
بسم الله الرحمن الرحيم
إنه من سليمان وإنه

Screening (Diagnostic Tests)

Shaker Salarilak

Outline

- Screening basics
- Evaluation of screening programs

Where we are?

- Definition of screening?
- Whether it is always beneficial?
- Types of bias in screening?
- Principles for the development of screening.
  - The test: Validity, LR, ROC curve, Kappa
  - The disease;
- Evaluation of a screening program

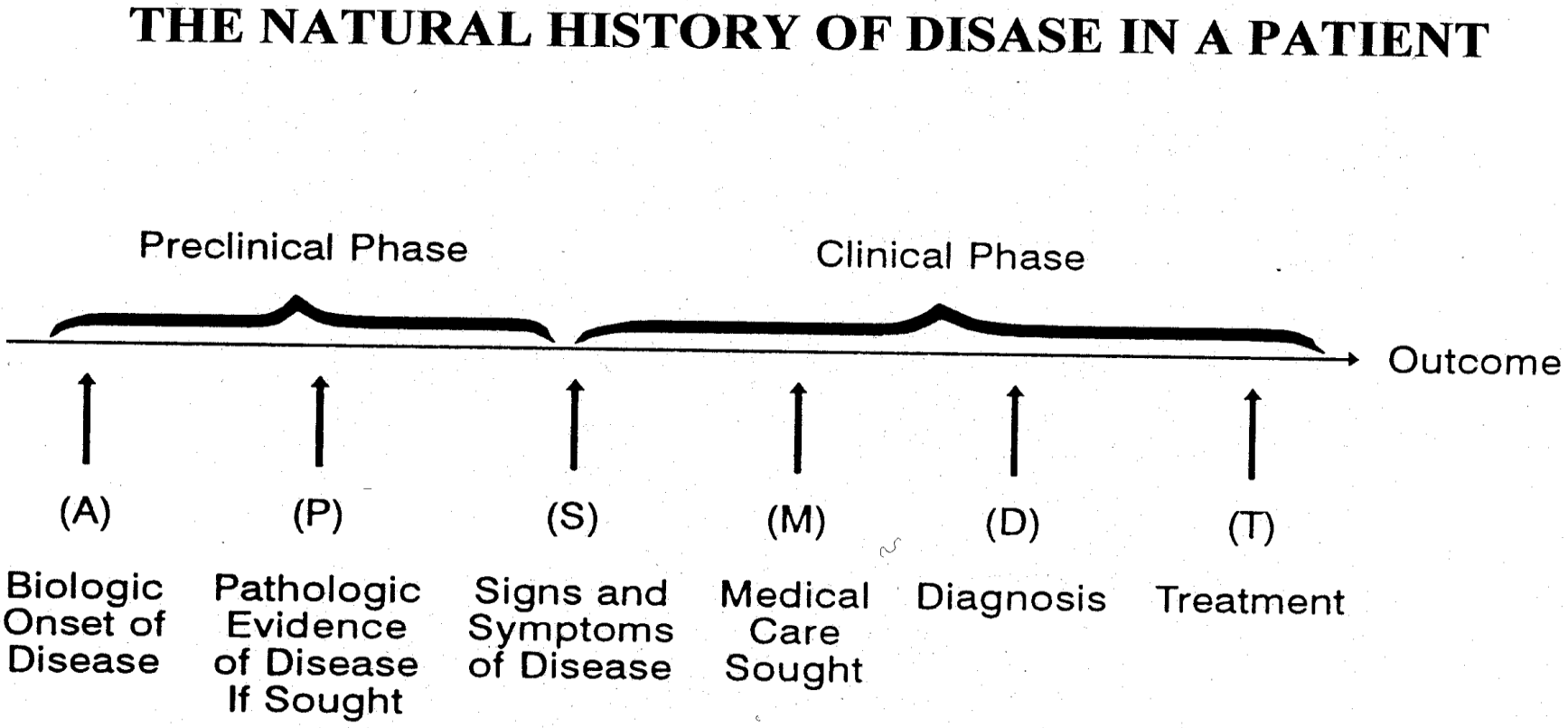
Screening Basics

- What does "screening" mean?
- What do we screen for (objective)?
- What makes a disease an appropriate target for screening?
- What makes a test a good screening test?

Levels of Prevention (Mausner and Kramer 1985)

- **Primary Prevention** - Prevention of the occurrence of disease (reduce incidence of disease)
- ***Secondary Prevention*** - Early detection and prompt treatment of disease for cure, to slow progression, to prevent complications, or to limit disability (reduce prevalence of disease)
- **Tertiary Prevention** - Limitation of disability and rehabilitation where disease has already occurred and left residual damage

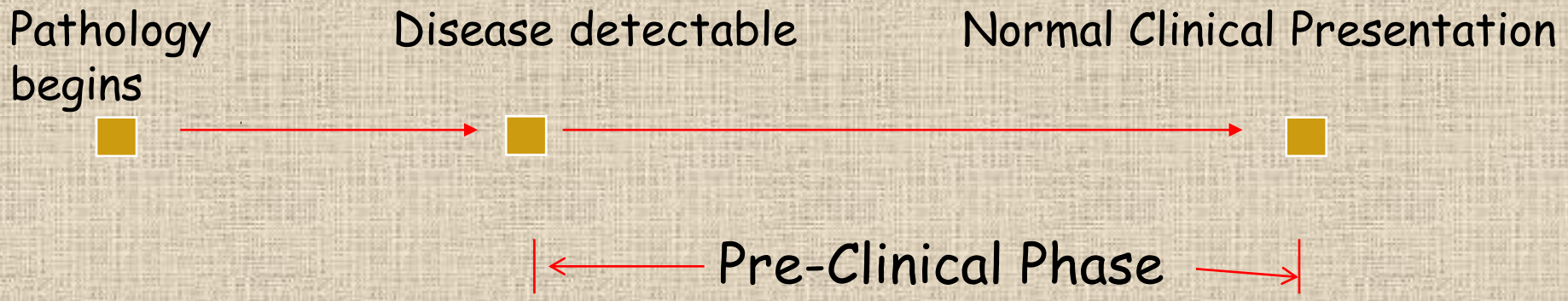
Natural History of Disease

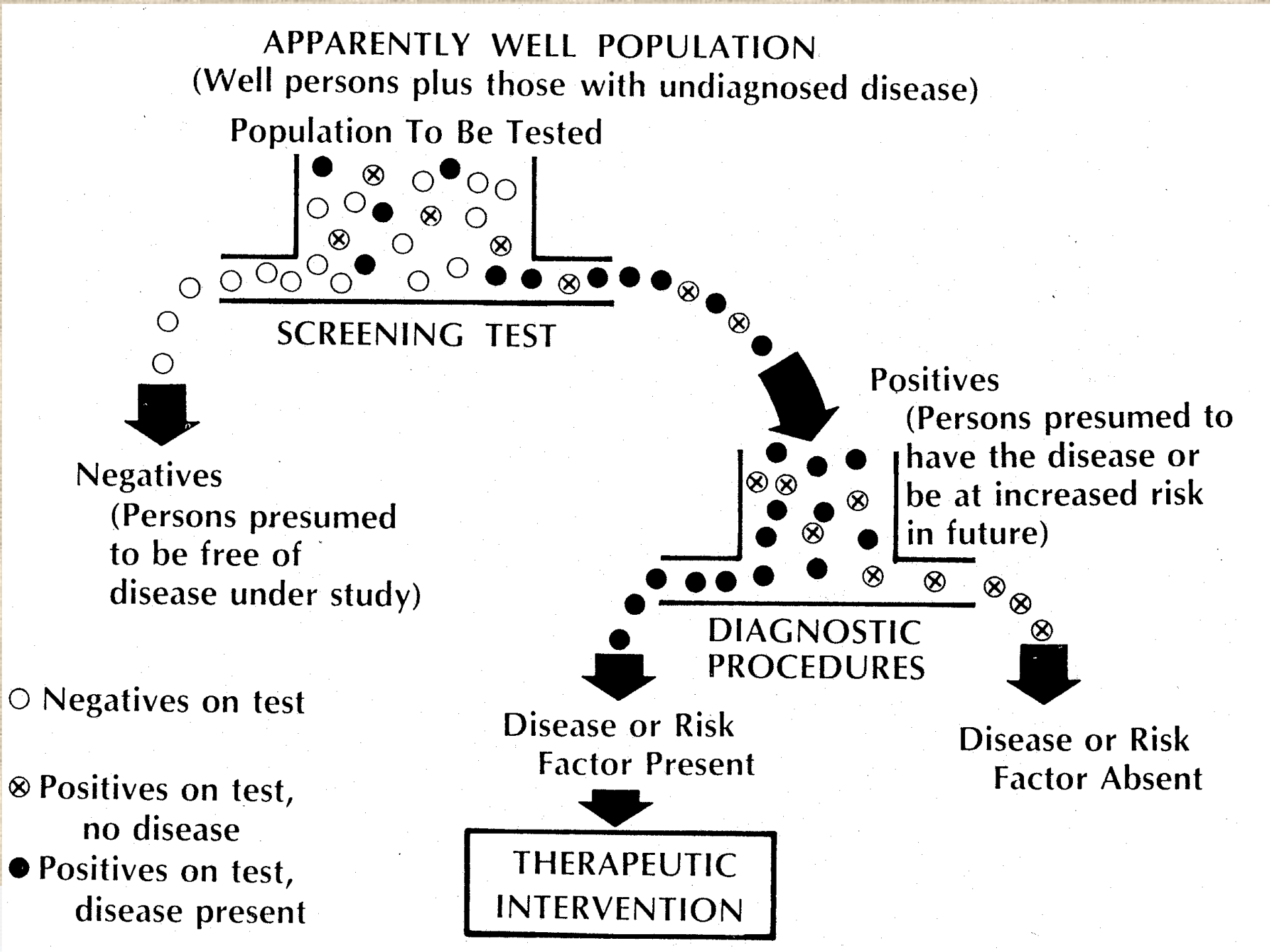


Pre-clinical Phase

The Pre-Clinical Phase (PCP) is

- the period between when early detection by screening is possible and when the clinical diagnosis would usually be made.




