Understanding Test Performance:

What does this test mean?

Dr Tom Johnston PhD FHEA

Lecturer in Population Health,

Hull York Medical School

Ultrasound 2024

12 Dec 2024

Outline

- What is a test?
- Test Performance
  - Sensitivity, Specificity
  - Positive Predictive Value, Negative Predictive Value
- Tests as part of a screening programme
  - Using tests together

What is a test?

What defines disease?

- Criteria that must be met for an individual to be considered a ‘case’.
  - Clinical sign or symptom?
    - Spots, fever
  - Biological indicator?
    - Antibodies, blood chemistry, kidney function, blood pressure
  - Survey responses?
    - Mental health

- Our set of criteria is a Test.

A *Gold Standard* test defines disease/exposure

When do we use a test?

- Surveillance (Active and Passive)
  - Estimate occurrence
  - Detect outbreaks and initiate response
  - Detect disease early to intervene
    - Screening
- Diagnosis
  - Initiate treatment
- Research
  - Estimate occurrence
  - Define individual status
    - Disease
    - Exposures (Risk factors)

What does a test look like?

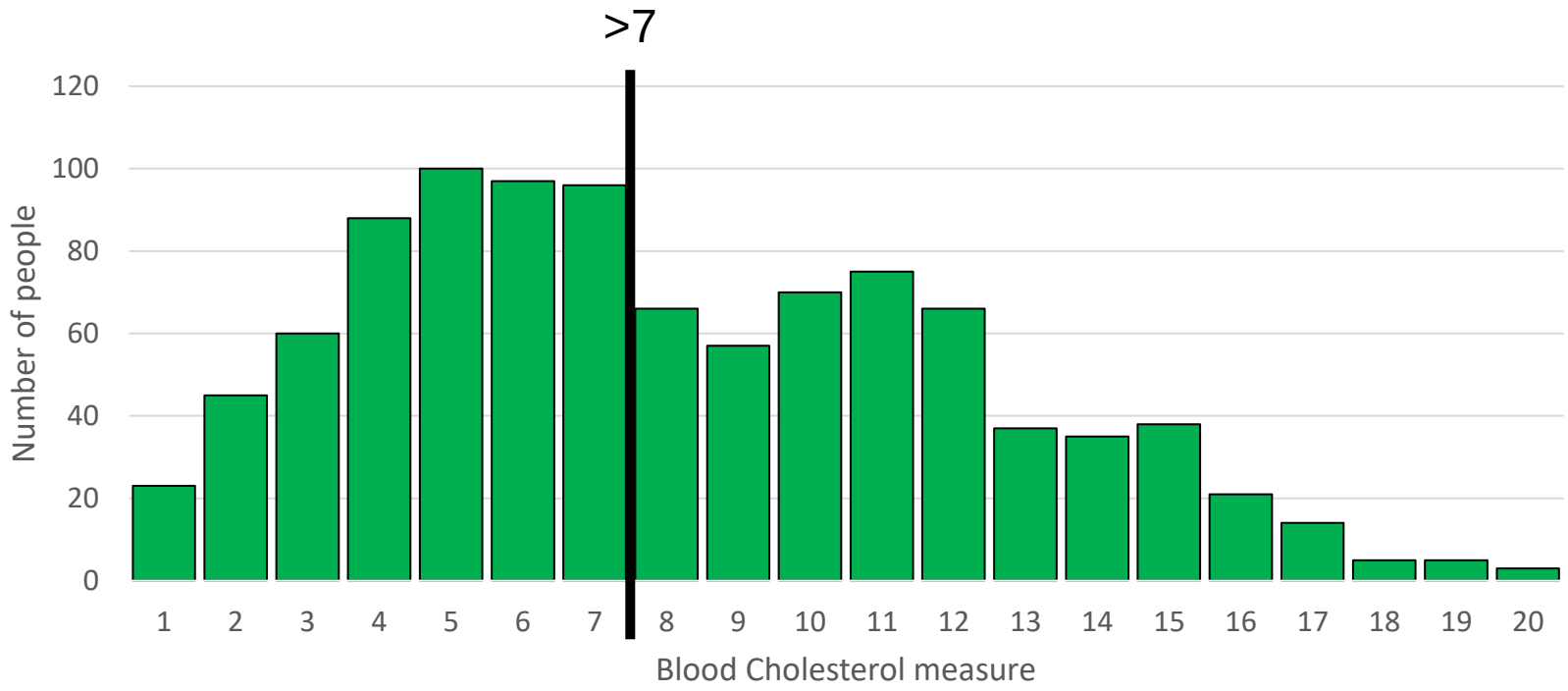
- A measurement or detection tool
  - Isolation/Culture
  - Visualization/Imagery
  - Physical Examination
  - Detection (Agglutination, PCR)
  - Questions
  - Biochemistry
  - Serology
  - Physical parameter
  - Questions

