Welcome and Recap from Evidence Boot Camp I

Vivian Coates
Vice President, Health Technology Assessment, ECRI Institute
ECRI Organizational Experience

- Nonprofit health services research institute with 46 years’ experience in laboratory evaluation of healthcare technology, devices and equipment
- 25 years’ experience in health technology assessment, comparative effectiveness research and forecasting of drugs, devices, procedures, including diagnostics
- Worldwide clients include: thousands of hospitals, health plans, national and regional governmental agencies
- For Agency for Healthcare Research and Quality (AHRQ): Evidence-based Practice Center, Patient Safety Organization, National Guideline Clearinghouse, National Quality Measures Clearinghouse, AHRQ Healthcare Horizon Scanning System
integrity

neither ecri nor any of its staff has a financial interest in the sale of any medical technology. ecri and its staff accept no royalties, gifts, finder’s fees, or commissions from the medical device or pharmaceutical industries and are not permitted to own stock in or undertake consulting work for such industries.

adhering to our conflict-of-interest rules - but also interacting with manufacturers and labs - are part of our culture.
HEALTH TECHNOLOGY ASSESSMENT EVIDENCE BOOTCAMP

CME/CEU Information

**Physicians:**
ECRI Institute designates this live activity for a maximum of **7.0 AMA PRA Category 1 credits™**.
All faculty members involved in this July 14-15, 2015 live activity have disclosed that there are no conflicts or financial affiliations.

**Nurses:**
This activity has been approved for up to 8.5 **California State** Nursing contact hours by the provider, Debora Simmons, who is approved by the California Board of Registered Nursing, Provider Number CEP 13677.

Details can be found in the credit handout, along with instructions for obtaining credit.

*All faculty members involved in this July 14-15, 2015, live educational event have disclosed in writing that they do not have any relevant conflicts or financial affiliations.*

In your packet, you should have received an evaluation form. We encourage you to fill out this form so we can make any necessary adjustments for future events.
Recap of Boot Camp I

- Overview of Health Technology Assessment Information Service (HTAIS) processes
- What constitutes a good evidence review?
- Interpreting the evidence
- Rapid reviews: opportunities and challenges
- Dealing with no evidence, partial evidence, and bad evidence
- Future directions for HTA
Health Technology Assessment Activity Scope

- Horizon scans & forecasts
- Rapid reviews
- HTAs with small evidence base
- Full-scale HTAs with meta-analysis
What Constitutes a Good Evidence Review – a Rigorous Literature Search

▶ A systematic review that misses critical publications may provide misleading results

▶ The literature search is an integral part of the systematic review process. It should be subject to the same scientific rigor as every other portion of the review.
What Constitutes a Good Evidence Review – Critiquing the Evidence

- Clinical research is easy in principle, difficult in practice.
- Proper comparison groups are essential to evaluating treatment effects.
- Assessing study quality is important in rating evidence for the eventual formation of treatment recommendations.
- Sources of bias, systematic error that can influence results, must be considered in assessing study quality.
What Constitutes a Good Evidence Review – Assessing Publication Bias and Other Types of Reporting Bias

- Publication bias is the selective publication of data
- This makes one suspect the accuracy of **published** data
- Evidence reviewers can be misled
- Therefore the users of such reviews can also be misled
- Detection is possible (e.g., clinicaltrials.gov, funnel plots)
- Need to downgrade strength of evidence rating, or estimate the impact using trim-and-fill

- Other types of reporting bias:
  - Selective outcome reporting
  - Selective analysis reporting
Rapid Reviews: Opportunities and Challenges

- Read “rapid reviews” carefully – what decisions were made to make the review “rapid”?
- Know how much uncertainty you can tolerate in your decision making.
- Recognize that short cuts on assessing the quality of the literature may introduce important bias.
- Be prepared to revisit decisions based on rapid reviews.
- New research on methods for creating reliable but more rapid reviews is in the works.
Dealing with No Evidence, Partial Evidence, and Bad Evidence

► “Gold Standard” evidence comes from well-designed RCTs, but trials may not exist that address your needs.

► In a time of evidence-based medicine, people still need to make decisions with little or no evidence

► What do you do if you have no “useful findings”? 
  ■ Use what limited information you do have
  ■ Use reasonable judgements about similar technologies
  ■ Use information from non-RCTs; recognize limitations from this evidence

► Remember: local factors in your setting may override what the evidence might suggest
  ■ The best evidence available is not helpful to you if your setting lacks the resource (e.g. Different imaging equipment, specific expert personnel, etc.)
Future Directions for Health Technology Assessment

- Impact of Patient Centered Outcomes Research (PCOR) and Comparative Clinical Effectiveness
- Role of AHRQ Healthcare Horizon Scanning System in Priority Setting for CER
- Use of Electronic Clinical Data
- Challenge of Genetic Tests
- For 2015: Increasing Importance of Value Analysis in HTA
Future Directions for Health Technology Assessment

- **Patient Centered Outcomes** need to be part of the entire drug and device development life cycle - don’t wait until the postmarket phase.

- **AHRQ Healthcare Horizon Scanning System** - inventory of innovations that address an unmet need and have the highest potential for impact.

- **Electronic Clinical Data** (“Big Data”) - subject to bias from many causes: need to assess risk of bias and exclude data at high risk of bias.

- **Challenge of Genetic Tests** - tests without clinical utility do not lead to improved outcomes but could impose unnecessary burdens on patients and society.

- **Value Analysis** - demonstrating value means providing evidence of superior comparative effectiveness and cost effectiveness, utilizing patient centered outcomes and a systematic process that engages all clinical stakeholders.
Diagnostic Technologies and Genetic Tests
July 14–15, 2015

Introduction to Boot Camp II

Jonathan R. Treadwell, Ph.D., Associate Director of the Evidence-based Practice Center, Senior Research Analyst
EVIDENCE BOOT CAMP II

Diagnostic Technologies and Genetic Tests
July 14–15, 2015

Evaluation Frameworks to Guide Analysis of Diagnostic Tests

Karen Schoelles MD, SM
Director, Evidence-based Practice Center and Health Technology Assessment Consulting
Project Director, AHRQ Healthcare Horizon Scanning System

©2015 ECRI INSTITUTE
Overview

- Why is appropriate use of diagnostic tests so challenging?
- The Prequel – vocabulary and concepts for diagnostic testing
- The 30,000-foot view - How did we get to our current methods of evaluating diagnostic tests?
Why aren’t diagnostic tests getting the respect they deserve?

[We] have the ironic situation in which important and painstakingly developed knowledge often is applied haphazardly and anecdotally. Such a situation, which is not acceptable in the basic sciences or in drug therapy, also should not be acceptable in clinical applications of diagnostic technology.

J. Sanford (Sandy) Schwartz, IOM, 1985
Why are diagnostic tests so easy to misuse?


A test that perfectly discriminates
Healthy 95%

2.5% False Positives  |  Reference Range  |  2.5% False Positives
<table>
<thead>
<tr>
<th>Index test result</th>
<th>Reference test results (&quot;truth&quot;)</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Disease positive</td>
<td>Disease negative</td>
</tr>
<tr>
<td>Index Test positive</td>
<td>True Positive (TP)</td>
<td>False positive (FP)</td>
</tr>
<tr>
<td>Index Test negative</td>
<td>False Negative (FN)</td>
<td>True Negative (TN)</td>
</tr>
<tr>
<td>Totals</td>
<td>(prevalence of disease) X (total population) = total with disease =TP+FN</td>
<td>(1-prevalence of disease) X (total population) = total without disease =TN+FP</td>
</tr>
<tr>
<td>Index test result</td>
<td>Reference test results (&quot;truth&quot;)</td>
<td>Totals</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td></td>
<td>Disease positive</td>
<td></td>
</tr>
<tr>
<td>Index Test positive</td>
<td>TP=</td>
<td></td>
</tr>
<tr>
<td>Index Test negative</td>
<td>FN=</td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>TP+FN=</td>
<td></td>
</tr>
</tbody>
</table>

| Disease negative |        |
| Index Test positive | FP= | TP+FP= |
| Index Test negative | TN= | TN+FN= |
| Totals            | TN+FP= | Total population= TP+FN+FP+TN= 1000 |

Sensitivity = 95%

Specificity = 90%
<table>
<thead>
<tr>
<th>Index test result</th>
<th>Reference test results (&quot;truth&quot;)</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Disease positive</td>
<td></td>
</tr>
<tr>
<td>Index Test positive</td>
<td>TP=</td>
<td>FP=</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TP+FP=</td>
</tr>
<tr>
<td>Index Test</td>
<td>FN=</td>
<td>TN=</td>
</tr>
<tr>
<td>negative</td>
<td></td>
<td>TN+FN=</td>
</tr>
<tr>
<td>Totals</td>
<td>TP+FN=</td>
<td>TN+FP=</td>
</tr>
<tr>
<td></td>
<td>Total population=</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TP+FN+FP+TN=</td>
<td>1000</td>
</tr>
</tbody>
</table>

Sensitivity = 95%

Specificity = 90%

Prevalence = 0.1%
<table>
<thead>
<tr>
<th>Index test result</th>
<th>Reference test results (&quot;truth&quot;)</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Disease positive</td>
<td>Disease negative</td>
</tr>
<tr>
<td>Index Test positive</td>
<td>TP=0.95</td>
<td>FP=100</td>
</tr>
<tr>
<td>Index Test negative</td>
<td>FN=0.05</td>
<td>TN=899</td>
</tr>
<tr>
<td>Totals</td>
<td>TP+FN=1</td>
<td>TN+FP=999</td>
</tr>
</tbody>
</table>

Sensitivity = 95%

Specificity = 90%

Prevalence = 0.1%
## Predictive values (post-test probabilities) of tests vary with prevalence

<table>
<thead>
<tr>
<th></th>
<th>Disease positive</th>
<th>Disease negative</th>
<th>Totals</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Index Test</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>positive</td>
<td>95.00</td>
<td>90.00</td>
<td>185.00</td>
<td>95%</td>
<td>90%</td>
<td>10.00%</td>
</tr>
<tr>
<td>negative</td>
<td>5.00</td>
<td>810.00</td>
<td>815.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>100.00</td>
<td>900.00</td>
<td>1000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Predictive Values

- **Positive predictive value** (PPV) = the number of people with a positive test who actually have disease divided by all who have a positive test:
  \[
  TP \div (TP+FP)
  \]

- **Negative predictive value** (NPV) = the number of people with a negative test who actually do not have disease divided by all who have a negative test:
  \[
  TN \div (FN+TN)
  \]

<table>
<thead>
<tr>
<th>Index Test positive</th>
<th>Disease positive</th>
<th>Disease negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index Test negative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TP</td>
<td></td>
<td>FP</td>
</tr>
<tr>
<td>FN</td>
<td></td>
<td>TN</td>
</tr>
</tbody>
</table>
Sensitivity = 99%; Specificity 99%

Prevalence = Pre-test probability

Probability Notation

- **Sensitivity** = \( P(T^+|D^+) \) = the probability of testing positive given that you have the disease

- **Specificity** = \( P(T^-|D^-) \) = the probability of testing negative given that you don’t have the disease
Probability Notation

- **Predictive Value Positive** = $P(D^+|T^+)$ = the probability of having the disease given that you test positive

- **Predictive Value Negative** = $P(D^-|T^-)$ = the probability of not having the disease given that you test negative
Bayes Theorem

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

• Or – the probability of having the disease given a positive test equals
• The probability of having a positive test when the disease is present (i.e., sensitivity) multiplied by the probability of disease (i.e., prevalence)
• Divided by that same quantity plus the probability of having a positive test when the disease is absent (i.e., false positive) multiplied by the probability of not having the disease (1-prevalence)
### Updating Probabilities: “Benign” Finding on MRI

<table>
<thead>
<tr>
<th>Pre-test Probability of the Lesion Being Malignant</th>
<th>Post-test Probability of the Lesion Being Malignant Despite a Finding of “Benign” on the MRI Exam</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesions in General</td>
<td>Lesions with Microcalcifications</td>
</tr>
<tr>
<td>1%</td>
<td>0% (0 to 0%)</td>
</tr>
<tr>
<td>5%</td>
<td>1% (0 to 1%)</td>
</tr>
<tr>
<td>10%</td>
<td>1% (1 to 2%)</td>
</tr>
<tr>
<td>20%</td>
<td>3% (2 to 4%)</td>
</tr>
<tr>
<td>30%</td>
<td>5% (3 to 6%)</td>
</tr>
<tr>
<td>40%</td>
<td>7% (5 to 9%)</td>
</tr>
<tr>
<td>50%</td>
<td>10% (7 to 13%)</td>
</tr>
<tr>
<td>60%</td>
<td>14% (11 to 18%)</td>
</tr>
<tr>
<td>70%</td>
<td>20% (16 to 26%)</td>
</tr>
<tr>
<td>80%</td>
<td>31% (24 to 38%)</td>
</tr>
<tr>
<td>90%</td>
<td>50% (42 to 57%)</td>
</tr>
</tbody>
</table>

Likelihood Ratios

- Probability of getting a test result in patients having the condition divided by the probability of getting that test result when they don’t

$$\Pr(T^+ \mid D^+) \div \Pr(T^+ \mid D^-)$$

- Positive likelihood ratio = sensitivity / (1-specificity) or 
  
  $$\frac{TP}{TP+FN} \div \frac{FP}{FP+TN}$$

  ■ the higher the result, the better the test is in ruling in the disease

- Negative likelihood ratio = (1-sensitivity) / specificity or 
  
  $$\frac{FN}{TP+FN} \div \frac{TN}{FP+TN}$$

  ■ the lower the result, the better the test is in ruling out the disease
Fagan TJ

Interactive version:
http://www.cebm.net/
Likelihood Ratios

- $LR = 1$  No new information
- $LR > 1$  Argues in favor of disease
- $LR = \infty$  Disease is certain
- $LR < 1$  Argues against disease
- $LR = 0$  Disease excluded
## PIOPED Study – V/Q Scanning vs. Angiography or Clinical Followup (1 yr)

<table>
<thead>
<tr>
<th>Scan Results</th>
<th>PE Present</th>
<th>PE Absent</th>
<th>Likelihood Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Probability</td>
<td>102</td>
<td>14</td>
<td>18.3</td>
</tr>
<tr>
<td>Intermediate Probability</td>
<td>105</td>
<td>217</td>
<td>1.20</td>
</tr>
<tr>
<td>Low Probability</td>
<td>39</td>
<td>273</td>
<td>0.36</td>
</tr>
<tr>
<td>Normal/near normal</td>
<td>5</td>
<td>126</td>
<td>0.10</td>
</tr>
<tr>
<td>Total</td>
<td>251</td>
<td>630</td>
<td></td>
</tr>
</tbody>
</table>

*JAMA. 1990;263:2753-2759*
Test Characteristics – Core-needle biopsy for breast abnormalities

<table>
<thead>
<tr>
<th>Test Results</th>
<th>Disease</th>
<th>Present</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>True positives (TP)</td>
<td>False positives (FP)</td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>False negatives (FN)</td>
<td>True negatives (TN)</td>
<td></td>
</tr>
</tbody>
</table>

- Likelihood ratio – useful for comparing tests
  - Positive likelihood ratio = \( \frac{TP}{(TP+FN)} \div \frac{FP}{(FP+TN)} \)
  - Negative likelihood ratio = \( \frac{FN}{(TP+FN)} \div \frac{TN}{(FP+TN)} \)

- For this evaluation, not missing a cancer was considered the most important outcome, reflected by:
  - sensitivity, negative predictive value and negative likelihood ratio

©2015 ECRI INSTITUTE

ECRI Institute
The Discipline of Science. The Integrity of Independence.
## Summary of key accuracy findings – hypothetical population

<table>
<thead>
<tr>
<th>Type of biopsy</th>
<th>Number of missed cancers expected for every 1,000 biopsies</th>
<th>Risk of malignancy following a “benign” test result</th>
<th>Number of malignancies expected per 1,000 biopsy diagnoses of “high risk” lesion</th>
<th>Number of invasive cancers expected per 1,000 biopsy diagnoses of DCIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open surgical⁴</td>
<td>3 to 6</td>
<td>0 to 1%</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Freehand automated gun</td>
<td>24 to 73</td>
<td>3.4 to 10%</td>
<td>Insufficient data to estimate</td>
<td></td>
</tr>
<tr>
<td>US guidance automated gun</td>
<td>6 to 9</td>
<td>1 to 2%</td>
<td>234 to 359</td>
<td>271 to 450</td>
</tr>
<tr>
<td>Stereotactic guidance automated gun</td>
<td>3 to 13</td>
<td>0.5 to 2%</td>
<td>357 to 517</td>
<td>180 to 321</td>
</tr>
</tbody>
</table>
## Summary of key accuracy findings

<table>
<thead>
<tr>
<th>Type of biopsy</th>
<th>Number of missed cancers expected for every 1,000 biopsies</th>
<th>Risk of malignancy following a “benign” test result</th>
<th>Number of malignancies expected per 1,000 biopsy diagnoses of “high risk” lesion</th>
<th>Number of invasive cancers expected per 1,000 biopsy diagnoses of DCIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open surgical</td>
<td>3 to 6</td>
<td>0 to 1%</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MRI guidance automated gun</td>
<td>Insufficient data to estimate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US guidance vacuum-assisted</td>
<td>2 to 56</td>
<td>0.3 to 8%</td>
<td>Insufficient data to estimate</td>
<td></td>
</tr>
<tr>
<td>Stereotactic guidance vacuum-assisted</td>
<td>1 to 6</td>
<td>0.1 to 1%</td>
<td>177 to 264</td>
<td>111 to 151</td>
</tr>
</tbody>
</table>
History lesson – the evolution of diagnostic test assessment

➢ As early as the 40’s, the terms “sensitivity” and “specificity” were being used in the medical literature

■ Sensitivity – the probability of a correct diagnosis in people with the disease

■ Specificity – the probability of a correct [non]diagnosis in people without the disease
1959: Robert Ledley and Lee Lusted explored the process of medical diagnosis using probability theory and game theory

- Bayes’ theorem applied to diagnostic problems
- Expected value theory to the choice of treatments given multiple diagnostic possibilities
- Game theory to create an optimal decision making strategy

The logical aspect of the medical diagnosis problem is to determine the diseases $f$ such that if medical knowledge $E$ is known, then:

If the patient presents symptoms $G$, he has diseases $f$:

$$ E \rightarrow (G \rightarrow f) $$

Fig. 12. Cards notched to indicate columns of logical basis.
Fig. 13. Sorting the cards.
Fig. 14. For a patient presenting symptom complex $C^1$, the conditional probabilities for diagnoses $C_2$ and $C_3$ are read from the respective thicknesses of the decks as $P(C_2|C^1) = 5/(5 + 2)$ and $P(C_3|C^1) = 2/(5 + 2)$. 
1970’s

- Medicare and Medicaid
- Medical costs rising
- Nixon’s managed care proposal
- Computerized tomography becomes available
- Physicians react to CT images:
Response to concerns about health care spending

- 1971: American College of Radiology Efficacy Studies Committee
- Evaluated IVP efficacy
  - **Outcome efficacy**: Patient outcomes: Was the patient better off as a result of the procedure having been performed?
  - **Therapeutic efficacy**: To what extent did the test change patient management?
  - **Diagnostic efficacy**: To what degree did the X-ray result influence the clinician’s diagnostic thinking?

Center for the Analysis of Health Practices at the Harvard School of Public Health - 1978

1. Technical performance
2. Clinical efficacy
3. Resource costs, charges and efficiency
4. Safety
5. Acceptability to patients, physicians, and other users
6. Research benefits for the future
7. Larger effects on the organization of health services
8. Larger effects on society.

Fryback and Thornbury Hierarchical Model of Efficacy - 1991 – Expanding Our Vantage Points

Level 1: Technical accuracy

In the *laboratory setting*, does the test measure what it purports to measure?

Level 2: Diagnostic accuracy

What are the diagnostic test characteristics of the test (e.g. sensitivity, specificity)? 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?

Level 3: Diagnostic thinking

Does the diagnostic test help clinicians come to a diagnosis? Does the test change clinician’s pre-test estimate of the probability of a specific disease? (impact on the clinician)

Fryback and Thornbury Hierarchical Model of Efficacy - 1991

Level 4: Therapeutic efficacy

Does the diagnostic test aid in planning treatment? Does the diagnostic test change or cancel planned treatments?

Level 5. Patient outcomes

Do patients benefit from the use of the test? Do patients who undergo this diagnostic test fare better than similar patients who are not tested?

Level 6. Societal efficacy

Cost-benefit and cost-effectiveness

## Kent and Larson – Organizational Framework

<table>
<thead>
<tr>
<th>Quality of Research Methods</th>
<th>Technical Capacity</th>
<th>Diagnostic Accuracy</th>
<th>Diagnostic Impacts</th>
<th>Therapeutic Impacts</th>
<th>Patient Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>&gt;20</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>Many</td>
<td>11</td>
<td>6</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>Many</td>
<td>54 studies and claims</td>
<td>48 studies and claims</td>
<td>No studies, many claims</td>
<td>Claims</td>
</tr>
</tbody>
</table>

Mackenzie and Dixon – Donabedian’s Structure-Process-Outcomes Framework

- **Structure:** Do clinicians have access to CT? What is the equipment’s technical capability? Is it appropriately located, equipped and staffed?

- **Process:** Do clinicians and hospitals make appropriate use of CT?

- **Outcomes:** Do applications of the imaging improve patients’ health status

Drug development framework applied to diagnostics

- Phase 1: Studies of the analytical precision, accuracy, sensitivity, and specificity of a laboratory test
- Phase 2: Studies examining the usual range of results in healthy persons, or studies comparing the usual range in healthy persons to that in persons with a variety of disease states
- Phase 3: Prospective, blinded, controlled studies for answering a specific clinical question, with use of an independent method of answering the question in all patients.

Muin Khoury (CDC) – Research translation model

► Phase 1 (T1) studies move a basic genome-based discovery into a candidate health application (e.g., genetic test)

► Phase 2 (T2) studies assess the validity and utility of a developed genomic application for health practice, which leads to development of evidence-based guidelines

► Phase 3 (T3) research examines the movement of guidelines into practice

► Phase 4 (T4) studies evaluate the “real-world health outcomes” of genomic applications in practice

# NCI’s Early Detection Research Network (EDRN): Phases of Cancer Biomarker Development

<table>
<thead>
<tr>
<th>Phase</th>
<th>Objective</th>
<th>Study design</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Preclinical exploratory</td>
<td>Identification of new directions</td>
<td>Convenience sample case-control</td>
</tr>
<tr>
<td>2. Clinical assay and validation</td>
<td>Detection of known disease states</td>
<td>Population-based case-control</td>
</tr>
<tr>
<td>3. Retrospective longitudinal</td>
<td>Define a positive test and determine whether disease can be detected in preclinical stage</td>
<td>Nested case-control within a population cohort</td>
</tr>
<tr>
<td>4. Prospective screening</td>
<td>Determine characteristics of detected disease and false positive rate</td>
<td>Cross-sectional cohort</td>
</tr>
<tr>
<td>5. Cancer control</td>
<td>Determine population-level reduction in cancer burden</td>
<td>Randomized trial</td>
</tr>
</tbody>
</table>

Analytic Framework and PICO(TS)

- Population of interest
- Intervention being assessed
- Comparator
- Outcome
- Time point
- Setting
EGAPP draft framework: disease screening
= USPSTF framework

- Individuals at risk
  - Genetic testing
    - Early detection of target condition
    - Adverse effects of genetic testing
    - Adverse effects of treatment/other interventions

- Treatment
  - Intermediate outcome
    - Association
      - Mortality, morbidity, and other outcomes
Population representative of clinical practice

Diagnostic Accuracy Efficacy
- Sensitivity in Disease Positive Cohort
- Sensitivity/Specificity in Typical Clinical Population
- Meta-analysis of accuracy studies

Diagnostic Thinking Efficacy
- Change in diagnostic thinking

Reference Standard

New Test

Test-related Harms

No test

Diagnostic Thinking Favors A

Diagnostic Thinking Favors B

Other scenarios:
- Test as add-on to reference test
- Test as triage prior to more invasive reference test

Therapeutic Efficacy (Change in Management)
- Change in choice of next intervention
- Intervention A applied to test + patients
- Intervention B applied to test - patients

Societal Efficacy
- Cost-effectiveness
- Population health
- Legal implications
- Ethical implications

Technical Efficacy
- Feasibility
- Analytic Validity
- Algorithm development

Patient Outcome Efficacy

Diagnostic Accuracy Efficacy
- Sensitivity in Disease Positive Cohort
- Sensitivity/Specificity in Typical Clinical Population
- Meta-analysis of accuracy studies

Test development and evaluation with multiple feedback loops
Fig 2 Simplified test-treatment pathway showing each component of a patient’s management that can affect health outcomes.

Lavinia Ferrante di Ruffano et al. BMJ 2012;344:bmj.e686
A simple question can reveal as much as a test.

"WHAT ARE OUR GOALS FOR TODAY?"

Ask your patients about their health priorities at each visit. When you do, both you and your patient can make the most out of the time you have together, and they'll feel more invested in their own care. Not only does that improve efficiencies, but it also helps improve health outcomes.

For tools and tips to share with your patients, visit www.ahrq.gov/questions

ECRI Ins
The Discipline of Science. The Integrity

©2015 ECRI INSTITUTE
Optimizing Searches for Evidence on Diagnostic Tests

Eileen Erinoff, MSLIS
Director, Health Technology Assessment and Evidence-based Practice Information Center, ECRI Institute
Optimizing Searches for Evidence on Diagnostic Tests

- Review information retrieval processes
- Understand how to search for evidence on diagnostics
- Understand how searches for diagnostic-related evidence differ from other searches
Types of Searches

- Balancing precision vs. recall
  - Comprehensive
  - Targeted
  - Ready Reference
Types of Searches – precision vs. recall

- Comprehensive – systematic review
  - “Shotgun”
  - Very sensitive
  - Maximizes recall
Types of Searches – precision vs recall

- Targeted – rapid turn-around review
  - “Rifle” – very precise search
  - Very specific
  - Maximizes precision
Types of Searches

- Ready Reference
  - Any good answer will do
    - What is the incidence of diabetes in the U.S.?
    - Can you find me a recent review on subject xyz?
Scientific Approach to Information Retrieval

- Unbiased and systematic data collection
- Transparency
- Reproducibility
Search Protocol

“A search protocol is an explicit, structured procedure for tackling the task of searching. It sets out the sources to be searched, providing a logical set of steps to work through in the course of the search in a detailed and transparent way, so that it is possible to run the search and get the same results at a later time”

Bidwell & Jensen, 2000

Resources

- Bibliographic databases
  - Medline
  - Embase
  - PsycINFO
  - CINAHL

- Hand-searches of journals and reference lists

- Gray literature
  - Ongoing research
  - National Guideline Clearinghouse
  - Internet searches
  - Regulatory data
  - Reimbursement data
  - Cost/charge data
  - Statistics: incidence, mortality, prevalence, vital
  - Technology Assessments
Core Bibliographic Resources

- MEDLINE
- EMBASE
- The Cochrane Library
- National Guideline Clearinghouse
How and where does ECRI find Gray Literature?

