Evaluation of diagnostic tests

Biostatistics and informatics

Miklós Kellermayer

1

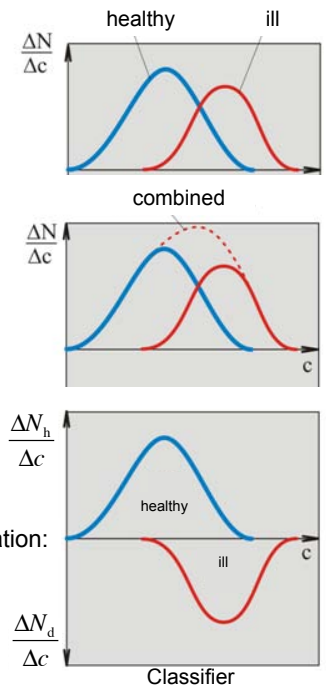
Overlapping distributions

Assumption:

A classifier value (e.g., diagnostic parameter, a measurable quantity, e.g., serum concentration) changes (e.g., increases) in disease.

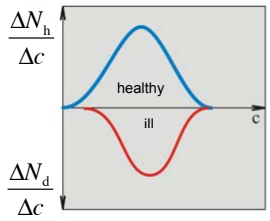
Diagnostic objective:

Predict the outcome (healthy versus ill) based on the classifier value.

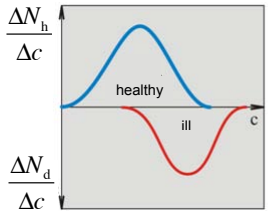


Novel representation:

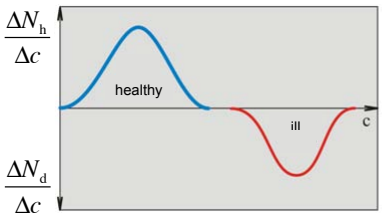
2



Full overlap



Partial overlap



Complete separation

Based on overlap magnitude:

