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

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

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 Eflornithine 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

7/14/2014
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>Incorrect (systematic error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precise</td>
<td><img src="image1.png" alt="Diagram" /></td>
<td><img src="image2.png" alt="Diagram" /></td>
</tr>
<tr>
<td>Standard (population level)</td>
<td><img src="image3.png" alt="Diagram" /></td>
<td><img src="image4.png" alt="Diagram" /></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>“Gold Standard”</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>A (True Positive)</td>
<td>B (False Positive)</td>
</tr>
<tr>
<td>Negative</td>
<td>C (False Positive)</td>
<td>D (True Negative)</td>
</tr>
</tbody>
</table>

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

Test Accuracy

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

- Sensitivity = A/(A+C)
- Specificity = D/(B+D)
- PPV = A/(A+B)
- NPV = 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 Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>890</td>
</tr>
<tr>
<td></td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

**80**

- Sensitivity = $\frac{70}{70+30} = 70\%$
- Specificity = $\frac{890}{10+890} = 98.9\%$
- PPV = $\frac{70}{70+10} = 87.5\%$
- NPV = $\frac{890}{30+890} = 96.7\%$

**920**

**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 Positive</td>
<td>9</td>
</tr>
<tr>
<td>New Test Negative</td>
<td>1</td>
</tr>
</tbody>
</table>

Take a test with 90% sensitivity and 99% specificity.
- Prevalence of condition here = $\frac{10}{110} = 9\%$
- PPV = $\frac{9}{10} = 90\%$
- NPV = $\frac{99}{100} = 99\%$

**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 Positive</td>
<td>90</td>
</tr>
<tr>
<td>New Test Negative</td>
<td>10</td>
</tr>
</tbody>
</table>

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

**ROC Curves**

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

<table>
<thead>
<tr>
<th>Test Method</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>GFR, creatinine</td>
<td>0.70±0.10</td>
<td>0.70±0.10</td>
</tr>
<tr>
<td>Inulin, clearance</td>
<td>0.70±0.10</td>
<td>0.70±0.10</td>
</tr>
<tr>
<td>Captopril, systolic blood pressure</td>
<td>0.70±0.10</td>
<td>0.70±0.10</td>
</tr>
<tr>
<td>Captopril, diastolic blood pressure</td>
<td>0.70±0.10</td>
<td>0.70±0.10</td>
</tr>
<tr>
<td>Captopril, pulse pressure</td>
<td>0.70±0.10</td>
<td>0.70±0.10</td>
</tr>
<tr>
<td>Captopril, heart rate</td>
<td>0.70±0.10</td>
<td>0.70±0.10</td>
</tr>
<tr>
<td>Captopril, renal function</td>
<td>0.70±0.10</td>
<td>0.70±0.10</td>
</tr>
<tr>
<td>Captopril, urine volume</td>
<td>0.70±0.10</td>
<td>0.70±0.10</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Truth</th>
<th>0</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pos</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Neg</td>
<td>100</td>
<td>50</td>
</tr>
</tbody>
</table>
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>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>1</td>
</tr>
<tr>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
</tbody>
</table>

Next person tests negative.

<table>
<thead>
<tr>
<th>Truth</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>1</td>
</tr>
<tr>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
</tbody>
</table>

Next 4 people test positive.

<table>
<thead>
<tr>
<th>Truth</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>5</td>
</tr>
<tr>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>39</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>

\[
\text{Likelihood Ratios} = \frac{\text{Post-test odds}}{\text{Pre-test odds}}
\]

- \(+LR = \frac{\text{Sensitivity}}{(1 - \text{Specificity})}\)
- \(-LR = \frac{(1 - \text{Sensitivity})}{\text{Specificity}}\)

**Example of nuclear stress test for CAD:**
Sensitivity = 90%, Specificity = 80%

\[+LR = 4.5, -LR = 0.13\]
**Likelihood Ratios**

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

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

> Apply test

<table>
<thead>
<tr>
<th></th>
<th>CAD</th>
<th>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
Likelihood Ratios

\[ (\text{Pre-test odds}) \times \text{LR} = (\text{Post-test odds}) \]