Discrete Outcomes *(often)*

Continuous or Ordinal Outcomes *(often)*

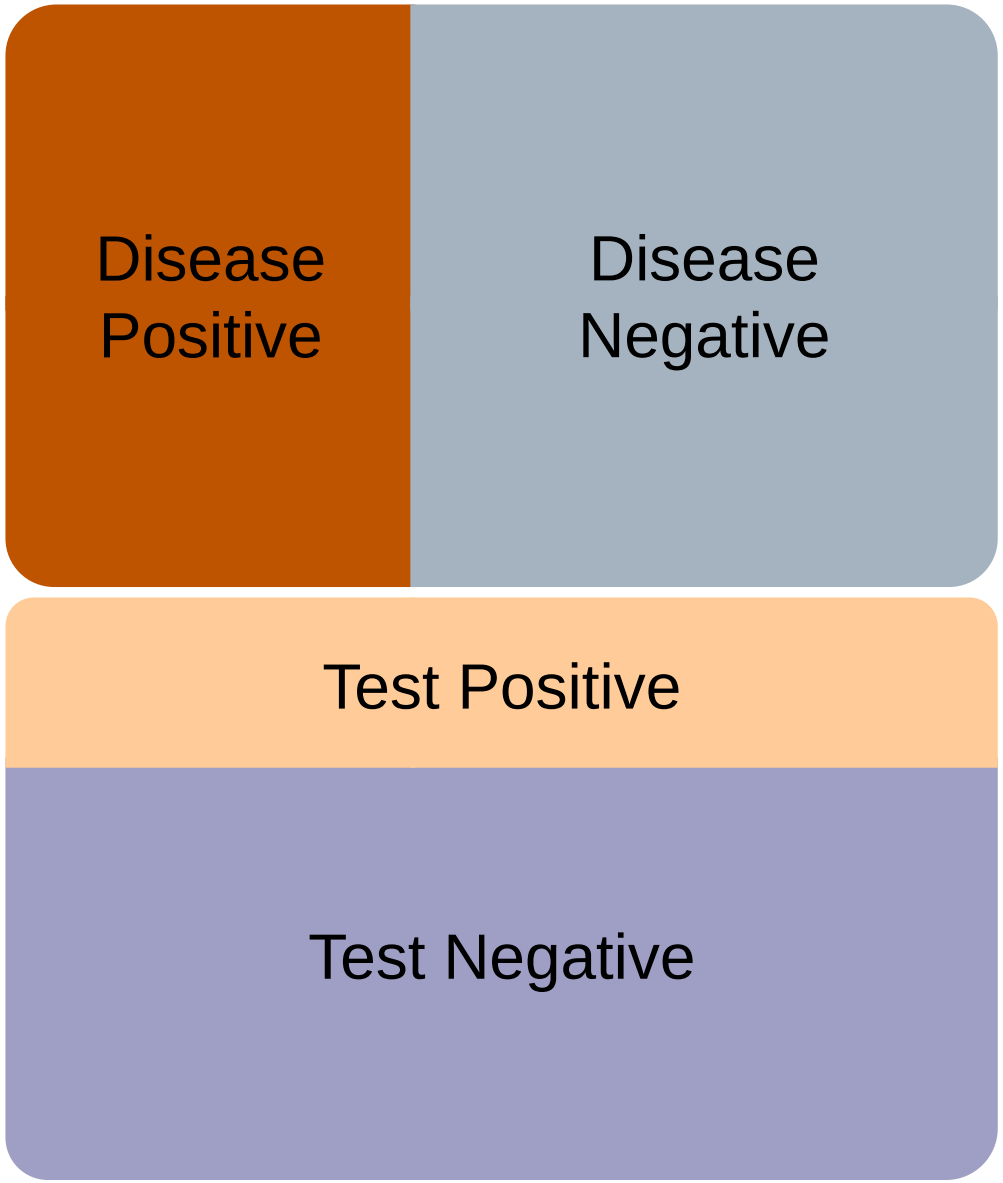
What does a test do for us?

- Occurrence
- Distribution
- Classifies individuals
  - Positive/Negative



However we use them, we need to remember that tests are imperfect: Misclassification bias

If we could apply the *Gold Standard* to all individuals…



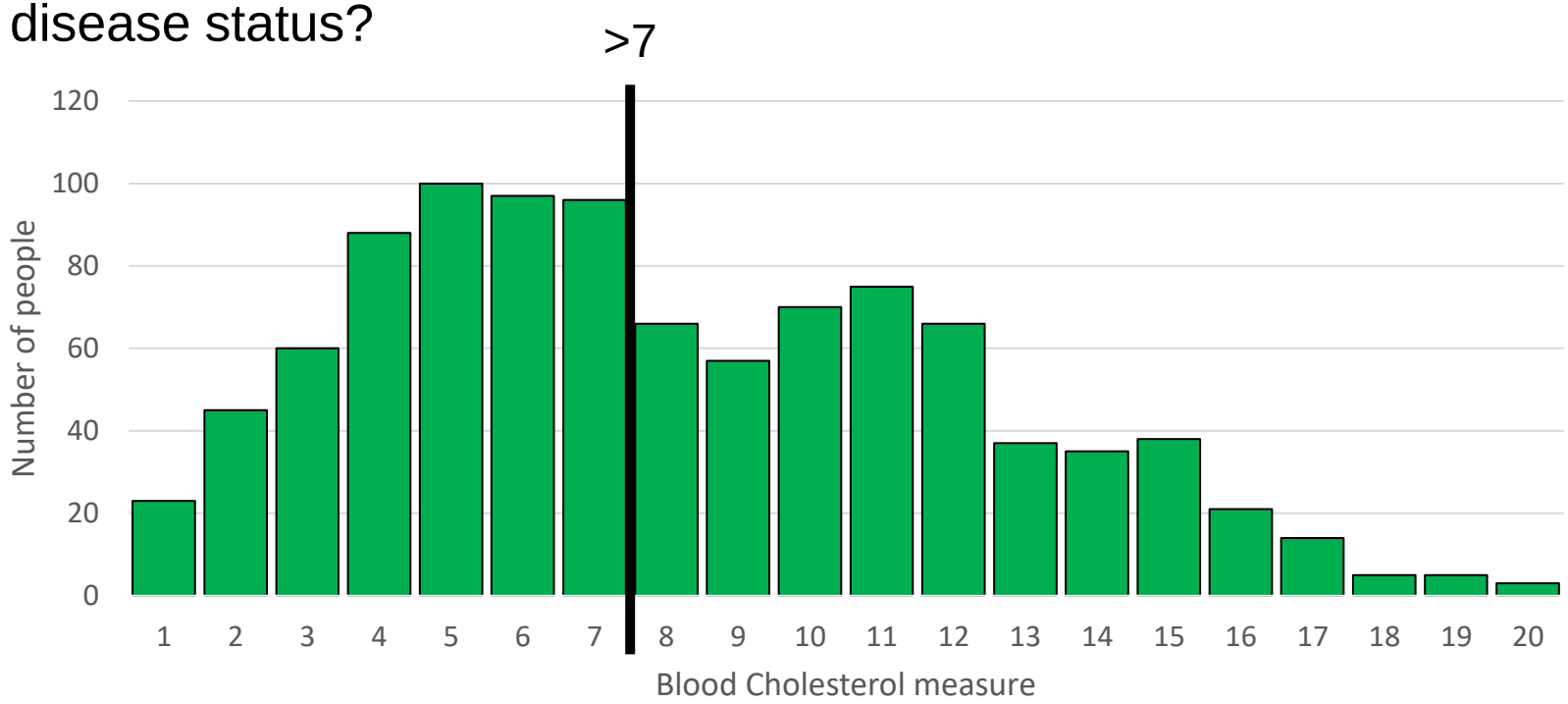
Test results may not completely align with the truth

