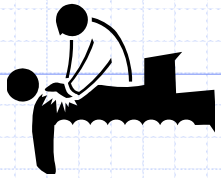
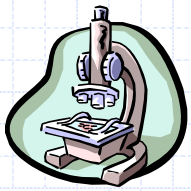
Diagnostic test accuracy design





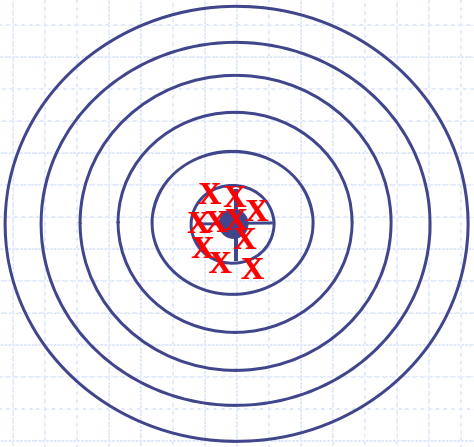
Madhukar Pai, MD, PhD
Assistant Professor of Epidemiology, McGill University
Montreal, Canada

Email: madhukar.pai@mcgill.ca

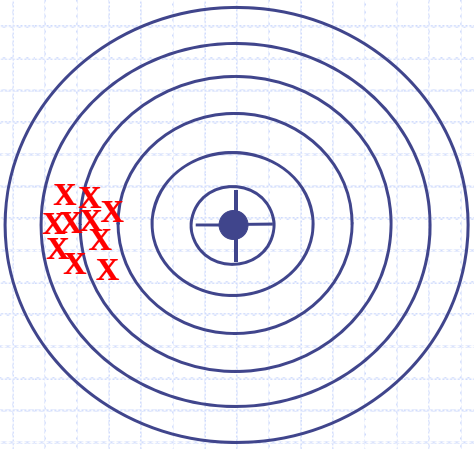
Two key properties of any test

- Accuracy (also called 'validity')
- Precision (also called 'reliability' or 'reproducibility')

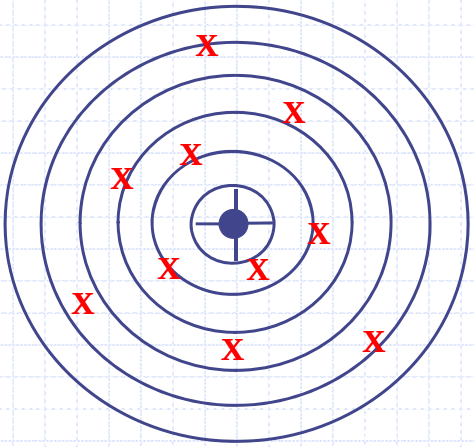
Precision and Accuracy



Precision and Accuracy



Precision and Accuracy



Quantifying precision/reliability

Observer Variation

- **Intraobserver agreement**

  Does the same clinician get the same result when repeating a symptom or sign on a patient who is clinically unchanged?

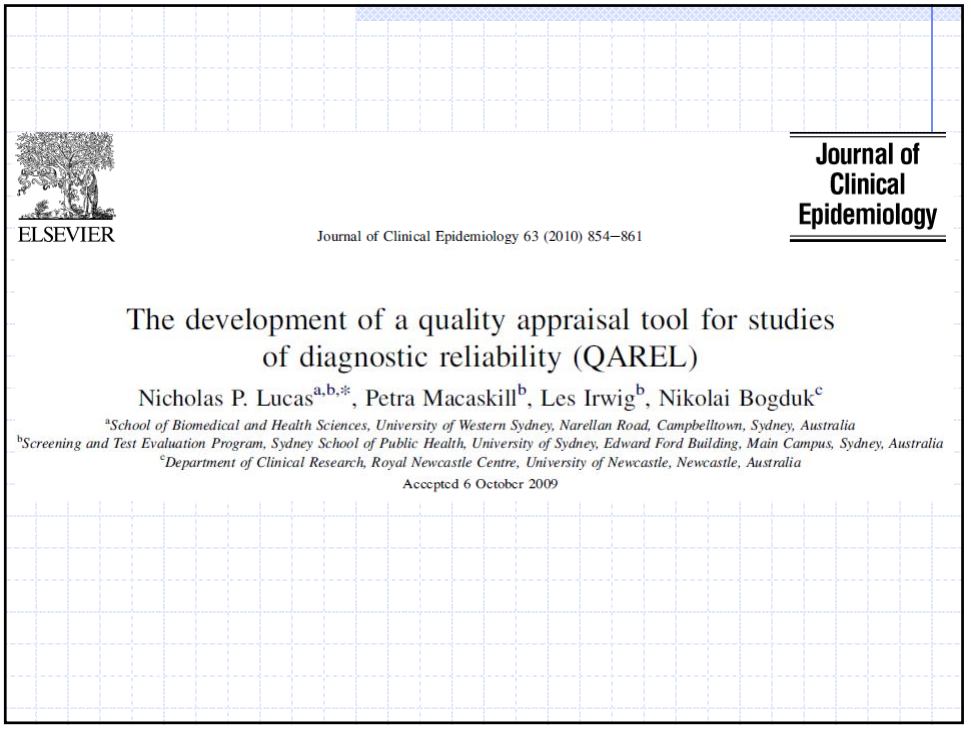
- **Interobserver agreement**

  Do 2 or more observers agree on the presence or absence of a finding in a patient who experienced no change in condition?

