JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Evaluation of Diagnostic and Screening Tests: Validity and Reliability

Sukon Kanchanaraksa, PhD
Johns Hopkins University

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section A

Sensitivity and Specificity

Correctly Classifying Individuals by Disease Status

- Tests are used in medical diagnosis, screening, and research
- How well is a subject classified into disease or non-disease group?
  - Ideally, all subjects who have the disease should be classified as "having the disease" and vice versa
  - Practically, the ability to classify individuals into the correct disease status depends on the accuracy of the tests, among other things

Diagnostic Test and Screening Test

- A **diagnostic test** is used to determine the presence or absence of a disease when a subject shows signs or symptoms of the disease
- A **screening test** identifies asymptomatic individuals who may have the disease
- The diagnostic test is performed **after** a positive screening test to establish a definitive diagnosis

Some Common Screening Tests

- Pap smear for cervical dysplasia or cervical cancer
- Fasting blood cholesterol for heart disease
- Fasting blood sugar for diabetes
- Blood pressure for hypertension
- Mammography for breast cancer
- PSA test for prostate cancer
- Fecal occult blood for colon cancer
- Ocular pressure for glaucoma
- PKU test for phenolketonuria in newborns
- TSH for hypothyroid and hyperthyroid

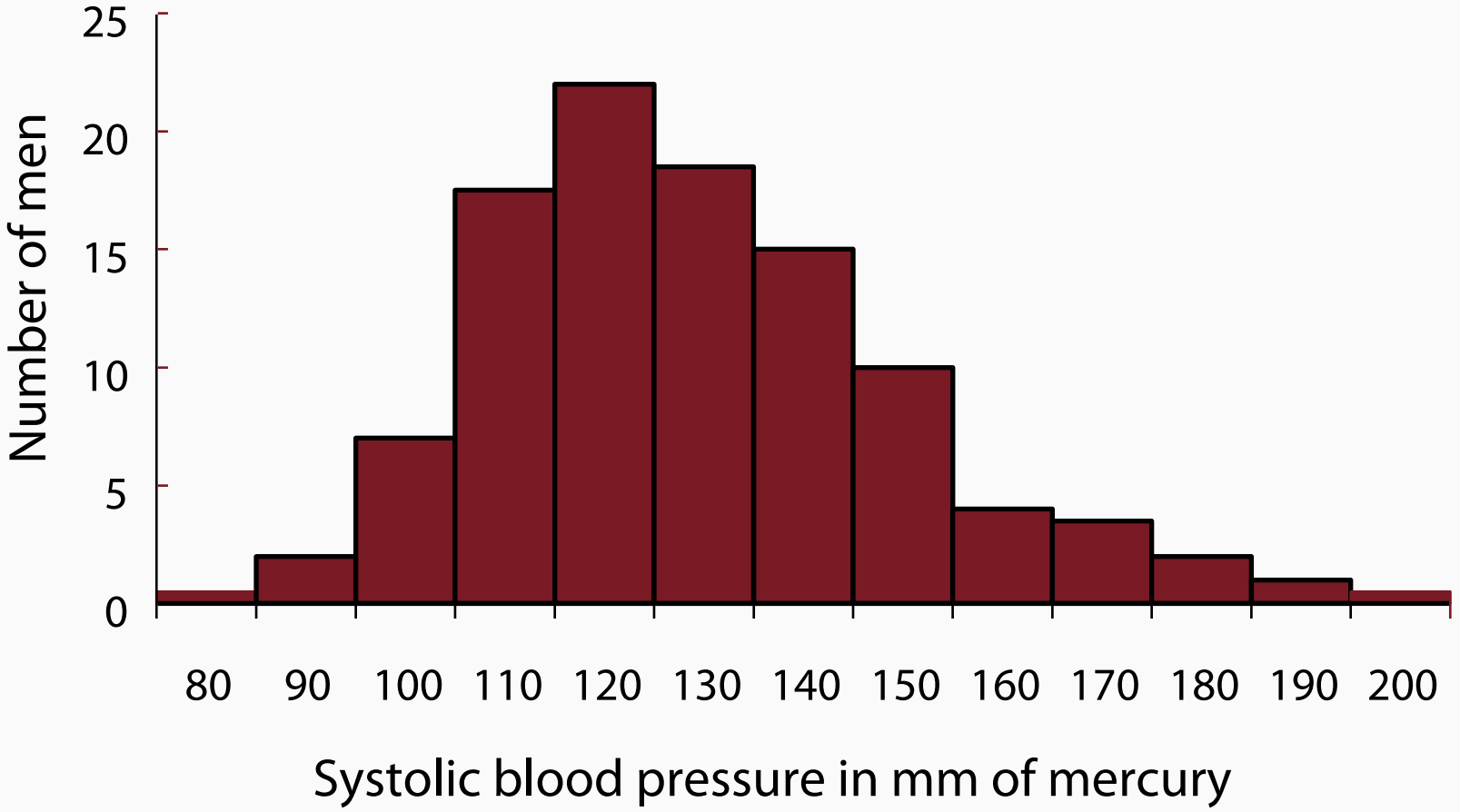
Variation in Biologic Values

- Many test results have a continuous scale (are continuous variables)
- Distribution of biologic measurements in humans may or may not permit easy separation of diseased from non-diseased individuals, based upon the value of the measurement

Distribution of Tuberculin Reactions

Source: Edwards et al, WHO Monograph 12, 1953

Distribution of Systolic Blood Pressures: 744 Employed White Males, Ages 40–64



Source: Data of W.W. Holland

Validity

- **Validity** is the ability of a test to indicate which individuals have the disease and which do not

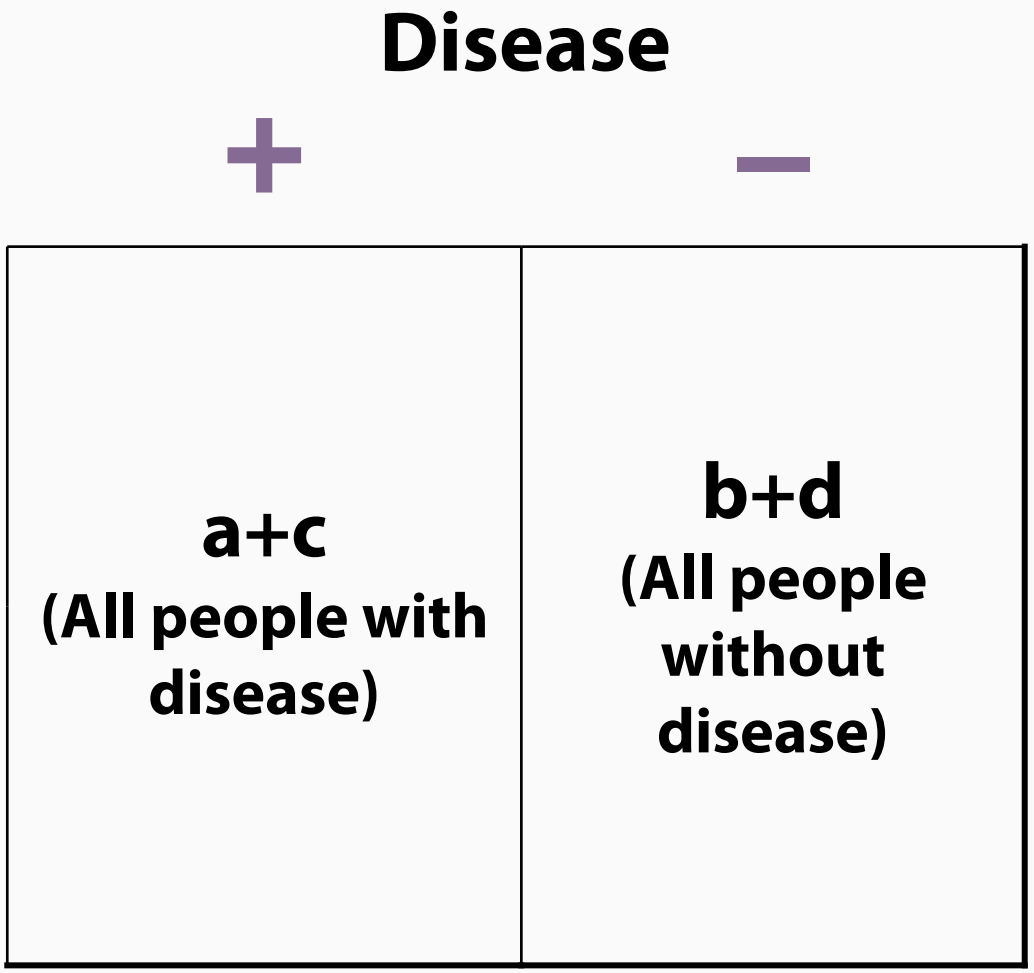
Sensitivity and Specificity

- **Sensitivity**
  - The ability of the test to identify correctly those who have the disease
- **Specificity**
  - The ability of the test to identify correctly those who **do not have** the disease

Determining the Sensitivity, Specificity of a New Test

- Must know the correct disease status prior to calculation
- **Gold standard test** is the best test available
  - It is often invasive or expensive
- A **new test** is, for example, a new screening test or a less expensive diagnostic test
- Use a 2 x 2 table to compare the performance of the new test to the gold standard test

Gold Standard Test



Comparison of Disease Status: Gold Standard Test and New Test

New test	Disease +	Disease −
+	**a** (True positives)	**b**
−	**c**	**d** (True negatives)

Sensitivity

- **Sensitivity** is the ability of the test to identify correctly those who have the disease (a) from all individuals with the disease (a+c)

$$\text{sensitivity} = \frac{a}{a+c} = \frac{\text{true positives}}{\text{disease}+}$$
$$= \Pr(T+ \mid D+)$$

- Sensitivity is a fixed characteristic of the test

Specificity

- **Specificity** is the ability of the test to identify correctly those who do not have the disease (d) from all individuals free from the disease (b+d)

$$\text{sensitivity} = \frac{d}{b+d} = \frac{\text{true negatives}}{\text{disease}-} = \Pr(T- \mid D-)$$

- Specificity is also a fixed characteristic of the test

Applying Concept of Sensitivity and Specificity to a Screening Test

- Assume a population of 1,000 people
- 100 have a disease
- 900 do not have the disease
- A screening test is used to identify the 100 people with the disease
- The results of the screening appears in this table

Screening Results	True Characteristics in Population		Total
	Disease	No Disease	
Positive	80	100	180
Negative	20	800	820
Total	100	900	1,000

Calculating Sensitivity and Specificity

Screening Results	True Characteristics in Population		Total
	Disease	No Disease	
Positive	80	100	180
Negative	20	800	820
Total	100	900	1,000

Sensitivity = 80/100 = 80%

Specificity = 800/900 = 89%

Evaluating Validity

Screening Results	True Characteristics in Population		Total
	Disease	No Disease	
Positive	80	100	180
Negative	20	800	820
Total	100	900	1,000

