Howard Takiff

Gobierno Bolivariano de Venezuela | Ministerio del Poder Popular para la Ciencia y Tecnología | Instituto Venezolano de Investigaciones Científicas

Evaluation of Diagnostic tests

Slides with this background are
mostly the presentation given by Muriel Vray,
of the Pasteur Institute,
in the 2014 TB course.

The 3 phases of development of a diagnostic test

The phase I (proof-of-concept)

The objective is to verify that results of the test are different in healthy and ill subjects (« Laboratory conditions »)

Verification of the mechanism of action

Test under different conditions (temperature, humidity..), condition of use

Test under different categories of subjects or samples (differents levels of severity of the disease, amount of bacteria, ….) including healthy subjects with similar symptoms as ill subjects

Reproducibility : capacity of a test to produce an identical result when repeated

Example: Dipstick to diagnose shigella in stools

- Check
  - dipsticks + in stools containing shigella
  - dipsticks - in stools not containing shigella
- Define how to collect stool samples (timing and conditions of sterility)
- Verify timing and condition of reading (instructions)
- Verify the reproducibility of the reading (use two dipsticks, read by two different operators)
- Test under different storage conditions (humidity,temperature)

➔ This phase allows to evaluate if the test is sufficiently « valid » to be used under the required conditions

Good reproducibility is necessary

The phase II (case-control study)

The objective is to show that
- The probablity to have a + result is higher in subjects with the disease
- The probablity to have a - result is higher in subjects without the disease

➔ Accuracy of the test under controlled conditions (≠ real world conditions)

Select the cases and the controls, as well as evaluators (physicians, nurses ..) who may differ from those in real world conditions

Define in a PROTOCOL the conditions under which the test will be used (avoiding the main biases)

Try under different conditions (temperature, humidity..)

Estimate the % of false positive and false negative cases

For continuous tests, define the cut-off (Roc curves), identify the factors that affect the test (or those that make it uninterpretable)

Phase I and Phase II studies are **retrospective** studies conducted only for research purposes

The status of the subjects (with or without the disease) is known before the conduct of the test

The phase III (prospective study)

To assess the accuracy of the index test under real world conditions

➔ Conducted in pratical conditions under which the test will be used

Concerns all subjects for whom status in unknown (with or without the disease) ➔ « grey zone»

To verify that the results of the test allow to distinguish patients with or without the disease compared to phase II studies that give overestimated accuracies

How to evaluate a new test?

- **SENSITIVITY**
- **SPECIFICITY**
- POSITIVE PREDICTIVE VALUE (PPV)
- NEGATIVE PREDICTIVE VALUE (NPV)
- POSITIVE LIKELIHOOD RATIO (LR+)
- NEGATIVE LIKELIHOOD RATIO (LR-)

Sensitivity and Specificity

- **Sensitivity**
  - The ability of the test to identify correctly those who have the disease
- **Specificity**
  - The ability of the test to identify correctly those who **do not have** the disease

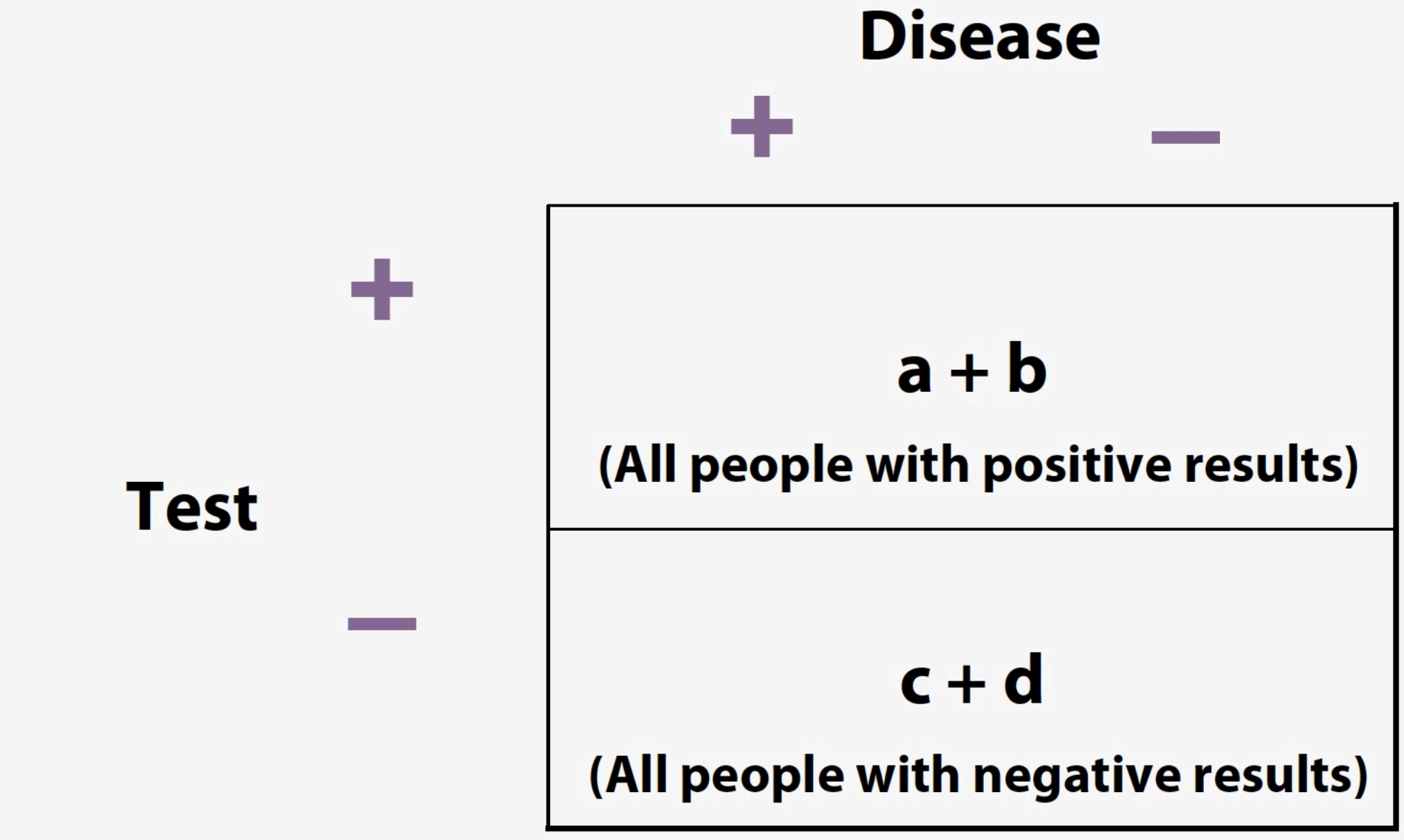
Determining the Sensitivity, Specificity of a New Test

- Must know the correct disease status prior to calculation
- **Gold standard test** is the best test available
  - It is often invasive or expensive
- A **new test** is, for example, a new screening test or a less expensive diagnostic test
- Use a 2 x 2 table to compare the performance of the new test to the gold standard test

Gold Standard Test

	Disease +	Disease −
	a+c (All people with disease)	b+d (All people without disease)

What the Test Shows



Comparison of Disease Status: Gold Standard Test and New Test

New test	Disease +	Disease −
+	**a** (True positives)	**b**
−	**c**	**d** (True negatives)

Comparison of Disease Status: Gold Standard Test and New Test

New test	Disease +	Disease −
+ Test is positive for Presence of disease	**a** (True positives)	**b** (False positives)
− Test is negative	**c** (False negatives)	**d** (True negatives)

Sensitivity

- **Sensitivity** is the ability of the test to identify correctly those who have the disease (a) from all individuals with the disease (a+c)

$$\text{sensitivity} = \frac{a}{a+c} = \frac{\text{true positives}}{\text{disease}+}$$
$$= \Pr(T+ \mid D+)$$

- Sensitivity is a fixed characteristic of the test

Test	Disease +	Disease –
+	**a** True Positive	**b** False Positive
–	**c** False Negative	**d** True Negative

Specificity

- **Specificity** is the ability of the test to identify correctly those who do not have the disease (d) from all individuals free from the disease (b+d)

$$\text{sensitivity} = \frac{d}{b+d} = \frac{\text{true negatives}}{\text{disease}-} = \Pr(T-\,|\,D-)$$

- Specificity is also a fixed characteristic of the test

Test	Disease +	Disease –
+	**a** True Positive	**b** False Positive
–	**c** False Negative	**d** True Negative

Applying Concept of Sensitivity and Specificity to a Screening Test

- Assume a population of 1,000 people
- 100 have a disease
- 900 do not have the disease
- A screening test is used to identify the 100 people with the disease
- The results of the screening appears in this table

Screening Results	True Characteristics in Population		Total
	Disease	No Disease	
Positive	80	100	180
Negative	20	800	820
Total	100	900	1,000

Calculating Sensitivity and Specificity

Screening Results	True Characteristics in Population		Total
	Disease	No Disease	
Positive	80	100	180
Negative	20	800	820
Total	100	900	1,000

Sensitivity = 80/100 = 80%

Specificity = 800/900 = 89%

Behind the Test Results

Test \ Disease	+	−
+	**a** (True positives)	**b** (False positives)
−	**c** (False negatives)	**d** (True negatives)

Review

- Fill in the missing cells and calculate sensitivity and specificity for this example

Screening Results	True Characteristics in Population		Total
	Disease	No Disease	
Positive	240		
Negative		600	
Total	300	700	1,000

Review

- Fill in the missing cells and calculate sensitivity and specificity for this example

Screening Results	True Characteristics in Population		Total
	Disease	No Disease	
Positive	240	?	?
Negative	?	600	?
Total	300	700	1,000

Review

- Fill in the missing cells and calculate sensitivity and specificity for this example

Screening Results	True Characteristics in Population		Total
	Disease	No Disease	
Positive	240	**100**	**340**
Negative	**60**	600	**660**
Total	300	700	1,000

Sensitivity : the test detected 240/300 patients **with** disease = 80% **sensitivity**

Specificity : the test was negative in 600/700 **without** disease = 86% **sensitivity**

How to evaluate a new test?

- SENSITIVITY
- SPECIFICITY
- **POSITIVE PREDICTIVE VALUE (PPV)**
- **NEGATIVE PREDICTIVE VALUE (NPV)**
- POSITIVE LIKELIHOOD RATIO (LR+)
- NEGATIVE LIKELIHOOD RATIO (LR-)

How to Estimate the Value of a Test

Predictive Values

- **Positive predictive value (PPV)**
  - The proportion of patients who test positive who actually have the disease
- **Negative predictive value (NPV)**
  - The proportion of patients who test negative who are actually free of the disease
- Note: PPV and NPV are not fixed characteristics of the test

The PPV and NPV are strongly influenced
by the prevalence of the disease
in the population where the test was administered

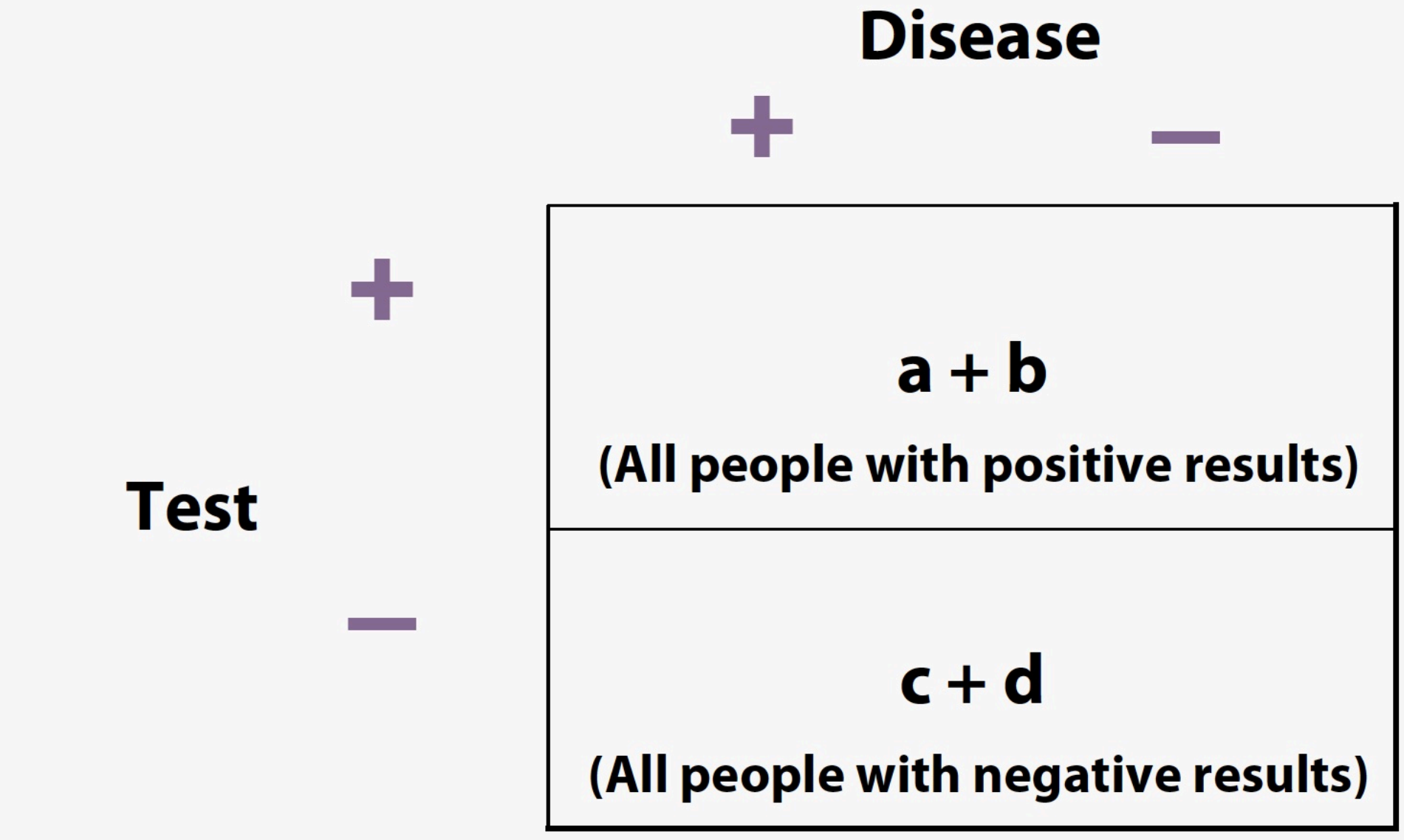
Another Interpretation of PPV

- If a person tests positive, what is the probability that he or she has the disease?
- (And if that person tests negative, what is the probability that he or she does not have the disease?)

Behind the Test Results

Test	Disease +	Disease −
+	**a** (True positives)	**b** (False positives)
−	**c** (False negatives)	**d** (True negatives)

What the Test Shows



Predictive Value

- Positive predictive value $= \dfrac{a}{a+b}$

Test \ Disease	+	–
+	**a** True Positive	**b** False Positive
–	**c** False Negative	**d** True Negative

$= \dfrac{\text{True Positives}}{\text{Test}+}$

$= P(D+ \mid T+)$

- Negative predictive value $= \dfrac{d}{c+d}$

$= \dfrac{\text{True Negatives}}{\text{Test}-}$

$= P(D- \mid T-)$

Applying Concept of Predictive Values to Screening Test

- Assume a population of 1,000 people
- 100 have a disease
- 900 do not have the disease
- A screening test is used to identify the 100 people with the disease
- The results of the screening appear in this table

Screening Results	True Characteristics in Population		Total
	Disease	No Disease	
Positive	80	100	180
Negative	20	800	820
Total	100	900	1,000

Calculating Predictive Values

Screening Results	True Characteristics in Population		Total
	Disease	No Disease	
Positive	80	100	180
Negative	20	800	820
Total	100	900	1,000

Calculating Predictive Values

Positive predictive value = 80/180 = 44%

Screening Results	True Characteristics in Population		Total
	Disease	No Disease	
Positive	80	100	180
Negative	20	800	820
Total	100	900	1,000

