Vanderbilt Center for Quantitative Sciences

Risk Assessment and Evaluation of Predictions

Zhiguo (Alex) Zhao

Division of Cancer Biostatistics
Department of Biostatistics
Vanderbilt Center for Quantitative Sciences

February 20, 2012
Outline

1. Conference on Risk Assessment and Evaluation of Predictions

2. Absolute Risk Prediction
   - Definition and examples of absolute risk
   - Some statistical aspects
   - Estimating absolute risk

   - Evaluating a risk model — basic concepts
   - Comparing risk models — including the role of risk reclassification metrics
   - Estimation and inference from data
About the Conference

- Oct. 12-14, 2011, Silver Spring, MD
- Co-sponsored by NCI and U. Maryland
- Biostatistics and Risk Assessment Center of the Department of Epidemiology and Biostatistics at the University of Maryland
- Major conference theme topics:
  - Applications of Risk Models in Cancer Studies
  - Models for Early Detection of Cancer
  - Assessing the Accuracy of Risk Models
  - Competing Risks Models
  - Evaluation of Prediction Models
  - Individualized Disease Risk Prediction
  - Reliability Methods and Applications
  - Methodology and Advances in Risk Assessment
  - Risk Reclassification Methods
  - ...
- Speakers and audiences
Without explicit citing, the materials of this presentation are from the following two documents:

Absolute Risk Prediction

Definition and examples of absolute risk

Definitions

- Relative risk

- Absolute risk (crude risk, cumulative incidence)
  - $\text{Prob}(c_1 \text{ occurs in } [t, t + \delta) | \text{ at risk at } t \text{ with risk factors } X \text{ in presence of competing risks, } c_2)$

- Pure risk
  - $\text{Prob}(c_1 \text{ occurs in } [t, t + \delta) | \text{ at risk at } t \text{ with risk factors } X \text{ and there are NO competing risks})$
Absolute risk vs. Pure risk

- Absolute risk: no competing risk assumptions like “independence”
- Absolute risk clinically relevant, because eliminating other deaths is not realistic
- Absolute risk nearly equals pure risk if death from competing causes is rare (e.g. short intervals)
- Pure risk has etiologic interest as a description related to cumulative cause specific hazard

\[ 1 - \exp \left\{ - \int_0^t h_1(u) \, du \right\} \]
Examples of absolute risk and pure risk

### Absolute ("Crude") and "Pure" Risk in 1000 60-Year Old Women

<table>
<thead>
<tr>
<th>Age at Start of Interval</th>
<th># At Risk</th>
<th># Incident Breast Cancer</th>
<th># Deaths from Other Causes</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>1000</td>
<td>17</td>
<td>44</td>
</tr>
<tr>
<td>65</td>
<td>939</td>
<td>20</td>
<td>63</td>
</tr>
<tr>
<td>70</td>
<td>856</td>
<td>22</td>
<td>89</td>
</tr>
<tr>
<td>75</td>
<td>745</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Absolute risk of breast cancer to age 75**

\[
\frac{(17+20+22)}{1000} = 5.9\%
\]

**“Pure” risk**

\[
1 - \left(1 - \frac{17}{1000}\right) \left(1 - \frac{20}{939}\right) \left(1 - \frac{22}{856}\right) = 6.3\%
\]
Example of absolute risk and pure risk

Framingham Model for Coronary Heart Disease (CHD)

- **Data**: Framingham cohort of 2489 men and 2856 women aged 30-74 in 1971-74
- **Time scale**: time since baseline exam
- **Modeling Pure risk of CHD**
  \[ = 1 - S_0(t)^{\exp{\beta^T(X - \text{mean } X)}} \]

- **X** includes age, age^2, TC, LDL, HDL, BP, Diabetes, Smoking

*Wilson et al, Circulation 1998;97:1837-47*
Examples of absolute risk and pure risk

- Cancer absolute risk
  - Bladder, Breast, Colon, Lung, Melanoma, Ovary, Pancreas

- Absolute risk of death from prostate cancer following diagnosis (e.g. Albertson, Hanley, Fine, JAMA 2005)
Absolute risk calculation

For a woman with breast cancer risk factor $X$

- $\int_{a}^{a+\delta} h_1(t) rr(t; X) \exp \left[ - \int_{a}^{t} [h_1(u)rr(u, X) + h_2(u)] du \right] dt$

- $h_1(t)$ is baseline hazard of breast cancer incidence
- $h_2(t)$ is mortality hazard from competing risks
- $rr(t; X) = \exp(\beta^T X(t))$ is relative risk of breast cancer for covariates $X(t)$
Choice of time scale

