Choosing Among Continuously Scaled Tests

Henry Glick

Epi 550

February 17, 2016

Some Last Comments on Graphing Results

Properties of an Ideal Statistic (Ideal Statistics)

1. Single statistic that identifies best test for all patients, no matter what their pretest probability (p) nor what treatment threshold (p*) applies to them
   – Ideal \textbf{OBtainable} statistic(s) – single graph or table that provides statistics for all relevant pretest probabilities and treatment thresholds
2. Statistics are characteristics of tests whose properties have same stability (instabilities) as sensitivity, specificity, and likelihood ratios
3. Statistics independent of development of new cost-effective treatments
   – Latter should affect appropriate DS for a particular patient, but not height of curve at any particular DS
Properties of an Ideal Statistic (Ideal Statistics) (2)

4. Statistics allow determination of complete ranking of testing strategies
   – e.g., that testing is superior to treating no one which is superior to treating everyone
5. Statistics allow determination of relative (or absolute) difference in outcomes among testing strategies
   – Latter in part to address issues related to inclusion of cost of test
6. Statistics unaffected by pre-test probability
   – “…index is independent of the relative sizes of the control and diseased groups” (Youden)
7. Statistics unaffected by treatment threshold
8. Possible to calculate a standard error for statistics (Youden)

Comparison of Methods

<table>
<thead>
<tr>
<th>Properties of Ideal Statistics for Test Strategies</th>
<th>PK</th>
<th>DC</th>
<th>FB</th>
<th>DS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single statistic for each test</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(Obtainable) Single graph or table for each test</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Stable characteristics of test (if test cost excluded)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Independent of development of new therapies</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Report complete ranking of test strategies</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Report (relative) difference in value of test strategies</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Unaffected by pre-test probability</td>
<td>+</td>
<td>+/-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Unaffected by treatment threshold</td>
<td>+/-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Possible to calculate a standard error for statistic</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

PK = Pauker & Kassirer test thresholds; DC = Decision Curves; FB = Football; DS = Decision Slopes

Notes for Selected - and +/- Categories

- Report complete ranking: Test thresholds and football report highest ranked therapy, but don’t report rankings for 2+ options that aren’t best
- Unaffected by pre-test probability: For decision curves, complete set of graphs for every pre-test probability is unaffected, but any single graph is affected
- Unaffected by treatment threshold: For test thresholds, complete set of graphs for every treatment threshold is unaffected, but any single graph is affected
Choosing Among Tests

Focus

• Choosing among tests for an individual patient
• Mainly address 2x2 approach
• Consider whether comparison of SSLR provides sufficient information to choose among tests

Choice Criteria

• Obtain maximum clinical information at lowest cost
  – Greatest net benefit from a positive test
• Elements of decision
  – Value of clinical information (post-test probabilities that are either appropriately or inappropriately above or below pre-test probability of disease)
  – Cost of test itself (i.e., technician, reagents, etc.)
  – Cost of delay in treatment due to test
  – Risk of test
• Discuss first element -- value of clinical information -- because it varies within a single test based on test characteristics we select
• Remaining three elements test specific, but constant for different test characteristics
Choice for Individual Patient

• If choice is for individual and 2x2 approach is used:
  – Choose between single operating points from two tests for like patients
  • By “like” we mean equal OOS (DS)
    – Either because the different patients’ p, \( \Delta O_{O_1} \) and \( \Delta O_{O_2} \) are all equal or because the different patients’ p, \( \Delta O_{O_1} \) and \( \Delta O_{O_2} \) differ, but their \((1-p)\frac{\Delta O_{O_1}}{p \Delta O_{O_2}}\) are equal

Choice for Formulary

• If choice is for formulary:
  – Which of several tests do we make available to clinicians?
  • In 2x2 approach, choose between pairs of operating points, where pairs represent optimal operating point from each test for sets of “like” patients
    – e.g., 25% of patients may require an OOS of 0.25; 50% may require an OOS of 1; and 25% may require an OOS of 4

Two Choice Methods

• Method 1. Compare tests’ intercepts of tangent OOS (decision curves or decision slopes)
  – Uncommon in literature
• Method 2. Compare tests’ ROC area
  – Common in literature
Method 1. Intercept of Tangent OOS and ROC Curve

• Already shown that expected net benefit for a positive test for any operating point can be expressed as function of intercept of OOS that is tangent to ROC curve and equals:
  \[ p \times \text{Intercept} \quad \text{or} \quad p \times \frac{\text{NB}}{p} \]

• Comparison of intercepts of tangent OOS (or of NB/p plotted for decision slopes) for two tests provides measure of difference in net benefit from use of two tests
  – Net benefit can be combined with cost of tests themselves, cost of delay in treatment due to tests, and risk from tests to identify appropriate test

Simplistic Definition of Net Benefit?

