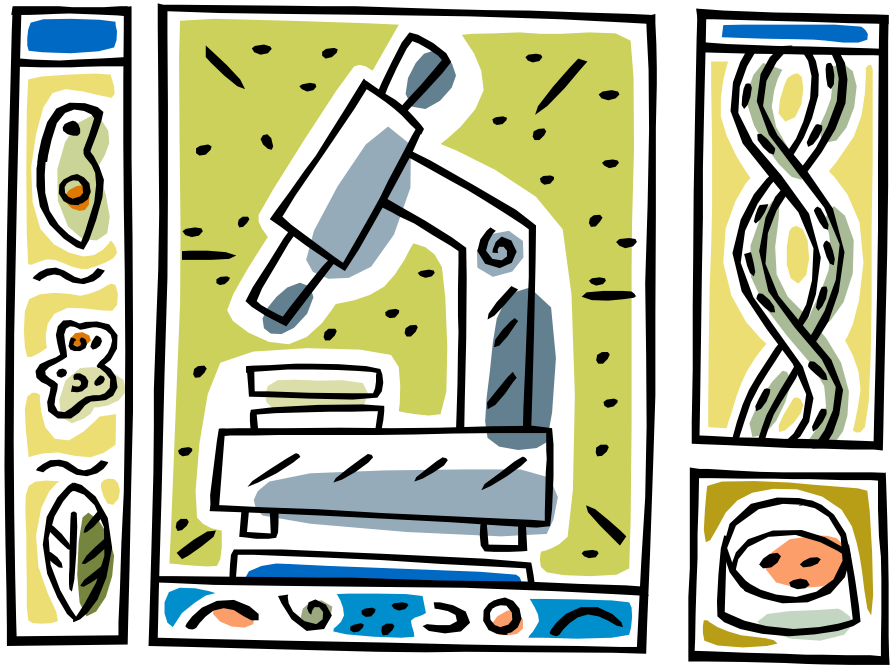


# PBIO 504

## Screening and Diagnostic Tests

## Sensitivity, Specificity, and Predictive Value

# Screening and Diagnostic Tests

- Diagnostic and screening tests are used to detect the presence or severity of disease in individuals.
- Clinicians rely on these tests to make decisions in treating patients.
- Screening can be selective, applying the screening tests to high risk groups of people, or it can be intended for mass screening.

# Screening and Diagnostic Tests

- Screening is useful for early detection of disease, when treatment has the greatest chance of being effective

- Importance:
  - improved morbidity and mortality rates

# Screening vs Diagnostic Tests

- Papanikolaou or Pap screening test for cervical cancer
- PSA test (prostate-specific antigen is a protein produced by cells of the prostate gland and it can be measured in blood)
- Mammogram
- Tuberculin skin test

- Biopsy
- B-cell chronic lymphocytic leukemia diagnosis is made through a combination of the microscopic examination of the peripheral blood and the flow cytometry analysis of the lymphocytes to confirm clonality and marker molecule expression

# Questions asked by clinical epidemiologists about screening and diagnostic tests:

Examples:

- Is the screening or diagnostic test accurate

- Who are those benefiting most from a certain screening or diagnostic test. How are they different (those who benefit and those who do not)

- Are there risks associated with a particular screening or diagnostic test

# Who Gets Screened?

- The goal is to identify symptomatic or asymptomatic individuals who might have a particular disease or condition.
- The decision to use a screening test on a given group of people is made on the basis of cost/benefit analysis.
- If the cost of testing all the individuals suspected is outweighed by the benefit of early detection to the few patients who have the disease, then a screening test is used.
- This means that there must be a treatment or some sort of medical care to offer to those with the disease.

# Types and Uses of Screening Tests

- **Establish a diagnosis in symptomatic patients**. E.g. an ECG to diagnose myocardial infarction in patients with chest pain.
- **Screen for disease in asymptomatic patients**. E.g. a prostate-specific antigen (PSA) test in men older than 50 years.
- **Provide prognostic information in patients with established disease**. E.g. a CD4 count in patients with HIV.
- **Monitor therapy by either benefits or side effects**. E.g. by measuring the international normalized ratio (INR) in patients taking warfarin.
- **A test may be performed to confirm that a person is free from a disease**. E.g. a pregnancy test to exclude the diagnosis of ectopic pregnancy.

