



Part 4

Lecture 3a Diagnostic Accuracy

1



Who I am...

Pascal Tyrrell, PhD

Associate Professor

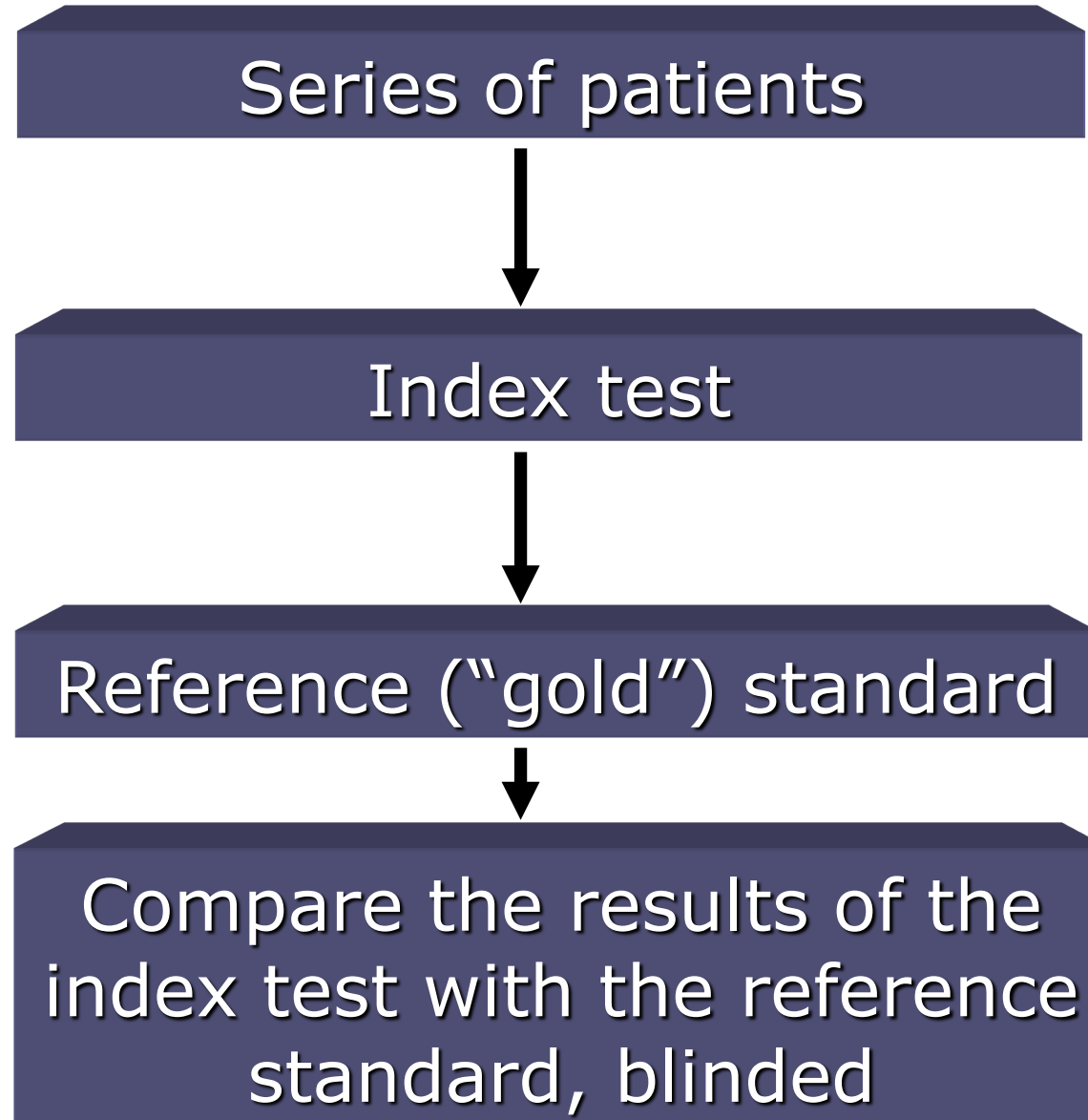
Department of Medical Imaging, Faculty of Medicine

