Diagnostic Tests

Dr. Diana Tichy (thanks to Dr. Christina Kunz)

dkfz. DEUTSCHES KREBSFORSCHUNGSZENTRUM IN DER HELMHOLTZ-GEMEINSCHAFT

News …

theguardian **October 30, 2012**

"Breast cancer screening causes more damage than previously thought

Around 4,000 women have unnecessary treatment for a disease that will never threaten their health, though tests should continue.

Breast cancer screening causes more harm than has previously been recognised, even though it saves lives, according to an independent review set up following years of scientific controversy surrounding the NHS programme.

Around 1,300 lives are saved every year by mammography, which women are invited to undergo between the ages of 50 to 70, said the review, which recommends that screenings should continue.

But 4,000 women will undergo unnecessary treatment, including surgery, radiotherapy and chemotherapy, for a cancer they would not otherwise have known about and which would have done them no harm in their lifetime. Some breast cancers are so tiny and slow growing that they would never be a threat to a woman's health, the review says. […]"

News …

theguardian **October 30, 2012**

"Breast cancer screening causes more damage than previously thought

Around 4,000 women have unnecessary treatment for a disease that will never threaten their health, though tests should continue

Breast cancer screening causes more harm than has previously been recognised, even though it saves lives, according to an independent review set up following years of scientific controversy surro programme.

true positives

Around 1,300 lives are saved every year by mammography, which women are invited to undergo between the ages of 50 to 70, said the review, which recommends that screenings should continue.

false positives

But 4,000 women will undergo unnecessary treatment, including surgery, radiotherapy and chemotherapy, for a cancer they would not otherwise have known about and which would have done them no harm in their lifetime. Some breast cancers are so tiny and slow growing that they would never be a threat to a woman's health, the review says. […]"

Outline

- What is a diagnostic test?
- Measures for diagnostic tests (sensitivity, specificity, predictive values, likelihood ratios)
- Sample size in reference-controlled diagnostic trials
- ROC analysis for diagnostic tests based on quantitative values
- Summary measures for ROC curves
- Using SigmaPlot for ROC analysis
- Selection of markers

Diagnostic test

A diagnostic test is any kind of medical test performed to aid in the diagnosis or detection of disease, injury or any other medical condition. For example, such a test may be used to confirm that a person is free from disease, or to fully diagnose a disease, including to sub-classify it regarding severity and susceptibility to treatment.

Here: The disease is present if it is verified by a criterion defined beforehand (=gold standard, reference diagnostics)

Examples of diagnostic tests

- Oral glucose tolerance test for diabetes mellitus
- Mammography when screening for breast cancer
- Haemoccult test when screening for colon cancer
- Tumor marker (e.g. CEA, CA50, CA19-9) in serum to detect cancer or to control the course of the disease (relapse/ progression of diagnosed cancer)

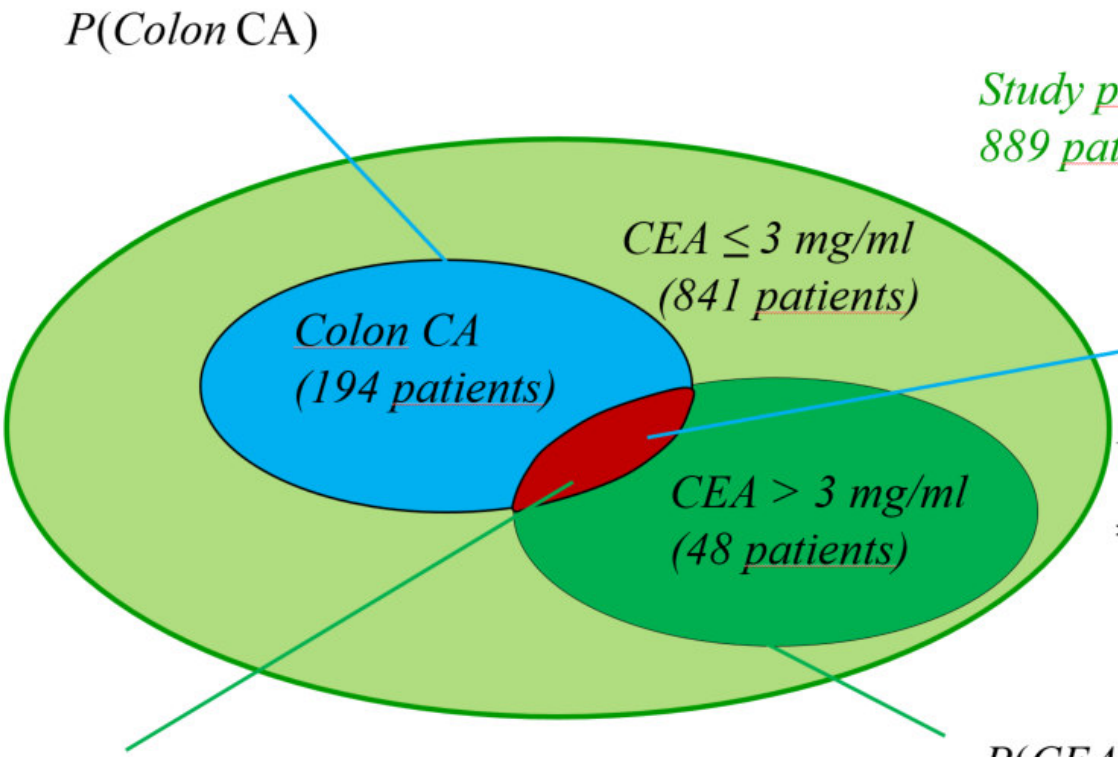
Diagnostic test example:
Carcino-embryonic antigen (CEA) in colon cancer (McCartney & Hoffer, 1974)

	CEA	
	> 3 mg/ml	≤ 3 mg/ml
Colon cancer	32	162
No colon cancer	16	679

reference/gold standard: diagnostic findings of the pathologist

Question: How „good“ is the test?
Is this a valid test to confine disease/no disease?
Is there a gain in information for the diagnostic process?
What are the characteristics of this test?

