"Designing Studies of Medical Tests"

Farhad Hosseinpanah
Obesity Research Center
Research Institute for Endocrine sciences
Shahid Beheshti University of Medical Sciences
September 20, 2023
Tehran

Agenda

- Introduction
- Studies of test **reproducibility**
- Studies of the **accuracy** of tests

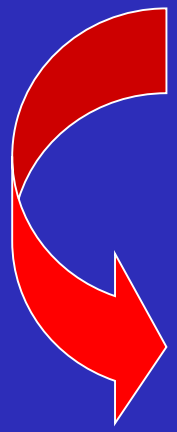
Introduction

Most designs for studies of

medical tests are **Descriptive**

Study of Diagnostic Test:

- **Causality is generally irrelevant**



- **Main goal is to determine usefulness in clinical practice**

Usefulness in clinical practice:

- **Accuracy**
- **Reproducibility**
- **Feasibility**
- **Effects on clinical decisions and outcomes**

■ TABLE 12.1
Questions to Determine Usefulness of a Medical Test, Possible Designs to Answer Them, and Statistics for Reporting Results

Question	Possible Designs	Statistics for Results*
How reproducible is the test?	Studies of intra- and interobserver and intra- and interlaboratory variability	Proportion agreement, kappa, coefficient of variation, mean and distribution of differences (avoid correlation coefficient)
How accurate is the test?	Cross-sectional, case control, or cohort-type designs in which test result is compared with a ``gold standard''	Sensitivity, specificity, positive and negative predictive value, ROC curves, and likelihood ratios
How often do test results affect clinical decisions?	Diagnostic yield studies, studies of pre- and posttest clinical decision making	Proportion abnormal, proportion with discordant results, proportion of tests leading to changes in clinical decisions; cost per abnormal result or per decision change
What are the costs, risks, and acceptability of the test?	Prospective or retrospective studies	Mean costs, proportions experiencing adverse effects, proportions willing to undergo the test
Does doing the test improve clinical outcome or have adverse effects?	Randomized trials, cohort or case-control studies in which the predictor variable is receiving the test and the outcome includes morbidity, mortality, or costs related either to the disease or to its treatment	Risk ratios, odds ratios, hazard ratios, number needed to treat, rates and ratios of desirable and undesirable outcomes

*Most statistics in this table should be presented with confidence intervals.

TABLE 4.3 The Precision and Accuracy of Measurements

	Precision	Accuracy
Definition	The degree to which a variable has nearly the same value when measured several times	The degree to which a variable actually represents what it is supposed to represent
Best way to assess	Comparison among repeated measures	Comparison with a reference standard
Value to study	Increase power to detect effects	Increase validity of conclusions
Threatened by	Random error (chance) contributed by The observer The subject The instrument	Systematic error (bias) contributed by The observer The subject The instrument

Studies of test reproducibility

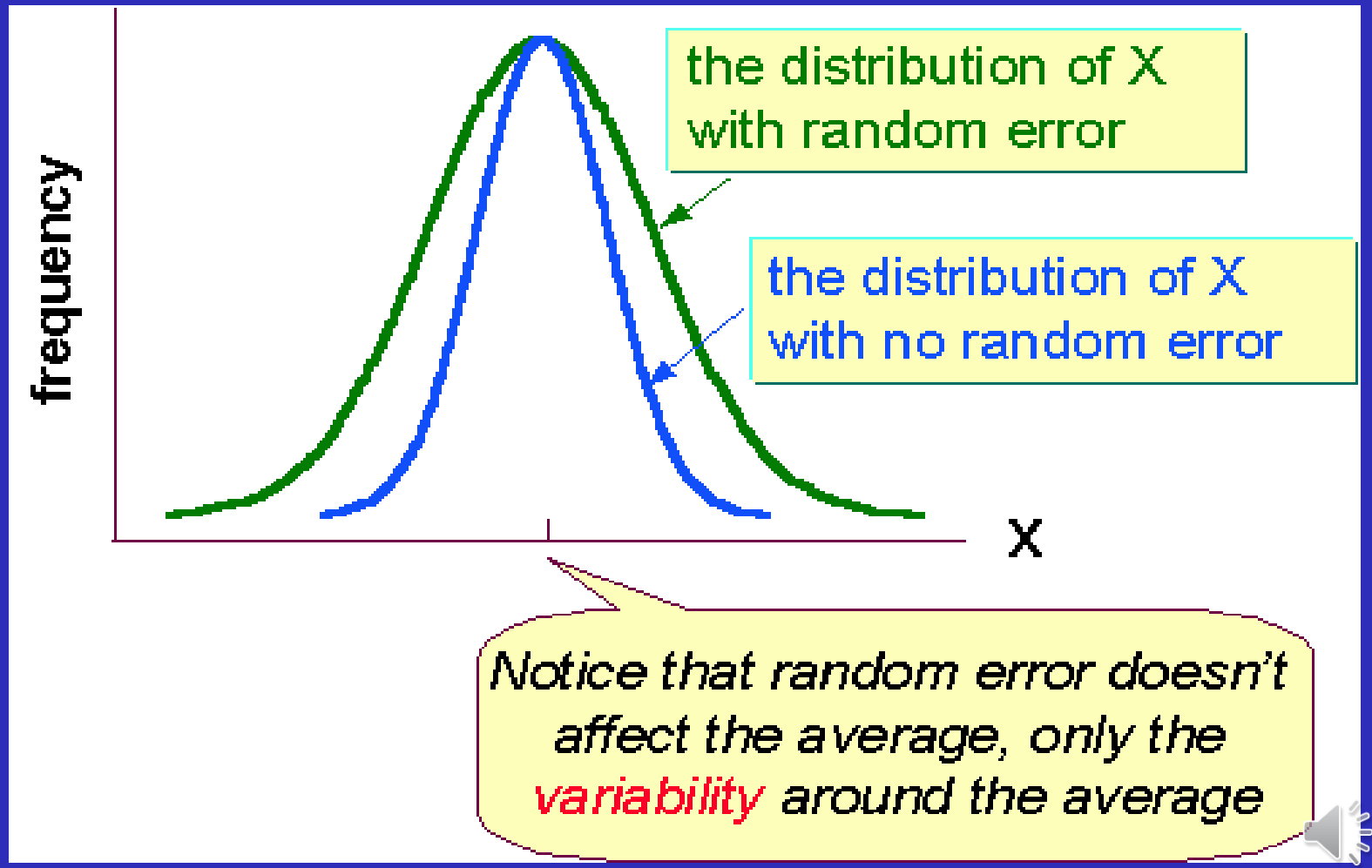
Precision
(or *reliability* or *reproducibility*)

- The extent that repeated measurements of a phenomenon tend to yield the **same results**
- **Precision refers to the *lack of random error***

– *Precision* ~ 1 / *random error*

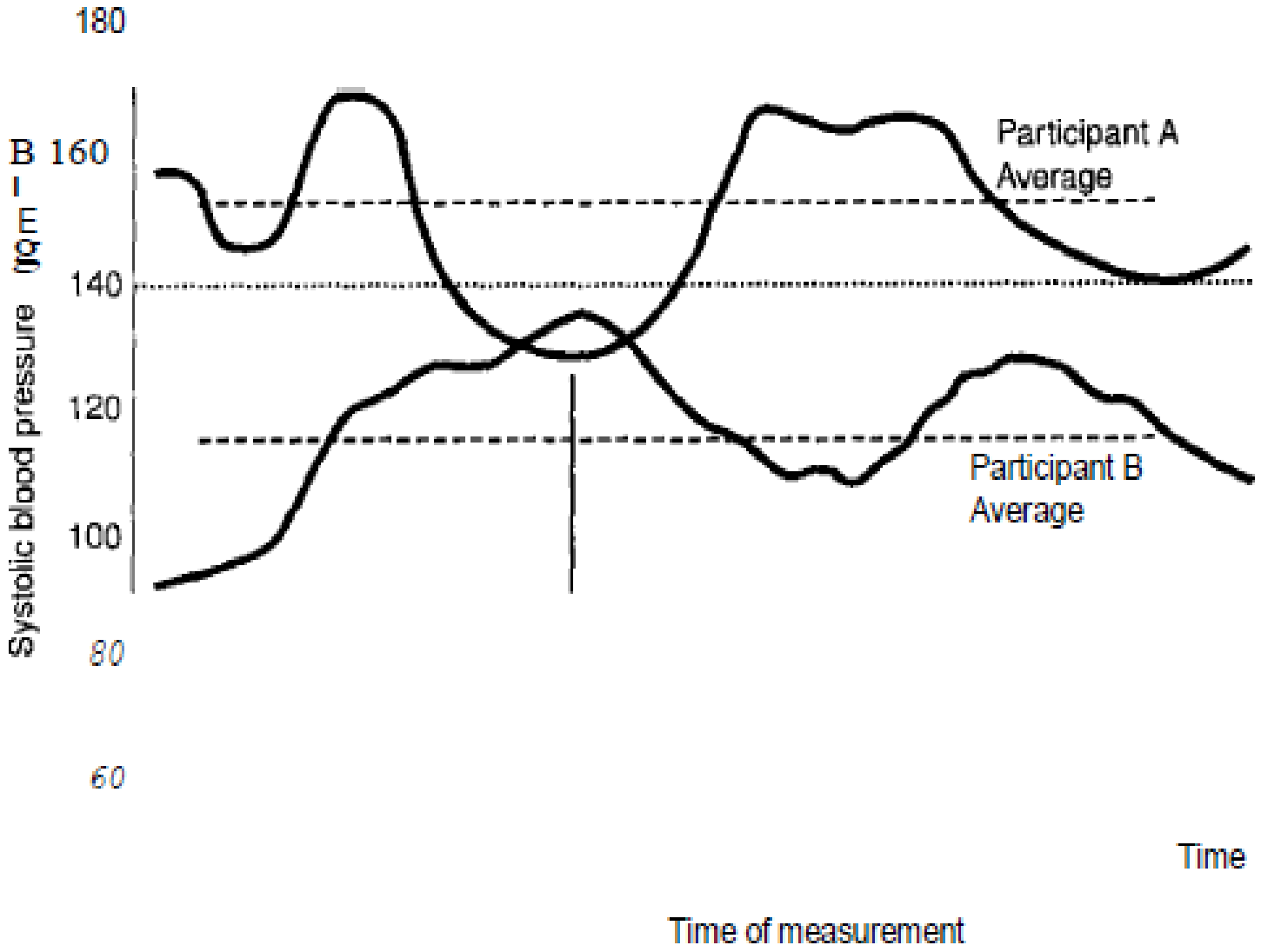
Measurement Error

Reliability



Variation in Clinical Data

- 1. ***Biologic Variation****= variation in the actual entity being measured*
  - derives from the dynamic nature of physiology, homeostasis and pathophysiology.
  - within (*intra-person*) *biologic* variability and,
  - between *(inter-person*) *biologic* variability


180
160
140
120
100
80
60
Systolic blood pressure (mm Hg)
Participant A
Average
Participant B
Average
Time
Time of measurement

Variation in Clinical Data

- 2. *Measurement Variation= variation due to the measurement process*

  – inaccuracy of the instrument (*instrument error*), and/or,

  – inaccuracy of the person (*operator error*)

Main sources of error

- Subject or biological variability (inter & intra person)
- Observer variability (Inter & Intra observer)
- Instrument variability (Within & Between Instrument)

Intraobserver variability

- Describes the lack of reproducibility in results when the same observer or laboratory performs the test at different times

