Evaluation of diagnostic tests

Biostatistics and informatics

Miklós Kellermayer

1

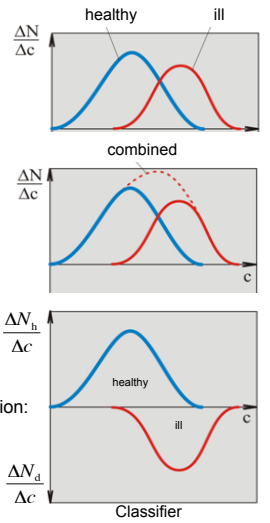
Overlapping distributions

Assumption:

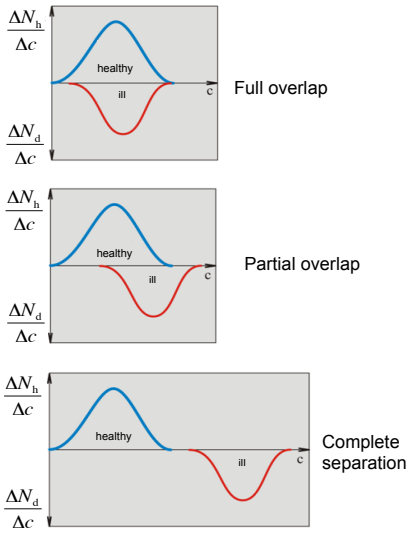
A classifier value (e.g., diagnostic parameter, a measurable quantity, e.g., serum concentration) changes (e.g., increases) in disease.

Diagnostic objective:

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



2



Full overlap

Partial overlap

Complete separation

Based on overlap magnitude:

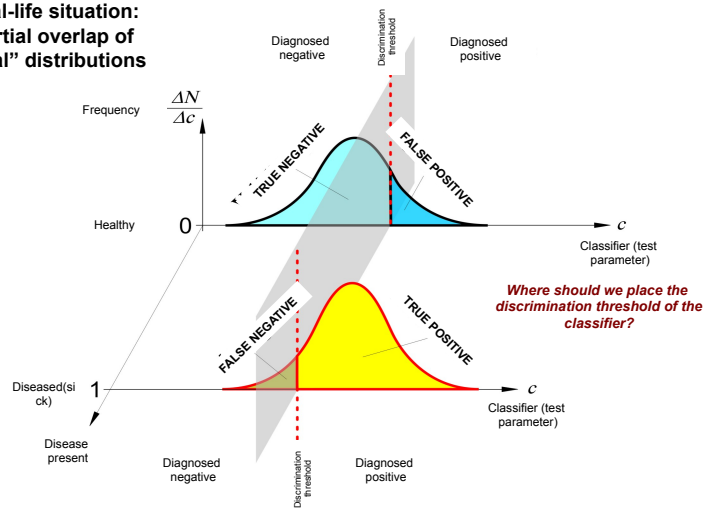
Useless method

Real-life situation

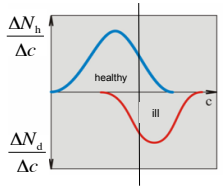
Perfect method

3

Real-life situation: Partial overlap of „real" distributions



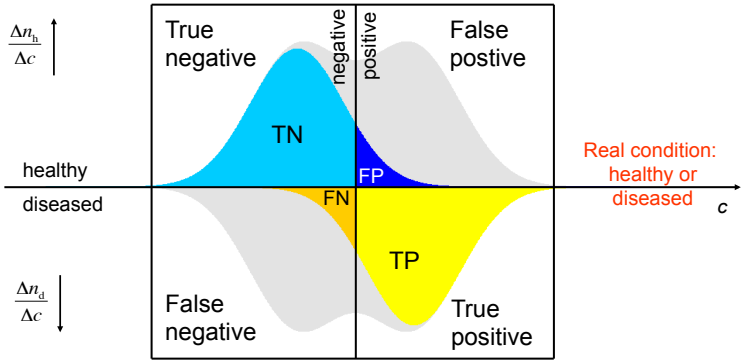
Construction of a contingency table



	Negative	Positive
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")

