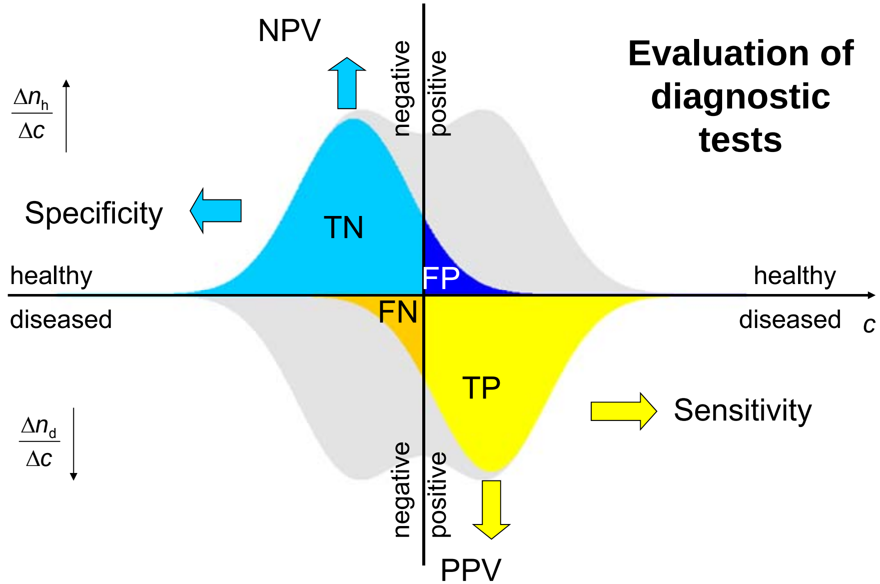
Evaluation of diagnostic tests

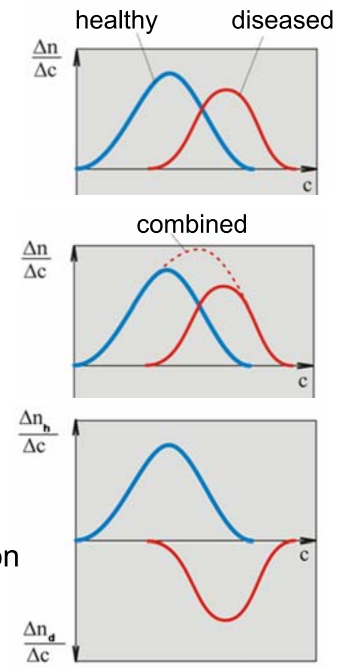


KAD 2014.11.27

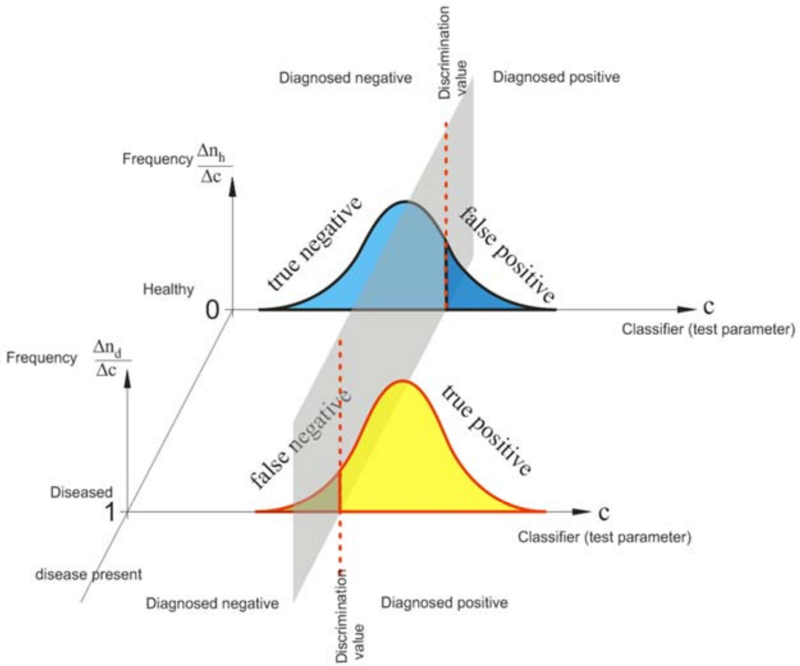
Overlapping distributions

assumption:

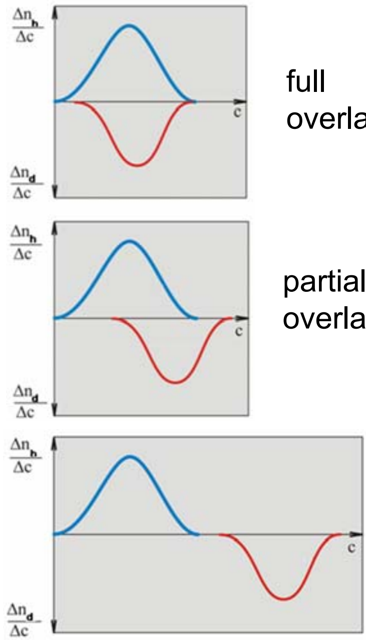
a classifier value
(e.g. serum concentration)
changes (e.g. increases)
in diseased subpopulation



novel
representation



Based on overlap magnitude:

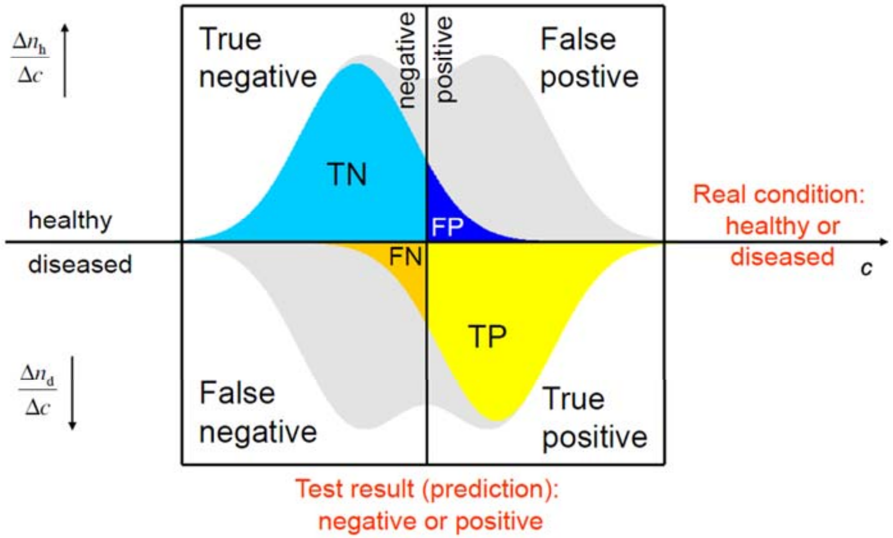


full overlap — useless method

partial overlap — real-life situation

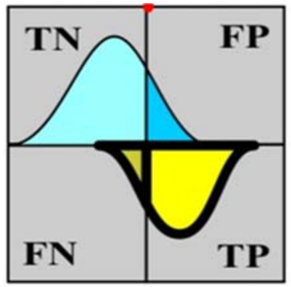
complete separation — perfect method

Contingency table: Confusion matrix (binary classification)

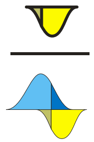


Prevalence

= frequency of diseased in examined population
= probability prior to test
= a-priori-probablity

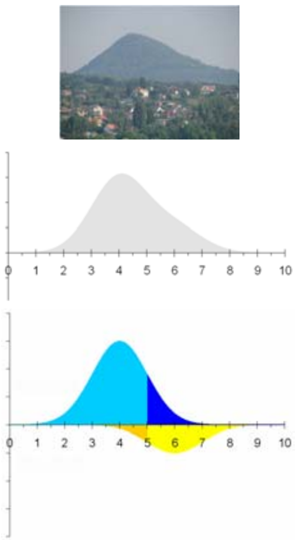


measure of how common the disease is

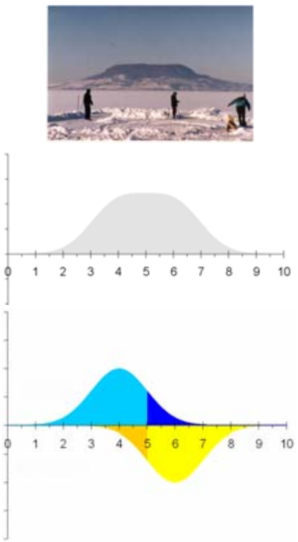


$$= \boxed{w} = \frac{\text{diseased}}{\text{total}} = \boxed{\frac{\mathrm{TP}+\mathrm{FN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}}} = \frac{\mathrm{de}-\mathrm{sp}}{\mathrm{se}-\mathrm{sp}}$$

