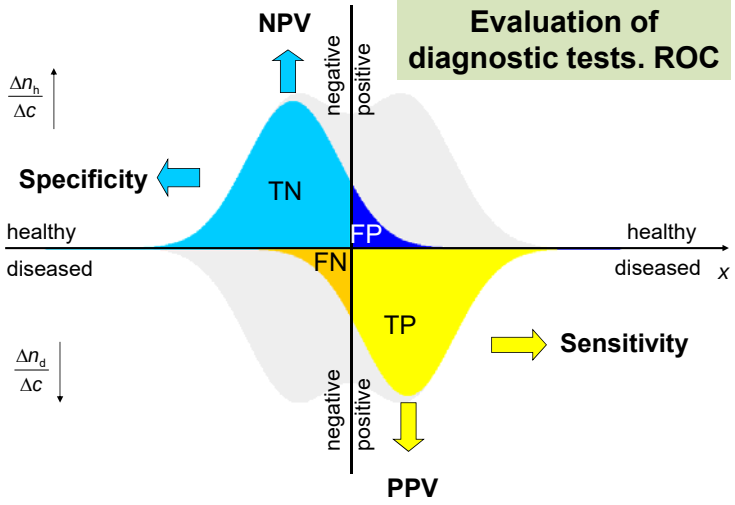
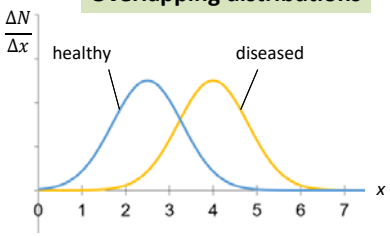
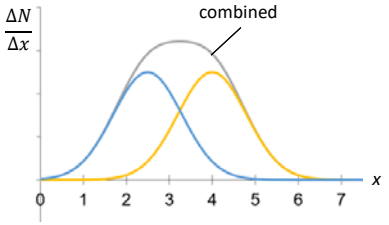
## Evaluation of diagnostic tests. ROC




KAD 2023.04.05

## Overlapping distributions




everything has a **distribution** (eye color, height, cholesterol level,…)

consider a continuous variable with a different distribution in the diseased population than in the healthy one

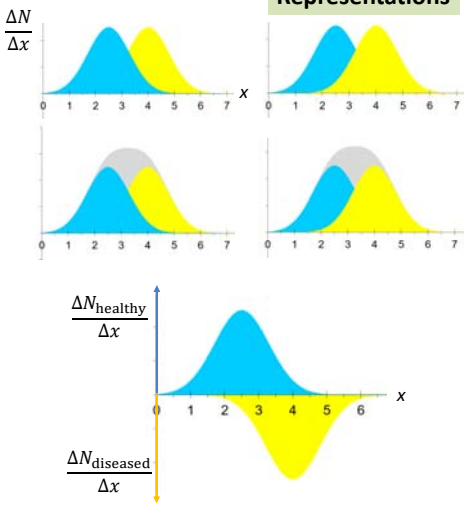
suppose that the measurable parameter is typically **larger** in the **diseased** population than in the healthy one (if smaller, the reciprocal of the original parameter can be used)

the figure shows two such density functions; the area under the curve corresponds to the number of individuals

in the present example the number of healthy and diseased is the same (the standard deviation of the parameter is the same)



## Representations



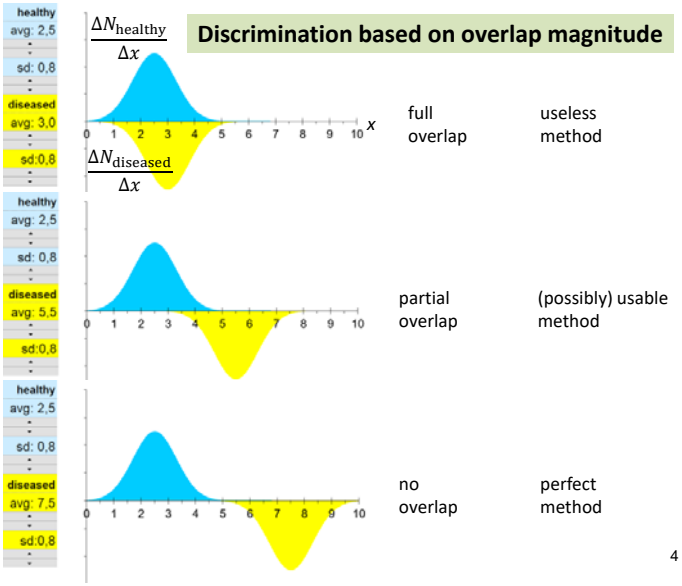

due to the great importance of the areas under the curve, we prefer an image that colors the areas instead of a line drawing

due to the overlaps, correct coloring is difficult (or impossible) in the usual representation

proposed new representation: instead of the negative axis, another positive axis, for the diseased



## Discrimination based on overlap magnitude



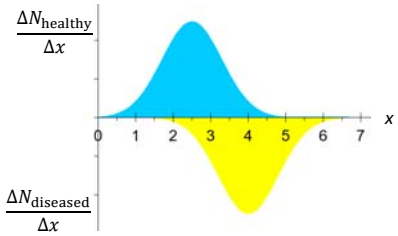

| | |
|---|---|
| full overlap | useless method |
| partial overlap | (possibly) usable method |
| no overlap | perfect method |

