Net Reclassification Risk: a graph to clarify the potential prognostic utility of new markers

Ewout Steyerberg
Professor of Medical Decision Making
Dept of Public Health, Erasmus MC

Birmingham
July, 2013
Ideas from

- Colleagues
  - Michael Pencina
  - Stuart Baker
  - Douwe Posmus
  - Ben van Calster
- PhD students:
  - Moniek Vedder
  - Maarten Leening
Overview

- Performance evaluation:
  - Prediction vs classification
  - Graphical displays
  - Traditional and novel measures

- Additional value of markers: NRI and other measures

- Apparent vs external validation
Performance evaluation of predictions

- First requirement: statistical significance
  - Likelihood ratio test
    - Model vs no model
    - Marker vs without marker
  - Do not use other test (DeLong for ROC area; NRI test; ..)

- Traditional: focus on prediction
  - Model fit (R²/ Brier/..)
  - Calibration (testing; recalibration; graphics)
  - AUC (c statistic / ..)
  - ...

*Steyerberg, Vickers, …, Kattan, Epidemiology 2010*
Performance evaluation: value of markers

- Better decision making by adding a marker
  - Net Benefit and decision curves (Vickers, *MDM* 2006)
  - Relative Utility (Baker, *JNCI* 2009)

- Focus shifts from continuous predictions to categorization
  - NRI often considered in CVD
    - 3 categories (thresholds: 6%, 20%)
  - NB popular in (prostate) cancer
    - 2 categories (biopsy threshold 10-40%)
Aim

- Compare graphical formats and summary measures to quantify usefulness of adding a marker to a model
Framingham case study

- Cohort: 3264 participants in Framingham Heart Study
- Age 30 to 74 years
- 183 developed CHD (10 year risk: 5.6%)
- Incremental value of adding HDL to Framingham model
  - HR=0.65; p<.0001

- Data used in
  - Pencina, Stat Med 2008, >1000 citations
  - Steyerberg, Rev Esp Cardiol 2011, 7 citations
Basic idea for discrimination

![Histograms showing distributions of Non-events and Events](image-url)

Logit
Density plots: marker leads to higher prediction for events, lower for non-events?
Box plots

Predicted risk without HDL

Predicted risk with HDL
Summarize differences in distributions

- Difference in mean + SD → delta c statistic: +0.012
- Log based scoring rules; Nagelkerke $R^2$: +1.6%
- Mean prediction by outcome:
  discrimination slope ≈ Pearson $R^2$ ≈ scaled Brier: +1.9%
Classic:
ROC curve plot (but often a waste of journal space)

AUC increases from 0.762 to 0.774

How to interpret 0.01 increase?
Interpretation of AUC

- Typically small improvement in discriminative ability according to AUC (or c statistic)
- c stat blamed for being insensitive

Letter by Pepe et al Regarding Article, “Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction”

To the Editor:

Current statistical approaches for evaluation of risk prediction markers are unsatisfactory. We applaud Cook’s criticisms of the c-index, or area under the receiver operating characteristic curve. This index is based on the notion of pairing subjects, one with poor outcome (eg, cardiovascular event within 10 years) and one without, and determination of whether the risk for the former (ie, the case) is larger than the risk for the latter (ie, the control). This probability of correct ordering of risks is not a relevant measure of clinical value. It should not play a central role in evaluation of risk markers.
A new perspective: reclassification
Figure. Reclassification of Risk Using the Reynolds Risk Score for a Representative Population of 100,000 Intermediate-Risk US Women Without Diabetes

100,000 Women With Intermediate CVD Risk

10-Year CVD Risk Stratification Using Adult Treatment Panel III Covariates\(^{45}\)

- 80%
- 20%

5% to <10% CVD Risk
- 80,000 Women

10% to <20% CVD Risk
- 20,000 Women

Reclassification of 10-Year CVD Risk Using Reynolds Risk Score

- 15.9%
- 3.8%
- 55.7%
- 19.9%
- 26.9%
- 55.1%
- 1.5%
- 21.2%

- 12,720 Women
- 44,560 Women
- 21,520 Women
- 1,200 Women

- 760 Women
- 3,980 Women
- 11,020 Women
- 4,240 Women

- <5% CVD Risk
- 5% to <10% CVD Risk
- 10% to <20% CVD Risk
- ≥20% CVD Risk

Low Risk | Low to Moderate Risk | Moderate to High Risk | High Risk

Percentages shown in ovals indicate the proportion of women distributed to risk categories based on Adult Treatment Panel III (top) and the Reynolds Risk Score (bottom). Reclassification using the Reynolds Risk Score is based on data shown in Table 5, Model B. CVD indicates cardiovascular disease.
Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond

Michael J. Pencina\textsuperscript{1,*},†, Ralph B. D’Agostino Sr\textsuperscript{1}, Ralph B. D’Agostino Jr\textsuperscript{2} and Ramachandran S. Vasan\textsuperscript{3}

\textsuperscript{1}Department of Mathematics and Statistics, Framingham Heart Study, Boston University, 111 Cummington St., Boston, MA 02215, U.S.A.
\textsuperscript{2}Department of Biostatistical Sciences, Wake Forest University School of Medicine, Medical Center Boulevard, Winston-Salem, NC 27157, U.S.A.
\textsuperscript{3}Framingham Heart Study, Boston University School of Medicine, 73 Mount Wayte Avenue, Suite 2, Framingham, MA 01702-5803, U.S.A.