- A radiologist , same chest radiograph , on two occasions

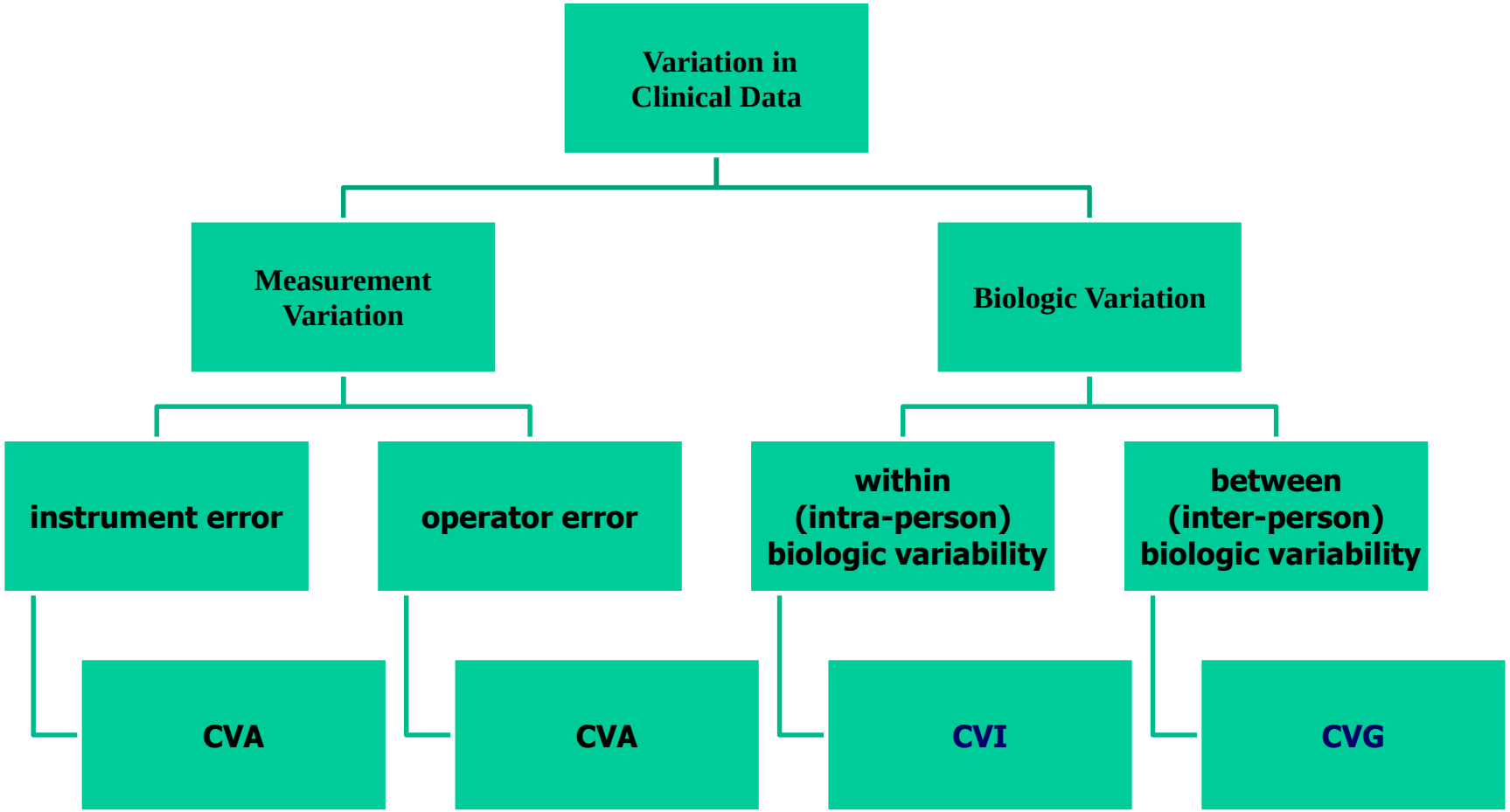
Interobserver variability

- Describes the lack of reproducibility among two or more observers
- Two radiologist , same chest radiograph

Important point

- Ideally, the ***only*** source of variability in a study should be that *between study participants.*

$$RCV = 2^{½} * Z * (CV_A^2 + CV_I^2)^{1/2}$$



Reproducibility study

- Cross sectional design
- Do not require a gold standard
- Categorical vs. continuous variables
- Main source of error

Reliability indices

Mean difference between paired measurements
Overall percent agreement
Kappa statistic
Weighted kappa
Coefficient of variation
Intraclass coefficient of variation

Categorical variables

- Overall percent agreement
- Kappa statistic
- Weighted kappa

Continous variables

- Mean difference between paired measurements
- Coefficient of variation
- Correlation coefficient
- Intraclass coefficient of variation

Agreement

A perfect standard is not available

Agreement

		imperfect standard positive	imperfect standard negative
New test	positive	a	b
	negative	c	d

overall percent agreement :

100% x (a+d)/(a+b+c+d)

Kappa

$$\frac{\text{\% observed agreement} - \text{\% agreement expected by chance}}{100\% - \text{\% agreement expected by chance}}$$

new test	Imperfect standard positive	negative	total
positive	41	3	44 (58.6%)
negative	4	27	31 (41.4%)
total	45 (60%)	30 (40%)	75 (100%)

percent observed agreement = $\frac{41 + 27}{75} * 100 = 90.7\%$

new test	Imperfect standard positive	Imperfect standard negative	total
positive	26.4		44 (58.6%)
negative			31 (41.4%)
total	45 (60%)	30 (40%)	75 (100%)

$$\frac{45 * 44}{75} = 26.4\%$$

new test	Imperfect standard positive	Imperfect standard negative	total
positive	**26.4**	17.6	44 (58.6%)
negative	18.6	**12.4**	31 (41.4%)
total	45 (60%)	30 (40%)	75 (100%)

percent agreement expected by chance = $\frac{26.4 + 12.4}{75}$ *100=51.7%

Kappa =

$$\frac{\text{\% observed agreement - \% agreement expected by chance}}{\text{100\% - \% agreement expected by chance}}$$

$$K = \frac{90.7\% - 51.7\%}{100\% - 51.7\%} = \frac{39\%}{48.3\%} = 0.81$$

Level of agreement

- > 0.75 excellent agreement
- 0.4-0.75 intermediate
- < 0.4 poor

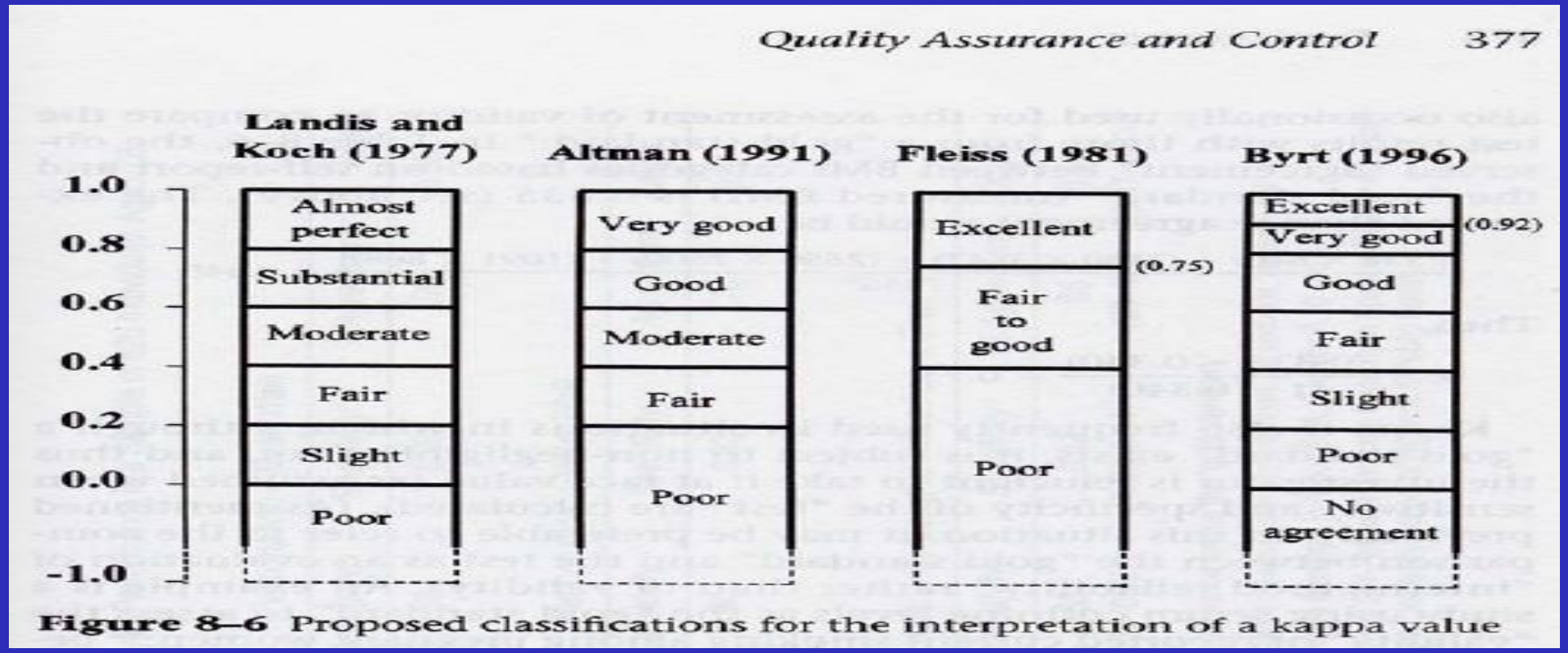


Figure 8–6 Proposed classifications for the interpretation of a kappa value

- ***Co-efficient of Variation (CV)***

$$\frac{SD}{\overline{X}} \%$$

– Represents the % variation of a set of measurements around their mean

– seful index for comparing the precision of different instruments, individuals and/or laboratories.

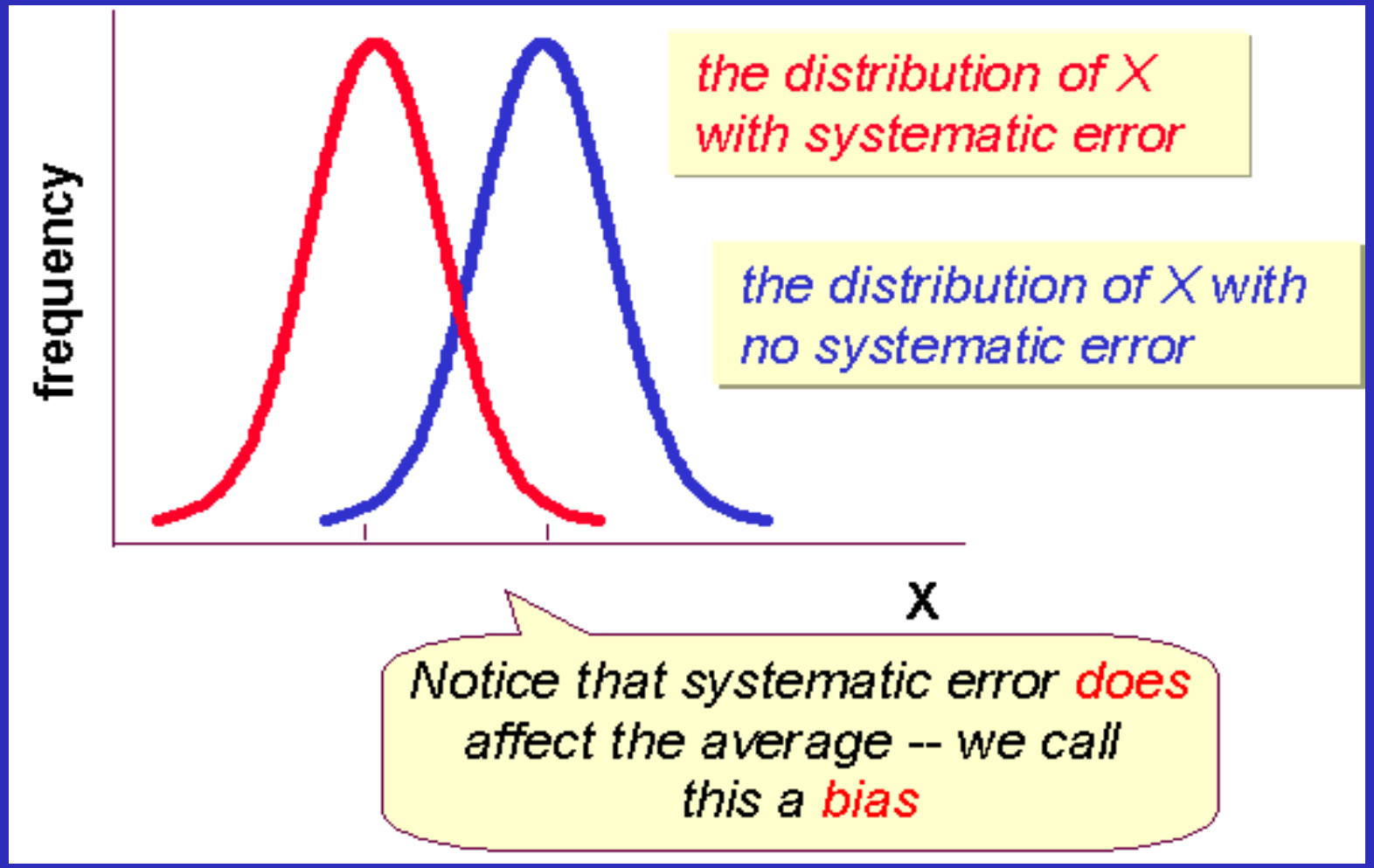
Studies of the accuracy of tests

Validity

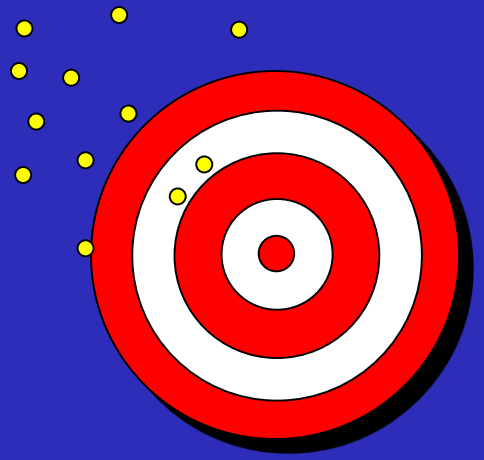
- Degree to which a measurement process measures what is intended i.e., accuracy.
- Lack of *systematic error* or *bias*.
- A valid instrument will, on average, be close to the underlying true value.
- Assessment of validity requires a **"gold standard"** (a reference).

