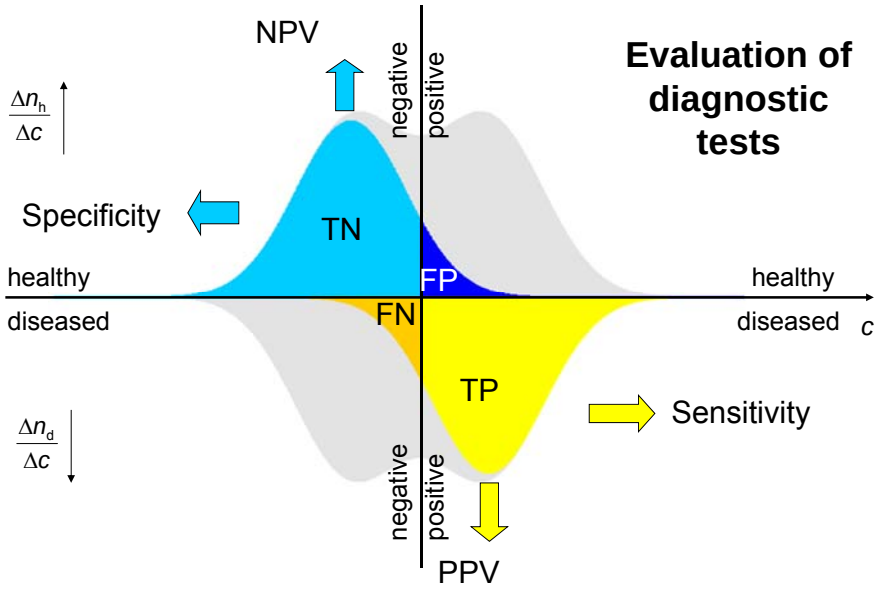


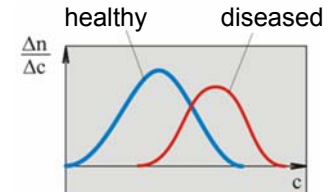
Evaluation of diagnostic tests

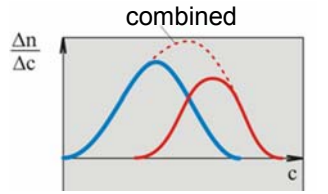
KAD 2013.11.26

Overlapping distributions

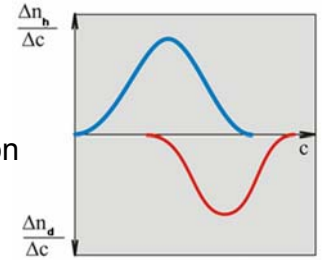
assumption:

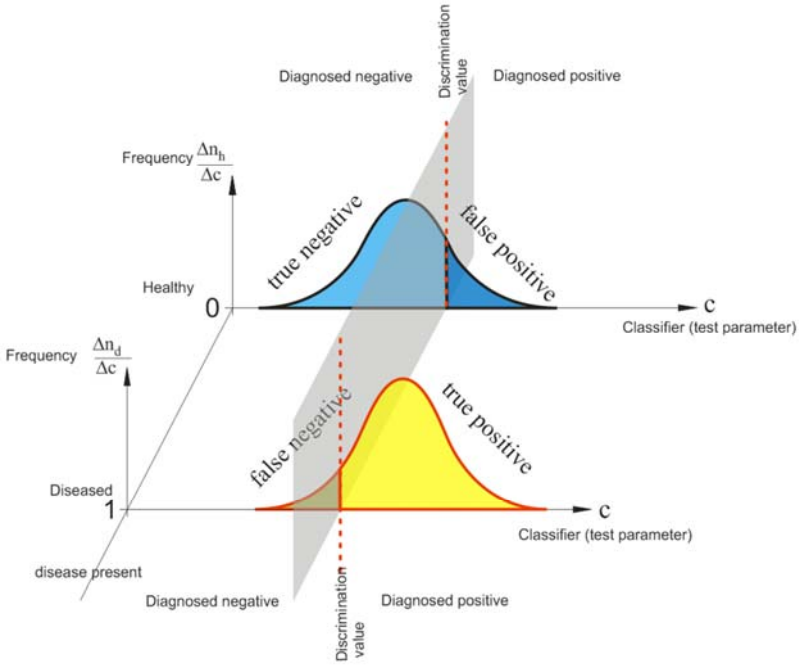
a classifier value
(e.g. serum concentration)
changes (e.g. increases)
in diseased subpopulation

