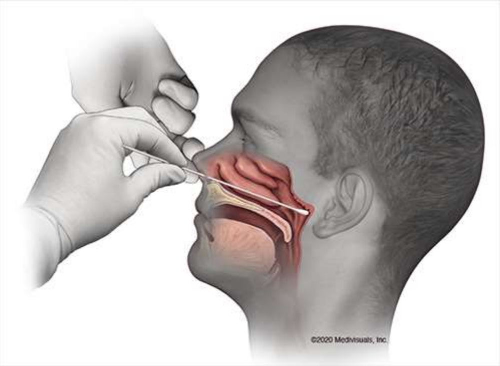
UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

Biostatistics

Diagnostic Process Evaluation: sensitivity, specificity & predictive values

Paola Rebora

A diagnostic test ...



... is any procedure ...
for identifying the state of a disease:

Examples: Glycemia for diabetes,
Viral test (i.e. swab) for SARS-CoV-2
Proteinuria for Kidney Disease
…

a threshold value is defined to classify patients as negative/positive.

NEGATIVE: The outcome of a test is negative if it leads to the exclusion of the presence of the disease

POSITIVE: The outcome of a test is positive if it raises suspicions of the presence of the disease

A useful diagnostic test...

… tends to provide positive results in subjects who have the diseases
…tends to give negative results in subjects who do not have the disease

Remarks
When the doctor examines the results of a screening test he does not know whether the patient is healthy or diseased, but would like that:

a positive result means diseased;
a negative result means healthy

THIS IS NOT ALWAYS TRUE

Example – breast cancer (BC) screening

When a woman turns 40 the gynecologist usually reminds her that it is time to begin to get a mammogram every 2 years.

On the advice of the doctor a woman 40 years old gets the first mammogram:
the result is positive…

This means absolute certainty of breast cancer?

Or it means probability of 99%, of 95, 90, 50%,
or even lower?

Validity/accuracy of the diagnostic procedure

An independent comparison is required to analyze the validity of a diagnostic procedure.

GOLD STANDARD
A diagnostic procedure of reference that classifies individuals with respect to their true state of illness: sick (D +) or healthy (D-).

Examples:
Tumors: histological examination
Diabetes: blood glucose analysis
PCR based viral test for SARS-CoV-2

Basics

Health Status (binary outcome)

Test results (binary outcome)

	Health Status	
Test results	DISEASE (D+)	NON-DISEASE (D-)
POSITIVE (T+)	TP: true positive	FP: false positive
NEGATIVE (T-)	FN: false negative	TN: true negative

FN: false negative imply missed cases, so potentially bad outcomes if untreated.

FP: entail the cost and danger of further investigations, worry for the patient.

Sensitivity: concerns only diseased subjcts (D^+):

	D+	**D-**	**total**
T+	**a**	b	a+b
T-	**c**	d	c+d
total	**a+c**	b+d	N=a+b+c+d

Sensitivity (Sn), also called True Positive Fraction:

is the probability that test is positive (T+) in diseased subjects (D+)

$$P(T+|D+)= \frac{true\ positive}{diseased\ subjects} = \frac{a}{a+c}$$

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

Note on basics concepts of Probability:
Conditional Probability

	D+	**D-**	**total**
T+	TP a	FP b	a+b
T-	FN c	TN d	c+d
total	a+c	b+d	N=a+b+c+d

First set the primary information or **condition of B** (eg. health status: D+), then define the component of **interest A** (test result: T+)

INTUITIVE APPROACH:

Sn (true positive fraction) = Pr(T+|D+) = a/(a+c)

FORMAL APPROACH: P(T+|D+)= $\frac{P(T+ \, and \, D+)}{P(D+)} = \frac{\frac{a}{N}}{\frac{a+c}{N}} = a/(a+c)$

Specificity: concerns only subjcts without (D^-):

	D+	**D-**	**total**
T+	**a**	b	a+b
T-	**c**	d	c+d
total	**a+c**	b+d	N=a+b+c+d

Specificity (Sp):

is the probability that test is negative (T-) in diseased-free subjects (D-)

$$P(T-|D-)= \frac{true\ negative}{non\ diseased\ subjects} = \frac{d}{b+d}$$

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)

Specifiity is the complement to 1 of the False Positive Fraction

Example Breast Cancer screening

The probability that a 40 years old woman has breast cancer is about 1%. If she has BC, the likelihood that a mammogram gives a positive result is 90%, and if she does not have it, the probability that the results of the test is still positive is 9%.

Think of 1000 women:

Example Breast Cancer screening

The probability that a 40 years old woman has breast cancer is about 1%. If she has BC, the likelihood that a mammogram gives a positive result is 90%, and if she does not have it, the probability that the results of the test is still positive is 9%.

Think about 1000 women:

	D+	**D-**	**total**
T+			
T-			
total	10	990	1000

Example Breast Cancer screening

The probability that a 40 years old woman has breast cancer is about 1%. If she has BC, the likelihood that a mammogram gives a positive result is 90%, and if she does not have it, the probability that the results of the test is still positive is 9%.

Think of 1000 women:

	D+	D-	total
T+	9		
T-	1		
total	10	990	1000

Example Breast Cancer screening

The probability that a 40 years old woman has breast cancer is about 1%. If she has BC, the likelihood that a mammogram gives a positive result is 90%, and if she does not have it, the probability that the results of the test is still positive is 9%.

Think of 1000 women:

	D+	**D-**	**total**
T+	9	89	
T-	1	901	
total	10	990	1000

Example Breast Cancer screening

The probability that a 40 years old woman has breast cancer is about 1%. If she has BC, the likelihood that a mammogram gives a positive result is 90%, and if she does not have it, the probability that the results of the test is still positive is 9%.

