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 (Clinical Medicine)
- Etio-gnosis
- Pro-gnosis

- Dia-gnosis
- Etio-gnosis
- Pro-gnosis (Public and community health)

Miettinen OS
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

<table>
<thead>
<tr>
<th>Pattern recognition</th>
<th>Probabilistic diagnostic reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>See it and recognize disorder</td>
<td>Clinical assessment generates pretest probability</td>
</tr>
<tr>
<td>▼</td>
<td>▼</td>
</tr>
<tr>
<td>Compare posttest probability with thresholds</td>
<td>New information generates posttest probability</td>
</tr>
<tr>
<td>(usually pattern recognition implies probability near 100% and so above threshold)</td>
<td>(may be interactive)</td>
</tr>
<tr>
<td></td>
<td>▼</td>
</tr>
<tr>
<td></td>
<td>Compare posttest probability with thresholds</td>
</tr>
</tbody>
</table>


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?

Above this point, treat

Below this point, no further testing

Disease ruled IN

Disease not ruled in or out

Disease ruled OUT
**FIGURE 3.1** Diagnostic Testing.
The Perfect Diagnostic Test

X

No Disease

Y

Diseased
Variations In Diagnostic Tests

Overlap

Range of Variation in Disease free

Range of Variation in Diseased
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!

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

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.

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 anemia 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

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. Geetadevi¹, ¹, Rajnish Joshi², ¹, Madhukar Pai³, ², S.P. Kalantri¹, *

¹ Department of Medicine, Mahatma Gandhi Institute of Medical Sciences, Sevagram, Wardha 442102, India

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

Shriprakash Kalantri, Rajnish Joshi, Trunal Lokhande, Amandeep Singh, Maureen Morgan, John M. Colford Jr, Madhukar Pai

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

Nitika Pant Pai, Rajnish Joshi, Sandeep Dogra, Bharati Taksand, S. P. Kalantri, Madhukar Pai, Pratibha Narang, Jacqueline P. Tulsky, Arthur L. Reingold

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

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

Nitika Pant Pai1, 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 preference. 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.

For author affiliations, see end of text.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Diagnosis</th>
<th>Risk Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Approach</strong></td>
<td>Patients are given a diagnosis: Either they have the disease or they do not</td>
<td>Patients are given a probability of a future event</td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td>Syphilitic hepatitis</td>
<td>Cardiovascular event within 10 years</td>
</tr>
<tr>
<td><strong>Lesion</strong></td>
<td>Unambiguous</td>
<td>Nonexistent or equivocal</td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td>Torn aorta</td>
<td>Depression</td>
</tr>
<tr>
<td><strong>Treatment effectiveness</strong></td>
<td>Often highly effective</td>
<td>Helpful, but patients may have event with treatment or avoid the event even if untreated</td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td>Antibiotics for syphilis</td>
<td>Statins for high cholesterol level</td>
</tr>
<tr>
<td><strong>Course of treatment</strong></td>
<td>Dictated by diagnosis</td>
<td>Open to discussion</td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td>Surgical treatment of a torn aorta</td>
<td>Treatment of early-stage prostate cancer</td>
</tr>
<tr>
<td><strong>Patient preference</strong></td>
<td>Generally of minor importance</td>
<td>Often of major importance</td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td>Antibiotics for syphilis</td>
<td>Treatment of early-stage prostate cancer</td>
</tr>
<tr>
<td><strong>Symptoms</strong></td>
<td>Patient has distressing symptoms</td>
<td>Patient is often asymptomatic; Disorder is a risk factor for a future event</td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td>Syphilitic hepatitis</td>
<td>Hyperlipidemia</td>
</tr>
</tbody>
</table>
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 shall take assessment of the value of the plasma concentration of B-type natriuretic peptide (BNP) in the diagnosis of left ventricular dysfunction. 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 combinations of heart failure. These results are shown in Table 1.

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 untreated 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.
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?

<table>
<thead>
<tr>
<th>Patients known to have disorder</th>
<th>Normal controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median (range) concentration of BNP precursor (pg/ml)</td>
<td>493.5 (248.9-909.0)</td>
</tr>
</tbody>
</table>

*BMJ* 2002;324:539–41
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?