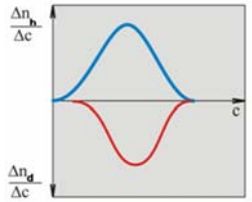




novel
representation



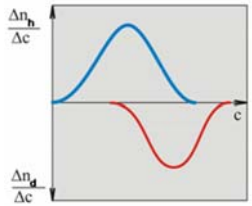




full
overlap

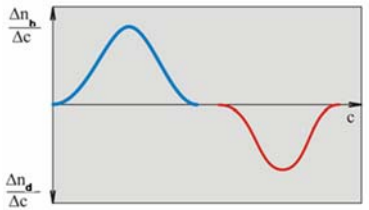
Based on overlap magnitude:

useless method



partial
overlap

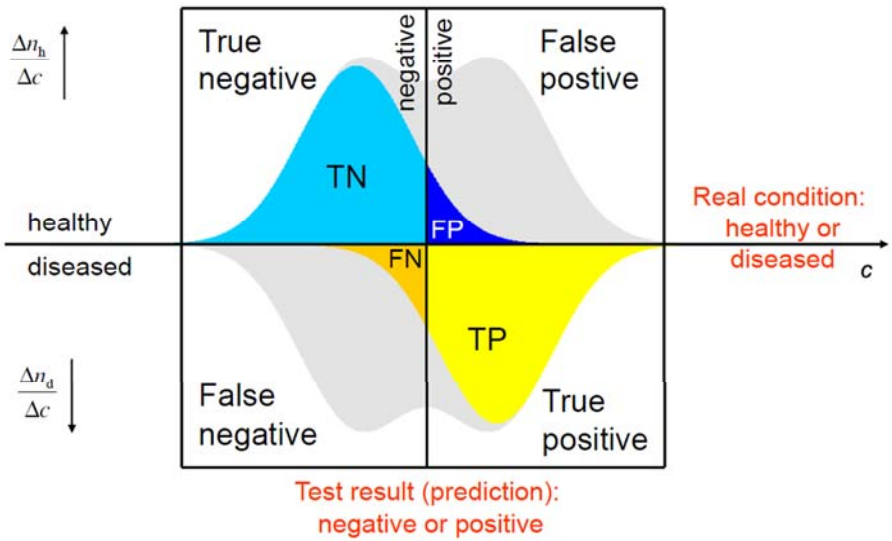
real-life situation



complete
separation

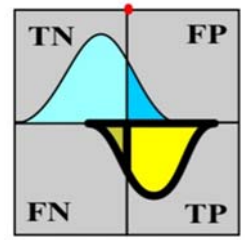
perfect method

Contingency table:
Confusion matrix (binary classification)

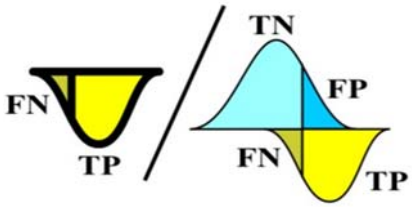


Prevalence

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



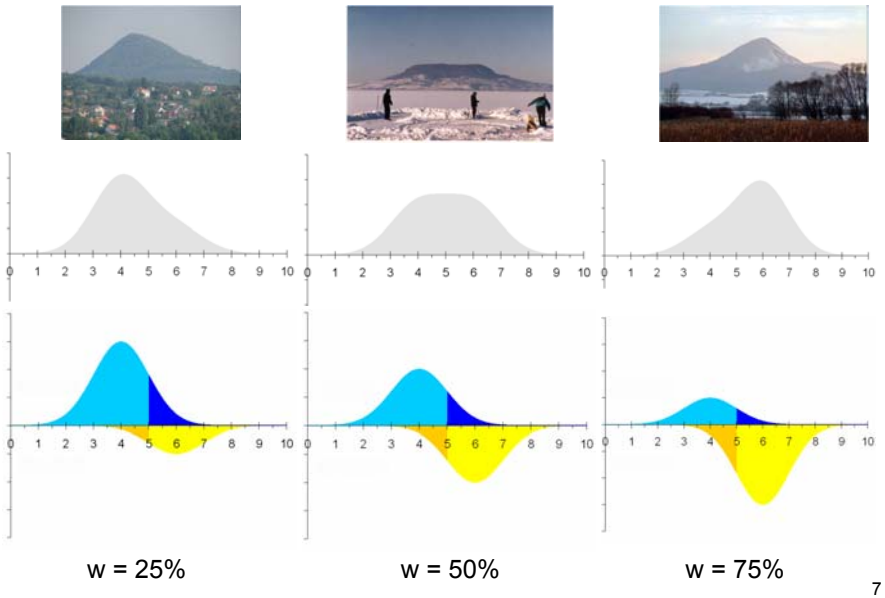
Prevalence: measure of how common the disease is; frequency of diseased in examined population, probability prior to test (a-priori-probablity),