APPARENTLY WELL POPULATION
(Well persons plus those with undiagnosed disease)
Population To Be Tested
SCREENING TEST
Negatives
(Persons presumed to be free of disease under study)
Positives
(Persons presumed to have the disease or be at increased risk in future)
DIAGNOSTIC PROCEDURES
Disease or Risk Factor Present
Disease or Risk Factor Absent
THERAPEUTIC INTERVENTION
○ Negatives on test
⊗ Positives on test, no disease
● Positives on test, disease present

Principles for the development of screening;

1. The condition screened for is an important cause of morbidity, disability, or mortality.
2. The natural history of the disease is sufficiently well known.
3. The test must have high levels performance.
4. The test must be acceptable to the target population and their health care providers, and appropriate follow-up of positive findings must be ensured.

Consequence of a screening test:

- Beneficence
- Non-beneficence
  - Do harm;

    Clofibrate in US

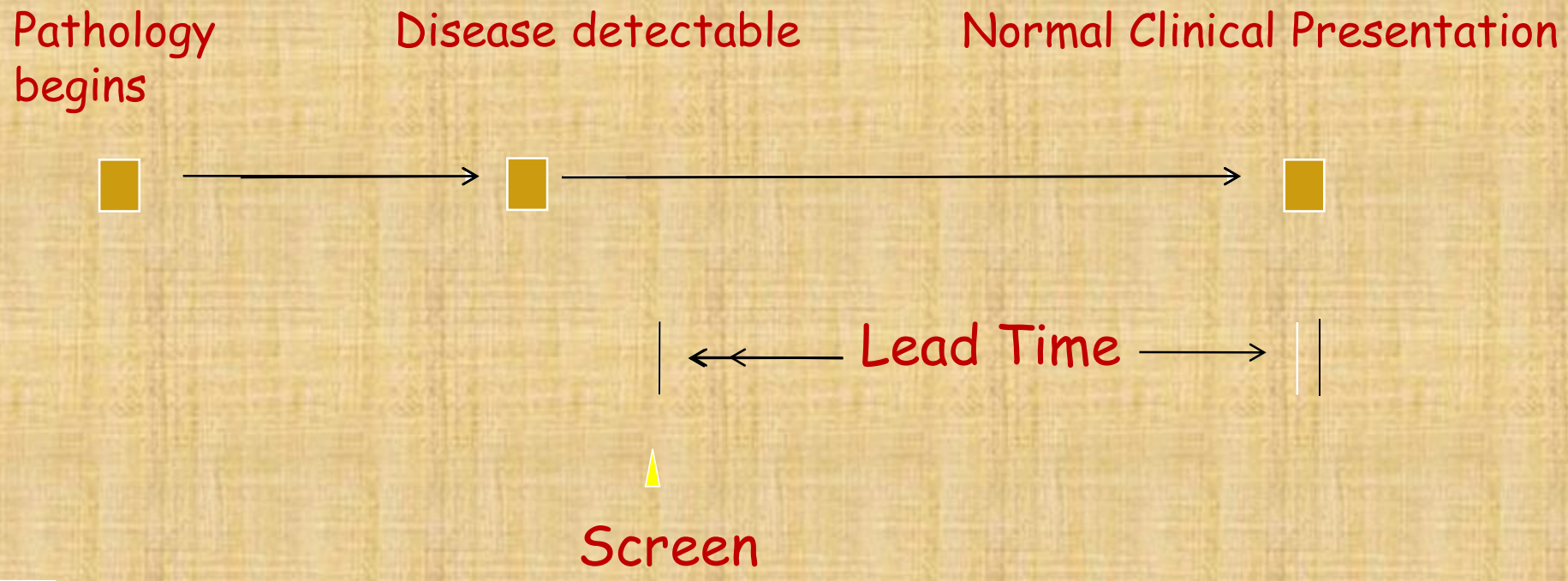
    Labeling effect; Social psychology

Biases in assessing efficacy of screening

- Two major biases affect these data:
  - lead time bias
  - length bias

Lead Time

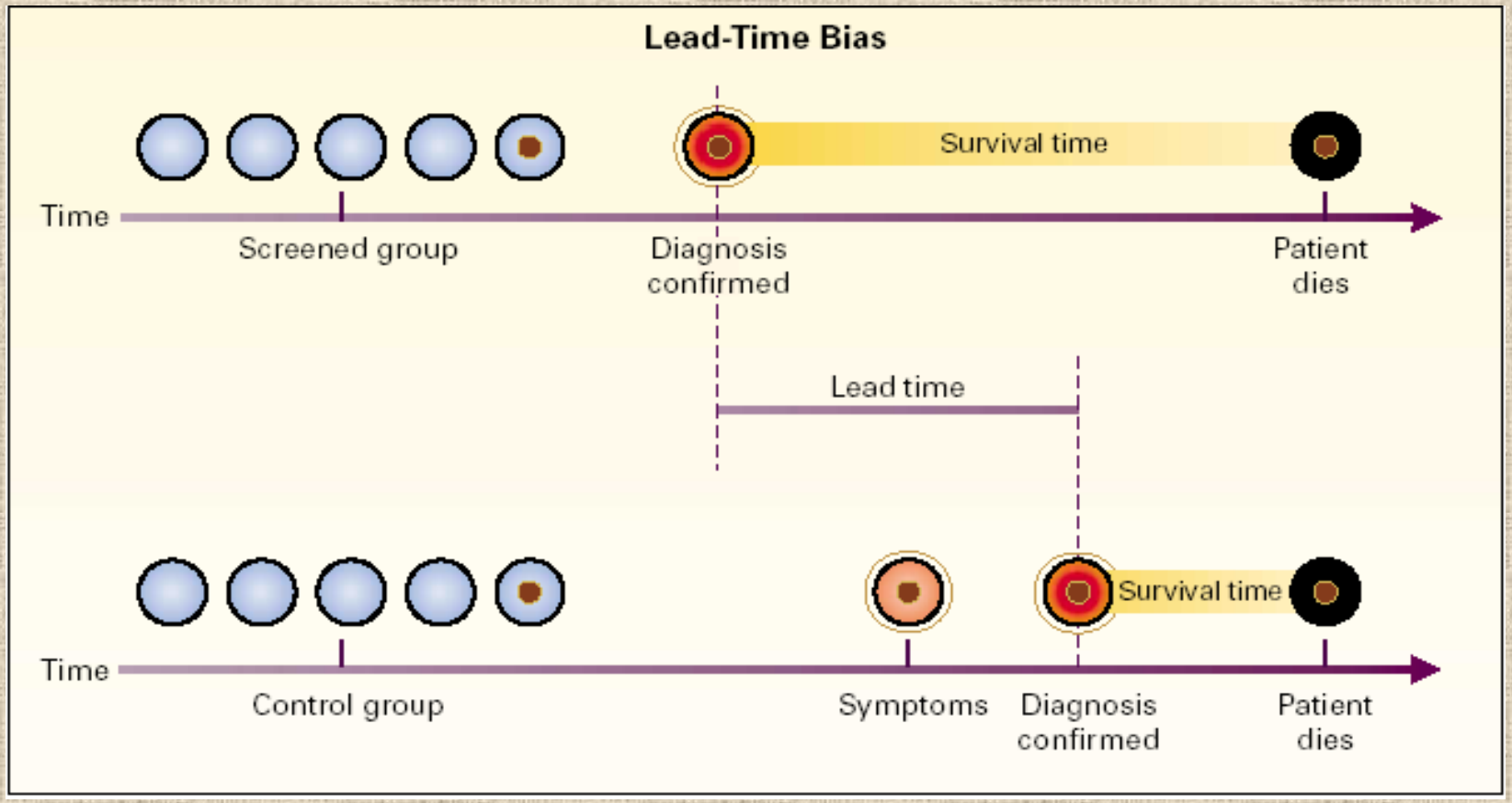
Lead time = amount of time by which diagnosis is advanced or made earlier



Lead time bias

- We *think* early detection has increased survival
  - in fact all it has done is increase the time the patient is aware of his disease!
  - treatment could even *hasten* death and it might appear survival is longer post diagnosis!!
- Cannot just look at survival time post diagnosis.

