Evaluation of Diagnostic Tests
July 22, 2013
Introduction to Clinical Research: A Two-week Intensive Course
David W. Dowdy, MD, PhD
Department of Epidemiology
Johns Hopkins
Bloomberg School of Public Health

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 study design and evaluation of diagnostics

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

The NEW ENGLAND JOURNAL OF MEDICINE

Rapid Molecular Detection of Tuberculosis and Rifampin Resistance

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

Summary

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
Part I: Epidemiological Tools for Evaluation of Diagnostics

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.

<table>
<thead>
<tr>
<th></th>
<th>Accurate</th>
<th>Inaccurate (systematic error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precise</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Inaccurate</strong> (systematic error)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

### Test Accuracy

<table>
<thead>
<tr>
<th>New Test</th>
<th>“Gold Standard”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
</tr>
<tr>
<td>Positive</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>True</td>
</tr>
<tr>
<td>Negative</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>False</td>
</tr>
</tbody>
</table>

- Sensitivity = \( \frac{A}{A+C} \)
- Specificity = \( \frac{D}{B+D} \)
- PPV = \( \frac{A}{A+B} \)
- NPV = \( \frac{D}{C+D} \)

### TB Culture

<table>
<thead>
<tr>
<th>Xpert MTB/RIF</th>
<th>TB Culture</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
</tr>
<tr>
<td>Positive</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>True</td>
</tr>
<tr>
<td>Negative</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>False</td>
</tr>
</tbody>
</table>

100 900

- Sensitivity = \( \frac{A}{A+C} \)
- Specificity = \( \frac{D}{B+D} \)
- PPV = \( \frac{A}{A+B} \)
- NPV = \( \frac{D}{C+D} \)
Test Accuracy

<table>
<thead>
<tr>
<th></th>
<th>TB Culture</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Xpert MTB/RIF</td>
<td>70</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>True</td>
<td>False</td>
</tr>
<tr>
<td>Negative</td>
<td>30</td>
<td>890</td>
</tr>
<tr>
<td></td>
<td>False</td>
<td>True</td>
</tr>
</tbody>
</table>

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

Effect of Prevalence on PPV and NPV

<table>
<thead>
<tr>
<th></th>
<th>“Gold Standard”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
</tr>
<tr>
<td>New Test</td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>9</td>
</tr>
<tr>
<td>Negative</td>
<td>1</td>
</tr>
</tbody>
</table>

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

Effect of Prevalence on PPV and NPV

<table>
<thead>
<tr>
<th></th>
<th>“Gold Standard”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
</tr>
<tr>
<td>New Test</td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>90</td>
</tr>
<tr>
<td>Negative</td>
<td>10</td>
</tr>
</tbody>
</table>

Now increase prevalence to 50%.
PPV = 90/91 = 98.9%
NPV = 99/109 = 90.8%
As prevalence increases, PPV increases and NPV decreases.
Sensitivity & Specificity in the Real World

<table>
<thead>
<tr>
<th>Test Method</th>
<th>Small Population</th>
<th>Large Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROC Curves</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Threshold higher than any value in the dataset:
Everyone tests negative
Lower threshold so that one person tests positive: This person has the condition (e.g., RAS)

<table>
<thead>
<tr>
<th>Truth</th>
<th>Test</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>49</td>
<td>100</td>
</tr>
</tbody>
</table>

Next person tests negative

<table>
<thead>
<tr>
<th>Truth</th>
<th>Test</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>49</td>
<td>99</td>
</tr>
</tbody>
</table>

Next 4 people test positive

<table>
<thead>
<tr>
<th>Truth</th>
<th>Test</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>45</td>
<td>99</td>
</tr>
</tbody>
</table>
One reasonable threshold: Sensitivity = 58%, Specificity = 85%

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

Threshold now very low: Virtually everyone tests positive
Area under the ROC curve ("c-statistic"):  
0.5 = random chance  
1.0 = all true-positives have higher values than any true-negatives

**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.

**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!
Diagnostic tests in a Bayesian framework

Example: Xpert MTB/RIF for diagnosis of active TB

<table>
<thead>
<tr>
<th>Pre-test probability</th>
<th>Diagnostic test</th>
<th>Post-test probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>20%</td>
<td>GeneXpert neg</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td>GeneXpert pos</td>
<td>95%</td>
</tr>
</tbody>
</table>

Likelihood Ratios

- (Pre-test odds) * LR = (Post-test odds)
- +LR = Sensitivity/(1 – Specificity)
- -LR = Specificity/(1 – Sensitivity)

Often reported as an inverse

<table>
<thead>
<tr>
<th>TB</th>
<th>No TB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test positive</td>
<td>90</td>
</tr>
<tr>
<td>Test negative</td>
<td>10</td>
</tr>
</tbody>
</table>