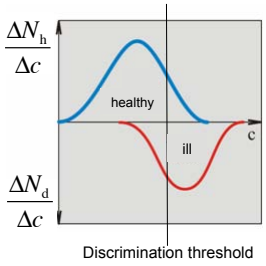
Useless method

Real-life situation

Perfect method

3

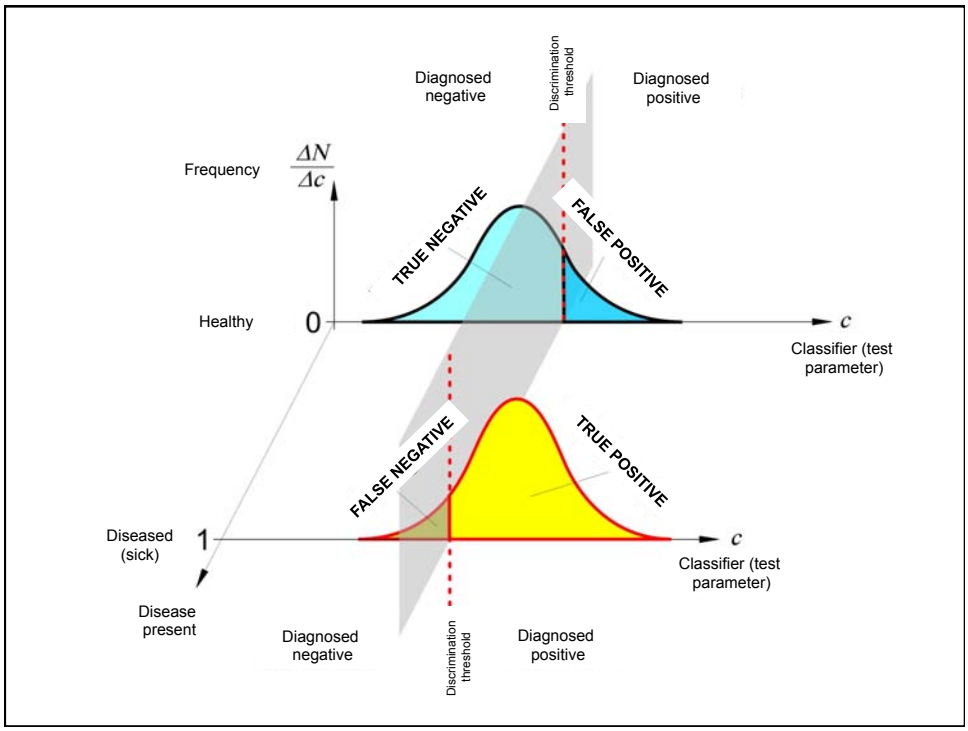
Real-life situation: Partial overlap



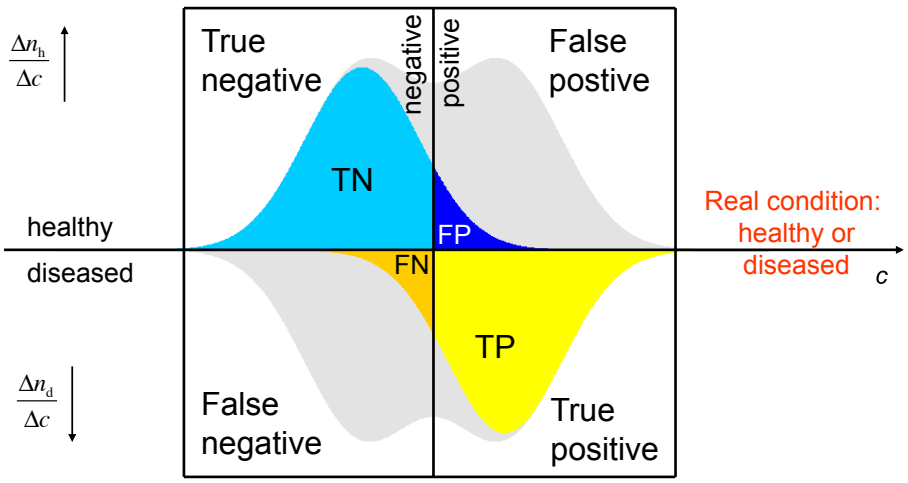
Actual value of classifier (diagnostic method)

Prediction	**N**egative	**P**ositive
Healthy	True Negative (TN) („correct rejection")	False Positive (FP) („false alarm", Type I error)
Diseased (sick)	False Negative(FN) („miss", Type II error)	True Positive (TP) („hit")

4



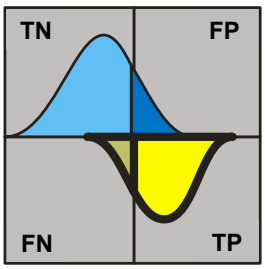
Contingency table: Confusion matrix (binary classification)



6

Prevalence

-Measure of how common the disease is
-Probability prior to test (*a priori* probability)



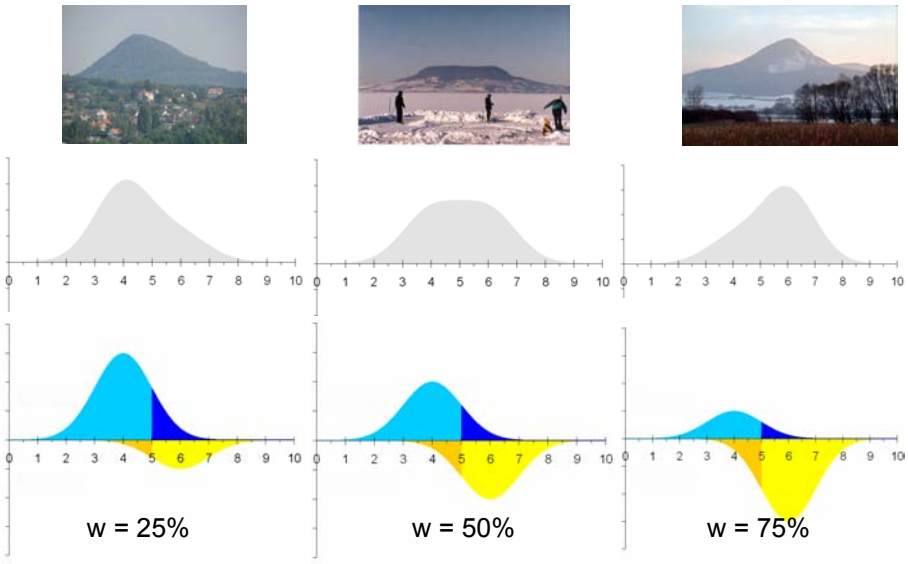
Frequency of diseased in examined population

$$\frac{\text{▽}}{\text{▽ ▲}} = w = \frac{diseased}{\text{total}} = \frac{FN + TP}{TN + FP + FN + TP} = \frac{ACC - SPC}{TPR - SPC}$$

