SENSITIVITY, SPECIFICITY, PPV, NPV

And the matrix of confusion

Goals for this lecture

- Understand the use of measure like Sensitivity and Specificity in validating tests/models etc
- How the setting matters – screening, diagnosis, prediction,prognosis
- To understand how to work out sens/spec/ppv/npv from a basic 2x2 table
- To understand the interpretation and limitations of these measures
- Errors in classification – in basic Epi and Machine Learning
- The relationship between sens/spec and classification AUC/ROC

Screening

Screening is the process in which we use a test to determine whether an individual likely has a particular health indicator or not or is likely to develop a particular health indicator or not.

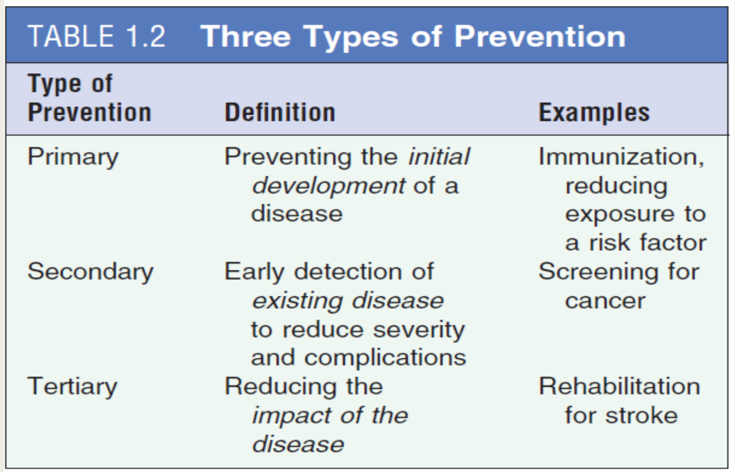
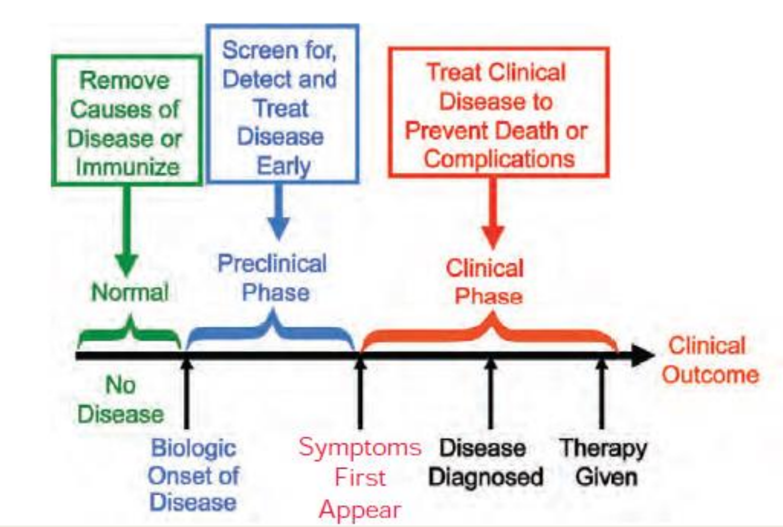


TABLE 1.2 **Three Types of Prevention**

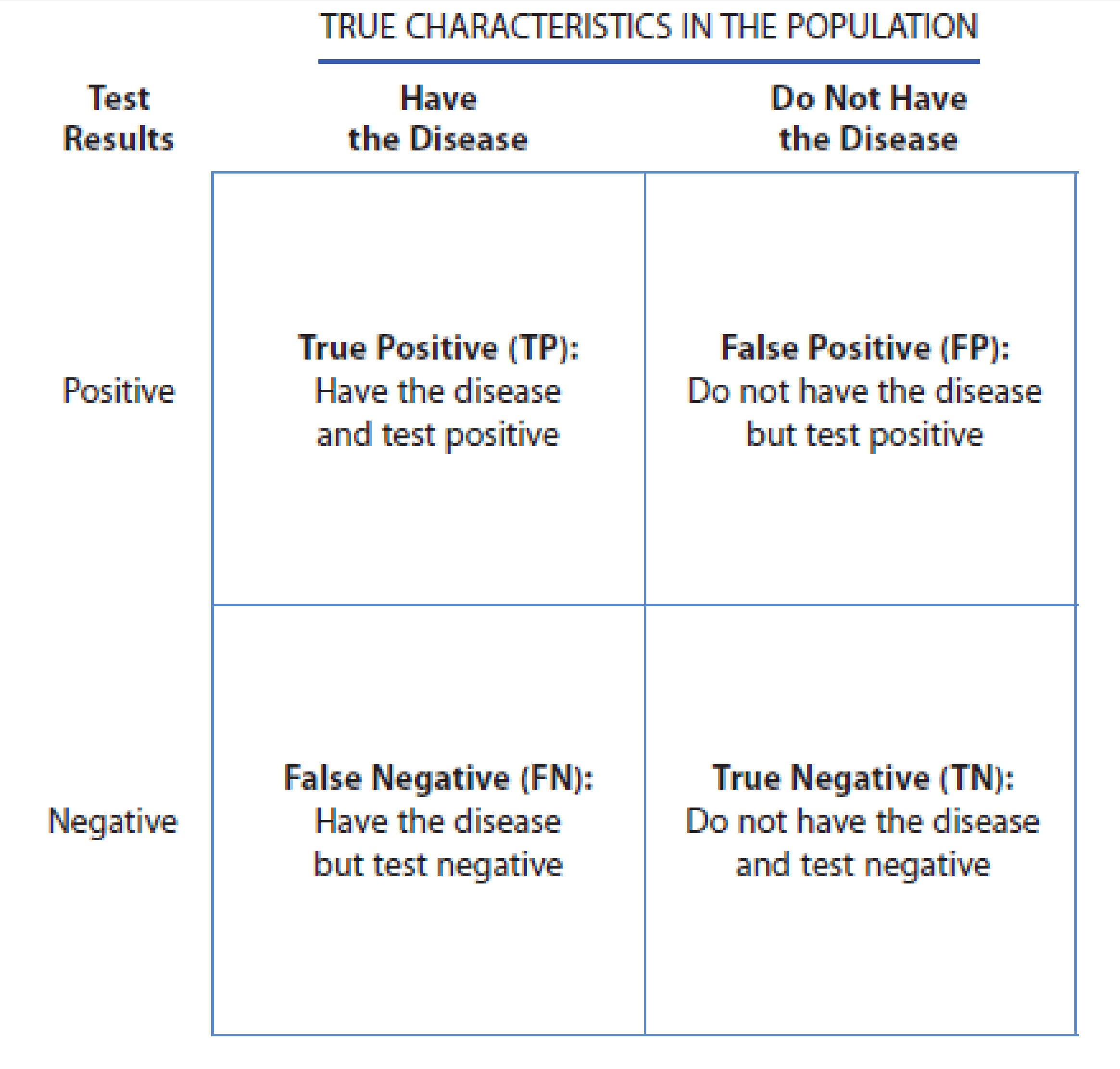
Type of Prevention	Definition	Examples
Primary	Preventing the *initial development* of a disease	Immunization, reducing exposure to a risk factor
Secondary	Early detection of *existing disease* to reduce severity and complications	Screening for cancer
Tertiary	Reducing the *impact of the disease*	Rehabilitation for stroke

Screening is <u>**not**</u> the same as diagnosis; screening tests give us information about whether the disease is <u>**likely**</u> to be present

Screening can ideally detect the presence of the disease **earlier** so that so that early detection can improve the lives of affected by any available treatment.


Remove Causes of Disease or Immunize
Screen for, Detect and Treat Disease Early
Treat Clinical Disease to Prevent Death or Complications
Normal
Preclinical Phase
Clinical Phase
Clinical Outcome
No Disease
Biologic Onset of Disease
Symptoms First Appear
Disease Diagnosed
Therapy Given

2 by 2 convention in epidemiology



	TRUE CHARACTERISTICS IN THE POPULATION	
Test Results	**Have the Disease**	**Do Not Have the Disease**
Positive	**True Positive (TP):** Have the disease and test positive	**False Positive (FP):** Do not have the disease but test positive
Negative	**False Negative (FN):** Have the disease but test negative	**True Negative (TN):** Do not have the disease and test negative

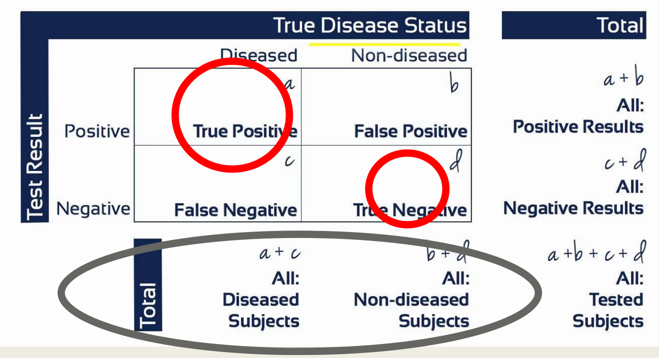
- 2 x 2 table in epidemiology
- Often called a confusion matrix
- "True" characteristics of population
- Convention for positive results to be in the top left corner

Two different perspectives: First perspective (A)

A1. Sensitivity (Se)

A2. Specificity (Sp)

Almost like inventor’s perspective.
“How well did the new screening invention do in terms of identifying diseased and non-diseased individuals?”

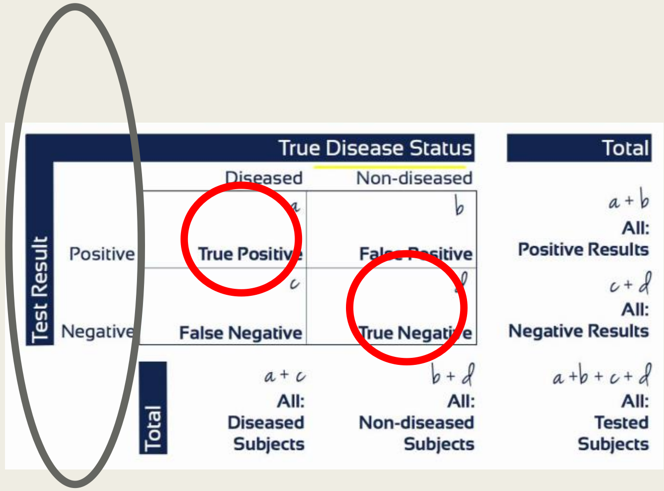


		True Disease Status		Total
		Diseased	Non-diseased	
Test Result	Positive	a True Positive	b False Positive	a + b All: Positive Results
	Negative	c False Negative	d True Negative	c + d All: Negative Results
Total		a + c All: Diseased Subjects	b + d All: Non-diseased Subjects	a + b + c + d All: Tested Subjects

B1. Positive predictive value (PPV)

B2. Negative predictive value (NPV)

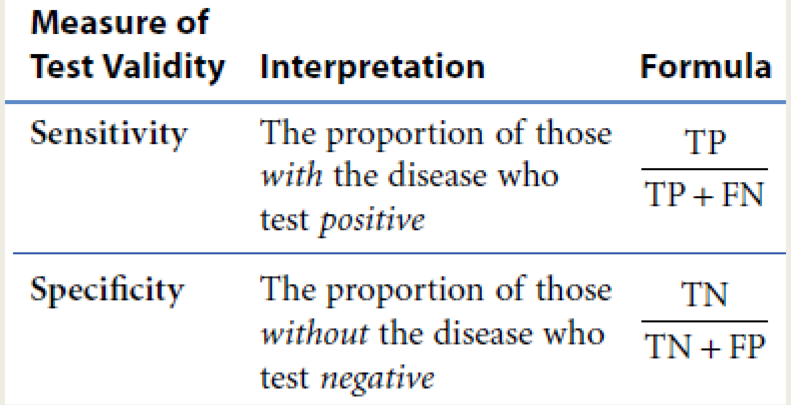
“Patient perspective.
What is the probability that I actually have the disease when I informed that I am screen positive?”



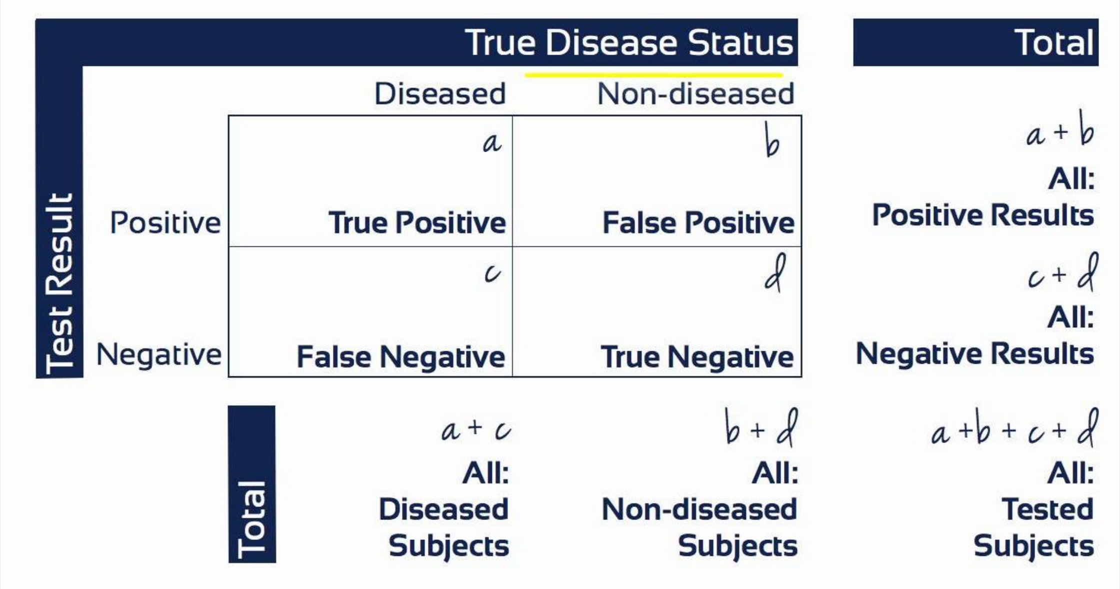
		True Disease Status		Total
		Diseased	Non-diseased	
Test Result	Positive	a True Positive	b False Positive	a + b All: Positive Results
	Negative	c False Negative	d True Negative	c + d All: Negative Results
Total		a + c All: Diseased Subjects	b + d All: Non-diseased Subjects	a + b + c + d All: Tested Subjects

Story starts from Blue. Your desire answer in Red.