Institute of Medical Science, Faculty of Medicine

Department of Statistical Sciences, Faculty of Arts and Science



Basic structure of diagnostic studies



Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study



Peter Ström*, Kimmo Kartasalo*, Henrik Olsson, Leslie Solórzano, Brett Delahunt, Daniel M Berney, David G Bostwick, Andrew J Evans, David J Grignon, Peter A Humphrey, Kenneth A Iczkowski, James G Kench, Glen Kristiansen, Theodoros H van der Kwast, Katia RM Leite, Jesse K McKenney, Jon Oxley, Chin-Chen Pan, Hemamali Samarasinghe, John R Srigley, Hiroyuki Takahashi, Toyonori Tsuzuki, Murali Varma, Ming Zhou, Johan Lindberg, Cecilia Lindskog, Pekka Ruusuvaara, Carolina Wahlby, Henrik Grönberg, Mattias Rantalainen, Lars Egevad, Martin Eklund

Summary

Background An increasing volume of prostate biopsies and a worldwide shortage of urological pathologists puts a strain on pathology departments. Additionally, the high intra-observer and inter-observer variability in grading can result in overtreatment and undertreatment of prostate cancer. To alleviate these problems, we aimed to develop an artificial intelligence (AI) system with clinically acceptable accuracy for prostate cancer detection, localisation, and Gleason grading.

Methods We digitised 6682 slides from needle core biopsies from 976 randomly selected participants aged 50–69 in the Swedish prospective and population-based STHLM3 diagnostic study done between May 28, 2012, and Dec 30, 2014 (ISRCTN84445406), and another 271 from 93 men from outside the study. The resulting images were used to train deep neural networks for assessment of prostate biopsies. The networks were evaluated by predicting the presence, extent, and Gleason grade of malignant tissue for an independent test dataset comprising 1631 biopsies from 246 men from STHLM3 and an external validation dataset of 330 biopsies from 73 men. We also evaluated grading performance on 87 biopsies individually graded by 23 experienced urological pathologists from the International Society of Urological Pathology. We assessed discriminatory performance by receiver operating characteristics and tumour extent predictions by correlating predicted cancer length against measurements by the reporting pathologist. We quantified the concordance between grades assigned by the AI system and the expert urological pathologists using Cohen's kappa.

Findings The AI achieved an area under the receiver operating characteristics curve of 0.997 (95% CI 0.994–0.999) for distinguishing between benign (n=910) and malignant (n=721) biopsy cores on the independent test dataset and 0.986 (0.972–0.996) on the external validation dataset (benign n=108, malignant n=222). The correlation between cancer length predicted by the AI and assigned by the reporting pathologist was 0.96 (95% CI 0.95–0.97) for the independent test dataset and 0.87 (0.84–0.90) for the external validation dataset. For assigning Gleason grades, the AI achieved a mean pairwise kappa of 0.62, which was within the range of the corresponding values for the expert pathologists (0.60–0.73).

Interpretation An AI system can be trained to detect and grade cancer in prostate needle biopsy samples at a ranking comparable to that of international experts in prostate pathology. Clinical application could reduce pathology workload by reducing the assessment of benign biopsies and by automating the task of measuring cancer length in positive biopsy cores. An AI system with expert-level grading performance might contribute a second opinion, aid in standardising grading, and provide pathology expertise in parts of the world where it does not exist.

Funding Swedish Research Council, Swedish Cancer Society, Swedish eScience Research Center, EIT Health.

Copyright © 2020 Elsevier Ltd. All rights reserved.

Lancet Oncol 2020

Published Online
January 8, 2020
[https://doi.org/10.1016/S1470-2045\(19\)30738-7](https://doi.org/10.1016/S1470-2045(19)30738-7)

See Online/Comment/
[https://doi.org/10.1016/S1470-2045\(19\)30793-4](https://doi.org/10.1016/S1470-2045(19)30793-4)