Source: http://emedicine.medscape.com

# PKU Example

- **The PKU (phenyketonuria) test that babies receive by law in the nursery screens for a preventable form of mental retardation.**
- **It must be identified, and treatment started in the first few days of life, or the affected child will suffer irreversible brain damage.**
- **The test costs only a few dollars.**
- **The disease is caused by a defective gene which regulates metabolism of the amino acid phenylalanine. A metabolite of phenylalanine builds up in the blood stream because the enzyme that is supposed to break it down is defective. This substance accumulates in the brain where it is toxic.**
- **The disease is treated by a diet containing little or no phenylalanine, and is easily controlled by educating the parents about the types of food that the child must avoid. Growth and development are then normal.**
- **PKU affects between 1 in 10,000 and 1 in 20,000 newborns, depending on the country of origin**

# Gold Standard Tests

- The criterion (reference), or “gold-standard” test, decides either presence or absence of disease and the performance of a new screening test.
- Examples include pathological specimens for malignancies and pulmonary angiography for pulmonary embolism.
- However, criterion tests come with drawbacks; they are usually expensive, less widely available, more invasive, and riskier. The price most of the other diagnostic tests pay for their ease of use compared with their criterion standard is a decrease in accuracy.

# Examples of Gold Standard Tests

- Venography, the criterion test for vein thrombosis, is an invasive procedure with significant complications including renal failure, allergic reaction, and clot formation.
- These risks make it less desirable than the alternative diagnostic test – venous ultra-sonography - which is inexpensive and non-invasive.

# The "Gold Standard" for Evaluating True Disease Status

1. The most definitive diagnostic procedure to date
   e.g. microscopic examination of a tissue specimen

2. The best current available laboratory test
   e.g. serum antibodies to HIV

3. A comprehensive clinical evaluation
   e.g. clinical assessment of arthritis

# Validity and Reproducibility of Examinations and Medical Tests

- Clinical examinations, laboratory tests, and biological markers are thought to yield more reliable information, in general, than self reported information obtained through questionnaires.

- How can we check if that is true? We check the:
- **Validity** (or accuracy) of the test, and the
- **Reproducibility** (or precision) of the test.

# Validity and Reproducibility of Diagnostic and Screening Tests

- **Validity**: ability of a test to distinguish between those who have a disease and those who do not

  - *The validity of a test depends on the inherent quality of the test.*

- **Reproducibility**: will the same test results be obtained if the test is repeated, either by the same instrument or observer, or by two different instruments or observers?

  - *Reproducibility depends on the various external conditions in which the test is administered or analyzed, e.g. observer variability, biological variability of the subject, quality control of the instrument for recording test results.*

# Validity and Reproducibility of Diagnostic and Screening Tests

- Assess the validity and reproducibility of any given screening test before implementing.

- A good screening test should obviously be both valid and reliable.

- We must assess its performance before making it available in clinical setting or propose as a screening test for the general population.

# Measures of Test Validity

***Sensitivity*:**

*Percent of diseased persons who test positive*

Example: what proportion of women with breast cancer will have a positive mammogram?

***Specificity*:**

*Percent of non-diseased persons who test negative*

Example: what proportion of women without breast cancer will have a negative mammogram?