- **Kappa (κ)**

  Agreement beyond chance and can be used to describe both intra- and interobserver agreement

  Note: Other measures are used for continuous measurements (e.g. correlation coefficient, limits of agreement, etc)



ELSEVIER

Journal of Clinical Epidemiology 63 (2010) 854–861

Journal of Clinical Epidemiology

The development of a quality appraisal tool for studies of diagnostic reliability (QAREL)

Nicholas P. Lucas[a,b,*], Petra Macaskill[b], Les Irwig[b], Nikolai Bogduk[c]

[a]*School of Biomedical and Health Sciences, University of Western Sydney, Narellan Road, Campbelltown, Sydney, Australia*
[b]*Screening and Test Evaluation Program, Sydney School of Public Health, University of Sydney, Edward Ford Building, Main Campus, Sydney, Australia*
[c]*Department of Clinical Research, Royal Newcastle Centre, University of Newcastle, Newcastle, Australia*

Accepted 6 October 2009

Quantifying accuracy

- **Sensitivity and Specificity**
- **Likelihood ratios**
- **Positive and Negative Predictive Value**
- **Diagnostic Odds Ratio**

Tests with dichotomous results

A standard Phase II/III diagnostic design for accuracy estimation

- Define gold standard
- Recruit consecutive patients in whom the test is indicated (in whom the disease is suspected)
- Perform gold standard and separate diseased and disease free groups
- Perform test on all and classify them as test positives or negatives
- Set up 2 x 2 table and compute:
  - Sensitivity
  - Specificity
  - Predictive values
  - Likelihood ratios
  - Diagnostic odds ratio

Evaluating a diagnostic test

•Diagnostic 2 X 2 table*:

	Disease +	Disease -
Test +	True Positive	False Positive
Test -	False Negative	True Negative

*When test results are not dichotomous, then can use ROC curves [see later]

Sensitivity [true positive rate]

	Disease present	Disease absent
Test positive	True positives	False positives
Test negative	False negative	True negatives

The proportion of patients with disease who test positive = P(T+|D+) = TP / (TP+FN)

Specificity [true negative rate]

	Disease present	Disease absent
Test positive	True positives	False positives
Test negative	False negative	True negatives

The proportion of patients without disease who test negative: P(T-|D-) = TN / (TN + FP).

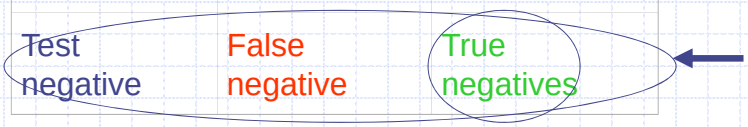
Predictive value of a positive test

	Disease present	Disease absent
Test positive	True positives	False positives
Test negative	False negative	True negatives

Proportion of patients with positive tests who have disease = P(D+|T+) = TP / (TP+FP)

Predictive value of a negative test

	Disease present	Disease absent
Test positive	True positives	False positives
Test negative	False negative	True negatives



Proportion of patients with negative tests who do not have disease = P(D-|T-) = TN / (TN+FN)

Example: Serological test for TB

		Culture (gold standard) Yes	No	
Serological Test	Positive	14	3	17
	Negative	54	28	82
		68	31	99

Sensitivity = 21%
Specificity = 90%

Clin Vacc Immunol 2006;13:702-03

All accuracy measures must be reported with confidence intervals!!

Sensitivity 20.6% (95%CI 12.7, 31.6)

Specificity 90.3% (75.1, 96.7)

Positive Predictive Value 82.4% (58.9, 93.8)

Negative Predictive Value 34.2% (24.8, 44.9)

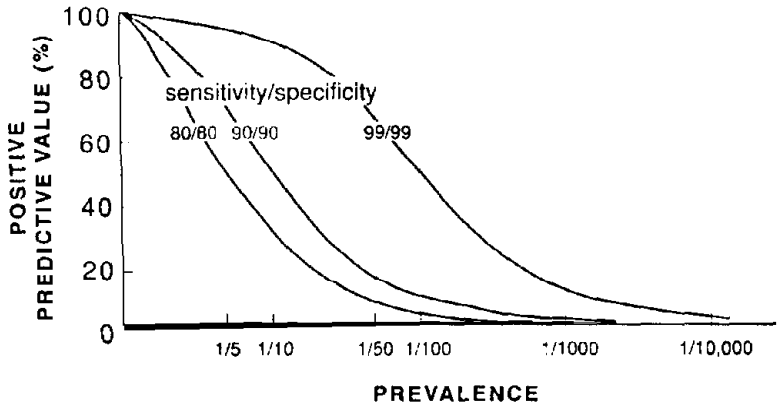
For a given test, predictive values will depend on prevalence

Effect of Prevalence on Predictive Value: Positive Predictive Value of Prostatic Acid Phosphatase for Prostatic Cancer (Sensitivity = 70%, Specificity = 90%) in Various Clinical Settiings*

Setting	Prevalence (Cases/100,000)	Postive Predictive Value (%)
General population	35	0.4
Men, age 75 or greater	500	5.6
Clinically suspicious prostatic nodule	50,000	93.0

* From: Watson RA, Tang DB. *N Engl J Med*, 1980; 303:497–499.