Validity of screening test (Sensitivity/Specificity)



Measure of Test Validity	Interpretation	Formula
Sensitivity	The proportion of those *with* the disease who test *positive*	$\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$
Specificity	The proportion of those *without* the disease who test *negative*	$\frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$

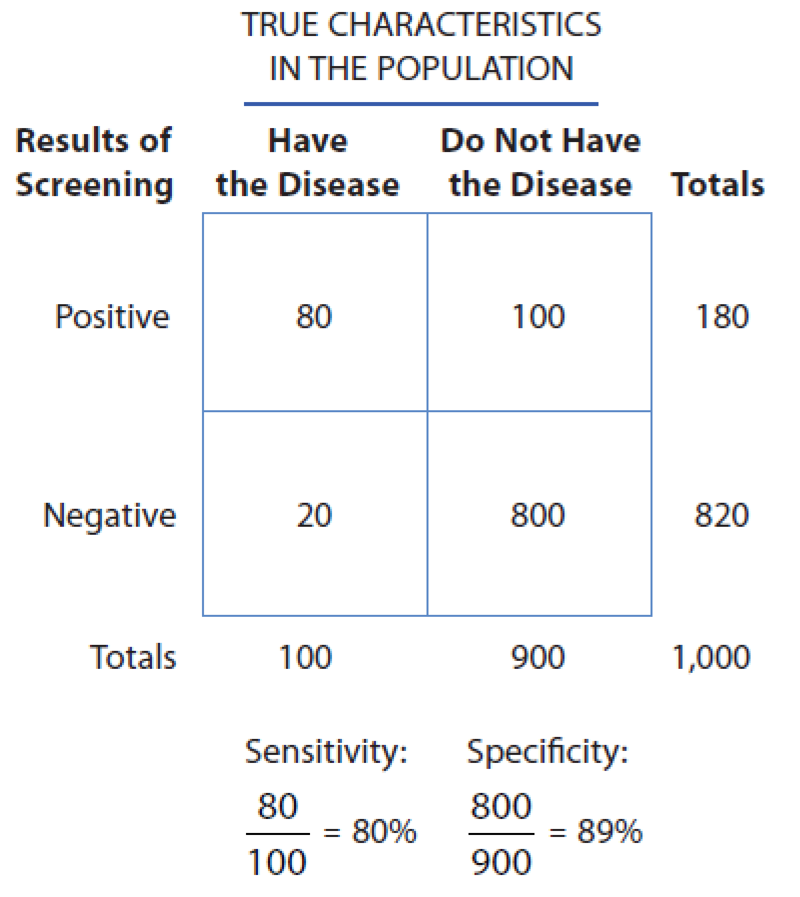


		True Disease Status		Total
		Diseased	Non-diseased	
Test Result	Positive	a True Positive	b False Positive	a + b All: Positive Results
	Negative	c False Negative	d True Negative	c + d All: Negative Results
Total		a + c All: Diseased Subjects	b + d All: Non-diseased Subjects	a +b + c + d All: Tested Subjects

$$\mathrm{Se} = \frac{a}{a+c} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{Sp} = \frac{d}{b+d} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

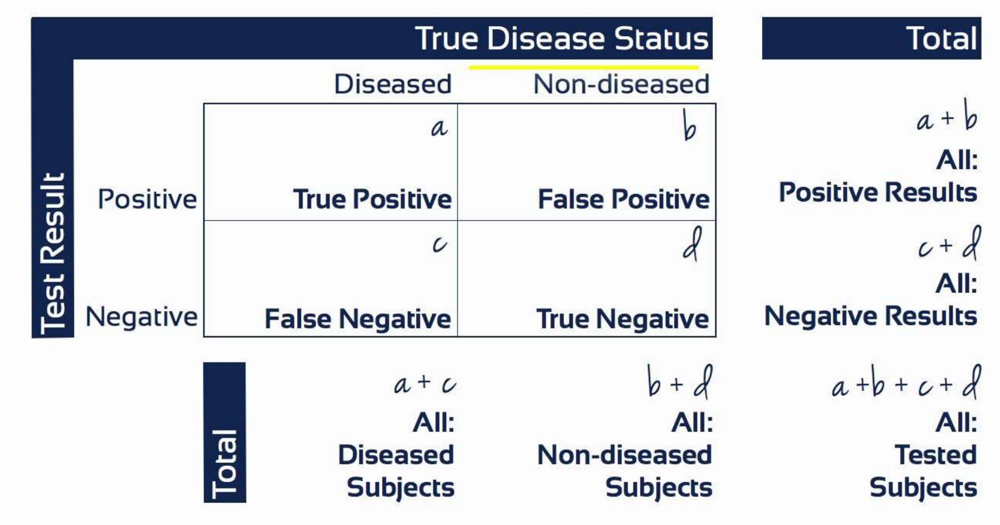
Calculation: Validity of screening test (Sensitivity/Specificity)



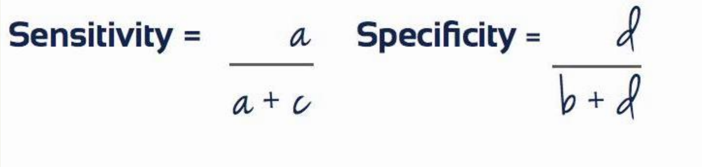
	TRUE CHARACTERISTICS IN THE POPULATION		
Results of Screening	Have the Disease	Do Not Have the Disease	Totals
Positive	80	100	180
Negative	20	800	820
Totals	100	900	1,000

Sensitivity: $\frac{80}{100} = 80\%$

Specificity: $\frac{800}{900} = 89\%$



Test Result	True Disease Status: Diseased	True Disease Status: Non-diseased	Total
Positive	a True Positive	b False Positive	a + b All: Positive Results
Negative	c False Negative	d True Negative	c + d All: Negative Results
Total	a + c All: Diseased Subjects	b + d All: Non-diseased Subjects	a + b + c + d All: Tested Subjects

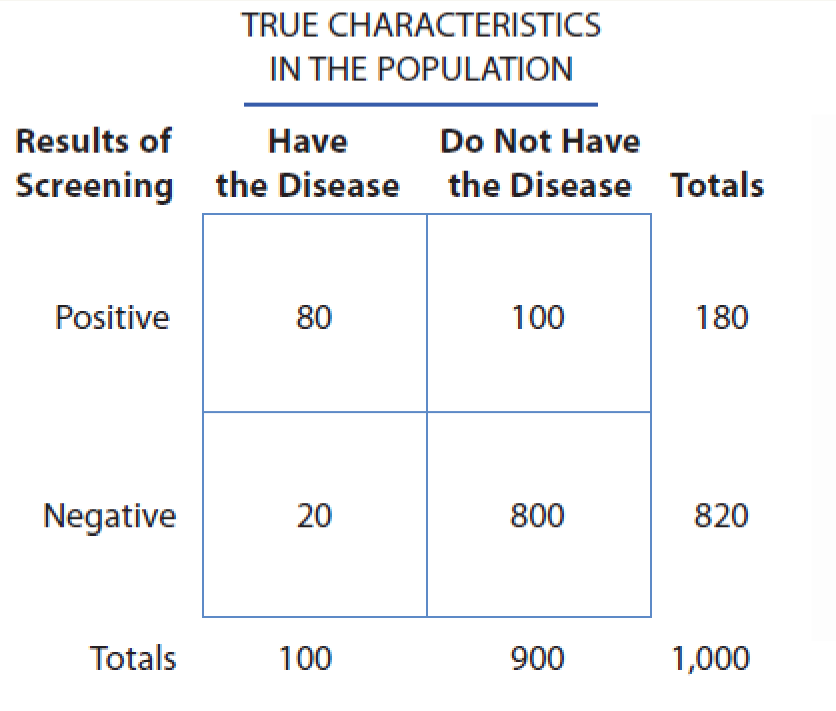


Sensitivity = $\frac{a}{a + c}$ **Specificity** = $\frac{d}{b + d}$

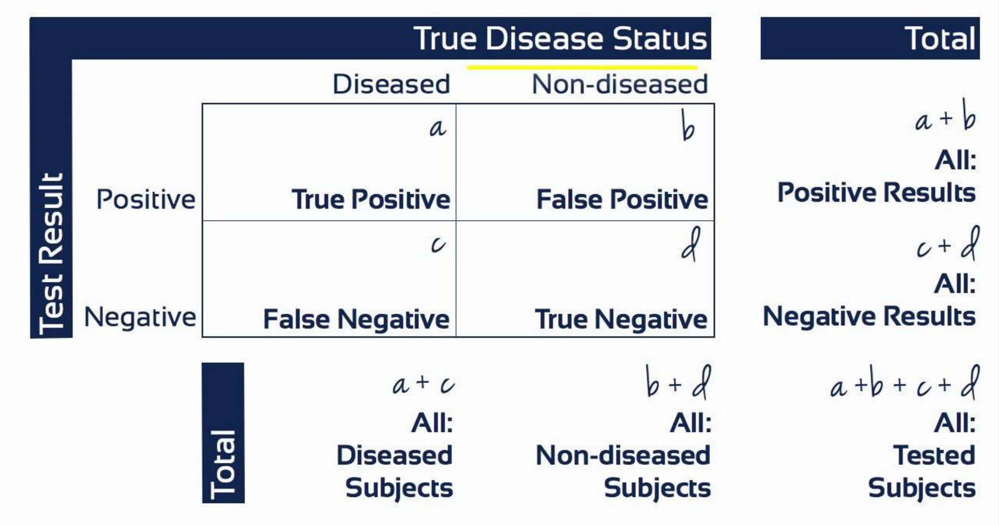
Sensitivity: 80/100 or 80% of diseased people were correctly identified as positive by the screening test.

Specificity: 800/900 or 89% of non-diseased people were correctly identified as negative by the screening test

False Negative and False Positive



Results of Screening	TRUE CHARACTERISTICS IN THE POPULATION Have the Disease	 Do Not Have the Disease	Totals
Positive	80	100	180
Negative	20	800	820
Totals	100	900	1,000



		True Disease Status		Total
		Diseased	Non-diseased	
Test Result	Positive	a True Positive	b False Positive	a + b All: Positive Results
	Negative	c False Negative	d True Negative	c + d All: Negative Results
Total		a + c All: Diseased Subjects	b + d All: Non-diseased Subjects	a + b + c + d All: Tested Subjects

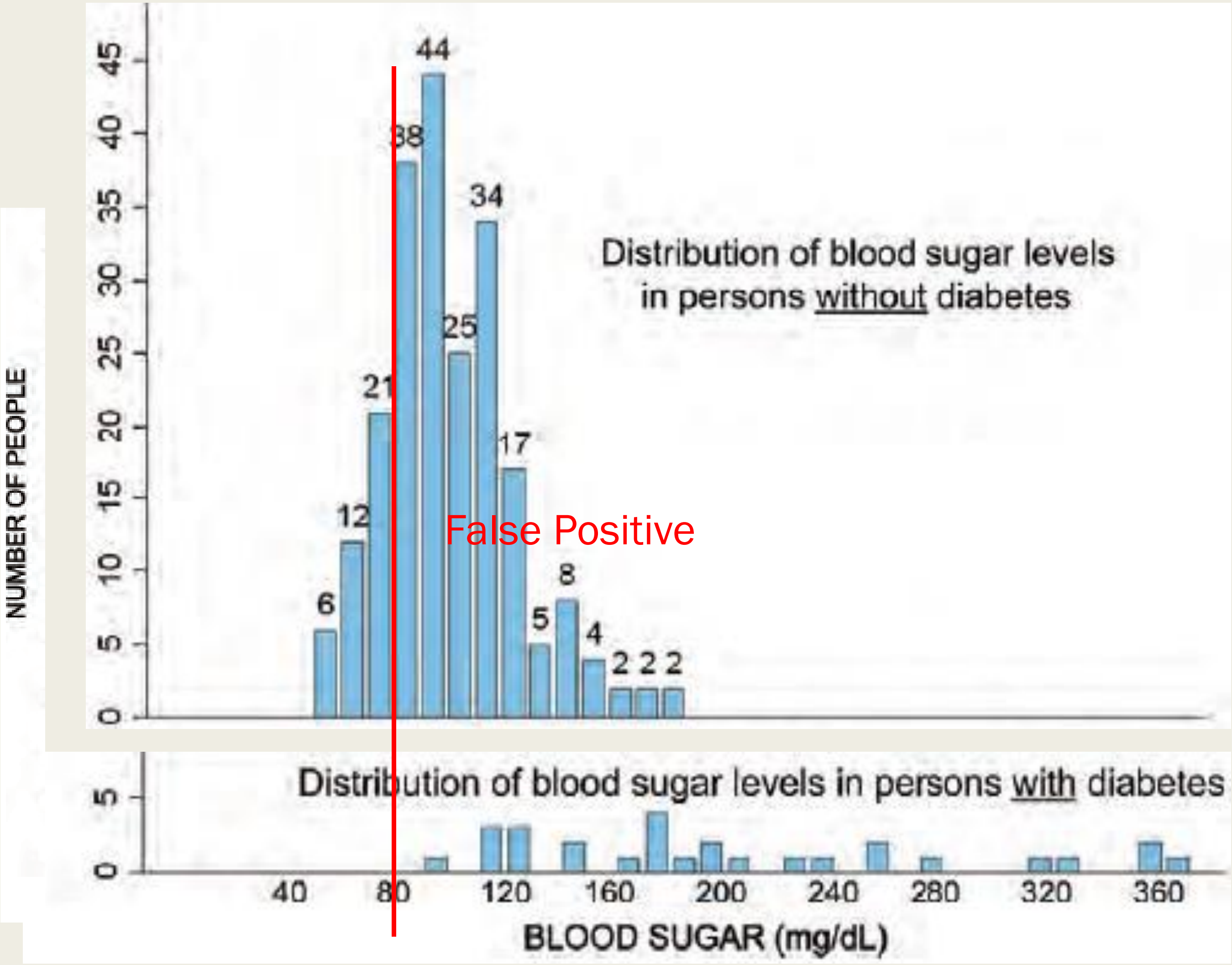
False negative: 20/100 or 20% of diseased people were incorrectly classified as "disease-negative" by the screening test