Think of 1000 women:

	D+	**D-**	**total**
T+	9	89	98
T-	1	901	902
total	10	990	1000

Breast cancer example: sensitivity

	D+	**D-**	**total**
T+	a=9	b=89	a+b=98
T-	c=1	d=901	c+d=902
total	a+c = 10	b+d = 990	N=1000

Among those **with disease**, what proportion does the test detect?

$$\mathbf{Sn} = \Pr(T+|D+) = a/(a+c) = 9/10 = 90\%$$

Interpretation: we have a test with 90% sensitivity.

Among those with breast cancer, the test captures 90% of all cases.

Breast cancer example: specificity

	D+	D-	total
T+	a=9	b=89	a+b=98
T-	c=1	d=901	c+d=902
total	a+c = 10	b+d = 990	N=1000

Among those **without disease**, what proportion does the test correctly identify as disease free?

Sp = Pr(T-|D-) = d/(b+d) = 901/990= 91%

Interpretation: test classifies 91% of women without breast cancer.

Sensitivity and Specificity

- are internal and specific characteristics of a diagnostic test, since each refers to an homogeneous set (diseased or not diseased) predefined;
- are experimentally measurable characteristics (as the relative frequency of positive or negative outcomes) on samples of patients suffering from the disease or healthy subjects
- are between 0 and 1: indeed, they express probability values
- generally, for a single screening test are not simultaneously equal to 1: that is to say that screening tests do not provide certainty

Example:

diagnosis of death rigor mortis is a very specific symptom no alive presents it!

$$P(T+|D-)=0 \text{ thus } P(T-|D-)=1$$

However it is not present in the dead just after death or after long time from death

$$P(T+|D+)<1$$

Flat Electroencephalography (EEG) is a very sensitive symptom all the dead have flat EEG!

$$P(T+|D+)=1$$

However, the EEG may occur transiently plate in subjects in deep coma

$$P(T+|D-)>0 \text{ thus } P(T-|D-)<1$$

Example Breast Cancer screening

The probability that a 40 years old woman has breast cancer is about 1%. If she has BC, the likelihood that a mammogram gives a positive result is 90%, and if she does not have it, the probability that the results of the test is still positive is 9%.

Think of 1000 women:

	D+	**D-**	**total**
T+	9	89	98
T-	1	901	902
total	10	990	1000

How likely is it that a woman 40 years old with a positive mammogram actually has breast cancer?

Positive predictive value:

	D+	**D-**	**total**
T+	**a**	b	a+b
T-	**c**	d	c+d
total	**a+c**	b+d	N=a+b+c+d

… of a positive test result (PPV):

probability that a positive subjects (T+) has disease (D+)

$$P(D+|T+)= \frac{true\ positive}{positive\ subjetcs} = \frac{a}{a+b}$$

The proportion of patients who test positive who actually have the disease

Breast cancer example

	D+	**D-**	**total**
T+	a=9	b=89	a+b=98
T-	c=1	d=901	c+d=902
total	a+c = 10	b+d = 990	N=1000

Among those who have **positive test**, what proportion actually has the disease?

PPV = Pr(D+|T+) = a/(a+b) = 9/98 = 9%

Interpretation: about 90% of women screening positive falsely.

Negative predictive value:

	D+	**D-**	**total**
T+	**a**	b	a+b
T-	**c**	d	c+d
total	**a+c**	b+d	N=a+b+c+d

… of a negative test result (NPV):
probability that a negative subjects (T-) is disease-free (D-)

$$\mathrm{P(D-|T-)} = \frac{true\ negative}{negative\ subjetcs} = \frac{d}{c+d}$$

The proportion of patients who test negative who actually have not the disease

Breast cancer example

	D+	**D-**	**total**
T+	a=9	b=89	a+b=98
T-	c=1	d=901	c+d=902
total	a+c = 10	b+d = 990	N=1000

Among those who have **negative test**, what proportion actually do not have the disease?

NPV = Pr(D-|T-) = d/(d+c) = 901/902 =99.9%

Interpretation: the test has almost perfect negative predictive value. Among those who are negative on the test, we can be perfectly confident that none of those individuals actually have the disease.

In summary:

The ratio D+/(D+ U D-) is the disease prevalence.

Fixed characteristic of the test:

- ratio TN/D-: **specificity**;
- ratio TP/D+: **sensitivity**;

Not fixed characteristic of the test:

- ratio TP/T+ is the predictive value of a positive result (PPV), increases with the prevalence.
- ratio TN/T- is the predictive value of a negative result (NPV), increases when prevalence decreases.

Example:

To evaluate ...

... **a diagnostic test** you need
Sensitivity (Sn = TP/diseased)
Specificity (Sp = TN/healthy)

... **a patient** you need
Positive Predictive Value (PPV = TP/positives)
Negative Predictive Value (NPV = TN/negatives)