*These authors contributed equally

Department of Medical Epidemiology and Biostatistics (P Ström MSc, H Olsson MSc, J Lindberg PhD, Prof H Grönberg MD, M Rantalainen PhD, M Eklund PhD) and Department of Oncology and Pathology (Prof L Egevad MD), Karolinska Institutet, Stockholm, Sweden; Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland (K Kartasalo MSc, P Ruusuvaara PhD); Centre for Image Analysis, Department of Information Technology (L Solórzano MSc, Prof C Wahlby PhD) and Department of Immunology, Genetics, and Pathology (C Lindskog PhD), Uppsala University, Uppsala, Sweden; Department of Pathology and Molecular Medicine, Wellington School of Medicine and Health Sciences, University of Otago, Wellington, New Zealand (Prof B Delahunt MD); Barts Cancer Institute, Queen Mary University of London, London, UK (Prof DM Berney MD);



Series of patients: “...We digitised 6682 slides from needle core biopsies from 976 randomly selected participants aged 50–69 in the Swedish prospective and population-based STHLM3 diagnostic study done between May 28, 2012, and Dec 30, 2014 (ISRCTN84445406), and another 271 from 93 men from outside the study...”

Index test: “...The resulting images were used to train deep neural networks for assessment of prostate biopsies. The networks were evaluated by predicting the presence, extent, and Gleason grade of malignant tissue...”

Gold standard: “...an independent test dataset comprising 1631 biopsies from 246 men from STHLM3 and an external validation dataset of 330 biopsies from 73 men. We also evaluated grading performance on 87 biopsies individually graded by 23 experienced urological pathologists from the International Society of Urological Pathology...”

Accuracy: “...The AI achieved an area under the receiver operating characteristics curve of 0·997 (95% CI 0·994–0·999) for distinguishing between benign (n=910) and malignant (n=721) biopsy cores on the independent test dataset and 0·986 (0·972–0·996) on the external validation dataset (benign n=108, malignant n=222). The correlation between cancer length predicted by the AI and assigned by the reporting pathologist was 0·96 (95% CI 0·95–0·97) for the independent test dataset and 0·87 (0·84–0·90) for the external validation dataset. For assigning Gleason grades, the AI achieved a mean pairwise kappa of 0·62, which was within the range of the corresponding values for the expert pathologists (0·60–0·73).”

Series of patients



Index test



Reference ("gold") standard



Compare the results of the index test with the reference standard

Men 50-69 with suspected prostate cancer

AIML classifier:
Malignant Y/N

Board certified urological pathologist

External test set of 330 cores:
AUC of 98.6%

What we will cover

- ❑ Diagnostic reasoning
- ❑ Basic design of diagnostic studies
- ❑ Appraising a diagnostic study in 3 easy steps
- ❑ What do all the numbers mean?



Appraising diagnostic tests: 3 easy steps

1. Are the results valid?



2. What are the results?



**3. Will they help me
look after my patients?**

Appraising diagnostic tests: 3 easy steps

1. Are the results valid?



2. What are the results?



**3. Will they help me
look after my patients?**

- Appropriate spectrum of patients?
- Does everyone get the gold standard?
- Is there an independent, blind or objective comparison with the gold standard?

Appropriate spectrum of patients?

- Ideally, test should be performed on group of patients in whom it will be applied in the real world clinical setting

Series of patients: “...We digitised 6682 slides from needle core biopsies from 976 randomly selected participants aged 50–69 in the Swedish prospective and population-based STHLM3 diagnostic study done between May 28, 2012, and Dec 30, 2014 (ISRCTN84445406), and another 271 from 93 men from outside the study...”

***ALL* patients have the gold standard?**

- Ideally all patients get the gold /reference standard test

Gold standard: “...an independent test dataset comprising 1631 biopsies from 246 men from STHLM3 and an external validation dataset of 330 biopsies from 73 men. We also evaluated grading performance on 87 biopsies individually graded by 23 experienced urological pathologists from the International Society of Urological Pathology...”

Independent, blind or objective comparison with the gold standard?

- ❑ Ideally, the gold standard is independent, blind and objective

Empirical Evidence of Design-Related Bias in Studies of Diagnostic Tests

Jeroen G. Lijmer, MD

Ben Willem Mol, MD, PhD

Siem Heisterkamp, PhD

Gouke J. Bonsel, MD, PhD

Martin H. Prins, MD, PhD

Jan H. P. van der Meulen, MD, PhD

Patrick M. M. Bossuyt, PhD

DURING RECENT DECADES, THE number of available diagnostic tests has been rapidly increasing. As for all new medical technologies, new diagnostic tests should be thoroughly evaluated prior to their introduction into daily practice. The number of test evaluations in the literature is increasing but the methodological quality of these studies is on av-

Context The literature contains a large number of potential biases in the evaluation of diagnostic tests. Strict application of appropriate methodological criteria would invalidate the clinical application of most study results.