The goal is to have both of these measures as close to 100% as possible

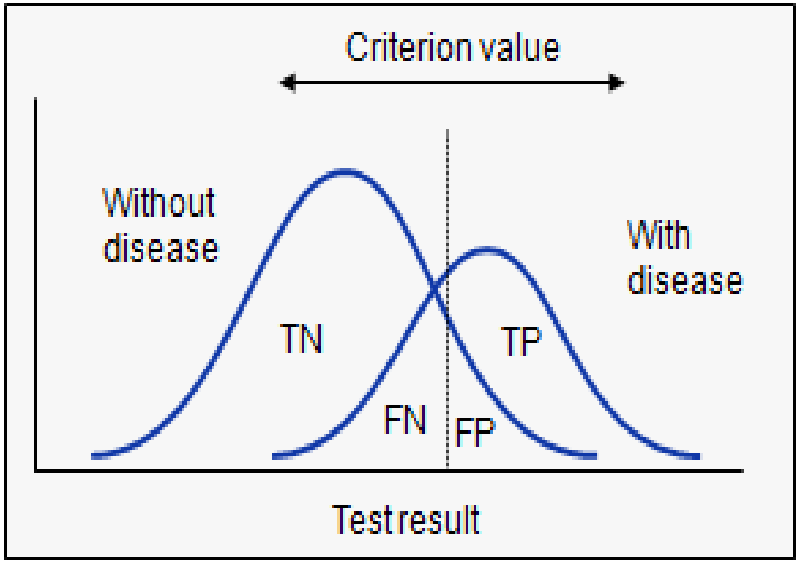
# Truth and Errors in Testing

| *Contrast between the new test result for a disease and the true state of disease* | | Truth about the disease: (as measured by the current gold standard test) | |
|---|---|---|---|
| | | Disease present | Disease absent |
| New Test Results: | Positive | **True positive** | **False positive** |
| | Negative | **False negative** | **True negative** |