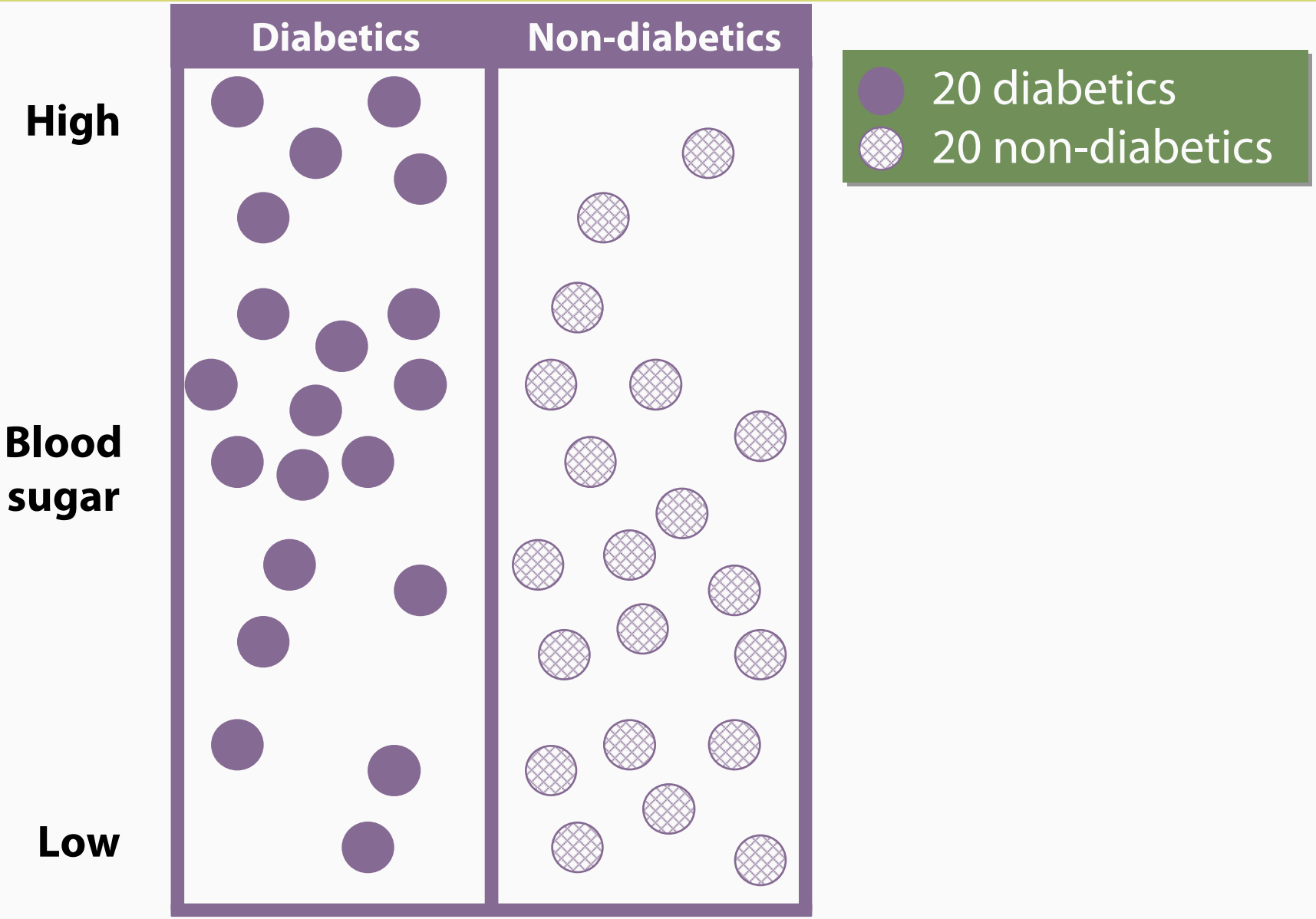
Sensitivity = 80/100 = 80% **Specificity** = 800/900 = 89%

Examining the Effect of Changing Cut-Points

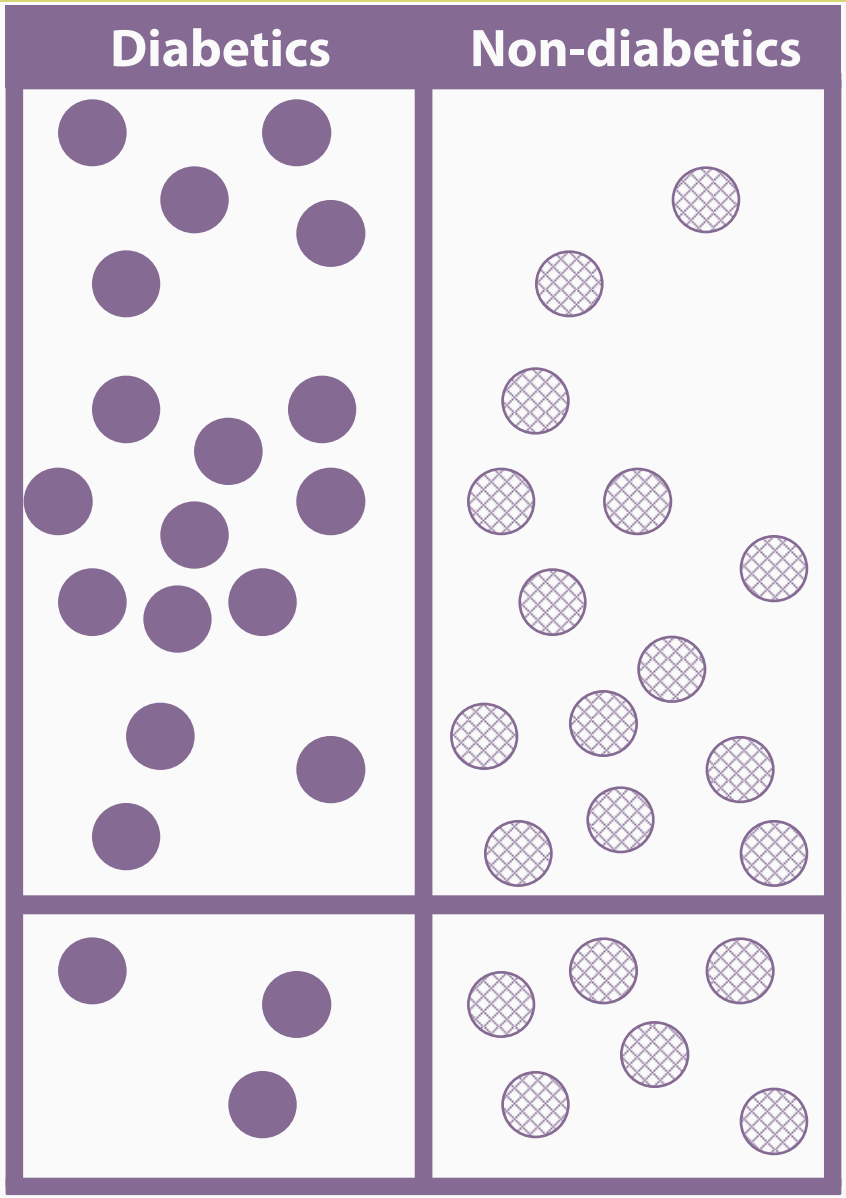
- Example: type II diabetes mellitus
  - Highly prevalent in the older, especially obese, U.S. population
  - Diagnosis requires oral glucose tolerance test
  - Subjects drink a glucose solution, and blood is drawn at intervals for measurement of glucose
  - Screening test is fasting plasma glucose
    - Easier, faster, more convenient, and less expensive

http://en.wikipedia.org/wiki/Oral_glucose_tolerance_test
http://care.diabetesjournals.org/cgi/content/full/25/suppl_1/s21

Concept of Sensitivity and Specificity



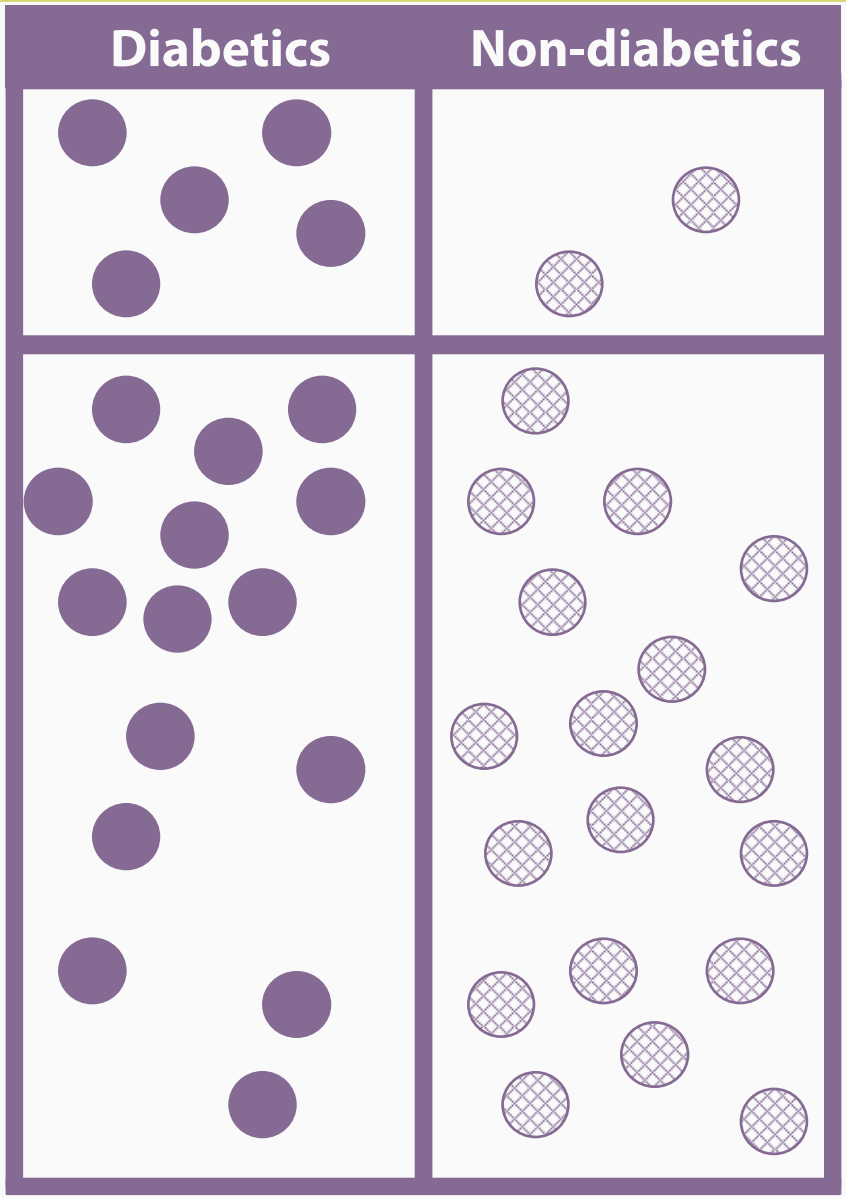
Concept of Sensitivity and Specificity



Subjects are screened using fasting plasma glucose with a low (blood sugar) cut-point

