Choosing Among Continuously Scaled Tests

Henry Glick

Epi 550

February 8, 2012

Outline

• Choosing among tests
  – Two-by-two approach and choice
    • Choice for individual
    • Choice for formulary
  – SSLR approach and choice

Choice Criteria

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

• Is the choice being made for an individual patient?
  – In the 2x2 approach, we are choosing between single operating points from two tests for "like" patients (where "like" means equivalent OOS)

• Is the choice being made for a formulary (i.e., which one of several tests do we want to make available to clinicians for testing for a given disease?)
  – In the 2x2 approach, we are choosing between multiple pairs of operating points, one from each of two tests, where each pair represents a set 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
• Method 2. Compare tests' ROC area

Method 1. Intercept of Tangent OOS and ROC Curve

• As we have already indicated, the cost of mistakes for any operating point is a function of the intercept of the OOS that is tangent to the ROC curve and equals:
  \[ P \times CFN \times (1 - \text{Intercept}) \]

• Thus, comparison of intercepts of the tangent OOS for the two tests provides a measure of the difference in the costs of mistakes made by use of the two tests
  – Costs of mistakes can be combined with the cost of tests themselves, the cost of delay in treatment due to the tests, and the risk from tests to identify the appropriate test
Simplistic Definition of Costs of Mistakes?

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

Method 2. ROC Area

- Common method proposed in literature for comparing tests (i.e., select the test with greatest area)
- Inappropriate for choices among tests for the individual, given that the measure of clinical information of interest relates to the 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 the curves cross, each test often has some operating points appropriate for certain individuals, whether or not the ROC area of one test is the same, less than, or greater than the 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

- Suppose we wish to choose between WBC and IL-6 for a patient for whom we believe the pre-test probability for bacteremia \(= 0.2 \) and \( C_{FP} = 0.25 \) \( C_{FN} \) (OOS = 1).

ROC Areas for WBC and IL-6

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

<table>
<thead>
<tr>
<th>test</th>
<th>Obs</th>
<th>ROC Area</th>
<th>Std Err</th>
<th>-Asymptotic Normal- [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 - 0.86537</td>
</tr>
<tr>
<td>1 (IL-6)</td>
<td>68</td>
<td>0.7717</td>
<td>0.0573</td>
<td>0.65938 - 0.88409</td>
</tr>
</tbody>
</table>

Ho: area(0)=area(1)
\( \chi^2(1) = 0.01 \)  Prob >\( \chi^2 = 0.9320 \)
Comparison of Intercepts for WBC and IL-6

- Suppose we instead use the intercepts of tangent lines to choose between WBC and IL-6 for a patient for whom we believe the pre-test probability = 0.2 and the ratio of CFP to CFN = 0.25 (i.e., OOS = 1.0)
- We obtain 2 tangencies
- Arrows on Y axis represent difference in clinical information

Assessing the Magnitude of the Difference Between 2 Intercepts: Equations

- We can use the formula for the tangent lines to compare the difference in the 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 the tangent lines is given by:
  \[ \text{Sensi} = (\text{OOS} \times [1-\text{Speci}]) + \text{Inti} \]
- Rearranging:
  \[ \text{Inti} = \text{Sensi} - (\text{OOS} \times [1-\text{Speci}]) \]
- Thus:
  \[ \text{Int}_1 = \text{Sensi}_1 - \text{OOS} \times (1-\text{Spec}_1) \]
  \[ \text{Int}_2 = \text{Sensi}_2 - \text{OOS} \times (1-\text{Spec}_2) \]

Subtracting \text{Int}_2 from \text{Int}_1:
\[ \text{Int}_1 - \text{Int}_2 = \text{Sensi}_1 - \text{Sensi}_2 + \text{OOS} \times (\text{Spec}_1 - \text{Spec}_2) \]
Example: Assessing the 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>

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

\[ (\text{Sens}_1-\text{Sens}_2) + \text{OOS}(\text{Spec}_1-\text{Spec}_2) \]

- Given that at the intercept, mistakes are made among people with disease only, the value of this difference is:

\[ 0.2 \times C_{FN} \times 0.111 = 0.0222 C_{FN} \]

- i.e., for this patient, IL-6 reduces the costs of mistakes by 0.0222 \( C_{FN} \)

Statistical Significance

- Can test the statistical significance of this difference by bootstrapping the intercepts of the tangent lines (programs available at www.uphs.upenn.edu/dgimhsr)

<table>
<thead>
<tr>
<th>Test</th>
<th>Intercept</th>
<th>SE</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL-6, ≥2</td>
<td>0.534</td>
<td>0.1089</td>
<td>0.41</td>
</tr>
<tr>
<td>WBC, ≥15</td>
<td>0.423</td>
<td>0.0806</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>0.111</td>
<td>0.1354</td>
<td></td>
</tr>
</tbody>
</table>

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

Do SSLR Provide Sufficient Information?