- Internet searches
- Mining specialty organization sites
- Conference abstracts
- Press releases
- Ongoing clinical trials
Searching the Gray Literature

- Requires a different approach
- Much more dependent upon keywords
- Determine a priori how much time you will spend on this part of the process
  - The most difficult thing to learn is knowing when to stop
Important elements of a search strategy

- Key concepts
- Controlled vocabularies
- Text words (a.k.a. “keywords”)
- Limiters
- Logic used to combine concepts
Key concepts used during search process

P population
I intervention
C comparators
O outcomes
T time
S settings
Controlled vocabularies

- Categorize concepts
- Standardize concepts
- Establish relationships between concepts
- Facilitate information retrieval

©2015 ECRI INSTITUTE
Controlled Vocabularies

- Vocabulary terms are assigned to citations by professional indexers (subjective)
- Several terms are selected to represent the main concept of the article
- Some concepts, such as age group, language of publication, and publication type are applied to all indexed articles
Controlled vocabularies

- Many controlled vocabularies are hierarchies
- In databases that support “explosion” searching on a broader term will automatically include all narrower terms associated with that concept
- PubMed automatically “explodes” MeSH terms
- Use the rubric [mh:noexp] to limit to the broader term only
Diagnosis - MeSH

- Diagnosis [E01]
  - Delayed Diagnosis [E01.110]
  - Diagnosis, Computer-Assisted [E01.158] +
  - Diagnosis, Differential [E01.171]
  - Diagnosis, Dual (Psychiatry) [E01.190]
  - Diagnostic Errors [E01.354] +
  - Diagnostic Techniques and Procedures [E01.370] +
  - Early Diagnosis [E01.390] +
  - Incidental Findings [E01.410]
  - Prodromal Symptoms [E01.599]
  - Prognosis [E01.789] +
Diagnostic Techniques and Procedures [E01.370]

Age Determination by Skeleton [E01.370.049]

Autopsy [E01.370.060]

Breath Tests [E01.370.100]

Clinical Laboratory Techniques [E01.370.225] +

Diagnostic Imaging [E01.370.350] +

Diagnostic Self Evaluation [E01.370.360]

Diagnostic Techniques, Cardiovascular [E01.370.370] +

Diagnostic Techniques, Digestive System [E01.370.372] +

Diagnostic Techniques, Endocrine [E01.370.374] +

Diagnostic Techniques, Neurological [E01.370.376] +

Diagnostic Techniques, Obstetrical and Gynecological [E01.370.378] +

Diagnostic Techniques, Ophthalmological [E01.370.380] +

Diagnostic Techniques, Otological [E01.370.382] +

Diagnostic Techniques, Radioisotope [E01.370.384] +

Diagnostic Techniques, Respiratory System [E01.370.386] +

Diagnostic Techniques, Surgical [E01.370.388] +

Diagnostic Techniques, Urological [E01.370.390] +

Diagnostic Tests, Routine [E01.370.395]

Disability Evaluation [E01.370.400] +

Electrodiagnosis [E01.370.405] +

Insufflation [E01.370.450]

Kymography [E01.370.475] +

Mass Screening [E01.370.500] +

Medical History Taking [E01.370.510] +

Monitoring, Physiologic [E01.370.520] +

Myography [E01.370.530] +

Photoacoustic Techniques [E01.370.565]

Physical Examination [E01.370.600] +

Plethysmography [E01.370.610] +

Premarital Examinations [E01.370.620]

Psychophysics [E01.370.685] +

Sex Determination Analysis [E01.370.701] +

Speech Production Measurement [E01.370.760] +

Symptom Assessment [E01.370.872]

Visual Analog Scale [E01.370.928]
Controlled vocabularies

- Diagnostic Techniques, Obstetrical and Gynecological
  - Prenatal Diagnosis
    - Amniocentesis
    - Chorionic Villi Sampling
    - Fetoscopy
    - Maternal Serum Screening Tests
    - Ultrasonography, Prenatal
Controlled Vocabularies - Subheadings

- Components of controlled vocabularies that allow searchers to further refine an aspect of a search
- Also called Qualifiers
- Can be attached to a term or used independently ("floated")
  - DNA/blood[mh] – attached
    - Used for the presence or analysis of substances in the blood; also for examination of, or changes in, the blood in disease states. It excludes serodiagnosis, for which the subheading "diagnosis" is used, and serology, for which "immunology" is used.
  - Diagnostic use[sh] – floated
    - du[sh]
Useful subheadings for searches of diagnostic topics

- Analysis
- Blood
- Cerebrospinal fluid
- Diagnosis
- Diagnostic Use
- Genetics
- Pathology
- Radiography
- Radionuclide imaging
- Ultrasonography
- Urine
Controlled Vocabularies

- Limiters
  - Method of further refining the scope of a search
  - Common limiters:
    - Age
    - Sex
    - Date of Publication
    - Publication Type
**Fields**

- Within databases information is stored in separate fields or tables
- Searches can be limited to these individual elements
  - Available from PubMed Advanced Search
- **Examples:**
  - Title
  - Author
  - Abstract
  - Descriptors (controlled vocabulary)
  - Publication type
Controlled Vocabularies - Caveats

► Journals that meet NLM’s inclusion criteria have to reach their third year of publication before they are indexed in PubMed

► Not all journals indexed in PubMed are indexed comprehensively

► Check when your terms were added to the vocabulary
  ■ Using only a recently added term de facto limits your search to the date the term was added to the vocabulary
Truncation characters

One method of increasing retrieval is removing letters from the end of a word and replacing them with a “wildcard” or truncation character

- Decis* will retrieve decision, decisions, decisive, etc.
- Decide* will retrieve decide, decides, decided, etc.
- Deci* is too short – it will retrieve decimal, decimate, and many other words you may not wish to include in your search
- Examples of truncation characters:

  ?  *  $
Where should you truncate the word diagnosis?

- Diagnos*
  - Diagnose
  - Diagnosed
  - Diagnosis
  - Diagnoses
  - Diagnostic
What are useful synonyms for diagnosis and where should you truncate them?

- Detect*
- Identif*
Boolean logic

- **AND** operator narrows the scope of the search
- **OR** operator broadens the scope of the search
- **NOT** operator narrows the scope of the search
- Many search engines support nested logic. For example: \(((a \text{ OR } b) \text{ AND } (c \text{ OR } d)) \text{ NOT } e\)
Boolean logic – Venn Diagrams

Composed Tomography (CT)

CT AND Lung Cancer

Lung Cancer
Boolean Logic

▶ Google

- Supports Boolean logic
  ▶ “AND” is assumed
  ▶ Use OR to expand the scope of the search
- Limit by domain
  ▶ site:.gov, site:www.fda.gov, site:.org, site:.edu
- Supports proximity operators
  ▶ AROUND
    □ ((pulmonary OR lung) AROUND(2)(nodule OR nodules)) (“CT” OR “CAT” OR “computed tomography”) site:.edu
Constructing search strategies

- Conduct search and evaluate results
- Revise search based on retrieval
  - Review the indexing of relevant citations to see which controlled vocabulary terms had been used to represent the concepts of interest and add them to the strategy
  - Note whether there are trends in the types and numbers of irrelevant citations retrieved by the search strategy
Search filters

What are search filters?

- Also called hedges
- Preconstructed search strategies that can be used to identify the same concept in multiple searches
Examples of diagnostic test methods

- Biopsy
- Clinical laboratory
  - Blood tests
  - Urinalysis
- Endoscopic procedures
- Genetic/Molecular testing
- Imaging
- Pulmonary function studies
- Urodynamic studies
How can we frame a systematic approach? What do diagnostic tests have in common?

- Accuracy
- Precision
- Prognostic ability
- Sensitivity and specificity
- Validity
  - **Analytic validity** – test’s ability to accurately and reliably measure the properties or characteristics it is intended to measure
  - **Clinical validity** - how well a test predicts the presence or absence of a clinical condition
  - **Clinical utility** – test’s usefulness in affecting patient outcomes or clinical decisions
Techniques

Include technical terms and keywords that frequently appear in diagnostic studies in your search strategy:

- Accuracy
- False negative, false positive, true negative, true positive
- Likelihood
- Maximum likelihood method
- Positive predictive value (PPV)
- Precision
- Prediction and forecasting
- Receiver operating characteristic
- ROC curve
- Sensitivity and specificity
It is important to use all known variants of a test name, as in the examples below that refer to hematocrit:

- Abbreviations (Hct, Crit, PCV)
- Generic names (hematocrit, packed cell volume)
- Proprietary names (e.g., LighTouch® HCT)
- International terms/spellings (haematocrit)
- Analyte plus subheadings
How do searches for diagnostic topics differ from other types of searches?

Unique challenges

- Indexing for diagnostic topics can be inconsistent
  - Search for both the disease with general diagnosis terms and the disease with the specific intervention
- Diagnostic methods are frequently mentioned in the methods section of the abstract even when they are not the focus of the article. Example: Is the article focusing on CT as a means of diagnosing lung cancer or does it mention the technology in passing in the methods section?
  - Use major heading, keyword in title or diagnosis subheadings

Less focus on study type

- Far fewer randomized controlled trials. Observational studies are frequently included in the search protocol.
  - Don’t use restrictive study filters
## Example – Imaging tests for the staging of colorectal cancer

<table>
<thead>
<tr>
<th></th>
<th>Colorectal cancer</th>
<th>exp Colorectal Neoplasms/ or exp colon cancer/ or exp colon tumor/ or exp rectum cancer/ or exp rectum tumor/ or ((Colon$ or colorectal or rect$) adj2 (cancer$ or tumo$ or neoplas$ or carcinoma$ or adenocarcinoma$)).ti,ab.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Staging</td>
<td>neoplasm staging/ or cancer staging/ or (stag$ or restag$ or re-stag$).ti,ab.</td>
</tr>
<tr>
<td>3</td>
<td>Imaging</td>
<td>exp Diagnostic Imaging/ or exp Tomography, Emission-Computed/ or exp Tomography, X-Ray Computed/ or exp Magnetic Resonance Imaging/ or exp Ultrasonography/ or Radiography, Thoracic/ or exp computer assisted tomography/ or positron emission tomography/ or multidetector computed tomography/ or exp nuclear magnetic resonance imaging/ or Thorax radiography/ or exp echography/ or computer assisted emission tomography/ or Endoscopy, Gastrointestinal/ or gastrointestinal endoscopy/ or (“computed tomography” or “computerized tomography” or “multidetector computerized tomography” or “magnetic resonance imaging” or “positron emission tomography” or (CT or PET or MRI or TRUS or TUS or ERUS or EUS or MD-CT or x-ray) or ((endorectal or endoscop$ or transrectal or transabdominal) and ultrasound) or imag$).mp</td>
</tr>
<tr>
<td>4</td>
<td>Combine sets</td>
<td>#1 AND #2 AND #3</td>
</tr>
</tbody>
</table>
Example – 2015 Evidence-based Practice Center Technical Brief


- This Technical Brief collects and summarizes information on genetic tests clinically available in the United States to detect genetic markers that predispose to DDs. It also identifies but does not systematically review, existing evidence addressing the tests’ clinical utility. This Brief primarily focuses on patients with idiopathic or unexplained DDs, particularly intellectual disability, global developmental delay, and autism spectrum disorder. Several better-defined DD syndromes, including Angelman syndrome, fragile X syndrome, Prader-Willi syndrome, Rett syndrome, Rubinstein-Taybi syndrome, Smith-Magenis syndrome, velocardiofacial syndrome, and Williams syndrome are also included. Patient-centered health outcomes (e.g. functional or symptomatic improvement) and intermediate outcomes (e.g. changes in clinical decisions or family reproductive decisions, the tests’ diagnostic accuracy and analytic validity) are examined.
Sample concept sheet

<table>
<thead>
<tr>
<th>Genetic Testing</th>
<th>Medline (MeSH)</th>
<th>Embase (EMTREE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>‘chromosome disorders’/exp</td>
<td>‘chromosome aberration’/exp – note: this is a large category that encompasses the entire scope of this report.</td>
</tr>
<tr>
<td></td>
<td>‘genetic techniques’/exp</td>
<td>’epigenetics’:de</td>
</tr>
<tr>
<td></td>
<td>‘genetic testing’/exp</td>
<td>’exome’:de</td>
</tr>
<tr>
<td></td>
<td>‘microarray analysis’/exp</td>
<td>‘gene mutation’/exp</td>
</tr>
<tr>
<td></td>
<td>‘oligonucleotide array sequence analysis’:de</td>
<td>‘gene sequencing’:de</td>
</tr>
<tr>
<td></td>
<td>‘comparative genomic hybridization’:de</td>
<td>‘genetic screening’:de</td>
</tr>
<tr>
<td></td>
<td>‘molecular sequence data’</td>
<td>‘genetic procedures’/exp</td>
</tr>
<tr>
<td></td>
<td>‘sequence analysis, DNA’:de</td>
<td>‘genome’:de</td>
</tr>
<tr>
<td></td>
<td>‘sequence deletion’/genetics</td>
<td>‘genome imprinting’:de</td>
</tr>
<tr>
<td></td>
<td>aCGH</td>
<td>‘microarray analysis’:de</td>
</tr>
<tr>
<td></td>
<td>Array CGH</td>
<td>‘molecular diagnosis’:de</td>
</tr>
<tr>
<td></td>
<td>Array genomic hybridization</td>
<td>‘nucleic acid analysis’/exp</td>
</tr>
<tr>
<td></td>
<td>cDNA array</td>
<td></td>
</tr>
</tbody>
</table>
# Sample Strategy – Genetic testing concepts

<table>
<thead>
<tr>
<th>Set Number</th>
<th>Concept</th>
<th>Search statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Genetic testing</td>
<td>‘Chromosome aberration’/exp or (chromosom* NEAR/2 (duplicat* or deletion or ‘copy number’ or insertion))</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>‘microarray analysis’:de or ‘nucleic acid analysis’/exp or ‘molecular diagnosis’:de or ‘genetic screening’:de or ‘genetic procedures’/exp or ‘array cgh’ or ‘aCGH’ or ‘CMA’ or ‘comparative genomic hybridization’ or ‘array genomic hybridization’ or microarray or (molecular NEAR/2 diagnos*) or snp or ‘single nucleotide polymorphism array’ or (genetic NEAR/2 test*)</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>(exome:de OR genome:de) and ‘gene sequencing’:de</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>(‘whole exome’ or ‘whole genome’) NEAR/3 sequencing</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>‘next generation sequencing’ or ‘NGS’</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>‘gene expression assay’/exp or ‘gene chips’ or ‘cDNA array’ or ‘cDNA microarray’ or ‘genome imprinting’:de or imprinting</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Methylation or ‘epigenetics’:de or epigenetic*</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>#1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7</td>
</tr>
</tbody>
</table>
## Sample Strategy – Conditions

<table>
<thead>
<tr>
<th></th>
<th>Conditions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Development* NEAR/2 (delay* or disabilit*)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>‘mental deficiency’/exp or (mental* NEAR/2 retard*) or (intellect* NEAR/2 (disabilit* or delay*))</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>(Neurocognitive NEAR/2 impair*) or ‘cognitive defect’:de or ‘intellectual impairment’:de</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>‘Fragile X’ or ‘fragile-x’ or ‘mental retardation malformation syndrome’/exp</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>‘autism’/exp or autistic* or autism or Asperger*:ti,ab or ‘asd’:ti,ab or ‘rett syndrome’ or ‘pervasive developmental disorder’ or ‘PDD’</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Specific syndromes (original)</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Specific syndromes – KI suggested</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Specific genes</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Combine sets</td>
<td></td>
</tr>
</tbody>
</table>

- ‘Fragile X’ or ‘fragile-x’ or ‘mental retardation malformation syndrome’/exp
- ‘autism’/exp or autistic* or autism or Asperger*:ti,ab or ‘asd’:ti,ab or ‘rett syndrome’ or ‘pervasive developmental disorder’ or ‘PDD’
- ‘angelman syndrome’/exp OR ‘happy puppet’ OR ‘prader-willi’/exp OR ‘rubinstein-taybi’/exp OR ‘smith magenis’/exp OR ‘velocardo facial syndrome’/exp OR ‘digeorge syndrome’/exp OR ‘shprrintzen syndrome’ OR ‘conotruncal anomaly face syndrome’ OR ‘williams syndrome’/exp OR ‘williams-beuuren syndrome’/exp
- ‘kleefstra syndrome’ OR ‘miller-dieker syndrome’ OR ‘koolen-de vries syndromre’ OR ‘wagr syndrome’ OR ‘langer gideon syndrome’ OR ‘cri du chat syndrome’ OR ‘wolf-hirschorn syndrome’ OR ‘jacobson syndrome’ OR ‘alagille syndromre’ OR ‘1p36 deletion syndromre’ OR ‘9q deletion syndromre’ OR ‘17q21.31 deletion syndromre’ OR ‘18p minus syndromre’ OR ‘18q minus syndromre’ OR ‘sry deletion’ OR ‘pten deletion’ OR ‘charcot-marie-tooth syndrome’
- #9 OR #10 OR #11 OR #12 OR #13 OR #14 OR #15 OR #16

**Specific genes**
- ube3a OR fmr1 OR mecp2 OR cdkl5 OR foxg1 OR crebbp OR ep300
## Sample Strategy - Diagnosis

<table>
<thead>
<tr>
<th>19</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>'diagnostic test accuracy':de OR 'diagnosis':lnk OR 'receiver operating characteristic':de OR 'roc curve'/exp OR 'roc curve' OR 'sensitivity and specificity':de OR 'sensitivity' OR 'specificity' OR 'accuracy':de OR 'precision'/exp OR precision OR 'prediction and forecasting'/exp OR 'prediction and forecasting' OR 'diagnostic error'/exp OR 'diagnostic error' OR 'maximum likelihood method':de OR 'likelihood' OR 'predictive value'/exp OR 'predictive value' OR ppv OR (false OR true) NEAR/1 (positive OR negative)</td>
</tr>
</tbody>
</table>
## Add limiters

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>21</strong></td>
<td>Limit by keywords</td>
<td>#18 AND (idiopathic or (clinical NEAR/2 (valid* or util* or relevanc*)) )</td>
</tr>
<tr>
<td><strong>22</strong></td>
<td>Combine sets</td>
<td>#20 OR #21</td>
</tr>
<tr>
<td><strong>23</strong></td>
<td>Limits</td>
<td>#22 NOT (prenatal:ti or maternal:ti)</td>
</tr>
<tr>
<td><strong>24</strong></td>
<td>Limit by publication and study type</td>
<td>#23 AND ('clinical article'/de OR 'clinical trial'/de OR 'cohort analysis'/de OR 'comparative study'/de OR 'controlled study'/de OR 'diagnostic test accuracy study'/de OR 'intermethod comparison'/de OR 'major clinical study'/de OR 'medical record review'/de OR 'practice guideline'/de OR 'prospective study'/de OR 'retrospective study'/de OR 'validation study'/de) AND ('Article'/it OR 'Article in Press'/it OR 'Conference Abstract'/it OR 'Conference Paper'/it OR 'Review'/it)</td>
</tr>
</tbody>
</table>
Combine sets

- Main conceptual groups
  - Set #19 = #8 (genetic testing) AND #17 (conditions) AND #19 (diagnosis)

- Apply limiters
  - Idiopathic OR clinical validity/utility
  - NOT (prenatal:ti OR maternal:ti)
  - Publication types
Combine sets

- The intersection at the center is set #24
- The limits included articles published from August 2014 through January 2015 and English language publications
Sample Strategy – Product Brief - Cologuard

PubMed (note that Exact Sciences retrieves many false hits, so it requires additional descriptive terms)

S1 cologuard[tiab]
S2 ("exact sciences"[tiab] OR "exact sciences"[ad]) AND (colorectal neoplasms[mh] OR colon[tiab] OR colorectal[tiab])
S4 #1 OR #2 OR #3

EMBASE

S1 cologuard:ab,ti,dn
S2 'exact sciences':ab,ti,df,ad AND ('colon tumor'/exp OR colon OR colorectal)
S3 ('faecal DNA' OR 'fecal DNA' OR 'stool DNA'):ti AND (screen OR test*):ti AND ('colon tumor'/exp OR colon OR colorectal)
S4 #1 OR #2 OR #3
Sample Strategy – Product Brief - Cologuard

Cochrane Library
S1  cologuard [in title, abstract, or keywords fields]
S2  "exact sciences" AND (colon OR colorectal) [in title, abstract, or keywords fields]
S3  ("faecal DNA" OR "fecal DNA" OR "stool DNA") AND (screen OR test*) AND (colon OR colorectal) [in title, abstract, or keywords fields]
S4  #1 OR #2 OR #3

Google Scholar
Cologuard
Challenges

- Searching by product name is difficult when products are not specifically named in abstracts or articles.

<table>
<thead>
<tr>
<th>No.</th>
<th>Query</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>#21</td>
<td>#20 AND (therap* OR treat*)</td>
<td>253</td>
</tr>
<tr>
<td>#20</td>
<td>#19 AND [2010-2015]py</td>
<td>272</td>
</tr>
<tr>
<td>#19</td>
<td>#18 NOT (animal* OR model OR simulation)</td>
<td>272</td>
</tr>
<tr>
<td>#18</td>
<td>#15 OR #17</td>
<td>292</td>
</tr>
<tr>
<td>#17</td>
<td>#16 AND gene*</td>
<td>289</td>
</tr>
<tr>
<td>#16</td>
<td>#14 AND (cancer OR myeloma OR glioma OR oncolog* OR tumor OR tumour OR glioblastoma OR nsclc)</td>
<td>303</td>
</tr>
<tr>
<td>#15</td>
<td>#14 AND (genetic OR genomic OR gene*) AND (profile OR profiling)</td>
<td>162</td>
</tr>
<tr>
<td>#14</td>
<td>foundationone:dn OR foundationone OR &quot;foundation medicine&quot;:df OR 'foundation medicine'</td>
<td>308</td>
</tr>
<tr>
<td>#13</td>
<td>#11 OR #12 AND [humans]/lim AND [english]/lim</td>
<td>22</td>
</tr>
<tr>
<td>#12</td>
<td>341 NEAR/1 gene</td>
<td>5</td>
</tr>
<tr>
<td>#11</td>
<td>#10 NOT #1</td>
<td>25</td>
</tr>
<tr>
<td>#10</td>
<td>#4 OR #9</td>
<td>34</td>
</tr>
<tr>
<td>#9</td>
<td>#7 AND #8</td>
<td>8</td>
</tr>
<tr>
<td>#8</td>
<td>hybrid NEAR/2 ('capture-based' OR 'capture based')</td>
<td>17</td>
</tr>
<tr>
<td>#7</td>
<td>#5 OR #6</td>
<td>15,809</td>
</tr>
<tr>
<td>#6</td>
<td>'ngs'</td>
<td>5,167</td>
</tr>
<tr>
<td>#5</td>
<td>('next generation' OR 'next-generation') NEAR/1 sequencing</td>
<td>14,291</td>
</tr>
<tr>
<td>#4</td>
<td>#2 AND #3</td>
<td>26</td>
</tr>
<tr>
<td>#3</td>
<td>'massively parallel' NEAR/2 sequencing</td>
<td>1,988</td>
</tr>
<tr>
<td>#2</td>
<td>genomic NEAR/2 profiling*</td>
<td>2,976</td>
</tr>
<tr>
<td>#1</td>
<td>foundationone:dn OR foundationone OR &quot;foundation medicine&quot;:df OR 'foundation medicine'</td>
<td>308</td>
</tr>
</tbody>
</table>
Challenges

- Refining topics can be a challenge.
  - Example – excluding citations pertaining to non-small cell lung cancer from a search for small cell lung cancer diagnostics.
  - Problem – the phrase we want is embedded in the phrase we want to exclude.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Search Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small cell lung cancer</td>
<td>'small cell lung carcinoma'/exp OR 'small cell lung carcinoma' OR 'carcinoma small cell'/exp OR 'carcinoma small cell' OR 'lung small cell cancer':de</td>
</tr>
<tr>
<td></td>
<td>'small-cell lung cancer' OR 'small cell lung cancer' OR 'oat cell' OR sclc</td>
</tr>
<tr>
<td></td>
<td>#1 OR #2</td>
</tr>
<tr>
<td></td>
<td>#3 NOT ('non-small cell':ti OR 'non-small-cell':ti OR 'non small cell':ti OR 'nonsmall cell':ti OR nsclc:ti)</td>
</tr>
</tbody>
</table>
Challenges

► Searching for diagnosis in the gray literature generates a lot of false positive results.

► Many treatment studies note that the patient “has a diagnosis”.

► Indexing of diagnostic concepts not consistent – you need to search for the related concepts using keywords and controlled vocabulary terms.
  ■ Even when a study claims to focus on clinical utility it frequently is reporting on clinical validity.
Challenges

- Can be difficult to distinguish between a lack of information and the failure of a strategy to identify information

- Searching is an iterative process – requires time
  - Comprehensive (sensitive) search – more time consuming
  - Targeted (specific) search – less time

- Trade-offs
  - When you use a very targeted strategy you inherently exclude citations
Take home messages

- Searching for diagnostic topics is tricky
- You need to use more than one bibliographic database
- You need to search the gray literature
- You need to include both controlled vocabulary terms and keywords in your searches.
Take home messages

- Consult with an information professional
  - Librarians do more than shelve books!
Risk of bias of diagnostic test evidence

Amy Tsou, MD, MSc
Senior Research Analyst
Why does it matter?

- Diagnostic tests play a key role in medicine
- There can be a lot at stake when tests get things wrong.
Prenatal Tests Have High Failure Rate, Triggering Abortions

by SUSAN DONALDSON JAMES

Stacie and Lincoln Chapman with their son Lincoln Sam. A screening test suggested Sam had Trisomy 18, but he was born healthy. © Lauren Owens/ECRI
Getting at a true estimate of a test’s accuracy

Understanding potential limitations of diagnostic test evidence *matters!*
Overview

- What are some distinctive features of diagnostic studies?
- What are common sources of bias in diagnostic studies?
- One tool for systematic assessment of risk of bias in diagnostic studies

Scope of this talk: diagnostic *accuracy* studies
Distinctive Features: Comparators

What is being compared?

Intervention Studies

- Intervention vs. No intervention
  - Medication vs. No medication
  - Stent placement vs. Medical Management

Diagnostic Accuracy Studies

- Diagnostic Test (Index) vs. Reference Test
  - Cognitive test vs. Autopsy for Alzheimer’s
  - MRI vs. CT for Stroke Detection
Distinctive Features: Different Outcomes

**Intervention Studies**

**Diagnostic Accuracy Studies**

**Clinical Outcome**
- Change in Blood Pressure
- Mortality
- Surrogate Measures (Readmissions, Hospital days)

**Accuracy, Predictive Values**
- Sensitivity, Specificity
- Positive / negative predictive value
# Diagnostic Study Measures

<table>
<thead>
<tr>
<th>Test results</th>
<th>Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>With Disease</td>
</tr>
<tr>
<td></td>
<td>Without Disease</td>
</tr>
<tr>
<td>Positive</td>
<td>True positives</td>
</tr>
<tr>
<td></td>
<td>False Positives</td>
</tr>
<tr>
<td>Negative</td>
<td>False negatives</td>
</tr>
<tr>
<td></td>
<td>True negatives</td>
</tr>
</tbody>
</table>

- **Sensitivity**: Probability that an individual with disease gets a *positive* test result \(\frac{TP}{TP + FN}\)

- **Specificity**: Probability that an individual without disease gets a *negative* test result \(\frac{TN}{TN + FP}\)
# Diagnostic Study Measures

<table>
<thead>
<tr>
<th>Test results</th>
<th>Participants</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>With Disease</td>
<td>Without Disease</td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>True positives</td>
<td>False Positives</td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>False negatives</td>
<td>True negatives</td>
<td></td>
</tr>
</tbody>
</table>

- **Positive Predictive Value**: Probability that a person with a positive result actually *has* the disease \((TP/TP + FP)\)

- **Negative Predictive Value**: Probability that a person with a negative result *does not* have the disease \((TN/TN + FN)\)

**Predictive values are affected by disease prevalence in the population in which a test is being used.**
Distinctive Features: Best Trial Design

Intervention Studies

Prospective, double blind randomized controlled trial (RCT)

Diagnostic Accuracy Studies

Prospective blind comparison (of test/reference test) in a consecutive series of patients from the relevant patient population

The Challenge

Diagnostic Accuracy Studies

How accurate is diagnostic test X compared to the reference standard (test Y)?

- What factors may cause a study to systematically OVERESTIMATE or UNDERESTIMATE a test’s diagnostic accuracy?
Overview

► What are some distinctive features of diagnostic studies?

► What are common sources of bias in diagnostic studies?