For a given test, predictive values will depend on prevalence



Positive predictive value according to sensitivity, specificity, and prevalence of disease.

Fletcher 1996

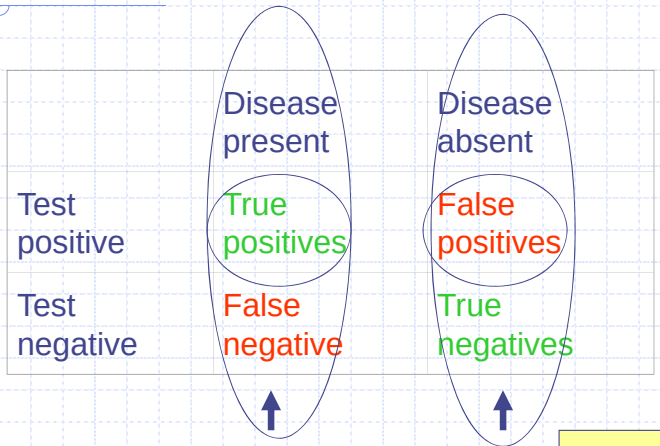
Likelihood Ratios (also called 'Bayes Factor')

- Likelihood ratio of a positive test: is the test more likely to be positive in diseased than non-diseased persons?
  - LR+ = TPR / FPR

$$LR+ = \frac{\Pr(T+ \mid D+)}{\Pr(T+ \mid D-)}$$

  - High LR+ values help in RULING IN the disease
  - Values close to 1 indicate poor accuracy
  - E.g. LR+ of 10 means a diseased person is 10 times more likely to have a positive test than a non-diseased person

Likelihood Ratio of a Positive Test



How more often a positive test result occurs in persons with compared to those without the target condition

LR+ = TPR / FPR

$$LR+ = \frac{\Pr(T+ \mid D+)}{\Pr(T+ \mid D-)}$$

Likelihood Ratios

- Likelihood ratio of a negative test: is the test less likely to be negative in the diseased than non-diseased persons?

- LR- = FNR / TNR

$$LR- = \frac{\Pr(T- \mid D+)}{\Pr(T- \mid D-)}$$

  - Low LR- values help in RULING OUT the disease
  - Values close to 1 indicate poor accuracy
  - E.g. LR- of 0.5 means a diseased person is half as likely to have a negative test than a non-diseased person

Likelihood Ratio of a Negative Test

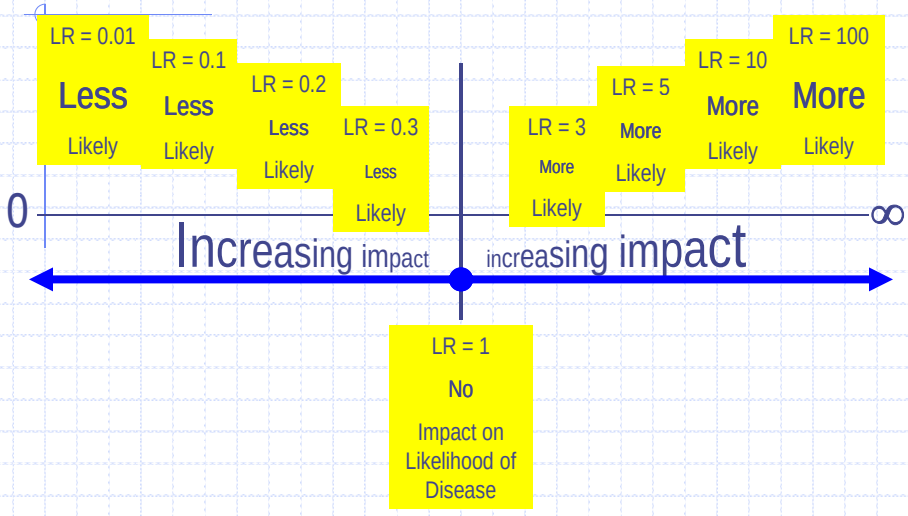
	Disease present	Disease absent
Test positive	True positives	False positives
Test negative	False negative	True negatives

How less likely a negative test result is in persons with the target condition compared to those without the target condition

LR- = FNR / TNR

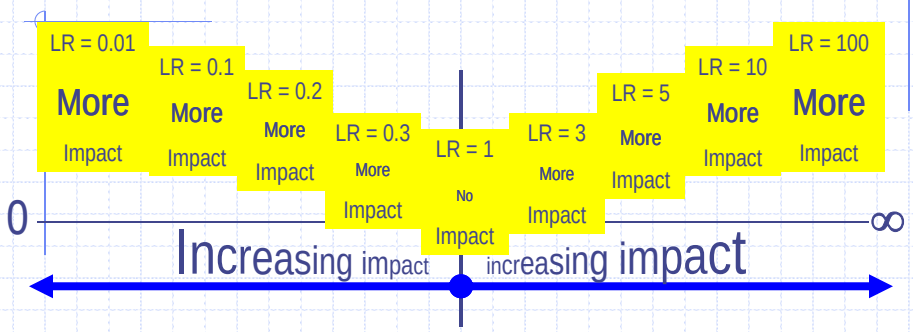
$$LR- = \frac{\Pr(T- \mid D+)}{\Pr(T- \mid D-)}$$