Measurement Error

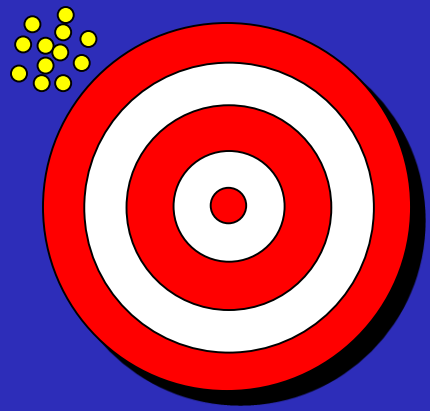
validity



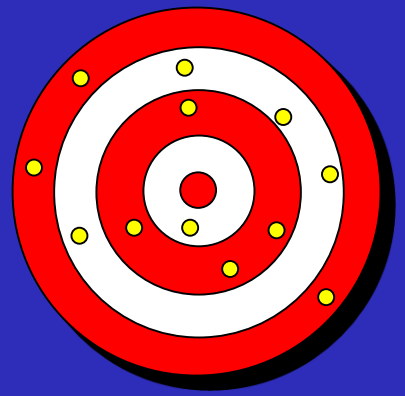
RELIABILITY AND VALIDITY



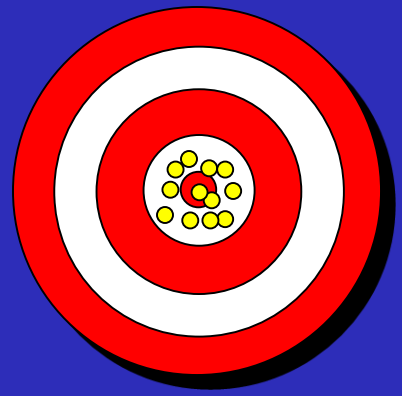
Not Reliable
Biased
(Not Valid)



Reliable
Biased
(Not Valid)



Not Reliable
Valid



Reliable
Valid

Random error – measurement not reliable

Systematic error – measurement biased (not valid)

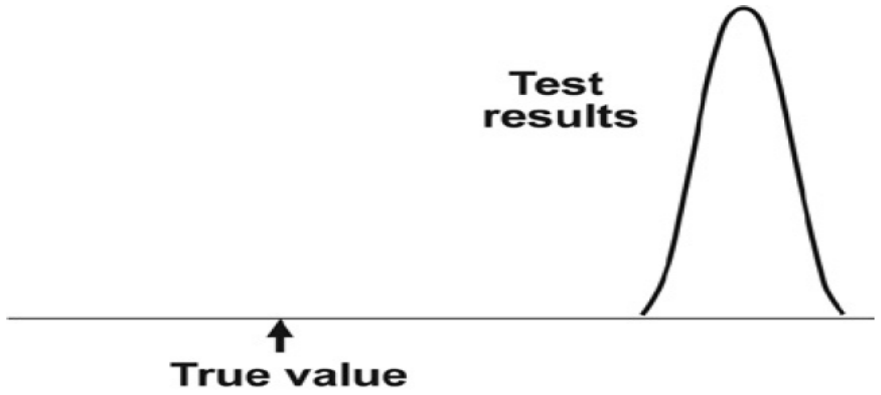


Fig. 5.18 Graph of hypothetical test results that are reliable, but not valid.

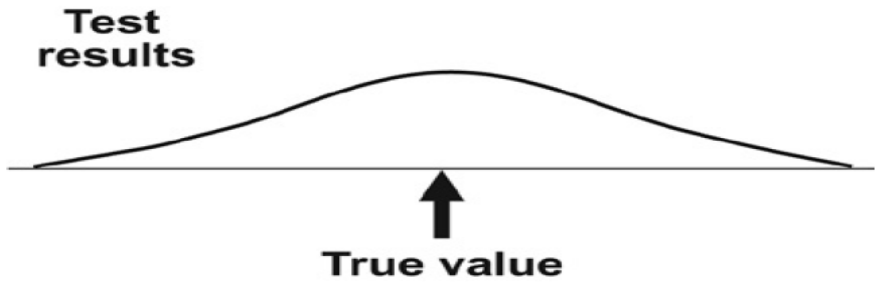


Fig. 5.19 Graph of hypothetical test results that are valid, but not reliable.

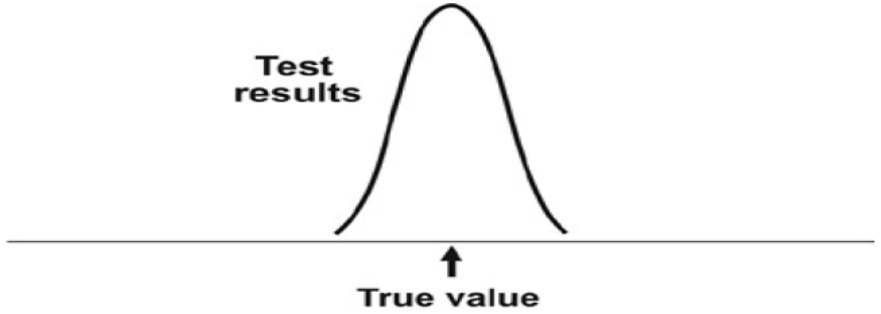


Fig. 5.20 Graph of hypothetical test results that are both valid and reliable.

- **Different phases of diagnostic studies**

Case-referent approach

Phase I questions:

Do test results in patients with the target disorder differ from those in normal people?

Table 1 Answering a phase I question: do patients with left ventricular dysfunction have higher concentrations of B-type natriuretic peptide (BNP) precursor than normal individuals?

	Patients known to have disorder	Normal controls
Median (range) concentration of BNP precursor (pg/ml)	493.5 (248.9-909.0)	129.4 (53.6-159.7)

test based enrolment

Phase II questions:

Are patients with certain test results more likely to have the target disorder than patients with other test results?

Table 2 Answering a phase II question: are patients with higher concentrations of B-type natriuretic peptide (BNP) more likely to have left ventricular dysfunction than patients with lower concentrations?

	Patients known to have target disorder	Normal controls
High BNP concentration	39	2
Normal BNP concentration	1	25

Test characteristics (95% CI):
- Sensitivity=98% (87% to 100%)
- Specificity=92% (77% to 98%)
- Positive predictive value=95% (84% to 99%)
- Negative predictive value=96% (81% to 100%)
- Likelihood ratio for an abnormal test result=13 (3.5 to 50.0)
- Likelihood ratio for a normal test result=0.03 (0.0003 to 0.19)

Survey of total study population

Phase III questions

Does the test result distinguish patients with and without the target disorder among patients in whom it is clinically **reasonable to suspect** that the disease is present?

- ✓**Must be prospective**
- ✓**Less clinical contrast**
- ✓**Larger sample size**

Table 3 Answering a phase III question: among patients in whom it is clinically sensible to suspect left ventricular dysfunction (LVD), does the concentration of B-type natriuretic peptide (BNP) distinguish patients with and without left ventricular dysfunction?

	Patients with LVD on echocardiography	Patients with normal results on echocardiography
Concentration of BNP:		
High (>17.9 pg/ml)	35	57
Normal (<18 pg/ml)	5	29
Prevalence (pretest probability) of LVD	40/126=32%	

Test characteristics (95% CI):
- Sensitivity=88% (74% to 94%)
- Specificity=34% (25% to 44%)
- Positive predictive value=38% (29% to 48%)
- Negative predictive value=85% (70% to 94%)
- Likelihood ratio for an abnormal test result=1.3 (1.1 to 1.6)
- Likelihood ratio for a normal test result=0.4 (0.2 to 0.9)

Phase IV questions

Do patients who undergo this diagnostic test fare better (in their ultimate health outcomes) than similar patients who are not tested?

Validity indices

Sensitivity
Specificity
PPV
NPV
Likelihood ratio
ROC curve

Outcome of diagnostic process

- **Dichotomous result**
  - **diseased or non-diseased**
- **Dichotomous measure**
  - **presence or absence of bacteria**
- **Continuous measure**
  - **Blood glucose, white cell count, antibody titre**
  - **convert to dichotomous result**
  - **choice of ‘cut-off’ or ‘normal range’**

Biologic Variation of Human Populations

- **Bimodal curve:** distribution with 2 peaks
  - Relatively easy to separate most of the population into 2 groups (e.g., ill & not ill; have condition or abnormality & do NOT have condition or abnormality)
  - Some fall into "gray zone" – may belong to either curve
  - Most human characteristics are NOT distributed bimodally

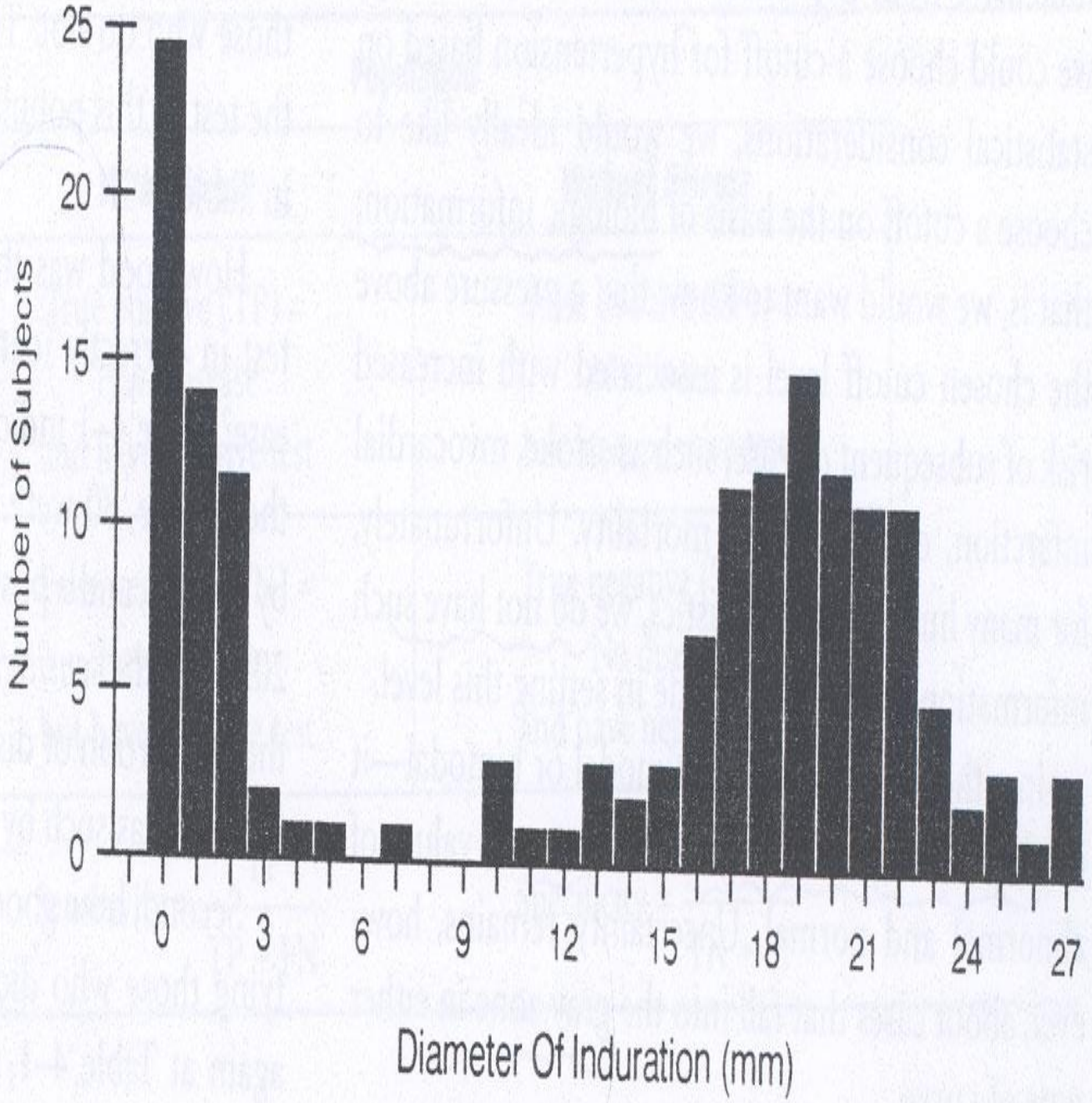
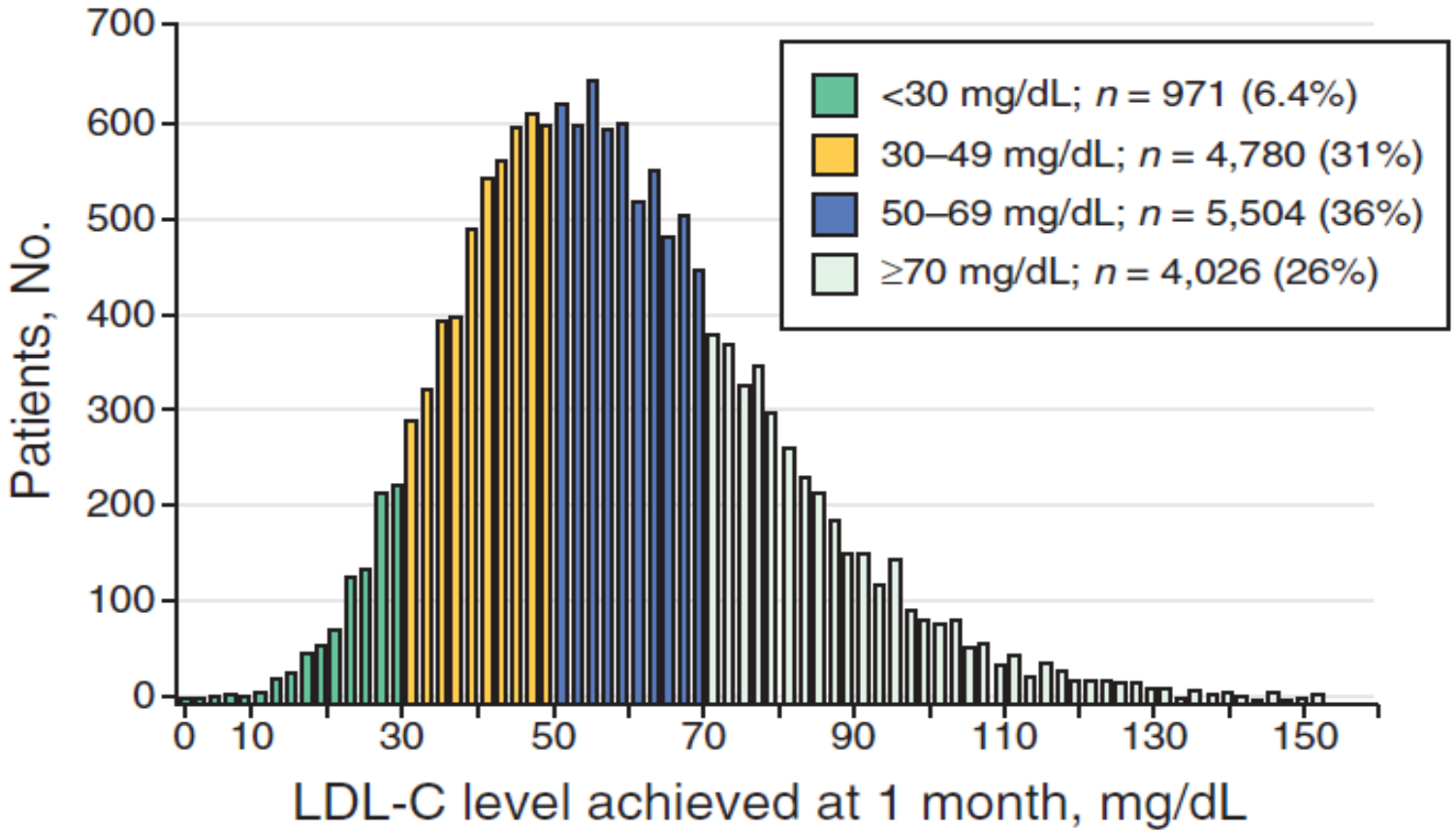


