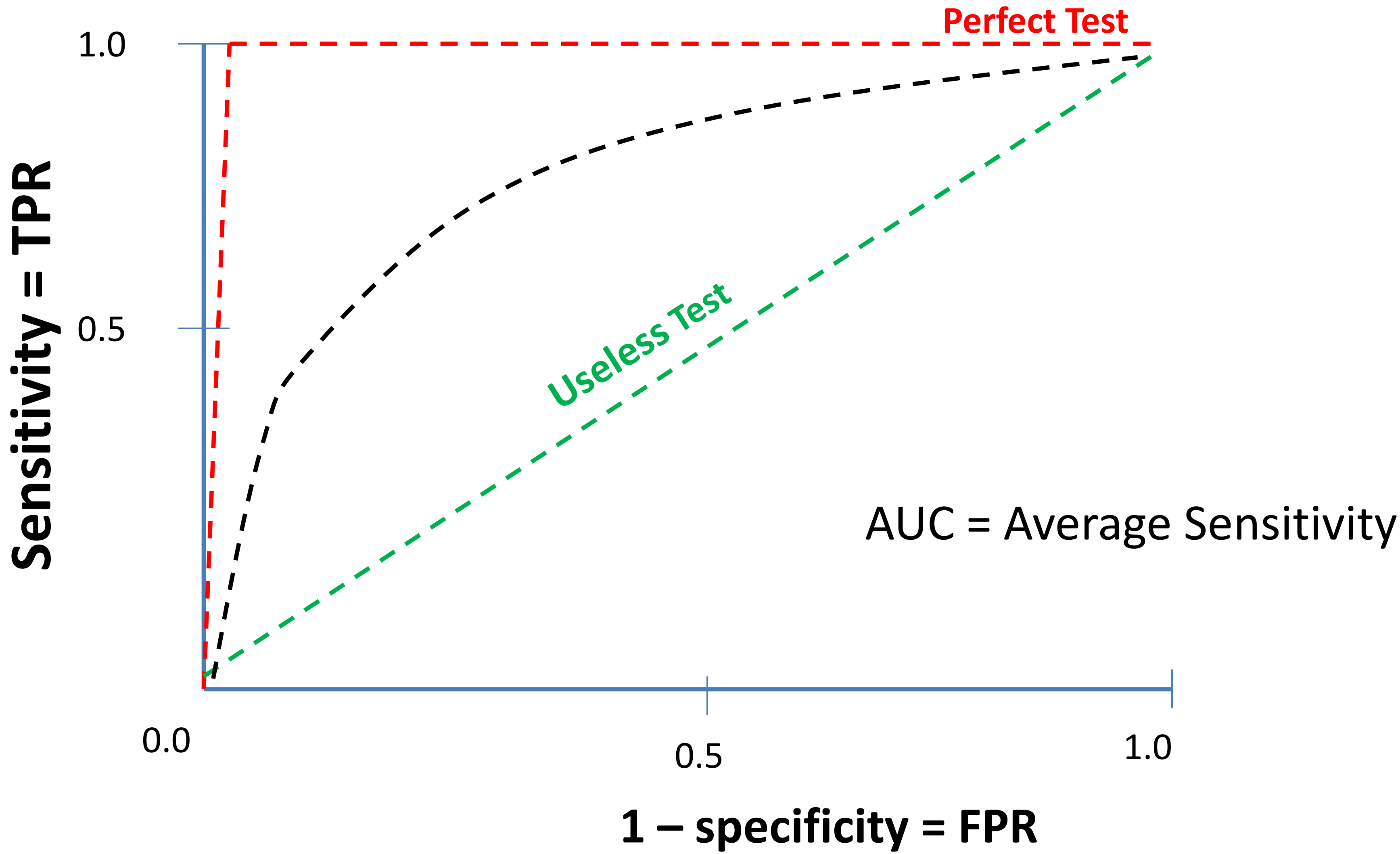
Real Test



Useless Test
AUC = 0.5

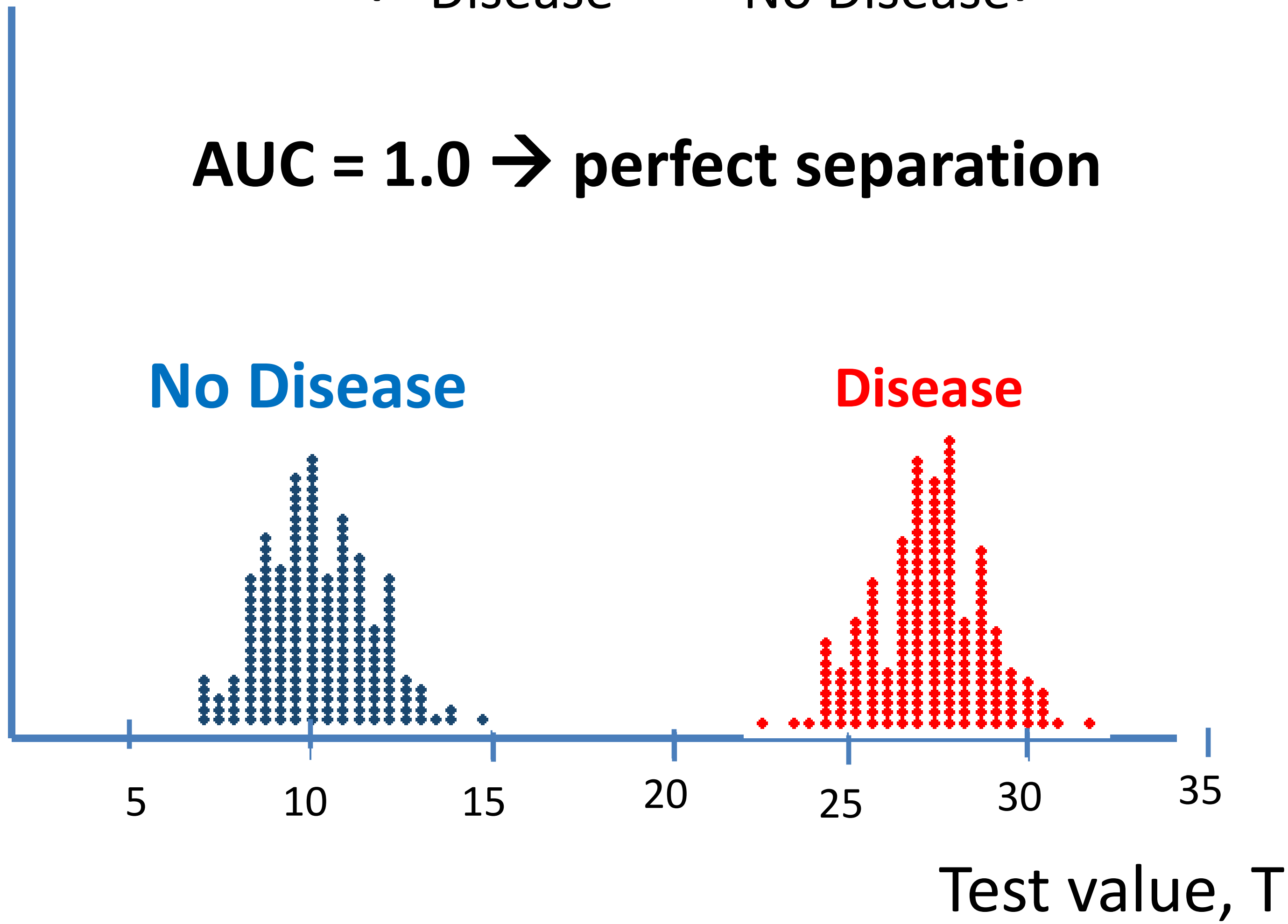
Does the AUC mean anything?

Does the AUC Mean Anything?



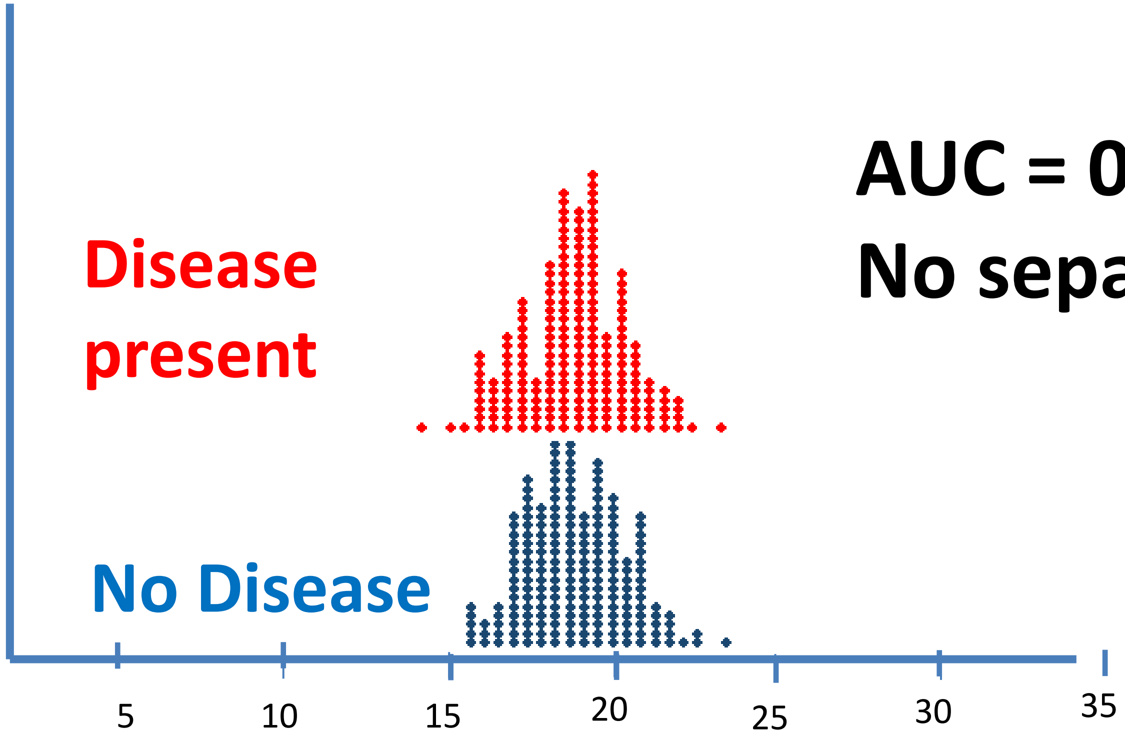
Meaning of AUC

Prob ($T_{\text{Disease}} > T_{\text{No Disease}}$)



Meaning of AUC

$$\text{Prob}\ (T_{\text{Disease}} > T_{\text{No Disease}})$$



AUC = 0.50

No separation

PSA vs PSA velocity

	Area Under the Curve (AUC)	
Study	**PSA velocity**	**PSA**
Eggener, 2005	0.91	0.88
Ciatto, 2004	0.74	0.67
Berger, 2007	0.87	0.65

Take home message:

Sensitivity
Specificity

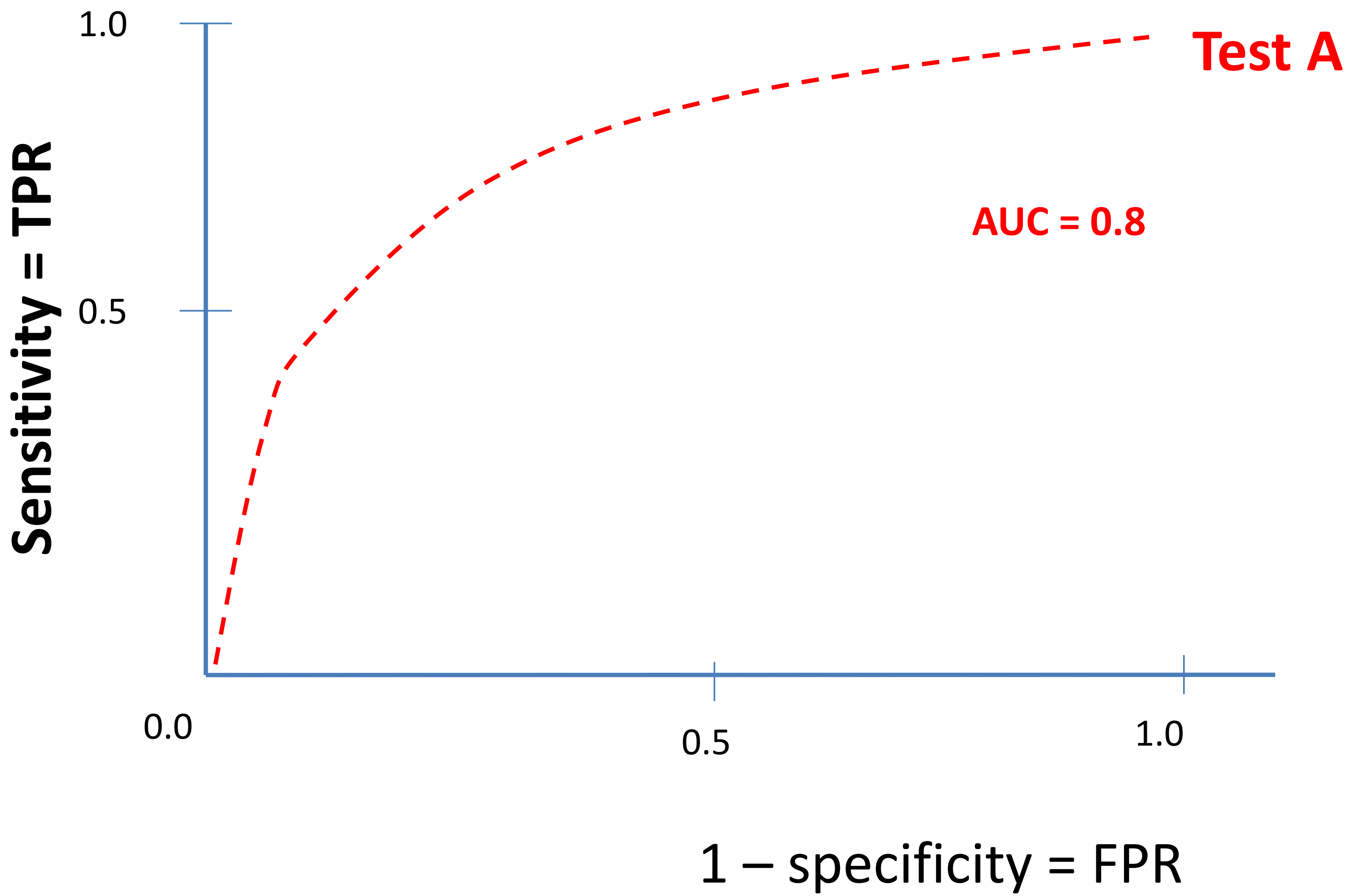
ROC Curves
AUC

Threshold Effects

No Threshold Effects

How do I know if a test is useful?