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

\[
\begin{array}{c|c|c}
\text{Test positive} & \text{CAD} & \text{CAD} \\
18 & 16 & \text{PPV: 92%} \\
\text{Test negative} & 2 & 64 & \text{NPV: 89%} \\
\end{array}
\]

Sens: 90% Spec: 80%

\[ +\text{LR: 4.5} \]

\[ -\text{LR: 0.13} \]

Apply test

\[
\begin{array}{c|c|c}
\text{Test positive} & \text{CAD} & \text{CAD} \\
18 & 16 & \text{PPV: 92%} \\
\text{Test negative} & 2 & 64 & \text{NPV: 89%} \\
\end{array}
\]

Sens: 90% Spec: 80%

\[ +\text{LR: 4.5} \]

\[ -\text{LR: 0.13} \]
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

<table>
<thead>
<tr>
<th>Test</th>
<th>CAD</th>
<th>CAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>18</td>
<td>16</td>
</tr>
<tr>
<td>Negative</td>
<td>2</td>
<td>64</td>
</tr>
</tbody>
</table>

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

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
- If the post-test probability will not lead to different management, do not order the test.
  - It’s OK to be uncertain!!

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):
  \[ \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} \]
  where \( \sigma_a^2 \) is between-group variance and \( \sigma_e^2 \) is within-group variance.

- 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?

(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.
Measures of Precision/Repeatability

- 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>

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

Kappa Example

<table>
<thead>
<tr>
<th></th>
<th>Reader 1</th>
<th>Reader 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TB</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>No TB</td>
<td>2</td>
<td>93</td>
</tr>
</tbody>
</table>

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

➤ First, calculate expected agreement

<table>
<thead>
<tr>
<th></th>
<th>Reader 1</th>
<th>Reader 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TB</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>No TB</td>
<td>2</td>
<td>85</td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>88</td>
</tr>
</tbody>
</table>

Kappa Example

➤ Calculate expected agreement

<table>
<thead>
<tr>
<th></th>
<th>Reader 1</th>
<th>Reader 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TB</td>
<td>0.12*0.15 = 0.018</td>
<td>0.88*0.15 = 0.132</td>
</tr>
<tr>
<td>No TB</td>
<td>0.12*0.85 = 0.102</td>
<td>0.88*0.85 = 0.748</td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>88</td>
</tr>
</tbody>
</table>

Kappa Example

➤ Multiply by the total

➤ Expected agreement = 74.8 + 1.8 = 76.6/100

<table>
<thead>
<tr>
<th></th>
<th>Reader 1</th>
<th>Reader 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TB</td>
<td>1.8</td>
<td>15</td>
</tr>
<tr>
<td>No TB</td>
<td>10.2</td>
<td>85</td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>88</td>
</tr>
</tbody>
</table>
Kappa Example

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

$$\text{Kappa} = \frac{\text{observed} - \text{expected}}{1 - \text{expected}}$$
$$= \frac{(0.93 - 0.766)/(1 - 0.766)}{0.70}$$

“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?
    - Know your pre-test probability, LRs, and management thresholds!
- **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 evaluation of diagnostic tests

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

- 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?
“Phase I-IV” for Diagnostics?

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

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

**Phase III**
Randomized controlled trial?
Diagnoses 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>Tests to be performed</th>
<th>Total</th>
<th>95% CI</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Positive Predictive Value</th>
<th>Negative Predictive Value</th>
<th>Positive Likelihood Ratio</th>
<th>Negative Likelihood Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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.
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?
Proof From Clinical Trials

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, LTID 2013; 5:641)
    - Slow
    - Results not trusted
    - Empiric therapy had already been initiated
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.

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.
A Recent Example

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

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 incremental value.
  - What is the intended benefit of the test to patients and society?

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 incremental value.
  - Consider effects on decision-making, patient outcomes, and society.