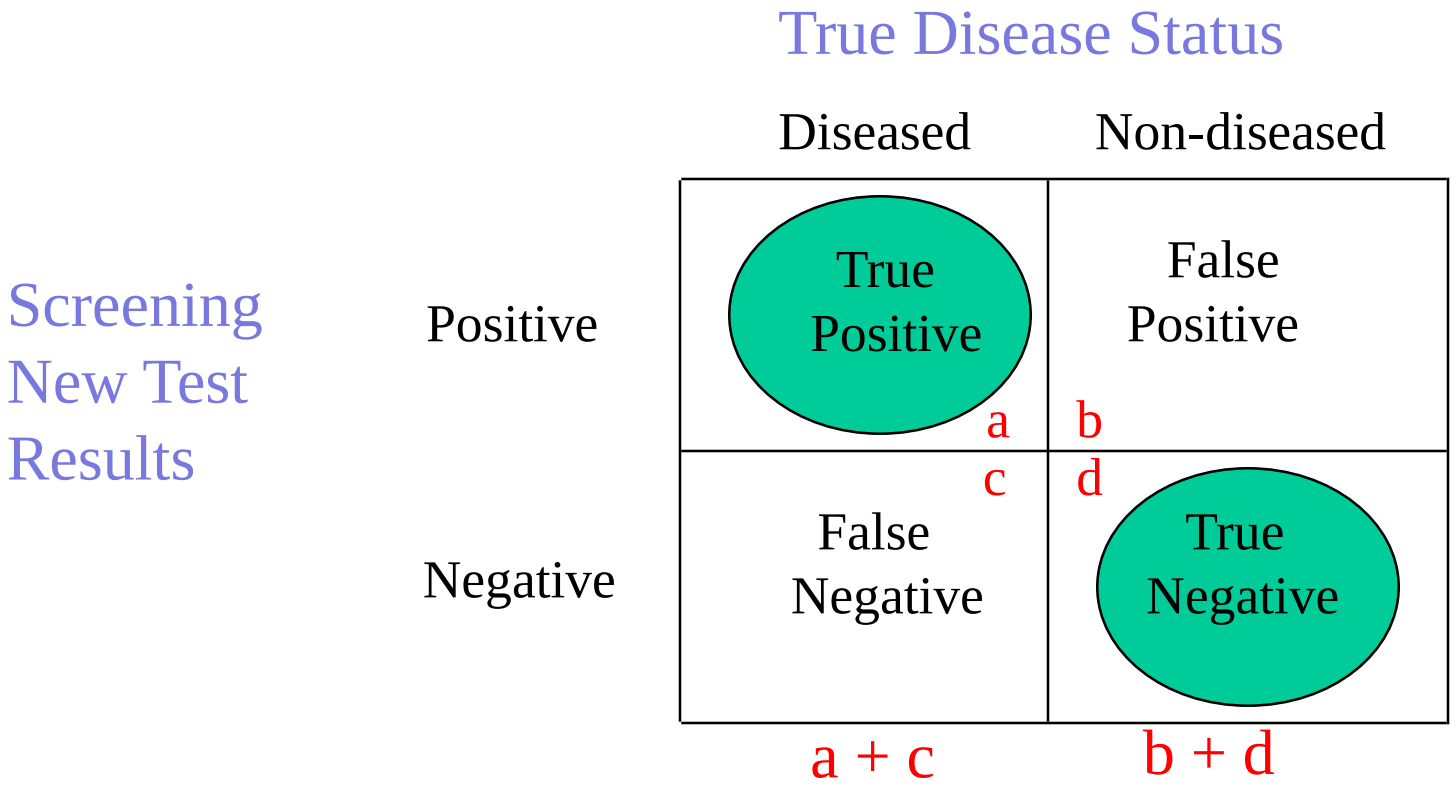
Gordis pp. 87-92

# The impact of the chosen cut-off point used to define diagnosis with the new test

- Consider two populations, one with and one without the disease.
- Notice the overlapping distributions of test results according to the new test and the cut-off point (criterion value).
- If we move the cut-off point in either direction, we are still misclassifying people from both populations.



# Sensitivity and Specificity




$$\text{Sensitivity} = \frac{\text{True positives}}{\text{All diseased}} = \frac{a}{a + c}$$

(Se and Sp are expressed as %)

$$\text{Specificity} = \frac{\text{True negatives}}{\text{All non-diseased}} = \frac{d}{b + d}$$

# Sensitivity and Specificity

- **Se**nsitivity tries to minimize the problem of **false negatives**

  -High Se: Low *false negative rate* (1-Se) is the probability of someone with the disease to have a negative test

- **Sp**ecificity tries to minimize the problem of the **false positives**

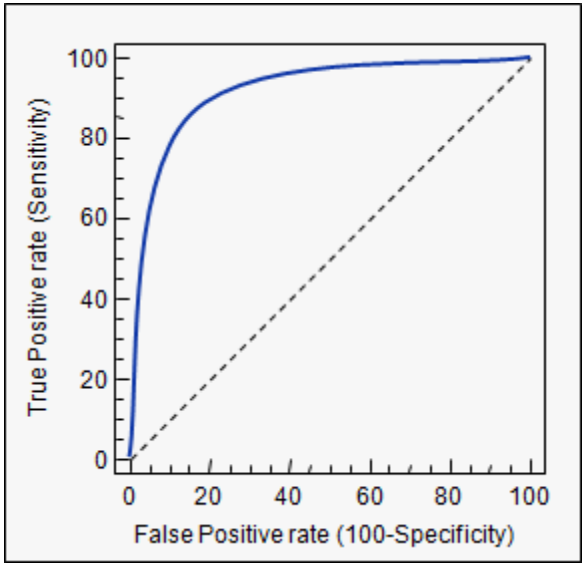
  -High Sp: Low *false positive rate* (1-Sp) is the probability of someone non diseased to have a positive test

- **Note**: when screening populations, a test with higher specificity is preferred, even if less sensitive.

# Receiver Operating Characteristics Curves

The ROC curve is a graph of true positives (Sensitivity) on the Oy-axis versus false positives (1 – Specificity) on the Ox-axis for different cut-off points or critical values indicating normal test results (no disease). It is a graphical display of trade-off between both measures for each cut-off point value considered for the new test.

The area under the curve (AUC) is associated with the accuracy of the new test and it corresponds to the probability of the test correctly classifying patients as True Positive or True Negative.




*Homework Reading Assignment:*
*Ashwell et.al. 2012 posted on Canvas*

# High Se for Blood Bank Screening

- To protect the blood supply from viruses in donated blood bags, the blood is tested for HIV, HBV, and HCV, as these viruses can infect the recipient of the tainted blood. If any virus is detected then the blood is discarded.
- Screening tests to detect these viruses should have extremely high **Sensitivity**.
- **Specificity** can be less rigorous (i.e. some blood will be falsely discarded).