- Net Reclassification Improvement:
  - (move up | event– move down | event) +
  - (move down | non-event – move up | non-event)
Table II. Reclassification among people who experience a CHD event and those who do not experience a CHD event on follow-up.

<table>
<thead>
<tr>
<th>Model without HDL</th>
<th>Model with HDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency (Row per cent)</td>
<td>&lt;6 per cent</td>
</tr>
<tr>
<td>Participants who experience a CHD Event</td>
<td></td>
</tr>
<tr>
<td>&lt;6 per cent</td>
<td>39 (72.22)</td>
</tr>
<tr>
<td>6–20 per cent</td>
<td>4 (3.81)</td>
</tr>
<tr>
<td>&gt;20 per cent</td>
<td>0 (0.00)</td>
</tr>
<tr>
<td>Total</td>
<td>43</td>
</tr>
</tbody>
</table>

22/183 = 12%

Participants who do not experience a CHD Event |

<table>
<thead>
<tr>
<th>Frequency (Row per cent)</th>
<th>&lt;6 per cent</th>
<th>6–20 per cent</th>
<th>&gt;20 per cent</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants who do not experience a CHD Event</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;6 per cent</td>
<td>1959 (93.24)</td>
<td>142 (6.76)</td>
<td>0 (0.00)</td>
<td>2101</td>
</tr>
<tr>
<td>6–20 per cent</td>
<td>148 (16.78)</td>
<td>703 (79.71)</td>
<td>31 (3.51)</td>
<td>882</td>
</tr>
<tr>
<td>&gt;20 per cent</td>
<td>1 (1.02)</td>
<td>25 (25.51)</td>
<td>72 (73.47)</td>
<td>98</td>
</tr>
<tr>
<td>Total</td>
<td>2108</td>
<td>870</td>
<td>103</td>
<td>3081</td>
</tr>
</tbody>
</table>

1/3081 = .03%

HDL cholesterol is routinely used in CHD prediction models [3–5]. Both IDI and NRI suggest that including it in the prediction model results in significant improvement in performance. That conclusion could not have been drawn relying solely on the increase in AUC. The increase in IDI, albeit significant, was of small magnitude—0.009 on the absolute scale or 7 per cent relative increase. It can be interpreted as equivalent to the increase in average sensitivity given no changes in specificity. Based on the NRI and its components, we conclude that addition of HDL improved classification for a net of 12 per cent of individuals with events, with no net loss for non-events. Even though the NRI results look convincing, caution needs to be given to their interpretation, as it is dependent on the somewhat arbitrary choice of categories.
NRI

- Introduced with 3 categories
- Extended to “any category”: continuous NRI
- Simplified to 2 categories: delta Youden index
Extension to any threshold: continuous NRI ([Stat Med 2011](https://journals.lww.com/statistics-in-medicine))
NRI events 24.6%; NRI non-e 5.5%; NRI +0.301
AUC increases from 0.550 to 0.579

How to interpret 0.029 increase?

Make it twice as big: $NRI = 0.058$
Criticism on NRI

- ‘Nothing new’; related to other measures
  - for binary classification $\text{NRI} = \text{delta sens} + \text{delta spec}$
  - AUC for binary classification $= \frac{\text{sens} + \text{spec}}{2}$
  - $\text{NRI} = 2 \times \text{delta AUC}$

- Interpretation:
  - delta AUC larger with fewer cut-offs;
  - NRI larger with more cut-offs
  - Interpretation as ‘net reclassified in population’ (%)
"The net reclassification index increased significantly after addition of intima-media thickness of the internal carotid artery (7.6%, P<0.001)"

Recent example, Gulati \textit{et al}, JAMA 2013;
- LTTE Leening
- Reply: NRI components more important than their sum; and erratum
Limitation of NRI

- Nothing new
- Interpretability / misleadingness
- Utility / weighting
COMMENTARY

The need for reorientation toward cost-effective prediction: Comments on ‘Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond’ by M. J. Pencina et al., *Statistics in Medicine* (DOI: 10.1002/sim.2929)

Sander Greenland

Departments of Epidemiology and Statistics, University of California, Los Angeles, CA 90095-1772, U.S.A.