False positive: 100/900 or 11% of non-diseased people were incorrectly classified as "disease-positive" by the screening test

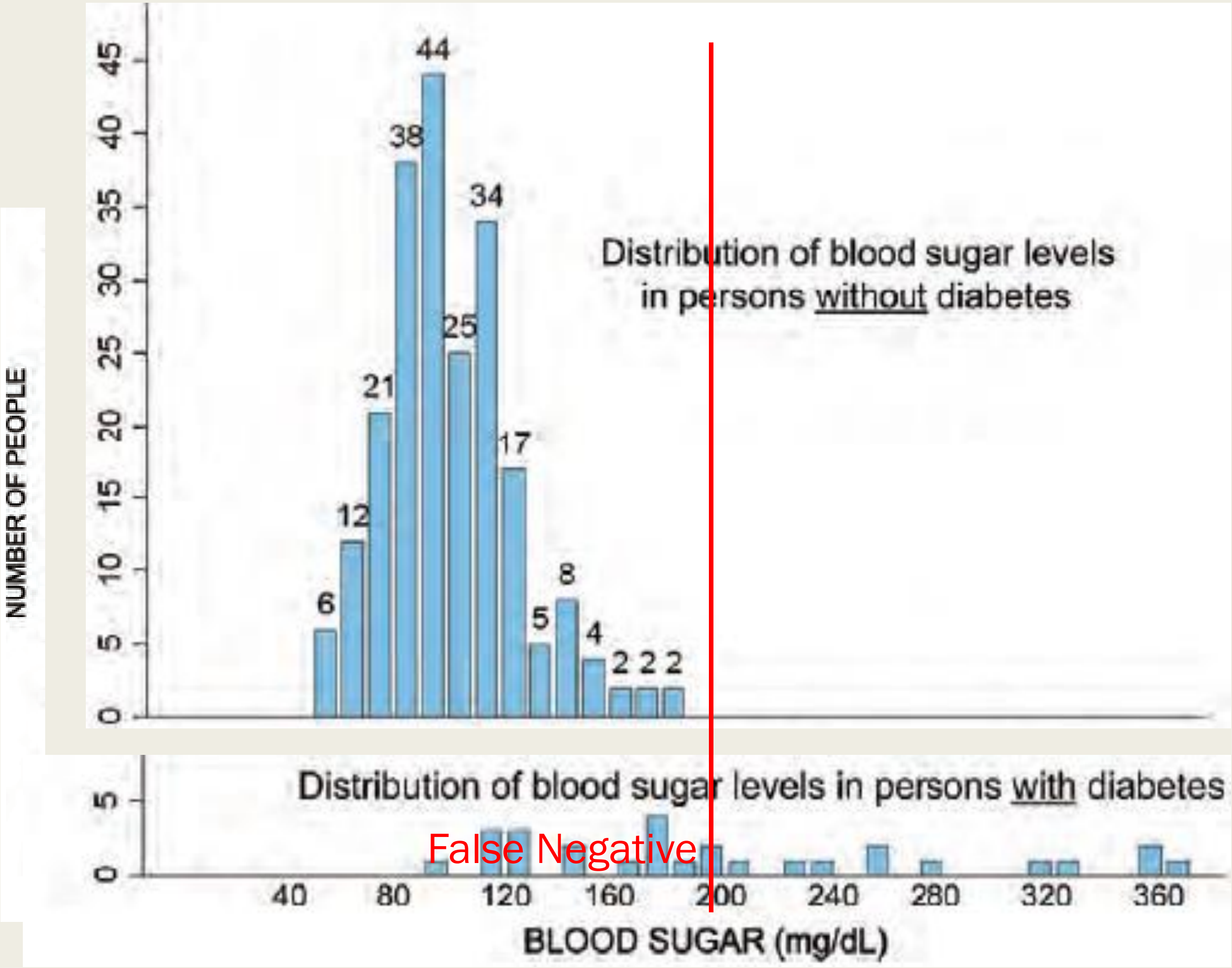
Trade off between Sensitivity and Specificity

- Higher sensitivity and specificity of screening tests are ideal.
- Unfortunately, there is often trade off between sensitivity and specificity that can influence false positive and false positive.
- So far we have discussed a test with only two possible results: positive or negative, but we often test for a <u>continuous</u> variable, such as blood pressure.
- In the following section, we will demonstrate trade off between sensitivity and specificity for screening test that involves deciding a **numeric cut-off** to establish screen positive and screen negative.

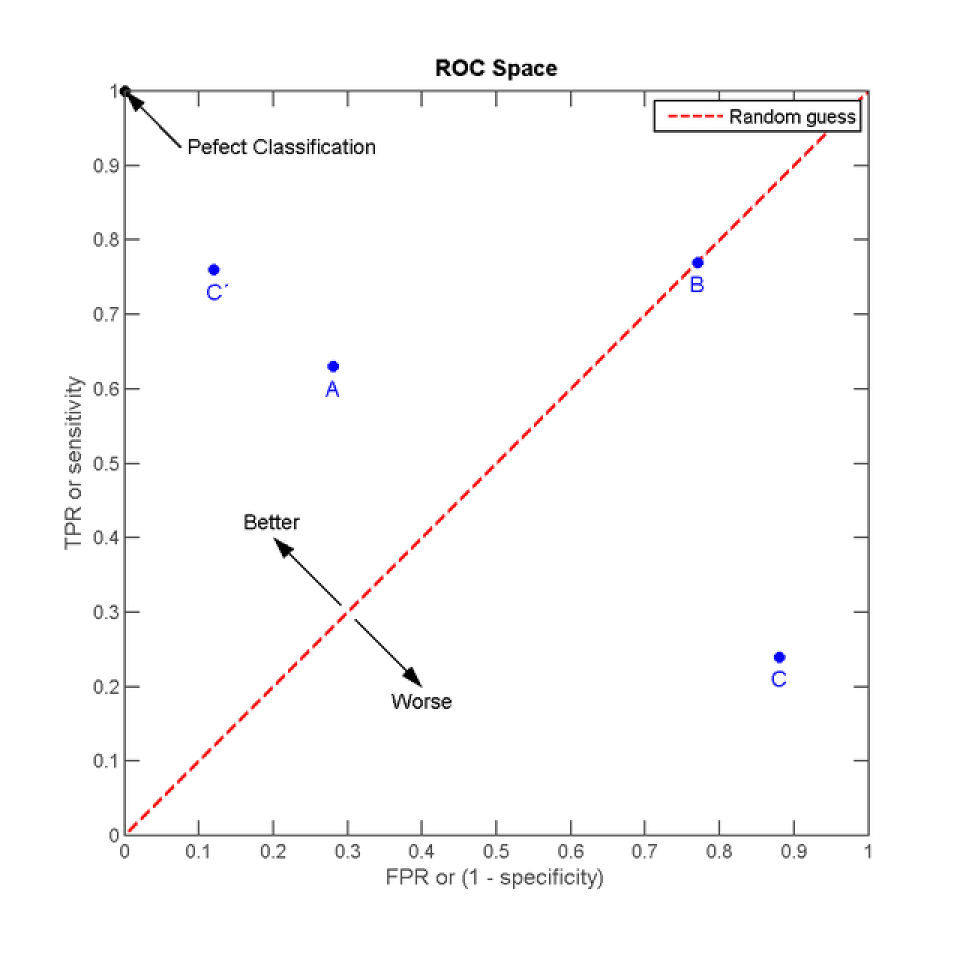
Screening test (Scenario #1): Continuous predictor for sensitivity

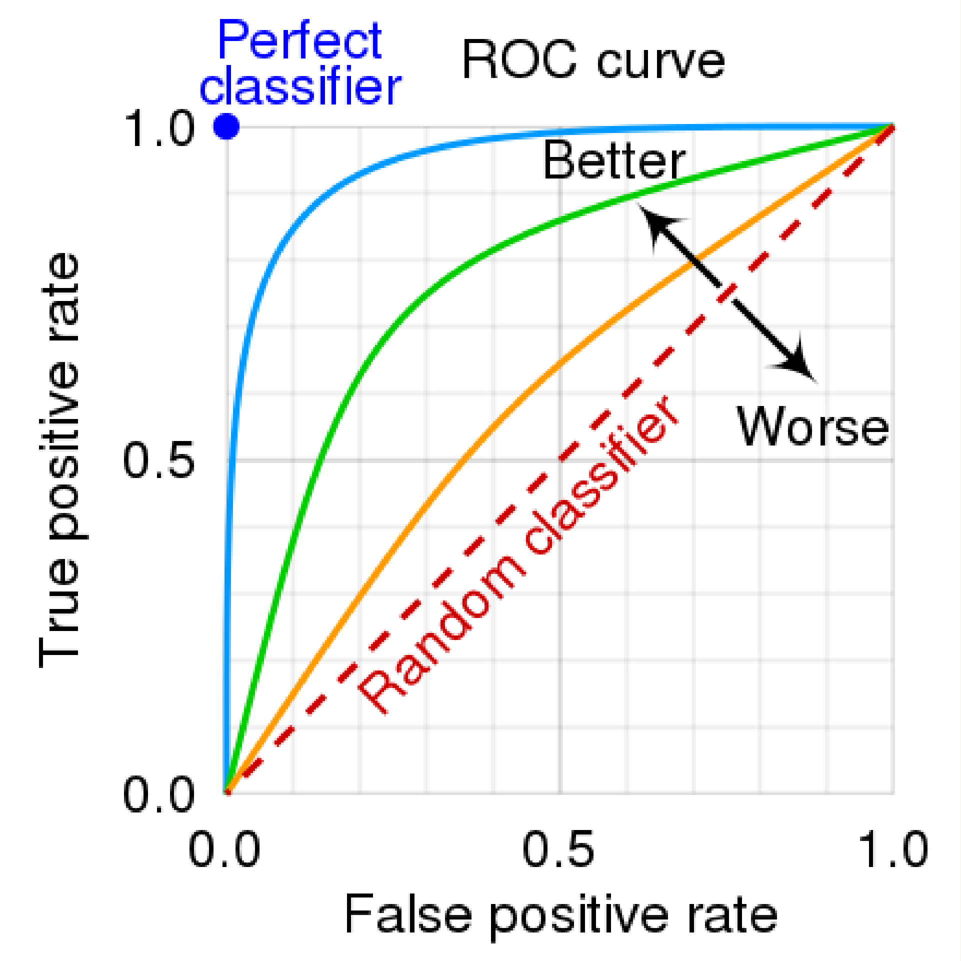


Screening test (Scenario #2): Continuous predictor for specificity



ROC Curve


ROC Space
Random guess
Pefect Classification
C´
A
B
C
Better
Worse
TPR or sensitivity
FPR or (1 - specificity)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Perfect classifier
ROC curve
Better
Worse
Random classifier
True positive rate
False positive rate
0.0
0.5
1.0

Implications of False Positive and Negative

- Choice of cut-off for screening also depends on the importance we placed on false positive and false negative.