LR: Impact on Likelihood of Disease



The Rational Clinical Examination

LR: Impact on Likelihood of Disease



Quick review of odds vs. probability

- odds = probability / (1 – probability)

$$\text{Odds}(D+) = \frac{\Pr(D+)}{1-\Pr(D+)}$$

- probability = odds / (1 + odds)

$$\Pr(D+) = \frac{\text{Odds}(D+)}{1+\text{Odds}(D+)}$$

Diagnostic Odds Ratio (DOR)

	Disease present	Disease absent
Test positive	True positives (a)	False positives (b)
Test negative	False negative (c)	True negatives (d)

Odds of positive test result in persons with the target condition compared to those without the target condition

DOR = (a/c) / (b/d)
DOR = ad / bc
DOR = Odds of T+|D+ / Odds of T+|D-

Example: Serological test for TB

		Culture (gold standard) Yes	No	
Serological Test	Positive	14	3	17
	Negative	54	28	82
		68	31	99

LR+ = 2
LR- = 0.9
DOR = 2.4

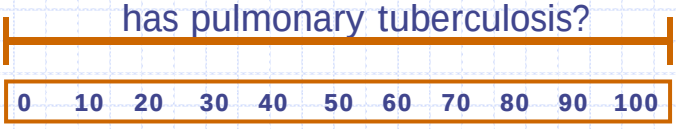
Clin Vacc Immunol 2006;13:702-03

Using LRs in practice

- Scenario:
  - Mr. A, a 27-year old male factory worker
  - Fever and productive cough for the past 3 weeks
  - Lost weight

Assess the patient and estimate the baseline risk (pre-test probability)

Based on initial history, how likely is it that Mr. A has pulmonary tuberculosis?

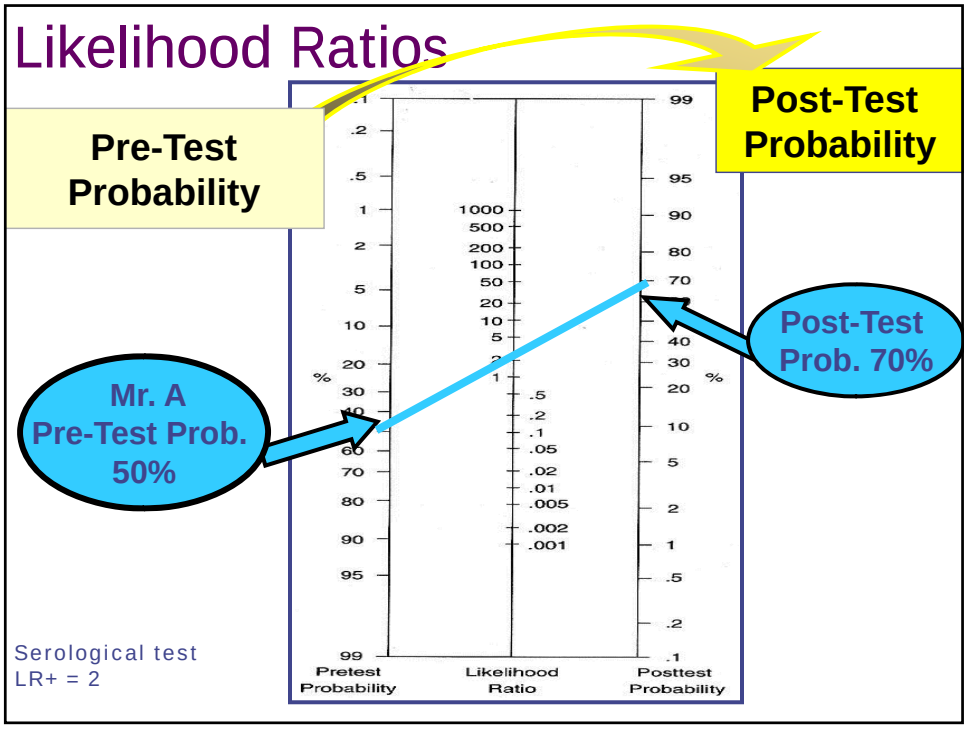


Pre-Test Probability

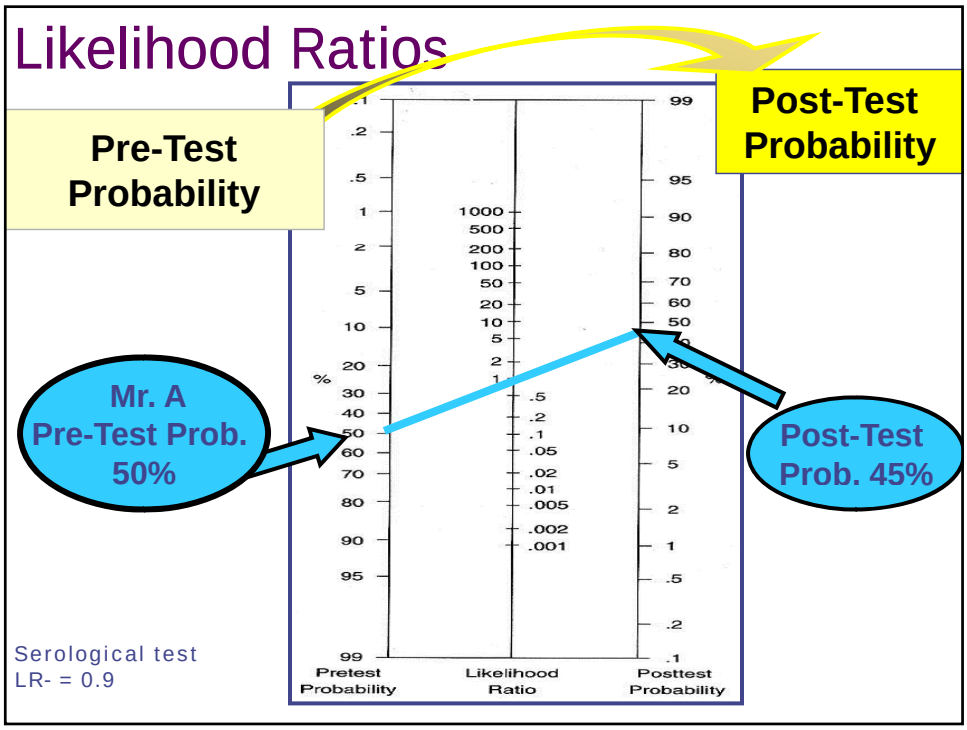
How might the result of a serological test change the likelihood of TB in this patient?