## Prevalence

**frequency of diseased** in examined population

measure of how common the disease is

= probability prior to test
= a-priori-probablity

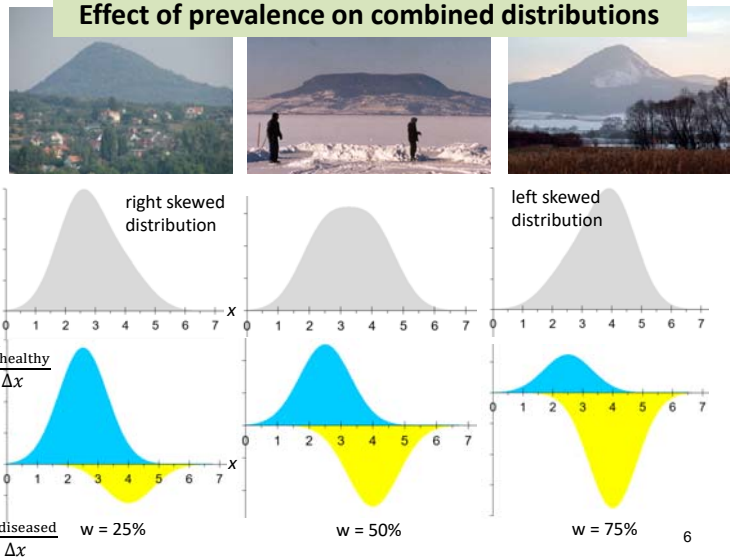



$$w = \frac{\text{diseased}}{\text{total}} = \frac{\text{diseased}}{\text{diseased} + \text{healthy}} = \frac{\text{de} - \text{sp}}{\text{se} - \text{sp}}$$

cf: incidence = the number of new cases in a given period and in a given number of population, e.g. 29 per year, per 10 000 people



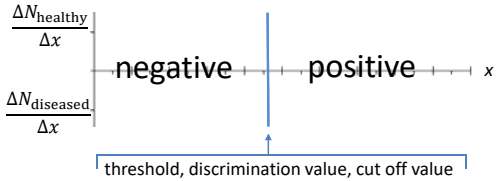
## Effect of prevalence on combined distributions






## A negative test result below the threshold and a positive test result above it

among the possible measurement parameter values, by designating a **threshold** value, we decide which will be the positive values and which are the negative ones according to the test method



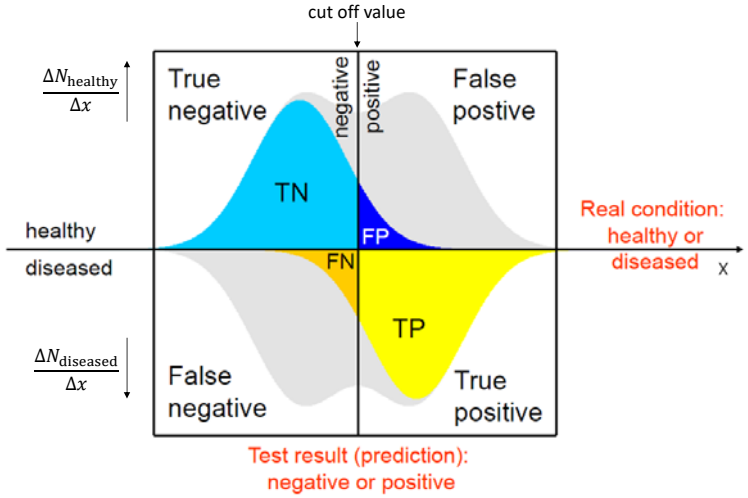

the wish/**desire**/request that the diseased and the positive, respectively healthy and negative match each other as much as possible

however, the **classification** is almost never perfect:

there will be diseased who are positive: true positive, TP ✓
there will be diseased who are negative: false negative, FN 💣
there will be healthy who are negative: true negative, TN ✓
there will be healthy who are positive : false positive, FP 💣



## Confusion matrix



## Parameters of diagnostic „**goodness**”

based on one (or more) measured parameters diagnostic tests divide the examined into (test) **positive** and (test) **negative** groups

the **“goodness”** of grouping **cannot** be characterized by a single number

(a) how well does it catch those **to be caught**?
e.g. the probability of a COVID infected stating/determining to be positive

(b) how well does it leave those **to be left alone**?
e.g. the probability of claiming to be negative for a person not infected with a COVID

(c) how reliable is a **positive test result**?
in the case of a positive test result, how certain the patient is diseased
e.g. in the case of a positive COVID test, how certain it is that the person is infected with COVID

(d) how reliable the **negative test result** is?
in the case of a negative test result, how certain the person is healthy
e.g. in the case of a negative COVID test, how certain it is that the person is not infected with COVID



The goodness of a test can be described in terms of the following diagnostic parameters

Sensitivity

Specificity

PPV, relevance

NPV, segregation

only 3 independent!

Every method must be compared with a reference-method: **gold standard** method known to always work (sometimes only the result of an autopsy)

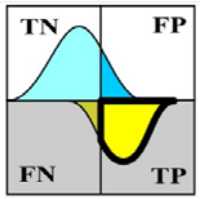


## Diagnostic **sensitivity**

= positive within diseased

= true postive rate

= recall rate




probability that the test finds the diseased positive

$$\frac{\text{▽}}{\text{▽}} = \text{se} = \frac{\text{true positive}}{\text{diseased}} = \frac{\text{TP}}{\text{TP}+\text{FN}} = p(\text{positive}|\text{diseased})$$

discr. threshold↓ sens. ↑

Large-sensitivity tests are required:

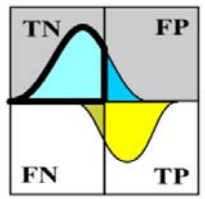
In early diagnosis (screening) so that few patients remain unrecognized.
If the risk of disease is higher than the risk of treatment.