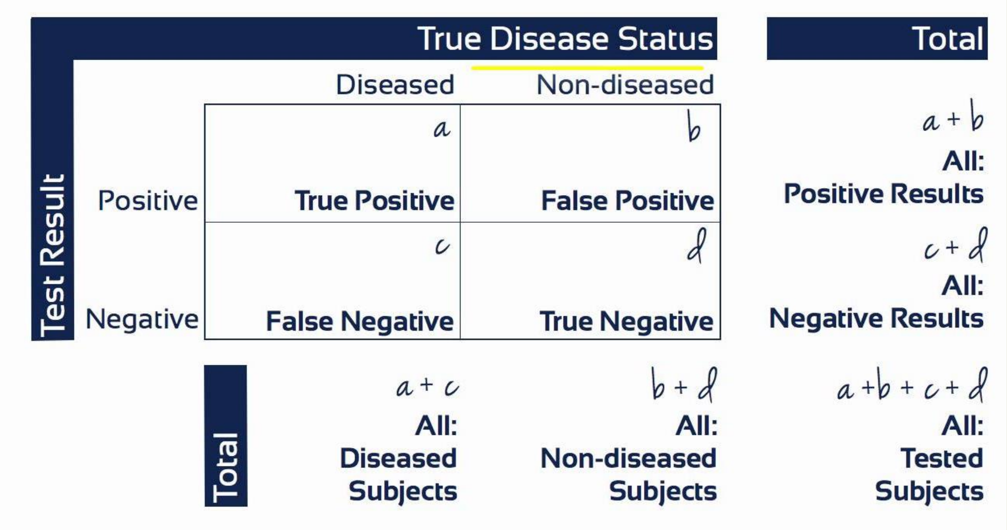
- <u>Issues of false positive</u>:

  All people who screened positive are brought back for more sophisticated and more expensive tests.

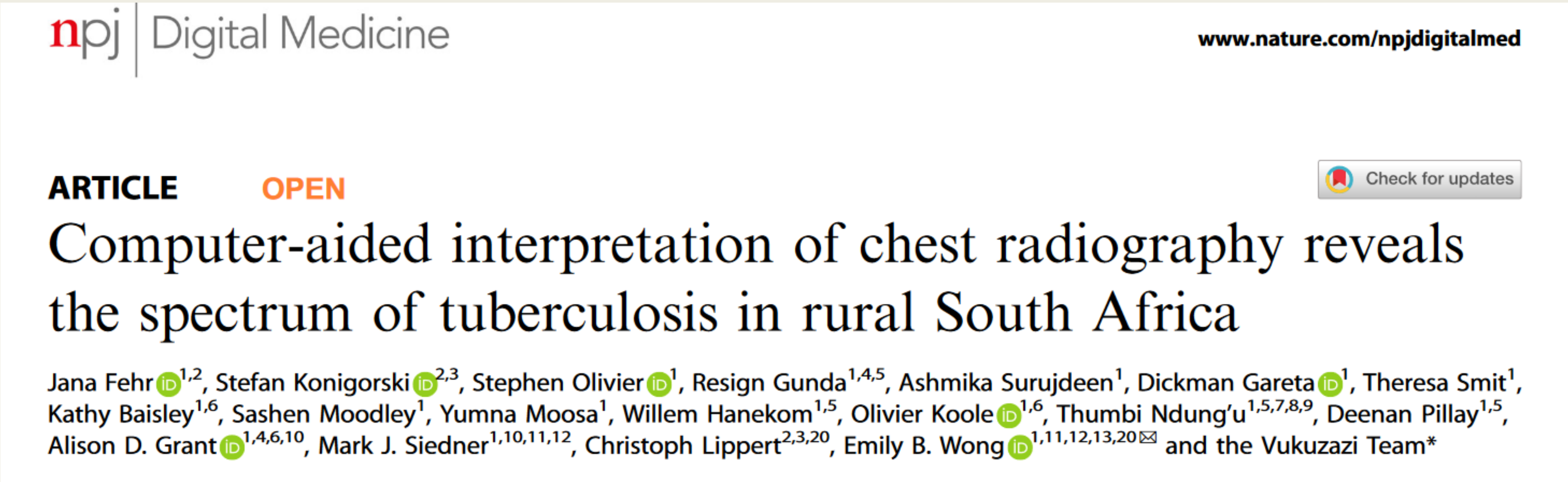
  - Burden on health care system
  - Anxiety and emotional cost

- <u>Issues of false negative</u>:

  – *Treatment delay for potentially serious nature of disease where early intervention may be crucial.*



		True Disease Status		Total
		Diseased	Non-diseased	
Test Result	Positive	a True Positive	b False Positive	a + b All: Positive Results
	Negative	c False Negative	d True Negative	c + d All: Negative Results
Total		a + c All: Diseased Subjects	b + d All: Non-diseased Subjects	a + b + c + d All: Tested Subjects



npj | Digital Medicine

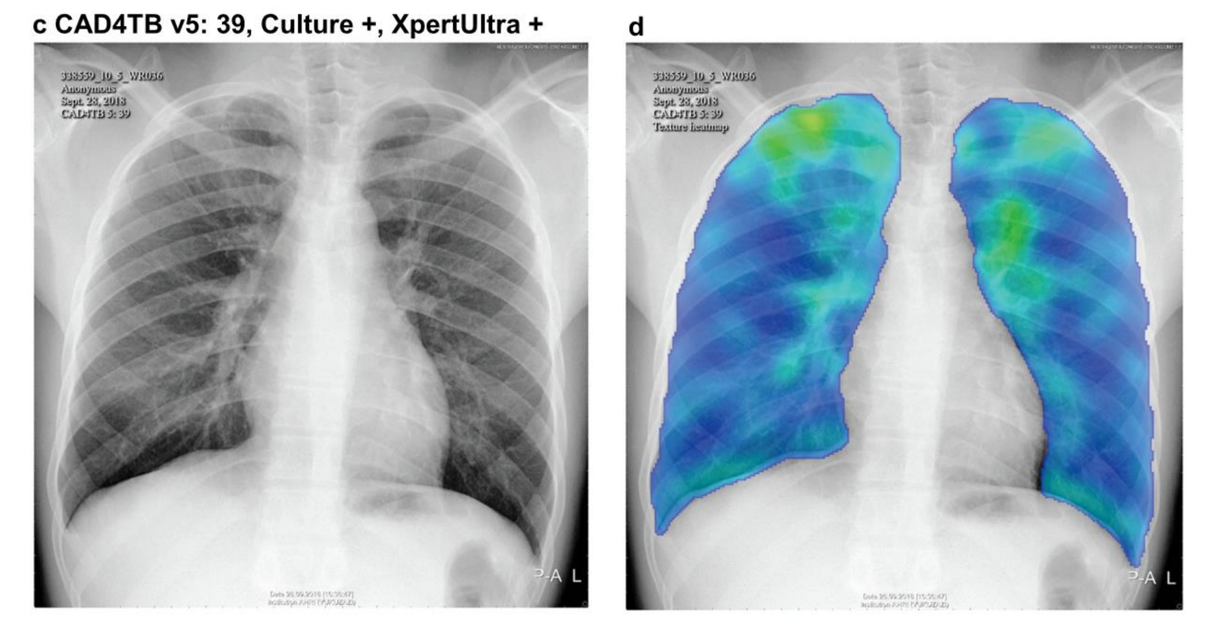
www.nature.com/npjdigitalmed

ARTICLE OPEN

Check for updates

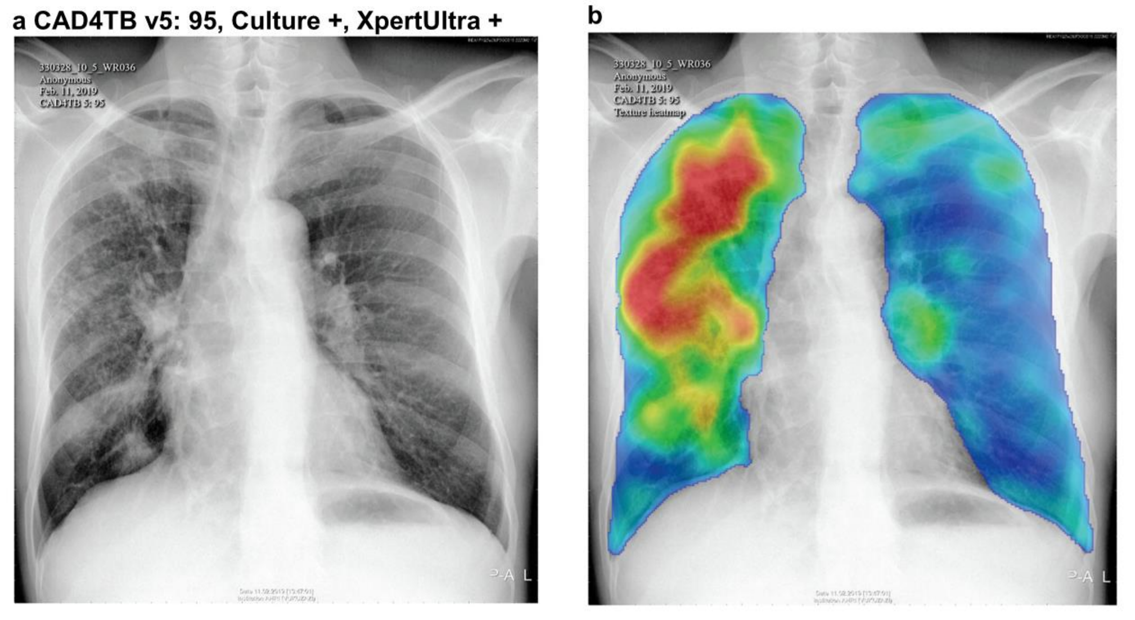
Computer-aided interpretation of chest radiography reveals the spectrum of tuberculosis in rural South Africa

Jana Fehr[1,2], Stefan Konigorski[2,3], Stephen Olivier[1], Resign Gunda[1,4,5], Ashmika Surujdeen[1], Dickman Gareta[1], Theresa Smit[1], Kathy Baisley[1,6], Sashen Moodley[1], Yumna Moosa[1], Willem Hanekom[1,5], Olivier Koole[1,6], Thumbi Ndung'u[1,5,7,8,9], Deenan Pillay[1,5], Alison D. Grant[1,4,6,10], Mark J. Siedner[1,10,11,12], Christoph Lippert[2,3,20], Emily B. Wong[1,11,12,13,20] ✉ and the Vukuzazi Team*



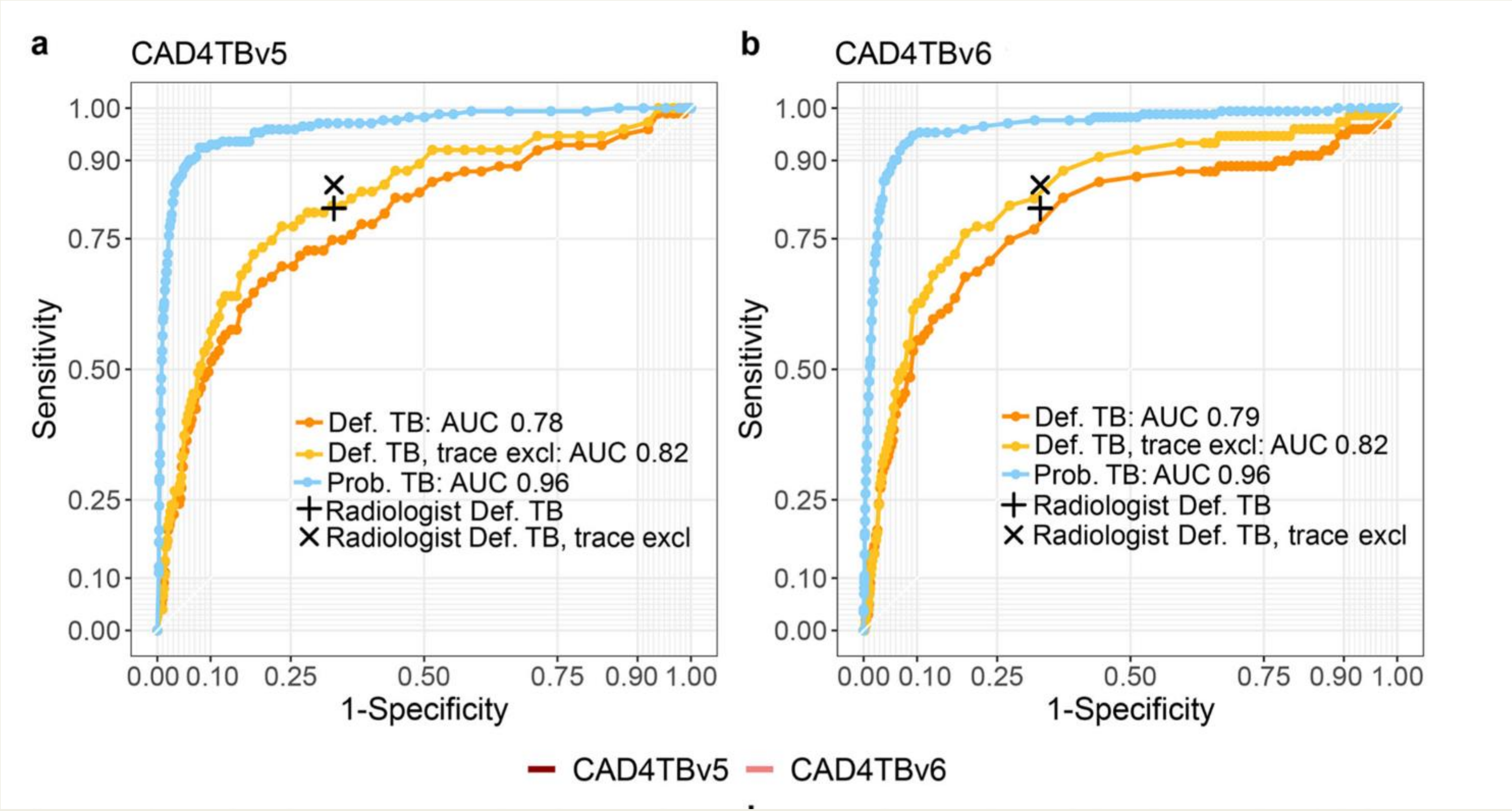
c CAD4TB v5: 39, Culture +, XpertUltra +

d



a CAD4TB v5: 95, Culture +, XpertUltra +

b

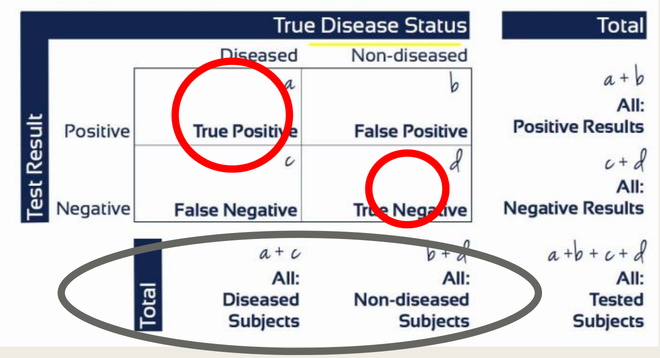

a
CAD4TBv5
Sensitivity
1-Specificity
1.00
0.90
0.75
0.50
0.25
0.10
0.00
0.00 0.10 0.25 0.50 0.75 0.90 1.00
Def. TB: AUC 0.78
Def. TB, trace excl: AUC 0.82
Prob. TB: AUC 0.96
Radiologist Def. TB
Radiologist Def. TB, trace excl
b
CAD4TBv6
Sensitivity
1-Specificity
1.00
0.90
0.75
0.50
0.25
0.10
0.00
0.00 0.10 0.25 0.50 0.75 0.90 1.00
Def. TB: AUC 0.79
Def. TB, trace excl: AUC 0.82
Prob. TB: AUC 0.96
Radiologist Def. TB
Radiologist Def. TB, trace excl
CAD4TBv5
CAD4TBv6