Negative predictive value = 800/820 = 98%

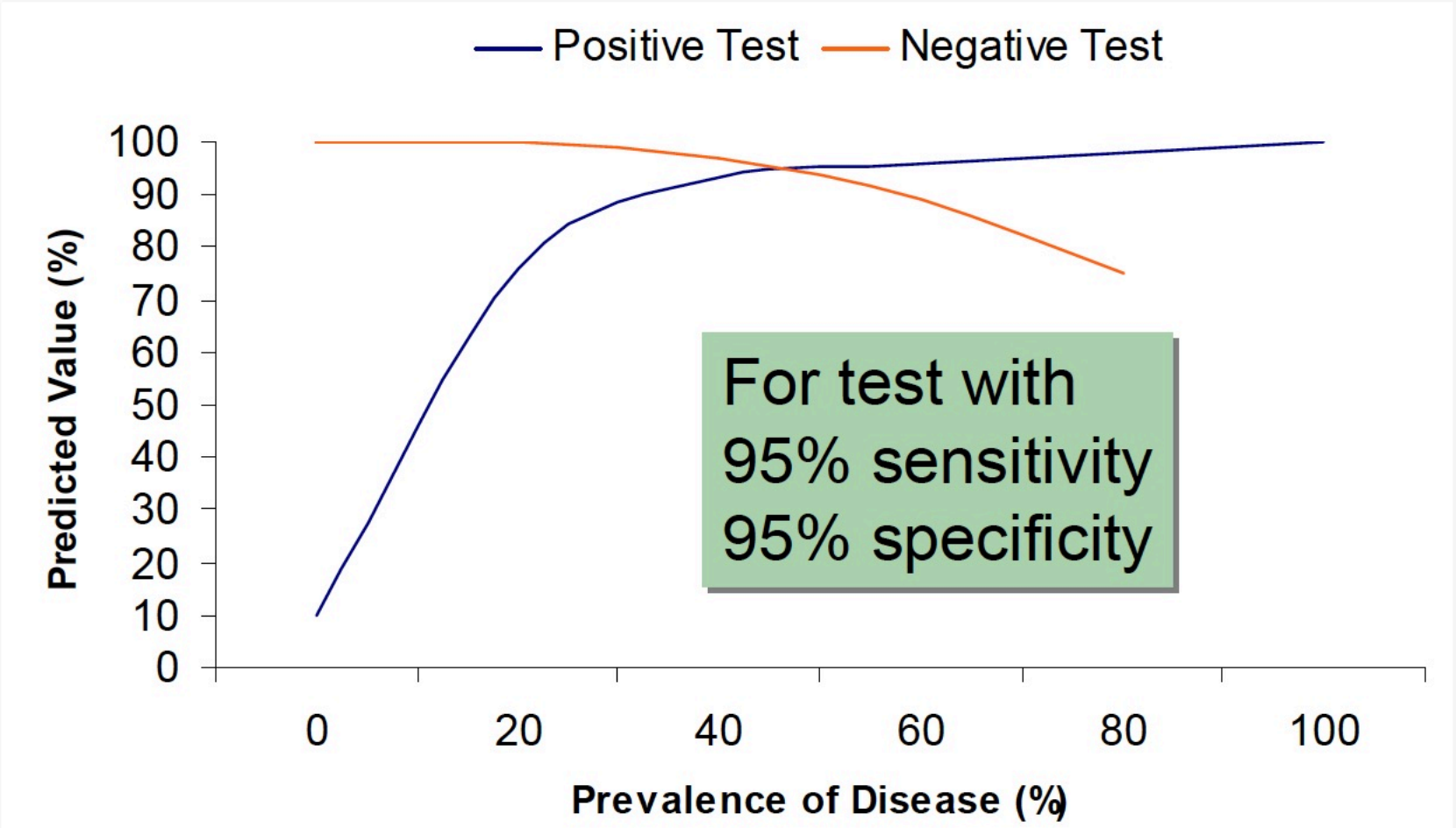
PPV Primarily Depends On . . .

- The prevalence of the disease in the population tested, and the test itself (sensitivity and specificity)
  - In general, it depends more on the specificity (and less on the sensitivity) of the test (if the disease prevalence is low)

Relationship of Disease Prevalence to Predictive Value

Example: Sensitivity = 99%; Specificity = 95%					
Disease Prevalence	Test Results	Sick	Not Sick	Totals	Positive Predictive Value
1%	+	99	495	594	$\frac{99}{594} = 17\%$
	–	1	9,405	9,406	
	Totals	100	9,900	10,000	
5%	+	495	475	970	$\frac{495}{970} = 51\%$
	–	5	9,025	9,303	
	Totals	500	9,500	10,000	

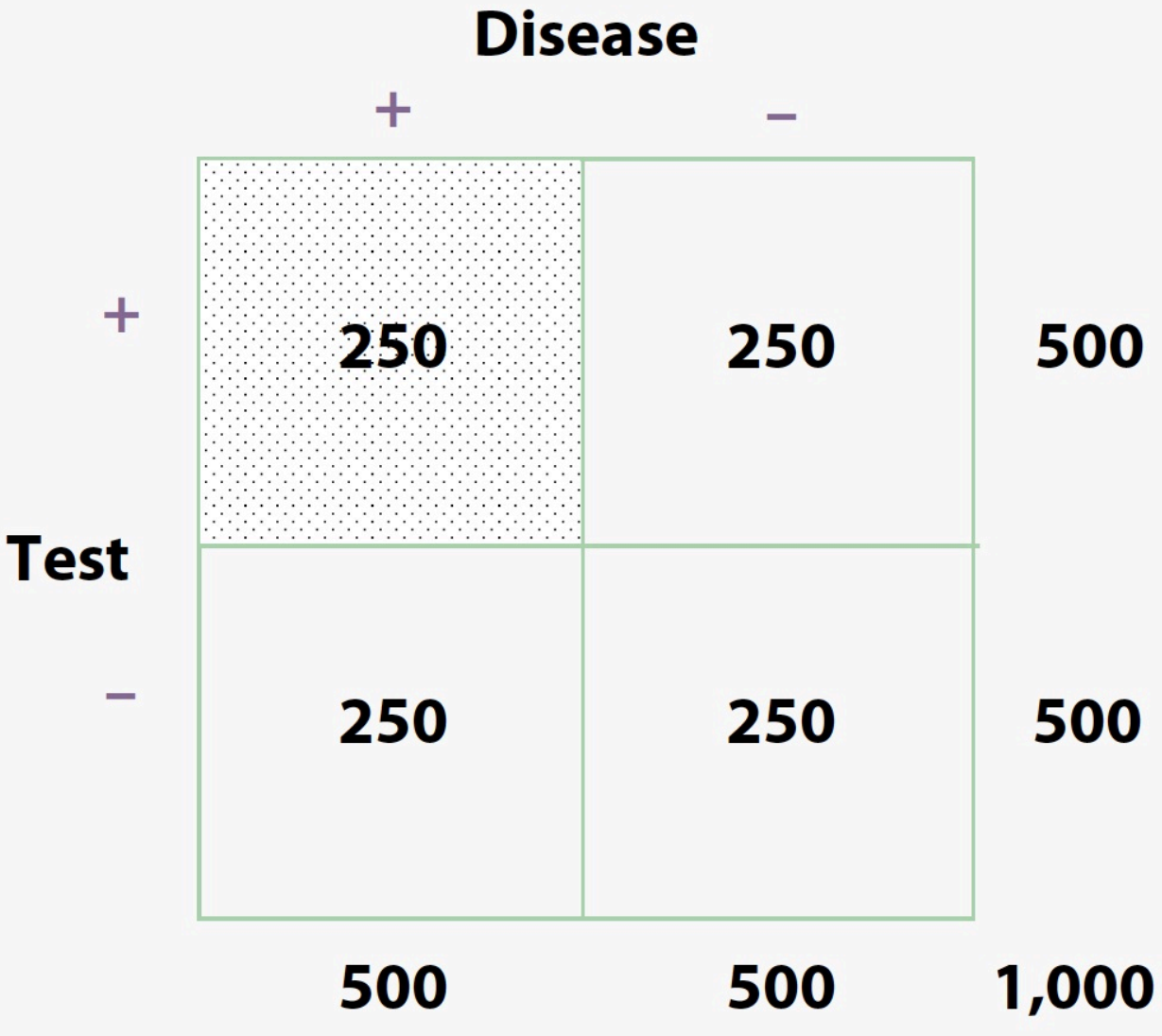
Prevalence of Disease



So If a Person Tests Positive . . .

- The probability that he or she has the disease depends on the prevalence of the disease in the population tested and the validity of the test (sensitivity and specificity)
- In general, specificity has more impact on predictive values

The Relationship of Specificity to Predictive Value



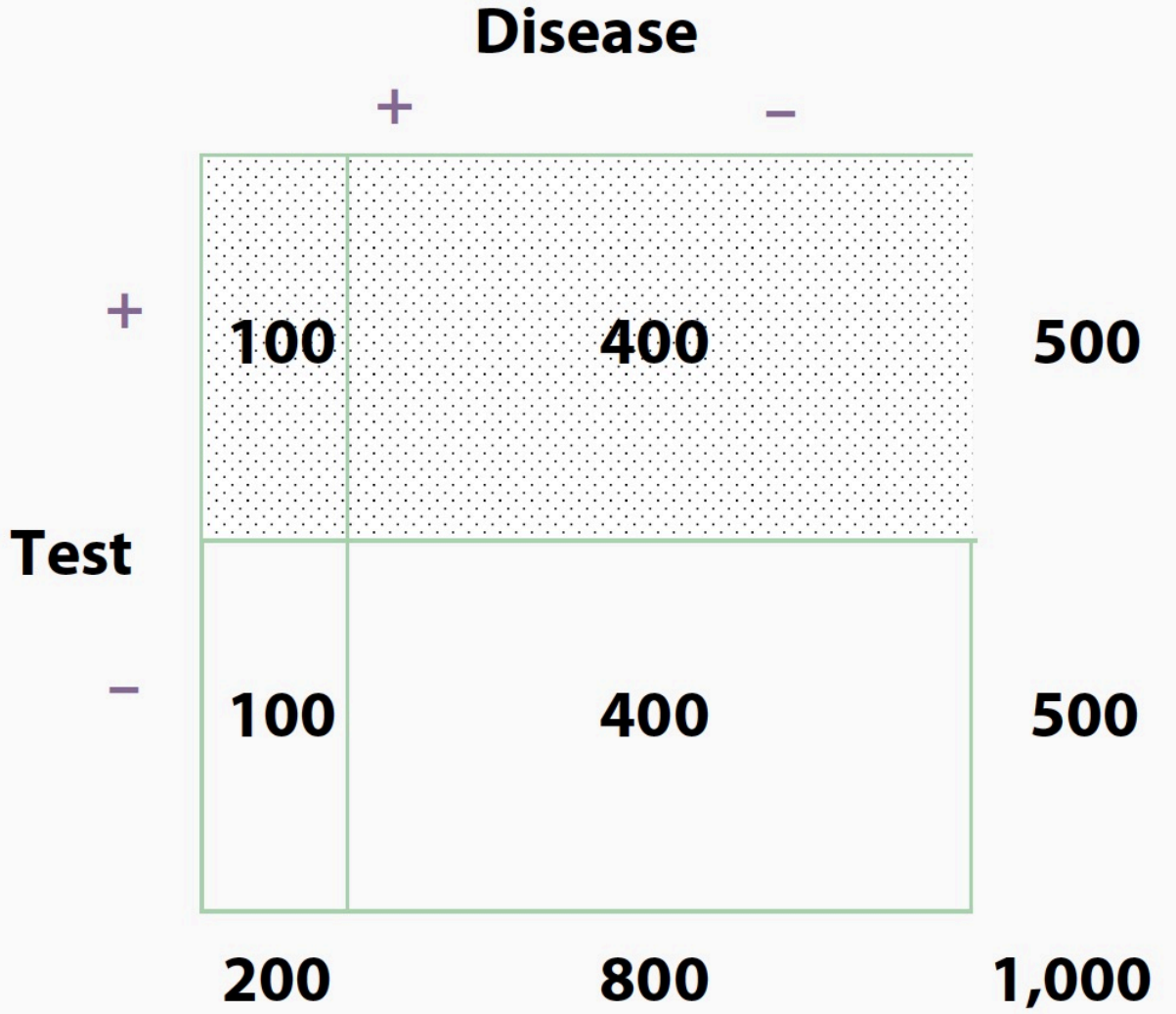
Prevalence = 50%

Sensitivity = 50%

Specificity = 50%

$$PPV = \frac{250}{500} = 50\%$$

The Relationship of Specificity to Predictive Value



Prevalence = 20%

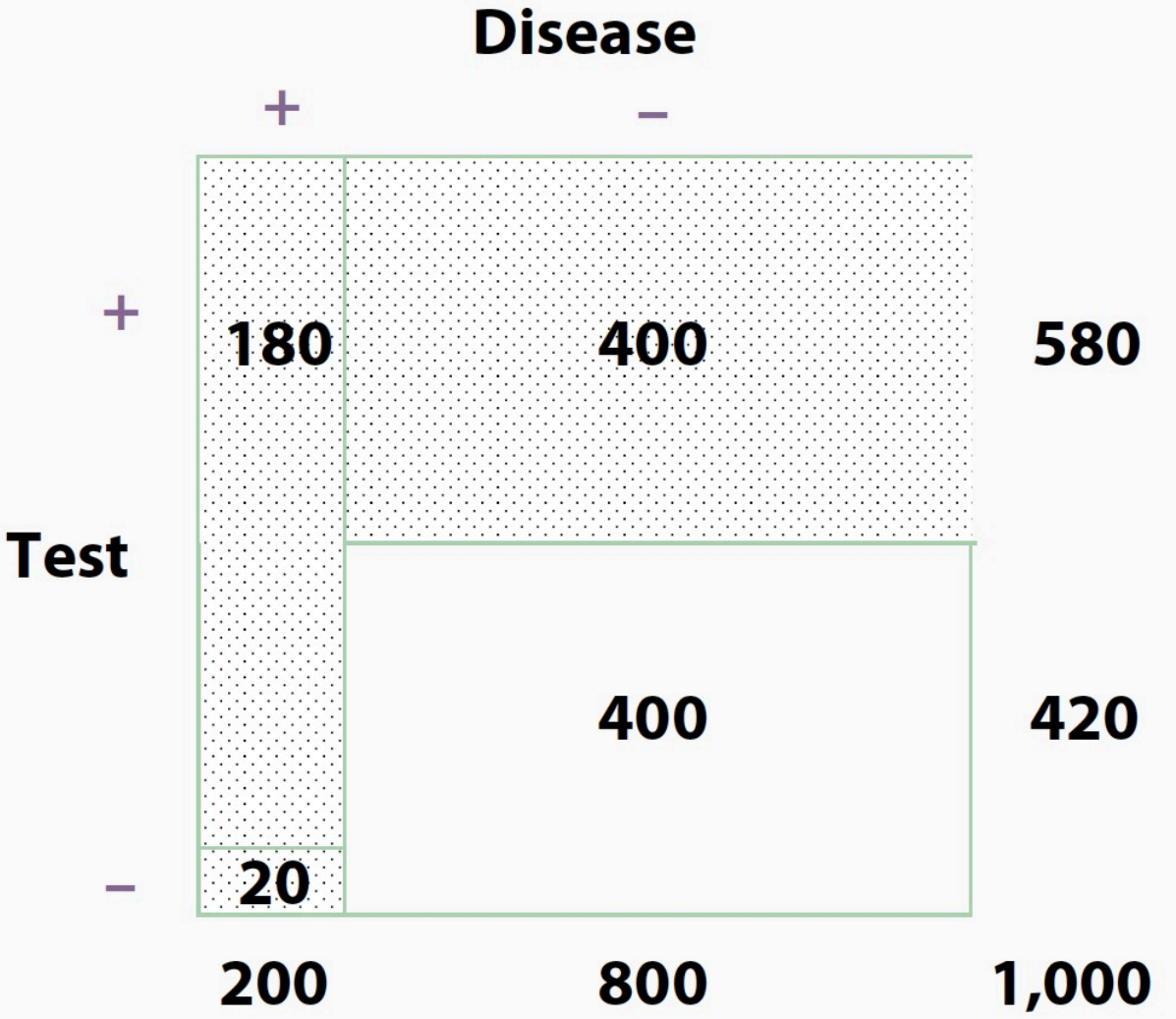
Sensitivity = 50%

Specificity = 50%

$$PPV = \frac{100}{500} = 20\%$$

Change prevalence

The Relationship of Specificity to Predictive Value



Prevalence = 20%

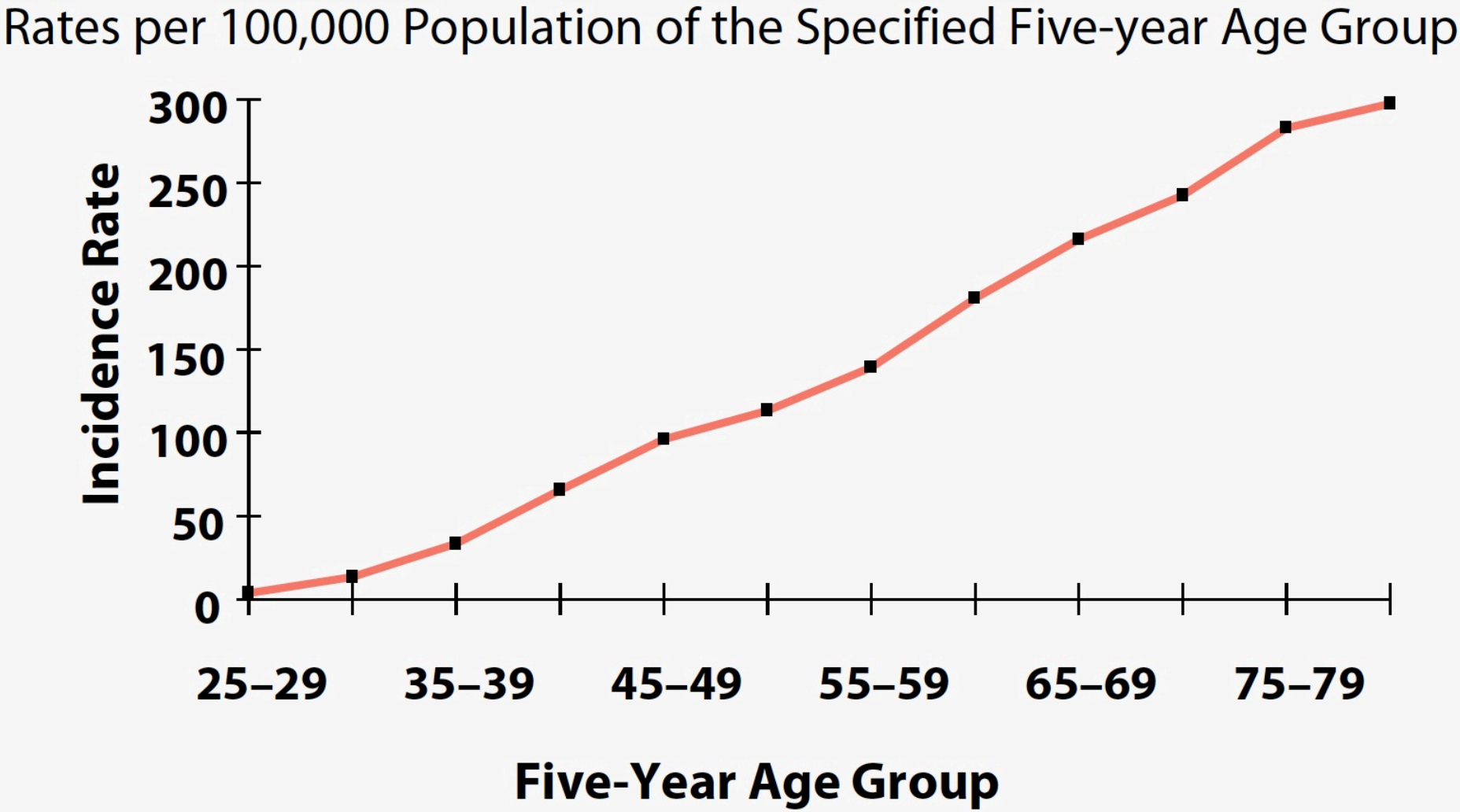
Sensitivity = 90%

Specificity = 50%

$$PPV = \frac{180}{580} = 31\%$$

Change sensitivity

Age-Specific Breast Cancer Incidence Rates U.S., All Races (SEER 1984-88)



