Evaluation of Diagnostic Tests

July 21, 2014
Introduction to Clinical Research:
A Two-week Intensive Course

David W. Dowdy, MD, PhD
Department of Epidemiology
Johns Hopkins
Bloomberg School of Public Health

1

Learning objectives

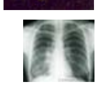
- **Part I: Recap basic epidemiological tools for evaluating diagnostics**
  - Accuracy
    - *Sensitivity & Specificity*
    - *Positive & Negative Predictive Value*
    - *Receiver Operating Curve (ROC) Analysis*
    - *Bayesian Approaches (Likelihood Ratio)*
  - Precision
    - *Intra-Class Correlation*
    - *Kappa Statistic*
- **Part II: Discuss challenges in evaluation of diagnostic tools**
  - Recognize differences between diagnostics and therapeutics
  - Understand challenges in evaluation of diagnostic tests

2

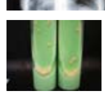
Motivating Example:
Diagnostic Tests for Tuberculosis (TB)

- Sputum Smear Microscopy
  - Simple, fast, detects the most infectious
  - Misses at least 30-40% of cases


- Chest X-ray
  - Almost always abnormal in TB
  - Abnormal CXR can be many things

- TB Culture
  - Closest we have to a “gold standard”
  - Takes weeks, high contamination rate




- PCR: Xpert MTB/RIF
  - Detects more cases than smear, less than culture
  - Minimal infrastructure, but expensive
  - FDA application pending…

GeneXpert: A History

Journal of Clinical Microbiology, July 2010, p. 2495–2501 Vol. 48, No. 7
0095-1137/10/$12.00 doi:10.1128/JCM.00128-10

Evaluation of the Analytical Performance of the Xpert MTB/RIF Assay▽†

Robert Blakemore,[1] Elizabeth Story,[1] Danica Helb,[1]‡ JoAnn Kop,[2] Padmapriya Banada,[1] Michelle R. Owens,[2] Soumitesh Chakravorty,[1] Martin Jones,[2] and David Alland[1]*

culosis and rifampin (RIF) resistance in under 2 h. The sensitivity of the assay was tested with 79 phylogenetically and geographically diverse *M. tuberculosis* isolates, including 42 drug-susceptible isolates and 37 RIF-resistant isolates containing 13 different *rpoB* mutations or mutation combinations. The specificity of the assay was tested with 89 nontuberculosis bacteria, fungi, and viruses. The Xpert MTB/RIF assay correctly identified all 79 *M. tuberculosis* isolates and correctly excluded all 89 nontuberculosis isolates. RIF resistance was correctly identified in all 37 resistant isolates and in none of the 42 susceptible isolates. Dynamic range was assessed by adding 10^2 to 10^7 CFU of *M. tuberculosis* into *M. tuberculosis*-negative sputum samples. The assay showed a log-linear relationship between cycle threshold and input CFU over the entire concentration

4

The NEW ENGLAND JOURNAL of MEDICINE
ESTABLISHED IN 1812 SEPTEMBER 9, 2010 VOL. 363 NO. 11

Rapid Molecular Detection of Tuberculosis and Rifampin Resistance

Catharina C. Boehme, M.D., Pamela Nabeta, M.D., Doris Hillemann, Ph.D., Mark P. Nicol, Ph.D., Shubhada Shenai, Ph.D., Fiorella Krapp, M.D., Jenny Allen, B.Tech., Rasim Tahirli, M.D., Robert Blakemore, B.S., Roxana Rustomjee, M.D., Ph.D., Ana Milovic, M.S., Martin Jones, Ph.D., Sean M. O'Brien, Ph.D., David H. Persing, M.D., Ph.D., Sabine Ruesch-Gerdes, M.D., Eduardo Gotuzzo, M.D., Camilla Rodrigues, M.D., David Alland, M.D., and Mark D. Perkins, M.D.

RESULTS

Among culture-positive patients, a single, direct MTB/RIF test identified 551 of 561 patients with smear-positive tuberculosis (98.2%) and 124 of 171 with smear-negative tuberculosis (72.5%). The test was specific in 604 of 609 patients without tuberculosis (99.2%). Among patients with smear-negative, culture-positive tuberculosis, the addition of a second MTB/RIF test increased sensitivity by 12.6 percentage points and a third by 5.1 percentage points, to a total of 90.2%. As compared with phenotypic drug-susceptibility testing, MTB/RIF testing correctly identified 200 of 205 patients (97.6%) with rifampin-resistant bacteria and 504 of 514 (98.1%) with rifampin-sensitive bacteria. Sequencing resolved all but two cases in favor of the MTB/RIF assay.