	Diabetics	Non-Diabetics
+	17	14
–	3	6
	20	20
	Sens=85%	Spec=30%

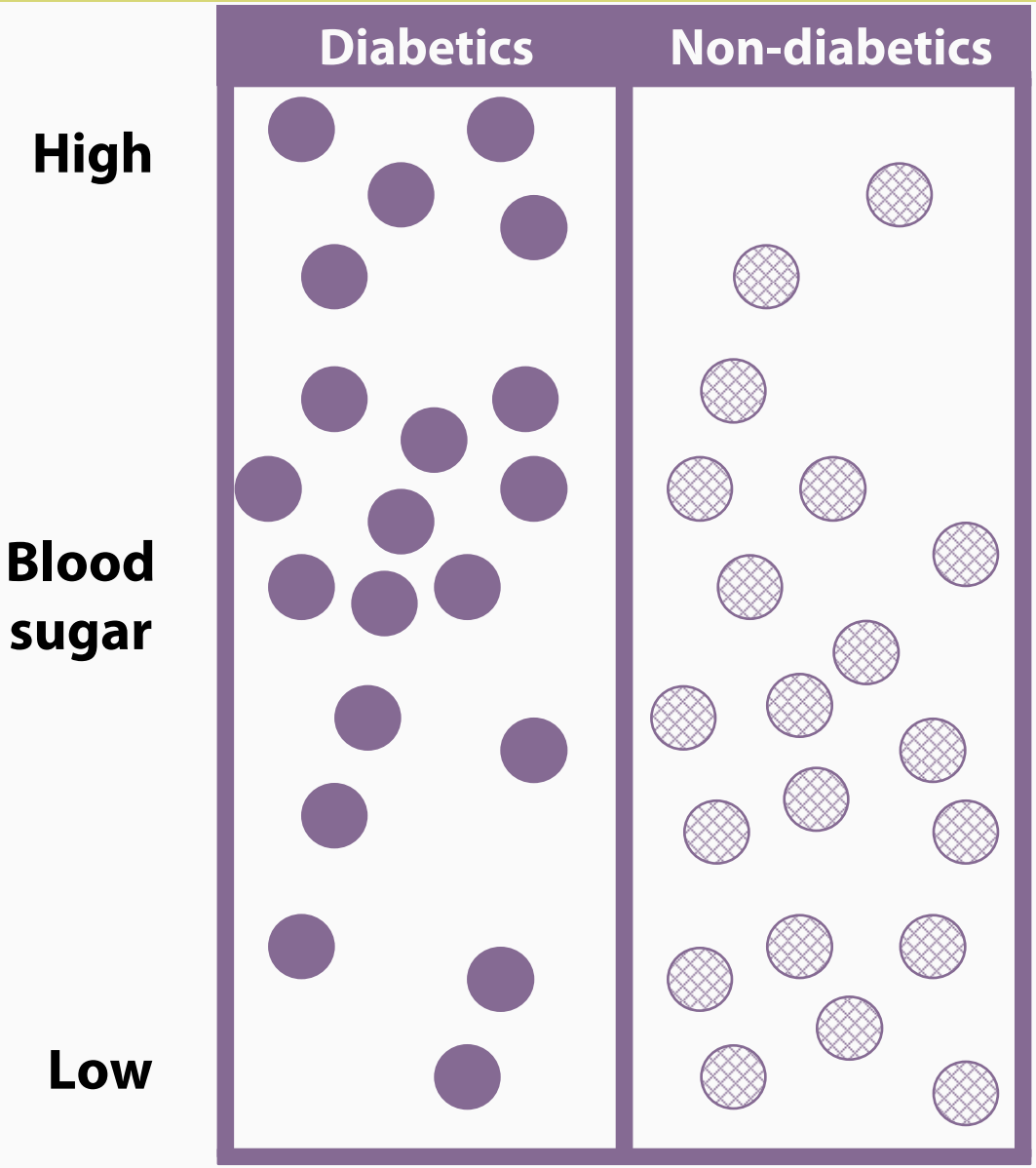
Concept of Sensitivity and Specificity



Subjects are screened using fasting plasma glucose with a high cut-point

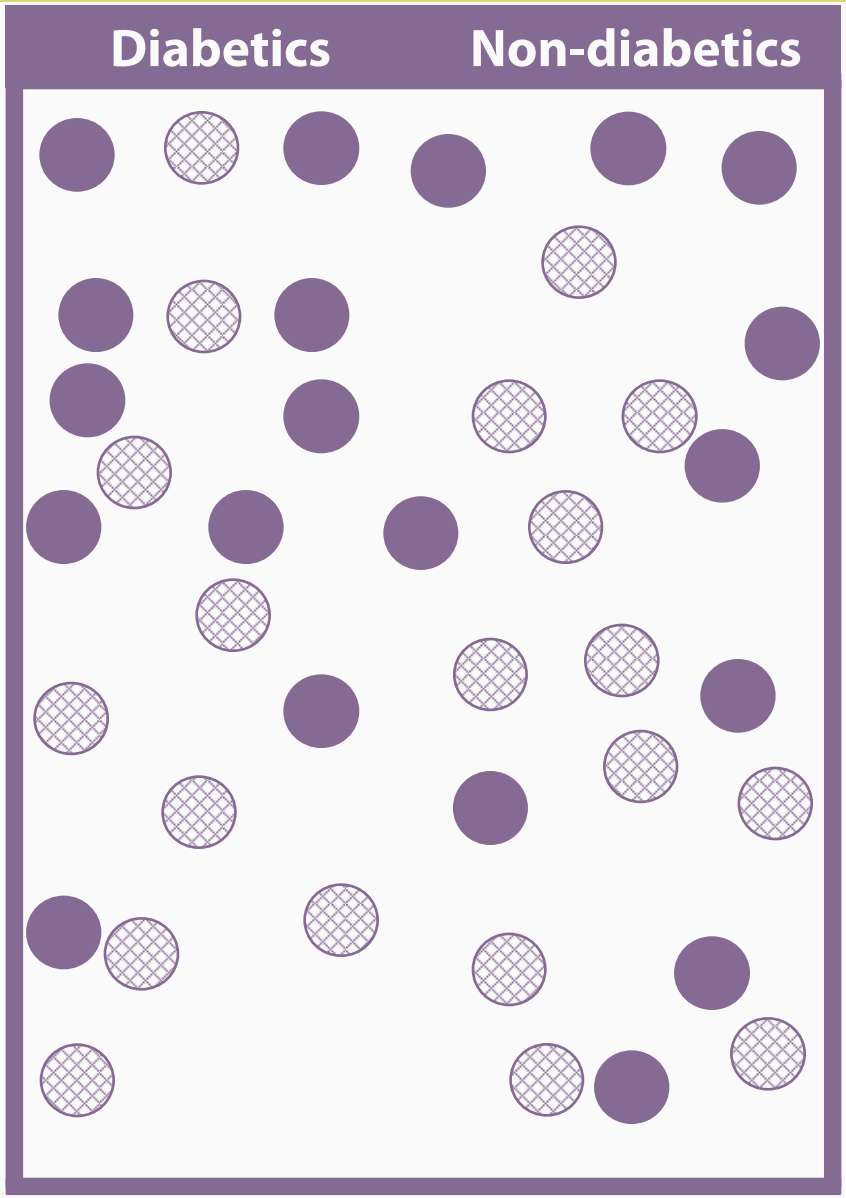
	Diabetics	Non-Diabetics
+	5	2
–	15	18
	20	20
	Sens=25%	Spec=90%

Concept of Sensitivity and Specificity



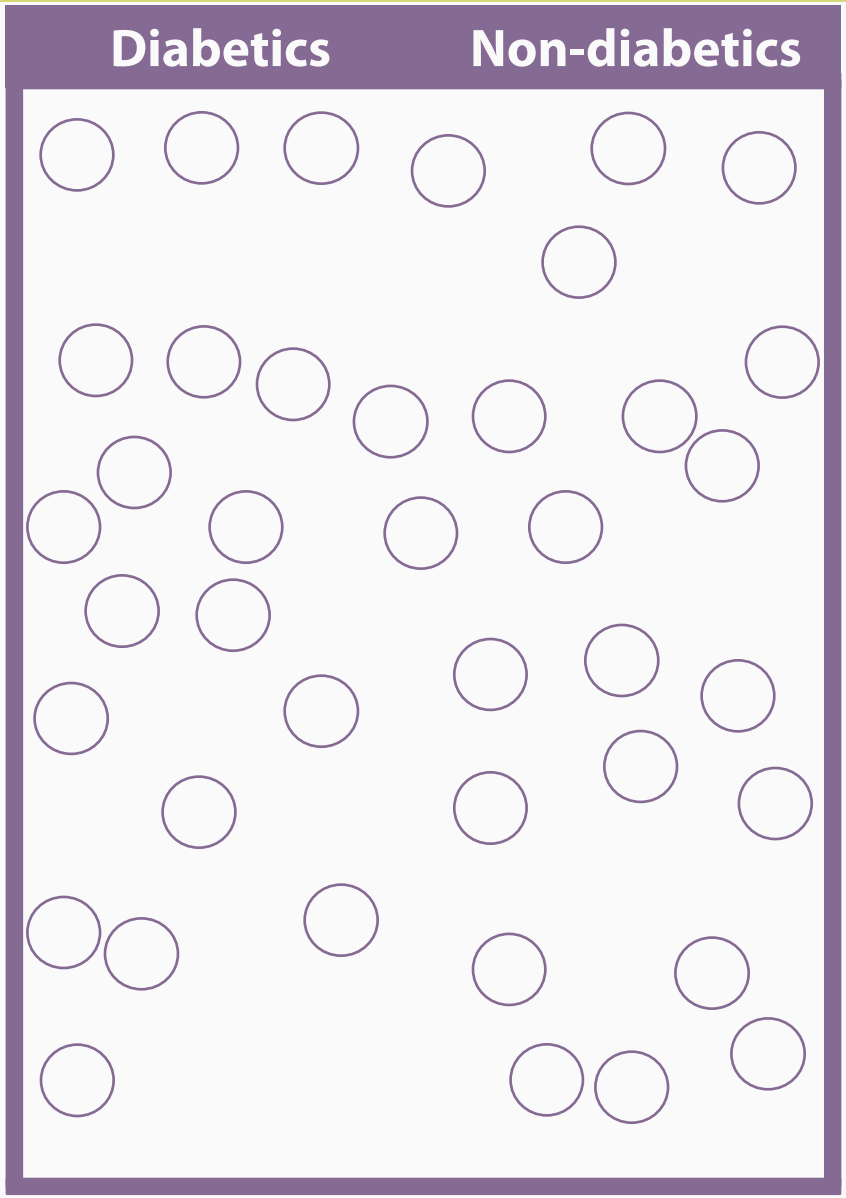
In a typical population, there is no line separating the two groups, and the subjects are mixed