## Diagnostic **specificity**

= negative among healthy

= true negative rate




probability that the test finds a healthy negative

$$\frac{\text{△}}{\text{△}} = \text{sp} = \frac{\text{true negative}}{\text{healthy}} = \frac{\text{TN}}{\text{TN}+\text{FP}} = p(\text{negative}|\text{healthy})$$

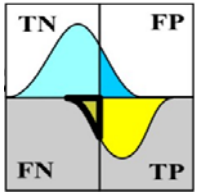
discr. thresold ↑ spec. ↑

High-specificity tests are important:

When the false positive values have severe consequences (e.g. surgery).
When the risk of treatment is higher than the risk of disease.

## Diagnostic False Negative Rate

Type-II error




the probability that the test will find a diseased negative

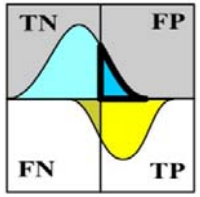
negative among diseased

$$= 1 - \text{se} = \frac{\text{FN}}{\text{diseased}} = \frac{\text{FN}}{\text{FN} + \text{TP}} = p(\text{negative}|\text{diseased})$$



## Diagnostic False Positive Rate

Type-I error




the probability that the test will find a healthy positive

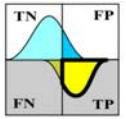
positive among the healthy

$$= 1 - \text{sp} = \frac{\text{FP}}{\text{healthy}} = \frac{\text{FP}}{\text{TN} + \text{FP}} = p(\text{positive}|\text{healthy})$$



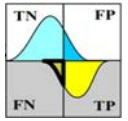
## Horizontal rates are independent of prevalence

**sensitivity**
(se)



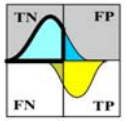

$$\text{se} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

false negative rate
(1-se)



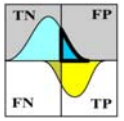

$$1 - \text{se} = \frac{\text{FN}}{\text{FN} + \text{TP}}$$

**specificity**
(sp)




$$\text{sp} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

false positive rate
(1-sp)




$$1 - \text{sp} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$



Predictive values (vertical rates)

a-posteriori-probabilities; they depend strongly on prevalence

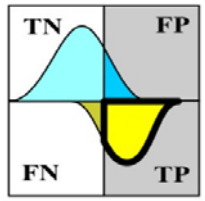
## Positive predictive value

= **PPV**

= predictive value positive

= PVP

= diagnostic **relevance**




probablity of disease if test is positive

diseased among positives

$$= \text{PPV} = \frac{\text{TP}}{\text{positive}} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{se} \cdot \text{w}}{\text{se} \cdot \text{w} + (1 - \text{sp}) \cdot (1 - \text{w})} = p(\text{diseased}|\text{positive})$$

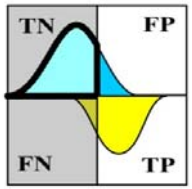
## Negative predictive value

= **NPV**

= predictive value negative

= PVN

= diagnostic **segregation**




probabilty of health if test is negative

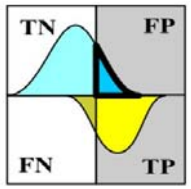
healthy among negatives

$$\frac{\text{healthy}}{\text{negative}} = \boxed{\text{NPV}} = \frac{\text{TN}}{\text{negative}} = \boxed{\frac{\text{TN}}{\text{TN}+\text{FN}}} = \frac{\text{sp}\cdot(1-\text{w})}{\text{sp}\cdot(1-\text{w})+(1-\text{se})\cdot\text{w}} = p(\text{healthy}|\text{negative})$$



## False alarm rate

=1-PPV




the probability of the absence of the disease if the test is positive

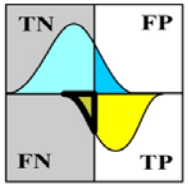
healthy among positives

$$\frac{\text{healthy}}{\text{positive}} = \boxed{1-\text{PPV}} = \frac{\text{FP}}{\text{positive}} = \boxed{\frac{\text{FP}}{\text{FP}+\text{TP}}} = p(\text{healthy}|\text{positive})$$



## False reassurance rate

=1-NPV




the probability of the presence of the disease if the test is negative

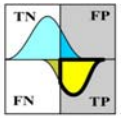
diseased among negatives

$$\frac{\text{diseased}}{\text{negative}} = \boxed{1-\text{NPV}} = \frac{\text{FN}}{\text{negative}} = \boxed{\frac{\text{FN}}{\text{FN}+\text{TN}}} = p(\text{diseased}|\text{negative})$$



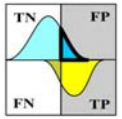
## Vertical rates are dependent of prevalence

positive predictive value (PPV)



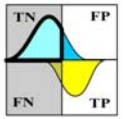

$$\frac{\text{TP}}{\text{TP}+\text{FP}} = \text{PPV}$$

false alarm rate (1-PPV)



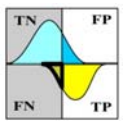

$$\frac{\text{FP}}{\text{TP}+\text{FP}} = 1-\text{PPV}$$

negative predictive value (NPV)




$$\frac{\text{TN}}{\text{TN}+\text{FN}} = \text{NPV}$$

false reassurance rate (1-NPV)



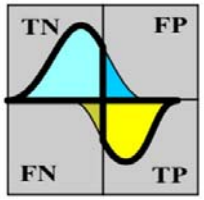

$$\frac{\text{FN}}{\text{TN}+\text{FN}} = 1-\text{NPV}$$

## Diagnostic **accuracy**

= da=de

= efficacy/efficiency

= correct classification rate




probability of correct diagnosis

$$= de = \frac{TP + TN}{total} = \frac{TP + TN}{TN + FP + FN + TP} = se \cdot w + sp \cdot (1 - w)$$

often: discrimination thresold is chosen so that accuracy is maximized



## **Effect of prevalence**

NPV = 90%

case1: ***w* = 50%**

| | | Test | |
|---|---|---|---|
| | | negative | positive |
| Gold-standard | healthy | 90 | 10 |
| | diseased | 10 | 90 |

sp = 90%

se = 90%

(de = 90%)

