Evaluations of Predictions

Changyu Shen
Richard A. and Susan F. Smith Center for Outcomes Research in Cardiology
Beth Israel Deaconess Medical Center
Harvard Medical School
“Sweetheart, my neural net predicts that you and I are 98.9% compatible. Will you be my Valentine?”
Outline

• Criteria of the performance of a prediction model
  – Discrimination
  – Calibration

• Estimation of the performance measure
Criteria of performance
Discrimination

• How well a prediction rule discriminate those who will have the event from those who will not

• Global accuracy: how accurate is the prediction for the entire population

• Local accuracy: how accurate is the prediction for a sub-population
  – 80% accuracy for the entire population and 70% accuracy for a sub-population

• Global accuracy may not represent the prediction accuracy for a subgroup
Calibration

• Calibration means the estimated probability of event agrees with the empirical proportions
  – If we predict people with biomarker A will have 20% incidence rate of cancer, then it has good calibration if indeed 20% of the subjects with the biomarker in our sample has cancer

• Calibration is a property related to goodness of fit of a model
Binary outcome
Notations

• $Y = 1$ (event) or $0$ (no event): the actual outcome
• $\hat{Y} = 1$ or $0$: the predicted value
• $p(X_0)$: estimated probability of event
• $Pr[A|B]$: the probability of $A$ given that $B$ is true
• Overall event rate: $\rho = Pr[Y = 1]$
Discrimination measure: sensitivity and specificity

• **Sensitivity** (also known as **recall**): the proportion of subjects predicted to have the event among those who actually have the event ($Pr[\hat{Y} = 1|Y = 1]$)

• **Specificity**: the proportion of subjects predicted not to have the event among those who actually do not have the event ($Pr[\hat{Y} = 0|Y = 0]$)
Discrimination measure: PPV and NPV

- **Positive predictive value** (PPV, also known as precision): the proportion of subjects having the event among those who are predicted to have the event ($\Pr[Y = 1|\hat{Y} = 1]$)

- **Negative predictive value** (NPV): the proportion of subjects not having the event among those who are predicted not to have the event ($\Pr[Y = 0|\hat{Y} = 0]$)
<table>
<thead>
<tr>
<th>( \hat{Y} = 1 )</th>
<th>( Y = 1 )</th>
<th>( Y = 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{Y} = 1 ) True positive (TP)</td>
<td>False positive (FP)</td>
<td></td>
</tr>
<tr>
<td>( \hat{Y} = 0 ) False negative (FN)</td>
<td>True negative (TN)</td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity\(=\frac{TP}{TP+FN}\)  
Specificity\(=\frac{TN}{TN+FP}\)  
PPV\(=\frac{TP}{TP+FP}\)  
NPV\(=\frac{TN}{TN+FN}\)
PPV and NPV cannot be estimated in a case-control design!

\[
PPV = \frac{\rho \times \text{Sensitivity}}{\rho \times \text{Sensitivity} + (1 - \rho) \times (1 - \text{specificity})}
\]

\[
NPV = \frac{(1 - \rho) \times \text{specificity}}{(1 - \rho) \times \text{specificity} + \rho \times (1 - \text{sensitivity})}
\]
Connection with hypothesis testing

• Prediction of a binary outcome can be viewed as a hypothesis testing problem:

\[ H_0: Y = 0 \quad \text{versus} \quad H_A: Y = 1 \]

Type I error rate: \( \alpha = \Pr[\text{claim } Y = 1 | Y = 0] \)

Power: \( 1 - \beta = \Pr[\text{claim } Y = 1 | Y = 1] \)

• \( \alpha = 1 - \text{specificity} \); Power = sensitivity
Receiver Operating Characteristic (ROC) curve

• A threshold is applied to a risk score for predictions
• Sensitivity and specificity change in opposite directions as the threshold varies
• The ROC curve describes the relationship between sensitivity and false positive rate (1-specificity) as the threshold changes continuously
Less stringent threshold
Area Under the Curve (AUC)

- AUC takes value between 0 and 1
- 1 means perfect prediction at some threshold
- If a risk score is independent of the event, the AUC is 0.5 (ROC is the diagonal line)
- An alternative interpretation of AUC: the probability that the risk score of a randomly chosen subject with event is higher than the risk score of a randomly chosen subject without the event
FIGURE 2. Probability distributions of a marker, X, in cases (solid curves) and controls (dashed curves) consistent with the logistic model logit \( P(D = 1|X) = \alpha + \beta X \). It has been assumed that X has a mean of 0 and a standard deviation of 0.5 in controls so that a unit increase represents the difference between the 84th and 16th percentiles of X in controls. The marker is normally distributed, with the same variance in cases. The odds ratio (OR) per unit increase in X is shown.