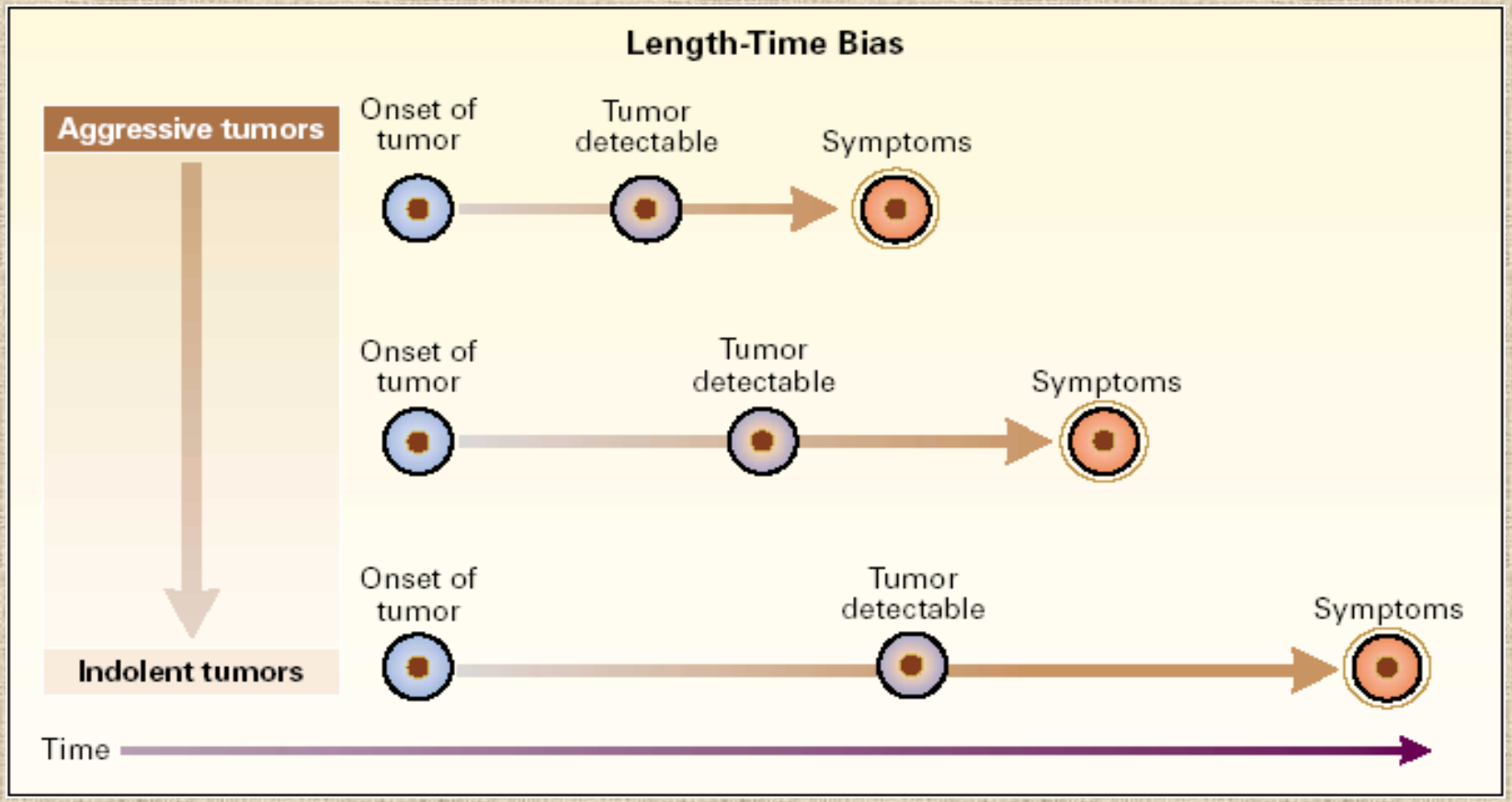
Lead-time Bias



Length bias

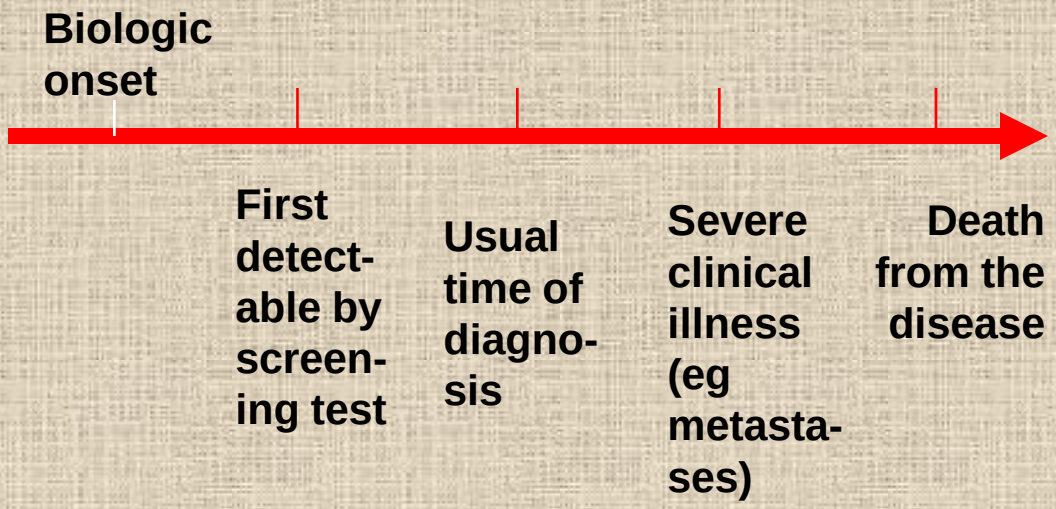
- **Survival due to screening and treatment may be over rated because screening will tend to discover more slow-growing disease.**

Length-time Bias



Suppose there are two subtypes of the disease:

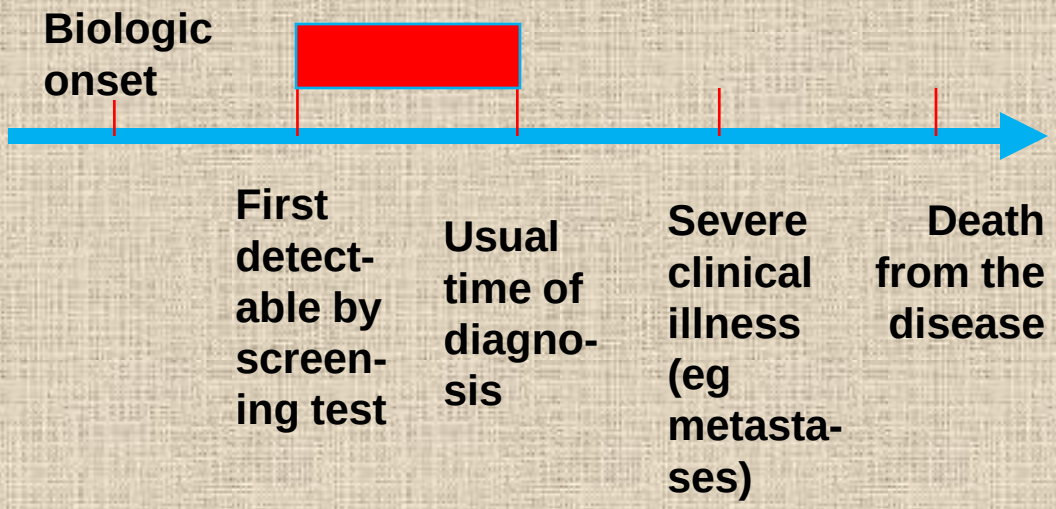
Type 1: fast progression



Type 2: slow progression

Length of time in pre-clinical phase longer in Type 2 than in Type 1

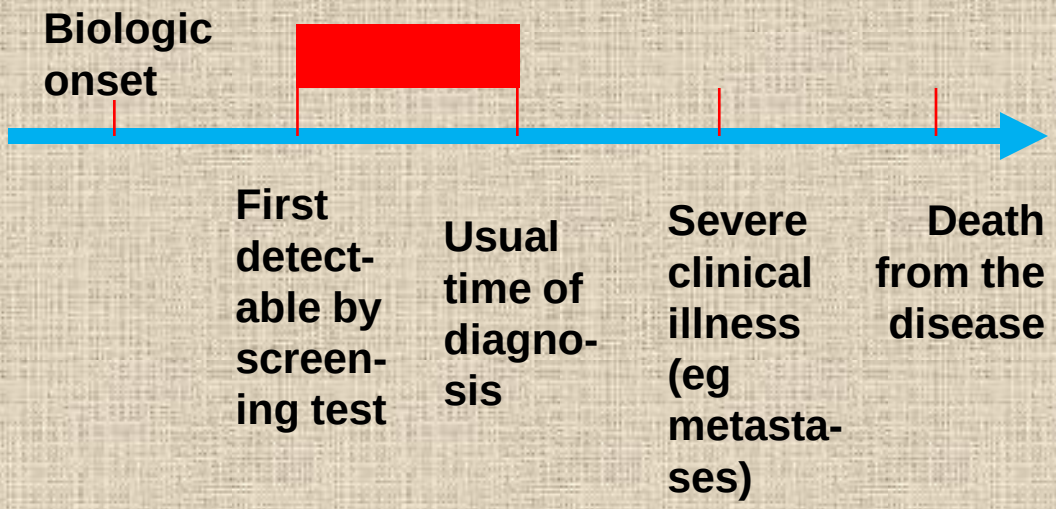
Type 1



Type 2

Periodic screening will tend to detect more of Type 2, as these have longer “exposure” in the critical interval for screening.