► One tool for systematic assessment of risk of bias in diagnostic studies
Risk of Bias: 3 Factors to Consider

Study Design

Study Conduct

Study Reporting

Santaguida et al., Assessing Risk of Bias as a Domain of Quality in Medical Test Studies, Chapter 5 of Methods Guide for Medical Test Reviews; Agency for Healthcare Research and Quality; June 2012
Risk of Bias: 3 Factors to Consider

Study Design

Study Conduct

Study Reporting

Spectrum Bias

Santaguida et al., Assessing Risk of Bias as a Domain of Quality in Medical Test Studies, Chapter 5 of Methods Guide for Medical Test Reviews; Agency for Healthcare Research and Quality; June 2012
Spectrum Bias

- Flawed estimate of accuracy because the test was validated in patients that aren’t representative

- Official Definition: “Demographic features or disease severity may lead to variations in estimates of test performance”

- Example: Diagnostic Imaging

Santaguida et al., Assessing Risk of Bias as a Domain of Quality in Medical Test Studies, Chapter 5 of Methods Guide for Medical Test Reviews; Agency for Healthcare Research and Quality; June 2012
A new kind of diagnostic imaging

How accurate are these apps?

How accurate are smartphone apps for detecting melanoma?

Database of 188 skin photos
(60 melanoma, 128 benign)

Images uploaded to 4 mobile melanoma detection apps

Primary Outcome: Sensitivity

Wolf et al. Diagnostic Inaccuracy of Smartphone Applications for Melanoma Detection; JAMA Dermatology, 2013;149(4):422-426
Melanoma Detection:

<table>
<thead>
<tr>
<th>Mobile App Number</th>
<th>Sensitivity (%)</th>
<th>95% Confidence Interval</th>
<th>Specificity (%)</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>70</td>
<td>56 to 80.8</td>
<td>39.3</td>
<td>30.7 to 48.6</td>
</tr>
<tr>
<td>2</td>
<td>69</td>
<td>55.3 to 80.1</td>
<td>37</td>
<td>28.7 to 46.1</td>
</tr>
<tr>
<td>3</td>
<td>6.8</td>
<td>2.2 to 17.3</td>
<td>93.7</td>
<td>87 to 97.2</td>
</tr>
<tr>
<td>4</td>
<td>98.1</td>
<td>88.8 to 99.9</td>
<td>30.4</td>
<td>22.1 to 40.3</td>
</tr>
</tbody>
</table>

Wide range of sensitivities: 6.8% to 98.1%

Only app # 4 involved sending the photo for evaluation by a dermatologist.

Wolf et al. Diagnostic Inaccuracy of Smartphone Applications for Melanoma Detection; JAMA Dermatology, 2013;149(4):422-426
Study Design: Selected Study Population

No Excision! Reassurance or Monitoring

Biopsy

©2015 ECRI INSTITUTE
Study Design Can Lead to Bias

Overestimation of accuracy

Spectrum Bias

- ↑ Prevalence of melanoma
- ↑ Disease severity

All patients presenting to dermatologist

No Biopsy

Biopsy
Spectrum Bias

- Differentially studying patients with more severe disease may lead to consistent OVERESTIMATION of accuracy.

- Differentially studying patients with mild disease may lead to consistent UNDERESTIMATION of accuracy.
## Melanoma Detection:

<table>
<thead>
<tr>
<th>Mobile App Number</th>
<th>Sensitivity (%)</th>
<th>95% Confidence Interval</th>
<th>Specificity (%)</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>70</td>
<td>56 to 80.8</td>
<td>39.3</td>
<td>30.7 to 48.6</td>
</tr>
<tr>
<td>2</td>
<td>69</td>
<td>55.3 to 80.1</td>
<td>37</td>
<td>28.7 to 46.1</td>
</tr>
<tr>
<td>3</td>
<td>6.8</td>
<td>2.2 to 17.3</td>
<td>93.7</td>
<td>87 to 97.2</td>
</tr>
<tr>
<td>4</td>
<td>98.1</td>
<td>88.8 to 99.9</td>
<td>30.4</td>
<td>22.1 to 40.3</td>
</tr>
</tbody>
</table>

These estimates, probably too high!

In this case, further evidence that smartphone apps are even worse!

Wolf et al. Diagnostic Inaccuracy of Smartphone Applications for Melanoma Detection; JAMA Dermatology, 2013;149(4):422-426
Spectrum Bias: The MMSE

Mini Mental Status Exam (MMSE): Used to evaluate cognitive function and diagnose dementia

- 11 tasks, scored from 0 to 30 (perfect score)
  - Easy: What is the date, month, year, season and day of the week?
  - Hard: Serial 7’s. Start at 100 and keep subtracting 7

- Test performance varies with # of years of education

Crum et al. Population-Based Norms for the Mini-Mental State Examination by Age and Educational Level, JAMA 1993; 269:2386-2391
Spectrum Bias: The MMSE

- Variation of test performance by schooling

- When the study population only includes a particular part of the spectrum, this limits the study’s ability to accurately describe the test’s performance

Crum et al. Population-Based Norms for the Mini-Mental State Examination by Age and Educational Level, JAMA 1993; 269:2386-2391
In fact, many factors impact test performance.

Age, Gender, Schooling all affect test performance.

Mini-Mental State Examination norms by age, gender, and education level

http://www.uptodate.com/contents/image?imageKey=PC%2FF79818&topicKey=DRUG_GEN%2FF9268&rank=1%7E150&source=see_link&search=mmse+dementia
Problematic Study Designs: Case Control

Case Control Studies

Patients chosen based on whether they are:

- Cases (With disease)
- Controls (No disease)

High Risk for Spectrum Bias!
Case Control Design

All patients presenting to dermatologist

60 cases (melanoma)
128 controls (benign)
Case control studies for diagnostic test accuracy = BAD

But how bad are they really? What’s the evidence?
What’s the effect on accuracy?

- Lijmer et al. performed a systematic review, meta-analysis of studies evaluating 218 diagnostic tests

- How do estimates of accuracy from case-control studies compare to cohort studies (from more representative patient samples)?

- Case-control studies were significantly more likely to overestimate a test’s accuracy, reporting diagnostic odds ratios that were 3 times higher compared to non-case control studies
  - Probably because they tended to exclude patients with less severe disease

Risk of Bias: 3 Factors to Consider

**Study Design**

- Spectrum Bias

**Study Conduct**

- Partial Verification Bias
- Clinical Review Bias
- Observer/Instrument Variation

**Study Reporting**

Santaguida et al., Assessing Risk of Bias as a Domain of Quality in Medical Test Studies, Chapter 5 of Methods Guide for Medical Test Reviews; Agency for Healthcare Research and Quality; June 2012
Partial Verification Bias

- Only a selected sample of patients undergoing the index test are verified by the reference standard

- In other words: Not all patients go on to have reference test

- Example: Imaging for staging in small cell lung cancer (SCLC)

Santaguida et al., Assessing Risk of Bias as a Domain of Quality in Medical Test Studies, Chapter 5 of Methods Guide for Medical Test Reviews; Agency for Healthcare Research and Quality; June 2012
Partial Verification Bias: Example

- How accurate are imaging modalities like PET-CT for identifying SCLC metastases?
- Reference standard: Biopsy

![Diagram showing the verification process of SCLC patients and PET-CT results.

SCLC Patients → PET-CT

- If + metastases on PET-CT: Verified by Biopsy
- If no mets on PET-CT: No Verification, or Verified by different standard

Treadwell et al. Imaging for Staging in SCLC; under review
Partial Verification Bias: cont’d

- Good reasons why patients do not always end up getting the reference test
  - If PET-CT does not identify any potential metastases, where would you biopsy??
  - As a surgical procedure, biopsy has risks
  - Depending on location, biopsy might not be feasible
  - May not be important for clinical decision-making: staging and treatment don’t change if one of the potential mets in the brain turns out to be a false positive.

- Even if there are “good” reasons, partial verification bias can affect estimates of accuracy
Partial Verification Bias: Example

- Introduces Spectrum Bias
- Patients getting a biopsy are more likely to be abnormal
Risk of Bias: 3 Factors to Consider

- Study Design
- Study Conduct
- Study Reporting

Spectrum Bias

- Partial Verification Bias
- Clinical Review Bias
- Observer/Instrument Variation

Santaguida et al., Assessing Risk of Bias as a Domain of Quality in Medical Test Studies, Chapter 5 of Methods Guide for Medical Test Reviews; Agency for Healthcare Research and Quality; June 2012
Clinical review bias

Definition: Availability of clinical data such as age, sex, and other symptoms, during interpretation of test may affect estimates of test performance

In other words: Having access to other information about the patient could bias how the test gets interpreted

Example: One study compared PET-CT to “standard staging” protocols in SCLC patients.

Santaguida et al., Assessing Risk of Bias as a Domain of Quality in Medical Test Studies, Chapter 5 of Methods Guide for Medical Test Reviews; Agency for Healthcare Research and Quality; June 2012
Clinical review bias: Example

- Radiologists interpreting PET-CTs, *not* blinded to patient’s clinical data.

- Knowing the patient complained of severe back pain might bias radiologist towards concluding that a “borderline” abnormality in the spine is a metastases.

- Conversely, knowing the patient denied any pain might lead a radiologist to conclude something is NOT abnormal.
Clinical Review Bias

• Different from using a test in *clinical practice*, where it’s important to consider the clinical picture

• In context of a trial of accuracy, important to get at *how well does the test perform by itself*
Risk of Bias: 3 Factors to Consider

Spectrum Bias

- Partial Verification Bias
- Clinical Review Bias
- Observer/Instrument Variation

Santaguida et al., Assessing Risk of Bias as a Domain of Quality in Medical Test Studies, Chapter 5 of Methods Guide for Medical Test Reviews; Agency for Healthcare Research and Quality; June 2012
Observer variability bias

- For a test to be accurate, the results have to be consistently reproducible, even when the test is performed on different equipment or by different people.

- **Intraobserver variability**: When the test is performed again by the same observer, but with different results.

- **Interobserver variability**: Test performed by different observers with different results.
Observer variability bias: Examples

- Imaging for SCLC: An experienced radiologist might correctly interpret something as artifact, while a new radiology resident (on July 1) might think it’s abnormal.

- Particularly problematic for instruments administered by people or requiring subjective judgment:
  - How a test is administered can bias results: Variation in survey introductions
  - Subjective assessments: Capturing dysarthria (slurred speech) in patients with neurodegenerative disease
Risk of Bias: 3 Factors to consider

- **Study Design**
  - Spectrum Bias

- **Study Conduct**
  - Partial Verification Bias
  - Clinical Review Bias
  - Observer/Instrument Variation

- **Study Reporting**

Santaguida et al., Assessing Risk of Bias as a Domain of Quality in Medical Test Studies, Chapter 5 of Methods Guide for Medical Test Reviews; Agency for Healthcare Research and Quality; June 2012

©2015 ECRI INSTITUTE
Study Reporting

- How well did the study describe its design and how it was conducted?

- Study authors often fail to report key aspects of the study

- Examples:
  - No description of what reference standard was used, or whether all tests were verified by the same reference standard
  - Unclear if test readers were blinded or not
  - Unclear consecutive patients enrolled, or what criteria for selection were
Study Reporting

- Inadequate reporting does not necessarily mean the risk of bias is high!
- But without information, hard to assess whether bias could be present or not

*Studies shouldn’t necessarily be penalized, but also not appropriate to rate the risk of bias as LOW*

Santaguida et al., Assessing Risk of Bias as a Domain of Quality in Medical Test Studies, Chapter 5 of Methods Guide for Medical Test Reviews; Agency for Healthcare Research and Quality; June 2012
Risk of Bias: 3 Factors to Consider

Study Design

Study Conduct

Study Reporting

Spectrum Bias

- Partial Verification Bias
- Clinical Review Bias
- Observer/Instrument Variation

Particularly Problematic for Diagnostic Studies

Whiting et al. Sources of Variation and Bias in Studies of Diagnostic Accuracy. Annals of Internal medicine, 2004;140;189-202
## Effects of Various Biases on Accuracy

<table>
<thead>
<tr>
<th>Type of Bias</th>
<th>Clear, consistent Effect?</th>
<th># of studies</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Spectrum</td>
<td>✓</td>
<td>7</td>
<td>Increase</td>
<td>Mixed</td>
</tr>
<tr>
<td>• Partial Verification</td>
<td>✓</td>
<td>33</td>
<td>Increase</td>
<td>Decrease</td>
</tr>
<tr>
<td>• Clinical Review</td>
<td>✓</td>
<td>15</td>
<td>Increase</td>
<td>Variable</td>
</tr>
<tr>
<td>• Observer Variation (Interobserver)</td>
<td>✓</td>
<td>14</td>
<td>Increase for experts</td>
<td>NR</td>
</tr>
</tbody>
</table>

Whiting et al. A systematic review classified sources of bias and variation in diagnostic test accuracy studies; J. Clinical Epidemiology; 66(2013); 1093-1104
Types of Bias in Diagnostic Studies

Population
- Spectrum bias/spectrum effect
- Context Bias

Test Protocol
- Variation in text execution
- Variation in test technology
- Treatment paradox
- Disease progression bias

Reference Standard and Verification Procedure
- Inappropriate reference standard
- Differential verification bias
- Partial verification bias

Interpretation
- Review bias
- Clinical review bias
- Incorporation bias
- Observer variability

Analysis
- Handling of indeterminate results
- Arbitrary choice of threshold value

Whiting et al. Sources of Variation and Bias in Studies of Diagnostic Accuracy. Annals of Internal medicine, 2004;140;189-202
Overview

- What are some distinctive features of diagnostic studies?
- What are common sources of bias in diagnostic studies?
- One tool for systematic assessment of risk of bias in diagnostic studies
A quality assessment tool for diagnostic accuracy studies

The development of QUADAS-2 was led by a team based at the School of Social and Community Medicine at the University of Bristol.

QUADAS-2 Tool

4 Domains for Risk of Bias

Is the risk of bias: LOW  HIGH  UNCLEAR

www.quadas.org
Suggested graphical display of QUADAS-2 results

- **Flow and timing**
- **Reference Standard**
- **Index Standard**
- **Patient Population**

Proportion of studies with low, high, or unclear risk of bias, %

- Low
- High
- Unclear

Conclusion

- Assessing risk of bias is important! Clear and consistent evidence that spectrum bias, partial verification bias, clinical review bias and observer/instrument variation bias can distort estimates of accuracy

- Avoid case control study designs if at all possible!

  Case-control studies = 😞

- Validated instruments like the QUADAS-2 provide a helpful framework for assessing risk of bias
And last, but not least

- For now, better stick to getting your moles checked out by a dermatologist
References

- Santaguida et al., Assessing Risk of Bias as a Domain of Quality in Medical Test Studies, Chapter 5 of Methods Guide for Medical Test Reviews; Agency for Healthcare Research and Quality; June 2012
- Whiting et al. Sources of Variation and Bias in Studies of Diagnostic Accuracy. Annals of Internal medicine, 2004;140;189-202
- Whiting et al. A systematic review classified sources of bias and variation in diagnostic test accuracy studies; J. Clinical Epidemiology; 66(2013); 1093-1104
- Whiting, Penny et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Annals of Internal Medicine 2011;155;529-536
- Mulherin et al. Spectrum Bias or Spectrum Effect? Subgroup Variation in Diagnostic Test Evaluation; Annals of Internal Medicine, 2002;137:598-602
EVIDENCE BOOT CAMP II

Diagnostic Technologies and Genetic Tests
July 14–15, 2015

Meta-analysis of Diagnostic Tests

Kristen D'Anci, PhD
Senior Research Analyst, Health Technology Assessment and Evidence-based Practice Center, ECRI Institute
Background

There are two goals for a meta-analysis in a systematic review:

- Provide summary estimates, get an idea for the magnitude of the observed effect
- Identify, and hopefully explain heterogeneity in the results of studies included in the review

Image from http://omerad.msu.edu/ebm/Meta-analysis/Meta2.html
Background

- For systematic reviews of medical tests, a meta-analysis often focuses on synthesis of test performance data or accuracy
  - Remember: Accuracy is a surrogate outcome!
  - Diagnostic tests do not cure patients

- Tests compared to what other test?
  - Meta-analysis allows you to compare the accuracy of two or more tests to a standard comparator
  - The type of comparator test matters

Lavinia Ferrante di Ruffano et al. BMJ 2012;344:bmj.e686
Standards and tests

► Gold Standard: A “perfect” test that definitively defines the presence or absence of the condition of interest (disease)
  ■ Usually considered the “ideal test”
  ■ However, may be invasive
    ▶ Alzheimer's disease, the firm diagnosis is made with pathological exam of the brain at autopsy—but removing a brain from a living person is not a good treatment goal
    ▶ Celiac disease, the gold standard is biopsy of the small intestine, preparation for the process is unpleasant for the patient.

► “Gold Standard” may not yet exist for a condition
  ▶ e.g. OSA or fibromyalgia

Standards and tests

► **Reference Standard**: A standard with (at least some!) demonstrated accuracy
  - “Imperfect reference standards” misclassify patients
  - Type of reference standard may vary according to setting (e.g. diagnosis of concussion on the sports field versus diagnosis of concussion in the ER)
  - May differ according to goal (e.g. differentiating between concussion and no concussion vs. differentiating between uncomplicated concussion and concussion requiring possible neurosurgical intervention.

► **Index Test**: Our diagnostic test of interest
Clinical Problem
Memory loss and other signs of dementia

At least two of the following core mental functions must be significantly impaired to be considered dementia:

- Memory
- Communication and language
- Ability to focus and pay attention
- Reasoning and judgment
- Visual perception

What is the best way to determine these changes in function?

Different cognitive tests used in screening, many take less than 20 minutes to administer

- Most well-known MMSE
- Others: ACE-R, MoCA, Mini-cog

http://www.alz.org/what-is-dementia.asp
Cognitive tests to detect dementia (Tsoi et al. 2015)

PICOTS:
- All older patients
- Index test: Cognitive tests (e.g. MMSE, Mini-cog)
- Reference test: DSM or ICD diagnosis
- Accurate diagnosis of dementia
- Timing (n/a)
- Setting (n/a)

149 trials examining 11 screening tests
- Over 49,000 patients
- Risk of bias assessed with QUADAS2
- Bivariate model
- Hierarchical summary ROC curves
Diagnostic meta-analysis differs from meta-analysis of intervention studies

► Traditional meta-analysis focuses on one intervention and one outcome
  ▪ Antianxiety medications → Reductions in anxiety scores

► Diagnostic meta-analyses examine two factors that are not independent of each other across trials
  ▪ Sensitivity
  ▪ Specificity

► Mathematically and conceptually more complex
Diagnostic meta-analysis differs from meta-analysis of intervention studies

- May or may not see an overall summary effect estimate
  - Depending on the data, a pooled estimate may not be useful
- Data are more often presented in paired Forest plots or in various ROC curves
Familiar “Traditional” Forest Plot Comparing Treatment to Control

Review: Hardy & Thompson (1996)
Comparison: 01 Effects of diuretics on pre-eclampsia
Outcome: 01 Diuretics for pre-eclampsia (data from Hardy & Thompson, 1996)

<table>
<thead>
<tr>
<th>Study or sub-category</th>
<th>Treatment n/N</th>
<th>Control n/N</th>
<th>OR (fixed) 95% CI</th>
<th>Weight %</th>
<th>OR (fixed) 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14/131</td>
<td>14/136</td>
<td>3.86 [0.48, 2.28]</td>
<td>1.04</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>21/385</td>
<td>17/134</td>
<td>7.50 [0.20, 0.78]</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>14/57</td>
<td>24/48</td>
<td>6.18 [0.14, 0.74]</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6/38</td>
<td>18/40</td>
<td>4.64 [0.08, 0.67]</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>12/1011</td>
<td>35/760</td>
<td>12.42 [0.13, 0.48]</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>138/1370</td>
<td>175/1336</td>
<td>50.11 [0.59, 0.94]</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>15/506</td>
<td>20/524</td>
<td>6.00 [0.39, 1.52]</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>6/108</td>
<td>2/103</td>
<td>0.61 [0.59, 15.07]</td>
<td>2.97</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>65/153</td>
<td>40/102</td>
<td>8.68 [0.69, 1.91]</td>
<td>1.14</td>
<td></td>
</tr>
</tbody>
</table>

Total (95% CI): 3759 [3183] 100.00 [0.67] [0.56, 0.79]
Total events: 291 (Treatment), 345 (Control)
Test for heterogeneity: Chi² = 27.27, df = 8 (P = 0.0006), I² = 70.7%
Test for overall effect: Z = 4.60 (P < 0.00001)
Forest Plots for Pooled Sensitivity and Specificity

A. Studies of Mini-Cog test scores for dementia

<table>
<thead>
<tr>
<th>Study</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Borson et al., 2000</td>
<td>0.99 (0.96-1.00)</td>
<td>0.93 (0.87-0.97)</td>
</tr>
<tr>
<td>Borson et al., 2003</td>
<td>0.76 (0.65-0.85)</td>
<td>0.89 (0.87-0.91)</td>
</tr>
<tr>
<td>Borson et al., 2005</td>
<td>0.77 (0.66-0.86)</td>
<td>0.83 (0.78-0.87)</td>
</tr>
<tr>
<td>Borson et al., 2006</td>
<td>0.99 (0.95-1.00)</td>
<td>0.95 (0.92-0.98)</td>
</tr>
<tr>
<td>Carnero-Pardo et al., 2013</td>
<td>0.78 (0.71-0.84)</td>
<td>0.59 (0.50-0.67)</td>
</tr>
<tr>
<td>Fuchs et al., 2012</td>
<td>1.00 (0.84-1.00)</td>
<td>0.85 (0.81-0.89)</td>
</tr>
<tr>
<td>Holsinger et al., 2012</td>
<td>0.80 (0.56-0.94)</td>
<td>0.73 (0.69-0.77)</td>
</tr>
<tr>
<td>Kaufer et al., 2008</td>
<td>0.87 (0.76-0.95)</td>
<td>0.54 (0.43-0.64)</td>
</tr>
<tr>
<td>Millan et al., 2012</td>
<td>0.87 (0.83-0.90)</td>
<td>1.00 (0.94-1.00)</td>
</tr>
<tr>
<td>Combined</td>
<td>0.91 (0.80-0.96)</td>
<td>0.86 (0.74-0.93)</td>
</tr>
</tbody>
</table>

Q = 71.76, P < .001

I² = 88.85 (82.94-94.76)

B. Studies of ACE-R scores for dementia

<table>
<thead>
<tr>
<th>Study</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexopoulos et al., 2010</td>
<td>0.93 (0.83-0.98)</td>
<td>0.96 (0.92-0.98)</td>
</tr>
<tr>
<td>Bastide et al., 2012</td>
<td>0.92 (0.86-0.96)</td>
<td>0.69 (0.61-0.75)</td>
</tr>
<tr>
<td>Carvalho et al., 2010</td>
<td>1.00 (0.89-1.00)</td>
<td>0.82 (0.70-0.91)</td>
</tr>
<tr>
<td>dos Santos Kawala et al., 2012</td>
<td>0.94 (0.88-0.97)</td>
<td>0.94 (0.87-0.98)</td>
</tr>
<tr>
<td>Fang et al., 2014</td>
<td>0.92 (0.74-0.99)</td>
<td>0.86 (0.78-0.91)</td>
</tr>
<tr>
<td>Konstantinopoulou et al., 2011</td>
<td>0.91 (0.77-0.98)</td>
<td>0.82 (0.70-0.90)</td>
</tr>
<tr>
<td>Kwaik et al., 2010</td>
<td>0.93 (0.85-0.96)</td>
<td>0.95 (0.88-0.99)</td>
</tr>
<tr>
<td>Mioshi et al., 2006</td>
<td>0.94 (0.89-0.97)</td>
<td>0.89 (0.78-0.95)</td>
</tr>
<tr>
<td>Pigliautile et al., 2011, study 1</td>
<td>0.90 (0.76-0.97)</td>
<td>0.80 (0.65-0.91)</td>
</tr>
<tr>
<td>Pigliautile et al., 2011, study 2</td>
<td>0.82 (0.71-0.90)</td>
<td>1.00 (0.89-1.00)</td>
</tr>
<tr>
<td>Terpening et al., 2011</td>
<td>0.85 (0.76-0.92)</td>
<td>0.80 (0.64-0.91)</td>
</tr>
<tr>
<td>Torralba et al., 2011</td>
<td>0.98 (0.92-1.00)</td>
<td>0.88 (0.73-0.96)</td>
</tr>
<tr>
<td>Wong et al., 2013</td>
<td>0.93 (0.82-0.98)</td>
<td>0.95 (0.84-0.99)</td>
</tr>
<tr>
<td>Combined</td>
<td>0.92 (0.90-0.94)</td>
<td>0.89 (0.84-0.93)</td>
</tr>
</tbody>
</table>

Q = 25.54, P = .01

I² = 53.02 (23.44-82.60)
Dependence of sensitivity and specificity across studies

Meta-analysis aims to provide a meaningful summary of sensitivity and specificity across studies.

- **Within each study**, sensitivity and specificity are independent — they are estimated from different patients (those with a disease or those who are healthy).

- **Across studies**, sensitivity and specificity are generally *negatively* correlated — as one increases the other is expected to decrease.
  - This negative correlation is most obvious with varying thresholds (known as “threshold effect”), varying time from onset of symptom to test, et cetera.
  - Positive correlations are often due to a missing covariate in the analysis.

Trikalinos TA, Coleman CI, Griffith L, et al. Meta-analysis of test performance when there is a “gold standard.” In: Methods guide for medical test reviews. Available at www.effectivehealthcare.ahrq.gov/medtestsguide.cfm. ©2015 ECRI INSTITUTE
Paired Forest Plots

- This is an example with 11 studies using D-dimer tests to diagnose acute coronary events, showing that sensitivity increases as specificity decreases:

- Summarizing the two correlated variables is a multivariate problem, and multivariate methods should be used to address it.

A passing note on thresholds

Different studies may incorporate different thresholds for a diagnostic test

- e.g. MMSE could be a score of 23 or 24 for probable Alzheimer’s or 26 for MCI (Remembering higher scores are better scores)
- Not all tests have a specific threshold (e.g. imaging studies)

Changing the threshold for a measure impacts sensitivity and specificity

- Lower thresholds tend to classify more patients with a given condition
Changing the threshold for a measure impacts sensitivity and specificity

- Lower thresholds tend to classify more patients with a given condition
Meta-analysis considerations: Pooled sensitivity and specificity

- Simplest analysis; treat Se and Sp as separate outcomes (univariate analyses) and get an estimate of an “average” effect.

- This is naïve because they are related to each other via threshold
  - To use this approach the test threshold must be consistent across studies

- Beware studies with dissimilar results...
Example: 3 studies with different values of sensitivity and specificity

Study 1: 10% & 90% -- not very sensitive, but high specificity
Study 2: 80% and 80% -- Okay sensitivity and specificity
Study 3: 90% and 10%. – High sensitivity, but low specificity

Simply pooling these gives sensitivity of 60% and specificity of 60% - which does not really tell us anything useful about these data
What, then, to do with data from different studies?

- With a “Gold Standard” you need to incorporate the variation between studies
  - The bivariate random-effects model gives “average” sensitivity and specificity
  - Hierarchical summary ROC curves – gives you the line of “best fit” for your data on one plot

- Imperfect reference standard is handled a little differently
  - Assess the ability of the index test to predict patient outcomes
  - Assess agreement of the index and reference test results
  - Go ahead with “Gold Standard” paradigm, calculate “naïve” estimates of the index test’s sensitivity and specificity, but qualify study findings to avoid misinterpretation.