- Previously indicated we don’t need to construct an ROC curve to identify the optimal cut-off for a positive test
- If the OOS is 1.0, the lines below define positive and negative strata

<table>
<thead>
<tr>
<th>Test</th>
<th>SSLR</th>
<th>SSLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBC</td>
<td>Score</td>
<td>Score</td>
</tr>
<tr>
<td>25+</td>
<td>7.624</td>
<td>3+</td>
</tr>
<tr>
<td>20-25</td>
<td>3.073</td>
<td>2-3</td>
</tr>
<tr>
<td>15-20</td>
<td>1.793</td>
<td>0-2</td>
</tr>
<tr>
<td>10-15</td>
<td>0.792</td>
<td>0.299</td>
</tr>
<tr>
<td>&lt;10</td>
<td>0.179</td>
<td></td>
</tr>
</tbody>
</table>

- Is the data above sufficient to identify the better test?
SSLR and Choice

• Suppose you are comparing 2 tests
  – Each has 3 strata
  – The 3 SSLR for test 1 are identical to the 3 SSLR for test 2 (0.25, 1, and 4)
  – The cost of the tests themselves (i.e., technician, reagents, etc.), the cost of delay in treatment due to the tests, and the risk from tests themselves (e.g., adverse reactions) are identical

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

SSLR Alone Not Sufficient for Choice

SSLR Alone Not Sufficient for Choice (cont.)

• The 3 SSLR for the "inner" dotted test are identical to the 3 SSLR for the "outer" dashed test
  – Thus, YES, one test can be better than the other
  – The fact that two tests share SSLRs doesn’t necessarily mean they have the same sensitivities and specificities

• And NO, the SSLR alone do not provide sufficient information to choose between the tests
  – Because SSLR are identical, but the tests differ in the clinical information they provide
SSLR and Choice (2)

- Suppose you are again comparing 2 tests
  - Each has 3 strata
  - The 3 SSLR for test 1 are all “better” than the 3 SSLR for test 2
    - For each of the 2 strata with SSLR > 1, each of test 1’s SSLR are larger than test 2’s
      - e.g., 1.69 and 10 VS 1 and 4
    - For the strata with SSLR < 1, test 1’s is smaller than test 2’s
      - e.g., 0.167 vs 0.308
- Must test 1 be better than test 2, and if so, do the SSLR alone provide the information we need to choose between the tests?

SSLR Alone Still Not Sufficient

Where Do the Cut-Offs Come From?
The quality of the test depends both on the magnitude of the SSLR and the differences in the proportions of the population having the test results.

Is There a Prevalence-Independent Indicator That Allows Us to Compare Tests?

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

Summary, 2x2 Approach, Choice for the Individual

- The comparison of interest is between single operating points on two ROC curves.
- Differences in areas refer to the set of possible 2x2 tables, not to the optimal 2x2 table, and provide no quantitative measure of the difference in cost of mistakes between the two tests.
- Comparisons of intercepts of OOS evaluate the two single operating points and provide a quantitative measure (and statistical test) of the difference in cost of mistakes.
2x2 Choice for the Formulary
Method 1. Comparison of Intercepts

- When comparing intercepts, how does the choice problem change when choosing for the individual vs the formulary?
  - Choice for the individual is based on 1 common OOS whereas choice for the formulary is based on multiple OOS

Method 1. Comparison of Intercepts (cont.)

- There is no requirement that all OOS will be used equifrequently
  - 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

2x2 Choice for the Formulary
Method 2. ROC Area

- Information of interest: Costs of mistakes arising from multiple pairs of operating points one from each of two tests
- ROC area potentially more appropriate for choices among tests for the formulary (i.e., across all possible operating points), but:
  - Would need to use all the potential test operating points equifrequently (unlikely)
  - Doesn't provide a quantitative measure of costs of mistakes
Multiple "Individuals", Multiple Intercepts

• Suppose we used two tests to diagnose 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

<table>
<thead>
<tr>
<th>Obs</th>
<th>ROC</th>
<th>Area</th>
<th>Std Err</th>
<th>-Asymptotic Normal-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[95% Conf. Interval]</td>
</tr>
<tr>
<td>1000</td>
<td>0.8025</td>
<td>0.0125</td>
<td>0.77793</td>
<td>0.82707</td>
</tr>
</tbody>
</table>

. roctab dis score if test==2

<table>
<thead>
<tr>
<th>Obs</th>
<th>ROC</th>
<th>Area</th>
<th>Std Err</th>
<th>-Asymptotic Normal-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[95% Conf. Interval]</td>
</tr>
<tr>
<td>1000</td>
<td>0.8025</td>
<td>0.0125</td>
<td>0.77793</td>
<td>0.82707</td>
</tr>
</tbody>
</table>
Estimating Intercepts
Int = Sens - (OOS * (1-spec))