• Definition of value assumed by intercept of tangent line is that costs of mistakes are related to test moving us to wrong side of treatment threshold (i.e., one (test) and done decision making)
  – Potentially could account for differential mistake costs related to magnitude of distance we move above or below threshold (if using “continuous updating decision making”)
  • For example, larger post-test probabilities above threshold that, for those without disease, take you outside testing range may have greater costs than smaller probabilities above threshold that leave you within testing range

Method 2: Comparison of ROC Areas

• Area under ROC curve is a commonly reported statistic for diagnostic tests

• Many commentators believe it is an appropriate method of choosing between tests
  – Choose the test with the larger area

• ROC area is an indirect measure of the test’s “discriminating ability”
Discrimination

- Discrimination: Ability to assign different scores to those with and without disease
  - e.g., to assign generally lower scores to those without disease and to assign generally higher scores to those with disease
  - Discrimination is a property of scores
  - Given that predicted probabilities can be interpreted as scores, it applies to probabilities as well
  - For example, if a test score was 0.49 for everyone with disease and 0.51 for everyone without disease, the predictions would be perfectly discriminating
    - That is, everyone with disease has a higher score than everyone without disease

Bounds of ROC Areas

- ROC areas can range between 0.5 (area under 45° line of no information) and 1.0 (area under ROC curve of a dichotomous test that has 100% sensitivity and specificity)
  - Area of 0.5 represents no ability to discriminate risk
    - Test assigns a similar distribution of scores to those in whom disease is present and those in whom disease is absent
  - Area of 1.0 represents perfect discrimination
    - No overlap in the distribution of the scores assigned to those in whom disease is present and those in whom disease is absent

What Do Differences in ROC Areas Mean?

- Although curves with ROC areas of 0.5 and 1.0 are clearly distinguishable, there is little systematic information available about benefit of small increases in area under ROC curve (e.g., an increase from 0.75 to 0.77)
  - But, tests with larger areas under their ROC curve in general are more discriminating than are tests with smaller areas
Technical Interpretation of ROC Area

- Technically, ROC area equals probability that rule will correctly rank any randomly selected pair of persons, one in whom outcome of interest is present and one in whom it is absent
  - Nonparametric area represents the p-value we derive from a Wilcoxon rank sum test
  - How often do pairs of patients walk into a provider's office; declare that one has disease while the other does not, and then ask "which of us has a higher test score?"

ROC Area Reported in Many Applications

- The ROC area is used as a measure of discrimination in many applications other than diagnostic test evaluation
  - C-statistic that is routinely reported by SAS as an index of discriminating ability of fitted logistic regressions models equals nonparametric area under logistic regression's ROC curve
  - Similarly, lroc command in STATA that can be run after logistic regression reports same area

ROC Area and Choice for Individual

- ROC area inappropriate for choice for individual, given that area relates to curve as a whole, but choice problem focuses on choice between 2 optimal operating points, one for each test
- Difference in areas not a measure of costs of mistakes, and cannot be used to generate such a measure
Difficulties Interpreting ROC Areas: Curves Cross

- If curves cross, each test often has some operating points appropriate for certain individuals, whether or not ROC area of one test is same, less than, or greater than area of another.

Difficulties Interpreting ROC Areas: Dominance

- If areas are significantly different, doesn’t imply significance for all operating points.
- If not significantly different, doesn’t imply non-significance for all operating points.