# Example: Blood Sugar Testing

| *Blood glucose level of 5.8 mM or above as a screening test for diabetes* | | True Disease Status | |
|---|---|---|---|
| | | Diabetes present | Diabetes absent |
| Screening Test Results | Positive | **True positive a=62** | **False positive b=125** |
| | Negative | **False negative c=1** | **True negative d=215** |
| | **Totals:** | **N=63** | **N=340** |

**Sensitivity**= a/(a+c) = 62/63 = .984 (or 98.4%)

**Specificity**= d/(b+d) = 215/340 = .632 (or 63.2%)

***Systematic Error*** = (a+b)/(a+c)

# Predictive Value of a Test

- **Positive predictive value:** percent of positive tests that are truly positive, or

  PPV: percent of those diseased among all those who tested positive
- **Negative predictive value:** percent of negative tests that are truly negative, or

  NPV: percent of those non-diseased among all those who tested negative

# Predictive Value of a Test

In addition to Se and Sp, the performance of a screening test is also measured by its Predictive Value

- **Positive predictive value:**

  PPV: the probability that one has the Dz, given he tested positive

- **Negative predictive value:**

  NPV: the probability that one does not have the Dz, given he tested negative

# Predictive Value of a Test

- Bayes Theorem is applied when calculating conditional probabilities. We use this formula to obtain the positive and negative predictive values of a test.

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A|B) \times P(B) + P(A|\overline{B}) \times P(\overline{B})}$$

# Positive (Negative) Predictive Value: Formulae

- **Definitions**

  **Prevalence: the proportion of those with the disease (P(D))**

  **Positive predictive value P(D | +): the proportion of patients with positive test results who are correctly diagnosed**

  **Negative predictive value P(~D | -): the proportion of patients with negative test results who are correctly diagnosed**

- **Relationship**

$$PPV = \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + (1 - P(- \mid \sim D))(1 - P(D))}$$

$$= \frac{sensitivity * prevalence}{sensitivity * prevalence + (1 - specificty) * (1 - prevalence)}$$

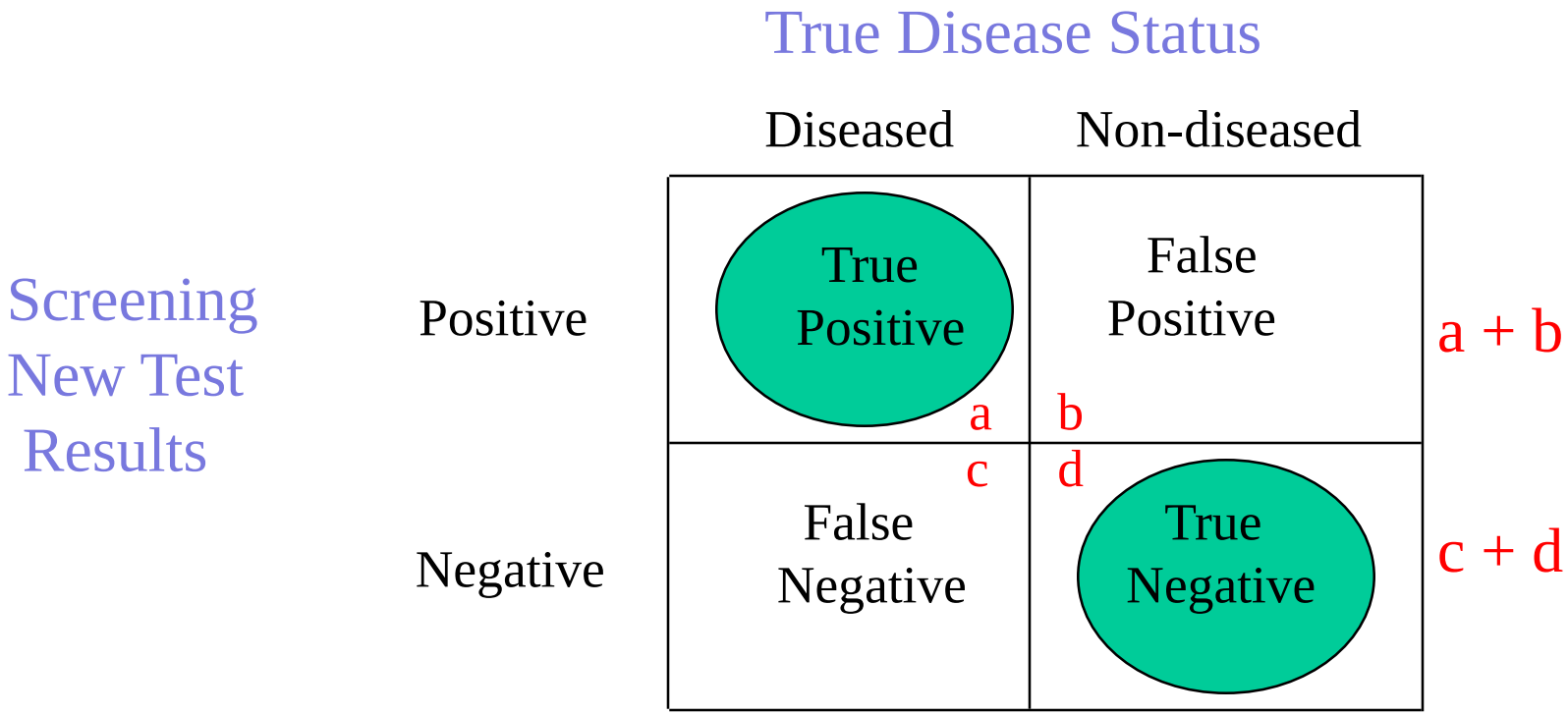
- **PPV depends on prevalence. Ask "what patient population is the test intended for?"**