Bivariate Analysis

- A bivariate approach preserves the two-dimensional nature of the original data.
- Pairs of sensitivity and specificity are jointly analyzed
  - Correlation between the two measures is addressed by using a random effects model
  - Covariates can be added to the model (Multivariate analysis)
- Allows you to report a summary estimate
- Bivariate and multivariate approaches to diagnostic tests are an evolving area of meta-analytic methodology
Forest Plots for Pooled Sensitivity and Specificity

A. Studies of Mini-Cog test scores for dementia

<table>
<thead>
<tr>
<th>Study</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Borson et al,10</td>
<td>0.99 (0.96-1.00)</td>
<td>0.93 (0.87-0.97)</td>
</tr>
<tr>
<td>Borson et al,34</td>
<td>0.76 (0.65-0.85)</td>
<td>0.89 (0.87-0.91)</td>
</tr>
<tr>
<td>Borson et al,35</td>
<td>0.77 (0.66-0.86)</td>
<td>0.83 (0.78-0.87)</td>
</tr>
<tr>
<td>Borson et al,36</td>
<td>0.99 (0.95-1.00)</td>
<td>0.95 (0.92-0.98)</td>
</tr>
<tr>
<td>Carnero-Pardo et al,37</td>
<td>0.78 (0.71-0.84)</td>
<td>0.59 (0.50-0.67)</td>
</tr>
<tr>
<td>Fuchs et al,38</td>
<td>1.00 (0.84-1.00)</td>
<td>0.85 (0.81-0.89)</td>
</tr>
<tr>
<td>Holsinger et al,39</td>
<td>0.80 (0.56-0.94)</td>
<td>0.73 (0.69-0.77)</td>
</tr>
<tr>
<td>Kaufer et al,40</td>
<td>0.87 (0.76-0.95)</td>
<td>0.54 (0.43-0.64)</td>
</tr>
<tr>
<td>Milian et al,41</td>
<td>0.87 (0.83-0.90)</td>
<td>1.00 (0.94-1.00)</td>
</tr>
<tr>
<td><strong>Combined</strong></td>
<td><strong>0.91 (0.80-0.96)</strong></td>
<td><strong>0.86 (0.74-0.93)</strong></td>
</tr>
</tbody>
</table>

Q_5 = 71.76, P < .001
I^2 = 88.85 (82.94-94.76)

B. Studies of ACE-R scores for dementia

<table>
<thead>
<tr>
<th>Study</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexopoulos et al,42</td>
<td>0.93 (0.83-0.98)</td>
<td>0.96 (0.92-0.98)</td>
</tr>
<tr>
<td>Bastide et al,43</td>
<td>0.92 (0.86-0.96)</td>
<td>0.69 (0.61-0.75)</td>
</tr>
<tr>
<td>Carvalho et al,44</td>
<td>1.00 (0.89-1.00)</td>
<td>0.82 (0.70-0.91)</td>
</tr>
<tr>
<td>dos Santos Kawala et al,45</td>
<td>0.94 (0.89-0.97)</td>
<td>0.94 (0.87-0.98)</td>
</tr>
<tr>
<td>Fang et al,46</td>
<td>0.92 (0.74-0.99)</td>
<td>0.86 (0.78-0.91)</td>
</tr>
<tr>
<td>Konstantinopoulos et al,47</td>
<td>0.91 (0.77-0.98)</td>
<td>0.82 (0.70-0.90)</td>
</tr>
<tr>
<td>Kwek et al,48</td>
<td>0.93 (0.85-0.96)</td>
<td>0.95 (0.88-0.99)</td>
</tr>
<tr>
<td>Mioshi et al,49</td>
<td>0.94 (0.89-0.97)</td>
<td>0.89 (0.78-0.95)</td>
</tr>
<tr>
<td>Pigliautile et al,50</td>
<td>0.90 (0.76-0.97)</td>
<td>0.80 (0.65-0.91)</td>
</tr>
<tr>
<td>Pigliautile et al,51</td>
<td>0.82 (0.71-0.90)</td>
<td>1.00 (0.89-1.00)</td>
</tr>
<tr>
<td>Terpening et al,52</td>
<td>0.85 (0.76-0.92)</td>
<td>0.80 (0.64-0.91)</td>
</tr>
<tr>
<td>Torralva et al,53</td>
<td>0.98 (0.92-1.00)</td>
<td>0.88 (0.73-0.96)</td>
</tr>
<tr>
<td>Wong et al,54</td>
<td>0.93 (0.82-0.98)</td>
<td>0.95 (0.84-0.99)</td>
</tr>
<tr>
<td><strong>Combined</strong></td>
<td><strong>0.92 (0.90-0.94)</strong></td>
<td><strong>0.89 (0.84-0.93)</strong></td>
</tr>
</tbody>
</table>

Q_5 = 258.11, P < .001
I^2 = 96.91 (95.80-98.03)
HSROC Curve: Sensitivity and Specificity of MMSE for the Detection of Dementia

Table 3. Meta-analyses for Diagnostic Accuracy on Dementia

<table>
<thead>
<tr>
<th>Screening Test</th>
<th>No. of Study Cohorts</th>
<th>Pooled (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMSE²</td>
<td>108</td>
<td>Sensitivity: 0.81 (0.78-0.84)</td>
</tr>
<tr>
<td>Very brief (≤5 min)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
HSROC Curve: Sensitivity and Specificity of MMSE for the Detection of Dementia

probability that the patients will be correctly classified
Sensitivity and Specificity of ACE-R, Mini-Cog Test and MMSE for the Detection of Dementia

Confidence ellipses clearly show the differences in sensitivity and specificity of the ACE-R, Mini-cog, and the MMSE.
Sensitivity and Specificity of MMSE and MoCA for the Detection of MCI

Tsoi et al. JAMA Intern Med. Published online June 08, 2015
Estimates of Heterogeneity in Diagnostic Meta-analyses

- $I^2$ statistic: expressed as a percentage, is independent of scale
  - Might be more useful to conceptualize as a measure of inconsistency across study findings

- Q statistic (Cochrane Q statistic or Chi-squared test)
  - Statistically significant $p$ value indicates heterogeneity
  - Has been argued to be underpowered

- Likely to see one or both measures with Forest plots
Forest Plots for Pooled Sensitivity and Specificity

A. Studies of Mini-Cog test scores for dementia

<table>
<thead>
<tr>
<th>Study</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Borson et al.</td>
<td>0.99 (0.96-1.00)</td>
<td>0.93 (0.87-0.97)</td>
</tr>
<tr>
<td>Borson et al.</td>
<td>0.76 (0.65-0.85)</td>
<td>0.89 (0.87-0.91)</td>
</tr>
<tr>
<td>Borson et al.</td>
<td>0.77 (0.66-0.86)</td>
<td>0.83 (0.78-0.87)</td>
</tr>
<tr>
<td>Borson et al.</td>
<td>0.99 (0.95-1.00)</td>
<td>0.95 (0.92-0.98)</td>
</tr>
<tr>
<td>Carnero-Pardo et al.</td>
<td>0.78 (0.71-0.84)</td>
<td>0.59 (0.50-0.67)</td>
</tr>
<tr>
<td>Fuchs et al.</td>
<td>1.00 (0.84-1.00)</td>
<td>0.85 (0.81-0.89)</td>
</tr>
<tr>
<td>Holsinger et al.</td>
<td>0.80 (0.56-0.94)</td>
<td>0.73 (0.69-0.77)</td>
</tr>
<tr>
<td>Kauf et al.</td>
<td>0.87 (0.76-0.95)</td>
<td>0.54 (0.43-0.64)</td>
</tr>
<tr>
<td>Millan et al.</td>
<td>0.87 (0.83-0.90)</td>
<td>1.00 (0.94-1.00)</td>
</tr>
<tr>
<td>Combined</td>
<td>0.91 (0.88-0.96)</td>
<td>0.86 (0.74-0.93)</td>
</tr>
</tbody>
</table>

Q_\chi^2 = 71.76, P < .001
I^2 = 88.85 (82.94-94.76)

B. Studies of ACE-R scores for dementia

<table>
<thead>
<tr>
<th>Study</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexopoulos et al.</td>
<td>0.93 (0.83-0.98)</td>
<td>0.96 (0.92-0.98)</td>
</tr>
<tr>
<td>Bastide et al.</td>
<td>0.92 (0.86-0.96)</td>
<td>0.69 (0.61-0.75)</td>
</tr>
<tr>
<td>Carvalho et al.</td>
<td>1.00 (0.89-1.00)</td>
<td>0.82 (0.70-0.91)</td>
</tr>
<tr>
<td>dos Santos Kawala et al.</td>
<td>0.94 (0.89-0.97)</td>
<td>0.94 (0.87-0.98)</td>
</tr>
<tr>
<td>Fang et al.</td>
<td>0.92 (0.74-0.99)</td>
<td>0.86 (0.78-0.91)</td>
</tr>
<tr>
<td>Konstantinopoulou et al.</td>
<td>0.91 (0.77-0.98)</td>
<td>0.82 (0.70-0.90)</td>
</tr>
<tr>
<td>Kwa et al.</td>
<td>0.93 (0.85-0.96)</td>
<td>0.95 (0.88-0.99)</td>
</tr>
<tr>
<td>Mioh et al.</td>
<td>0.94 (0.89-0.97)</td>
<td>0.89 (0.78-0.95)</td>
</tr>
<tr>
<td>Piglia et al.</td>
<td>0.90 (0.76-0.97)</td>
<td>0.80 (0.65-0.91)</td>
</tr>
<tr>
<td>Piglia et al.</td>
<td>0.82 (0.71-0.90)</td>
<td>1.00 (0.89-1.00)</td>
</tr>
<tr>
<td>Terpenning et al.</td>
<td>0.85 (0.76-0.92)</td>
<td>0.80 (0.64-0.91)</td>
</tr>
<tr>
<td>Torralba et al.</td>
<td>0.98 (0.92-1.00)</td>
<td>0.88 (0.73-0.96)</td>
</tr>
<tr>
<td>Wong et al.</td>
<td>0.93 (0.82-0.98)</td>
<td>0.95 (0.84-0.99)</td>
</tr>
<tr>
<td>Combined</td>
<td>0.92 (0.90-0.94)</td>
<td>0.89 (0.84-0.93)</td>
</tr>
</tbody>
</table>

Q_\chi^2 = 25.54, P = .01
I^2 = 53.02 (23.44-82.60)
Possible Sources of Heterogeneity: Possible Subgroup Analyses (If you have sufficient data)

- Patient population/selection
- Methods to verify/interpret results
  - Variation in test readers
- Clinical setting
  - Could also be location specific, such as tests given in different countries or in different health care groups
- Disease severity
- Study quality/potential bias
Diagnostic Technologies and Genetic Tests
July 14–15, 2015

Grading the Evidence on Diagnostic Tests

James Reston, PhD, MPH
Associate Director, Health Technology Assessment and Evidence-based Practice Center, ECRI Institute
Overview

- Why an evidence-grading system is important
- The GRADE system
- Challenges specific to grading diagnostic evidence
- Choosing diagnostic accuracy outcomes
- Impact of accuracy outcomes on clinical outcomes
- GRADE domains applied to diagnostic studies
- Worked examples
Why is a Grading System Important?

- Reduces variability among different reviewers
- Improves transparency in methods
- Ensures that no important facets are overlooked
- Encourages researchers to conduct better research on important questions
- Provides users greater clarity as to the reviewer’s confidence in the evidence to support their conclusions
Graded Evidence Statements

► “The strength of evidence for diagnosing condition X with technology Y is moderate.” How was that determined?
► Let’s look under the hood.
GRADE*

Grades of Recommendation Assessment, Development and Evaluation

• See www.gradeworkinggroup.org

• Key separation between:
  – Quality of the evidence for each outcome and
  – Strength of recommendation for the technology


GRADE: Quality of the Evidence for a Single Outcome for a Single Comparison

1) Study design
2) Risk of bias
3) Inconsistency
4) Indirectness
5) Imprecision
6) Publication bias
7) Controlling for all plausible confounders would increase the effect
8) Large magnitude of effect
9) Dose-response gradient
GRADE: Quality of the Evidence for a Single Outcome for a Single Comparison

**High**
We are very confident that the true effect lies close to that of the estimate of the effect.

**Moderate**
We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different.

**Low**
Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect.

**Very Low**
We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect.
Challenges to Grading Diagnostic Evidence

- Evidence-grading tools designed for interventions are not easily applied to diagnostic test evidence.
  - Diagnostic evidence often indirectly related to key questions

- Applying strength-of-evidence domains to diagnostic studies is challenging when assessing diagnostic accuracy outcomes.
  - Difficult to determine when to downgrade for indirectness. Linking diagnostic accuracy outcomes to clinical outcomes partly depends on benefits and harms of treatment
  - Precision is difficult to determine for diagnostic accuracy outcomes because the impact on clinical outcomes is often unclear.

- Relative importance of outcomes depends on clinical context

Choosing Diagnostic Accuracy Outcomes

- Diagnostic accuracy outcomes include sensitivity, specificity, PPV and NPV, likelihood ratios, diagnostic odds ratios, post-test probabilities

- Clinical context determines diagnostic accuracy outcomes most likely to impact clinical outcomes

Choosing Diagnostic Accuracy Outcomes

- Sometimes disease diagnosis is less important than ruling out a disease with severe consequences

- **Triage tests** with high sensitivity and/or high NPV are useful (e.g. a negative plasma D-dimer test can rule out pulmonary embolism [PE] in patients with a low probability of PE.)

Accurate disease diagnosis is important when disease treatment has high risks (e.g. cancer).

- **Single test** needs both high sensitivity and specificity (or high PPV and NPV). If no adequate single test exists, consider **add-on test** (with high specificity or high PPV the most important outcomes). (e.g. PET to help identify distant metastases in small cell lung cancer).

---

Choosing Diagnostic Accuracy Outcomes

- Is it an invasive test?
Choosing Diagnostic Accuracy Outcomes

More invasive tests have greater harms, with further harms resulting from misdiagnosis.

False-positive and false-negative measurements for a test become important. The degree of harms depends on:

- **False-negative results**
  - Severity of disease (for missed diagnosis)
  - Risks of testing (if test is invasive and has harms itself)

- **False-positive results**
  - Invasiveness of further testing/treatment
  - Cognitive/emotional effects of inaccurate disease labeling

Impact of Accuracy Outcomes on Clinical Outcomes

Sometimes utility or impact of accuracy measures upon patients is unclear or irrelevant and will depend upon intermediary steps (especially treatment plans)

- PET/CT for staging primary cervical cancer in pelvic lymph nodes

ECRI Institute evidence report. PET/CT for cervical cancer. 2010.
GRADE: Quality of the Evidence for a Single Outcome for a Single Comparison

1) Study design
2) Risk of bias
3) Inconsistency
4) Indirectness
5) Imprecision
6) Publication bias
7) Controlling for all plausible confounders would increase the effect
8) Large magnitude of effect
9) Dose-response gradient
Study Design

- Determines the starting GRADE

- The other 8 domains are used to either increase or decrease from the starting grade
Diagnostic Studies that Evaluate Clinical Outcomes

- Trials that randomly assigned patients to groups start at **High Quality**
- Studies that did not randomly assign patients to groups (observational studies) start at **Low Quality**
- Same criteria as used for intervention/treatment studies
- Most diagnostic studies do not evaluate the effect of the test on clinical outcomes

The GRADE handbook chapter 7. Available at http://www.guidelinedevelopment.org/handbook/#h.f71c8w9c3nh8
Diagnostic Studies that Evaluate Diagnostic Accuracy Outcomes

- Cross-sectional or cohort studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard start at **High Quality**

- Other studies (e.g. diagnostic case-control studies, diagnostic case series) start at **Low Quality**

The GRADE handbook chapter 7. Available at http://www.guidelinedevelopment.org/handbook/#h.f7lc8w9c3nh8
GRADE: Quality of the Evidence for a Single Outcome for a Single Comparison

1) Study design
2) Risk of bias
3) Inconsistency
4) Indirectness
5) Imprecision
6) Publication bias
7) Controlling for all plausible confounders would increase the effect
8) Large magnitude of effect
9) Dose-response gradient
Risk of Bias

- Can only result in a downgrade (1 or 2 levels)
- “Serious limitations” in the studies means a 1-level downgrade (e.g. spectrum bias)
- “Very serious limitations” in the studies mean a 2-level downgrade (e.g. spectrum bias plus clinical review bias)
- Risk of bias is based on individual study evaluation of risk of bias; one can take an average or use only higher-quality studies when grading
GRADE: Quality of the Evidence for a Single Outcome for a Single Comparison

1) Study design
2) Risk of bias
3) Inconsistency
4) Indirectness
5) Imprecision
6) Publication bias
7) Controlling for all plausible confounders would increase the effect
8) Large magnitude of effect
9) Dose-response gradient

High
Moderate
Low
Very Low
Inconsistency

- Inconsistency refers to heterogeneity in the direction and magnitude of test results across studies.

- Inconsistency in test performance can be visually assessed on a receiver-operating characteristics (ROC) curve showing true-positive versus false-positive rates in ROC space.

Example: Anti-CCP for Diagnosis of RA

Inconsistency

- Heterogeneity across studies may be explained by different study designs, study quality, differences in reference standards or diagnostic test cutoffs, different patient characteristics etc.

- Unexplained heterogeneity should result in a downgrade

GRADE: Quality of the Evidence for a Single Outcome for a Single Comparison

1) Study design
2) Risk of bias
3) Inconsistency
4) Indirectness
5) Imprecision
6) Publication bias
7) Controlling for all plausible confounders would increase the effect
8) Large magnitude of effect
9) Dose-response gradient

High
Moderate
Low
Very Low
Indirectness

- Can only result in a downgrade (1 or 2 levels)

- Four types of indirectness
  - Indirectness of comparisons
  - Indirectness of outcomes
  - Indirectness of interventions
  - Indirectness of populations
Indirectness of Comparisons

- **Direct comparison** – tests A and B are compared against each other and a reference standard in the same study.

  1 study

  A vs B vs Reference

- **Indirect comparison** – test A is compared to the reference standard in one study, test B is compared to the reference standard in another study, and inferences are made about the relative performance of tests A and B.

  2 studies

  A vs Reference  B vs Reference

©2015 ECRI INSTITUTE
Indirectness of Outcomes

- **Direct outcomes** – generally patient-centered health outcomes (e.g. mortality, bone fracture, QOL)

- **Indirect outcomes** – surrogate or intermediate outcomes (e.g. diagnostic accuracy outcomes).
Indirectness of Outcomes

- Often there is no direct linkage between diagnostic accuracy and clinical outcomes.
  - Example: When tests are used as triage, accuracy of risk classification is more important than accuracy of diagnosis (e.g. D-Dimer to rule out PE in patients at low risk of PE).

- Sometimes reviewers may only be interested in diagnostic accuracy. In these cases there would be no downgrade for indirectness.

Indirectness of Interventions and Populations

- A test may differ slightly from the test of interest

- A study population may differ from the target population (e.g. a low risk vs. high risk of disease). Different settings (e.g. primary versus tertiary care) often have a different spectrum of patients.

- If there is evidence that these differences substantially impact outcomes, downgrade; otherwise do not downgrade
GRADE: Quality of the Evidence for a Single Outcome for a Single Comparison

1) Study design
2) Risk of bias
3) Inconsistency
4) Indirectness
5) Imprecision
6) Publication bias
7) Controlling for all plausible confounders would increase the effect
8) Large magnitude of effect
9) Dose-response gradient
Imprecision

- Can only result in a downgrade (1 or 2 levels)

- Random error, which can be caused by:
  - Large variability among patients
  - Small number of studies
  - Small number of patients

- Evaluating imprecision requires assessment of confidence intervals around diagnostic accuracy outcomes
Imprecision

Judging the precision of a particular confidence interval in estimates of test performance is challenging.

- This difficulty is due to the logarithmic nature of diagnostic performance measurements such as sensitivity, specificity, likelihood ratios, and diagnostic odds ratios.

- Relatively wide confidence intervals (suggesting imprecision) may not translate into clinically meaningful impacts.

- Clinical impact can be assessed by calculating post-test probabilities over a range of sensitivity/specificity values.

Example: Impact of the Precision of Sensitivity on Negative Predictive Value

- Core-needle biopsy for diagnosis of breast lesions
- Assume a 10% reduction in the sensitivity of freehand automated gun biopsy (98% → 88%)
- Estimated probability of having cancer after a negative test changes from 6% → 9%


©2015 ECRI INSTITUTE
GRADE: Quality of the Evidence for a Single Outcome for a Single Comparison

1) Study design
2) Risk of bias
3) Inconsistency
4) Indirectness
5) Imprecision
6) Publication bias
7) Controlling for all plausible confounders would increase the effect
8) Large magnitude of effect
9) Dose-response gradient

High
Moderate
Low
Very Low
Publication Bias

- Can only result in a downgrade (1 or 2 levels)

- Use when negative or no-difference findings appear to be unpublished/unavailable

- Publication bias can be assessed by testing for asymmetry in funnel plots that display outcomes from multiple studies. However, consensus is lacking on the best method to use.

- A study of 28 meta-analyses of diagnostic accuracy found evidence of asymmetry in the majority (smaller studies were associated with greater diagnostic accuracy)*

Publication Bias

Funnel Plots

Results of smaller negative trials may have been suppressed
GRADE: Quality of the Evidence for a Single Outcome for a Single Comparison

1) Study design
2) Risk of bias
3) Inconsistency
4) Indirectness
5) Imprecision
6) Publication bias
7) Controlling for all plausible confounders would increase the effect
8) Large magnitude of effect
9) Dose-response gradient
Example 1: Multislice Spiral CT vs Conventional Coronary Angiography for Diagnosing CAD

- CA is costly and invasive with potential complications; MSCT is non-invasive
- Meta-analysis of 21 studies with 1570 patients
- All patients were selected for conventional CA and generally had high probability of CAD (median prevalence in included studies 63.5%, range 6.6-100%)
- Graded outcomes included diagnostic measures; clinical outcomes not reported in the evidence base

GRADE Assessment: MSCT vs Conventional Coronary Angiography for Diagnosing CAD

▶ **Study design**: cross-sectional studies

▶ **Risk of bias**: no serious limitations

▶ **Indirectness**: True-positive, true-negative, and false-positive results were considered direct evidence with little uncertainty about clinical implications.

▶ Some uncertainty about directness for false negatives related to detrimental effects from delayed diagnosis or myocardial insult, resulting in one-level downgrade.

GRADE Assessment: MSCT vs Conventional Coronary Angiography for Diagnosing CAD

**Inconsistency**: Statistically significant, unexplained heterogeneity of results for sensitivity, specificity, likelihood ratios, and diagnostic odds ratios. All downgraded one level.

Inconsistency in Forest Plots

▶ **Specificity** (bottom graph) is clearly inconsistent among studies, so definite downgrade

▶ **Sensitivity** (top graph) is quantitatively inconsistent ($I^2 = 65.5\%$), but less obvious visually. Downgrade requires more judgment.

GRADE Assessment: MSCT vs Conventional Coronary Angiography for Diagnosing CAD

- **Imprecision**: No serious imprecision for any outcomes (95% CIs were not wide enough to change clinical impact)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Test findings (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pooled sensitivity</td>
<td>0.96 (0.94 to 0.98)</td>
</tr>
<tr>
<td>Pooled specificity</td>
<td>0.74 (0.65 to 0.84)</td>
</tr>
<tr>
<td>Positive likelihood ratio</td>
<td>5.4 (3.4 to 8.3)</td>
</tr>
<tr>
<td>Negative likelihood ratio</td>
<td>0.05 (0.03 to 0.09)</td>
</tr>
</tbody>
</table>

- **Publication bias**: Considered unlikely for all outcomes

# GRADE Summary Table: MSCT vs Conventional Coronary Angiography for Diagnosing CAD

<table>
<thead>
<tr>
<th>No of studies</th>
<th>Design</th>
<th>Limitations</th>
<th>Indirectness</th>
<th>Inconsistency</th>
<th>Imprecise data</th>
<th>Publication bias</th>
<th>Quality of evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True positives (patients with coronary artery disease)</strong></td>
<td>21 studies (1570 patients)</td>
<td>Cross sectional studies</td>
<td>No serious limitations</td>
<td>Little or no uncertainty</td>
<td>Serious inconsistency</td>
<td>No serious imprecision</td>
<td>Unlikely</td>
</tr>
<tr>
<td><strong>True negatives (patients without coronary artery disease)</strong></td>
<td>21 studies (1570 patients)</td>
<td>Cross sectional studies</td>
<td>No serious limitations</td>
<td>Little or no uncertainty</td>
<td>Serious inconsistency</td>
<td>No serious imprecision</td>
<td>Unlikely</td>
</tr>
<tr>
<td><strong>False positives (patients incorrectly classified as having coronary artery disease)</strong></td>
<td>21 studies (1570 patients)</td>
<td>Cross sectional studies</td>
<td>No serious limitations</td>
<td>Little or no uncertainty</td>
<td>Serious inconsistency</td>
<td>No serious imprecision</td>
<td>Unlikely</td>
</tr>
<tr>
<td><strong>False negatives (patients incorrectly classified as not having coronary artery disease)</strong></td>
<td>21 studies (1570 patients)</td>
<td>Cross sectional studies</td>
<td>No serious limitations</td>
<td>Some uncertainty</td>
<td>Serious inconsistency</td>
<td>No serious imprecision</td>
<td>Unlikely</td>
</tr>
</tbody>
</table>

Example 2: Cologuard for Colorectal Cancer Screening

- A stool-based test for detection of CRC-associated genetic markers and occult hemoglobin
- Intended as a non-invasive screening option for average-risk patients age 50 or older unwilling to undergo the invasive gold standard colonoscopy
- Evidence base: one multicenter prospective diagnostic cohort study with 12,776 average-risk asymptomatic patients scheduled for screening colonoscopy. Patients were also screened with Cologuard and fecal immunochemical test (FIT). 9,989 were analyzed.

GRADE-based Assessment: Cologuard for Colorectal Cancer Screening

- Graded outcomes included sensitivity for CRC, sensitivity for advanced precancerous lesions, and specificity for absence of CRC and advanced precancerous lesions.

- **Study design**: Diagnostic cohort study

- **Risk of bias**: Low (no serious limitations) for all 3 outcomes using modified QUADAS instrument

- **Indirectness**: Direct because diagnostic accuracy outcomes were the focus of specific KQs.

GRADE-based Assessment: Cologuard for Colorectal Cancer Screening

- **Inconsistency**: Unknown (single study), one-level downgrade
- **Imprecision**: Precise (no serious imprecision for all 3 outcomes).

<table>
<thead>
<tr>
<th>Measure</th>
<th>Cologuard Test findings (95% CI)</th>
<th>FIT findings (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity for CRC</td>
<td>92.3% (83% to 97.5%)</td>
<td>73.8% (61.5% to 84%)</td>
</tr>
<tr>
<td>Sensitivity for advanced precancerous lesions</td>
<td>42.4% (38.9% to 46%)</td>
<td>23.8% (20.8% to 27%)</td>
</tr>
<tr>
<td>Specificity for absence of CRC and advanced precancerous lesions</td>
<td>86.6% (85.9% to 87.2%)</td>
<td>94.9% (94.4% to 95.3%)</td>
</tr>
</tbody>
</table>

## Summary Table: Cologuard for Colorectal Cancer Screening

<table>
<thead>
<tr>
<th>Evidence base</th>
<th>Outcome</th>
<th>Risk of bias</th>
<th>Indirectness</th>
<th>Inconsistency</th>
<th>Imprecision</th>
<th>Evidence favors</th>
<th>Evidence grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 diagnostic cohort study</td>
<td>Sensitivity for CRC</td>
<td>Low</td>
<td>Direct</td>
<td>Unknown</td>
<td>Precise</td>
<td>Cologuard</td>
<td>Moderate</td>
</tr>
<tr>
<td>Cologuard vs. FIT</td>
<td>Sensitivity for advanced precancerous lesions</td>
<td>Low</td>
<td>Direct</td>
<td>Unknown</td>
<td>Precise</td>
<td>Cologuard</td>
<td>Moderate</td>
</tr>
<tr>
<td></td>
<td>Specificity for absence of CRC and advanced precancerous lesions</td>
<td>Low</td>
<td>Direct</td>
<td>Unknown</td>
<td>Precise</td>
<td>FIT</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

Summary

- Grading the evidence from diagnostic studies presents some unique challenges.
- Using GRADE for diagnostic accuracy outcomes requires a different approach than using GRADE for clinical outcomes.
- Assessing indirectness and imprecision is more complicated for diagnostic accuracy outcomes.
- However, the same GRADE domains should be used for intervention and diagnostic studies.
- Transparency of judgments in grading and the process of combining different domains for a summary grade is still important.
The Hospital Perspective:
Role of TA in Evaluating Dx Tech to Achieve Value-based Care

Joe Cummings, PhD
UHC Technology Assessment Program
cummings@uhc.edu

July 15, 2015
Disclaimers

I have no financial conflict of interest in any technologies discussed.