Two different perspectives: Second perspective (B)

A1. Sensitivity

A2. Specificity

Almost like inventor’s perspective.
How well did the new screening invention do in terms of identifying diseased and non-diseased individuals?

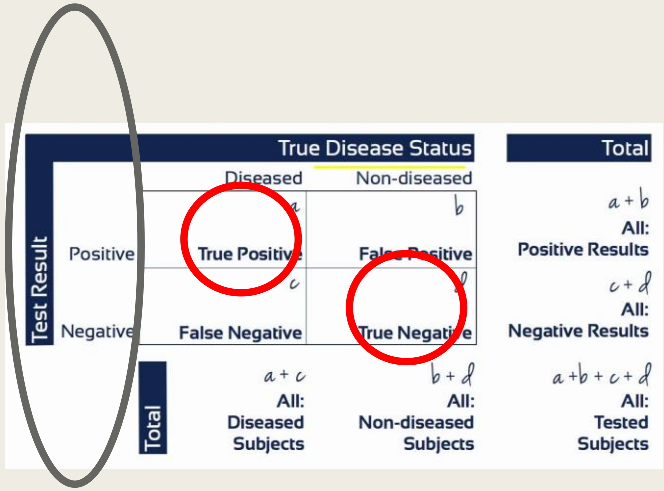


		True Disease Status		Total
Test Result		Diseased	Non-diseased	
	Positive	a True Positive	b False Positive	a + b All: Positive Results
	Negative	c False Negative	d True Negative	c + d All: Negative Results
Total		a + c All: Diseased Subjects	b + d All: Non-diseased Subjects	a +b + c + d All: Tested Subjects

B1. Positive predictive value (PPV)

B2. Negative predictive value (NPV)

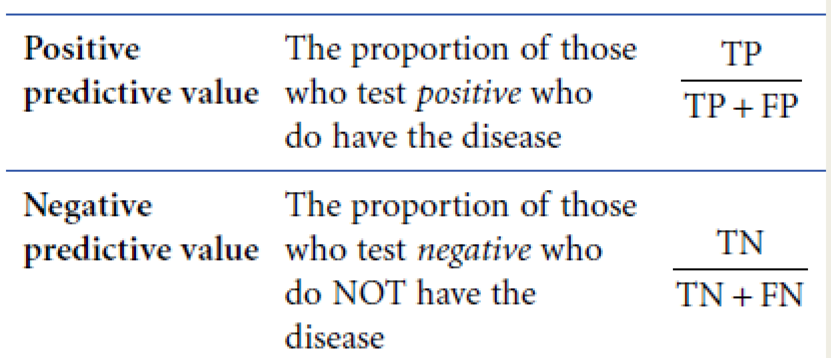
Patient perspective.
What is the probability that I actually have the disease when I informed that I am screen positive?



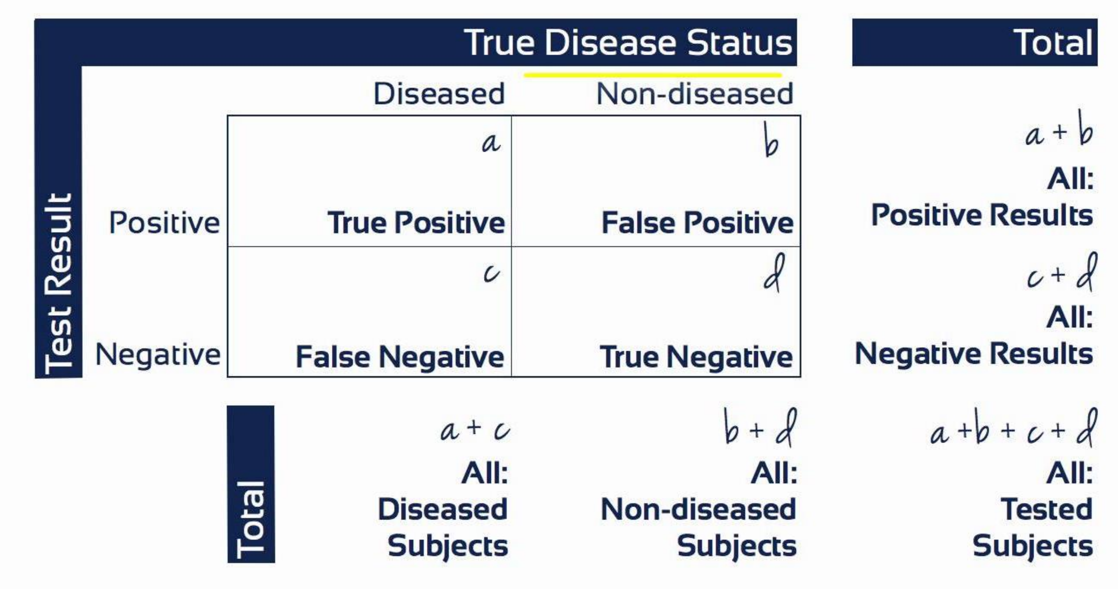
		True Disease Status		Total
Test Result		Diseased	Non-diseased	
	Positive	a True Positive	b False Positive	a + b All: Positive Results
	Negative	c False Negative	d True Negative	c + d All: Negative Results
Total		a + c All: Diseased Subjects	b + d All: Non-diseased Subjects	a +b + c + d All: Tested Subjects

Story starts from Blue. Your desire answer in Red.

Predictability of screening test (PPV/NPV)



Positive predictive value	The proportion of those who test *positive* who do have the disease	$\frac{TP}{TP + FP}$
Negative predictive value	The proportion of those who test *negative* who do NOT have the disease	$\frac{TN}{TN + FN}$



	True Disease Status		Total
Test Result	Diseased	Non-diseased	
Positive	a **True Positive**	b **False Positive**	a + b **All: Positive Results**
Negative	c **False Negative**	d **True Negative**	c + d **All: Negative Results**
Total	a + c **All: Diseased Subjects**	b + d **All: Non-diseased Subjects**	a + b + c + d **All: Tested Subjects**

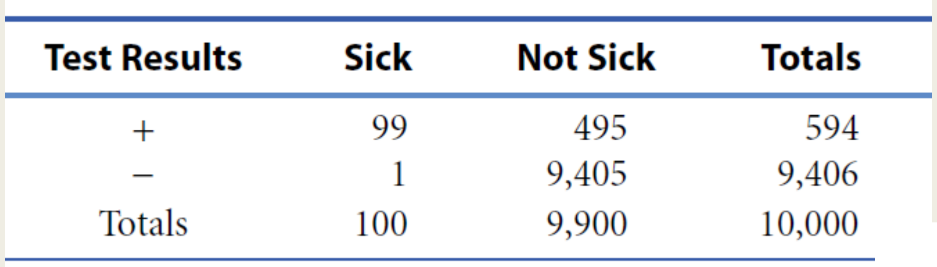
$$\text{PPV} = \frac{a}{a+b}$$

$$= \frac{TP}{TP + FP}$$

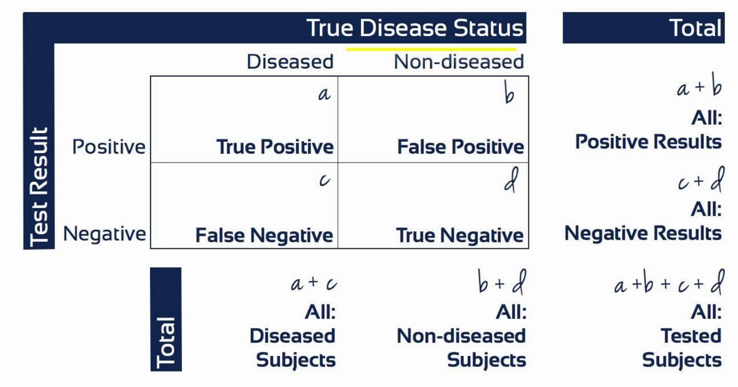
$$\text{NPV} = \frac{d}{c+d}$$

$$= \frac{TN}{TN + FN}$$

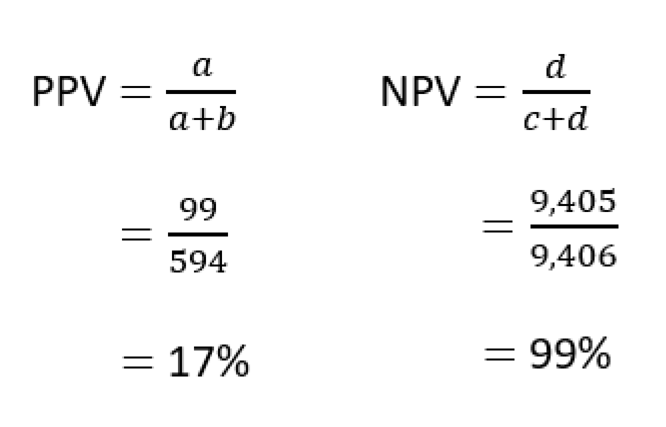
Calculation Example #1: Predictability of screening test (PPV/NPV)



Test Results	Sick	Not Sick	Totals
+	99	495	594
–	1	9,405	9,406
Totals	100	9,900	10,000



Test Result	True Disease Status: Diseased	True Disease Status: Non-diseased	Total
Positive	a True Positive	b False Positive	a + b All: Positive Results
Negative	c False Negative	d True Negative	c + d All: Negative Results
Total	a + c All: Diseased Subjects	b + d All: Non-diseased Subjects	a + b + c + d All: Tested Subjects



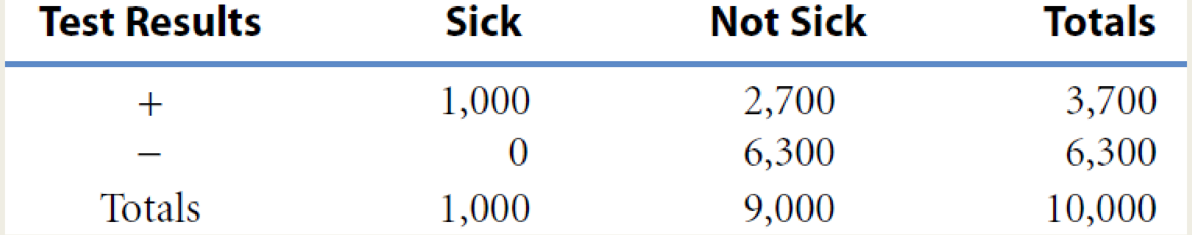
$$\text{PPV} = \frac{a}{a+b} = \frac{99}{594} = 17\%$$

$$\text{NPV} = \frac{d}{c+d} = \frac{9{,}405}{9{,}406} = 99\%$$

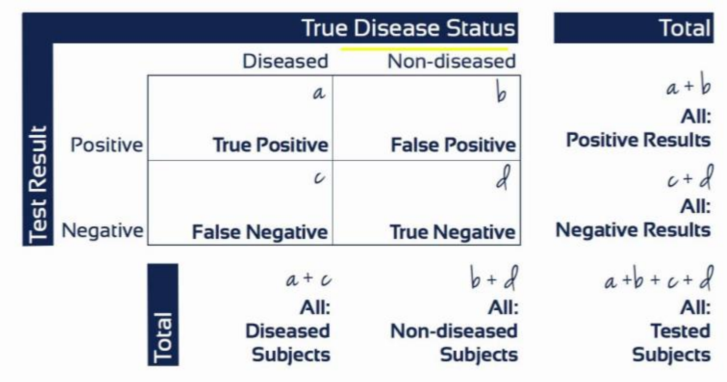
PPV of 17% interpreted as – probability that you will have the disease if you *test positive* on screening test is 17%