Performance is how we quantify how good this alignment is.

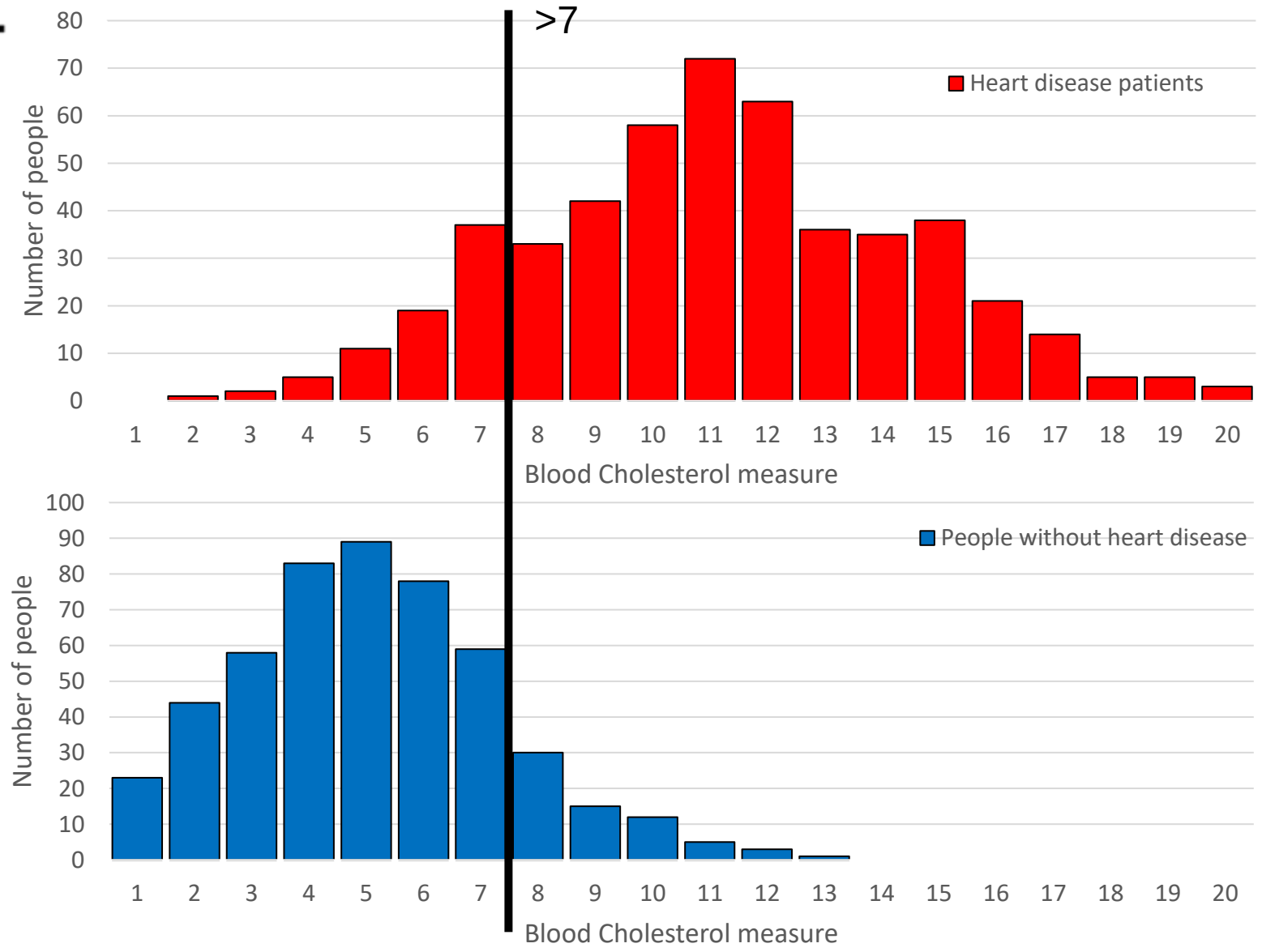
What is test performance?

What do we mean by performance?