Sens 90% Spec 80%

PPV 90% NPV 89%

+LR: 4.5 
-LR: 0.13

Likelihood Ratios

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

<table>
<thead>
<tr>
<th>CAD</th>
<th>No CAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test positive</td>
<td>90</td>
</tr>
<tr>
<td>Test negative</td>
<td>10</td>
</tr>
</tbody>
</table>

Sens 90% Spec 80%

PPV 82% NPV 89%

+LR: 4.5 
-LR: 8.0 or 0.13
## Likelihood Ratios

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

<table>
<thead>
<tr>
<th></th>
<th>CAD</th>
<th>No CAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test positive</td>
<td>45</td>
<td>10</td>
</tr>
<tr>
<td>Test negative</td>
<td>5</td>
<td>40</td>
</tr>
</tbody>
</table>

- PPV: 82%
- NPV: 89%
- Sens: 90%
- Spec: 80%
- +LR: 4.5
- ‐LR: 0.13

### Pre-test odds

\[
\text{Pre-test odds} = \frac{1}{1 + LR} = \frac{1}{4.5 + 1} = 0.22
\]

**Post-test odds (among those testing positive)**

\[
\text{Post-test odds} = \frac{\text{Post-test odds}}{\text{Pre-test odds}} = \frac{45}{10} = 4.5
\]

**Post-test odds (among those testing negative)**

\[
\text{Post-test odds} = \frac{\text{Post-test odds}}{\text{Pre-test odds}} = \frac{5}{40} = 0.13
\]

---

**Post-test odds (among those testing positive)**

\[
\text{Post-test odds} = \frac{45}{10} = 4.5
\]

**Post-test odds (among those testing negative)**

\[
\text{Post-test odds} = \frac{5}{40} = 0.13
\]
Likelihood Ratios

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

<table>
<thead>
<tr>
<th>Test positive</th>
<th>Test negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAD</td>
<td>No CAD</td>
</tr>
<tr>
<td>CAD</td>
<td>No CAD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>PPV: 82%</th>
<th>NPV: 89%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test positive</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>Test negative</td>
<td>80</td>
<td>20</td>
</tr>
</tbody>
</table>

Sens: 90% Spec: 80%
+LR: 4.5 -LR: 8.0 or 0.13

Apply test

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

<table>
<thead>
<tr>
<th>Test positive</th>
<th>Test negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAD</td>
<td>No CAD</td>
</tr>
<tr>
<td>CAD</td>
<td>No CAD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>PPV: 82%</th>
<th>NPV: 89%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test positive</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>Test negative</td>
<td>16</td>
<td>64</td>
</tr>
</tbody>
</table>

Sens: 90% Spec: 80%
+LR: 4.5 -LR: 8.0 or 0.13

Apply test

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

Post-test odds (positive) = 18/16 = 1.13 = 0.25 * 4.5

<table>
<thead>
<tr>
<th>Test positive</th>
<th>Test negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAD</td>
<td>No CAD</td>
</tr>
<tr>
<td>CAD</td>
<td>No CAD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>PPV: 82%</th>
<th>NPV: 89%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test positive</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>Test negative</td>
<td>16</td>
<td>64</td>
</tr>
</tbody>
</table>

Sens: 90% Spec: 80%
+LR: 4.5 -LR: 8.0 or 0.13

Apply test

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

Post-test odds (positive) = 18/16 = 1.13 = 0.25 * 4.5

<table>
<thead>
<tr>
<th>Test positive</th>
<th>Test negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAD</td>
<td>No CAD</td>
</tr>
<tr>
<td>CAD</td>
<td>No CAD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>PPV: 82%</th>
<th>NPV: 89%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test positive</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>Test negative</td>
<td>16</td>
<td>64</td>
</tr>
</tbody>
</table>

Sens: 90% Spec: 80%
+LR: 4.5 -LR: 8.0 or 0.13
Likelihood Ratios

- (Pre-test odds) * LR = (Post-test odds)
- Pre-test odds = 1 (i.e., probability 50%)
- Apply test
- Post-test odds (negative) = 2/64 = 0.03 = 0.25 * 0.13

<table>
<thead>
<tr>
<th></th>
<th>CAD</th>
<th>No CAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>18</td>
<td>16</td>
</tr>
<tr>
<td>positive</td>
<td>2</td>
<td>64</td>
</tr>
</tbody>
</table>

PPV: 82%
NPV: 89%
Sens: 90%
Spec: 80%
+LR: 4.5
-LR: 0.13

Measures of Precision/Repeatability

- Intraclass correlation coefficient (ICC):
  \[
  \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}
  \]
  where \( \sigma_a^2 \) = variance of a "random effect" and \( \sigma_e^2 \) = variance of "error/noise"

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.
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 each measurement to be of the same quantity.
  - 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?