NPV of 99% interpreted as – probability that you won’t have the disease if you *test negative* on screening test is 99%

Calculation Example #2:
Predictability of screening test (PPV/NPV)



Test Results	Sick	Not Sick	Totals
+	1,000	2,700	3,700
–	0	6,300	6,300
Totals	1,000	9,000	10,000



		True Disease Status		Total
		Diseased	Non-diseased	
Test Result	Positive	a True Positive	b False Positive	a + b All: Positive Results
	Negative	c False Negative	d True Negative	c + d All: Negative Results
Total		a + c All: Diseased Subjects	b + d All: Non-diseased Subjects	a + b + c + d All: Tested Subjects

$$\text{PPV} = \frac{a}{a+b} \qquad \text{NPV} = \frac{d}{c+d}$$

$$= \frac{1000}{3700} \qquad = \frac{6{,}300}{6{,}300}$$

$$= 27\% \qquad = 100\%$$

PPV of 27% interpreted as – probability that you will have the disease if you *test positive* on screening test is 27%

NPV of 100% interpreted as – probability that you won't have the disease if you *test negative* on screening test is 100%

Relationship between PPV and Disease Prevalence

- Point #1: PPV is influenced by prevalence of disease
- Implication about targeting (e.g. where) of screening program.
- Below from example #1:

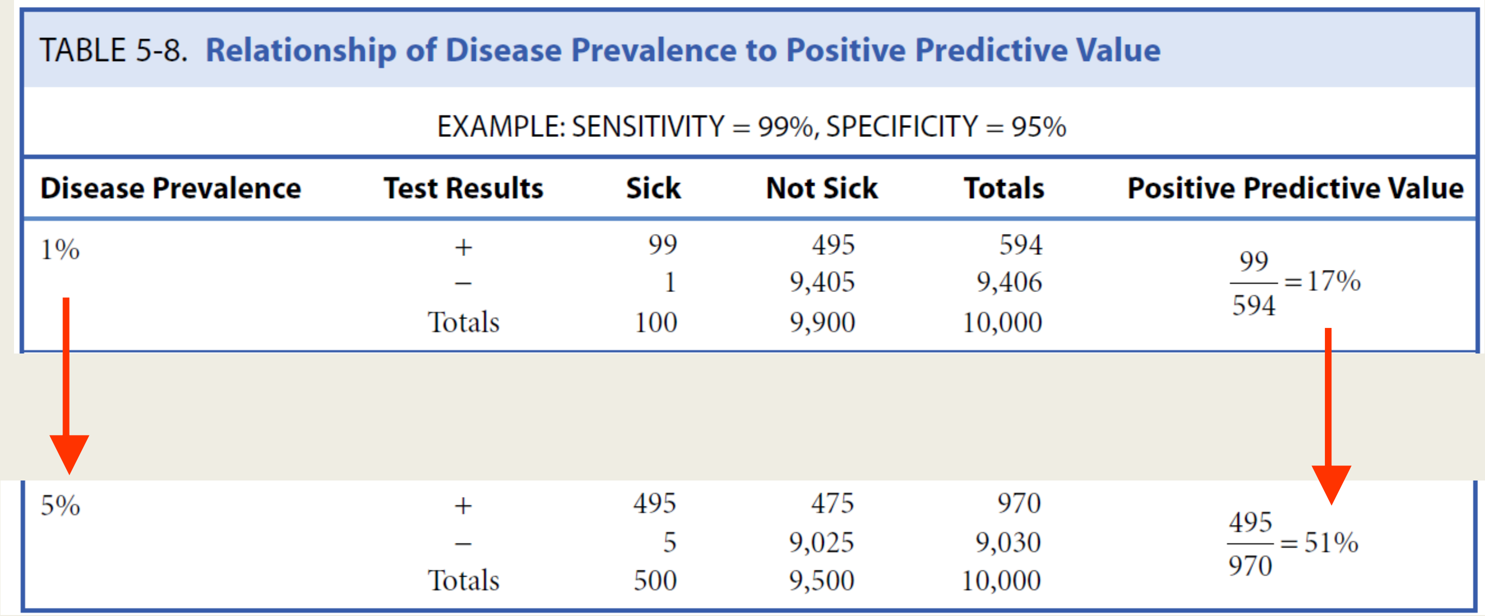


TABLE 5-8. **Relationship of Disease Prevalence to Positive Predictive Value**

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	9,030	
	Totals	500	9,500	10,000	

Note: Virtually no changes to NPV

Relationship between PPV and Specificity

- Point #2: PPV is influenced by specificity.
- Implication about which tool to adopt for screening.
- Below from example #2:

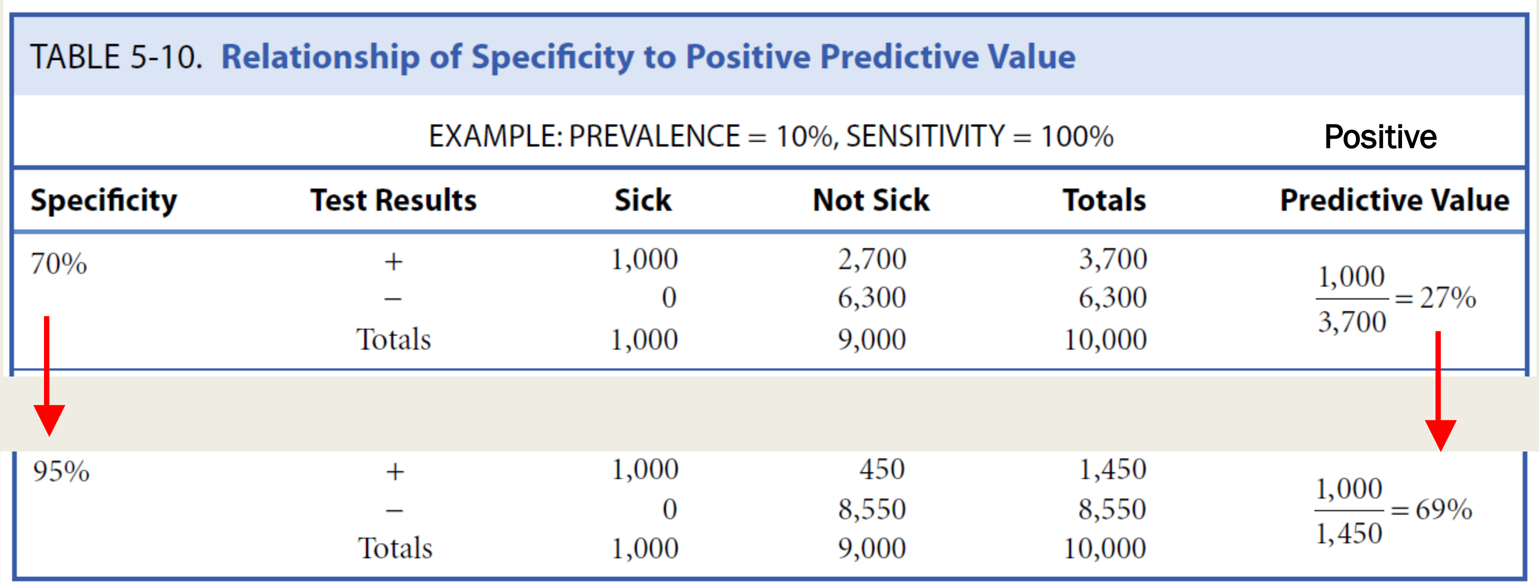


TABLE 5-10. **Relationship of Specificity to Positive Predictive Value**

EXAMPLE: PREVALENCE = 10%, SENSITIVITY = 100%

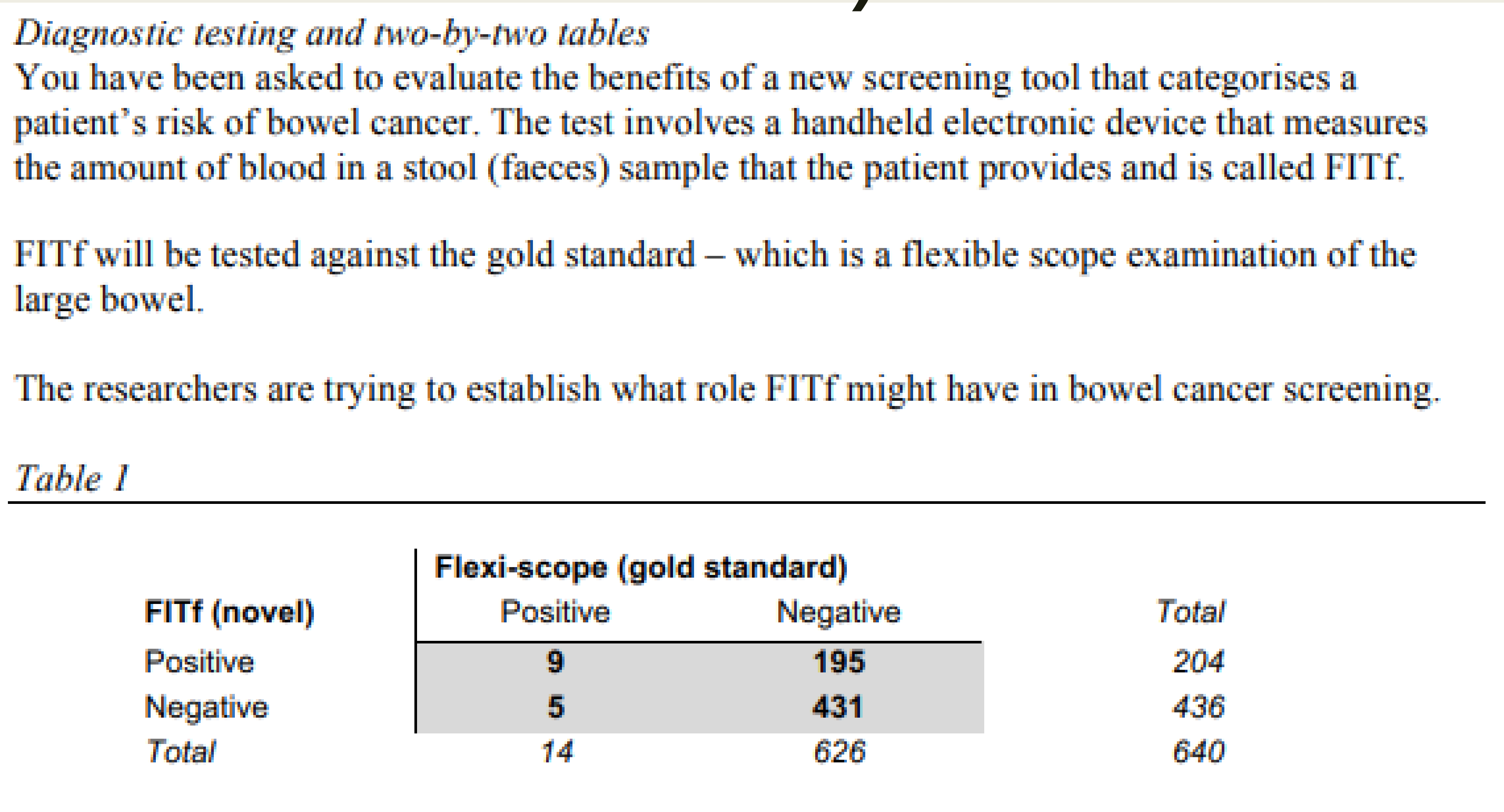
					Positive
Specificity	**Test Results**	**Sick**	**Not Sick**	**Totals**	**Predictive Value**
70%	+	1,000	2,700	3,700	$\frac{1,000}{3,700} = 27\%$
	–	0	6,300	6,300	
	Totals	1,000	9,000	10,000	
95%	+	1,000	450	1,450	$\frac{1,000}{1,450} = 69\%$
	–	0	8,550	8,550	
	Totals	1,000	9,000	10,000	

We should also care about specificity.