Concept of conditional probality



Condition: „colon CA“

$$P(\text{"CEA} > 3mg/ml\text{"} \mid \text{"colon CA"}) = \frac{P(\text{"CEA} > 3mg/ml\text{" and "colon CA"})}{P(\text{"colon CA"})}$$

$P(CEA > 3mg/ml)$

Condition: „CEA > 3 mg/ml“

$$P(\text{"colon CA"} \mid \text{"CEA} > 3mg/ml\text{"}) = \frac{P(\text{"CEA} > 3mg/ml\text{" and "colon CA"})}{P(\text{"CEA} > 3mg/ml\text{"})}$$

Diagnostic quality criteria

		TEST	
		positive (T^+)	negative (T^-)
DISEASE	yes (D^+)	true positive (TP)	false negative (FN)
	no (D^-)	false positive (FP)	true negative (TN)

sensitivity (Se) = probability of a positive test for a diseased patient
= $P(T^+|D^+) = P(T^+ \cap D^+) / P(D^+)$

estimate (Se) = proportion of diseased patients with positive test
= TP/(TP+FN)

specificity (Sp) = probability of a negative test for a healthy patient
= $P(T^-|D^-) = P(T^- \cap D^-) / P(D^-)$

estimate (Sp) = proportion of healthy patients with negative test
= TN/(FP+TN)

Sensitivity and specificity in the example

	CEA	
	>3 mg/ml (T^+)	≤3 mg/ml (T^-)
Colon cancer (D^+)	32	162
No colon cancer (D^-)	16	679

$Se = TP/ (TP+FN) = 32/ (32+162) = 0.165 = 16.5\%$

$Sp = TN/ (TN+FP) = 679/ (16+679) = 0.977 = 97.7\%$

Confidence intervals for Se and Sp

Approximate 95% confidence intervals:

For sensitivity

$$Se \; +/- \; 1.96\sqrt{Se\,(1-Se)/\underbrace{(TP+FN)}_{D^+}}$$

standard error of Se

For specificity

$$Sp \; +/- \; 1.96\sqrt{Sp\,(1-Sp)/\underbrace{(FP+TN)}_{D^-}}$$

standard error of Sp

Example: CEA measurements

- Se = 0.165 → 95% confidence interval : [0.113, 0.217]
- Sp = 0.977 → 95% confidence interval : [0.966, 0.988]

Properties of sensitivity (Se) and specificity (Sp)

1. Sensitivity and specificity are meaningful only in combination.
2. A diagnostic test is better than a random decision if and only if Se + Sp > 1.
3. Sensitivity and specificity do **not** depend on the prevalence of the disease.

prevalence (Prev) = probability of the disease in the total (considered) study population

$= P(D^+)$

$$\text{Prev} = \frac{\text{No. of diseased persons}}{\text{No. of persons in total population}}$$

Predictive values

Question: What can the physician tell the patient when he knows the test result?

positive predictive value

PPV = probability, that a patient with positive test has the specific disease
= $P(D^+|T^+)$

negative predictive value

NPV = probability, that a patient with negative test **doesn‘t** have the specific disease
= $P(D^-|T^-)$

Estimates for the predictive values

Use the contingency table for the test

		TEST	
		positive (T^+)	negative (T^-)
Disease	yes (D^+)	true positive (TP)	false negative (FN)
	no (D^-)	false positive (FP)	true negative (TN)

Estimates for PPV = TP/(TP+FP)

NPV = TN/(TN+FN)

Relationship between sensitivity, specificity, prevalence and predictive values (via the „Bayes theorem“)

$$PPV = \frac{Se \cdot Prev}{Se \cdot Prev + (1-Sp) \cdot (1-Prev)}$$

$$NPV = \frac{Sp \cdot (1-Prev)}{Sp \cdot (1-Prev) + (1-Se) \cdot Prev}$$

- PPV and NPV : a-posteriori probabilities („post-test“ probabilities) (information gained by test?)
- prevalence: a-priori probability („pre-test“ probability)
- If the test is independent from the disease, then no information is gained by the test → PPV = Prev

Predictive values depend on prevalence

Example: CEA measurements

	CEA	
	>3 mg/ml (T^+)	≤3 mg/ml (T^-)
Colon cancer (D^+)	32	162
No colon cancer (D^-)	16	679

Se = 0.165 = 16.5%,

Sp = 0.977 = 97.7%

Prev = $P(D^+)$ = (32+162) / (32+162+16+679) = 0.218 = 21.8 %

$$\rightarrow \text{PPV} = \frac{32}{32+16} = 0.667 = 66.7\%$$

$$= \frac{0.165 \cdot 0.218}{0.165 \cdot 0.218 + (1-0.977) \cdot (1-0.218)}$$

$$\text{NPV} = \frac{679}{679+162} = 0.807 = 80.7\%$$

$$= \frac{0.977 \cdot (1-0.218)}{0.977 \cdot (1-0.218) + (1-0.165) \cdot 0.218}$$

Other prevalence: Prev=0.001 = 0.1%

$$\rightarrow \text{PPV} = \frac{0.165 \cdot 0.001}{0.165 \cdot 0.001 + (1-0.977) \cdot (1-0.001)}$$

$$= 0.007 = 0.7\%$$

$$\text{NPV} = \frac{0.977 \cdot (1-0.001)}{0.977 \cdot (1-0.001) + (1-0.165) \cdot 0.001}$$

$$= 0.999 = 99.9\%$$

Since PPV, NPV depend on prevalence …

1. PPV increases when Prev increases

2. NPV decreases when Prev increases.

3. The estimates can not be transferred to other populations (only locally valid).

4. The direct estimate is not valid if samples of „healthy“ and „diseased“ patients are drawn separately in the study („stratified“ studies).