# Generalizability of the Predictive Value of a Test

Predictive values depend on the prevalence rate of the disease in the sample selected to check the performance of the new screening test.

For the predictive values to be meaningful at the population level (or population of interest level), *the sample used* to test the new screening test against the current gold standard test, *should have the true distribution of that particular disease as it is in the general population*, (i.e the Pe of Dz in the selected sample has to be the same as the Pe of Dz in the general population (or the population of interest).

# Calculate the Predictive Value of a Test




$$\text{Positive Predictive Value} = \frac{\text{True positives}}{\text{All positive tests}} = \frac{a}{a + b}$$

Expressed as %

$$\text{Negative Predictive Value} = \frac{\text{True negatives}}{\text{All negative tests}} = \frac{d}{c + d}$$

# Predictive Values from the Diabetes Testing Example

- **PPV** = 62 / (62+125) = 0.332 (33.2%)
- **NPV** = 215 / (215+1) = 0.995 (99.5%)
- For the sample tested these percentages answer questions like:

  -What is the probability that someone with a positive test result to actually be diabetic?

  -What is the proportion of  people without diabetes among those with negative test results?

- Note that in the diabetes testing example only 62 (15%) of the 403 people tested have the disease: positive predictive value is generally low for rare diseases.

# Effect of Prevalence on Positive Predictive Value with Constant Sensitivity and Specificity

| Prevalence (%) | PPV (%) | Se (%) | Sp (%) |
|---|---|---|---|
| 0.1 | 1.8 | 90 | 95 |
| 1.0 | 15.4 | 90 | 95 |
| 5.0 | 48.6 | 90 | 95 |
| 50.0 | 94.7 | 90 | 95 |

# Example of the Effect of Prevalence on Positive Predictive Value

## A. Without palpable mass in breast

Surgical biopsy (Gold Standard)

| Fine Needle Aspiration | Cancer | No Cancer |
|---|---|---|
| Positive | 14 | 8 |
| Negative | 1 | 91 |

Prevalence=13%
Se = 93%
Sp = 92%
PPV = 64%
NPV = 99%

## B. With palpable mass in breast

Surgical biopsy (Gold Standard)

| Fine Needle Aspiration | Cancer | No Cancer |
|---|---|---|
| Positive | 113 | 15 |
| Negative | 8 | 181 |

Prevalence = 38%
Se = 93%
Sp = 92%
PPV = 88%
NPV = 96%

# Measures of Reliability or Repeatability of Tests

1. Will the same results be obtained if two different observers, or two different laboratory instruments, perform the same measurement?

2. Will the same results be obtained if the same test is repeated on the same subject, using the same method of measurement?

# Sources of Variability Examples

- **Observer variability** may be reduced for instance by standardizing the measurement methods in an operations manual, and training and certifying the observer

- **Instrument variability** may be addressed by refining the instrument (calibrate)

- **Subject variabilit**y may be reduced by standardizing the measurement methods, automating the instrument used, and repeating the measurements

# Variability (or Reproducibility)

- Several factors, other than purely random errors, can affect the results of a screening or biomarker test.