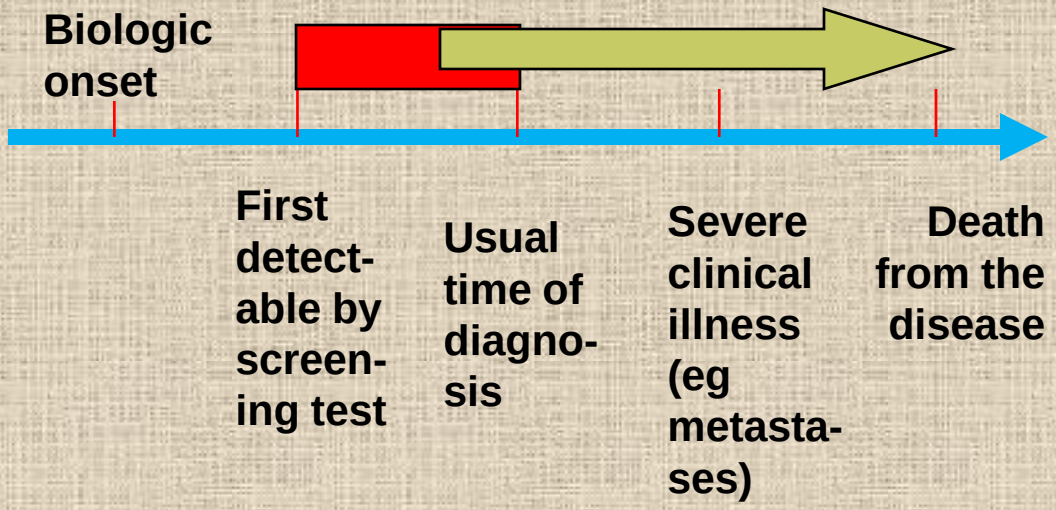
Type 1



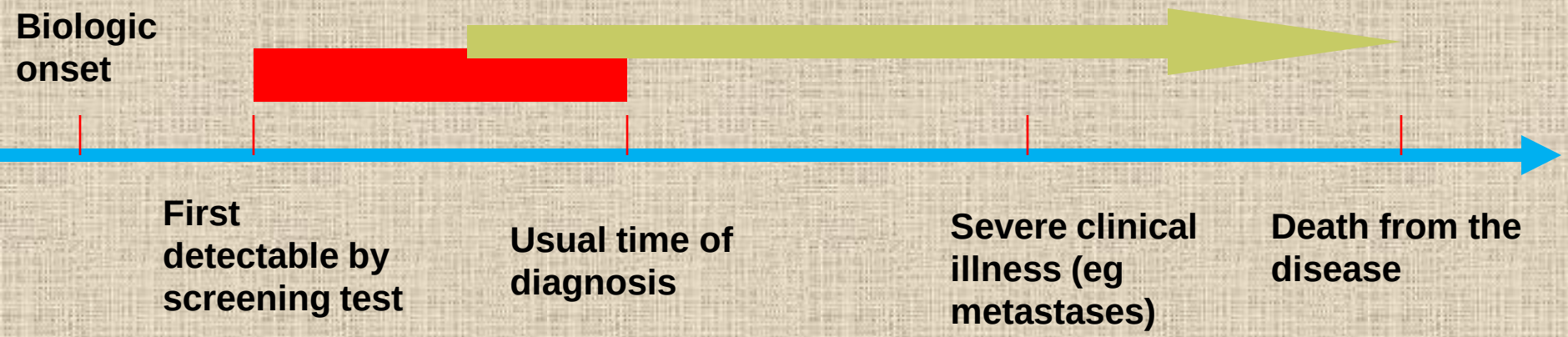
Type 2

But look!! Type 2 individuals have a longer survival time from time of diagnosis than do Type 1.

Type 1



Type 2



Length bias

- Without screening, suppose type 1 and type 2 were equal fractions of the population
  - average survival time is 50:50 mixture of the short and long survival times.

- *With* screening, the screen-detected population has a higher fraction of type 2 (slow) individuals
  - mix will be proportional to ratio of the two intervals
  - suppose it is 70:30 in favor of long interval
    - average survival time will be longer in screen detected individuals!

Length bias

- Even if the treatment tended to be harmful and shorten life, because more longer interval individuals tend to be detected by screening, the screening program will *appear* to be effective!!

Principles for the development of screening

1. The test must have high levels performance.
2. The condition screened for is an important cause of morbidity, disability, or mortality.
3. The natural history of the disease is sufficiently well known.
4. The test must be acceptable to the target population and their health care providers, and appropriate follow-up of positive findings must be ensured.

Characteristics of Test

- Safety
- Cost
- Acceptability
- Validity
- Reliability

Diagnostic tests

- When looking at a paper about a diagnostic test we ask ourselves three questions.

Diagnostic tests

- Is this test useful?
- Is it reliable?
- Is it valid?

Is this test useful?

- The test should have been researched in a study population **relevant** to the individual or population in whom it is to be used.

Reliability

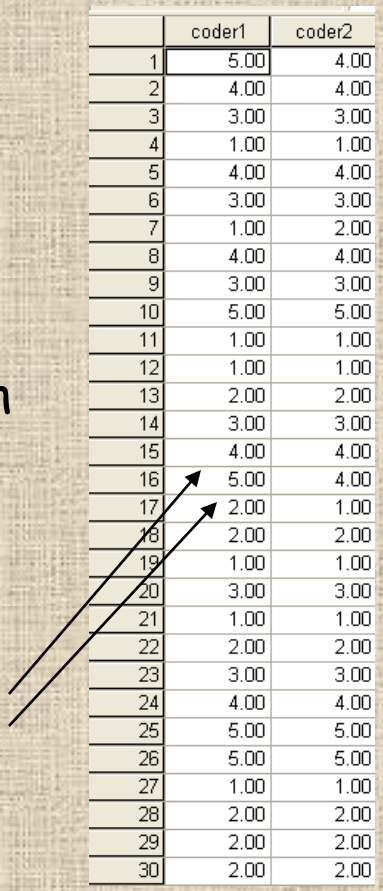
- Reliability refers to the **repeatability** or reproducibility of a test.
- It can be assessed by repeating the test using the same or different observers.

Calculating Inter-coder Reliability

- Suppose you had thirty message segments or photos and you wanted to apply to them a coding scheme which had five categories
- You had each of two coders examine each the thirty message segments and assign it to one of the five categories
- You want to know how reliable this coding scheme is in practice. Another way to say this is, "what is the inter-coder reliability?"

Here's What your Data Look Like

- You enter your data into SPSS as shown on the right, where each of the thirty lines represents one of your messages or message units that was analyzed, and the two columns contain the categories which coder 1 and then coder 2 assigned that message to. If both assigned the message to the same category, then that indicates inter-coder agreement, and that's good. Note that in the data there are a few messages on which the coders did not agree as what category it should be placed in



	coder1	coder2
1	5.00	4.00
2	4.00	4.00
3	3.00	3.00
4	1.00	1.00
5	4.00	4.00
6	3.00	3.00
7	1.00	2.00
8	4.00	4.00
9	3.00	3.00
10	5.00	5.00
11	1.00	1.00
12	1.00	1.00
13	2.00	2.00
14	3.00	3.00
15	4.00	4.00
16	5.00	4.00
17	2.00	1.00
18	2.00	2.00
19	1.00	1.00
20	3.00	3.00
21	1.00	1.00
22	2.00	2.00
23	3.00	3.00
24	4.00	4.00
25	5.00	5.00
26	5.00	5.00
27	1.00	1.00
28	2.00	2.00
29	2.00	2.00
30	2.00	2.00

The numbers stand for the message's being assigned to one of the five categories in your coding scheme (nominal-level data)

Kappa

$$\kappa = \frac{N_o - N_e}{1 - N_e}$$

N_o = Observed number of agreement

Ne= Number of agreement expected to occur by chance alone

Varies from -1 to 1

Population One (Prevalence = 0.05)
Table for true positives

		Observer A		
		Positive	**Negative**	
Observer B	**Positive**	**36**	**9**	***45***
	Negative	**4**	**1**	***5***
		40	***10***	***50***