The assessments and opinions herein are my own and not affiliated with ECRI Institute or any other entity.

This presentation has been reviewed and contains no Protected Health Information.
Outline:

I. Technology Significance
II. Dx Evaluation Theory
III. Hospital Evaluation Paradigm
IV. Examples
V. Conclusions
Dx Technology Significance
## Hospital Costs—Principal Dx Breast Cancer (174.0 - 174.9)

<table>
<thead>
<tr>
<th>Service Group</th>
<th>Service</th>
<th>% Utilization</th>
<th>Mean Direct Cost (Cases using)</th>
<th>Mean Direct Cost (All cases)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accommodations</td>
<td>ICU</td>
<td>8.06</td>
<td>3,455</td>
<td>279</td>
</tr>
<tr>
<td></td>
<td>Other Accommodations</td>
<td>18.41</td>
<td>1,320</td>
<td>243</td>
</tr>
<tr>
<td></td>
<td>Routine Accommodations</td>
<td>94.71</td>
<td>1,505</td>
<td>1,425</td>
</tr>
<tr>
<td>Ancillary Services</td>
<td>Other Ancillary Services</td>
<td>15.42</td>
<td>165</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Physical Therapy</td>
<td>24.12</td>
<td>129</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Respiratory</td>
<td>11.90</td>
<td>214</td>
<td>26</td>
</tr>
<tr>
<td>Cardiac Dx Services</td>
<td>EKG/Telemetry</td>
<td>22.98</td>
<td>29</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Other Cardiac Services</td>
<td>4.60</td>
<td>224</td>
<td>10</td>
</tr>
<tr>
<td>Diagnostic Imaging</td>
<td>CT/MRI</td>
<td>11.25</td>
<td>287</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Nuclear Medicine</td>
<td>35.81</td>
<td>184</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>Other Diagnostic Imaging</td>
<td>9.31</td>
<td>210</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>X-Ray</td>
<td>31.17</td>
<td>106</td>
<td>33</td>
</tr>
<tr>
<td>Laboratory</td>
<td>Laboratory</td>
<td>98.72</td>
<td>484</td>
<td>477</td>
</tr>
<tr>
<td>Other Spec Dx Services</td>
<td>Other Spec Dx Svcs</td>
<td>15.02</td>
<td>151</td>
<td>23</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>Miscellaneous</td>
<td>38.74</td>
<td>207</td>
<td>80</td>
</tr>
<tr>
<td>Surgical Services</td>
<td>Anesthesia</td>
<td>76.92</td>
<td>425</td>
<td>327</td>
</tr>
<tr>
<td></td>
<td>Med. Surg. Supplies</td>
<td>94.54</td>
<td>3,161</td>
<td>2,989</td>
</tr>
<tr>
<td></td>
<td>OR Services</td>
<td>89.92</td>
<td>3,344</td>
<td>3,007</td>
</tr>
<tr>
<td></td>
<td>Other Surgical Services</td>
<td>83.12</td>
<td>435</td>
<td>361</td>
</tr>
<tr>
<td>Treatment</td>
<td>Blood</td>
<td>8.28</td>
<td>674</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>Dialysis</td>
<td>0.28</td>
<td>1,140</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Oncology &amp; Chemotherapy</td>
<td>0.49</td>
<td>76</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Other Treatment</td>
<td>2.60</td>
<td>934</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>Pharmacy &amp; IV Therapy</td>
<td>99.86</td>
<td>1,062</td>
<td>1,061</td>
</tr>
</tbody>
</table>

~7% of total costs Dx-related

Total cost = $10,653

## Hospital Costs - Principal Procedure Total Knee (81.54)

<table>
<thead>
<tr>
<th>Service Group</th>
<th>Service</th>
<th>Mean Direct Cost (Cases Using Service)</th>
<th>Mean Direct Cost (All Cases)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accommodations</td>
<td>ICU</td>
<td>3,120</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>Other Accommodations</td>
<td>1,034</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>Routine Accommodations</td>
<td>1,633</td>
<td>1,585</td>
</tr>
<tr>
<td>Ancillary Services</td>
<td>Other Ancillary Services</td>
<td>147</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>Physical Therapy</td>
<td>327</td>
<td>326</td>
</tr>
<tr>
<td></td>
<td>Respiratory</td>
<td>127</td>
<td>21</td>
</tr>
<tr>
<td>Cardiac Dx Services</td>
<td>EKG/Telemetry</td>
<td>26</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Other Cardiac Services</td>
<td>304</td>
<td>7</td>
</tr>
<tr>
<td>Diagnostic Imaging</td>
<td>CT/MRI</td>
<td>134</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Nuclear Medicine</td>
<td>315</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Other Diagnostic Imaging</td>
<td>100</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>X-Ray</td>
<td>52</td>
<td>37</td>
</tr>
<tr>
<td>Laboratory</td>
<td>Laboratory</td>
<td>150</td>
<td>149</td>
</tr>
<tr>
<td>Other Spec Dx Srvcs</td>
<td>Other Spec Dx Svcs</td>
<td>93</td>
<td>22</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>Miscellaneous</td>
<td>535</td>
<td>203</td>
</tr>
<tr>
<td>Surgical Services</td>
<td>Anesthesia</td>
<td>204</td>
<td>180</td>
</tr>
<tr>
<td></td>
<td>Med. Surg. Supplies</td>
<td>6,625</td>
<td>6,606</td>
</tr>
<tr>
<td></td>
<td>OR Services</td>
<td>1,932</td>
<td>1,896</td>
</tr>
<tr>
<td></td>
<td>Other Surgical Services</td>
<td>384</td>
<td>364</td>
</tr>
<tr>
<td>Treatment</td>
<td>Blood</td>
<td>395</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Dialysis</td>
<td>1,317</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Oncology &amp; Chemotherapy</td>
<td>47</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Other Treatment</td>
<td>86</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Pharmacy &amp; IV Therapy</td>
<td>710</td>
<td>710</td>
</tr>
</tbody>
</table>

Total cost = $12,391

~3.5% of total Costs Dx-related
Technology significance

Dx expense <10% of hospital costs

But, non-linearly affects other 90% of costs

Dx is like chaos theory…

Small changes

Large impacts

The butterfly effect
Dx Technology Evaluation Theory
Idealized evaluation paradigm

- Dx test
  - Screening
  - Diagnosis
  - Tx selection
  - Monitoring
  - Prognosis

Impact on outcomes

- Devices
- Procedures
- Diagnostics
- Drugs
Important outcomes

Improving outcomes in any of these tiers can create value

Dimensions

Survival

Degree of health or recovery

Time to recovery and time to return to normal activities

Disutility of care or treatment process (e.g., diagnostic errors, ineffective care, treatment-related discomfort, complications, adverse effects)

Sustainability of health or recovery and nature of recurrences

Long-term consequences of therapy (e.g., care-induced illnesses)
Looking inside the box….

Dx test

Impact on outcomes
Where are the outcomes measured?

Robust clinical outcomes are not always available...

Outcomes as defined by Fryback and Thornbury.*

<table>
<thead>
<tr>
<th>Level of Efficacy</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Technical efficacy</td>
<td>Performability, timing, interperatibility</td>
</tr>
<tr>
<td>2. Dx accuracy efficacy</td>
<td>Sensitivity/specificity</td>
</tr>
<tr>
<td>3. Dx thinking efficacy</td>
<td>Pre- and post-test changes in subjective outcome</td>
</tr>
<tr>
<td>4. Therapeutic efficacy</td>
<td>Effects of Dx on choice of Tx</td>
</tr>
<tr>
<td>5. Patient outcome efficacy</td>
<td>Value of test, including measure of morbidity, mortality, and QOL</td>
</tr>
<tr>
<td>6. Societal efficacy</td>
<td>Cost-benefit societal perspective</td>
</tr>
</tbody>
</table>

Qualitative acceptance of lesser evidence

Dimensions of an evidence strategy:

<table>
<thead>
<tr>
<th>At Risk Population</th>
<th>Anticipated Clinical Impact</th>
<th>Economic Impact</th>
<th>Level of Evidence (Fryback &amp; Thornbury)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>Large</td>
<td>Small</td>
<td>Level 1: Technical efficacy</td>
</tr>
<tr>
<td>Medium</td>
<td>Medium</td>
<td>Medium</td>
<td>Level 2: Diagnostic accuracy efficacy</td>
</tr>
<tr>
<td>Large</td>
<td>Small</td>
<td>Large</td>
<td>Level 3: Diagnostic thinking efficacy</td>
</tr>
<tr>
<td>Large</td>
<td>Small</td>
<td>Large</td>
<td>Level 4: Therapeutic efficacy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Level 5: Patient outcome efficacy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Level 6: Societal Efficacy</td>
</tr>
</tbody>
</table>

Level 3: FDG PET

Level 4: CT angiography

Level 5: CT colonography

**Developmental technology lifecycle**

It is always too early to assess a technology, until suddenly it is too late. - Buxton 1987

<table>
<thead>
<tr>
<th>Development</th>
<th>Introductory</th>
<th>Mature</th>
<th>Obsolescence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Establishment of technical</td>
<td>Comparison to alternatives</td>
<td>Multi-institutional studies</td>
<td>Replacement by newer technologies</td>
</tr>
<tr>
<td>parameters, algorithms, criteria</td>
<td>Assessment outside investigational settings</td>
<td>Outcomes/ Cost-effective</td>
<td></td>
</tr>
<tr>
<td>Quantification of performance in clinical settings</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Development**
  - Establishment of technical parameters, algorithms, criteria
  - Quantification of performance in clinical settings

- **Introductory**
  - Comparison to alternatives
  - Assessment outside investigational settings

- **Mature**
  - Multi-institutional studies
  - Outcomes/ Cost-effective

- **Obsolescence**
  - Replacement by newer technologies

- **FDA approval**

- **Time**

- **Usage**
Disruptive change is transforming the industry

Health Care Reform is creating a shift from Fee-for-Service to Value-based Care

What is value-based care?

Value-based Care (VBC) ↔ Value-based Purchasing (VBP)

**Value-based Purchasing** is a **Demand-side** strategy used by employers and the Federal government to maximize their market power as a force to promote quality and value of health care services - NBCH.org

Example: VBP as part of ACA [Section 3001(a)]

2015: 1.5% at risk

2016: 1.75% at risk

Source: CMS.gov: Hospital-based VBP program.
VBP means hospital pay is tied to performance

More and more payment is tied to outcomes

Solution to disruptive changes in healthcare

Focus on what the hospitals are getting paid for…Outcomes

Hospital supply chain perspective ➔ Link products to outcomes

Hospitals should be buying outcomes not devices…
How to “Buy Outcomes”

Determine level of evidence available/needed
• Accuracy, impact on process of care, patient outcomes

Pick the right outcomes…
• Avoid intermediate/surrogate endpoints

Manage data gaps

Options if there is no good data...
• Wait for more rigorous data
• Develop own data - provisional use w/ pilot
• Modeling - linked financial/clinical databases
• Expert opinion
  • Manufacturer submitted info/models

Determine the impact on the Health System…
• Cost, resource utilization, LOS, reimbursement
Value = outcomes relative to costs

Note: May be highly dependent on the choice of outcome and comparator.

Examples

- Lack of data
- Other factors at play
### Examples

<table>
<thead>
<tr>
<th>Cost</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
</tr>
</tbody>
</table>

**ADOPT**

In CEA, called dominant technologies

- **CO**
- **2**
- **BNP**
Examples

<table>
<thead>
<tr>
<th>Cost</th>
<th>Outcome</th>
</tr>
</thead>
</table>
| +    | +       | Disruptive technologies
| -    | -       |
| +    | -       | ↑ availability
| -    | +       | ↑ convenience
| -    | +       | ↓ decision time
| +    | +       | ↑ ease of use
Examples

- OUTCOME
- COST

Decision Threshold

Majority of devices fit into this category
Better outcomes, higher costs
Progress, but can we afford it.

Conclusions
Take home message

Start or solidify an internal TA process

Use evidence-based analysis

Focus on buying outcomes

Base decisions on value
Decision Trees for Diagnostic Tests

David Samson, MS
Senior Associate Director, HTA & EPC Group
Learning Objectives

▶ Understand how analytic frameworks and estimates of diagnostic performance contribute to construction of decision trees

▶ Distinguish between different uses of tests (replacement, add-on, triage) and how these uses affect the structure of decision trees

▶ Learn how to apply an algorithm to decide whether modeling should be part of the evaluation of a test
Analytic Frameworks

Graphical representation that:

- Starts from point of testing/intervention
- Leads through surrogate/intermediate outcomes to all important clinical/health outcomes
- Reveals potential Key Questions
- Developed with stakeholders
- Strengthens systematic review process by:
  - Clarifying links between testing, intermediate outcomes, health outcomes
  - Clarifying Key Questions, illustrating decisional dilemmas
Establishing Clinical Context: PICO(TS)

- Crucial for planning meaningful decision analysis
- PICOTS define the clinical context (Population, Intervention, Comparator, Outcomes, Timing) and
  - Setting (screening, diagnosis, treatment guidance, patient monitoring, or prognosis)
- Also: role/use of test (replacement, add-on, triage)

Replacement, Add-on, Triage Tests

- Replacement: new test replaces existing test
  - New test alone instead of existing test:
    - ↑ diagnostic accuracy (at least = on 1 index, > on other)
      - Plus ↓ lower costs, ↓ invasiveness
Replacement, Add-on, Triage Tests

► Add-on: combine new test with existing test

■ Two tests vs one test: ↑ diagnostic accuracy
  ▶ ↑ sensitivity, either test positive rule
  ▶ ↑ specificity, both tests positive rule

▶ Threshold costs/tradeoffs:

□ ↑ sensitivity → ↓ specificity
□ ↑ specificity → ↓ sensitivity
Replacement, Add-on, Triage Tests

- Triage: new test determines who undergoes existing test

  - Decision rules
    - New test positive → do existing test
    - New test negative → do existing test

  - Not to ↑ diagnostic accuracy, but ↓ invasive/costly testing
Background: Why is Decision Analysis Needed?

- Many systematic reviews focus only on test performance (limited literature)

- Test performance not sufficient to assess usefulness
  - Complex links between testing, test results, and patient outcomes (analytic framework)
  - Uncertainty:
    - doctors may not act on test results,
    - patients may not follow recommendations, and
    - interventions may not lead to a benefit

- Studies comparing test-and-treat strategies ideal but rare

- Need to assemble evidence from different sources
Background

▶ Modeling (decision/economic/cost-effectiveness analysis) can:
  - Link evidence from different sources
  - Explore impact of uncertainty
  - Make assumptions clear
  - Evaluate tradeoffs in benefits, harms, and costs
  - Assess multiple test-and-treat strategy comparisons without direct evidence
  - Explore hypothetical scenarios

▶ Modeling links testing to patient outcomes, aids understanding, aids interpreting systematic reviews of medical tests
What is Decision Modeling?

A model is a “simplified representation of reality that captures some of that reality’s essential properties and relationships (e.g. logical, quantitative, cause/effect)“. (Stahl Phamacoeconomics 2008 26(2):131)
What is Decision Modeling?

- Types of models:
  - decision trees,
  - state-transition models (STMs, e.g., Markov models),
  - discrete event simulations (DESs),
  - dynamic transition models,
  - agent-based models (Archimedes),
  - combination models and
  - hybrid models
Decision Trees

- Intended for modeling relatively simple problems over short time horizons
- Defined by:
  - square decision nodes,
  - branches,
  - strategies,
  - circular chance nodes (probabilities)
  - triangular terminal nodes
  - payoffs: life expectancies, costs, utilities (0-1)
  - evaluation of the tree by folding back process, producing expected values for each strategy, facilitating choice
Diagnostic Accuracy Indices

- Sensitivity (positive in disease, set is disease present)
- Specificity (negative in health, set is disease absent)
- Positive predictive value (PPV, diseased if positive, set is test positive)
- Negative predictive value (NPV, healthy if negative, set is test negative)
Decision Trees

Test 1
- Positive test
  - True positive (disease present)
  - False positive (disease absent)
- Negative test
  - True negative (disease absent)
  - False negative (disease present)

Test 2
- Positive test
  - True positive (disease present)
  - False positive (disease absent)
- Negative test
  - True negative (disease absent)
  - False negative (disease present)
Decision Trees

Decision node

Test 1
- Positive test: $p(T+)$
  - Positive predictive value: $\frac{TP}{TP + FP}$
  - 1 - Positive predictive value: $\frac{TN}{TN + FN}$
- Negative test: $p(T-)$
  - Negative predictive value: $\frac{TN}{TN + FN}$
  - 1 - Negative predictive value: $\frac{TP}{TP + FP}$

Test 2
- Positive test: $p(T+)$
  - Positive predictive value: $\frac{TP}{TP + FP}$
  - 1 - Positive predictive value: $\frac{TN}{TN + FN}$
- Negative test: $p(T-)$
  - Negative predictive value: $\frac{TN}{TN + FN}$
  - 1 - Negative predictive value: $\frac{TP}{TP + FP}$
Decision Trees

- **Decision node**
  - **Test 1**
    - **Disease present**
      - True positive (test positive)
      - False negative (test negative)
    - **Disease absent**
      - True negative (test negative)
      - False positive (test positive)
  - **Test 2**
    - **Disease present**
      - True positive (test positive)
      - False negative (test negative)
    - **Disease absent**
      - True negative (test negative)
      - False positive (test positive)
Decision Trees

Terminal branch quality of life payoff → utility: 0 (immediate death) to 1 (perfect health)
Example: Replacement Test

Sydney Breast Imaging Accuracy Study: Comparative Sensitivity and Specificity of Mammography and Sonography in Young Women with Symptoms

AJR: 180, April 2003

Nehmat Houssami 1-3
Les Irwig 2
Judy M. Simpson 2
Merran McKessar 1,4
Steven Blome 1,5
Jennie Noakes 4
Replacement Test PICO

- **P:** women with palpable breast masses

- **I:** test-and-treat strategy 1: ultrasonography, downstream tests/treatments and outcomes

- **C:** test-and-treat strategy 2: mammography, downstream tests/treatments and outcomes

- **O:** direct test-related outcomes (discomfort, anxiety), indirect test/treatment decision-related outcomes
## Breast Cancer Diagnosis

<table>
<thead>
<tr>
<th>Test positive</th>
<th>Reference standard positive</th>
<th>Reference standard negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive:</td>
<td>Test+, breast cancer present</td>
<td>False positive:</td>
</tr>
<tr>
<td></td>
<td>Receive needed treatment</td>
<td>Test+, breast cancer absent</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Receive unneeded procedures</td>
</tr>
<tr>
<td>False negative:</td>
<td>False negative:</td>
<td>True negative:</td>
</tr>
<tr>
<td></td>
<td>Test-, breast cancer present</td>
<td>Test-, breast cancer absent</td>
</tr>
<tr>
<td></td>
<td>Forgo/delay needed treatment</td>
<td>Avoid unneeded procedures</td>
</tr>
</tbody>
</table>
Replacement Test Decision Tree

**Decision node**

- **Mammography**
  - **Positive**
    - **True positive** → Needed treatment outcomes
    - **False positive** → Unneeded procedures outcomes
  - **Negative**
    - **True negative** → Avoid unneeded procedures outcomes
    - **False negative** → Forgo needed treatment outcomes

- **Ultrasonography**
  - **Positive**
    - **True positive** → Needed treatment outcomes
    - **False positive** → Unneeded procedures outcomes
  - **Negative**
    - **True negative** → Avoid unneeded procedures outcomes
    - **False negative** → Forgo needed treatment outcomes
<table>
<thead>
<tr>
<th></th>
<th>RS+</th>
<th>RS-</th>
<th>Sens</th>
<th>Spec</th>
<th>PPV</th>
<th>1-PPV</th>
<th>NPV</th>
<th>1-NPV</th>
<th>Prev Ca</th>
<th>Prev nCa</th>
<th>Prev MM+</th>
<th>Prev MM-</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM+</td>
<td>182</td>
<td>29</td>
<td>0.7583</td>
<td>0.8755</td>
<td>0.8626</td>
<td>0.1374</td>
<td>0.7786</td>
<td>0.2214</td>
<td>0.5074</td>
<td>0.4926</td>
<td>0.4461</td>
<td>0.5539</td>
</tr>
<tr>
<td>MM-</td>
<td>58</td>
<td>204</td>
<td>0.7786</td>
<td>0.4736</td>
<td>0.1767</td>
<td>0.8233</td>
<td>0.1250</td>
<td>0.8750</td>
<td>0.8167</td>
<td>0.8798</td>
<td>0.8750</td>
<td>0.1250</td>
</tr>
<tr>
<td>US+</td>
<td>196</td>
<td>28</td>
<td>0.8167</td>
<td>0.8798</td>
<td>0.8750</td>
<td>0.1250</td>
<td>0.8233</td>
<td>0.1767</td>
<td>0.5074</td>
<td>0.4926</td>
<td>0.4736</td>
<td>0.5264</td>
</tr>
<tr>
<td>US-</td>
<td>44</td>
<td>205</td>
<td>0.8626</td>
<td>0.4461</td>
<td>0.1374</td>
<td>0.8750</td>
<td>0.1250</td>
<td>0.8233</td>
<td>0.1767</td>
<td>0.7786</td>
<td>0.2214</td>
<td>0.5539</td>
</tr>
</tbody>
</table>

**Decision node**

- **Mammography**
  - Positive: 0.4461, 0.8626, 0.85, needed treatment
  - False positive: 0.1374, 0.90, unneeded procedures
  - True negative: 0.7786, 1.00, avoid unneeded procedures
  - False negative: 0.2214, 0.75, forgo needed treatment
- **Ultrasonography**
  - Positive: 0.4736, 0.8750, 0.85, needed treatment
  - False positive: 0.1250, 0.90, unneeded procedures
  - True negative: 0.8233, 1.00, avoid unneeded procedures
  - False negative: 0.1767, 0.75, forgo needed treatment

ECRI Institute
The Discipline of Science. The Integrity of Independence.

©2015 ECRI INSTITUTE
<table>
<thead>
<tr>
<th>Decision node</th>
<th>Mammography</th>
<th>Ultrasonography</th>
</tr>
</thead>
</table>
| **EU(MM+)**  | U(TP) x p(TP) + U(FP) x p(FP) | EU(MM+) x p(MM+) + EU(MM-) x p(MM-)
|              | 0.85 x 0.8626 + 0.90 x 0.1374 | 0.8569 x 0.4461 + 0.9447 x 0.5539 |
|              | 0.8569                                      | 0.9055 |
| **EU(MM-)**  | U(TN) x p(TN) + U(FN) x p(FN) | EU(MM-) x p(MM-) + EU(MM) x p(MM+)
|              | 1.00 x 0.7786 + 0.75 x 0.2214 | 1.00 x 0.75 + 0.75 x 0.25 |
|              | 0.9447                                      | 0.9447 |
| **EU(US+)**  | U(TP) x p(TP) + U(FP) x p(FP) | EU(US+) x p(US+) + EU(US-) x p(US-)
|              | 0.85 x 0.8750 + 0.90 x 0.1250 | 0.8563 x 0.4736 + 0.9558 x 0.5264 |
|              | 0.8563                                      | 0.9087 |
| **EU(US-)**  | U(TN) x p(TN) + U(FN) x p(FN) | EU(US-) x p(US-) + EU(US) x p(US+)
|              | 1.00 x 0.8233 + 0.75 x 0.1767 | 1.00 x 0.75 + 0.75 x 0.25 |
|              | 0.9558                                      | 0.9558 |

**Utilities**

**Probabilities**

- Positive: True positive
- False positive
- True negative
- False negative

**Expected utilities**

- Mammography
  - Positive: 0.8626, needed treatment
  - Negative: 0.5539
- Ultrasonography
  - Positive: 0.8750, needed treatment
  - Negative: 0.5364
<table>
<thead>
<tr>
<th>Decision node</th>
<th>Mammography</th>
<th>Ultrasonography</th>
</tr>
</thead>
<tbody>
<tr>
<td>EU(MM+)</td>
<td>0.85 x 0.8626 + 0.90 x 0.1374</td>
<td>EU(MM) = EU(MM+) x p(MM+) + EU(MM-) x p(MM-)</td>
</tr>
<tr>
<td>EU(MM-)</td>
<td>1.00 x 0.7786 + 0.75 x 0.2214</td>
<td>= 0.9055</td>
</tr>
<tr>
<td>EU(US+)</td>
<td>0.85 x 0.8750 + 0.90 x 0.1250</td>
<td>EU(US) = EU(US+) x p(US+) + EU(US-) x p(US-)</td>
</tr>
<tr>
<td>EU(US-)</td>
<td>1.00 x 0.8233 + 0.75 x 0.1767</td>
<td>= 0.9087</td>
</tr>
</tbody>
</table>

**EU(MM+)** = 0.85 x 0.8626 + 0.90 x 0.1374 = 0.8569

**EU(MM-)** = 1.00 x 0.7786 + 0.75 x 0.2214 = 0.9447

**EU(US+)** = 0.85 x 0.8750 + 0.90 x 0.1250 = 0.8563

**EU(US-)** = 1.00 x 0.8233 + 0.75 x 0.1767 = 0.9558

**Positive**
- True positive: 0.8626, **0.85**, needed treatment
- False positive: 0.1374, **0.90**, unneeded procedures

**Negative**
- True negative: 0.7786, **1.00**, avoid unneeded procedures
- False negative: 0.2214, **0.75**, forgo needed treatment

**Positive**
- True positive: 0.8750, **0.85**, needed treatment
- False positive: 0.1250, **0.90**, unneeded procedures

**Negative**
- True negative: 0.8233, **1.00**, avoid unneeded procedures
- False negative: 0.1767, **0.75**, forgo needed treatment
<table>
<thead>
<tr>
<th>Decision node</th>
<th>Mammography</th>
<th>Ultrasonography</th>
</tr>
</thead>
<tbody>
<tr>
<td>EU(MM+)</td>
<td>$0.85 \times 0.8626 + 0.90 \times 0.1374$</td>
<td>$0.8569 \times 0.4461 + 0.9447 \times 0.5539$</td>
</tr>
<tr>
<td>EU(MM-)</td>
<td>$1.00 \times 0.7786 + 0.75 \times 0.2214$</td>
<td>$0.9055$</td>
</tr>
<tr>
<td>EU(US+)</td>
<td>$0.85 \times 0.8750 + 0.90 \times 0.1250$</td>
<td>$0.8563 \times 0.4736 + 0.9558 \times 0.5264$</td>
</tr>
<tr>
<td>EU(US-)</td>
<td>$1.00 \times 0.8233 + 0.75 \times 0.1767$</td>
<td>$0.9558$</td>
</tr>
</tbody>
</table>

**Mammography**

- **Positive**
  - True positive: $0.85$, needed treatment
  - False positive: $0.90$, unneeded procedures
- **Negative**
  - True negative: $1.00$, avoid unneeded procedures
  - False negative: $0.75$, forgo needed treatment

**Ultrasonography**

- **Positive**
  - True positive: $0.85$, needed treatment
  - False positive: $0.90$, unneeded procedures
- **Negative**
  - True negative: $1.00$, avoid unneeded procedures
  - False negative: $0.75$, forgo needed treatment
### Decision Node: Mammography

<table>
<thead>
<tr>
<th>EU(MM+)</th>
<th>=</th>
<th>U(TP) x p(TP) + U(FP) x p(FP)</th>
<th></th>
<th>EU(MM)</th>
<th>=</th>
<th>EU(MM+) x p(MM+) + EU(MM-) x p(MM-)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.85 x 0.8626 + 0.90 x 0.1374</td>
<td></td>
<td></td>
<td></td>
<td>0.8569 x 0.4461 + 0.9447 x 0.5539</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.8569</td>
<td></td>
<td></td>
<td></td>
<td>0.9055</td>
</tr>
</tbody>
</table>