Concept of Sensitivity and Specificity



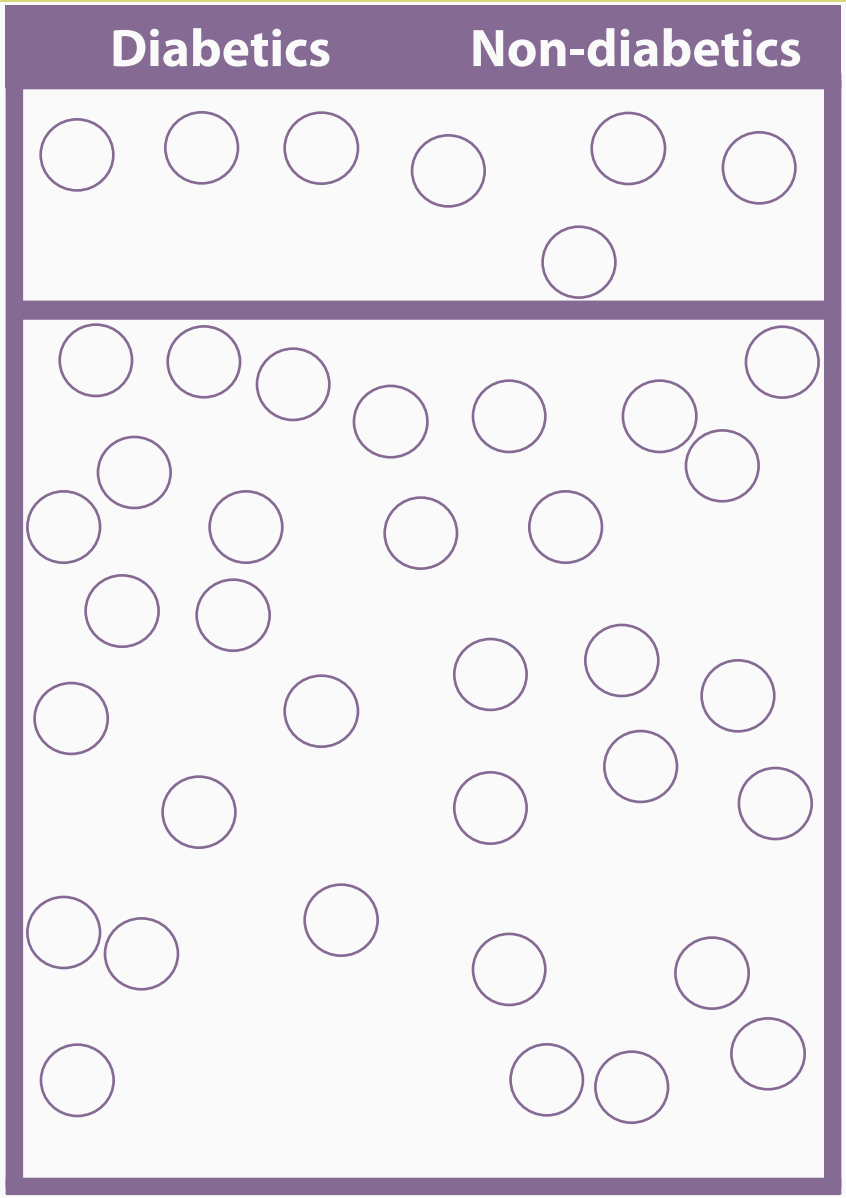
In a typical population, there is no line separating the two groups, and the subjects are mixed

Concept of Sensitivity and Specificity



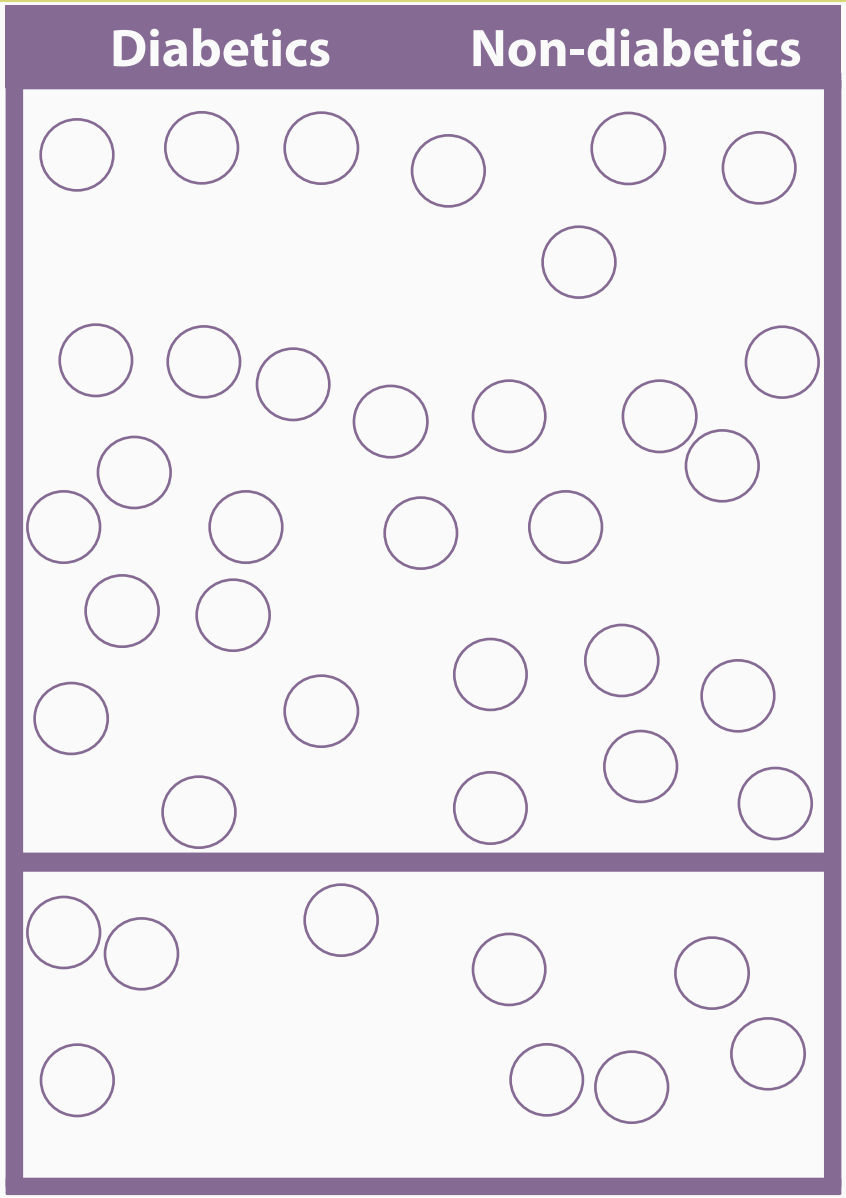
In fact, there is no color or label

Concept of Sensitivity and Specificity



A screening test using a high cut-point will treat the bottom box as normal and will identify the 7 subjects above the line as having diabetes

Concept of Sensitivity and Specificity



A screening test using a high cut-point will treat the bottom box as normal and will identify the 7 subjects above the line as having diabetes;

But a low cut-point will result in identifying 31 subjects as having diabetes

Lessons Learned

- Different cut-points yield different sensitivities and specificities
- The cut-point determines how many subjects will be considered as having the disease
- The cut-point that identifies more true negatives will also identify more false negatives
- The cut-point that identifies more true positives will also identify more false positives

Where to Draw the Cut-Point

- If the diagnostic (confirmatory) test is expensive or invasive:
  - Minimize false positives
    or
  - Use a cut-point with high specificity
- If the penalty for missing a case is high (e.g., the disease is fatal and treatment exists, or disease easily spreads):
  - Maximize true positives
    - That is, use a cut-point with high sensitivity
- Balance severity of false positives against false negatives

Behind the Test Results

Test	Disease +	Disease −
+	**a** (True positives)	**b** (False positives)
−	**c** (False negatives)	**d** (True negatives)

Review

- Fill in the missing cells and calculate sensitivity and specificity for this example

Screening Results	True Characteristics in Population		Total
	Disease	No Disease	
Positive	240		
Negative		600	
Total	300	700	1,000

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section B

Multiple Testing

Use of Multiple Tests

- Commonly done in medical practice
- Choices depend on cost, invasiveness, volume of test, presence and capability of lab infrastructure, urgency, etc.
- Can be done **sequentially** or **simultaneously**

Sequential Testing (Two-Stage Screening)

- After the first (screening) test was conducted, those who tested **positive** were brought back for the second test to further reduce false positives
- Consequently, the overall process will increase specificity but with reduced sensitivity

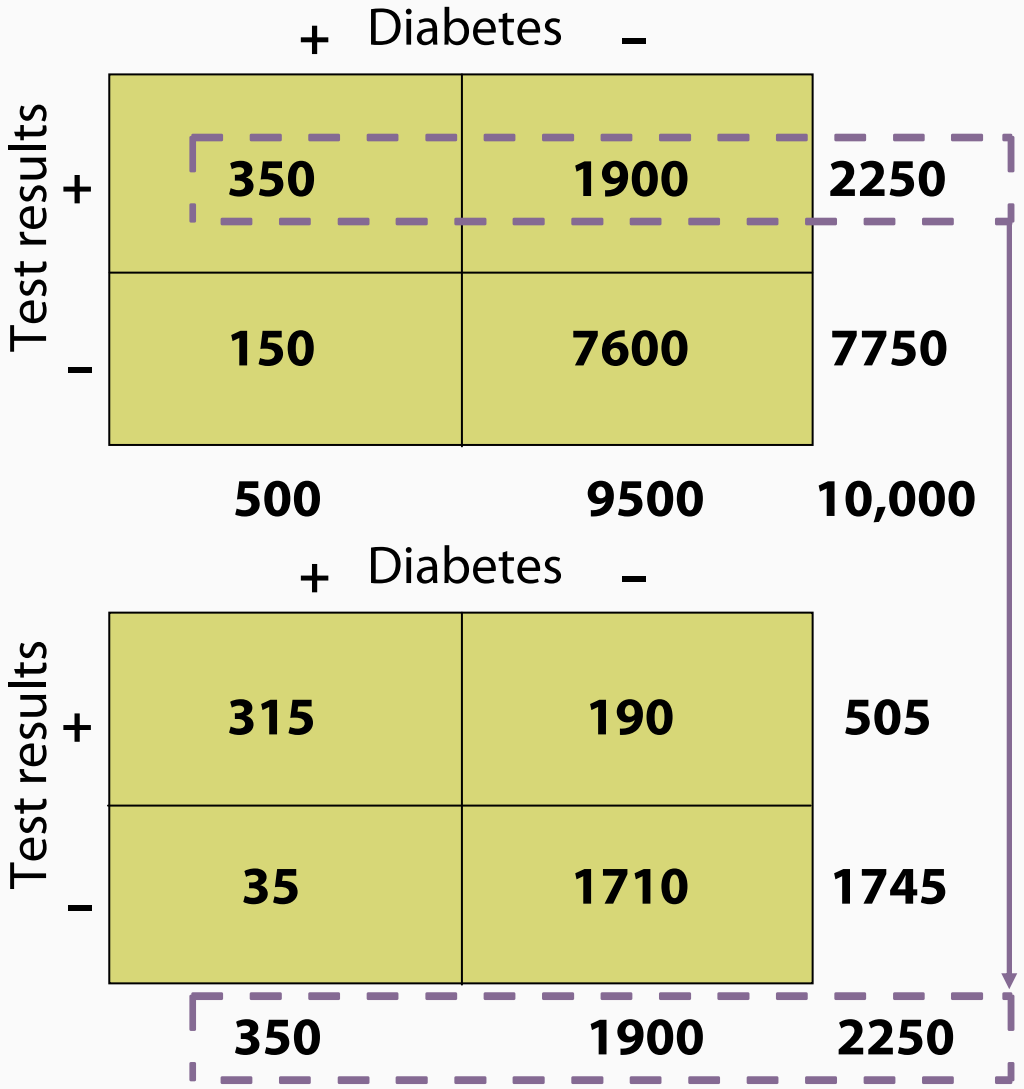
Example of a Two-Stage Screening Program: Test 1 (Blood Sugar)