5

Feasibility, diagnostic accuracy, and effectiveness of decentralised use of the Xpert MTB/RIF test for diagnosis of tuberculosis and multidrug resistance: a multicentre implementation study

Catharina C Boehme, Mark P Nicol, Pamela Nabeta, Joy S Michael, Eduardo Gotuzzo, Rasim Tahirli, Ma Tarcela Gler, Robert Blakemore, William Worodria, Christen Gray, Laurence Huang, Tatiana Caceres, Rafail Mehdiyev, Lawrence Raymond, Andrew Whitelaw, Kalaiselvan Sagadevan, Heather Alexander, Heidi Albert, Frank Cobelens, Helen Cox, David Alland, Mark D Perkins

Summary

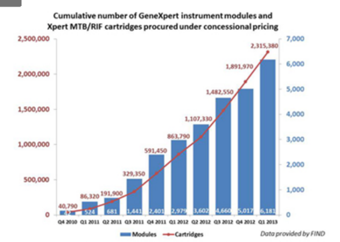
Background The Xpert MTB/RIF test (Cepheid, Sunnyvale, CA, USA) can detect tuberculosis and its multidrug-

Methods We assessed adults (≥18 years) with suspected tuberculosis or multidrug-resistant tuberculosis consecutively presenting with cough lasting at least 2 weeks to urban health centres in South Africa, Peru, and India, drug-resistance screening facilities in Azerbaijan and the Philippines, and an emergency room in Uganda. Patients were excluded from the main analyses if their second sputum sample was collected more than 1 week after the first sample, or if no valid reference standard or MTB/RIF test was available. We compared one-off direct MTB/RIF testing in nine microscopy laboratories adjacent to study sites with 2–3 sputum smears and 1–3 cultures, dependent on site, and drug-susceptibility testing. We assessed indicators of robustness including indeterminate rate and between-site performance, and compared time to detection, reporting, and treatment, and patient dropouts for the techniques used.

Findings We enrolled 6648 participants between Aug 11, 2009, and June 26, 2010. One-off MTB/RIF testing detected 933 (90·3%) of 1033 culture-confirmed cases of tuberculosis, compared with 699 (67·1%) of 1041 for microscopy. MTB/RIF test sensitivity was 76·9% in smear-negative, culture-positive patients (296 of 385 samples), and 99·0% specific (2846 of 2876 non-tuberculosis samples). MTB/RIF test sensitivity for rifampicin resistance was 94·4% (236 of 250) and specificity was 98·3% (796 of 810). Unlike microscopy, MTB/RIF test sensitivity was not significantly lower in patients with HIV co-infection. Median time to detection of tuberculosis for the MTB/RIF test was 0 days (IQR 0–1), compared with 1 day (0–1) for microscopy, 30 days (23–43) for solid culture, and 16 days (13–21) for liquid culture.

Lancet 2011; 377: 1495–1505

6



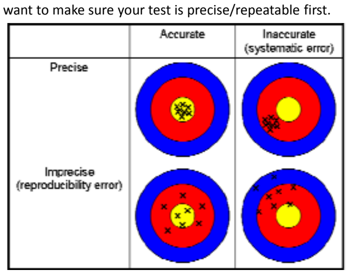
As of 31 March 2013, a total of 1,123 GeneXpert instruments (comprising 6,181 modules) and 2,315,380 Xpert MTB/RIF cartridges had been procured in the public sector in 83 of the 145 countries eligible for concessional pricing.

Part I:
Epidemiological Tools for Evaluation of Diagnostics

8

Accuracy vs. Precision

- **Accuracy:** How close diagnostic test results are to the "truth"
  - More a measure of effectiveness/appropriateness
- **Precision:** How close diagnostic test results are to each other
  - More a measure of technical specification
  - Usually want to make sure your test is precise/repeatable first.



9

Measures of Accuracy

- **Sensitivity**
  - *Proportion of people with the condition who test positive*
- **Specificity**
  - *Proportion of people without the condition who test negative*
- **Positive Predictive Value**
  - *Proportion of people testing positive who have the condition*
- **Negative Predictive Value**
  - *Proportion of people testing negative who do not have the condition*
- Sensitivity and specificity are characteristics of the test; PPV and NPV depend on the prevalence of the condition in the population tested.

10

Test Accuracy

		"Gold Standard"	
		Positive	Negative
New Test	Positive	A True Positive	B False Positive
	Negative	C False Negative	D True Negative

- **Sensitivity = A/(A+C)**
- **Specificity = D/(B+D)**
  - **PPV = A/(A+B)**
  - **NPV = D/(C+D)**

11

Test Accuracy

		TB Culture		
		Positive	Negative	
Xpert MTB/RIF	Positive	70 True Positive	10 False Positive	**80**
	Negative	30 False Negative	890 True Negative	**920**
		100	**900**	

- **Sensitivity = A/(A+C)**
- **Specificity = D/(B+D)**
  - **PPV = A/(A+B)**
  - **NPV = D/(C+D)**

12

Test Accuracy

		TB Culture		
		Positive	Negative	
Xpert MTB/RIF	Positive	70 True Positive	10 False Positive	**80**
	Negative	30 False Negative	890 True Negative	**920**
		100	**900**	

- **Sensitivity = 70/(70+30) = 70%**
- **Specificity = 890/(10+890) = 98.9%**
- **PPV = 70/(70+10) = 87.5%**
- **NPV = 890/(30+890) = 96.7%**

13

Effect of Prevalence on PPV and NPV

		"Gold Standard"	
		Positive	Negative
New Test	Positive	9	1
	Negative	1	99

- Take a test with 90% sensitivity and 99% specificity.
- Prevalence of condition here = 10/110 = 9%
  - PPV = 9/10 = 90%
  - NPV = 99/100 = 99%

