Discrimination and Reclassification

in

Statistics and Study Design

AACC/ASN 30th Beckman Conference

Michael J. Pencina, PhD

Duke Clinical Research Institute
Duke University
Department of Biostatistics and Bioinformatics
Acknowledgments

Disclosures
Conflicts related to this presentation: none

Acknowledgement
Research has been conducted in collaboration with Dr. Ralph B. D’Agostino, Sr.
Discrimination...
and Reclassification

STATISTICS

LIES

DAMN LIES

1st Circle: Limbo
2nd Circle: Lust
3rd Circle: Gluttony
4th Circle: Avarice & Prodigality
5th Circle: Wrath & Sullenness
6th Circle: Heresy
7th Circle: Violence
8th Circle: Fraud
9th Circle: Treachery
Definitions

- Risk prediction FUNCTION: method of assigning risks of disease to individuals based on (baseline) variables known as risk factors.

- Risk prediction RULE: method of assigning individuals into clinically meaningful categories based on their predicted risk of disease.
Wang et al. Multiple Biomarkers for the Prediction of First Major Cardiovascular Events and Death. NEJM 2006

“In summary, biomarkers from multiple, biologically distinct pathways are associated with the risks of death and major cardiovascular events. Nonetheless, the use of contemporary biomarkers adds only moderately to standard risk factors for risk assessment of individual persons.”
Damn lies (not really)

Zethelius et al. Use of Multiple Biomarkers to Improve the Prediction of Death from Cardiovascular Causes. NEJM 2008

“Our data suggest that in elderly men with or without prevalent cardiovascular disease, the simultaneous addition of several biomarkers of cardiovascular and renal abnormalities substantially improves the risk stratification for death from cardiovascular causes beyond that of a model that is based only on established risk factors.”
and Statistics (really)...

- All-cause mortality
  - Wang et al.: C statistic changes from 0.80 to 0.82
  - Zethelius et al.: C statistic changes from 0.604 to 0.676

- CV events (Wang) or mortality (Zethelius)
  - Wang et al.: C statistic changes from 0.76 to 0.77
  - Zethelius et al.: C statistic changes from 0.69 to 0.75
What is C statistic?

- Area under ROC curve, i.e. plot of Sensitivity versus 1-Specificity

- Probability that given two randomly selected subjects, one with event and one without, model-based risk is higher for the one with event.
Change in C depends where we start

<table>
<thead>
<tr>
<th>Effect size of added marker</th>
<th>Baseline AUC</th>
<th>0.50</th>
<th>0.55</th>
<th>0.60</th>
<th>0.65</th>
<th>0.70</th>
<th>0.75</th>
<th>0.80</th>
<th>0.85</th>
<th>0.90</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>ΔAUC</td>
<td>0.21</td>
<td>0.17</td>
<td>0.13</td>
<td>0.10</td>
<td>0.08</td>
<td>0.06</td>
<td>0.04</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>0.5</td>
<td>ΔAUC</td>
<td>0.14</td>
<td>0.10</td>
<td>0.07</td>
<td>0.05</td>
<td>0.04</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>0.2</td>
<td>ΔAUC</td>
<td>0.06</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Would other measures give different results?

- Discrimination Slope defined as difference in means of model-based probabilities for events minus nonevents
- Asymptotically equivalent to $R^2$ for binary outcomes
- Equal to integrated sensitivity plus integrated specificity minus 1 (hence difference called IDI)
- Alternative to C statistic as measure of discrimination but influenced by model calibration
Discrimination Slope

Discrimination histograms
0=non-events, 1=events

Estimated Probability

Mean 0.100
Std Dev 0.087

Mean 0.200
Std Dev 0.134
Net Reclassification Improvement

• IDI weights each change by its magnitude

• Instead we can weight upward movements with 1, and downward with -1

• Resulting statistic called continuous net reclassification improvement: NRI(>0)

• NRI(>0) minimally influenced by strength of baseline model
Visualizing impact of new marker