ACC = accuracy
SPC = specificity
TPR = true positive rate (sensitivity)

7

Shape of combined distributions



8

Important parameters of diagnostic „goodness"

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

1. True Positive Rate, TPR (sensitivity)
2. True Negative Rate, TNR (specificity, SPC)
3. Positive Predictive Value, PPV (precision, diagnostic relevance)
4. Negative Predictive Value (diagnostic segregation)

Only three of them are independent!

Every method must be compared with a reference („Gold standard")

Gold standard: method known to work; often autopsy

9

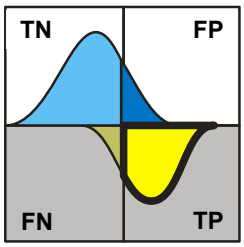
***A priori* (before test) probabilities are independent of prevalence**

Diagnostic **sensitivity**

-True Positive Rate (TPR)

-Hit Rate

-Recall



Probability that the test finds the diseased positive.

Positive within diseased.

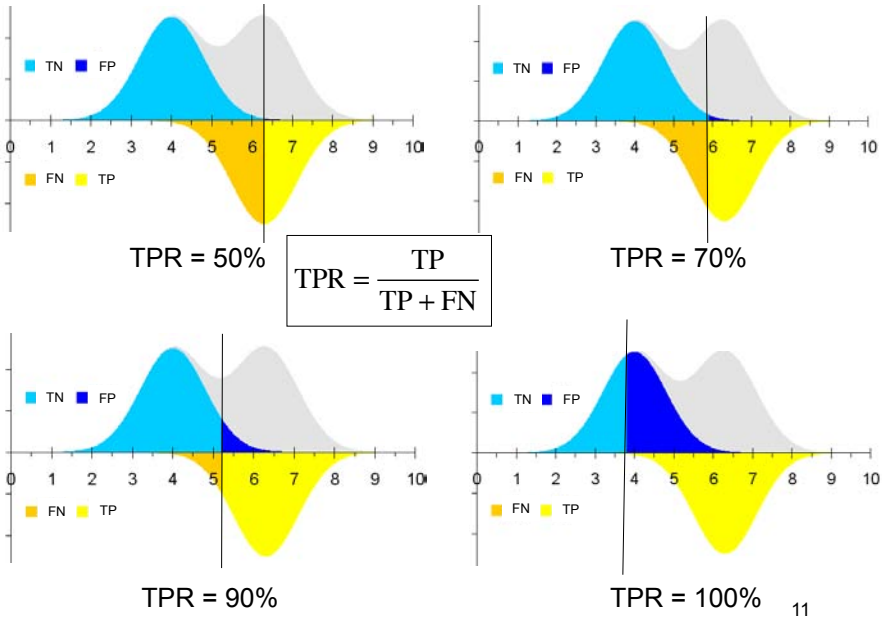
$$P(\text{positive}\,|\,\text{diseased})$$

$$\frac{\text{TP}}{\text{diseased}} = \text{TPR} = \frac{\text{TP}}{\text{FN} + \text{TP}}$$

Large-sensitivity tests (100%) are required in early diagnosis (screening) so that few patients remain unrecognized.

10

Discrimination threshold ↓ sensitivity ↑



TPR = 50%

TPR = 70%

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

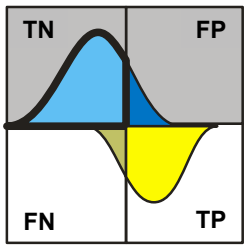
TPR = 90%

TPR = 100%

11

Diagnostic **specificity (SPC)**

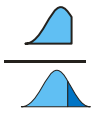
-True Negative Rate (TNR)



Probability that the test finds a healthy negative.