14

Effect of Prevalence on PPV and NPV

		"Gold Standard"	
		Positive	Negative
New Test	Positive	90	1
	Negative	10	99

- Now increase prevalence to 50%.
  - PPV = 90/91 = 98.9%
  - NPV = 99/109 = 90.8%
- *As prevalence increases, PPV increases and NPV decreases.*

15

Sensitivity & Specificity in the Real World

The NEW ENGLAND JOURNAL of MEDICINE

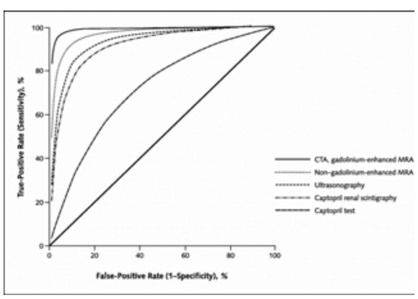
Table 2. Overall Sensitivity and Specificity of the MTB/RIF Test, According to the Number of Tests per Patient, as Compared with Three Smears and Four Cultures.

Site and No. of Tests	Sensitivity			Specificity
	All Culture-Positive	Smear-Positive and Culture-Positive	Smear-Negative and Culture-Positive	No Tuberculosis
Site				
Lima, Peru				
Correct — no./total no. (%)	209/211 (99.1)	199/199 (100)	10/12 (83.3)	102/102 (100)
95% CI	96.6–99.7	98.1–100.0	55.2–95.3	96.4–100.0
Baku, Azerbaijan				
Correct — no./total no. (%)	144/149 (96.6)	80/80 (100.0)	64/69 (92.8)	68/70 (97.1)
95% CI	92.4–98.6	95.4–100.0	84.1–96.9	90.2–99.2
Cape Town, South Africa				
Correct — no./total no. (%)	142/148 (95.9)	95/96 (99.0)	47/52 (90.4)	186/189 (98.4)
95% CI	91.4–98.1	94.3–99.8	79.4–95.8	95.4–99.5
Durban, South Africa				
Correct — no./total no. (%)	43/45 (95.6)	30/30 (100.0)	13/15 (86.7)	213/219 (97.3)
95% CI	85.2–98.8	88.6–100.0	62.1–96.3	94.2–98.7
Mumbai, India				
Correct — no./total no. (%)	185/188 (98.4)	162/162 (100.0)	23/26 (88.5)	35/36 (97.2)
95% CI	95.4–99.5	99.7–100.0	71.0–96.0	85.8–99.5

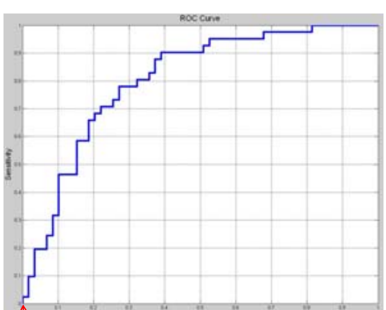
16

ROC Curves

Diagnostic Tests for Renal Artery Stenosis in Patients Suspected of Having Renovascular Hypertension: A Meta-Analysis



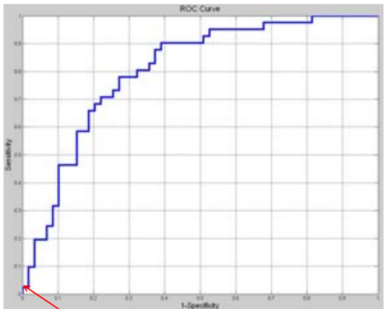
17



Threshold higher than any value in the dataset: Everyone tests negative

		Truth	
		+	-
Test	+	0	0
	-	50	100

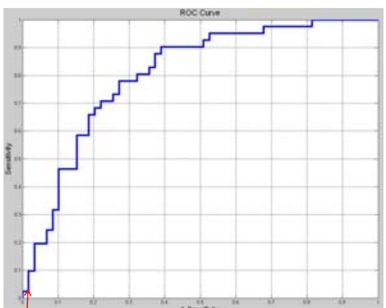
18



Lower threshold so that one person tests positive:
This person has the condition (e.g., RAS)