Figure 1: Reclassification plot
## Impact of uncorrelated new predictor

<table>
<thead>
<tr>
<th>$d$</th>
<th>$\Delta$AUC</th>
<th>0.50</th>
<th>0.55</th>
<th>0.60</th>
<th>0.65</th>
<th>0.70</th>
<th>0.75</th>
<th>0.80</th>
<th>0.85</th>
<th>0.90</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>$\Delta$AUC</td>
<td>0.21</td>
<td>0.17</td>
<td>0.13</td>
<td>0.10</td>
<td>0.08</td>
<td>0.06</td>
<td>0.04</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>IDI(0.1)</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>NRI(&gt;0)</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
</tr>
<tr>
<td>0.5</td>
<td>$\Delta$AUC</td>
<td>0.14</td>
<td>0.10</td>
<td>0.07</td>
<td>0.05</td>
<td>0.04</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>IDI(0.1)</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>NRI(&gt;0)</td>
<td>0.40</td>
<td>0.40</td>
<td>0.40</td>
<td>0.40</td>
<td>0.40</td>
<td>0.40</td>
<td>0.40</td>
<td>0.40</td>
<td>0.40</td>
</tr>
<tr>
<td>0.2</td>
<td>$\Delta$AUC</td>
<td>0.06</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>IDI(0.1)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>NRI(&gt;0)</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
</tr>
</tbody>
</table>
Are new metrics better?

- Not necessarily; they have a different focus.
- AUC’s main focus is on improving model at hand.
- NRI(>0)’s main focus is on new predictor.
- IDI falls in the middle.
- When do they disagree? For extreme baseline models.
More lies…

• Lies: Markers that do not lead to statistically significant increases in AUC statistic cannot improve model performance

• Damn lies: Markers that lead to statistically significant increases in AUC, IDI or NRI always improve model performance

• Statistics: Focus on magnitude and clinical impact!
More statistics...

- Hypothesis testing based on change in AUC (or any other measure of performance) is incorrect (usually overly conservative) for nested models.

- On the other hand, for large samples anything will be statistically significant.

- Present only likelihood ratio test for new biomarker and estimate change in AUC, IDI or NRI with confidence interval.
How to interpret magnitude?

• Some heuristic benchmarks based on normally distributed data:
  
  • At fixed Specificity of 0.75, each 0.010 increase in C statistic increases Sensitivity by about 0.017

  • NRI(>0) <0.20 corresponds to adding weak independent predictor, >0.60 strong, 0.40 intermediate

  • Relative IDI = IDI/baseline slope \( \approx \frac{1}{\#\text{predictors}} \) suggests strength equal to average of those included
Applied researchers argue that performance metrics quantifying usefulness of new markers should be tied to their impact on clinical decisions.

This suggests evaluation of prediction rules.

In some settings there exist meaningful cut-offs for assignment of risk categories.
Example: CACS in MESA

Polonsky et al. Coronary Artery Calcium Score and Risk Classification for Coronary Heart Disease Prediction. JAMA 2010

Model performance metrics (for 5-year risk):

- C statistic increases from 0.76 to 0.81
- IDI = 0.026
- relative IDI = 0.81 (1/#predictors = 0.14)
## Reclassification Table

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Up</th>
<th>Down</th>
<th>NRI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Events</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No CACS model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model with CACS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 3%</td>
<td>51</td>
<td>81</td>
<td>77</td>
<td>209</td>
</tr>
<tr>
<td>3-10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt; 10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>51</td>
<td>81</td>
<td>77</td>
<td>209</td>
</tr>
</tbody>
</table>

|                  |       |      |      |     |
| No CACS model    |       |      |      |     |
| Model with CACS  |       |      |      |     |
| < 3%             | 34    | 22   | 1    | 57  |
| 3-10%            | 15    | 52   | 48   | 115 |
| > 10%            | 2     | 7    | 28   | 37  |

|                  |       |      |      |     |
| No CACS model    |       |      |      |     |
| Model with CACS  |       |      |      |     |
| < 3%             | 3276  | 408  | 5    | 3689|
| 3-10%            | 697   | 791  | 244  | 1732|
| > 10%            | 30    | 63   | 155  | 248 |

|                  |       |      |      |     |
| **Total**        | 4003  | 1262 | 404  | 5669|
Net Reclassification Improvement

• NRI for events = (71 – 24) * (1/209) = 0.23

• NRI for nonevents = (790 – 657) * (1/5669) = 0.02

• NRI combined = 0.23 + 0.02 = 0.25

• Simple addition as above implies weight between events and nonevents of (1-p)/p (non-event odds), 27:1 in this case
Net Reclassification Improvement

• Reporting separate components of categorical NRI more informative than simple sum:
  
  • Allows researchers to apply their own weights
  
  • Allows to discern relative impact on events vs. non-events
  
• Weighted NRI can be constructed allowing for weights corresponding to given clinical situation
Lies: Apparent improvement in model performance based on matched case-control studies should translate into similar improvement in prospective studies.

