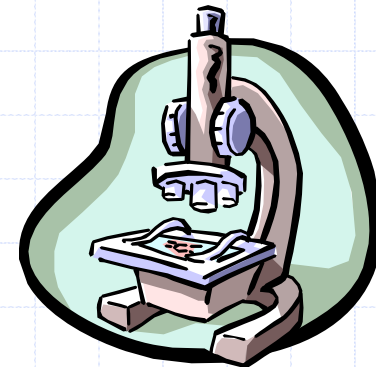
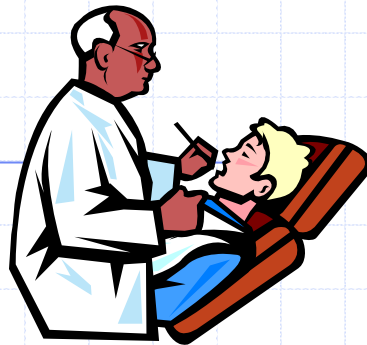
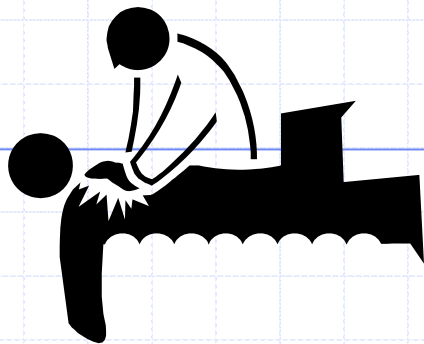


Diagnostic research: an introductory overview



Madhukar Pai, MD, PhD

Assistant Professor of Epidemiology, McGill University

Montreal, Canada

Professor Extraordinary, Stellenbosch University, S Africa

Email: madhukar.pai@mcgill.ca

Diagnosis: why does it matter?

- ◆ To effectively practice medicine and public health, we need evidence/knowledge on 3 fundamental types of professional knowing “gnosis”:

Dia-gnosis	Etio-gnosis	Pro-gnosis	For individual (Clinical Medicine)
Dia-gnosis	Etio-gnosis	Pro-gnosis	For community (Public and community health)

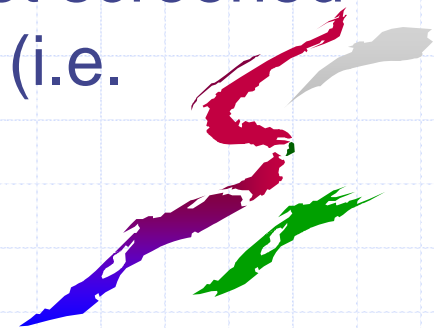
Dia-gnosis

The word diagnosis is derived through Latin from Greek:

- “dia” meaning *apart*, and “gnosis” meaning *to learn*.

Diagnosis Vs Screening

- A diagnostic test is done on sick people
 - patient presents with symptoms
 - pre-test probability of disease is high (i.e. disease prevalence is high)
- A screening test is usually done on asymptomatic, apparently healthy people
 - healthy people are encouraged to get screened
 - pre-test probability of disease is low (i.e. disease prevalence is low)



Approaches to Diagnosis

◆ Consider the following diagnostic situations:

- A 43-year-old woman presents with a painful cluster of vesicles grouped in the T3 dermatome of her left thorax.
- A 78-year-old man returns to the office for follow-up of hypertension. He has lost 10 kg since his last visit 4 months ago. He describes reduced appetite, but otherwise, there are no localizing symptoms. You recall that his wife died a year ago and consider depression as a likely explanation, yet his age and exposure history (ie, smoking) suggest other possibilities.

Approaches to Diagnosis

Pattern recognition

See it and recognize disorder



Compare posttest probability
with thresholds

(usually pattern recognition implies
probability near 100% and
so above threshold)

Probabilistic diagnostic reasoning

Clinical assessment generates pretest
probability



New information generates posttest
probability

(may be interactive)



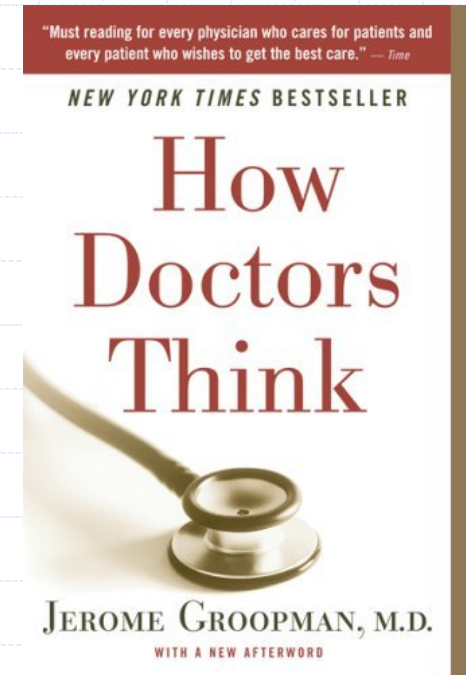
Compare posttest probability with thresholds

Source: Guyatt G, Rennie D, Meade MO, Cook DJ: *Users' Guides to the Medical Literature: A Manual for Evidence-Based Practice*, 2nd Edition: <http://www.jamaevidence.com>

Copyright © American Medical Association. All rights reserved.

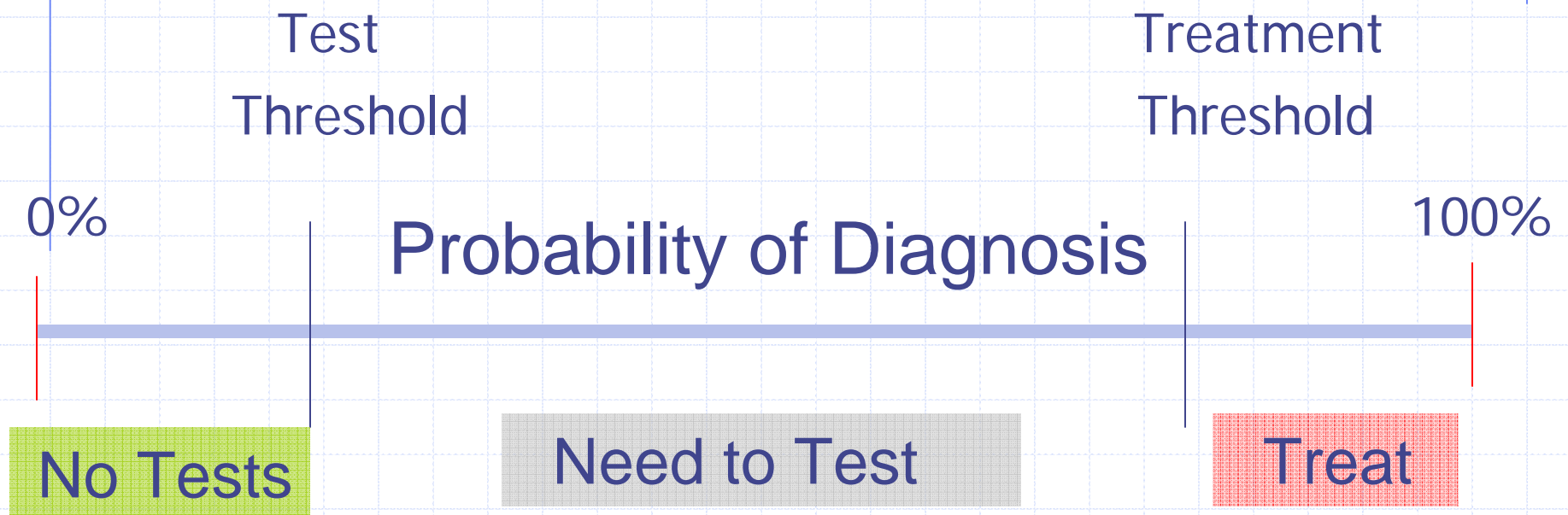
Misdiagnosis is common!

- ◆ Most misguided care results from thinking errors rather than technical mistakes.
- ◆ Major thinking traps: “three As”
 - **Anchoring**
 - ◆ Shortcut in thinking when a person doesn't consider multiple possibilities but quickly latches on to a single one.
 - **Availability**
 - ◆ Tendency to judge the likelihood of an event by the ease with which relevant examples come to mind.
 - **Attribution**
 - ◆ Based on stereotypes that are based on someone's appearance, emotional state or circumstances
- ◆ Key question to avoid these traps:
“What else can it be?”

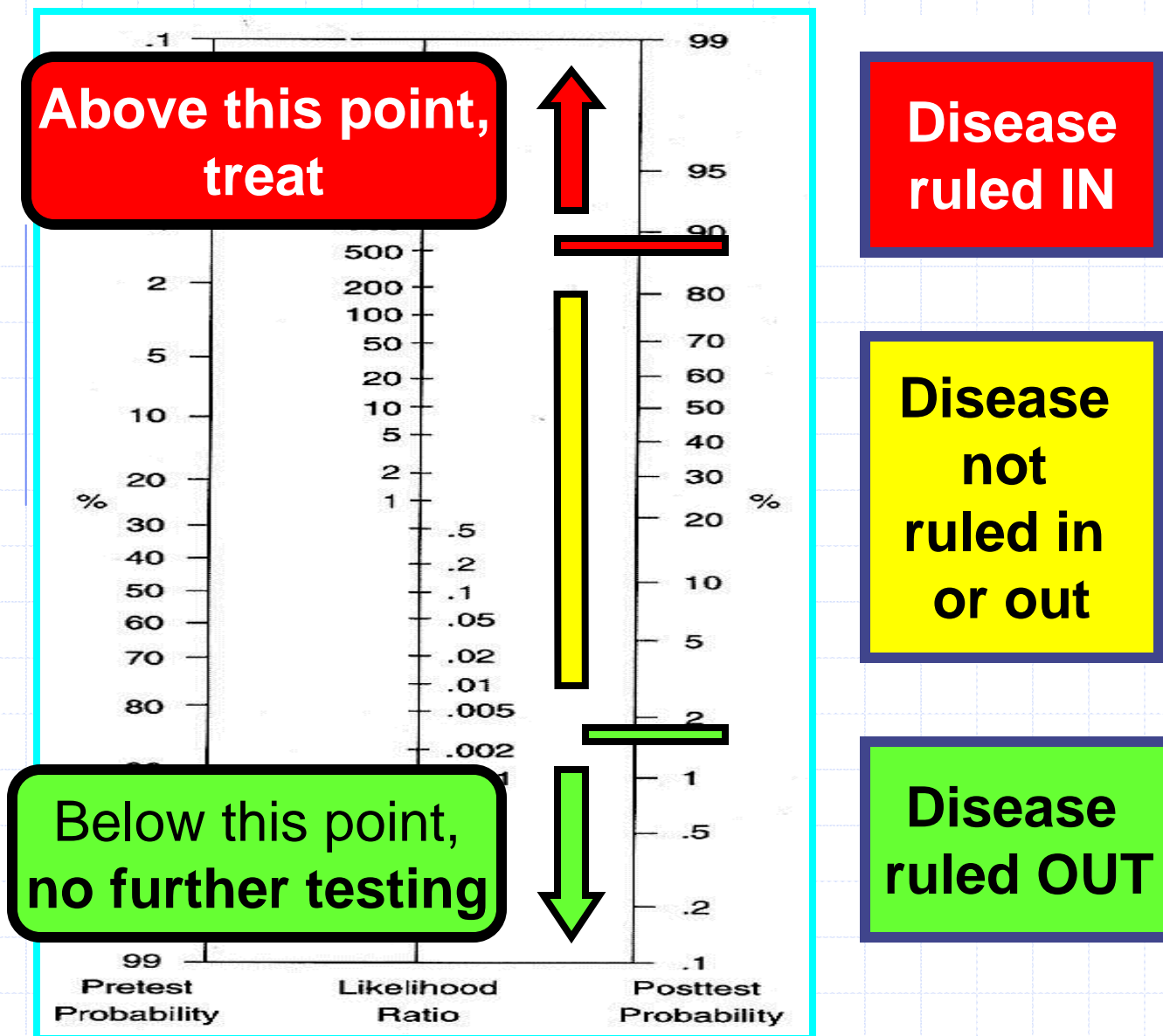


"Usually doctors are right, but conservatively about 15 percent of all people are misdiagnosed. Some experts think it's as high as 20 to 25 percent," - Groopman

Process of diagnosis: all about
probability and decision making
under uncertainty!



Thresholds for decision-making: when will you stop investigating?
when will you test further? when will you rule out disease?



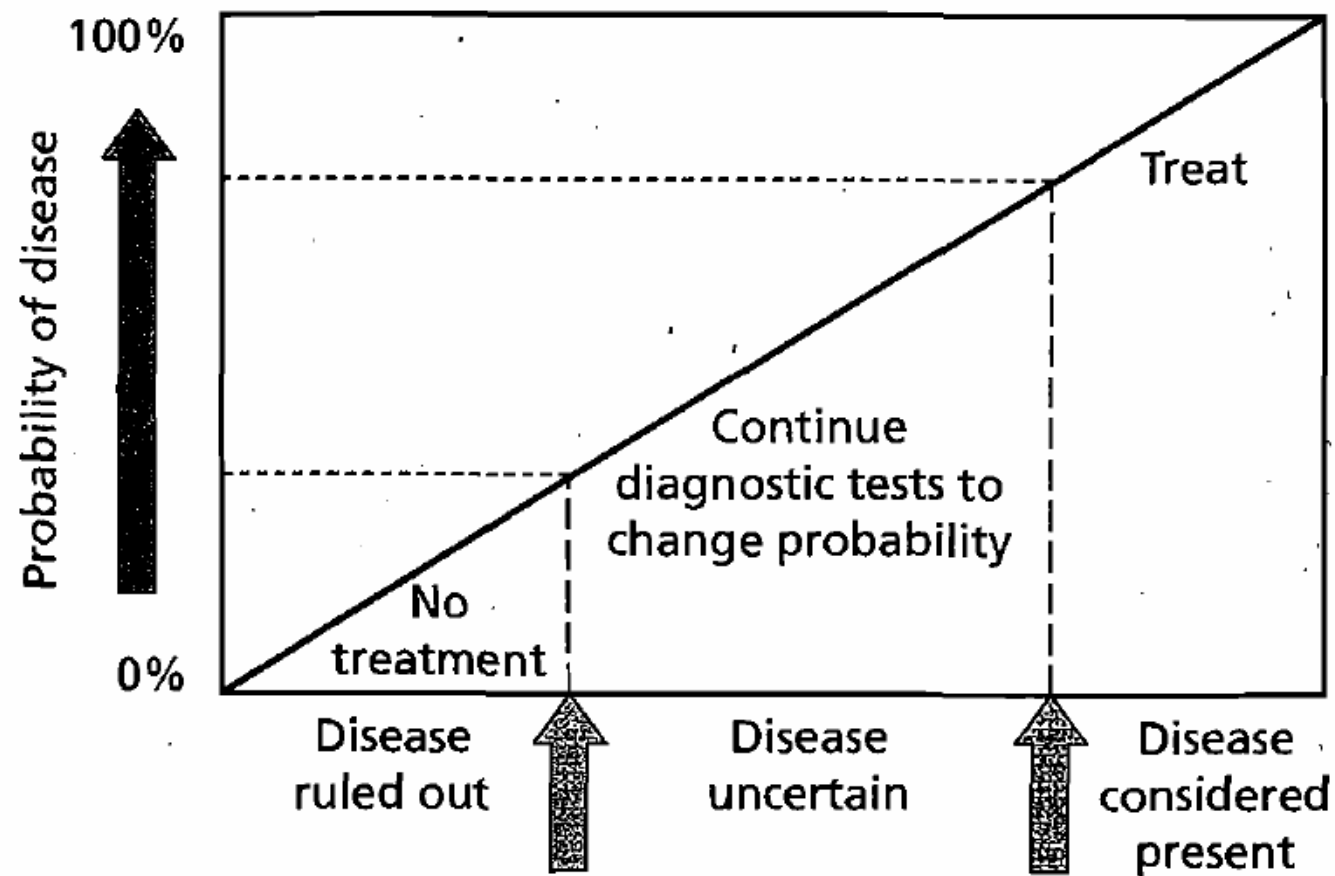
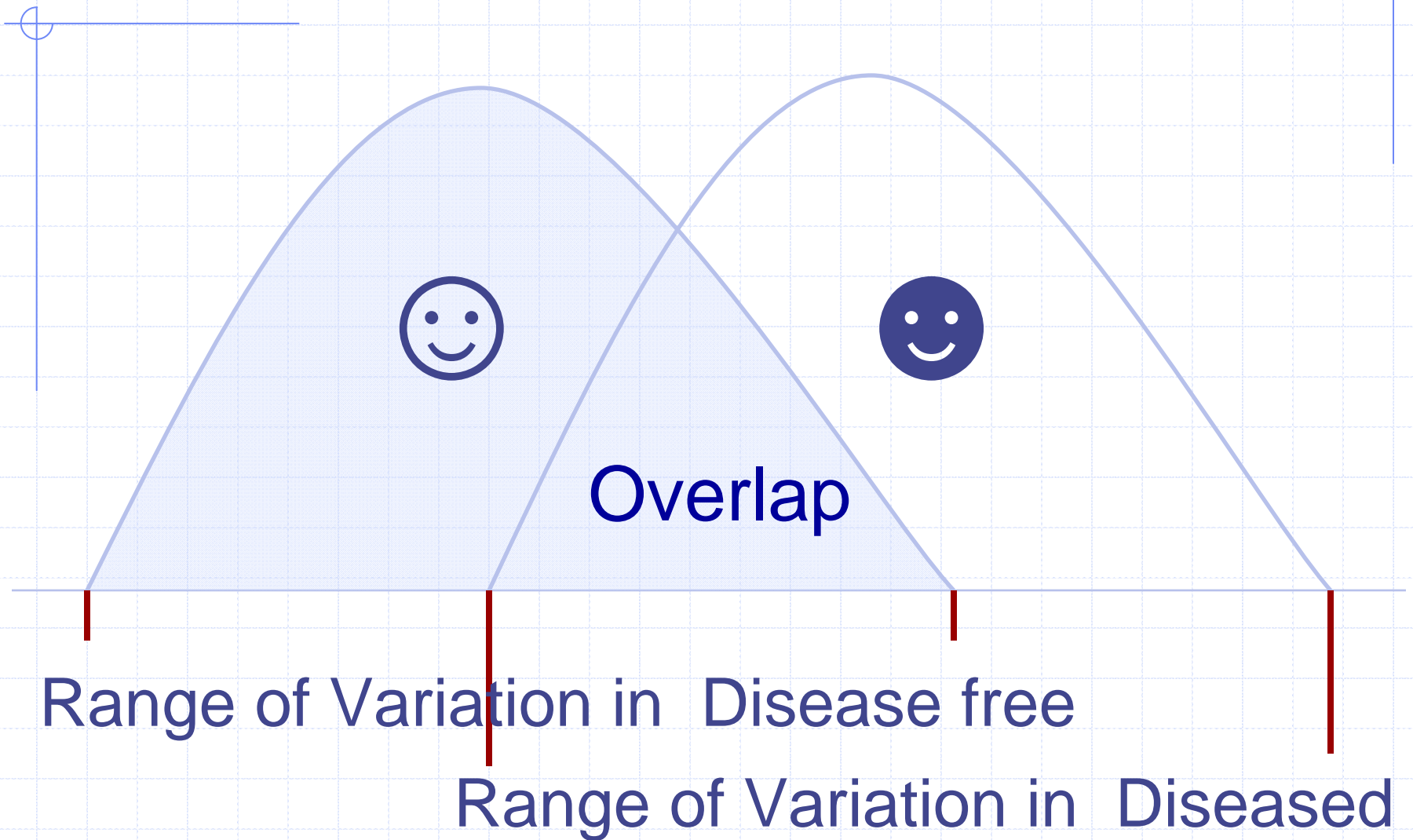


FIGURE 3.1 Diagnostic Testing.

The Perfect Diagnostic Test

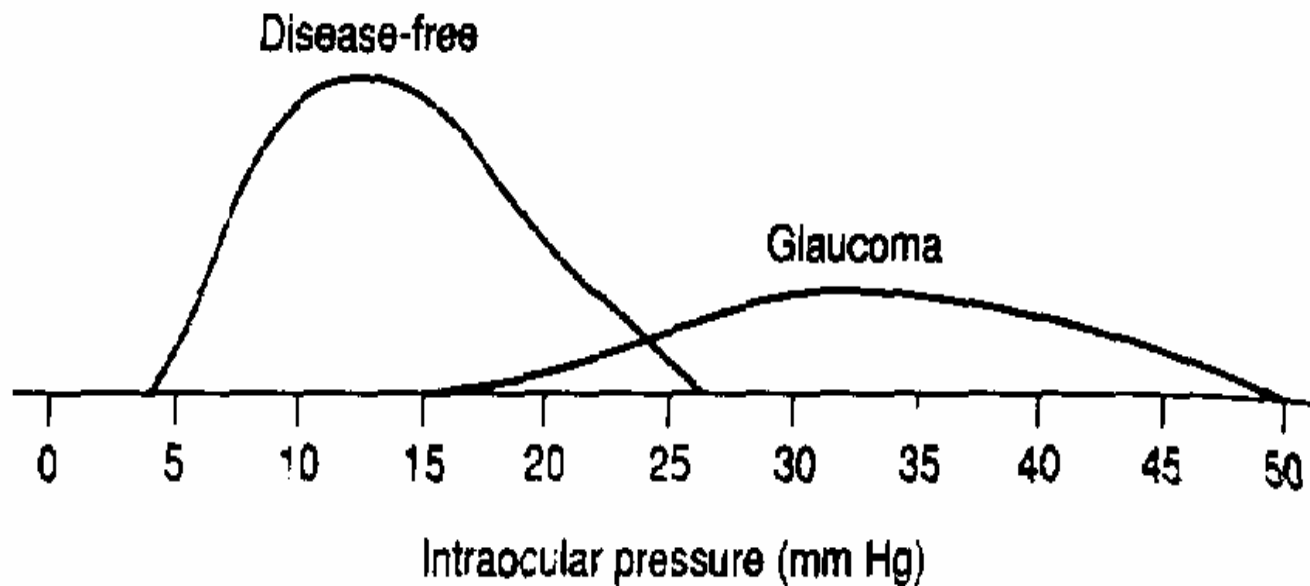


Variations In Diagnostic Tests



Example: intra-ocular pressure

Fig 2



Overlap of distributions of intraocular pressure among those with glaucoma and those without glaucoma

Riegelman & Hirsch 1996

Example: WBC count in bacteremia

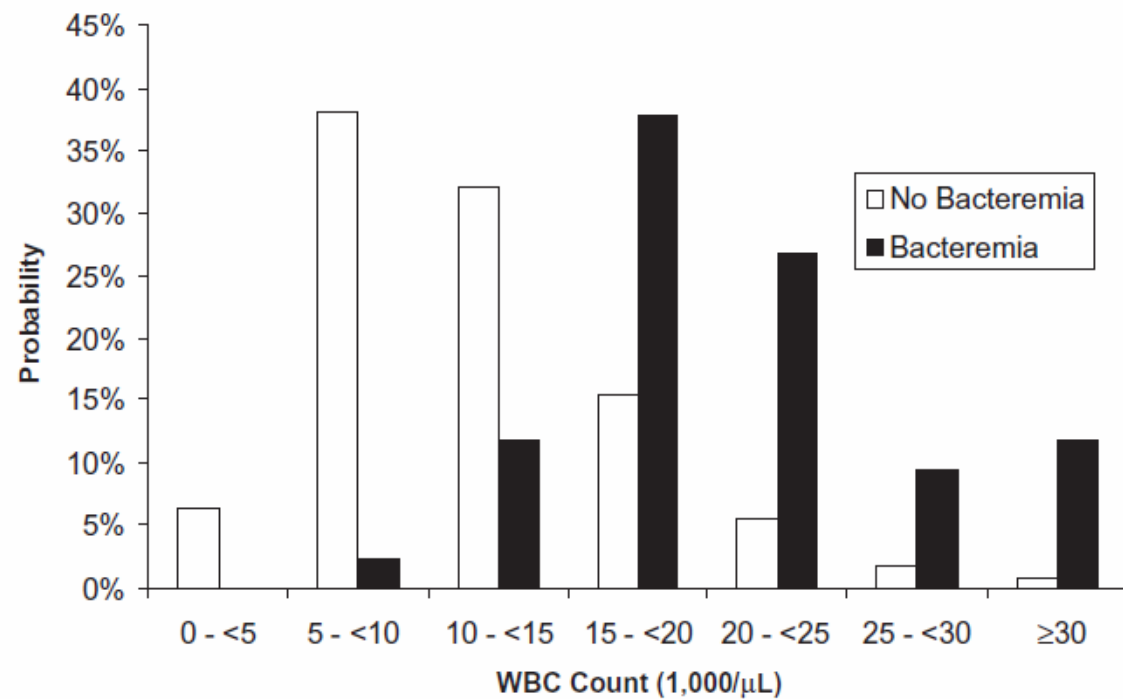


Figure 4.4 Histogram showing distributions of the nonbacteremic and bacteremic populations across the WBC count intervals.