FIGURE 4–1. Distribution of tuberculin reactions. (Adapted from Edwards LB, Palmer CE, Magnus K: BCG Vaccination: Studies by the WHO Tuberculosis Research Office, Copenhagen, WHO Monograph No. 12. WHO, Geneva, 1953.)

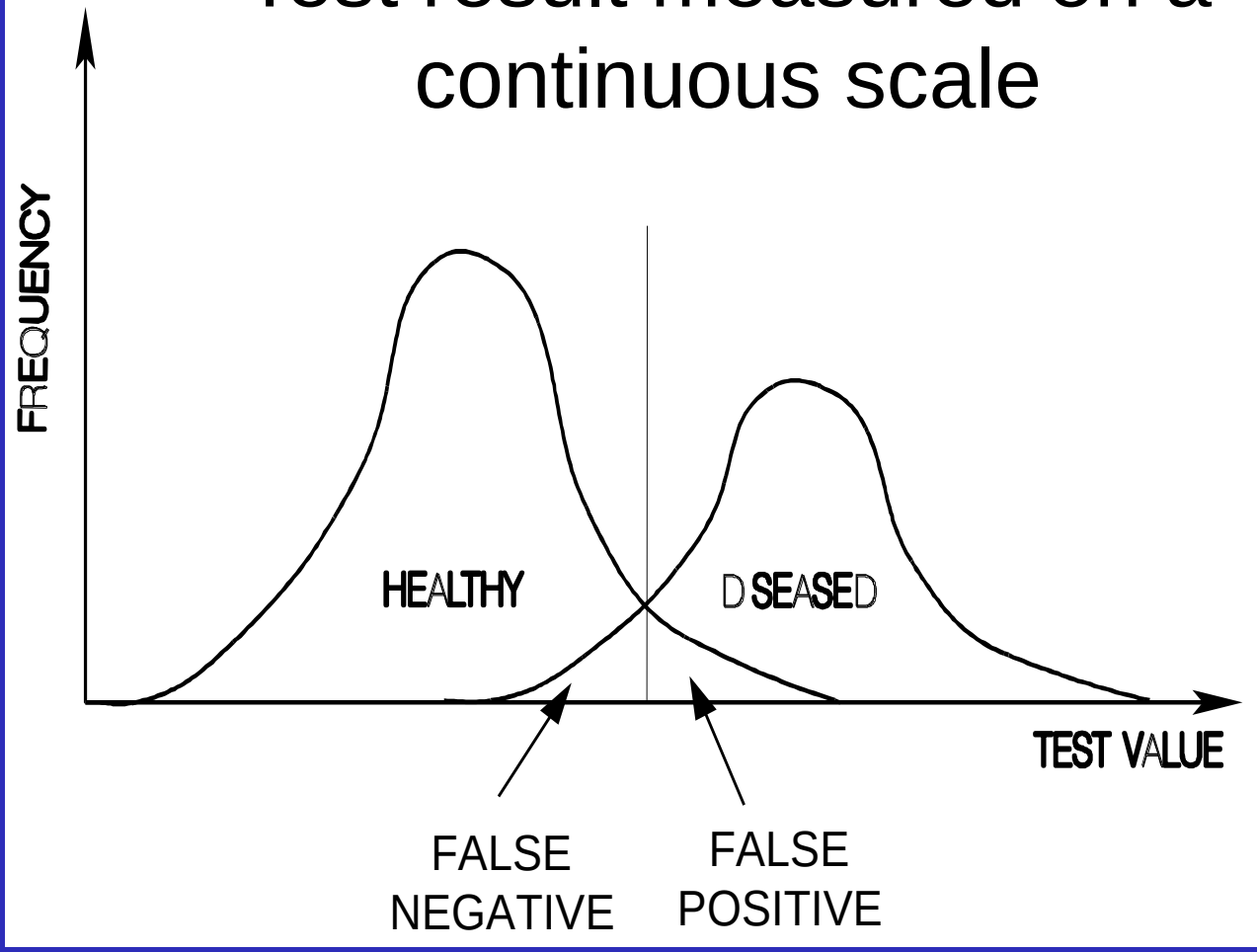
Biologic Variation of Human Populations

- **Unimodal curve:** distribution with 1 peak
  - Must set a cutoff level to distinguish those with condition & those without condition
  - Relatively easy to to distinguish the extreme values of abnormal & normal
  - Uncertainty about those cases in the gray zone



The median LDL-C level was 56 mg/dL (interquartile range, 43–70 mg/dL).
to convert LDL-C to millimoles per liter, multiply by 0.0259.

Test result measured on a continuous scale



Gold standard

- In any study of diagnosis, the method being evaluated has to be compared to something
- The best available test that is used as comparison is called the GOLD STANDARD

2 X 2 tables

- To evaluate results of diagnostic studies, we use a 2 X 2 table
- By convention, the gold standard goes across the top and the new test goes to the side
- The four quadrants of the 2 X 2 represent true positives, false positives, false negatives, and true negatives

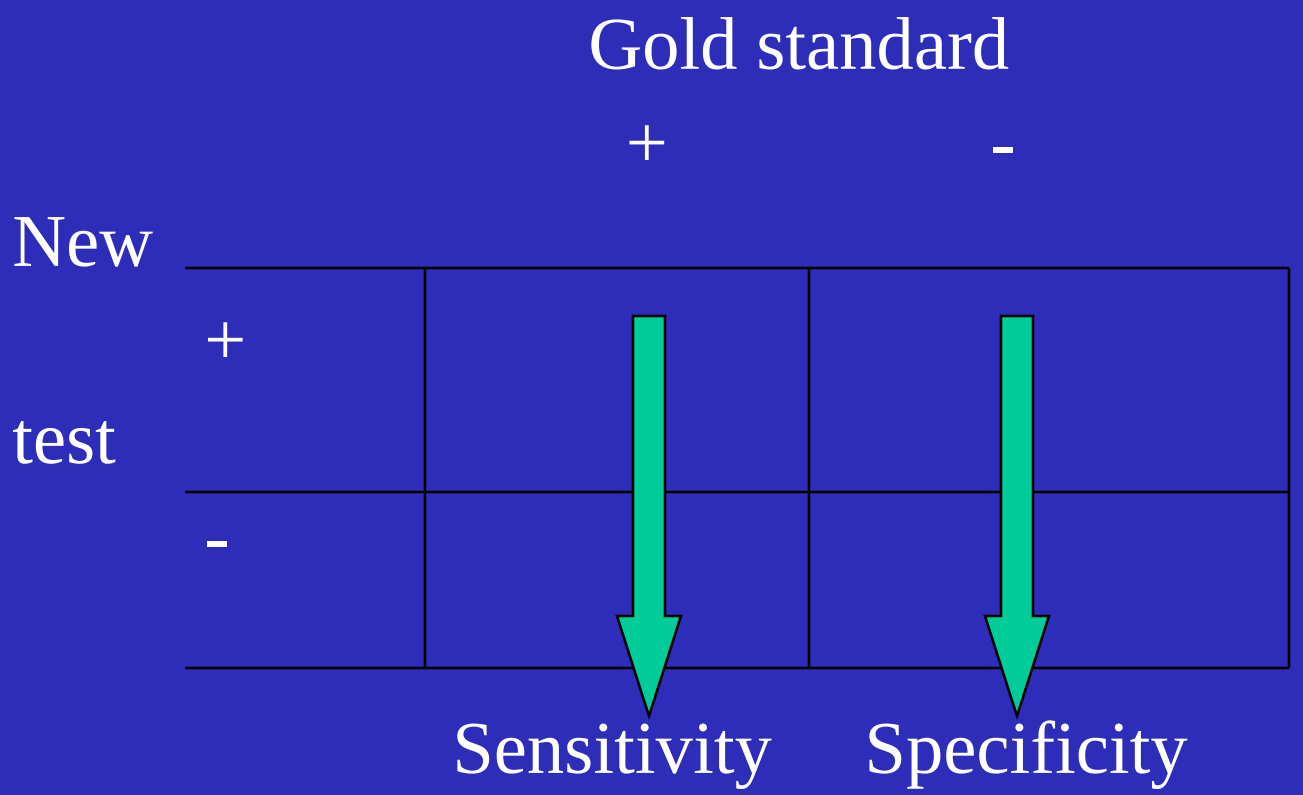
Information for a dichotomous test

		Disease	
		Present	Absent
Test Result	Positive	True positive A	False positive B
	Negative	False negative C	True negative D

Sensitivity = A / (A+C)

Specificity = D / (B+D)

Sensivity and specificity



Information for a dichotomous test

		Disease	
		Present	Absent
Test Result	Positive	True positive A = 103	False positive B = 16
	Negative	False negative C = 12	True negative D = 211

Information for a dichotomous test

		Disease: Present	Disease: Absent
Test Result	Positive	True positive A = 103	False positive B = 16
	Negative	False negative C = 12	True negative D = 211

Sensitivity=103/(103+12)=89%

Specificity=211/(16+211)=93%

Sensitivity and specificity

Limitations:

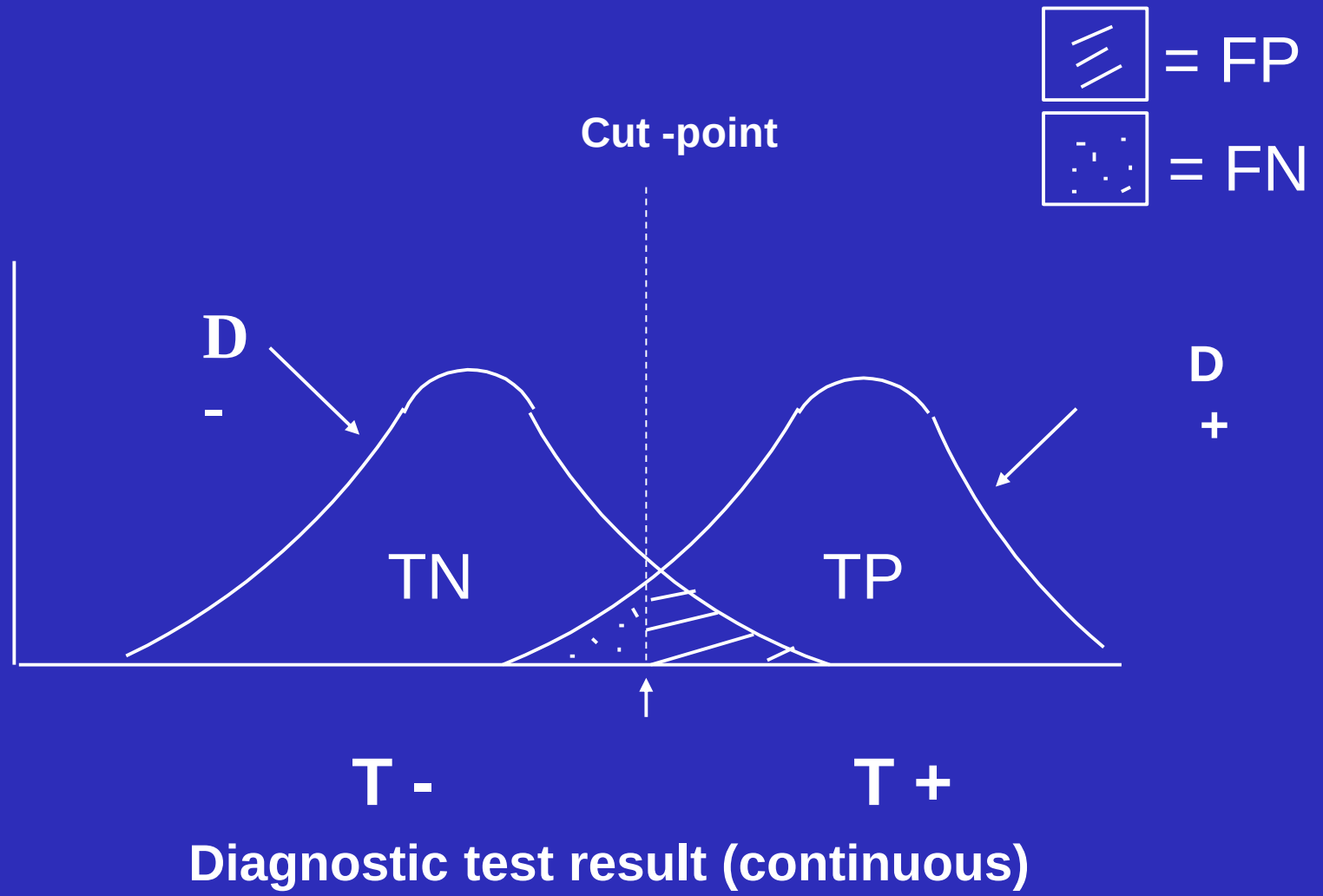
- we don’t know who has the disease before the test ! Otherwise we wouldn’t need to order the diagnostic test.

Sensitivity & Specificity of Tests

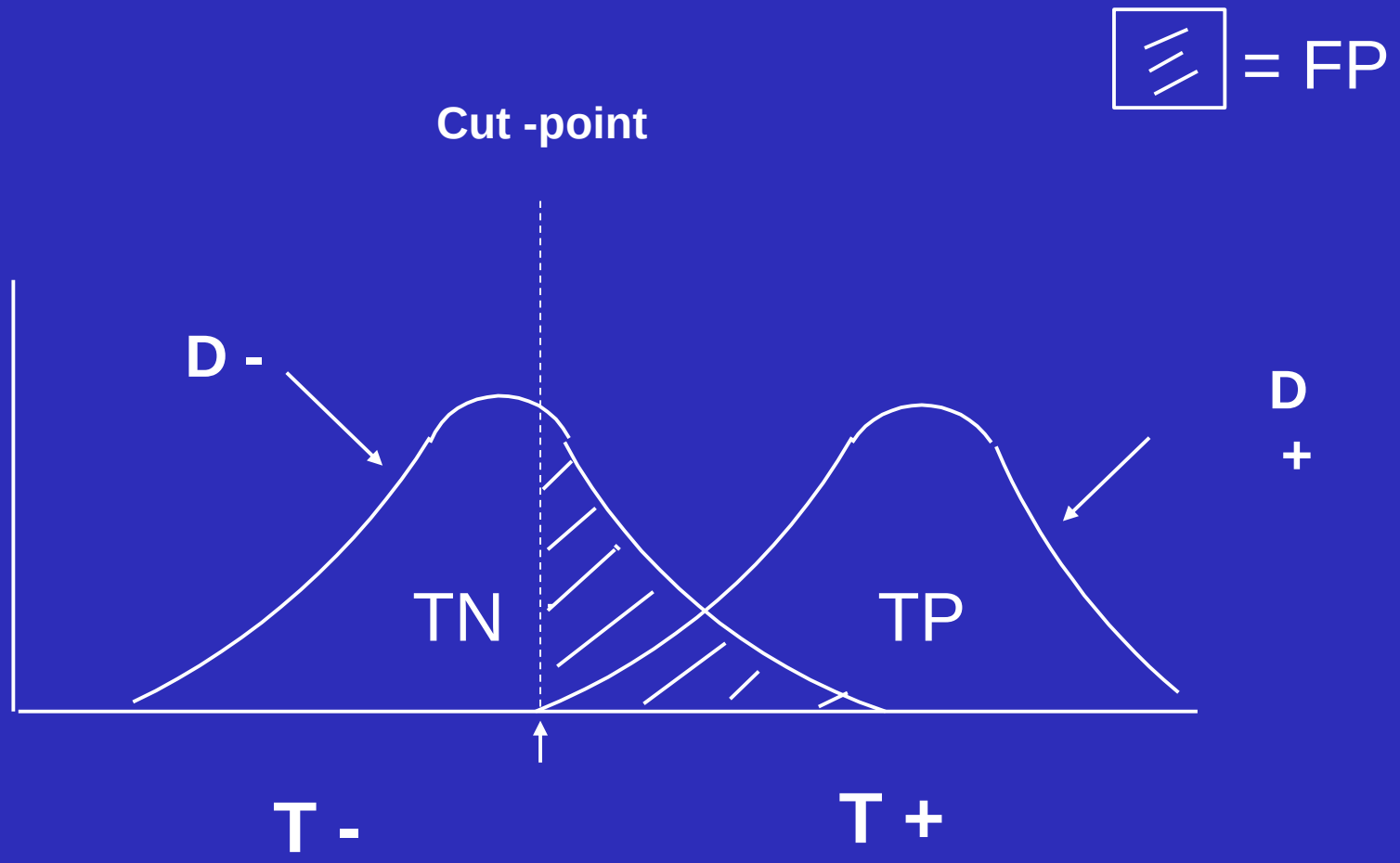
Limitations:

- Demand only 2 test results: + or -,hence all test data cannot be used, but must be collapsed by a “cut-point”.
- The “cut-point” can be reset and sen. & spec. recalculated.
- Plotting the results of several “cut-points” provides an ROC curve that reveals the overall quality of a test.

Results for a Typical Diagnostic Test Illustrating Overlap Between Disease (D+) and Non-disease (D-) Populations



Example of a Perfectly Sensitive Test



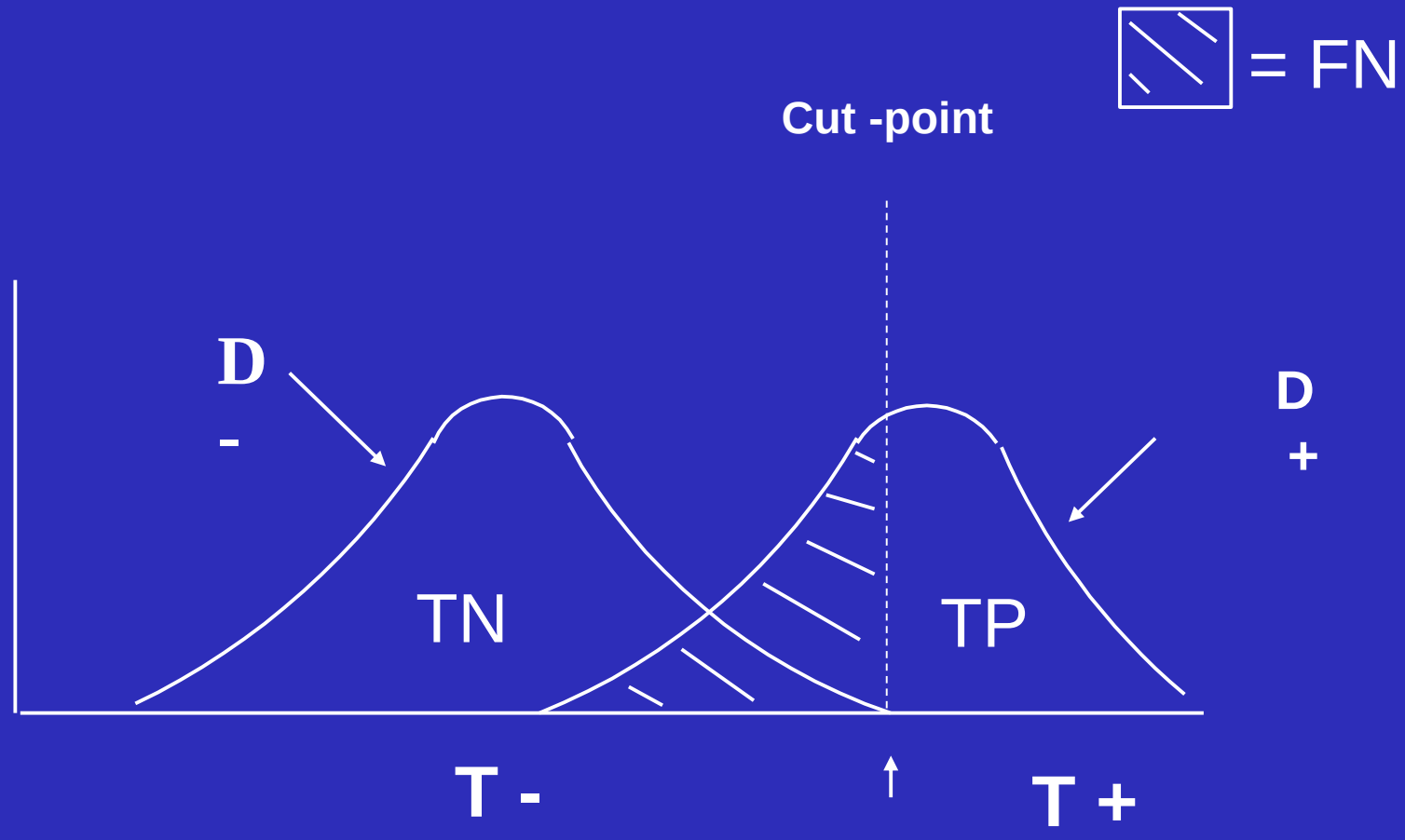
Tests with High Sensitivity

- perfectly sensitive test (Se = 100%), all diseased patients are test positive (no FN's)
- all test negative patients are disease free (TNs) but usually many FPs exist.
- highly sensitive tests are used to **rule-out** disease - if the test is negative you can be confident that disease is **absent** (FN results are rare!).
- highly sensitive tests **do not tell you if disease is present**, because they provides no information regarding FP's (see Sp).
- ***SnNout*** = if a sign, symptom or other diagnostic tests has a sufficiently high *Sen*sitivity, a *N*egative result rules *out* disease.

Tests with High Sensitivity

- Three clinical scenarios where high sensitivity test should be used:
  - 1) **Early stages of a diagnostic work-up**.
    - large number of potential diseases are being considered.
    - a negative result indicates a particular disease can be dropped (i.e., ruled out).
  - 2) **Important penalty for missing a disease**.
    - Examples - TB, syphilis - dangerous but treatable conditions.
    - don’t want to miss cases, hence avoid false negative results
  - 3) **Screening tests.**
    - the probability of disease is relatively low (i.e., low prevalence)
    - want to find as many asymptomatic cases as possible (incr. yield)