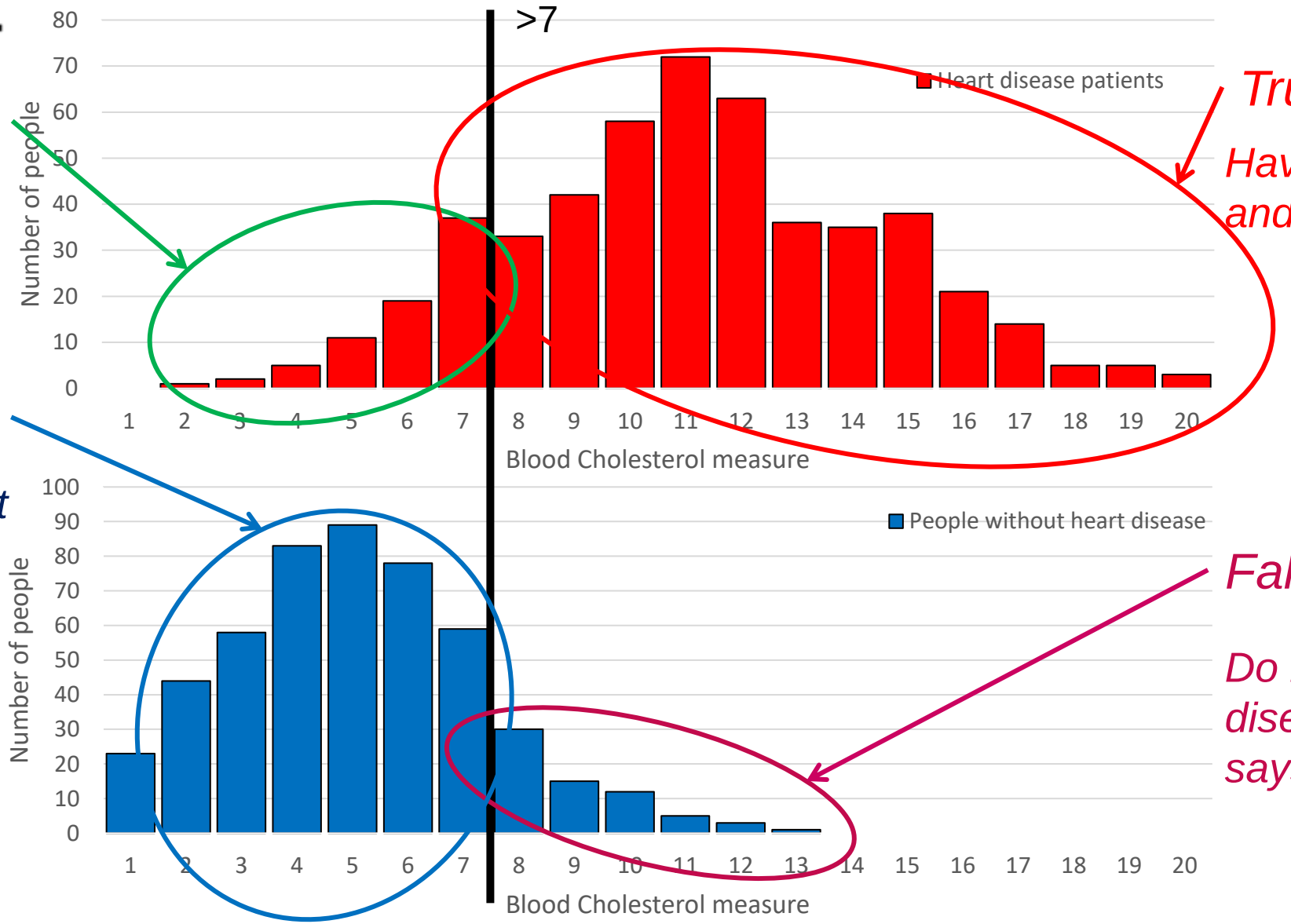
- How ‘good’ the test is
  - Can it identify those with disease?
  - Can it identify those without disease?
  - Does the test result predict disease status?
    - Positive result
    - Negative result



Distribution of a test parameter



Classification based on the test



False negative

Have heart disease but test says they do not

True Positive

Have heart disease and test says they do

True Negative

Do not have heart disease and test says they don't

False positive

Do not have heart disease but test says they do

Classification Summarised in a 2x2 table

		True disease status Measured by Gold Standard	
		Positive	Negative
Test	Positive	*True positive* has got 'it' and has a positive test result	*False positive* has not got 'it' and has a positive test result
	Negative	*False negative* has got 'it' and has a negative test result	*True negative* has not got 'it' and has a negative test result

Defining Performance…

		Gold Standard	
		+ve	-ve
Test Result	+ve	True Positive	
	-ve		True Negative
		Total Diseased	Total Free of Disease