There is no perfect test!

Thomas Bayes



Thomas Bayes (The correct identification of this portrait has been [1] questioned.)

Born	c. 1702 London
Died	April 17th 1761 Tunbridge Wells
Nationality	British

LII. *An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S.*

Dear Sir,

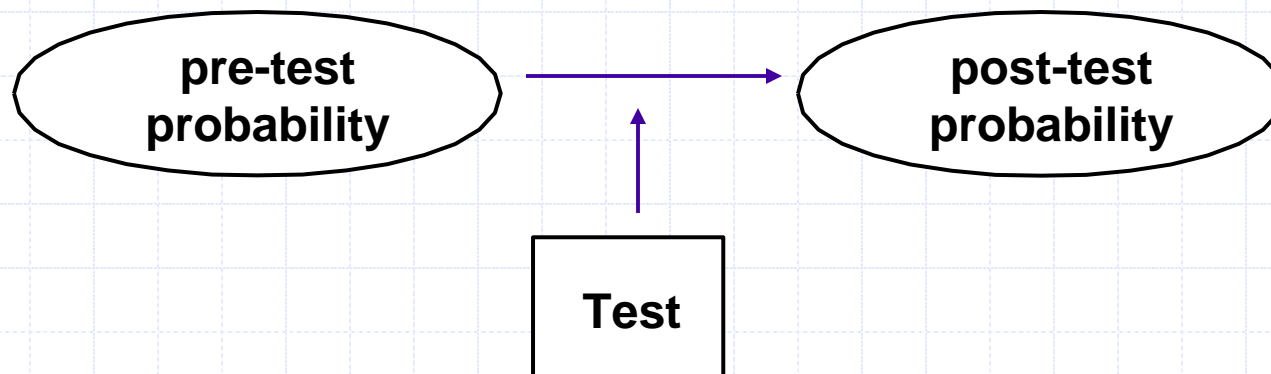
Read Dec. 23, 1763. I now send you an essay which I have found among the papers of our deceased friend Mr. Bayes, and which, in my opinion, has great merit, and well deserves to be preserved. Experimental philosophy, you will find, is nearly interested in the subject of it; and on this account there seems to be particular reason for thinking that a communication of it to the Royal Society cannot be improper.

He had, you know, the honour of being a member of that illustrious Society, and was much esteemed by many as a very able mathematician. In an introduction which he has writ to this Essay, he says, that his design at first in thinking on the subject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumstances, upon supposition that we know nothing concerning it but that, under the same circumstances, it has happened a certain number of times, and failed a certain other number of times. He adds, that he soon perceived that it would not be very difficult to do this, provided some rule could be found, according to which we ought to estimate the chance that the probability for the happening of an event perfectly unknown, should lie between any two named degrees of prob-

All we can hope to do is increase or decrease probabilities, and Bayes' theorem helps with this process

Bayes' theory

- Bayes' Theorem is a simple mathematical formula used for calculating conditional probabilities
- every test is done with a certain probability of disease - degree of suspicion [pre-test or prior probability]
- the probability of disease after the test result is the post-test or posterior probability



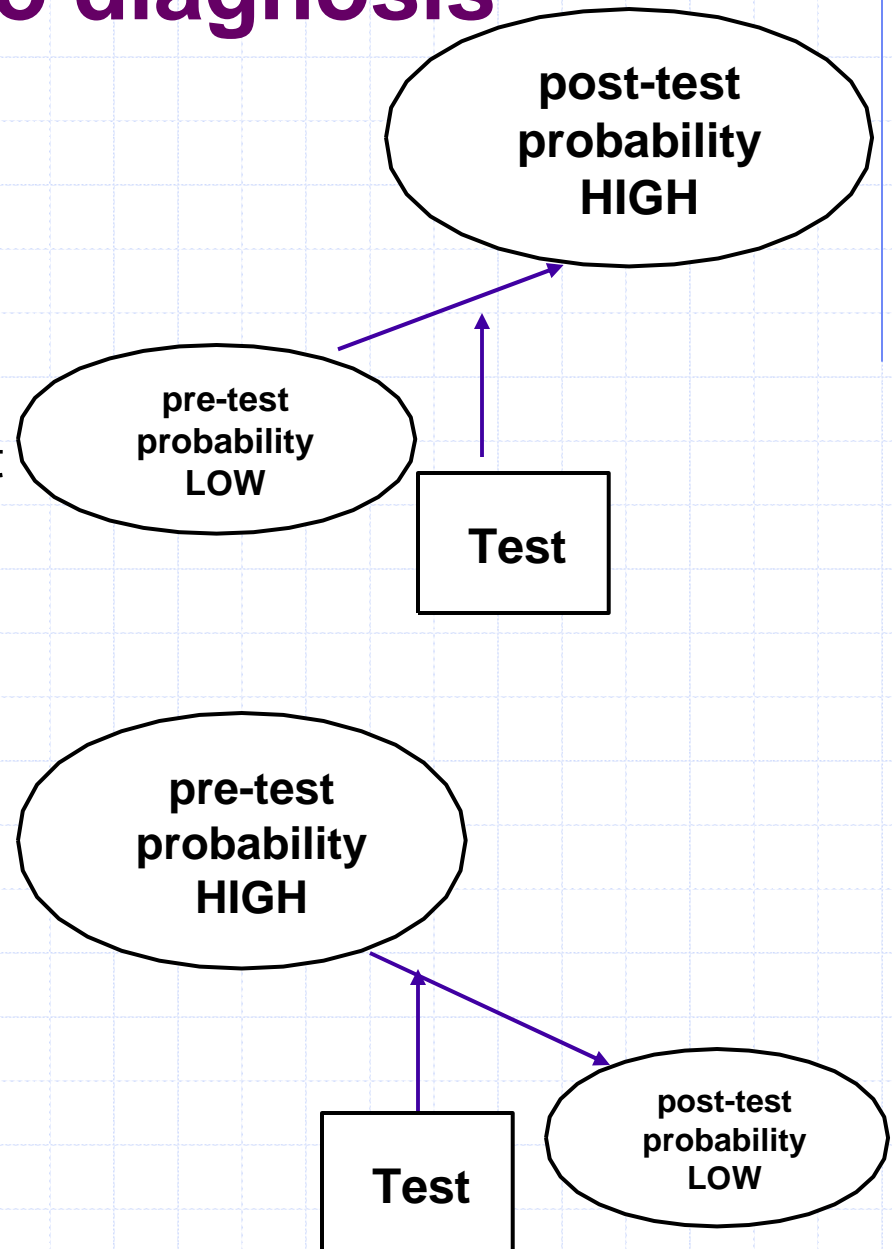
$$\text{Post-test odds} = \text{Pre-test odds} \times \text{Likelihood ratio}$$

The most simplistic way of explaining Bayes' theorem

What you thought before + New information = What you think now

Bayesian approach to diagnosis

- An accurate test will help reduce uncertainty
- The pre-test probability is revised using test result to get the post-test probability
- Tests that produce the biggest changes from pretest to post-test probabilities are most useful in clinical practice [very large or very small likelihood ratios]
- LR also called “Bayes Factor”



Why clinicians are natural bayesians

Christopher J Gill, Lora Sabin, Christopher H Schmid

Thought you didn't understand bayesian statistics? Read on and find out why doctors are expert in applying the theory, whether they realise it or not

Center for
International
Health and
Development,
Department of
International
Health, Boston
University School of
Public Health,
Boston, MA 02118,
USA

Christopher J Gill
assistant professor
Lora Sabin
assistant professor

Biostatistics
Research Center,
Division of Clinical
Care Research,
Department of
Medicine, Tufts
University—New
England Medical
Center, Boston,
MA 02111, USA
Christopher H
Schmid
associate professor

Correspondence to:
C J Gill
cgill@bu.edu

Two main approaches are used to draw statistical inferences: frequentist and bayesian. Both are valid, although they differ methodologically and perhaps philosophically. However, the frequentist approach dominates the medical literature and is increasingly applied in clinical settings. This is ironic given that clinicians apply bayesian reasoning in framing and revising differential diagnoses without necessarily undergoing, or requiring, any formal training in bayesian statistics. To justify this assertion, this article will explain how bayesian reasoning is a natural part of clinical decision making, particularly as it pertains to the clinical history and physical examination, and how bayesian approaches are a powerful and intuitive approach to the differential diagnosis.

A sick child in Ethiopia

On a recent trip to southern Ethiopia, my colleagues and I encountered a severely ill child at a rural health clinic. The child's palms, soles, tongue, and conjunctivae were all white from severe anaemia and his spleen was swollen and firm; he was breathing rapidly, had bilateral pulmonary rales, and was semiconscious. It

did clinical judgments prove superior to the algorithm, a diagnostic tool carefully developed over two decades of research? Was it just a lucky guess?

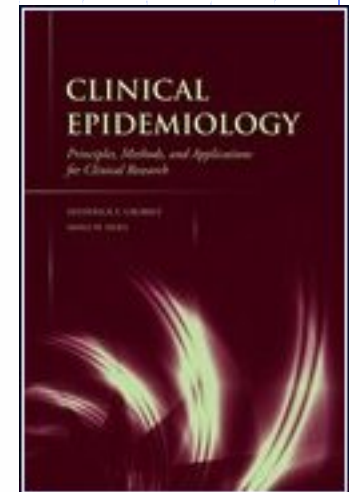
Interpreting diagnostic test results from the bayesian perspective

Clinical diagnosis ultimately rests on the ability to interpret diagnostic test results. But what is a diagnostic test? Clearly blood tests, radiography, biopsies, and other technology based evaluations qualify. However, this view is far too restrictive. In truth, any patient feature that varies in a given disease also qualifies. This definition would include each step in the clinical algorithm above, and, importantly, virtually all elements of the clinical history and physical examination.

Bayesians interpret the test result not as a categorical probability of a false positive but as the degree to which a positive or negative result adjusts the probability of a given disease. In this way, the test acts as an opinion modifier, updating a prior probability of disease to generate a posterior probability. In a sense, the bayesian approach asks, "What is the probability that this patient has the disease, given this test result?"

The diagnostic process is probabilistic, multivariable and sequential

1. A diagnosis starts with a patient presenting a complaint (symptom and/or sign) suggestive of a certain disease to be diagnosed.
2. The subsequent work-up is a multivariable process. It involves multiple diagnostic determinants (tests) that are applied in a logical order: from age, gender, medical history, and signs and symptoms, to more complicated, invasive, and costly tests.
3. Setting or ruling out a diagnosis is a probabilistic action in which the probability of the presence or absence of the disease is central. This probability is continuously updated based on subsequent diagnostic test results.
4. The true diagnostic value of a test is determined by the extent to which it provides diagnostic information beyond earlier tests, that is, materially changes the probability estimation of disease presence based on previous test results.
5. The goal of the diagnostic process is to eventually rule in or out the disease with enough confidence to take clinical decisions. This requires precise estimates of the probability of the presence of the target disease(s).



A diagnostic 'test' can be:

- ◆ A question (e.g. asking about a symptom)
- ◆ A simple physical sign
- ◆ A laboratory or imaging or other test
- ◆ A combination of many tests (e.g. a risk score or clinical prediction rule)
- ◆ An entire algorithm

Accuracy of perception and touch for detecting fever in adults: a hospital-based study from a rural, tertiary hospital in Central India

Manoj Singh¹, Madhukar Pai² and S. P. Kalantri¹

1 Department of Medicine, Mahatma Gandhi Institute of Medical Sciences, Sevagram, India

2 Division of Epidemiology, University of California at Berkeley, Berkeley, CA 94720, USA

Original Article

www.jpgmonline.com

Accuracy of physical examination in the diagnosis of hypothyroidism: A cross-sectional, double-blind study

Indra R, Patil SS, Joshi R, Pai M,* Kalantri SP

Simple clinical predictors of brain lesions in patients with impaired consciousness: a cross sectional study from a rural, tertiary hospital in central India[☆]

Y. Geetadevi^{a, 1}, Rajnish Joshi^{a, 1}, Madhukar Pai^{b, 2}, S.P. Kalantri^{a, *}

^a Department of Medicine, Mahatma Gandhi Institute of Medical Sciences, Sevagram, Wardha 442102, India

8

THE NATIONAL MEDICAL JOURNAL OF INDIA

VOL. 16, NO. 1, 2003

Original Articles

Poor accuracy of the Siriraj and Guy's hospital stroke scores in distinguishing haemorrhagic from ischaemic stroke in a rural, tertiary care hospital

PRIYA BADAM, VAISHALI SOLAO, MADHUKAR PAI, S. P. KALANTRI



ELSEVIER

respiratoryMEDICINE

Accuracy and reliability of physical signs in the diagnosis of pleural effusion

Shriprakash Kalantri^a, Rajnish Joshi^{a,c}, Trunal Lokhande^a, Amandeep Singh^a, Maureen Morgan^b, John M. Colford Jr^c, Madhukar Pai^{d,*}

OPEN ACCESS Freely available online



Evaluation of Diagnostic Accuracy, Feasibility and Client Preference for Rapid Oral Fluid-Based Diagnosis of HIV Infection in Rural India

Nitika Pant Pai^{1*}, Rajnish Joshi², Sandeep Dogra³, Bharati Taksande², S. P. Kalantri², Madhukar Pai⁴, Pratibha Narang², Jacqueline P. Tulskey⁵, Arthur L. Reingold⁶

1 Immunodeficiency Service, Montreal Chest Institute, McGill University Health Center, Montreal, Canada, 2 Mahatma Gandhi Institute of Medical Sciences, Sevagram, Maharashtra, India, 3 Acharya Shri Chander College of Medical Sciences, Jammu, India, 4 Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada, 5 Department of Internal Medicine, University of California at San Francisco, San Francisco, California, United States of America, 6 Division of Epidemiology, University of California at Berkeley, Berkeley, California, United States of America

Sensitivity of a Whole-Blood Interferon-Gamma Assay Among Patients with Pulmonary Tuberculosis and Variations in T-Cell Responses During Anti-Tuberculosis Treatment

M. Pai, R. Joshi, M. Bandyopadhyay, P. Narang, S. Dogra,
B. Taksande, S. Kalantri

OPEN ACCESS Freely available online

PLOS MEDICINE

Impact of Round-the-Clock, Rapid Oral Fluid HIV Testing of Women in Labor in Rural India

Nitika Pant Pai^{1*}, Ritu Barick², Jacqueline P. Tulsky³, Poonam V. Shivkumar², Deborah Cohan³, Shriprakash Kalantri²,
Madhukar Pai⁴, Marina B. Klein¹, Shakuntala Chhabra²

1 Division of Infectious Diseases and Immunodeficiency Service, Montreal Chest Institute, McGill University Health Center, Montreal, Canada, **2** Mahatma Gandhi Institute of Medical Sciences, Sevagram, Wardha, Maharashtra, India, **3** Positive Health Program, Division of Internal Medicine, University of California San Francisco, San Francisco, California, United States of America, **4** Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada

Diagnosis vs. prediction

◆ Diagnosis:

- Disease has already occurred and we are trying to detect its presence

◆ Prognosis:

- Disease has not occurred and we want to know who is most likely to develop the disease

◆ Both are amenable to multivariable approaches and prediction models

◆ They are often mixed up

- Sometimes a diagnostic test itself can be used to predict future outcomes (e.g. PSA)

Against Diagnosis

Andrew J. Vickers, PhD; Ethan Basch, MD; and Michael W. Kattan, PhD

The act of diagnosis requires that patients be placed in a binary category of either having or not having a certain disease. Accordingly, the diseases of particular concern for industrialized countries—such as type 2 diabetes, obesity, or depression—require that a somewhat arbitrary cut-point be chosen on a continuous scale of measurement (for example, a fasting glucose level >6.9 mmol/L [>125 mg/dL] for type 2 diabetes). These cut-points do not adequately reflect disease biology, may inappropriately treat patients on either side of the cut-point as 2 homogenous risk groups, fail to incorporate other risk factors, and are invariable to patient prefer-

ence. This article discusses risk prediction as an alternative to diagnosis: Patient risk factors (blood pressure, age) are combined into a single statistical model (risk for a cardiovascular event within 10 years) and the results are used in shared decision making about possible treatments. The authors compare and contrast the diagnostic and risk prediction approaches and attempt to identify the types of medical problem to which each is best suited.

Ann Intern Med. 2008;149:200-203.

For author affiliations, see end of text.

www.annals.org

Table. Comparison of Typical Features of Diagnostic and Risk Prediction Approaches

Variable	Diagnosis	Risk Prediction
Approach	Patients are given a diagnosis: Either they have the disease or they do not	Patients are given a probability of a future event
Example	Syphilitic hepatitis	Cardiovascular event within 10 years
Lesion	Unambiguous	Nonexistent or equivocal
Example	Torn aorta	Depression
Treatment effectiveness	Often highly effective	Helpful, but patients may have event with treatment or avoid the event even if untreated
Example	Antibiotics for syphilis	Statins for high cholesterol level
Course of treatment	Dictated by diagnosis	Open to discussion
Example	Surgical treatment of a torn aorta	Treatment of early-stage prostate cancer
Patient preference	Generally of minor importance	Often of major importance
Example	Antibiotics for syphilis	Treatment of early-stage prostate cancer
Symptoms	Patient has distressing symptoms	Patient is often asymptomatic: Disorder is a risk factor for a future event
Example	Syphilitic hepatitis	Hyperlipidemia



Types of diagnostic study designs

Evidence base of clinical diagnosis

The architecture of diagnostic research

D L Sackett, R B Haynes

Considerable effort has been expended at the interface between clinical medicine and scientific methods to achieve the maximum validity and usefulness of diagnostic tests. This article focuses on the specific kinds of questions that arise in diagnostic research and the study architectures (the conversions of these clinical questions into appropriate research designs) used to answer them. As an example we shall take assessment of the value of the plasma concentration of B-type natriuretic peptide (BNP) in the diagnosis of left ventricular dysfunction.¹ Randomised controlled trials are dealt with elsewhere.

As in other forms of clinical research, there are several different ways studying the potential or real diagnostic value of a physical sign or laboratory test, and each is appropriate to one kind of question and inappropriate for others. Among the possible questions about the relation between a putative diagnostic test and a target disorder (for example, the concentration of BNP and left ventricular dysfunction), four are most relevant.

Types of question

Phase I questions

Do test results in patients with the target disorder differ from those in normal people? Table 1 shows the architecture of this question.

For example, investigators at a British university hospital measured concentrations of BNP precursor in non-systematic ("convenience") samples from normal controls and from patients who had various combina-

Summary points

Diagnostic studies should match methods to diagnostic questions

- Do test results in affected patients differ from those in normal individuals?
- Are patients with certain test results more likely to have the target disorder?
- Do test results distinguish patients with and without the target disorder among those in whom it is clinically sensible to suspect the disorder?
- Do patients undergoing the diagnostic test fare better than similar untested patients?

The keys to validity in diagnostic test studies are

- independent, blind comparison of test results with a reference standard among a consecutive series of patients suspected (but not known) to have the target disorder
- inclusion of missing and indeterminate results
- replication of studies in other settings

Both specificity and sensitivity may change as the same diagnostic test is applied in primary, secondary, and tertiary care

This is the second in a series of five articles

Trout Research and Education Centre at Irish Lake, RR1, Markdale, ON, Canada N0C 1H0

D L Sackett
professor

Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada L8N 3Z5

R B Haynes
director

Correspondence to: D L Sackett
sackett@bmts.com

BMJ 2002;324:539-41

BMJ 2002;324:539-41

Phase I to IV diagnostic studies

◆ Phase I questions

- Do test results in patients with the target disorder differ from those in normal people?

Table 1 Answering a phase I question: do patients with left ventricular dysfunction have higher concentrations of B-type natriuretic peptide (BNP) precursor than normal individuals?

	Patients known to have disorder	Normal controls
Median (range) concentration of BNP precursor (pg/ml)	493.5 (248.9-909.0)	129.4 (53.6-159.7)

Phase I to IV diagnostic studies

◆ Phase II questions

- Are patients with certain test results more likely to have the target disorder than patients with other test results?

Table 2 Answering a phase II question: are patients with higher concentrations of B-type natriuretic peptide (BNP) more likely to have left ventricular dysfunction than patients with lower concentrations?

	Patients known to have target disorder	Normal controls
High BNP concentration	39	2
Normal BNP concentration	1	25

Test characteristics (95% CI):

Sensitivity=98% (87% to 100%)

Specificity=92% (77% to 98%)

Positive predictive value=95% (84% to 99%)

Negative predictive value=96% (81% to 100%)

Likelihood ratio for an abnormal test result=13 (3.5 to 50.0)

Likelihood ratio for a normal test result=0.03 (0.0003 to 0.19)

Phase I to IV diagnostic studies

◆ Phase III questions

- Does the test result distinguish patients with and without the target disorder among patients in whom it is clinically reasonable to suspect that the disease is present?

Table 3 Answering a phase III question: among patients in whom it is clinically sensible to suspect left ventricular dysfunction (LVD), does the concentration of B-type natriuretic peptide (BNP) distinguish patients with and without left ventricular dysfunction?