WBC and IL-6

- Should we choose WBC or IL-6 for a patient for whom pre-test probability for bacteremia = 0.2 and ∆Od− = 0.25 ∆Od+ (OOS = 1)
ROC Areas for WBC and IL-6

• What information does following Stata roccomp area output provide for choosing between WBC and IL-6?

```
n . roccomp bact score, by(test)

<table>
<thead>
<tr>
<th>test</th>
<th>Obs</th>
<th>ROC Area</th>
<th>Std Err</th>
<th>-Asymptotic Normal-</th>
<th>95% Conf. Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (wbc)</td>
<td>885</td>
<td>0.7779</td>
<td>0.0446</td>
<td>0.69051</td>
<td>0.86537</td>
</tr>
<tr>
<td>1 (IL-6)</td>
<td>68</td>
<td>0.7717</td>
<td>0.0573</td>
<td>0.65938</td>
<td>0.88409</td>
</tr>
</tbody>
</table>
```

Ho:area(0=area(1)

chi²(1) = 0.01     Prob >chi²=0.9320

Comparison of Intercepts for WBC and IL-6

• If instead use intercepts of tangent lines (e.g., OOS =1):
  – Obtain 2 tangencies
  • Arrows on Y axis represent difference in clinical information

Assessing Magnitude of Difference Between 2 Intercepts: Equations

• Can use formula for tangent lines to compare difference in expected costs of mistakes yielded by two tests (e.g., one with a higher sensitivity [1] and one with a lower sensitivity [2])
• The formula for tangent lines is given by:
  Sensᵢ = (DS × [1-Specᵢ]) + Intᵢ
  Rearranging:
  Intᵢ = Sensᵢ - (DS × [1-Specᵢ])
  Thus:
  Intᵢ₁ = Sensᵢ₁ - DS (1-Specᵢ₁)
  Intᵢ₂ = Sensᵢ₂ - DS (1-Specᵢ₂)
Subtracting 2 intercepts, we get:

\[ \text{Int}_1 - \text{Int}_2 = \text{Sens}_1 - \text{Sens}_2 + \text{DS} (\text{Spec}_1 - \text{Spec}_2) \]

Example: Assessing Difference

- Test characteristics at tangency

<table>
<thead>
<tr>
<th>Test</th>
<th>Sens</th>
<th>Spec</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL-6, ≥2</td>
<td>0.773</td>
<td>0.761</td>
</tr>
<tr>
<td>WBC, ≥15</td>
<td>0.654</td>
<td>0.769</td>
</tr>
</tbody>
</table>

\[ \text{Int}_1 - \text{Int}_2 = (0.773-0.654) + 1(0.761-0.769) = 0.111 \]

- Multiplying 0.111 times \( \Delta \text{Od}^+ \) yields estimate of absolute net benefit of a positive test: 0.0222 \( \Delta \text{Od}^+ \)
  - i.e., for this patient, IL-6 reduces costs of mistakes by 0.0222 \( \Delta \text{Od}^+ \)

Statistical Significance

- Can test statistical significance of this difference in one of two ways:
  - Using equation:

\[ \text{SE}_{\text{Net}} = \left( \text{SE}_{\text{Sens}}^2 + (\text{DS} \times \text{SE}_{\text{Spec}})^2 - 2 \rho \times \text{SE}_{\text{Sens}} \times \text{DS} \times \text{SE}_{\text{Spec}} \right)^{0.5} \]

- IL-6, SE_{\text{Sens}} = .0893; SE_{\text{Spec}} = .0629, Nb_{\text{IL-6,ds=1}} = .1092
- WBC, SE_{\text{Sens}} = .0933; SE_{\text{Spec}} = .0144, Nb_{\text{WBC,ds=1}} = .0944
- Joint Nb_{\text{IL-6,ds=1}} = ((.1092)^2 + (.0944)^2)^{0.5} = 0.1444

- Bootstrapping intercepts of tangent lines (programs available at [www.uphs.upenn.edu/dgimhsr](http://www.uphs.upenn.edu/dgimhsr))
### Statistical Significance

- Given small numbers of patients with bacteremia in WBC (N = 26) and IL-6 (N = 22) studies, would not expect to have sufficient power to conclude 0.111 differs from 0