Post-Test Probability

Likelihood Ratios



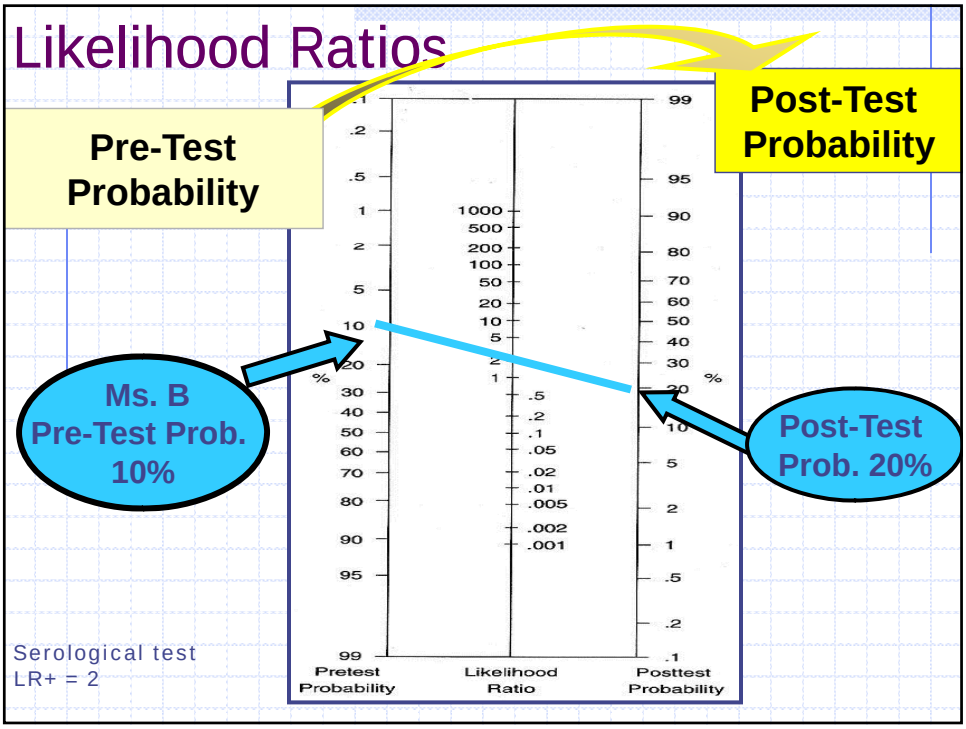
Likelihood Ratios



Using LRs in practice

- Scenario:
  - Ms. B, a 18 year old engineering student
  - Fever and non-productive cough for the past 4 days
  - Nobody in the household has had TB

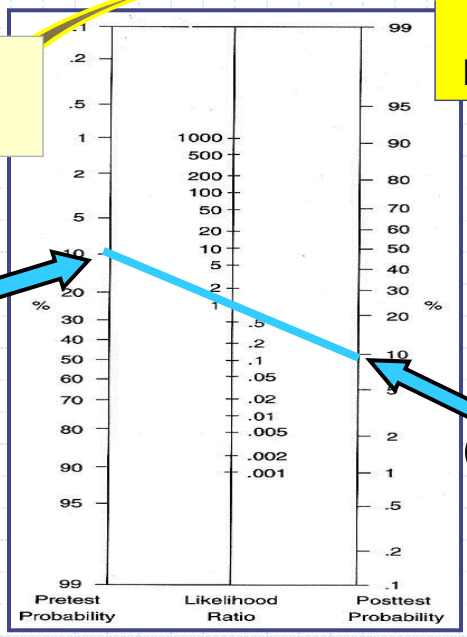
Likelihood Ratios



Likelihood Ratios

Pre-Test Probability

Post-Test Probability

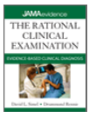


Ms. B Pre-Test Prob. 10%

Post-Test Prob. 10%

Serological test
LR- = 0.9

Where do we get LRs from?



The Rational Clinical Examination: Evidence-Based Clinical Diagnosis >
Pretest Probabilities and Likelihood Ratios for Clinical Findings
Quick Reference Online Only
http://jamaevidence.com

Note: Large images and tables on this page may necessitate printing in landscape mode.