Negative among healthy

$P(\text{negative} \mid \text{healthy})$

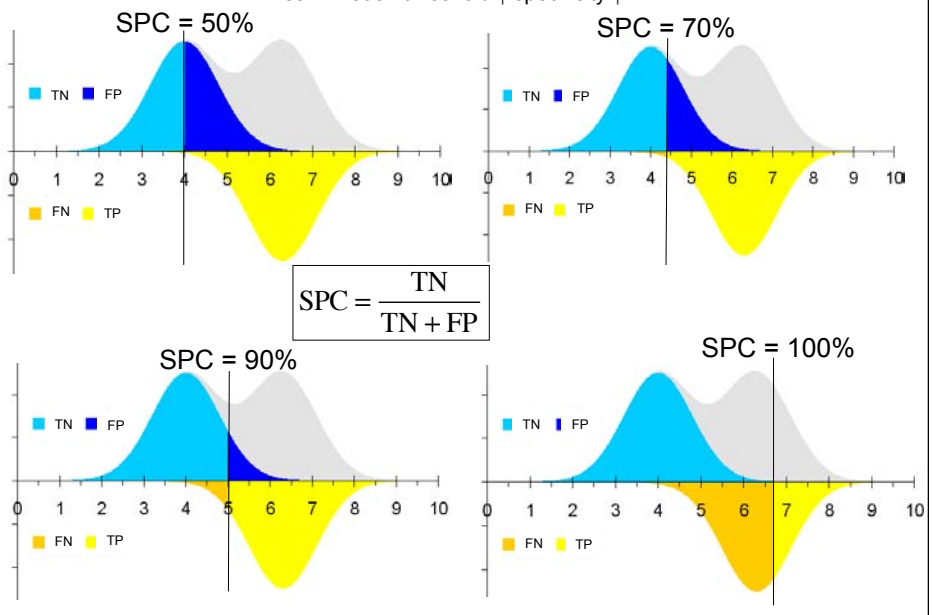


$$= \text{SPC} = \frac{\text{TN}}{\textit{healthy}} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

High-specificity tests (near 100 %) are important when the false positive values have severe consequences (e.g., surgery).

12

Discrimination threshold ↑ specificity ↑

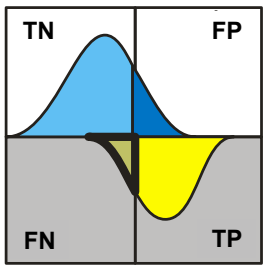


$$SPC = \frac{TN}{TN + FP}$$

13

False Negative Rate (FNR)

-Rate of Type-II error



Probability that the test finds the diseased negative.

Negative among diseased.

$P(\text{negative} \mid \text{diseased})$

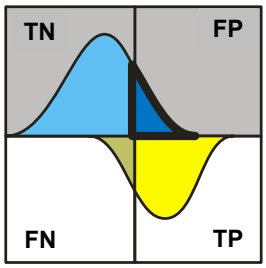
$$\frac{\text{FN}}{\text{diseased}} = 1 - \text{TPR} = \frac{\text{FN}}{\text{diseased}} = \frac{\text{FN}}{\text{FN} + \text{TP}}$$

TPR = True Positive Rate, sensitivity

14

False Positive Rate (FPR)

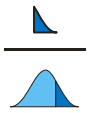
-Rate of Type-I error



Probability that the test finds a healthy positive.

Positive among healthy.

$P(\text{positive}|\text{healthy})$



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

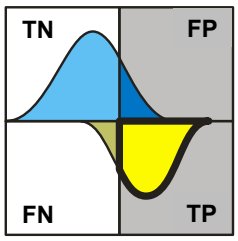
15

***A posteriori* (after test) probabilities depend strongly on prevalence**

Diagnostic **precision**

-Positive Predictive Value (PPV)

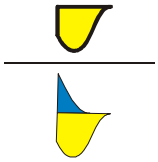
-Relevance



Probability of diseas if test is positive.

Diseased among positive.

$P(\text{diseased}|\text{positive})$



$$= \text{PPV} = \frac{\text{TP}}{\text{total positive}} = \frac{\text{TP}}{\text{FP} + \text{TP}} = \frac{\text{TPR} \cdot \text{w}}{\text{TPR} \cdot \text{w} + (1 - \text{SPC}) \cdot (1 - \text{w})}$$

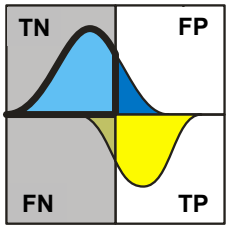
SPC = specificity
TPR = True Positive Rate, sensitivity

16

Negative Predictive Value (NPV)

-Correct negativity

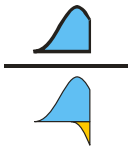
-Segregation



Probability of healthiness if test is negative.

