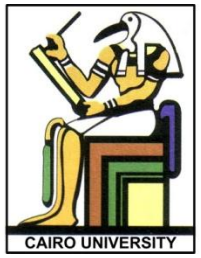


ADVANCED TOPICS IN BIOMEDICAL ENGINEERING

Topic 2: Diagnostic Performance

Prof. Yasser Mostafa Kadah

Lecture 2 - 2012

Diagnostic Performance Definitions

- The *performance* of a diagnostic examination can be basically considered as its degree of accuracy, namely its ability to find the subjects affected with a given disease as positive and the subjects not affected with same disease as negative
- The indices which in different ways measure this performance are defined *measures of diagnostic performance* and the studies aimed at measuring the diagnostic performance of an examination or, more often, at comparing the diagnostic performance of two or more examinations, are defined *studies of diagnostic performance*.

Results of an Examination Compared to Reference Standard

- To evaluate the performance of a diagnostic examination, we need to compare its results to a reference standard
  - "Gold standard"
- Typical example: to verify each result of a diagnostic examination for a sample of n patients with pathology report
- Suppose that both radiologist and pathologist are required to give a dichotomous judgment (yes/no) about malignancy:
  - True positive
  - False positive
  - True negative
  - False negative

Two-by-Two Contingency Table

		Reference standard	
		Positive	Negative
Radiologic examination	Positive	True positives (TP)	False positives (FP)
	Negative	False negatives (FN)	True negatives (TN)

		Reference standard		
		Affected	Nonaffected	Total
Radiologic examination	Positive	True positives (TP)	False positives (FP)	All positives (TP + FP)
	Negative	False negatives (FN)	True negatives (TN)	All negatives (FN + TN)
Total		All affected (TP + FN)	All nonaffected (FP + TN)	Grand total (TP + FP + FN + TN)

Terminology

- Different terms: *cases, lesions, findings, patients*, and *subjects*
- Consider the study subjects as *patients* when they present with symptoms or signs for a disease
- Name the asymptomatic persons enrolled in population screening program only as *subjects*
- *Statistical Unit* to be considered
  - Patient, organ, segment, or lesion
- Avoid the term *case* in a scientific context
  - Ambiguous because it can be used for both patients and lesions

Measures of Diagnostic Performance

Index	Definition	Formula	Dependence on disease prevalence
1. Sensitivity (or TP rate)	Ability to identify the presence of disease	TP/(TP+FN)	No
2. Specificity (or TN rate)	Ability to identify the absence of disease	TN/(TN+FP)	No
3. Positive predictive value (PPV)	Reliability of the positive result	TP/(TP+FP)	Yes
4. Negative predictive value (NPV)	Reliability of the negative result	TN/(TN+FN)	Yes
5. Overall accuracy	Global reliability	(TP+TN)/(TP+TN+FP+FN)	Yes

Measures of Diagnostic Performance

Index	Definition	Formula	Dependence on disease prevalence
6. FN rate	Proportion between FN and all affected	FN/(FN+TP) = (1 – Sensitivity)	No
7. FP rate	Proportion between FP and all nonaffected	FP/(FP+TN) = (1 – Specificity)	No
8. Positive likelihood ratio	Increase in disease probability when the result is positive	Sensitivity/(1 – Specificity)	No
9. Negative likelihood ratio	Decrease in disease probability when the result is negative	(1 – Sensitivity)/Specificity	No

Note that *disease prevalence* is equal to (TP+FN) / (TP+TN+FP+FN), being the ratio between the number of subjects affected by the disease and the grand total of sample of subjects under investigation.

Sensitivity

- Sensitivity: the ability to identify the presence of a disease
- Example [SARDANELLI ET AL, 2004]: Sensitivity of mammography and dynamic contrast enhanced MRI for the detection of malignant lesions in patients candidate for mastectomy. The authors investigate 99 breasts in 90 candidates for unilateral (n = 81) or bilateral (n = 9) mastectomy. The reference standard, i.e. the pathology exam of the whole excised breast, establishes the presence of 188 malignant lesions. Mammography has 124 true positives and 64 false negatives, MR imaging 152 true positives and 36 false negatives.
  - Sensitivity is 124/(124+64) = 0.66 (66%) for mammography and 152/(152+36) = 0.809 (80.9%) for MRI.
  - The FN rate is 0.340 (34.0%) and 0.191 (19.1%), respectively.
  - Note that the statistical unit is the lesion and not the patient or the breast

Specificity

- *Specificity*: the ability to identify the absence of a disease
- Example [SOBUE ET AL, 2002]: Low-dose CT screening for lung cancer: Of a total of 1611 asymptomatic subjects who undergo the first screening event, 186 are found to be positive and are further studied with high-resolution scanning; 21 of these undergo biopsy. Thirteen subjects are found to be affected by lung cancer. There are no interval cancers (cancers detected between the first and the second screening event). As a result there are 1425 true negatives (=1611-186) and 173 false positives (=186-13)
  - Specificity is 1425/(1425+173)= 0.892 = 89.2%.
  - In this series only one possible lesion is considered for each subject. Lesion and subject are coincident as a statistical unit.