Results of First Screening Mammography by Age Group — UCSF Mobile Mammography Program

Age (Years)	Cancer Detected	No Cancer Detected	Total Abnormal	Positive Predictive Value
30–39	9	273	282	3%
40–49	26	571	597	4%
50–59	30	297	327	9%
60–69	46	230	276	17%
70	26	108	134	19%

Consequence of Different PPVs

Age	< 50 Years	≥ 50 Years
Positive predictive value	4%	14%

So Screening tests for diseases with a low prevalence require a confirmatory test with a very high specificity.

How to evaluate a new test?

- SENSITIVITY
- SPECIFICITY
- POSITIVE PREDICTIVE VALUE (PPV)
- NEGATIVE PREDICTIVE VALUE (NPV)

How to summarize the accuracy of a test

- **POSITIVE LIKELIHOOD RATIO (LR+)**
- **NEGATIVE LIKELIHOOD RATIO (LR-)**

Positive Likelihood Ratio (LR+)

Definition: ratio of the probability of a positive test in subjects with the disease compared to subjects without the disease

- Probability of a positive test in subjects with the disease : sensibility (Se)
- Probability of a positive test in subjects without the disease : 1-specificity (1-Sp)
- LR+= Se / (1-Sp)

True positives/False positives

How much more likely is a positive test to be found in a person with, as opposed to without, the disease

The higher the value of LR+, the more important the diagnostic gain of a positive result

Negative Likelihood Ratio (LR-)

Definition: ratio of the probability of a negative test in subjects with the disease compared to subjects without the disease

- Probability of a negative test in subjects with the disease :
  = 1- sensitivity (1-Se)
- Probability of a negative test in subjects without the disease :
  = Specificity (Sp)

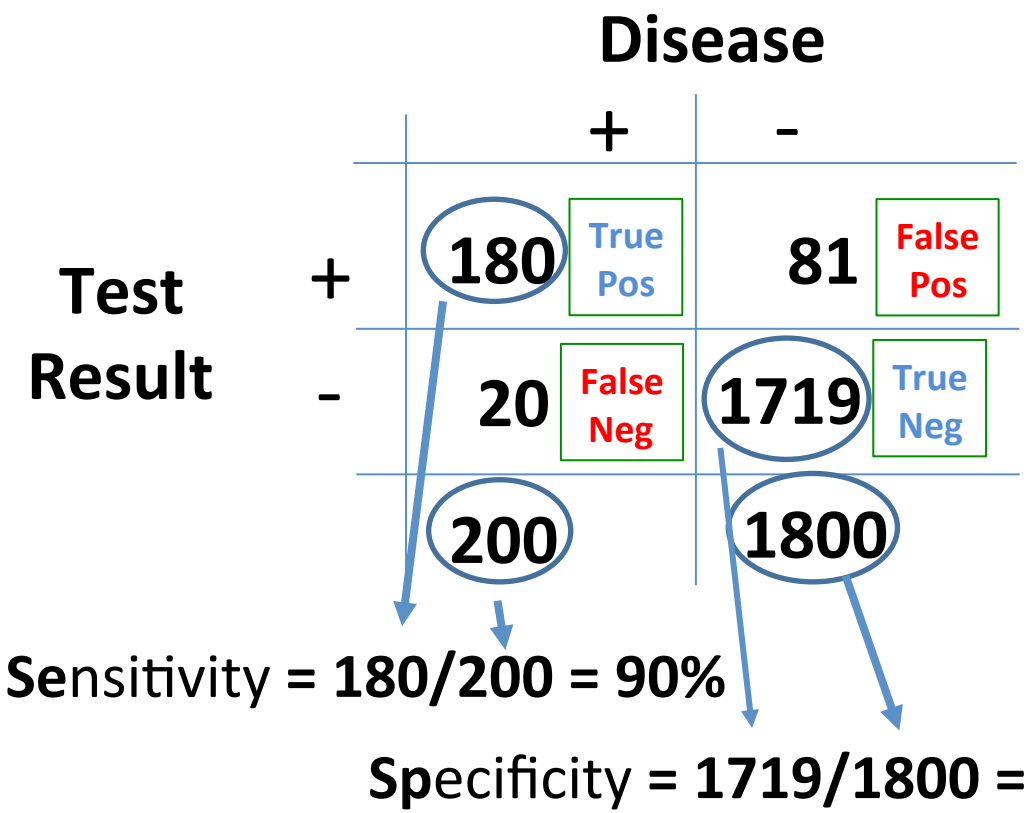
- LR- = (1-Se)/ Sp

% False negatives/%True Negatives

How much more likely is a negative test to be found in a person with, as opposed to without, the disease

The lower the value of LR-, the more important the diagnostic gain of a negative result

Likelihood Ratio LR (An Example)



Se = 180/200 = 90%

Sp = 1719/1800 = 95.5%

LR+ = Se/(1-Sp)

= 0.9/(0.045) = 20

LR- = (1-Se)/(Sp)

= 0.1/ (0.955) = 0.1

Likelihood Ratio **LR**

- Takes into account both Sensitivity **(Se)** & Specificity **(Sp)**
- Independent of the prevalence of the disease
- Calculates the probability that the subjects with a **positive test** really **HAVE** the disease (LR+)
- Calculates the probability that subjects with a **negative test** really **DON'T HAVE** the disease (LR-)
- A "**GOOD**" test:

LR+ > 9
LR- < 0,1

Evaluation of a

Continuous test : ROC curves (Receiver operating characteristic)

How to determine the Cut-off values of a test above which the test is positive

Variation in Biologic Values

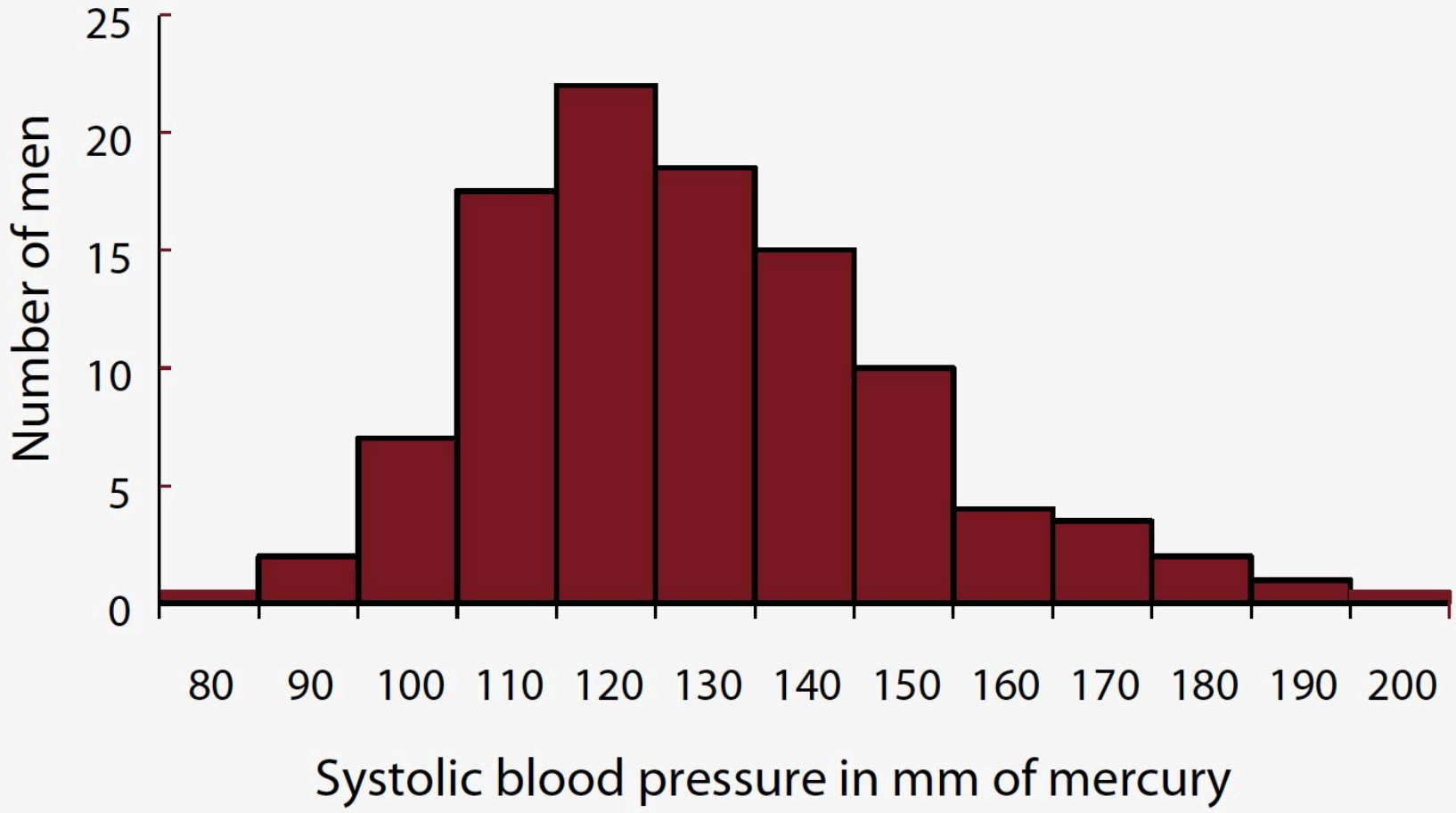
- Many test results have a continuous scale (are continuous variables)
- Distribution of biologic measurements in humans may or may not permit easy separation of diseased from non-diseased individuals, based upon the value of the measurement

Distribution of Tuberculin Reactions

EXAMPLE:
At what diameter should the PPD test be considered positive?

Source: Edwards et al, WHO Monograph 12, 1953

Distribution of Systolic Blood Pressures: 744 Employed White Males, Ages 40–64

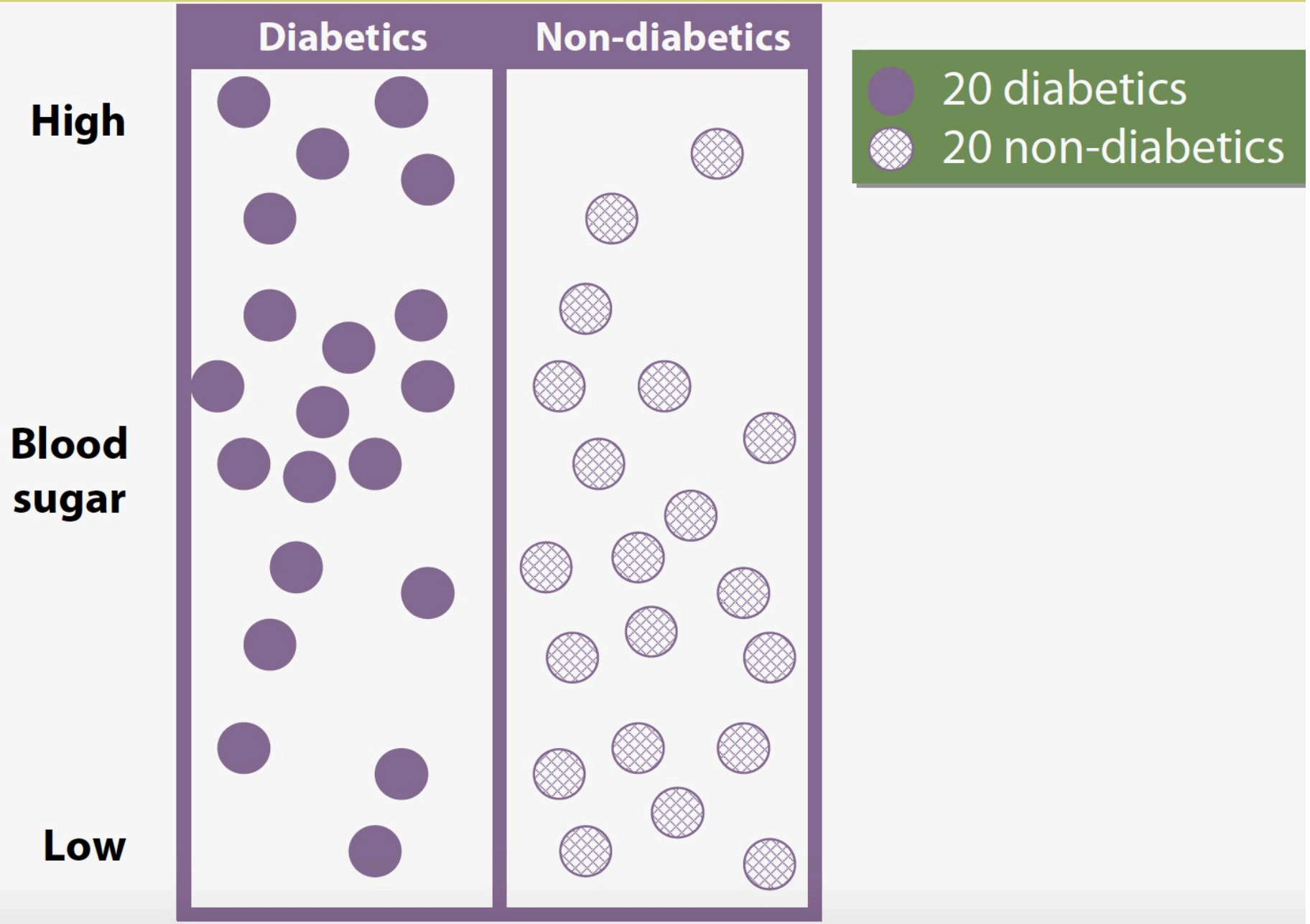


Source: Data of W.W. Holland

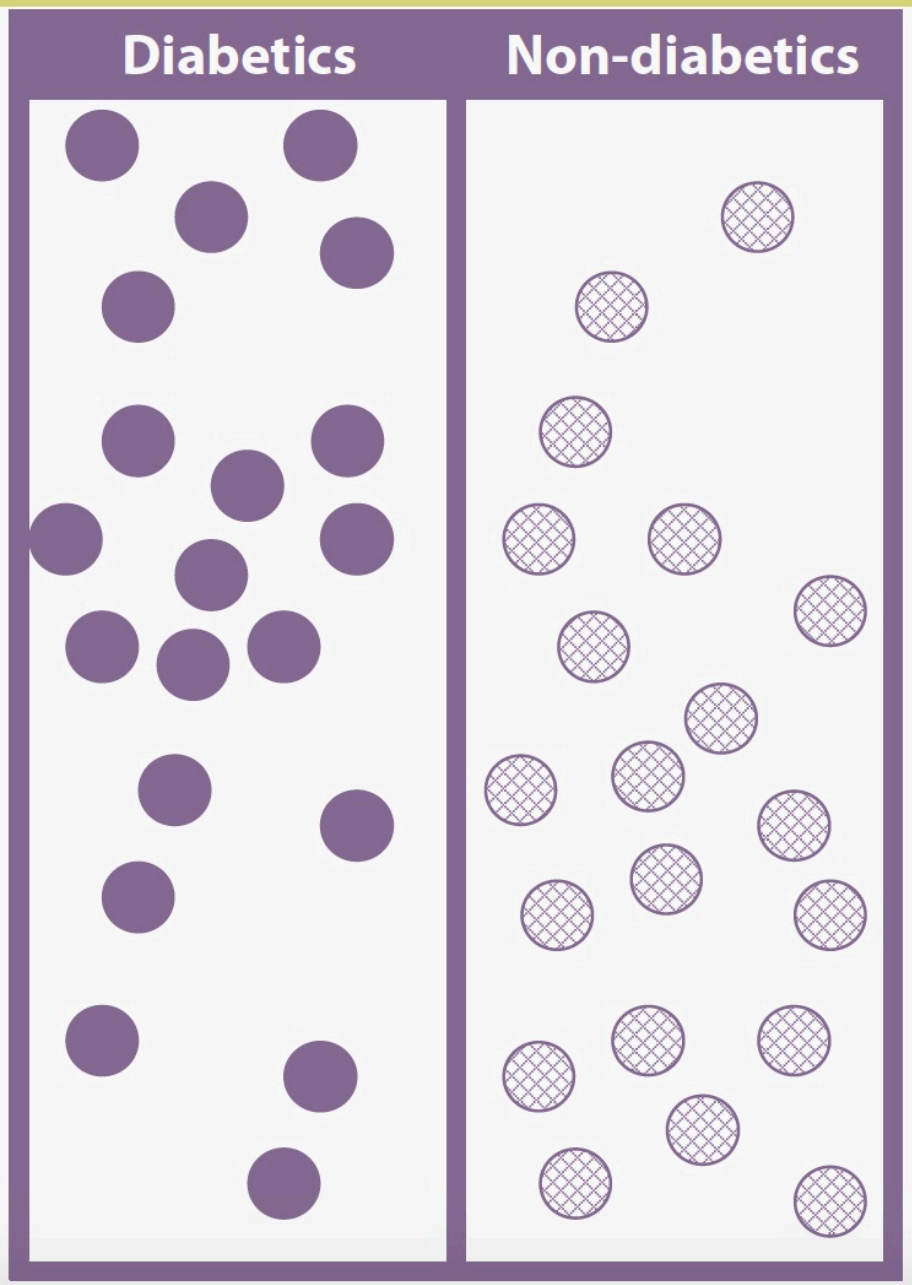
Examining the Effect of Changing Cut-Points