- **Test 1 (blood sugar),** assume:
  - Disease prevalence = 5%, population = 10,000
  - Sensitivity = 70%, specificity = 80%
  - Screen **positives** from the first test

Test results	+ Diabetes	– Diabetes	
+	70% x 500=350	9500–7600=1900	**2250**
–	500–350=150	80% x 9500=7600	**7750**
	5% x 10,000 = 500	**9500**	**10,000**

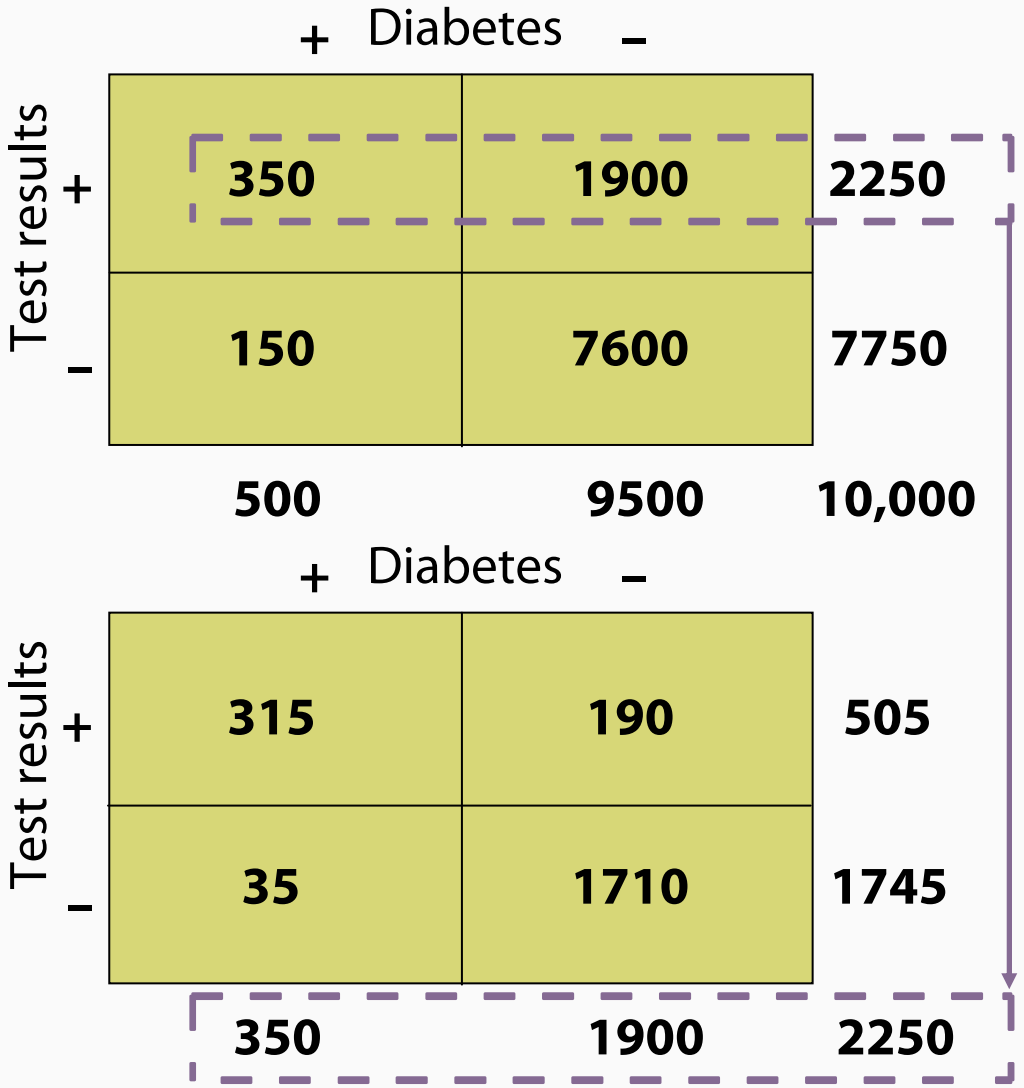
Example of a Two-Stage Screening Program: Test 2 (Glucose Tolerance Test)

- **Test 1 (blood sugar)**
  - Sensitivity = 70%
  - Specificity = 80%
- **Test 2 (glucose tolerance test)**
  - Sensitivity = 90%
  - Specificity = 90%



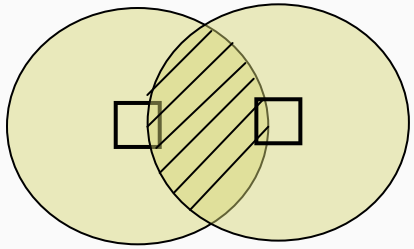
Example of a Two-Stage Screening Program: Test 2 (Glucose Tolerance Test)

- **Test 1 (blood sugar)**
  - Sensitivity = 70%
  - Specificity = 80%
- **Test 2 (glucose tolerance test)**
  - Sensitivity = 90%
  - Specificity = 90%
- Net sensitivity = $\frac{315}{500} = 63\%$
- New specificity = $\frac{7600+1710}{9500} = 98\%$



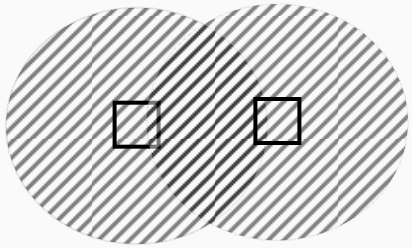
Two-Stage Screening: Re-Screen the Positives from the First Test

- Subject is disease positive when test positive in **both** tests



$A \cap B$

- Subject is disease negative when test negative in either test



$A \cup B$

Net Sensitivity in a Two-Stage Screening when Test + in the First Test Are Re-Screened

The multiplication rule of probability is

$$P(A \cap B) = P(B) * P(A \mid B)$$

When events are independent (two tests are independent), then

$$P(A \mid B) = P(A)$$

thus

$$P(A \cap B) = P(A) * P(B)$$

- Net sensitivity = Sensitivity 1 x Sensitivity 2

Net Specificity in a Two-Stage Screening when Test + in the First Test Are Re-Screened

Use addition rule of probability

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- Net specificity = Spec1 + Spec2 – (Spec1 x Spec2)

Other Two-Stage Screening

- Screen the **negatives** from the first test to identify any missed true positives from the first test
  - Net sensitivity and net specificity calculation follows similar but different logical algorithms
  - What is the net effect of testing the negatives from the first test?
    - Find more true positives => net sensitivity will be higher than sensitivity from the individual tests
    - Also find more false positives => net specificity will be lower than specificity from the individual tests

Simultaneous Testing

- When two (or more) tests are conducted in parallel
- The goal is to maximize the probability that subjects with the disease (true positives) are identified (increase sensitivity)
- Consequently, more false positives are also identified (decrease specificity)

Simultaneous Testing: Calculate Net Sensitivity

- When two tests are used simultaneously, disease positives are defined as those who test positive **by either one test or by both tests**
- We use the addition rule of probability to calculate the net sensitivity

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- Net sensitivity = sens 1 + sens 2 – (sens 1 x sens 2)

Simultaneous Testing: Calculate Net Specificity

- When two tests are used simultaneously, disease negatives are defined as those who test negative by both tests
- We use the multiplication rule of probability to calculate the net specificity

$$P(A \cap B) = P(A) * P(B)$$

- Net specificity = specificity test 1 x specificity test 2

Example of a Simultaneous Testing

- In a population of 1000, the prevalence of disease is 20%
- Two tests (A and B) are used at the same time
- Test A has sensitivity of 80% and specificity of 60%
- Test B has sensitivity of 90% and specificity of 90%
- Calculate net sensitivity and net specificity from using Test A and Test B simultaneously

Net sensitivity = sens 1 + sens 2 – sens1 x sens 2
= 80% + 90% – (80% x 90%)
= 98%

Net specificity = spec 1 x spec 2
= 60% x 90%
= 54%

Net Gain and Net Loss

- In simultaneous testing, there is a net gain in sensitivity but a net loss in specificity, when compared to either of the tests used
- In sequential testing when positives from the first test are re-tested, there is a net loss in sensitivity but a net gain in specificity, compared to either of the tests used

Review

- Test A is known to have the following characteristics:
  - Sensitivity of 80%
  - Specificity of 90%
  - Cost of $15 per test
- *Suppose the following:*
  - Test A is used in a population of 10,000 to identify individuals who have the disease
  - The prevalence of the disease is 5%
- **What are the net sensitivity, net specificity, and cost per positive case when:** (1) Test A is used twice simultaneously and when (2) a single Test A is used first, and individuals who test positive with Test A are tested again with Test A (sequentially)

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section C

Predictive Values

Predictive Values

- **Positive predictive value (PPV)**
  - The proportion of patients who test positive who actually have the disease
- **Negative predictive value (NPV)**
  - The proportion of patients who test negative who are actually free of the disease
- Note: PPV and NPV are not fixed characteristics of the test