Notes on Different Measures

- Sensitivity and specificity: answers to pretest questions
  - If the patient is affected by the disease, what is the probability that the examination produces a positive result (sensitivity)?
  - If the patient is not affected by the disease, what is the probability that the examination produces a negative result (specificity)?
- Differentiation between *sensitivity* and *specificity* as answers to pre-examination questions and *predictive values* as answers to post-examination questions
- Sensitivity and specificity do not depend on disease prevalence
  - *Prevalence* indicates proportion between number of subjects affected by disease and total number of subjects of an entire population for a defined time interval
  - *Incidence* indicates the number of subjects newly diagnosed as affected by the disease during a defined time interval

Notes on Different Measures

- Optimal situation in clinical practice is when a single diagnostic examination is available with levels of sensitivity or specificity high enough to produce conclusive decision-making
- An examination is SNOUT when its negative result excludes the possibility of the presence of the disease
  - When a test has a very high Sensitivity, a Negative result rules OUT the diagnosis
- An examination is SPIN when its positive result definitely confirms the presence of the disease
  - When a test has a very high Specificity, a positive result rules IN the diagnosis
- In most situations, a certain degree of certainty can be reached with a single diagnostic examination but not a definitive conclusion
  - More than one examination is generally needed

Predictive Values

- Indicate the reliability of positive or negative result and answer questions posed after having performed the examination
  - If the result of the examination is positive, what is the probability that the patient really is affected by the disease (positive predictive value)?
  - If the result of the examination is negative, what is the probability that the patient is really not affected by the disease (negative predictive value)?
- Predictive values depend on disease prevalence
  - Positive predictive value is directly related to disease prevalence
  - Negative predictive value is inversely related to disease prevalence
- Reliability of reports also depends on patient selection by the referring physicians

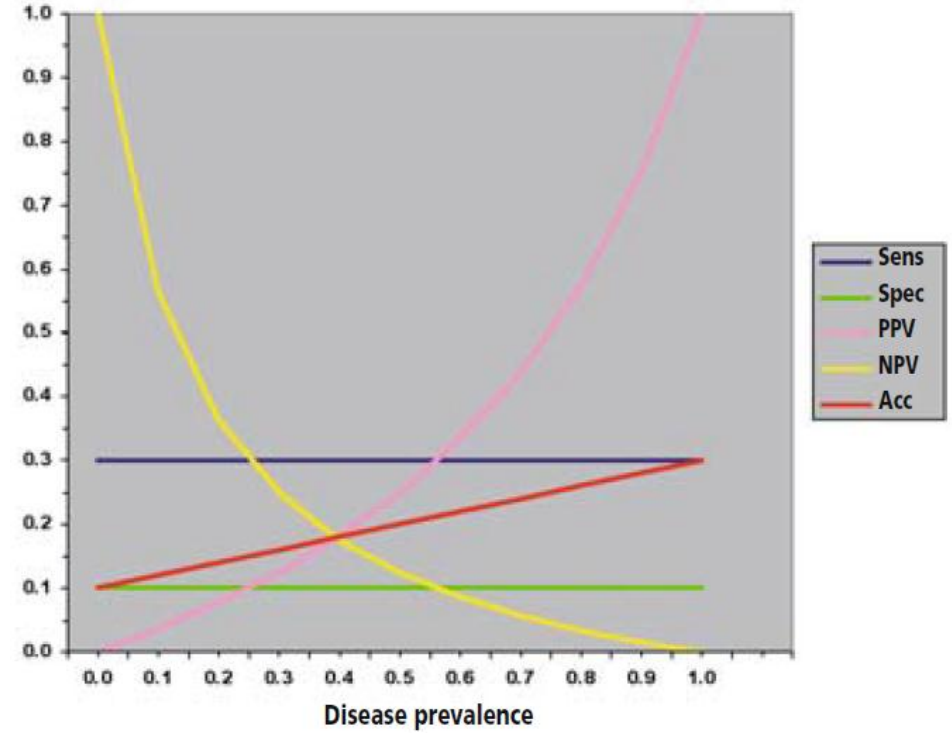
Predictive Values

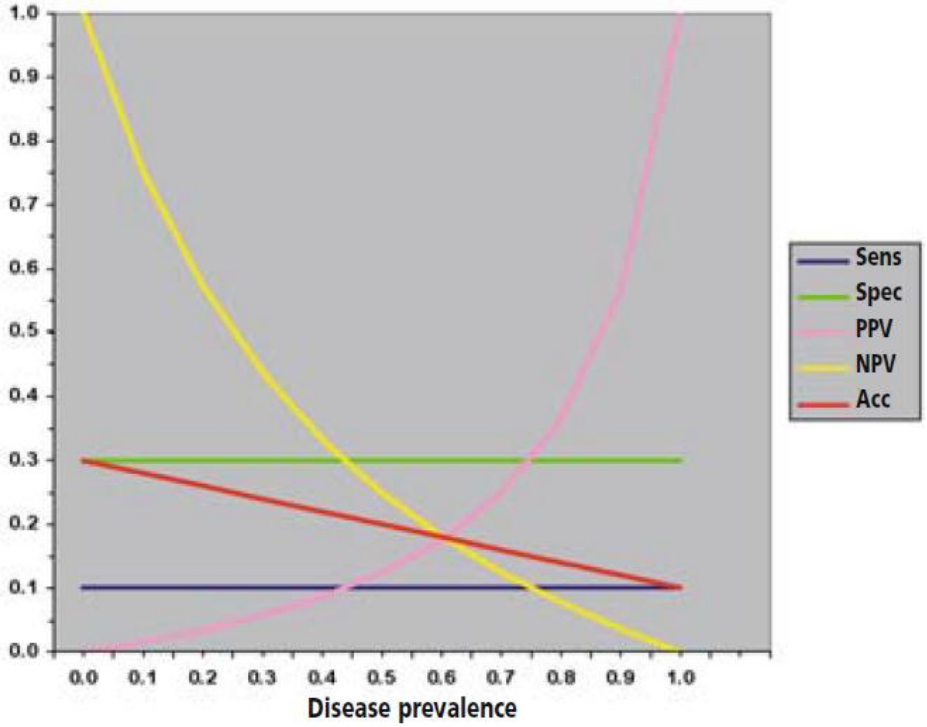
- A disease can affect a patient with different levels of severity (or stage) and the probability of a positive result of an examination increases with the level of severity.
  - Level of severity should be lower in subjects in whom the disease is diagnosed with periodic screening than that found in symptomatic subjects in whom the disease is diagnosed in clinical practice
- In this way we observe a direct influence on sensitivity and specificity: they are higher in symptomatic subjects than in asymptomatic subjects in whom the disease is more likely in an early stage