<table>
<thead>
<tr>
<th>Test</th>
<th>Intercept</th>
<th>Bootstrap SE</th>
<th>Se (from formula)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL-6, &gt;2</td>
<td>0.534</td>
<td>0.1089</td>
<td>0.1092</td>
</tr>
<tr>
<td>WBC, &gt;15</td>
<td>0.423</td>
<td>0.0806</td>
<td>0.0944</td>
</tr>
<tr>
<td>Difference</td>
<td>0.111</td>
<td>0.1354</td>
<td>0.1443</td>
</tr>
</tbody>
</table>

| p-value | 0.41 | 0.44 |
Intercepts of Lines Intersecting Single Operating Point

- Have previously focused on tangents
- Can instead think about intercepts of lines intersecting a single operating point
  - E.g., for $>15$ with sens = 0.654 and spec = 0.769

\[
\text{Intercept} = \text{Sensi} - (\text{OOS} \times (1 - \text{Speci}))
\]

\[\text{e.g., } 0.539 = 0.654 - 0.5 \times 0.231\]

Can Construct DS Graph for Each Operating Point From Both Tests

\[
\text{NB} = \frac{\text{sens} - \text{DS}}{\text{1-spec}}
\]
Plotting NB/p / Intercepts for WBC >15

- Horizontal axis = DS
- Vertical axis = intercept (NB/p)
- Intercept = Sens
- Slope = -(1-Spec)

Intercepts for ≥15

NB/p / Intercepts for All WBC Operating Points

Intercept = Sensi - (OOS × [1-Speci])

Largest WBC NB/p / Intercepts for All DS/OOS

Tangency between ROC curve and line defined with OOS occurs at operating point with largest intercept
Intercepts for All IL-6 Operating Points

Intercept = Sens = (OOS × [1-Spec])

Largest IL-6 NB/p / Intercepts for All DS/OOS

Tangency between ROC curve and line defined with OOS occurs at operating point with largest intercept

Largest NB/p / Intercepts for All DS/OOS

DS/OOS Crossovers

<table>
<thead>
<tr>
<th>DS/OOS Crossovers</th>
<th>DS/OOS Crossovers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Blue to Magenta</td>
<td>0.179</td>
</tr>
<tr>
<td>Magenta to Rose</td>
<td>0.459</td>
</tr>
<tr>
<td>Rose to Black</td>
<td>2.444</td>
</tr>
<tr>
<td>Black to Violet</td>
<td>3.073</td>
</tr>
<tr>
<td>Violet to Red</td>
<td>7.624</td>
</tr>
</tbody>
</table>

WBC 1.0, 0
WBC .923, 43
WBC 285, 920
WBC 231, 970
WBC 0, 1.0
**NB/p / Intercepts as a Summary Measure**

- NB/p / intercepts closest thing to a prevalence-free statistic for a test
  - Because any prevalence can be used to define a single DS/OOS
    - E.g., DS = 2 when $\Delta O_{d+} = 10,000$, $\Delta O_{d-} = 5000$, and $p = 0.2$
    - DS also equals 2 when $\Delta O_{d+} = 5000$, $\Delta O_{d-} = 10,000$, and $p = 0.5$
  - Because any difference in value can be used to define a single DS/OOS
- But intercepts don’t satisfy goal of single statistic for a test
  - Instead, provide a single statistic for each optimal operating slope for test

**Expected NB for Given DS Affected by Combinations of P, $\Delta O_{d+}$, and $\Delta O_{d-}$**

<table>
<thead>
<tr>
<th>$\Delta O_{d+}$</th>
<th>$\Delta O_{d-}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,000</td>
<td>5,000</td>
</tr>
<tr>
<td>10,000</td>
<td>15,000</td>
</tr>
</tbody>
</table>

**Do SSLR Provide Sufficient Information?**

- Previously indicated don’t need to construct an ROC curve to identify optimal cut-off for a positive test
  - If SSLR greater than OOS, test result is positive
  - IF SSLR less than OOS, test result is negative
- Is a comparison of SSLR sufficient for choosing among tests?
SSLR and Choice

• Suppose you are comparing 2 tests
  – Each has 3 strata
  – SSLR for test 1 identical to SSLR for test 2
    • SSLR = 0.1, 1.1, and 10
  – Cost of tests themselves (i.e., technician, reagents, etc.), cost of delay in treatment due to tests, and risk from tests themselves (e.g., adverse reactions) are identical

• Can one test be better than other?
  – If so, do SSLR alone provide information we need to choose between tests?