### Decision Node: Ultrasonography

<table>
<thead>
<tr>
<th>EU(US+)</th>
<th>=</th>
<th>U(TP) x p(TP) + U(FP) x p(FP)</th>
<th></th>
<th>EU(US)</th>
<th>=</th>
<th>EU(US+) x p(US+) + EU(US-) x p(US-)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.85 x 0.8750 + 0.90 x 0.1250</td>
<td></td>
<td></td>
<td></td>
<td>0.8563 x 0.4736 + 0.9558 x 0.5264</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.8563</td>
<td></td>
<td></td>
<td></td>
<td>0.9087</td>
</tr>
</tbody>
</table>

### Decision Node: 

- **Positive**
  - True positive: 0.8626, needed treatment
  - False positive: 0.90, unneeded procedures
- **Negative**
  - True negative: 1.00, avoid unneeded procedures
  - False negative: 0.75, forgo needed treatment

### Decision Node: 

- **Positive**
  - True positive: 0.8750, needed treatment
  - False positive: 0.90, unneeded procedures
- **Negative**
  - True negative: 1.00, avoid unneeded procedures
  - False negative: 0.75, forgo needed treatment
### Decision Node: Mammography vs. Ultrasonography

#### Mammography
- **True positive (TP)**: 0.8626
- **False positive (FP)**: 0.1374
- **Expected Utility (EU)**:
  - **EU(MM+) = 0.85 x 0.8626 + 0.90 x 0.1374 = 0.8569**
  - **EU(MM-) = 1.00 x 0.7786 + 0.75 x 0.2214 = 0.9447**

#### Ultrasonography
- **True positive (TP)**: 0.8750
- **False positive (FP)**: 0.1250
- **Expected Utility (EU)**:
  - **EU(US+) = 0.85 x 0.8750 + 0.90 x 0.1250 = 0.8563**
  - **EU(US-) = 1.00 x 0.8233 + 0.75 x 0.1767 = 0.9558**
<table>
<thead>
<tr>
<th>Decision node</th>
<th>Mammography</th>
<th>Ultrasonography</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EU(MM+)</strong> = &amp; U(TP) x p(TP) + &amp; U(FP) x p(FP) &amp; EU(MM) = &amp; EU(MM+) x p(MM+) + &amp; EU(MM-) x p(MM-)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>= &amp; 0.85 x 0.8626 &amp; + &amp; 0.90 x 0.1374 &amp; = &amp; 0.8569 &amp; 0.4461 &amp; + &amp; 0.9447 &amp; 0.5539</td>
<td></td>
<td></td>
</tr>
<tr>
<td>= &amp; 0.8569 &amp; &amp; &amp; &amp; = &amp; 0.9055 &amp;</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>EU(MM-)</strong> = &amp; U(TN) x p(TN) + &amp; U(FN) x p(FN) &amp; = &amp; 1.00 x 0.7786 &amp; + &amp; 0.75 x 0.2214 &amp; = &amp; 0.9447 &amp;</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>EU(US+)</strong> = &amp; U(TP) x p(TP) + &amp; U(FP) x p(FP) &amp; EU(US) = &amp; EU(US+) x p(US+) + &amp; EU(US-) x p(US-)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>= &amp; 0.85 x 0.8750 &amp; + &amp; 0.90 x 0.1250 &amp; = &amp; 0.8563 &amp; 0.4736 &amp; + &amp; 0.9558 &amp; 0.5264</td>
<td></td>
<td></td>
</tr>
<tr>
<td>= &amp; 0.8563 &amp; &amp; &amp; &amp; = &amp; 0.9087 &amp;</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>EU(US-)</strong> = &amp; U(TN) x p(TN) + &amp; U(FN) x p(FN) &amp; = &amp; 1.00 x 0.8233 &amp; + &amp; 0.75 x 0.1767 &amp; = &amp; 0.9558 &amp;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **True positive** for mammography: 0.8569 (needed treatment)
- **False positive** for mammography: 0.4461 (unneeded procedures)
- **True negative** for mammography: 0.5539 (avoids unneeded procedures)
- **False negative** for mammography: 0.1374 (forgo needed treatment)

- **True positive** for ultrasonography: 0.4736 (needed treatment)
- **False positive** for ultrasonography: 0.0875 (unneeded procedures)
- **True negative** for ultrasonography: 0.5364 (avoids unneeded procedures)
- **False negative** for ultrasonography: 0.1767 (forgo needed treatment)
### Decision Node: Mammography

- **False Positive Rate:** 0.1374 (90%)
- **True Positive Rate:** 0.8626 (85%)
- **Avoid Unneeded Procedures:** 1.00
- **Forgo Needed Treatment:** 0.75
- **True Positive:** 0.8569
- **False Positive:** 0.4461
- **EU(MM+):** 0.90
- **EU(MM-):** 0.75

### Decision Node: Ultrasonography

- **False Positive Rate:** 0.1250 (90%)
- **True Positive Rate:** 0.8750 (85%)
- **Avoid Unneeded Procedures:** 1.00
- **Forgo Needed Treatment:** 0.75
- **True Positive:** 0.8563
- **False Positive:** 0.4736
- **EU(US+):** 0.9558
- **EU(US-):** 0.9087

---

**EU(MM+):**

\[
EU(MM+) = U(TP) \times p(TP) + U(FP) \times p(FP)
\]

\[
= 0.85 \times 0.8626 + 0.90 \times 0.1374
\]

\[
= 0.8569
\]

**EU(MM-):**

\[
EU(MM-) = U(TN) \times p(TN) + U(FN) \times p(FN)
\]

\[
= 1.00 \times 0.7786 + 0.75 \times 0.2214
\]

\[
= 0.9447
\]

**EU(US+):**

\[
EU(US+) = U(TP) \times p(TP) + U(FP) \times p(FP)
\]

\[
= 0.85 \times 0.8750 + 0.90 \times 0.1250
\]

\[
= 0.8563
\]

**EU(US-):**

\[
EU(US-) = U(TN) \times p(TN) + U(FN) \times p(FN)
\]

\[
= 1.00 \times 0.8233 + 0.75 \times 0.1767
\]

\[
= 0.9558
\]
<table>
<thead>
<tr>
<th>EU(MM+)</th>
<th>U(TP) x p(TP) + U(FP) x p(FP)</th>
<th>EU(MM) = EU(MM+) x p(MM+) + EU(MM-) x p(MM-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>= 0.85</td>
<td>x 0.8626 + 0.90 x 0.1374</td>
<td>= 0.8569 x 0.4461 + 0.9447 x 0.5539</td>
</tr>
<tr>
<td>= 0.8569</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EU(MM-)</td>
<td>U(TN) x p(TN) + U(FN) x p(FN)</td>
<td>= 1.00 x 0.7786 + 0.75 x 0.2214</td>
</tr>
<tr>
<td>= 0.9447</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EU(US+)</td>
<td>U(TP) x p(TP) + U(FP) x p(FP)</td>
<td>EU(US) = EU(US+) x p(US+) + EU(US-) x p(US-)</td>
</tr>
<tr>
<td>= 0.85</td>
<td>x 0.8750 + 0.90 x 0.1250</td>
<td>= 0.8563 x 0.4736 + 0.9558 x 0.5264</td>
</tr>
<tr>
<td>= 0.8563</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EU(US-)</td>
<td>U(TN) x p(TN) + U(FN) x p(FN)</td>
<td>= 1.00 x 0.8233 + 0.75 x 0.1767</td>
</tr>
<tr>
<td>= 0.9558</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Decision node**
- **Mammography**
  - **Positive**
    - True positive: 0.8569
    - False positive: 0.9447
  - **Negative**
    - True negative: 0.9558

- **Ultrasonography**
  - **Positive**
    - True positive: 0.8563
    - False positive: 0.4461
  - **Negative**
    - True negative: 0.5364

True positive: 0.85, needed treatment
False positive: 0.90, unneeded procedures
True negative: 1.00, avoid unneeded procedures
False negative: 0.75, forgo needed treatment

True positive: 0.8750
False positive: 0.90, unneeded procedures
True negative: 1.00, avoid unneeded procedures
False negative: 0.75, forgo needed treatment

True positive: 0.4736
False positive: 0.1250
True negative: 0.9558
False negative: 0.8233

**ECRI Institute**
The Discipline of Science. The Integrity of Independence.
<table>
<thead>
<tr>
<th>EU(MM+)</th>
<th>U(TP) x p(TP) + U(FP) x p(FP)</th>
<th>EU(MM)</th>
<th>EU(MM+) x p(MM+) + EU(MM-) x p(MM-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.85</td>
<td>0.8626</td>
<td>0.8569</td>
<td>0.85 x 0.8626 + 0.90 x 0.1374</td>
</tr>
<tr>
<td>0.90</td>
<td>0.1374</td>
<td>0.4461</td>
<td>0.90 x 0.1374 + 0.75 x 0.2214</td>
</tr>
<tr>
<td>0.8569</td>
<td>0.4461</td>
<td>0.9055</td>
<td>0.85 x 0.8626 + 0.90 x 0.1374</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EU(MM-)</th>
<th>U(TN) x p(TN) + U(FN) x p(FN)</th>
<th>EU(MM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>0.7786</td>
<td>0.9447</td>
</tr>
<tr>
<td>0.75</td>
<td>0.2214</td>
<td></td>
</tr>
<tr>
<td>0.9447</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EU(US+)</th>
<th>U(TP) x p(TP) + U(FP) x p(FP)</th>
<th>EU(US)</th>
<th>EU(US+) x p(US+) + EU(US-) x p(US-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.85</td>
<td>0.8750</td>
<td>0.8563</td>
<td>0.85 x 0.8750 + 0.90 x 0.1250</td>
</tr>
<tr>
<td>0.90</td>
<td>0.1250</td>
<td>0.4736</td>
<td>0.90 x 0.1250 + 0.75 x 0.1767</td>
</tr>
<tr>
<td>0.8563</td>
<td>0.4736</td>
<td>0.9087</td>
<td>0.85 x 0.8750 + 0.90 x 0.1250</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EU(US-)</th>
<th>U(TN) x p(TN) + U(FN) x p(FN)</th>
<th>EU(MM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>0.8233</td>
<td>0.9558</td>
</tr>
<tr>
<td>0.75</td>
<td>0.1767</td>
<td></td>
</tr>
<tr>
<td>0.9558</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EU(MM+)</td>
<td>EU(MM)</td>
<td>EU(MM-)</td>
</tr>
<tr>
<td>---------</td>
<td>--------</td>
<td>---------</td>
</tr>
<tr>
<td>U(TP)</td>
<td>U(TN)</td>
<td>U(TP)</td>
</tr>
<tr>
<td>0.85</td>
<td>0.8569</td>
<td>0.0.8569</td>
</tr>
<tr>
<td>x p(TP)</td>
<td>x p(TN)</td>
<td>x p(TP)</td>
</tr>
<tr>
<td>0.8626</td>
<td>0.8750</td>
<td>0.90</td>
</tr>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>0.90</td>
<td>0.90</td>
<td>0.1374</td>
</tr>
<tr>
<td>0.1374</td>
<td>0.1374</td>
<td>0.1374</td>
</tr>
</tbody>
</table>

EU(MM) = EU(MM+) x p(MM+) + EU(MM-) x p(MM-)

EU(MM+) = 0.85 x 0.8626 + 0.90 x 0.1374 = 0.8569

EU(MM-) = 1.00 x 0.7786 + 0.75 x 0.2214 = 0.9447

EU(US+) = 0.85 x 0.8750 + 0.90 x 0.1250 = 0.8563

EU(US-) = 1.00 x 0.8233 + 0.75 x 0.1767 = 0.9558

EU(US+): True positive
0.85, needed treatment
0.90, unneeded procedures
0.85, forgo needed treatment
0.90, unneeded procedures
1.00, avoid unneeded procedures

EU(US-): True negative
0.75, forgo needed treatment
0.90, unneeded procedures
1.00, avoid unneeded procedures

Decision node: Mammography

Positive
0.8569
True positive
0.85, needed treatment
0.90, unneeded procedures

Negative
0.5539
False negative
0.75, forgo needed treatment

Decision node: Ultrasonography

Positive
0.4736
True positive
0.85, needed treatment
0.90, unneeded procedures

Negative
0.5364
False negative
0.75, forgo needed treatment

ECRI Institute
The Discipline of Science. The Integrity of Independence.

©2015 ECRI INSTITUTE
## Decision Node Evaluation

### Mammography

<table>
<thead>
<tr>
<th>EU(MM+)</th>
<th>U(TP) × p(TP) + U(FP) × p(FP)</th>
<th>EU(MM)</th>
<th>EU(MM+) × p(MM+) + EU(MM-) × p(MM-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.85</td>
<td>0.8569</td>
<td>0.8569</td>
<td>0.85 × 0.8626 + 0.90 × 0.1374</td>
</tr>
<tr>
<td>0.1374</td>
<td></td>
<td></td>
<td>0.90 × 0.1374</td>
</tr>
<tr>
<td>0.90</td>
<td></td>
<td>0.9055</td>
<td>0.8569 × 0.8626 + 0.9055 × 0.7786</td>
</tr>
<tr>
<td>0.7786</td>
<td></td>
<td></td>
<td>0.9055 × 0.7786</td>
</tr>
<tr>
<td>0.2214</td>
<td></td>
<td></td>
<td>0.9055 × 0.2214</td>
</tr>
<tr>
<td>0.8626</td>
<td></td>
<td></td>
<td>0.8569 × 0.8626</td>
</tr>
<tr>
<td>0.4461</td>
<td></td>
<td></td>
<td>0.8569 × 0.4461</td>
</tr>
<tr>
<td>0.1374</td>
<td></td>
<td></td>
<td>0.9055 × 0.1374</td>
</tr>
<tr>
<td>0.7786</td>
<td></td>
<td></td>
<td>0.9055 × 0.7786</td>
</tr>
<tr>
<td>0.2214</td>
<td></td>
<td></td>
<td>0.9055 × 0.2214</td>
</tr>
<tr>
<td>0.4461</td>
<td></td>
<td></td>
<td>0.8569 × 0.4461</td>
</tr>
<tr>
<td>0.1374</td>
<td></td>
<td></td>
<td>0.9055 × 0.1374</td>
</tr>
<tr>
<td>0.7786</td>
<td></td>
<td></td>
<td>0.9055 × 0.7786</td>
</tr>
<tr>
<td>0.2214</td>
<td></td>
<td></td>
<td>0.9055 × 0.2214</td>
</tr>
<tr>
<td>0.4736</td>
<td></td>
<td>0.8563</td>
<td>0.8563 × 0.8750 + 0.90 × 0.1250</td>
</tr>
<tr>
<td>0.1250</td>
<td></td>
<td>0.9558</td>
<td>0.9558 × 0.75</td>
</tr>
<tr>
<td>0.75</td>
<td>forgo needed treatment</td>
<td></td>
<td>0.75 × 0.8233 + 0.75 × 0.1767</td>
</tr>
<tr>
<td>0.1767</td>
<td></td>
<td></td>
<td>0.9558 × 0.1767</td>
</tr>
</tbody>
</table>

### Ultrasonography

<table>
<thead>
<tr>
<th>EU(US+)</th>
<th>U(TP) × p(TP) + U(FP) × p(FP)</th>
<th>EU(US)</th>
<th>EU(US+) × p(US+) + EU(US-) × p(US-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.85</td>
<td>0.8563</td>
<td>0.8563</td>
<td>0.85 × 0.8750 + 0.90 × 0.1250</td>
</tr>
<tr>
<td>0.1250</td>
<td></td>
<td>0.9558</td>
<td>0.9558 × 0.75</td>
</tr>
<tr>
<td>0.75</td>
<td>forgo needed treatment</td>
<td></td>
<td>0.75 × 0.8233 + 0.75 × 0.1767</td>
</tr>
<tr>
<td>0.1767</td>
<td></td>
<td></td>
<td>0.9558 × 0.1767</td>
</tr>
<tr>
<td>0.8750</td>
<td></td>
<td>0.9087</td>
<td>0.9087 × 0.8233 + 0.75 × 0.1767</td>
</tr>
<tr>
<td>0.1250</td>
<td></td>
<td>0.9558</td>
<td>0.9558 × 0.75</td>
</tr>
<tr>
<td>0.75</td>
<td>forgo needed treatment</td>
<td></td>
<td>0.75 × 0.8233 + 0.75 × 0.1767</td>
</tr>
<tr>
<td>0.1767</td>
<td></td>
<td></td>
<td>0.9558 × 0.1767</td>
</tr>
</tbody>
</table>

**Decision Node:**

- **Mammography**
  - Positive: 0.8569 (True positive: 0.85, needed treatment; False positive: 0.90, unneeded procedures)
  - Negative: 0.9447 (True negative: 1.00, avoid unneeded procedures; False negative: 0.75, forgo needed treatment)

- **Ultrasonography**
  - Positive: 0.8563 (True positive: 0.85, needed treatment; False positive: 0.90, unneeded procedures)
  - Negative: 0.9558 (True negative: 1.00, avoid unneeded procedures; False negative: 0.75, forgo needed treatment)
<table>
<thead>
<tr>
<th>EU(MM+)</th>
<th>EU(MM)</th>
<th>EU(MM-)</th>
<th>EU(US+)</th>
<th>EU(US)</th>
<th>EU(US-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ EU(MM+) = U(TP) \times p(TP) + U(FP) \times p(FP) ]</td>
<td>[ EU(MM) = EU(MM+) \times p(MM+) + EU(MM-) \times p(MM-) ]</td>
<td>[ EU(MM-) = U(TN) \times p(TN) + U(FN) \times p(FN) ]</td>
<td>[ EU(US+) = U(TP) \times p(TP) + U(FP) \times p(FP) ]</td>
<td>[ EU(US) = EU(US+) \times p(US+) + EU(US-) \times p(US-) ]</td>
<td>[ EU(US-) = U(TN) \times p(TN) + U(FN) \times p(FN) ]</td>
</tr>
<tr>
<td>[ = 0.85 \times 0.8626 + 0.90 \times 0.1374 ]</td>
<td>[ = 0.8569 ]</td>
<td>[ = 1.00 \times 0.7786 + 0.75 \times 0.2214 ]</td>
<td>[ = 0.85 \times 0.8750 + 0.90 \times 0.1250 ]</td>
<td>[ = 0.8563 ]</td>
<td>[ = 1.00 \times 0.8233 + 0.75 \times 0.1767 ]</td>
</tr>
<tr>
<td>[ = 0.8569 ]</td>
<td>[ = 0.9555 ]</td>
<td>[ = 0.9447 ]</td>
<td>[ = 0.8563 ]</td>
<td>[ = 0.9087 ]</td>
<td>[ = 0.9558 ]</td>
</tr>
</tbody>
</table>

- **Mammography**
  - **Positive**
    - True positive: 0.8626, needed treatment (0.85)
    - False positive: 0.1374, unneeded procedures (0.90)
  - **Negative**
    - True negative: 0.7786, avoid unneeded procedures (1.00)
    - False negative: 0.2214, forgo needed treatment (0.75)

- **Ultrasonography**
  - **Positive**
    - True positive: 0.8750, needed treatment (0.85)
    - False positive: 0.1250, unneeded procedures (0.90)
  - **Negative**
    - True negative: 0.8233, avoid unneeded procedures (1.00)
    - False negative: 0.1767, forgo needed treatment (0.75)
### Decision node

<table>
<thead>
<tr>
<th></th>
<th>EU(MM+)</th>
<th>EU(MM-)</th>
<th>EU(US+)</th>
<th>EU(US-)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EU(MM+)</strong></td>
<td>$0.85 \times p(TP) + 0.90 \times p(FP)$</td>
<td>$1.00 \times p(TN) + 0.75 \times p(FN)$</td>
<td>$0.85 \times 0.8750 + 0.90 \times 0.1250$</td>
<td>$1.00 \times 0.8233 + 0.75 \times 0.1767$</td>
</tr>
<tr>
<td><strong>EU(MM-)</strong></td>
<td>$0.1374 \times p(TP) + 0.1374 \times p(FP)$</td>
<td>$0.75 \times p(TN) + 0.2214 \times p(FN)$</td>
<td>$0.1250 \times 0.8750 + 0.90 \times 0.1250$</td>
<td>$0.75 \times 0.8233 + 0.2214 \times 0.1767$</td>
</tr>
<tr>
<td><strong>EU(US+)</strong></td>
<td>$0.4461 \times p(US+) + 0.1374 \times p(US-)$</td>
<td>$0.5539 \times p(US+) + 0.2214 \times p(US-)$</td>
<td>$0.8750 \times 0.8750 + 0.90 \times 0.1250$</td>
<td>$0.8233 \times 0.8750 + 0.90 \times 0.1250$</td>
</tr>
<tr>
<td><strong>EU(US-)</strong></td>
<td>$0.9055 \times p(US+) + 0.2214 \times p(US-)$</td>
<td>$0.9087 \times p(US+) + 0.2214 \times p(US-)$</td>
<td>$0.9558 \times 0.8750 + 0.90 \times 0.1250$</td>
<td>$0.9087 \times 0.8750 + 0.90 \times 0.1250$</td>
</tr>
</tbody>
</table>

- **Mammography**
  - Positive: $0.8569$, needed treatment
  - False positive: $0.8626$, unneeded procedures
  - True negative: $0.9447$, avoid unneeded procedures
  - False negative: $0.5539$, forgo needed treatment

- **Ultrasonography**
  - Positive: $0.9055$, needed treatment
  - False positive: $0.4461$, unneeded procedures
  - True negative: $0.9087$, avoid unneeded procedures
  - False negative: $0.5364$, forgo needed treatment
<table>
<thead>
<tr>
<th>EU(MM+) = U(TP) x p(TP) + U(FP) x p(FP)</th>
<th>EU(MM) = U(MM+) x p(MM+) + EU(MM-) x p(MM-)</th>
</tr>
</thead>
<tbody>
<tr>
<td>= 0.85 x 0.8626 + 0.90 x 0.1374</td>
<td>= 0.8569 x 0.4461 + 0.9447 x 0.5539</td>
</tr>
<tr>
<td>= 0.8569</td>
<td></td>
</tr>
<tr>
<td>EU(MM-) = U(TN) x p(TN) + U(FN) x p(FN)</td>
<td></td>
</tr>
<tr>
<td>= 1.00 x 0.7786 + 0.75 x 0.2214</td>
<td>= 0.9447</td>
</tr>
<tr>
<td>= 0.9447</td>
<td></td>
</tr>
<tr>
<td>EU(US+) = U(TP) x p(TP) + U(FP) x p(FP)</td>
<td>EU(US) = EU(US+) x p(US+) + EU(US-) x p(US-)</td>
</tr>
<tr>
<td>= 0.85 x 0.8750 + 0.90 x 0.1250</td>
<td>= 0.8563 x 0.4736 + 0.9558 x 0.5264</td>
</tr>
<tr>
<td>= 0.8563</td>
<td>= 0.9087</td>
</tr>
<tr>
<td>EU(US-) = U(TN) x p(TN) + U(FN) x p(FN)</td>
<td></td>
</tr>
<tr>
<td>= 1.00 x 0.8233 + 0.75 x 0.1767</td>
<td>= 0.9558</td>
</tr>
</tbody>
</table>

### Decision Node Diagram

- **Mammography**
  - Positive
    - True positive
      - 0.8569, needed treatment
    - False positive
      - 0.8626, 0.90, unneeded procedures
  - Negative
    - True negative
      - 0.9447, 1.00, avoid unneeded procedures
    - False negative
      - 0.1374, 0.75, forgo needed treatment

- **Ultrasonography**
  - Positive
    - True positive
      - 0.8563, 0.85, needed treatment
    - False positive
      - 0.8750, 0.90, unneeded procedures
  - Negative
    - True negative
      - 0.9558, 1.00, avoid unneeded procedures
    - False negative
      - 0.1250, 0.75, forgo needed treatment

**Positive**

**Negative**

**EU(MM+) = 0.8569**

**EU(MM-) = 0.9447**

**EU(US+) = 0.9087**

**EU(US-) = 0.9558**
Example: Add-on Test

DaTscan Ioflupane I-123 Injection (GE Healthcare) for Diagnosing Parkinson's Disease

We examine the evidence on DaTscan™ (GE Healthcare), which uses ioflupane I-123 injection during single photon emission computed tomography to diagnose Parkinsonian syndrome and reportedly can inform or alter the initial clinical diagnosis.

Impact of DaTscan SPECT imaging on clinical management, diagnosis, confidence of diagnosis, quality of life, health resource use and safety in patients with clinically uncertain parkinsonian syndromes: a prospective 1-year follow-up of an open-label controlled study

Andreas R Kupsch,¹ Nin Bajaj,² Frederick Weiland,³ Antonio Tartaglione,⁴ Susanne Klutmann,⁵ Melanie Buitendyk,⁶ Paul Sherwin,⁷ Ann Tate,⁷ Igor D Grachev⁷

J Neurol Neurosurg Psychiatry 2012;83:620–628.
Example: Add-on Test

Parkinson’s Disease is Overdiagnosed Clinically at Baseline in Diagnostically Uncertain Cases: A 3-Year European Multicenter Study with Repeat $[^{123}I]$FP-CIT SPECT

Vicky L. Marshall, MD,1,2,3* Cornelia B. Reininger, MD,4 Moritz Marquardt, MSc,4 Jim Patterson, PhD,1,2,3 Donald M. Hadley, PhD,1,2,3 Wolfgang H. Oertel, MD,5 Hani T.S. Benamer, PhD,6 Paul Kemp, MD,7 David Burn, MD,8 Eduardo Tolosa, MD,9 Jamie Kulisevsky, MD,10 Luis Cunha, MD,11 Durval Costa, MD,12 Jan Booij, MD,13 Klaus Tatsch, MD,14 K. Ray Chaudhuri, MD,15 Gudrun Ulm, MD,16 Oliver Pogarell, MD,17 Helmut Höffken, MD,5 Anja Gerstner, MD,5 and Donald G. Grosset, MD1,2,3
Add-on Test PICO

- **P:** adults with clinically uncertain Parkinsonian syndrome

- **I:** test-and-treat strategy 1: DaTscan + clinical info, downstream tests/treatments and outcomes

- **C:** test-and-treat strategy 2: clinical info only, downstream test/treatments and outcomes

- **O:** direct test-related outcomes (discomfort, anxiety), indirect test/treatment decision-related outcomes
# Parkinson’s Disease Diagnosis

<table>
<thead>
<tr>
<th>Reference standard positive</th>
<th>Reference standard negative</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test positive</strong></td>
<td></td>
</tr>
<tr>
<td>True positive:</td>
<td>False positive:</td>
</tr>
<tr>
<td>Test+, PD present</td>
<td>Test+, PD absent</td>
</tr>
<tr>
<td>Receive needed treatment</td>
<td>Receive unneeded tests/treatments</td>
</tr>
<tr>
<td>False negative:</td>
<td>True negative:</td>
</tr>
<tr>
<td>Test-, PD present</td>
<td>Test-, PD absent</td>
</tr>
<tr>
<td>Forgo/delay needed treatment</td>
<td>Avoid unneeded tests/treatments</td>
</tr>
</tbody>
</table>
Add-on Test Decision Tree

Decision node

- **DaTscan + clinical info**
  - Both positive
    - True positive: Needed treatment outcomes
    - False positive: Unneeded test/treatment outcomes
  - Not both positive
    - True negative: Avoid unneeded test/treatment outcomes
    - False negative: Forgo/delay needed treatment outcomes

- **clinical info alone**
  - PD suspected
    - True positive: Needed treatment outcomes
    - False positive: Unneeded test/treatment outcomes
  - PD not suspected
    - True negative: Avoid unneeded test/treatment outcomes
    - False negative: Forgo/delay needed treatment outcomes
Decision node

DaTscan + clinical info

0.8789

Both positive

0.8509

True positive

0.9821

0.85, needed treatment

0.9130

False positive

0.0179

0.90, unneeded tests/treatments

Not both positive

0.5490

True negative

0.6522

1.00, avoid unneeded tests/treatments

0.9130

False negative

0.3478

0.75, forgo/delay needed treatment

PD suspected

0.8593

True positive

0.8148

0.85, needed treatment

0.9306

False positive

0.1852

0.90 unneeded tests/treatments

PD not suspected

0.8182

True negative

0.7222

1.00, avoid unneeded tests/treatments

0.1818

False negative

0.2778

0.75, forgo/delay needed treatment

clinical info alone

0.8722
The diagram illustrates the decision-making process involving DaTscan and clinical info. The decision node is split into two paths: DaTscan + clinical info and clinical info alone.