Overall Accuracy

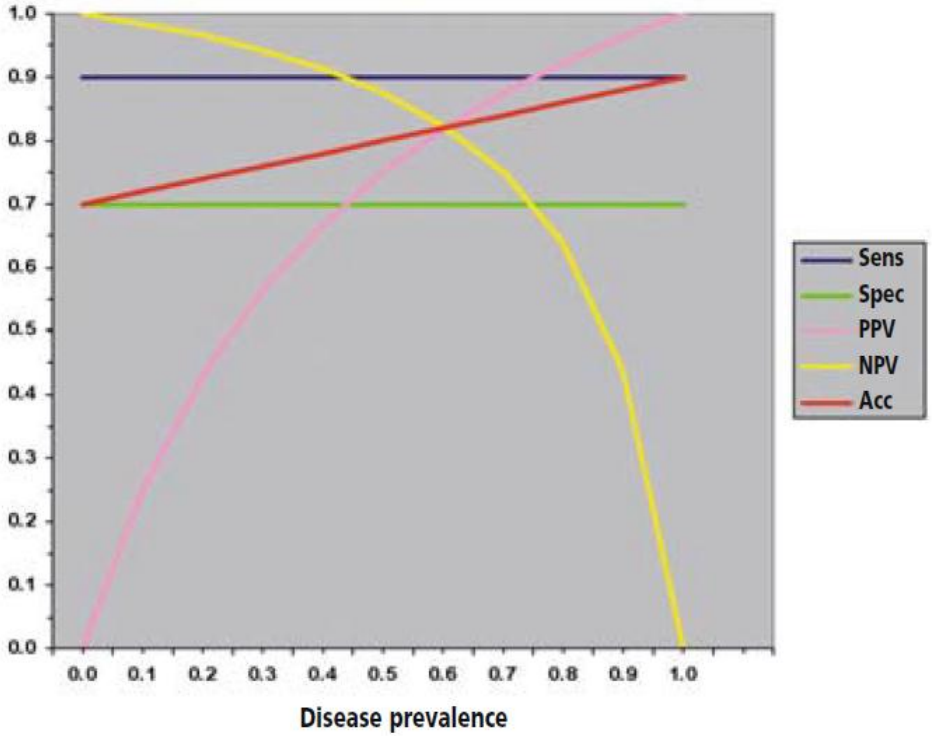
- Ability to correctly identify the presence and the absence of a disease
- It answers the question: what is the probability of a correct result?
  - Somewhat like a global index of diagnostic performance, but its linear distribution ranges between the sensitivity value and the specificity value.
  - It approaches the higher of the two with increasing disease prevalence and approaches the lower of the two with decreasing disease prevalence.
- In practice, it is a kind of "mean" between sensitivity and specificity which is weighted for disease prevalence
  - Dependence on disease prevalence is the feature shared with the predictive values