		Truth	
		+	-
Test	+	1	0
	-	49	100

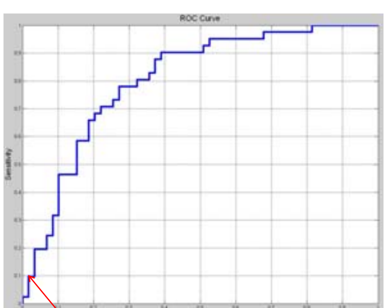
19



Next person tests negative

		Truth	
		+	-
Test	+	1	1
	-	49	99

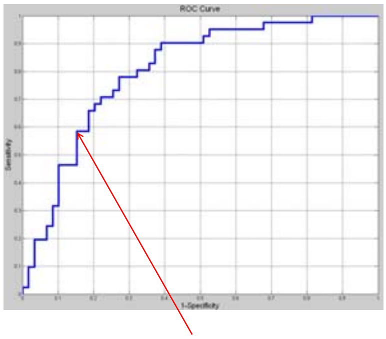
20



Next 4 people test positive

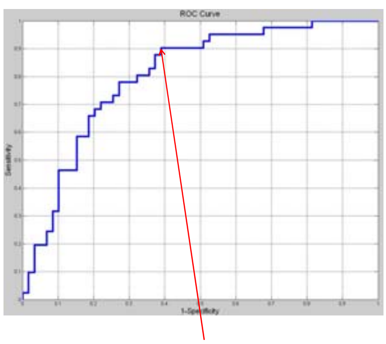
		Truth	
		+	-
Test	+	5	1
	-	45	99

21



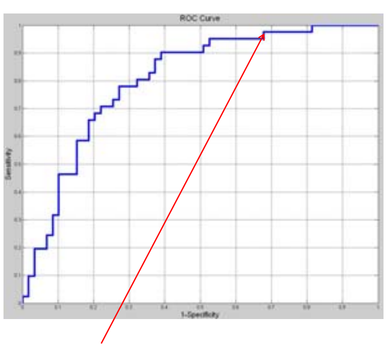
One reasonable threshold: Sensitivity = 58%, Specificity = 85%

22



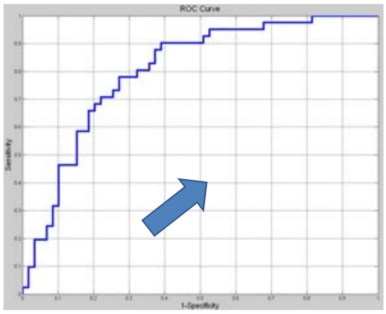
Another reasonable threshold: Sensitivity = 90%, Specificity = 61%

23



Threshold now very low:
Virtually everyone tests positive

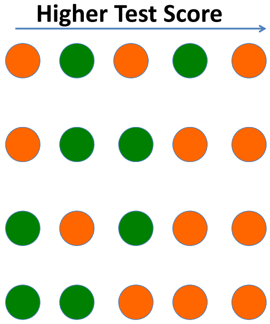
24



Area under the ROC curve ("c-statistic"):
0.5 = random chance
1.0 = all true-positives have higher values than any true-negatives

25

Higher Test Score



c = 0.5

c = 0.67 (4/6)

c = 0.83 (5/6)

c = 1.0

Area under the ROC curve ("c-statistic"):
Probability that, if you drew two observations at random, the one with true disease would have the higher score.

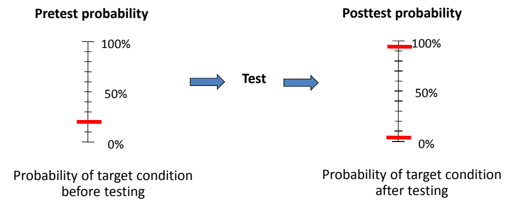
26

ROC Curves

- Convert numerical data into sensitivity and specificity at each possible threshold
- Give some idea of "separation" between people with and without a given condition
- Useful for determining appropriate thresholds for testing
- Not as useful if the threshold has already been determined
  - *Just calculate sensitivity and specificity instead!*

27

Diagnostic tests in a Bayesian framework



Example: Xpert MTB/RIF for diagnosis of active TB

Pre-test probability	Diagnostic test	Post-test probability
20%	GeneXpert neg	2%
	GeneXpert pos	95%

28

Likelihood Ratios

- (Pre-test odds) * LR = (Post-test odds)

- +LR = Sensitivity/(1 – Specificity)
- -LR = (1 – Sensitivity)/Specificity

Example of nuclear stress test for CAD:
Sensitivity = 90%, Specificity = 80%
+LR = 4.5, -LR = 0.13

29

Likelihood Ratios

- (Pre-test odds) * LR = (Post-test odds)
- Pre-test odds = 1 (i.e., probability 50%)

50 50

30

Likelihood Ratios

- (Pre-test odds) * LR = (Post-test odds)

- Pre-test odds = 1 (i.e., probability 50%)
- Apply test

	50 CAD	50 CAD		
Test positive	45	10	PPV: 82%	
Test negative	5	40	NPV: 89%	+LR: 4.5
	Sens: 90%	Spec: 80%		-LR: 0.13

31

Likelihood Ratios

- (Pre-test odds) * LR = (Post-test odds)

- Pre-test odds = 1 (i.e., probability 50%)
- Apply test
- Post-test odds (among those testing positive) = 45/10 = 4.5

	50 CAD	50 CAD		
Test positive	45	10	PPV: 82%	
Test negative	5	40	NPV: 89%	+LR: 4.5
	Sens: 90%	Spec: 80%		-LR: 0.13

32

Likelihood Ratios

- (Pre-test odds) * LR = (Post-test odds)

- Pre-test odds = 1 (i.e., probability 50%)
- Apply test
- Post-test odds (among those testing negative) = 5/40 = 0.13

	50 CAD	50 CAD		
Test positive	45	10	PPV: 82%	
Test negative	5	40	NPV: 89%	+LR: 4.5
	Sens: 90%	Spec: 80%		-LR: 0.13

33

Likelihood Ratios

- (Pre-test odds) * LR = (Post-test odds)
- Pre-test odds = 0.25 (i.e., probability 20%)