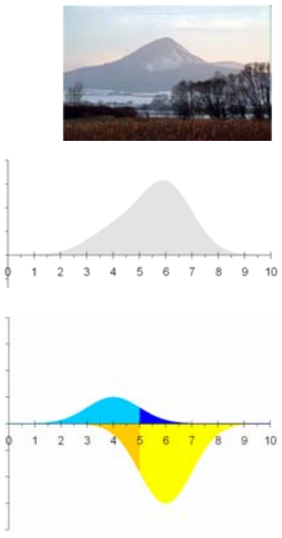
Shape of combined distributions



w = 25%



w = 50%



w = 75%

Parameters of diagnostic „goodness"

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

- Sensitivity
- Specificity
- PPV, relevance
- NPV, segregation

Every method must be compared with a reference-method (gold standard)

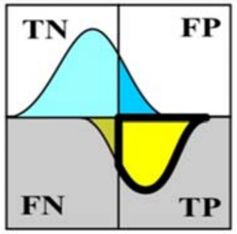
Gold standard: method known to always work; often autopsy

Diagnostic **sensitivity**

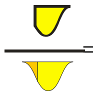
= positive within diseased

= true postive rate

= recall rate



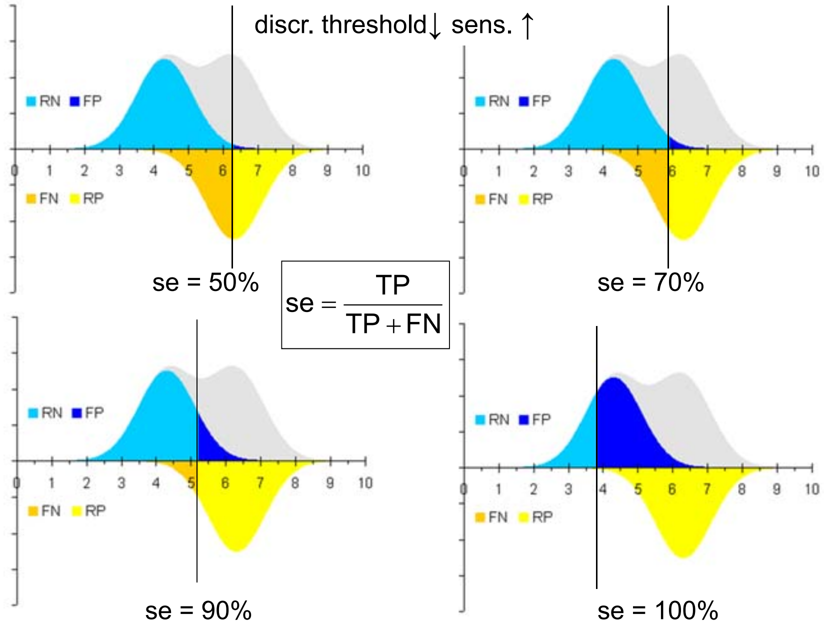
probability that the test finds the diseased positive



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

Large-sensitivity tests are required:

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

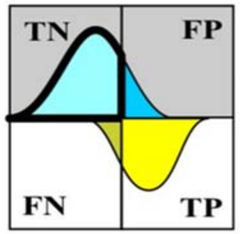


$$se = \frac{TP}{TP + FN}$$

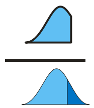
Diagnostic **specificity**

= negative among healthy

= true negative rate



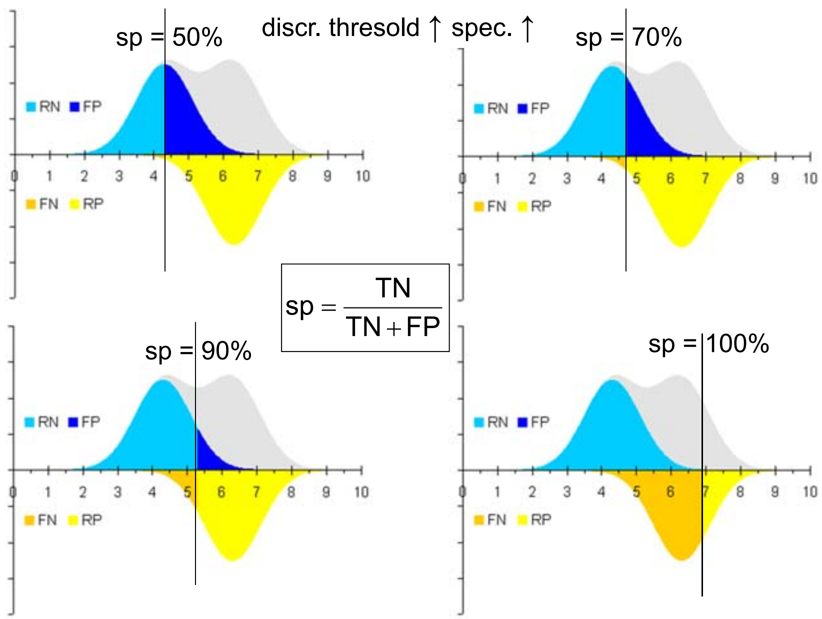
probability that the test finds a healthy negative



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

High-specificity tests are important:

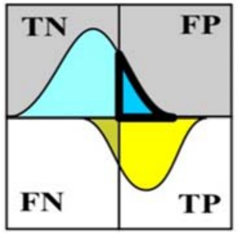
When the false positive values have severe consequences (e.g. surgery).
When the risk of treatment is greater than the risk of disease.

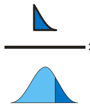


$$sp = \frac{TN}{TN + FP}$$

Diagnostic False Positive Rate

(Type-I error)

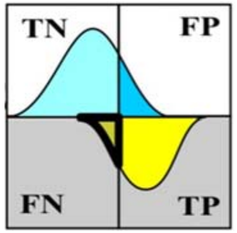


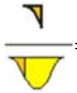


$$= 1 - \mathrm{sp} = \frac{\mathrm{FP}}{\mathrm{healthy}} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}} = p(\mathrm{positive}|\mathrm{healthy})$$

Diagnostic False Negative Rate

(Type-II error)

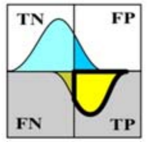


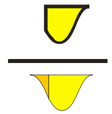


$$= 1 - \mathrm{se} = \frac{\mathrm{FN}}{\mathrm{diseased}} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} = p(\mathrm{negative}|\mathrm{diseased})$$

Horizontal rates are independent of prevalence

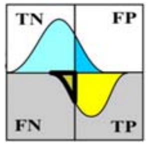
sensitivity
(se)

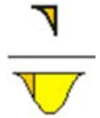




$$\mathrm{se} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

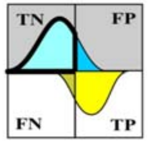
false negative rate
(1-se)

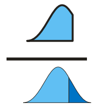




$$1 - \mathrm{se} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}}$$

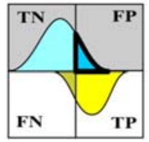
specificity
(sp)

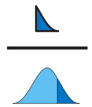




$$\mathrm{sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

false positive rate
(1-sp)





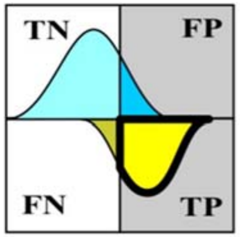
$$1 - \mathrm{sp} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}$$

Predictive values (vertical rates)

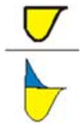
a-posteriori-probabilities; they depend strongly on prevalence

Positive predictive value

= PPV
= predictive value positive
= PVP
= diagnostic **relevance**
= diseased among positive



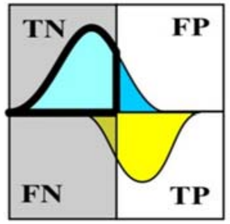
probablity of disease if test is positive



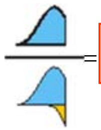
$$= \boxed{\mathrm{PPV}} = \frac{\mathrm{TP}}{\mathrm{positive}} = \boxed{\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}} = p(\mathrm{diseased}|\mathrm{positive}) = \frac{\mathrm{se} \cdot \mathrm{w}}{\mathrm{se} \cdot \mathrm{w} + (1 - \mathrm{sp}) \cdot (1 - \mathrm{w})}$$