Measures of performance

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

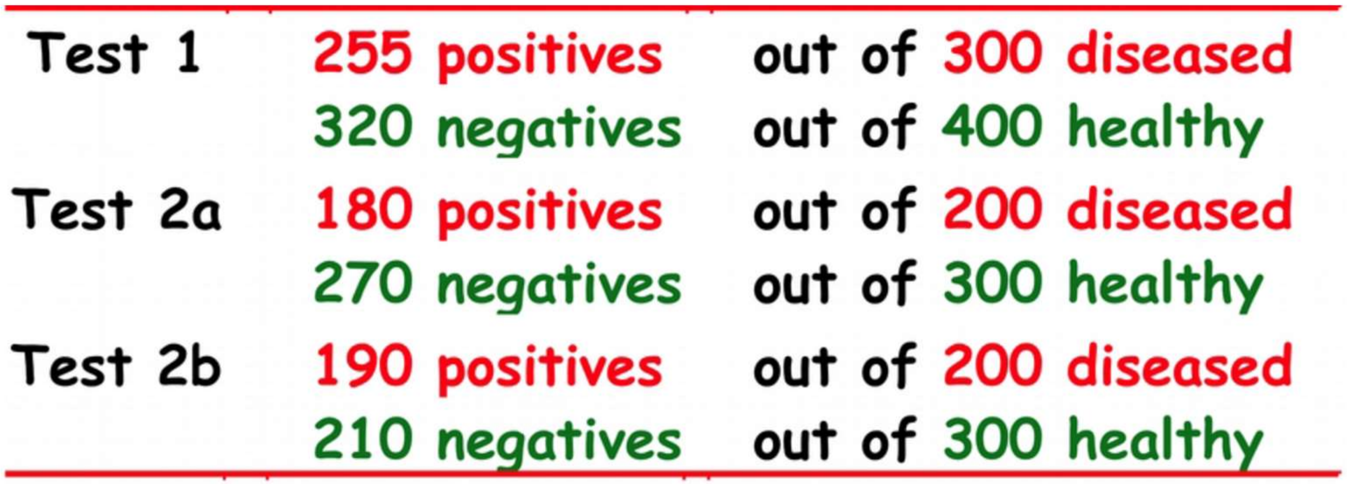
Specificity (Sp)
Specificity is the ability of the test to identify correctly those who do not have the disease (d) from all individuals free from the disease (b+d)

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

Exercise:

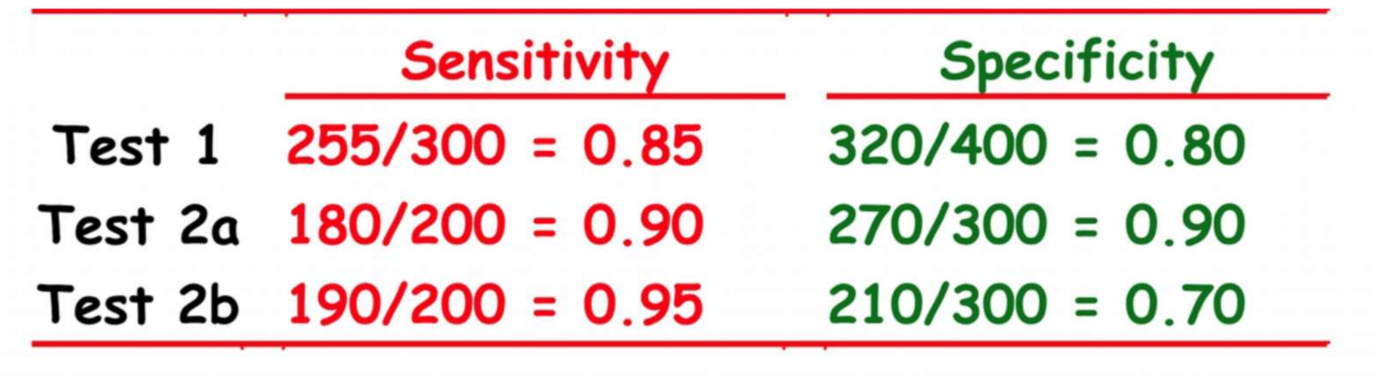
Consider the results obtained with three tests for the diagnosis of M disease, one already in use (1), and the other two (2a and 2b) proposed as an alternative to the first one:



Test 1	255 positives	out of 300 diseased
	320 negatives	out of 400 healthy
Test 2a	180 positives	out of 200 diseased
	270 negatives	out of 300 healthy
Test 2b	190 positives	out of 200 diseased
	210 negatives	out of 300 healthy

Calculate sensitivity and specificity for each test

Exercise:



	Sensitivity	Specificity
Test 1	255/300 = 0.85	320/400 = 0.80
Test 2a	180/200 = 0.90	270/300 = 0.90
Test 2b	190/200 = 0.95	210/300 = 0.70

Test 2a has greater sensitivity and specificity of the test 1:
it is better than test 1

Test 2b has greater sensitivity of Test 1 (+10%), but less specificity (-10%):
Test 2b is better or worse than the one already in use?

If the goal is: to identify the largest possible number of patients the best test has greater sensitivity
If the goal is: identify with certainty diseased patients, the best test has higher specificity
BUT also the prevalence of the disease should be taken in consideration

Prevalence of the disease

The disease prevalence is the ratio between D+/(D+ U D-)

Before doing any test, the best guide you have to a diagnosis is based on prevalence:

-Common conditions (in this population) are the more likely diagnoses

Prevalence indicates the “**pre-test probability of disease**” = $\frac{(a+c)}{N}$

Low prevalence (P(D+)=10%):

Test 1	**D^+**	**D^-**	***Totale***
T^+	85	180	265
T^-	15	720	735
Total	100	900	**1000**

Test 1

Sn = 0.85 Sp = 0.80

PPV = 85/265 = 0.321

NPV = 720/735 = 0.980

Test 2b	**D^+**	**D^-**	***Totale***
T^+	95	270	365
T^-	5	630	635
Total	100	900	**1000**

Test 2b

Sn = 0.95 Sp = 0.70

PPV = 95/365 = 0.260

NPV = 630/635 = 0.992

If the goal is to identify the largest possible number of patients, the best test has greater sensitivity, which involves:

- A better predictive value of a negative result (a negative result almost certainly indicates a healthy subject)
- But a lower predictive value of a positive result (a positive outcome may correspond a healthy subject).