20 80

34

Likelihood Ratios

- (Pre-test odds) * LR = (Post-test odds)
- Pre-test odds = 0.25 (i.e., probability 20%)
- Apply test

20 80

	CAD	CAD		
Test positive	**18**	**16**	PPV: 82%	
Test negative	**2**	**64**	NPV: 89%	+LR: 4.5
	Sens: 90%	Spec: 80%		-LR: 0.13

35

Likelihood Ratios

- (Pre-test odds) * LR = (Post-test odds)
- Pre-test odds = 0.25 (i.e., probability 20%)
- Apply test
- Post-test odds (positive) = 18/16 = 1.13 = 0.25 * 4.5

20 80

	CAD	CAD		
Test positive	**18**	**16**	PPV: 82%	
Test negative	**2**	**64**	NPV: 89%	+LR: 4.5
	Sens: 90%	Spec: 80%		-LR: 0.13

36

Likelihood Ratios

- (Pre-test odds) * LR = (Post-test odds)

- Pre-test odds = 0.25 (i.e., probability 20%)
- Apply test
- Post-test odds (negative) = 2/64 = 0.03 = 0.25 * 0.13

	20 CAD	80 CAD		
Test positive	**18**	**16**	PPV: 82%	
Test negative	2	64	NPV: 89%	+LR: 4.5
	Sens: 90%	Spec: 80%		-LR: 0.13

37

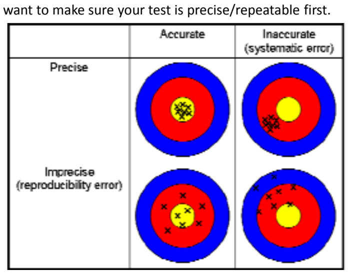
LR's: Bottom Line

- Any time you perform a test, you should be able to specify:
  - ***Pre-test probability***
  - ***Likelihood ratios of the test***
  - ***Post-test probability if test is positive/negative***
  - ***Management thresholds***
- <u>If the post-test probability will not lead to different management, *do not order the test.*</u>
  - It's OK to be uncertain!!

38

Accuracy vs. Precision

- **Accuracy:** How close diagnostic test results are to the "truth"
  - More a measure of effectiveness/appropriateness
- **Precision:** How close diagnostic test results are to each other
  - More a measure of technical specification
  - Usually want to make sure your test is precise/repeatable first.



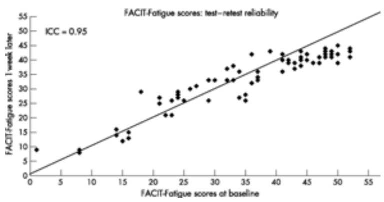
39

Measures of Precision/Repeatability

- **Intraclass correlation coefficient (ICC):**

$$\frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

where σ_a^2 = between-group variance and σ_e^2 = within-group variance



Chandran V, *Ann Rheum Dis* 2007; 66:936

40

Measures of Precision/Repeatability

- **Intraclass correlation coefficient (ICC):**
- Wide variety of uses (and statistical forms)
- Similar to the standard (Pearson) correlation coefficient
  - *But uses a pooled mean and s.d. – in other words, considers groups/pairs of measurements.*
- Easily calculable with most statistical packages
- Helpful for describing reliability/precision of diagnostic tests with continuous scales
- *What if your test is a binary measurement?*

41

Measures of Precision/Repeatability

- **(Cohen's) Kappa statistic:**

$$\frac{\text{(observed agreement)} - \text{(expected agreement)}}{1 - \text{(expected agreement)}}$$

- "Where does agreement fall, on a scale from 0 = random chance, to 1 = perfect agreement?"
  - Landis & Koch (Biometrics, 1977):
  - 0-0.2 = slight agreement
  - 0.21-0.4 = fair
  - 0.41-0.6 = moderate
  - 0.61-0.8 = substantial
  - 0.81-1.0 = almost perfect
- These categories are completely arbitrary, may be more useful for some measurements than others.

42

Measures of Precision/Repeatability

- **Kappa example:**
- Reading CXR as TB vs. not TB

		Reader 2	
		TB	No TB
Reader 1	TB	10	5
	No TB	2	83

43

Kappa Example

		Reader 2	
		TB	No TB
Reader 1	TB	10	5
	No TB	2	83

- **Could measure simple percent agreement: (83+10)/100**
- *But this is artificially inflated by the fact that most people do not have TB.*

44

Kappa Example

		Reader 2	
		TB	No TB
Reader 1	TB	0	5
	No TB	2	93

- *For example, percent agreement here is 93%, but the two readers don't agree on a single TB case!*

45

Kappa Example

- **First, calculate expected agreement**

		Reader 2		
		TB	No TB	
Reader 1	TB	10	5	15
	No TB	2	83	85
		12	88	Total = 100

46

Kappa Example

- **Calculate expected agreement**

		Reader 2		
		TB	No TB	
Reader 1	TB	0.12*0.15 = 0.018	0.88*0.15 = 0.132	15
	No TB	0.12*0.85 = 0.102	0.88*0.85 = 0.748	85
		12	88	Total = 100