PPV = 90%

NPV = 99%

Case 2: ***w* = 10%**

| | | Test | |
|---|---|---|---|
| | | negative | positive |
| Gold-standard | healthy | 810 | 90 |
| | diseased | 10 | 90 |

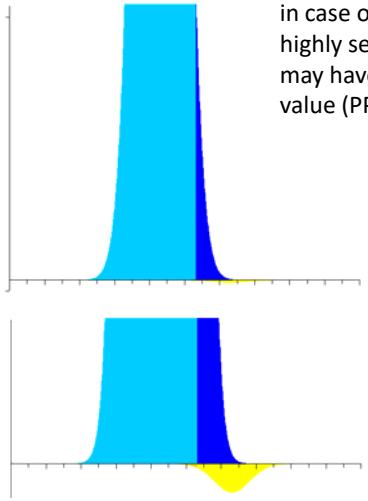
sp = 90%

se = 90%

(de = 90%)

PPV = 50%





in case of very small prevalence a highly sensitive and specific test may have low positive predictive value (PPV)

prevalence = 0.1 %

sensitivity = 98 %

specificity = 98 %

↓

PPV = 4 %



## **Overview**

| | | | | | |
|---|---|---|---|---|---|
| **Sensitivity** | se | $\frac{TP}{TP+FN}$ | $p(P\|D)$ | positive within diseased | True Positive Rate |
| **Specificity** | sp | $\frac{TN}{TN+FP}$ | $p(N\|H)$ | negative among healthy | True Negative Rate |
| **False Negative Rate** | 1-se | $\frac{FN}{TP+FN}$ | $p(N\|D)$ | negative among diseased | |
| **False Positive Rate** | 1-sp | $\frac{FP}{TN+FP}$ | $p(P\|H)$ | positive among the healthy | |
| **Positive Predictive Value** | PPV | $\frac{TP}{TP+FP}$ | $p(D\|P)$ | diseased among positives | Relevance |
| **Negative Predictive Value** | NPV | $\frac{TN}{TN+FN}$ | $p(H\|N)$ | healthy among negatives | Segregation |
| **False alarm rate** | 1-PPV | $\frac{FP}{TP+FP}$ | $p(H\|P)$ | healthy among positives | |
| **False reassurance rate** | 1-NPV | $\frac{FN}{TN+FN}$ | $p(D\|N)$ | diseased among negatives | |

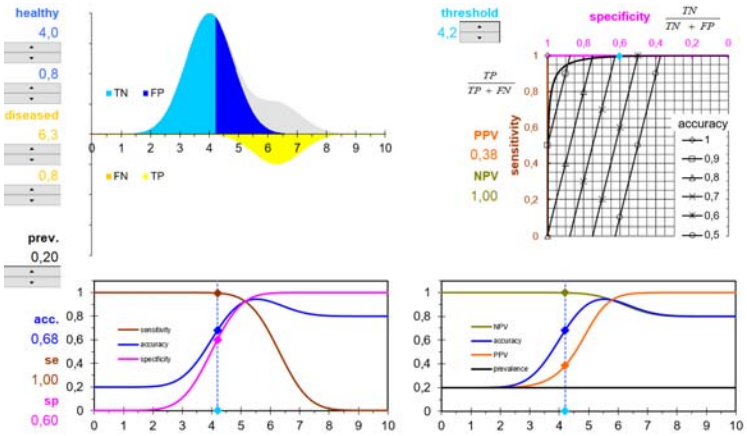
prevalence-independent

prevalence-dependent

conditional probability (Bayes)

how well does it catch those **to be caught**?

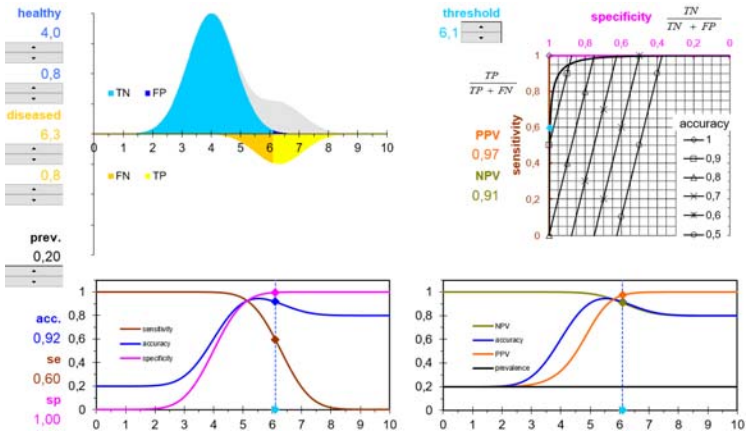
## **Maximize** diagnostic **sensitivity**






how well does it leave those **to be left alone**?