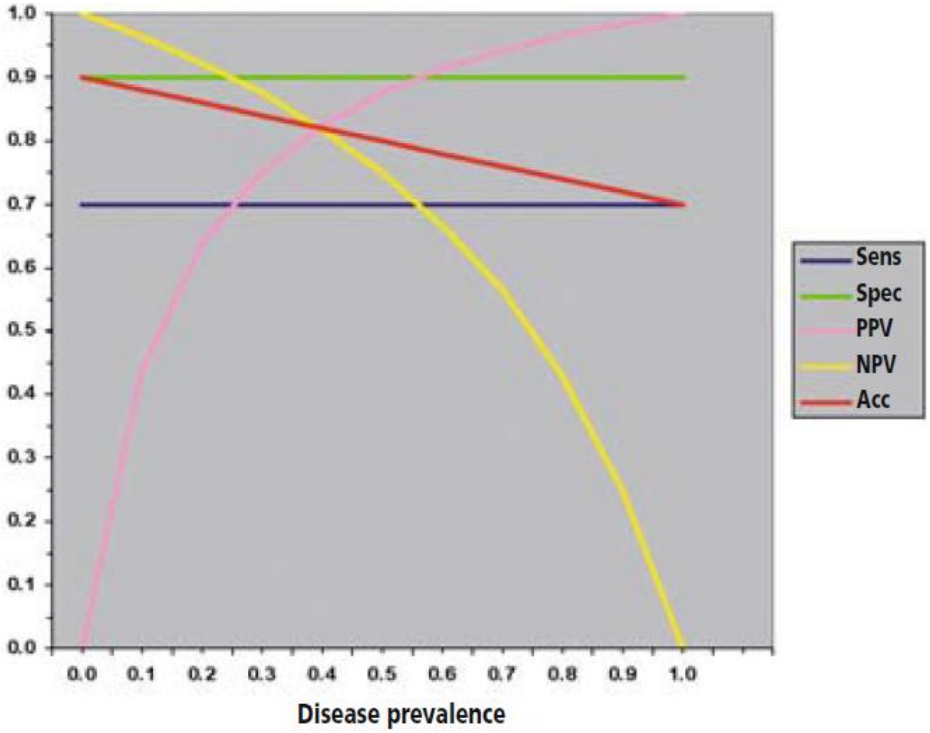
Measures vs. Disease Prevalence





Measures vs. Disease Prevalence





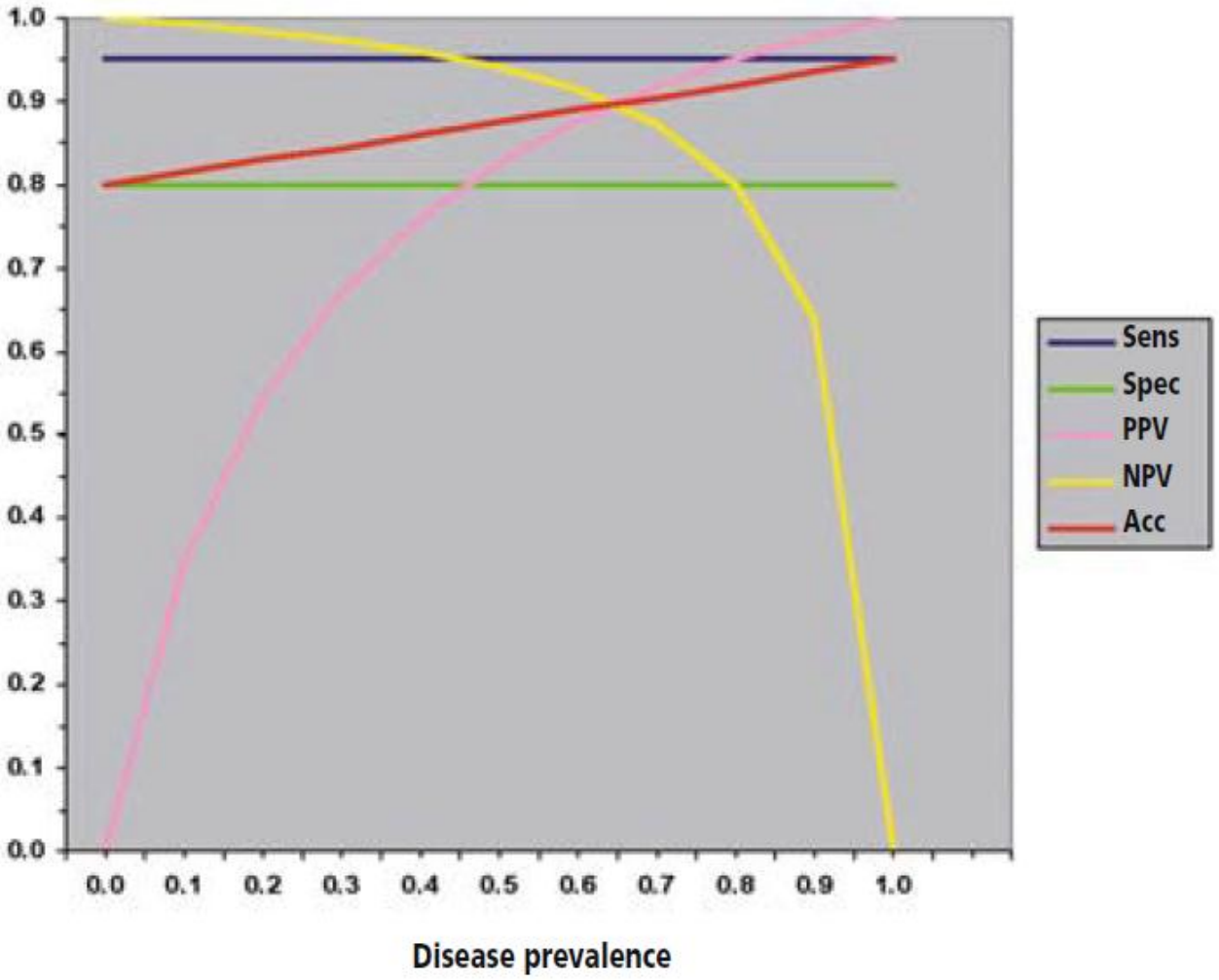
Example: Predictive Values of Clinical and Screening Mammography

- Imagine 10,000 women with a palpable lump are studied (clinical mammography), with 95% sensitivity and 80% specificity. With a disease prevalence of 50%, we would have 4,750 true positives, 4,000 true negatives, 1,000 false positives, and 250 false negatives. The PPV would be 4,750/(4,750+1,000) = 0.826; the NPV 4,000/(4,000+250) = 0.941.
  - For nearly every 5 women affected with cancer there would be a healthy woman who undergoes diagnostic work-up with possible needle biopsy (4,750/1000 = 4.75).
  - This woman with a benign palpable lump is unlikely to consider invasive examinations as useless or dangerous.