$$w = \frac{TP + FN}{TP + TN + FN + FP}$$

$$w = \frac{\text{diseased}}{\text{total}}$$

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

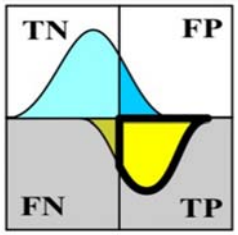
- Sensitivity
- Specificity
- PPV, relevance
- NPV, segregation

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

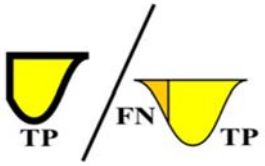
Gold standard: method known to always work; often autopsy

Diagnostic
sensitivity

(se)



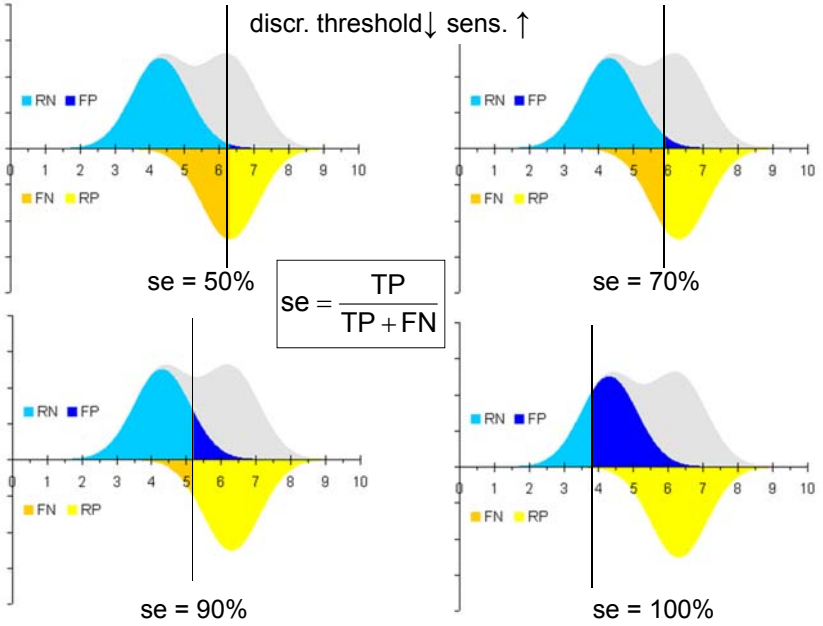
Probability that the test finds the diseased positive;
positive within diseased



$$se = \frac{\text{true positive}}{\text{diseased}} = \frac{TP}{TP + FN}$$

Large-sensitivity tests are required:

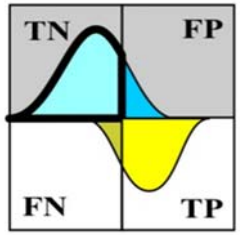
In early diagnosis (screening) so that few patients remain unrecognized.
If the risk of disease is greater than the risk of treatment.



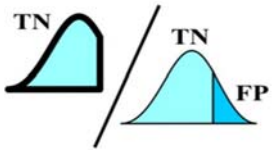
$$se = \frac{TP}{TP + FN}$$

Diagnostic
specificity

(sp)