- Example: type II diabetes mellitus
  - Highly prevalent in the older, especially obese, U.S. population
  - Diagnosis requires oral glucose tolerance test
  - Subjects drink a glucose solution, and blood is drawn at intervals for measurement of glucose
  - Screening test is fasting plasma glucose
    - Easier, faster, more convenient, and less expensive

Concept of Sensitivity and Specificity

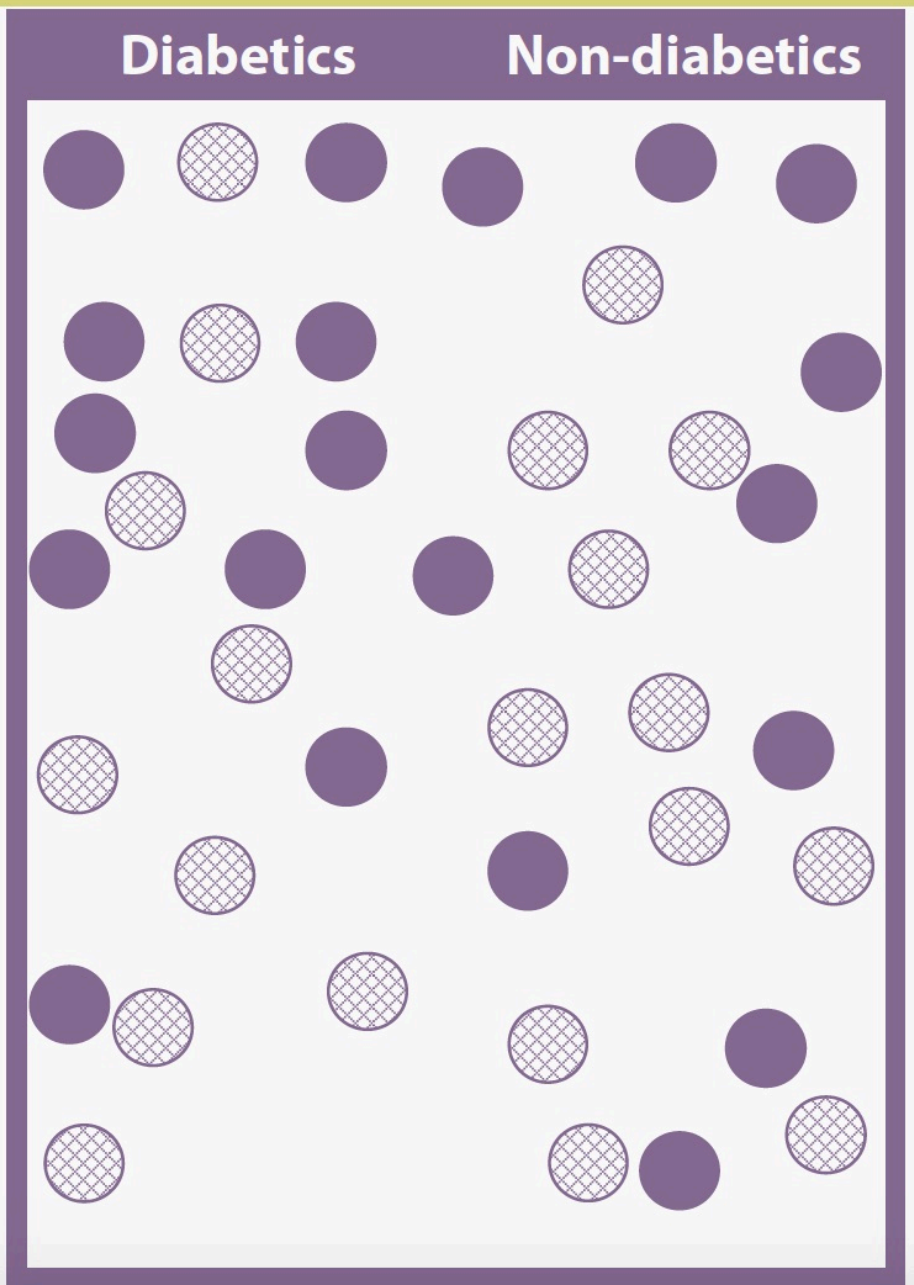


Concept of Sensitivity and Specificity



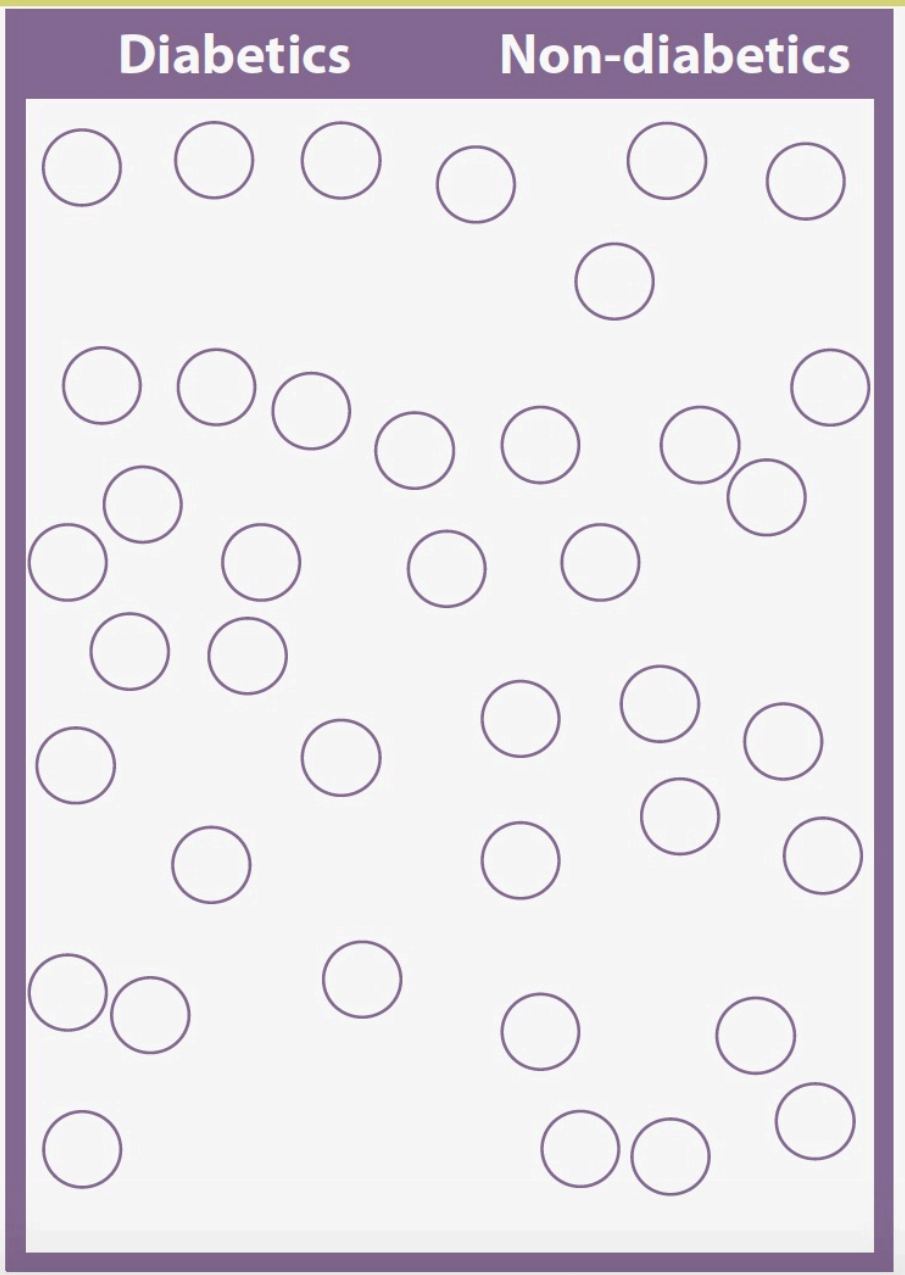
In a typical population, there is no line separating the two groups, and the subjects are mixed

Concept of Sensitivity and Specificity



In a typical population, there is no line separating the two groups, and the subjects are mixed

Concept of Sensitivity and Specificity



In fact, there is no color or label

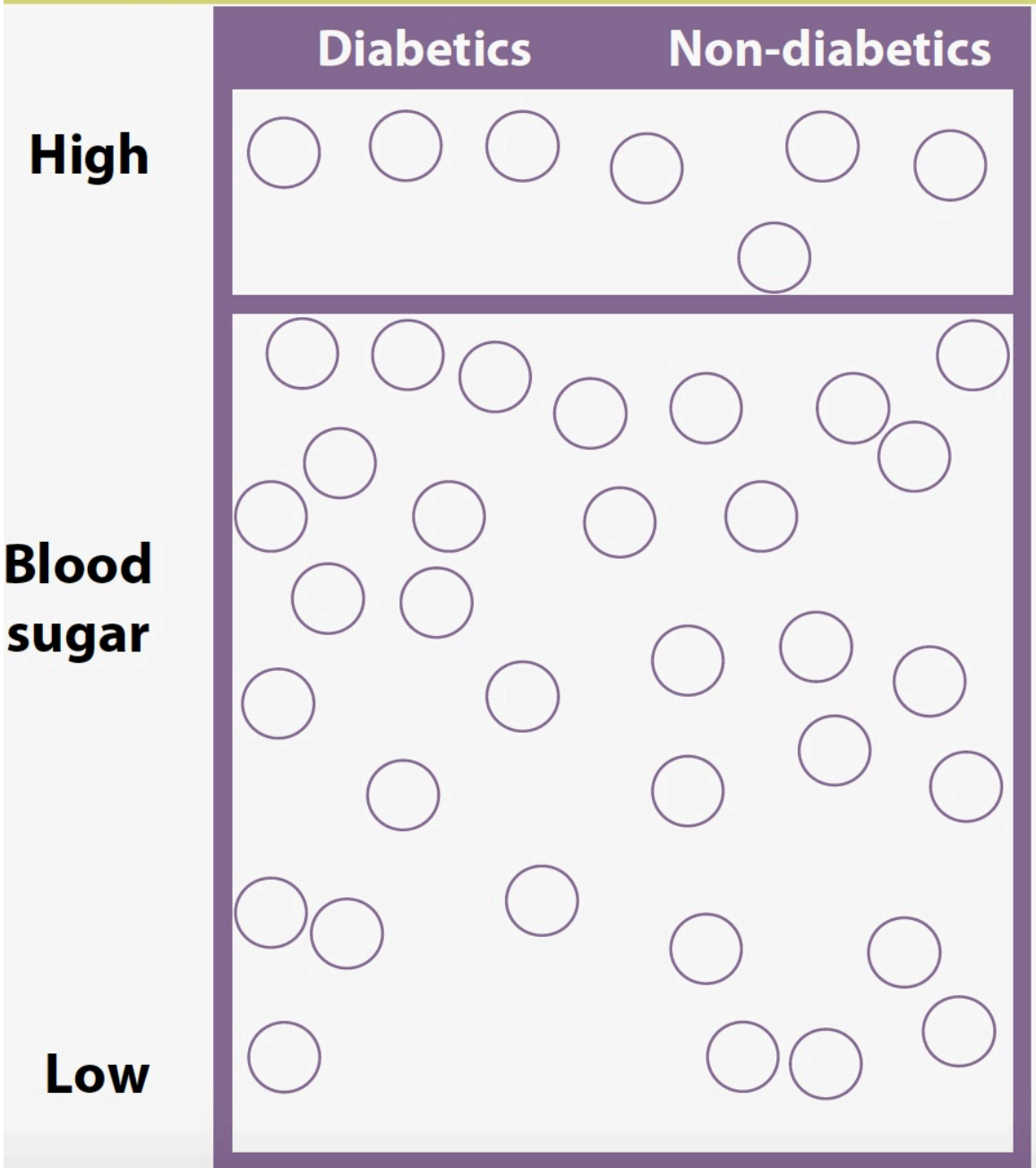
Diagnostic Test and Screening Test

- A **diagnostic test** is used to determine the presence or absence of a disease when a subject shows signs or symptoms of the disease
- A **screening test** identifies asymptomatic individuals who may have the disease
- The diagnostic test is performed **after** a positive screening test to establish a definitive diagnosis

Some Common Screening Tests

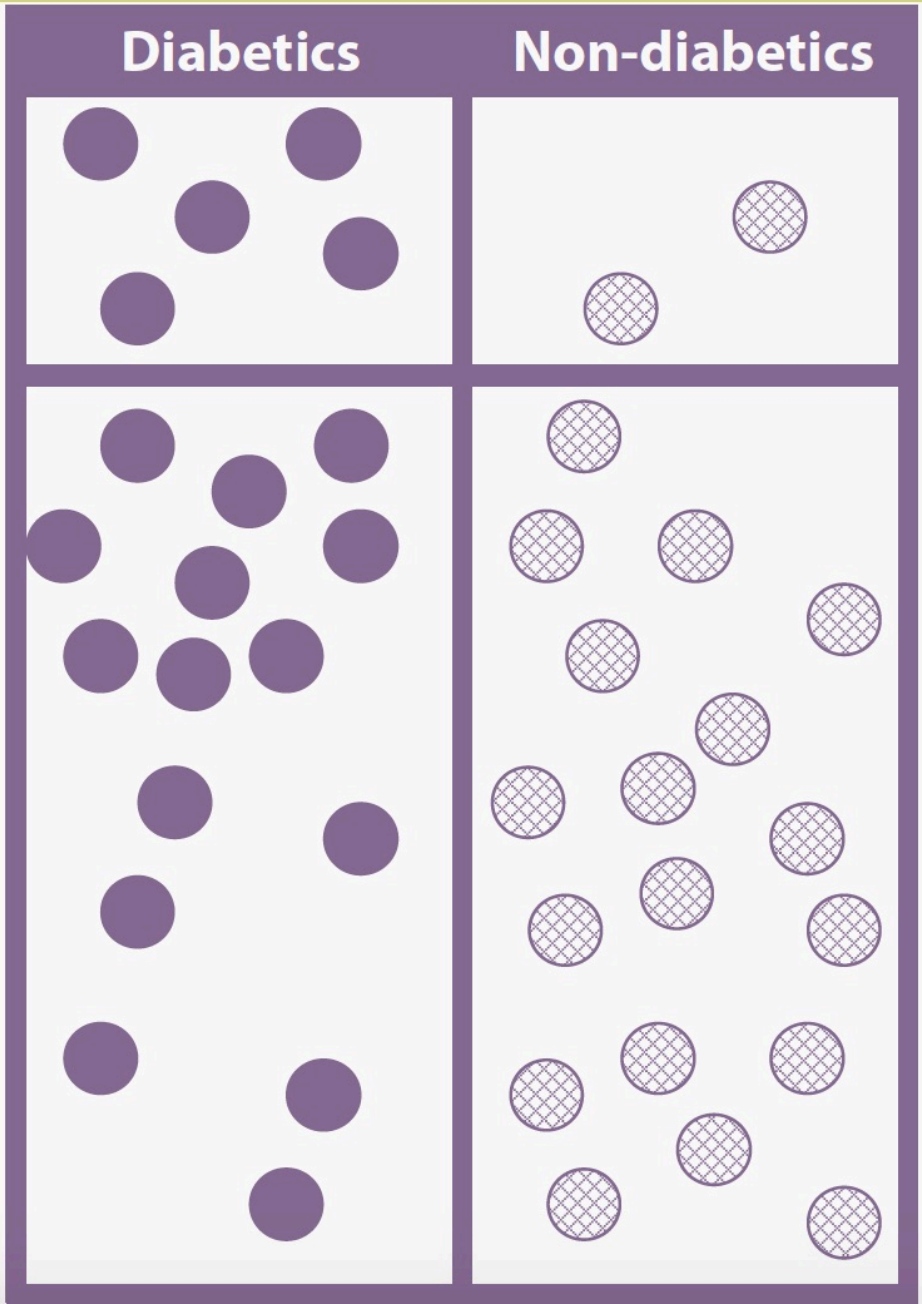
- Pap smear for cervical dysplasia or cervical cancer
- Fasting blood cholesterol for heart disease
- Fasting blood sugar for diabetes
- Blood pressure for hypertension
- Mammography for breast cancer
- PSA test for prostate cancer
- Fecal occult blood for colon cancer
- Ocular pressure for glaucoma
- PKU test for phenolketonuria in newborns
- TSH for hypothyroid and hyperthyroid