SSLR Alone Not Sufficient for Choice

Blue and Red Test Operating Points
Blue Test Has Larger or Equal Intercepts

OOS = 2 and 0.6, respectively

Blue and Red Test Intercepts

Identical SSLR / Different Clinical Information

- 3 SSLR for "inner" dotted red test are identical to 3 SSLR for "outer" dashed blue test
  - 0.1, 1.1, 10
- But blue test has intercepts of OOS that are greater than or equal to red test's intercepts
- Thus, YES, blue test is better than red test
  - Sharing SSLRs doesn't mean identical sensitivities and specificities
Identical SSLR / Different Clinical Information (2)

- And thus, NO, knowing tests’ SSLR not sufficient for determining if tests have same or different information
- Test’s information depends on:
  a) Test characteristics (e.g., SSLR) and
  b) Fraction of people with and without disease who will receive different test results
  - Can be calculated from sensitivity and specificity
    - Blue test has substantially higher proportions of patients who have test results with SSLR of 0.1 and 10
    - Red test has substantially higher proportion of patients who have test results with SSLR of 1.1

SSLR and Choice (2)

- Suppose you are again comparing 2 tests
  - Each has 3 strata
  - SSLR for test 1 are all “better” than SSLR for test 2

<table>
<thead>
<tr>
<th>SSLR</th>
<th>Test result</th>
<th>Test 1</th>
<th>Test 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>10</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>1.168</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>0.167</td>
<td>0.308</td>
<td></td>
</tr>
</tbody>
</table>

- Is test 1 better than test 2?
- If so, do SSLR alone provide information we need to choose between tests?
Blue and Red Test Intercepts

Cut-offs For Using Tests

(REPETITION) The quality of test depends both on magnitude of SSLR and differences in proportions of population having test results

SSLR throw away latter information
Is There a Prevalence-Independent Indicator That Allows Us to Compare Tests?

• Presuming choice between tests includes information about expected costs of mistakes, and
• Given that expected cost of mistakes is a function of prevalence of disease, test characteristics and cost of mistakes when they occur:
  → There can't be a single prevalence-independent indicator that would allow us to choose between tests for all individuals
• Proposed prevalence independent indicators (e.g., Youden index and diagnostic odds ratio) are blind alley!!!

Summary, 2x2 Approach, Choice for Individual

• The comparison of interest is between single operating points on two ROC curves
• Differences in areas refer to set of possible 2x2 tables, not to optimal 2x2 table, and provide no quantitative measure of difference in cost of mistakes between two tests
• Comparisons of intercepts of OOS evaluate two single operating points and provide a quantitative measure (and statistical test) of difference in cost of mistakes

2x2 Choice for a Formulary

• When comparing intercepts, how does the choice problem change when choosing for a formulary?
  – Choice for the individual is based on a single OOS whereas choice for a formulary is based on multiple OOS required for heterogeneous population
OOS Used Equifrequently?

- No requirement that all OOS will be used equi-frequently
  - e.g., based on subgroups with different pre-test probabilities and costs of mistakes, might use OOS as follows:

<table>
<thead>
<tr>
<th>OOS</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
</tr>
<tr>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>0.5</td>
<td>20</td>
</tr>
</tbody>
</table>

- The 40% with an OOS of 2 need not all have the same p or cost, just the same ratios of p and cost

Multiple "Individuals", Multiple Intercepts

- Suppose we can use two tests to make diagnosis for 2 types of patients, one with an OOS of 0.5, the other with an OOS of 2.0?