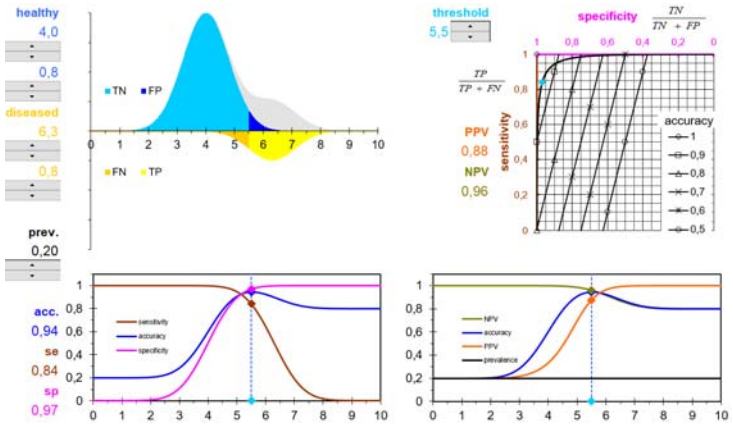
## **Maximize** diagnostic **specificity**






it is equally important to catch those **to be caught** and to leave those **to be left alone**?

## **Maximize** diagnostic **accuracy**






## **Take-home message**

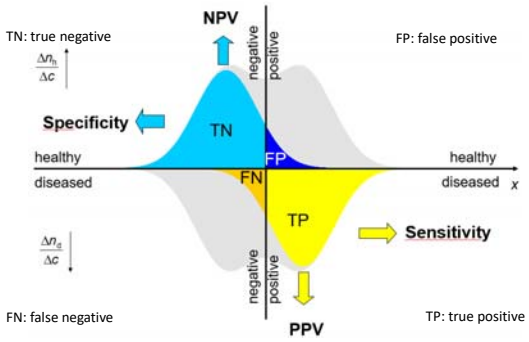
everything has a **distribution**; the distribution of sick and healthy values **overlap**

whether it is possible to decide which is more important :

to detect the disease in as many patients as possible in order to receive treatment (**maximizing sensitivity**), or

to assume a false positive value (minimizing false-positive ratio or **maximizing specificity**) in as few healthy people as possible so that they do not receive unnecessary therapy

if they cannot be decided, they are equally important: **maximizing accuracy**




TN: true negative

FP: false positive

FN: false negative

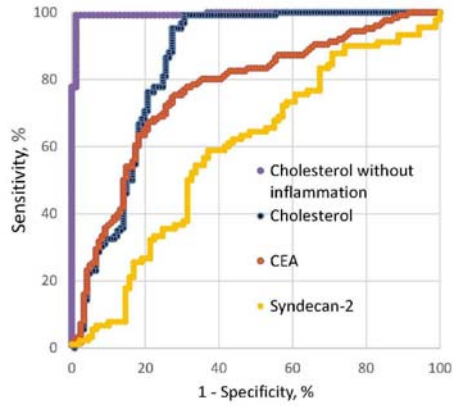
TP: true positive

diagnostic accuracy:

$$de = \frac{TP + TN}{total} = \frac{TP + TN}{TN + FP + FN + TP}$$

in case of several possible methods: ROC

## Comparison of diagnostic tests: the ROC space




Which marker has the best diagnostic performance in detecting neoplastic pleural involvement (figure 1)?

Miklos Gulyas, Janos Fillinger, Andras D Kaposi, Miklos Molnar, Use of cholesterol and soluble tumour markers CEA and syndecan-2 in pleural effusions in cases of inconclusive cytology, . J Clin Pathol 2019;72:529–535

$$p(\text{positive}\,|\,\text{diseased}) = \frac{p(\text{positive and diseased})}{p(\text{diseased})} = \frac{\frac{TP}{n}}{\frac{TP+FN}{n}} = \frac{TP}{TP+FN}$$ Sensitivity

**(1) The Total Probability Theorem:**

If the events $B_1$, $B_2$, ..., $B_n$ form a complete system of events $p(B_i) > 0$, i = 1, 2, ... , n, then $p(A) = \sum_{i=1}^{n} p(A|B_i)p(B_i)$

$B_1$: diseased
$B_2$: not diseased
A: positive

e.g.

$$p(A) = p(A|B_1)\cdot p(B_1) + p(A|B_2)\cdot p(B_2) = se\cdot w + (1-sp)\cdot(1-w) =$$

$$= \frac{RP}{RP+FN}\cdot\frac{RP+FN}{RP+FN+RN+FP} + \frac{FP}{RN+FP}\cdot\frac{RN+FP}{RP+FN+RN+FP} = \frac{RP+FP}{RP+FN+RN+FP}$$

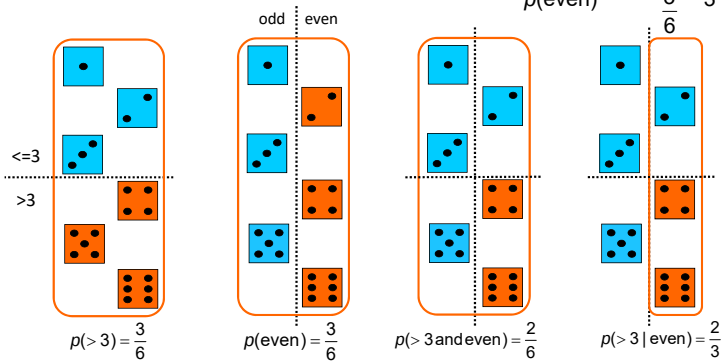


## Conditional probability

$$p(A|B) = \frac{p(A \text{ and } B)}{p(B)}$$

The probability that A is true given that B is true*.

e.g.: in a cube experiment : $p(>3\,|\,\text{even})$ $= \frac{p(>3 \text{ and even})}{p(\text{even})} = \frac{\frac{2}{6}}{\frac{3}{6}} = \frac{2}{3}$




*the conditional probability of *A* given *B*; the probability of *A* under the condition *B*

**(2) ) Bayes' theorem:**

Conversely, assuming event A, we are looking for the probability of event $B_k$.