Concept of Sensitivity and Specificity



A screening test using a high cut-point will treat the bottom box as normal and will identify the 7 subjects above the line as having diabetes

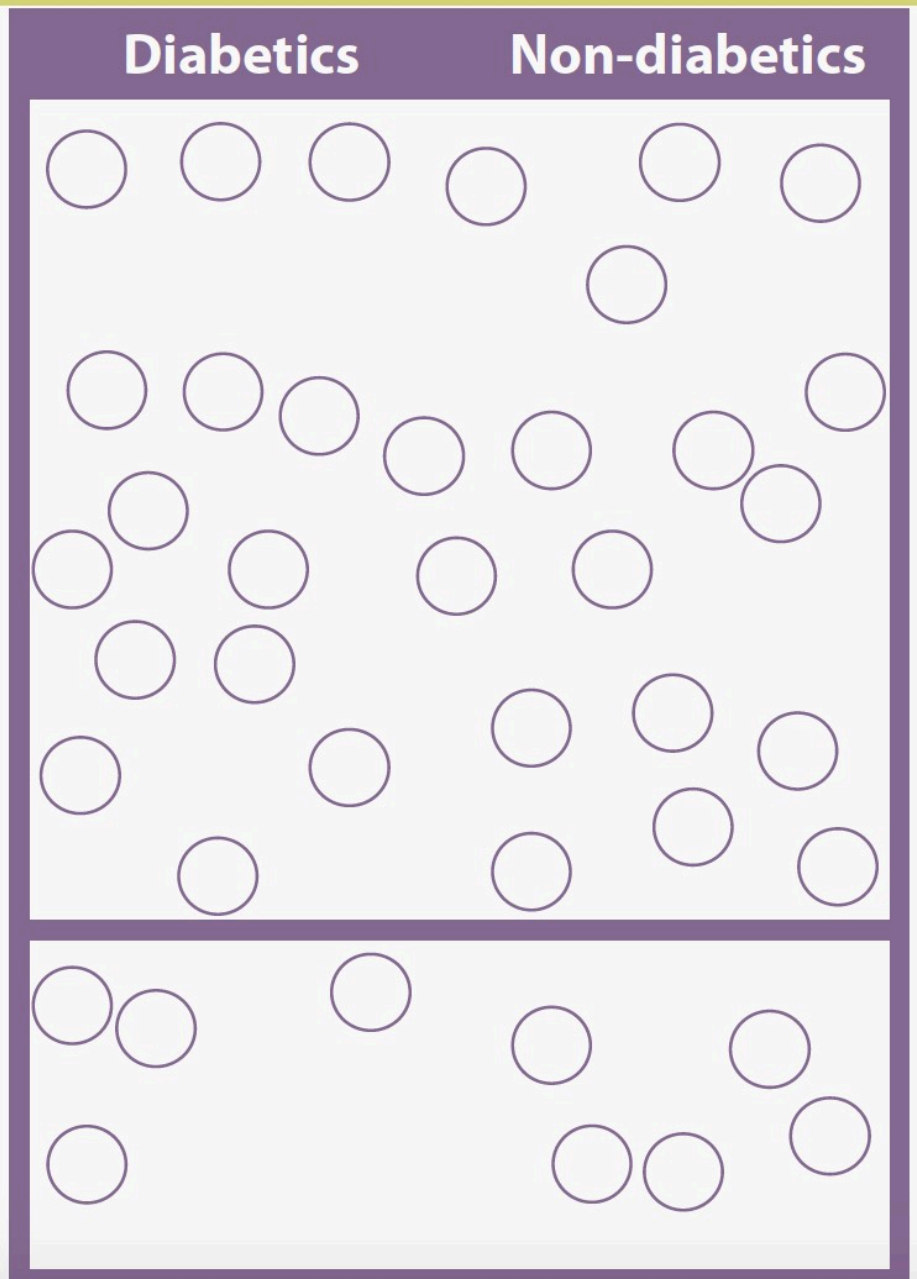
Concept of Sensitivity and Specificity



Subjects are screened using fasting plasma glucose with a high cut-point

	Diabetics	Non-Diabetics
+	5	2
–	15	18
	20	20
	Sens=25%	Spec=90%

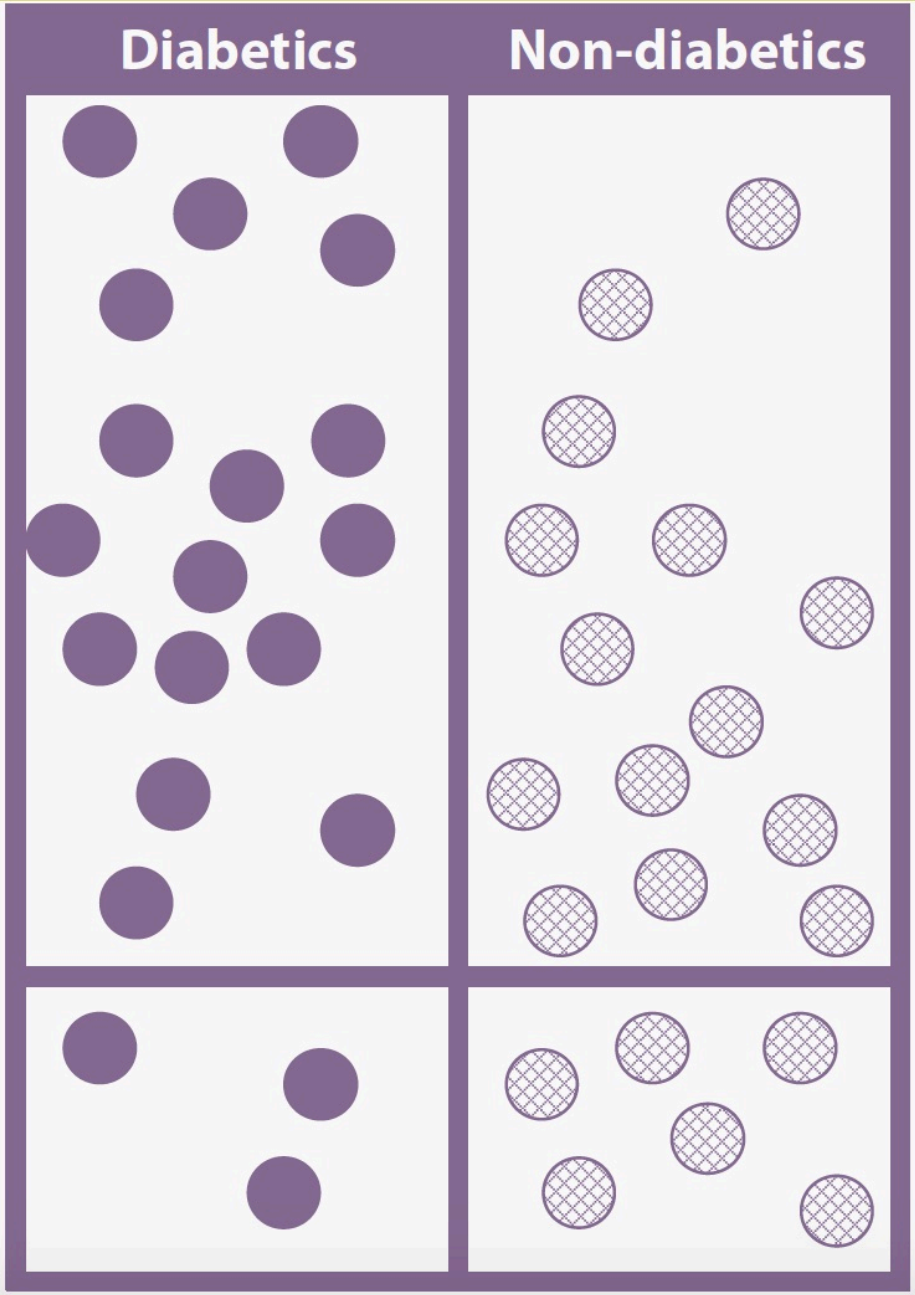
Concept of Sensitivity and Specificity



A screening test using a high cut-point will treat the bottom box as normal and will identify the 7 subjects above the line as having diabetes;

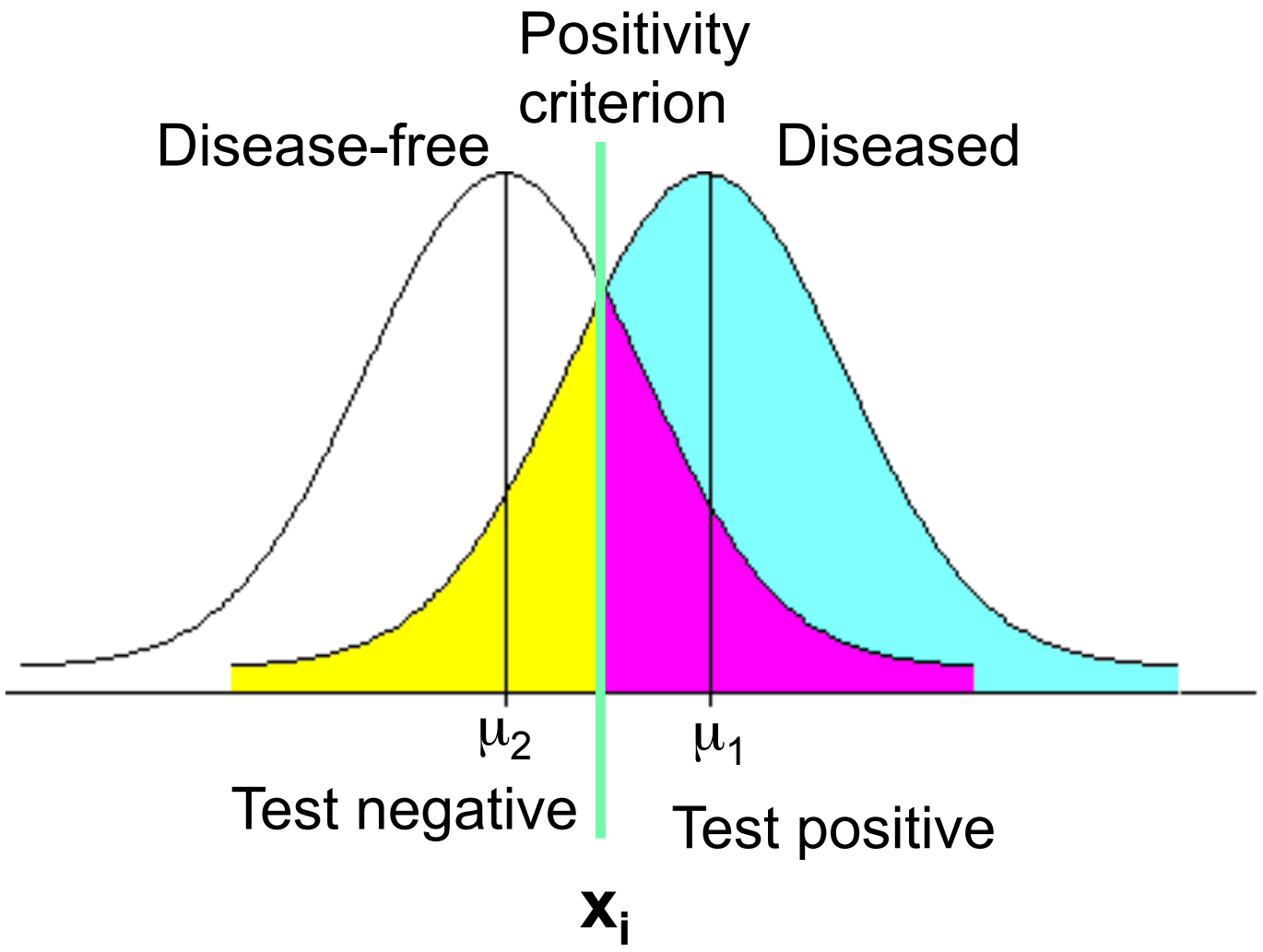
But a low cut-point will result in identifying 31 subjects as having diabetes

Concept of Sensitivity and Specificity



Subjects are screened using fasting plasma glucose with a low (blood sugar) cut-point

	Diabetics	Non-Diabetics
+	17	14
–	3	6
	20	20
	Sens=85%	Spec=30%


Positivity
criterion
Disease-free
Diseased
μ_2
μ_1
Test negative
Test positive
x_i
TP
FN
FP
TN

Lessons Learned

- Different cut-points yield different sensitivities and specificities
- The cut-point determines how many subjects will be considered as having the disease
- The cut-point that identifies more true negatives will also identify more false negatives
- The cut-point that identifies more true positives will also identify more false positives

Where to Draw the Cut-Point

- If the diagnostic (confirmatory) test is expensive or invasive:
  - Minimize false positives
    or
  - Use a cut-point with high specificity
- If the penalty for missing a case is high (e.g., the disease is fatal and treatment exists, or disease easily spreads):
  - Maximize true positives
    - That is, use a cut-point with high sensitivity
- Balance severity of false positives against false negatives

ROC Curves

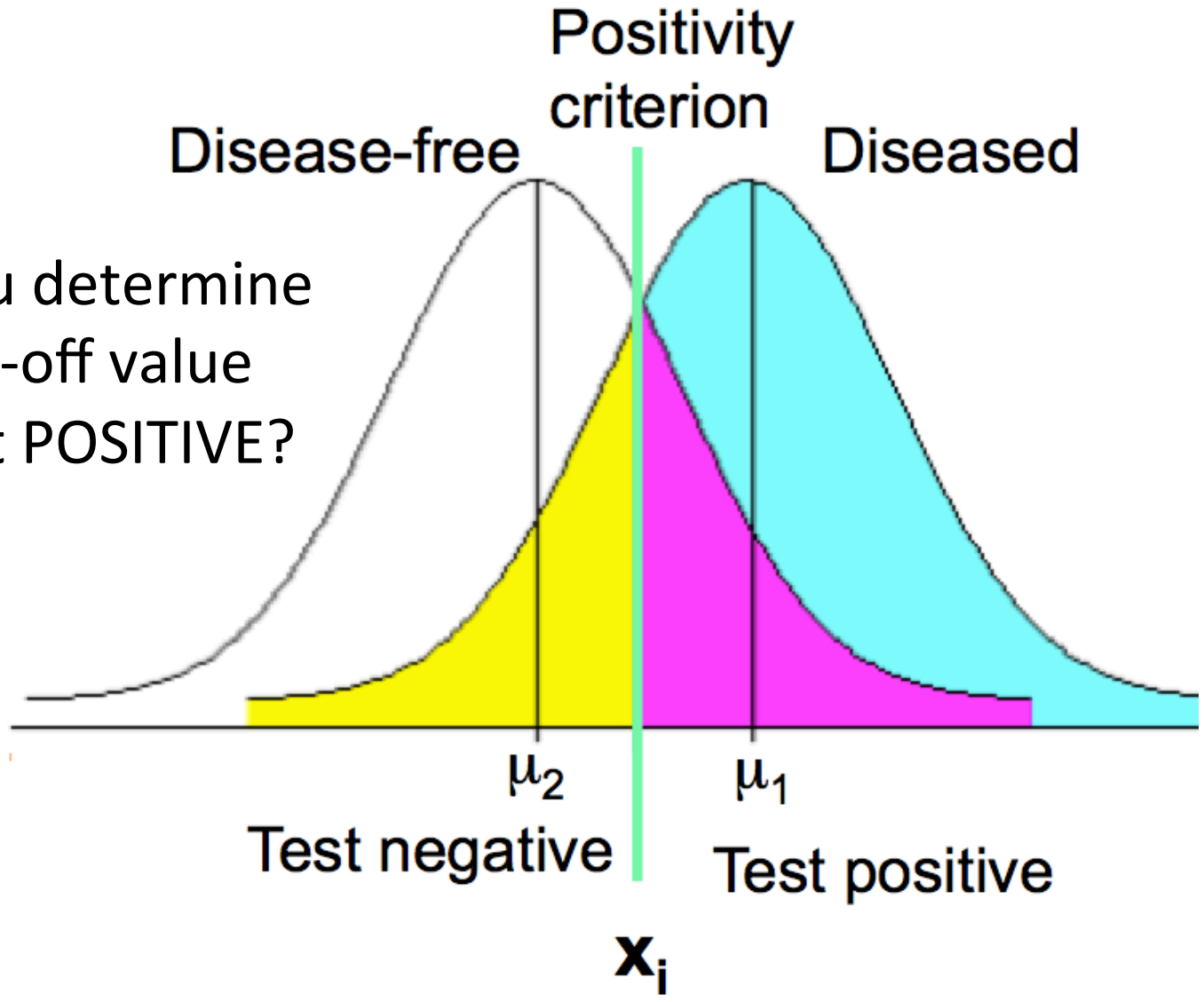
Continuous test → many values Se/Sp

The ROC curves = Graphical plot which illustrates the performance of a binary test as its discrimination threshold varied

It is created by plotting the True Positive Rate (Se) versus the False Positive Rate (1-Sp) at all the various threshold values

ROC analysis provided tools to select optimal models

How can you determine
The best cut-off value
To call a test POSITIVE?