Healthy among negative

$P(\text{healthy}\,|\,\text{negative})$



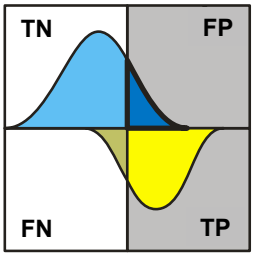
$$= \text{NPV} = \frac{\text{TN}}{\text{total negative}} = \frac{\text{TN}}{\text{FN} + \text{TN}} = \frac{\text{SPC} \cdot (1 - \text{w})}{\text{SPC} \cdot (1 - \text{w}) + (1 - \text{TPR}) \cdot \text{w}}$$

SPC = specificity
TPR = True Positive Rate, sensitivity

17

False Discovery Rate (FDR)

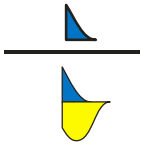
-False alarm rate



Probability of healthiness if test is positive.

Healthy among positive.

$P(\text{healthy}\,|\,\text{positive})$



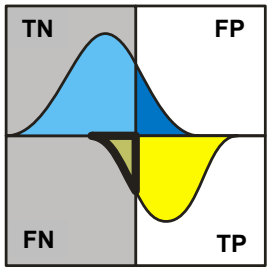
$$= 1 - \text{PPV} = \frac{\text{FP}}{\text{total positive}} = \frac{\text{FP}}{\text{FP} + \text{TP}}$$

PPV = Positive Predictive Value (precision)

18

False Reassurance Rate (FRR)

-False reassurance rate



Probability of disease if test is negative.

Diseased among negative.

$P(\text{diseased}|\text{negative})$

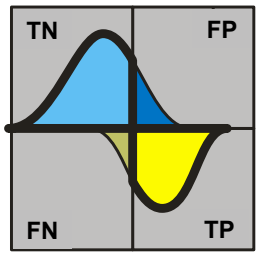
$$= 1 - \text{NPV} = \frac{\text{FN}}{\text{total negative}} = \frac{\text{FN}}{\text{FN} + \text{TN}}$$

NPV = Negative Predictive Value

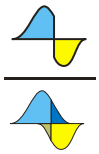
19

Diagnostic **efficiency**

-Accuracy (ACC)



Ratio of correct classification



$$= \text{ACC} = \frac{\text{TP} + \text{TN}}{\text{total}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} = \text{TPR} \cdot \text{w} + \text{SPC} \cdot (1 - \text{w})$$

SPC = specificity
TPR = True Positive Rate, sensitivity

Discrimination threshold is chosen so that accuracy is maximized.

20

Effect of prevalence

NPV = 90%

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

		test	
		negative	positive
Gold standard	healthy	90	10
	diseased	10	90

SPC = 90%

Sensitivity (TPR) = 90%

Precision, PPV = 90%

(ACC, de = 90%)

NPV = 99%

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

		test	
		negative	positive
Gold standard	healthy	810	90
	diseased	10	90

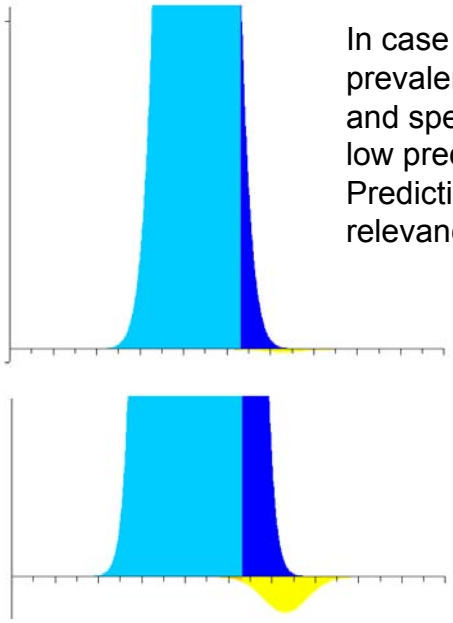
SPC = 90%

Sensitivity (TPR) = 90%

(ACC, de = 90%)