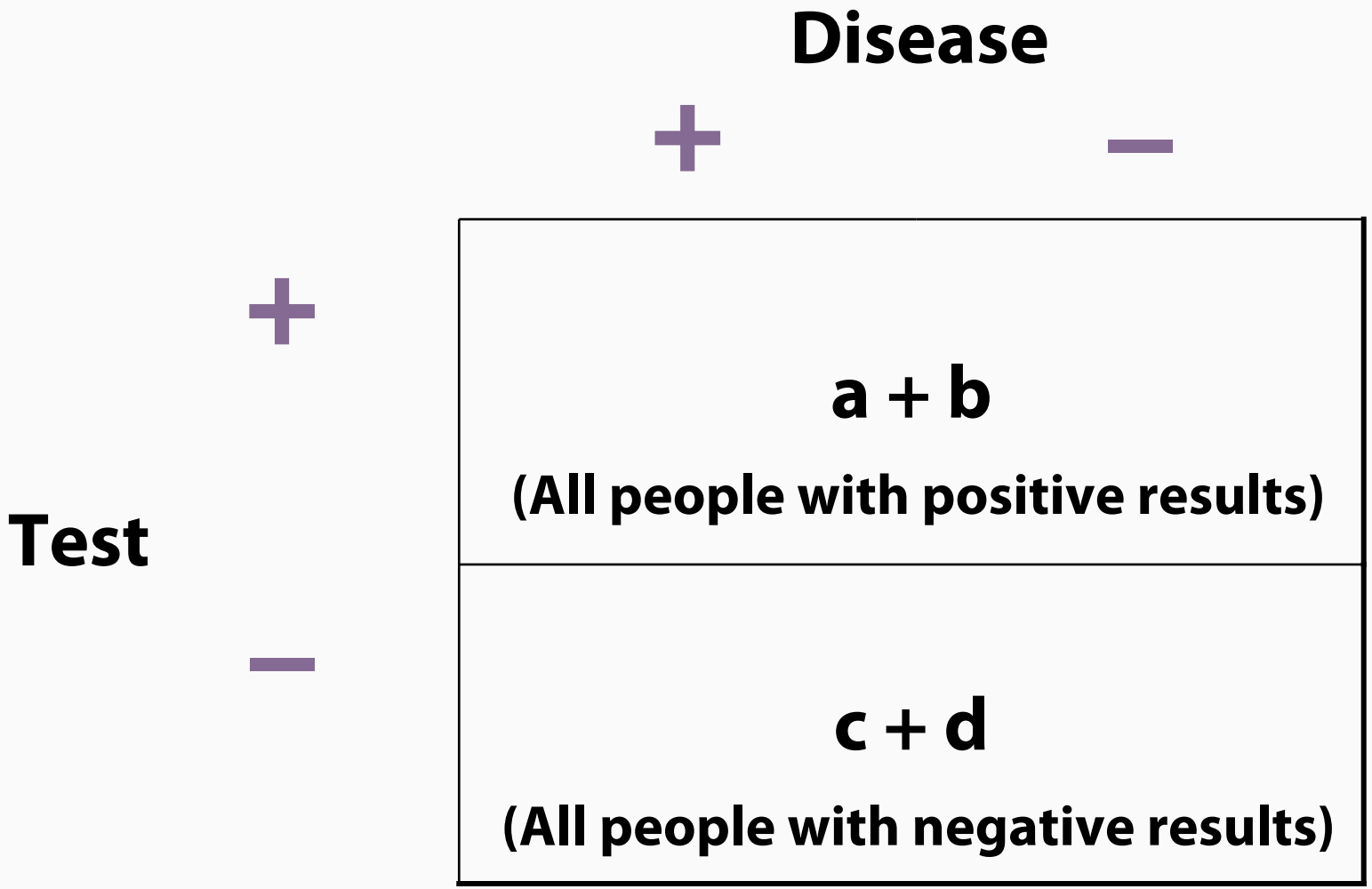
Another Interpretation of PPV

- If a person tests positive, what is the probability that he or she has the disease?
- (And if that person tests negative, what is the probability that he or she does not have the disease?)

Behind the Test Results

Test	Disease +	Disease −
+	**a** (True positives)	**b** (False positives)
−	**c** (False negatives)	**d** (True negatives)

What the Test Shows



Predictive Value

- Positive predictive value $= \dfrac{a}{a+b}$

$= \dfrac{\text{True Positives}}{\text{Test}+}$

$= P(D+ \mid T+)$

- Negative predictive value $= \dfrac{d}{c+d}$

$= \dfrac{\text{True Negatives}}{\text{Test}-}$

$= P(D- \mid T-)$

Applying Concept of Predictive Values to Screening Test

- Assume a population of 1,000 people
- 100 have a disease
- 900 do not have the disease
- A screening test is used to identify the 100 people with the disease
- The results of the screening appear in this table

Screening Results	True Characteristics in Population		Total
	Disease	No Disease	
Positive	80	100	180
Negative	20	800	820
Total	100	900	1,000

Calculating Predictive Values

Positive predictive value = 80/180 = 44%

Screening Results	True Characteristics in Population		Total
	Disease	No Disease	
Positive	80	100	180
Negative	20	800	820
Total	100	900	1,000

Negative predictive value = 800/820 = 98%

Calculating Predictive Values

Screening Results	True Characteristics in Population		Total
	Disease	No Disease	
Positive	80	100	180
Negative	20	800	820
Total	100	900	1,000

PPV Primarily Depends On . . .

- The prevalence of the disease in the population tested, and the test itself (sensitivity and specificity)
  - In general, it depends more on the specificity (and less on the sensitivity) of the test (if the disease prevalence is low)

PPV Formula

- For those who are interested

$$PPV = \frac{\text{sensitivity} \times \text{prevalence}}{(\text{sensitivity} \times \text{prevalence}) + (1 - \text{specificity}) \times (1 - \text{prevalence})}$$

$$NPV = \frac{\text{specificity} \times (1 - \text{prevalence})}{[(\text{specificity} \times (1 - \text{prevalence})] + [(1 - \text{sensitivity}) \times \text{prevalence}]}$$

- Use Bayes' theorem

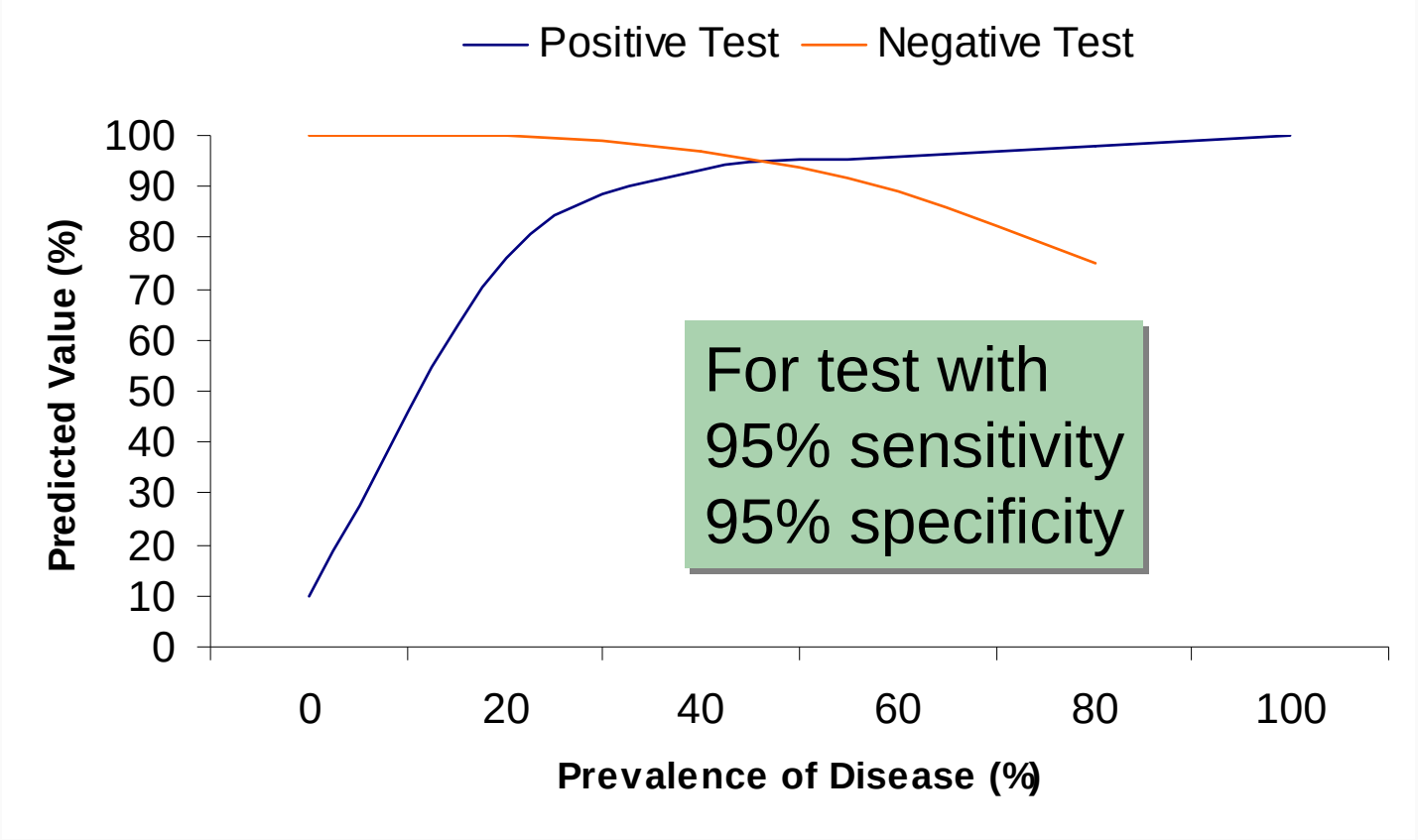
Calculation of PPV and NPV

- Construct the table and use the definition to guide the calculation of PPV and NPV
- [Or, use the formula]
- In a multiple testing situation, PPV and NPV are calculated for each test (not for the combined test)

Relationship of Disease Prevalence to Predictive Value

Example: Sensitivity = 99%; Specificity = 95%					
Disease Prevalence	Test Results	Sick	Not Sick	Totals	Positive Predictive Value
1%	+	99	495	594	$\frac{99}{594} = 17\%$
	–	1	9,405	9,406	
	Totals	100	9,900	10,000	
5%	+	495	475	970	$\frac{495}{970} = 51\%$
	–	5	9,025	9,303	
	Totals	500	9,500	10,000	

Prevalence of Disease



Adapted from Mausner JS, Kramer S.Epidemiology: an Introductory Text. Philadelphia, WB Saunders 1985, p221.

So If a Person Tests Positive . . .