Is this test useful?



Usefulness is defined by the customer

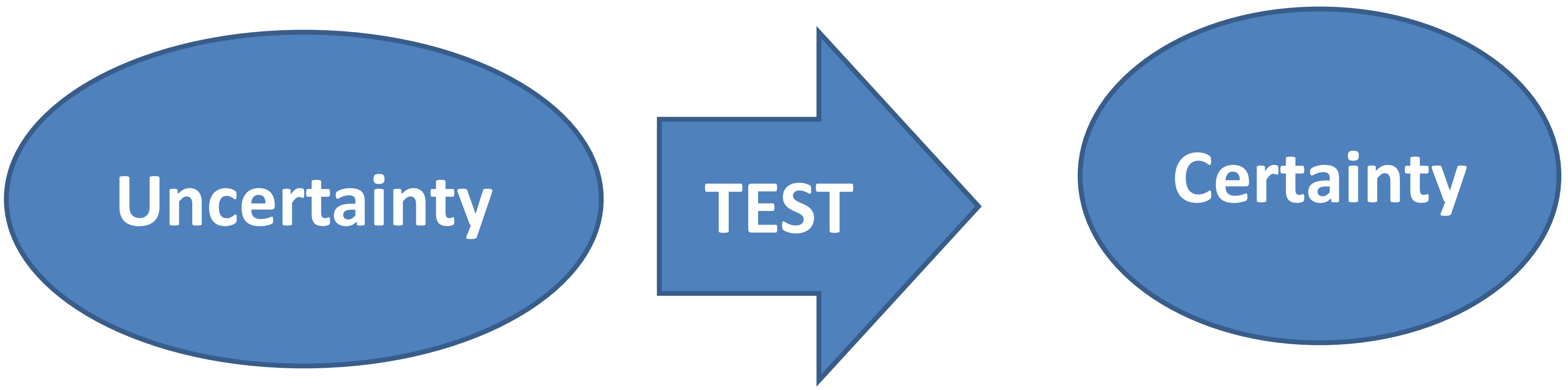
Will this test tell
me whether my
patient has
Prostate Cancer?

Usefulness is defined by the customer

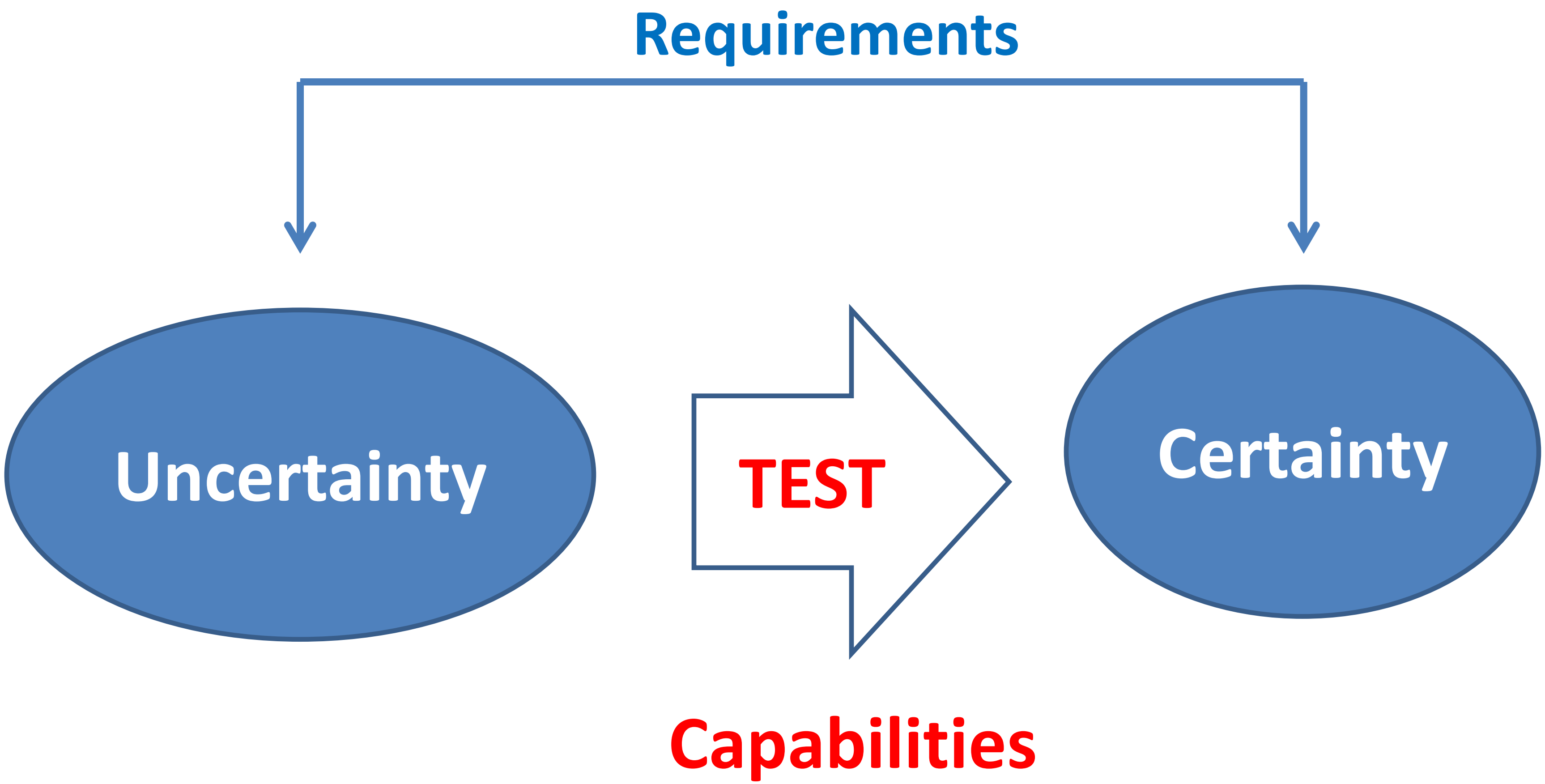
The AUC is 0.75!

Huh?

The customer's problem:

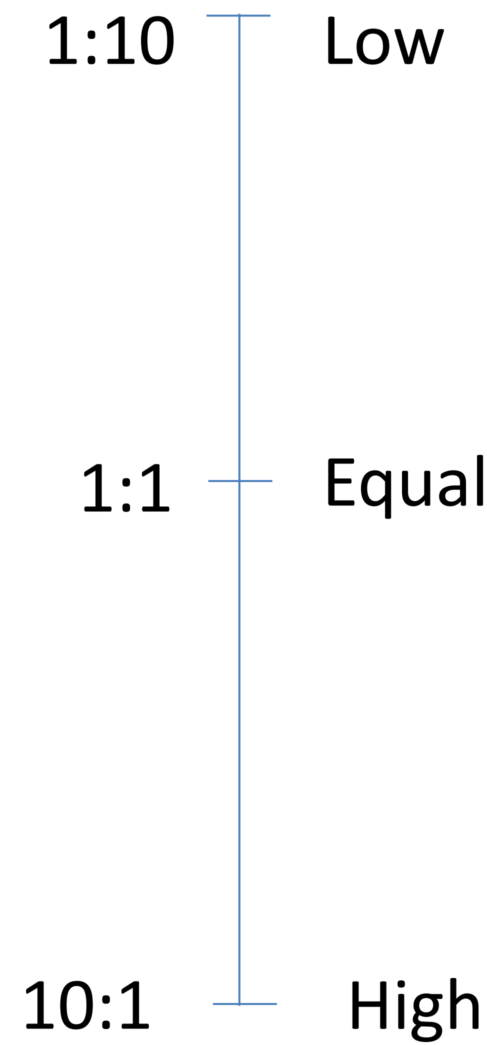


Usefulness = Capabilities - Requirements



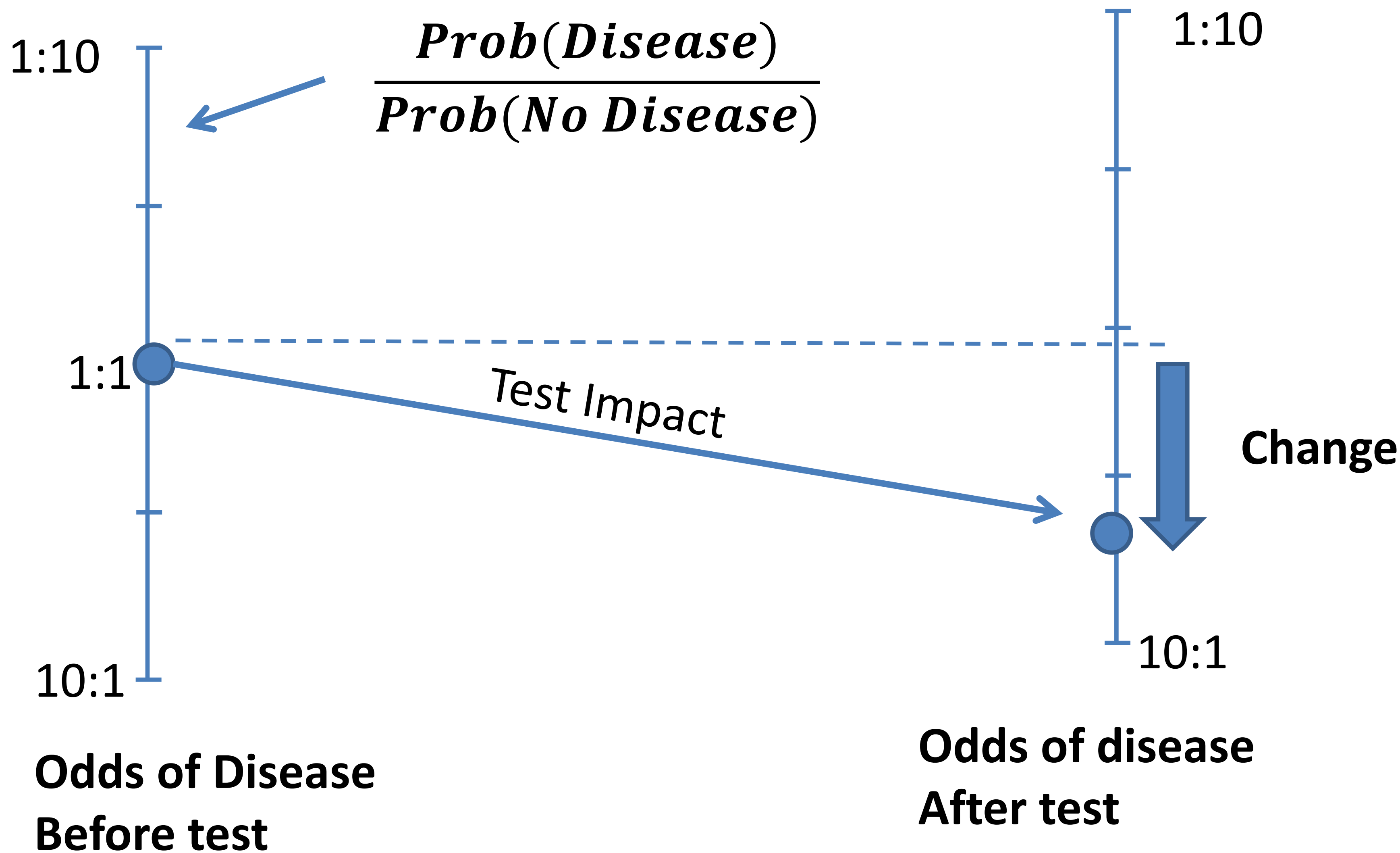
How to think about certainty

The Odds Scale



$$\textbf{Odds} = \frac{Prob(Disease)}{Prob(No\ Disease)} = \frac{P}{1-P}$$

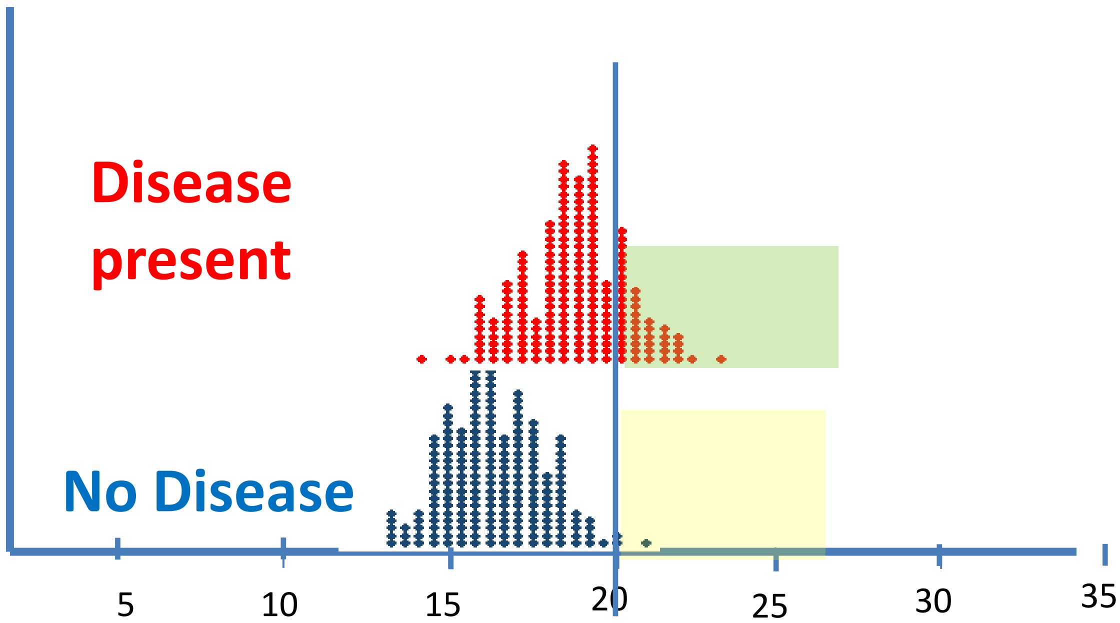
Capability = Change in Certainty



What is the impact of a ***positive*** result?