PPV = 50%

21



In case of very small prevalence a highly sensitive and specific test could be of low precision (Positive Predictive Value, PPV, relevance).

prevalence = 0.1 %

sensitivity = 98 %

specificity = 98 %

↓

precision = 4 %

22

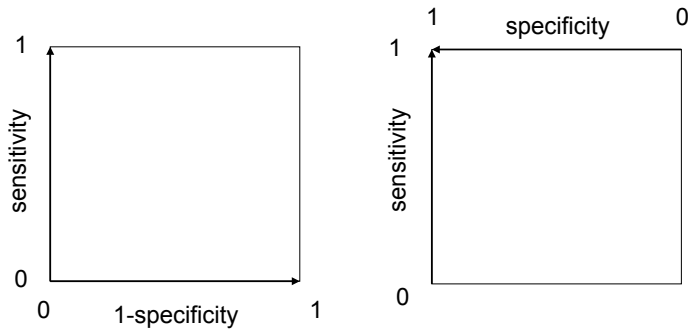
Comparison of diagnostic tests: the ROC space

ROC: Receiver Operating Characteristic

ROC curve is a graphical plot of the sensitivity (TPR) versus false positive rate (1-specificity) for a binary classifier system as its discrimination threshold is varied.

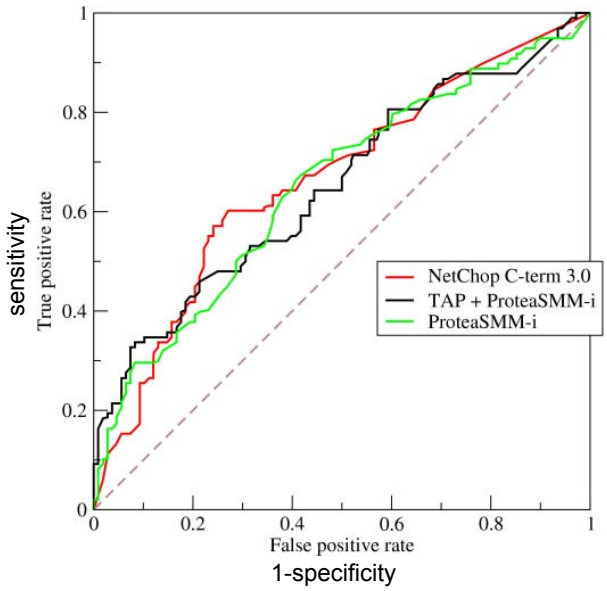
First ROC curve used in World War II for analysis of radar signals.

In the 1950s, ROC curves were employed in psychophysics to assess detection of weak signals, then later in medicine in the evaluation of diagnostic tests.



23

Example of a ROC curve



ROC curves of three epitope predictors

24

Application of the ROC space

A

TP=63	FP=28	91
FN=37	TN=72	109
100	100	200

TPR = 0.63
FPR = 0.28
ACC = 0.68

B

TP=77	FP=77	154
FN=23	TN=23	46
100	100	200

TPR = 0.77
FPR = 0.77
ACC = 0.50

C

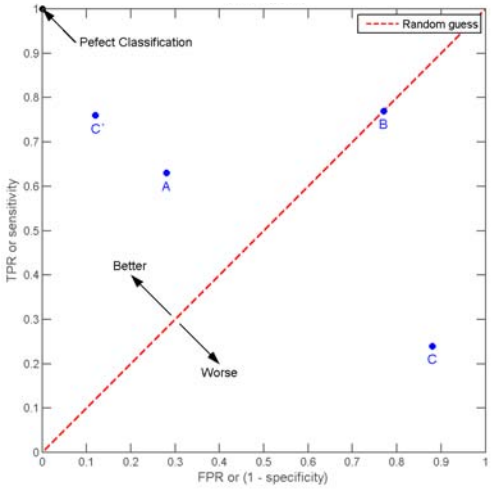
TP=24	FP=88	112
FN=76	TN=12	88
100	100	200

TPR = 0.24
FPR = 0.88
ACC = 0.18

C'