The Rational Clinical Examination > Pretest Probabilities and Likelihood Ratios for Clinical Findings >

Quick Reference

+ Add to my saved tables

Pretest Probabilities and Likelihood Ratios for Clinical Findings

	Prior Probability	Test/Finding	LR+	LR−
Chapter 1: Primer on Precision and Accuracy				
Chapter 2: Abdominal Aortic Aneurysm	Occur in 4% to 8% of older men. The prevalence in older women is less than 2%.	Physical examination for aneurysm > 4.0 cm	16 (8.6–29)	0.51 (0.38–0.67)
		Physical examination for aneurysm > 3.0 cm	12 (7.4–20)	0.72 (0.65–0.81)
Chapter 3:	Approximately 1% to 5% of the general population	Systolic–diastolic bruit	39 (10–145)	0.62 (0.49–0.73)

The Rational Clinical Examination

Epidemiology 3

Refining clinical diagnosis with likelihood ratios

Lancet 2005; 365: 1500–05

David A Grimes, Kenneth F Schulz

Family Health International, PO Box 13950, Research Triangle Park, NC 27709, USA (D A Grimes MD, K F Schulz PhD)

Correspondence to: Dr David A Grimes dgrimes@fhi.org

Likelihood ratios can refine clinical diagnosis on the basis of signs and symptoms; however, they are underused for patients' care. A likelihood ratio is the percentage of ill people with a given test result divided by the percentage of well individuals with the same result. Ideally, abnormal test results should be much more typical in ill individuals than in those who are well (high likelihood ratio) and normal test results should be most frequent in well people than in sick people (low likelihood ratio). Likelihood ratios near unity have little effect on decision-making; by contrast, high or low ratios can greatly shift the clinician's estimate of the probability of disease. Likelihood ratios can be calculated not only for dichotomous (positive or negative) tests but also for tests with multiple levels of results, such as creatine kinase or ventilation-perfusion scans. When combined with an accurate clinical diagnosis, likelihood ratios from ancillary tests improve diagnostic accuracy in a synergistic manner.

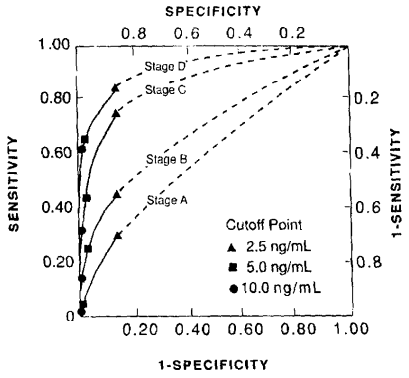
Are sens/spec and LRs inherent properties of a test?

- Most textbooks will say that sens and spec do not depend on disease prevalence
- This is not true
- In reality, sens/spec and LRs vary across populations because of differences in disease spectra (case-mix) and several other factors
- This is equivalent to "effect modification" in epidemiology

Example

Sens and Spec across populations

Ex:
Sensitivity+specificity of serum CEA For detection
of colorectal cancer, across stages

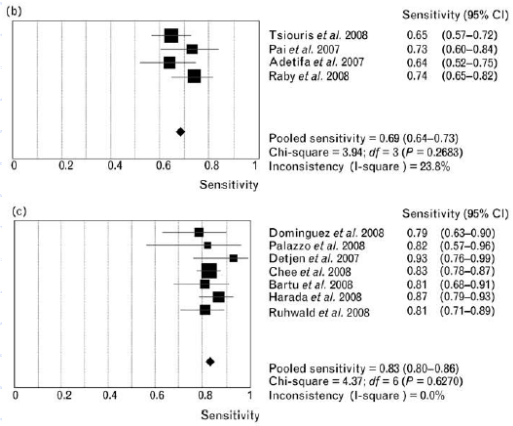


ROC curve for CEA as a diagnostic test for colorectal cancer, according to stage of disease. The sensitivity and specificity of a test vary with the stage of disease. (Redrawn from Fletcher RH. Carcinoembryonic antigen. Ann Intern Med 1986;104:66–73.)

Variation in performance in high vs low endemic countries: example

T-cell interferon-γ release assays for the rapid immunodiagnosis of tuberculosis: clinical utility in high-burden vs. low-burden settings

Keertan Dheda[a,b,c], Richard van Zyl Smit[a], Motasim Badri[a] and Madhukar Pai[d]