**DaTscan + clinical info**
- **Both positive**
  - True positive: 0.8509
  - False positive: 0.5490
- **Not both positive**
  - True negative: 0.9130
  - False negative: 0.4510

**Clinical info alone**
- **PD suspected**
  - True positive: 0.8593
  - False positive: 0.8182
  - False negative: 0.1818
- **PD not suspected**
  - True negative: 0.9306
  - False negative: 0.7222

The table below lists the probabilities for different outcomes:

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both positive</td>
<td>0.8509</td>
</tr>
<tr>
<td>Not both positive</td>
<td>0.4510</td>
</tr>
<tr>
<td>True positive</td>
<td>0.8509</td>
</tr>
<tr>
<td>False positive</td>
<td>0.5490</td>
</tr>
<tr>
<td>True negative</td>
<td>0.9130</td>
</tr>
<tr>
<td>False negative</td>
<td>0.4510</td>
</tr>
<tr>
<td>PD suspected</td>
<td>0.8593</td>
</tr>
<tr>
<td>False positive</td>
<td>0.8182</td>
</tr>
<tr>
<td>False negative</td>
<td>0.1818</td>
</tr>
<tr>
<td>PD not suspected</td>
<td>0.9306</td>
</tr>
<tr>
<td>True negative</td>
<td>0.9130</td>
</tr>
<tr>
<td>False negative</td>
<td>0.4510</td>
</tr>
</tbody>
</table>

The diagram indicates that:
- A true positive result (0.8509) leads to 0.85 needed treatment.
- A false positive result (0.5490) leads to 0.90 unneeded tests/treatments.
- A true negative result (0.9130) leads to 1.00 avoid unneeded tests/treatments.
- A false negative result (0.4510) leads to 0.75, forgo/delay needed treatment.

The numbers in bold indicate the probabilities associated with each outcome.
Example: Triage Test

Should FDG PET Be Used to Decide Whether a Patient with an Abnormal Mammogram or Breast Finding at Physical Examination Should Undergo Biopsy?

David J. Samson, BA, Carole Redding Flamm, MD, MPH, Etta D. Pisano, MD, Naomi Aronson, PhD

Acad Radiol 2002; 9:773-783
Triage Test PICO

- **P**: women with palpable breast mass/abnormal mammogram

- **I**: test-and-treat strategy 1: do biopsy if PET+, downstream tests/treatments and outcomes

- **C**: test-and-treat strategy 2: biopsy for all, downstream tests/treatments and outcomes

- **O**: direct test-related outcomes (discomfort, anxiety), indirect test/treatment decision-related outcomes
# Breast Biopsy

<table>
<thead>
<tr>
<th>Test positive</th>
<th>Reference standard positive</th>
<th>Reference standard negative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True positive:</td>
<td>False positive:</td>
</tr>
<tr>
<td></td>
<td>Test+, biopsy, breast cancer present</td>
<td>Test+, biopsy, breast cancer absent</td>
</tr>
<tr>
<td></td>
<td>Receive needed treatment</td>
<td>Biopsy AEs</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test negative</th>
<th>Reference standard positive</th>
<th>Reference standard negative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>False negative:</td>
<td>True negative:</td>
</tr>
<tr>
<td></td>
<td>Test-, no biopsy, breast cancer present</td>
<td>Test-, no biopsy, breast cancer absent</td>
</tr>
<tr>
<td></td>
<td>Forgo/delay needed treatment</td>
<td>Avoid biopsy AEs</td>
</tr>
</tbody>
</table>
Triage Test Decision Tree

- **Decision node**
  - **Biopsy if PET+**
    - **PET+, biopsy**
      - **Biopsy+ (TP)**
        - Biopsy AEs, needed treatment
      - **Biopsy- (FP)**
        - Biopsy AEs
      - **True negative**
        - Avoid biopsy AEs
      - **False negative**
        - Undetected cancer, forgo/delay treatment
  - **PET-, no biopsy**
    - Biopsy AEs
- **Biopsy all**
  - **Biopsy+**
    - Biopsy AEs, needed treatment
  - **Biopsy-**
    - Biopsy AEs
<table>
<thead>
<tr>
<th></th>
<th>RS+</th>
<th>RS-</th>
<th>Sens</th>
<th>Spec</th>
<th>PPV</th>
<th>1-PPV</th>
<th>NPV</th>
<th>1-NPV</th>
<th>Prev Ca</th>
<th>Prev nCa</th>
<th>Prev PET+</th>
<th>Prev PET-</th>
</tr>
</thead>
<tbody>
<tr>
<td>PET+</td>
<td>445</td>
<td>100</td>
<td>0.8900</td>
<td>0.8000</td>
<td>0.8165</td>
<td>0.1835</td>
<td>0.8791</td>
<td>0.1209</td>
<td>0.5</td>
<td>0.5</td>
<td>0.545</td>
<td>0.455</td>
</tr>
<tr>
<td>PET-</td>
<td>55</td>
<td>400</td>
<td>0.8165</td>
<td>0.1835</td>
<td>0.8791</td>
<td>0.1209</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

**Decision node**

- **Biopsy if PET+**
  - **PET+, biopsy**
    - **Biopsy+ (TP)**: 0.85, biopsy AEs, needed treatment
      - Biopsy- (FP): 0.99, biopsy AEs
      - True negative: 1.00, avoid biopsy AEs
      - False negative: 0.75, undetected cancer, forgo/delay treatment
    - PET-, no biopsy
      - **PET+, biopsy**
        - 0.545
      - **PET-, biopsy**
        - 0.455

- **Biopsy all**
  - **Biopsy+**
    - 0.5
  - **Biopsy-**
    - 0.5

- **Biopsy if PET-**
  - 0.5

**Notes:**

- Sens: Sensitivity
- Spec: Specificity
- PPV: Positive Predictive Value
- 1-PPV: 1 - PPV
- NPV: Negative Predictive Value
- 1-NPV: 1 - NPV
- Prev Ca: Prevalence of Cancer
- Prev nCa: Prevalence of No Cancer
- Prev PET+: Prevalence of PET+ Screening
- Prev PET-: Prevalence of PET- Screening

**ECRI Institute**

The Discipline of Science. The Integrity of Independence.

©2015 ECRI INSTITUTE
 Decision node

- **Biopsy if PET+**
  - PET+, biopsy
    - 0.8165: Biopsy+ (TP), 0.85, biopsy AEs, needed treatment
    - 0.1835: Biopsy- (FP), 0.99, biopsy AEs
  - PET-, no biopsy
    - 0.8791: True negative, 1.00, avoid biopsy AEs
    - 0.1209: False negative, 0.75, undetected cancer, forgo/delay treatment

- **Biopsy all**
  - 0.5: Biopsy+, 0.85, biopsy AEs, needed treatment
  - 0.5: Biopsy-, 0.99, biopsy AEs
Decision node

- Biopsy if PET+ (0.9185)
  - Biopsy+ (TP) (0.8757)
    - PET+, biopsy (0.8165)
    - PET-, no biopsy (0.9698)
      - PET- (0.1835)
      - True negative (0.8791)
      - False negative (0.1209)
  - Biopsy- (FP) (0.545)

- Biopsy all (0.9200)
  - Biopsy+ (0.5)
    - Biopsy+ (0.85, biopsy AEs, needed treatment)
  - Biopsy- (0.5)
    - Biopsy- (0.99, biopsy AEs)
Decision Modeling: A Five-Step Approach

A five-step approach to determine if modeling informative, worthwhile

1. Define how test will be used (PICOTS)
2. Use framework to identify test consequences, management strategies for each test result (downstream decision/ actions, outcomes)
3. Assess if modeling is useful (model when it will make a difference)
4. Evaluate previous modeling studies
5. Consider if modeling practically feasible in given time frame
Decision Modeling Step 3: Assess Whether Modeling Is Useful

In most cases, decision modeling is useful when evaluating medical testing because:

- Indirect links between testing and health outcomes
- Multitude of test-and-treat strategies can be contrasted

Modeling is not useful when:

1. One test “clear winner”
2. Information very scarce
Decision Modeling Step 3: Assess Whether Modeling Is Useful

1. Scenarios: one test-and-treat strategy can be a “clear winner”
   - Scenario A: direct comparative evidence
     - Evaluates all important test-and-treat strategies
     - From well-run randomized trials, nonrandomized studies
     - Applicable to clinical context, patient population
     - Shows one dominant strategy (both benefits and harms) with adequate statistical power
Decision Modeling Step 3: Assess Whether Modeling Is Useful

- Scenario B: One test-and-treat strategy clear winner by test accuracy alone
  - Same patient response to downstream treatments for all tests
  - Clear winner preferable in:
    1. Cost and safety
    2. Sensitivity — correctly identifying patients with disease
    3. Specificity — correctly identifying those without the disease
Decision Modeling Step 3: Assess Whether Modeling Is Useful

- Do patient groups have same response to treatment?
  - Randomized trials suggest same response
  - Inference between tests
    - If sensitivities of two tests very similar, can expect patients selected for treatment similar, respond to treatment similarly
  - Extrapolation between tests
    - Tests operate on same principle, so clinical/biological characteristics of additional cases expected to be same
Decision Modeling Step 3: Assess Whether Modeling Is Useful

2. Second case for not undertaking decision modeling: very scarce information

- Regarding:
  - Which modeling assumptions are reasonable
  - Downstream effects of testing
  - Plausible values of multiple influential parameters

- We do not understand the underlying disease processes well enough to credibly predict outcomes
Decision Modeling Step 5: Consider Whether Modeling Is Practically Feasible

- Feasibility considerations:
  - Time
  - Budget
  - Available personnel
  - Accessibility of pre-existing models
  - Modification needs for pre-existing models
  - Amount of out-of-scope literature required to develop/adapt a model

- If a model not currently feasible but would be useful, may be done later as a secondary project
Special Considerations for Molecular/Genetic Tests

Fang Sun, MD, PhD
Medical Director, Health Technology Assessment, ECRI Institute
Outline

- Overview of genetic tests
- Challenges in evaluating these tests
- How to deal with these challenges: cases
Different stakeholders may use the term “genetic test” differently.

“A genetic or genomic test involves an analysis of human chromosomes, deoxyribonucleic acid [DNA], ribonucleic acid [RNA], genes, and/or gene products (e.g., enzymes and other types of proteins), which is predominantly used to detect heritable or somatic mutations, genotypes, or phenotypes related to disease and health.”

—The Secretary's Advisory Committee on Genetics, Health, and Society
Genetic Tests

- **Cytogenetic tests**
  - Evaluate changes in the number or structure of chromosomes (e.g., karyotyping for Down syndrome)

- **Molecular tests**
  - Evaluate DNA or RNA for alterations
  - Constitute the majority of current genetic tests

- **Biochemical tests**
  - Measure products of genes (e.g., CA 125 test)
  - Proteomic tests
Common Testing Methods

- Karyotyping, fluorescence in situ hybridization (FISH)
- Polymerase chain reaction (PCR)
  - PCR variants (e.g., quantitative PCR, real-time PCR, multiplex ligation-dependent probe amplification [MLPA])
- Microarray (DNA chip)
- Array comparative genomic hybridization (aCGH)
- Sequencing (whole genome, whole exome, target sequencing)
  - Sanger method, next-generation sequencing (NGS)
Clinical Applications

- Diagnosis of symptomatic individuals
  - e.g., karyotyping for Down syndrome, DNA testing for fragile X syndrome

- Disease screening in asymptomatic individuals
  - e.g., molecular testing of stool samples for colorectal cancer screening (Cologuard test)
Clinical Applications

- Prenatal and newborn screening
  - e.g., analysis of cell-free DNA in maternal blood for fetal aneuploidies
- Risk/predisposition assessment
  - e.g., BRCA testing, Myriad myRisk™ Hereditary Cancer panel
Clinical Applications

- Prognosis assessment
  - e.g., ERBB2 testing for breast cancer, IgVH mutation analysis for chronic lymphocytic leukemia

- Treatment monitoring
  - e.g., CA-125 test for ovarian cancer monitoring
Clinical Applications

- Guiding drug selection or dosing
  - Testing for cytochrome P450 polymorphism in adults with nonpsychotic depression treated with selective serotonin reuptake inhibitors
  - *EGFR* testing to select patients for EGFR inhibitors (e.g., erlotinib, gefitinib) in patients with lung cancer
Clinical Applications

- To establish an “etiologic diagnosis”
  - A diagnosis has been established based on clinical manifestations
  - Targeted therapies may not be available
  - The main purpose of testing is to determine whether the patient carries a “pathogenic” genetic variant

The number of genetic tests has been growing fast

- According to Genetests.org
  - 53,071 tests are available worldwide (as of July 5, 2015)
    - For 4,375 disorders; involving 5,184 genes; offered by 655 laboratories
- The number is growing quickly
- Most tests are laboratory-developed tests (LDTs)
Two Regulatory Pathways

- **LDTs**
  - Performed only in the lab that developed the test
  - Historically, not actively regulated by FDA

- **FDA-cleared or approved test kits or systems**
  - Can be performed in multiple labs

- Arguably, the bar is lower for LDTs than for FDA-regulated tests

- FDA has determined it will regulate LDTs in the future
Quality, Regulation and Clinical Utility of Laboratory-developed Molecular Tests

Multigene panels are gaining in popularity

- May include hundreds of genes in a single panel
  - FoundationOne (Foundation Medicine, Inc.) Comprehensive Genomic Profiling Test for Guiding Targeted Therapy for Cancer (315 genes and introns from 28 additional genes for all types of solid tumor cancer)
  - myRisk Hereditary Cancer Panel (Myriad Genetics, Inc.) for Identifying Inherited Cancer Risk (25 genes for 8 types of cancer)
Whole genome/exome sequencing becomes increasingly available

- Cheaper
- Quicker
- Thanks to new technologies (e.g., NGS)
Direct evidence for clinical utility is rarely available

- Clinical utility (the test’s impact on health outcomes) is usually the ultimate interest of technology assessment
- Ideal type of evidence: studies that compare use versus no use of the test, reporting on patient-oriented health outcomes with sufficient follow-up
- Practical reasons for lack of direct evidence
  - Difficulty in patient recruitment, constant changes in technologies, long follow-up required
  - Some outcomes (e.g., psychological distress) are rarely studied
We need to develop a “chain of evidence” to assess clinical utility

- Analytic validity → Clinical validity → Clinical Utility
  - Does the test detect the genetic variant accurately/reliably?
  - Does the test detect the disorder accurately?
  - Does the test affect treatment decisions?
  - Does the treatment lead to improved health outcomes?
  - Are there any harms associated with the testing?
Challenges in addressing analytic/clinical validity

- Lack of transparency about the tests’ technical detail
- Lack of published data for analytic validity
- Data may be about a previous version of the test
  - Does the evidence apply to the current version?
- Lack of tools for assessing the quality of analytic validity studies
Genotype-phenotype associations are often the only evidence available

The test accurately/reliably detects the genetic variant

This genetic variant is strongly associated with the clinical condition

The test accurately/reliably detects the condition
Genotype-phenotype associations may not be well characterized

- Pathogenic (clinically significant) variants
- Natural (wild-type) variants
- Variants of uncertain or unknown significance (VUSs)
- Genotype-phenotype associations are highly complex and may be affected by environments or behaviors
Addressing Challenges in Genetic Test Evaluation: Evaluation Frameworks and Assessment of Analytic Validity

HTAIS Genetic Test Product Brief: FoundationOne (Foundation Medicine, Inc.) Comprehensive Genomic Profiling Test for Guiding Targeted Therapy for Cancer

FoundationOne

- A genomic profiling test intended to help physicians make treatment decisions for patients with all types of solid tumor cancers.
- Uses next-generation sequencing to simultaneously interrogate the entire coding region of 315 genes and select introns from 28 additional genes.
- To identify molecular growth drivers of cancers in these genes/introns and help oncologists match them with relevant targeted therapies.
FoundationOne (continued)

- The classes of genomic alterations assayed include single-base substitutions, insertions, deletions, copy number alterations, and rearrangements.

- The report highlights any relevant alteration(s) found in the genes or introns that FoundationOne interrogates and provides information about available targeted therapies and clinical trials.
The Main Challenge

The test includes a very large number of markers
- 315 genes and select introns from additional 28 genes
- For all solid tumor cancers

This Product Brief is not intended to separately evaluate the clinical significance of each of the genes/introns included in FoundationOne for guiding cancer treatment. This Product Brief focuses primarily on evaluating the FoundationOne test’s impact as a multigene panel on patient-oriented health outcomes.
The Main Issues (Key Questions)

- Does FoundationOne affect patient outcomes (e.g., overall or progression-free survival)?
  - Is there any direct evidence?
  - Can we develop a chain of evidence?
Is there any direct evidence?

- We searched PubMed, EMBASE, and selected web-based resources for studies evaluating the FoundationOne test’s clinical utility published in peer-reviewed journals between January 1, 2010, and May 26, 2015.

- Our search identified a small number of studies that reported cases in which FoundationOne’s results actually affected treatment decisions or clinical outcomes.
Is there any direct evidence?

- These studies are either single case reports or case series.
- We did not identify any comparative studies that directly evaluated FoundationOne’s impact on health outcomes.
- Validating the test’s clinical utility requires larger, longer-term comparative studies—ideally randomized controlled trials—that assess the test’s impact on patient-oriented health outcomes (e.g., overall or progression-free survival).
Can we develop a chain of evidence?

- Does FoundationOne detect the genetic markers accurately?
- Is each included marker a good predictor for drug response?
- Does FoundationOne affect treatment decisions?
- Does the treatment decision based on the FoundationOne results affect patient outcomes?
Does FoundationOne detect the genetic markers accurately?

- No analytic validity study for the current version of the test.
- One study evaluated a previous version of the test (sequencing 287 cancer-related genes).
  - The sensitivity and specificity reported in that study were high.
- According to Foundation Medicine—
  - “The technology platform for FoundationOne remained unchanged and internal company validation studies, also submitted to NY State, showed high concordance and similar performance between the two content versions.”
  - However, we did not identify any publicly accessible data to enable us to verify this claim.
Is each included marker a good indicator for drug response?

- Markers were selected based on literature, according to Foundation Medicine
  - About 80 FoundationOne-relevant studies are provided on the company’s website
- Some markers are considered well-established for guiding treatment decisions for certain cancers
  - e.g., *EGFR* mutations and *ALK* fusions for lung cancer (adenocarcinoma), *ERBB2* for breast cancer, *KRAS* mutations for colorectal cancer
- However, other markers included in the test may not carry the same clinical significance
Does FoundationOne affect treatment decisions?

- Yes, for some makers/cancer types
  - Based on a small number of case series and single case reports
- But not for all markers/cancer types
Does the treatment decision based on the FoundationOne results affect patient outcomes?

- FoundationOne is intended to identify actionable genomic alterations
  - Actionable genomic alterations—those for which a U.S. Food and Drug Administration (FDA)-approved drug for the cancer or another cancer type or a registered clinical trial on a drug for the cancer is available
  - Most of the actionable genomic alterations are for guiding off-label use of investigational drugs, which may not necessarily improve health outcomes and may even cause harm to patients
Does this chain of evidence help you come to any conclusion about the clinical utility of the test?
Relevant Clinical Guidelines

▶ The National Comprehensive Cancer Network (NCCN) guideline regarding non-small cell lung cancer (NSCLC)

- “The NCCN NSCLC Guidelines Panel strongly endorses broader molecular profiling with the goal of identifying rare driver mutations for which effective drugs may already be available, or to appropriately counsel patients regarding the availability of clinical trials. Broader molecular profiling is a key component of the improvement of the care of patients with NSCLC.”

▶ Our search did not identify any clinical practice guidelines regarding broader genomic profiling for other types of cancer
Coverage Policies

- No Medicare national coverage determination or any pending national coverage analyses regarding the test
  - One Local Coverage Determination (LCD) by Palmetto GBA
- We searched the websites of 11 major third-party payers that publish their coverage policies online
  - Five payers consider the test to be “experimental,” “investigational,” or “not medically necessary” and so do not reimburse its use
  - Six payers don’t have a specific policy
Diagnostic Technologies and Genetic Tests

Assessing Evidence on Genetic Tests

Jonathan R. Treadwell, PhD
Associate Director, Health Technology Assessment and Evidence-based Practice Center, ECRI Institute
The Plan

► Diagnosis vs. Prognosis
► Going Beyond
► 6 flavors of prognostic data
► Example: Oncotype DX 12-gene assay for assessing recurrence risk in colon cancer
► Example: VeriStrat® proteomics test for treatment planning in advanced non-small-cell lung cancer
► Special considerations
Diagnosis vs. prognosis

**Diagnosis**: Whether a patient has a disease at the time of the test

**Prognosis**: Whether a patient will later develop a disease, or experience a medical event
Diagnosis vs. prognosis

Diagnosis: Snapshot  Prognosis: Time Lapse
Types of prognostic questions

- How long will I live?
- What will my quality-of-life be?
- Will I get cancer?
- If I do, and I get treated, will the tumor respond?
- Even if it responds, will it someday come back?
Diagnosis vs prognosis

▶ Common threads

■ Is the test accurate?
■ Is it useful for clinical decision making?
■ Does it improve health?
Standard prognostic “tests”

- History, physical exam, family history, lab tests, imaging results, comorbidities
- Their purpose has always been to guide treatment decisions in an effort to improve outcomes.
For any new prognostic factor ...

We need to ask:

Does it improve our predictions beyond standard prognostic factors?
Take cancer

► Most genetic tests being marketed for prognosis are for cancer

► **Cancer stage** is the traditional prognostic factor. Further subdivisions are common (e.g., Stage IIIA or IIIB or IIC for breast cancer)

► Stage and treatment
  - Few treatment options: Only a few stages are necessary
  - Many treatment options: May need a complex staging system
Simple staging

► Small cell lung cancer

► 2 stages:
  ■ “Limited disease” (10%-20% of patients). Chemotherapy and radiation with curative intent
  ■ “Extensive disease” (80%-90% of patients). Chemotherapy, perhaps with palliative radiation
Complex staging

- Breast cancer

- TNM staging
  - Tx, T0, Tis, T1, T2, T3, T4
  - Nx, N0(i+), N0(mol+), N1mi, N1a, N1b, N1c, N2a, N2c, N3a, N3b, N3c
  - Mx, M0, cM0(i+), M1

- Converted to Stage IA, IB, IIA, IIB, IIIA, IIIB, IIIC, IV

- Treatments in several categories, each with options (surgery, radiation, chemotherapy, hormone therapy, targeted therapy, bone-directed therapy)

The 6 flavors of prognostic data

- **Cost-effectiveness**: Is worth the cost
- **Clinical outcomes**: Directly affects outcomes
- **Treatment impact**: Influences treatment decisions
- **Incremental value**: Is more predictive than standard prognostics alone
- **Prospective validation**: Has been confirmed prospectively
- **Proof of concept**: Is associated with outcomes

There is a 7th flavor

- Predicting response to treatment
- Those with a “Good” test result respond better to treatment than those with a “Poor” test result
Example 1: Oncotype DX® Colon Cancer Assay

- Colon cancer: 4th most prevalent cancer
- 66% of patients present at Stage II or III
- Stage II patients undergo surgery. Adjuvant chemotherapy is only recommended if there is a “high” recurrence risk
- Standard definition: High risk if any of the following:
  - T4 lesions
  - Fewer than 12 lymph nodes examined
  - Presence of bowel perforation or obstruction
  - Poorly differentiated tumors
  - Lymphatic or venous invasion
Example 1: Oncotype DX® Colon Cancer Assay

- 12-gene Oncotype DX® Colon Cancer Recurrence Score assay
- “In stage II patients with T3 MMR-P tumors, the Recurrence Score result informs whether additional therapy should be considered beyond surgery“
Example 1: Oncotype DX® Colon Cancer Assay

From the company website (http://www.oncotypedx.com/)

“The Oncotype DX® Colon Cancer Assay quantifies recurrence risk in stage II and stage III colon cancer, beyond traditional qualitative measures. This enables an individualized approach to treatment planning. The Oncotype DX test measures a group of cancer genes in the tumor, providing a quantitative Recurrence Score® result beyond traditional measures so physicians and patients can have a more complete discussion of recurrence risk.”
The 6 flavors of prognostic data

- **Cost-effectiveness**: Is worth the cost
- **Clinical outcomes**: Directly affects outcomes
- **Treatment impact**: Influences treatment decisions
- **Incremental value**: Is more predictive than standard prognostics alone
- **Prospective validation**: Has been confirmed prospectively
- **Proof of concept**: Is associated with outcomes

Incremental value

- Take a group of patients who, based only on standard prognostic tests, all have the same recurrence risk
- All with Stage II with T3 MMR-P tumors
- Among those patients, does the risk of recurrence vary according to the results of the Oncotype DX® Colon Cancer Assay
- This is incremental value
Incremental value

- **Results:**
  - “Low risk” Stage II T3 MMR-P patients: 3 yr. recurrence 12%
  - “Intermediate risk” Stage II T3 MMR-P patients: 3 yr. recurrence 18%
  - “High risk” Stage II T3 MMR-P patients: 3 yr. recurrence 22%

- Evidence of incremental prognostic value
- Not a huge effect
- Those with a high score on Oncotype DX were 83% more likely to have a recurrence than those with a low score \(\frac{(22-12)}{12}\)

Treatment impact

▶ How are you managed if the test is not available?
▶ How are you managed if the test result is available?
▶ If these differ, then the test has treatment impact
## Treatment impact

<table>
<thead>
<tr>
<th>Treatment plan before knowing test result</th>
<th>Observation</th>
<th>Fluoro-pyrimidine monotherapy</th>
<th>FOLFOX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observation</td>
<td><strong>Same</strong></td>
<td>More intensive</td>
<td>More intensive</td>
</tr>
<tr>
<td>Fluoro-pyrimidine monotherapy</td>
<td>Less intensive</td>
<td><strong>Same</strong></td>
<td>More intensive</td>
</tr>
<tr>
<td>FOLFOX</td>
<td>Less intensive</td>
<td>Less intensive</td>
<td><strong>Same</strong></td>
</tr>
</tbody>
</table>

**Treatment plan after knowing test result**
### Treatment impact

<table>
<thead>
<tr>
<th></th>
<th>Observation</th>
<th>Fluoro-pyrimidine monotherapy</th>
<th>11% more intensive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observation</td>
<td>38%</td>
<td>6%</td>
<td>4%</td>
</tr>
<tr>
<td>Fluoro-pyrimidine monotherapy</td>
<td>17%</td>
<td>6%</td>
<td>1%</td>
</tr>
<tr>
<td>FOLFOX</td>
<td>15%</td>
<td>1%</td>
<td>11%</td>
</tr>
</tbody>
</table>

Not all colon cancers are alike

45% of treatment recommendations changed after seeing the Oncotype DX® Recurrence Score® result

LEARN MORE ABOUT THE ONCOTYPE DX ASSAY
Clinical outcomes

Does getting the test vs. not getting the test affect patient outcomes?
Clinical outcomes

Stage II with T3 MMR-P tumors

Treated based on clinical judgment

Recurrence

No recurrence

Treated based on clinical judgment AND the test result

Recurrence

No recurrence

Overall survival, QOL

Overall survival, QOL

(based on treatment impact data, this will be less intensive)
Clinical outcomes

- No studies have made this direct comparison
- But, logically, it would make sense for outcomes to be better among those who got the test
  - Test result is associated with recurrence (incremental value)
  - Test result affects treatment choice (treatment impact)
  - Treatment choice affects recurrence
  - Recurrence affects survival/QOL
- Markov model by Alberts et al. (2014)

Clinical outcomes
Clinical outcomes

- a. No recurrence
- b. Recurrence
- c. Death
Clinical outcomes

- Alberts et al. combined survival and QOL into a single metric: Quality-Adjusted Life Years (QALYs).
- A year in perfect health is worth 1 QALY.
- A year in suboptimal health, such as having to undergo intensive chemotherapy, may only be worth 0.8 QALYs.
Clinical outcomes

- Results of Alberts et al. (2014):
  - Those who do not get the test accumulate ~8.001 QALYs
  - Those who do get the test accumulate ~8.115 QALYs
  - Thus the benefit is 0.114 QALYs
  - (Results not reported separately for survival vs. QOL)

- Indirect evidence of the test’s influence on clinical outcomes
Example 2: VeriStrat® for advanced NSCLC

- Lung cancer is the deadliest cancer
- 85% are NSCLC
- 70% of NSCLC are advanced
- Standard chemotherapy is platinum-based
- Newer treatment with tyrosine kinase inhibitors (TKIs) such as erlotinib (FDA clearance May 2013)
- Gregorc (2014)\(^1\) was a randomized trial providing data on whether VeriStrat predicts response to treatment

Example 2: VeriStrat® for advanced non-small-cell lung cancer

- From the company website (http://www.biodesix.com/products/veristrat/)

  - How can you tailor therapeutic strategies based on disease aggressiveness?
  - VeriStrat® is a blood-based predictive and prognostic proteomic test for patients with advanced non-small cell lung cancer who test negative for EGFR mutations (EGFR wild-type) or whose EGFR mutation status is unknown.
  - VeriStrat assesses disease aggressiveness, classifying patients as either VeriStrat Good or VeriStrat Poor.
  - Blood test, 72 hour results
  - VeriStrat classification is also predictive of differential treatment benefit for single agent therapy
Example 2: VeriStrat® for advanced NSCLC

VeriStrat®

“Good” response predicted
Chemotherapy
Erlotinib

“Poor” response predicted
Chemotherapy
Erlotinib

Survival duration?
12 months
62 months
6 months

Example 2: VeriStrat® for advanced NSCLC

- **VeriStrat “GOOD”, underwent chemotherapy**
- **VeriStrat “GOOD, took erlotinib”**
- **VeriStrat “POOR”, underwent chemotherapy**
- **VeriStrat “POOR”, took erlotinib**

If the VeriStrat result is “good”, you live longer, and treatment choice doesn’t matter

If it’s “poor”, avoid erlotinib

---

Example 2: VeriStrat® for advanced NSCLC

VeriStrat®

“Good” response predicted

Chemotherapy

12 months

Erlotinib

12 months

Interaction:

“Poor” response predicted

Chemotherapy

6 months

Erlotinib

3 months

Survival duration?