5

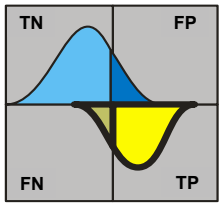
Contingency table: Confusion matrix (binary classification)



6

Parameter of the disease: Prevalence

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



Frequency of diseased in examined population

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

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

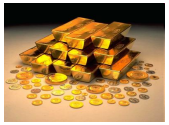
7

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***)

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



Gold standard: method **known to always work**; often autopsy

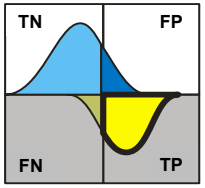
8

1. Diagnostic **sensitivity**

-True Positive Rate („TPR")

-Hit Rate

-Recall



Probability that the test finds the diseased positive.

Positive within diseased.

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

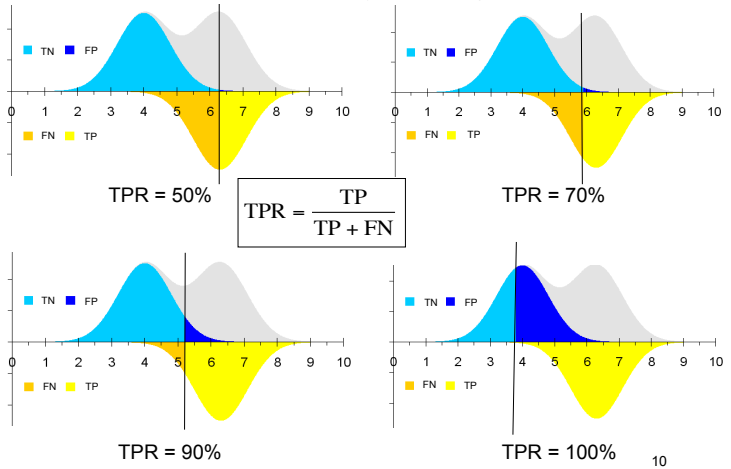
$$= \text{TPR} = \frac{\text{TP}}{\text{diseased}} = \frac{\text{TP}}{\text{FN} + \text{TP}}$$

Large-sensitivity tests (100%) are required:
-in early diagnosis (screening) so that few patients remain unrecognized.
-if the risk of disease is greater than the risk of treatment.

9

Effect of threshold on diagnostic **sensitivity**

Discrimination threshold ↓ sensitivity ↑

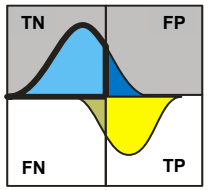


$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

10

2. Diagnostic **specificity („SPC")**

-True Negative Rate („TNR")



Probability that the test finds a healthy negative.

Negative among healthy

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

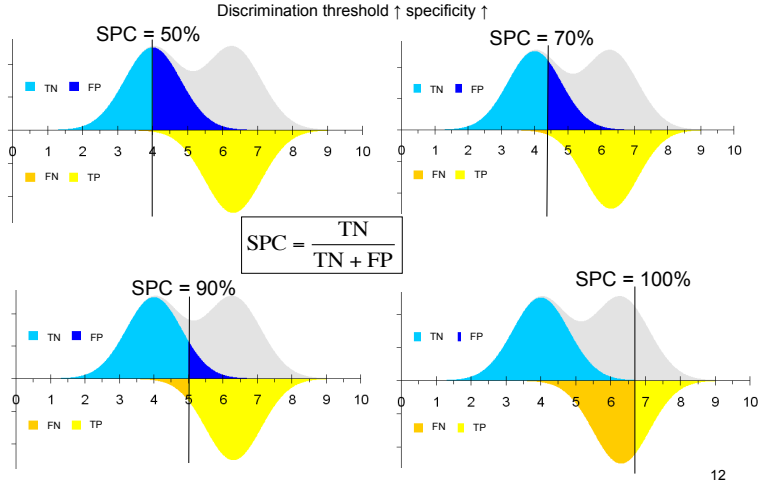
$$= \text{SPC} = \frac{\text{TN}}{\textit{healthy}} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

High-specificity tests (near 100 %) are important:
-when the false positive values have severe consequences (e.g., surgery).
-when the risk of treatment is greater than the risk of disease.