Negative predictive value

= NPV
= predictive value negative
= PVN
= diagnostic segregation
= healthy among positives

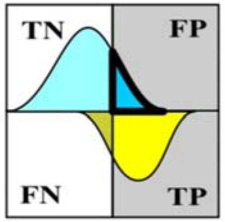


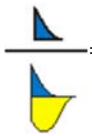
probabilty of healthiness if test is negative



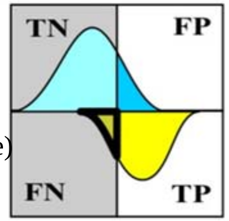
$$= \boxed{\mathrm{NPV}} = \frac{\mathrm{TN}}{\mathrm{negative}} = \boxed{\frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}} = p(\mathrm{healthy}|\mathrm{negative}) = \frac{\mathrm{sp} \cdot (1 - \mathrm{w})}{\mathrm{sp} \cdot (1 - \mathrm{w}) + (1 - \mathrm{se}) \cdot \mathrm{w}}$$

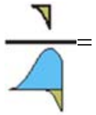
False alarm rate



 $= 1 - \mathrm{PPV} = \dfrac{\mathrm{FP}}{\mathrm{positive}} = \dfrac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TP}} = p(\mathrm{healthy}|\mathrm{positive})$

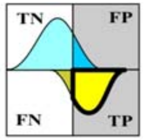
False reassurance rate

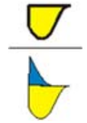


 $= 1 - \mathrm{NPV} = \dfrac{\mathrm{FN}}{\mathrm{negative}} = \dfrac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TN}} = p(\mathrm{diseased}|\mathrm{negative})$

Vertical rates are dependent of prevalence

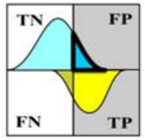
positive predictive value (PPV)





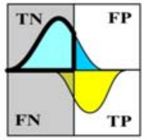
$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{FP}+\mathrm{TP}}$$

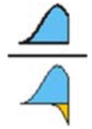
false alarm rate (1-PPV)



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

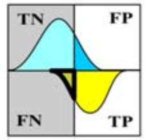
negative predictive value (NPV)

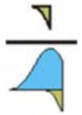




$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}}$$

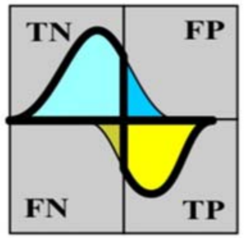
false reassurance rate (1-NPV)





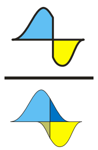
$$1 - \mathrm{NPV} = \frac{\mathrm{FN}}{\mathrm{TN}+\mathrm{FN}}$$

Diagnostic **efficacy/efficiency**



= accuracy

= correct classification rate

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

often: discrimination thresold is chosen so that de is maximized

Effect of prevalence

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

NPV = 90%

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

sp = 90%

se = 90%

PPV = 90%

(de = 90%)

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

NPV = 99%

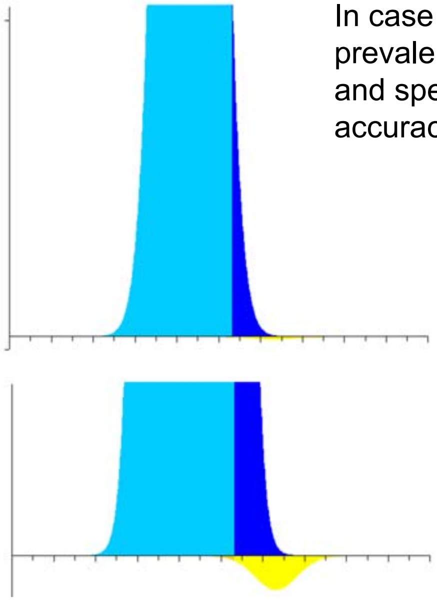
		Test	
		negative	positive
Gold-standard	healthy	810	90
	diseased	10	90

sp = 90%

se = 90%

(de = 90%)

PPV = 50%



In case of very small prevalence a highly sensitive and specific test have low accuracy (PPV).