- **(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]:
      - 0.00 – 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.

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

<table>
<thead>
<tr>
<th></th>
<th>Reader 1</th>
<th>Reader 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TB</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>No TB</td>
<td>2</td>
<td>83</td>
</tr>
</tbody>
</table>
Kappa Example

<table>
<thead>
<tr>
<th>Reader 1</th>
<th>Reader 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TB</td>
<td>10</td>
</tr>
<tr>
<td>No TB</td>
<td>2</td>
</tr>
</tbody>
</table>

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

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

Kappa Example

<table>
<thead>
<tr>
<th>Reader 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TB</td>
</tr>
<tr>
<td>No TB</td>
</tr>
</tbody>
</table>

\[
\text{Expected agreement} = \frac{10 \times 0 + 5 \times 2 + 2 \times 5 + 83 \times 93}{100}
\]

- First, calculate expected agreement

Kappa Example

<table>
<thead>
<tr>
<th>Reader 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TB</td>
</tr>
<tr>
<td>No TB</td>
</tr>
</tbody>
</table>

\[
\text{Actual agreement} = \frac{10 \times 0 + 5 \times 2 + 2 \times 5 + 83 \times 93}{100}
\]

- Total = 100
Kappa Example

- Calculate expected agreement

<table>
<thead>
<tr>
<th></th>
<th>Reader 2</th>
<th>Reader 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TB</td>
<td>No TB</td>
</tr>
<tr>
<td>Reader 2</td>
<td>0.12*0.15 = 0.018</td>
<td>0.88*0.15 = 0.132</td>
</tr>
<tr>
<td>Reader 1</td>
<td>0.12*0.85 = 0.102</td>
<td>0.88*0.85 = 0.748</td>
</tr>
</tbody>
</table>

\[ \text{Expected agreement} = 0.18 + 0.132 = 0.312 \]

\[ \text{Total} = 100 \]

- Multiply by the total
  
  \[ \text{Expected agreement} = 74.8 + 1.8 = 76.6/100 \]

<table>
<thead>
<tr>
<th></th>
<th>Reader 2</th>
<th>Reader 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TB</td>
<td>No TB</td>
</tr>
<tr>
<td>Reader 2</td>
<td>1.8</td>
<td>13.2</td>
</tr>
<tr>
<td>Reader 1</td>
<td>10.2</td>
<td>74.8</td>
</tr>
</tbody>
</table>

\[ \text{Total} = 100 \]

Kappa Example

- Kappa = \( \frac{\text{observed} - \text{expected}}{1 - \text{expected}} \)

\[ \text{Kappa} = \frac{0.93 - 0.766}{1 - 0.766} = 0.70 \]

"good/substantial," according to Landis & Koch
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?

- Measures of Precision/Agreement:
  - Intraclass correlation coefficient: continuous measures
  - Kappa statistic: binary measures

Part II: Evaluation of Diagnostic Tests

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
  - Intraclass Correlation
  - Kappa Statistic

- Part II: Discuss challenges in evaluation of diagnostic tools
  - Recognize differences between diagnostics and therapeutics
  - Understand challenges in study design and evaluation of diagnostics
**Reasonable Research Questions: The Example of GeneXpert**

- How reproducible is Xpert when performed in different conditions? When read by different readers? (Precision)
- How well does Xpert distinguish people with active TB from those without active TB? (Accuracy)

**A Few Notes on 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.

**Evaluating the Accuracy of GeneXpert**

- 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?
- What about early forms of TB?
- Are these even the most important questions?
Other Research Questions: More Relevant?
- How reproducible is Xpert when performed in different conditions? When read by different readers? (Precision)
- How well does Xpert distinguish people with active TB from those without active TB? (Accuracy)
- How does Xpert affect clinical management?
- How does Xpert affect patient-important outcomes (morbidity and mortality)?
- Is Xpert cost-effective?

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?