Special considerations

- Risk of bias
- Publication bias
- Communication of risk
- Strength of evidence
Special considerations: Risk of bias

- Overlapping datasets for developing vs. testing the prognostic factor
- Posthoc threshold for defining prognostic groups
- Different length of follow-up for different prognostic groups
- Failure to account for standard prognostic tests

Special considerations: Publication bias

▶ What if the test hadn’t been predictive of anything?
▶ Would the study have been published?
▶ Reviewer concerns:
  ■ How many unpublished studies might be out there?
  ■ Among published studies: Compare what was measured to what was reported

Special considerations: Communication of risk

- **Relative risk**: For predicting cancer recurrence, those who tested high on *AwesomeGeneTest* had a 67% higher risk than those who tested low.

- **Absolute risk**: For predicting cancer recurrence, those who tested high on *AwesomeGeneTest* had a 5% chance of recurrence, whereas those who tested low had a 3% chance of recurrence.

- These describe the same data \((5-3)/3=0.67\).

- Can be misleading to present only the relative risk.
Special considerations: Strength of evidence

- Prognostics: Nothing substantive yet from the GRADE working group
- Grading is similar to diagnostics
- Huguet et al. (2013)\(^1\): start with *phase of investigation*:
  - Start at High for phase 2 or 3 “explanatory research”
  - Start at Moderate for phase 1 “identifying associations”
- Unlikely the GRADE group will agree

Summary

- Prognosis is not diagnosis, but they share several concepts
- Standard prognostics already exist; what’s the value-add?
- 7 flavors of prognostic data
- 2 genetic test examples, and their supporting evidence
- Special considerations: Risk of bias, publication bias, risk communication, strength of evidence
Diagnostic Technologies and Genetic Tests
July 14–15, 2015

Exploring Genetic Test Evaluation: Some Examples

Jeff Oristaglio, Ph.D.
Research Analyst
ECRI Institute
Health Technology Assessment
Overview

- Genetic Testing: goals and limitations
- Genetic test evaluation
  - Example I: Cologuard (Exact Sciences)
  - Example II: Percepta Bronchial Genomic Classifier (Veracyte Inc.)
- General Summary and Closing remarks
The Overarching Goal: Personalized Medicine

- More effective screening and diagnosis for individual patients
- Customization of care
  - Identifying the safest and most effective treatments for each individual patient
- Prophylaxis
  - Identifying each individual’s unique constellation of risk factors and taking early action
But, ... genetics isn’t everything ...

- Environment plays an important role for many conditions
  - obesity
  - cardiovascular disease
  - mental illness, etc.

- Genetic testing will often predict risk, not provide definitive yes/no answers about health outcomes.

"Your weight problem is partly genetic and partly Boston Cream pie."
Genetic tests: What do we want to know?

Before we invest millions, maybe we should test it and see if it works.

freshspectrum.com
Challenges in Evaluating Genetic Tests (GTs)

- Evidence supporting most genetic tests stops at clinical validity.
- Is this good enough?
Example I

www.cologuardtest.com
Cologuard® Colon Cancer Screening (Exact Sciences Corporation)

- Non-invasive screening test for colon cancer
- Requires only a stool sample
- “No special preparation”
- “No diet or medication changes”
- “No time off needed”

*(quotes from manufacturer website)*
Cologuard® (Exact Sciences Corporation)

- **Intended purpose:** for simple, non-invasive detection of colorectal cancer (CRC) and precancerous lesions in stool samples
- Intended for subjects 50 years of age or older and at average risk for CRC
- **Not** intended as a replacement for diagnostic or surveillance colonoscopy in high-risk individuals
- Cologuard received FDA-approval (August, 2014)
- Cologuard is covered (once every 3 years) by Centers for Medicare & Medicaid Services (CMS)
  - Specified in national coverage determination titled “Screening for Colorectal Cancer-Stool DNA Testing”
# Existing CRC Screening Methods

<table>
<thead>
<tr>
<th>Screening Method</th>
<th>Recommended Screening Interval</th>
<th>Morbidity and Mortality Outcomes Reported in Clinical Studies</th>
<th>Intended Advantages Potential Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colonoscopy</td>
<td>10 years</td>
<td>Inferred 60%-70% reduction in CRC mortality</td>
<td>Advantages: Examines entire colon, allows immediate polypectomy, high accuracy, long screening interval Disadvantages: Risk of serious complications (e.g., perforation, bleeding), requires thorough bowel preparation, requires some sedation, performance may be operator dependent</td>
</tr>
<tr>
<td>Flexible Sigmoidoscopy</td>
<td>5 years</td>
<td>Reduces CRC mortality 28% and CRC incidence 18%</td>
<td>Advantages: Allows immediate polypectomy, requires enema based bowel preparation Disadvantages: Risk of serious complications (e.g., perforation, bleeding), doesn’t examine proximal colon, performance may be operator dependent</td>
</tr>
<tr>
<td>Computed Tomography Colonoscopy</td>
<td>5 years</td>
<td>None reported</td>
<td>Advantages: Minimally invasive, low complication rate compared with colonoscopy, no sedation required Disadvantages: Detects extracolonic abnormalities, requires colonic air insufflation, radiation exposure, performance may be operator dependent, requires thorough bowel preparation</td>
</tr>
<tr>
<td>Double Barium Enema (Lower GI Series)</td>
<td>5 years</td>
<td>None reported</td>
<td>Advantages: Inexpensive Disadvantages: Lower accuracy than other invasive methods</td>
</tr>
<tr>
<td>Fecal Immunochemical Test</td>
<td>1 year</td>
<td>None reported</td>
<td>Advantages: Noninvasive, inexpensive, widely available Disadvantages: Lower accuracy than colonoscopy, high testing frequency</td>
</tr>
<tr>
<td>High-sensitivity Guaiac Fecal Occult Blood Test</td>
<td>1 year</td>
<td>Reduces CRC related mortality 15%-33%</td>
<td>Advantages: Noninvasive, inexpensive, widely available Disadvantages: Lower accuracy than colonoscopy, high testing frequency</td>
</tr>
</tbody>
</table>
Cologuard: How it works (the patient perspective)

1. Patient visits provider who prescribes Cologuard
2. Patient receives Cologuard test package
3. Patient collects sample at home
4. Sample is shipped to ExactSciences
5. Doctor contacts patient with the test results
6. Follow up:
   - Negative results: retest in 3 years
   - Positive results: colonoscopy (potential follow-up with biopsy)
COLON CANCER SCREENING made easy, WITH NONINVASIVE COLOGUARD®

GET.

GO.

GONE.

WHAT IS COLOGUARD? »

For adults 50+ at average risk for colon cancer.

www.cologuardtest.com
YEP, YOU USE COLOGUARD in your own bathroom.

HOW DOES COLOGUARD WORK? »

For adults 50+ at average risk for colon cancer.

www.cologuardtest.com
Cologuard: How it works (the science)

- Cells slough off from lining of colon and are excreted in the stool
- Cologuard detects:
  - Altered DNA from abnormal cells that may be involved in cancer
  - Occult (hidden) blood in stool
- 3 separate analyses:
  - Methylated DNA from tumor-suppressing genes NDRG4 and BMP3 (methylation silences gene activity)
  - KRAS gene mutations (known to be present in CRCs and adenomas); specific mutations lead to uncontrolled cell proliferation
  - High-sensitivity immunochemical test to detect blood in stool samples
- Proprietary algorithm integrates these measures into risk score
- Predefined threshold value translates risk score to positive or negative result (negative meaning low risk for CRC)
Cologuard: Summary of findings from ECRI Emerging Technology Report and Product Brief

- Two reports of 1 study assessing *clinical validity* of Cologuard
- No studies found evaluating Cologuard’s clinical utility
Multitarget Stool DNA Testing for Colorectal-Cancer Screening

Thomas F. Imperiale, M.D., David F. Ransohoff, M.D., Steven H. Itzkowitz, M.D., Theodore R. Levin, M.D., Philip Lavin, Ph.D., Graham P. Lidgard, Ph.D., David A. Ahlquist, M.D., and Barry M. Berger, M.D.
12,776 asymptomatic patients (range 50 to 84 years of age) at average risk for CRC, and scheduled to undergo colonoscopy; 9989 participants evaluated

Patients provided stool samples and underwent colonoscopy no more than 90 days after enrollment

- Colonoscopy provided the definitive diagnosis

Cologuard test performed at one of 3 laboratories; all lab personnel were blinded to patient test results and clinical findings

**Primary outcome**: ability of the DNA test (Cologuard) to detect colorectal cancer

Cologuard test results compared to FIT (Fecal Immunochemical Testing)
9989 participants evaluated

65 (0.7%) had CRC; 757 (7.6%) had advanced precancerous lesions

Sensitivity (for CRC)
- Cologuard: 92.3% (NPV 99%)
- FIT: 73.8%

Specificity (for patients with negative results on colonoscopy)
- Cologuard: 89.8%
- FIT: 96.4%

Number patients needed to screen to detect one cancer
- Cologuard: 166
- FIT: 208
- Colonoscopy: 154
Imperiale et al. (2014) Multitarget stool DNA testing for colorectal cancer screening, NEJM 370, no. 14

Sensitivity: Cologuard vs. FIT

- Cologuard sensitivity equal to or better than FIT
Cologuard: Summary of ECRI Emerging Technology Report/Product Brief findings

- Cologuard detects DNA biomarkers associated with CRC and precancerous lesions
- Cologuard has *higher sensitivity* than fecal immunochemical testing (FIT)
  - Better at detecting CRC
  - Very high NPV (over 99% for absence of CRC)
- Cologuard has *lower specificity* than FIT
  - *More false-positives, but perhaps we can live with this!*
- Quality of evidence rated as moderate (using GRADE)
- Overall conclusions:
  - Current data indicates that Cologuard performs as intended as a screening test for CRC
  - Recommended testing every 3 years with Cologuard supported by indirect evidence (modeling study, submitted for publication)
  - Cologuard represents an additional choice for CRC screening
  - Relative benefit vs. FIT?

©2015 ECRI INSTITUTE
Example II

www.veracyte.com/percepta
Example II: Percepta Bronchial Genomic Classifier (Veracyte Inc.)

- For assessing lung nodules suspicious for malignancy
  - patients who are current or former smokers,
  - and at least 21 years of age
- Used in conjunction with bronchoscopy, a standard technique for assessing lung nodules
- **Intended purpose:** to reduce the number of costly, high-risk invasive diagnostic procedures following indeterminate bronchoscopy results

www.veracyte.com/percepta
Example II: Percepta Bronchial Genomic Classifier (Veracyte Inc.)

Assessing lung nodules

- Approximately 40% of bronchoscopies are indeterminate
- 20-25% of surgical biopsies are performed on patients with benign lesions

Patient with lung nodule (found with CT or chest x-ray) → BRONCHOSCOPY → Indeterminate (40%) → Further testing/treatment for LC

Positive

Negative

Watchful waiting: monitor with CT

Next steps uncertain:
Surgical biopsy?
Monitoring with CT?
Other choices???
Example II: Percepta Bronchial Genomic Classifier (Veracyte Inc.)

- **The unmet need**: A test that find patients at low risk for lung cancer, reducing the number of invasive diagnostic procedures.
Percepta: how it works (the patient perspective)

- Epithelial cells harvested during bronchoscopy are used for Percepta
- Samples are sent to CLIA-certified laboratory for processing
- If bronchoscopy is indeterminate, Percepta is run on the samples
- Results are reported to physician who then communicates to patient
- Percepta-negative patients can be subsequently referred for CT monitoring rather than more risky and inconvenient surgical biopsy

**Key points:**
- Percepta is designed to identify patients at low risk for lung cancer
- *Percepta fits neatly into the standard clinical progression (example of an add-on test)*
Percepta: how it works (the science)

- Percepta analyzes RNA expression of 23 genes associated with lung cancer risk using microarrays; includes genes involved in cell growth and proliferation, immune response, tracheal epithelial regeneration, and other functions.
- Genes selected for association with gender, tobacco use, and smoking history (*gene expression correlates*).
- A proprietary algorithm integrates gene expression levels, gene expression correlates, and patient age into a risk score.
- Percepta reports classify samples as “high-”, “intermediate-”, or “low-risk” for lung cancer.
Percepta: Summary of ECRI Product Brief findings

- Searched PubMed, EMBASE, Cochrane Library, selected web resources and documents published from January 1, 2010 to May 18, 2015
- 5 studies directly relevant to Percepta; two full-text articles (comprising 3 studies) and 3 conference abstracts; 4 of these 5 studies evaluated clinical validity
- Additional studies (n=32): academic research investigations of gene expression changes associated with lung cancer or exposure to cigarette smoke
- No studies assessing clinical utility were found
A Bronchial Genomic Classifier for the Diagnostic Evaluation of Lung Cancer

Gerard A. Silvestri, M.D., Anil Vachani, M.D., Duncan Whitney, Ph.D., Michael Elashoff, Ph.D., Kate Porta Smith, M.P.H., J. Scott Ferguson, M.D., Ed Parsons, Ph.D., Nandita Mitra, Ph.D., Jerome Brody, M.D., Marc E. Lenburg, Ph.D., and Avrum Spira, M.D., for the AEGIS Study Team*
Silvestri et al. (2015), A Bronchial Genomic Classifier for the Diagnostic Evaluation of Lung Cancer, NEJM

- Clinical validation study analyzing results from two independent, multicenter, prospective trials
- Data from a total of 639 current or former smokers undergoing bronchoscopy for suspected lung cancer
- 272 patients with non-diagnostic bronchoscopies
- Airway epithelial cells collected during bronchoscopy
- Percepta test run on collected samples; results were not reported to patients or physicians
- Patients followed until diagnosis was established or for 12 months following bronchoscopy
  - Diagnosis established with invasive procedure (surgical or transthoracic needle biopsy, additional bronchoscopy, or other invasive procedure
Percepta sensitivity by imaging characteristics

<table>
<thead>
<tr>
<th>Group</th>
<th>All Patients</th>
<th>Patients with Cancer</th>
<th>Sensitivity*</th>
<th>Classifier plus Bronchoscopy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no. of patients</td>
<td></td>
<td>Bronchoscopy</td>
<td>Classifier</td>
</tr>
<tr>
<td></td>
<td>percent (95% confidence interval)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All patients</td>
<td>639</td>
<td>487</td>
<td>75 (71–79)</td>
<td>89 (82–94)</td>
</tr>
<tr>
<td>Lesion size</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;2 cm</td>
<td>131</td>
<td>73</td>
<td>55 (43–66)</td>
<td>91 (76–98)</td>
</tr>
<tr>
<td>2 to 3 cm</td>
<td>80</td>
<td>60</td>
<td>58 (45–71)</td>
<td>92 (74–99)</td>
</tr>
<tr>
<td>&gt;3 cm</td>
<td>343</td>
<td>313</td>
<td>82 (78–86)</td>
<td>85 (74–93)</td>
</tr>
<tr>
<td>Infiltrate</td>
<td>60</td>
<td>25</td>
<td>84 (64–95)</td>
<td>100 (40–100)</td>
</tr>
<tr>
<td>Unknown</td>
<td>25</td>
<td>16</td>
<td>80 (54–96)</td>
<td>100 (29–100)</td>
</tr>
<tr>
<td>Lesion location</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central</td>
<td>225</td>
<td>174</td>
<td>84 (78–89)</td>
<td>81 (62–94)</td>
</tr>
<tr>
<td>Peripheral</td>
<td>194</td>
<td>133</td>
<td>55 (46–63)</td>
<td>90 (79–96)</td>
</tr>
<tr>
<td>Central and peripheral</td>
<td>192</td>
<td>164</td>
<td>82 (75–87)</td>
<td>97 (82–100)</td>
</tr>
<tr>
<td>Unknown</td>
<td>28</td>
<td>16</td>
<td>81 (54–96)</td>
<td>67 (9–99)</td>
</tr>
</tbody>
</table>

* The sensitivity of bronchoscopy was determined for patients with lung cancer in each category. The sensitivity of the classifier was determined for the patients with lung cancer whose cancer was not diagnosed during bronchoscopy. The sensitivity of the classifier combined with bronchoscopy was calculated for all patients with lung cancer in each category.
## Table 3. Performance of Bronchoscopy and the Classifier, Stratified According to the Pretest Probability of Cancer.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients with lung cancer — no. (%)</td>
<td>3 (5)</td>
<td>41 (41)</td>
<td>405 (95)</td>
<td>38 (76)</td>
</tr>
<tr>
<td>Patients with benign lesions — no. (%)</td>
<td>59 (95)</td>
<td>60 (59)</td>
<td>21 (5)</td>
<td>12 (24)</td>
</tr>
<tr>
<td>Bronchoscopy performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity — % (95% CI)</td>
<td>33 (1–91)</td>
<td>41 (26–58)</td>
<td>79 (74–82)</td>
<td>82 (66–92)</td>
</tr>
<tr>
<td>Patients with nondiagnostic bronchoscopic examination — no. (%) †</td>
<td>61 (98)</td>
<td>84 (83)</td>
<td>108 (25)</td>
<td>19 (38)</td>
</tr>
<tr>
<td>Classifier performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity — % (95% CI) ‡</td>
<td>100 (16–100)</td>
<td>88 (68–97)</td>
<td>89 (80–94)</td>
<td>100 (59–100)</td>
</tr>
<tr>
<td>Specificity — % (95% CI) §</td>
<td>56 (42–69)</td>
<td>48 (35–62)</td>
<td>29 (11–52)</td>
<td>33 (10–65)</td>
</tr>
<tr>
<td>Negative predictive value — % (95% CI) ¶</td>
<td>100 (89–100)</td>
<td>91 (75–98)</td>
<td>38 (15–65)</td>
<td>100 (40–100)</td>
</tr>
<tr>
<td>Positive predictive value — % (95% CI) ¶</td>
<td>7 (1–24)</td>
<td>40 (27–55)</td>
<td>84 (75–91)</td>
<td>47 (21–73)</td>
</tr>
<tr>
<td>Combined classifier and bronchoscopy sensitivity — % (95% CI)</td>
<td>100 (29–100)</td>
<td>93 (80–98)</td>
<td>98 (96–99)</td>
<td>97 (91–100)</td>
</tr>
</tbody>
</table>
### Table 3. Performance of Bronchoscopy and the Classifier, Stratified According to the Pretest Probability of Cancer.*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients with lung cancer — no. (%)</td>
<td>3 (5)</td>
<td>41 (41)</td>
<td>405 (95)</td>
<td>38 (76)</td>
</tr>
<tr>
<td>Patients with benign lesions — no. (%)</td>
<td>59 (95)</td>
<td>60 (59)</td>
<td>21 (5)</td>
<td>12 (24)</td>
</tr>
<tr>
<td>Bronchoscopy performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity — % (95% CI)</td>
<td>33 (1–91)</td>
<td>41 (26–58)</td>
<td>79 (74–82)</td>
<td>82 (66–92)</td>
</tr>
<tr>
<td>Patients with nondiagnostic bronchoscopic examination — no. (%)†</td>
<td>61 (98)</td>
<td>84 (83)</td>
<td>108 (25)</td>
<td>19 (38)</td>
</tr>
<tr>
<td>Classifier performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity — % (95% CI)‡</td>
<td>100 (16–100)</td>
<td>88 (68–97)</td>
<td>89 (80–94)</td>
<td>100 (59–100)</td>
</tr>
<tr>
<td>Specificity — % (95% CI)§</td>
<td>56 (42–69)</td>
<td>48 (35–62)</td>
<td>29 (11–52)</td>
<td>33 (10–65)</td>
</tr>
<tr>
<td>Negative predictive value — % (95% CI)¶</td>
<td>100 (89–100)</td>
<td>91 (75–98)</td>
<td>38 (15–65)</td>
<td>100 (40–100)</td>
</tr>
<tr>
<td>Positive predictive value — % (95% CI)¶</td>
<td>7 (1–24)</td>
<td>40 (27–55)</td>
<td>84 (75–91)</td>
<td>47 (21–73)</td>
</tr>
<tr>
<td>Combined classifier and bronchoscopy sensitivity — % (95% CI)</td>
<td>100 (29–100)</td>
<td>93 (80–98)</td>
<td>98 (96–99)</td>
<td>97 (91–100)</td>
</tr>
</tbody>
</table>

*Table 3. Performance of Bronchoscopy and the Classifier, Stratified According to the Pretest Probability of Cancer.*

‡: 100% indicates perfect sensitivity; 0% indicates no sensitivity.

†: Lower limit of 100% indicates absence of any nondiagnostic bronchoscopic examinations.

¶: 100% indicates perfect specificity; 0% indicates no specificity.

©2015 ECRI INSTITUTE
Percepta specificity by pretest cancer probability

<table>
<thead>
<tr>
<th>Table 3. Performance of Bronchoscopy and the Classifier, Stratified According to the Pretest Probability of Cancer.*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variable</strong></td>
</tr>
<tr>
<td>Patients with lung cancer — no. (%)</td>
</tr>
<tr>
<td>Patients with benign lesions — no. (%)</td>
</tr>
<tr>
<td>Bronchoscopy performance</td>
</tr>
<tr>
<td>Sensitivity — % (95% CI)</td>
</tr>
<tr>
<td>Patients with nondiagnostic bronchoscopic examination — no. (%)</td>
</tr>
<tr>
<td>Classifier performance</td>
</tr>
<tr>
<td>Sensitivity — % (95% CI)‡</td>
</tr>
<tr>
<td>Specificity — % (95% CI)§</td>
</tr>
<tr>
<td>Negative predictive value — % (95% CI)¶</td>
</tr>
<tr>
<td>Positive predictive value — % (95% CI)¶</td>
</tr>
<tr>
<td>Combined classifier and bronchoscopy sensitivity — % (95% CI)</td>
</tr>
</tbody>
</table>
# Negative Predictive Value

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients with lung cancer — no. (%)</td>
<td>3 (5)</td>
<td>41 (41)</td>
<td>405 (95)</td>
<td>38 (76)</td>
</tr>
<tr>
<td>Patients with benign lesions — no. (%)</td>
<td>59 (95)</td>
<td>60 (59)</td>
<td>21 (5)</td>
<td>12 (24)</td>
</tr>
<tr>
<td>Bronchoscopy performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity — % (95% CI)</td>
<td>33 (1-91)</td>
<td>41 (26-58)</td>
<td>79 (74-82)</td>
<td>82 (66-92)</td>
</tr>
<tr>
<td>Patients with nondiagnostic bronchoscopic examination — no. (%) †</td>
<td>61 (98)</td>
<td>84 (83)</td>
<td>108 (25)</td>
<td>19 (38)</td>
</tr>
<tr>
<td>Classifier performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity — % (95% CI) ‡</td>
<td>100 (16-100)</td>
<td>88 (68-97)</td>
<td>89 (80-94)</td>
<td>100 (59-100)</td>
</tr>
<tr>
<td>Specificity — % (95% CI) §</td>
<td>56 (42-69)</td>
<td>48 (35-62)</td>
<td>29 (11-52)</td>
<td>33 (10-65)</td>
</tr>
<tr>
<td>Negative predictive value — % (95% CI) ¶</td>
<td><strong>100 (89-100)</strong></td>
<td><strong>91 (75-98)</strong></td>
<td><strong>38 (15-65)</strong></td>
<td><strong>100 (40-100)</strong></td>
</tr>
<tr>
<td>Positive predictive value — % (95% CI) ¶</td>
<td>7 (1-24)</td>
<td>40 (27-55)</td>
<td>84 (75-91)</td>
<td>47 (21-73)</td>
</tr>
<tr>
<td>Combined classifier and bronchoscopy sensitivity — % (95% CI)</td>
<td>100 (29-100)</td>
<td>93 (80-98)</td>
<td>98 (96-99)</td>
<td>97 (91-100)</td>
</tr>
</tbody>
</table>
Percepta Bronchial Genomic Classifier: Conclusions

• Percepta has high sensitivity.
• Low specificity, but ... Percepta is intended to find patients at low risk for lung cancer; requires high sensitivity and high NPV.
• Limited evidence indicates that Percepta has these characteristics for patients with low to intermediate pretest probability for lung cancer. However, enrollment numbers in studies are small.
• Integrating results from both Percepta and bronchoscopy yields best overall predictive value.
• Limited evidence suggests that Percepta provides additional useful information for making clinical decisions regarding treatment of lung nodules (however, data also indicate a high false-positive rate).
• Studies specifically assessing Percepta’s clinical utility have yet to be reported.
• Methodological concern: 11% specimens produced insufficient quality RNA for testing (Silvestri et al., 2015)
Cologuard and Percepta: Key take-home points

- Cologuard and Percepta appear to be useful tests that serve their respective intended purposes.
- Both tests are supported by data from clinical validation studies. Studies highlight the strengths and limitations of these tests.
- Evaluating genetic tests requires analysis of test performance with careful regard for the test’s intended purpose. Performance need only be good enough to satisfy the test’s purpose!
- Special attention should be paid to the patient population to which the test is targeted.
  - Particularly important when PPV and NPV are used to assess performance!
Cologuard and Percepta: Key take-home points

- Sometimes, genetic tests will *complement*, not replace, standard tests (i.e., Percepta)
- Add-on tests
- Clinical decision making still requires careful integration of multiple pieces of evidence

*Let’s not throw the baby out with the bath water.*

*Picture from: www.uschamber.com*
Concerns for evaluating GTs

- Who conducted the studies?
  - Manufacturer sponsored or independent group?
  - Methodological bias?
- How many studies? How many patients were enrolled?
- Have validation studies been replicated? Independent groups?
- Spectrum bias
  - Validation test population
  - Algorithm development
What we want vs. What we have

- Does the evidence for clinical validity, in principle, support the likelihood of clinical utility?
- What considerations/concerns do we have for more widespread use of a test?
Diagnostic Technologies and Genetic Tests
July 14–15, 2015

Summary

Vivian Coates
Vice President, Health Technology Assessment, ECRI Institute