Positive Likelihood ratio, LR+

$$LR+ = \frac{Prob\,(Positive|disease)}{Prob(Positive|no\ disease)} = \frac{40/_{200}}{4/_{200}} = 10.0$$

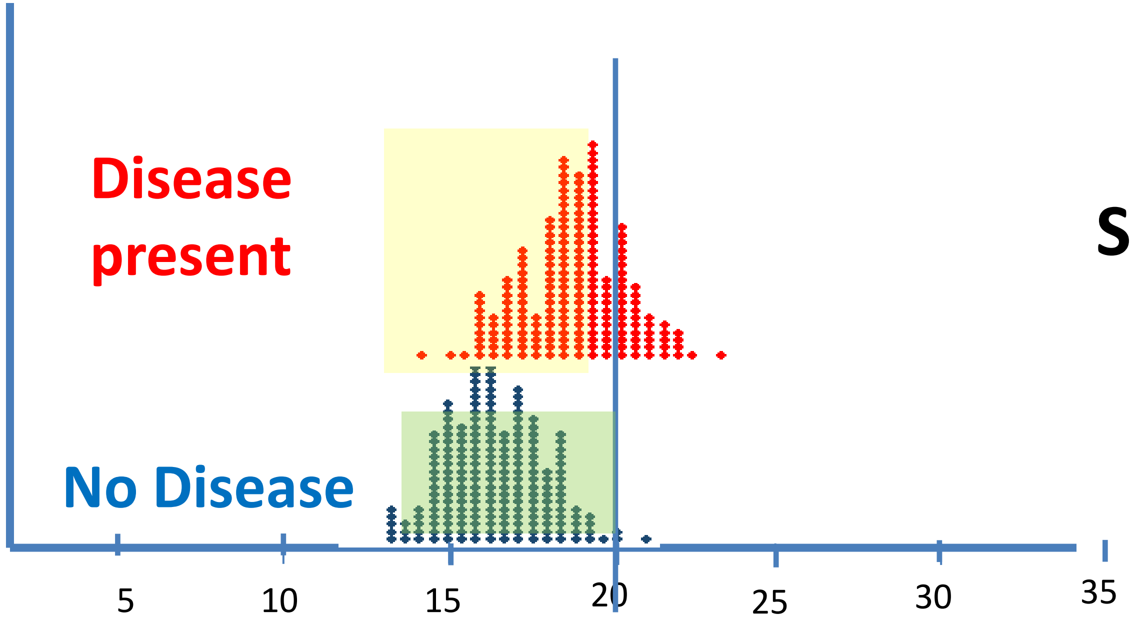


Bigger LR+ is better

What is the impact of a ***negative*** result?

Negative Likelihood ratio, LR-

$$\text{LR-}= \frac{\boldsymbol{Prob\ (Negative\ |disease)}}{\boldsymbol{Prob(Negative|no\ disease)}} = \frac{160/_{200}}{196/_{200}} = \mathbf{0.8}$$



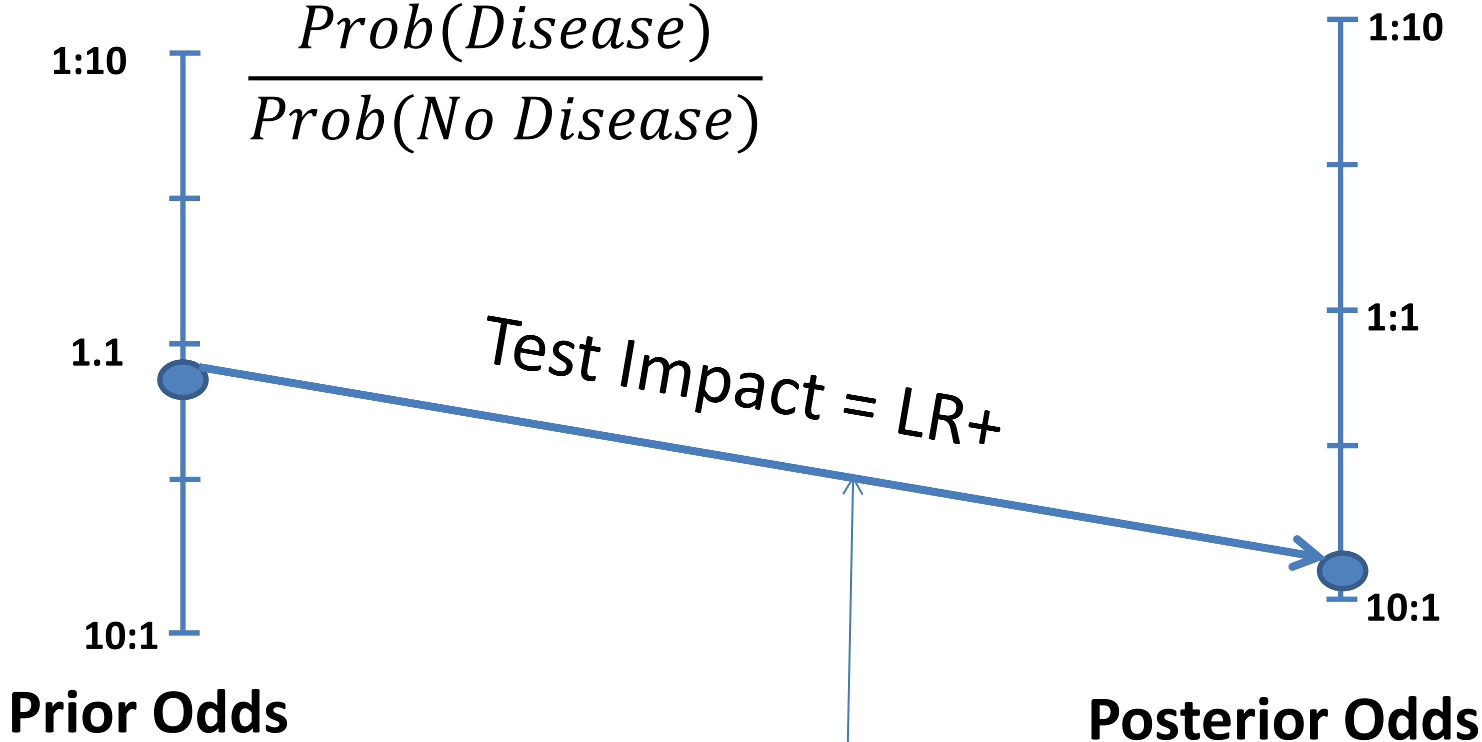
Smaller LR- is Better

Key Relationship

$$\text{Posterior Odds} = LR \times \text{Prior Odds}$$

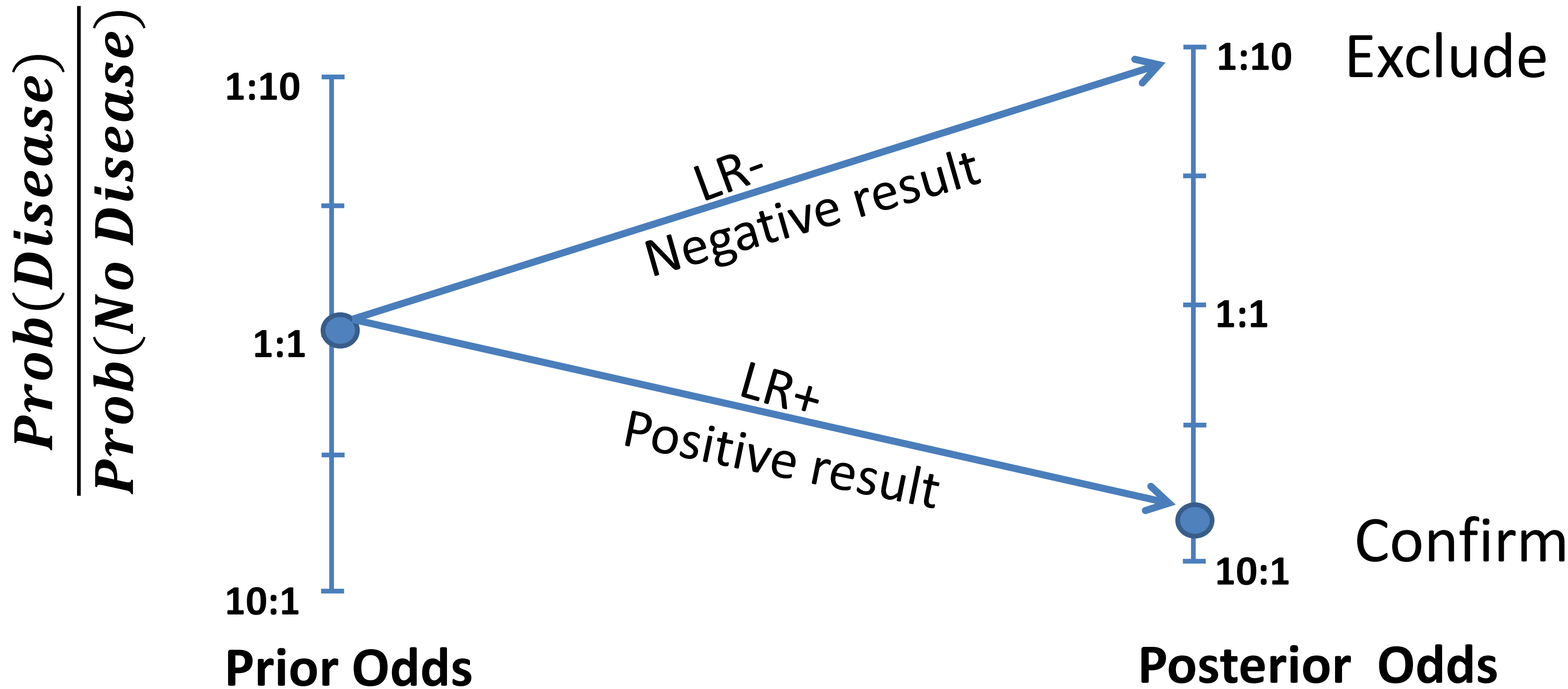
Likelihood = Impact Factor

$$\frac{Prob(Disease)}{Prob(No\ Disease)}$$



$$\boldsymbol{Prior\ odds\ x\ LR = \ \ Posterior\ odds}$$

Two ways to be certain:



$$Prior\ odds\ x\ Likelihood\ Ratio = \ Posterior\ odds$$

A test can solve the problem if:

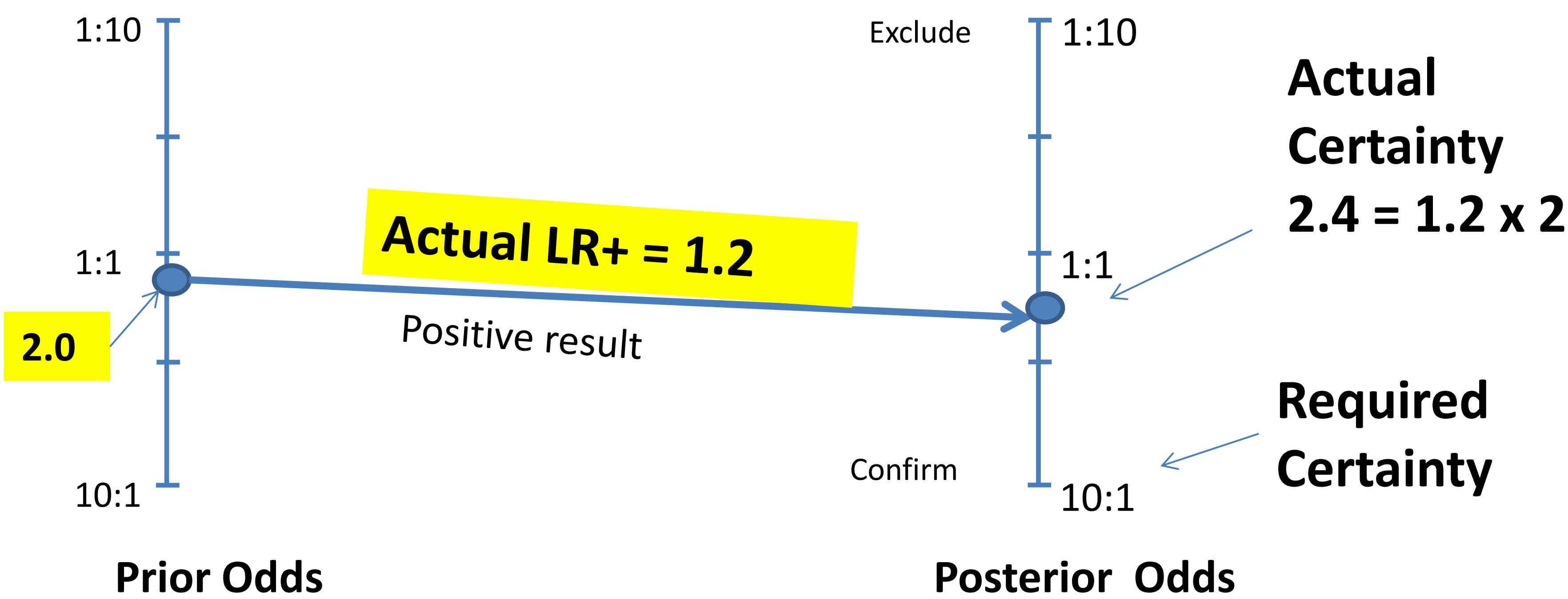
$$\mathrm{LR}_{\mathrm{Actual}} > \mathrm{LR}_{\mathrm{Required}}$$

$$LR^{+} > \frac{PosteriorOdds}{PriorOdds}$$ Confirm

Or:

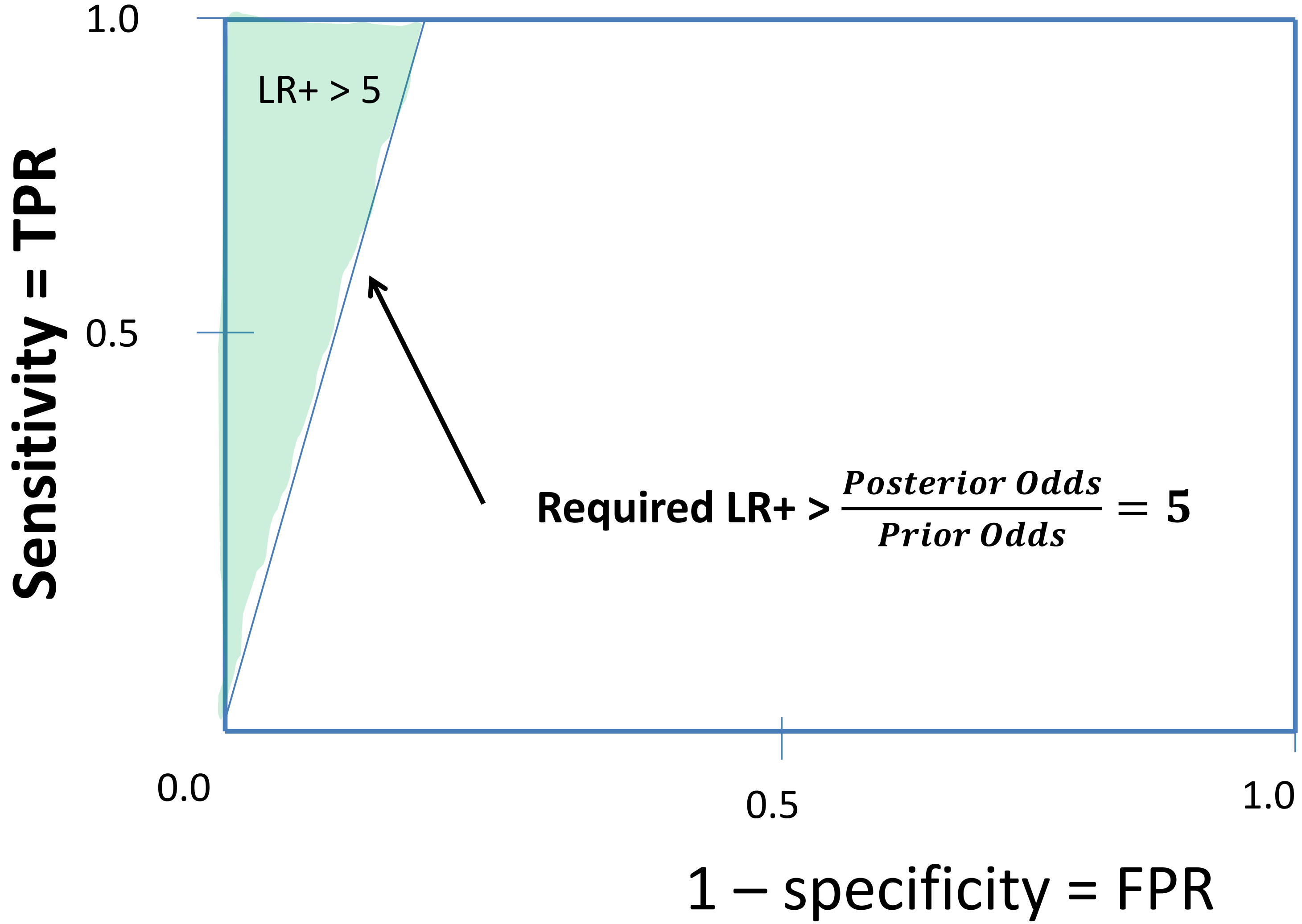
$$LR^{-} < \frac{PosteriorOdds}{PriorOdds}$$ Exclude

Futile test: *Required* LR+ $= \frac{PostOdds}{PriorOdds} = \frac{10}{2} = 5$



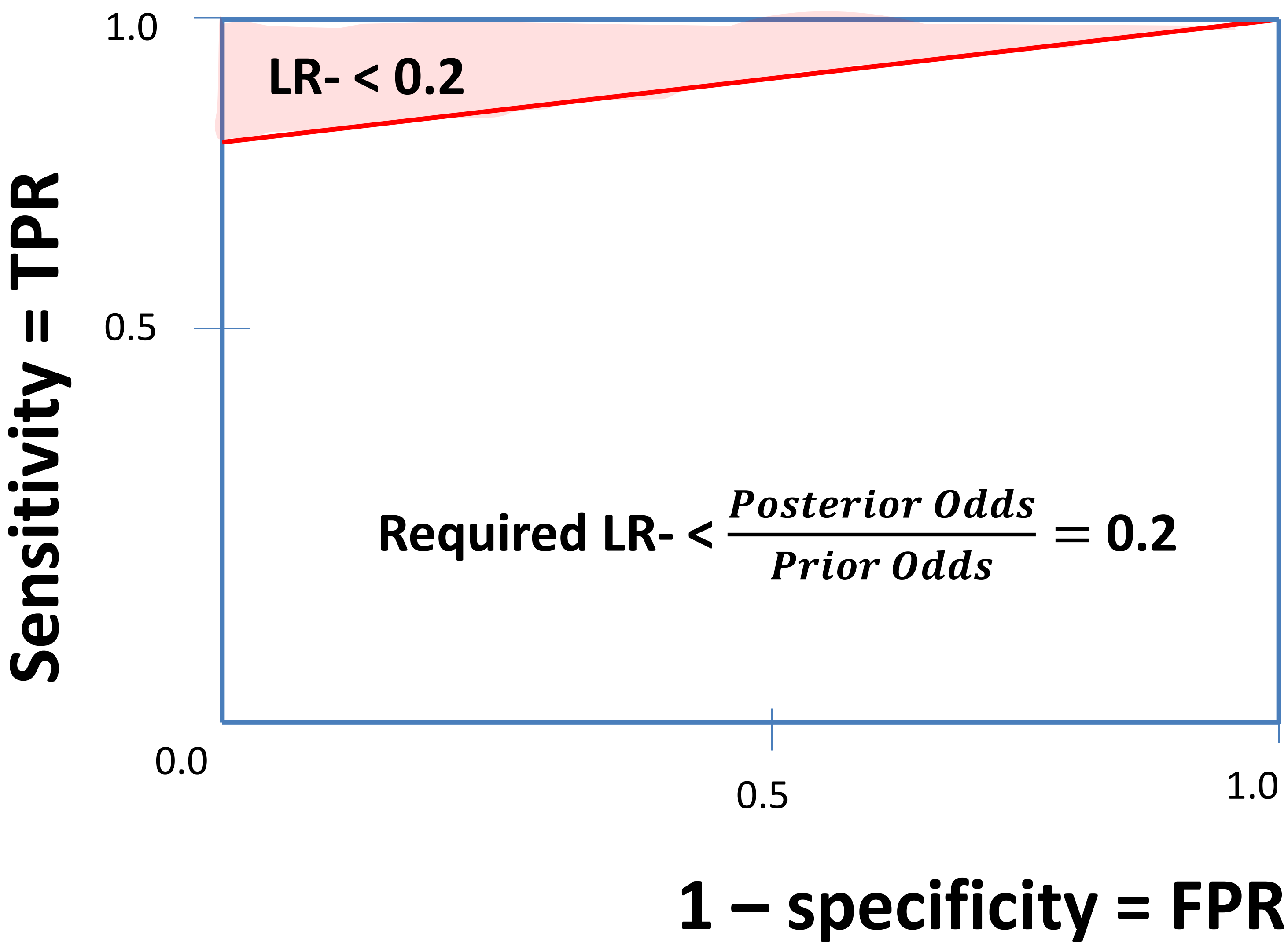
I have a slight hunch that this person may have Prostate CA. I will do a biopsy if the odds in favor of disease are around 10:1

Rule-in, Confirmation Zone
High LR+ is better



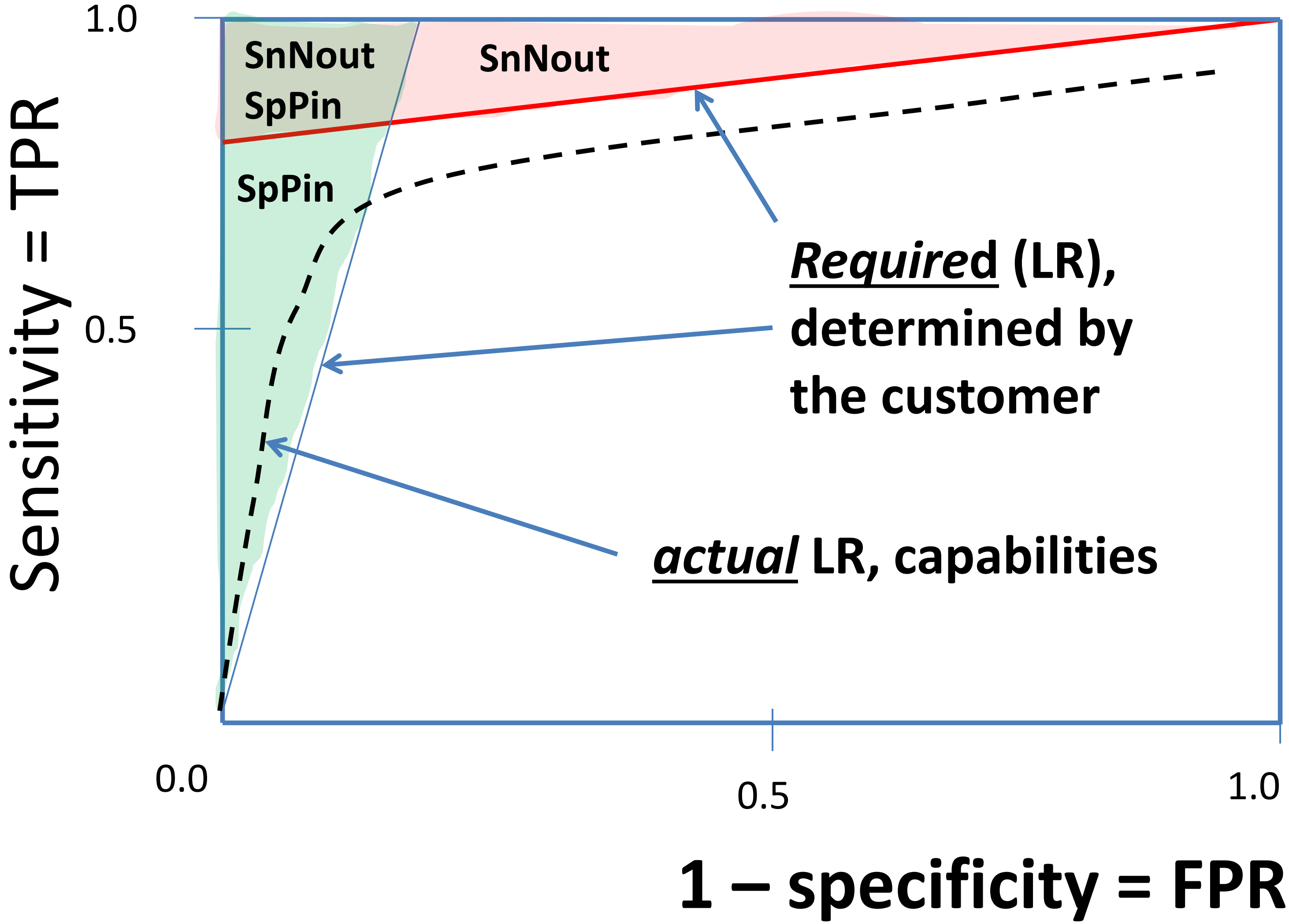
Rule-out, Exclusion Zone:

Low LR- is better



Required LR- < $\frac{\textit{Posterior Odds}}{\textit{Prior Odds}} = 0.2$

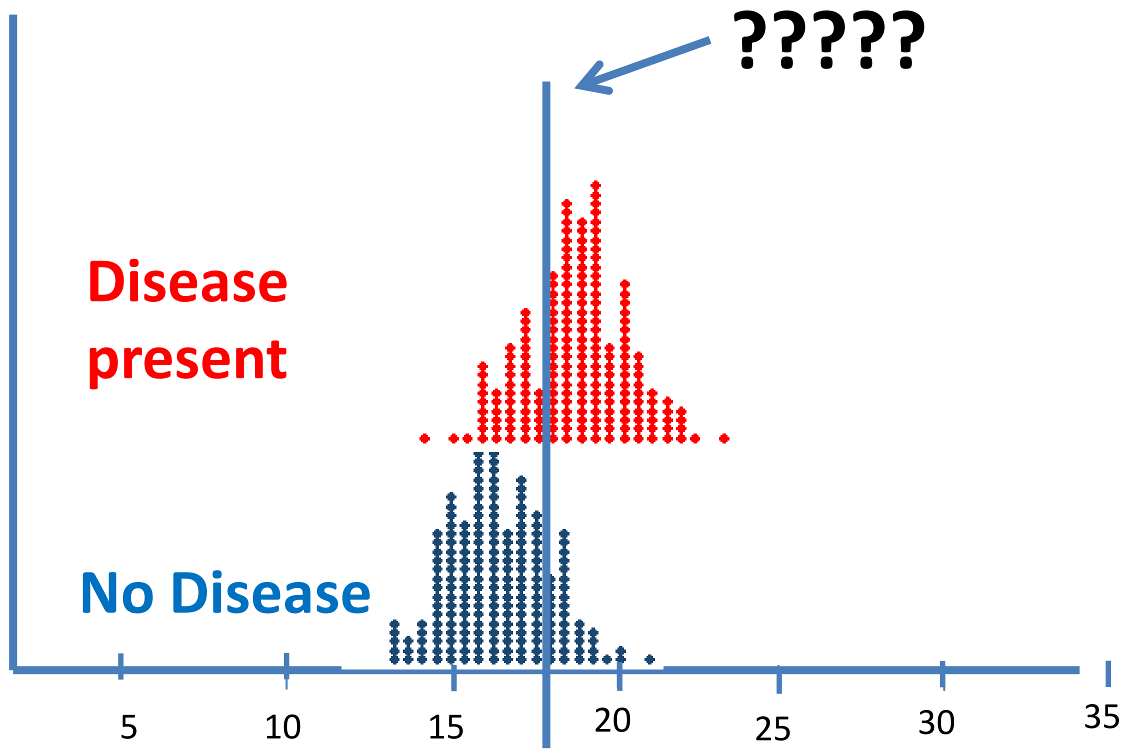
Capabilities > Requirements?