TP FN

FP TN

ROC (Receiver Operating Characteristic) CURVE

The ROC Curve is a graphic representation of the relationship between sensitivity and specificity for a diagnostic test. It provides a simple tool for applying the predictive value method to the choice of a positivity criterion.

ROC Curve is constructed by plotting the true positive rate (sensitivity) against the false positive rate (1-specificty) for several choices of the positivity criterion.

Construction of the ROC curve

- On the horizontal axis,

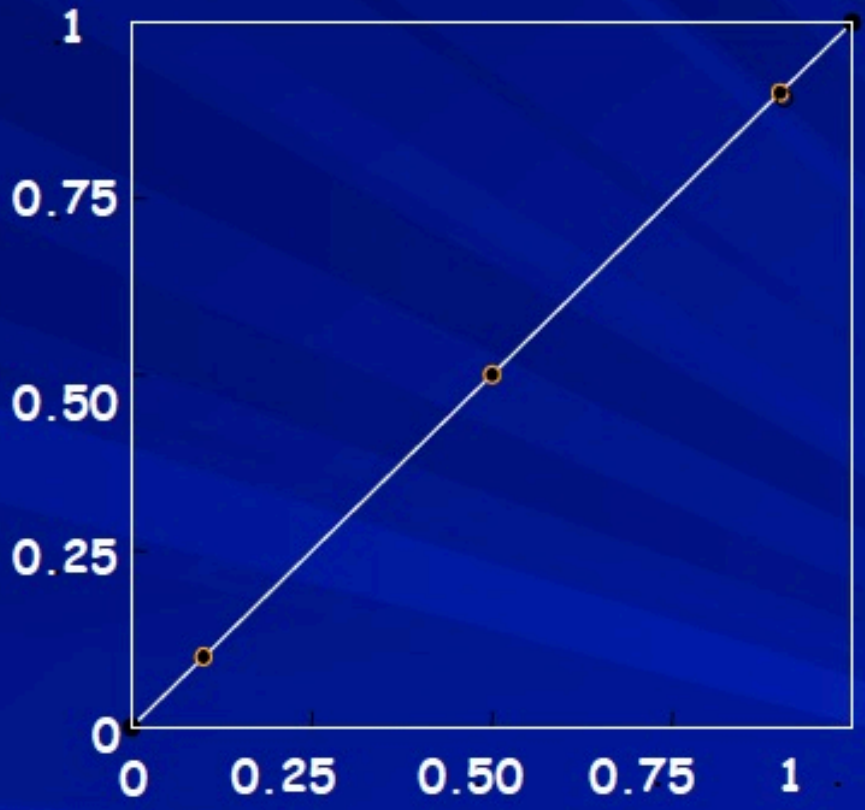
  we plot the value (1-Sp) = False positive among the subjects without the disease

- On the vertical axis

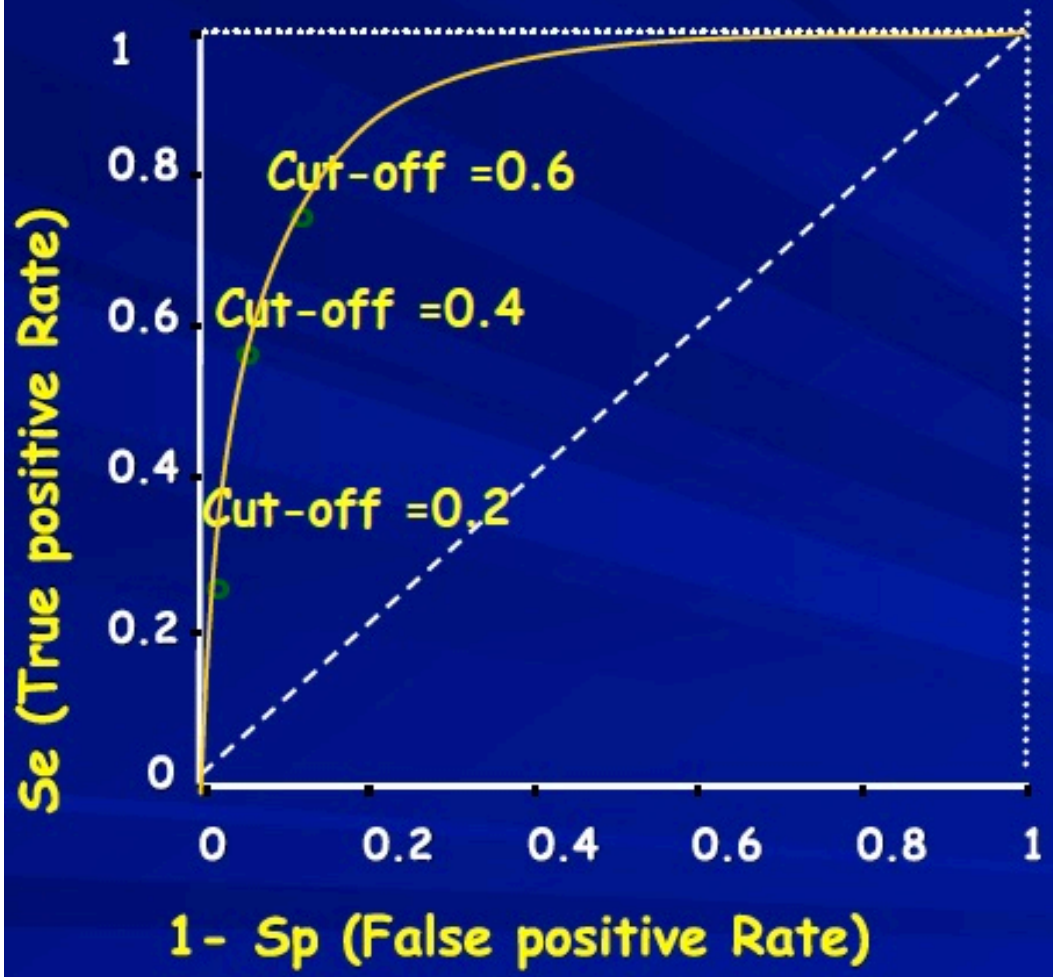
  we plot the value (Se) = True positive among the subjects with the disease

A threshold ideal allows to completely separate the positive and negative cases, with no false positive cases and no false negative cases

Se (True positive rate)



1-Sp (False positive rate)



$p_1 > S$
$p_1 \leq S$

D+	D-
TP	FP
FN	TN

Cut-off 0.2

$p_1 > S$
$p_1 \leq S$

D+	D-
TP	FP
FN	TN

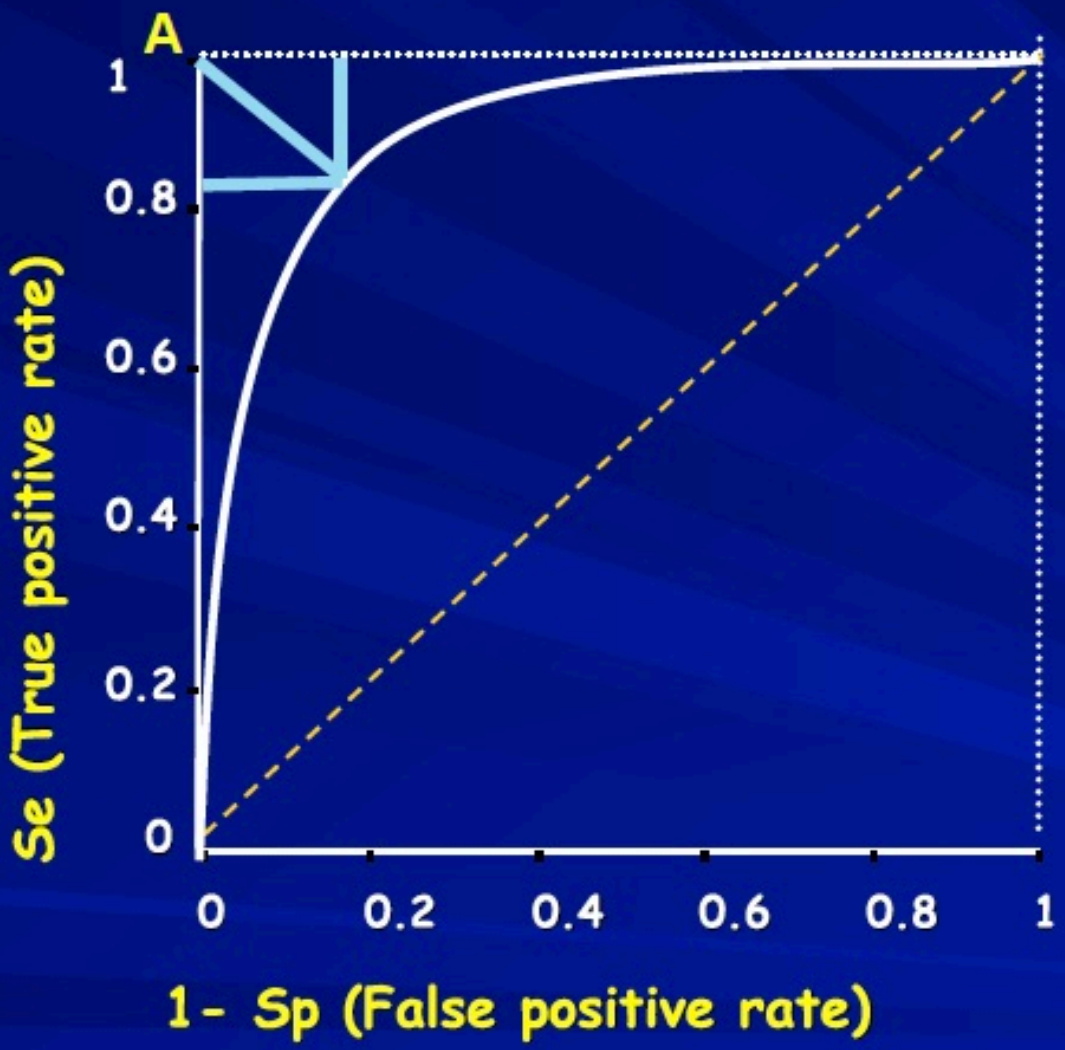
Cut-off 0.4

$p_1 > S$
$p_1 \leq S$

D+	D-
TP	FP
FN	TN

Cut-off 0.6

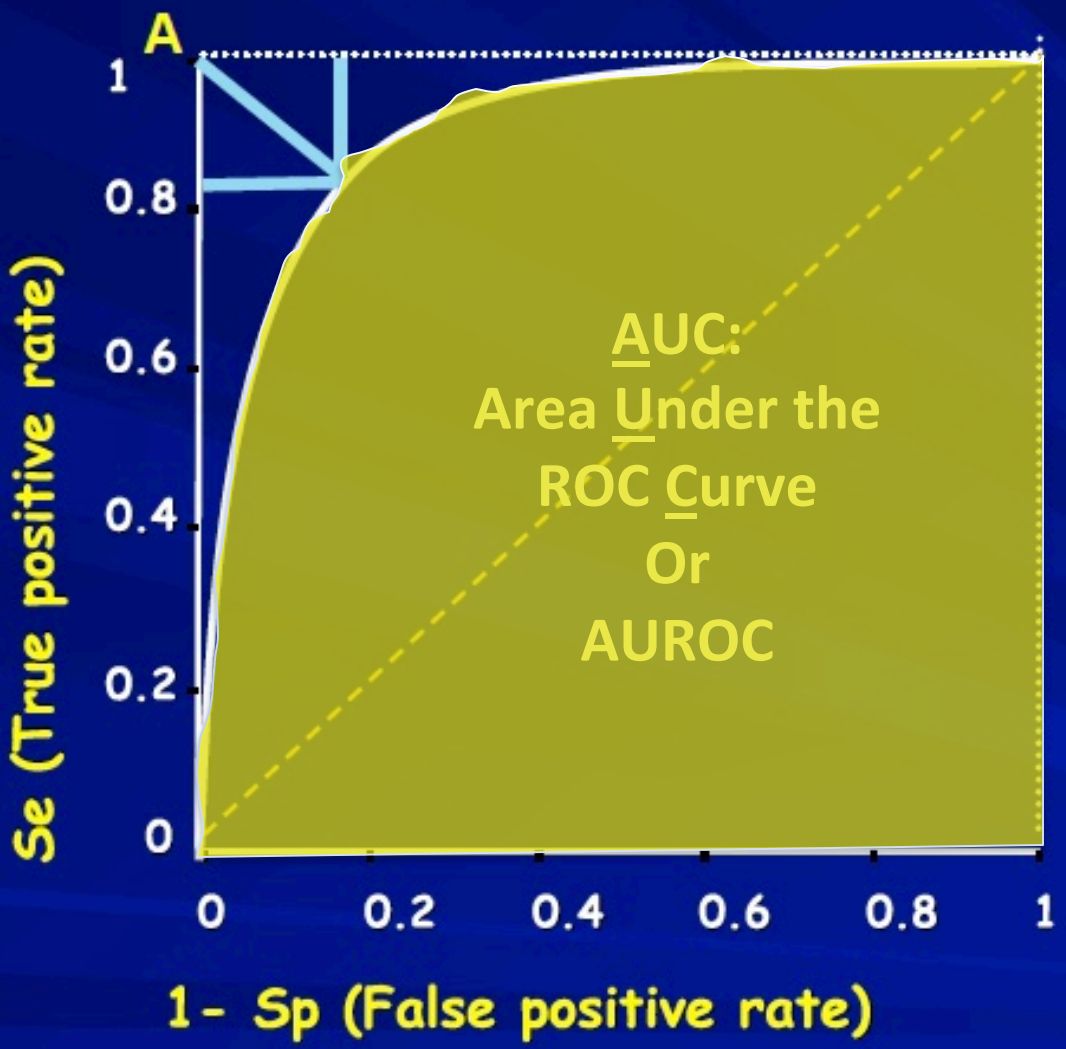
Construction of the ROC curve



Diagonal passing through 0
➔ Uninformative test

➔A test is more accurate when the curve is near A and far from the diagonal

Construction of the ROC curve



Diagonal passing through 0
➔ Uninformative test

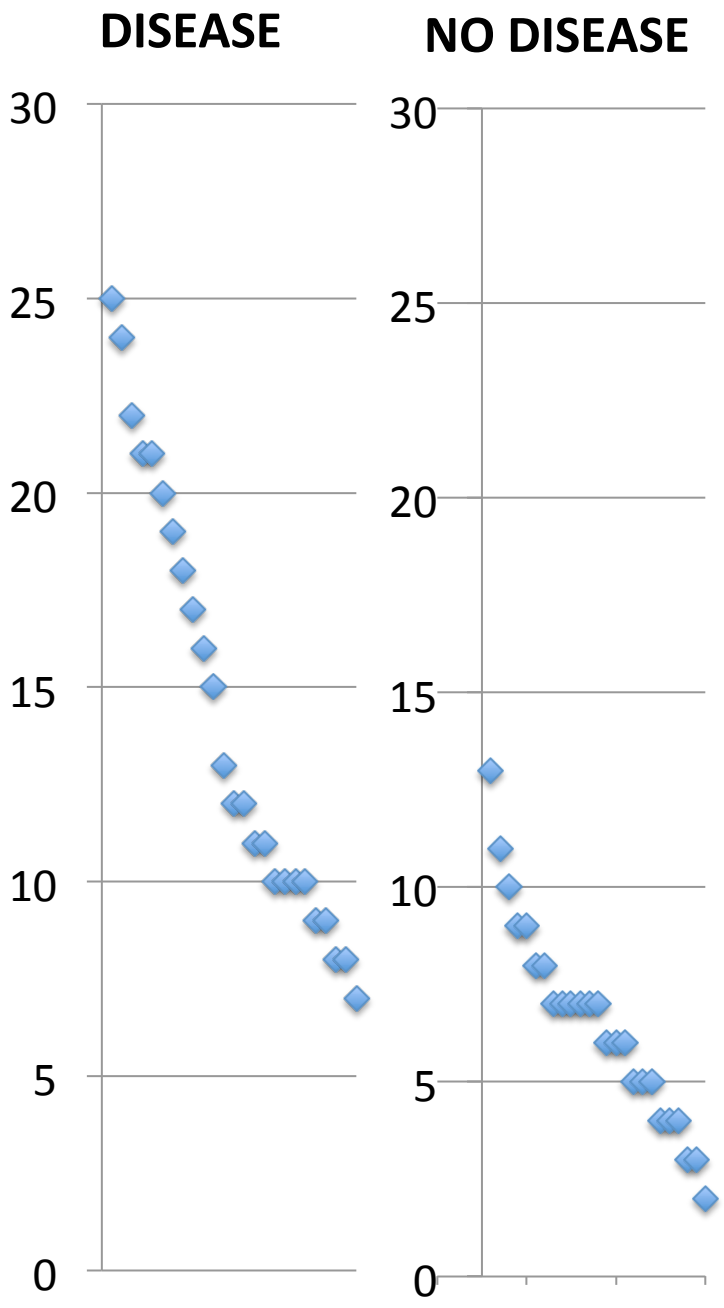
➔A test is more accurate when the curve is near A and far from the diagonal