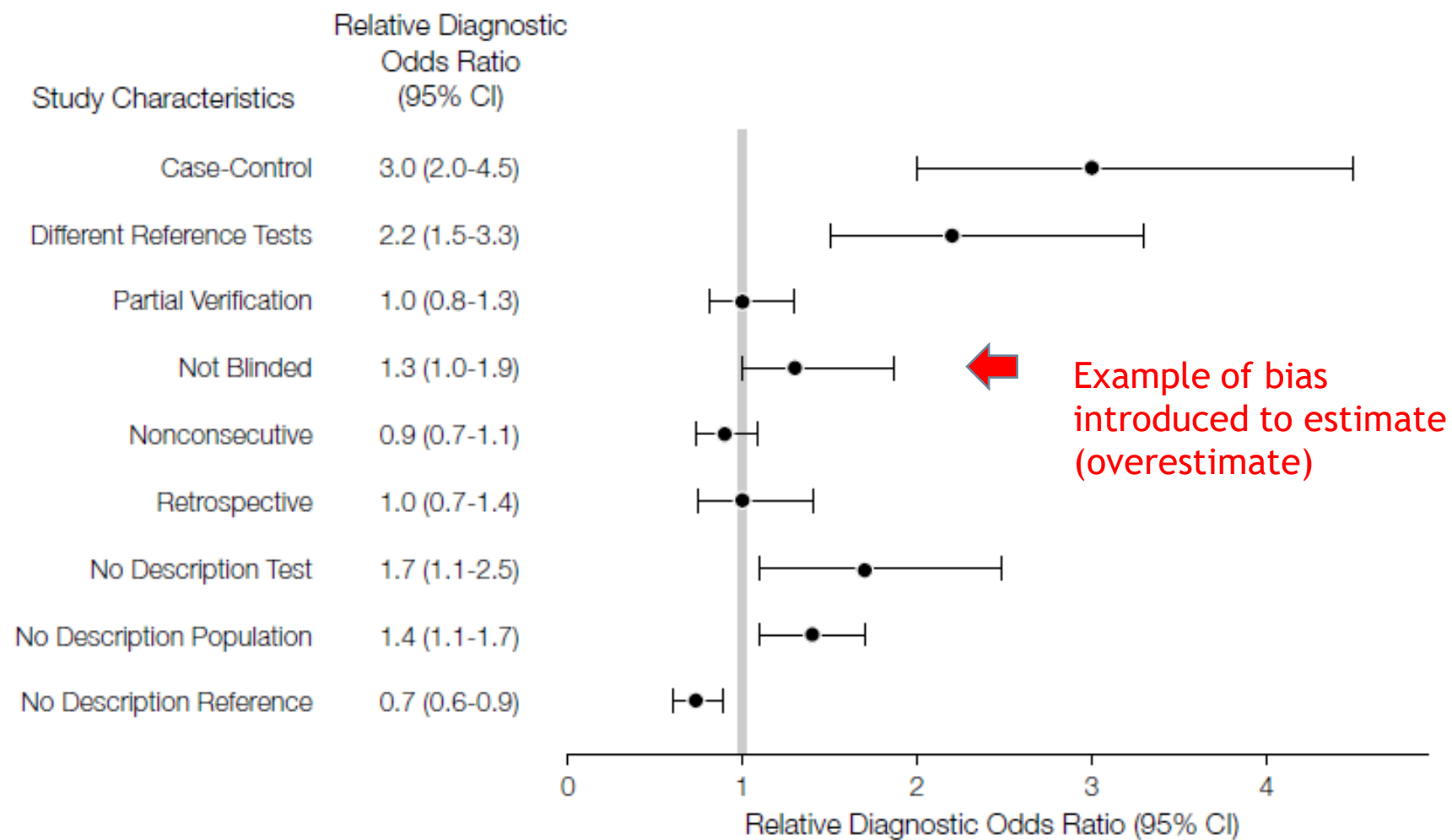
Objective To empirically determine the quantitative effect of study design shortcomings on estimates of diagnostic accuracy.