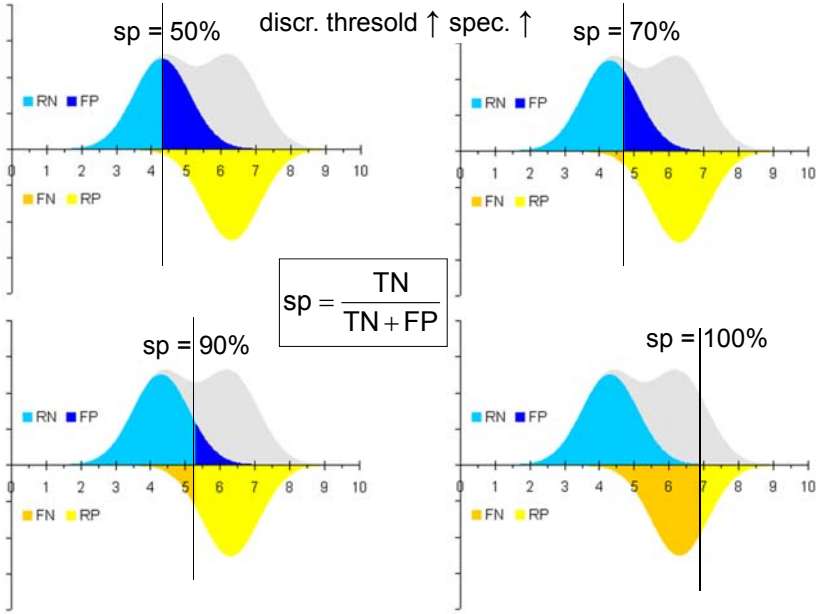
Probability that the test finds a healthy negative;
negative among healthy



$$sp = \frac{\text{true negative}}{\text{healthy}} = \frac{TN}{TN + FP}$$

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.

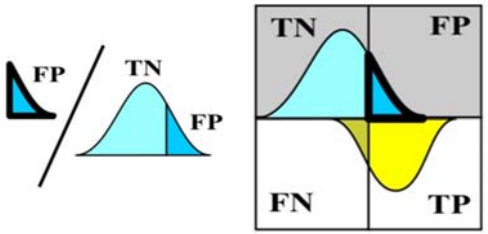


$$sp = \frac{TN}{TN + FP}$$

Diagnostic
False Positive Rate

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

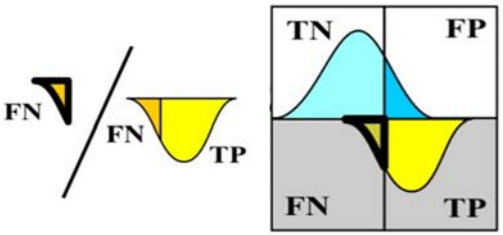
(Type-I error)



Diagnostic
False Negative Rate

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

(Type-II error)



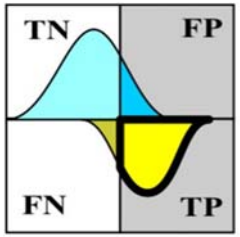
Predictive values (vertical rates)

a-posteriori-probabilities; they depend on prevalence

Diagnostic **relevance**

PPV,
positive predictive value

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

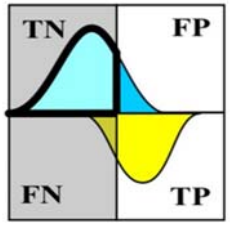


probablity of disease if test is positive,
diseased among positive

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

Diagnostic **segregation**

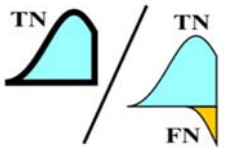
NPV,
negative predictive value



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

probabilty of healthiness if test is negative;

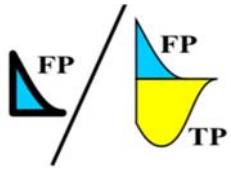
healthy among positives

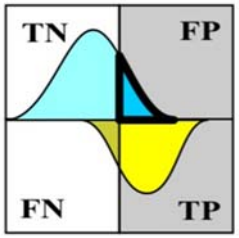


$$PVN = \frac{\text{true negative}}{\text{negative}} = \frac{TN}{TN + FN}$$

False alarm rate

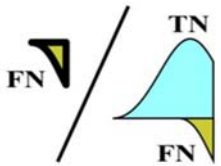
$$1\text{-}PPV = \frac{FP}{FP + TP}$$

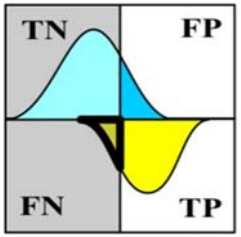




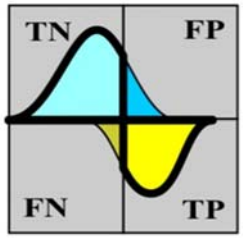
False reassurance rate

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





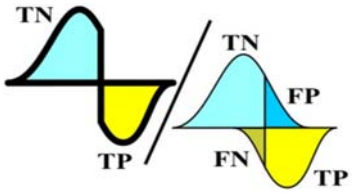
Diagnostic **efficacy/efficiency**



accuracy

correct classification rate

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



$$de = \frac{\text{true positive} + \text{true negative}}{\text{healthy} + \text{diseased}}$$