	Patients with LVD on echocardiography	Patients with normal results on echocardiography
Concentration of BNP:		
High (>17.9 pg/ml)	35	57
Normal (<18 pg/ml)	5	29
Prevalence (pretest probability) of LVD	40/126=32%	

Test characteristics (95% CI):

Sensitivity=88% (74% to 94%)

Specificity=34% (25% to 44%)

Positive predictive value=38% (29% to 48%)

Negative predictive value=85% (70% to 94%)

Likelihood ratio for an abnormal test result=1.3 (1.1 to 1.6)

Likelihood ratio for a normal test result=0.4 (0.2 to 0.9)

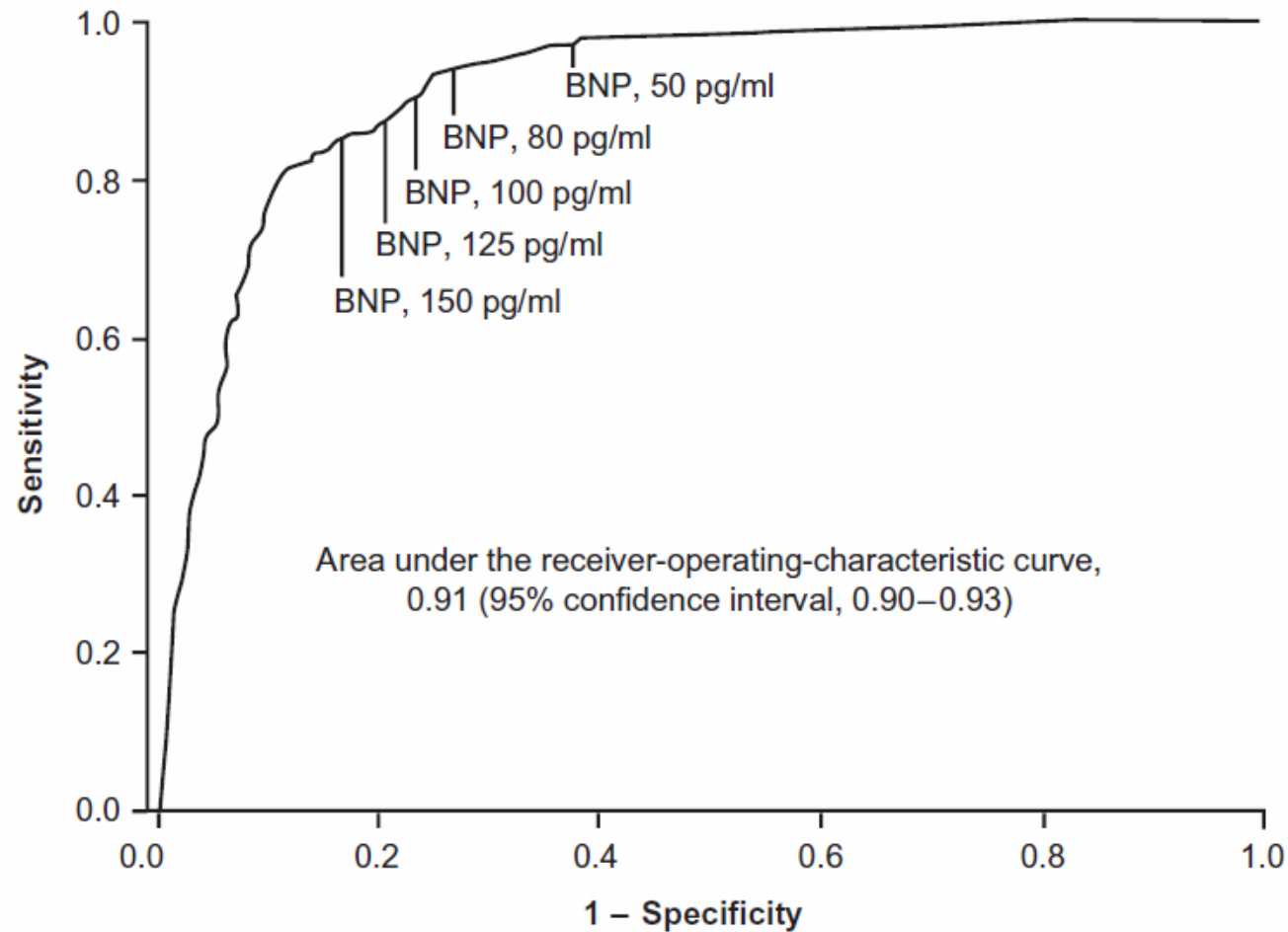
BMJ 2002;324:539–41

Phase I to IV diagnostic studies

◆ Phase IV questions

- Do patients who undergo this diagnostic test fare better (in their ultimate health outcomes) than similar patients who are not tested?

Rapid measurement of B-type natriuretic peptide in the emergency diagnosis of heart failure



Evaluation of Diagnostic Accuracy, Feasibility and Client Preference for Rapid Oral Fluid-Based Diagnosis of HIV Infection in Rural India

Nitika Pant Pai^{1*}, Rajnish Joshi², Sandeep Dogra³, Bharati Taksande², S. P. Kalantri², Madhukar Pai⁴, Pratibha Narang², Jacqueline P. Tulsy⁵, Arthur L. Reingold⁶

1 Immunodeficiency Service, Montreal Chest Institute, McGill University Health Center, Montreal, Canada, **2** Mahatma Gandhi Institute of Medical Sciences, Sevagram, Maharashtra, India, **3** Acharya Shri Chander College of Medical Sciences, Jammu, India, **4** Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada, **5** Department of Internal Medicine, University of California at San Francisco, San Francisco, California, United States of America, **6** Division of Epidemiology, University of California at Berkeley, Berkeley, California, United States of America

Background. Oral fluid-based rapid tests are promising for improving HIV diagnosis and screening. However, recent reports from the United States of false-positive results with the oral OraQuick[®] ADVANCE HIV1/2 test have raised concerns about their performance in routine practice. We report a field evaluation of the diagnostic accuracy, client preference, and feasibility for the oral fluid-based OraQuick[®] Rapid HIV1/2 test in a rural hospital in India. **Methodology/Principal Findings.** A cross-sectional, hospital-based study was conducted in 450 consenting participants with suspected HIV infection in rural India. The objectives were to evaluate performance, client preference and feasibility of the OraQuick[®] Rapid HIV-1/2 tests. Two Oraquick[®] Rapid HIV1/2 tests (oral fluid and finger stick) were administered in parallel with confirmatory ELISA/Western Blot (reference standard). Pre- and post-test counseling and face to face interviews were conducted to determine client preference. Of the 450 participants, 146 were deemed to be HIV sero-positive using the reference standard (seropositivity rate of 32% (95% confidence interval [CI] 28%, 37%)). The OraQuick test on oral fluid specimens had better performance with a sensitivity of 100% (95% CI 98, 100) and a specificity of 100% (95% CI 99, 100), as compared to the OraQuick test on finger stick specimens with a sensitivity of 100% (95% CI 98, 100), and a specificity of 99.7% (95% CI 98.4, 99.9). The OraQuick oral fluid-based test was preferred by 87% of the participants for first time testing and 60% of the participants for repeat testing. **Conclusion/Significance.** In a rural Indian hospital setting, the OraQuick[®] Rapid- HIV1/2 test was found to be highly accurate. The oral fluid-based test performed marginally better than the finger stick test. The oral OraQuick test was highly preferred by participants. In the context of global efforts to scale-up HIV testing, our data suggest that oral fluid-based rapid HIV testing may work well in rural, resource-limited settings.

Citation: Pant Pai N, Joshi R, Dogra S, Taksande B, Kalantri SP, et al (2007) Evaluation of Diagnostic Accuracy, Feasibility and Client Preference for Rapid Oral Fluid-Based Diagnosis of HIV Infection in Rural India. PLoS ONE 2(4): e367. doi:10.1371/journal.pone.0000367

Impact of Round-the-Clock, Rapid Oral Fluid HIV Testing of Women in Labor in Rural India

Nitika Pant Pai^{1*}, Ritu Barick², Jacqueline P. Tulsy³, Poonam V. Shivkumar², Deborah Cohan³, Shriprakash Kalantri², Madhukar Pai⁴, Marina B. Klein¹, Shakuntala Chhabra²

1 Division of Infectious Diseases and Immunodeficiency Service, Montreal Chest Institute, McGill University Health Center, Montreal, Canada, **2** Mahatma Gandhi Institute of Medical Sciences, Sevagram, Wardha, Maharashtra, India, **3** Positive Health Program, Division of Internal Medicine, University of California San Francisco, San Francisco, California, United States of America, **4** Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada

Methods and Findings

After they provided written informed consent, women admitted to the labor ward of a rural teaching hospital in India were offered two rapid tests on oral fluid and finger-stick specimens (OraQuick Rapid HIV-1/HIV-2 tests, OraSure Technologies). Simultaneously, venous blood was drawn for conventional HIV ELISA testing. Western blot tests were performed for confirmatory testing if women were positive by both rapid tests and dual ELISA, or where test results were discordant. Round-the-clock (24 h, 7 d/wk) abbreviated prepartum and extended postpartum counseling sessions were offered as part of the testing strategy. HIV-positive women were administered PMTCT interventions. Of 1,252 eligible women (age range 18 y to 38 y) approached for consent over a 9 mo period in 2006, 1,222 (98%) accepted HIV testing in the labor ward. Of these, 1,003 (82%) women presented with either no reports or incomplete reports of prior HIV testing results at the time of admission to the labor ward. Of 1,222 women, 15 were diagnosed as HIV-positive (on the basis of two rapid tests, dual ELISA and Western blot), yielding a seroprevalence of 1.23% (95% confidence interval [CI] 0.61%–1.8%). Of the 15 HIV test-positive women, four (27%) had presented with reported HIV status, and 11 (73%) new cases of HIV infection were detected due to rapid testing in the labor room. Thus, 11 HIV-positive women received PMTCT interventions on account of round-the-clock rapid HIV testing and counseling in the labor room. While both OraQuick tests (oral and finger-stick) were 100% specific, one false-negative result was documented (with both oral fluid and finger-stick specimens). Of the 15 HIV-infected women who delivered, 13 infants were HIV seronegative at birth and at 1 and 4 mo after delivery; two HIV-positive infants died within a month of delivery.

Conclusions

In a busy rural labor ward setting in India, we demonstrated that it is feasible to introduce a program of round-the-clock rapid HIV testing, including prepartum and extended postpartum counseling sessions. Our data suggest that the availability of round-the-clock rapid HIV testing resulted in successful documentation of HIV serostatus in a large proportion (82%) of rural women who were unaware of their HIV status when admitted to the labor room. In addition, 11 (73%) of a total of 15 HIV-positive women received PMTCT interventions because of round-the-clock rapid testing in the labor ward. These findings are relevant for PMTCT programs in developing countries.

A slightly different classification

Evidence based diagnostics

Christian Gluud, Lise Lotte Gluud

Diagnostic tests are often much less rigorously evaluated than new drugs. It is time to ensure that the harms and benefits of new tests are fully understood

Cochrane
Hepato-Biliary
Group,
Copenhagen Trial
Unit, Centre for
Clinical
Intervention
Research, H.S.
Rigshospitalet,
Copenhagen
University Hospital,
DK-2100
Copenhagen,
Denmark
Christian Gluud
head of department
Lise Lotte Gluud
specialist registrar

Correspondence to:
C Gluud
cgluud@ctu.rh.dk

BMJ 2005;330:724-6

No international consensus exists on the methods for assessing diagnostic tests. Previous recommendations stress that studies of diagnostic tests should match the type of diagnostic question.^{1 2} Once the specificity and sensitivity of a test have been established, the final question is whether tested patients fare better than similar untested patients. This usually requires a randomised trial. Few tests are currently evaluated in this way. In this paper, we propose an architecture for research into diagnostic tests that parallels the established phases in drug research.

Stages of research

We have divided studies of diagnostic tests into four phases (box). We use research on brain natriuretic peptide for diagnosing heart failure as an illustrative example.³ However, the architecture is applicable to a

measure brain natriuretic peptide in human plasma, phase I studies were done to establish the normal range of values in healthy participants.^{4 5}

Diagnostic phase I studies must be large enough to examine the potential influence of characteristics such as sex, age, time of day, physical activity, and exposure to drugs. The studies are relatively quick, cheap, and easy to conduct, but they may occasionally raise ethical problems—for example, finding abnormal results in an apparently healthy person.⁶

Diagnostic accuracy

In phase II, studies explore the diagnostic accuracy of a test in participants with both known and suspected relevant disease. Phase IIa studies compare test results in participants with disease diagnosed by a standard method with those in healthy participants (from

Four phases in architecture of diagnostic research

Phase I—Determining the normal range of values for a diagnostic test through observational studies in healthy people

Phase II—Determining the diagnostic accuracy through case-control studies, including healthy people and (a) people with known disease assessed by diagnostic standard and (b) people with suspected disease

Phase III—Determining the clinical consequences of introducing a diagnostic test through randomised trials

Phase IV—Determining the effects of introducing a new diagnostic test into clinical practice by surveillance in large cohort studies

Diagnostic RCTs

VIEWPOINT

Viewpoint

**Randomised comparisons of medical tests: sometimes invalid,
not always efficient**

Patrick M M Bossuyt, Jeroen G Lijmer, Ben W J Mol

Lancet 2000; 356: 1844–47

Diagnostic RCTs

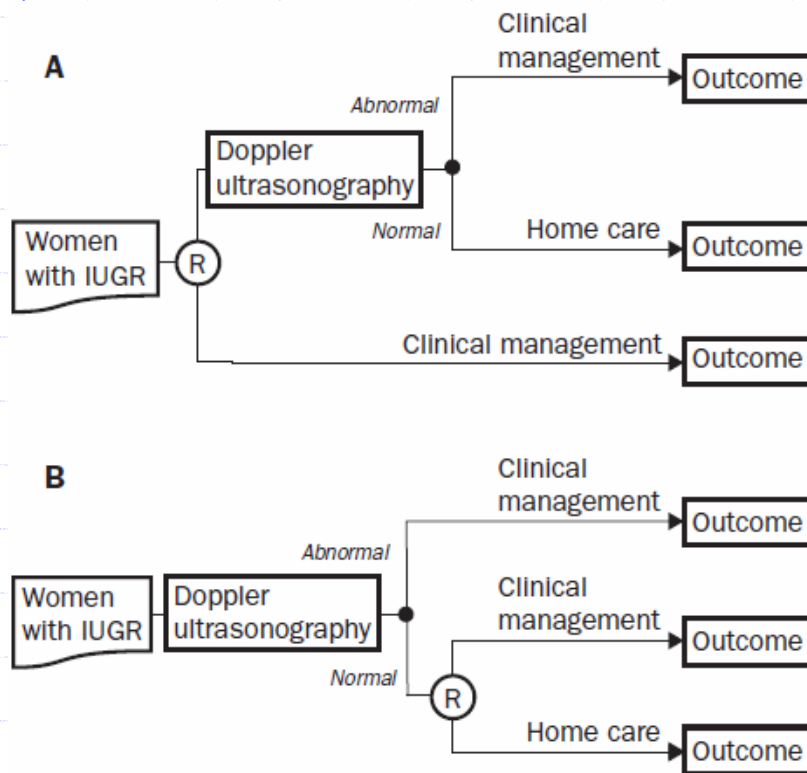


Figure 1: Trial designs of a single test
IUGR=intrauterine growth retardation; R=randomisation process.

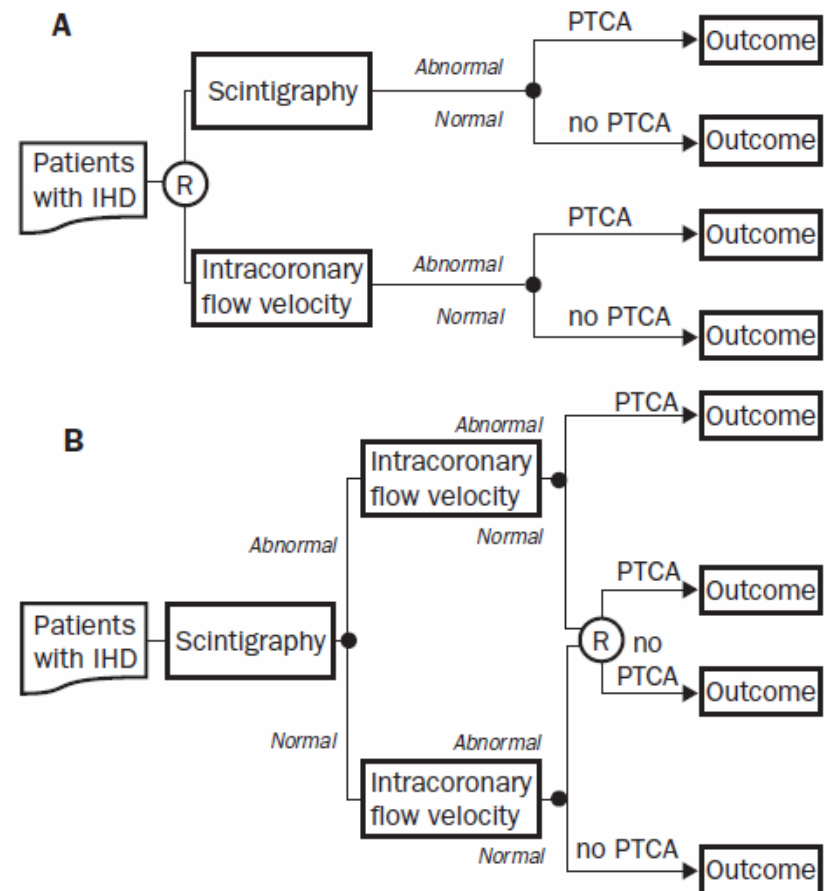
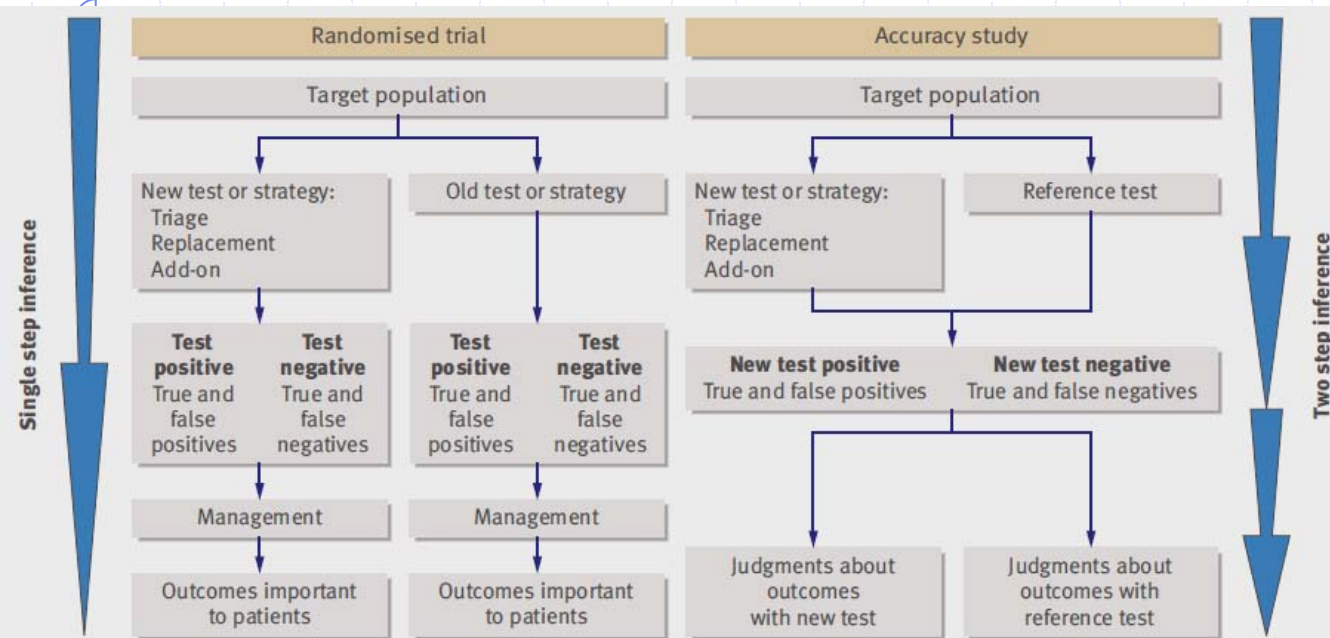


Figure 2: Trial designs to compare two tests
IHD=ischaemic heart disease; PTCA=percutaneous transluminal coronary angioplasty; R=randomisation process. Abnormal scintigraphy=reversible perfusion defect; abnormal intracoronary flow velocity=insufficient reserve.

Diagnostic RCT: is it really diagnostic?

When performing a randomized trial to determine the impact of a diagnostic test or strategy on patient outcome, an initially *diagnostic* research question is transformed into *therapeutic* research question (with the goal of establishing causality) with corresponding consequences for the design of the study. A disadvantage of a randomized approach to directly quantify the contribution of a diagnostic test and treatment on patient outcome is that it often addresses diagnosis and treatment as a single combined strategy, a “package deal.” This makes it impossible to determine afterwards whether a positive effect on patient outcome was attributed solely to the improved diagnosis by using the test under study or to the chosen (new) treatment strategies.