Note: Virtually no changes to NPV

Practical

In-class exercise (project during class)



Diagnostic testing and two-by-two tables

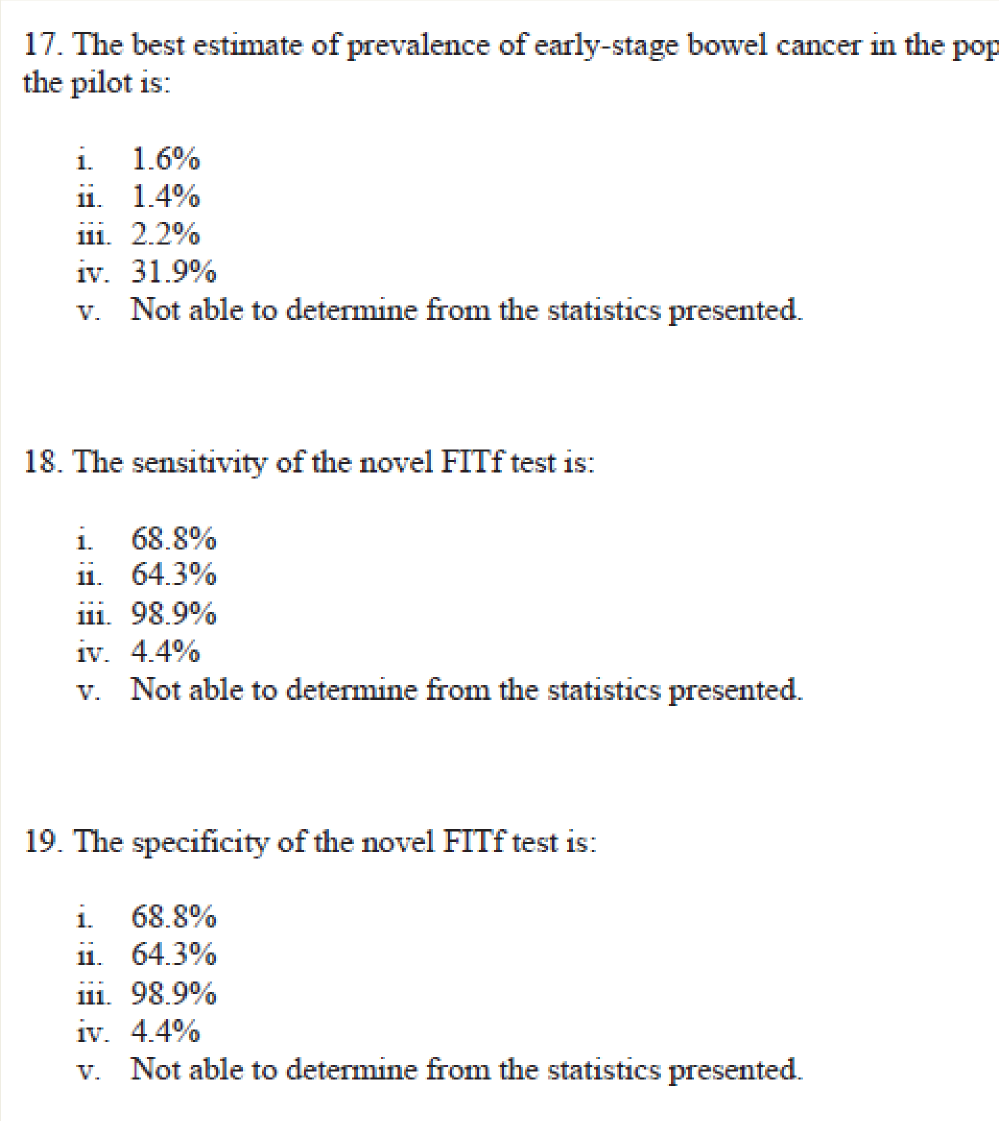
You have been asked to evaluate the benefits of a new screening tool that categorises a patient's risk of bowel cancer. The test involves a handheld electronic device that measures the amount of blood in a stool (faeces) sample that the patient provides and is called FITf.

FITf will be tested against the gold standard – which is a flexible scope examination of the large bowel.

The researchers are trying to establish what role FITf might have in bowel cancer screening.

Table 1

	Flexi-scope (gold standard)		
FITf (novel)	Positive	Negative	*Total*
Positive	**9**	**195**	*204*
Negative	**5**	**431**	*436*
Total	*14*	*626*	*640*



17. The best estimate of prevalence of early-stage bowel cancer in the population undergoing the pilot is:

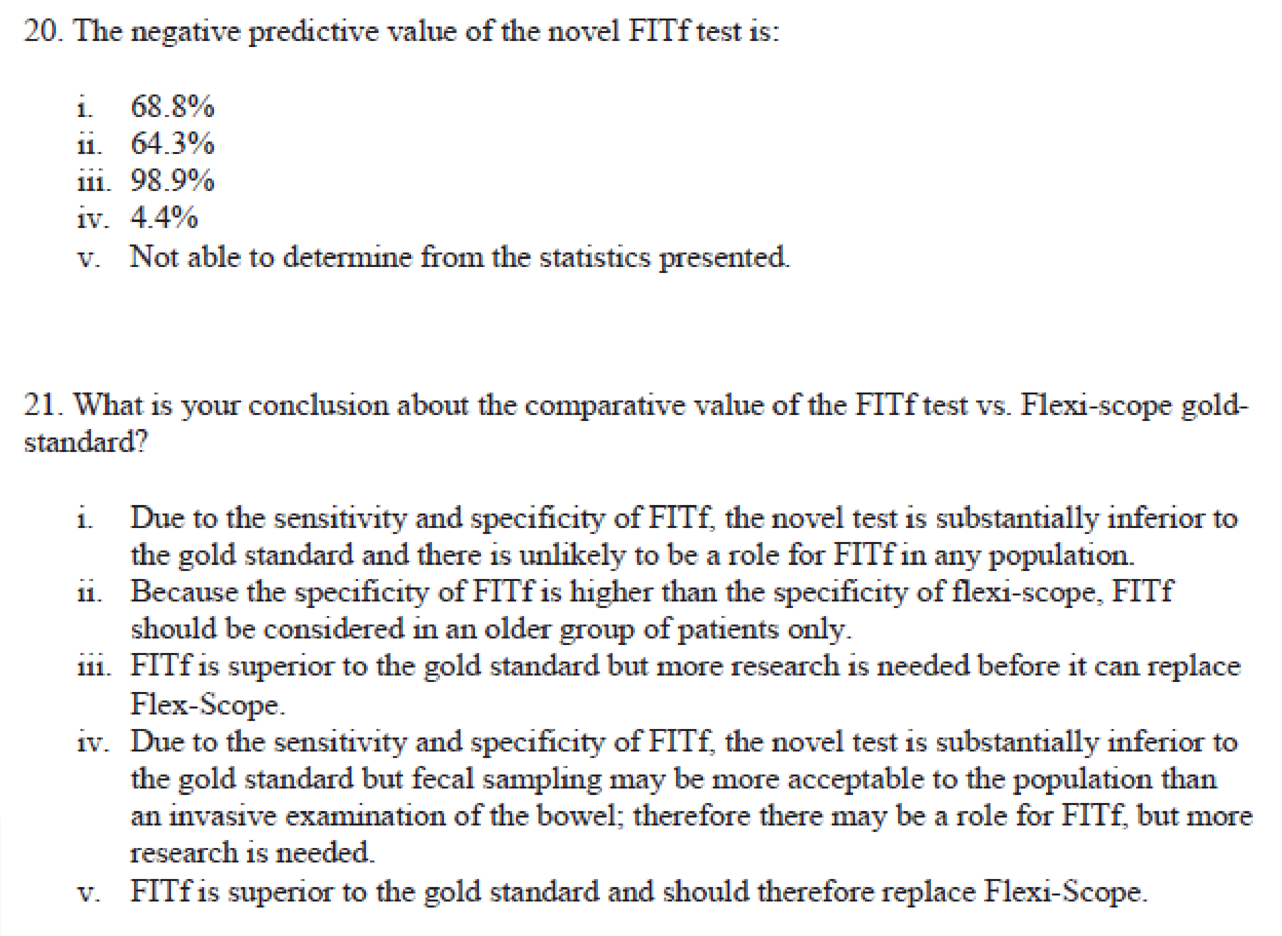
i. 1.6%
ii. 1.4%
iii. 2.2%
iv. 31.9%
v. Not able to determine from the statistics presented.

18. The sensitivity of the novel FITf test is:

i. 68.8%
ii. 64.3%
iii. 98.9%
iv. 4.4%
v. Not able to determine from the statistics presented.

19. The specificity of the novel FITf test is:

i. 68.8%
ii. 64.3%
iii. 98.9%
iv. 4.4%
v. Not able to determine from the statistics presented.



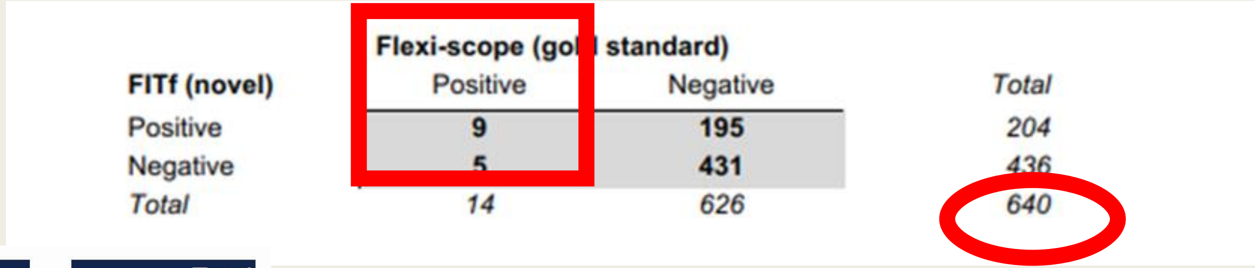
20. The negative predictive value of the novel FITf test is:

i. 68.8%
ii. 64.3%
iii. 98.9%
iv. 4.4%
v. Not able to determine from the statistics presented.

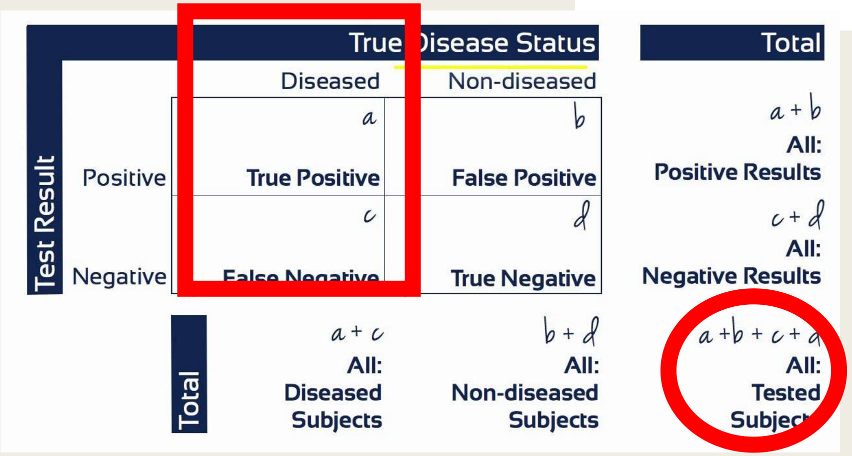
21. What is your conclusion about the comparative value of the FITf test vs. Flexi-scope gold-standard?

i. Due to the sensitivity and specificity of FITf, the novel test is substantially inferior to the gold standard and there is unlikely to be a role for FITf in any population.
ii. Because the specificity of FITf is higher than the specificity of flexi-scope, FITf should be considered in an older group of patients only.
iii. FITf is superior to the gold standard but more research is needed before it can replace Flex-Scope.
iv. Due to the sensitivity and specificity of FITf, the novel test is substantially inferior to the gold standard but fecal sampling may be more acceptable to the population than an invasive examination of the bowel; therefore there may be a role for FITf, but more research is needed.
v. FITf is superior to the gold standard and should therefore replace Flexi-Scope.

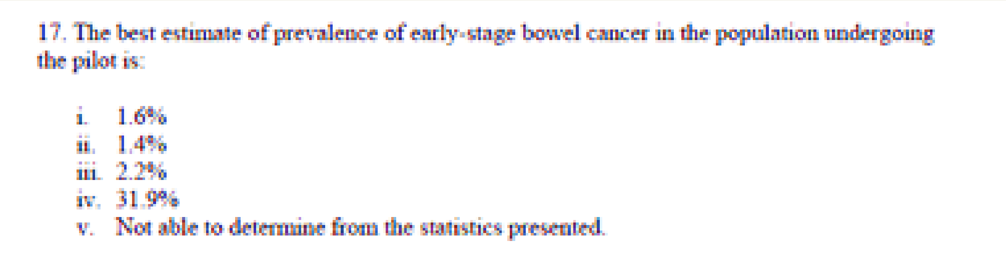
Q1 - Prevalence



FITf (novel)	Flexi-scope (gold standard) Positive	Negative	Total
Positive	**9**	**195**	*204*
Negative	**5**	**431**	*436*
Total	*14*	*626*	*640*



		True Disease Status		Total
		Diseased	Non-diseased	
Test Result	Positive	a **True Positive**	b **False Positive**	a + b **All: Positive Results**
	Negative	c **False Negative**	d **True Negative**	c + d **All: Negative Results**
Total		a + c **All: Diseased Subjects**	b + d **All: Non-diseased Subjects**	a + b + c + d **All: Tested Subjects**

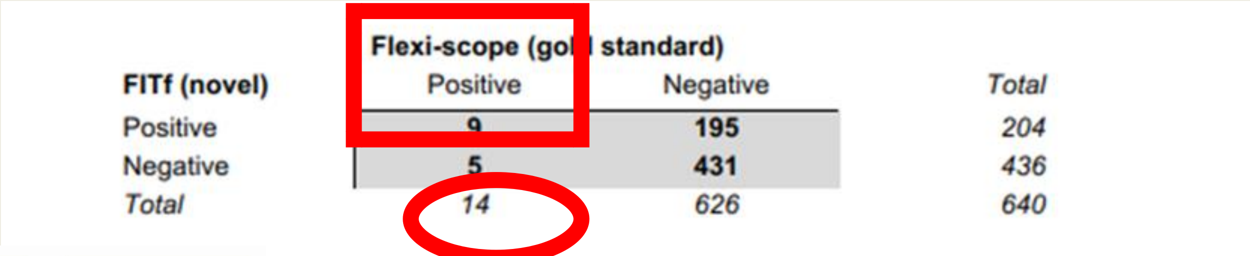


17. The best estimate of prevalence of early-stage bowel cancer in the population undergoing the pilot is:

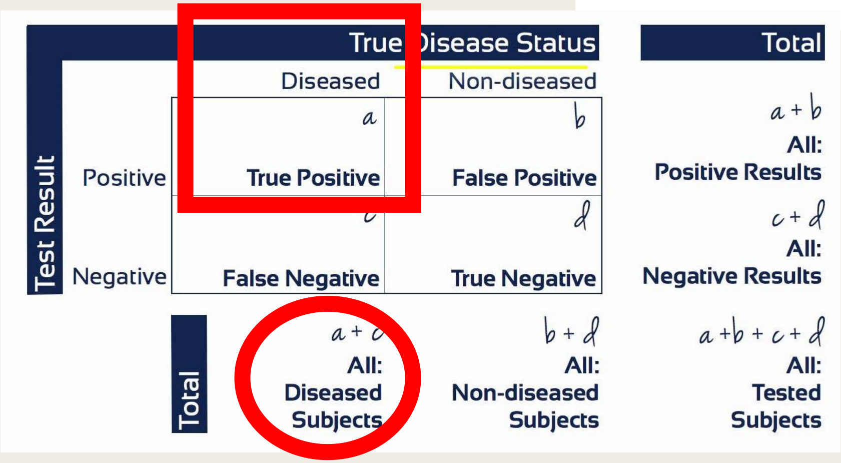
i. 1.6%
ii. 1.4%
iii. 2.2%
iv. 31.9%
v. Not able to determine from the statistics presented.