11

Effect of threshold on diagnostic **specificity**

Discrimination threshold ↑ specificity ↑



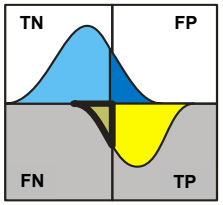
$$\text{SPC} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

12

Some further *a priori* parameters....

a. False Negative Rate („FNR")

-Rate of Type-II error



Probability that the test finds the diseased negative.

Negative among diseased.

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

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

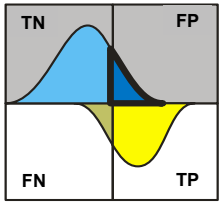
TPR = True Positive Rate, sensitivity

13

Some further *a priori* parameters....

b. False Positive Rate („FPR")

-Rate of Type-I error



Probability that the test finds a healthy positive.

Positive among healthy.

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

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

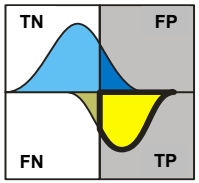
14

A posteriori probabilities – calculated following the diagnostic test
They depend strongly on prevalence

3. Diagnostic **precision**

-Positive Predictive Value (PPV)

-Relevance



Probability of disease if test is positive.

Diseased among positive.

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

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

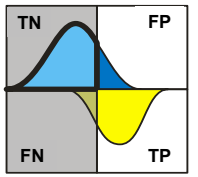
SPC = specificity
TPR = True Positive Rate, sensitivity

15

4. Negative Predictive Value („NPV")

-Correct negativity

-Segregation



Probability of healthiness if test is negative.

Healthy among negative

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

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

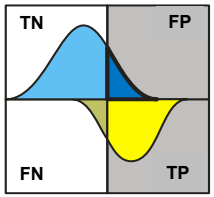
SPC = specificity
TPR = True Positive Rate, sensitivity

16

Some further *a posteriori* parameters....

a. False Discovery Rate („FDR")

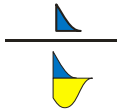
-False alarm rate



Probability of healthiness if test is positive.

Healthy among positive.

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



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

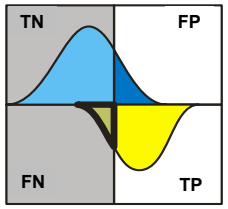
PPV = Positive Predictive Value (precision)

17

Some further *a priori* parameters....

b. False Reassurance Rate („FRR")

-False reassurance rate



Probability of disease if test is negative.

Diseased among negative.

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

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

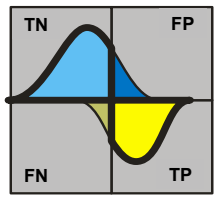
NPV = Negative Predictive Value

18

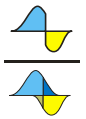
Some further *a priori* parameters....

c. Diagnostic **efficiency („de")**

-Accuracy (ACC)



Ratio of correct classification



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

SPC = specificity
TPR = True Positive Rate, sensitivity

Discrimination threshold is chosen so that accuracy is maximized.

19

Effect of prevalence – influences a posteriori probabilities

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

NPV = 90%

		test: negative	test: positive
Gold standard	healthy	90	10
	diseased	10	90

SPC = 90%

Sensitivity (TPR) = 90%

(ACC, de = 90%)

Precision, PPV = 90%

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

NPV = 99%

		test: negative	test: positive
Gold standard	healthy	810	90
	diseased	10	90

SPC = 90%

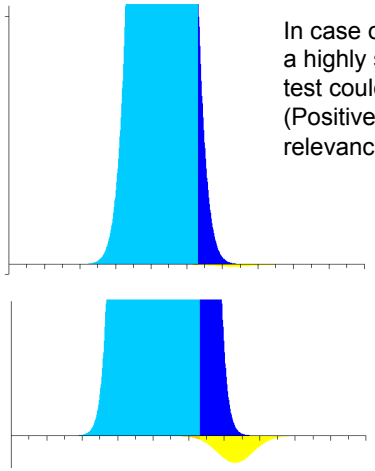
Sensitivity (TPR) = 90%

(ACC, de = 90%)