TP=76	FP=12	88
FN=24	TN=88	112
100	100	200

TPR = 0.76
FPR = 0.12
ACC = 0.82



25

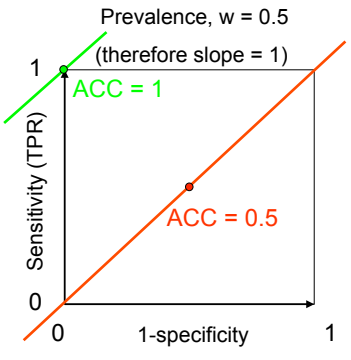
Dependence of ROC curve on diagnostic parameters I: accuracy

Equation of ROC curve:

$$TPR = \underbrace{\frac{1-w}{w}}_{\text{Slope}} \times \underbrace{(1-SPC)}_{\text{Independent variable}} + \underbrace{\frac{1}{w}ACC + \frac{w-1}{w}}_{\text{y-intercept}}$$

Dependent variable

TPR = Tre Positive Rate, senstivity
w = prevalence
SPC = specificity
ACC = Accuracy (diagnostic efficiency)



Increasing accuracy increases y-intercept, hence improves classification.

26

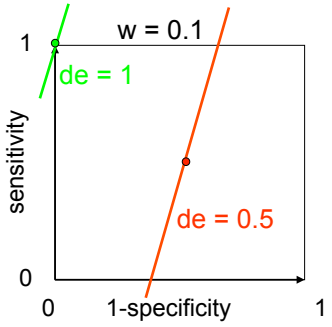
Dependence of ROC curve on diagnostic parameters II: prevalence

$$TPR = \underbrace{\frac{1-w}{w}}_{\text{Slope}} \times \underbrace{(1-SPC)}_{\text{Independent variable}} + \underbrace{\frac{1}{w}ACC + \frac{w-1}{w}}_{\text{y-intercept}}$$

Dependent variable

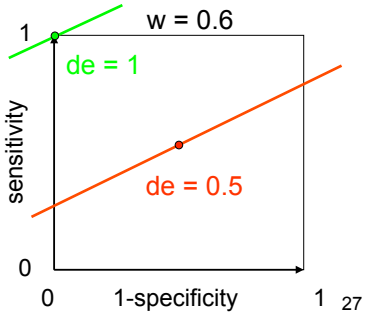
If w < 0.5, at identical accuracies the slope is greater than 1.

Case 1: *w* = 0.1, slope= 9



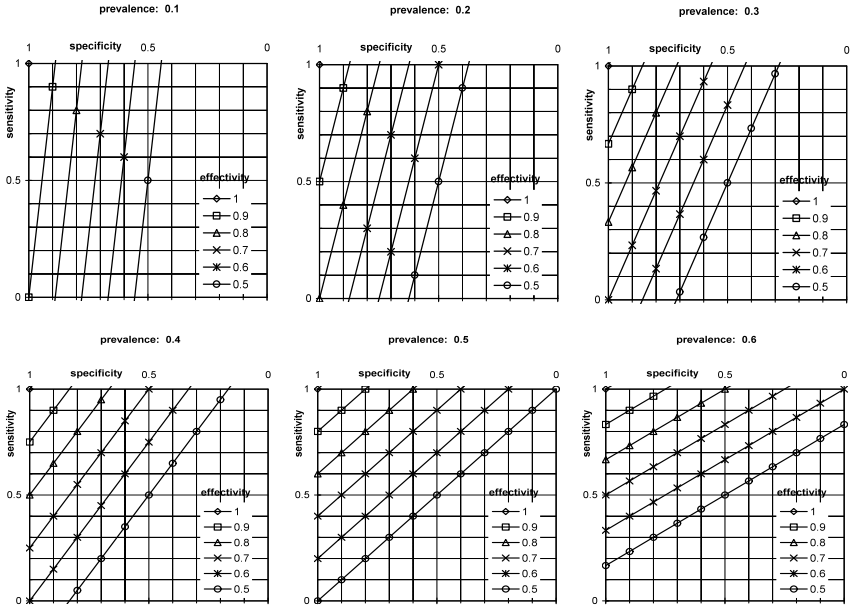
If w > 0.5, then at identical accuracies the slope is smaller than 1.