Diagnostic study design



Example

Randomised control trials (RCTs) explored a diagnostic strategy guided by the use of B type natriuretic peptide (BNP)—designed to aid diagnosis of heart failure—compared with no use of BNP in patients presenting to the emergency department with acute dyspnoea.^{8,9} As it turned out, the group randomised to receive BNP spent a shorter time in the hospital at lower cost, with no increased mortality or morbidity

Example

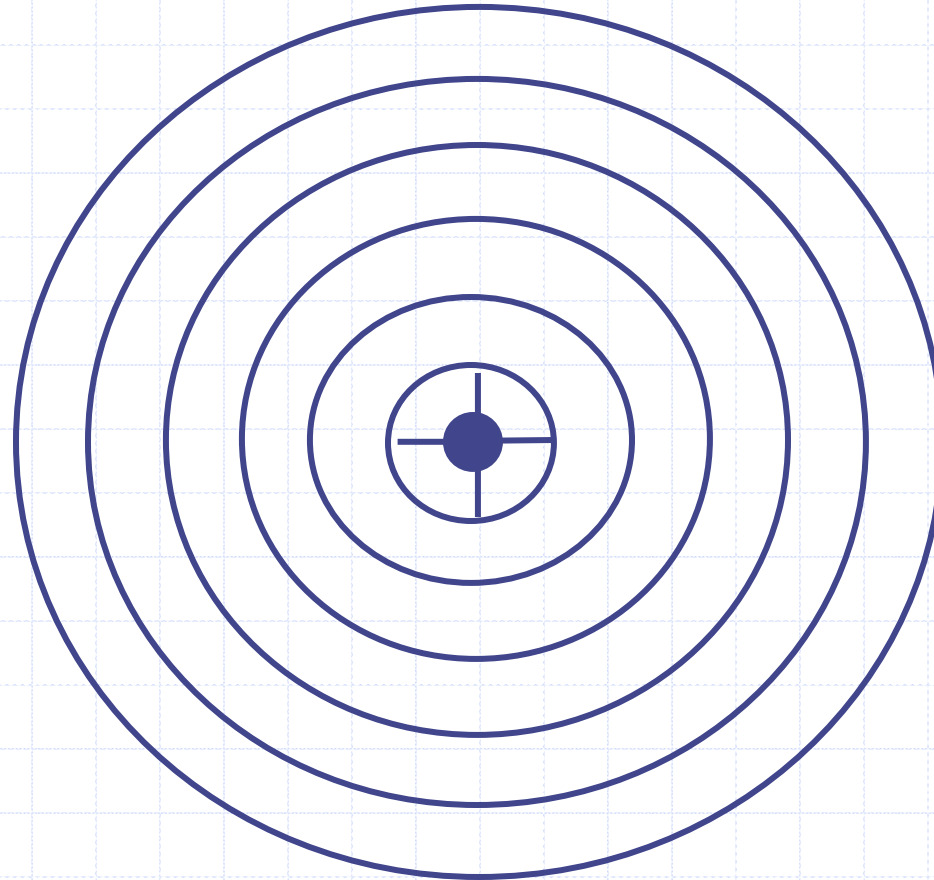
Consistent evidence from well designed studies shows fewer false negative results with non-contrast helical computed tomography (CT) than with intravenous pyelography (IVP) in the diagnosis of suspected acute urolithiasis.¹⁰ However, the stones in the ureter that CT detects but IVP “misses” are smaller, and hence are likely to pass more easily. As RCTs evaluating the outcomes in patients treated for smaller stones are not available, the extent to which reduction in cases that are missed (false negatives) and follow-up of incidental findings unrelated to renal calculi with CT have important health benefits remains uncertain¹¹

Two generic ways in which a test or diagnostic strategy can be evaluated. On the left, patients are randomised to a new test or strategy or to an old test or strategy. Those with a positive test result (cases detected) are randomised (or were previously randomised) to receive the best available management (second step of randomisation for management not shown). Investigators evaluate and compare patient-important outcomes in all patients in both groups.⁶ On the right, patients receive both a new test and a reference test (old or comparator test or strategy). Investigators can then calculate the accuracy of the test compared with the reference test (first step). To make judgments about importance to patients of this information, patients with a positive test (or strategy) in either group are (or have been in previous studies) submitted to treatment or no treatment; investigators then evaluate and compare patient-important outcomes in all patients in both groups (second step)

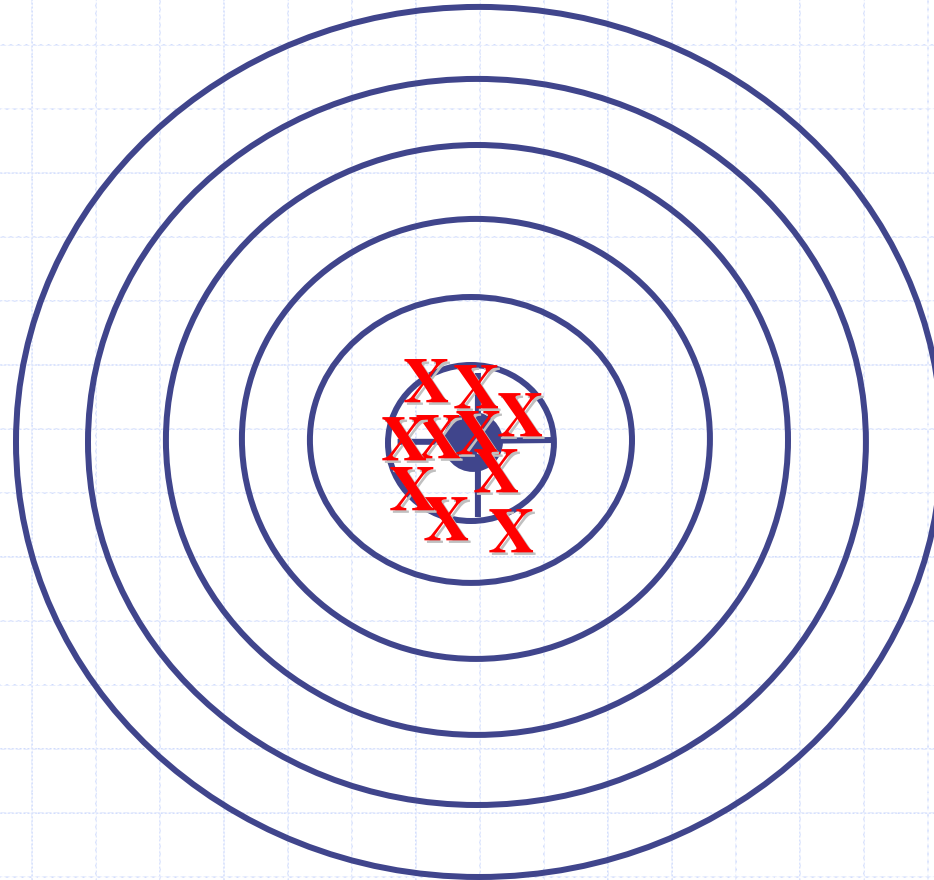
Two key properties of any test

- ◆ Accuracy (also called ‘validity’)
- ◆ Precision (also called ‘reliability’ or ‘reproducibility’)

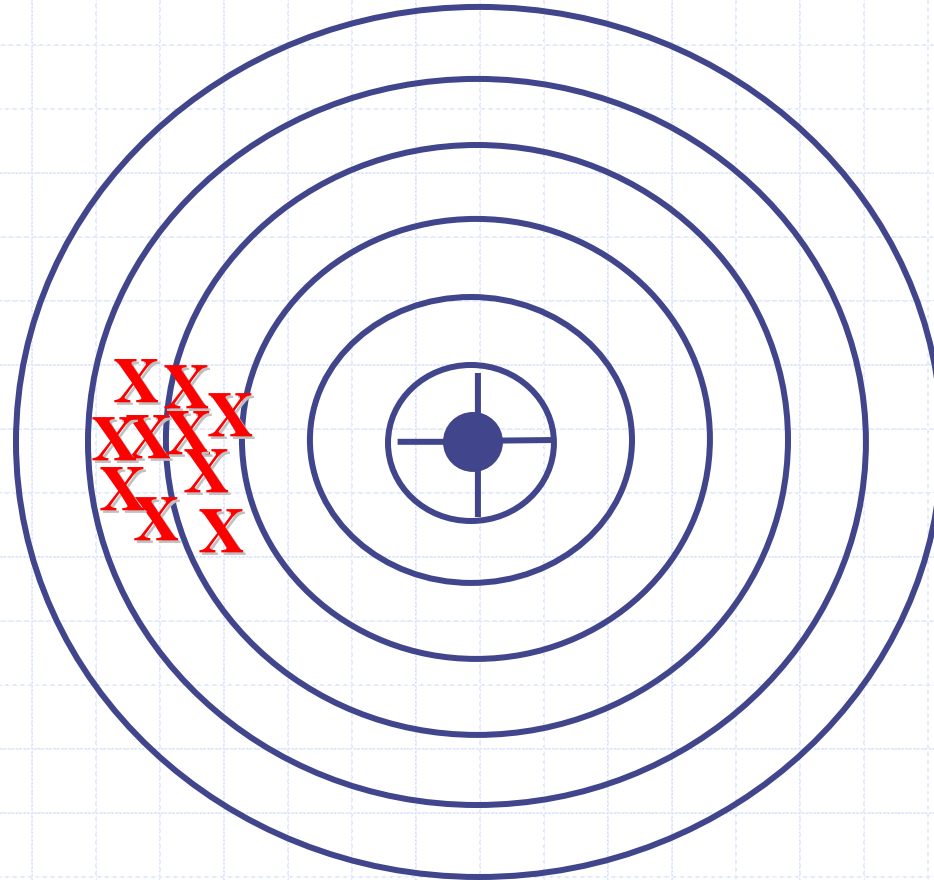
Precision and Accuracy



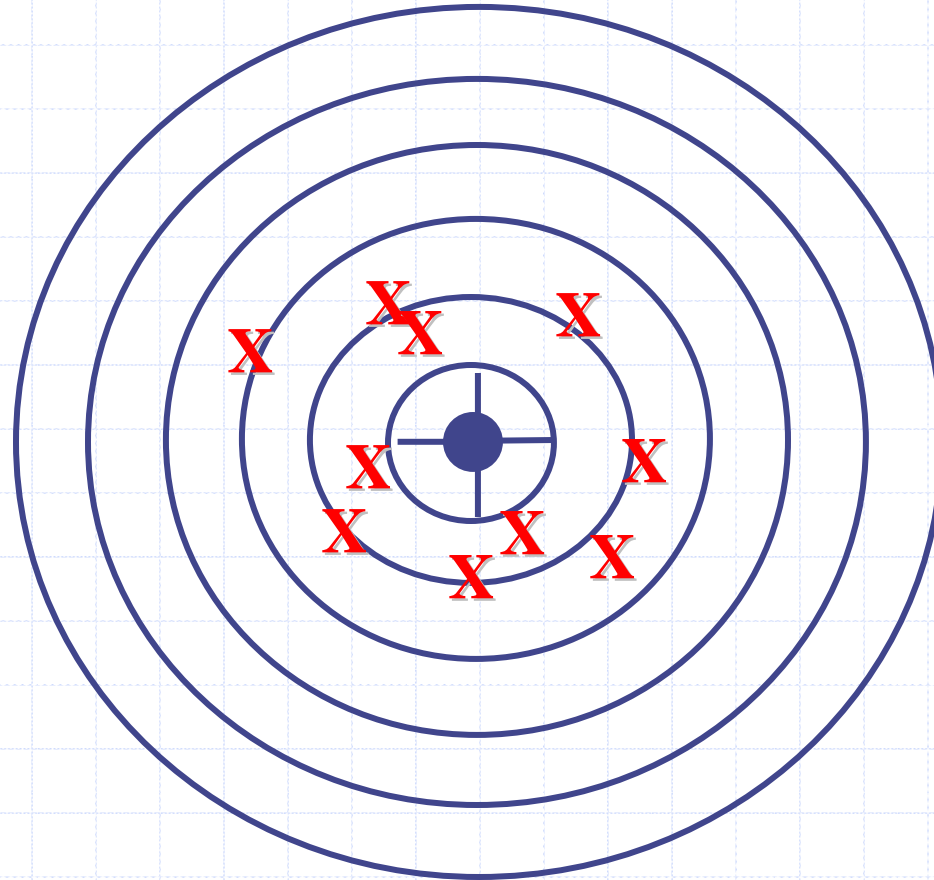
Precision and Accuracy



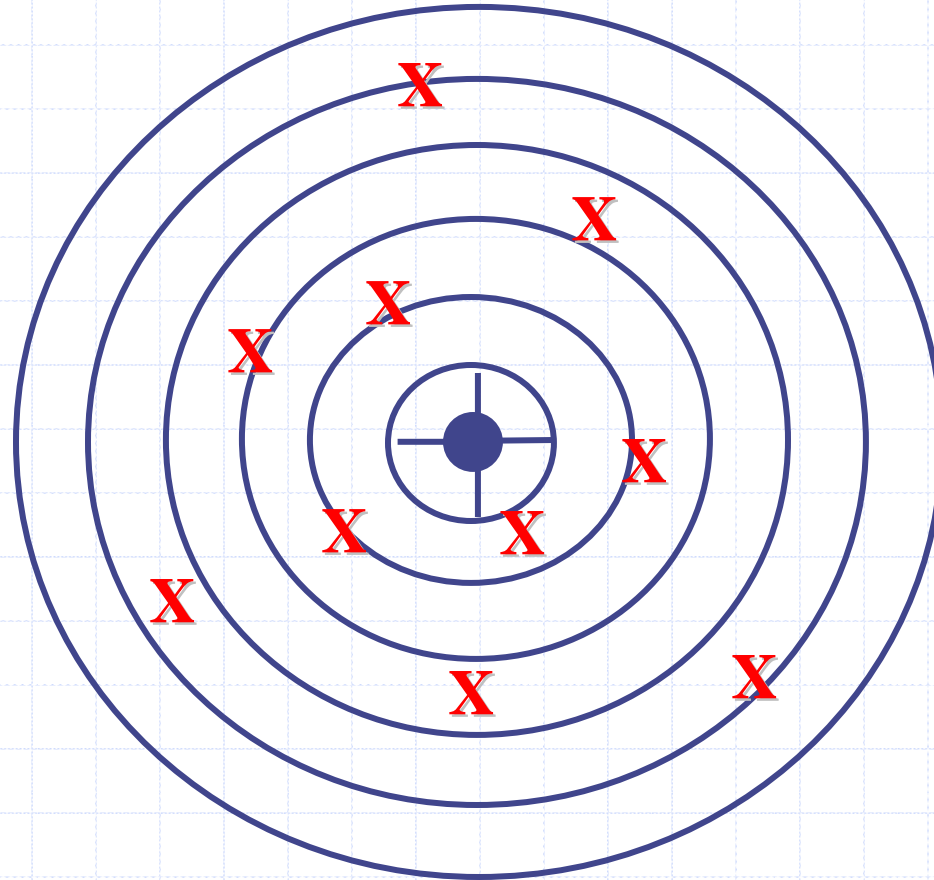
Precision and Accuracy



Precision and Accuracy



Precision and Accuracy



Quantifying precision

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)

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]

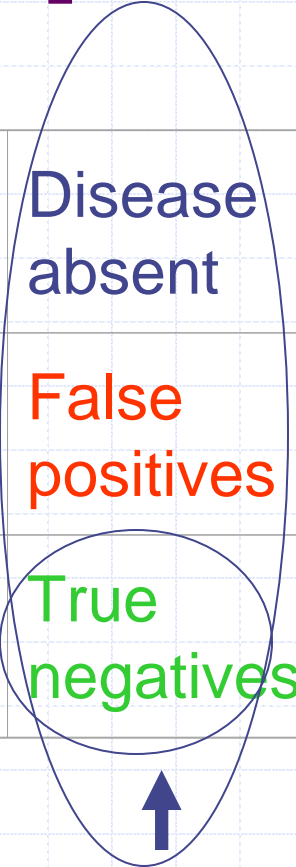
	Disease present	Disease absent
Test positive	True positives	False positives
Test negative	False negative	True negatives

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

Specificity

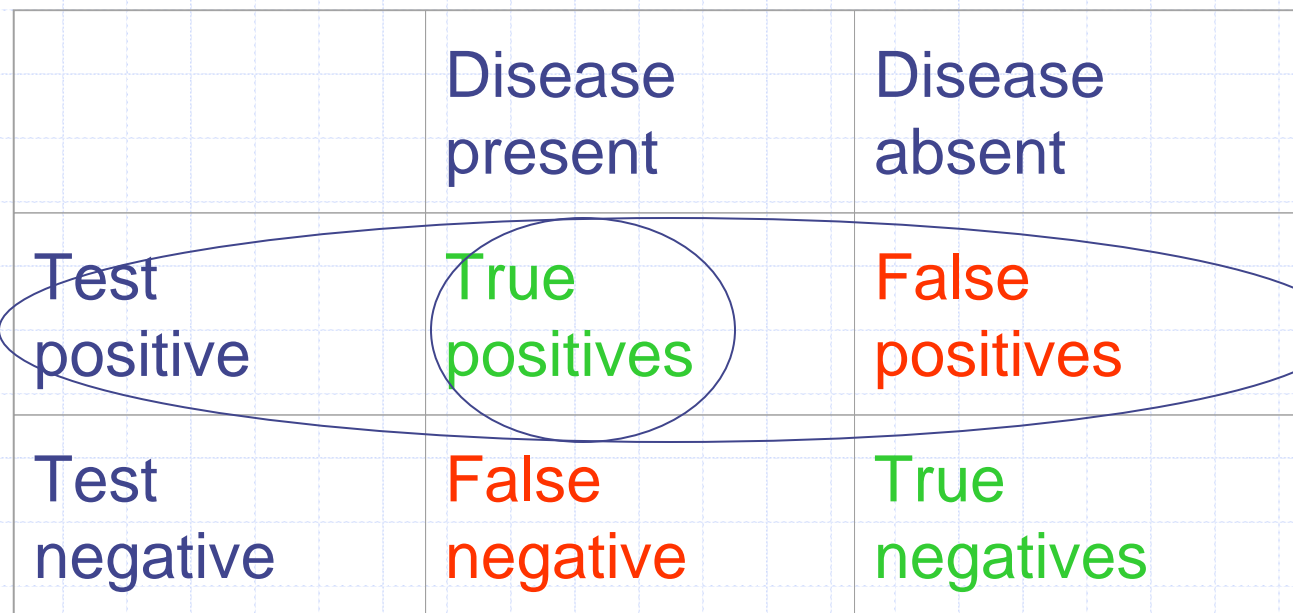
[true negative rate]

	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: $P(T-|D-) = TN / (TN + FP)$.

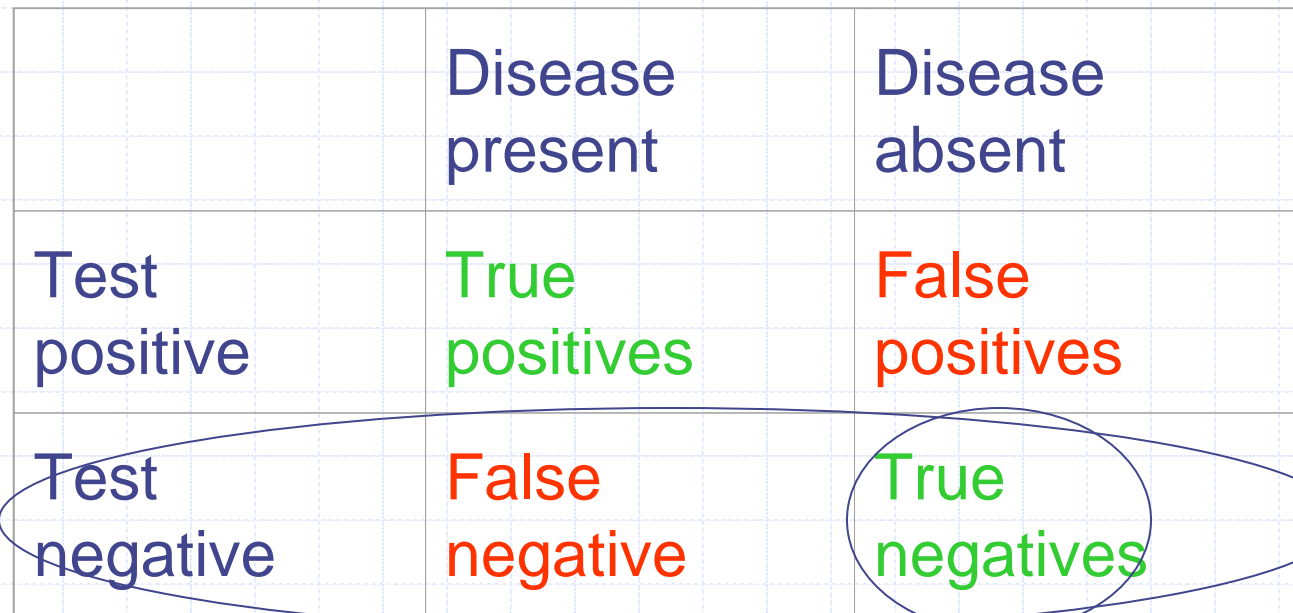
Predictive value of a positive test



	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 = $P(D-|T-) = TN / (TN+FN)$

If you hate formulae and numbers, then...

Understanding sensitivity and specificity with the right side of the brain

Tze-Wey Loong

Can you explain why a test with 95% sensitivity might identify only 1% of affected people in the general population? The visual approach in this article should make the reason clearer

Department of
Community,
Occupational, and
Family Medicine,
National University
of Singapore,
Singapore

Tze-Wey Loong
*clinical teacher
(part time)*

Correspondence to:
T-W Loong, King
George's Medical
Centre, Block 803
King George's
Avenue, #01-144,
Singapore 200803,
Singapore
tzewey@
singnet.com.sg

BMJ 2003;327:716-9

I first encountered sensitivity and specificity in medical school. That is, I remember my eyes glazing over on being told that “sensitivity = $TP / (TP + FN)$, where TP is the number of true positives and FN is the number of false negatives.” As a doctor I continued to encounter sensitivity and specificity, and my bewilderment turned to frustration—these seemed such basic concepts; why were they so hard to grasp? Perhaps the left (logical) side of my brain was not up to the task of comprehending these ideas and needed some help from the right (visual) side. What follows are diagrams that were useful to me in attempting to better visualise sensitivity, specificity, and their cousins positive predictive value and negative predictive value.

Sensitivity and specificity

I will be using four symbols in these diagrams (fig 1).