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, 2011; 5:641)
- Slow
- Results not trusted
- Empiric therapy had already been initiated
Diagnostics vs. Therapeutics

<table>
<thead>
<tr>
<th>Diagnostics</th>
<th>Therapeutics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work outside the body</td>
<td>Work inside the body</td>
</tr>
<tr>
<td>Designed to detect disease</td>
<td>Designed to treat disease</td>
</tr>
<tr>
<td>System-dependent</td>
<td>Direct biological effect</td>
</tr>
<tr>
<td>“Adverse event” = wrong result</td>
<td>Adverse event = direct toxicity</td>
</tr>
<tr>
<td>People with &amp; without disease</td>
<td>People with disease only</td>
</tr>
<tr>
<td>Cost depends on other factors</td>
<td>Cost often direct administration</td>
</tr>
<tr>
<td>Make drugs effective</td>
<td>Make diagnostics effective</td>
</tr>
</tbody>
</table>

Test phases for therapeutics

<table>
<thead>
<tr>
<th>Phase</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase I</td>
<td>Safety and Pharmacokinetics&lt;br&gt;Small studies of 10s of healthy volunteers</td>
</tr>
<tr>
<td>Phase II</td>
<td>Dose-Ranging, Adverse Events, Early Efficacy&lt;br&gt;Studies of 100s of volunteers, e.g., advanced disease</td>
</tr>
<tr>
<td>Phase III</td>
<td>Efficacy, Clinical Effectiveness&lt;br&gt;Randomized trials of 1,000s of representative individuals</td>
</tr>
<tr>
<td>Phase IV</td>
<td>Post-Marketing Surveillance (Rare Events, etc.)&lt;br&gt;Population-based evaluations</td>
</tr>
</tbody>
</table>

Does This System Work for Diagnostics?
“Phase I-IV” for Diagnostics?

Phase I  Safety? Pharmacokinetics? Can 10s of people give any data? Diagnostics do not have a direct biological effect

Phase II  Dose-Ranging = Setting Thresholds?; Early Efficacy = Accuracy? Is there a difference between CVD and “CVD 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?

Models of diagnostic test evaluation phases:

It’s complicated!

<table>
<thead>
<tr>
<th>Phase</th>
<th>Technical specifications</th>
<th>Test accuracy</th>
<th>Effects on decisions</th>
<th>Effects on patient outcomes</th>
<th>Effects on health care system</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


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.
Randomized Trials of Diagnostic Tests

- Ultimately, want to demonstrate that a diagnostic test has impact on patient outcomes, not just test results.

Phase V

Patients suspected of having TB

- Randomized trial

The critical questions when assessing patient outcomes

- What is the intended incremental value of the test on outcomes (short- and long-term patient outcomes and costs)?
- What type of evidence is needed to assess this incremental value?

Recommended approach

- Define the purpose of the test
- Display the existing test-treatment strategy
- Display the new test-treatment strategy
- Identify the critical comparison to assess the incremental value
- Assess whether existing evidence suffices or if RCTs are required

Test-treatment strategy for replacement tests
Example: Liquid-based cytology to replace Pap smear for cervical cancer screening in order to reduce repeated testing (poor Pap smear quality)

Target population

- Pap smear
  - Test again
  - Test result
  - Management
  - Test pos path
  - Test neg path

- LBC
  - Test again
  - Test result
  - No change in management
  - Test pos path
  - Test neg path

RCT to compare short-term effects from testing

No long-term RCT needed

Patient outcomes

---

Not all tests are replacements:

Triage D-Dimer test to reduce ultrasound in patients at low risk for DVT

Target population

- Prior tests
  - Ultrasound
    - Test result
    - Management
    - Test pos pathway A
    - Test neg pathway B

- D-Dimer
  - Test result
  - Management
  - Test pos path
  - Test neg path

RCT to compare short-term effects from testing

RCT to compare long-term effects may not be necessary

Patient outcomes

---

Use of randomized trials for test evaluation

- Define the purpose of the test
- Display the existing test-treatment strategy
- Display the new test-treatment strategy
- Identify the critical comparison to assess the incremental value
- Assess whether existing evidence suffices or if RCTs are required

- Note that diagnostic tests are part of a larger pathway: cluster-randomized trials are often required
Summary: Evaluation of Diagnostic Tests

- Accuracy of a test may differ according to the setting.
- Accurate test results may not imply better patient outcomes.
- Diagnostic tests are different from therapeutics.
  - Different process of evaluation
  - Technical specs/precision -> accuracy -> effects on decisions -> effects on patient outcomes -> effects on the healthcare system
- Selection of randomized trials should center on the question of a test's incremental value.
  - Compare the old strategy to a new strategy
  - Long-term RCT needed only if test results change long-term management and treatment effects
  - Cluster RCTs often needed for diagnostics because they are integrated into a larger system (which must be randomized).

Key messages

- An accurate test result does not imply a healthier patient.
  - Diagnostics require a special set of evaluation techniques.
- Diagnostic test evaluation process:
  - Technical specifications/precision
  - Test accuracy
  - Effects on management decisions
  - Effects on patient outcomes
  - Effects on the system
- Key epidemiological tools reflect that process:
  - ICC & Kappa
  - Sensitivity/specificity, ROC curves, LRs
  - Randomized trials
  - Economic evaluations and others

Thank You!

Contact me: ddowdy@jhsph.edu