From Szklo and Nieto, 2000

Interpretation of Kappa

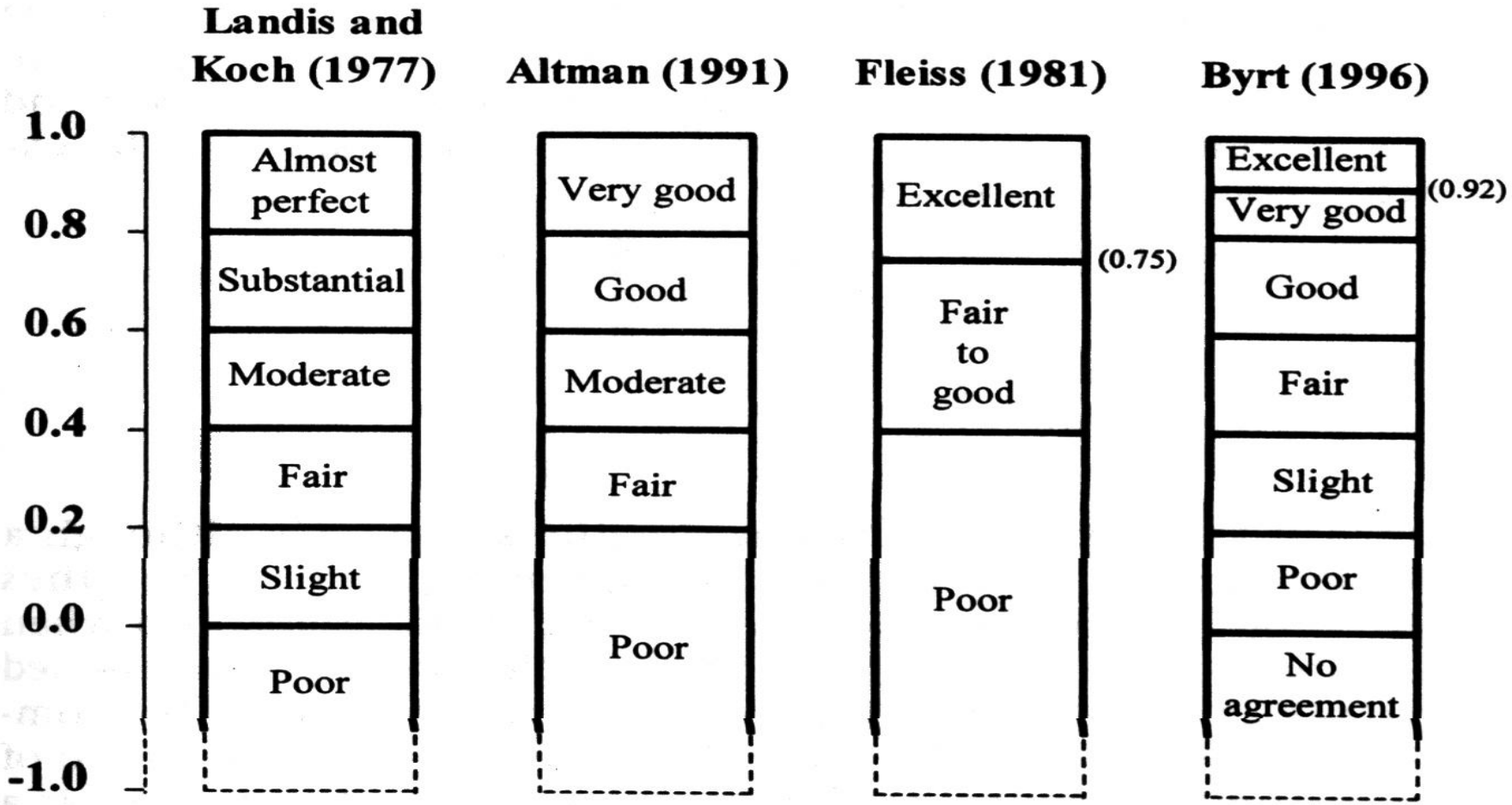


Figure 8–6 Proposed classifications for the interpretation of a kappa value

How to Compute Kappa, the Inter-coder Reliability

- In SPSS Data Editor, go to Analyze/ Descriptive/Crosstabs
- Move the Coder1 variable into the Column box and the Coder2 Variable into the row box (or vice versa, doesn't matter)
- Click on Statistics, select Kappa, then Continue and then OK
- You will obtain output as shown on the next slide

SPSS Output for Kappa, the Inter-coder Reliability Coefficient

CODER2 * CODER1 Crosstabulation

Count

		CODER1					Total
		1.00	2.00	3.00	4.00	5.00	
CODER2	1.00	6	1	0	0	0	7
	2.00	1	6	0	0	0	7
	3.00	0	0	6	0	0	6
	4.00	0	0	0	5	2	7
	5.00	0	0	0	0	3	3
Total		7	7	6	5	5	30

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	.832	.077	9.084	.000
N of Valid Cases		30			

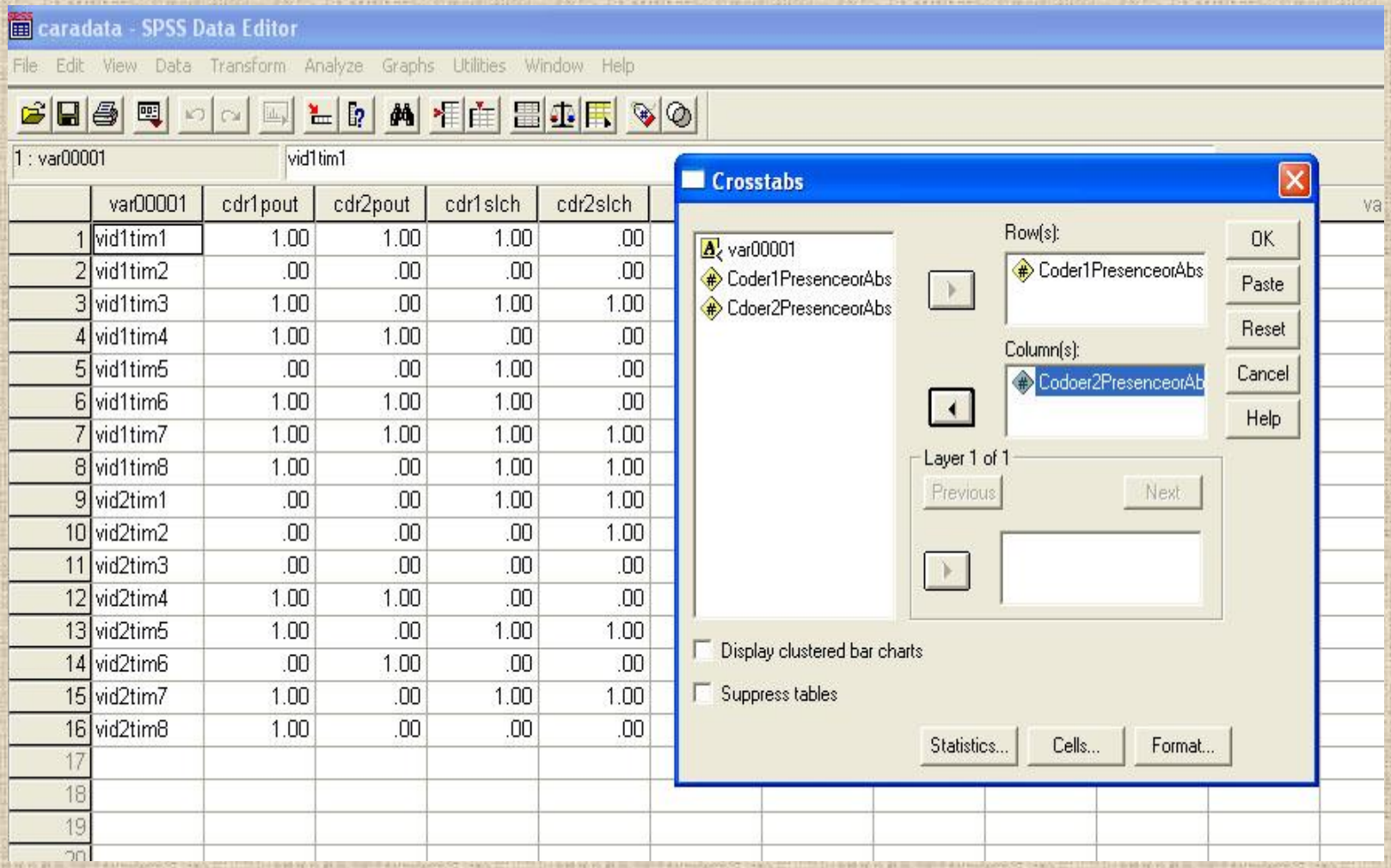
a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

The off-diagonal elements show you where the raters disagreed. See the colored dots, which shows they had problems between categories 4 and 5 and categories 1 and 2. You could work more on distinguishing those and recode some of the items on which they disagreed after a little retraining

Here is your value of kappa: .832

Another Example Assessing Intercoder Reliability for Two Variables



Output of SPSS Calculation of Kappa

Coder1PresenceorAbsenceofPout * Codoer2PresenceorAbsenceofPout Crosstabulation

Count

		Codoer2PresenceorAbsenceofPout		
		absent	present	Total
Coder1PresenceorAbsenceofPout	absent	5	1	6
	present	5	5	10
Total		10	6	16

Coder disagreements

A low obtained value of kappa

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	.294	.205	1.333	.182
N of Valid Cases		16			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Validity

- Relates to whether the test measures what it purports to measure. Is the result true?

- It can be assessed by comparing the test results with a Gold Standard.

Validity

- For example if you measure blood pressure in an obese patient and use a cuff that is too small you are likely to get a falsely high reading. The reading maybe reliable (you get the same blood pressure if you do it again) but it lacks validity.

Gold standard

- The gold standard is the test or battery of tests that will most accurately diagnose a particular disease or condition.
  - The OGTT for diabetes
  - Fluoroscein angiography for diabetic retinopathy (too expensive or invasive)
  - The Jones criteria for rheumatic fever (a battery of tests or symptoms)

What is your variable?

Table 8–3 Summary of Indices or Graphic Approaches Most Frequently Used for the Assessment of Validity and Reliability

		Mostly Used to Assess . . .	
Type of Variable	*Index or Technique*	*Validity*	*Reliability*
Categorical	Sensitivity/specificity	++	
	Percent agreement	+	++
	Percent positive agreement	+	++
	Kappa statistic	+	++
Continuous	Scatter plot (correlation graph)	+	++
	Linear correlation coefficient (Pearson)	+	+
	Ordinal correlation coefficient (Spearman)	+	+
	Intraclass correlation coefficient	+	++
	Coefficient of variation		++
	Bland-Altman plot	++	++

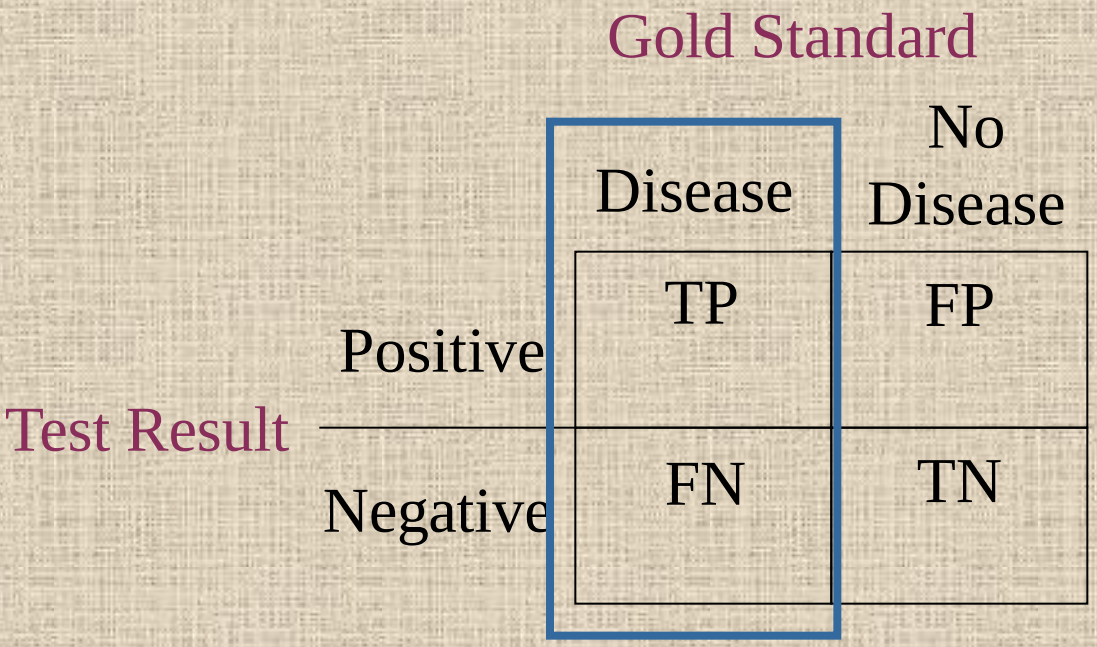
Note: ++, the index is indicated and used to measure the magnitude of validity or reliability; +, although the index is used to measure the magnitude of either validity or reliability, its indication is somewhat questionable.