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

often: discrimination thresold is chosen so that de is maximized

Effect of prevalence

Case 1: ***w* = 50%**

			NPV = 90% test: negative	positive
SP = 90%	Gold standard	healthy	90	10
		diseased	10	90

Sensitivity (TPR) = 90%

Precision, PPV = 90%

(ACC, de = 90%)

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

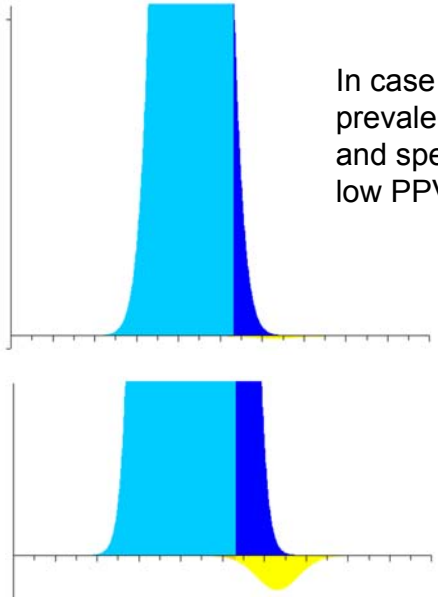
			NPV = 99% test: negative	positive
SP = 90%	Gold standard	healthy	810	90
		diseased	10	90

Sensitivity (TPR) = 90%

(ACC, de = 90%)

PPV = 50%

Effect of prevalence



In case of very small prevalence a highly sensitive and specific test could be of low PPV.

prevalence = 0.1 %

sensitivity = 98 %

specificity = 98 %

↓

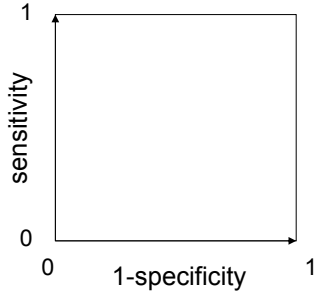
PPV = 4 %

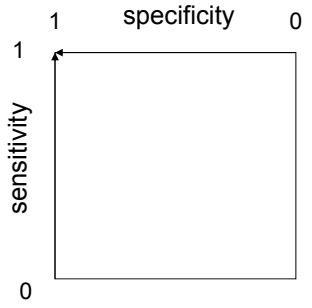
Comparison of diagnostic tests: the ROC space

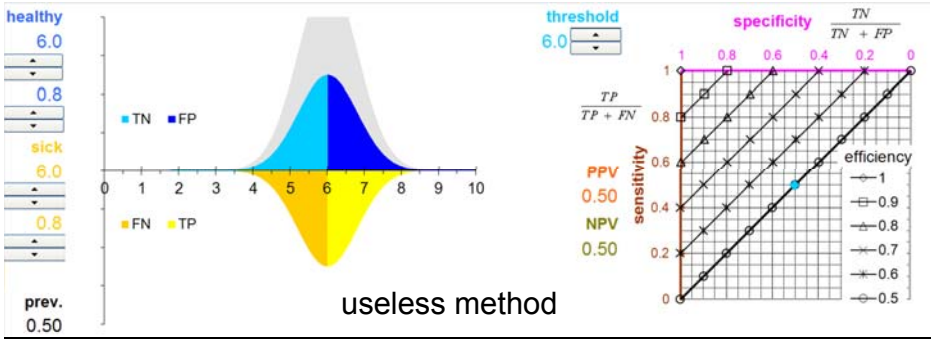
ROC: receiver-operator (operating) characteristic

~ 1950: first ROC Analysis (receiver: Radar)