Clinical applications:
GP: first attempt at diagnosing

- Low prevalence
- higher sensitivity: Not to fail to identify who has the disease ... at the cost of having false positives
- A better predictive value of a negative result: a negative result almost certainly indicates a healthy subject

SnNout rule

When a sign, test or symptom has a high sensitivity (Sn), a negative (N) result rules out the diagnosis.

High prevalence (P(D+)=80%):

Test 1	**D^+**	**D^-**	***Totale***
T^+	680	40	720
T^-	120	160	280
Total	800	200	**1000**

Test 1

Sn = 0.85 Sp = 0.80

PPV = 680/720 = 0.944

NPV = 160/280 = 0.571

Test 2b	**D^+**	**D^-**	***Totale***
T^+	760	60	820
T^-	40	140	180
Total	800	200	**1000**

Test 2b

Sn = 0.95 Sp = 0.70

PPV = 760/820 = 0.927

NPV = 140/180 = 0.778

If the goal is: identify with certainty patients, the best test has higher specificity, which involves:

- a better predictive value of a positive result (almost certainly a positive result indicates a sick person),
- but a lower predictive value of a negative result (a negative result may correspond to a sick person).

Clinical applications
Specialised doctor: last confirmation of a strong suspicion

- High prevalence
- Higher specificity : to confirm a diagnosis already suggested by other data
- a better predictive value of a positive result (almost certainly a positive result indicates a sick person)

SpPin rule

When a sign, test or symptom has an extremely high specificity (Sp) (say, over 95%), a positive (P) result tends to rule in the diagnosis.

Characteristics of Prevalence

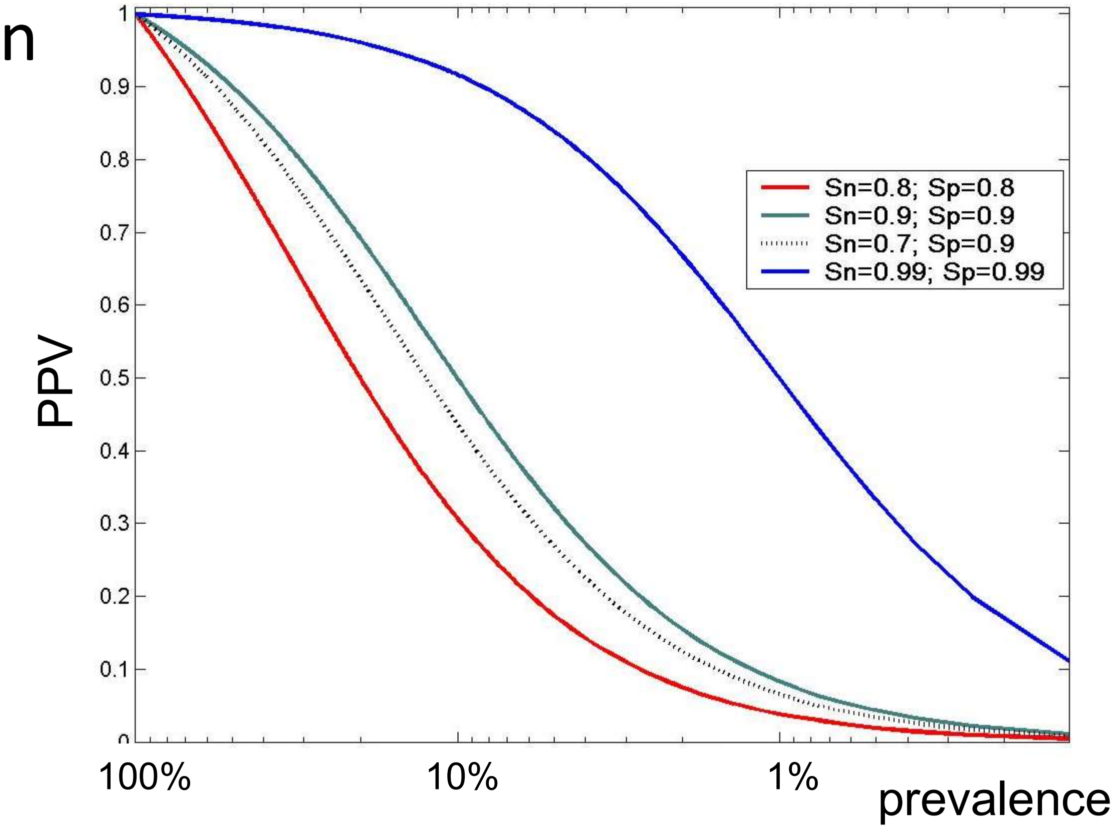
- ✓ PPV is dependent on the prevalence of a health indicator in population
- ✓ As prevalence ↑ the probability that individual who screens positive will be true case ↑
- ✓ As prevalence ↑ probability of being true negative case ↓
- ✓ Sensitivity and specificity are not directly influenced by prevalence
- ✓ Sensitivity and specificity look *among* those who have the health indicator versus those who do not

- ✓ PPV and NPV are dependent on **prevalence:**

  In evaluating these indices it is therefore necessary to know which population it refers to and to know what the prevalence of the disease is in this population.