Got it? Please mark the correct answers

1. Find for each measure A. – D. of diagnostic tests the correct definition in a. – d.

   A. Sensitivity
   B. Positive predictive value
   C. Specificity
   D. Negative predictive value

   a. The probability that a test correctly classifies as positive those who have the disease
   b. The probability that a test correctly classifies individuals without disease as negative
   c. The probability that those with positive test have the disease
   d. The probability that those with negative test do not have the disease

2. Based on all the information currently available, you estimate that the patient in your office has a one in four chance of having a serious disease. You order a diagnostic test with sensitivity of 95% and specificity of 90%. The result comes back positive. Based on all the information now available, the chance your patient really has the disease is closest to
   a. 100%b. 95% c. 90% d. 75% e. 60% f. roughly 30%

3. If the same screening test is conducted in two populations, one with a high prevalence of the disease and one with a low prevalence of the disease, assuming the sensitivity and specificity of the screening test are the same, which of the following statements about positive predictive value (PPV) applies:
   A. PPV is higher in the screened population with higher prevalence
   B. PPV is lower in the screened population with higher prevalence
   C. PPV is the same in both populations
   D. It cannot be determined

Youden Index

$$Y = Se + Sp - 1$$

- prevalence-independent measure for diagnostic test
- high Y indicate good diagnostic tests

  → used to find optimal cut-off value for quantitative marker

  → Se and Sp equally weighted

  → maximum Y not unique

Example: CEA measurements

Y = 0.165 + 0.977 -1 = 0.142

Requirements for sensitivity and specificity

- depend on consequences of misclassification (ethical and financial costs)
- false positive (FP) more severe, then high specificity required (affirmative test)
- false negative (FN) more severe, then high sensitivity required (detection test = screening test)
- exclusion of a disease, then high sensitivity required (exclusion test)
- for low prevalence, the test needs high specificity to reach a sufficiently high positive predictive value.

Diagnostic Likelihood Ratios

		TEST	
		positive (T^+)	negative (T^-)
Disease	yes (D^+)	true positive (TP)	false negative (FN)
	no (D^-)	false positive (FP)	true negative (TN)

Positive Diagnostic Likelihood Ratio (DLR+)

"How much more likely is it that a diseased person is test positive than that a healthy person is test positive"

$$DLR^+ = \frac{P(T+|D+)}{P(T+|D-)} = \frac{Se}{1-Sp}$$

Negative Diagnostic Likelihood Ratio (DLR-)

"How much more likely is it that a diseased person is test negative than that a healthy person is test negative"

$$DLR^- = \frac{P(T-|D+)}{P(T-|D-)} = \frac{1-Se}{Sp}$$

→ Diagnostic Likelihood ratios > 0

Diagnostic Likelihood Ratios

	CEA		
	>3 mg/ml (T^+)	≤3 mg/ml (T^-)	Total
Colon cancer (D^+)	32	162	194
No colon cancer (D^-)	16	679	695
Total	48	841	889

Positive Diagnostic Likelihood Ratio (DLR+) = (32/194)/(16/695) = 7.16

Negative Diagnostic Likelihood Ratio (DLR-) = (162/194)/(679/695) = 0.85

Measures of diagnostic accuracy for binary tests

	Classification probabilities	Predictive values	Diagnostic likelihood ratios
Parameter definitions	Se = P(T+|D+) Sp = P(T-|D-)	PPV = P(D+|T+) NPV = P(D-|T-)	DLR+ = P(T+|D+)/P(T+|D-) DLR- = P(T-|D+)/P(T-|D-)
Scale	[0,1]	[0,1]	[0,∞)
Perfect test	Se = 1 Sp = 1	PPV = 1 NPV = 1	DLR+ = ∞ DLR- = 0
Useless test	Se + Sp = 1	PPV = Prevalence NPV = 1-Prevalence	DLR+ = 1 DLR- = 1
Context for use	Accuracy	Clinical prediction	Test informativeness
Question addressed	To what degree does the test reflect the true disease state?	How likely is disease given the test result?	By how much does the test change knowledge of disease status?
Affected by disease prevalence?	No	Yes	No

Pepe 2003

Choosing a cut-off value for a quantitative test: Example

Numerical Example (Pepe, 2003, example 5.1)

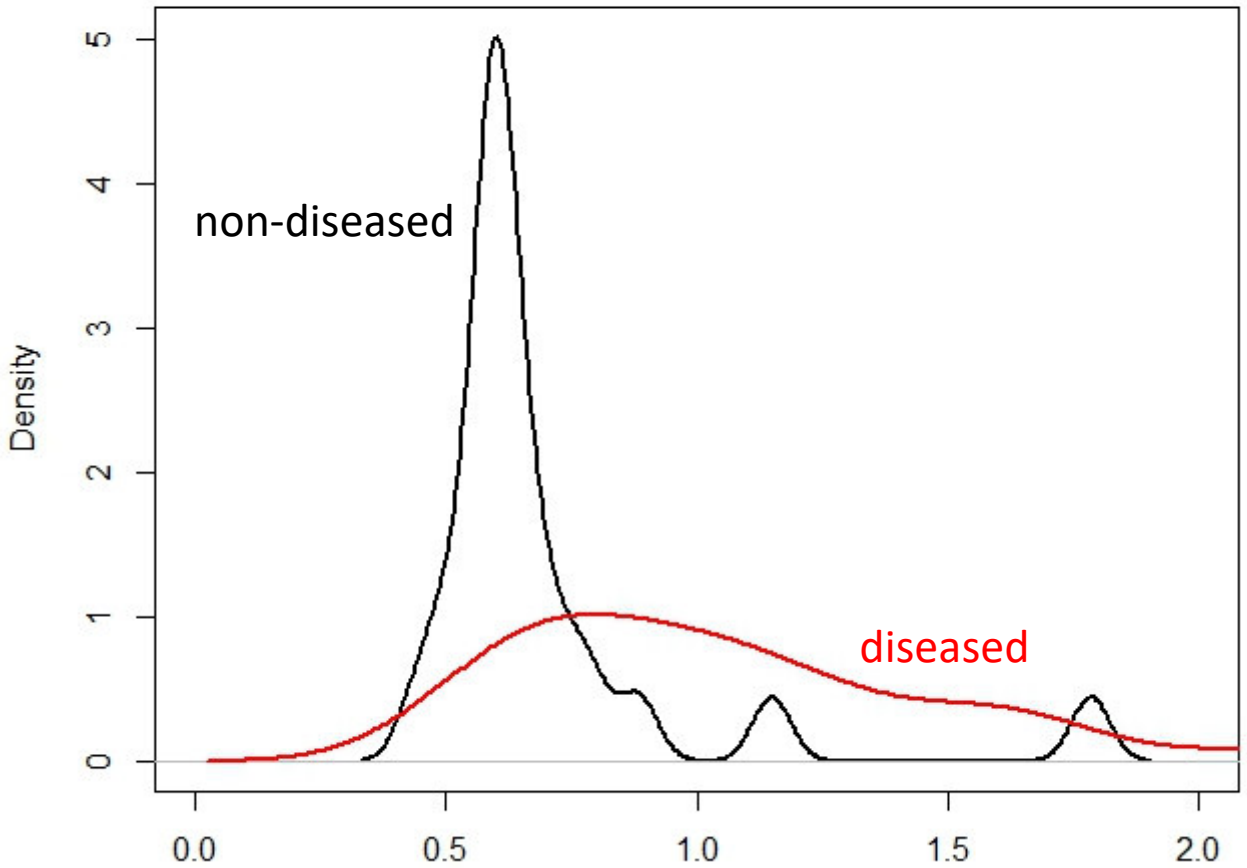
- 23 non-diseased ovarian tissues, 30 ovarian tumor tissues
- specific gene expression for all 53 patients:

Normal tissues	0.442 0.500 0.510 0.568 0.571 0.574 0.588 0.595 0595 0.595 0.598 0.606 0.617 0.628 0.641 0.641 0.680 0.699 0.746 0.793 0.884 1.149 1.785
Cancer tissues	0.543 0.571 0.602 0.609 0.628 0.641 0.666 0.694 0.769 0.800 0.800 0.847 0.877 0.892 0.925 0.943 1.041 1.075 1.086 1.123 1.136 1.190 1.234 1.315 1.428 1.562 1.612 1.666 1.666 2.127

Question: Can the gene serve as diagnostic test for ovarian cancer?

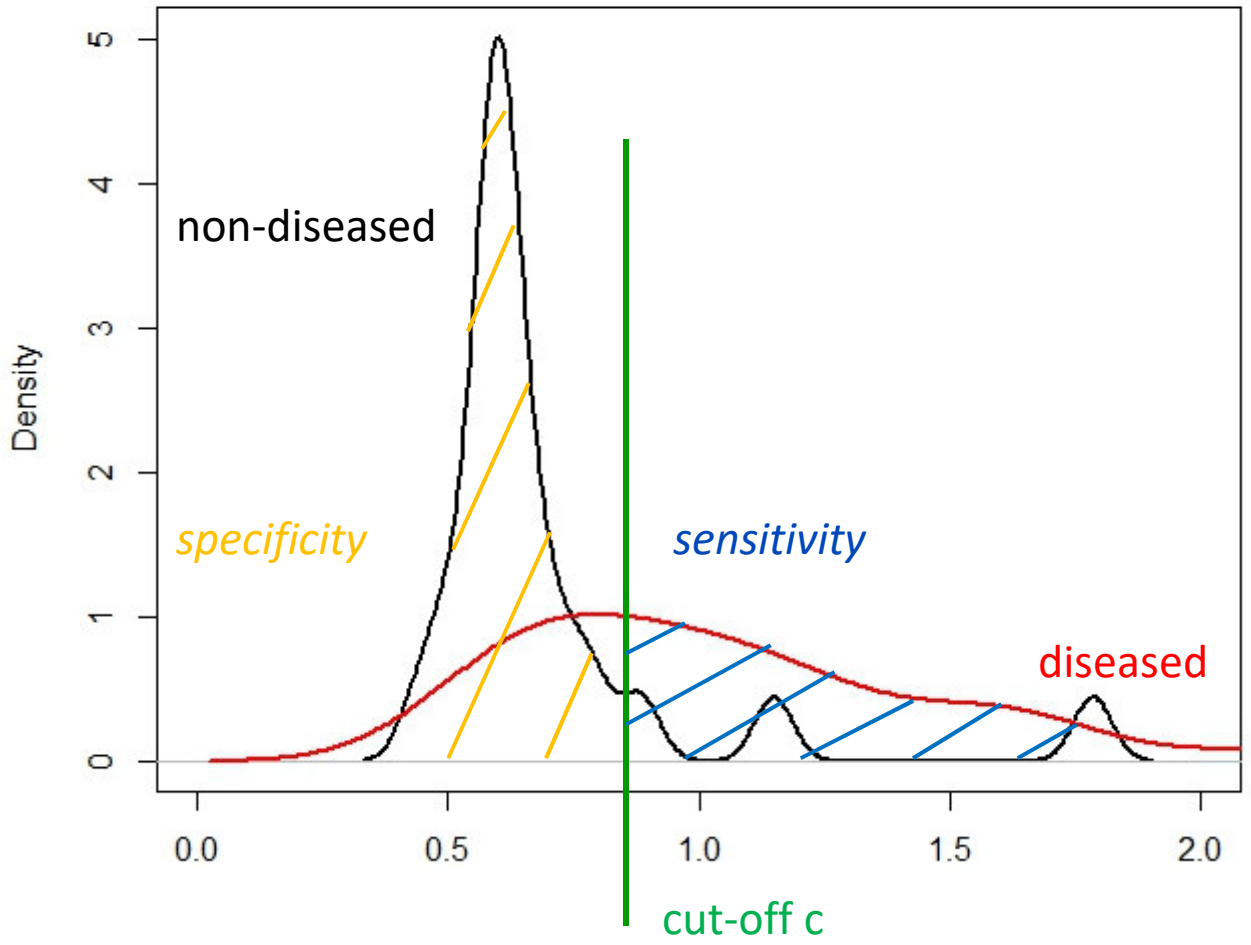
Comparing the distribution of diseased and non-diseased

Density estimates for 23 non-diseased ovarian tissues, 30 ovarian tumor tissues:

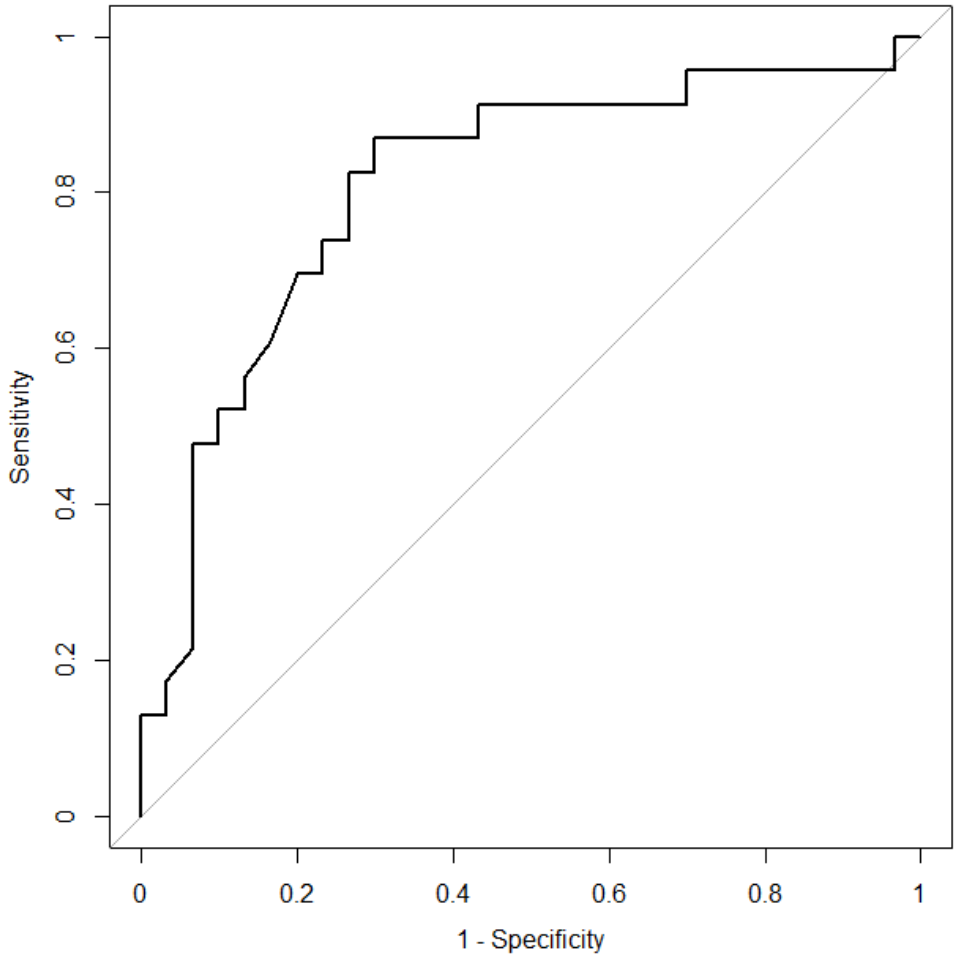


Comparing the distribution of diseased and non-diseased

Dichotomize at a cut-off: Test positive if value > c



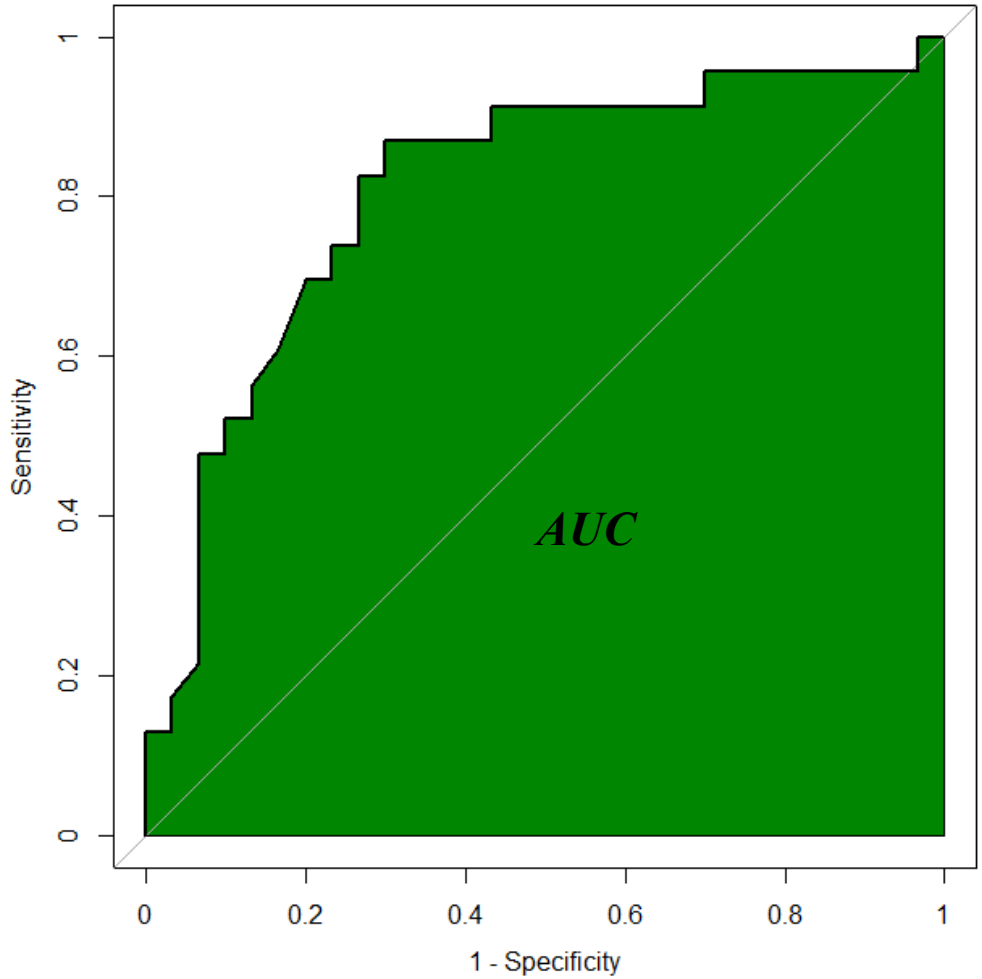
ROC curve for gene
(ROC = „Receiver Operating Characteristic“)



ROC curve shows

- family of binary tests
- for varying cut-off values
- pairs (Se, 1-Sp) of binary tests