<table>
<thead>
<tr>
<th>Patients known to have target disorder</th>
<th>Normal controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>High BNP concentration</td>
<td>39</td>
</tr>
<tr>
<td>Normal BNP concentration</td>
<td>1</td>
</tr>
</tbody>
</table>

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>
<thead>
<tr>
<th>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?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients with LVD on echocardiography</td>
</tr>
<tr>
<td>Concentration of BNP:</td>
</tr>
<tr>
<td>High (&gt;17.9 pg/ml)</td>
</tr>
<tr>
<td>Normal (&lt;18 pg/ml)</td>
</tr>
<tr>
<td>Prevalence (pretest probability) of LVD</td>
</tr>
<tr>
<td>Test characteristics (95% CI):</td>
</tr>
<tr>
<td>Sensitivity = 88% (74% to 94%)</td>
</tr>
<tr>
<td>Specificity = 34% (25% to 44%)</td>
</tr>
<tr>
<td>Positive predictive value = 38% (29% to 48%)</td>
</tr>
<tr>
<td>Negative predictive value = 85% (70% to 94%)</td>
</tr>
<tr>
<td>Likelihood ratio for an abnormal test result = 1.3 (1.1 to 1.6)</td>
</tr>
<tr>
<td>Likelihood ratio for a normal test result = 0.4 (0.2 to 0.9)</td>
</tr>
</tbody>
</table>

*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

Area under the receiver-operating-characteristic curve, 0.91 (95% confidence interval, 0.90–0.93)
Evaluation of Diagnostic Accuracy, Feasibility and Client Preference for Rapid Oral Fluid-Based Diagnosis of HIV Infection in Rural India

Nitika Pant Pai1, Rajnish Joshi2, Sandeep Dogra3, Bharati Taksande2, S. P. Kalantri2, Madhukar Pai4, Pratibha Narang2, Jacqueline P. Tulsky5, 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 HIV-1/2 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.

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

Nitika Pant Pai1, Ritu Barick2, Jacqueline P. Tulsky3, Poonam V. Shivkumar4, Deborah Cohan3, Shriprakash Kalantri2, Madhukar Pai1, 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 Biostatistic, 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.
Evidence based diagnostics
Christian Glund, Lise Løtze Glund

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.

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. 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. 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.

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.

Diagnostic accuracy
In phase II, studies explore the diagnostic accuracy of a test in participants with both known and suspected relevant disease. Phase IIIa 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 though 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
Randomised comparisons of medical tests: sometimes invalid, not always efficient

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

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

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. 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:

<table>
<thead>
<tr>
<th></th>
<th>Disease +</th>
<th>Disease -</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test +</strong></td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td><strong>Test -</strong></td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

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

### [true positive rate]

<table>
<thead>
<tr>
<th></th>
<th>Disease present</th>
<th>Disease absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test positive</td>
<td>True positives</td>
<td>False positives</td>
</tr>
<tr>
<td>Test negative</td>
<td>False negative</td>
<td>True negatives</td>
</tr>
</tbody>
</table>

The proportion of patients with disease who test positive = \( P(T+|D+) = \frac{TP}{TP+FN} \)
Specificity
[true negative rate]

<table>
<thead>
<tr>
<th></th>
<th>Disease present</th>
<th>Disease absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test positive</td>
<td>True positives</td>
<td>False positives</td>
</tr>
<tr>
<td>Test negative</td>
<td>False negative</td>
<td>True negatives</td>
</tr>
</tbody>
</table>

The proportion of patients without disease who test negative: \( P(T^{-|D^{-}}) = \frac{TN}{TN + FP} \).
### Predictive value of a positive test

<table>
<thead>
<tr>
<th></th>
<th>Disease present</th>
<th>Disease absent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test positive</strong></td>
<td>True positives</td>
<td>False positives</td>
</tr>
<tr>
<td><strong>Test negative</strong></td>
<td>False negative</td>
<td>True negatives</td>
</tr>
</tbody>
</table>

Proportion of patients with positive tests who have disease:  
\[
P(D+|T+) = \frac{TP}{TP + FP}
\]
### Predictive value of a negative test

<table>
<thead>
<tr>
<th></th>
<th>Disease present</th>
<th>Disease absent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test positive</strong></td>
<td>True positives</td>
<td>False positives</td>
</tr>
<tr>
<td><strong>Test negative</strong></td>
<td>False negative</td>
<td>True negatives</td>
</tr>
</tbody>
</table>

Proportion of patients with negative tests who do not have disease = \( P(D^{-}\mid T-) = \frac{TN}{(TN+FN)} \)
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.