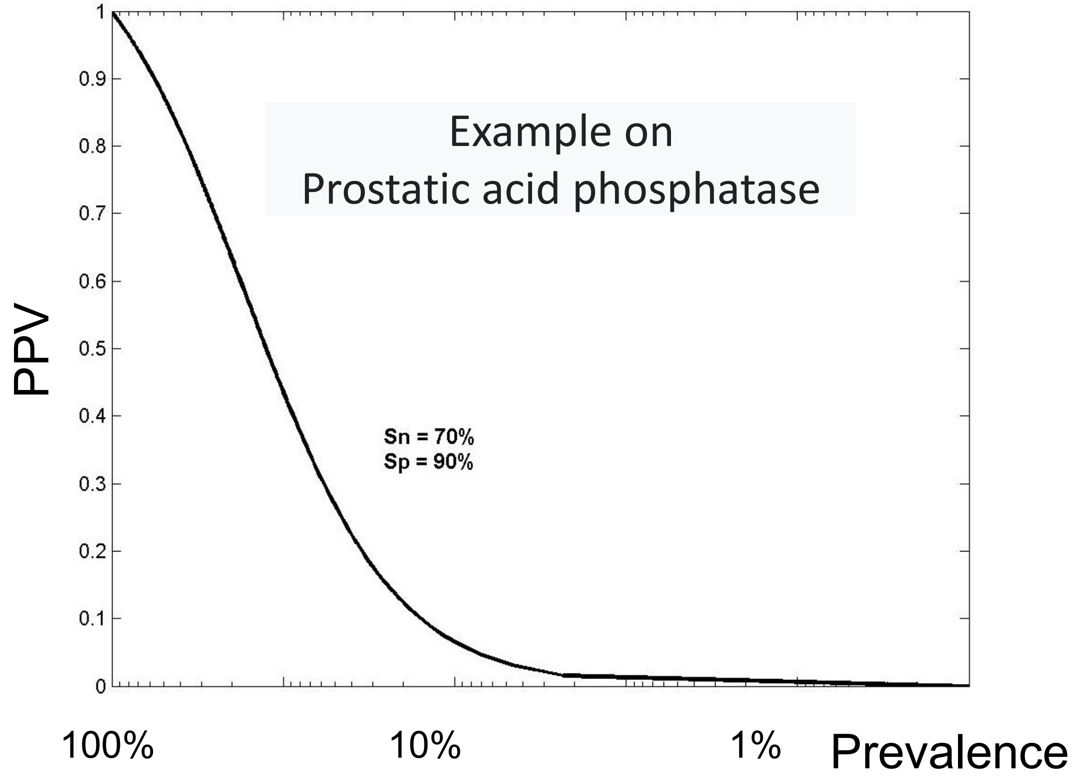
Effect of the prevalence on the predictive value of the population

When the prevalence of the disease in the population is high, the performance of all tests is good.



However, for very low prevalence values, the predictive value of all tests approaches zero; under these conditions, any test becomes virtually useless for diagnostic purposes. The influence of prevalence on the predictive value is proportional to the decrease in sensitivity and specificity of the test.

Effect of the prevalence on the predictive value of the population



	Prevalence	PPV
General Pop.	35/100000	0.4%
Males, age≥ 75	500/100000	5.6%
Suspicious prostate node	50000/100000	93.0%

Post test probability: predictive value

PPV: Probability of presence of the disease in the evaluated subject, when the diagnostic test result is positive P(D+|T+).
NPV: Probability of absence of the disease in the evaluated subject, when the diagnostic test result is negative P(D-|T-).

Calculating from Bayes’ Theorem:

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

$$= \frac{P(D+)P(T+|D+)}{P(D+)P(T+|D+)+P(D-)P(T+|D-)} = \frac{Prev*Sn}{Prev*Sn+(1-Prev)*(1-Sp)}$$

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

$$= \frac{P(D-)P(T-|D-)}{P(D-)P(T-|D-)+P(D+)P(T-|D+)} = \frac{(1-Prev)*Sp}{(1-Prev)*Sp+Pr \quad *(1-Sn)}$$

Exercise: compute sensitivity, specificity, predictive values

	Iron deficiency anemia		**Tot**
Serum ferritin	present	absent	
Positive (<65mmol/L)	731	270	1001
Negative (≥65mmol/l)	78	1500	1578
Tot	809	1770	2579

Exercise: compute sensitivity, specificity, predictive values

	Iron deficiency anemia		Tot
Serum ferritin	present	absent	
Positive (<65mmol/L)	731	270	1001
Negative (≥65mmol/l)	78	1500	1578
Tot	809	1770	2579

TP= 731
FP = 270
Prevalence = 809/2579 = 0.31 (31%)
Sensitivity = 731/809 = 90.4%
Specificity = 1500/1770 = 84.7%
PPV = 731/1001 = 73.0%
NPV = 1500/1578 = 95.1%

What if another threshold is used (e.g 50)?

	Iron deficiency anemia		**Tot**
Serum ferritin	present	absent	
Positive (<50mmol/L)	631	170	801
Negative (≥50mmol/l)	178	1600	1778
Tot	809	1770	2579

Prevalence = 809/2579 = 0.31 (31%)
Sensitivity = 631/809 = 78.0%
Specificity = 1600/1770 = 90.4%

What if another threshold is used (e.g 70)?