A good test…

Places most people here

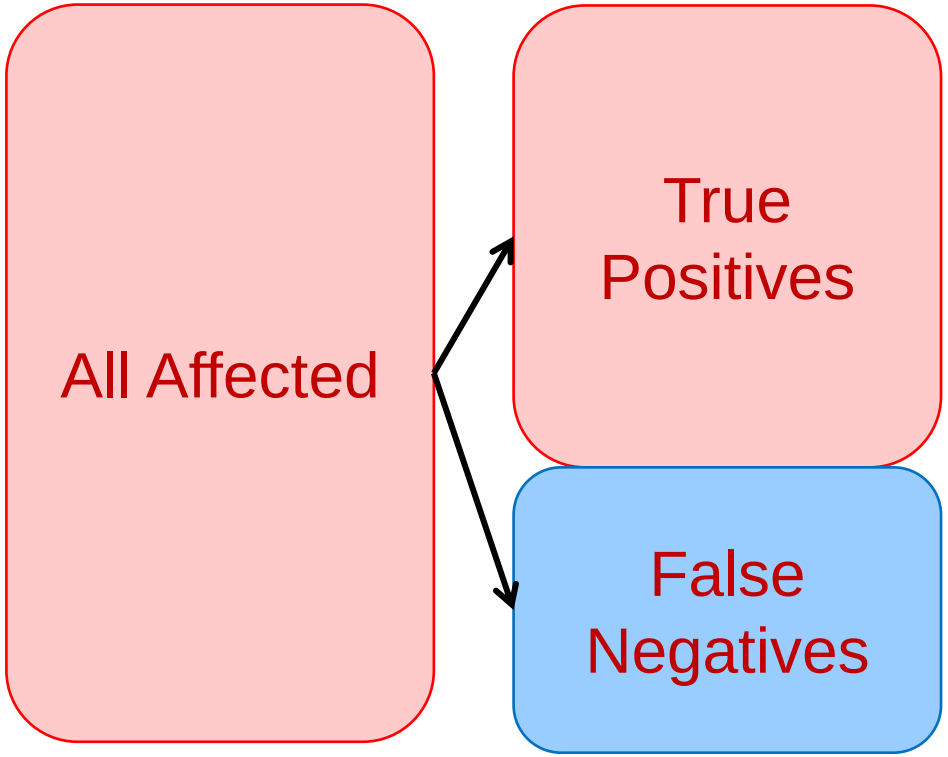
Places few people here

- How good is the test at finding Disease?
  - Sensitivity
- How good is the test at excluding Disease?
  - Specificity

Test Performance: Sensitivity

How well does the test identify *diseased* individuals?

Sensitivity = The proportion of affected individuals *correctly identified* by the test



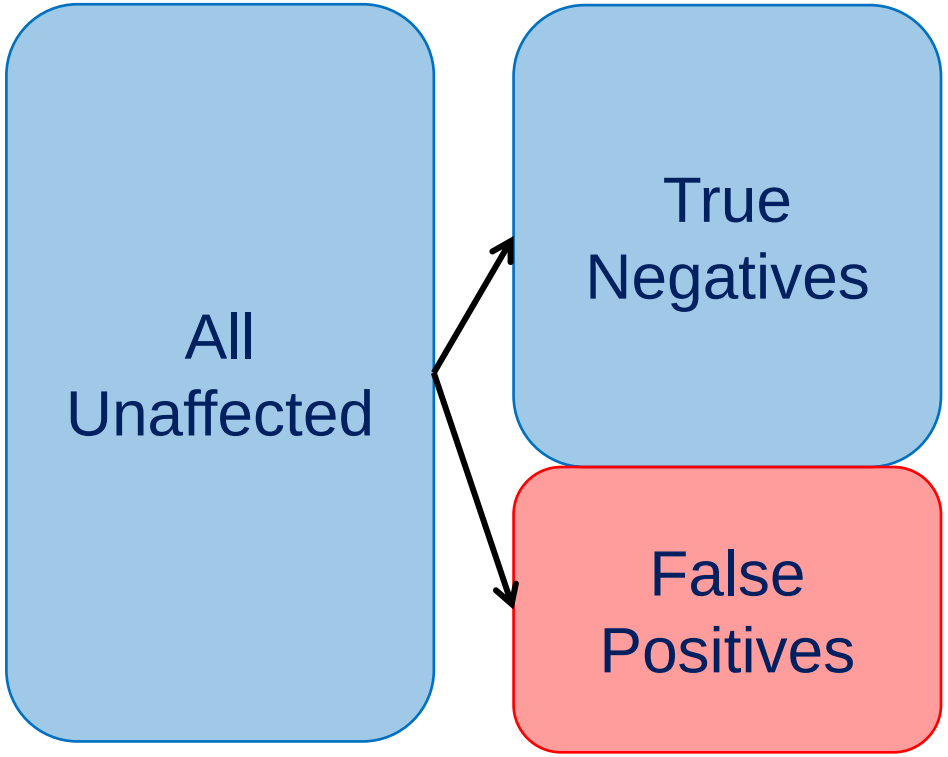
$$Sn = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

$$= \frac{\text{True Positives}}{\text{All Affected}}$$

Test Performance: Specificity

How well does the test identify *disease-free* individuals?

Specificity = The proportion of unaffected individuals *correctly identified* by the test