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

<table>
<thead>
<tr>
<th>Culture (gold standard)</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serological Test</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>14</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>Negative</td>
<td>54</td>
<td>28</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>68</td>
<td>31</td>
<td>99</td>
</tr>
</tbody>
</table>

Sensitivity = 21%
Specificity = 90%

*Clin Vacc Immunol 2006;13: 702-03*
For a given test, predictive values will depend on prevalence

Test with 80% sensitivity and 90% specificity:

<table>
<thead>
<tr>
<th>pre-test probability (disease prevalence)</th>
<th>1%</th>
<th>10%</th>
<th>50%</th>
<th>90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPV</td>
<td>7.5%</td>
<td>47.1%</td>
<td>88.9%</td>
<td>98.6%</td>
</tr>
<tr>
<td>NPV</td>
<td>99.8%</td>
<td>97.6%</td>
<td>81.8%</td>
<td>33.3%</td>
</tr>
</tbody>
</table>

[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*  

<table>
<thead>
<tr>
<th>Setting</th>
<th>Prevalence (Cases/100,000)</th>
<th>Positive Predictive Value (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>General population</td>
<td>35</td>
<td>0.4</td>
</tr>
<tr>
<td>Men, age 75 or greater</td>
<td>500</td>
<td>5.6</td>
</tr>
<tr>
<td>Clinically suspicious prostatic nodule</td>
<td>50,000</td>
<td>93.0</td>
</tr>
</tbody>
</table>

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?

• $LR^+ = \frac{TPR}{FPR}$

• 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

\[ LR^+ = \frac{\Pr(T^+ \mid D^+)}{\Pr(T^+ \mid D^-)} \]

- **True positives** (Test positive and Disease present)
- **False positives** (Test positive and Disease absent)
- **False negatives** (Test negative and Disease present)
- **True negatives** (Test negative and Disease absent)

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

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

\[ LR^- = \frac{\Pr(T- \mid D+)}{\Pr(T- \mid D-)} \]
LR: Impact on Likelihood of Disease

LR = 0.01
Less Likely

LR = 0.1
Less Likely

LR = 0.2
Less Likely

LR = 0.3
Less Likely

LR = 1
No Impact on Likelihood of Disease

LR = 3
More Likely

LR = 5
More Likely

LR = 10
More Likely

LR = 100
More Likely

Increasing impact
LR: Impact on Likelihood of Disease

LR: Impact on Likelihood of Disease

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact

More Impact
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)

<table>
<thead>
<tr>
<th>Test result</th>
<th>Disease present</th>
<th>Disease absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test positive</td>
<td>True positives (a)</td>
<td>False positives (b)</td>
</tr>
<tr>
<td>Test negative</td>
<td>False negative (c)</td>
<td>True negatives (d)</td>
</tr>
</tbody>
</table>

\[
\text{DOR} = \frac{(a/c)}{(b/d)}
\]

\[
\text{DOR} = \frac{ad}{bc}
\]

\[
\text{DOR} = \frac{\text{Odds of } T^+|D^+}{\text{Odds of } T^+|D^-}
\]

Odds of positive test result in persons with the target condition compared to those without the target condition.
Example: Serological test for TB

<table>
<thead>
<tr>
<th>Culture (gold standard)</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>Negative</td>
<td>54</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>68</td>
<td>31</td>
</tr>
</tbody>
</table>

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 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?

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

Pre-Test Probability

Mr. A
Pre-Test Prob. 50%

Serological test
LR+ = 2

Post-Test Probability

Post-Test Prob. 70%
Likelihood Ratios

Pre-Test Probability

Mr. A
Pre-Test Prob. 50%

Post-Test Probability

Post-Test Prob. 45%

Serological test
LR- = 0.9
Using LRs 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

Ms. B
Pre-Test Prob. 10%

Post-Test Prob. 20%

Serological test
LR+ = 2
Ms. B
Pre-Test Prob. 10%

Serological test
LR- = 0.9
Example: Ultrasonography for Down Syndrome
Another example: Nuchal fold & Down Syndrome

<table>
<thead>
<tr>
<th></th>
<th>Down Syndrome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>21</td>
</tr>
<tr>
<td>No</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nuchal fold</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>21</td>
<td>7</td>
</tr>
<tr>
<td>No</td>
<td>4</td>
<td>188</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>4</td>
<td>25</td>
</tr>
<tr>
<td>7</td>
<td>188</td>
<td>195</td>
</tr>
</tbody>
</table>

Sensitivity = 75%
Specificity = 98%
LR+ = 36
LR- = 0.26
DOR = 141

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

Mrs. A
Pre-Test Prob. 10%

Post-Test Prob. 80%

Nuchal fold abnormal
LR = 36
Likelihood Ratios

Pre-Test Probability

Mrs. A Pre-Test Prob. 10%

Nuchal fold normal
LR = 0.26

Post-Test Probability

Post-Test Prob. 3%
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

Mrs. B
Pre-Test Prob. 0.5%

Nuchal fold abnormal
LR = 36

Post-Test Probability

Post-Test Prob. 10%
Likelihood Ratios

Pre-Test Probability

Mrs. B
Pre-Test Prob. 0.5%

Post-Test Probability

Nuchal fold normal
LR = 0.26

Post-Test Prob. 0.1%
Where do we get LRs from?