	Iron deficiency anemia		**Tot**
Serum ferritin	present	absent	
Positive (<70mmol/L)	781	320	1101
Negative (≥70mmol/l)	28	1450	1478
Tot	809	1770	2579

Prevalence = 809/2579 = 0.31 (31%)

Sensitivity = 781/809 = 96.5%

Specificity = 1450/1770 = 81.9%

Summarising

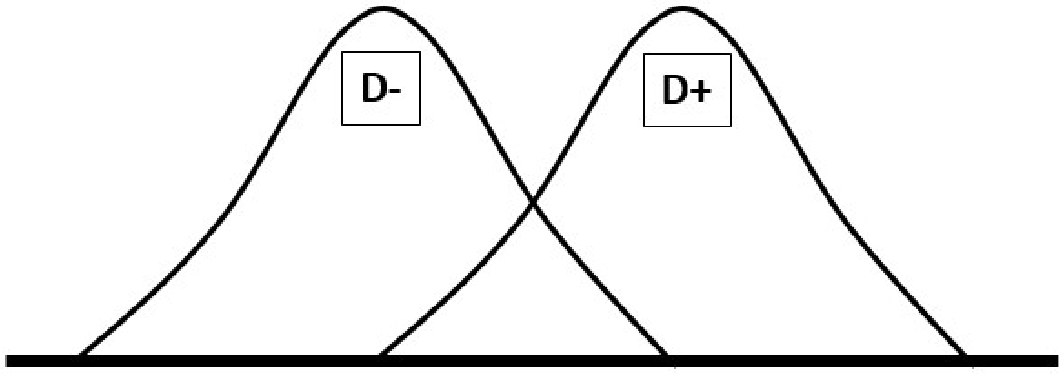
Cut-off (mmol/L)	Sensitivity (%)	Specificity (%)
50	78.0	90.4
65	90.4	84.7
70	96.5	81.9

Sensitivity and Specificity are intrinsic characteristics of the test, with the property that, as the cut-off changes, the increase of one implies the decrease of the other.

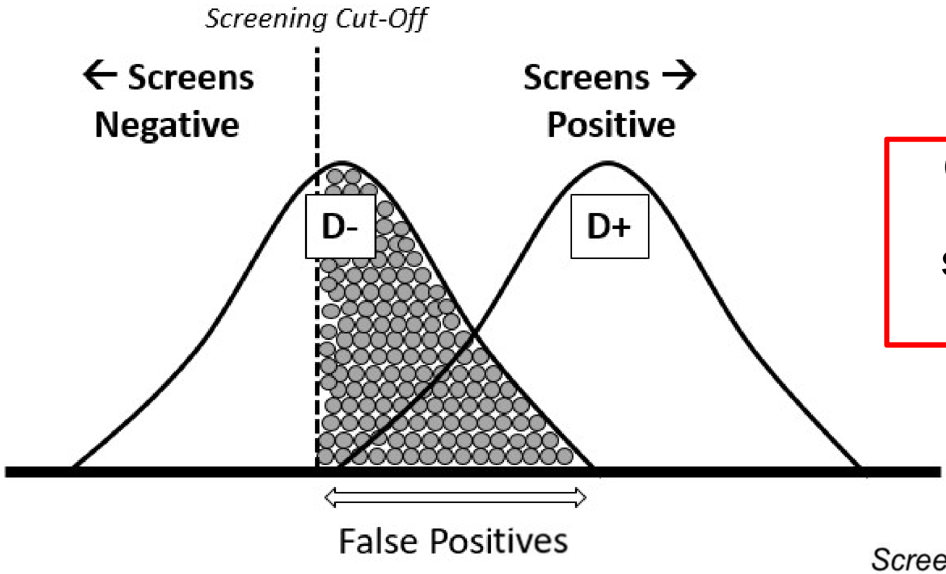
Selecting a Cutting Point

Different cut-points yield different sensitivities and specificities:

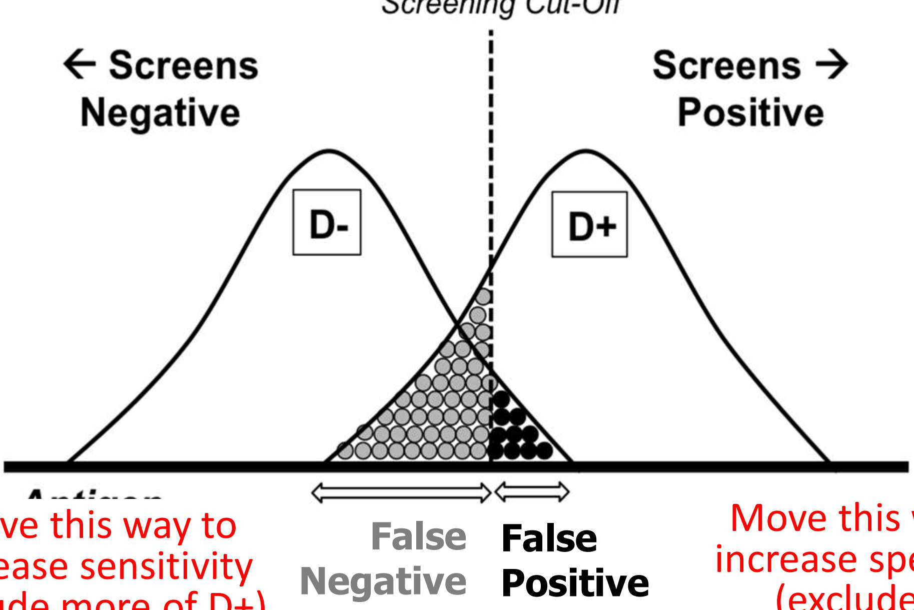
- ✓ The cut-point determines how many subjects will be considered as having the disease
- ✓ The cut-point that identifies more true negatives will also identify more false negatives
- ✓ The cut-point that identifies more true positives will also identify more false positives