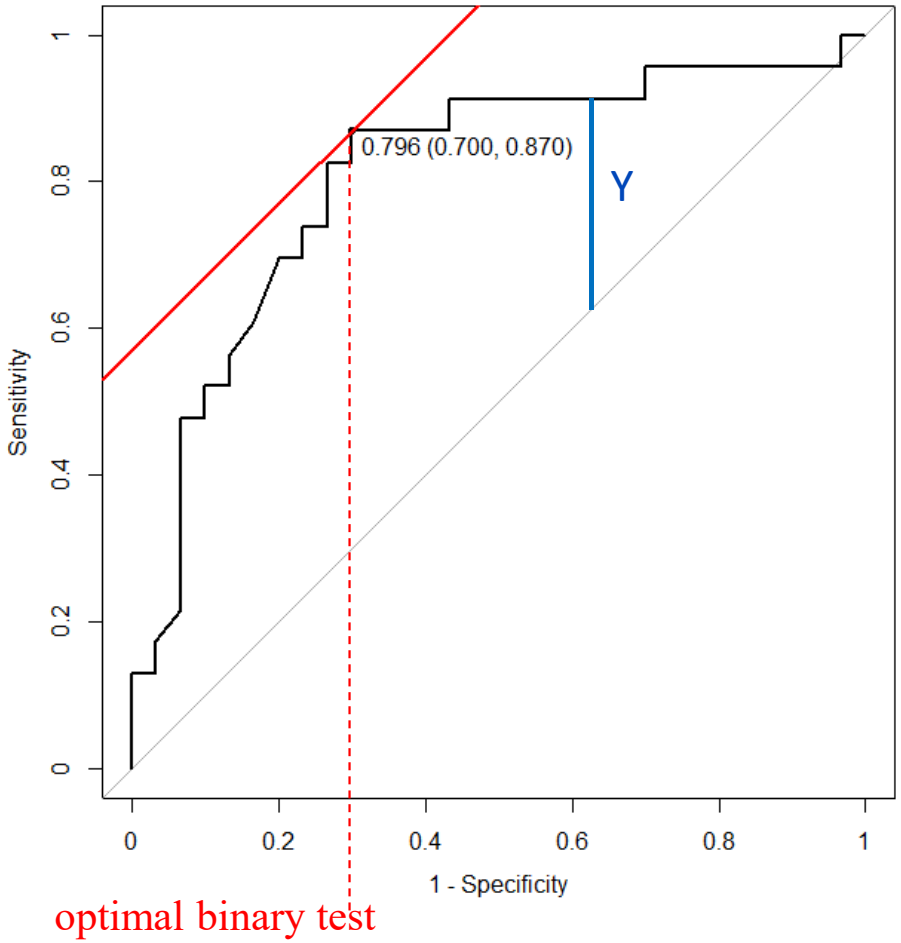
ROC curve for gene
(ROC = „Receiver Operating Characteristic“)



AUC („area under the curve“) is used to compare different diagnostic tests

- comparison of intersecting ROC curves is difficult to interprete
- AUC = 1 for „perfect“ test
- AUC = 0.5 for „random“ diagnosis

Optimal cut-off value using Youden Index



- Youden Index: Y = Se – (1-Sp)
- Sensitivity and specificity are equally important
- maximize Y
  → optimal binary test/ cut-off value
- maximal Youden Index: 0.5696
- optimal threshold value: 0.7965
  Specificity = 87%, Sensitivity = 70%
- graphically:
  test corresponding to point on ROC curve with slope 1

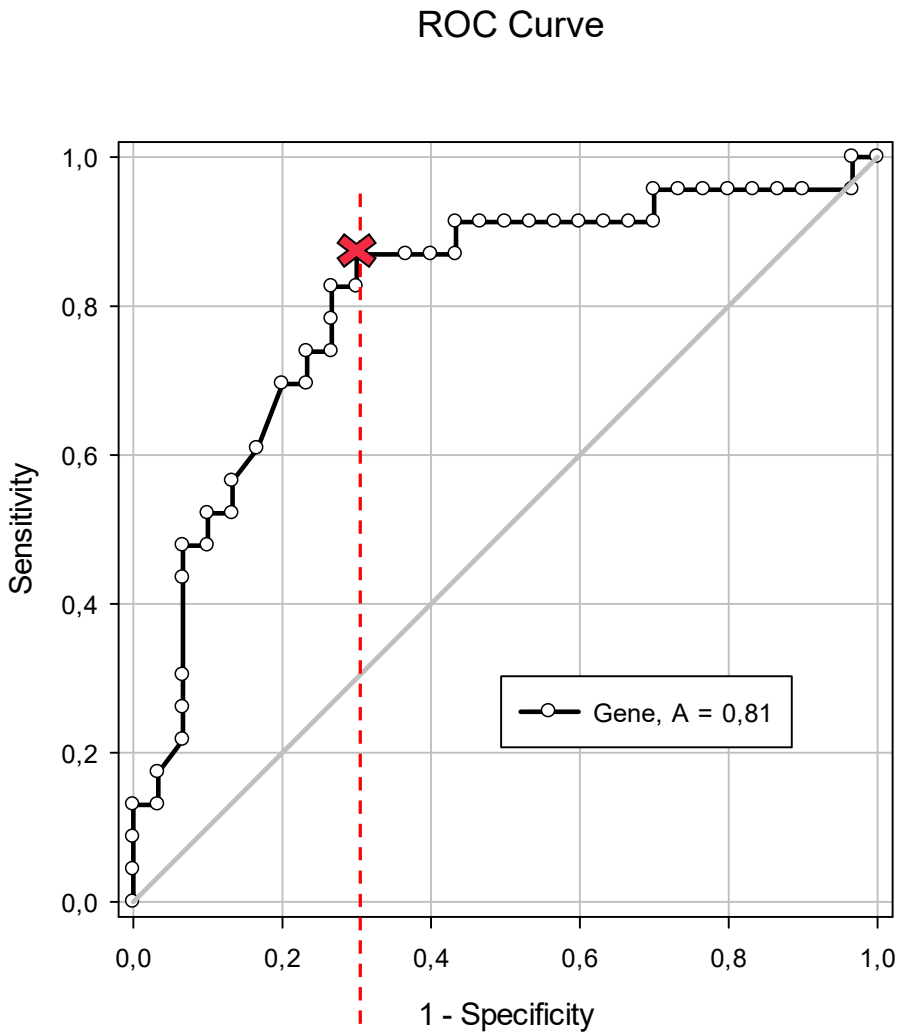
Be aware of potential overfitting:

An optimal cut-off must be validated in an **independent data set**

How to choose the cut-off value (discrimination value/ threshold) of a quantitative test (Galen and Gambino 1979)