Any decision rule entails an implicit loss function, and the loss functions implicit in rules that appear to neglect loss functions are usually clinically absurd. One property of the loss function

The test criterion $\Delta$ involves cost parameters that can be far beyond the scope of statistical expertise, involving matters of valuation and quality of life. It is then natural and may often suffice to focus statistical efforts on maximizing the accuracy of the risk score with and without $X$, to provide an accurate basis for further evaluations. Nonetheless, by including costs as free parameters in a loss function, a statistician can (with the aid of contextual experts) perform a sensitivity analysis over a range of reasonable values, rather than rely on potentially absurd implicit defaults. Occasionally, it may even be deemed worthwhile to statistically estimate costs as well as risks from available data, to provide a complete health-service evaluation.
Loss function in NRI

- Binary NRI = sum (delta(sens) + delta(spec))
- Improvements for events / non-events are weighted equally; implies weighting TP (or FN, for events) and FP (or TN, for non-events) by odds(non-events vs events)
- E.g. incidence 5.6%, relative weight 94.4 vs 5.6 ≈ 17
- But cut-off of 20% implies 80:20 = 4

Inconsistent?

- Choosing a cut-off on the probability scale implies a relative weight of TP vs FP, or harm vs benefit; and vice versa

(Peirce Science 1884; Pauker NEJM 1975;
Decision-analytic

- Decision analytic, e.g. Net Benefit in Decision Curve
  - Net Benefit = (TP – w FP) / N
    - w = threshold probability / (1-threshold probability)
      - e.g.: threshold 50%: w = .5/.5=1; threshold 20%: w=.2/.8=1/4

- Number of true-positive classifications, penalized for false-positive classifications

Vickers & Elkin, MDM 2006
### Reclassification table

<table>
<thead>
<tr>
<th></th>
<th>Model without HDL</th>
<th>Model with HDL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Without event: n=3081</strong></td>
<td>Low: 2952</td>
<td>High: 31</td>
</tr>
<tr>
<td></td>
<td>High: 26</td>
<td>72</td>
</tr>
<tr>
<td><strong>Model without HDL</strong></td>
<td>Low: 145</td>
<td>High: 14</td>
</tr>
<tr>
<td><strong>With event: n=183</strong></td>
<td>Low: 3</td>
<td>High: 21</td>
</tr>
</tbody>
</table>

### Reclassification risk table

<table>
<thead>
<tr>
<th>Reclassification</th>
<th>Events</th>
<th>No event</th>
<th>N</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-low</td>
<td>145</td>
<td>2952</td>
<td>3097</td>
<td>4.7%</td>
</tr>
<tr>
<td>High-Low</td>
<td>3</td>
<td>26</td>
<td>29</td>
<td>10%</td>
</tr>
<tr>
<td>Low-High</td>
<td>14</td>
<td>31</td>
<td>45</td>
<td>31%</td>
</tr>
<tr>
<td>High-high</td>
<td>21</td>
<td>72</td>
<td>93</td>
<td>23%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>183</td>
<td>3081</td>
<td>3264</td>
<td>5.6%</td>
</tr>
</tbody>
</table>

*NRI events = 14-3=11 on 183 events = 6.0%*

*NRI non events = 26 – 31 = –5 on 3081 non events = –0.2%*

*NB = (11 - 0.25*5) / 3264 = 0.003*

*Test tradeoff = 1/0.003 = 335*
Net Reclassification Risk graph

- Area = N * risk
- Preventable events = area L/H – area H/L
- Overtreatment = complementary area L/H – complementary H/L
- Large prognostic utility: large areas and high risk differences
External validation: miscalibration

Predictions twice too high

Predictions twice too low
Sensitivity analyses for cut-off: decision curve
Conclusions

- Evaluation of improvement in predictive ability requires different graphs and measures for continuous predictions than for classifications
  - Overall discriminative ability: delta AUC

- Usefulness for decision making
  - Utility-respective measures such as Net Benefit and test tradeoff
  - Net Reclassification Risk graph for visual assessment and Decision Curve as a sensitivity analysis
Limitations of utility-respecting measures

- Properties of Net Reclassification Risk need more work
  - External validation: model misspecification / miscalibration
  - Extension to survival / competing risk
  - Indicators of uncertainty

- Definition of ‘important gain’ remains arbitrary; vs p<.05 well accepted; formal cost-effectiveness at the end of evaluation pyramid
References

Assessing the performance of prediction models: a framework for traditional and novel measures.

Assessing the incremental value of diagnostic and prognostic markers: a review and illustration.

Performance measures for prediction models and markers: evaluation of predictions and classifications.

Evaluating a new marker for risk prediction using the test tradeoff: an update.