Example of a Perfectly Specific Test



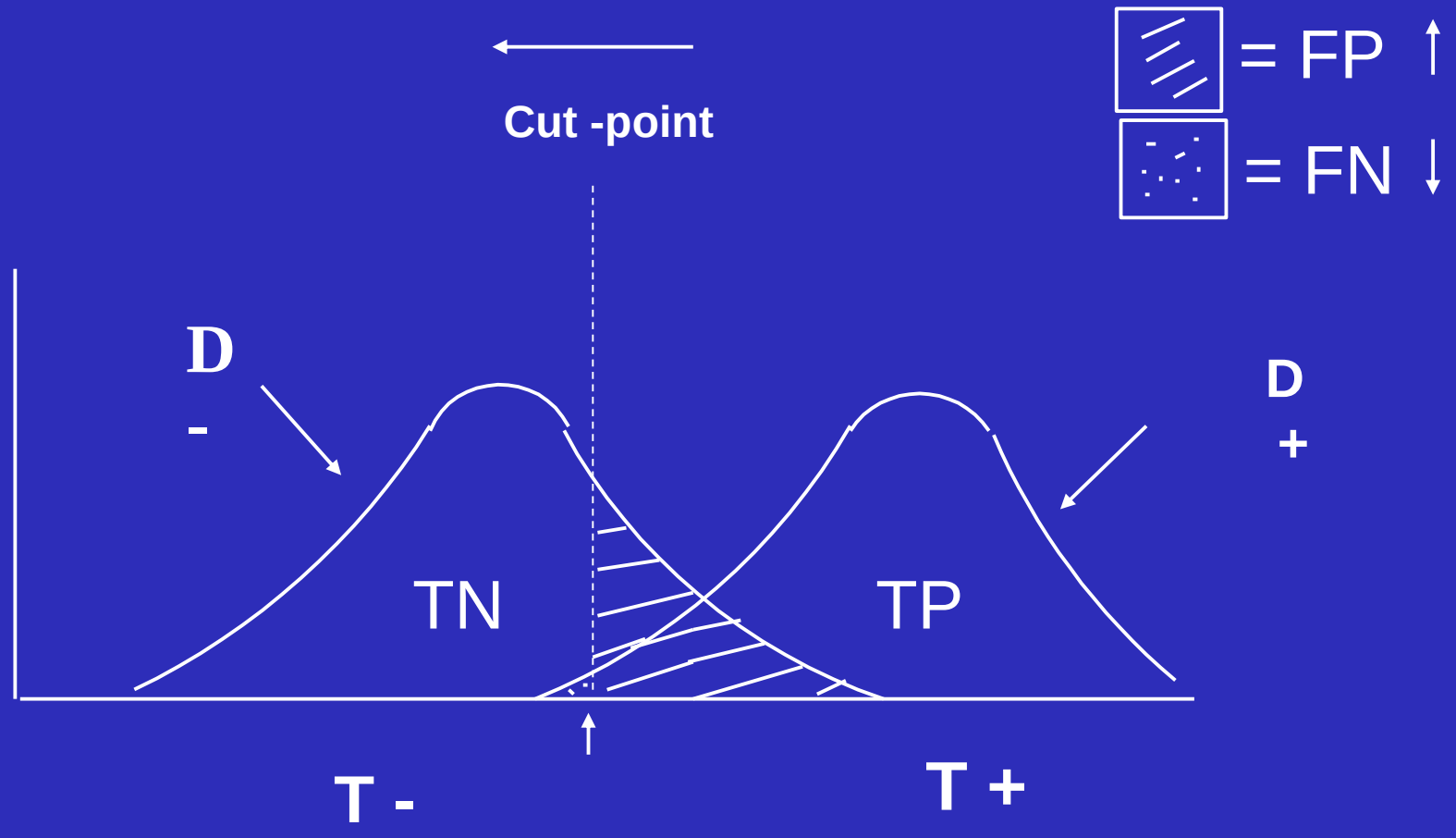
Tests with High Specificity

- perfectly specific test (Sp = 100%), all non-diseased patients test negative (no FP's)
- all test positive patients have disease (TPs) but usually sizeable number of FNs
- highly specific tests are used to **rule-in** disease - if the test is positive you can be confident that disease is **present** (FPs are rare).
- highly specific tests **do not tell you if disease is absent**, because Sp provides no information regarding FNs (see Se).
- ***SpPin*** = if a sign, symptom or other diagnostic tests has a sufficiently high *Sp*ecificity, a *P*ositive result rules *in* disease.

Tests with High Specificity

- Clinical scenarios when high specificity tests should be used:
  - 1) To rule-in a diagnosis suggested by other tests
    - specific tests are therefore used at the end of a work-up to rule-in a final diagnosis e.g., biopsy, culture, CT scan.
  - 2) False positive tests results can harm patient
    - want to be absolutely sure that disease is present.
    - example, the confirmation of HIV positive status or the confirmation of cancer prior to chemotherapy.

Trade-off Between Se and Sp: Lowering the Test Cut-point Increases Se but Decreases Sp



Using sensitivity and specificity

- As a consequence, you tend to believe a sensitive test when it is negative (rules out the disease)
- You tend to believe a specific test when it is positive (rules in the disease)
- Way to remember:
  - **Sensitive rules out (SNOUT)**
  - **Specific rules in (SPIN)**

Information for a dichotomous test

		Disease	
		Present	Absent
Test Result	Positive	True positive A = 103	False positive B = 16
	Negative	False negative C = 12	True negative D = 211

Information for a dichotomous test

		Disease Present	Absent
Test Result	Positive	True positive A	False positive B
	Negative	False negative C	True negative D

Sensitivity = A / (A+C)

Specificity = D / (B+D)

PPV = A / (A+B) $= \frac{\text{sen} * \text{prevalence}}{(\text{sen}*\text{prev}) + (1-\text{spec})*(1-\text{prev})}$

NPV = D / (C+D) $= \frac{\text{Spec}*(1-\text{prev})}{(1-\text{sen})*\text{prev} + \text{spec}*(1-\text{prev})}$

Information for a dichotomous test

Test Result	Disease: Present	Disease: Absent
Positive	True positive A = 103	False positive B = 16
Negative	False negative C = 12	True negative D = 211

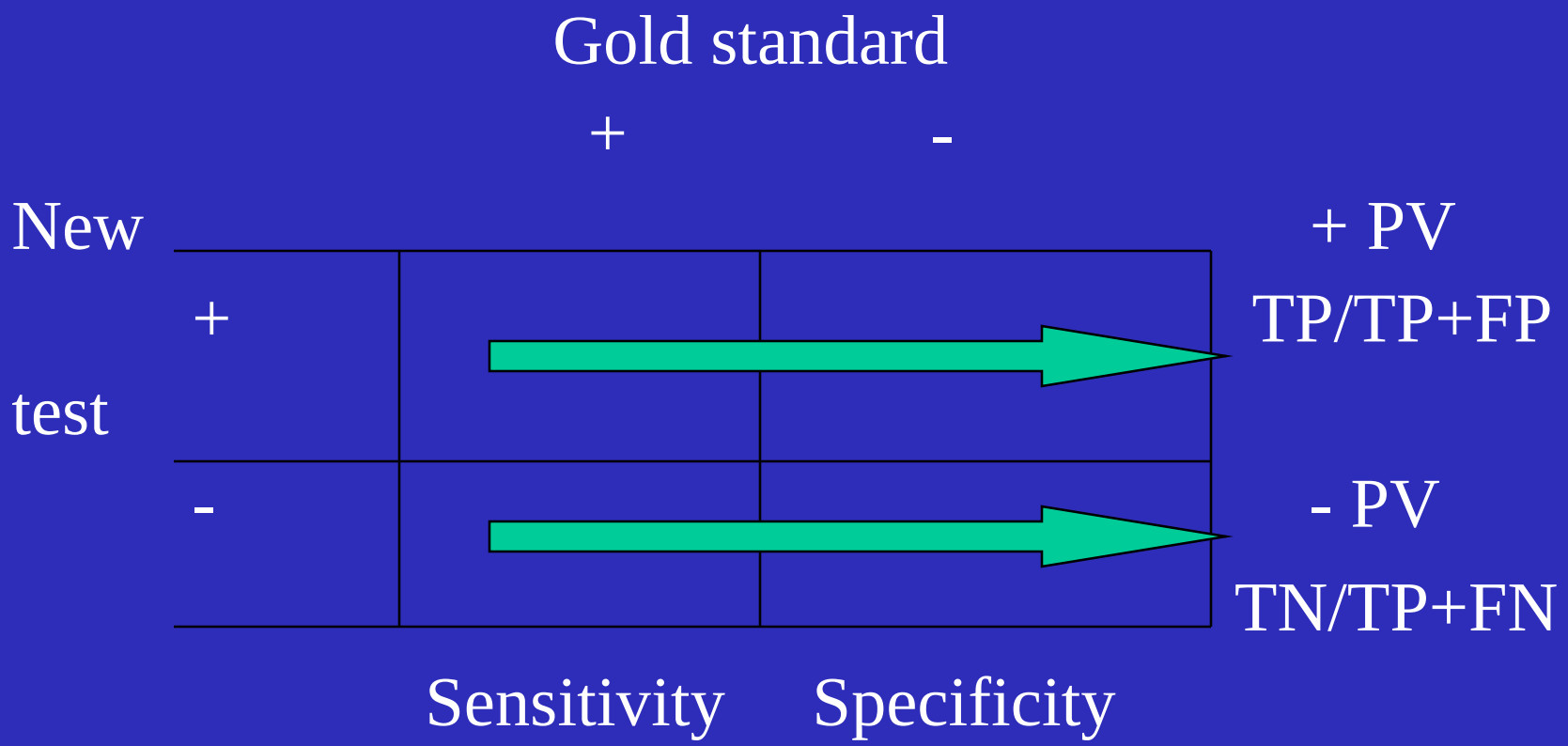
Sensitivity=103/(103+12)=89%

Specificity=211/(16+211)=93%

PPV = 103 / (103+16) = 86%

NPV = 211 / (12+211) = 94%

Pos and Neg Predictive Value



Using predictive values

- Tell you whether you should believe your test results: very clinically useful!!
- Unlike sens and spec, predictive values are dependent upon the test *and the population that you are testing*
- While sensitivity and specificity are constants, Predictive Values change depending upon who you are testing

Predictive values

- Limitation: predictive values are dependent on the fixed prevalence (pretest probability) of disease in the studied population. If the pretest probability of the disease is equal to prevalence of disease then the post test probability of disease will be equal to PPV (e.g in screening)

Example: low prevalence

1% of people have disease out of 1,000 tested

New	+	-	
+	9	99	+ PV 8.3%
-	1	891	- PV 99.9%
	Sensitivity 90%	Specificity 90%	

Example 2: High prevalence

99% of people have disease out of 1,000 tested

New	+	-	
+	891	1	+ PV 99.9%
-	99	9	- PV 8.3%
	Sensitivity 90%	Specificity 90%	

Example 3: Medium prevalence

50% of people have disease out of 1,000 tested

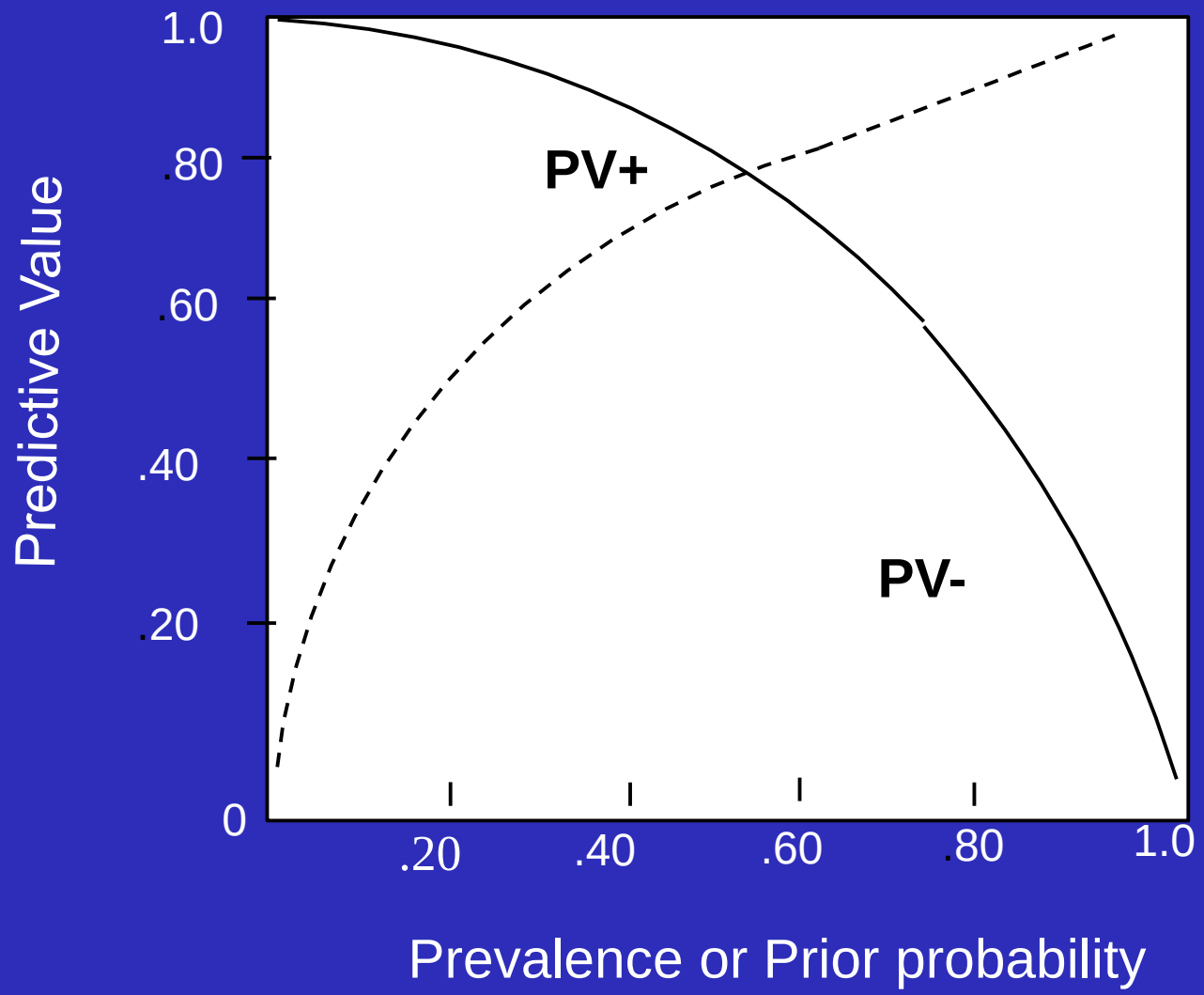
New	+	-	
+	450	50	+ PV 90%
-	50	450	- PV 90%
	Sensitivity 90%	Specificity 90%	