- The probability that he or she has the disease depends on the prevalence of the disease in the population tested and the validity of the test (sensitivity and specificity)
- In general, specificity has more impact on predictive values

The Relationship of Specificity to Predictive Value

Test	Disease +	Disease −	
+	250	250	500
−	250	250	500
	500	500	1,000

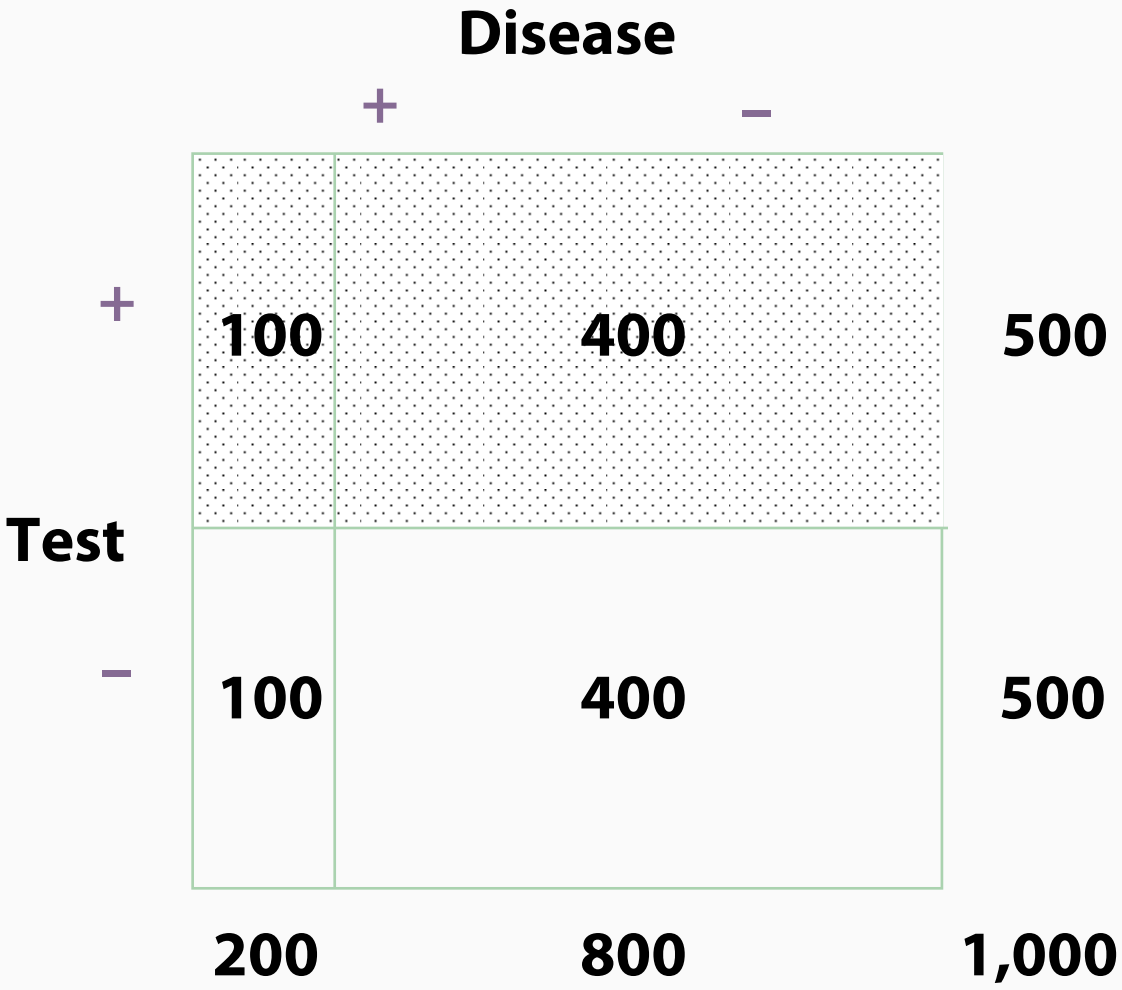
Prevalence = 50%

Sensitivity = 50%

Specificity = 50%

$$PPV = \frac{250}{500} = 50\%$$

The Relationship of Specificity to Predictive Value



Prevalence = 20%

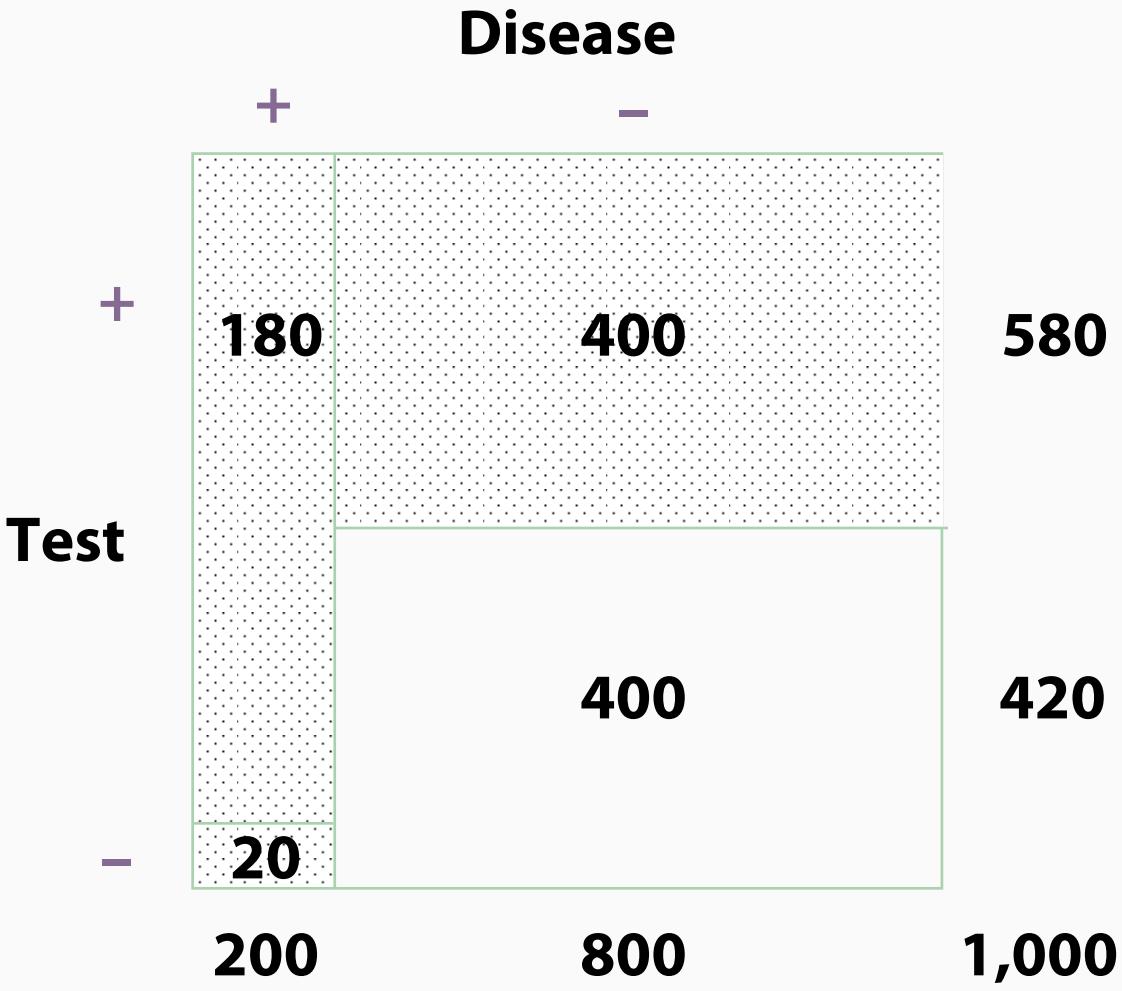
Sensitivity = 50%

Specificity = 50%

$$PPV = \frac{100}{500} = 20\%$$

Change prevalence

The Relationship of Specificity to Predictive Value



Prevalence = 20%

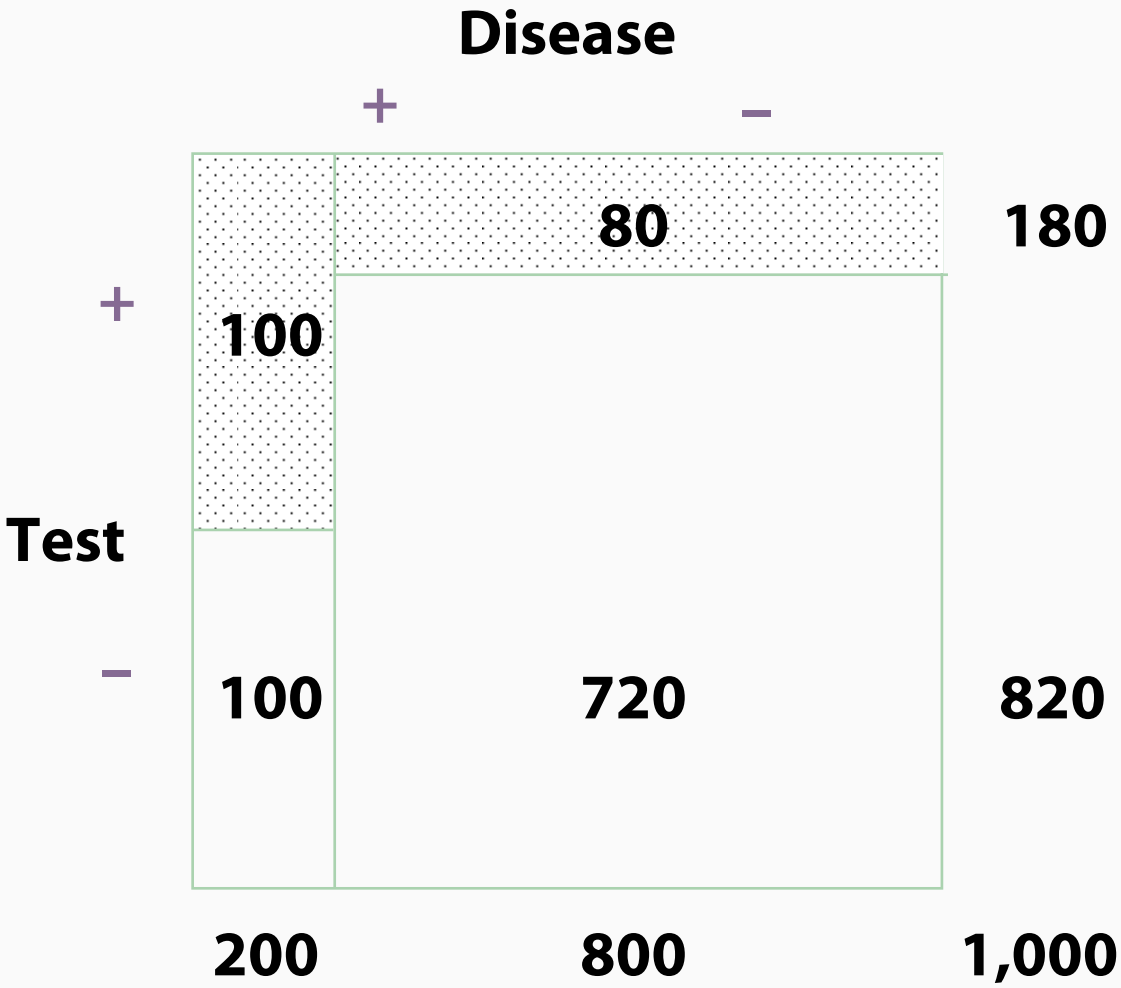
Sensitivity = 90%

Specificity = 50%

$$PPV = \frac{180}{580} = 31\%$$

Change sensitivity

The Relationship of Specificity to Predictive Value



Prevalence = 20%

Sensitivity = 50%

Specificity = 90%

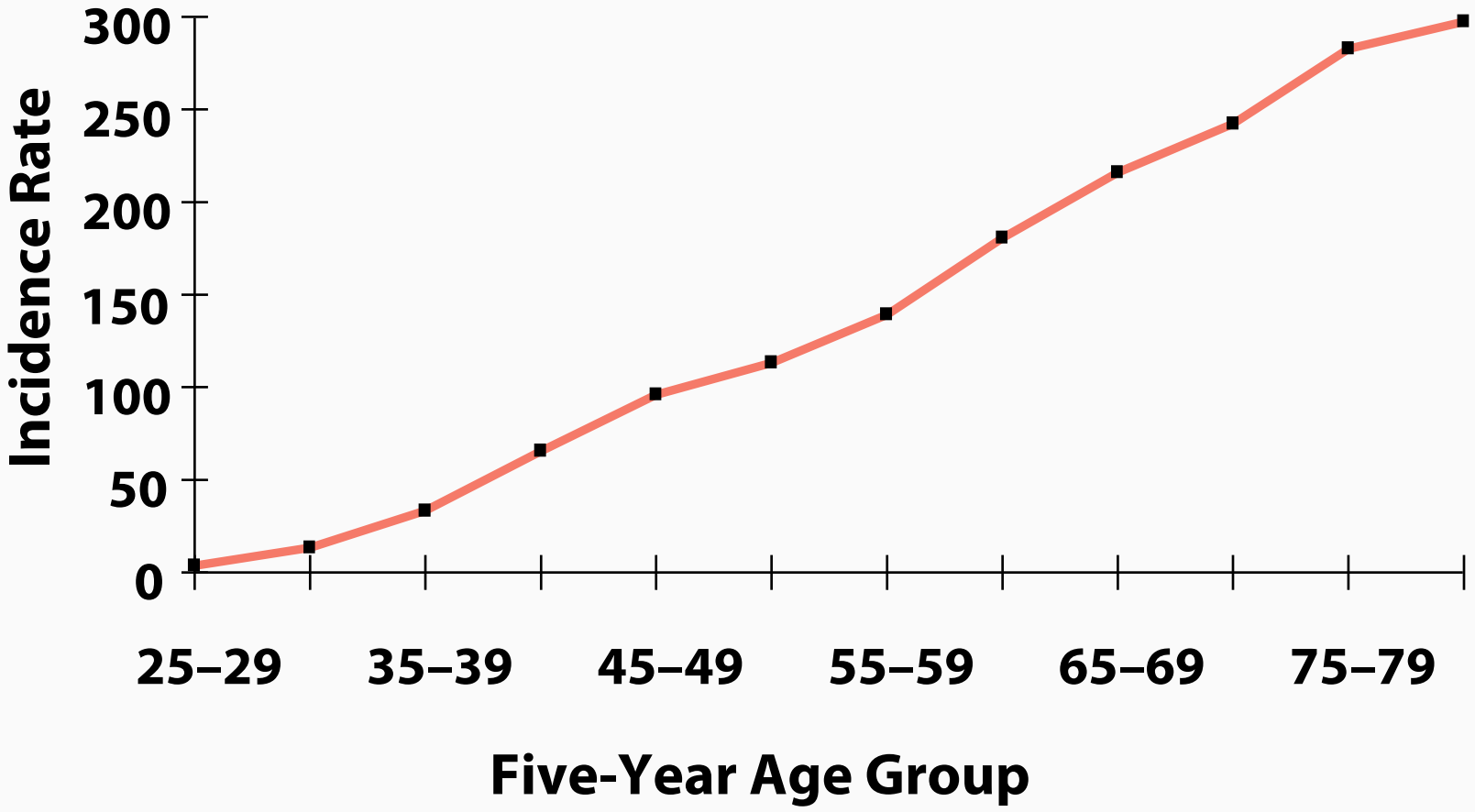
$$PPV = \frac{100}{180} = 56\%$$

Restore sensitivity
Change specificity

More impact than changing sensitivity

Age-Specific Breast Cancer Incidence Rates U.S., All Races (SEER 1984-88)

Rates per 100,000 Population of the Specified Five-year Age Group



Source: SEER Program.

Results of First Screening Mammography by Age Group — UCSF Mobile Mammography Program

Age (Years)	Cancer Detected	No Cancer Detected	Total Abnormal	Positive Predictive Value
30–39	9	273	282	3%
40–49	26	571	597	4%
50–59	30	297	327	9%
60–69	46	230	276	17%
70	26	108	134	19%

PPV of First Screening Mammography by Age and Family History of Breast Cancer

Age (Years)	Women without a Family History of Breast Cancer	Women with a Family History of Breast Cancer
30–39	3%	4%
40–49	4%	13%
50–59	9%	22%
60–69	17%	14%
70	19%	24%

Consequence of Different PPVs

Age	< 50 Years	≥ 50 Years
Positive predictive value	4%	14%

Source: Kerlikowske et al. (1993). *JAMA*, 270:2444–2450.

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section D

Reliability (Repeatability)

Reproducibility, Repeatability, Reliability

- **Reproducibility, repeatability, reliability** all mean that the results of a test or measure are identical or closely similar each time it is conducted