Disease treatable?	Consequences of false positive diagnostic findings	Weighting
yes	not severe	Se high
no	severe	Sp high
yes	as severe as the consequences of false negative diagnostic findings	Se $\approx$ Sp

ROC analysis with SigmaPlot: Ovarian cancer example



optimal binary test

max. Youden Index:
0.5696

optimal cut-off value:
0.7965

(1-Sp, Se):
(0.3, 0.8696)

AUC=0.81,
95%-CI [0.69, 0.93]

Selection of markers

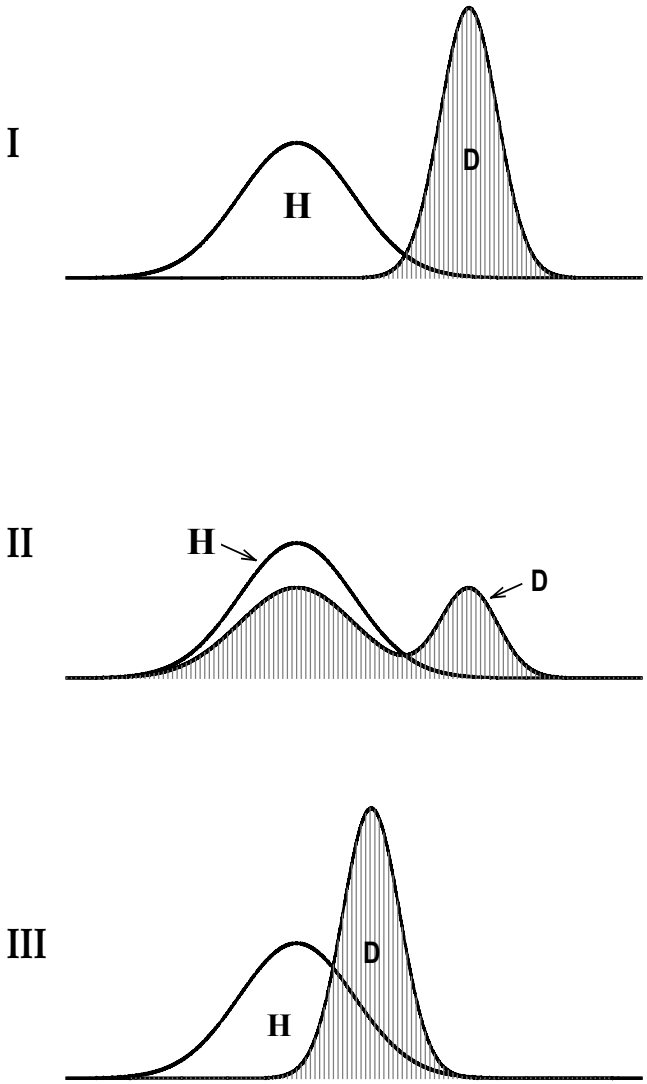
Aim

- Identify potential promising markers among a set of possible markers

  E.g., identify genes that are differentially expressed in ovarian cancer tissue compared with normal ovarian tissue.

- Rank candidate markers to some statistical measure

Separation of marker value distribution



Panel I:

Almost complete separation between the distributions of healthy/controls (H) and diseased (D).

Classify with almost 100% accuracy.

Panels II and III:

Overlapping distributions.

Cancer screening:

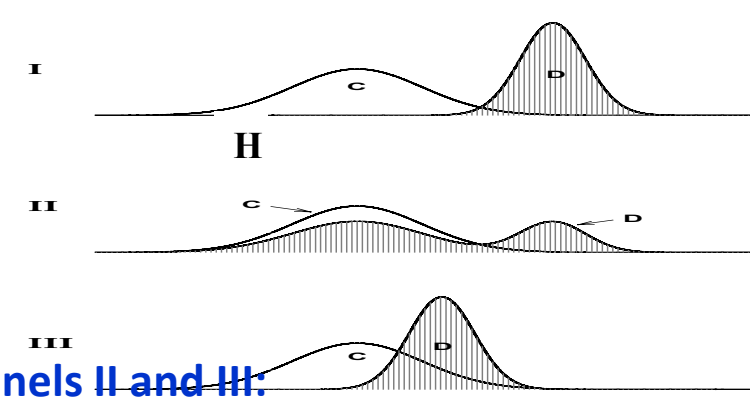
Panel II is of more practical interest than panel III.

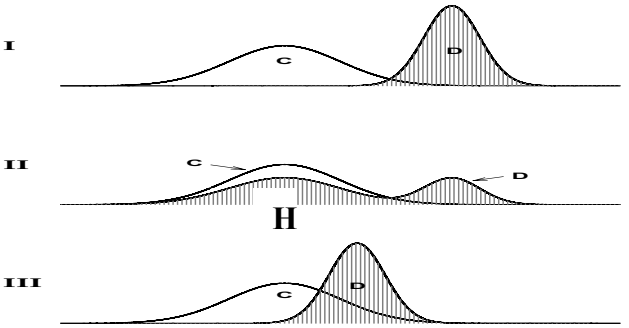
Panel II: marker clearly distinguishes a subset of D from H

Panel III: marker values for D are entirely within the range of those for H.

(Pepe et al., Biometrics 2003)