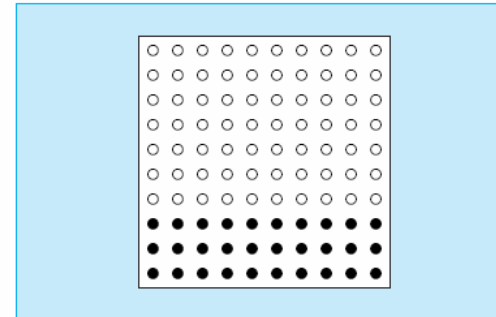
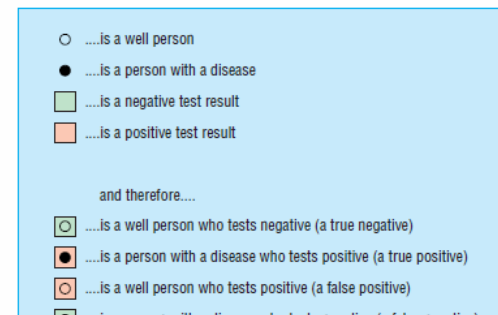


Fig 2 Hypothetical population

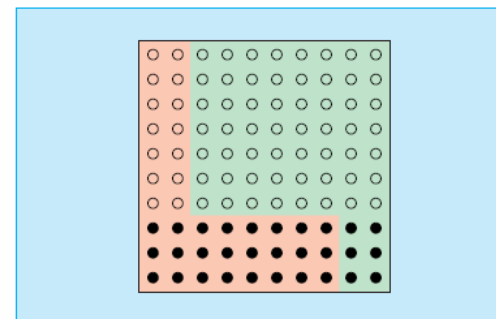


Fig 3 Results of diagnostic test on hypothetical population

Sensitivity refers to how good a test is at correctly identifying people who have the disease. When

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%

For a given test, predictive values will depend on prevalence

Test with 80% sensitivity and 90% specificity:

	pre-test probability (disease prevalence)			
	1%	10%	50%	90%
PPV	7.5%	47.1%	88.9%	98.6%
NPV	99.8%	97.6%	81.8%	33.3%

[Reigelman 1996]

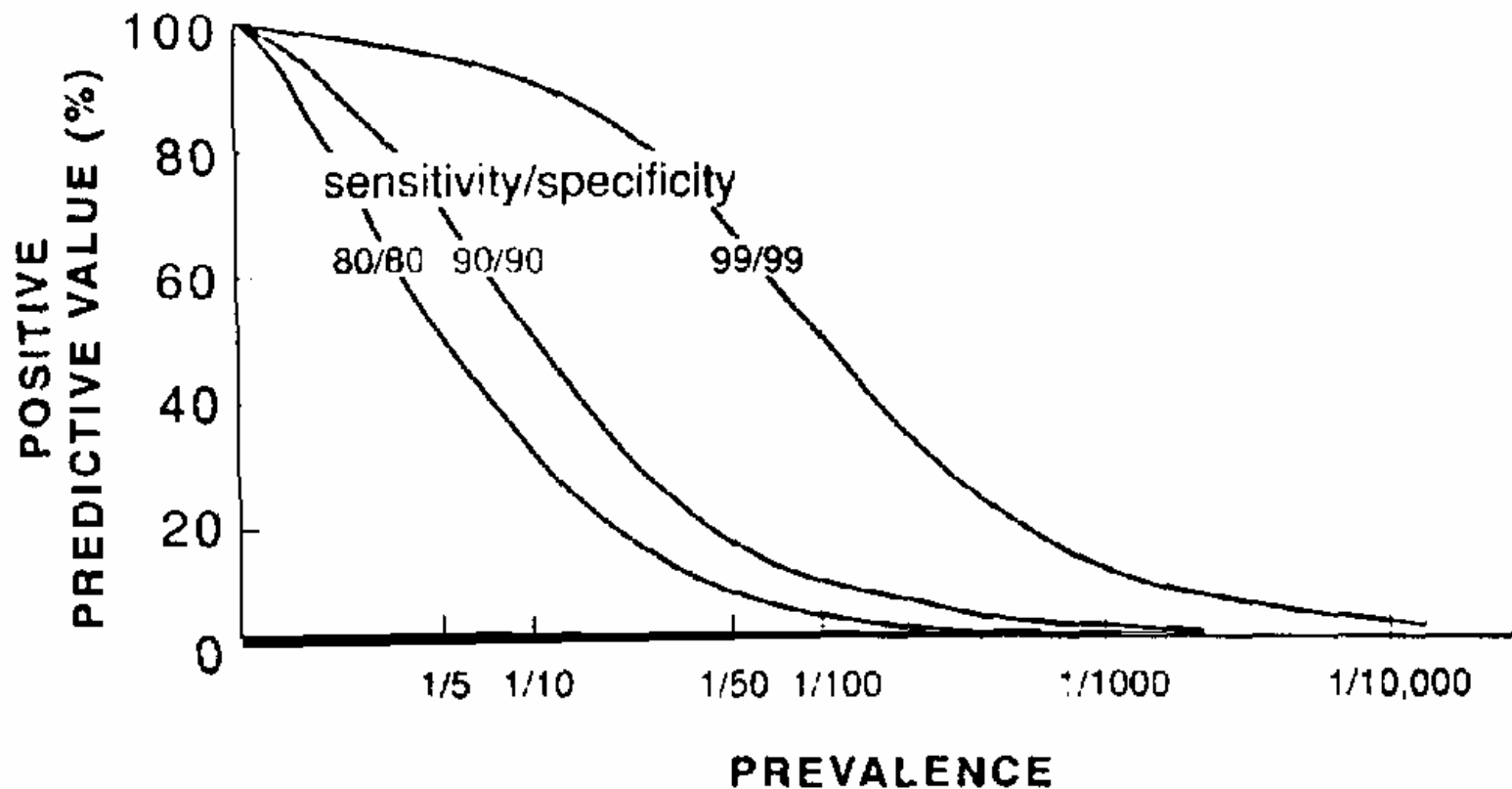
For a given test, predictive values will depend on prevalence

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

Setting	Prevalence (Cases/100,000)	Positive 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')

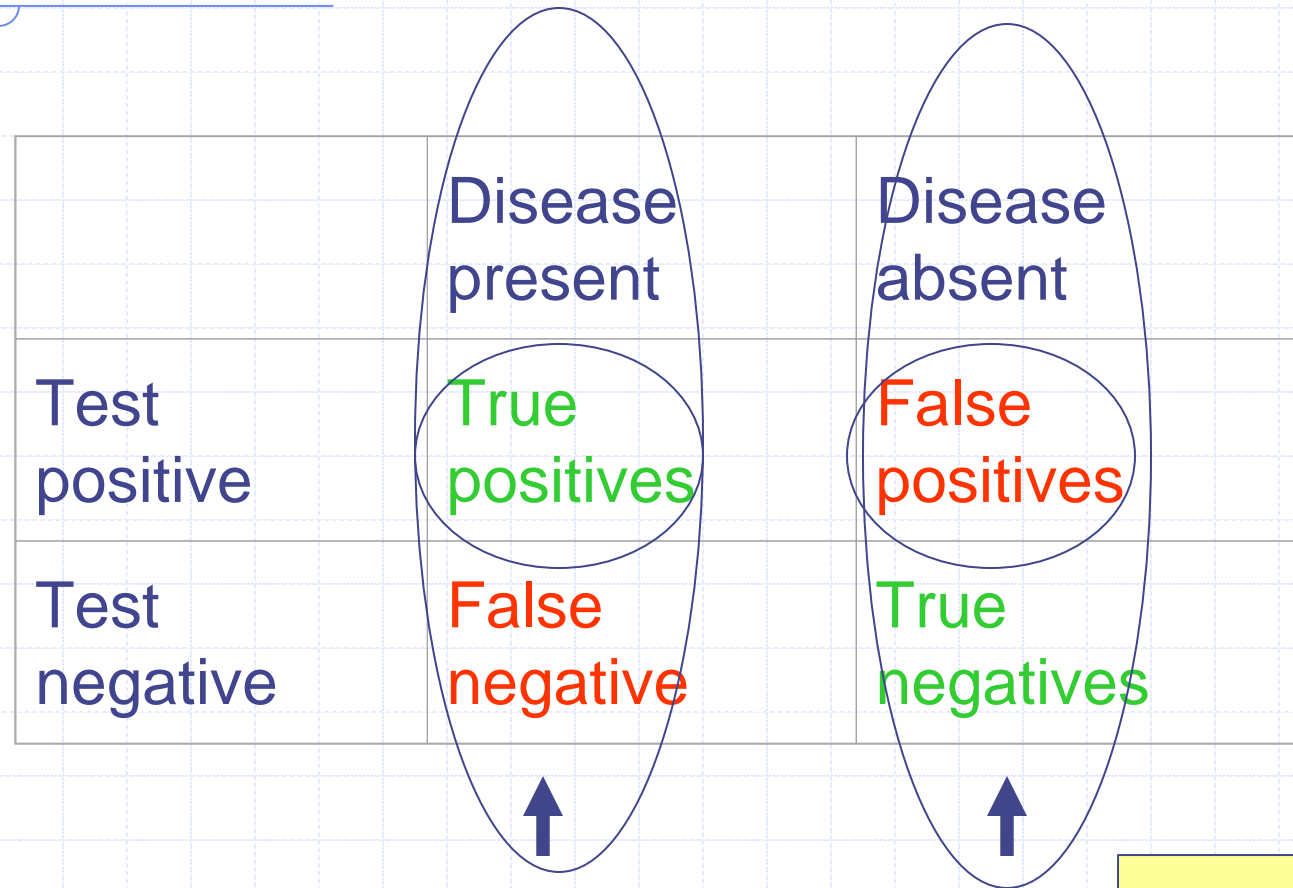
- Likelihood ratio of a positive test: is the test more likely to be positive in diseased than non-diseased persons?

- $LR+ = TPR / FPR$

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

- 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

$$LR+ = TPR / FPR$$

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

Likelihood Ratios

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

- $LR- = FNR / TNR$

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

- 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

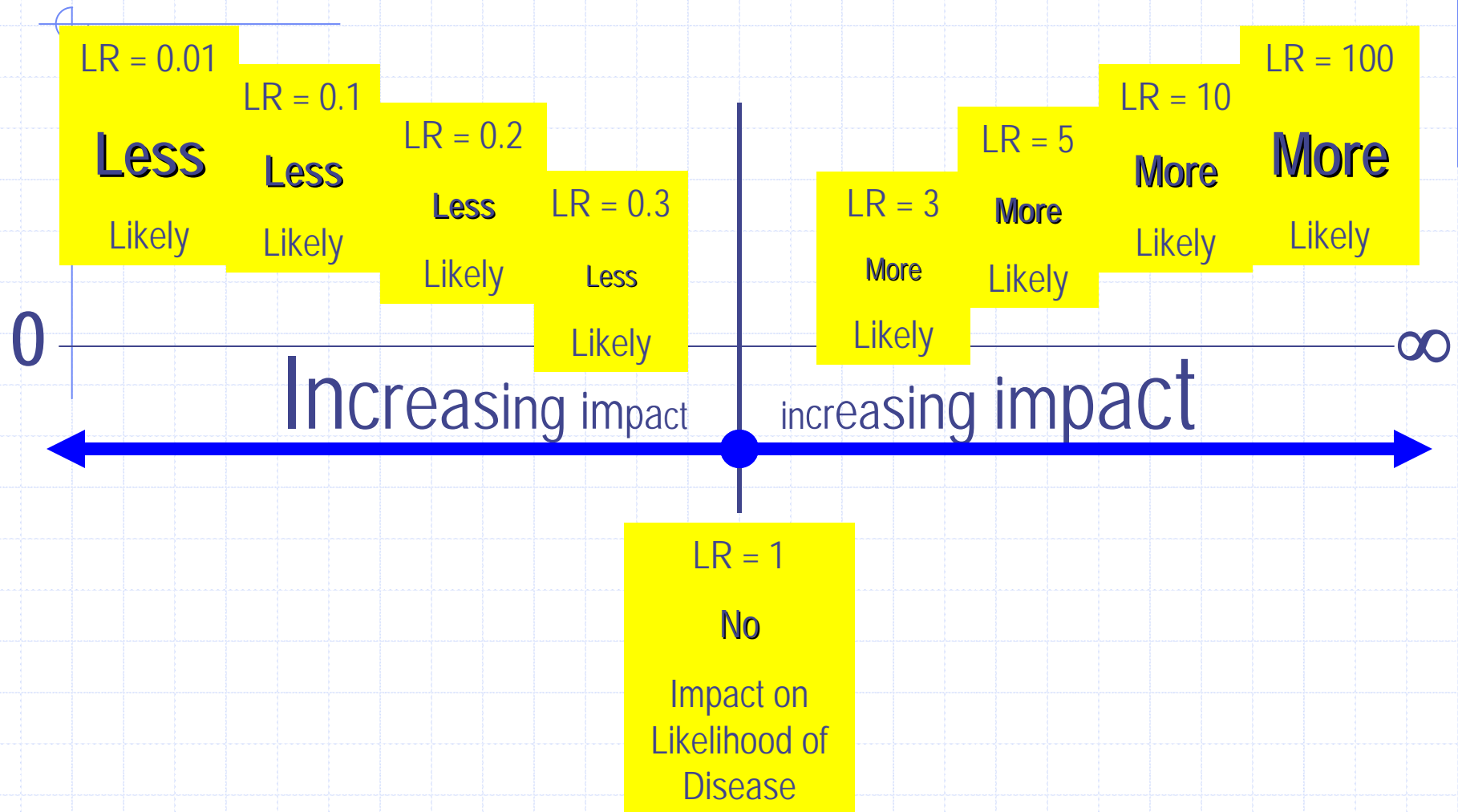
	Disease present	Disease absent
Test positive	True positives	False positives
Test negative	False negative	True negatives

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

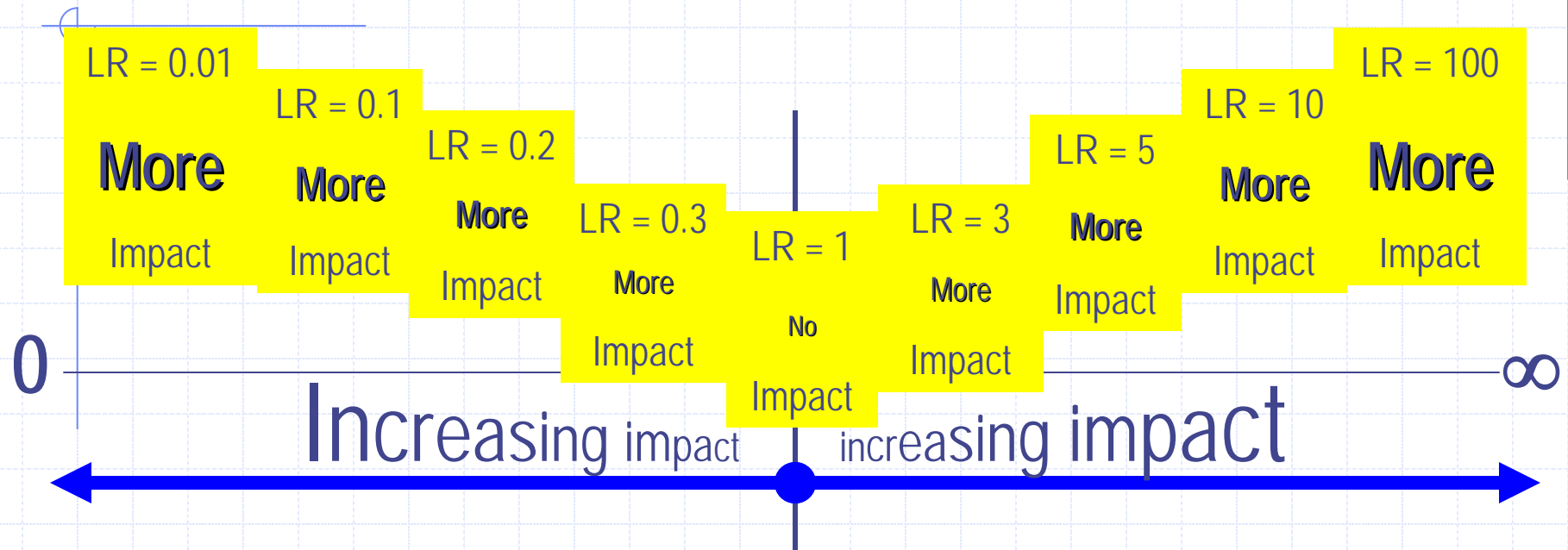
$$LR- = FNR / TNR$$

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

LR: Impact on Likelihood of Disease



LR: Impact on Likelihood of Disease



Quick review of odds vs. probability

◆ odds = probability / (1 – probability)

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

◆ probability = odds / (1 + odds)

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

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

$$\text{DOR} = (a/c) / (b/d)$$

$$\text{DOR} = ad / bc$$

$$\text{DOR} = \text{Odds of } T+|D+ / \text{Odds of } T+|D-$$

Example: Serological test for TB

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

$$LR+ = 2$$

$$LR- = 0.9$$

$$DOR = 2.4$$

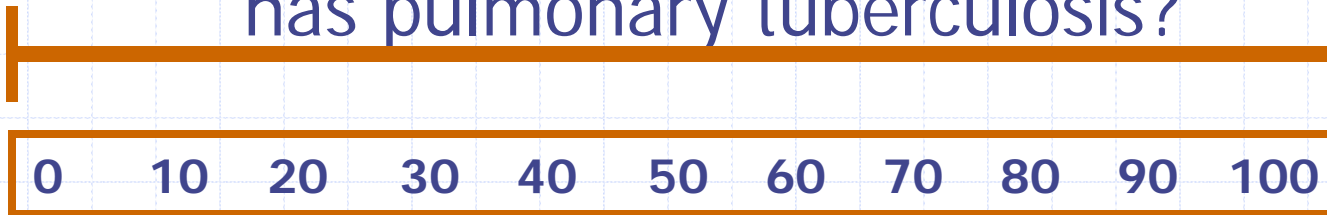
Using LRs in practice

◆ Scenario:

- Mr. A, a 27-year old S African black male
- 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?

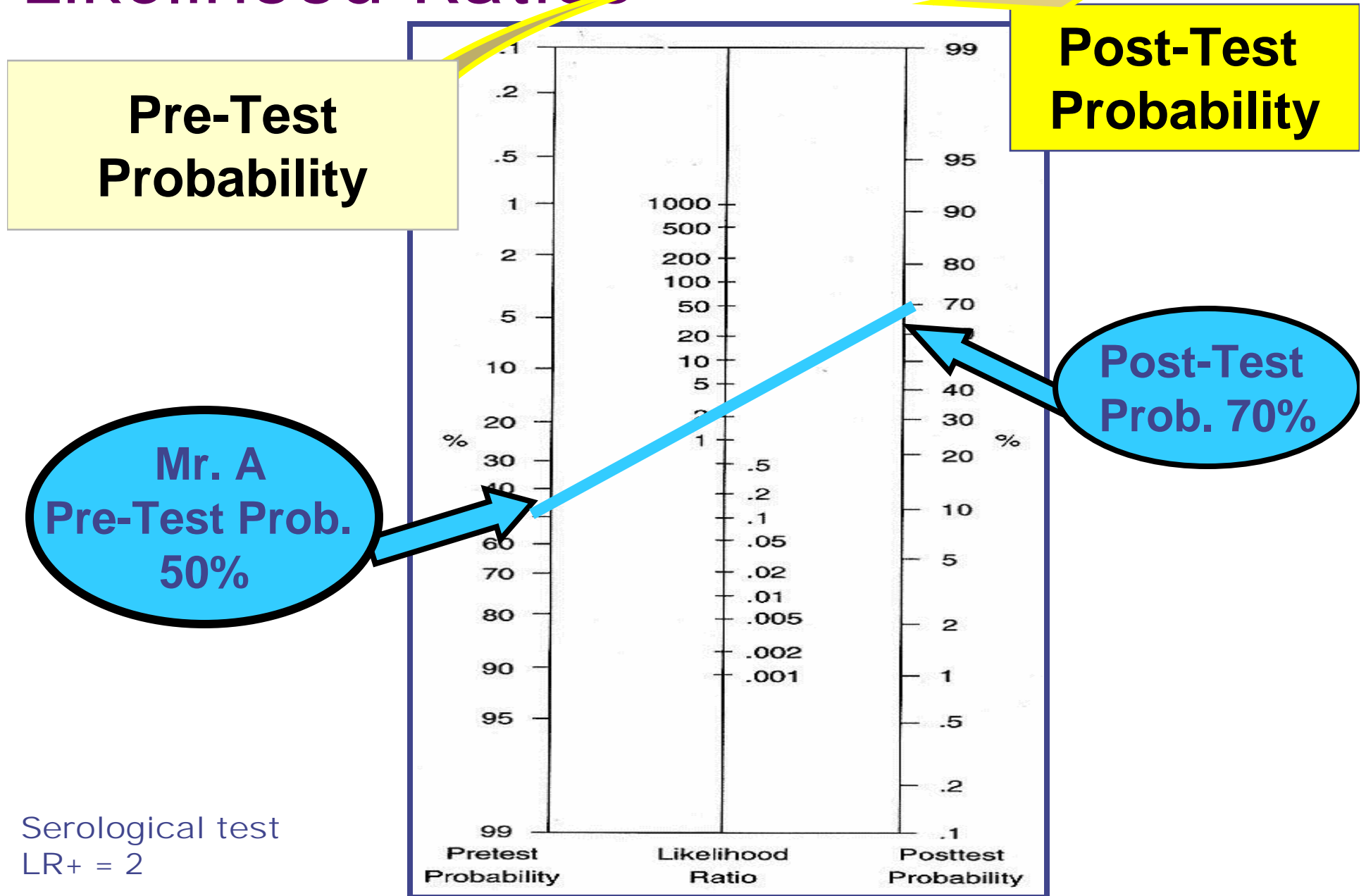


Pre-Test Probability

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

Post-Test Probability

Likelihood Ratios



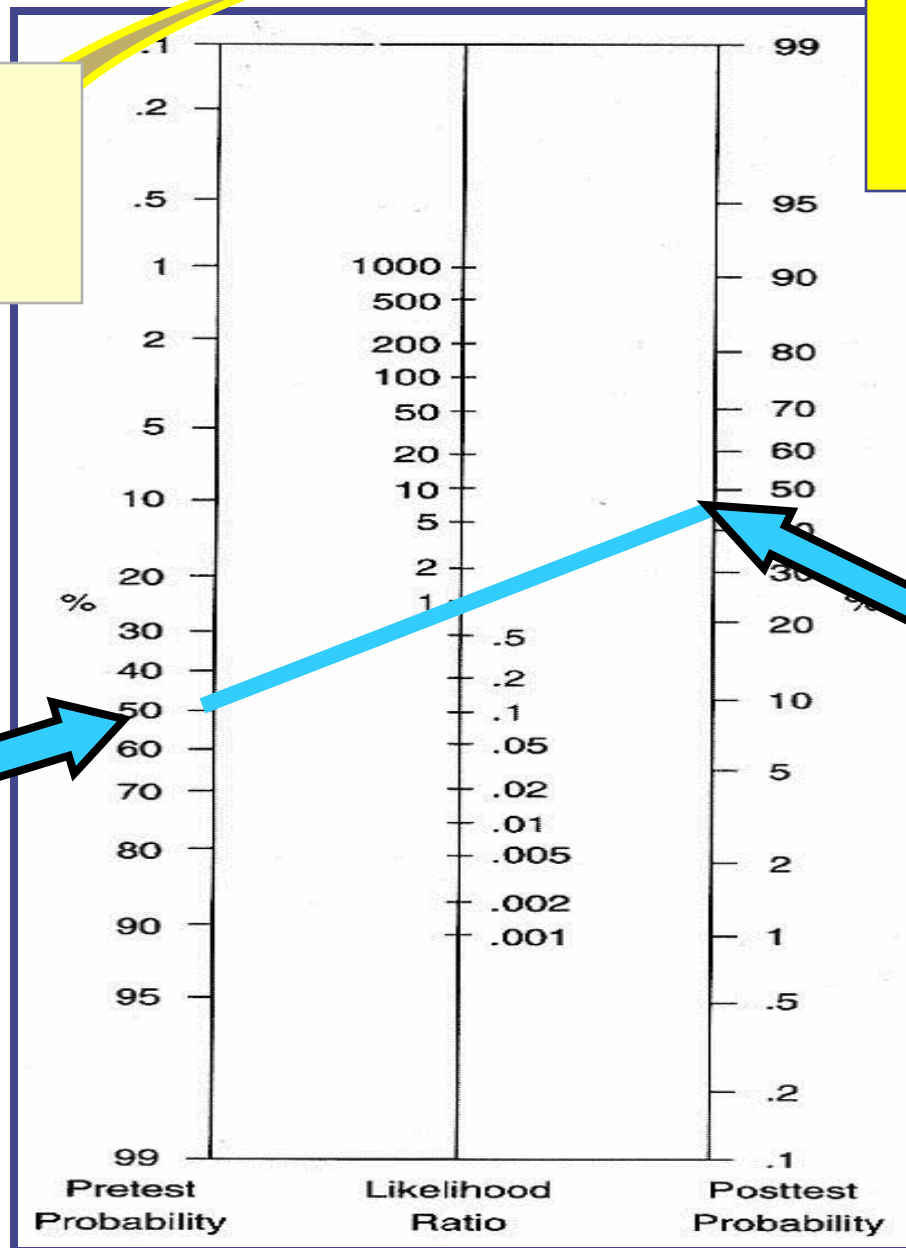
Likelihood Ratios

**Pre-Test
Probability**

**Post-Test
Probability**

**Mr. A
Pre-Test Prob.
50%**

**Post-Test
Prob. 45%**



Serological test
 $LR^- = 0.9$

Using LR in practice

◆ Scenario:

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

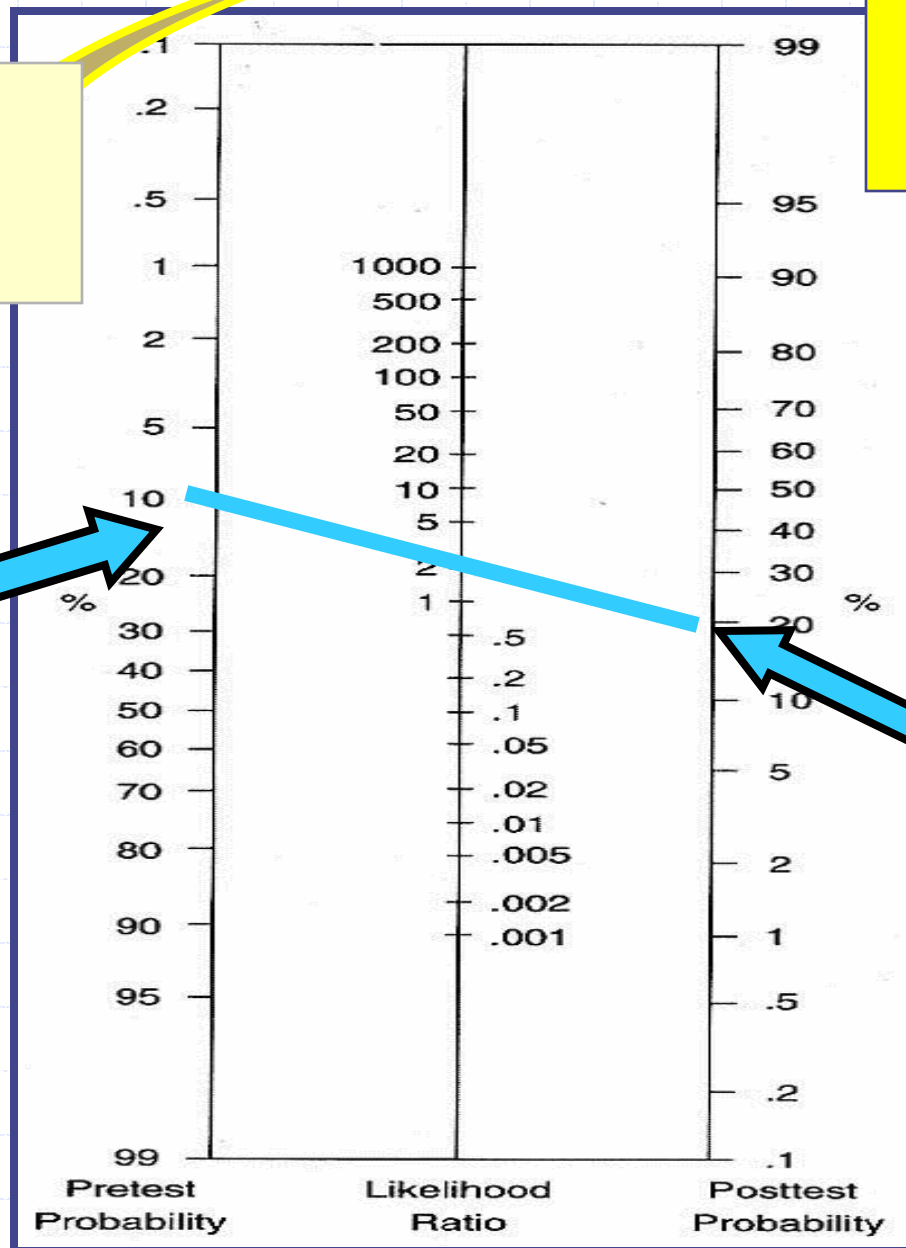
Likelihood Ratios

**Pre-Test
Probability**

**Post-Test
Probability**

**Ms. B
Pre-Test Prob.
10%**

**Post-Test
Prob. 20%**



Serological test
 $LR+ = 2$

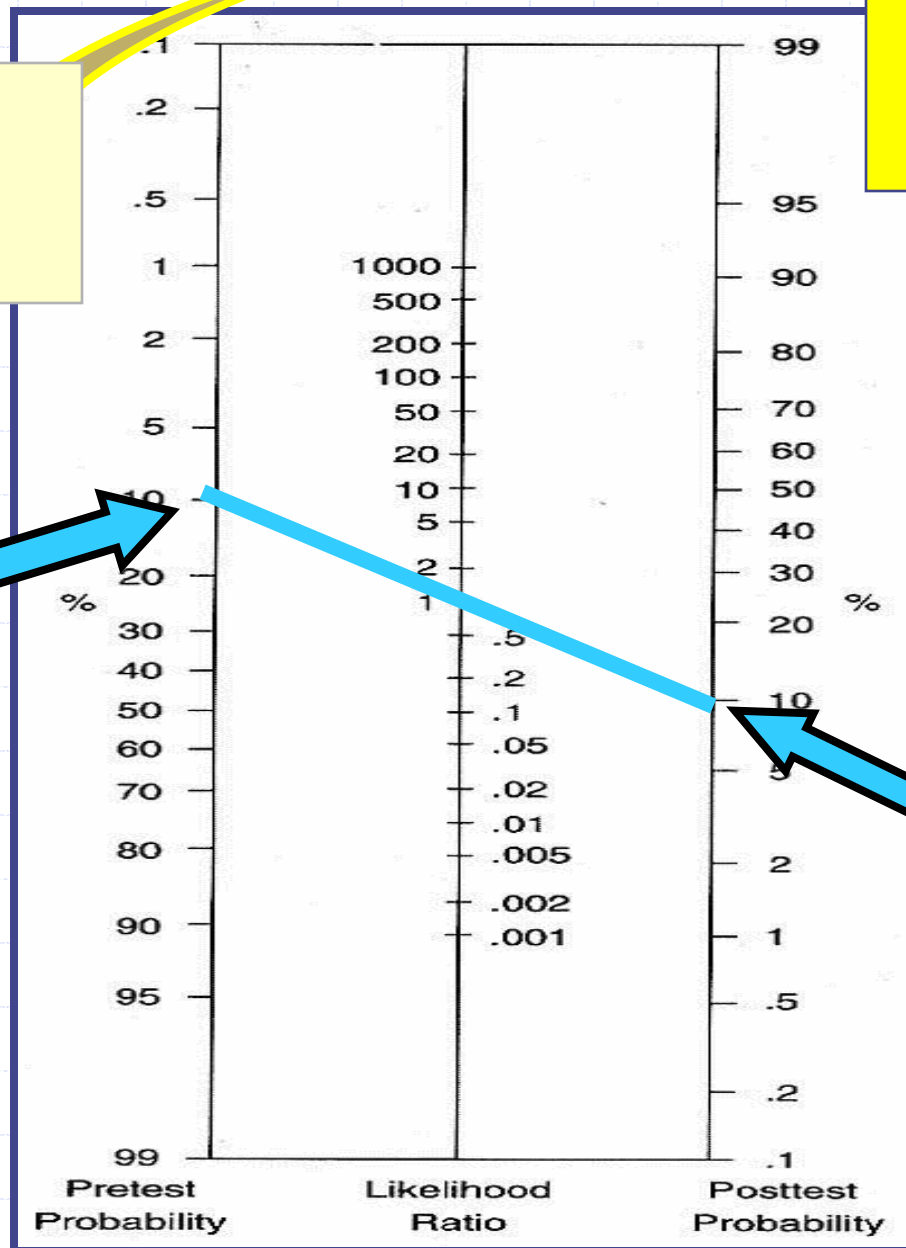
Likelihood Ratios

**Pre-Test
Probability**

**Post-Test
Probability**

**Ms. B
Pre-Test Prob.
10%**

**Post-Test
Prob. 10%**



Serological test
 $LR^- = 0.9$

Example: Ultrasonography for Down Syndrome



Another example: Nuchal fold & Down Syndrome

Down Syndrome

		Yes	No	
Nuchal fold	Positive	21	4	25
	Negative	7	188	195
		28	192	220

Sensitivity = 75%

Specificity = 98%

LR+ = 36

LR- = 0.26

DOR = 141

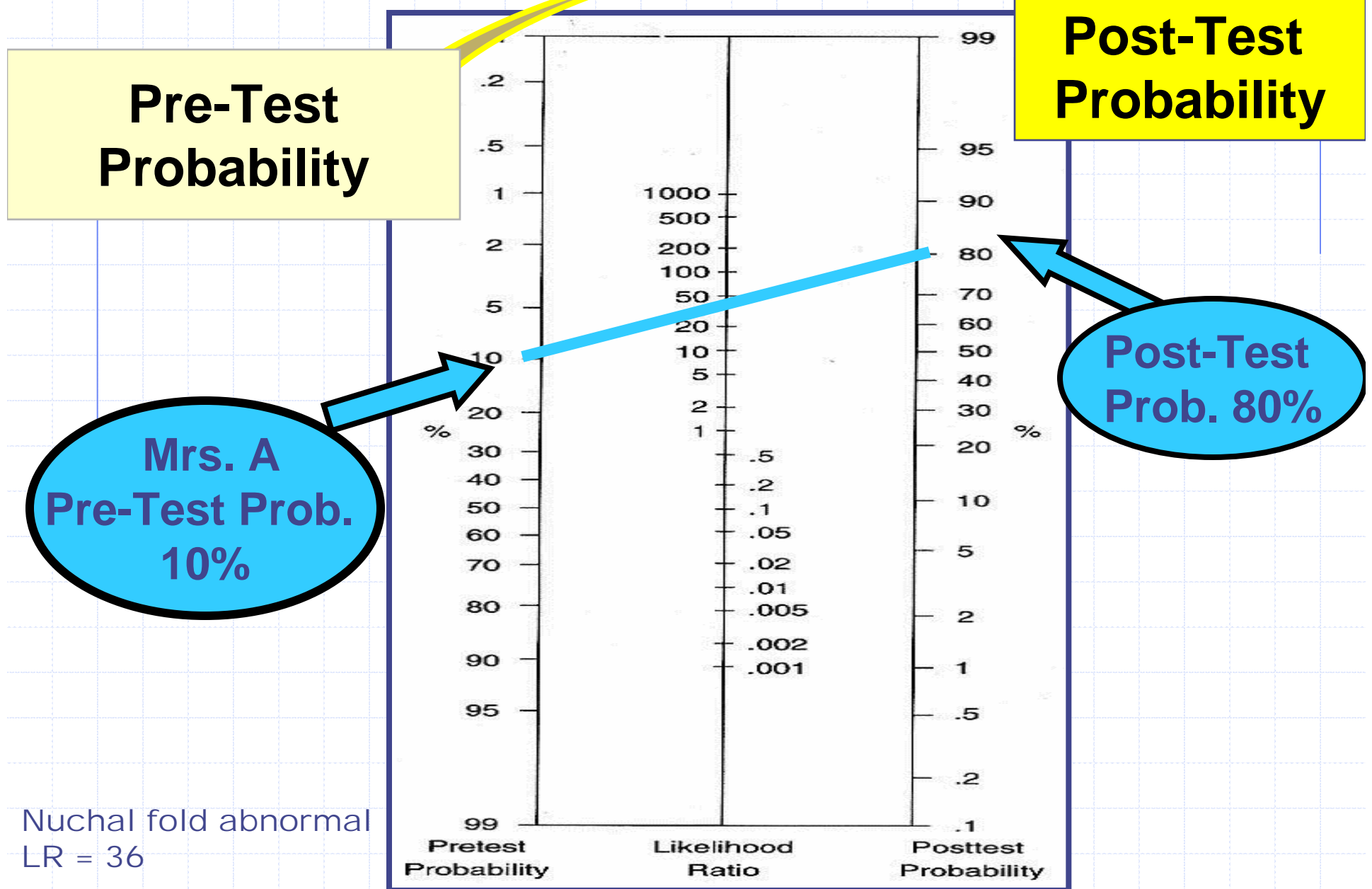
N Engl J Med 1987;317:1371

Using LRs in practice

◆ Scenario:

- Mrs. A, a 37-year old woman with a previous affected pregnancy, seen at a high-risk clinic in a tertiary, referral hospital
- What is the pretest probability of Down syndrome in this case?

Likelihood Ratios



Likelihood Ratios

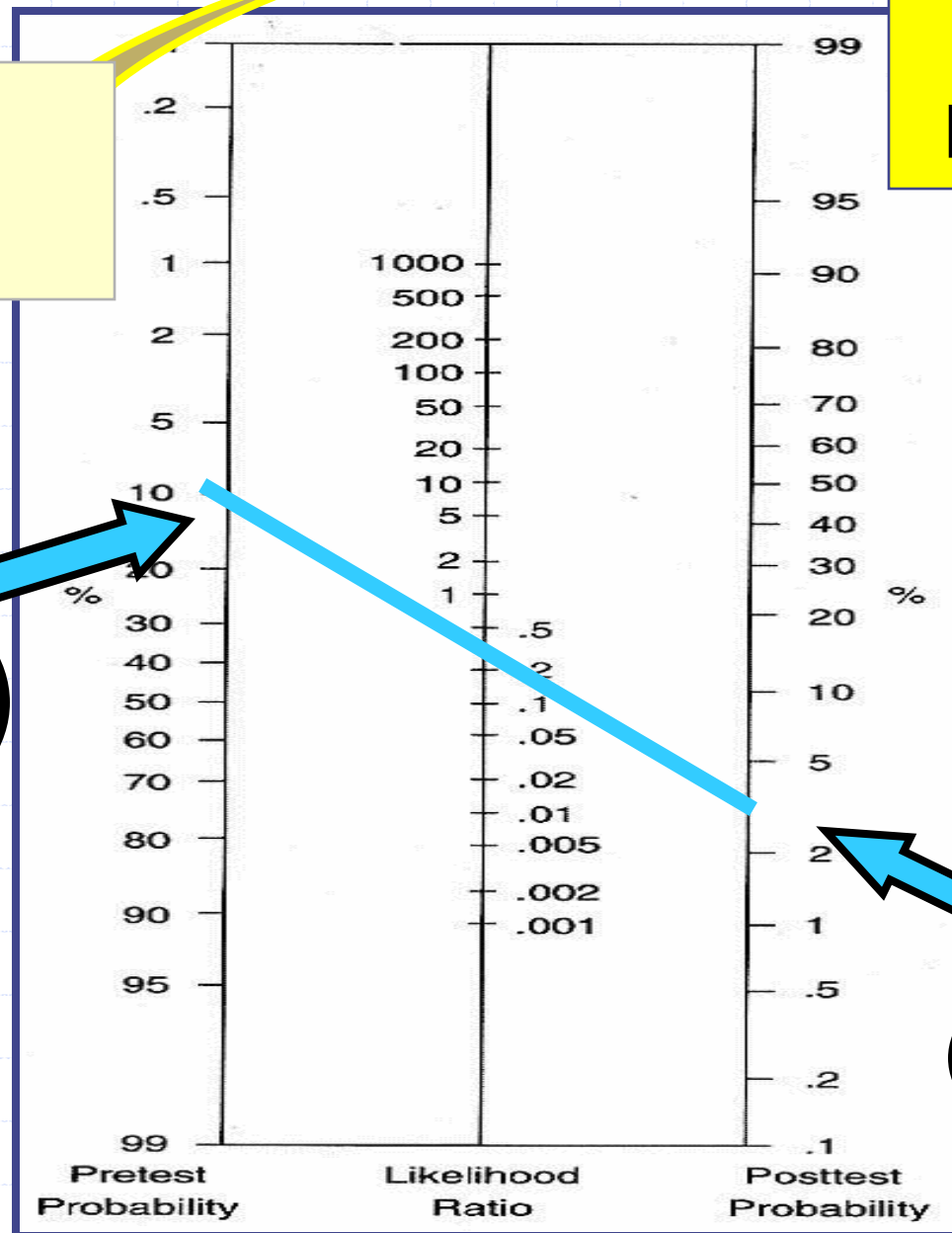
**Pre-Test
Probability**

**Post-Test
Probability**

**Mrs. A
Pre-Test Prob.
10%**

**Post-Test
Prob. 3%**

Nuchal fold normal
LR = 0.26



Using LRs in practice

◆ Scenario:

- Mrs. B, a 20-year old woman with a previous normal pregnancy, seen at a community hospital
- What is the pretest probability of Down syndrome in this case?

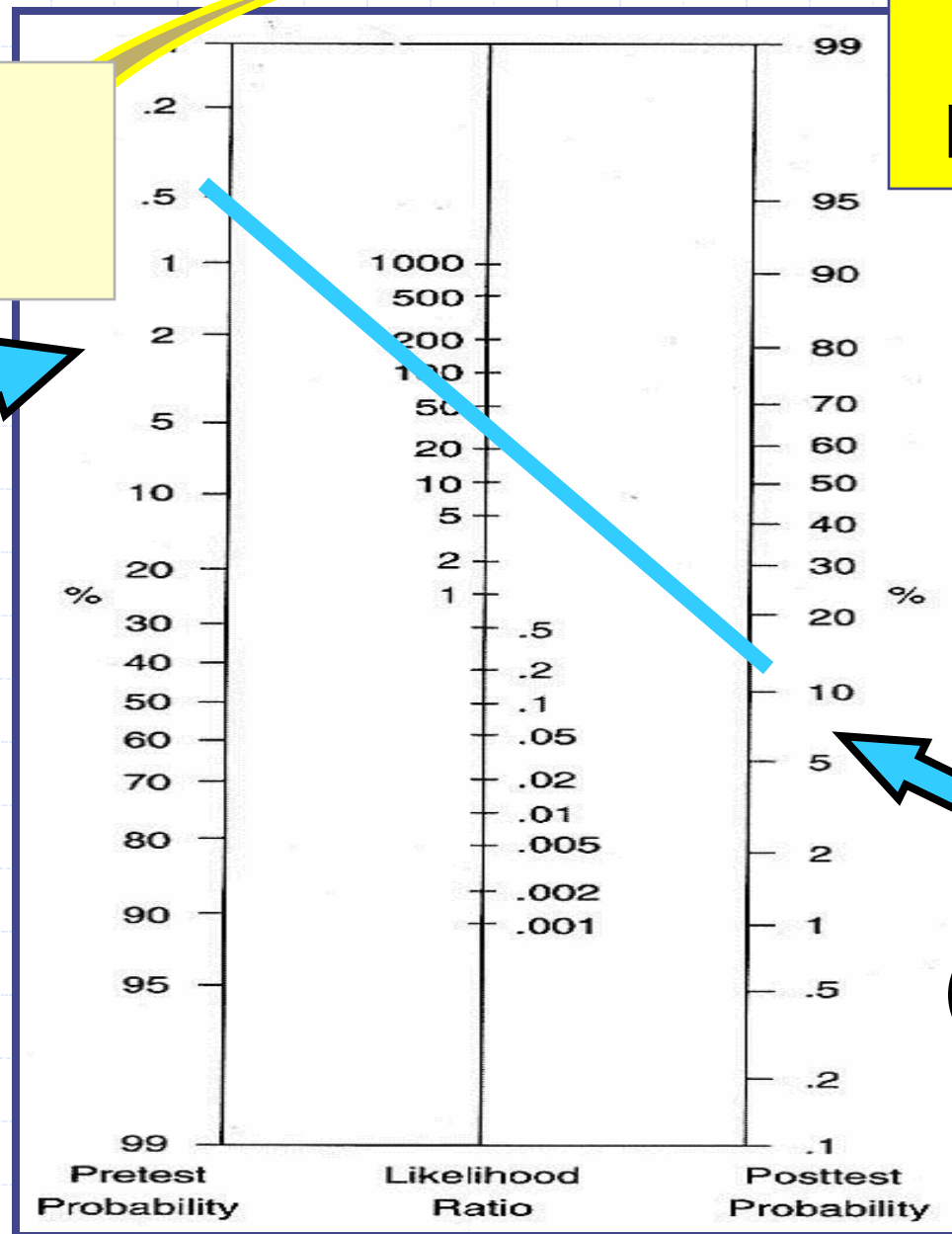
Likelihood Ratios

**Pre-Test
Probability**

**Post-Test
Probability**

**Mrs. B
Pre-Test Prob.
0.5%**

**Post-Test
Prob. 10%**



Nuchal fold abnormal
LR = 36

Likelihood Ratios

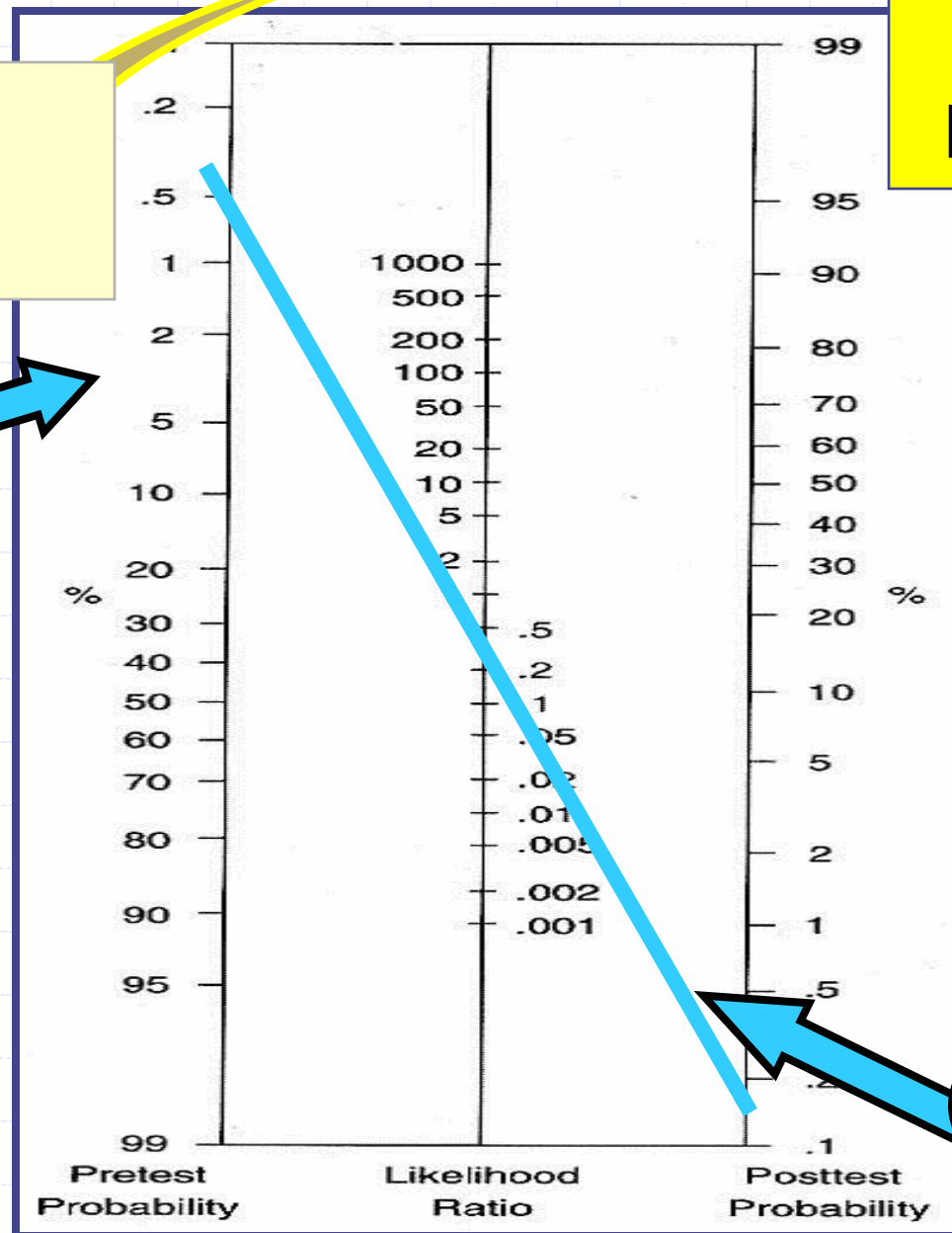
**Pre-Test
Probability**

**Post-Test
Probability**

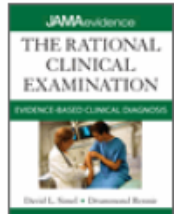
**Mrs. B
Pre-Test Prob.
0.5%**

**Post-Test
Prob. 0.1%**

Nuchal fold normal
LR = 0.26



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

Copyright © American Medical Association. All rights reserved. | JAMA | The McGraw-Hill Companies, Inc.