Change the cut-off point



Changing cut-point can improve sensitivity *or* specificity, but never both



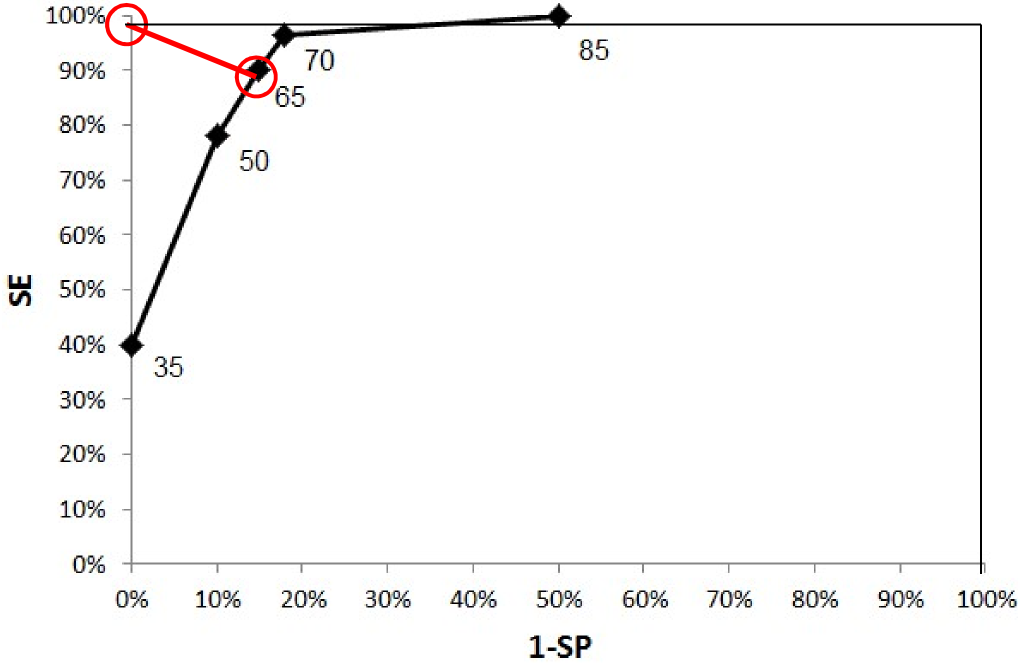
Move this way to increase sensitivity (include more of D+)

Move this way to increase specificity (exclude D-)

How to define a cut-off?

Receiver Operating Characteristic (ROC) curves

A curve that combines sensitivity and specificity at different cut-off points.



The "best" threshold value is the one corresponding to the point closest to the upper left corner of the graph that identifies the perfect test, ie the one that has SE = 100 and SP = 100 (1-SP = 0). Other criteria exist.

How to define the cut-off?

If the diagnostic (confirmatory) test is expensive or invasive:
-Minimize false positives; use a cut-point with high specificity

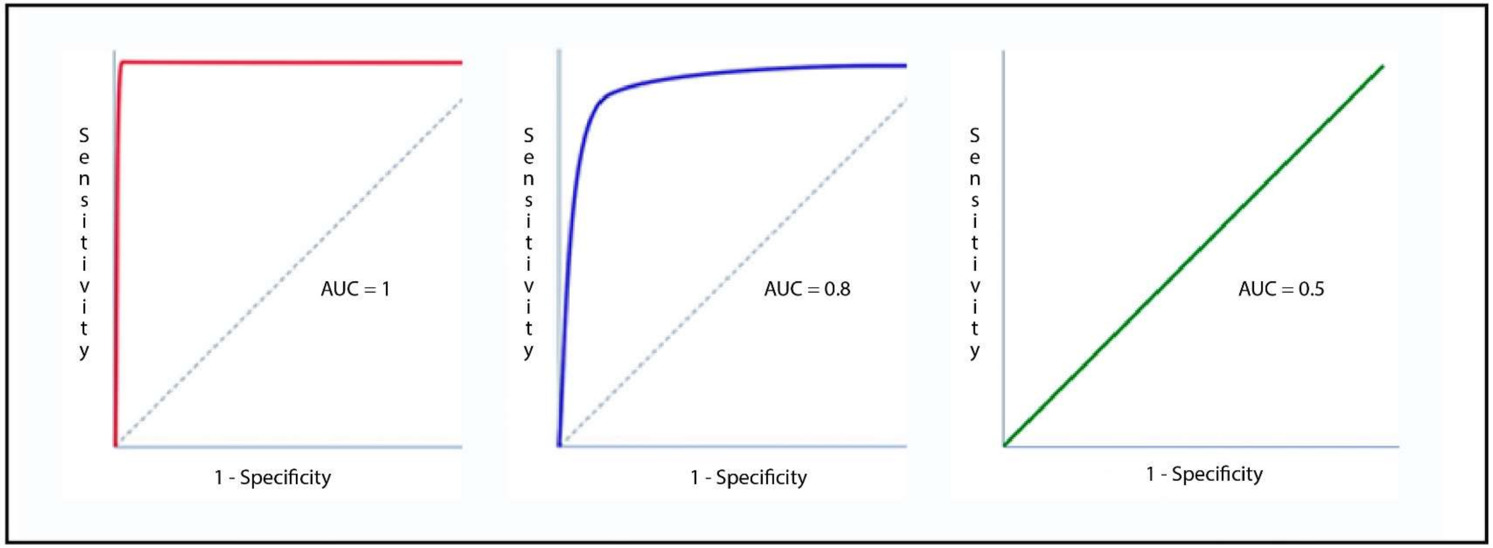
If the penalty for missing a case is high (e.g., the disease is fatal and treatment exists, or disease easily spreads):
- Minimize false negative; that is, use a cut-point with high sensitivity

Choice depends on relative implications of false positive and false negative errors!!!

Receiver Operating Characteristic (ROC) curves

The AUC (Area Under the Curve) index can be used to evaluate the diagnostic ability of the test as a whole. Its value is equal to the area under the ROC curve.