$$Sp = \frac{\text{True Negatives}}{\text{True Negatives + False Positives}}$$

$$= \frac{\text{True Negatives}}{\text{All Unaffected}}$$

Diagnostic tool for Depression

Klinkman et al. 1998. Arch Fam Med

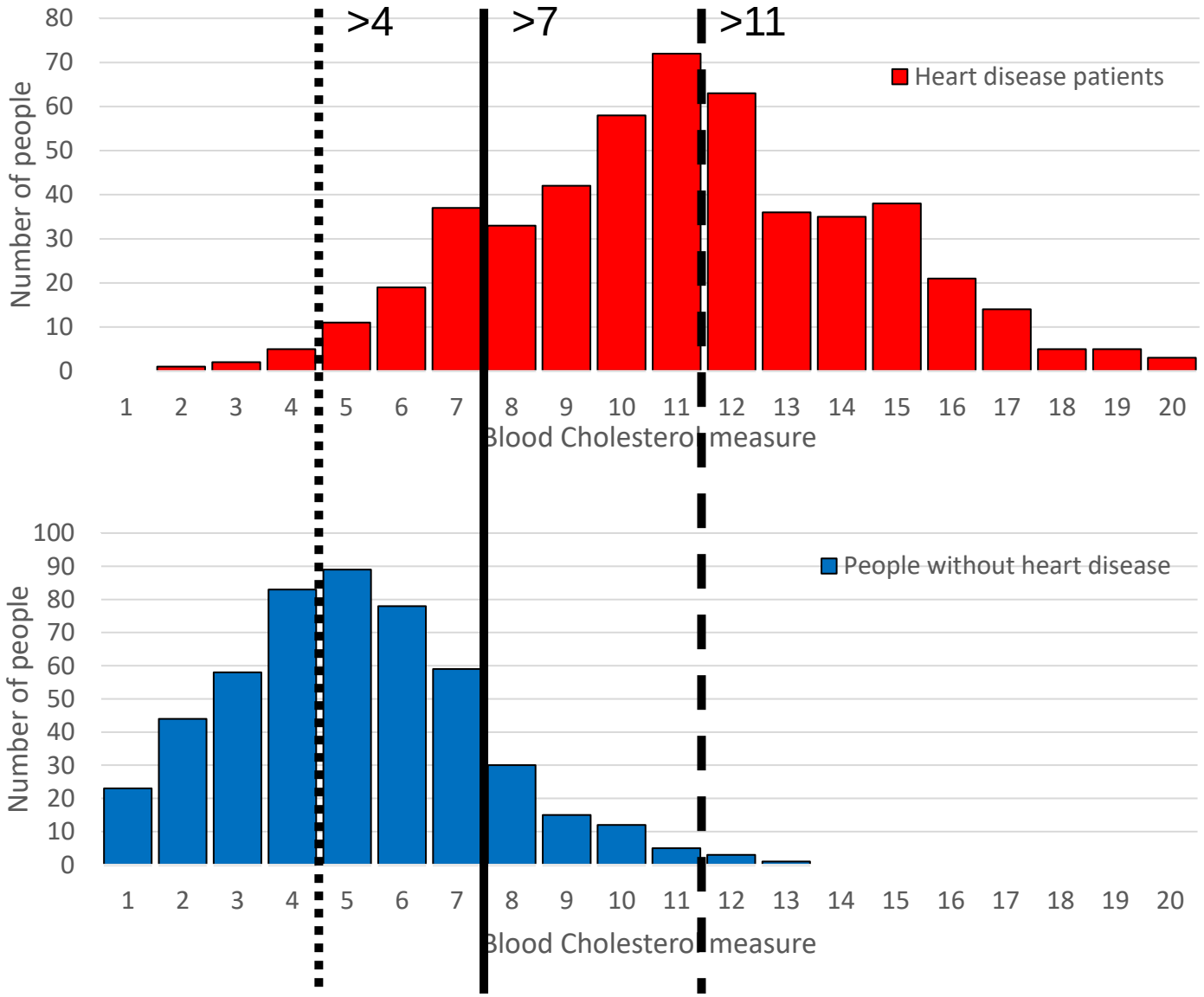
		GP-assessed		
		+	-	
Tool	+	31	34	65
	-	50	257	307
		81	291	372

- Sensitivity = 31/81 = 38%
  - Therefore 38% of DEPRESSED people will be identified as BEING DEPRESSED
    - Or 62% of cases WILL BE MISSED
- Specificity = 257/291 = 88%
  - 88% of NOT-DEPRESSED people will be identified as not being depressed
    - but 12% will be identified as being DEPRESSED

Remember: we set the cut off value at 7

What if we used 4?

- Very few cases missed
- Higher Sensitivity
- More misclassification of disease-free
- Lower Specificity



What if we used 11?

- Very few disease-free missed
- Higher Specificity
- More misclassification of cases
- Lower Sensitivity

Impact in our example

Diagnostic tool for Depression

– People with depression are expected to score higher on the screening tool

- To maximise Sensitivity you could DECREASE the cutoff score
  - More people would be classed as being depressed
- Consequences:
  - More people with depression would be flagged for therapy
    - Good thing
  - More people without depression would also be flagged
    - Not so good

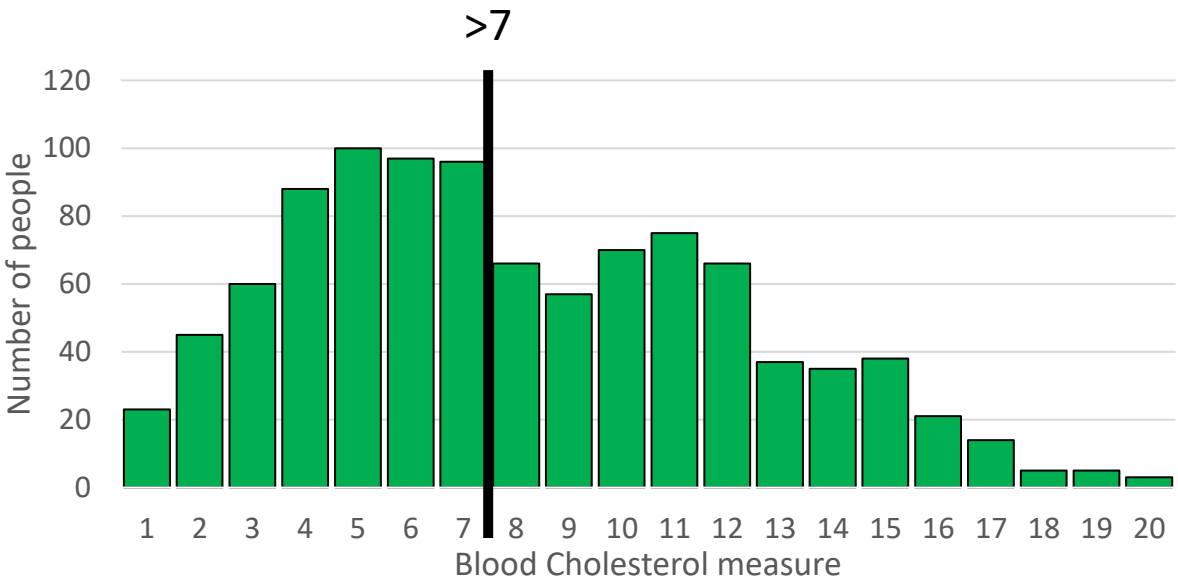
- To maximise Specificity you could INCREASE the cutoff score
  - Fewer people would be classed as being depressed
- Consequences:
  - Fewer people without depression would be flagged for therapy
    - Good thing
  - Fewer people with depression would also be flagged
    - Not so good

What does our test result mean?