47

Kappa Example

- **Multiply by the total**
  - Expected agreement = 74.8 + 1.8 = 76.6/100

		Reader 2		
		TB	No TB	
Reader 1	TB	1.8	13.2	15
	No TB	10.2	74.8	85
		12	88	Total = 100

48

Kappa Example

		Reader 2	
		TB	No TB
Reader 1	TB	10	5
	No TB	2	83

- **Kappa = (observed – expected)/(1 – expected)**
  = (0.93 – 0.766)/(1 – 0.766)
  = 0.70
  "good/substantial," according to Landis & Koch

49

Part I: Summary

- **Accuracy vs. Precision**
- **Measures of Accuracy:**
  - Sensitivity/specificity: characteristics of the test
  - PPV/NPV: depend on prevalence
  - ROC curve: summary measure of accuracy using different cutoffs
  - Likelihood ratios: how are tests used in decision-making?
    - *Know your pre-test probability, LRs, and management thresholds!*
- **Measures of Precision/Agreement:**
  - Intraclass correlation coefficient: continuous measures
  - Kappa statistic: binary measures

50

Part II: Evaluation of Diagnostic Tests

51

Learning objectives

- Part I: Recap basic epidemiological tools for evaluating diagnostics
  - Accuracy
    - *Sensitivity & Specificity*
    - *Positive & Negative Predictive Value*
    - *Receiver Operating Curve (ROC) Analysis*
    - *Bayesian Approaches (Likelihood Ratio)*
  - Precision
    - *Intra-Class Correlation*
    - *Kappa Statistic*

- **Part II: Discuss challenges in evaluation of diagnostic tools**
  - Recognize differences between diagnostics and therapeutics
  - Understand challenges in evaluation of diagnostic tests

52

Diagnostics vs. Therapeutics

Diagnostics	Therapeutics
Work outside the body	Work inside the body
Designed to detect disease	Designed to treat disease
System-dependent	Direct biological effect
"Adverse event" = wrong result	Adverse event = direct toxicity
People with & without disease	People with disease only
Cost depends on other factors	Cost often direct administration
Make drugs effective	Make diagnostics effective

53

Test phases for therapeutics

Phase I Safety and Pharmacokinetics
Small studies of 10s of healthy volunteers

Phase II Dose-Ranging, Adverse Events, Early Efficacy
Studies of 100s of volunteers, e.g., advanced disease

Phase III Efficacy, Clinical Effectiveness
Randomized trials of 1,000s of representative individuals

Phase IV Post-Marketing Surveillance (Rare Events, etc.)
Population-based evaluations

Does This System Work for Diagnostics?

54

"Phase I-IV" for Diagnostics?

Phase I
Safety? Pharmacokinetics?
Diagnostics do not have a direct biological effect

Phase II
Dose-Ranging = Setting Thresholds?; Early Efficacy = Accuracy?
Is there a difference between CrCl and "CKD yes/no"?
Diagnostics will perform differently depending on setting

Phase III
Randomized controlled trial?
Diagnostics will change index of suspicion, treatment patterns, etc.
Do we need to know this before licensing a new test?

Phase IV
Post-deployment
How do you know if a diagnostic is performing well after it's deployed?
What rare "adverse events" would we look for?

55

Models of diagnostic test evaluation phases: It's complicated!

Table 1 Summary of Proposals for the Phased Evaluation of Medical Tests

Levels/Phases	Loop 1978[6]	Zweig 1982[23]	Guyatt 1986[17]	Freedman 1987[15]	Memorandum 1990[20]	Fryback 1991[4]	Kent 1992[9]	Taylor 1993[16]	Silverstein 1994[12]	van der Schouw 1995[18]	Mackenzie 1995[10]	Pearl 1999[13]	Houn 2000[21]	Gatsonis 2000[22]	Sackett 2002[24]	Haddow 2003[14]	Pepe 2005[25]	Taube 2005[26]	Van den Bruel 2007[19]
Technical efficacy		1	1	1	1	1	1	1–3		1	1	1	1	1		1	1–3	1–2	1
Intended use			2															3	2
Diagnostic accuracy			3	2	2	2	2	4	1		2	2	2			2	4	4–6	3
Usual range		2			3					2				1	1–2				
Subgroups																			
Clinical population		3			4					3				2	3			7	
Diagnostic thinking efficacy	1		4			3	3		2	4	3	3							
Therapeutic efficacy	2		5			4	4		3		4	4	3	3–4					
Patient outcome efficacy	3		6	3	5	5	5	5	4		5	5	4	3–4	4	3	5		4
Societal efficacy				4		6				5		6	5			4			5

Technical requirements → Test accuracy → Effects on decisions → Effects on patient outcomes → Effects on health care system