High incidence countries

Low incidence countries

40

Tests with continuous or multi-level results

Example: WBC count in bacteremia

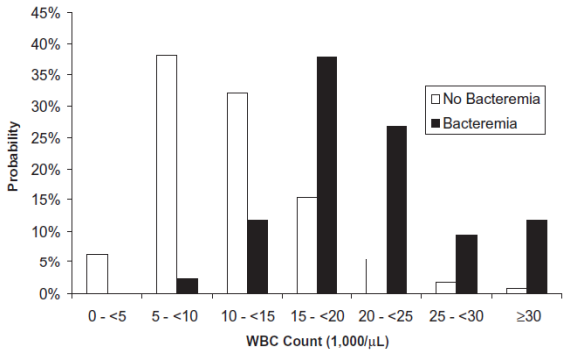


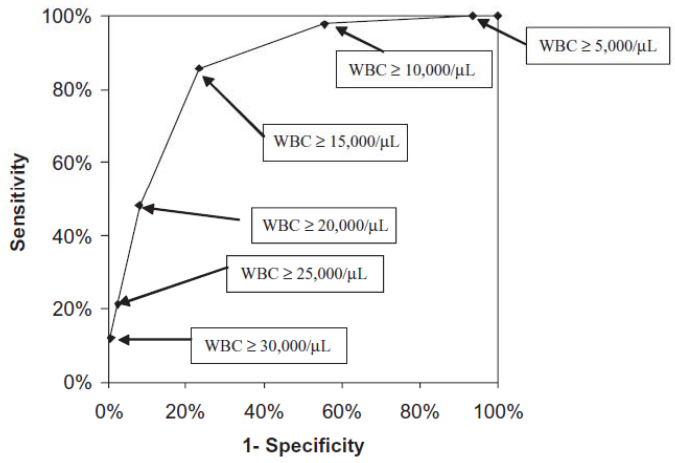
Figure 4.4 Histogram showing distributions of the nonbacteremic and bacteremic populations across the WBC count intervals.

Newman T, Kohn MA. Evidence-based diagnosis. 2009, Cambridge Univ Press

Table 4.3. Sensitivity and specificity of the WBC count as a predictor of bacteremia at different cut-offs for considering the test "positive" (data from Lee and Harper 1998)

WBC count interval (×1,000/μL)	Percent of bacteremia patients in interval	Percent of no bacteremia patients in interval	Sensitivity (using bottom of interval as cut-off)	1 − Specificity (using bottom of interval as cut-off)
≥30	11.8%	0.8%	11.8%	0.8%
25 to <30	9.4%	1.8%	21.3%	2.6%
20 to <25	26.8%	5.4%	48.0%	8.0%
15 to <20	37.8%	15.5%	85.8%	23.5%
10 to <15	11.8%	32.1%	97.6%	55.6%
5 to <10	2.4%	38.1%	100%	93.7%
0 to <5	0.0%	6.3%	100%	100%

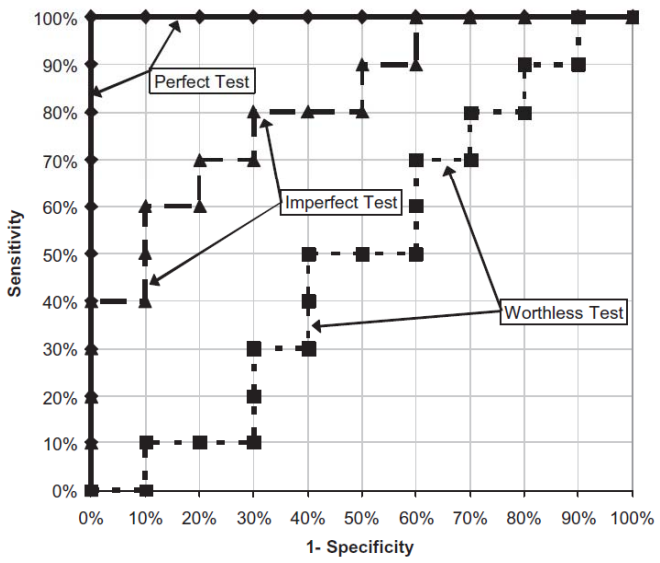
Newman T, Kohn MA. Evidence-based diagnosis. 2009, Cambridge Univ Press



Area Under Curve (AUC) = 0.86

Figure 4.5 ROC curve corresponding to the distributions in Figure 4.4.

Newman T, Kohn MA. Evidence-based diagnosis. 2009, Cambridge Univ Press



Newman T, Kohn MA. Evidence-based diagnosis. 2009, Cambridge Univ Press

Multi-level likelihood ratios

Table 4.4. Likelihood ratios for WBC and bacteremia (from Lee and Harper 1998)

WBC Count (×1,000/μL)	Bacteremia	No bacteremia	LR
30–35	11.8%	0.8%	15.2
25–30	9.4%	1.8%	5.3
20–25	26.8%	5.4%	4.9
15–20	37.8%	15.5%	2.4
10–15	11.8%	32.1%	0.37
5–10	2.4%	38.1%	0.06
0–5	0.0%	6.3%	0.00

Newman T, Kohn MA. Evidence-based diagnosis. 2009, Cambridge Univ Press

Using ROCs to compare tests

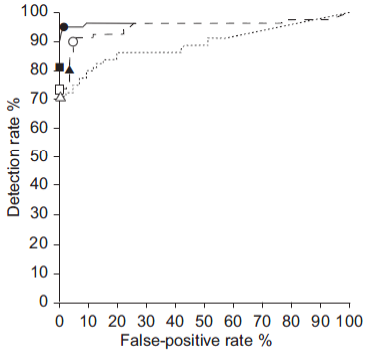


FIGURE 2. Whole blood was stimulated with *Mycobacterium tuberculosis*-specific antigens or saline. The diagnostic potential of interferon (IFN)-γ, IFN-γ inducible protein (IP)-10 and monocyte chemotactic protein (MCP)-2 was determined by receiver operating characteristic curve analysis using antigen-dependent values. Students were used as gold standard for noninfected; tuberculosis patients were used as gold standard for infected. ●: maximum Youden's index (YI) for IFN-γ (4 pg·mL^{-1}); △: maximum YI for MCP-2 (97 pg·mL^{-1}); ■: cut-off applied in the QuantiFERON In-Tube test (Cellestis, Carnagie, Australia; 17.5 pg·mL^{-1}); ○: maximum YI for IP-10 test, used as cut-off for the IP-10 test (237 pg·mL^{-1}); □: cut-off for the IP-10 test (673 pg·mL^{-1}); ▲: selected pragmatic cut-off for the IP-10 test (455 pg·mL^{-1}). ——: IFN-γ; – – – –: IP-10; ····: MCP-2.