Sensitivity 80% , specificity 90%

Pretest probability

	1%	10%	50%	90%
PPV	7.5%	47.1%	88.9%	98.6%
NPV	99.8%	97.6%	81.8%	33.3%

The PVP and PVN as a Function of Prevalence for a Typical Diagnostic Test



- *As prevalence falls, positive predictive value must fall along with it, and negative predictive value must rise. Conversely, as prevalence increases, positive predictive value will increase and negative predictive value will fall.*

Likelihood ratio

- Likelihood ratio = the likelihood of a test result in patients *with* the disease / the likelihood of a test result in patients *without* the disease
  - LR(+) = sensitivity/(1-specificity)
  - LR(-) = (1-sensitivity)/specificity

Likelihood ratios

- Another way of looking at whether you should believe a test
- Compares the odds of having the disease before the test and the odds of having the disease after the test
- Based on the ratio of true positives (or negative) to false positive (or negatives)
- **Larger the likelihood ratio, better the test**

Information for a dichotomous test

		Disease	
		Present	Absent
Test Result	Positive	True positive A = 103	False positive B = 16
	Negative	False negative C = 12	True negative D = 211

Information for a dichotomous test

Test Result	Disease: Present	Disease: Absent
Positive	True positive A	False positive B
Negative	False negative C	True negative D

Sensitivity = A / (A+C)

Specificity = D / (B+D)

PPV = A / (A+B)

NPV = D / (C+D)

$$LR(+) = \frac{A/(A+C)}{B/(B+D)} = sn / (1-sp)$$

$$LR(-) = \frac{C/(A+C)}{D/(B+D)} = (1-sn) / sp$$

Information for a dichotomous test

Test Result	Disease: Present	Disease: Absent
Positive	True positive A = 103	False positive B = 16
Negative	False negative C = 12	True negative D = 211

Sensitivity=103/(103+12)=89%

Specificity=211/(16+211)=93%

PPV = 103 / (103+16) = 86%

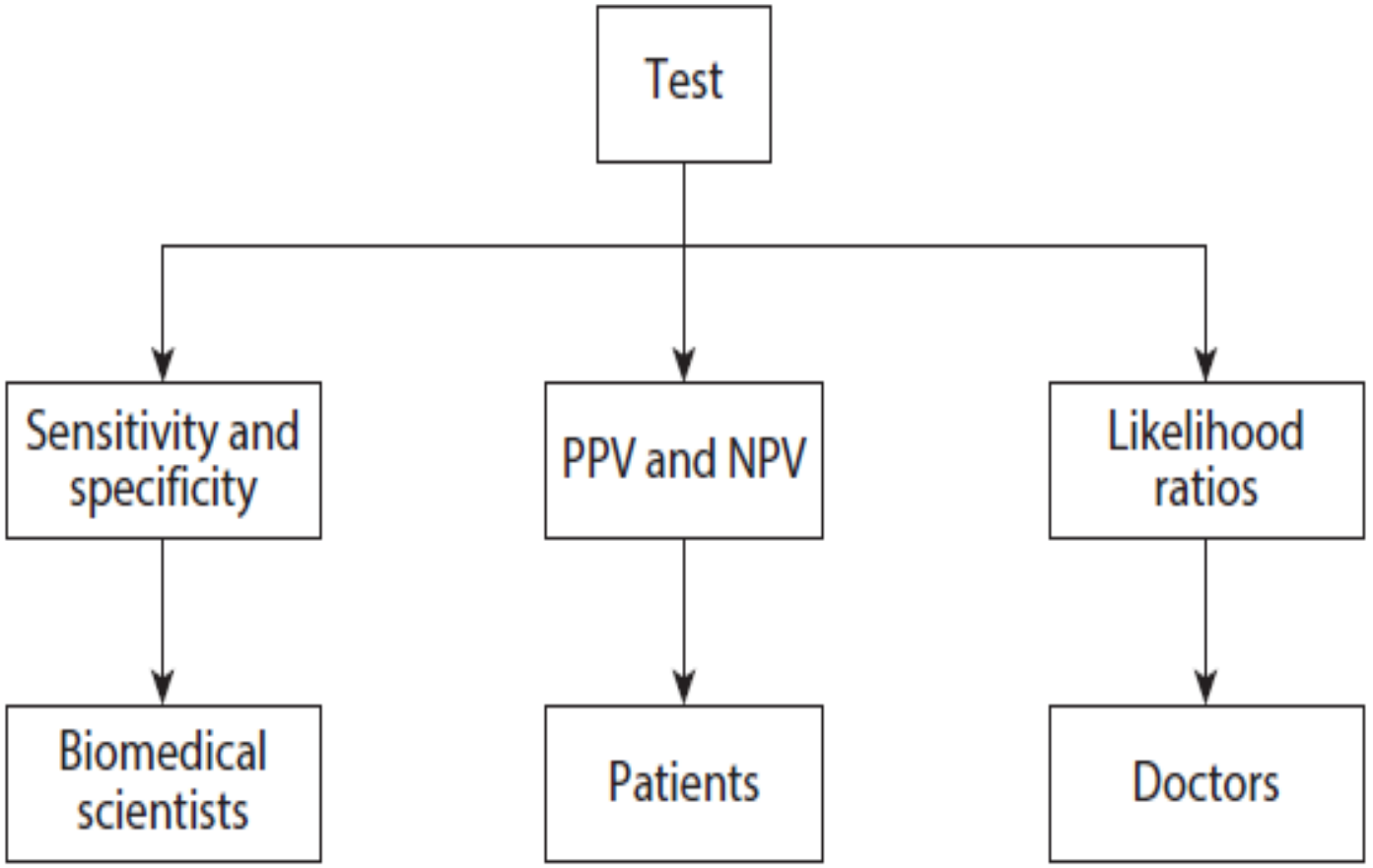
NPV = 211 / (12+211) = 94%

$$LR(+) = \frac{A/(A+C)}{B/(B+D)} = sn / (1-sp) = 12.7$$

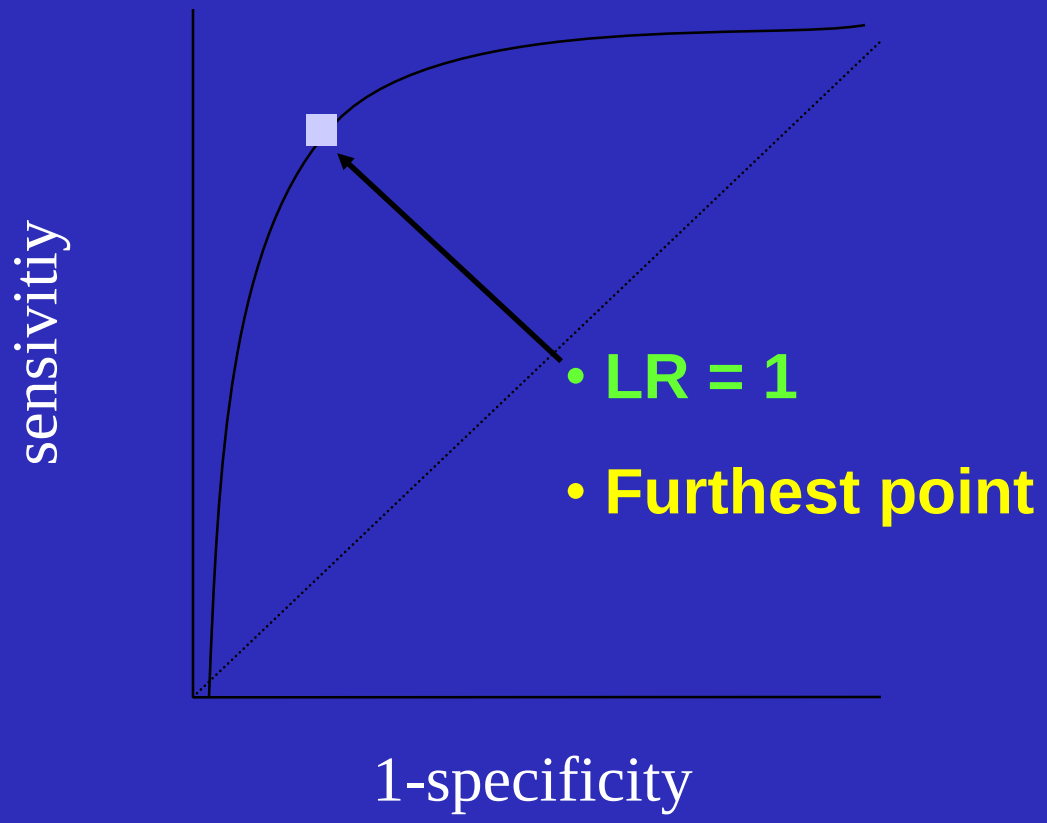
$$LR(-) = \frac{C/(A+C)}{D/(B+D)} = (1-sn) / sp = 0.11$$

LR	Interpretation
> 10	Large and often conclusive increase in the likelihood of disease
5 - 10	Moderate increase in the likelihood of disease
2 - 5	Small increase in the likelihood of disease
1 - 2	Minimal increase in the likelihood of disease

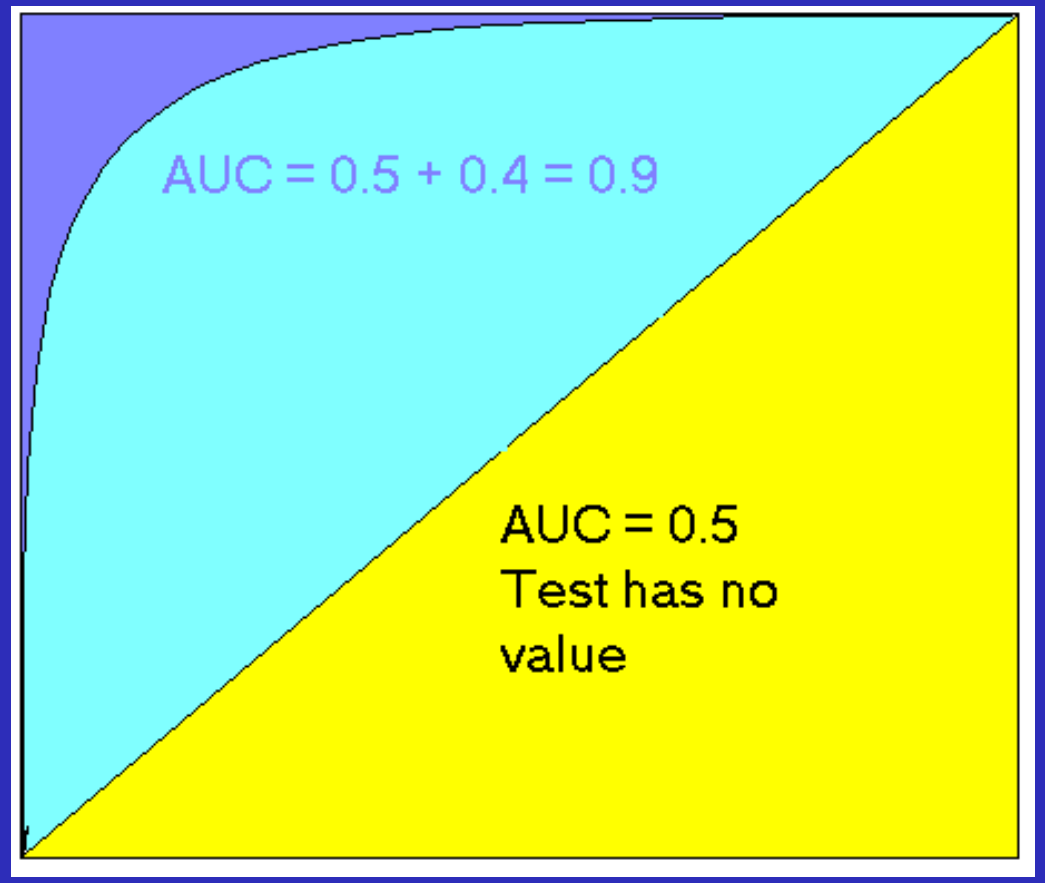
LR	INTERPRETATION
1	No change in the likelihood of disease
0.5 - 1.0	Minimal decrease in the likelihood of disease
0.2 - 0.5	Small decrease in the likelihood of disease
0.1 - 0.2	Moderate decrease in the likelihood of disease
< 0.1	Large and often conclusive decrease in the likelihood of disease


Test
Sensitivity and specificity
PPV and NPV
Likelihood ratios
Biomedical scientists
Patients
Doctors

Conventional cut-off using ROC curve



ROC for test evaluation



AUC = 1: perfect test

>0.9: high accuracy

0.7-0.9: moderate

0.5-0.7: less accurate

ROC curve

- Roc curve is simply a graph of the pairs of TPR and FPR
- Useful to determine cutoff
- ROC curves are an excellent way to compare diagnostic tests for same target disorder
- ROC curve can show global accuracy of test