Concordance (c statistic)

- C statistic: the probability that the risk score is higher for a randomly selected subject with event as compared with a randomly selected subject without the event
- AUC=c statistic for binary outcome
- Rank-based measure, relatively insensitive to systematic errors in calibration
Calibration plot
Perfect

Overly optimistic on certainty

Upward bias

Combined issue
Hosmer-Lemeshow goodness-of fit test

• A well-adopted test for logistic regression

• Procedure
  – Divide samples into about approximately 10 groups with similar number of subjects
  – Compute a Chi-square statistic

\[
\sum_{i=1}^{10} \frac{(O_i - E_i)^2}{E_i(N_i - E_i)/N_i}.
\]

\(O_i\): # of observed event in group \(i\)
\(E_i\): # of expected event based on the model in group \(i\)
\(N_i\): # of subjects in group \(i\)
Discrimination versus calibration

Good discrimination, good calibration

Poor discrimination, poor calibration
Discrimination versus calibration

Good discrimination, poor calibration

Poor discrimination, good calibration
Calibration parameters

• Fit a logistic regression model in a validation data where the only covariate is the risk score
• Intercept (calibration-in-the large): should ideally be 0
  – Intercept<0: predicted probabilities are systematically too high
  – Intercept>0: predicted probabilities are systematically too low
• Slope: should ideally be 1
  – Slope<1: predicted probabilities too extreme (too much certainty)
  – Slope>1: predicted probabilities not sufficiently extreme (not enough certainty)
C statistic (95% CI) 0.78 (0.76 to 0.81)
Calibration-in-the-large -2.361 P<0.001
Recalibration slope -0.323 P<0.001

Observed proportion with coronary artery disease

Predicted probability according to Duke clinical score

Coronary artery disease

No coronary artery disease
The Brier score

• Brier score is the average of the squared prediction error

\[ \sum_{i=1}^{n} (Y_i - p_i)^2 / n \]

• Values range from 0 (perfect) to 0.25 (worthless)

• Integrated measure of discrimination and calibration
Net Reclassification Improvement (NRI)

- Suppose there is an existing prediction model $M$ for some event. Model $M$ divides the population into several risk groups (e.g. low, mid, and high) based on predicted event probability.

- How do we assess the improvement in discrimination by adding one more risk factor?
Net Reclassification Improvement (NRI)

- $\text{NRI} = [Pr(up|Y = 1) - Pr(down|Y = 1)] - [Pr(up|Y = 0) - Pr(down|Y = 0)]$

  - $Pr(up|Y = 1) - Pr(down|Y = 1)$: increase in sensitivity
  - $Pr(up|Y = 0) - Pr(down|Y = 0)$: decrease in specificity
Integrated Discrimination Improvement (IDI)

• A “continuous” version of NRI
  – Instead of “up” and “down”, focusing on the actual change in estimated probability

• Let
  
  \[ S(\text{new}) = \text{Avg}[p(\text{new})|Y = 1] \]
  
  \[ T(\text{new}) = \text{Avg}[p(\text{new})|Y = 0] \]
  
  \[ S(\text{old}) = \text{Avg}[p(\text{old})|Y = 1] \]
  
  \[ T(\text{old}) = \text{Avg}[p(\text{old})|Y = 0] \]

• \( IDI = \{S(\text{new})-S(\text{old})\}-\{T(\text{new})-T(\text{old})\} \)
  
  \[ = \{S(\text{new})-T(\text{new})\}-\{S(\text{old})-T(\text{old})\} \]
  
  \[ = \text{Discrimination slope (new)} - \text{Discrimination slope (old)} \]
An Example

- Men with metastatic non-seminomatous testicular cancer can often be cured by cisplatin based chemotherapy
- After chemotherapy, a decision needs to be made on whether or not to receive surgical resection to remove residual tumor
- Objective: compute risk of presence of residual tumor
Two models

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Without LDH</th>
<th>With LDH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary tumor teratoma-positive?</td>
<td>2.7 [1.8 – 4.0]</td>
<td>2.5 [1.6 – 3.8]</td>
</tr>
<tr>
<td>Prechemotherapy AFP elevated?</td>
<td>2.4 [1.5 – 3.7]</td>
<td>2.5 [1.6 – 3.9]</td>
</tr>
<tr>
<td>Prechemotherapy HCG elevated?</td>
<td>1.7 [1.1 – 2.7]</td>
<td>2.2 [1.4 – 3.4]</td>
</tr>
<tr>
<td>Square root of postchemotherapy mass size (mm)</td>
<td>1.08 [0.95 – 1.23]</td>
<td>1.34 [1.14 – 1.57]</td>
</tr>
<tr>
<td>Reduction in mass size per 10%</td>
<td>0.77 [0.70 – 0.85]</td>
<td>0.85 [0.77 – 0.95]</td>
</tr>
<tr>
<td>Prechemotherapy LDH (log(LDH/upper limit of local normal value))</td>
<td>-</td>
<td>0.37 [0.25 – 0.56]</td>
</tr>
</tbody>
</table>