Ruhwald, ERJ 2008

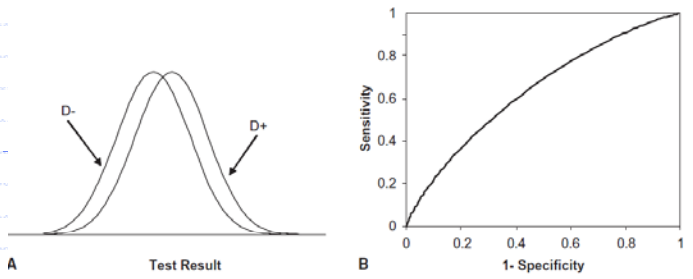


Figure 4.2 Test discriminates poorly between patients with disease (D+) and patient without disease (D−). **(A)** The distribution of test results in D+ patients is very similar to the distribution in D− patients. **(B)** This "bad" ROC curve approaches a 45-degree diagonal line.

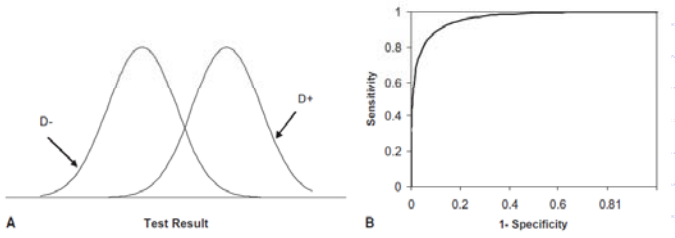


Figure 4.3 Test discriminates well between patients with the disease (D+) and patients without the disease (D−). **(A)** The distribution of test results in D+ patients differs substantially from the distribution in D− patients. **(B)** This "good" ROC curve nears the upper left corner of the grid.

Newman T, Kohn MA. Evidence-based diagnosis. 2009, Cambridge Univ Press

After understanding ROC curves, it should be obvious that

- the case of a dichotomous test accuracy (i.e. the usual 2 x 2 table) is merely a single point on some underlying ROC curve
- in other words, all tests have some underlying ROC curve
- we can easily change the sens/spec by shifting the point on the ROC curve

ROC: pros and cons

- Pros:
  - Provides a wholistic picture (a global assessment of a test's accuracy)
  - Not dependent on disease prevalence
  - Does not force us to pick a single cut-off point
  - Shows the trade off between sens and spec
  - Great for comparing accuracy of competing tests
  - Can be applied to any diagnostic system: weather forecasting, lie detectors, medical imaging, to detection of cracks in metals!

ROC: pros and cons

- Cons:
  - Not very intuitive for clinicians; the ROC and AUC cannot be directly used for any given patient
  - Clinicians prefer simple yes/no test results
  - You can have the same AUC, but different shapes
  - Does not fit into the EBM framework of working with LRs and probabilities
  - Very hard to meta-analyze

Articles

Measuring the Accuracy of Diagnostic Systems

JOHN A. SWETS

Diagnostic systems of several kinds are used to distinguish between two classes of events, essentially "signals" and "noise." For them, analysis in terms of the "relative operating characteristic" of signal detection theory provides a precise and valid measure of diagnostic accuracy. It is the only measure available that is uninfluenced by decision biases and prior probabilities, and it places the performances of diverse systems on a common, easily interpreted scale. Representative values of this measure are reported here for systems in medical imaging, materials testing, weather forecasting, information retrieval, polygraph lie detection, and aptitude testing. Though the measure itself is sound, the values obtained from tests of diagnostic systems often require qualification because the test data on which they are based are of unsure quality. A common set of problems in testing is faced in all fields. How well these problems are handled, or can be handled in a given field, determines the degree of confidence that can be placed in a measured value of accuracy. Some fields fare much better than others.

one or another inadequate or misleading way, a good way is available for general use. The preferred way quantifies accuracy independently of the relative frequencies of the events (conditions, objects) to be diagnosed ("disease" and "no disease" or "rain" and "no rain," for instance) and also independently of the diagnostic system's decision bias, that is, its particular tendency to choose one alternative over another (be it "disease" over "no disease," or vice versa). In so doing, the preferred measure is more valid and precise than the alternatives and can place all diagnostic systems on a common scale.

On the other hand, good test data can be very difficult to obtain. Thus, the "truth" against which diagnostic decisions are scored may be less than perfectly reliable, and the sample of test cases selected may not adequately represent the population to which the system is applied in practice. Such problems occur generally across diagnostic fields, but with more or less severity depending on the field. Hence our confidence in an assessment of accuracy can be higher in some fields than in others—higher, for instance, in weather forecasting than in polygraph lie detection.

Two classic papers on ROC

[Reprinted from RADIOLOGY, Vol. 143, No. 1, Pages 29–36, April, 1982.]

James A. Hanley, Ph.D.
Barbara J. McNeil, M.D., Ph.D.

The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve[1]