  Some examples include:

- **Inter-individual variability** in the test
- **Intra-individual variability** in the test
- **Inter-observer reliability** in scoring the test
- **Intra-observer reliability** in scoring the test

# Measuring Overall Agreement Between Two Independent Observers

| | | *Observer 1* | |
|---|---|---|---|
| | | Positive | Negative |
| *Observer 2* | Positive | a<br>41 | b<br>3 |
| | Negative | c<br>4 | d<br>27 |

**Overall Agreement**:

Including ‘d’ negative agreement = $\frac{a + d}{a + b + c + d}$ = 0.907 (91%)

Excluding ‘d’ negative agreement = $\frac{a}{a + b + c}$ = 0.85 (85%)

# Measuring Agreement Beyond Chance:

## The Kappa Statistic

$$\text{Kappa} = \frac{\text{Overall agreement - Agreement expected by chance}}{\text{1.0 - Agreement expected by chance}}$$

| | | Observer 1: Positive | Observer 1: Negative | |
|---|---|---|---|---|
| Observer 2 | Positive | a = 41 | b = 3 | $m_1$=44 |
| | Negative | c = 4 | d = 27 | $m_2$=31 |
| | | $n_1$=45 | $n_2$=30 | T=75 |

Agreement expected by chance = $\frac{26.4 + 12.4}{75} = 0.517$

Cell a:$(m_1/T * n_1/T) * T=(m_1/T)n_1=26.4$

For cell b: $(m_1/T)n_2 = 17.6$

For cell c: $(m_2/T)n_1 = 18.6$

For cell d: $(m_2/T)n_2 = 12.4$

$$\text{Kappa} = \frac{0.907 - 0.517}{1.0 - 0.517}$$

Kappa = 0.81

# Interpretation of Kappa Values

- Excellent agreement:  Kappa from 0.75-1.0
- Good agreement:  Kappa from 0.40-0.74
- Poor agreement: Kappa less than 0.40
- No agreement:  Kappa at or near 0.0

# Exercise 1

| | | Disease Status: Positive | Disease Status: Negative |
|---|---|---|---|
| Test | Positive | a<br>99 | b<br>90 |
| | Negative | c<br>1 | d<br>810 |

Prevalence is 10%

Calculate:

Se, Sp, PPV, NPV

# Exercise 2

| | | Disease Status: Positive | Disease Status: Negative |
|---|---|---|---|
| Test | Positive | a<br>495 | b<br>50 |
| | Negative | c<br>5 | d<br>450 |

Prevalence is 50%

Calculate:

Se, Sp, PPV, NPV

# Prostate Cancer Screening

**The American Cancer Society did issue five years ago new guidelines for patients and doctors on screening for prostate cancer. Officials said they felt that an update was needed in response to the recent findings of a large study involving 76,600 men in the United States who received annual PSA blood tests and digital rectal exams compared to those who had "usual care." Researchers found little difference in prostate cancer death rates between the two groups. In Europe, a study involving 182,000 men is finding similar results. You would think that widespread screenings would make sense given the fact that prostate cancer is probably the most common cancer among men. However, it is an age-related type of cancer (two-thirds of all men 65 and older suffer from it).**

**Overall, seventeen percent of men 50 years and older will develop prostate cancer during their lifetime.**

Adapted from "Our patients, ourselves: Guidelines for prostate cancer screening tests" (Linda Rhodes: blog.pennlive.com/life/2010)

Annals of Internal Medicine
Established in 1927 by the American College of Physicians

**From: Benefits and Harms of Prostate-Specific Antigen Screening for Prostate Cancer: An Evidence Update for the U.S. Preventive Services Task Force**