prevalence = 0.1 %

sensitivity = 98 %

specificity = 98 %

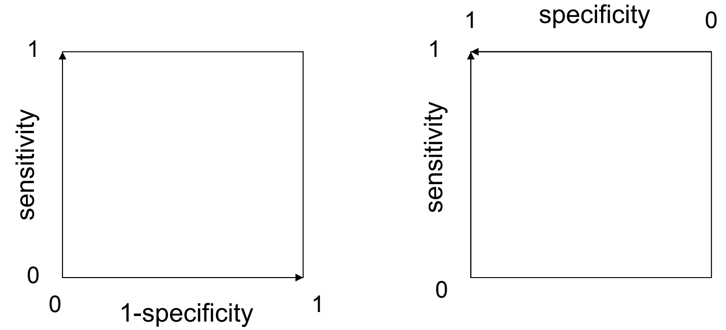
PPV = 4 %

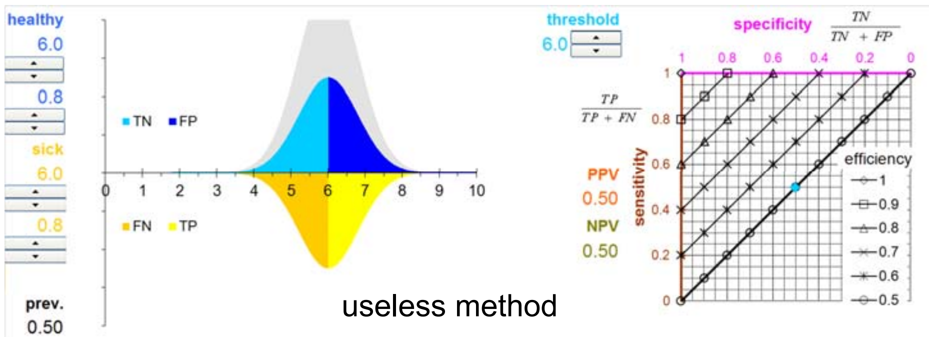
Comparison of diagnostic tests: the ROC space

ROC: receiver-operator (operating) characteristic

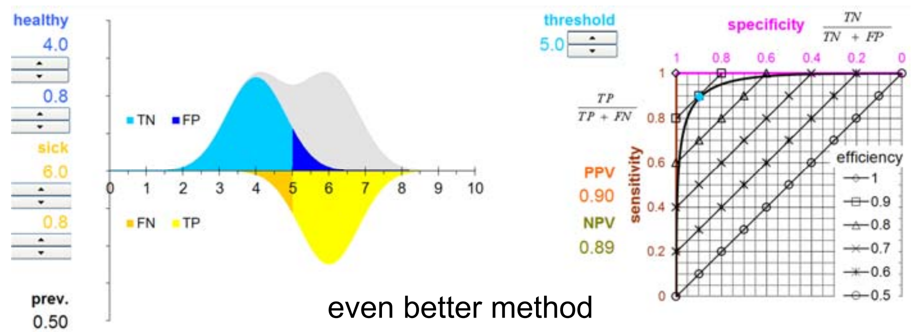
~ 1950: first ROC Analysis (receiver: Radar)