PPV = 50%

20

Effect of prevalence



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 %

21

Comparison of diagnostic tests: the ROC space

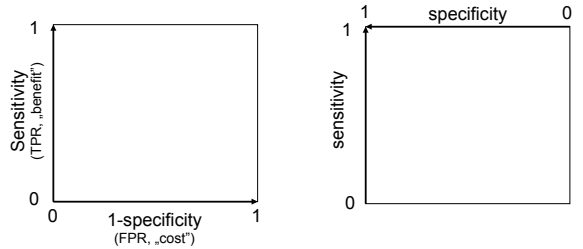
ROC: Receiver Operating Characteristic

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

Related to "cost-benefit analysis".

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.



22

Application of the ROC space

(Example of discrete classifier)

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

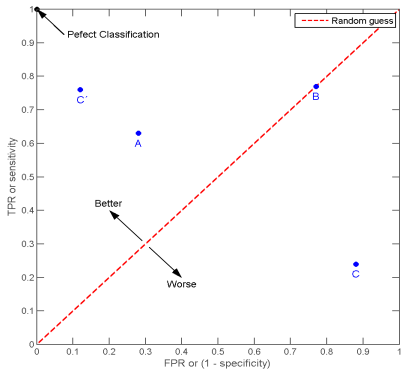
TP=24	FP=88	112
FN=76	TN=12	88
100	100	200

TPR = 0.24
FPR = 0.88
ACC = 0.18

C'

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

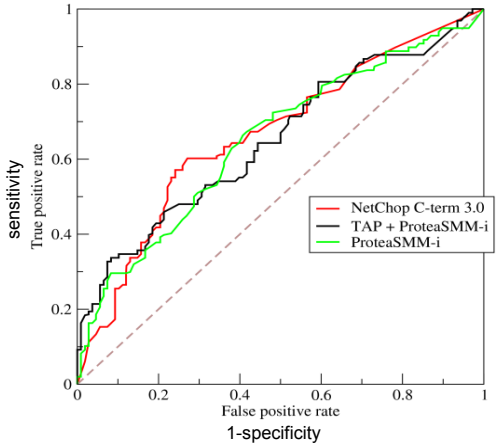
TPR = 0.76
FPR = 0.12
ACC = 0.82



23

Application of the ROC space: the ROC curve

(Example of continuus classifier)



ROC curves of three epitope (antigenic determinant) predictors.

Question:
Which method has the largest predictive power?
This depends on *accuracy* and *prevalence*.

24

Dependence of ROC curve on diagnostic parameters I: accuracy

Equation of ROC curve:

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

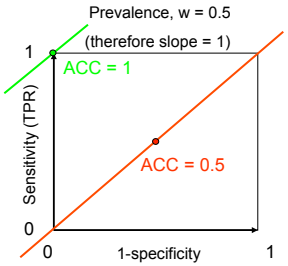
Dependent variable

TPR = Tre Positive Rate, senstivity

w = prevalence

SPC = specificity

ACC = Accuracy (diagnostic efficiency, de)



Increasing accuracy increases y-intercept, hence improves classification.

25

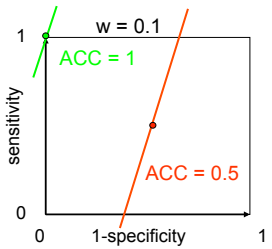
Dependence of ROC curve on diagnostic parameters II: prevalence

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

Dependent variable

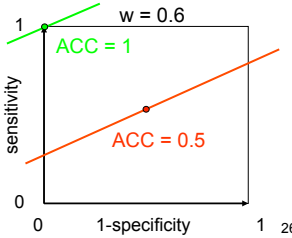
If w < 0.5, at identical accuracies the slope is greater than 1.

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



26