Example: Predictive Values of Clinical and Screening Mammography

- If we were to study 10,000 asymptomatic women (screening mammography) with the same levels of sensitivity and specificity (95% and 80%, respectively) with a disease prevalence of 3%, we would have 285 true positives, 7,760 true negatives, 1,940 false positives, and 15 false negatives. The NPV would go up to 7,760/(7,760+15) = 0.998, PPV would go down to 285/(285+1,940) = 0.128.
  - This means that nearly 7 healthy women would be sent for diagnostic work-up with a possible needle biopsy for every woman effectively diagnosed with cancer (1,940/285 = 6.8).
  - Recall rate would be very high, equivalent to 22.25% (2,225/10,000).
  - Overall effect would be a false alarm (if at every round we recall 20-25% of the women, after 4-5 rounds on average all women would be recalled).
  - Work-flow and economic costs would be huge. Above all, the women would lose confidence with the screening program

Example: Predictive Values of Clinical and Screening Mammography



Notes

- Sensitivity and specificity may appear to be properties intrinsic to the examination and independent of the disease we would like to confirm or to exclude, which is not the case
  - Always relate the measures of diagnostic performance to a defined disease
- *Clinical Radiology* vs. *Screening Radiology*
  - Clinical radiology (symptomatic subjects): try to use examinations with a high sensitivity, even in the presence of a relatively low specificity
  - Screening radiology (asymptomatic subjects): try to use examinations with a high specificity, also accepting a trade-off for sensitivity
  - While in clinical radiology the major priority is to diagnose a symptomatic disease (possibly in an advanced stage), in screening radiology the diagnosis of an asymptomatic disease must be balanced by the need of a limited amount of useless diagnostic work-up in the screened population

Bayesian vs. Frequentist Statistics

- Concept of probability as a degree of our believing that an event happens (subjective probability) is the foundation of Bayesian statistics
- Frequentist statistics: classic viewpoint based on frequencies and proportions (objective probability)
- Frequentist methods are today mainly used in medical research
  - In part due to the possibility of presenting the reliability of an investigated hypothesis as a number (the well-known p value)
- With regard to the evaluation of diagnostic performance, Bayes' theorem has a basic conceptual relevance

Bayes' Theorem

$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)}$$

- P(y) is the *a priori* probability of y
- P(x | y) is the likelihood function;
- P(x) is the marginal probability (probability of observing x event without any previous information
- P(y | x) is the *a posteriori* probability of y, given x

- Theorem allows us to calculate the disease probability (the y event) after having obtained a positive result
  - That is, the post-test probability
- Concept of odds
  - Odds of disease is the ratio between the subjects with the disease and the subjects without the disease

Bayesian Statistics

- If odds = a/b then frequency in the whole sample = a/(a+b)
- Conversely, if frequency in the whole sample = x then odds = x/(1-x)
- According to Bayes' theorem:

  odds of post-test disease = positive LR × odds of pretest disease

  - Positive LR = sensitivity/(1-specificity)

  odds of post-test disease absence = negative LR × odds of pretest disease

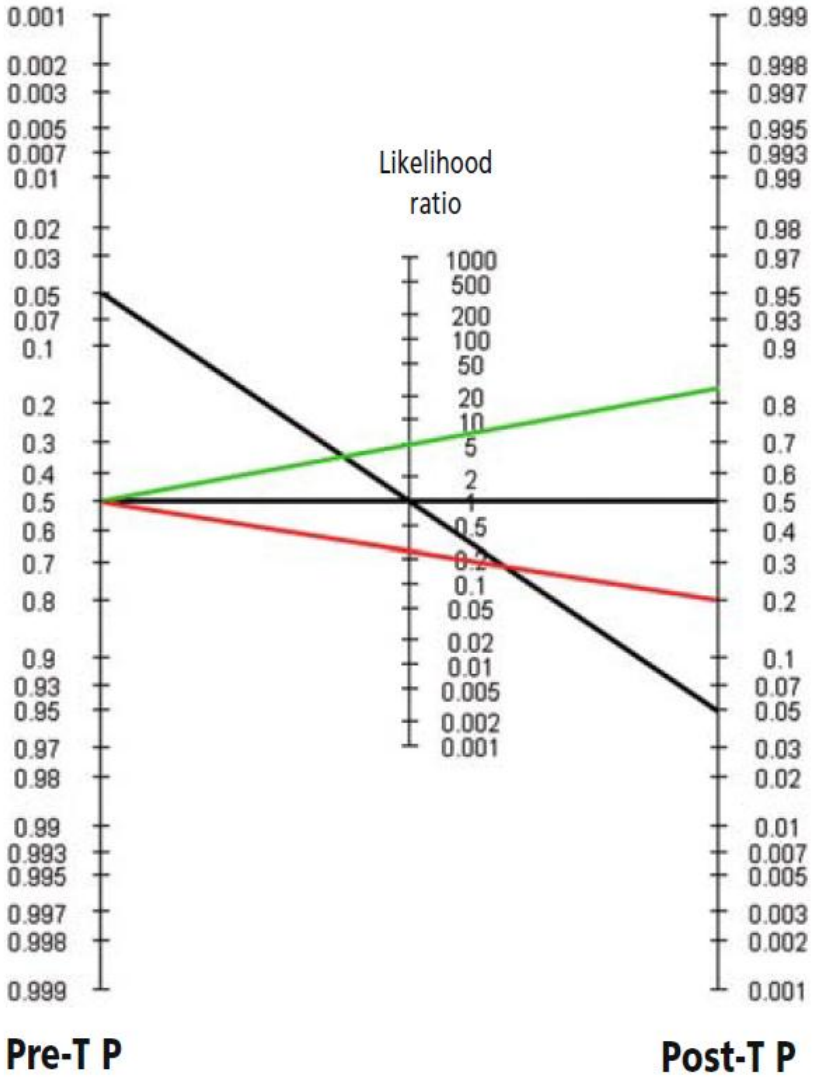
  - Negative LR = (1-sensitivity)/specificity
- In practice, when the positive LR of a test is known, the clinician can change the pretest probability into post-test probability
  - i.e. into the real diagnostic performance supplied by the test