Ann Intern Med. 2008;149(3):192-199. doi:10.7326/0003-4819-149-3-200808050-00009

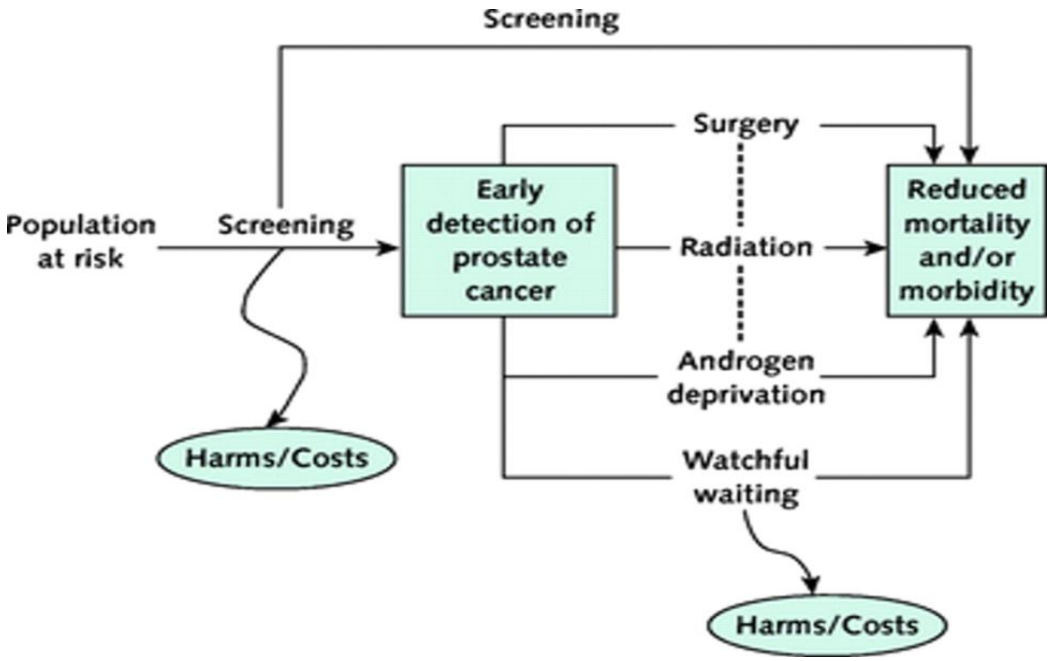



**Figure Legend:**

**Analytic framework for screening for prostate cancer.**

# Get Screened with the PSA Test?

- Most forms of prostate cancer are very slow moving — so slow that a man diagnosed with it in his 70s or 80s is more likely to die from another illness such as heart attack or stroke.
- Experts have found that aggressively treating a cancer that might never cause real problems for the rest of a man's natural life can cause more harm than good.
- Treatment for prostate cancer has pretty tough side effects for many men including incontinence, bowel problems and impotence.
- That's why the American Cancer Society (ACS) thinks doctors and male patients should talk about the benefits and risks of tests and treatments for prostate cancer. Factors like family history have to be weighed in.

# ACS Advice

- ACS officials also want men to be aware of the uncertainty of tests results before making irreversible life-altering decisions.
- According to the ACS: “No screening test is perfect, but the degree of over-diagnosis and associated over-treatment appears to be greater for prostate screening than for any other of the cancers for which routine screening currently occurs.”

# Over-Diagnosis Bias

Over-diagnosis bias means that the screening test identifies a condition that would not show clinical signs or symptoms and does not have an impact on the quality of life of an individual probably for a very long time, and may not even be the eventual cause of death

- Screening identifies a large number of affected individuals otherwise not having a chance of being detected
- As a result, an individual may undergo treatment and be subjected to all risks associated with it

# Current Recommendations for Prostate Cancer Screening

In May 2012, the U.S. Preventive Services Task Force (USPSTF) recommended against PSA-based screening for prostate cancer for everyone based upon the following reasons:

- The benefit from PSA-based screening for prostate cancer is rather small since only 1 in 1,000 man can avoid death from prostate cancer because of PSA-based screening and there is no evidence supporting that PSA testing could improve outcomes of treating prostate cancer
- Patients screened with PSA test may suffer from more harms because 3~4% of them may develop erectile dysfunction or urinary incontinence and they may encounter pain or infection
- PSA-based screening could lead to considerable over-diagnosis of prostate cancer

**Source;**

**http://www.uspreventiveservicestaskforce.org/prostatecancerscreening.htm**