- We don't know 'The Truth'
- All we see are the results
  - Positive
  - Negative

Question:

How well do the test results predict the true status?



		Disease Status		
		+	-	
Test Status	+	?	?	Test +ve
	-	?	?	Test -ve

Predictive Values

How often is the test result correct?

Positive Predictive Value (PPV)

- Proportion of positive tests that are correct

Negative Predictive Value (NPV)

- Proportion of negative tests that are correct

		Gold Standard		
		+ve	-ve	
Test Result	+ve	TP	FN	All Test +ve
	-ve	FN	TN	All Test -ve

$$PPV = \frac{\text{True positives}}{\text{Test positives}}$$

$$NPV = \frac{\text{True negatives}}{\text{Test negatives}}$$

What do predictive values mean?

Depression Screening Tool

- Sensitivity: 38%
- Specificity: 88%

Prevalence = 21.8%		Depressed (GP)		
		+	-	
Screening Tool	+	31	34	65
	-	50	257	307
		81	291	372

PPV:

65 people test positive,
31 of which are truly positive

Therefore: 31/65 = 47.7%

48% of *positive* tests are correct

and 52 % are not!

NPV:

307 people test negative,
257 of which are truly negative

Therefore: 257/307 = 83.7%

84% of *negative* tests are correct

and 16% are not

Some situations where prevalence changes...

- Between primary care and secondary care
- Across age groups
- Between countries

What if disease is less frequent?

Depression Screening Tool

- Sensitivity: 38%
- Specificity: 88%

Prevalence = 5.1%		Depressed (GP)		
		+	-	
Screening Tool	+	7	41	48
	-	12	312	324
		19	353	372

PPV:

48 people test positive,
7 of which are truly positive

Therefore: 7/48 = 14.6%

15% of *positive* tests are correct

and 85% are not!

NPV:

324 people test negative,
312 of which are truly negative

Therefore: 312/324 = 96.3%

96% of *negative* tests are correct

and 4% are not

∴ Prevalence *decreases* →

Positive Predictive Value *decreases* & Negative Predictive Value *increases*

Why do the Predictive Values change like this?

As prevalence decreases:

- Absolute numbers of True Positives gets smaller
  - Fewer cases around
- Absolute numbers of False Positives gets bigger
  - More people without the disease

Prevalence = 21.8%		Depressed (GP)		
		+	-	
Screening Tool	+	31	34	65
	-	50	257	307
		81	291	372

Prevalence = 5.1%		Depressed (GP)		
		+	-	
Screening Tool	+	7	41	48
	-	12	312	324
		19	353	372

Tests as part of a screening programme

What is screening?

‘the systematic application of a test or inquiry, to identify individuals at sufficient risk of a specific disorder to warrant further investigation or direct preventive action, amongst persons who have not sought medical attention on account of symptoms of that disorder.’

(Wald, 1994, p.76)