~ 1970: first medical applications

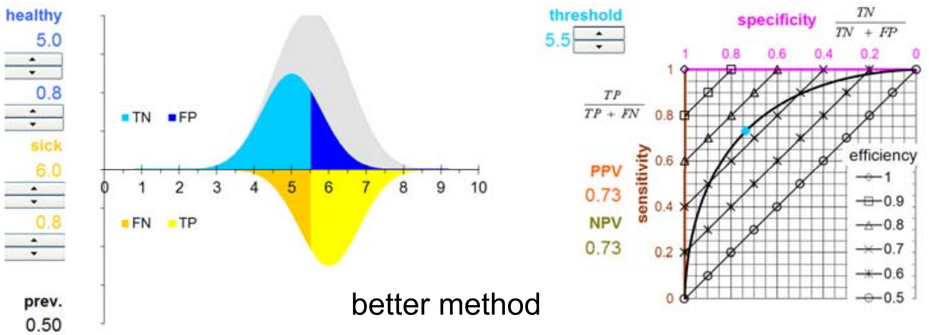




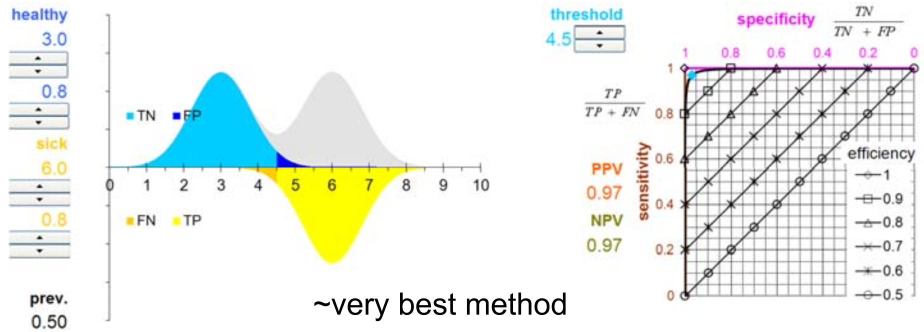
useless method



even better method



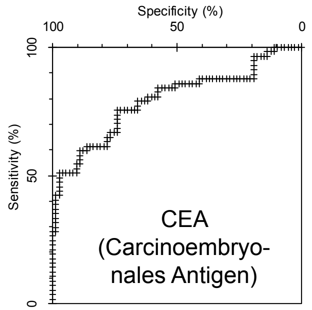
better method

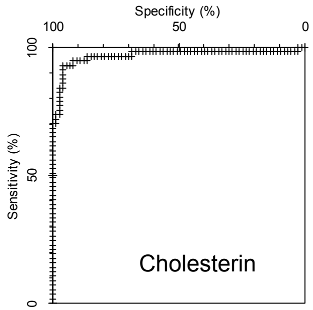


~very best method

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?

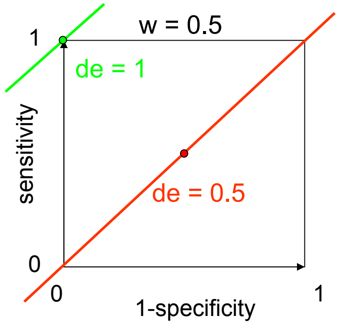
Gulyás M, Kaposi AD, Elek G, Szollár LG, Hjerpe A, Value of carcinoembryonic antigen (CEA) and cholesterol assays of ascitic fluid in cases of inconclusive cytology, J Clinical Pathology 2001 (54) 831-835

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

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

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

$$se = \underbrace{\frac{1-w}{w}}_{\text{slope}}(1-sp) + \underbrace{\frac{1}{w} de + \frac{w-1}{w}}_{\text{intercept}}$$



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

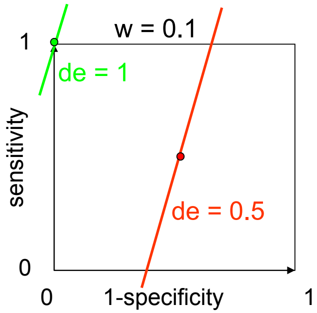
The points have the same diagnostic efficiency belong to a line with a slope of 1.

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

$$\boxed{se = \underbrace{\frac{1-w}{w}}_{\text{slope}}(1-sp) + \underbrace{\frac{1}{w} de + \frac{w-1}{w}}_{\text{intercept}}}$$

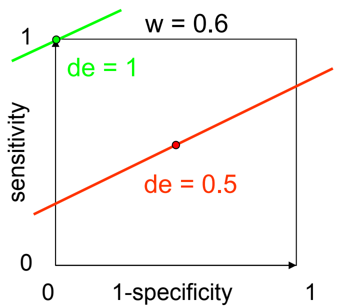
if w < 0.5: the slope of lines at identical diagnostic efficiencies is greater than 1.