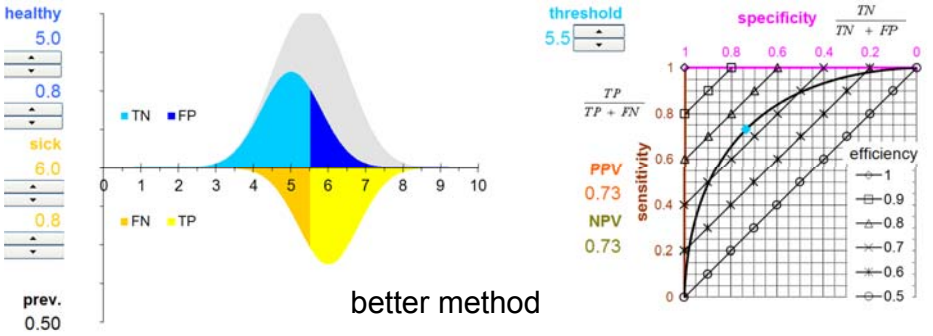
~ 1970: first medical applications



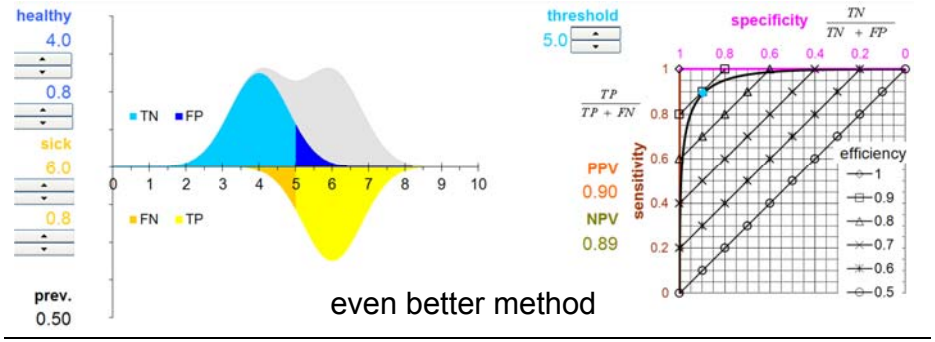




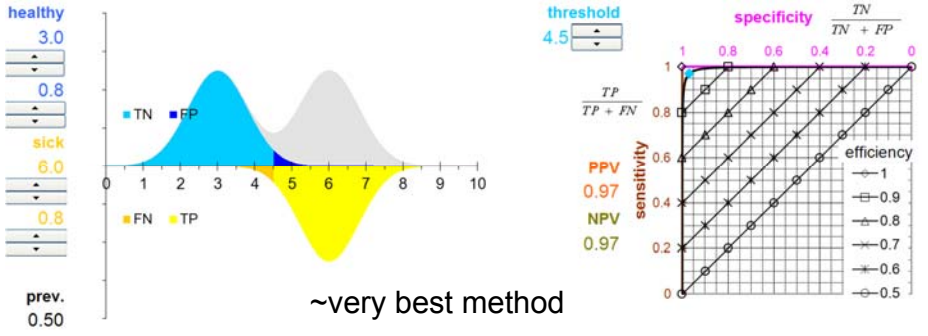
useless method



better method



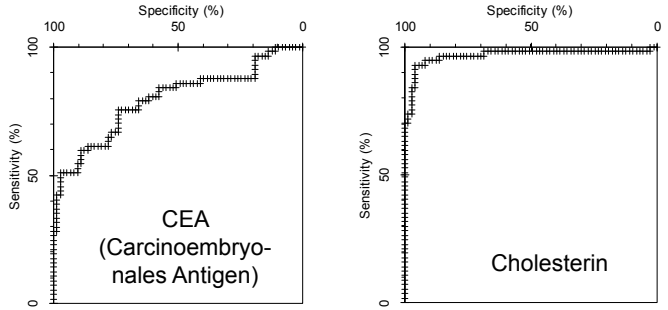
even better method



~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

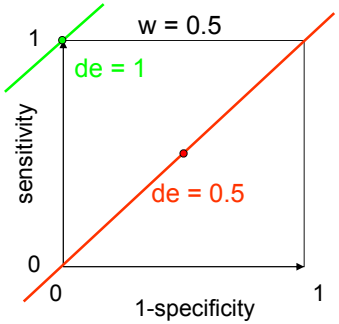
$$de = se \cdot w + sp \cdot (1-w)$$

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

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

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

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.