Design and Setting Observational study of the methodological features of 184 original studies evaluating 218 diagnostic tests. Meta-analyses on diagnostic tests were identified through a systematic search of the literature using MEDLINE, EMBASE, and DARE databases and the Cochrane Library (1996-1997). Associations between study characteristics and estimates of diagnostic accuracy were evaluated with a regression model.

Main Outcome Measures Relative diagnostic odds ratio (RDOR), which compared the diagnostic odds ratios of studies of a given test that lacked a particular methodological feature with those without the corresponding shortcomings in design.

Results Fifteen (6.8%) of 218 evaluations met all 8 criteria; 64 (30%) met 6 or more. Studies evaluating tests in a diseased population and a separate control group overestimated the diagnostic performance compared with studies that used a clinical population (RDOR, 3.0; 95% confidence interval [CI], 2.0-4.5). Studies in which different reference tests were used for positive and negative results of the test under study overestimated the diagnostic performance compared with studies using a single reference

Figure. Relative Diagnostic Odds Ratios and 95% Confidence Intervals (CIs) of the 9 Study Characteristics Examined With a Multivariate Regression Analysis



RDORs indicate the diagnostic performance of a test in studies failing to satisfy the methodological criterion, relative to its performance in studies with the corresponding feature

Appraising diagnostic tests: 3 easy steps

1. Are the results valid?



2. What are the results?



**3. Will they help me
look after my patients?**

- Sensitivity, specificity
- Likelihood ratios
- Predictive values

Appraising diagnostic tests: 3 easy steps

1. Are the results valid?



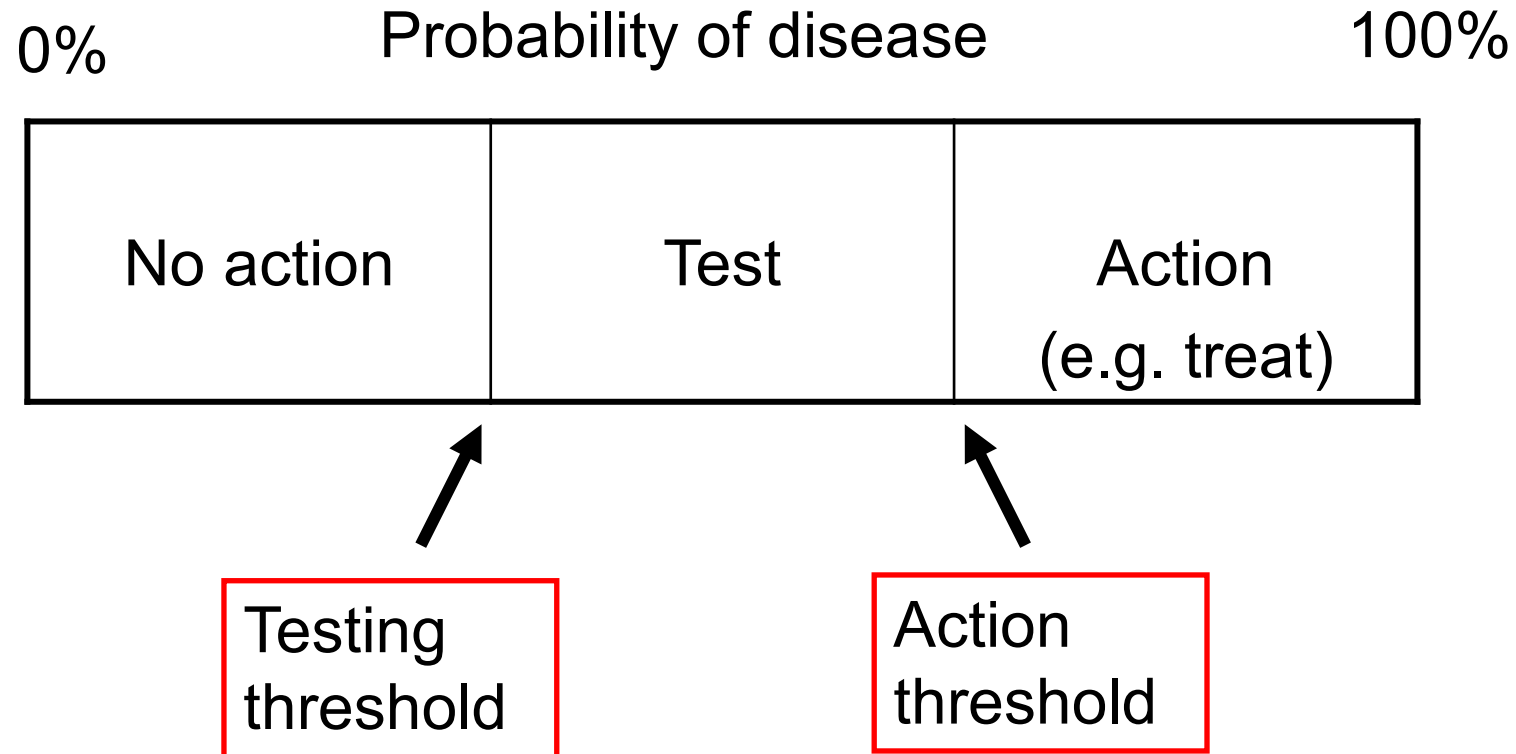
2. What are the results?