56

Lijmer et al. *Med Decis Making 2009; 29; E13*

Evaluating Accuracy

- We think of accuracy as being an intrinsic characteristic of the test, but it often is not.
  - *Depends on quality of lab using the test, population characteristics, etc.*
- Sensitivity and specificity require the presence of a "gold standard," which is often hard to define.
  - *If your new test claims to be better than your old test, how do you distinguish a false-positive new test from a false-negative old test?*
- Sensitivity and specificity are only useful when tests are being used in binary fashion (presence vs. absence of condition).
  - *Many tests (e.g., WBC count) are used in a way that the numerical value has meaning, and contributes partial information.*
  - *Other tests (e.g., CXR) provide data in many different domains.*

57

Example: Evaluating the Accuracy of Xpert

- How well does Xpert distinguish people with active TB from those without active TB?
  - *Is this the same question at JHH lab vs. Delhi, India?*
  - *How do you determine who has active TB when 20% of TB is culture-negative?*
  - *Xpert sensitivity for smear-negative TB: 70% in Uganda, 20% in Canada*
- Are these even the most important questions?

Why Might An Accurate Test Not Improve Outcomes?
Let Me Count the Ways…(There are More)

- Test result is too slow to change management.
- Test result doesn't make it back to the ordering physician.
- Patient is already too sick/too healthy for the test result to matter.
- Test is performed inappropriately.
- Result of test is acted upon inappropriately.
- The test in question is only one of many tests ordered.
- Treatment is not available (too expensive, out of stock, etc.).
- Treatment is not delivered.
- Patient declines the treatment.
- Another condition co-exists, and the patient suffers outcomes based on the other condition instead.

- *Should we hold diagnostic tests to a standard of making accurate diagnoses, or improving patient outcomes?*

59

Diagnostic Test Result ≠ Clinical Outcomes

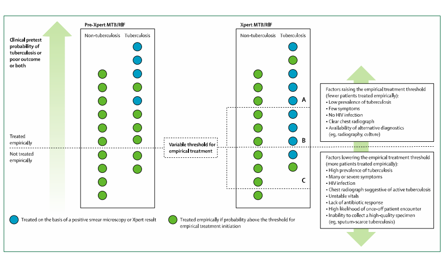


Figure: **Factors affecting the empirical treatment threshold**
The threshold for empirical treatment might rise, stay the same, or lower after Xpert implementation. In the pre-Xpert scenario, six of ten patients without tuberculosis and eight of ten patients with tuberculosis are treated. Xpert implementation could change the threshold for empirical treatment according to one of three different scenarios: threshold raised (A), Xpert will reduce false-positive treatment of people without tuberculosis and increase true-positive treatment; threshold remains constant (B), Xpert will not change the rates of false-positive treatments, but will increase true-positive treatments; or threshold lowered (C), Xpert will increase false-positive treatment of people without tuberculosis and increase true-positive treatment. Tests are assumed to have high specificity.

Proof From Clinical Trials

Articles

Feasibility, accuracy, and clinical effect of point-of-care Xpert MTB/RIF testing for tuberculosis in primary-care settings in Africa: a multicentre, randomised, controlled trial

*Grant Theron, Lynn Zijenah, Duncan Chanda, Petra Clowes, Andrea Rachow, Maia Lesosky, Wilbert Bara, Stanley Mungofa, Madhukar Pai, Michael Hoelscher, David Dowdy, Alex Pym, Peter Mwaba, Peter Mason, Jonny Peter, Keertan Dheda, for the TB-NEAT team**

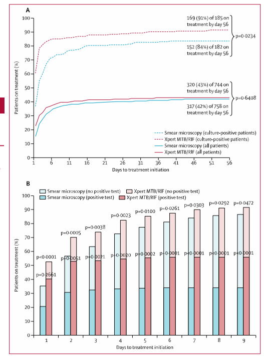
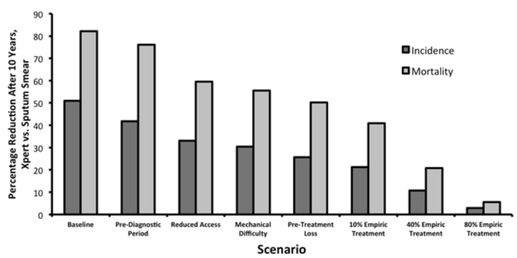


Figure 3: Initiation of anti-tuberculosis treatment in all patients and culture-positive patients, per allocation group

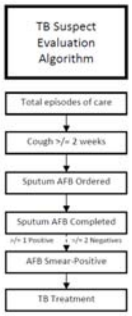
Impact of The Same Diagnostic Test on Mortality in Different Systems



Diagnostic Tests and Impact on Decision-Making: The TB Example

- TB tests (like all tests) are not performed in isolation but are part of a system.
- A test that gives you the "right" result may not change clinical management.
  - *Example: TB culture (most sensitive test) changed management of <1% of patients in an Indian hospital.* (Stall N, IJTLD 2011; 5:641)
    - Slow
    - Results not trusted
    - Empiric therapy had already been initiated

63



Indicator	Numbers, by quarter				Proportions*, by quarter				p-Value
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	
	14,852	14,652	17,369	16,036	--	--	--	--	--
	365	280	349	294	2.5%	2.0%	2.1%	1.8%	0.27
1	75	111	211	155	21%	40%	60%	53%	0.014
2	55	90	168	119	73%	81%	80%	77%	0.85
	7	19	30	25	13%	21%	18%	21%	0.25
3	5	13	23	21	71%	68%	77%	84%	0.016
Cumulative Probability of Being Diagnosed with and Treated for TB*					11%	22%	37%	34%	0.005