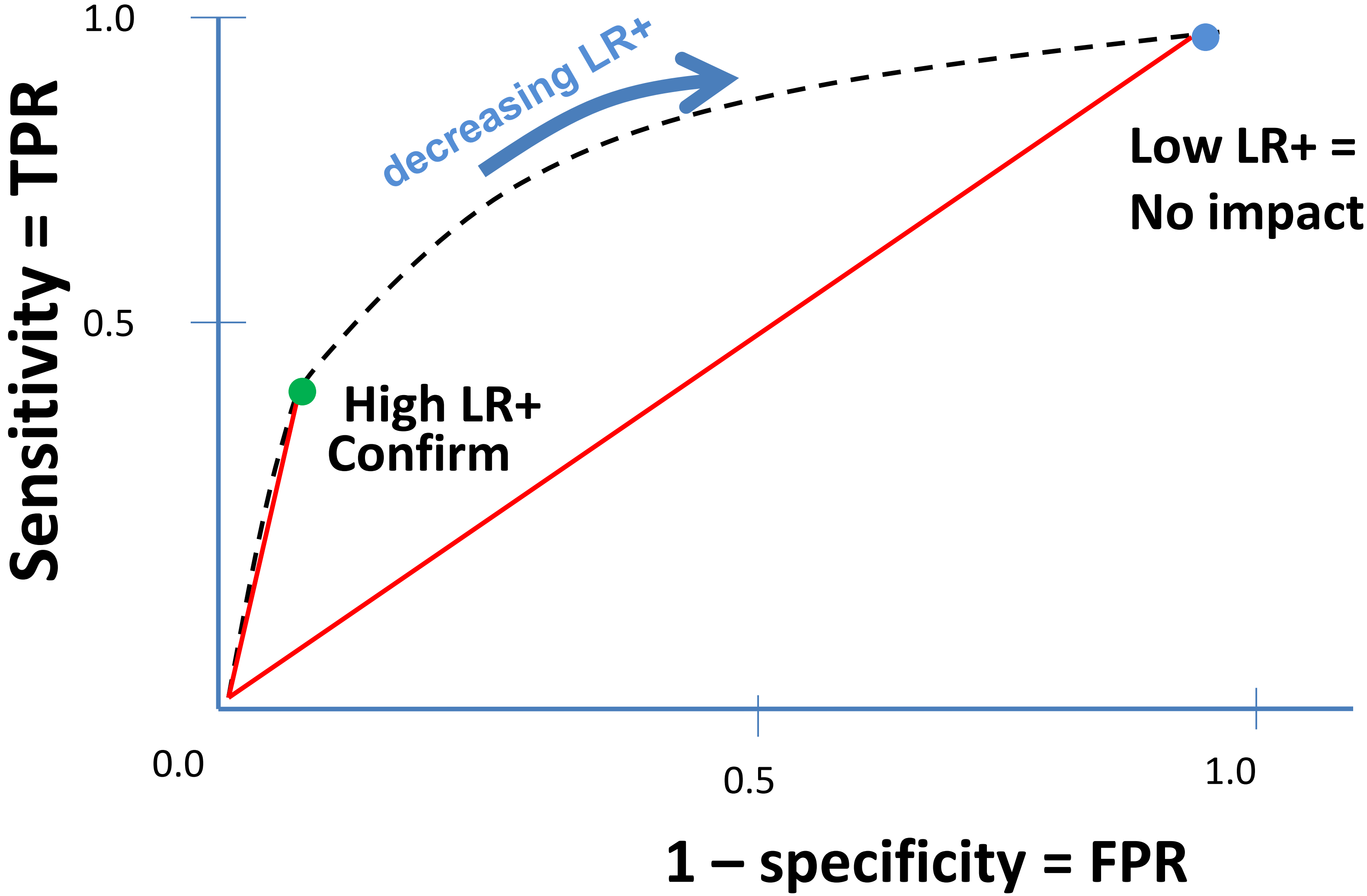
Key Points:

- **Accuracy ≠ Usefulness**
- **Potential Usefulness = LR = f(Sn,Sp)**
- **Usefulness = Capabilities - Requirements:**
  - The objective (exclude, confirm)
  - Prior uncertainty
  - Required certainty
  - The Test Impact (*actual* LR vs *required* LR)

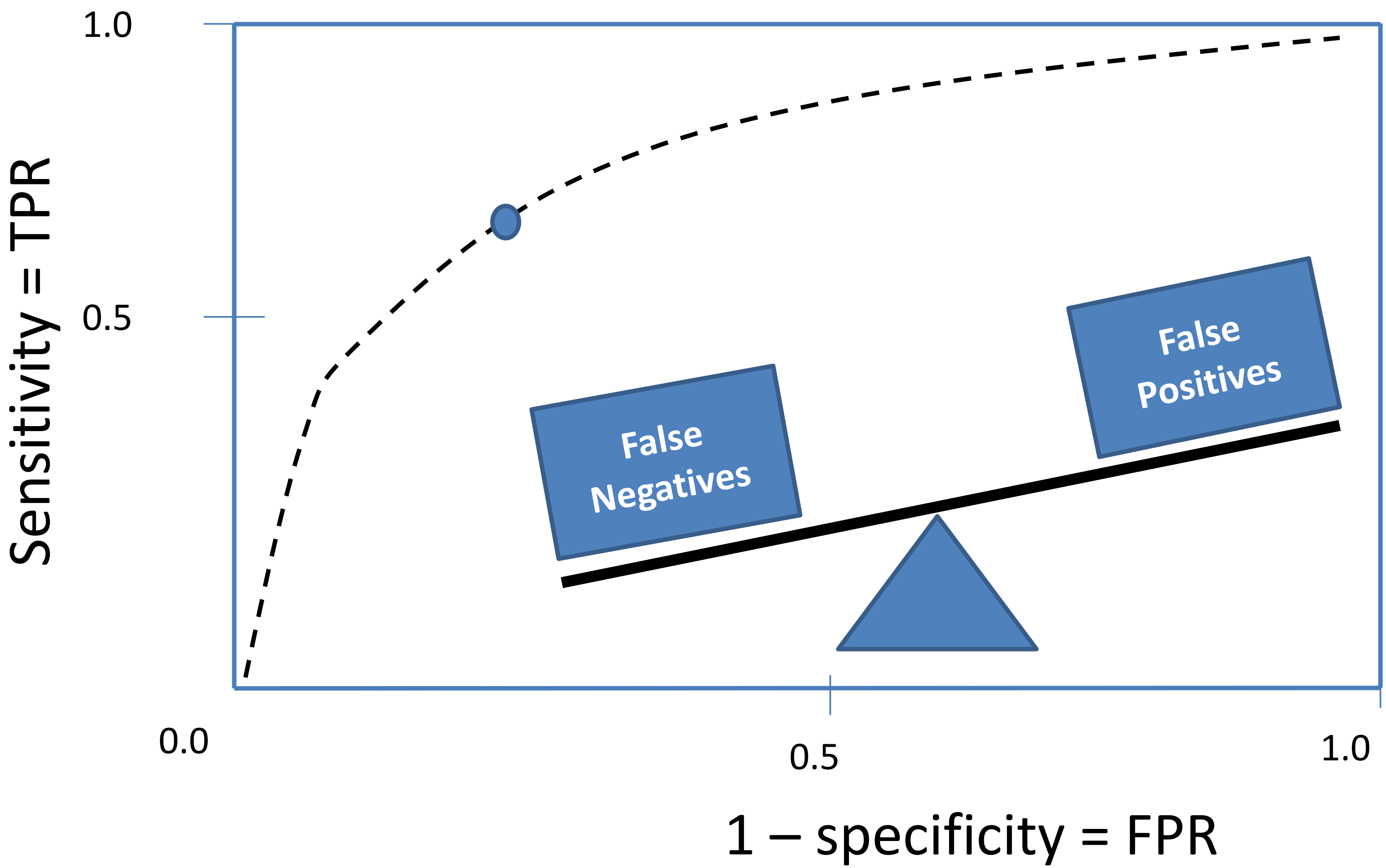
Setting Test Thresholds



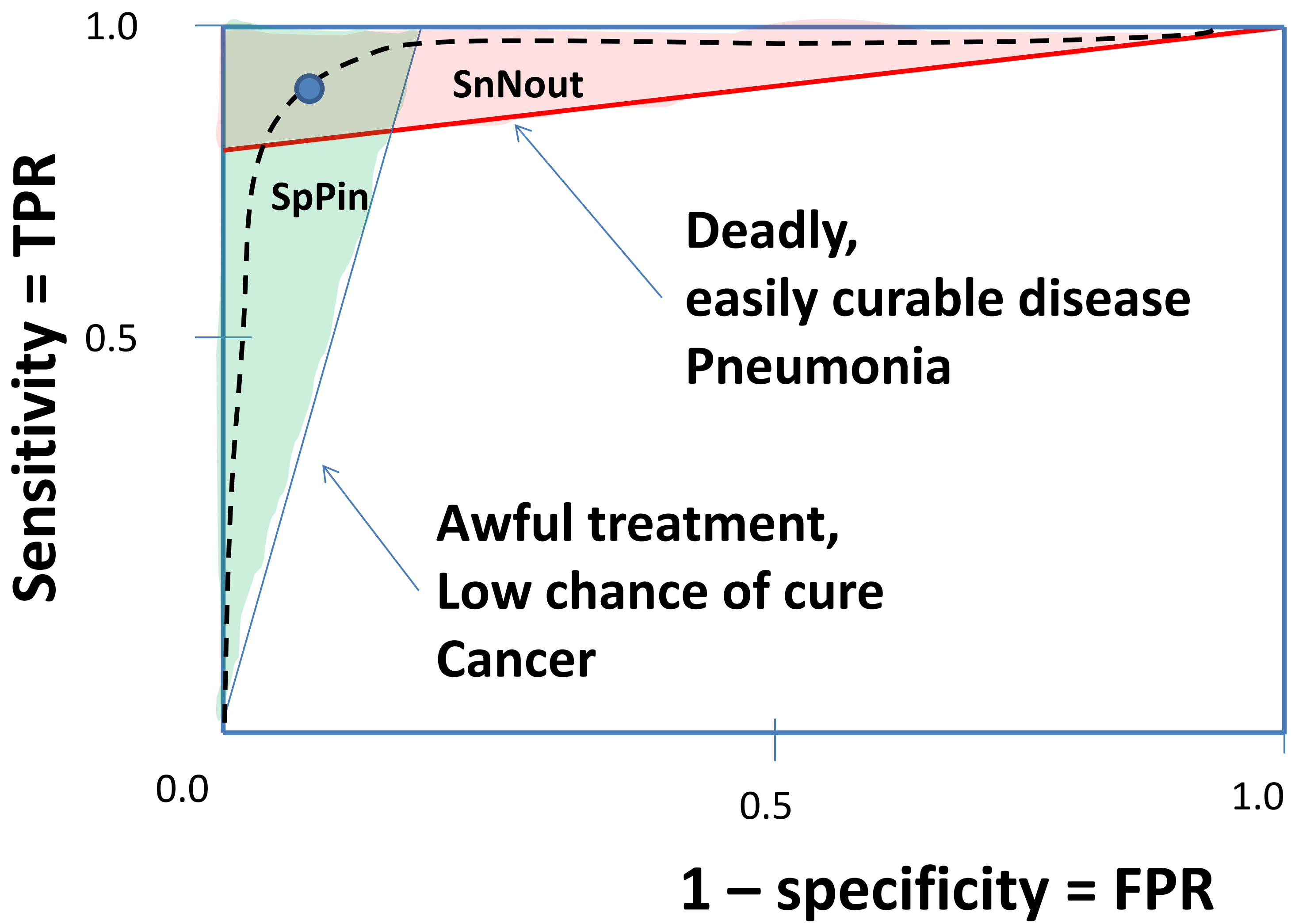
ROC curve = set of available Likelihood Ratios



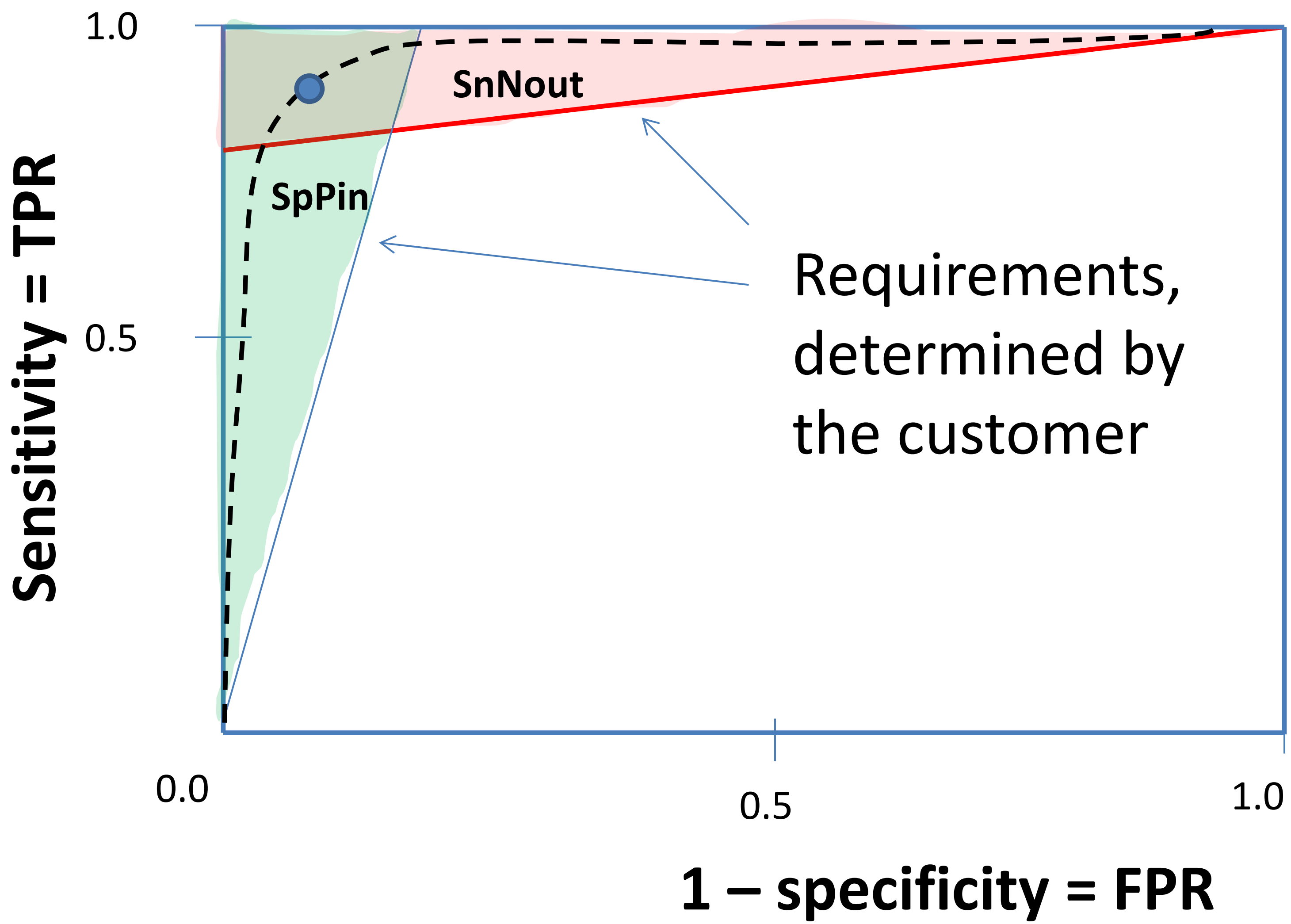
How to select a threshold value



Diagnostic Zones for Thresholds



Accuracy , Usefulness and Optimality



Key Points – Setting Thresholds

- **Comparing Tests**
  - Thresholds are a nuisance
  - ROC/AUC facilitates comparisons of *diagnostic* accuracy

- **Using Tests**
  - Thresholds are required
    - Define a test
    - Link capabilities and requirements
    - Can be set to optimize performance
      - Optimum is context dependent
      - Depends on error costs

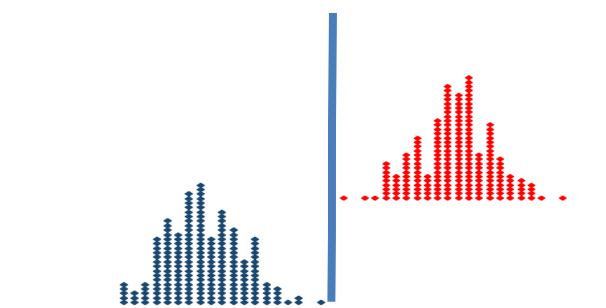
Comparing Test Performance

Why do test results differ?