**3. Will they help me
look after my patients?**

- Can I do the test in my setting?
- Do results apply to the mix of patients I see?
- Will the result change my management?
- Costs to patient/health service?

Will the result change management?



Will the test apply in my setting?

- ☐ Will the results change my management?
- ☐ Reproducibility of the test and interpretation in my setting
- ☐ Impact on outcomes that are important to patients?
- ☐ Where does the test fit into the diagnostic strategy?
- ☐ Costs to patient/health service?

- In the **ideal** world, a test would have perfect discrimination...

i.e. all the patients who HAVE the disease are identified by the test

AND all the patients who DO NOT have the disease have a negative test

2 by 2 table

		Disease	
		+	-
Test	+		
	-		

2 by 2 table


		Disease	
		+	-
Test	+	True positives	
	-		True negatives

2 by 2 table

		Disease	
		+	-
Test	+	True positives	False positives
	-	False negatives	True negatives

2 by 2 table: sensitivity

		Disease	
		+	-
Test	+	a True positives	
	-	c False negatives	




$$\text{Sensitivity} = a / a + c$$

Proportion of people with the disease who have a positive test result.

So, a test with 84% sensitivity....means that the test identifies 84 out of 100 people **WITH** the disease

2 by 2 table: sensitivity

		Disease	
		+	-
Test	+	84	
	-	16	



$$\text{Sensitivity} = 84 / 100$$

Proportion of people with the disease who have a positive test result.

So, a test with 84% sensitivity....means that the test identifies 84 out of 100 people **WITH** the disease

2 by 2 table: specificity

		Disease	
		+	-
Test	+		b False positives
	-		d True negatives

Proportion of people without the disease who have a negative test result.

$$\text{Specificity} = d / b + d$$

2 by 2 table: specificity

		Disease	
		+	-
Test	+		b 25
	-		d 75

Proportion of people without the disease who have a negative test result.

So, a test with 75% specificity will be **NEGATIVE** in 75 out of 100 people without the disease

$$\text{Specificity} = 75/100$$

Tip....

- ❑ Sensitivity is useful to me
 - ❑ 'The new rapid COVID19 test was positive in 47 out of 56 persons with COVID19 (sensitivity = 83.9%)'

- ❑ Specificity seems a bit confusing!
 - ❑ 'The new rapid COVID19 test was negative in 600 of the 607 persons who did not have COVID19 (specificity = 98.8%)'

- ❑ So...the false positive rate is sometimes easier
 - ❑ False positive rate = $1 - \text{specificity}$
 - ❑ So a specificity of 98.8% means that the new rapid test is wrong (or falsely positive) in 1.2% of people tested

2 by 2 table: specificity

		COVID19 Lab PCR	
		+	-
rapid test nasal swab	+		7
	-		600
			607

*There were 607
persons who did not
have COVID19...
the rapid test was
falsely positive in 7
of them*

$$\text{Specificity} = 600/607 = 98.8\%$$

$$\text{False positive rate} = 1 - \text{specificity} = 1.2\%$$