Case 2: *w* = 0.6, slope= 0.66



27

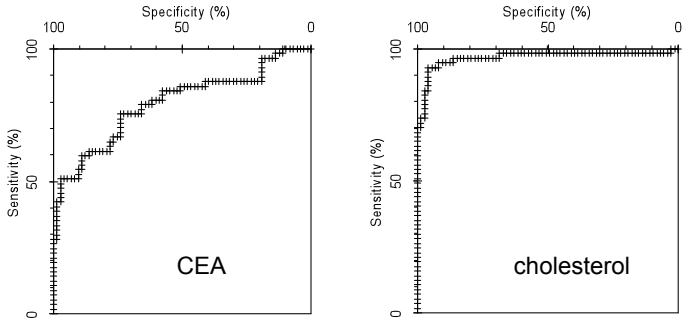
Identical accuracies define isoefficiency curves in the ROC space



28

Comparison of diagostic methods in ascites

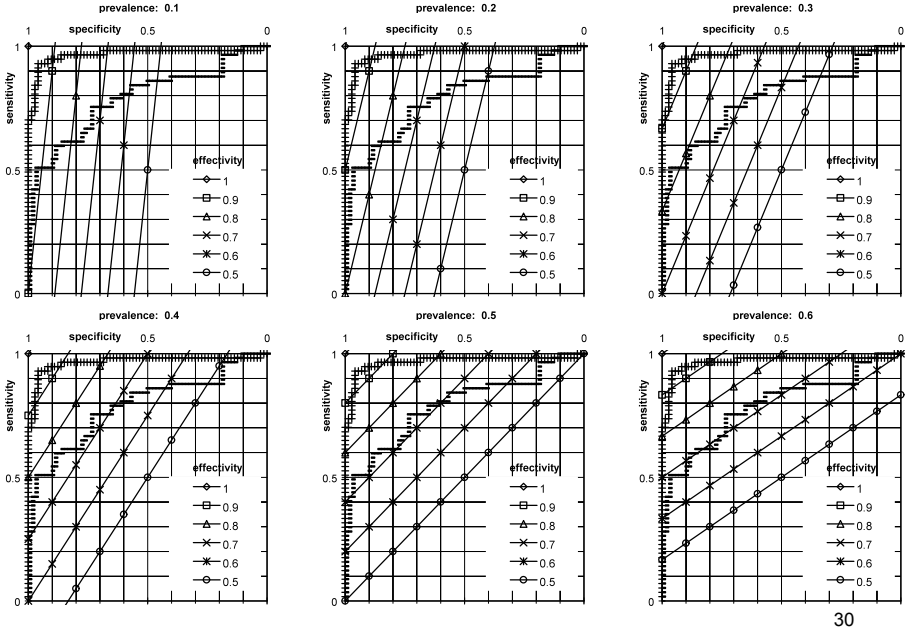
Carcinoembryonic antigen (CEA) and cholesterol are both increased in the ascites in carcinosis, raising the possibility of using either of them as diagnostic tools. Which one is better? What discrimination threshold should be used?



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

29

Cholesterol level in ascites provides better test across all prevalences

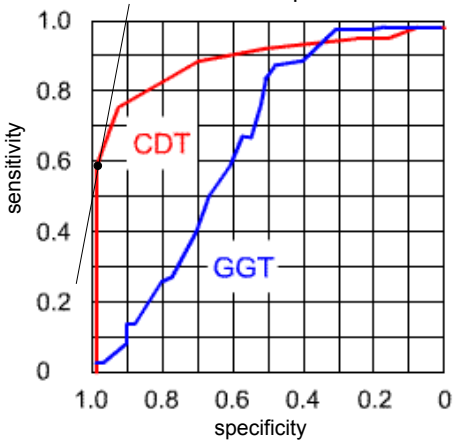


30

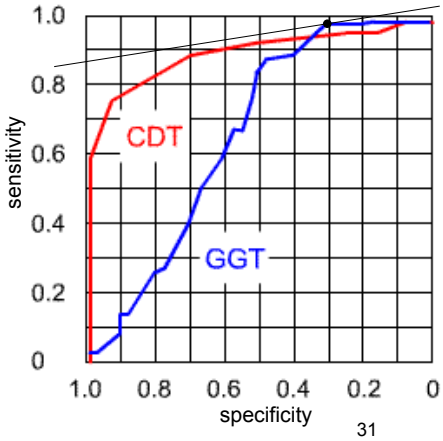
Twist in the ROC space...

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

CDT is better at low prevalences.



In case of high prevalences, the GGT test excels.



31