We calculate the area under the curve
AUROC (95% CI)

Higher the AUROC (→1), more accurate is the test

Advantages of the ROC curve

- Simple and easy to understand
- Takes into account all the values of the test (without arbitrary choice of a cut-off)
- Totally independent of the prevalence of the disease in the sample
- Allow a visual comparison between several tests on the same scale (+ statistical test to compare AUC)
- Calculation of the 95%CI of the AUC. To conclude to a « good » test, the lower limit must not include the value 0.5



An Example of an ROC curve

Results of a Test

to determine if a patient

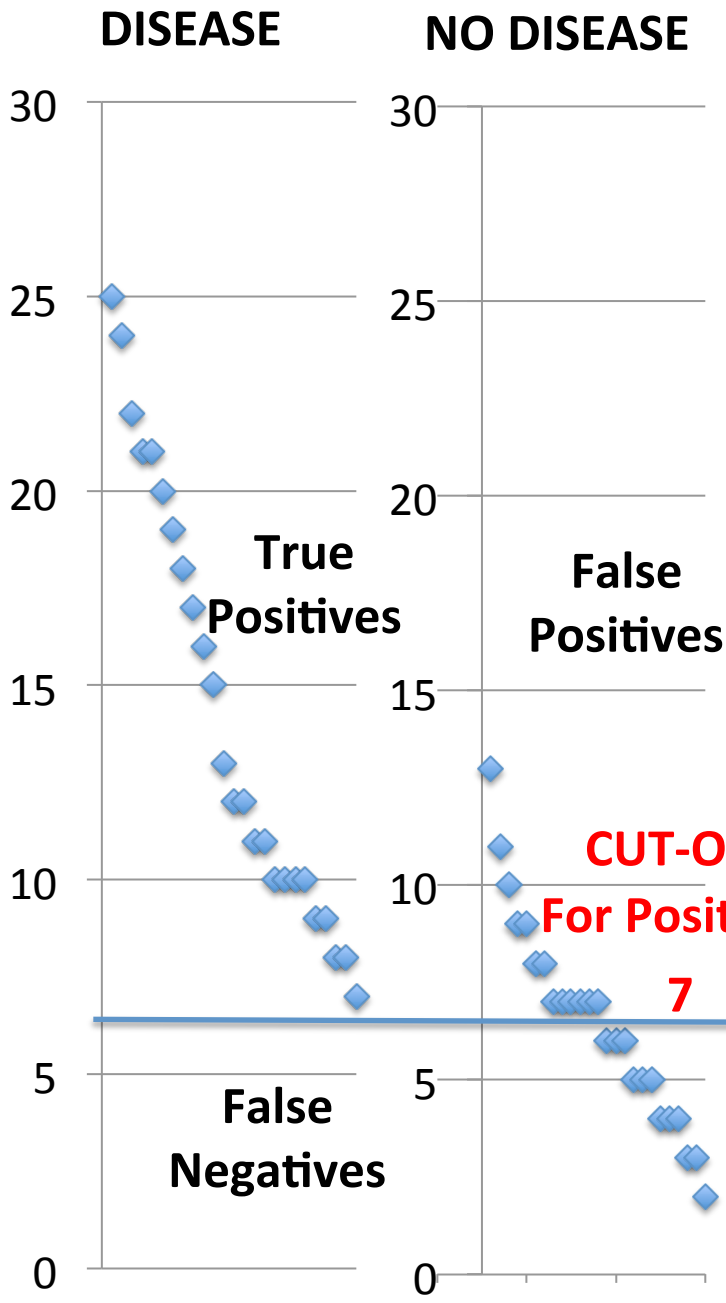
Has Disease X

What is the best cutoff level

To distinguish between

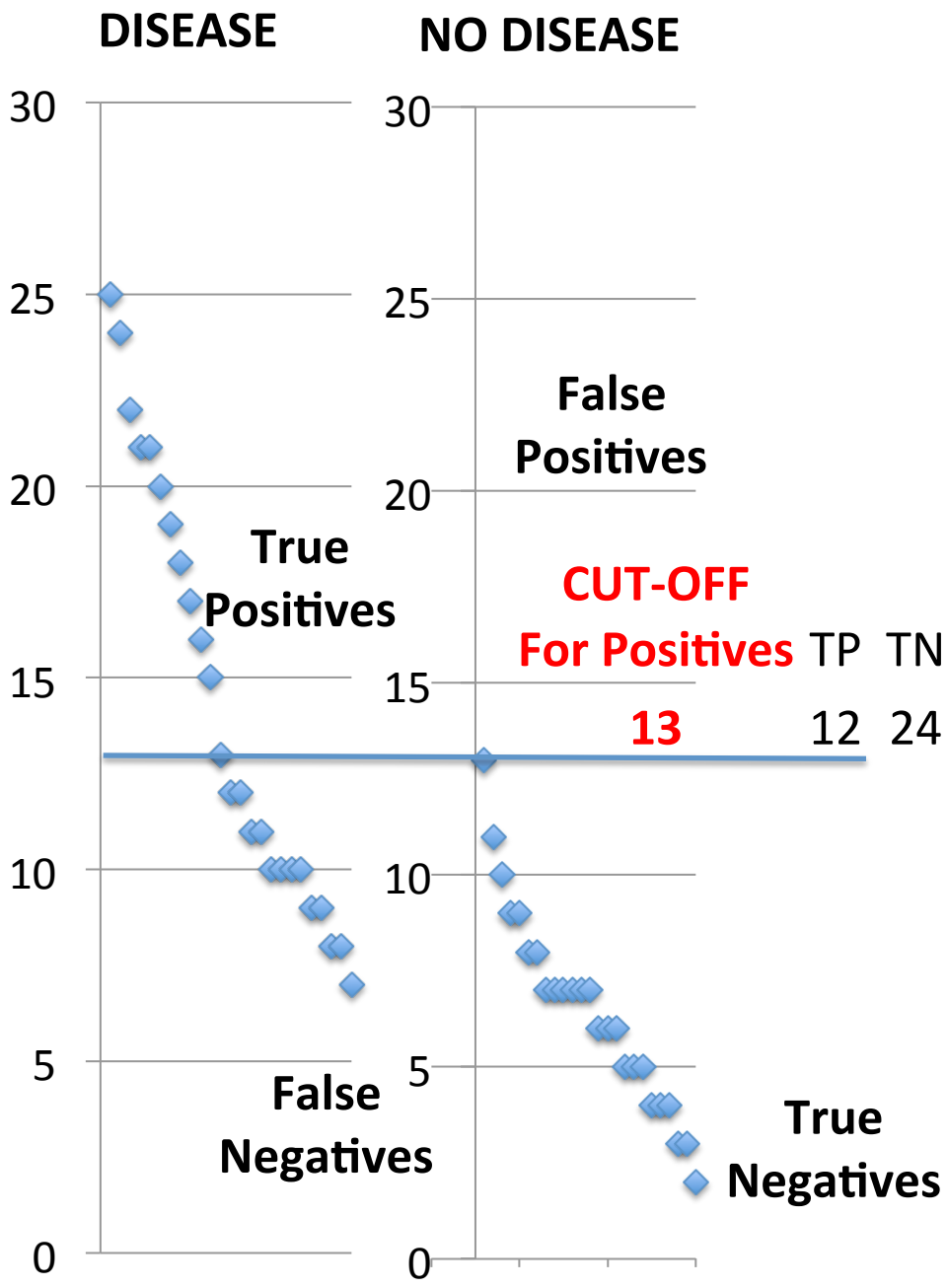
People with disease

From those without disease?



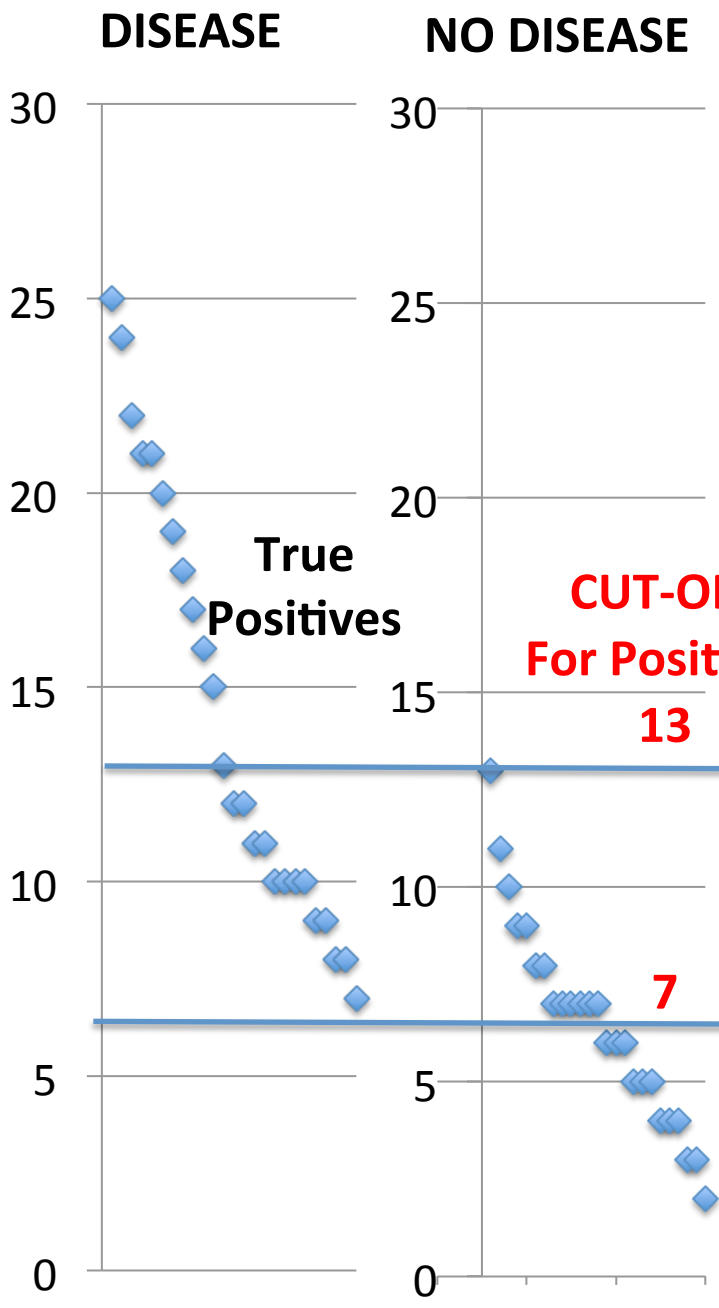
CUT-OFF For Positives 7

TP	TN	FP	FN	Sensitivity	Specificity	1-Specificity
25	12	13	0	1.00	0.48	0.52

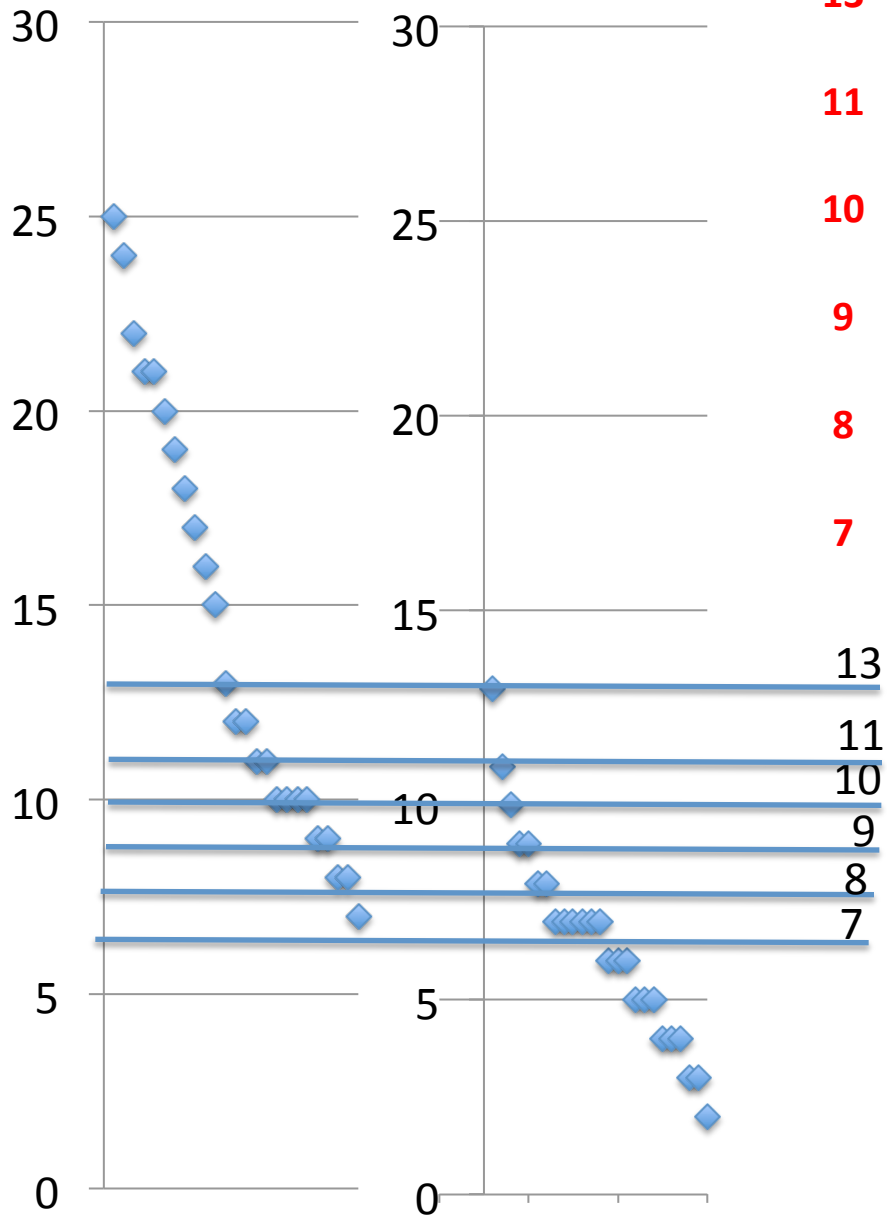


CUT-OFF For Positives

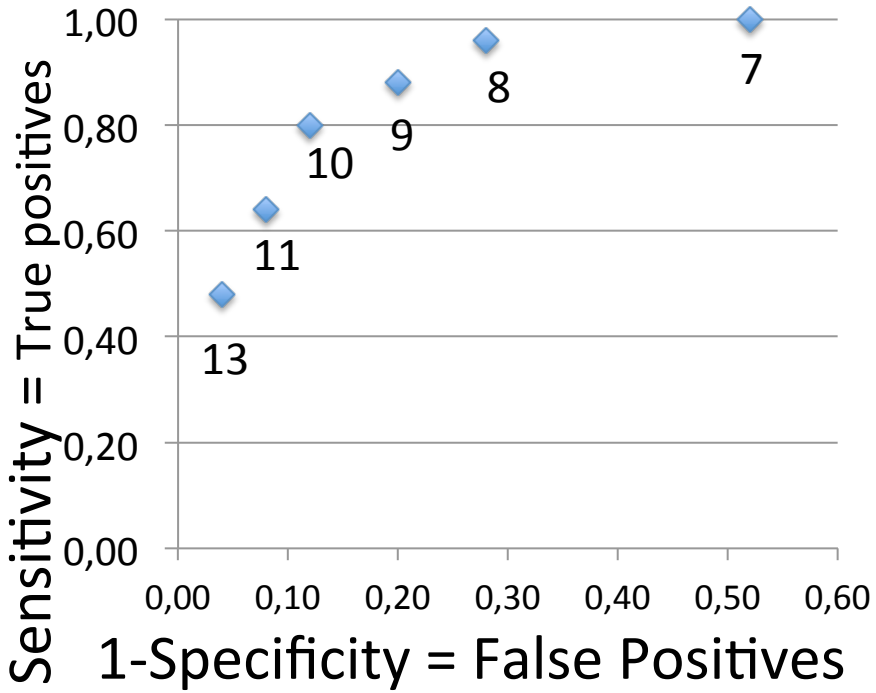
TP	TN	FP	FN	Sensitivity	Specificity	1-Specificity
13	12	24	1	13	0.48	0.96	0.04