Bayesian Statistics

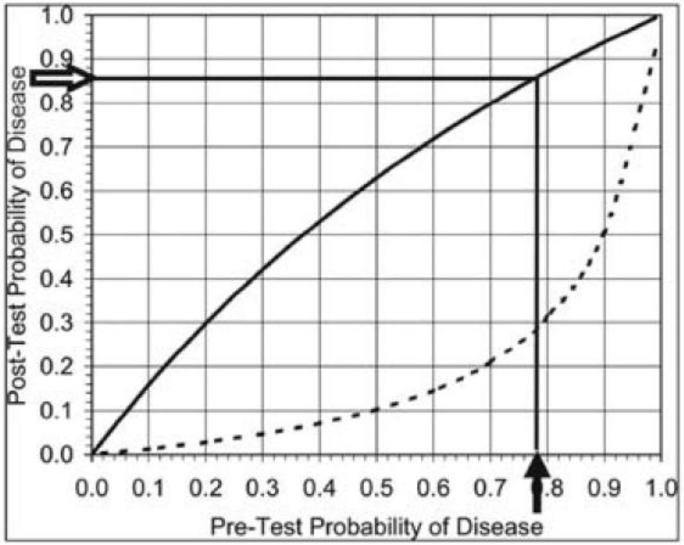
- Logical reasoning behind LRs is that they answer the questions:
  - To what extent does the positive result of the test increase disease probability (positive LR)?
  - To what extent does the negative result of the test reduce disease probability (negative LR)?
- Likelihood ratios quantify the "*power*" of an examination
  - When positive LR = negative LR = 1, no new information from test
  - When positive LR is high (or negative LR is low), diagnostic performance is high

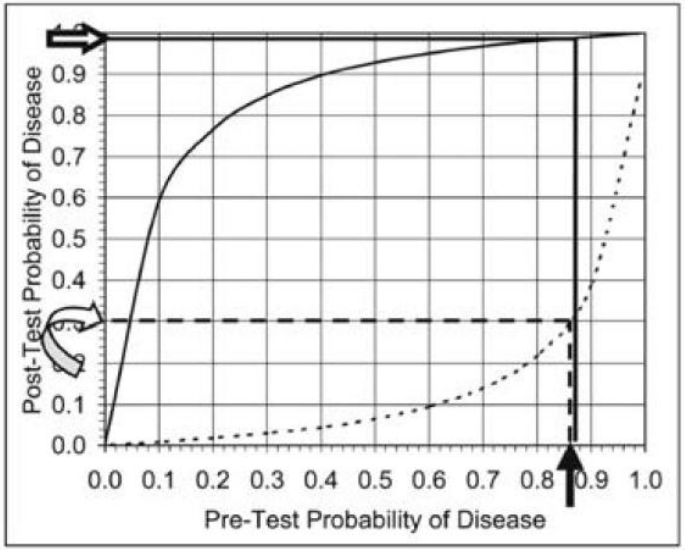
Fagan's Bayesian Nomogram

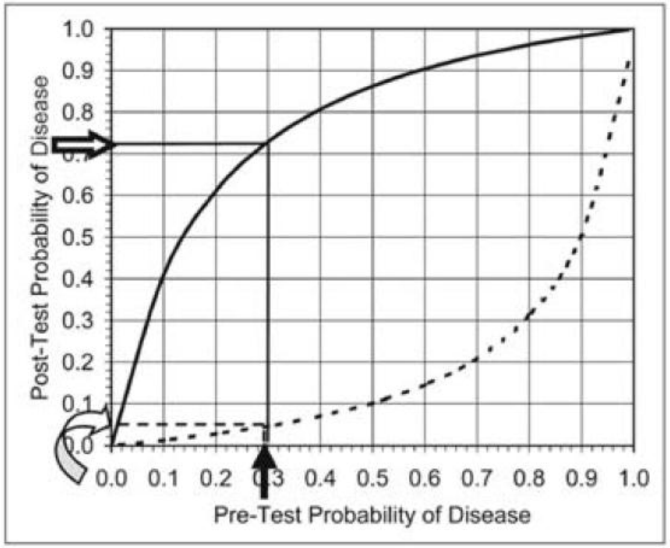
- Changes pretest disease probability into post-test disease probability using a geometric projection, without any need for calculation
- The slope of the straight line on the nomogram allows us to graphically see the power of the examination



Graphs of Conditional Probability (GCP)