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

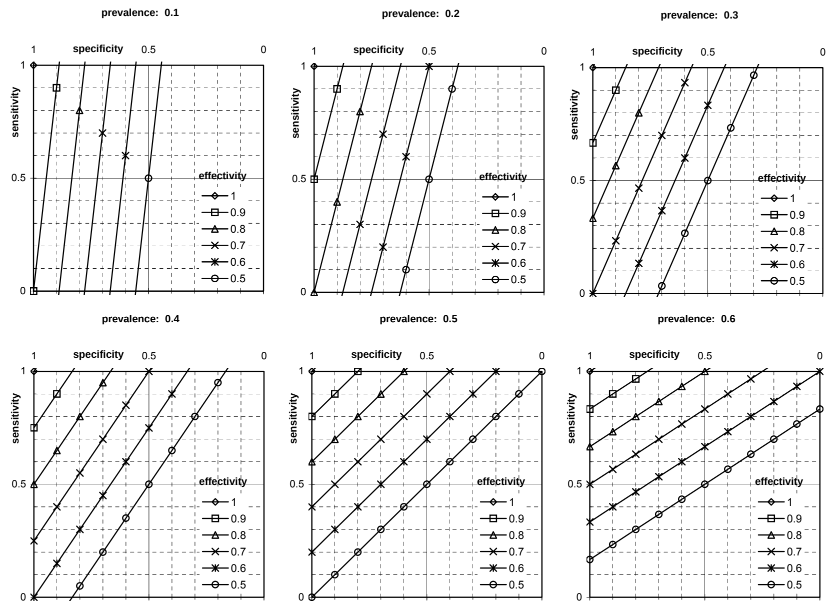


if w > 0.5: the slope of lines at identical diagnostic efficiencies is smaller than 1.

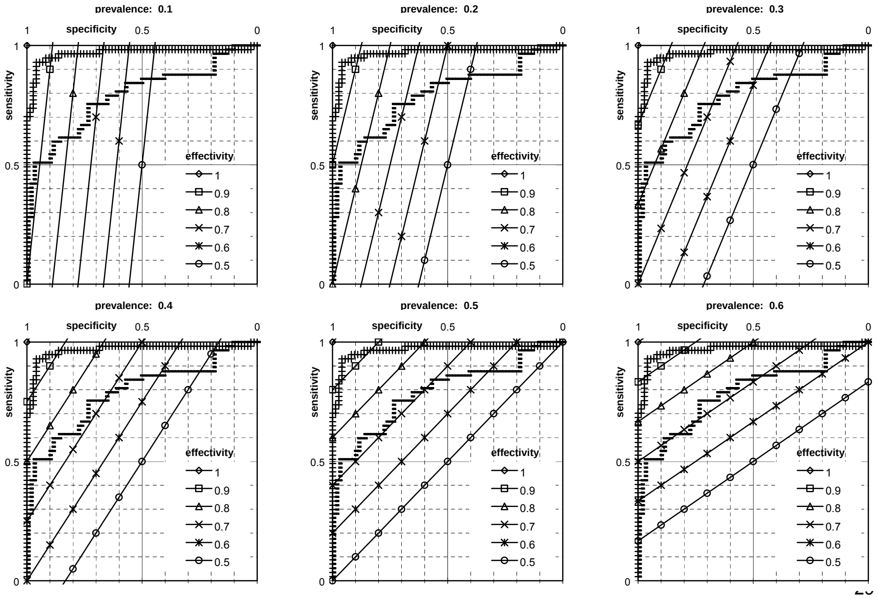
e.g. *w* = 0.6, slope: 0.66



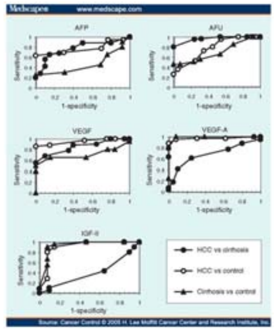
Isoeffective curves on the ROC

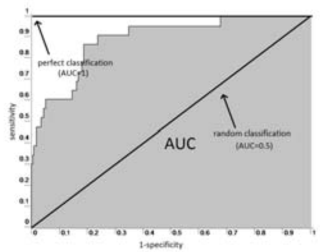


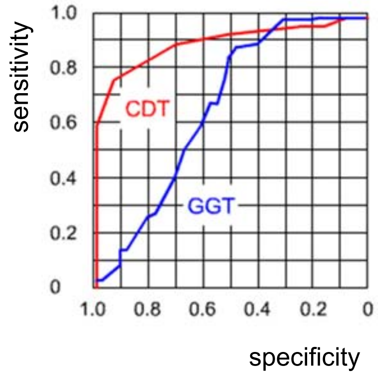
Ascites (+ Cholesterin, – CEA)



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…