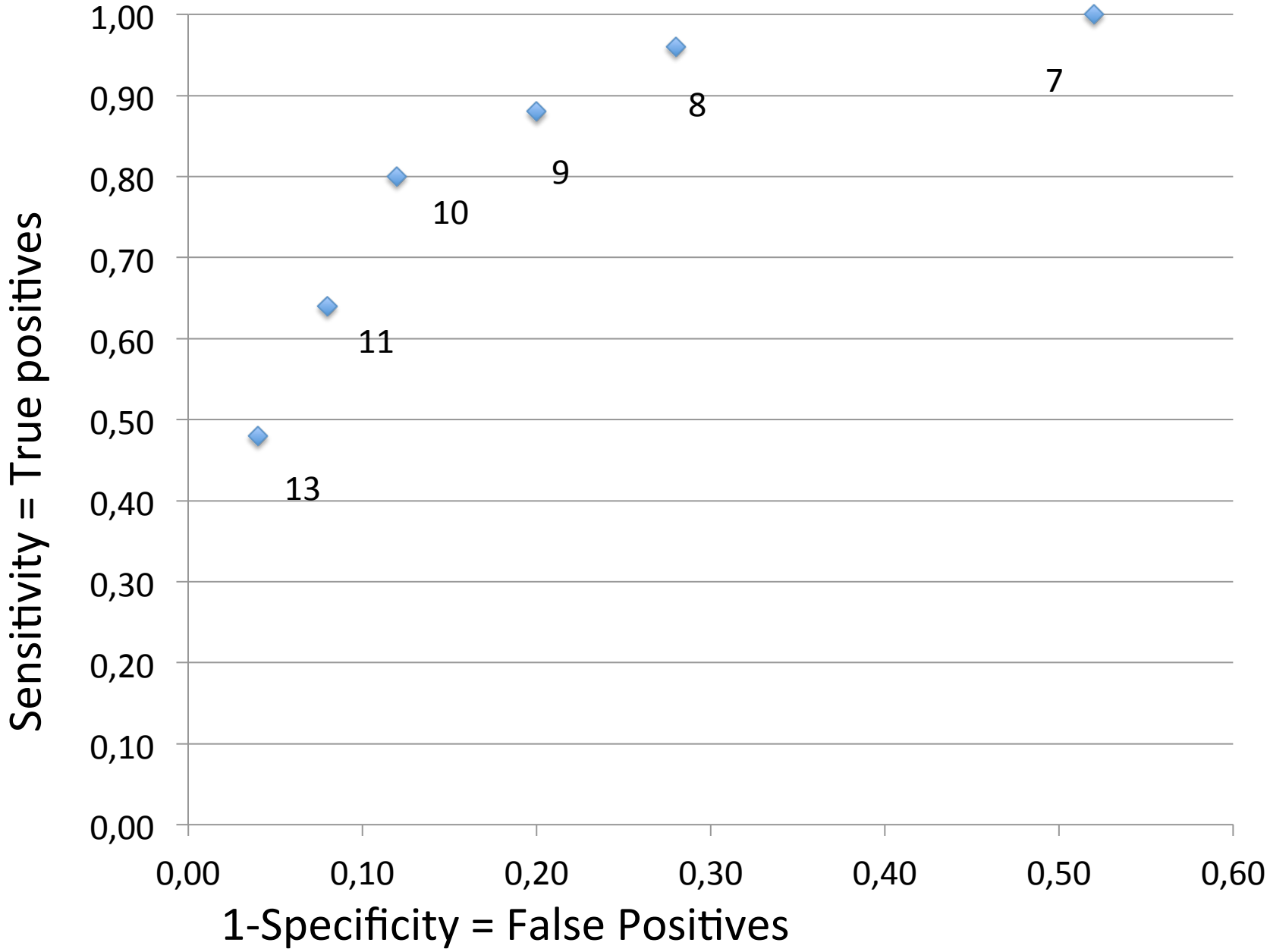


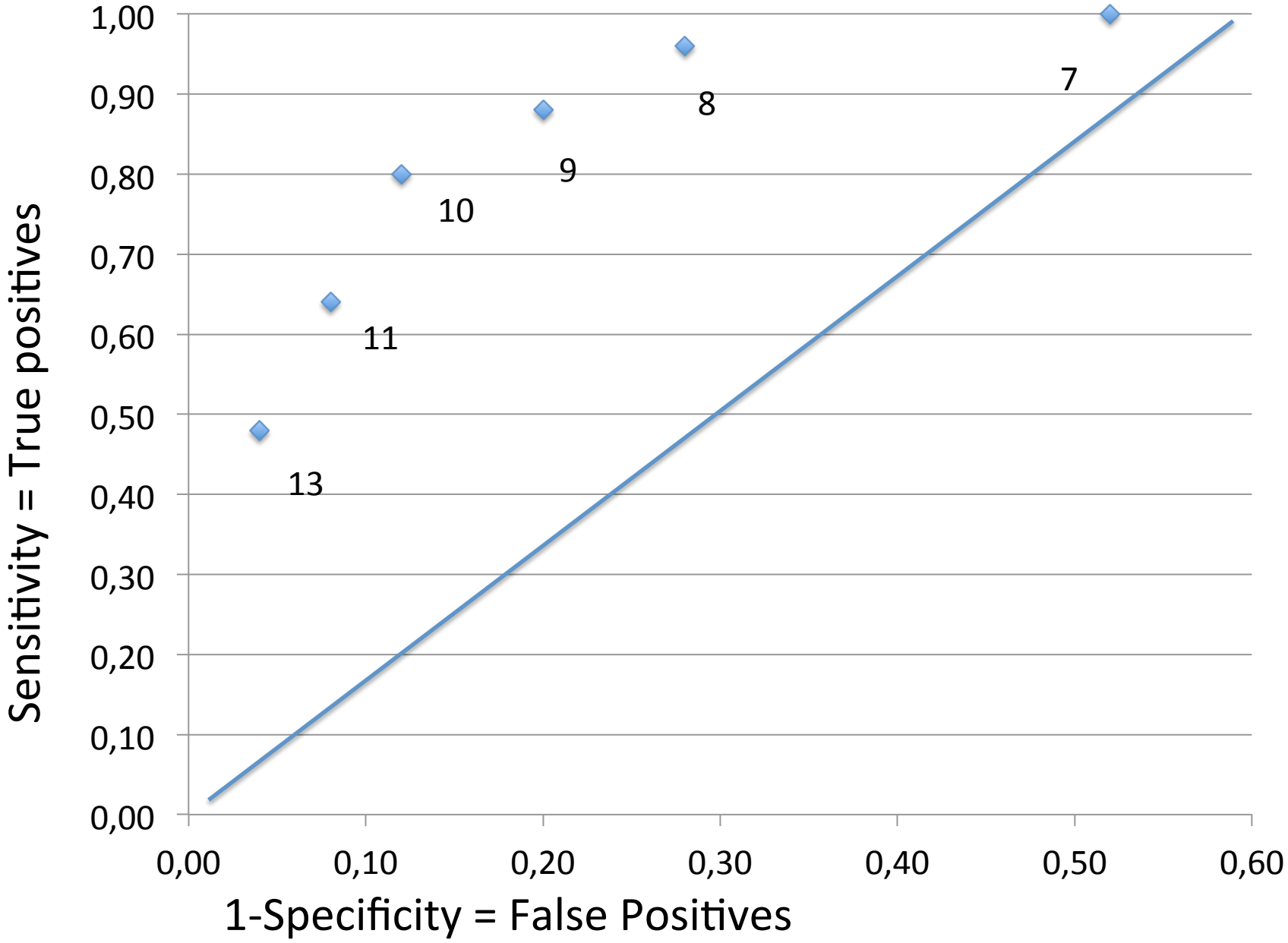
CUT-OFF For Positives	TP	TN	FP	FN	Sensitivity	Specificity	1-Specificity
13	12	24	1	13	0.48	0.96	0.04
7	25	12	13	0	1.00	0.48	0.52

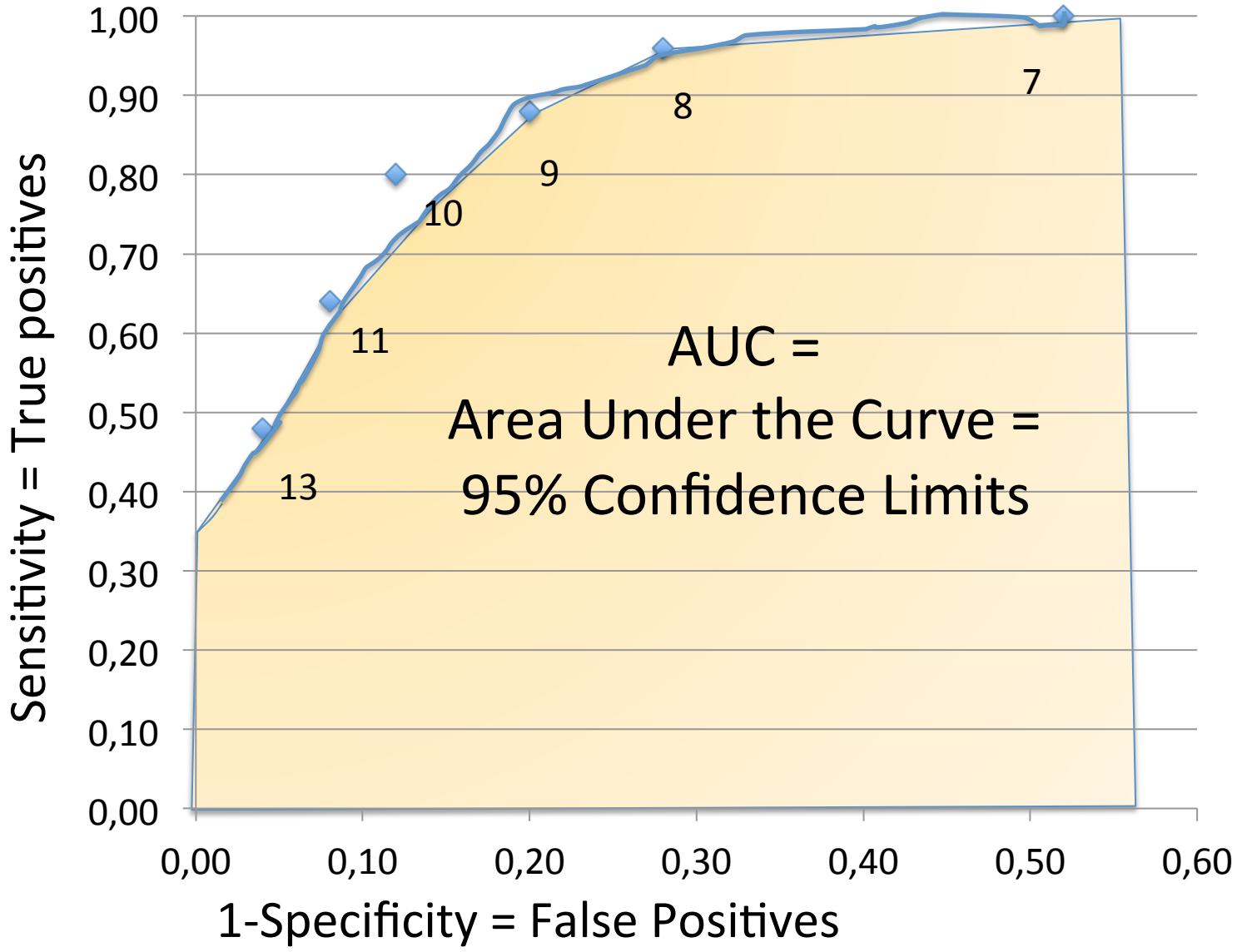


CUT-OFF For Positives	TP	TN	FP	FN	Sensitivity	Specificity	1-Specificity
13	12	24	1	13	0.48	0.96	0.04
11	16	23	2	9	0.64	0.92	0.08
10	20	22	3	5	0.8	0.88	0.12
9	22	20	5	3	0.88	0.80	0.20
8	24	18	7	1	0.96	0.72	0.28
7	25	12	13	0	1.00	0.48	0.52




1,00
0,90
0,80
0,70
0,60
0,50
0,40
0,30
0,20
0,10
0,00
Sensitivity = True positives
0,00
0,10
0,20
0,30
0,40
0,50
0,60
1-Specificity = False Positives
7
8
9
10
11
13


1,00
0,90
0,80
0,70
0,60
0,50
0,40
0,30
0,20
0,10
0,00
Sensitivity = True positives
0,00
0,10
0,20
0,30
0,40
0,50
0,60
1-Specificity = False Positives
7
8
9
10
11
13


1,00
0,90
0,80
0,70
0,60
0,50
0,40
0,30
0,20
0,10
0,00
Sensitivity = True positives
0,00
0,10
0,20
0,30
0,40
0,50
0,60
1-Specificity = False Positives
7
8
9
10
11
13
AUC =
Area Under the Curve =
95% Confidence Limits

How to evaluate the agreement of two tests:

Cohen's kappa coefficient

- It measures the agreement between two different tests for determining the same thing – for example a diagnosis
- Or the agreement of two people rating the same item
  -for example two people evaluating the same application
- It is a statistical measure that takes into account the agreement that would occur by chance between the two tests or the two raters.

Calculation

Cohen's kappa measures the agreement between two raters who each classify N items into C mutually exclusive categories. The first mention of a kappa-like statistic is attributed to Galton (1892),[1] see Smeeton (1985).[2]

The equation for κ is:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

where p_o is the relative observed agreement among raters, and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters other than what would be expected by chance (as given by p_e), $\kappa \leq 0$.

To measure the true agreement of two tests or two evaluators
you calculate the observed agreement and then
subtract the agreement that would be expected purely by chance

Suppose that you were analyzing data related to a group of 50 people applying for a grant. Each grant proposal was read by two readers and each reader either said "Yes" or "No" to the proposal. Suppose the dis/agreement count data were as follows, where A and B are readers, data on the main diagonal of the matrix (top left-bottom right) the count of agreements and the data off the main diagonal, disagreements:

		B	
		Yes	No
A	Yes	20	5
	No	10	15

Note that there were 20 proposals that were granted by both reader A and reader B, and 15 proposals that were rejected by both readers. Thus, the observed proportionate agreement is $p_o = (20 + 15) / 50 = 0.70$

		B	
		Yes	No
A	Yes	20	5
	No	10	15

To calculate p_e (the probability of random agreement) we note that:

- Reader A said "Yes" to 25 applicants and "No" to 25 applicants. Thus reader A said "Yes" 50% of the time.
- Reader B said "Yes" to 30 applicants and "No" to 20 applicants. Thus reader B said "Yes" 60% of the time.

Therefore the probability that both of them would say "Yes" randomly is $0.50 \cdot 0.60 = 0.30$ and the probability that both of them would say "No" is $0.50 \cdot 0.40 = 0.20$. Thus the overall probability of random agreement is $Pr(e) = 0.3 + 0.2 = 0.5$.

So now applying our formula for Cohen's Kappa we get:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{0.70 - 0.50}{1 - 0.50} = 0.40$$

Same percentages but different numbers

A case sometimes considered to be a problem with Cohen's Kappa occurs when comparing the Kappa calculated for two pairs of raters with the two raters in each pair having the same percentage agreement but one pair give a similar number of ratings while the other pair give a very different number of ratings.[5] For instance, in the following two cases there is equal agreement between A and B (60 out of 100 in both cases) so we would expect the relative values of Cohen's Kappa to reflect this. However, calculating Cohen's Kappa for each:

		B	
		Yes	No
A	Yes	45	15
	No	25	15

Po = 45/100 + 15/100 = 0.6 = observed agreement

Pe = ((45 + 15)/100) x (45 + 25)/100) + ((25+15)/100 x (15 + 15)/100)

= (.6 x .7) + (.4 x .3) = .42 + .12 = .54 = agreement expected by chance

$$\kappa = \frac{0.60 - 0.54}{1 - 0.54} = 0.1304$$

		B	
		Yes	No
A	Yes	25	35
	No	5	35

Po = 25/100 + 35/100 = 0.6

Pe = (.3 x .6) + (.4 x .7) = .18 + .28 = .46

$$\kappa = \frac{0.60 - 0.46}{1 - 0.46} = 0.2593$$

There is greater similarity between A &B
In the second case because
While the percentage agreement (0.6)
is the same, the percentage agreement
That would occur « by chance » is
Is significantly higher in the first case
(0.54 compared to 0.46)

This is the formula for the Standard Error of the Kappa value
In order to calculate confidence intervals.

$$SE_{kappa} = \sqrt{\frac{P_o(1-P_o)}{n(1-P_e)(1-P_e)}} = 0.07$$

$$IC_{95\%} = Kappa \pm 1.96 \times SE_{kappa}$$

$$IC_{95\%} = 0.8 \pm 1.96 \times 0.07 = 0.8 \pm 0.14$$

Cohen's kappa coefficient

- It expresses a relative difference between the proportion of observed agreement **Po** and the proportion of chance agreement **Pe**
- → **K** is a percentage of the maximum agreement corrected by the chance agreement
- **Kappa** is a value between **-1 and +1 (1= complete agreement and 0 = no agreement)**
- Kappa can be used to evaluate **the reproducibility** and **the validity of a test** when the test and the Gold standard have the same number of categories.

Concordance and Kappa
« Reference values » (Landis-Koch, 1977)

Kappa coefficient is used as a descriptive indicator of concordance

κ Value	Strength of Agreement beyond Chance
<0	Poor
0–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

Thank you for your attention

- Questions?

- There is a handout with the basic formulas