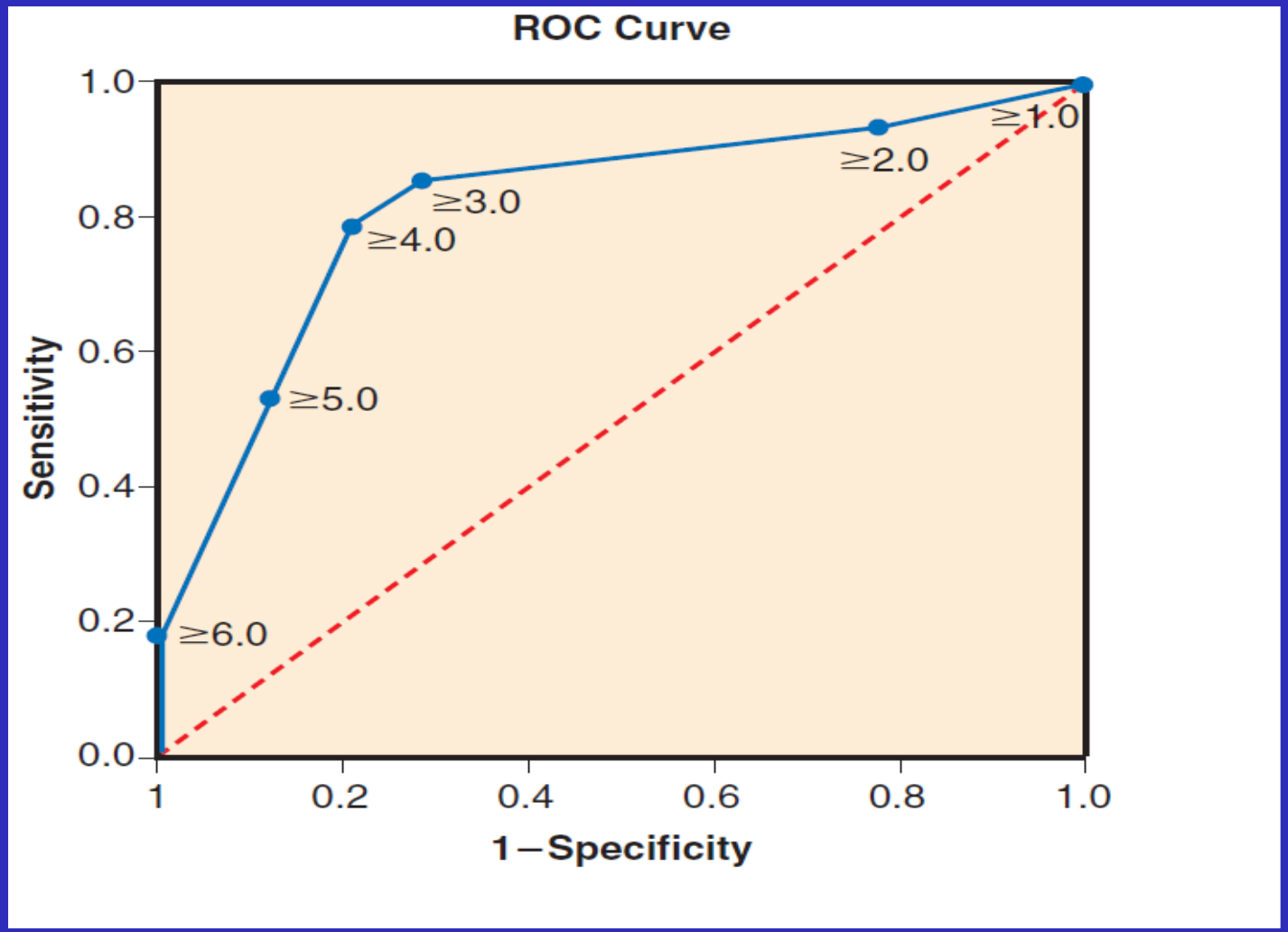
A. Data

Height Loss	Known Group 1	
	Fracture *n* = 62	No Fracture *n* = 89
1.0–1.9 cm	4	19
2.0–2.9 cm	5	44
3.0–3.9 cm	4	7
4.0–4.9 cm	17	8
5.0–5.9 cm	21	11
≥6.0 cm	11	0

B. Cutoff Points

Cutoff Point	True Positives	True Negatives	Coordinates of the ROC Curve 2			Youden Index
			Sensitivity	Specificity	1–Specificity	
≥1.0 cm	62	0	1.00	0.00	1.00	0.0
≥2.0 cm	58	19	.94	.21	.79	.15
≥3.0 cm	53	63	.86	.71	.29	.57
≥4.0 cm	49	70	.79	.79	.21	.58
≥5.0 cm	32	78	.52	.88	.12	.40
≥6.0 cm	11	89	.18	1.00	0.00	.18

An Individual has a positive test if height loss is equal to or greater than the cutoff point.



Area Under the Curve

Test Result Variables(s): HL

3 Area	Std. Error	Asymptotic Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
.815	.037	.000	.743	.888

The test result variable(s): HL has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

Validity

- Did selection lead to an appropriate **spectrum** of participants (like those assessed in practice)
- What was the **gold standard** of diagnosis?
- Did the results of the test being evaluated **influence** the decision to perform the reference standard?
- Was there a **blind** comparison between the test and an independent Gold (reference) Standard?
- Were the methods for performing the test described in **sufficient detail** to permit replication?

Spectrum

- Participants with the range of **common presentations** of the target disorder and with commonly **confused diagnosis**
- Florid cases & asymptomatic volunteers only
- **Spectrum bias**
- **Sensitivity up, specificity up**

Spectrum Bias

- Spectrum of disease and non disease differs from **clinical practice**.
- Sensitivity depends on **spectrum of disease.**
- Specificity depends on spectrum of non disease or of diseases that might **mimic** the disease of interest.

Spectrum Bias

When disease is skewed toward higher severity than in clinical practice , “sickest of the sick.”

Sensitivity 

When nondisease is skewed toward greater health, “wellest of the well.”

Specificity 

Spectrum Bias

- Diagnostic accuracy of low MCV for iron deficiency anemia in the US vs Africa ?

- Occult blood for diagnosis of colon cancer in two population with different prevalence of peptic ulcer?

- ✓**Severity of disease**
- ✓**Alternative diagnosis**

	Sensitivity	**specificity**
Case control	++++	+++++
Hospital practice	+++	++
General practice	++	++

Practical point

- When you read a paper that tries to measure sensitivity and specificity, think about whether the spectrums of disease and nondisease in the study subjects are **similar** to those in patients you are likely to see.

- *As a general rule, the more severe the disease in the patients who have it, the **greater** the sensitivity, whereas the healthier the "nondiseased" group, the **greater** the specificity*

Gold standard

- A different test that is known to give an accurate answer (but that may be more expensive or more invasive than the new test)
- A composite of several tests
- The result of another medical procedure (such as surgery)
- The outcome of a period of follow-up (indicating whether the person develops the condition in question).

Gold standard

- 1.Laboratory tests
  (Infectious & endocrine diseases)
- 2.Imaging
  (DVT, PTE)
- 3.Biopsy
  (Cancer , vasculitis)
- 4.Autopsy
  (neurologic diseases)
- 5.long-term follow-up
  (SLE , MS)

Independency

- The properties of a diagnostic test will be distorted if its result influences whether patients undergo confirmation by the reference standard.

- People who are positive on the index test are more likely to get the gold standard, and only those who receive the gold standard are included in the study(**Verification bias)**

- **Sensitivity up, specificity down**

Ankle swelling and x-ray

Box 5.2: Numerical example of verification bias

We examine two hypothetical studies of ankle swelling as a predictor of fractures in patients with ankle injuries. The first study is a consecutive sample of 200 patients. In this study, all patients presenting to the emergency department with ankle injuries get x-rays, regardless of swelling. The sensitivity and specificity of ankle swelling are 80% and 75%, as shown in the following table:

	Fracture	No Fracture
Swelling	32	40
No Swelling	8	120
Total	**40**	**160**
	Sensitivity = 32/40 = 80%	Specificity = 120/160 = 75%

The second study is a **selected** sample, in which only half the patients without ankle swelling are x-rayed. Thus, the numbers in the "No Swelling" row will be reduced by half. This raises the apparent sensitivity from 32/40 (80%) to 32/36 (89%) and lowers the apparent specificity from 120/160 (75%) to 60/100 (60%), as shown in the next table:

	Fracture	No Fracture
Swelling	32	40
No Swelling	4	60
Total	**36**	**100**
	Sensitivity = 32/36 = 89%	Specificity = 60/100 = 60%

Blinding

- Those applying and interpreting the reference standard should ideally be unaware of the result of the test .

- The pulmonary nodule on CT , and comparison to CXR

Sufficient details

- This description should cover all issues that are important in the preparation of the patient (diet, drugs to be avoided, precautions after the test), the performance of the test (technique, possibility of pain), and the analysis and interpretation of its results

Levels of evidence for studies of diagnostic methods

Level	Criteria
1	An independent, masked comparison with reference standard among an appropriate population of consecutive patients.
2	An independent, masked comparison with reference standard among non-consecutive patients or confined to a narrow population of study patients.
3	An independent, masked comparison with an appropriate population of patients, but reference standard not applied to all study patients
4	Reference standard not applied independently or masked
5	Expert opinion with no explicit critical appraisal, based on physiology, bench research, or first principles.

Seven standards

Standard 1: Spectrum composition

Standard 2: Pertinent subgroups

Standard 3: Avoidance of workup bias

Standard 4: Avoidance of review bias

Standard 5: Precision of results for test accuracy

Standard 6: Presentation of indeterminate test results

Standard 7: Test reproducibility

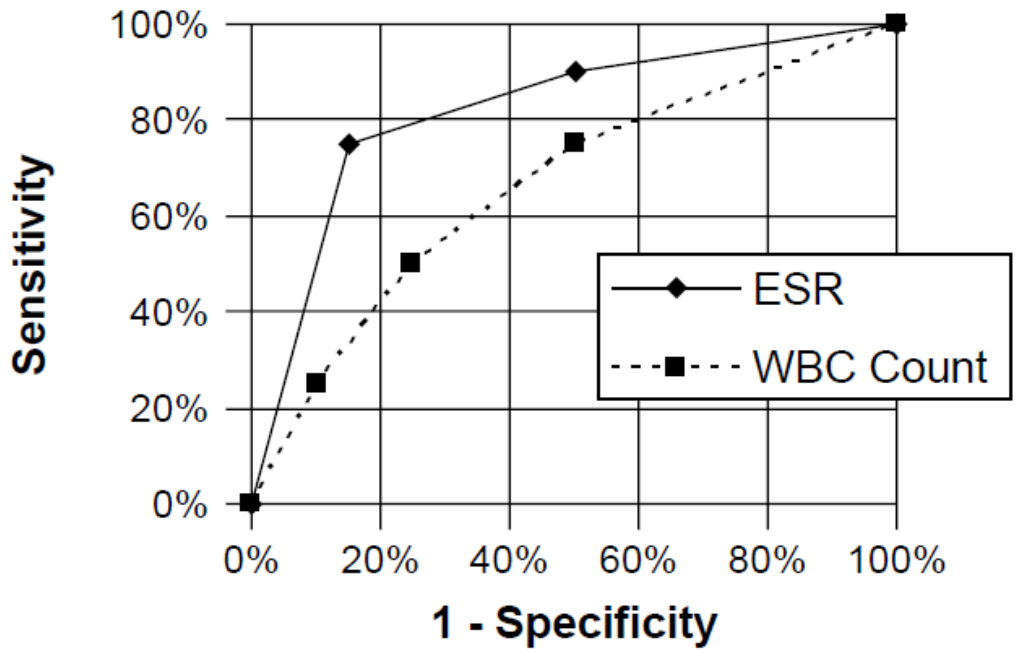
Exercise

- You are interested in studying the erythrocyte sedimentation rate (ESR) as a test for pelvic inflammatory disease (PID) in women with abdominal pain.
  - To do this, you will need to assemble groups of women who do and do not have PID. What would be the best way to sample these women?
  - How might the results be biased if you used final diagnosis of PID as the gold standard and those assigning that diagnosis were aware of the ESR?

Exercise

- You find that the sensitivity of an ESR of at least 20 mm/hr is 90%, but the specificity is only 50%. On the other hand, the sensitivity of an ESR of at least 50 mm/hr is only 75%, but the specificity is 85%. How should you present these results?

ESR	PID	No PID	Likelihood Ratio
≥ 50	75%	15%	5.00
20–49	15%	35%	0.43
< 20	10%	50%	0.20
	100%	100%	



Exercise

- The investigator undertakes a case control study to address the research question, Does eating more fruits and vegetables reduce the risk of coronary heart disease (CHD) in the elderly? Suppose that her study shows that people in the control group report a higher intake of fruits and vegetables than people with CHD.
- What are the possible explanations for this inverse association between intake of fruits and vegetables and CHD? How could each of these possibilities be altered in the design phase of the study? How could they be addressed in the analysis phase?

Exercise ….

- Give special attention to the possibility that the association between eating fruits and vegetables and CHD may be confounded by exercise (if people who eat more fruits and vegetables also exercise more, and this is the cause of their lower CHD rates). What approaches could you use to cope with exercise as a possible confounder, and what are the advantages and disadvantages of each plan?