$B_1$: diseased
$B_2$: not diseased
A: positive

$$p(B_k|A) = \frac{p(A|B_k)p(B_k)}{\sum_{i=1}^{n} p(A|B_i)p(B_i)} = \frac{p(A|B_k)p(B_k)}{p(A)}$$

e.g.

$$p(\text{diseased}|\text{positive}) = \frac{p(\text{positive}|\text{diseased})\cdot p(\text{diseased})}{p(\text{positive}|\text{diseased})\cdot p(\text{diseased}) + p(\text{positive}|\text{healthy})\cdot p(\text{healthy})} =$$

$$= \frac{se\cdot w}{se\cdot w + (1-sp)\cdot(1-w)} = PPV$$

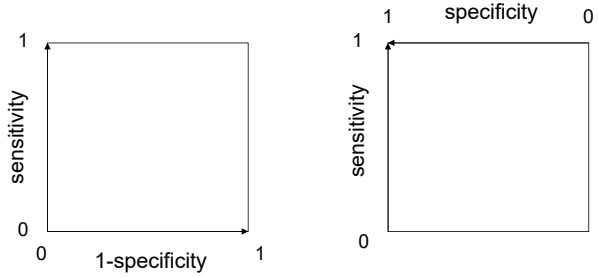
$$p(B_2|1-A) = \frac{p(1-A|B_2)\cdot p(B_2)}{p(1-A|B_1)\cdot p(B_1) + p(1-A|B_2)\cdot p(B_2)} = \frac{sp\cdot(1-w)}{(1-se)\cdot w + sp\cdot(1-w)} = NPV$$

## Comparison of diagnostic tests: the ROC space

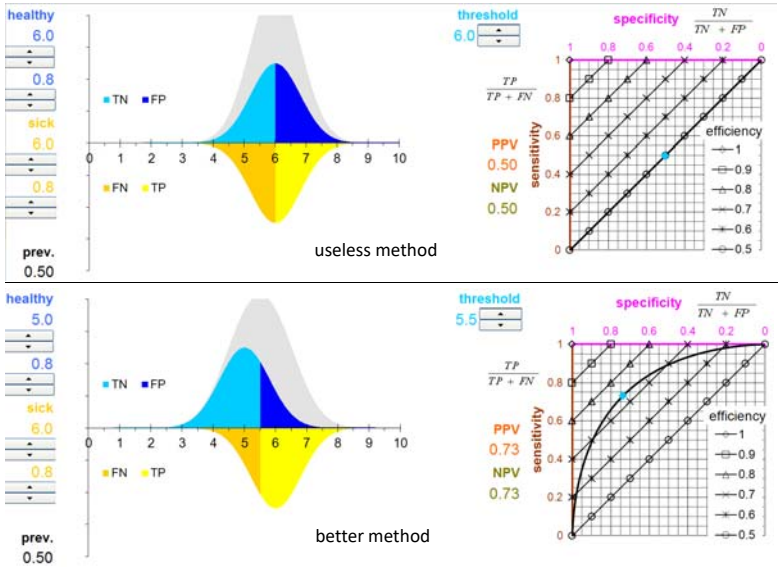
ROC: receiver-operator (operating) characteristic

~ 1950: first ROC Analysis (receiver: Radar)

~ 1970: first medical applications

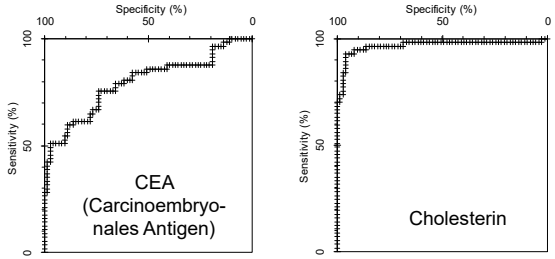





## E.g.: Tumor markers in the ascites

increased CEA and/or cholesterin concentrations in ascites are diagnostic markers for carcinomatosis




Which method is better? What discrimination threshold should be used?

Gulyás M, Kaposi AD, Elek G, Szollár LG, Hjerpe A, Value of carcinoembryonic antigen (CEA) and cholesterol assays of ascitic fluid in cases of inconclusive cytology, J Clinical Pathology 2001 (54) 831-835



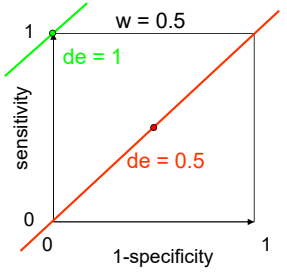
$$de = se \cdot w + sp \cdot (1-w)$$

$$\frac{de}{1-w} = \frac{w}{1-w} se + (sp-1) + 1$$

$$(1-sp) + \frac{de}{1-w} - 1 = \frac{w}{1-w} se$$

$$se = \frac{1-w}{w}(1-sp) + \left(\frac{1}{w} de + \frac{w-1}{w}\right)$$

slope intercept




if $w = 0.5$: $se = 1 \cdot (1-sp) + 2 \cdot de - 1$

The points have the same diagnostic efficiency belong to a line with a slope of 1.

If $de$ = 0.5 , the intercept is 0.