A Screening Programme…

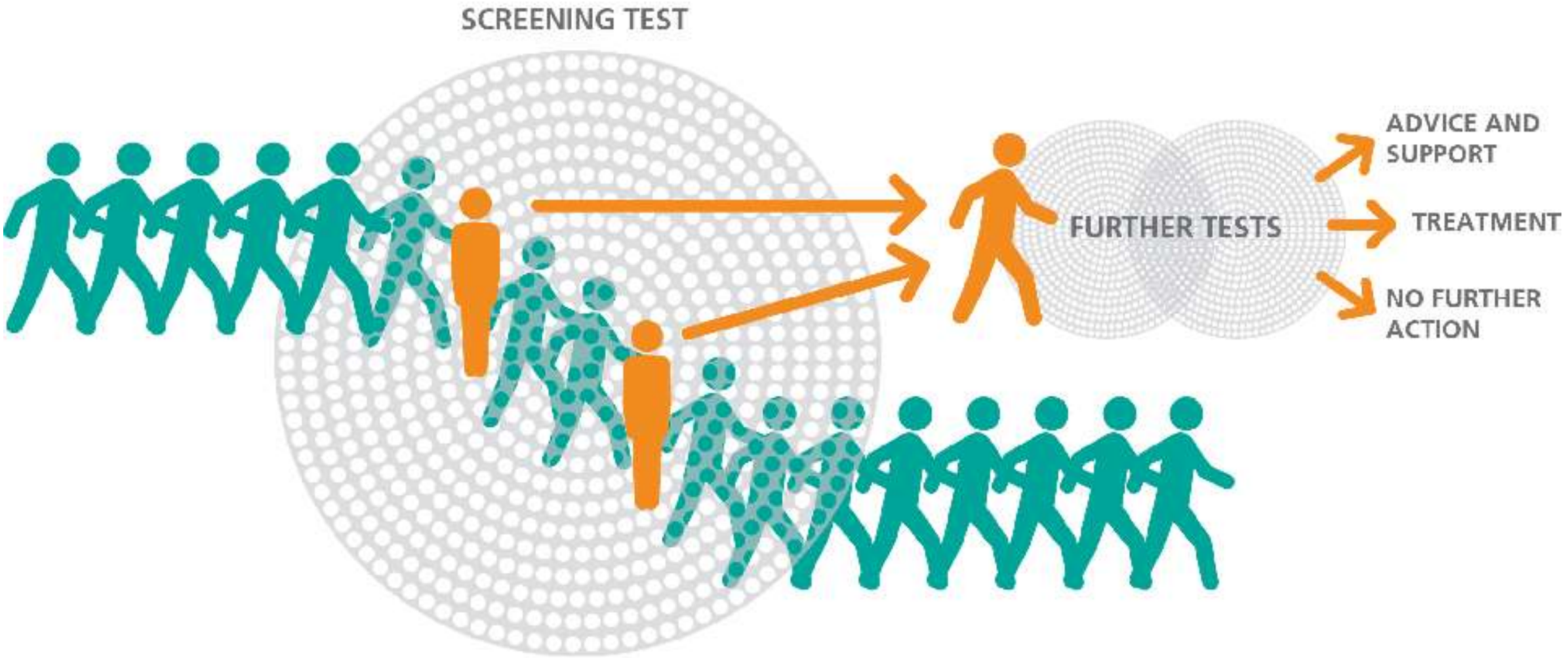


Image source: Population screening explained - GOV.UK

What do we need from tests in a screening programme?

- Identify as many that have condition or could benefit from intervention
  - Sensitivity
- Put forward the right people for the intervention
  - Positive Predictive Value
    - high Specificity

- How to achieve this?
  - Apply different tests with different attributes

Using different tests as part of a programme

- *Multiple tests or Multi-stage testing*
  - Common practice
  - Can tailor requirements of the programme
  - Maximise Sens, Spec, NPV or PPV as required

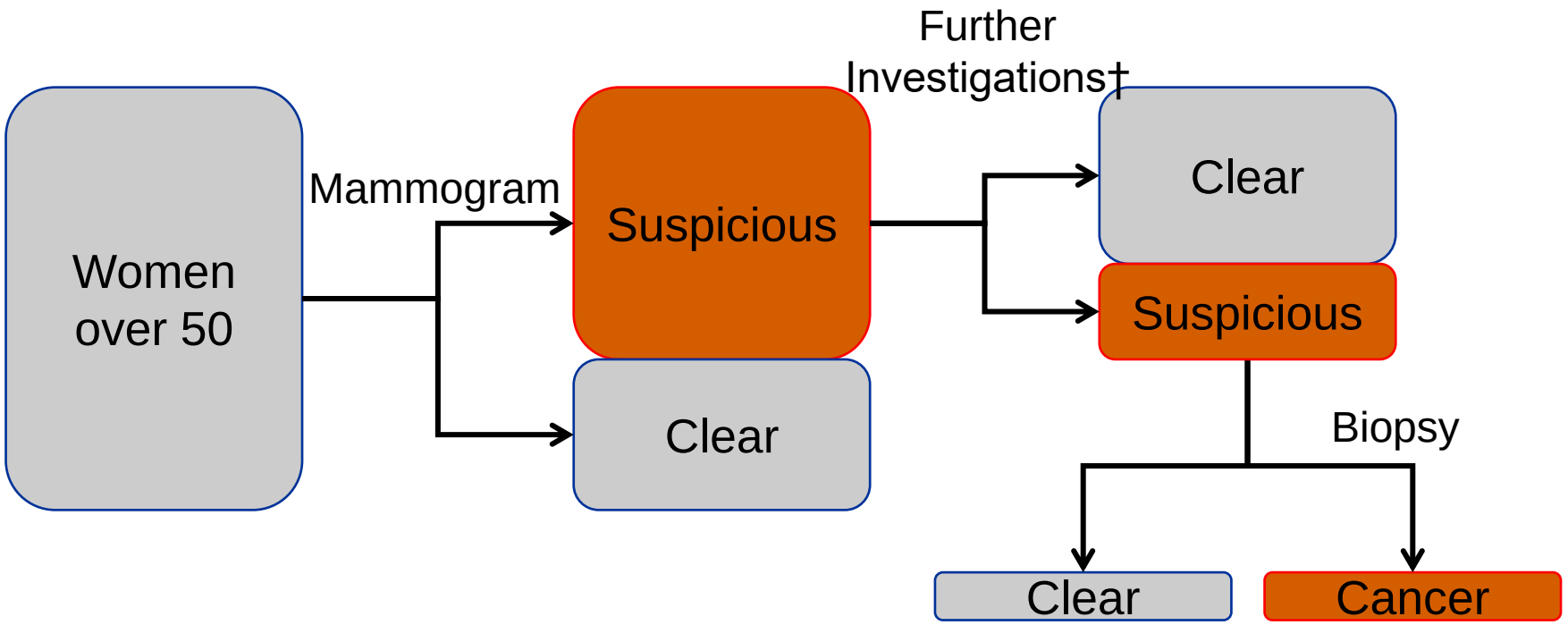
Increased Sensitivity

Increased Specificity

Parallel *Serial*

ID	Test A	Test B	Positive to Either test	Positive to Both tests
1	+	+	+	+
2	+	-	+	-
3	-	-	-	-
4	-	+	+	-
5	+	+	+	+
6	-	-	-	-

Breast Screening – a staged Serial



† More mammograms
Ultrasound

At each stage, the negatives aren't retested

Net result is that to be treated for cancer, patient must 'fail' every test

Summary

- We use Sensitivity and Specificity to describe test performance
  - We can modify performance by adjusting cutoff values
- Positive and Negative Predictive values inform our interpretation of test results
  - Are affected by disease frequency
- For screening programmes we need:
  - Sensitive tests to find as many who would benefit
  - High Positive Predictive Value tests so we only intervene where appropriate
  - Achieved by using tests in tandem

Thank you!

Dr Tom Johnston
Lecturer in Population Health
tom.johnston@hyms.ac.uk

Hull York Medical School

@HullYorkMed

@hullyorkmed