Sensitivity and specificity

Ability of a test to accurately diagnose diseased and healthy individuals

- Sensitivity
- Specificity
- Likelihood Ratio
- …

Sensitivity



Sensitivity: The capacity of the test to correctly identify **diseased** individuals in a population; "TRUE POSITIVES".

Specificity

Test Result	Gold Standard: Disease	Gold Standard: No Disease
Positive	TP	FP
Negative	FN	TN

Specificity: The capacity of the test to correctly exclude individuals who are free of the disease; "TRUE NEGATIVES".

Sensitivity and Specificity

Test Result	Gold Standard: Disease	Gold Standard: No Disease
Positive	TP	FP
Negative	FN	TN
	Sensitivity TP/TP+FN	Specificity TN/FP+TN

Example

Test Result	Gold Standard: Disease	Gold Standard: No Disease	
Positive	75	20	95
Negative	25	180	205
	100	200	300

Sensitivity = 75/100 = 75%

Specificity = 180/200 = 90%

Accuracy of the test

Test Result	Gold Standard: Disease	Gold Standard: No Disease	
Positive	a	b	a+b
Negative	c	d	c+d
	a+c	b+d	300

(a+d)/(a+b+c+d)

Positive Predictive Value

Test Result	Gold Standard: Disease	Gold Standard: No Disease
Positive	TP	FP
Negative	FN	TN

PPV = TP/TP+FP

PPV: The probability of the disease being present, among those with positive diagnostic test results

Negative Predictive Value

		Gold Standard: Disease	Gold Standard: No Disease
Test Result	Positive	TP	FP
	Negative	FN	TN

NPV = TN/TN+FN

NPV: The probability that the disease was absent, among those whose diagnostic test results were negative

The effect of Sense, Spec, and P on PPV and NPV

		PPV			NPV		
		Prevalence					
Sensitivity	Specificity	90%	50%	10%	90%	50%	10%
70%	60%	94%	64%	16%	18%	67%	95%
70%	90%	98.4%	88%	44%	25%	75%	96%
80%	90%	98.6%	89%	47%	33%	82%	98%
90%	90%	98.7%	90%	50%	50%	90%	99%
100%	5%	2%	51%	10%	100%	100%	100%
5%	100%	100%	100%	100%	98%	51%	90%

There are some predictors other than the prevalence:

What do we do in clinic?

Likelihood ratio

$$\text{LR Positive} = \frac{\text{Likelihood of (+) test in diseased persons}}{\text{Likelihood of (+) test in healthy persons}}$$

$$\text{LR Positive} = \frac{\text{Sensitivity}}{\text{1 - Specificity}}$$

$$\text{LR Negative} = \frac{\text{Likelihood of (-) test in diseased persons}}{\text{Likelihood of (-) test in healthy persons}}$$

$$\text{LR Negative} = \frac{\text{1 - Sensitivity}}{\text{Specificity}}$$

Likelihood ratio

Sensitivity = 90%

Specificity = 90%

$$\text{LR Positive} = \frac{\text{Sensitivity}}{1 - \text{Specificity}} = \frac{0.90}{1 - 0.90} = 9$$

$$\text{LR Negative} = \frac{1 - \text{Sensitivity}}{\text{Specificity}} = \frac{1 - 0.90}{0.90} = 1/9$$

Example

5000 pregnant women underwent a test for blood glucose at 24 weeks, following a glucose load. 243 women were found to have a blood glucose greater than 6.8 mmol/L and were referred for an OGTT. 186 were found to have gestational diabetes. Four women who initially had tested negative were diagnosed as having diabetes later in their pregnancy.

	Diabetes	No diabetes	Total
Positive	186	57	243
Negative	4	4753	4757
Total	190	4810	5000

Example

Prevalence	
Sensitivity	
Specificity	
Positive predictive value	
Negative predictive value	
Likelihood ratio + test	
Likelihood ratio - test	
Accuracy	

Example

Prevalence	190/5000
Sensitivity	186/190
Specificity	4753/4810
Positive predictive value	186/243
Negative predictive value	4753/4757
Likelihood ratio + test	(186/190)/(57/4810)
Likelihood ratio - test	(4/190)/(4753/4810)
Accuracy	(186+4753)/5000

Example

Prevalence	3.8%
Sensitivity	97.9%
Specificity	98.8%
Positive predictive value	76.5%
Negative predictive value	99.9%
Likelihood ratio + test	82.6
Likelihood ratio - test	.02
Accuracy	98.8%

Sequential (Two-stage) Tests

- In sequential or two-stage screening, a less expensive, less invasive, or less uncomfortable test is generally performed first, and those who screen positive are recalled for further testing with a more expensive, more invasive, or more uncomfortable test, which may have greater sensitivity and specificity.

Sequential (Two-stage) Tests

- In this method the net sensitivity decreased and the net specificity increased.

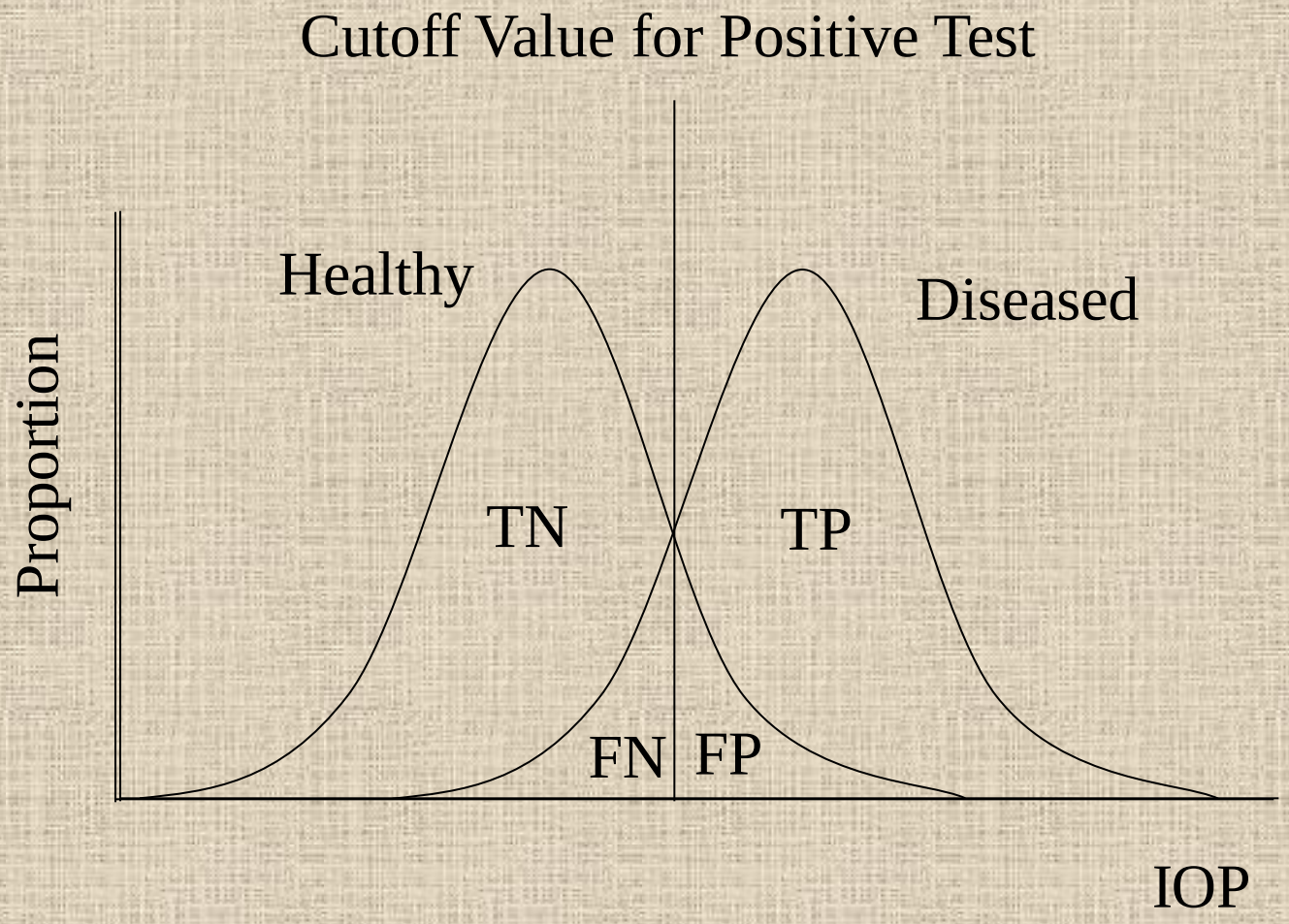
Simultaneous Tests

- In clinical setting, multiple tests are often used simultaneously. For example, patient admitted to a hospital may have an array of test performed at the time admission. When multiple test are used simultaneously to detect specific disease, the individual is generally considered to have tested "positive" if he or she has a positive result on any one or more of the tests. The individual is considered to have tested "negative" if he or she testes negative on all of the tests.