### Quick Reference

<table>
<thead>
<tr>
<th>Chapter 1: Primer on Precision and Accuracy</th>
<th>Test/Finding</th>
<th>LR+</th>
<th>LR-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 2: Abdominal Aortic Aneurysm</td>
<td>Occur in 4% to 8% of older men. The prevalence in older women is less than 2%.</td>
<td>Physical examination for aneurysm &gt; 4.0 cm</td>
<td>16 (8.6–29)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Physical examination for aneurysm &gt; 3.0 cm</td>
<td>12 (7.4–20)</td>
</tr>
<tr>
<td>Chapter 3:</td>
<td>Approximately 1% to 5% of the general population</td>
<td>Systolic–diastolic bruit</td>
<td>39 (10–145)</td>
</tr>
</tbody>
</table>
## Examples

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

<table>
<thead>
<tr>
<th>Prior Probability</th>
<th>Test/Finding</th>
<th>LR+</th>
<th>LR−</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chapter 51: Urinary Tract Infection, Women</strong></td>
<td>Dysuria</td>
<td>1.5 (1.2–2.0)</td>
<td>0.5 (0.3–0.7)</td>
</tr>
<tr>
<td>48% among women with compatible symptoms</td>
<td>Frequency</td>
<td>1.8 (1.1–3.0)</td>
<td>0.5 (0.4–1.0)</td>
</tr>
<tr>
<td></td>
<td>Vaginal discharge</td>
<td>0.3 (0.1–0.9)</td>
<td>3.1 (1.0–9.3)</td>
</tr>
<tr>
<td></td>
<td>Vaginal irritation</td>
<td>0.2 (0.1–0.9)</td>
<td>2.7 (0.9–8.5)</td>
</tr>
<tr>
<td></td>
<td>Dipstick result</td>
<td>4.2</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Epidemiology 3

Refining clinical diagnosis with likelihood ratios

David A Grimes, Kenneth F Schulz

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

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

Area Under Curve (AUC) = 0.86
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)**

<table>
<thead>
<tr>
<th>WBC Count (×1,000/μL)</th>
<th>Bacteremia</th>
<th>No bacteremia</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>30–35</td>
<td>11.8%</td>
<td>0.8%</td>
<td>15.2</td>
</tr>
<tr>
<td>25–30</td>
<td>9.4%</td>
<td>1.8%</td>
<td>5.3</td>
</tr>
<tr>
<td>20–25</td>
<td>26.8%</td>
<td>5.4%</td>
<td>4.9</td>
</tr>
<tr>
<td>15–20</td>
<td>37.8%</td>
<td>15.5%</td>
<td>2.4</td>
</tr>
<tr>
<td>10–15</td>
<td>11.8%</td>
<td>32.1%</td>
<td>0.37</td>
</tr>
<tr>
<td>5–10</td>
<td>2.4%</td>
<td>38.1%</td>
<td>0.06</td>
</tr>
<tr>
<td>0–5</td>
<td>0.0%</td>
<td>6.3%</td>
<td>0.00</td>
</tr>
</tbody>
</table>
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 1950s 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:
Cut-off is set very low (i.e. too sensitive)

Cut-off is set low (i.e. sensitive)
Cut-off is set where TPR and FRP are the same.

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
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.

[Reprinted from RADIOL OGY, Vol. 143, No. 1, Pages 29-38, April, 1982.]
Copyright 1982 by the Radiological Society of North America, Incorporated

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

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

Swets JA. Science 1988
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

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

* 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

Moons et al. JECH 2002

Opinion

Test Research versus Diagnostic Research

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 coursework during a summer term. Summer session website: http://www.mcgill.ca/epi/biostat/csc/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
Date: May 6 - 9, 2008
Class times: 8:30 AM - 3:30 PM
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 150 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.
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).

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).
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*

- **Clinical Epidemiology**
  How to Do Clinical Practice Research

- **Evidence-Based Diagnosis**

- **How Doctors Think**
  Jerome Groopman, M.D.
  *New York Times Bestseller*
  "Must reading for every physician who cares for patients and every patient who wishes to get the best care."

- **The Evidence Base of Clinical Diagnosis**
  Edited by J. André Kortmann