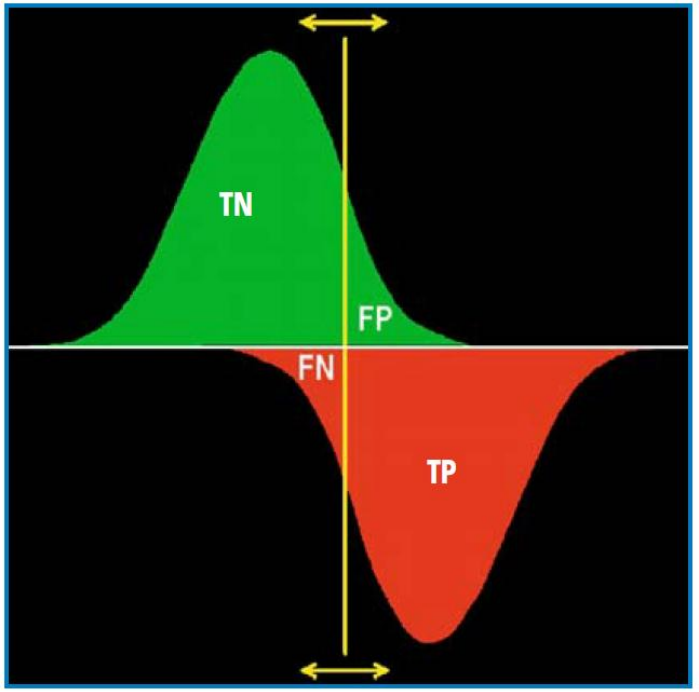




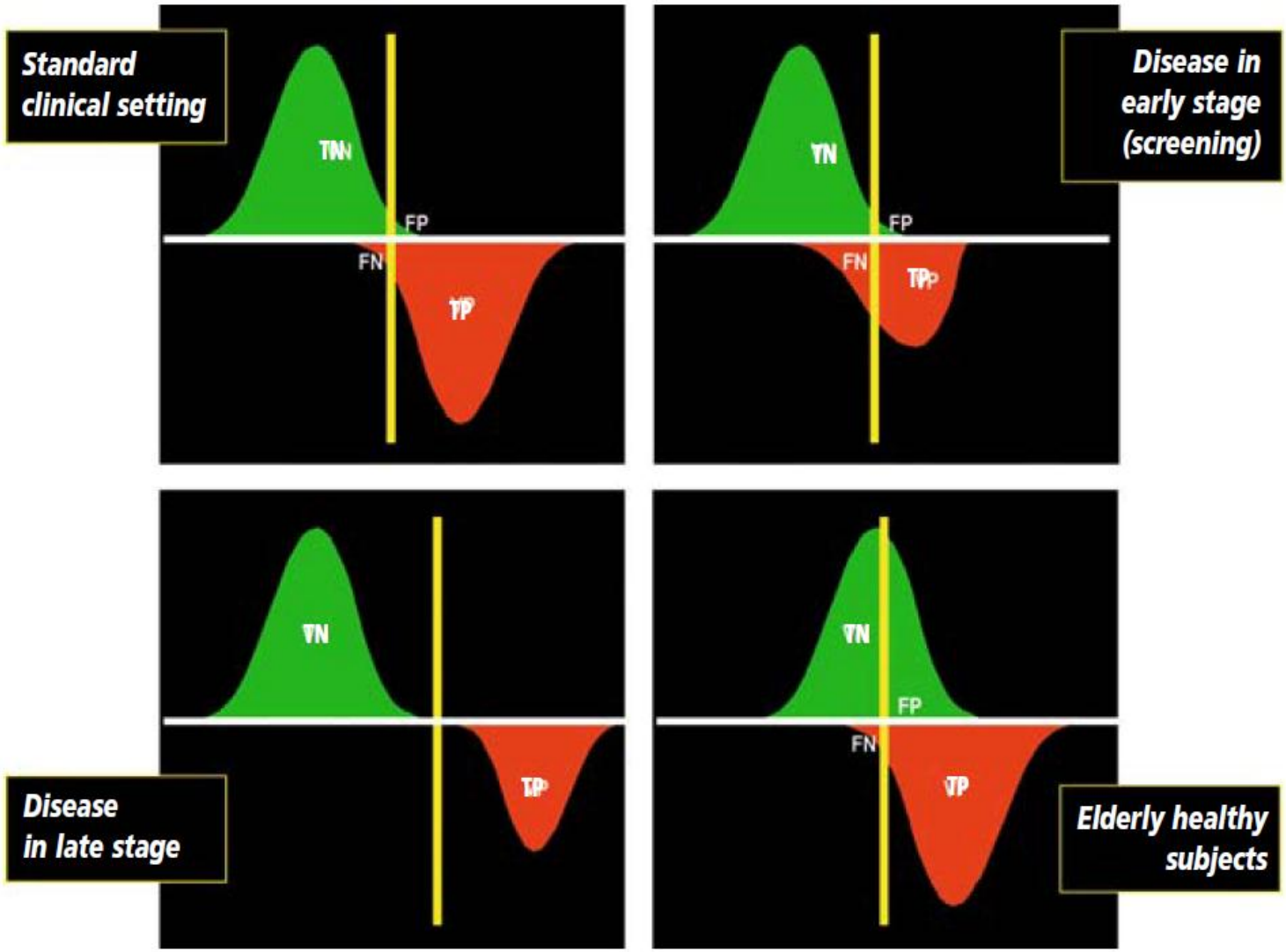


Thresholds and Cutoff

- Both the radiologist and the pathologist are required to give a dichotomous judgment (yes/no) about the malignancy of the lesion
  - Problem is related to the threshold we choose for our diagnostic decision, i.e. the cutoff
  - Above the cutoff a radiologic sign is considered predictive of a disease
  - If we lower the cutoff, we gain in sensitivity and lose in specificity
  - If we raise the cutoff, we gain in specificity and lose in sensitivity
- The cutoff could be optimized by choosing the level which minimizes total errors (sum of false negatives and false positives)