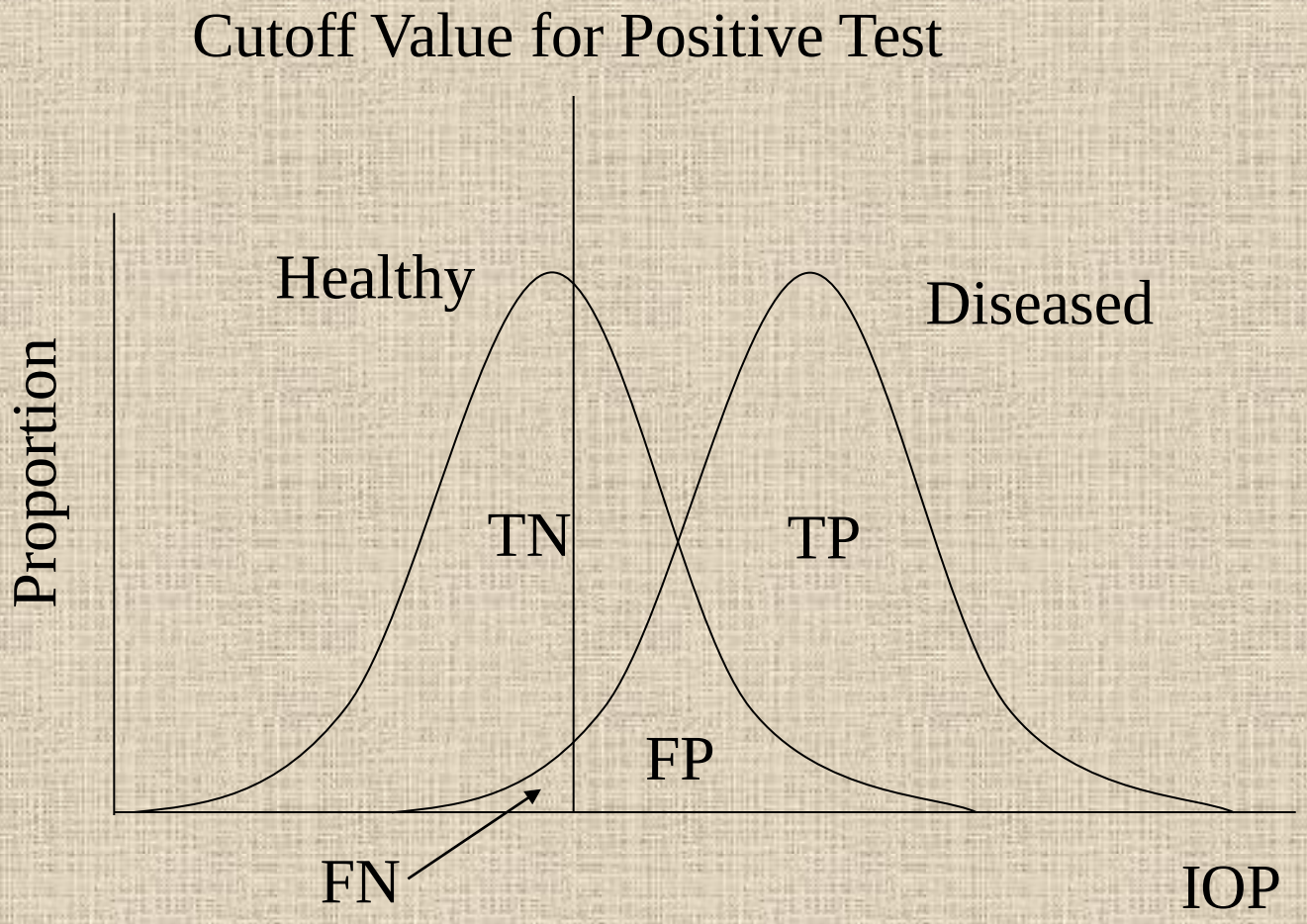
Simultaneous Tests

- In this method the net sensitivity increased and the net specificity decreased.