- *Davis JL et al, AJRCCM 2011*

64

Evaluation of Diagnostic Tests

- Diagnostics are different from therapeutics (or vaccines).
  - *A different system of evaluation is required.*
  - *Different tools are used for that evaluation.*
- Progression of evaluation for diagnostics:
  - *Technical specifications (e.g., precision)*
  - *Accuracy*
    - First in known positives vs. negatives
    - Then in the target population
  - *Effect on clinical decisions*
  - *Effect on patient outcomes*
  - *Utility to society*
- Evaluation of diagnostics requires evaluation of a system, not just a test.

65

The critical question when assessing the impact of diagnostic testing on patient outcomes

What is the intended **incremental value** of the test on outcomes (short- and long-term patient outcomes and costs)?

Examples of incremental value:

- Less use of more expensive testing (e.g., D-dimer for DVT)
- Patient convenience/more tx initiation (e.g. rapid strep test)
- Improved patient symptoms (e.g., CT urography/nephrolithiasis)
- Reduced mortality (e.g., colonoscopy)

Better accuracy may be appropriate for licensure of a test, but tests should not be recommended/performed if they do not add incremental value, either to patients or to the healthcare system.

66

Lord et al. Med Dec Making 2009;29:E1

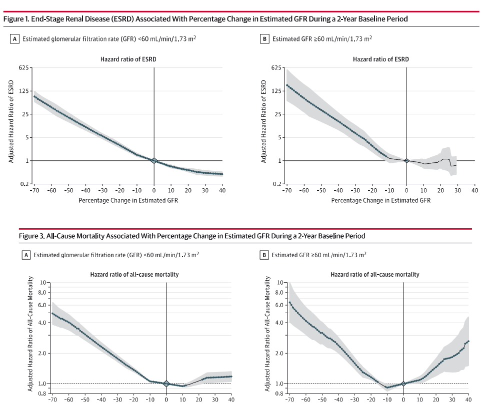
A Recent Example

Original Investigation

Decline in Estimated Glomerular Filtration Rate and Subsequent Risk of End-Stage Renal Disease and Mortality

Josef Coresh, MD, PhD; Tanvir Chowdhury Turin, MD, PhD; Kunihiro Matsushita, MD, PhD; Yingying Sang, MSc; Shoshana H. Ballew, PhD; Lawrence J. Appel, MD; Hisatomi Arima, MD; Steven J. Chadban, PhD; Massimo Cirillo, MD; Ognjenka Djurdjev, MSc; Jamie A. Green, MD; Gunnar H. Heine, MD; Lesley A. Inker, MD; Fujiko Irie, MD, PhD; Areef Ishani, MD, MS; Joachim H. Ix, MD, MAS; Csaba P. Kovesdy, MD; Angharad Marks, MBBCh; Takayoshi Ohkubo, MD, PhD; Varda Shalev, MD; Anoop Shankar, MD; Chi Pang Wen, MD, DrPH; Paul E. de Jong, MD, PhD; Kunitoshi Iseki, MD, PhD; Benedicte Stengel, MD, PhD; Ron T. Gansevoort, MD, PhD; Andrew S. Levey, MD; for the CKD Prognosis Consortium

CONCLUSIONS AND RELEVANCE Declines in estimated GFR smaller than a doubling of serum creatinine concentration occurred more commonly and were strongly and consistently associated with the risk of ESRD and mortality, supporting consideration of lesser declines in estimated GFR (such as a 30% reduction over 2 years) as an alternative end point for CKD progression.



Necessary Steps Before Using 30%ΔGFR as a Diagnostic?

- Validation of accuracy
- Effect on decision-making
  - *Do treatment decisions change based on this new knowledge?*
- Effect on patient outcomes
  - *Do these treatment decisions actually impact important outcomes?*
- Effect on society
  - *Is the test cost-effective, does it lead to overdiagnosis, improved CKD morbidity/mortality, etc.?*

Summary: Evaluation of Diagnostic Tests

- **Diagnostic tests are different from therapeutics.**
  - Different process of evaluation
- **Accurate test results may not imply better patient outcomes.**
  - Progression of evaluation:
    - Technical specs/precision
    - Accuracy
    - Effects on decisions
    - Effects on patient outcomes
    - Effects on the healthcare system
- **Evaluation should center on a test's <u>incremental value</u>.**
  - What is the intended benefit of the test to patients and society?

70

Diagnostic testing: Take-home messages

- Key epidemiological measures in evaluating diagnostics:
  - *Accuracy: Sensitivity and specificity, ROC curves*
  - *Clinical Utility: LRs (know your pre-test-probability!)*
  - *Precision/Reproducibility: ICC & Kappa*
- When evaluating diagnostic tests:
  - *Remember that accuracy does not imply better patient outcomes.*
  - *Clarify a test's intended <u>incremental value.</u>*
  - *Consider effects on decision-making, patient outcomes, and society.*

71