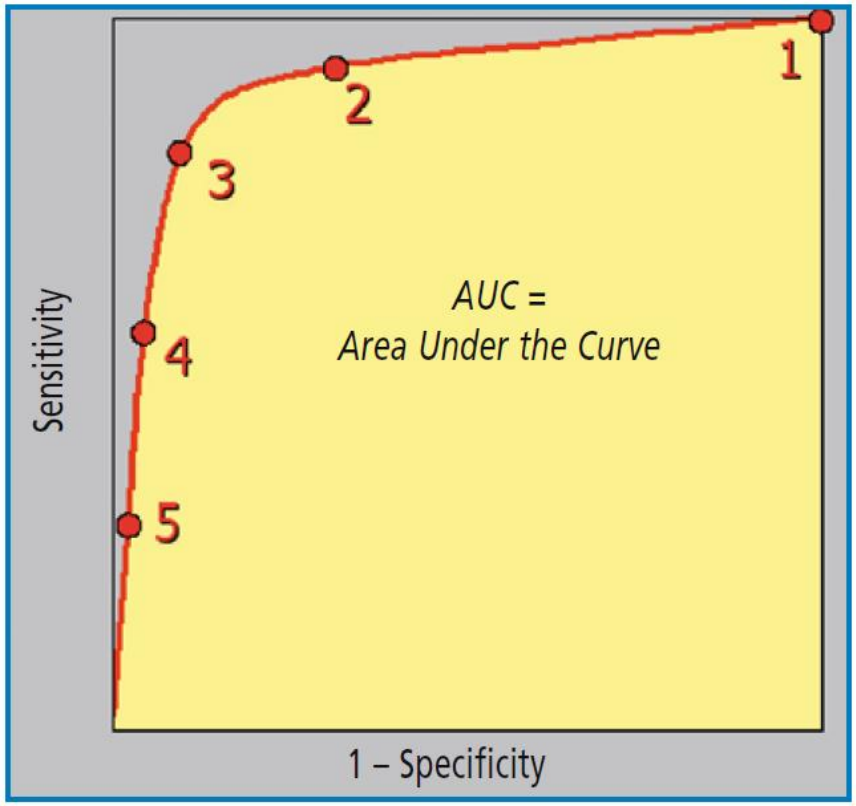


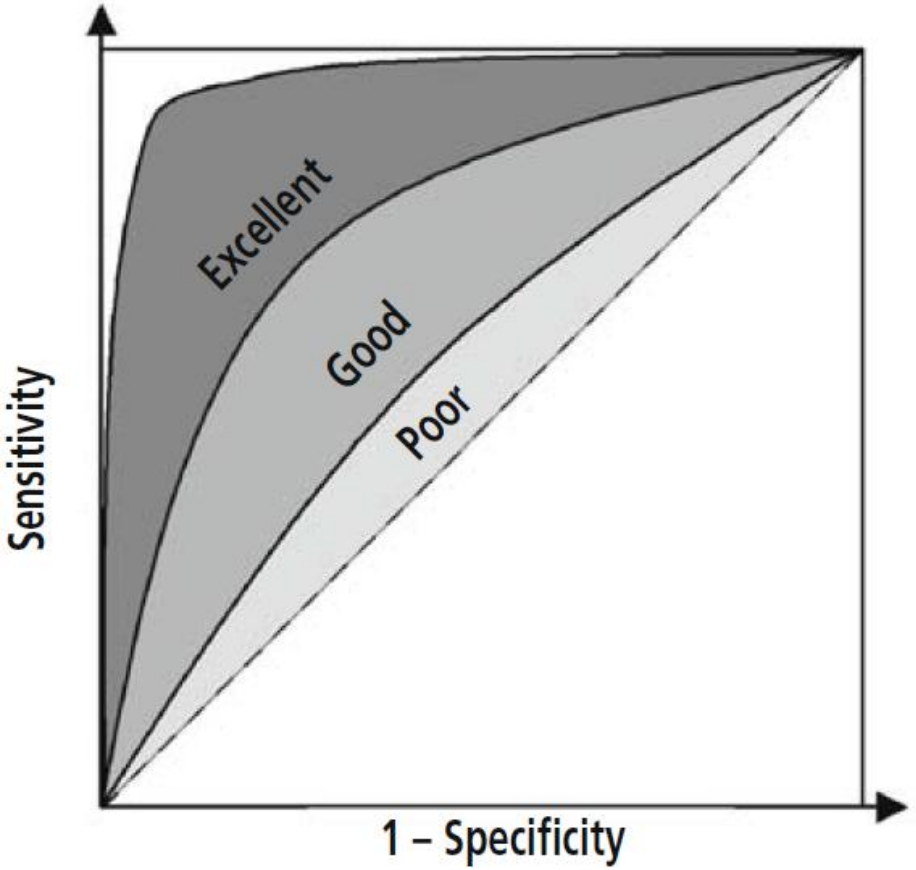
Role of Disease Spectrum



Receiver Operator Characteristic (ROC) Curve

- Sensitivity is graphed on the y-axis and (1– Specificity) on the x-axis using different cutoffs





Assignments

- Read the BIRADS system of mammography reporting. Are there any other scoring systems used in this area?
- Write code to implement all statistical diagnostic performance measures in this lecture.
- Download and read 1 paper on CAD in mammography and comment on their use of diagnostic performance measures to describe their technique.