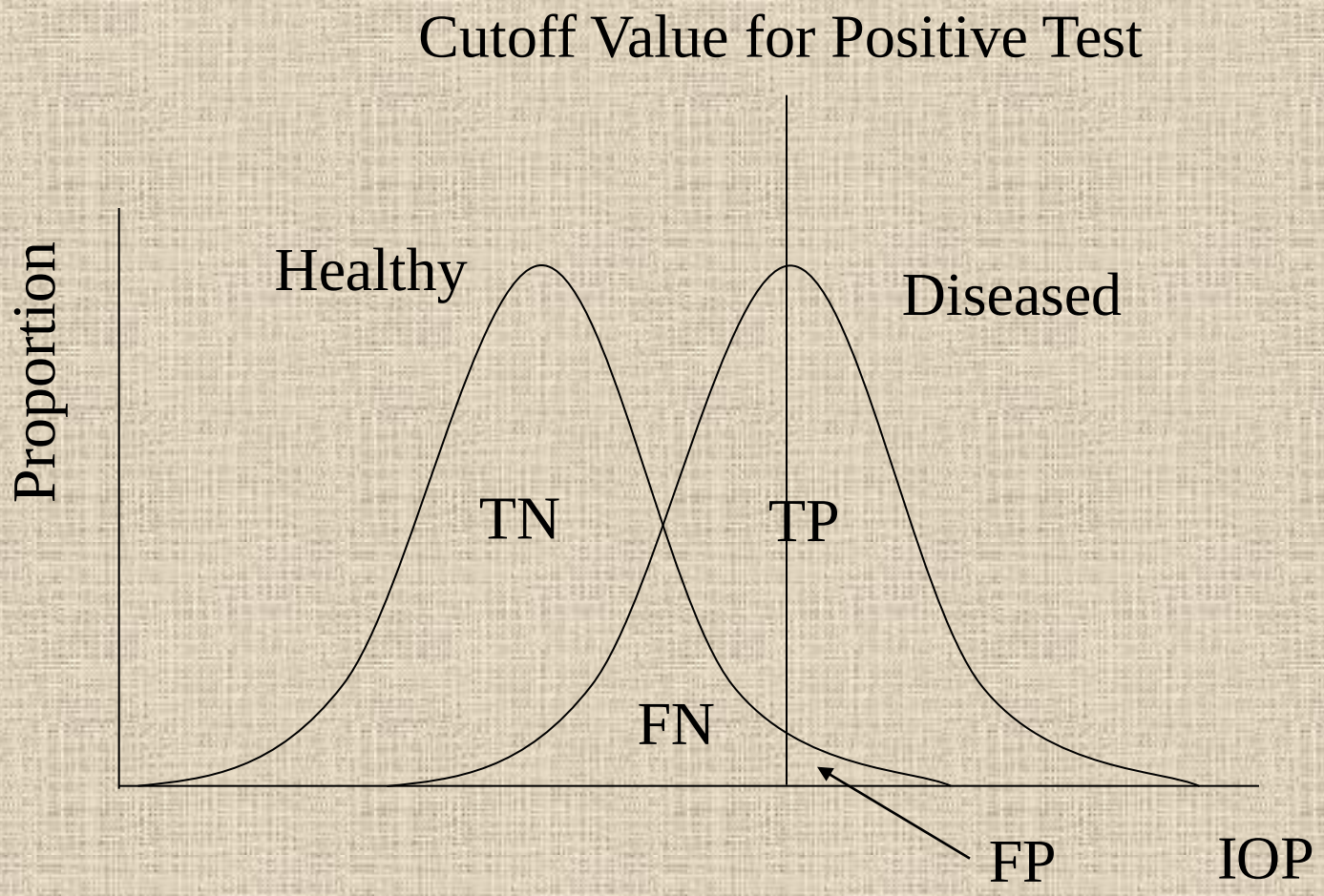
Continuous Measurements



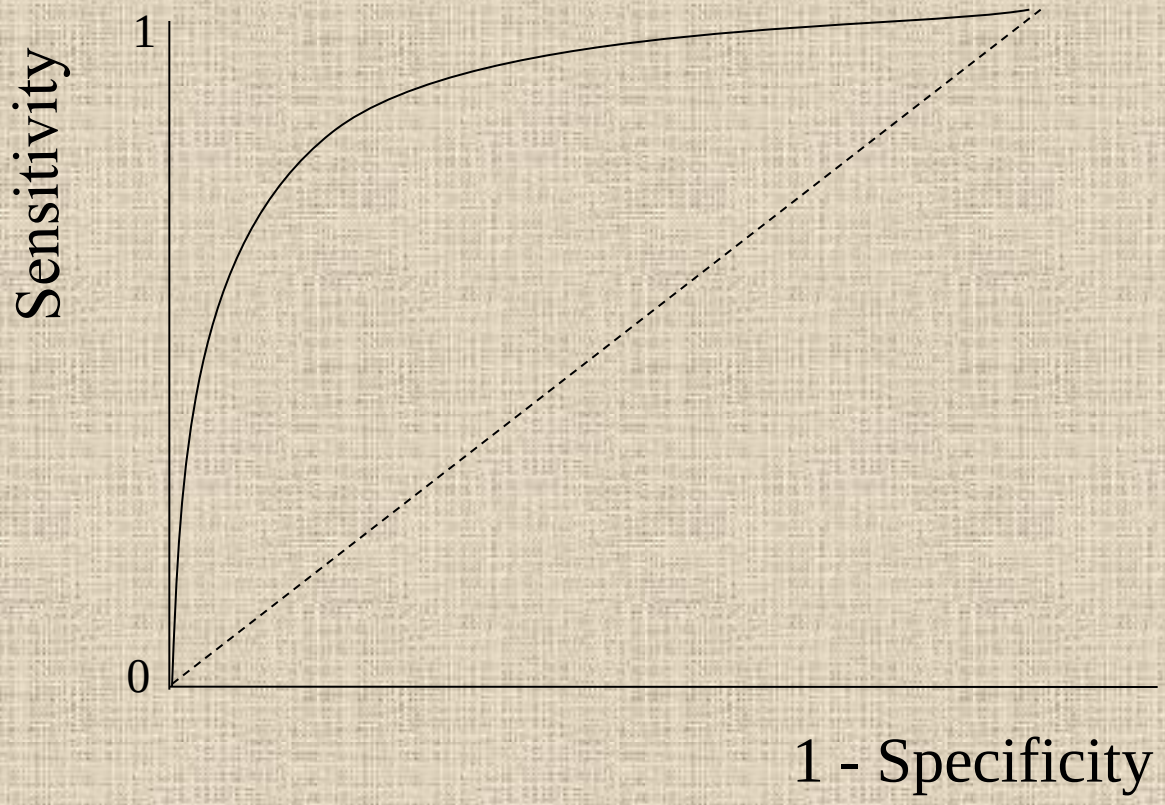
Continuous Measurements



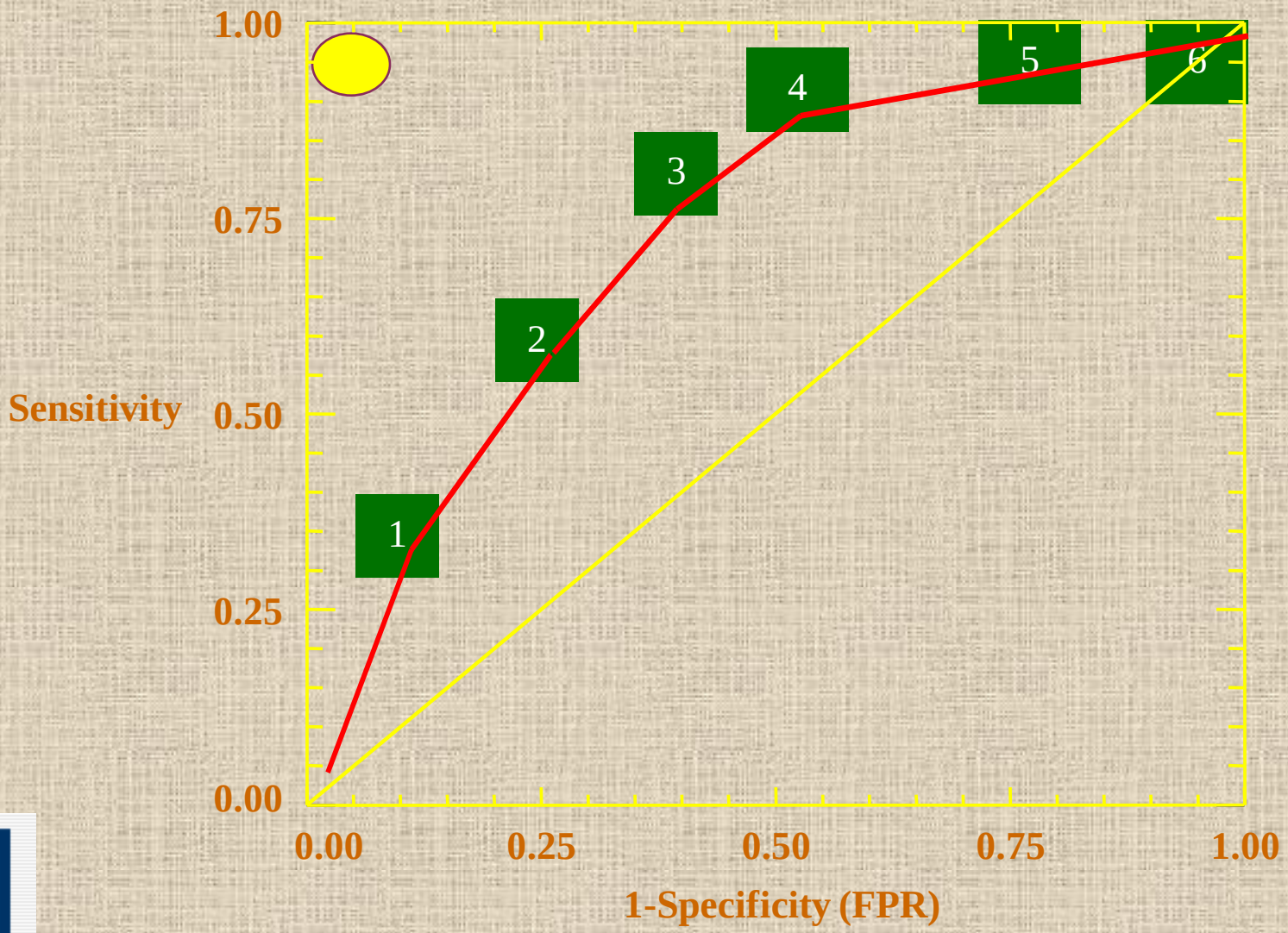
Continuous Measurements

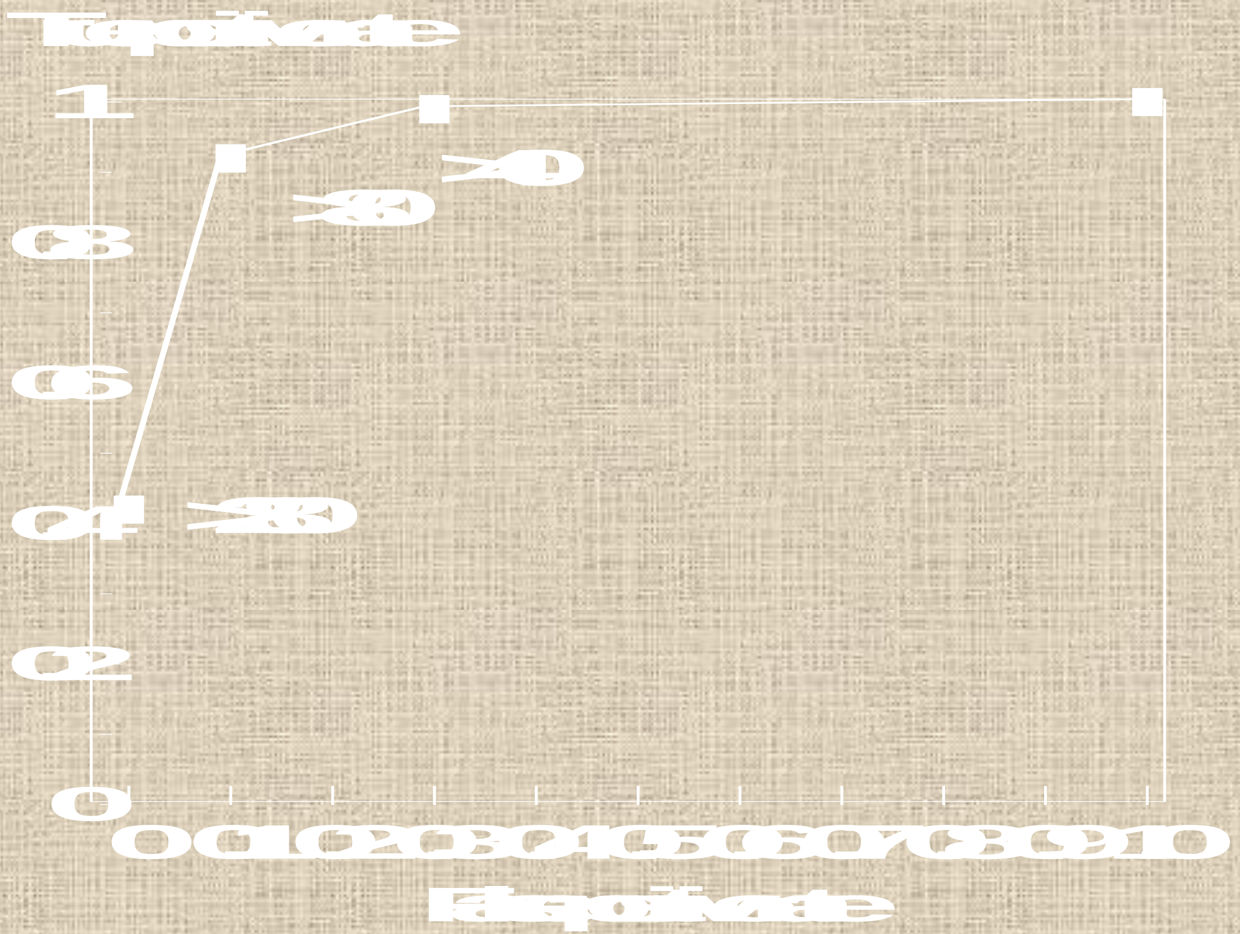


Receiver Operator Characteristic Curve
ROC Curve



ROC Curve Analysis





Thank You