- Because of variation in laboratory procedures, observers, or changing conditions of test subjects (such as time, location), a test may not consistently yield the same result when repeated
- Different types of variation
  - Intra-subject variation
  - Intra-observer variation
  - Inter-observer variation

Intra-Subject Variation

- **Intra-subject variation** is a variation in the results of a test conducted over (a short period of) time on the same individual
- The difference is due to the changes (such as physiological, environmental, etc.) occurring to that individual over that time period

Variation in Blood Pressure Readings: A 24-Hour Period

Blood Pressure (mm Hg)	Female 27 Years Old	Female 62 Years Old	Male 33 Years Old
Basal	110/70	132/82	152/109
Lowest hour	86/47	102/61	123/78
Highest hour	126/79	172/94	153/107
Casual	108/64	155/93	157/109

Inter-Observer and Intra-Observer Variation

- **Inter-observer variation** is a variation in the result of a test due to multiple observers examining the result (inter = between)
- **Intra-observer variation** is a variation in the result of a test due to the same observer examining the result at different times (intra = within)
- The difference is due to the extent to which observer(s) agree or disagree when interpreting the same test result

Agreement between Two Observers (Or Two Observations)

- A perfect agreement occurs when:
  - b=0
  - c=0

		Observer 1: Positive	Observer 1: Negative
Observer 2	Positive	a	b
	Negative	c	d

Percent Agreement

$$\text{Overall Percent Agreement} = \frac{a + d}{a + b + c + d} \times 100$$

$$\text{Percent Positive Agreement} = \frac{a}{a + b + c} \times 100$$

Note: This is a conditional probability

Example

		Radiologist A: Positive	Radiologist A: Negative
Radiologist B	Positive	4	5
	Negative	2	6

$$\text{Overall Percent Agreement} = \frac{4 + 6}{4 + 5 + 2 + 6} \times 100 = 58.8\%$$

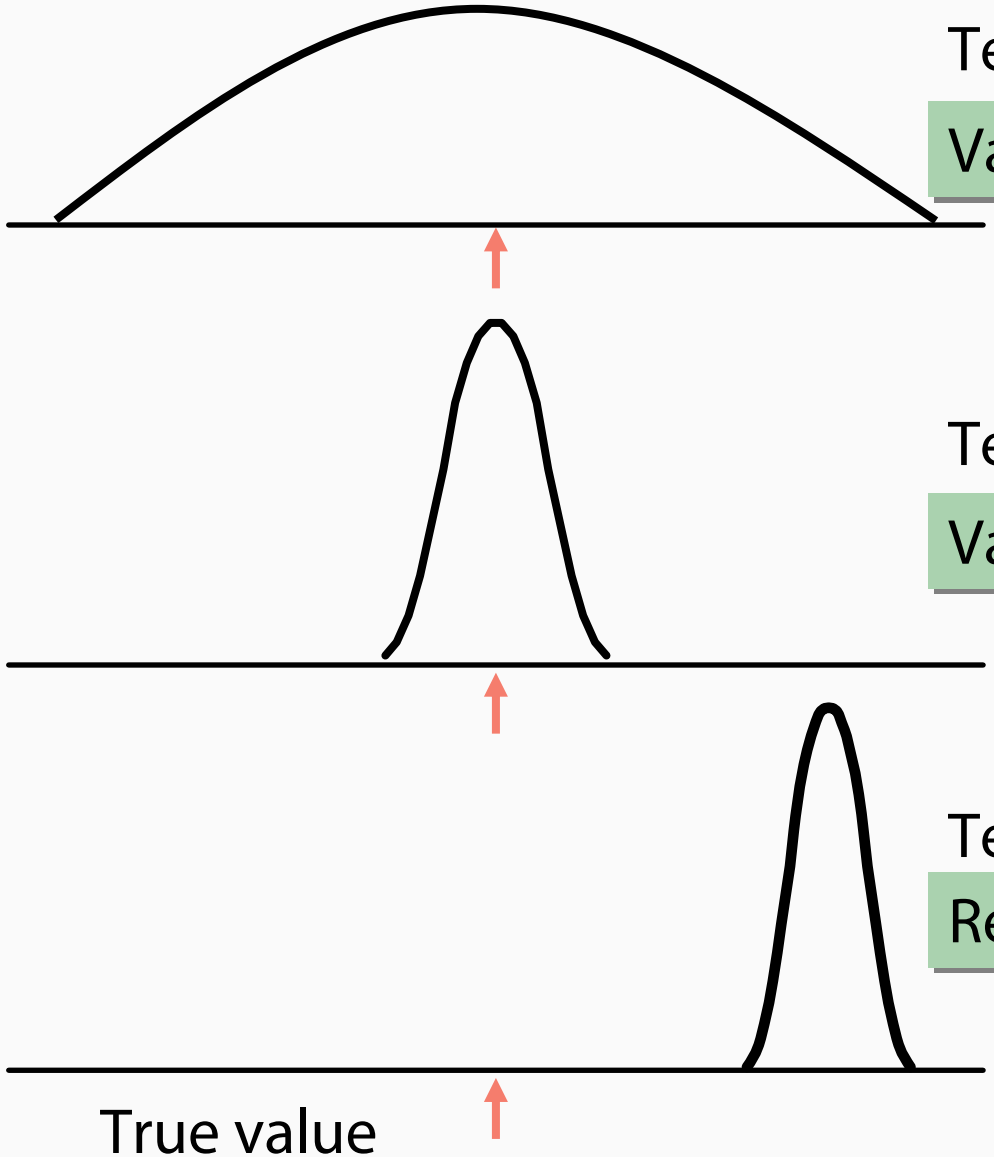
$$\text{Percent Positive Agreement} = \frac{4}{4 + 5 + 2} \times 100 = 36.4\%$$

Observer or Instrument Variation: Overall Percent Agreement

Reading #2	Reading #1			
	Abnormal	Suspect	Doubtful	Normal
Abnormal	A	B	C	D
Suspect	E	F	G	H
Doubtful	I	J	K	L
Normal	M	N	O	P

$$\text{Percent agreement} = \frac{A+F+K+P}{\text{Total}} \times 100$$

Outcome of Test Results



Review

- Define
  - Overall percent agreement
  - Percent positive agreement
- Contrast overall percent agreement and percent positive agreement