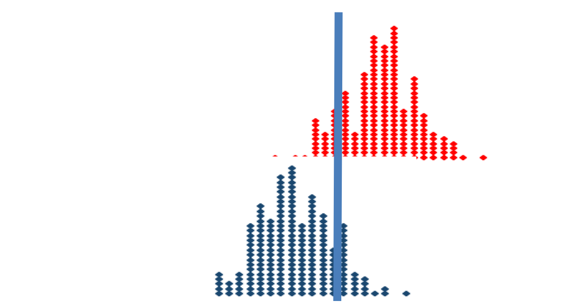
1. **True differences**
2. False differences
   - (bias)
   - thresholds
3. Random variation (imprecision)

Evaluating Test Accuracy: Ideal vs Realistic Conditions

Consecutive Patients → Index Test → Gold Standard



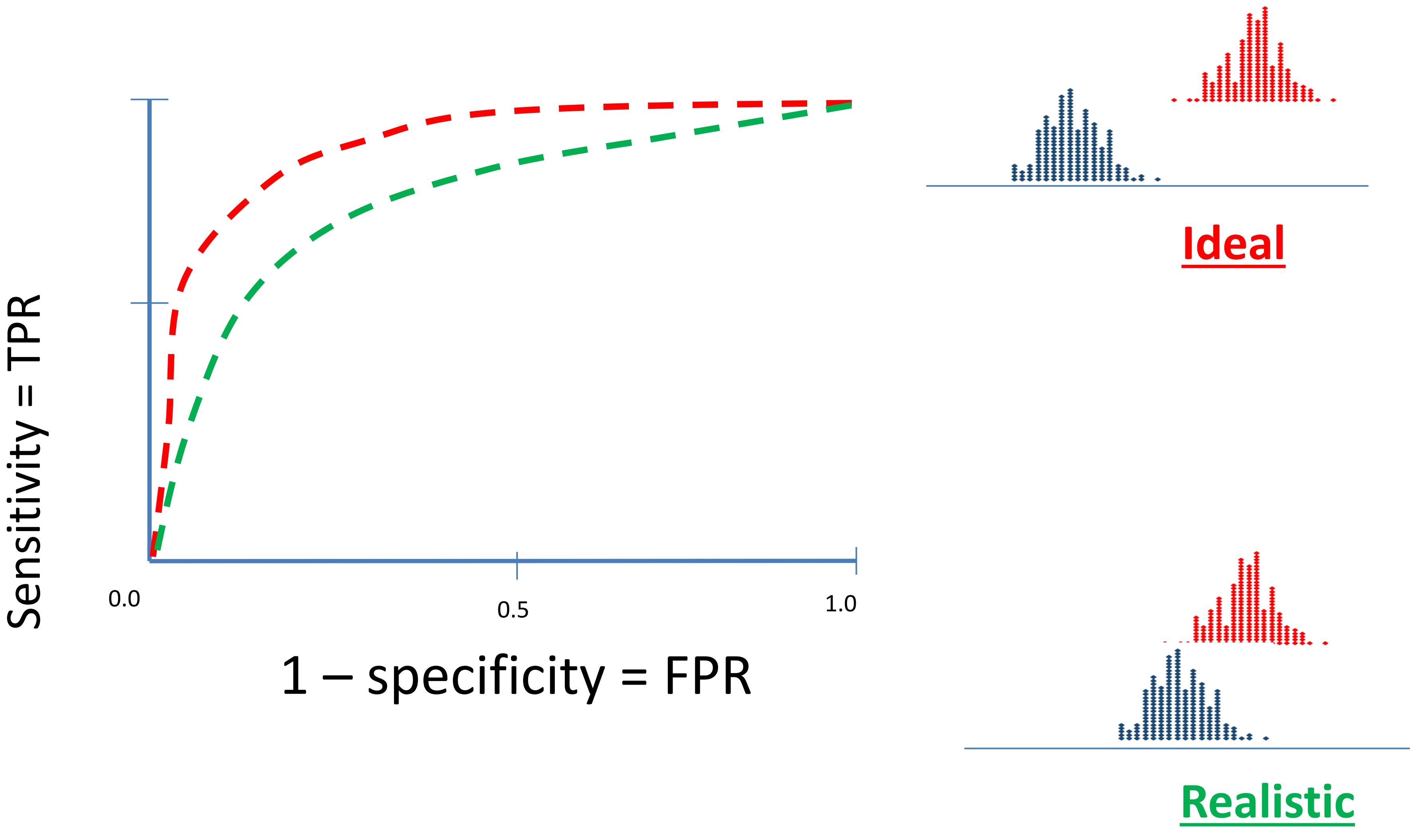
Ideal



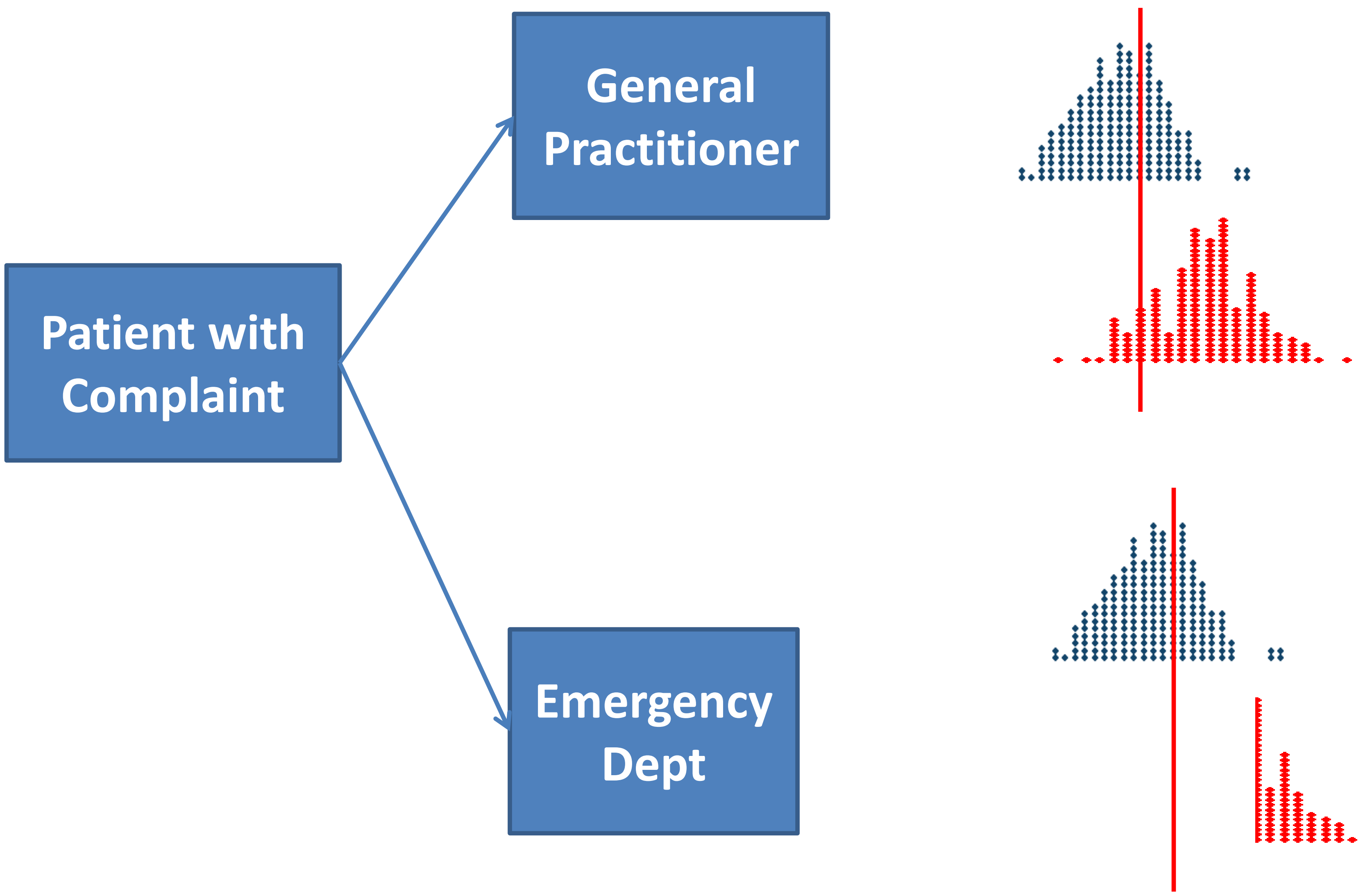
Realistic

Test Performance

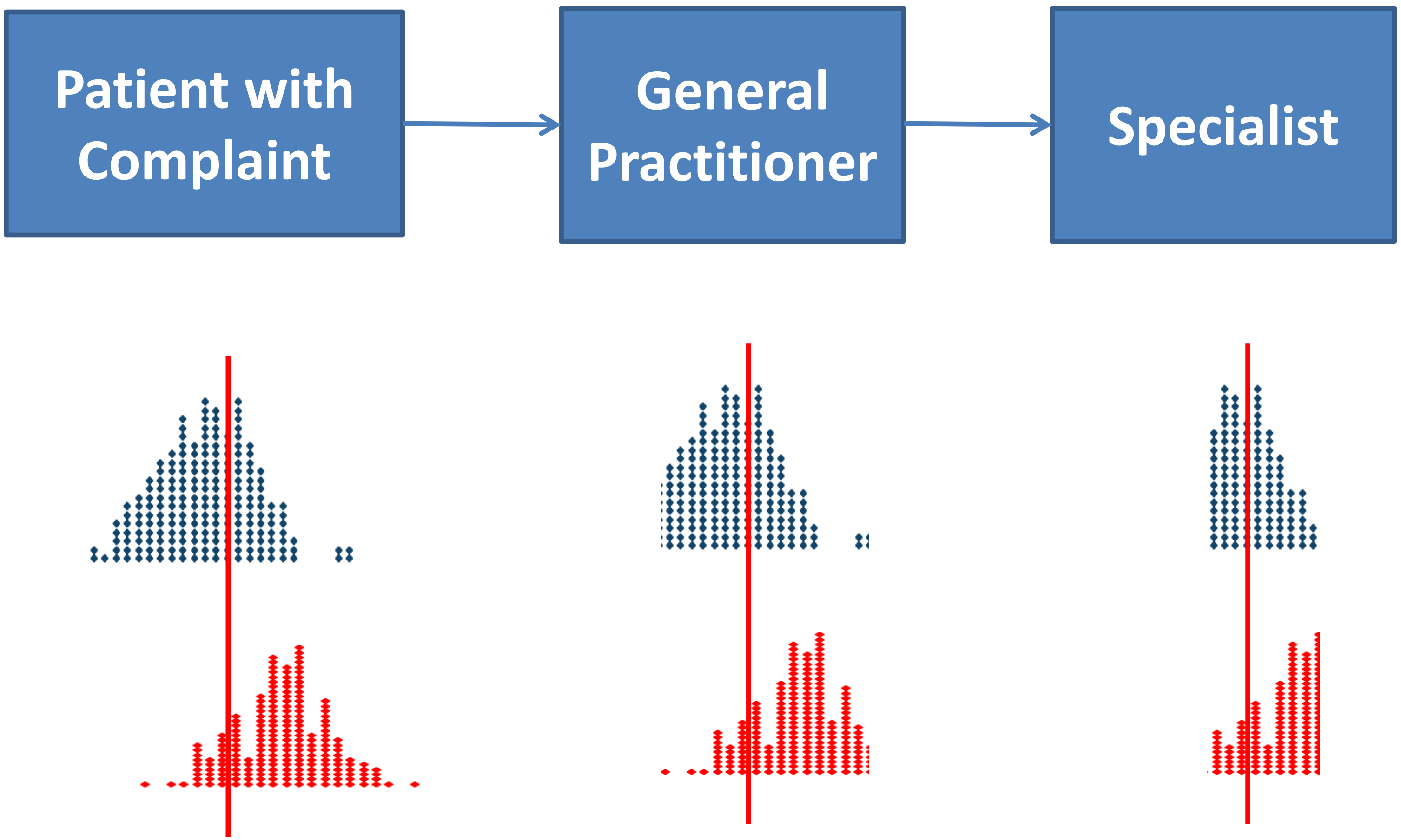
Ideal vs Realistic Conditions



Referral Pattern & Disease Spectrum

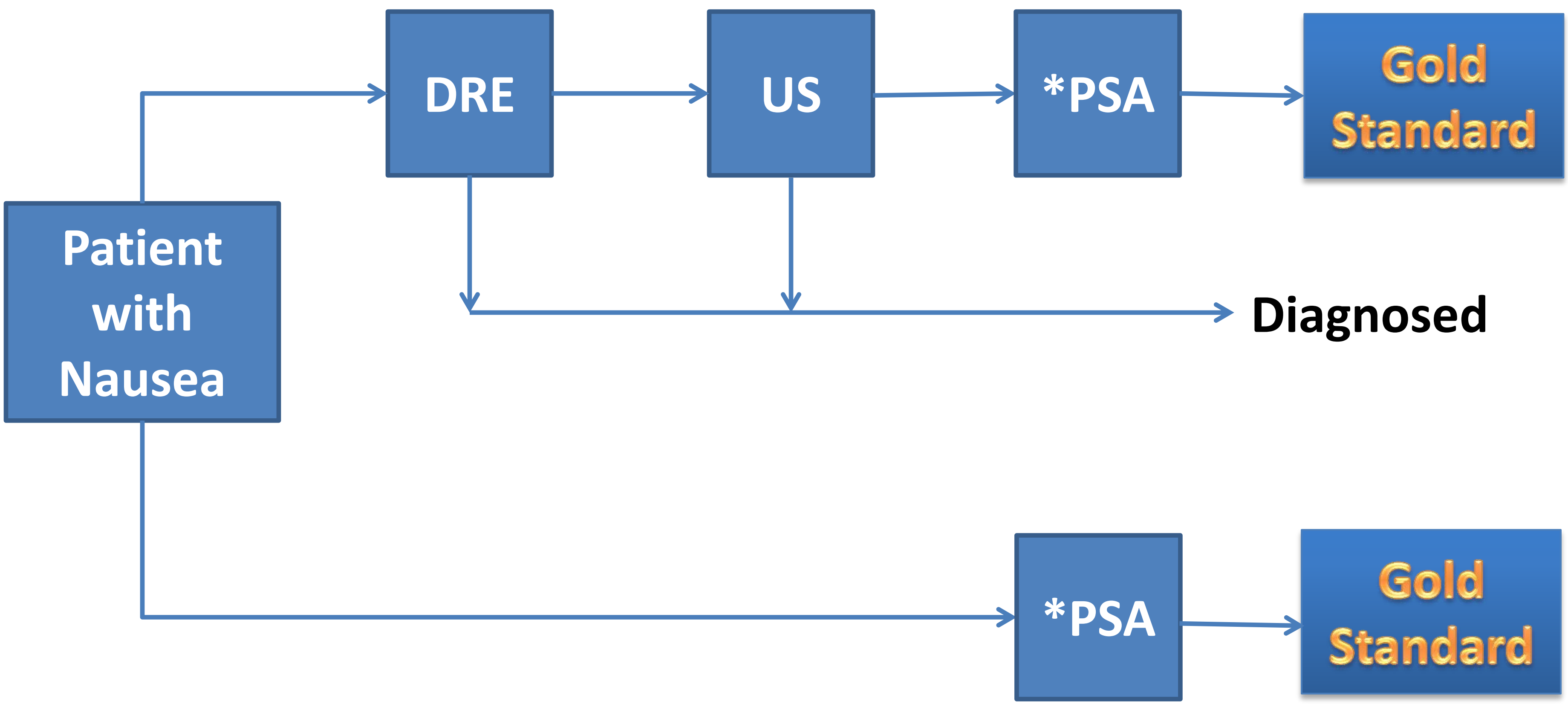


Referral Pattern & Disease Spectrum



Effect of Prior Testing

Will the index test perform differently?