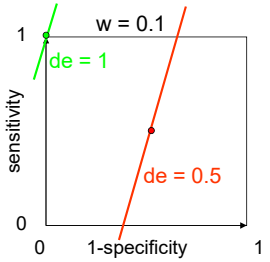
$$se = \frac{1-w}{w}(1-sp) + \frac{1}{w}de + \frac{w-1}{w}$$

slope intercept

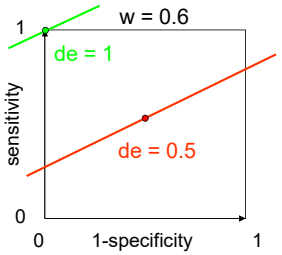
if w < 0.5: the slope of lines at identical diagnostic efficiencies is greater than 1.

if w > 0.5: the slope of lines at identical diagnostic efficiencies is smaller than 1.

e.g. *w* = 0.1, slope: 9

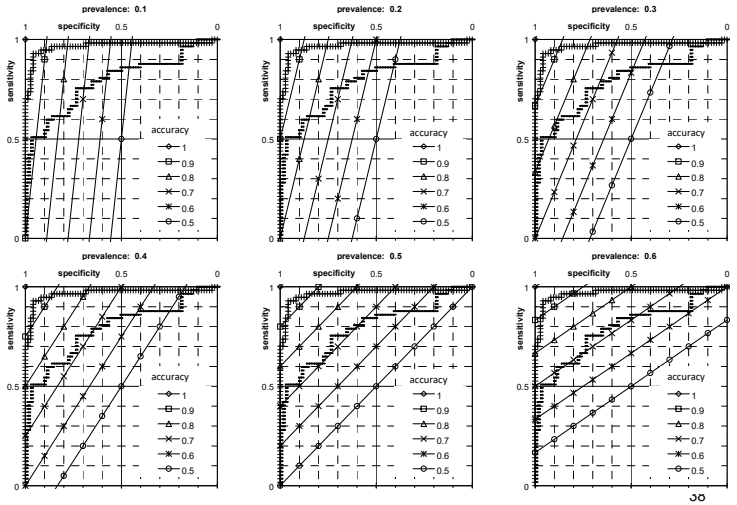



e.g. *w* = 0.6, slope: 0.66



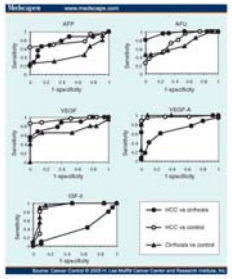


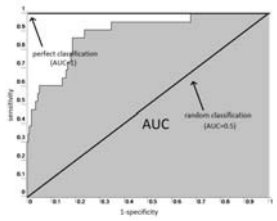

Ascites (+ Cholesterin, – CEA)

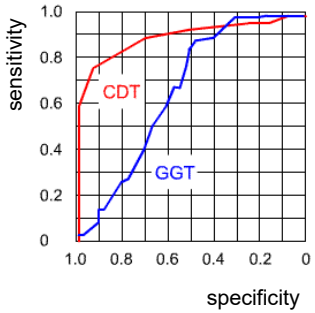





**Additional examples**



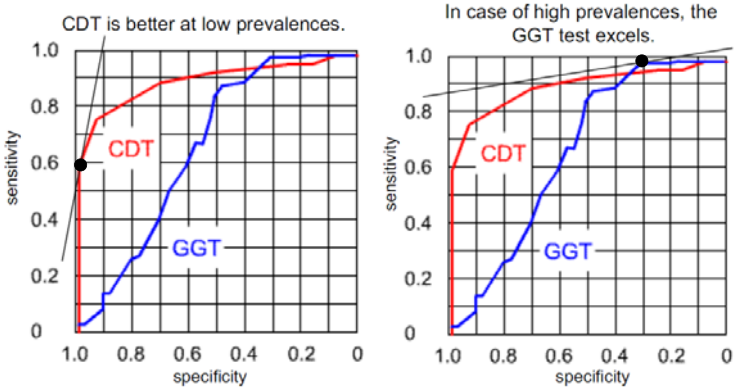



Alcoholism diagnostics with CDT (carbohydrate deficient transferrin) and GGT (gamma-glutamyltransferase). AUC of CDT is larger than of GGT. Is it a better method?

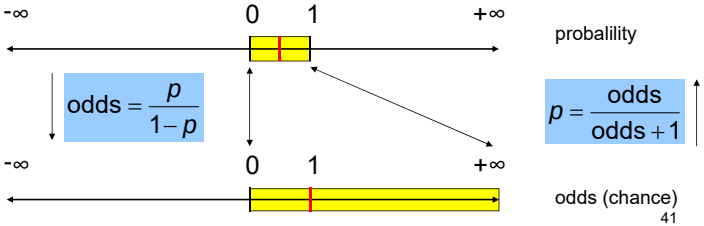


**If we maximize the diagnostic accuracy…**



## Metrics for quantifying the chances of events occurring

| event $E$ | **probability**, p($E$) | **odds** |
|---|---|---|
| impossible event | 0 | 0 |
| the occurrence and non-occurrence of the event have an equal chance | 0.5 | 1 |
| certain event | 1 | ∞ |






## Likelihood ratio

A ratio that indicates the extent to which a test method changes the chances of illness.

likelihood ratio of a positive test result:
(posttest odds/pretest odds):

$$LR_{pos} = \frac{\frac{TP}{FP}}{\frac{diseased}{healthy}} = \frac{\frac{TP}{FP}}{\frac{TP+FN}{TN+FP}} = \frac{TP}{TP+FN}\frac{1}{\frac{FP}{TN+FP}} = \frac{se}{1-sp}$$

likelihood ratio of a negative test result :

$$LR_{neg} = \frac{\frac{FN}{TN}}{\frac{diseased}{healthy}} = \frac{\frac{FN}{TN}}{\frac{TP+FN}{TN+FP}} = \frac{FN}{TP+FN}\frac{1}{\frac{TN}{TN+FP}} = \frac{1-se}{sp}$$



## $F_1$ score

$F_1$ score: is the **harmonic mean** of sensitivity and PPV

In statistical analysis of binary classification, the **F-score** or **F-measure** is a measure of a test's „accuracy".
It is one of the most important evaluation metrics in machine learning.