Examples

	Prior Probability	Test/Finding	LR+	LR-
Chapter 41: Pneumonia, Infant and Child	15% to 35% prevalence of pneumonia given cough or respiratory symptoms	Grunting among children with wheezing, < 18 mo	2.8 (1.6–4.4)	0.7 (0.55–0.89)
		Retraction	2.7 (1.1–6.9)	0.97 (0.93–1.0)
		Rales	1.8–15	0.69–0.86
		Tachypnea (use WHO adjusted criteria)	1.6–8.0	0.32–0.91

	Prior Probability	Test/Finding	LR+	LR-
Chapter 51: Urinary Tract Infection, Women	48% among women with compatible symptoms	Dysuria	1.5 (1.2–2.0)	0.5 (0.3–0.7)
		Frequency	1.8 (1.1–3.0)	0.5 (0.4–1.0)
		Vaginal discharge	0.3 (0.1–0.9)	3.1 (1.0–9.3)
		Vaginal irritation	0.2 (0.1–0.9)	2.7 (0.9–8.5)
		Dipstick result	4.2	0.3

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.

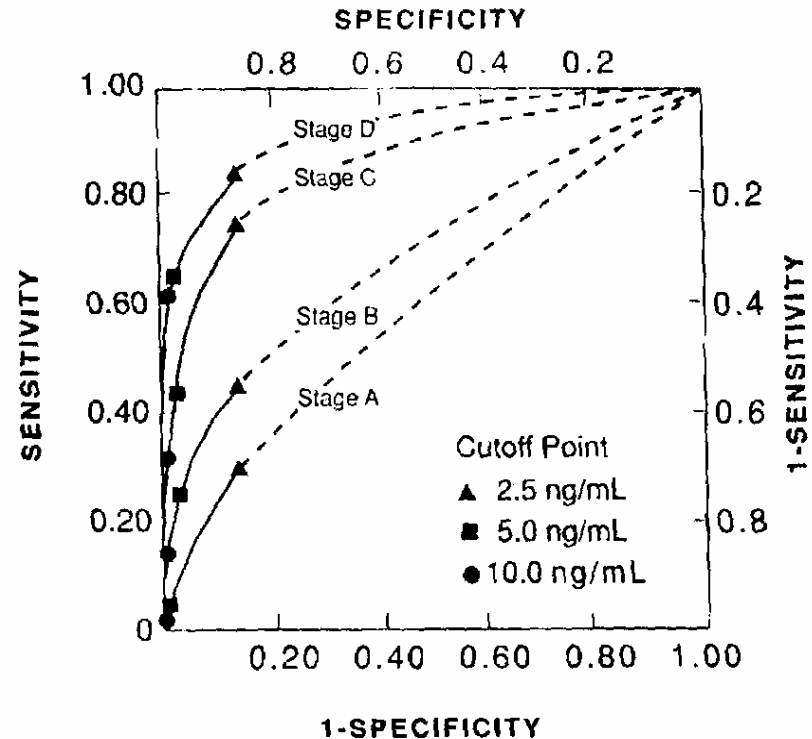
Are sens/spec and LRs inherent properties of a test?

- ◆ Most textbooks will say that sens and spec do not depend on disease prevalence
- ◆ This is partly true but oversimplified
- ◆ 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.)



Tests with continuous or multi-level results

Example: WBC count in bacteremia

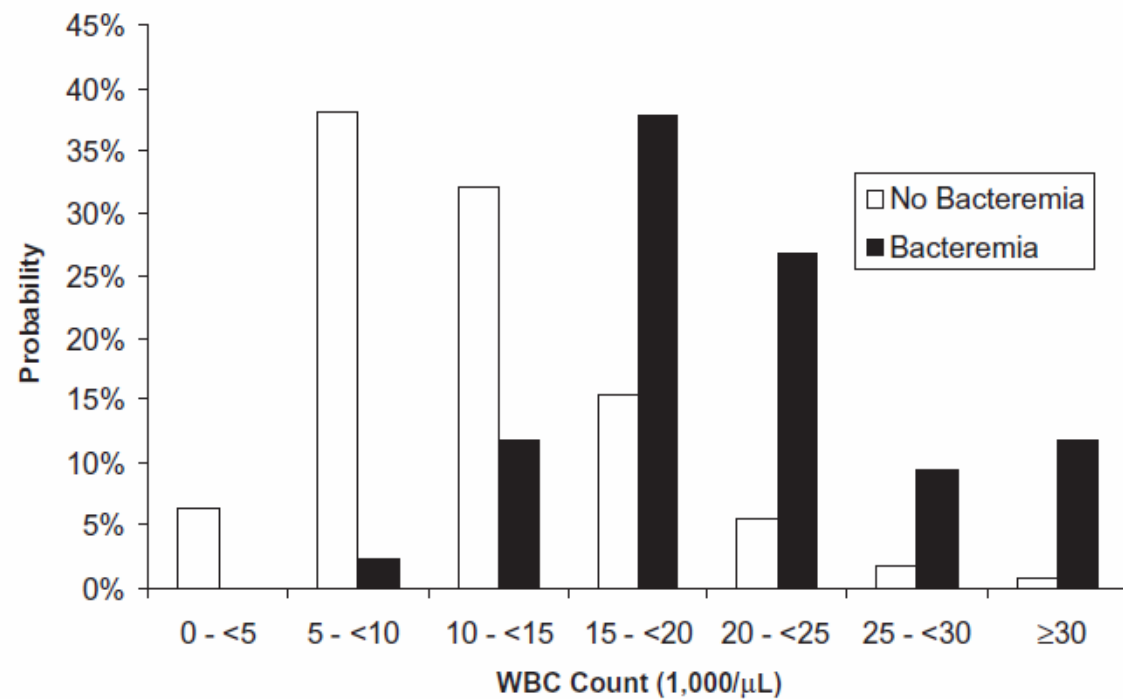


Figure 4.4 Histogram showing distributions of the nonbacteremic and bacteremic populations across the WBC count intervals.

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 ($\times 1,000/\mu\text{L}$)	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%

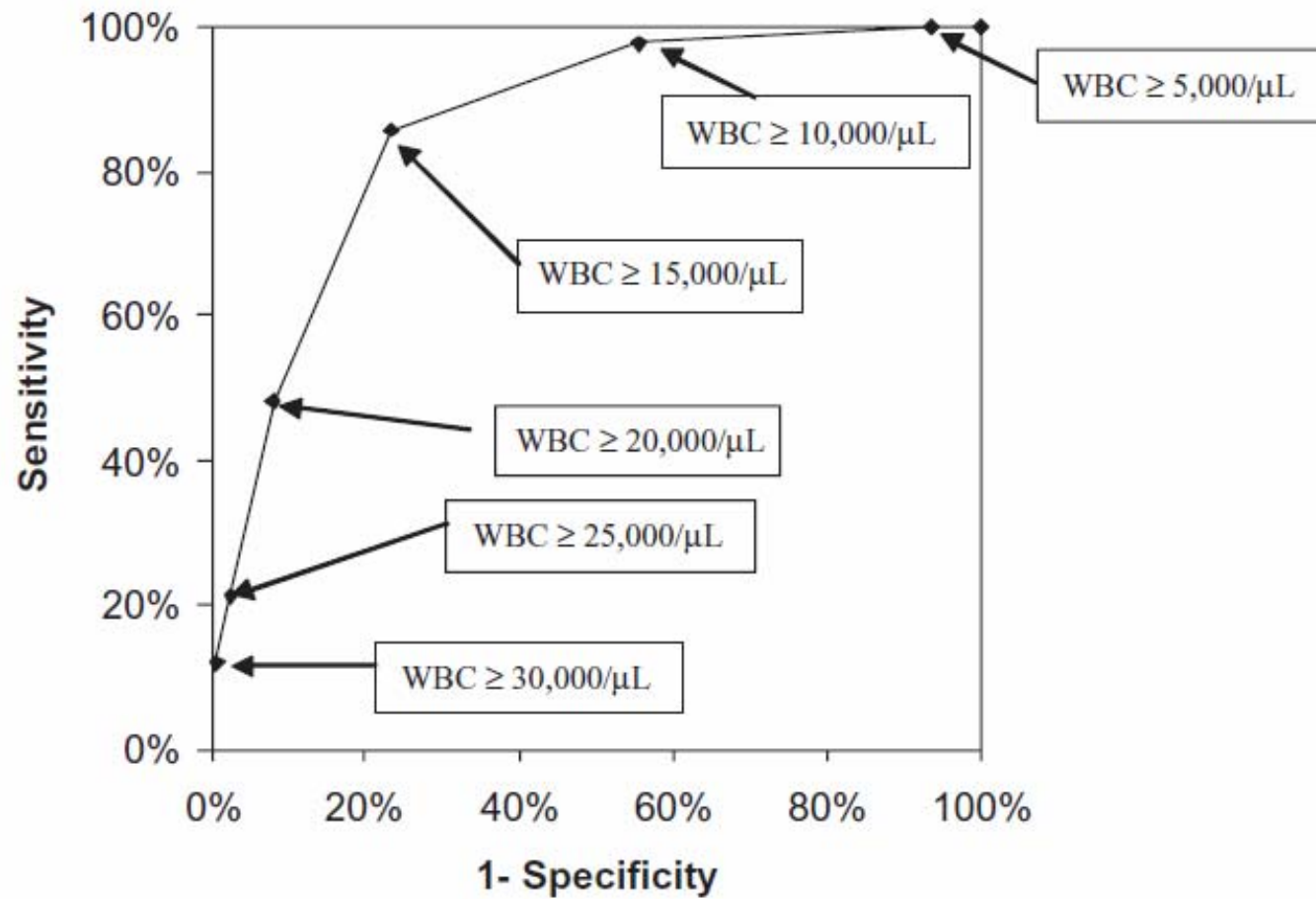
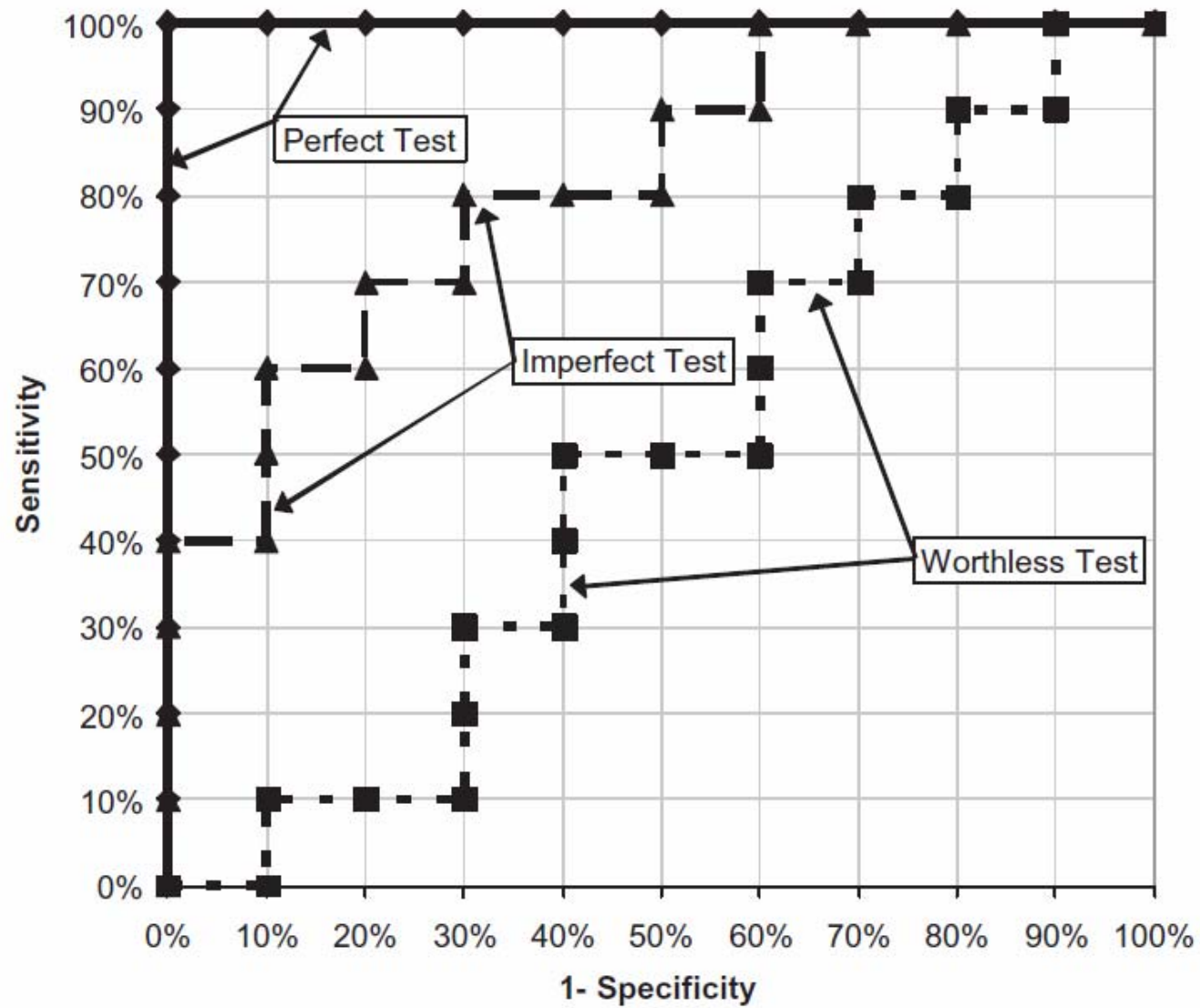


Figure 4.5 ROC curve corresponding to the distributions in Figure 4.4.



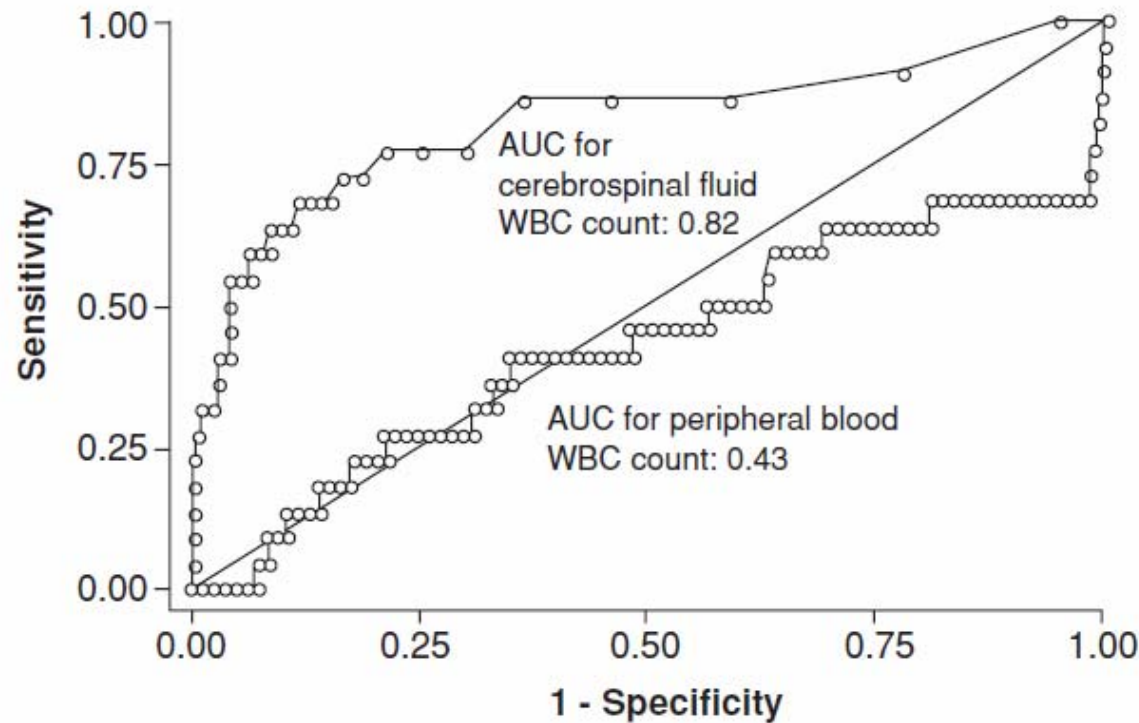


Figure 4.6 Example of computer-drawn ROC curves, in which the cut-off for considering the test “abnormal” is systematically decreased from the highest to the lowest values observed in infants with and without bacterial meningitis. Note that two different WBC counts are considered: the WBC count in the cerebrospinal fluid, which discriminates fairly well between those with and without bacterial meningitis; and the WBC count in the peripheral blood, which discriminates poorly. (From Bonsu and Harper 2003, with permission.) AUC = Area Under Curve.

Multi-level likelihood ratios

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

WBC Count ($\times 1,000/\mu\text{L}$)	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

[Curves: Official Site](#)

Discover a gym where women
change their lives - 30 minutes
at a time.
www.Curves.com

[EKG Training](#)

Search Healthcare Schools for
EKG Training & Certification
GuideToHealthcareSchools.com

[Free SQL Training CD](#)

Step-By-Step Presentations &
Walk Throughs. Learn SQL
W/AppDev Today!
www.AppDev.com

[Jaken Medical Inc.](#)

New, Demo and Refurbished
EKG Machines 888-559-5253
www.jakenmedical.com

Ads by Google

The magnificent ROC

(Receiver Operating Characteristic curve)

"There is no statistical test, however intuitive and simple, which will not be abused by medical researchers"

Introduction - A statistical prelude

ROC CURVES WERE DEVELOPED IN THE 1950'S AS A BY-PRODUCT OF RESEARCH INTO MAKING SENSE OF RADIO signals contaminated by noise. More recently it's become clear that they are remarkably useful in medical decision-making. That doesn't mean that they are always used appropriately! We'll highlight their use (and misuse) in our tutorial. We'll first try to move rapidly through basic stats, and then address ROC curves. We'll take a practical, medical approach to ROC curves, and give a few examples.

If you know all about the terms 'sensitivity', 'specificity', FPF, FNF, TPF and TNF, as well as understanding the terms 'SIRS' and 'sepsis', you can [click here to skip past the basics](#), but we wouldn't advise it! Once we've introduced ROCs, we'll [play a bit](#), and then look at two examples - [procalcitonin and sepsis](#), and also [tuberculosis and pleural fluid adenosine deaminase](#). Finally, in a [footnote](#), we examine accuracy, and positive and negative predictive values - such discussion will become important when we find out about costing, and [how to set a test threshold](#).

Consider patients in intensive care (ICU). One of the major causes of death in such patients is "sepsis". Wouldn't it be nice if we had a quick, easy test that defined early on whether our patients were "septic" or not? Ignoring for the moment what sepsis *is*, let's consider such a test. We imagine that we take a population of ICU patients, and do two things:



top



home



up

roc>



help

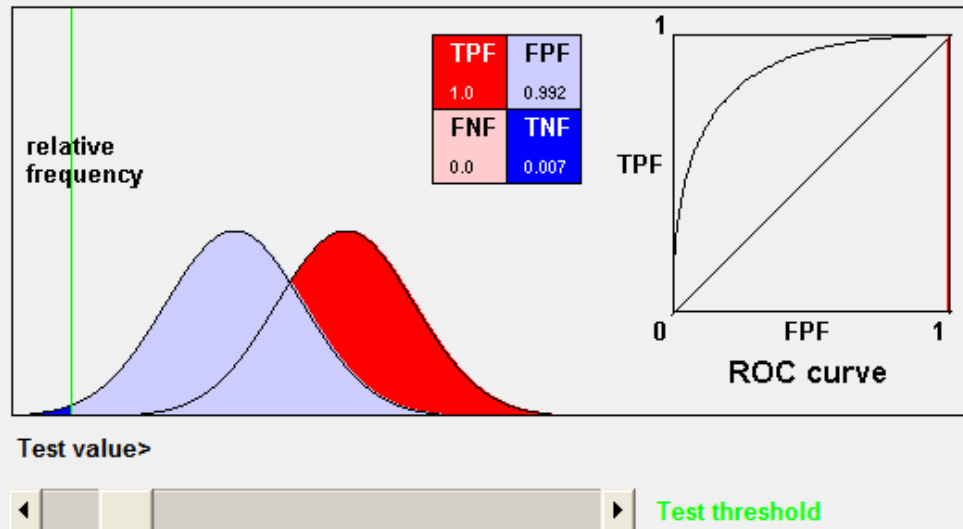


caution



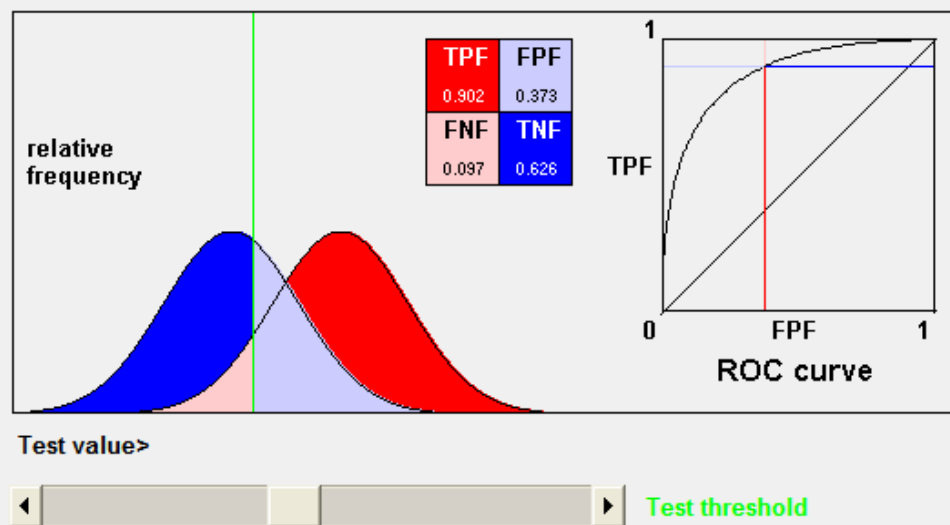
end

ROC CURVE DEMONSTRATION



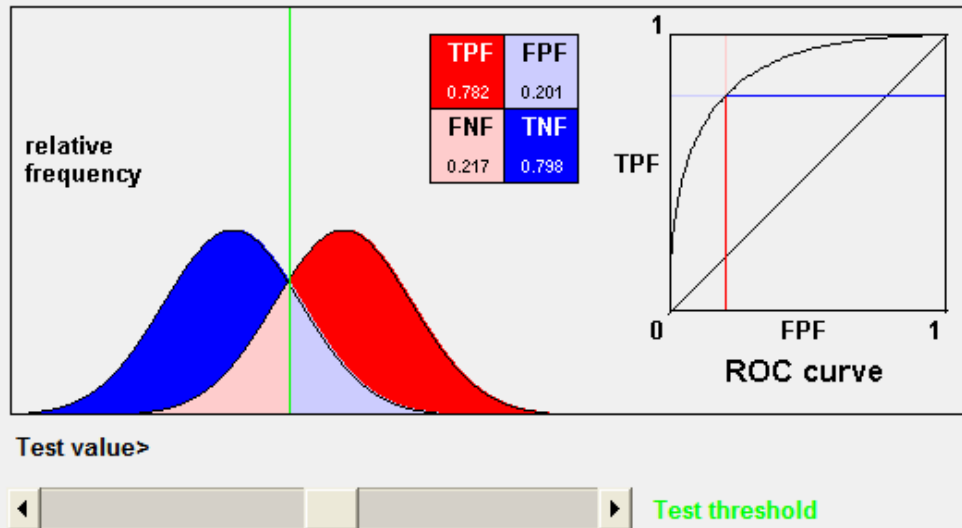
Cut-off is set very low
(i.e. too sensitive)