*Index Test = the test of interest

Defining a Test: PICCO

Sources of Real Differences: Context is Everything

P	Population	Setting Exclusion/Inclusion criteria Referral pattern Comorbidities Age, Gender
I	Index Test	Method (in detail) Cutoff Skill level
C	Condition	Disease of interest
C	Comparator (reference test)	Definition of disease
O	Outcome measure	Diagnostic accuracy Discomfort, adverse events Operational (TAT, Availability, cost, etc)

Comparing Test Performance

Why do test results differ?

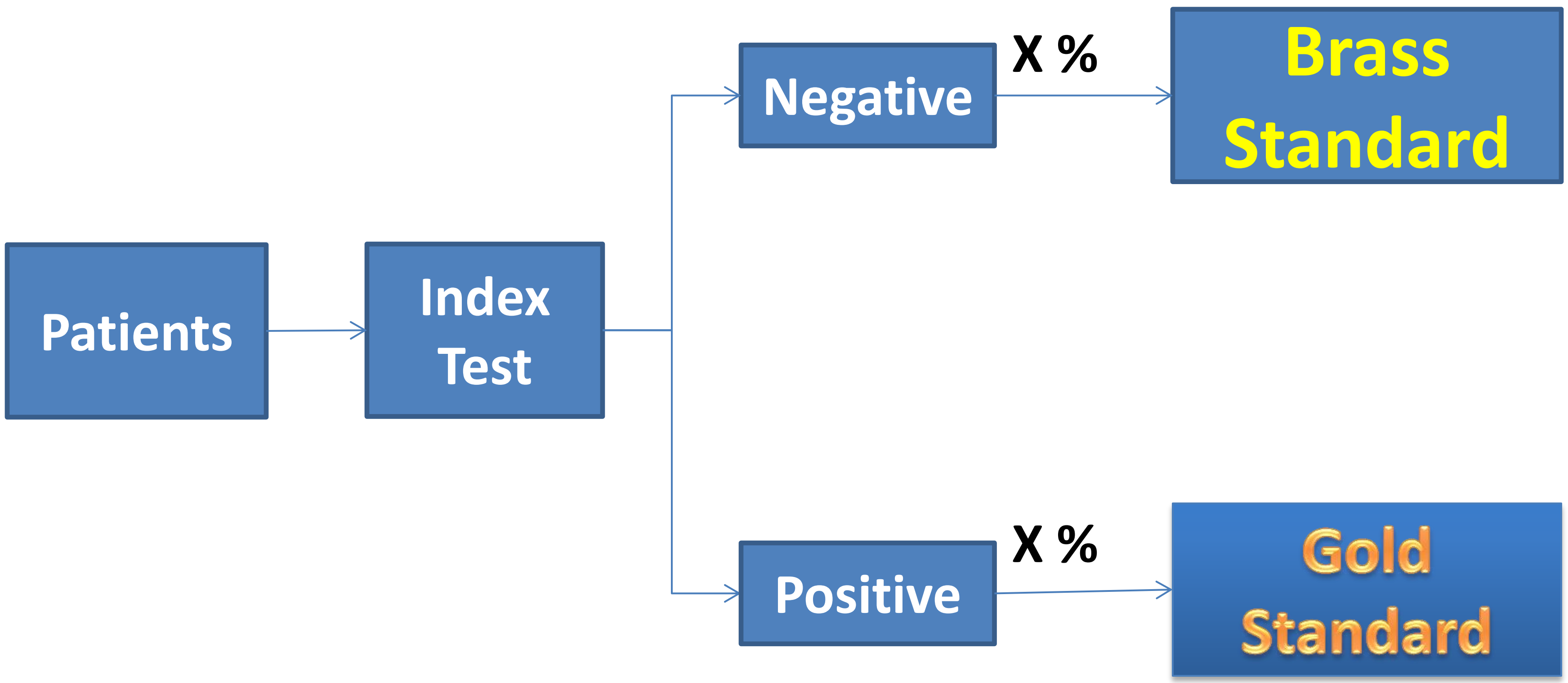
1. True differences
2. False differences
   - **bias**
   - thresholds
3. Random variation (imprecision)

Sources of Bias (Phantom Differences)

- Imperfect gold standard
- Verification Bias
- Indeterminate results
- Others….

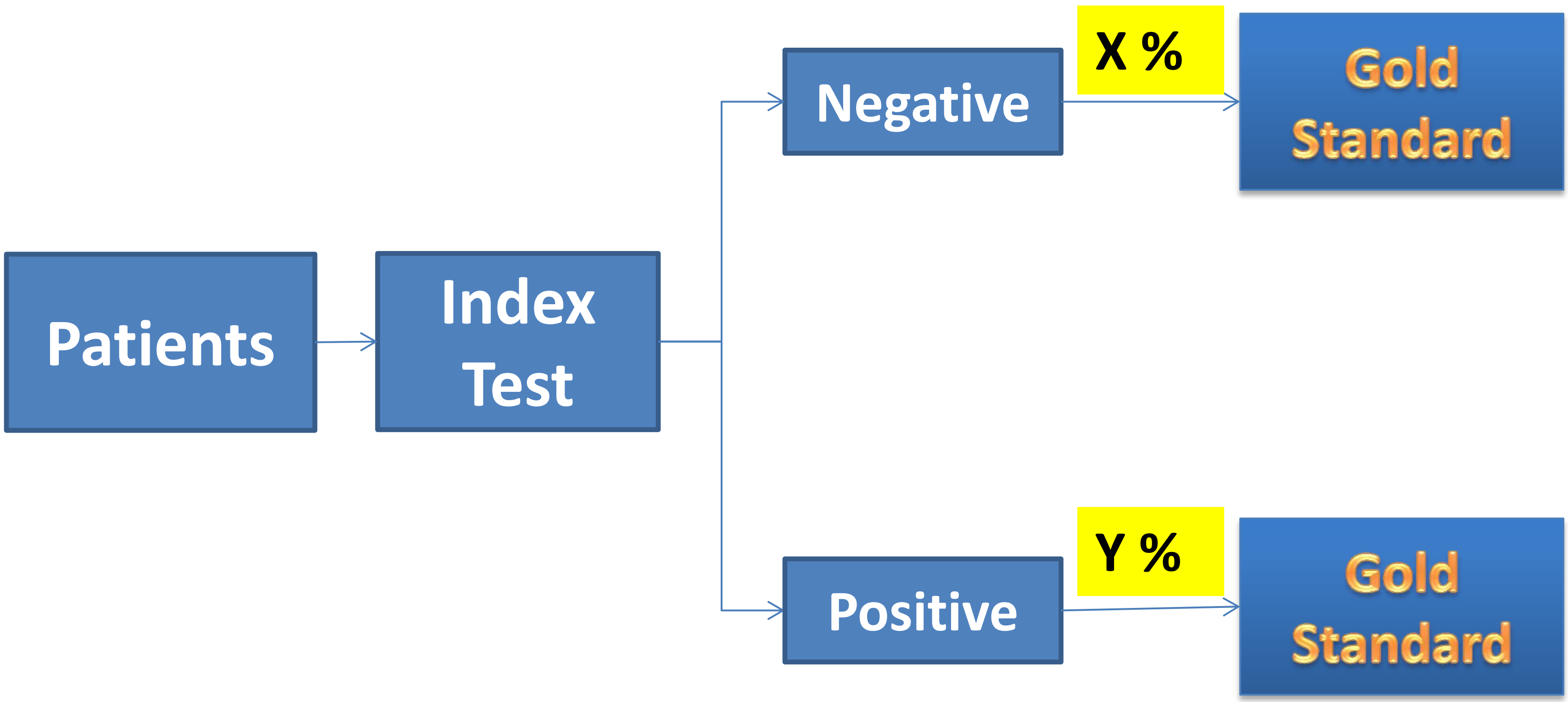
Imperfect Gold Standard

(Differential verification)



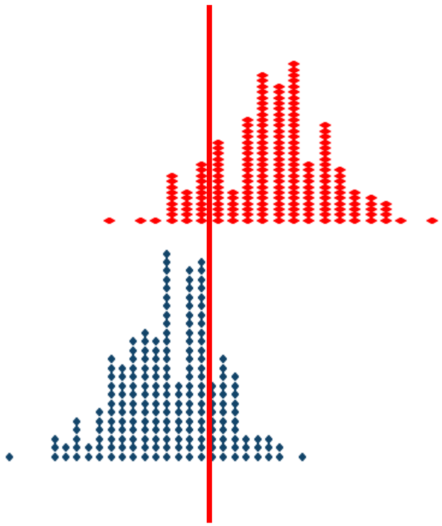
Verification bias

(Differential sampling)



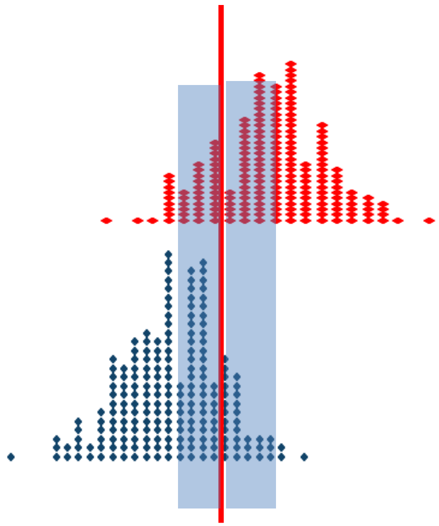
Bias due to indeterminates

Evaluator A:
No indeterminates



Low sensitivity
Low specificity

Evaluator B:
Many indeterminates



High sensitivity
High specificity

Indeterminates:
Where do these values go?

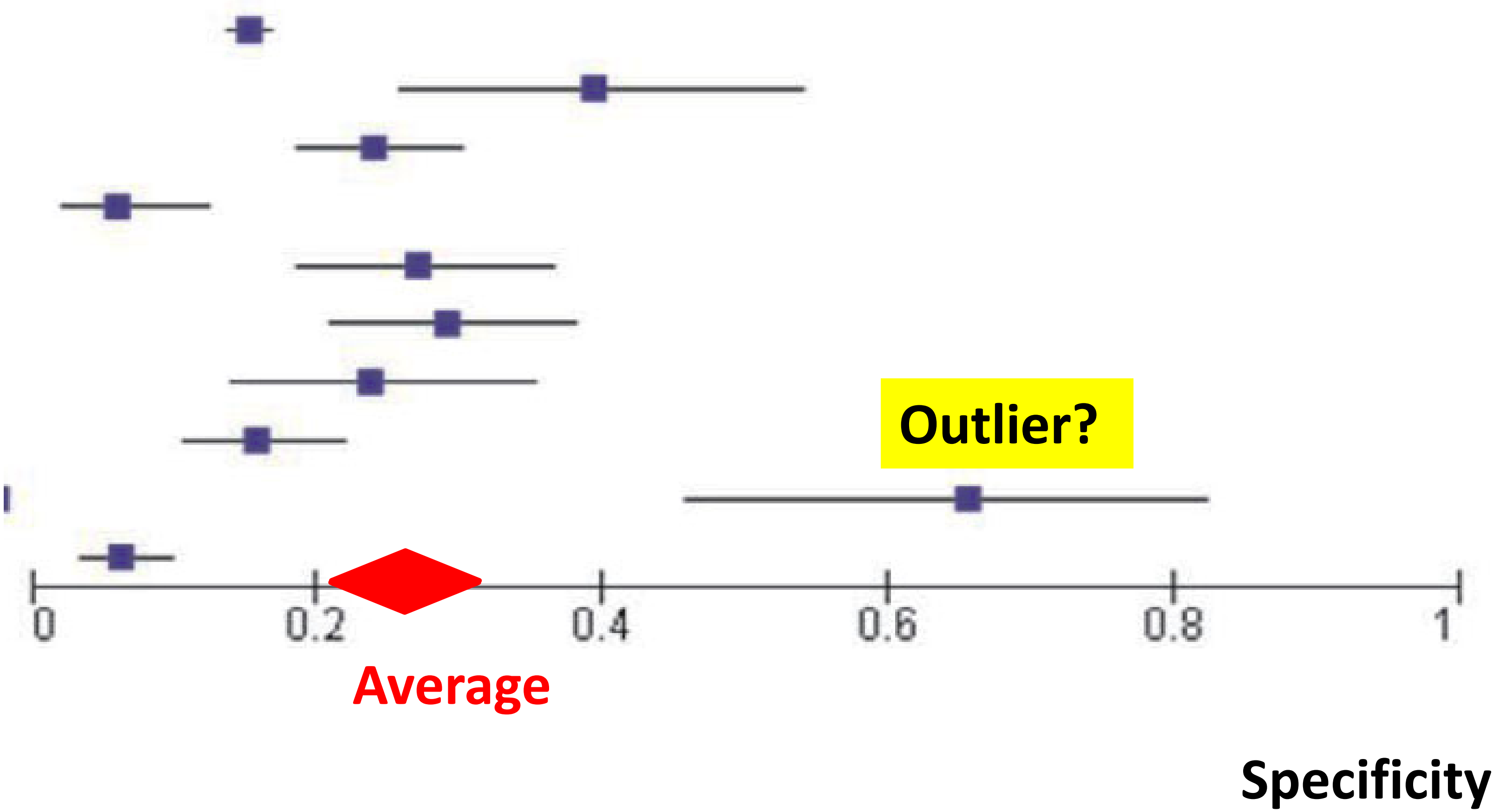
	Gold Standard		
Index Test	**Disease Present**	**Indeterminate**	**Disease Absent**
Positive		**X**	
Indeterminate	**U**	**Y**	**V**
Negative		**Z**	

Comparing Test Performance

Why do test results differ?