Statistics: Pepe et al. (Clin. Chem. 2012) have shown that incremental value of biomarkers can be grossly exaggerated in matched case-control studies.

Even more lies…
And more statistics…

• They give example of a marker which increased the AUC by 0.17 in matched case-control study but this increase became 0.02 in unmatched data

• Non-matched case-control studies give a more accurate assessment

• Problem partially explained by weaker performance of the baseline model in matched case-control setting

• Pepe et al. propose a statistical tool to correct the problem
Final Comments

- Model performance measures useful only if interpretable
- Rely on magnitude, not p-values
- Increase in C statistic heavily dependent on baseline model, NRI(>0) weakly dependent, IDI in the middle
- Prediction model and rule related but slightly different focus
- Unrealistic to expect that new markers will improve prediction rules without improving prediction models
Use of Biomarkers as Surrogates
Types of biomarkers
Based on FDA’s Qualification process for drug development tools

- A *prognostic* biomarker is a baseline patient or disease characteristic that categorizes patients by degree of risk for disease occurrence or progression. A prognostic biomarker informs about the natural history of the disorder in that particular patient in the absence of a therapeutic intervention.

- A *predictive* biomarker is a baseline characteristic that categorizes patients by their likelihood for response to a particular treatment. A predictive biomarker is used to identify whether a given patient is likely to respond to a treatment intervention in a particular way (favorable or not).
Types of biomarkers
Based on FDA’s Qualification process for drug development tools

• A *pharmacodynamic* (or activity) biomarker is a dynamic assessment that shows that a biological response has occurred in a patient after having received a therapeutic intervention. A pharmacodynamic biomarker may be treatment-specific or more broadly informative of disease response. Examples include blood pressure, cholesterol, HbA1c, intraocular pressure, radiographic measures, and C-reactive protein.

• A *surrogate endpoint* is defined as a biomarker intended to substitute for a clinical efficacy endpoint. Surrogate endpoints are expected to predict clinical benefit (or harm).
Surrogate Endpoints
Based on FDA’s Qualification process for drug development tools

- A clinical endpoint is defined as a characteristic or variable that reflects how a patient feels, functions, or survives.

- Surrogate endpoints are a subset of pharmacodynamic biomarkers; it is likely that only a few biomarkers will be appropriate for use as surrogate endpoints.

- Because there is substantial risk of adversely affecting the public health if a biomarker is falsely accepted as a surrogate endpoint, robust scientific evidence is needed to justify qualification of a biomarker for broad use as a surrogate.
# Desirable Characteristics of Surrogates

<table>
<thead>
<tr>
<th>Austin Bradford Hill's guidelines increasing the likelihood of causative association</th>
<th>Characteristics of useful biomarkers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Guidelines</strong></td>
<td><strong>Characteristics of useful biomarkers</strong></td>
</tr>
<tr>
<td>Strength</td>
<td>A strong association between marker and outcome, or between the effects of a treatment on each</td>
</tr>
<tr>
<td>Consistency</td>
<td>The association persists in different individuals, in different places, in different circumstances, and at different times</td>
</tr>
<tr>
<td>Specificity</td>
<td>The marker is associated with a specific disease</td>
</tr>
<tr>
<td>Temporality</td>
<td>The time-courses of changes in the marker and outcome occur in parallel</td>
</tr>
<tr>
<td>Biological gradient (dose-responsiveness)</td>
<td>Increasing exposure to an intervention produces increasing effects on the marker and the disease</td>
</tr>
<tr>
<td>Plausibility</td>
<td>Credible mechanisms connect the marker, the pathogenesis of the disease, and the mode of action of the intervention</td>
</tr>
<tr>
<td>Coherence</td>
<td>The association is consistent with the natural history of the disease and the marker</td>
</tr>
<tr>
<td>Experimental evidence</td>
<td>An intervention gives results consistent with the association</td>
</tr>
<tr>
<td>Analogy</td>
<td>There is a similar result to which we can adduce a relationship</td>
</tr>
</tbody>
</table>
Problems with Surrogate Endpoints

- Effect of treatment on biomarker does not translate into effect on clinical outcome:
  - Hemoglobin A1c and HDL are related to the risk of CVD, yet treatments that affect either marker do not affect clinical outcome

- Confounding: impact of treatment on other factors modifies its effect on the surrogate

- Regression to the mean: surrogate used as entry criterion and endpoint might lead to change that is not meaningful