Example

- Two hypothetical tests (which yield the ROC curves in the previous slide)

<table>
<thead>
<tr>
<th>Score</th>
<th>Test 1</th>
<th>Test 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D+</td>
<td>D-</td>
</tr>
<tr>
<td>3</td>
<td>300</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>1</td>
<td>150</td>
<td>425</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>Sens</th>
<th>1-Spec</th>
<th>Sens</th>
<th>1-Spec</th>
</tr>
</thead>
<tbody>
<tr>
<td>3+</td>
<td>0.6</td>
<td>0.05</td>
<td>0.85</td>
<td>0.3</td>
</tr>
<tr>
<td>2+</td>
<td>0.7</td>
<td>0.15</td>
<td>0.95</td>
<td>0.4</td>
</tr>
</tbody>
</table>
Stata Commands for ROC Areas

```
. roctab dis score if test==1

ROC Area Std Err  -Asymptotic Normal-
Obs 1000 0.8025 0.0125  [95% Conf. Interval] 0.77793 0.82707

. roctab dis score if test==2

ROC Area Std Err  -Asymptotic Normal-
Obs 1000 0.8025 0.0125  [95% Conf. Interval] 0.77793 0.82707
```

Estimating Intercepts

\[ \text{Int} = \text{Sens} - (\text{OOS} \times (1 - \text{spec})) \]

- **Intercepts**
  - **OOS = 0.5**
    - Test 1: \(0.70 - (0.5 \times 0.15) = 0.625\)
    - Test 2: \(0.95 - (0.5 \times 0.40) = 0.75\)
  - **OOS = 2**
    - Test 1: \(0.60 - (2 \times 0.05) = 0.50\)
    - Test 2: \(0.85 - (2 \times 0.30) = 0.25\)

Example *

```
• bootstrap "rocintercept dis score 0.5 if test==1" "cutoff intercept optsens optspec area", reps(1000) saving(bscomptest2) replace strata(dis) notable

<table>
<thead>
<tr>
<th>Test</th>
<th>N</th>
<th>Bootstrap Intercept</th>
<th>SE</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOS  = 0.5</td>
<td>1</td>
<td>1000</td>
<td>0.6249984</td>
<td>0.0221688</td>
<td>0.5515</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1000</td>
<td>0.7500025</td>
<td>0.0143058</td>
<td>0.7032</td>
</tr>
<tr>
<td>OOS  = 2.0</td>
<td>1</td>
<td>1000</td>
<td>0.5014063</td>
<td>0.0295849</td>
<td>0.4117</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1000</td>
<td>0.2500358</td>
<td>0.0436304</td>
<td>0.1176</td>
</tr>
</tbody>
</table>
```

* Command for row 1 of table
Statistical Test, Difference in Intercepts

- Slope = 0.5
  \[ \text{display } 2 \times \text{ttail}(68, (((0.75 - 0.625)/((0.0221688^2 + 0.0143058^2)^{0.5})))) \]
  \[ 0.000142 \]
- Slope = 2.0
  \[ \text{display } 2 \times \text{ttail}(1998, (((0.5 - 0.25)/((0.0295849^2 + 0.0436304^2)^{0.5})))) \]
  \[ 0.0001122 \]
- Conclusion: p<0.0000 that when the OOS = 0.5, test 2’s intercept is greater than test 1’s AND that when the OOS = 2.0, test 1’s intercept is greater than test 2’s

Combining Multiple Intercepts

- The weighted average of differences in the intercepts for each OOS (where weights are determined by expected frequency of use of each OOS) represents the quantitative measure of the difference in two tests’ expected costs of mistakes
  - Suppose we expect to use 0.5 60% of the time and 2.0 40%
    - Test 1: \((0.6 \times 0.625) + (0.4 \times 0.5) = 0.575\)
    - Test 2: \((0.6 \times 0.75) + (0.4 \times 0.25) = 0.55\)
  - Combined SEs? Sampling uncertainty for costs?
- Thus, comparison of the intercepts for each optimal operating slope allows the determination of the better test as well as the quantification of the costs of mistakes for each test

Summary, 2x2 Choice for the Formulary

- The comparison of interest is between the relevant (potentially all) pairs of operating points on two ROC curves
- A weighted average of comparisons of intercepts for pairs of operating points provides a quantitative measure (and statistical tests) of the difference in the costs of mistakes
Summary

• When selecting among tests, we should maximize clinical information while minimizing cost.
• If we are choosing among tests for individual, we are choosing between single operating points from two tests for "like" patients.
• If we are choosing among tests for the formulary, we are choosing between multiple pairs of operating points one from each of two tests, where each pair represents a set of "like" patients.
• Comparison of intercepts of tangent OOS for two tests provides a measure of difference in costs of mistakes made by use of two tests as well as a statistical test of this difference.