Separation of marker value distribution



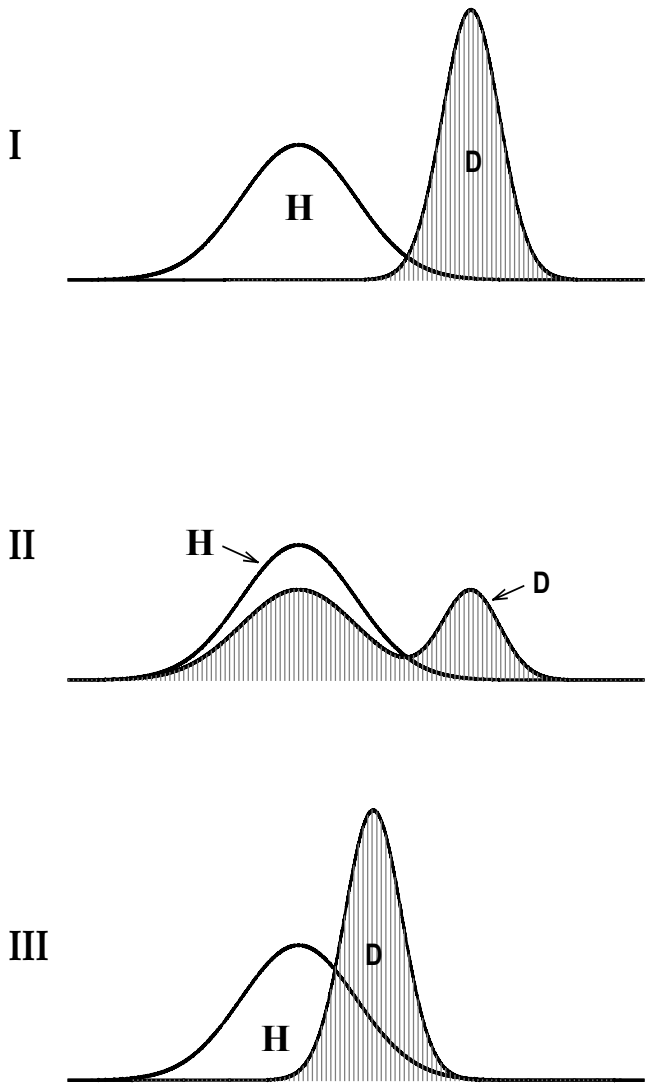


Panels II and III:

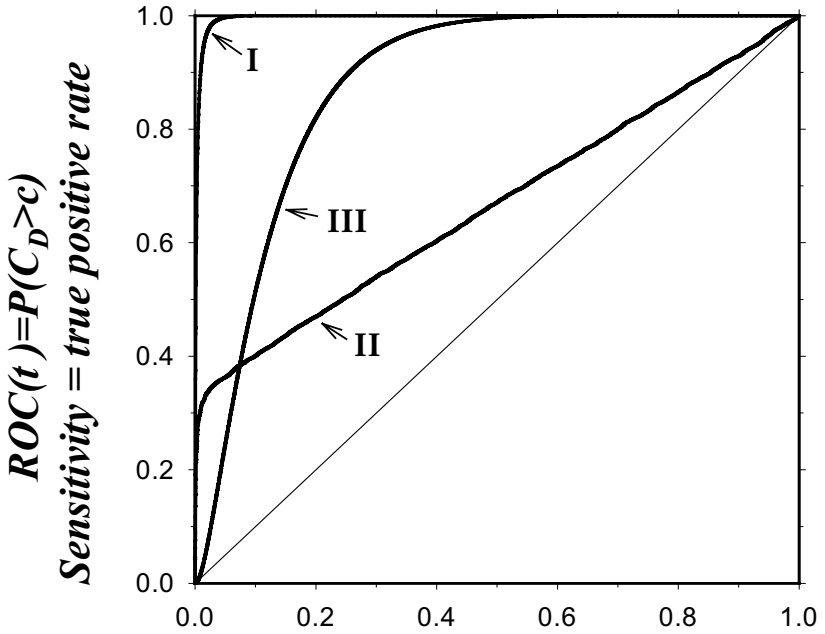
- in panel II there is a threshold for the screening test that provides detection of about 30% of diseased patients while falsely identifying only 1% of healthy patients as screen positive
- in screening it is important to keep false positive rates low because even a small false positive rate translates into a large numbers of people being subjected to diagnostic procedures that may be costly and invasive
- using a similar threshold in panel III corresponding to the 1% false positive rate, detects only 2% of diseased patients because the distributions overlap over the whole normative range

(Pepe et al., Biometrics 2003)

ROC curves



Resulting ROC curves for panels I - III



(Pepe et al., Biometrics 2003)

Tests for comparison of ROC curves

- ROC and AUC can be used to compare and rank potential markers
- AUC insufficient to compare intersecting ROC curves (panel II and III)
- To compare two ROC-curves (of non-nested models), deLong test can be used
- To compare two ROC-curves (of nested models), use the Likelihood-ratio test
- AUC = Wilcoxon ranksum statistics (with deLong 95% confidence interval!)

Summary

- Accuracy of diagnostic tests is judged by sensitivity and specificity
- Predictive values allow physicians to asses the test results
- Likelihood ratios show how informative the test is
- Sensitivity and specificity are independent of prevalence
- Predictive values depend on prevalence
- According to the aims, a diagnostic test requires high sensitivity, or high specificity, or equal sensitivity/specificity
- Accuracy requirements on sensitivity/specificity can be used to determine sample size of reference-controlled diagnostic trials
- The choice of cut-off value for quantitative diagnostic tests influences sensitivity and specificity
- ROC curve shows sensitivity and 1-specificity for varying cut-off values
- Optimal cut-off values and/or selected marker must be validated in an independent data set.

Software

- SigmaPlot
- R package pROC
- Sensitivity/Specificity: http://vassarstats.net/clin1.html
- Sensitivity/Specificity: https://www.medcalc.org/calc/diagnostic_test.php

References

Galen, Gambino (1979). *Norm und Normabweichung klinischer Daten. Gustav Fischer Verlag Stuttgart New York*

Hilgers RD, Bauer P, Scheiber V (2000). *Einführung in die medizinische Statistik. Springer-Verlag Berlin Heidelberg New York*

Krummenauer F, Kauczor HU (2002). *Fallzahlplanung in referenzkontrollierten Diagnosestudien/ Sample Size Determination in Reference-Controlled Diagnostic Trials. Fortschr Röntgenstr 174: 1438–1444*

McCartney WH, Hoffer PB (1974). *The value of carcinoembryonic antigen (CEA) as an adjunct to the radiological colon examination in the diagnosis of malignancy. Radiology 110(2):325-8*

Pepe MS, Longton G, Anderson GL, Schummer M (2003). *Selecting differentially expressed genes from microarray experiments. Biometrics 59:133-142*

Pepe MS (2004). *The statistical evaluation of medical tests for classification and prediction. Oxford University Press*

References on COVID-19 Testing

Kumleben et al. (2020). *Test, test, test for COVID-19 antibodies: the importance of sensitivity, specificity and predictive power. Public Health 185, 88-90.*

Vandenberg et al. (2020). *Considerations of diagnostic COVID-19 tests. Nature reviews. Microbiology.*

Next lecture

December 9th 2020

Design of clinical trials

PD Dr. Tim Holland-Letz