The more the AUC approaches 1 (maximum area), the more the test correctly classifies healthy and sick. The minimum value is instead equal to 0.5.



AUC> 0.9
very accurate

0.7-0.9
moderate accuracy
0.5-0.7
low accuracy

0.5
no ability to
discriminate

Requirements for SCREENING TEST...

Obviously not for all pathologies one can think of an effective screening program.

For an effective screening is necessary that the disease

- Has a high incidence and severity (the disease must be socially relevant, very widespread, to justify the cost of screening)
- has a long preclinical phase (the disease must be diagnosed in the asymptomatic period)
- Prognosis is likely to improve if diagnosed in the asymptomatic period

Furthermore, it is necessary that

- There is an effective therapy (for example, the removal by endoscopy of colorectal polyps or conservative surgery in cancer breast)
- the natural history of disease is known (including the development from latent stage to symptomatic stage).

Diagnosis and Screening

It is important to understand the difference between **diagnosis** a disease and **screening** for it:

- In the former case there are usually some symptoms, and so there may already be a suspicion that something is wrong (diagnosis).

- If a test is positive some action will be taken. In the latter case there are usually no symptoms and so if the test is negative the person will have no further tests (screening).

Low disease Prevalence:

If the goal is:
to identify the largest possible number of patients,
the best test has greater sensitivity, which involves:

- A better predictive value of a negative result
- But a lower predictive value of a positive result

High disease Prevalence:

If the goal is:
to identify the certainty patients,
the best test has higher specificity, which involves:

- A better predictive value of a positive result
- But a lower predictive value of a negative result

In term of Likelihood Ratios

Defined as the odds that a given level of a diagnostic test result would be expected in a patient with the disease, as opposed to a patient without.

Calculating from Bayes’ Theorem:

- Express pre-test probability as **odds**:

  Pre-test odds = prevalence/(1-prevalence)

- Convert sensitivity and specificity into **likelihood ratios:**

How much a positive test result increases the probability of disease?

$\boxed{LR+ = p(T+|D+)/p(T+|D-) = Sn/(1-Sp)}$ → bigger is better (>>1)

How much a negative test result decreases the probability of disease?

$\boxed{LR- = p(T-|D+)/p(T-|D-) = (1-Sn)/Sp}$ → smaller is better (<<1)

- Post-test odds = pre-test odds * LR (+ or -)

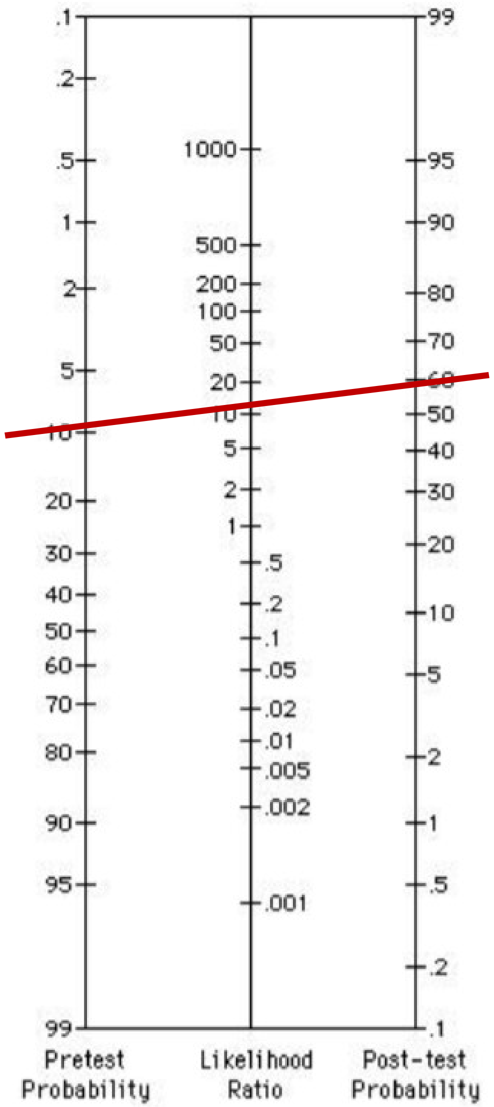
Fagan nomogram

Allows to turn pre-test probabilities into post-test probabilities. To use the nomogram, simply line up the **pre-test probability on the left** with the appropriate likelihood ratio in the center and read off the **post-test probability on the right.**

e.g. Pre-test prob = 10%

LR+ = 11

Post-test = ?



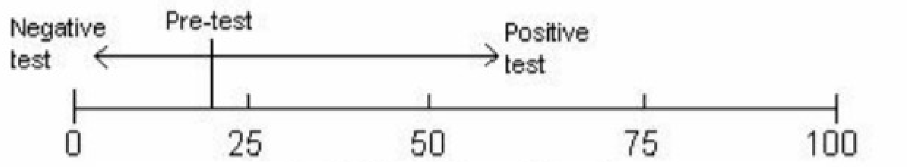
Properties of likelihood ratio (LR)

- ✓ LR indicate the powerful of a test
- ✓ Do not change with the prevalence of disease
- ✓ Combines sensitivity and specificity into one number
- ✓ Can be calculated for many levels of the test
- ✓ Can be turned into predictive values

Diagnostic tests and Likelihood Ratios

When we decide to order a diagnostic test, we want to know which test (or tests) will best help us “rule in” or “rule out” a disease in a given patient.

We take an initial assessment of the likelihood of disease (pre-test probability), do a test, then use the test results to shift our suspicion of disease one way or the other, thereby determining a final assessment of the likelihood of disease (post-test probability).



Revising the probability of disease