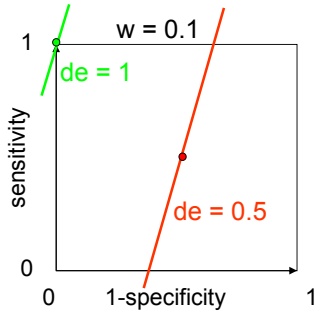
$$se = \frac{1-w}{w}(1-sp) + \frac{1}{w}de + \frac{w-1}{w}$$

slope intercept

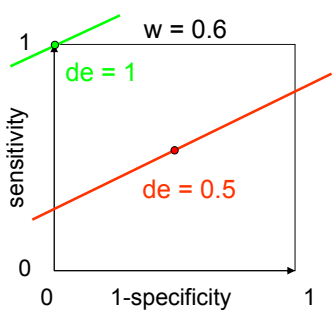
if w < 0.5: the slope of lines at identical diagnostic efficiencies is greater than 1.

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

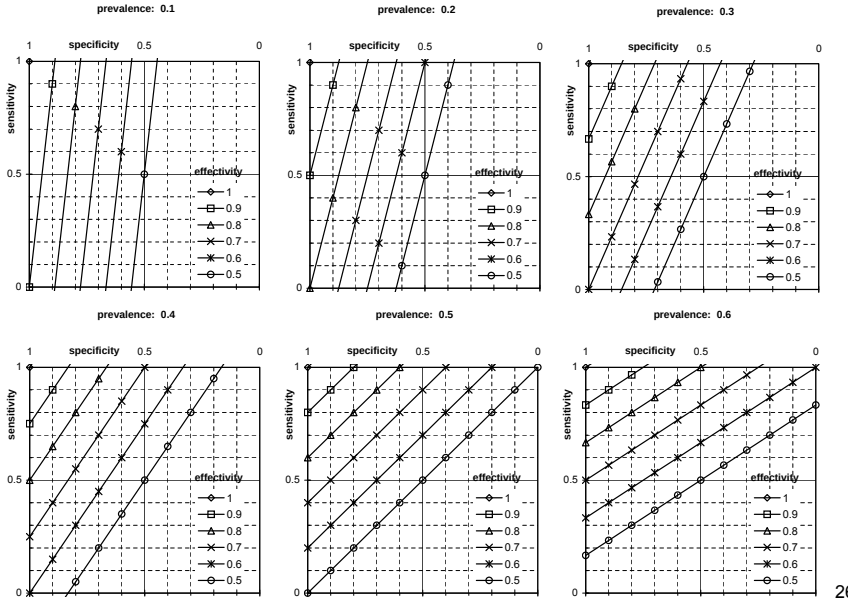
e.g. *w* = 0.1, slope: 9



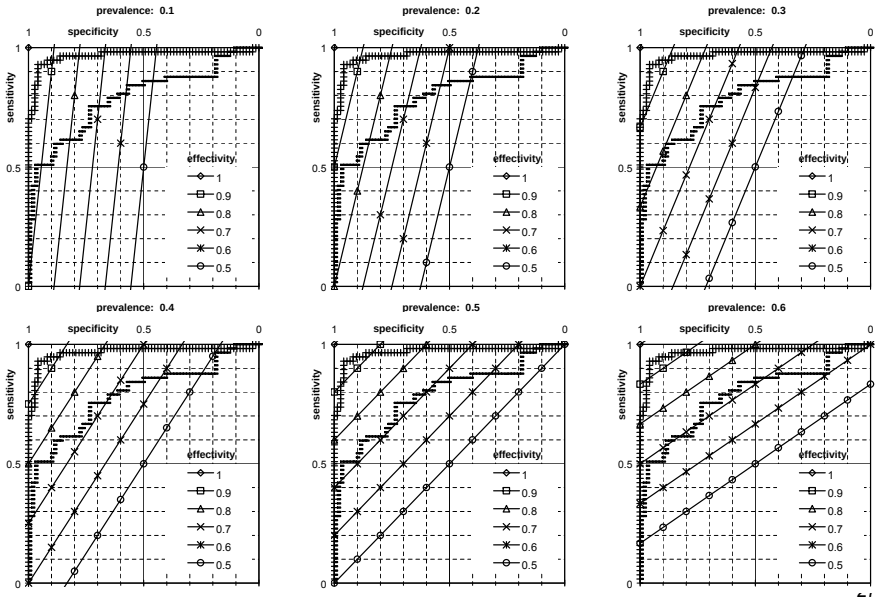
e.g. *w* = 0.6, slope: 0.66



Isoeffective curves on the ROC



Ascites (+ Cholesterin, – CEA)



Twist in the ROC space...

Alcoholism diagnostics with CDT (carbohydrate deficient transferrin) and γ-GT (gamma-Glutamyltransferase)