$$F_1 = \frac{2}{se^{-1} + PPV^{-1}} = 2\frac{se \bullet PPV}{se + PPV} = \frac{TP}{TP + \frac{1}{2}(TP + FN)}$$

harmonic mean:

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \left(\frac{\sum_{i=1}^{n} x_i^{-1}}{n}\right)^{-1}.$$

BEYOND NORMALITY: THE PREDICTIVE VALUE AND EFFICIENCY OF MEDICAL DIAGNOSES

ROBERT S. GALEN, M.D., M.P.H.
S. RAYMOND GAMBINO, M.D.
Columbia University
College of Physicians and Surgeons

A Wiley Biomedical Publication
JOHN WILEY & SONS, New York • London • Sydney • Toronto

Old version

published in 1975

Table 132

EFFICIENCY OF TEST (IN PERCENT)
PREVALENCE=50000 PER 100,000

prevalence: 50 %

| SPECIFICITY (%) | SENSITIVITY (%) 50 | 60 | 70 | 80 | 85 | 90 | 91 | 92 | 93 | 94 | 95 | 96 | 97 | 98 | 99 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 50.00 | 50 | 55 | 60 | 65 | 67 | 70 | 71 | 71 | 72 | 72 | 73 | 73 | 74 | 74 | 75 |
| 60.00 | 55 | 60 | 65 | 70 | 72 | 75 | 75 | 76 | 76 | 77 | 77 | 78 | 78 | 79 | 79 | 80 |
| 70.00 | 60 | 65 | 70 | 75 | 77 | 80 | 80 | 81 | 81 | 82 | 82 | 83 | 83 | 84 | 84 | 85 |
| 80.00 | 65 | 70 | 75 | 80 | 82 | 85 | 85 | 86 | 86 | 87 | 87 | 88 | 88 | 89 | 89 | 90 |
| 90.00 | 70 | 75 | 80 | 85 | 87 | 90 | 90 | 91 | 91 | 92 | 92 | 93 | 93 | 94 | 94 | 95 |
| 91.00 | 70 | 75 | 80 | 85 | 88 | 90 | 91 | 91 | 92 | 92 | 93 | 93 | 94 | 94 | 95 | 95 |
| 93.00 | 71 | 76 | 81 | 86 | 89 | 91 | 92 | 92 | 93 | 93 | 94 | 94 | 95 | 95 | 96 | 96 |
| 95.00 | 72 | 77 | 82 | 87 | 90 | 92 | 93 | 93 | 94 | 94 | 95 | 95 | 96 | 96 | 97 | 97 |
| 97.00 | 73 | 78 | 83 | 88 | 91 | 93 | 94 | 94 | 95 | 95 | 96 | 96 | 97 | 97 | 98 | 98 |
| 99.00 | 74 | 79 | 84 | 89 | 92 | 94 | 95 | 95 | 96 | 96 | 97 | 97 | 98 | 98 | 99 | 99 |
| 99.10 | 75 | 80 | 85 | 90 | 92 | 95 | 95 | 96 | 96 | 97 | 97 | 98 | 98 | 99 | 99 | 100 |
| 99.30 | 75 | 80 | 85 | 90 | 92 | 95 | 95 | 96 | 96 | 97 | 97 | 98 | 98 | 99 | 99 | 100 |
| 99.50 | 75 | 80 | 85 | 90 | 92 | 95 | 95 | 96 | 96 | 97 | 97 | 98 | 98 | 99 | 99 | 100 |
| 99.70 | 75 | 80 | 85 | 90 | 92 | 95 | 95 | 96 | 96 | 97 | 97 | 98 | 98 | 99 | 99 | 100 |
| 99.90 | 75 | 80 | 85 | 90 | 92 | 95 | 95 | 96 | 96 | 97 | 97 | 98 | 98 | 99 | 99 | 100 |
| 99.91 | 75 | 80 | 85 | 90 | 92 | 95 | 95 | 96 | 96 | 97 | 97 | 98 | 98 | 99 | 99 | 100 |
| 99.92 | 75 | 80 | 85 | 90 | 92 | 95 | 95 | 96 | 96 | 97 | 97 | 98 | 98 | 99 | 99 | 100 |
| 99.93 | 75 | 80 | 85 | 90 | 92 | 95 | 95 | 96 | 96 | 97 | 97 | 98 | 98 | 99 | 99 | 100 |
| 99.94 | 75 | 80 | 85 | 90 | 92 | 95 | 95 | 96 | 96 | 97 | 97 | 98 | 98 | 99 | 99 | 100 |
| 99.95 | 75 | 80 | 85 | 90 | 92 | 95 | 95 | 96 | 96 | 97 | 97 | 98 | 98 | 99 | 99 | 100 |
| 99.96 | 75 | 80 | 85 | 90 | 92 | 95 | 95 | 96 | 96 | 97 | 97 | 98 | 98 | 99 | 99 | 100 |
| 99.97 | 75 | 80 | 85 | 90 | 92 | 95 | 95 | 96 | 96 | 97 | 97 | 98 | 98 | 99 | 99 | 100 |
| 99.98 | 75 | 80 | 85 | 90 | 92 | 95 | 95 | 96 | 96 | 97 | 97 | 98 | 98 | 99 | 99 | 100 |
| 99.99 | 75 | 80 | 85 | 90 | 92 | 95 | 95 | 96 | 96 | 97 | 97 | 98 | 98 | 99 | 99 | 100 |
| 100.00 | 75 | 80 | 85 | 90 | 92 | 95 | 95 | 96 | 96 | 97 | 97 | 98 | 98 | 99 | 99 | 100 |