- Intercepts
  - OOS = 0.5
    Test 1: 0.70 - (0.5 * 0.15) = 0.625
    Test 2: 0.95 - (0.5 * 0.40) = 0.75
  - OOS = 2
    Test 1: 0.60 - (2 * 0.05) = 0.50
    Test 2: 0.85 - (2 * 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>OOS</th>
<th>N</th>
<th>Bootstrap Intercept</th>
<th>SE</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>1000</td>
<td>0.6249984</td>
<td>0.021688</td>
<td>0.5515</td>
<td>0.6931</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1000</td>
<td>0.750025</td>
<td>0.0143058</td>
<td>0.7032</td>
<td>0.7967</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1000</td>
<td>0.5014063</td>
<td>0.0295849</td>
<td>0.4117</td>
<td>0.5854</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1000</td>
<td>0.2500358</td>
<td>0.0436304</td>
<td>0.1176</td>
<td>0.3823</td>
</tr>
</tbody>
</table>

* Command for row 1 of table

Statistical Test, Difference in Intercepts
- Slope = 0.5
  - display 2*ttail(68,(.75-.625)/((.0221688^2)+(.0143058^2))^(.5)))
    0.000142
- Slope = 2.0
  - display 2*ttail(1998,(.5-.25)/((.0295849^2)+(.0436304^2))^(.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
- In a limited set of circumstances, comparison of areas under the ROC curves of two tests may identify the better test, but does not allow quantification of the difference in the tests' cost of mistakes
- 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

Comparison of Tests Summarized with Likelihood Ratios

- Will we make the same choice between tests if we use the tests' SSLR rather than their sensitivity and specificity?
  - YES for "One and Done" decision making
  - NOT NECESSARILY for "continuous updating" decision making
SSLR and "One and Done" Decision Making

- "One and done" decision making assumes the same definition of costs that we used above
  - The cost of mistakes is the same for any post-test probability that for diseased individuals is below the underlying treatment threshold and that for nondiseased individuals is above the underlying treatment threshold
  - i.e., for those with disease, mistakes marginally below the threshold have costs similar to mistakes substantially below the threshold
  - i.e., for those without disease, mistakes marginally above the threshold have costs similar to mistakes substantially above the threshold

SSLR and "One and Done" Decision Making (2)

- Given the equations that show that for one (test) and done decision making, the optimal 2×2 table and the SSLR approach yield identical treatment recommendations:
  - The cost of mistakes in the stratum-specific approach is identical to the cost of mistakes in the approach comparing intercepts of tangent lines
  - The comparison of tests for either the individual or for the formulary based on the intercepts of optimal operating slopes tangent to ROC curves can be used when the test is being characterized as a series of stratum-specific likelihood ratios

SSLR and "Continuous Updating" Decision Making

- "Continuous Updating" decision making requires a more complex definition of the cost of a mistake
  - Given we are interested in the result from a series of revisions of probabilities of disease, the cost should reflect how the final predicted probability -- and resulting FP and FN mistakes -- is affected by the individual tests we use in the chain
  - It should NOT reflect whether a particular test result moves you above or below the treatment threshold
SSLR and "World Enough" Decision Making (2)

• One version of such a definition:
  – Suppose we quantified the costs of absolute increases in the post-test probability of disease for people without disease and the costs of absolute decreases in the post-test probability of disease for people with disease?

More Complicated Definition of the Cost of Mistakes

• Under this definition, the total costs of mistakes for a test are defined as follows:

\[
\sum_{i} p_{i|\text{Dis}} \left( \frac{p \cdot \text{SSLR}}{(p \cdot \text{SSLR}) + (1-p)} \right) C_{\text{FN}i}\]

\[+\]

\[
\sum_{j} (1-p)_{j|\text{NoDis}} \left( \frac{p \cdot \text{SSLR}}{(p \cdot \text{SSLR}) + (1-p)} \right) C_{\text{FP}j}\]

where: \( p \) = pre-test probability of disease; \( i \) = strata with SSLR < 1; \( j \) = strata with SSLR > 1; SSLR = the likelihood ratio for test results in stratum \( i \); SSLR = the likelihood ratio for test results in stratum \( j \); \( p_{i|\text{Dis}} \) = the probability that individuals with disease will have a test result falling within stratum \( i \); \( p_{j|\text{NoDis}} \) = the probability that individuals without disease will have a test result falling within stratum \( j \); \( C_{\text{FN}i} \) = the cost of false negative results given the reduction in the probability of disease due to an LR < 1; \( C_{\text{FP}j} \) = the cost of false positive results given the increase in the probability of disease due to an LR > 1

Understanding the Formula

\[
\sum_{\text{prob} \text{ lower dis}} \text{prob} \text{ lower dis} \times \text{post-test} \times \text{change in p} \times \text{cost} \]

\[
\sum_{\text{prob} \text{ higher nodis}} \text{prob} \text{ higher nodis} \times \text{post-test} \times \text{change in p} \times \text{cost} \]
A More Complicated Definition of the Cost of Mistakes

- Finally, we can quantify both the costs of mistakes from absolute increases in the post-test probability of disease for people without disease and the costs of absolute decreases in the post-test probability of disease for people with disease (as above) as well as the benefits from correct increases and decreases in the post-test probability due to the test results.

Summary

- When selecting among tests, we should maximize clinical information while minimizing cost.
- If we are choosing among tests for the 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 the tangent OOS for the two tests provides a measure of the difference in the costs of mistakes made by use of the two tests as well as a statistical test of this difference.