1. True differences
2. False differences (bias)
3. **Random variation (imprecision)**

Understanding Statistical Variation in Studies
Meta-Analysis



Comparing tests

Source of Difference	Countermeasures
True Differences	
	Complete Reporting
	PICCO
	Meta-analysis
False Differences	
bias	Improved Study Design
thresholds	ROC Curves
Random variation	
	Study design
	Meta-analysis

Higher Levels of Test Evaluation

Societal Impact

Cost effectiveness

Clinical effectiveness

Clinical performance

Analytical performance

Problems with Test Evaluation

- Potentially useful ≠ Clinically useful
- Potential problems
  - Tests are not used properly
  - Tests do not change diagnosis
  - Tests do not change management
- Tests are not used in isolation
  - Incremental value

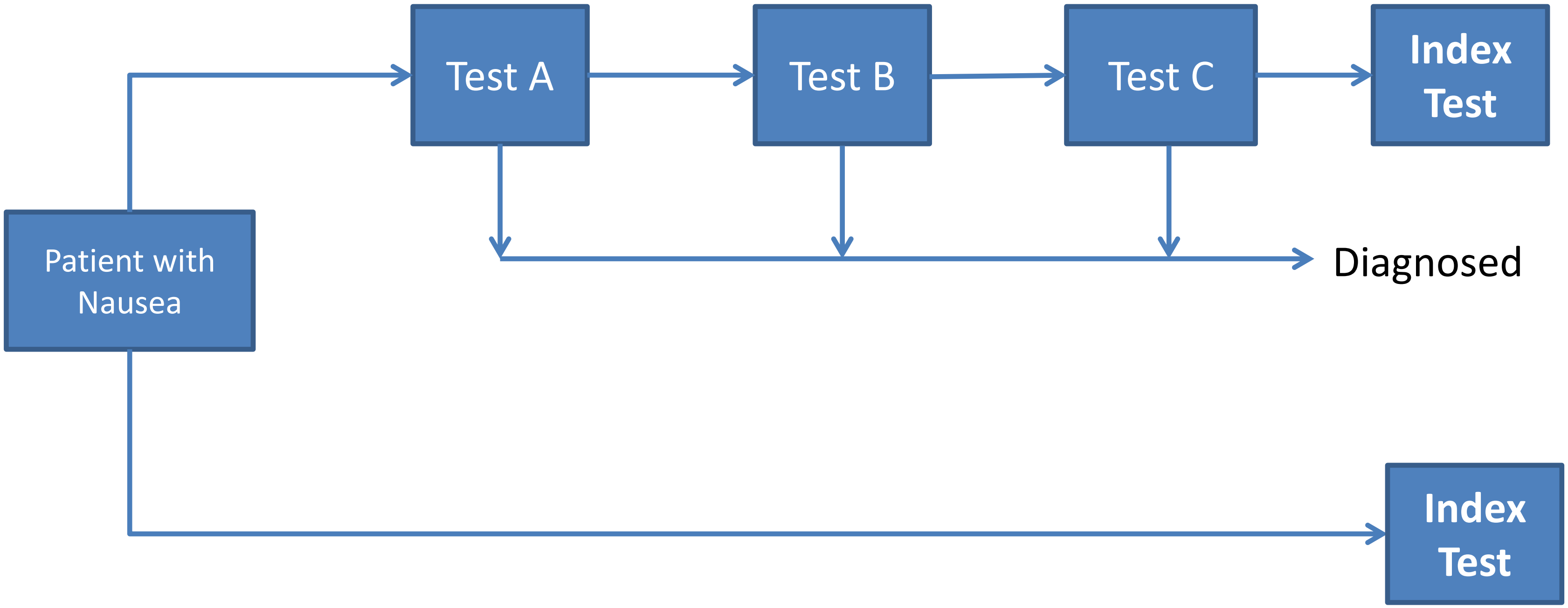
Clinical Trial Evaluation of Tests

Key Question:

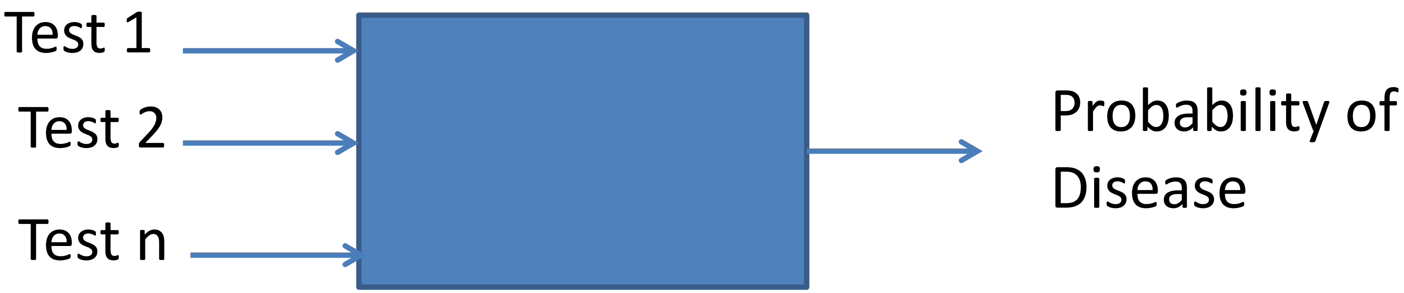
Do patients who receive this test have better outcomes?

Tests don’t exist in isolation

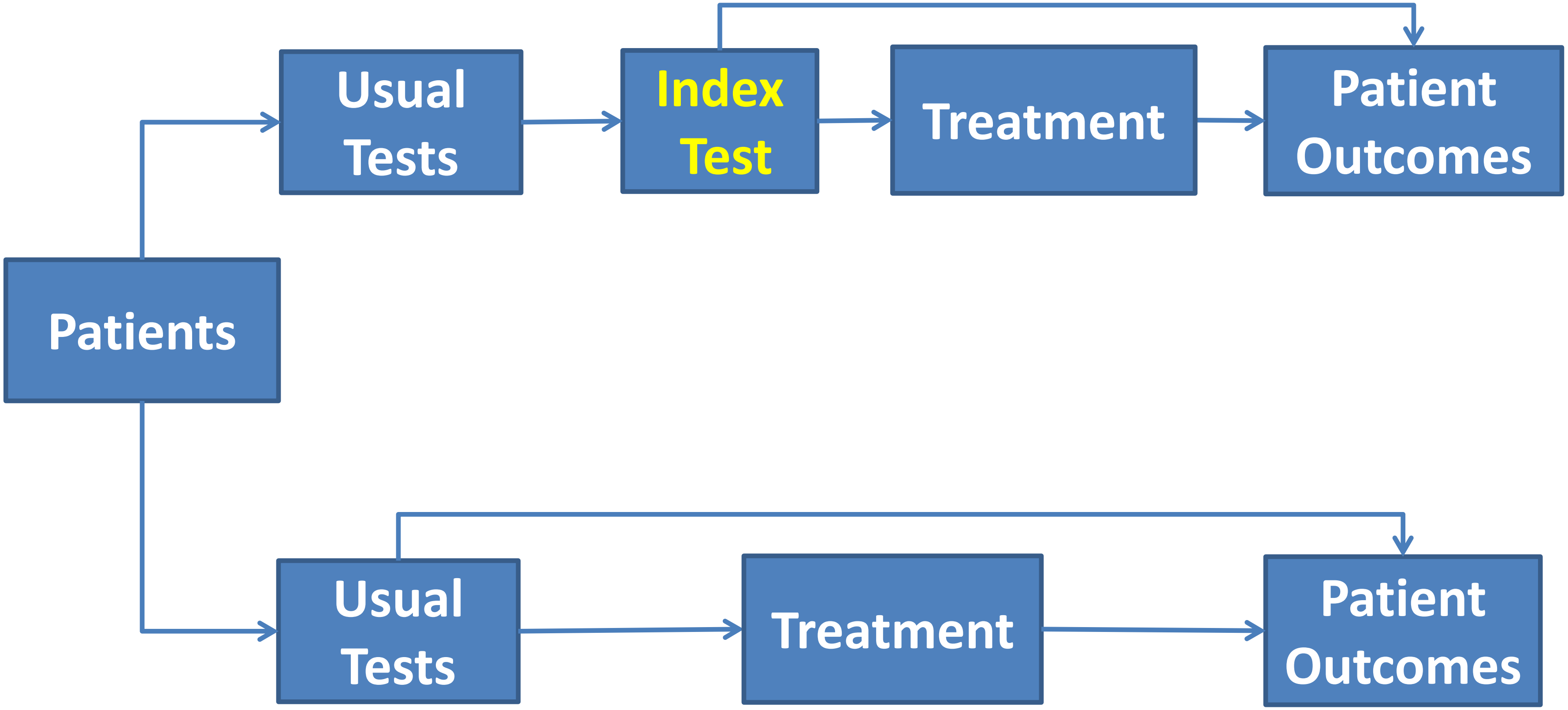
- Test Research vs Diagnostic Research
- What is the ***incremental*** value of a test?



Tests are often combined



The acid test



Clinical Trial Evaluation
Prostate Screening (PSA)

	Event Rate per 1000		
Outcomes	**No Screen**	**Screen**	**Relative Risk**
All cause mortality	200	198	0.99 [0.97-1.01]
Death from prostate CA	8	7	0.88 [0.71-1.09]
Prostate CA diagnosis	44	64	1.46 [1.21-1.77]

Levels of Evaluation

Therapeutics

Phase II/III Trial – Explanatory Trial

Scientific Perspective
Hypothesis: Does this drug affect outcomes?
As-Treated Analysis
Carefully controlled population, setting
Carefully controlled administration and monitoring

Phase III Trial – Pragmatic Trial

Policy Perspective
Hypothesis: Does *prescribing* this drug affect outcomes?
Intention-to-Treat Analysis
Patients seeking treatment for condition
Usual conditions

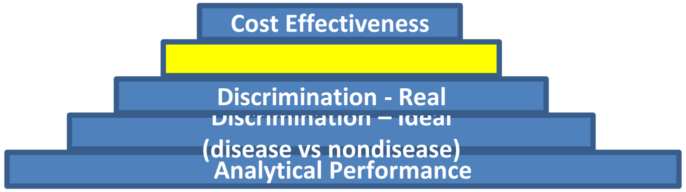
Diagnostics

Scientific Test Evaluation

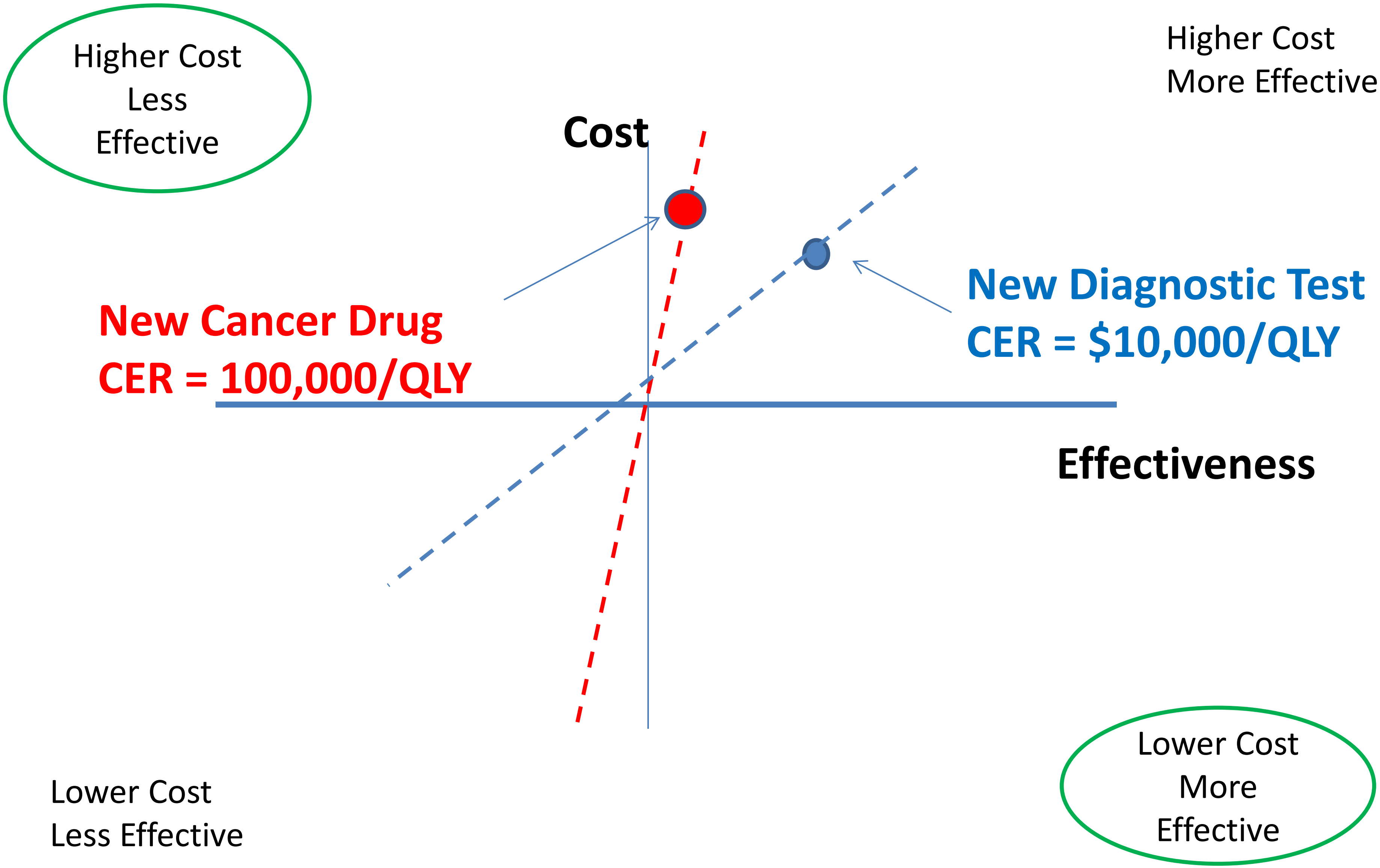
Single test
Idealized population
Expert administration
Expert interpretation

Pragmatic Test Evaluation

Multiple tests
Actual population
Usual conditions



Cost-Effectiveness Plane



Summary

- Many ways to assess performance
- Many reasons why studies differ
  - Real differences (PICCO)
  - False differences
    - Thresholds
    - Bias
  - Statistical variation
- Progress in Performance Evaluation
  - Quality of Reporting
  - Quality of studies
  - Types of studies
- Educating Clinicians

Testing A Test:
Beyond Sensitivity and Specificity