PREVALENCE = (a + c)/(a+b+c+d) = 0.022 or 2.2%

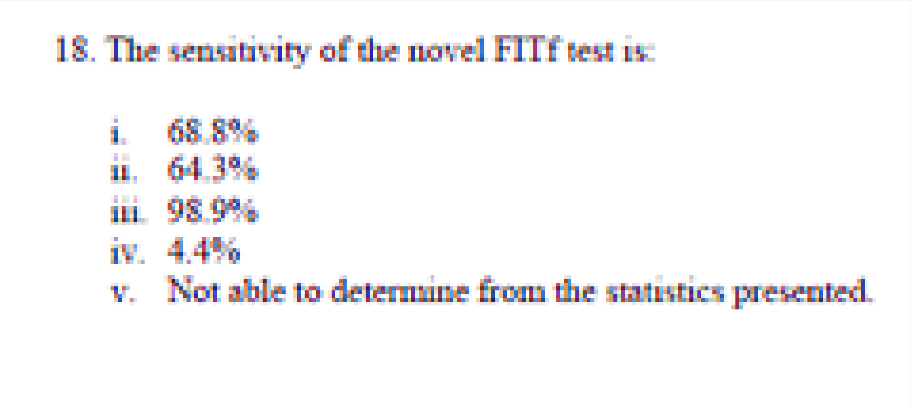
Q2 - Sensitivity



	Flexi-scope (gold standard)		
FITf (novel)	Positive	Negative	*Total*
Positive	**9**	**195**	*204*
Negative	**5**	**431**	*436*
Total	*14*	*626*	*640*



		True Disease Status		Total
		Diseased	Non-diseased	
Test Result	Positive	a **True Positive**	b **False Positive**	a + b **All:** **Positive Results**
	Negative	c **False Negative**	d **True Negative**	c + d **All:** **Negative Results**
Total		a + c **All:** **Diseased** **Subjects**	b + d **All:** **Non-diseased** **Subjects**	a + b + c + d **All:** **Tested** **Subjects**

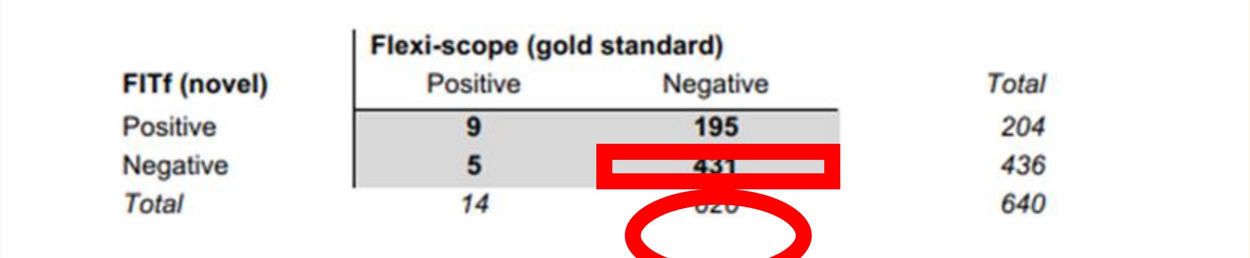


18. The sensitivity of the novel FITf test is:

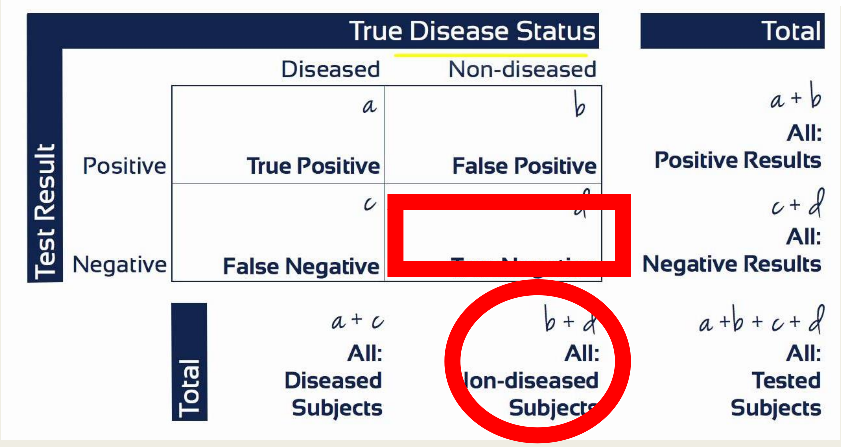
i. 68.8%
ii. 64.3%
iii. 98.9%
iv. 4.4%
v. Not able to determine from the statistics presented.

SENSITIVITY = a/(a+c) = 0.643 or 64.3%

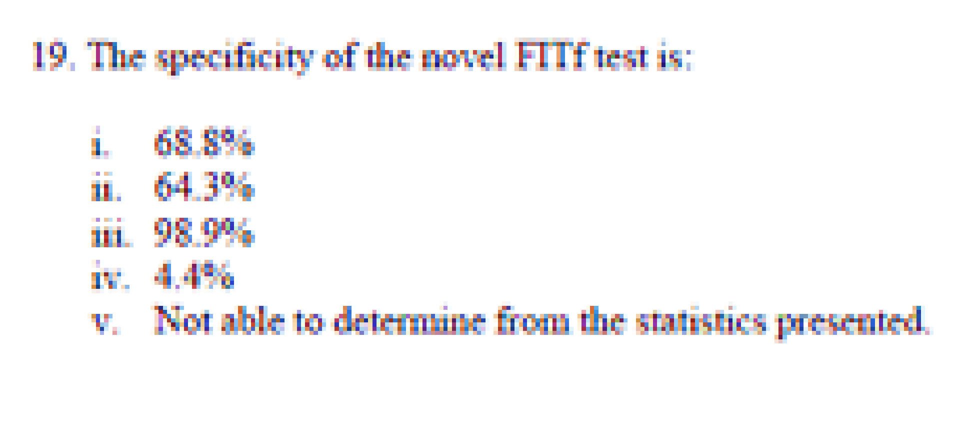
Q3 - Specificity



FITf (novel)	Flexi-scope (gold standard) Positive	Negative	*Total*
Positive	**9**	**195**	*204*
Negative	**5**	**431**	*436*
Total	*14*	*626*	*640*



Test Result	True Disease Status Diseased	Non-diseased	Total
Positive	a **True Positive**	b **False Positive**	a + b **All: Positive Results**
Negative	c **False Negative**	d **True Negative**	c + d **All: Negative Results**
Total	a + c **All: Diseased Subjects**	b + d **All: Non-diseased Subjects**	a + b + c + d **All: Tested Subjects**

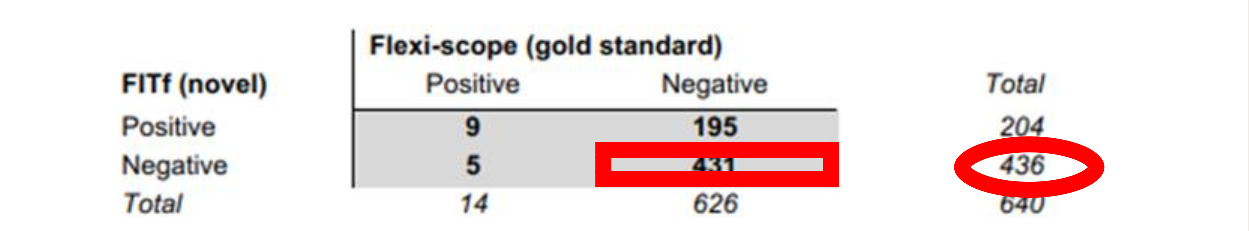


19. The specificity of the novel FITf test is:

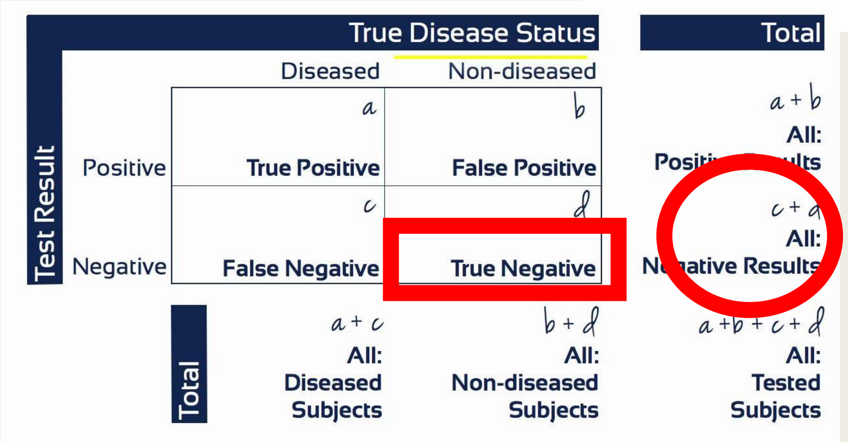
i. 68.8%
ii. 64.3%
iii. 98.9%
iv. 4.4%
v. Not able to determine from the statistics presented.

SPECIFICITY = d/(b+d) = 0.688 or 68.8%

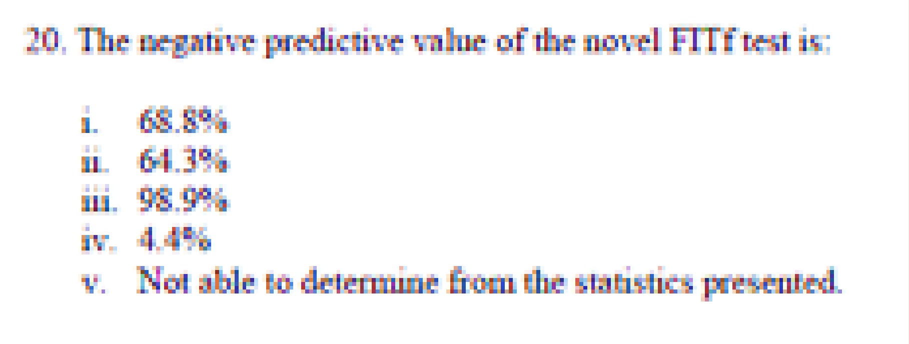
Q4 - NPV



FITf (novel)	Flexi-scope (gold standard) Positive	 Negative	*Total*
Positive	**9**	**195**	*204*
Negative	**5**	**431**	*436*
Total	*14*	*626*	*640*



		True Disease Status		Total
		Diseased	Non-diseased	
Test Result	Positive	*a* **True Positive**	*b* **False Positive**	*a + b* **All:** **Positive Results**
	Negative	*c* **False Negative**	*d* **True Negative**	*c + d* **All:** **Negative Results**
Total		*a + c* **All:** **Diseased Subjects**	*b + d* **All:** **Non-diseased Subjects**	*a +b + c + d* **All:** **Tested Subjects**

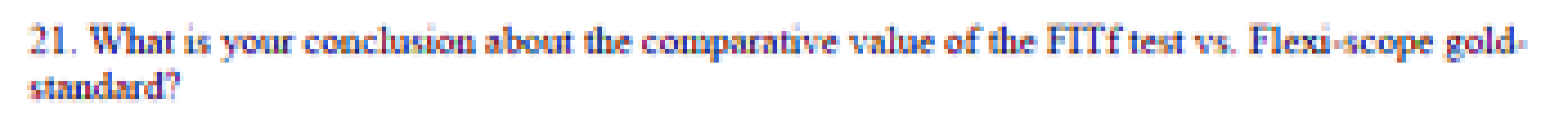


20. The negative predictive value of the novel FITf test is:

i. 68.8%
ii. 64.3%
iii. 98.9%
iv. 4.4%
v. Not able to determine from the statistics presented.

NPV = d/(c+d) = 0.989 or 98.9%

Q5 - Interpretation



21. What is your conclusion about the comparative value of the FITf test vs. Flexi-scope gold-standard?

- Usual target value for sensitivity or specificity – 80%
- FITf: Sen = 64.3%, Spec = 68.8%
- Sen/Spec FITf <80%
- However… Fecal blood testing – “home” testing, not invasive like biopsies, not waiting in lines for radiological exam
- FITf inferior to gold standard, but approach might be acceptable in some populations. Possible role for FITf in these populations, requires more research (increase sample size?, older individuals?, individuals who are working and cannot come in for other screening tests?).