ROC CURVE DEMONSTRATION



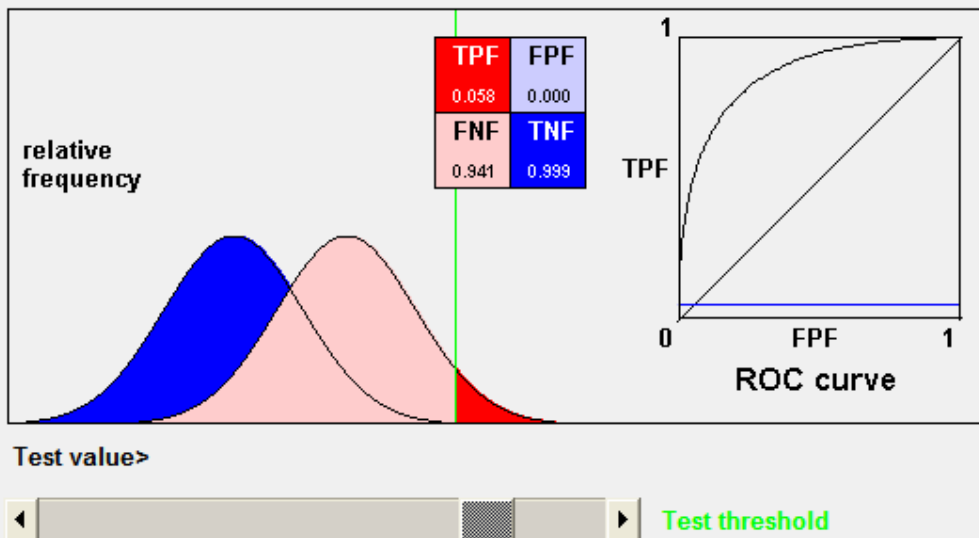
Cut-off is set low
(i.e. sensitive)

ROC CURVE DEMONSTRATION



Cut-off is set where TPR and FRP are the same

ROC CURVE DEMONSTRATION



Cut-off is set very high (i.e. too specific)

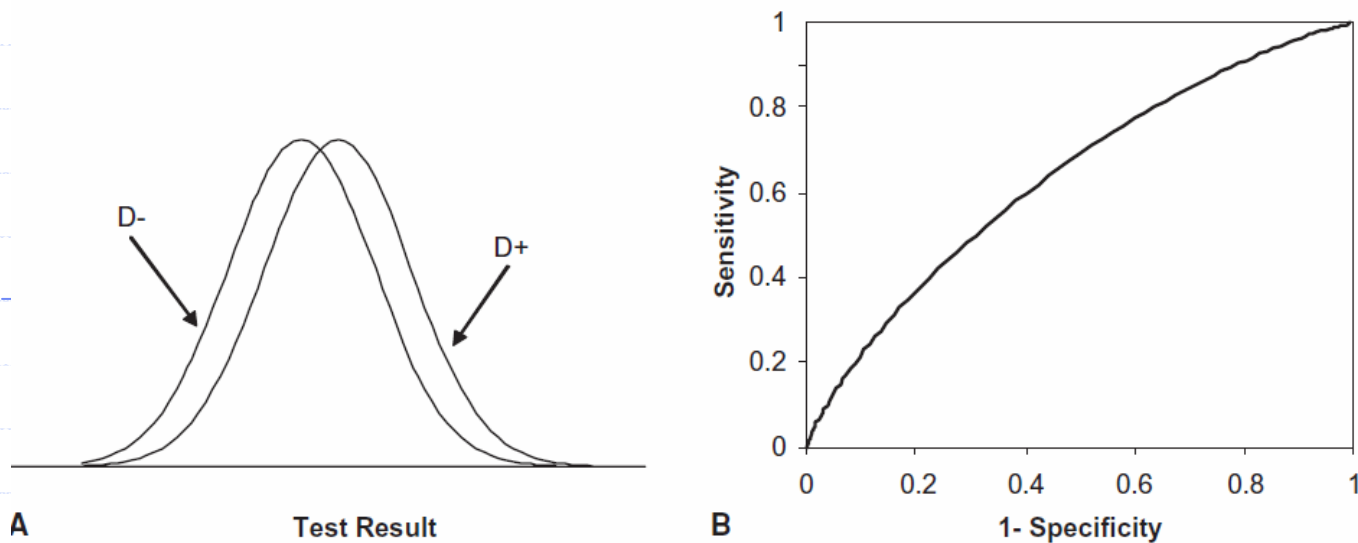


Figure 4.2 Test discriminates poorly between patients with disease (D+) and patient without disease (D-). (A) The distribution of test results in D+ patients is very similar to the distribution in D- patients. (B) This "bad" ROC curve approaches a 45-degree diagonal line.

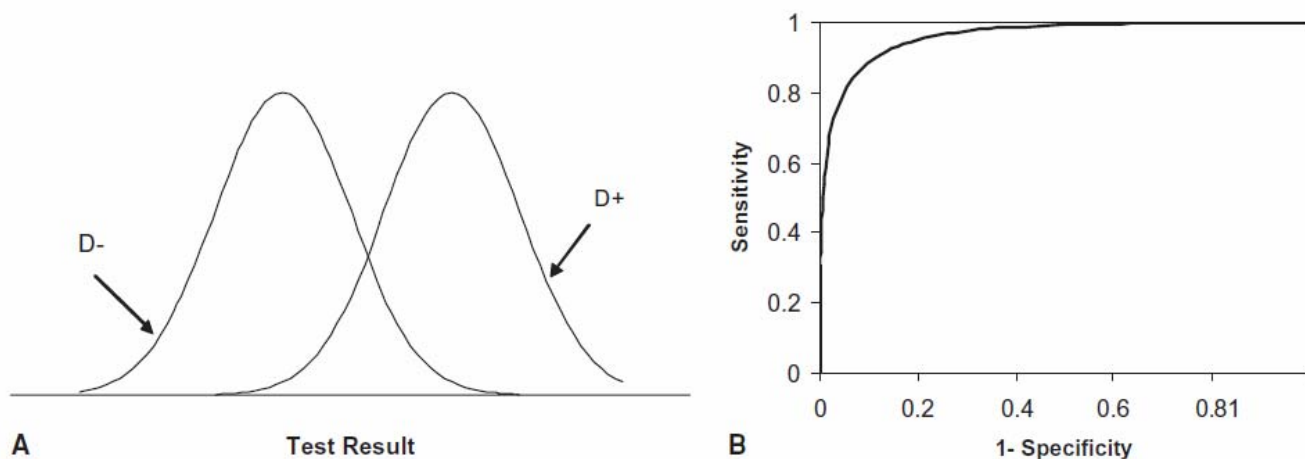


Figure 4.3 Test discriminates well between patients with the disease (D+) and patients without the disease (D-). (A) The distribution of test results in D+ patients differs substantially from the distribution in D- patients. (B) This "good" ROC curve nears the upper left corner of the grid.

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.

[Reprinted from RADIOLOGY, Vol. 143, No. 1, Pages 29-36, April, 1982.]
Copyright 1982 by the Radiological Society of North America, Incorporated

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) Curve¹

Two classic
papers on ROC

ROCs for various diagnostic systems

Fig. 3. Measured values of A for forecasts of several different weather conditions. Ranges are shown where multiple tests were made.

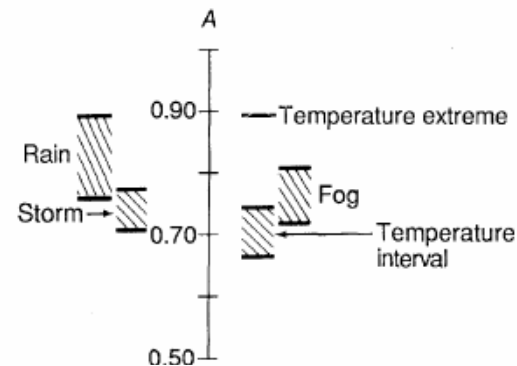


Fig. 5. Measured values of A for two aptitude tests (on the right) that were followed by schooling of all testees; a roughly adjusted range of A values for a test (on the left) that was followed by schooling only of those who achieved a criterion score on the test.

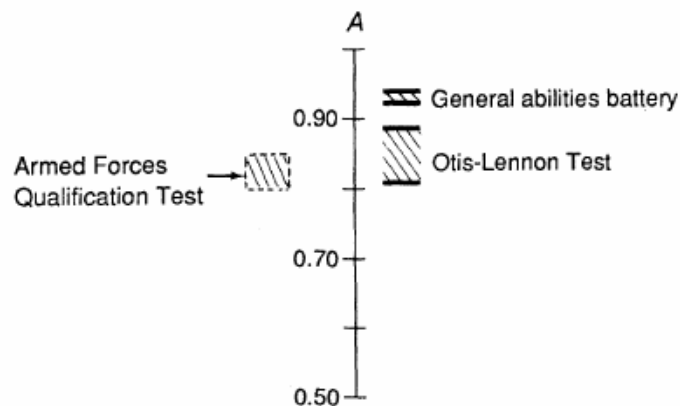
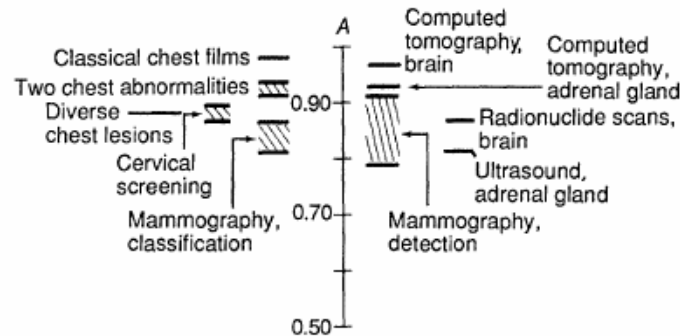


Fig. 6. Measured values of A for several imaging tests in clinical medicine.



ROCs for various diagnostic systems

Fig. 7. Measured values of A for detecting cracks in airplane wings by two techniques, from several Air Force bases.

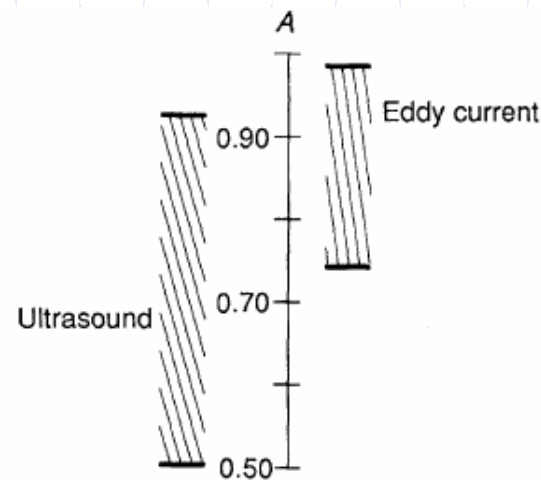
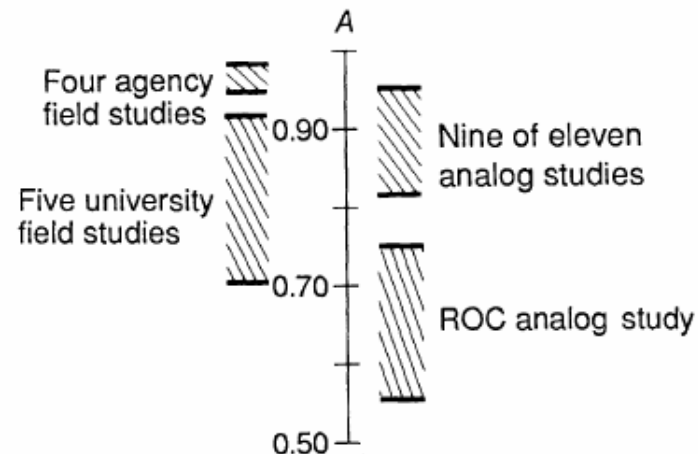


Fig. 8. Measured values of A for polygraph lie detection in several field studies (on the left) and several analog studies (on the right).





Beyond diagnostic accuracy

Are sensitivity and specificity the most meaningful measures?

Table 1. Hierarchy of Diagnostic Evaluation and the Number of Studies Available for Different Levels of Diagnostic Test in a Technology Assessment of Magnetic Resonance Spectroscopy for Brain Tumors*

Level	Description	Examples of Study Purpose or Measures	Studies Available, <i>n</i>	Patients, <i>n</i>
1	Technical feasibility and optimization	Ability to produce consistent spectra	85	2434
2	Diagnostic accuracy	Sensitivity and specificity	8	461
3	Diagnostic thinking impact	Percentage of times clinicians' subjective assessment of diagnostic probabilities changed after the test	2	32
4	Therapeutic choice impact	Percentage of times therapy planned before MRS changed after the test	2	105
5	Patient outcome impact	Percentage of patients who improved with MRS diagnosis compared with those without MRS (e.g., survival, quality of life)	0	0
6	Societal impact	Cost-effectiveness analysis (e.g., use to detect tumor in asymptomatic population)	0	0

* MRS = magnetic resonance spectroscopy.

RATING QUALITY OF EVIDENCE AND STRENGTH OF RECOMMENDATIONS

GRADE: grading quality of evidence and strength of recommendations for diagnostic tests and strategies

The GRADE system can be used to grade the quality of evidence and strength of recommendations for diagnostic tests or strategies. This article explains how patient-important outcomes are taken into account in this process

SUMMARY POINTS

As for other interventions, the GRADE approach to grading the quality of evidence and strength of recommendations for diagnostic tests or strategies provides a comprehensive and transparent approach for developing recommendations

Cross sectional or cohort studies can provide high quality evidence of test accuracy

However, test accuracy is a surrogate for patient-important outcomes, so such studies often provide low quality evidence for recommendations about diagnostic tests, even when the studies do not have serious limitations

Inferring from data on accuracy that a diagnostic test or strategy improves patient-important outcomes will require the availability of effective treatment, reduction of test related adverse effects or anxiety, or improvement of patients' wellbeing from prognostic information

Judgments are thus needed to assess the directness of test results in relation to consequences of diagnostic recommendations that are important to patients

Redundancy of Single Diagnostic Test Evaluation

Karel G.M. Moons,^{1,2,3} Gerri-Anne van Es,⁴ Bowine C. Michel,⁵ Harry R. Büller,⁶
J. Dik F. Habbema,³ and Diederick E. Grobbee¹

Moons et al. Epidemiology 1999

Diagnostic research

Diagnostic studies as multivariable,
prediction research

K G M Moons, D E Grobbee

Patient outcomes in diagnostic research

Moons et al. JECH 2002

Opinion

Test Research versus Diagnostic Research

Moons et al. Clin Chem 2004



McGill Summer Session in Epidemiology and Biostatistics 2008

The Summer Session in Epidemiology and Biostatistics at McGill offers health professionals the opportunity to gain familiarity with the principles of epidemiology and biostatistics. It also offers graduate students from McGill and other universities the opportunity to acquire academic credits and thereby accelerate course work during a summer term. **Summer session website:** <http://www.mcgill.ca/epi-biostat-occh/summer/>

Advanced Diagnostic Research

A special course jointly sponsored by Epidemiology & Biostatistics, and the CIHR Strategic Training Centre in Infectious Diseases and Autoimmunity, McGill Centre for the Study of Host Resistance

Academic credits: 2

Dates: May 6 - 9, 2008

Class times: 9 - 4:30 PM, Tuesday through Friday

Course instructor: Professor Karel Moons, MD, PhD (University Medical Center, Utrecht, The Netherlands)

Course coordinator: Dr. Madhukar Pai, MD, PhD (madhukar.pai@mcgill.ca)

Enrollment limit: 20



Description: Diagnostic research is often focused on estimating the sensitivity and specificity of diagnostic tests. This course will demonstrate that this so called 'test research' is not necessarily the same as diagnostic research. Furthermore, we will widen the horizon by proposing methods of diagnostic study design and of data analysis in which the patient's test result can be considered in the context of his or her set of individual characteristics and prior test results. These methods enable both direct estimation of individual probabilities of disease presence based on all diagnostic information and the evaluation of the extent to which a test can aid in the clinical setting. The course will include hands-on computer labs.

Course content: Principles of diagnostic research, design of diagnostic studies, data-analysis in diagnostic research and development of risk scores, and meta analyses of diagnostic studies.

Prerequisites: This is an advanced course, and prior coursework in intermediate epidemiology and biostatistics is required (specifically, knowledge of multivariable logistic regression). Students without prior coursework in multivariable methods will not be permitted to register.

Course materials: All participants will receive a course-pack with articles, readings, labs, etc.

Instructor: Karel G.M. Moons is Professor of Clinical Epidemiology at the Julius Center for Health Sciences and Primary Care at Utrecht, Netherlands. His main focus concerns the methodology of diagnostic research. His major expertise is testing existing and introducing innovative designs and analytical methods for the evaluation of diagnostic tests, and the development, validation and implementation of diagnostic and prognostic prediction rules. He teaches courses on advanced diagnostic research throughout the world. He has over 130 publications and has obtained numerous grants and awards in the field.

Note: The language of instruction is English, and students are advised that fluency in English is essential to benefit from the course. However, students may submit their course assignments and examinations in French. Courses may be taken for Academic Credit, Continuing Medical Education (CME) Credit, or for a Professional Interest Certificate.

Multivariable approach

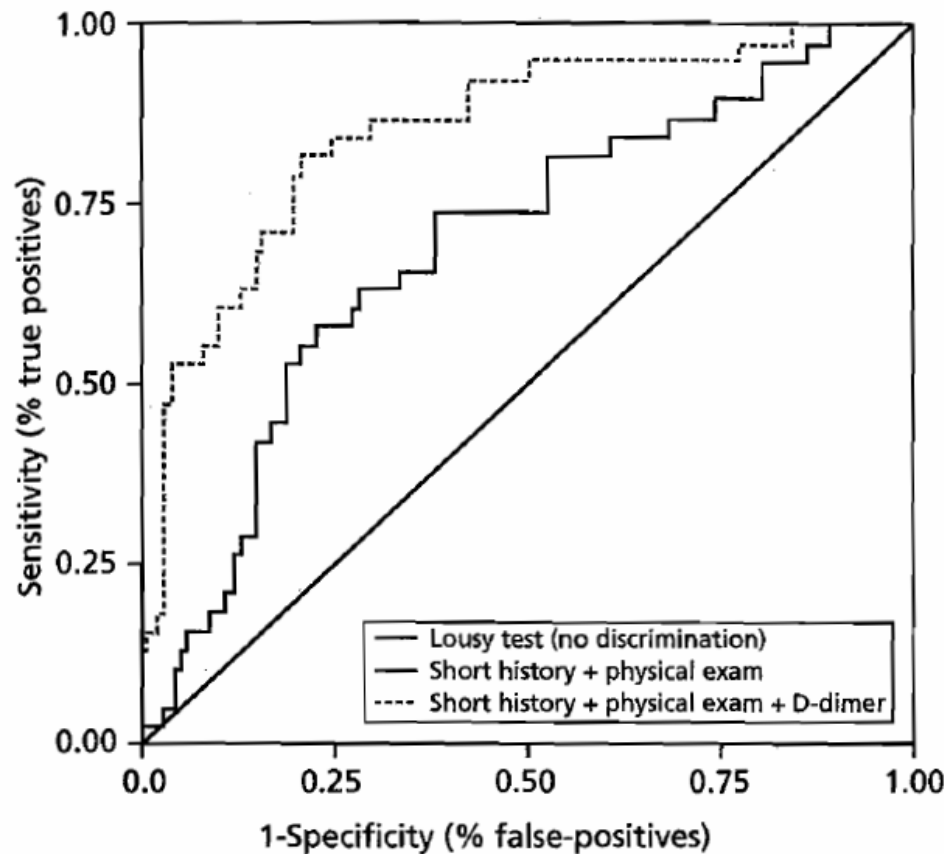
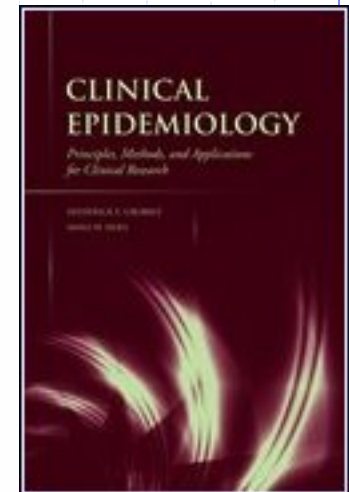


FIGURE 3.3 Example of an ROC curve of the reduced multivariable logistic regression model, including the same six determinants as in Figure 3.2. The ROC area of the "reduced history + physical model" (red) was 0.70 (95% confidence interval [CI], 0.66–0.74) and of the same model added with the D-dimer assay (green) 0.84 (95% CI, 0.80–0.88).

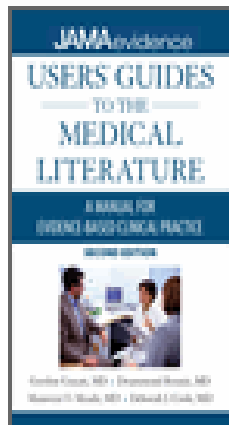
Key outcome here is what is the added value of a new test, beyond all the prior tests that may have been done (including history/physical)

The multivariable approach mimics the real life diagnostic process

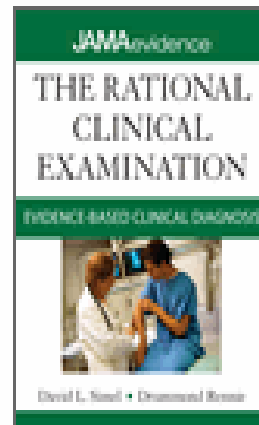
1. A diagnosis starts with a patient presenting a complaint (symptom and/or sign) suggestive of a certain disease to be diagnosed.
2. The subsequent work-up is a multivariable process. It involves multiple diagnostic determinants (tests) that are applied in a logical order: from age, gender, medical history, and signs and symptoms, to more complicated, invasive, and costly tests.
3. Setting or ruling out a diagnosis is a probabilistic action in which the probability of the presence or absence of the disease is central. This probability is continuously updated based on subsequent diagnostic test results.
4. The true diagnostic value of a test is determined by the extent to which it provides diagnostic information beyond earlier tests, that is, materially changes the probability estimation of disease presence based on previous test results.
5. The goal of the diagnostic process is to eventually rule in or out the disease with enough confidence to take clinical decisions. This requires precise estimates of the probability of the presence of the target disease(s).



Relevant books



Users' Guides to the Medical Literature
A Manual for Evidence-Based Clinical Practice, 2nd Edition



The Rational Clinical Examination
Evidence-Based Clinical Diagnosis
Includes online-only content