*Epidemiology, 2010; 21: 128-138*
Epidemiology, 2010; 21: 128-138
Epidemiology, 2010; 21: 128-138
<table>
<thead>
<tr>
<th>Performance measure</th>
<th>Development Without LDH</th>
<th>Development With LDH</th>
<th>External validation Without LDH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brier</td>
<td>0.174</td>
<td>0.163</td>
<td>0.161</td>
</tr>
<tr>
<td>Brier_scaled</td>
<td>29.8%</td>
<td>34.0%</td>
<td>20.0%</td>
</tr>
<tr>
<td>R² (Nagelkerke)</td>
<td>38.9%</td>
<td>43.1%</td>
<td>25.0%</td>
</tr>
<tr>
<td>Discrimination</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C stat</td>
<td>0.818 [0.78 – 0.85]</td>
<td>0.839 [0.81 – 0.87]</td>
<td>0.785 [0.73 – 0.84]</td>
</tr>
<tr>
<td>Discrimination slope</td>
<td>0.301</td>
<td>0.340</td>
<td>0.237</td>
</tr>
<tr>
<td>Calibration</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calibration-in-the-large</td>
<td>0</td>
<td>0</td>
<td>-0.03</td>
</tr>
<tr>
<td>Calibration slope</td>
<td>1</td>
<td>1</td>
<td>0.74</td>
</tr>
<tr>
<td>H-L test</td>
<td>Chi-square 6.2, p=0.63</td>
<td>Chi-square 12.0, p=0.15</td>
<td>Chi-square 15.9, p=0.07</td>
</tr>
<tr>
<td>Clinical usefulness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Net Benefit at threshold 20%*</td>
<td>0.2%</td>
<td>1.2%</td>
<td>0.1%</td>
</tr>
</tbody>
</table>
Time-to-event outcome
Dichotomization

• Define a binary outcome by applying a threshold to the time-to-event variable
  – Die within 5 years
  – Incidence of MI within 2 years
• Can apply all concepts developed for binary outcome
Time-dependent ROC

• There will be multiple ROC curves depending on the threshold
  – ROC for the outcome of 2-year mortality
  – ROC for the outcome of 5-year mortality
  – ...

• Estimation will need to account for right-censoring
Concordance measure

- C statistic: the probability that the subject with longer survival time from a randomly selected pair of subjects has higher risk score
  \[ c = \Pr[r(X_1) > r(X_2) | S_1 > S_2] \]
- When the outcome is subject to right censoring, estimation of \( c \) requires some model assumptions
- Alternative definition \( c^* = \Pr[r(X_1) > r(X_2) | S_1 > S_2 \text{ can be ascertained}] \) (depends on the censoring process)
Estimation of performance measures
Background

• The performance measure is defined on a population
  – Sensitivity is defined as the proportion of claimed event among ALL those with the event

• Performance measures is an unknown parameter

• Apparent estimators of the performance measure based on the same data used to develop the prediction model are overly optimistic
Two strategies

• External validation: Evaluation of the performance using new participant level data, external to the data used to develop the model

• Internal validation: Evaluation of the performance using the same data, in a “smart” way
Internal validation

• Split-sample: allocate x% of the data (training sample) to develop a prediction model and use the rest (100-x)% of the data (test sample) to estimate the performance
  – 50% training, 50% test
  – 67% training, 33% test
Internal validation

- Cross-validation: consecutively change the training and testing sample
  - 10 fold cross-validation:
    - Divide data into 10 equal size subsets
    - Use subset 2-10 to build prediction model and evaluate its performance on subset 1
    - Use subset 1, 3-10 to build prediction model and evaluate its performance on subset 2
    - ...
    - Average the performance measure on the 10 subsets
Bootstrap

• Compute the apparent performance measure (W)
• Generate m datasets, each of which is composed of n records that are drawn independently from the original data with replacement
• Construct prediction model on each bootstrap data set and obtain the apparent performance measure on the same bootstrap data
• Compute the average of (over the m bootstrap samples) the performance measure (V)
• The bias (or optimism) of the initial apparent performance measure can be estimated as V-W
• The final performance measure is estimated as W-(V-W)=2W-V
Original sample
(AUC=0.81)

Bootstrap samples

BS 1
AUC₁=0.85

BS 2
AUC₂=0.80

BS 3
AUC₃=0.78

BS m
AUCₘ=0.83

Average bootstrap AUC=0.83

Bias (optimism)=0.83-0.81=0.02
Bias corrected AUC=0.81-0.02=0.79
Summary

• Performance measure
  – Discrimination (sensitivity, specificity, PPV, NPV, ROC, AUC (c statistic), Brier score)
  – Incremental discrimination (NRI, IDI)
  – Calibration

• Estimation of performance measure
  – Apparent estimates are usually overly optimistic
  – For internal validation, bootstrap performs the best
Survey

https://www.surveymonkey.com/r/7Q3JBRY
Thank you