- Time since baseline evaluation
  - Require careful modeling of age as covariate for predicting incidence
  - May be more powerful than age for predicting (e.g. death from cancer following diagnosis)
  - Standard survival methods for right censoring can be used
- Age
  - Important predictor of incidence
  - Need to account for left-truncation and right censoring
    - Likelihood: \[
    \frac{S_1(t; \theta)h_1(t; \theta)^{\delta_1}}{S_1(\text{age at entry}; \theta)}
    \]
    - For non- and semi-parametric analyses, a person is at risk only for \( t \geq \text{her age at entry} \).
Choice of time scale

Left Truncation and Right Censoring on Age Scale
(e.g. age at breast cancer onset in BCDDP study)
Designs that can yield absolute risk

- **Cohort**
  - Prospective selected study population
  - Population-based (e.g., NHANES)
  - Retrospective

- **Sub-samples of a cohort or population base**
  - Sub-sampling of a cohort (nested case-control or case-cohort)
  - Population-based case-control combined with registry data (e.g., SEER)

- **Family-based**
  - Well defined ascertainment with a retrospective family cohort (e.g., Kin-cohort design, case-control family study)
  - Segregation analysis of pedigrees (absolute or relative risks)
  - Relative risks from family-based case-control study plus population rates
Estimating absolute risk from different design

- From cohort data
  - No covariates (Gaynor et al, JASA 1993)
  - Cumulative incidence regression
    - Fine-Gray model (Fine & Gary, JASA 1999, Fine, Biostatistics, 2001; implemented in cmprsk package in R)
    - Time-varying covariate (Scheike, Zhang & Gerds, 2008)
  - Modeling via cause-specific hazards
    - Cox proportional hazards model; (Benichou & Gail, 1990, gave variance of absolute risk estimate)
    - Additive hazard model (Aalen 1989, 1993 proposed the model; Lin & Ying 1994 gave explicit form for $\hat{\beta}_i$; Chen & Shen, 1999 gave estimation, prediction and simultaneous CI for $r_1(t, X)$ under the additive model)
Estimating absolute risk from different design

- From sub-samples of cohort
  - Nested case-control design (Langholz & Borgan 1997): at each time a case develops, sample individuals from risk set.
  - Case-cohort design (Prentice & Self, 1988, Langholz & Jiao 2007): analyze data from subcohort selected at start of follow-up and all cases observed during follow-up.

- From combining Relative Risk estimates with Registry Data
  - Estimate relative risk and attributable risk from cohort, Nested case-control, case-cohort, case-control
  - Obtain composite age-specific hazard and competing hazard
  - Estimate absolute risk
References for absolute risk prediction

References

- Cheng, SC; Fine, JP; Wei, LJ Prediction of cumulative incidence function under the proportional hazards model. BIOMETRICS 54 (1) 219-228 1998.
- Shen, Y; Cheng, SC Confidence bands for cumulative incidence curves under the additive risk model. BIOMETRICS 55 (4), 1093-1100 1999.
- LIN DY and Ying SEMIPARANMETRIC ANALYSIS OF THE ADDITIVE RISK MODEL. BIOMETRIKA 81 : 61 1994
- Scheike, TH; Zhang, MJ; Gerds, TA Predicting cumulative incidence probability by direct binomial regression. BIOMETRIKA 95 (1) 205-220, 2008.
The point of developing a risk model

- To help make medical decisions
- Offer new interventions particularly to those who might benefit (cases)
- Offer no new interventions but more peace of mind to those who will not benefit from intervention (controls)
- Opposite scenario is analogous: where reduction from standard intervention is the goal
- \( r(\mathbf{X}) = \text{Prob}(D=1 \mid \mathbf{X} = \mathbf{x}) \) is a frequency of events among the group of subjects with \( \mathbf{X} = \mathbf{x} \)
- Individual level risks are not well defined, not observable
Calibration — crucial!

- Is the risk calculator valid?
- Among people with \( r(X) = r \), is the fraction of events \( \approx r \) ?
- We are asking if \( P(D=1 \mid r(X) = r) = r \), instead of \( r(X) = P(D=1 \mid X) \).
- Validity of the risk calculator is crucial otherwise, we are engaged in evaluating a score for discriminatio/classification, not with the higher level task of evaluating risk prediction performance.
- Visual assessment with the predictiveness curve (Pepe 2011) or calibration plot (Steyerberg et. al, 2010)
- Assume henceforth that risk calculators are valid.
Net benefit (when have risk categories)

\[ NB(t) = B \times P(D = 1)HR_C(t) - Cost \times P(D = 0)HR_N(t) \]

- Ideally, \( HR_C(t) = 1 \) and \( HR_N(t) = 0 \)
- \( B \): expected benefit of treatment to a case
- \( Cost \): expected cost of treatment to a control
- \( Cost/B = t/(1-t) \)
- maximum value of \( NB = P(D = 1) \); \( B \) is the unit of measurement
- Define \( \rho \equiv P(D = 1) \)
- Relative utility (RU) = \( NB(t)/\rho = \% \) of maximum benefit; true positive rate discounted appropriately for the false positive rate
Plots (when have no risk categories or thresholds)

- Predictiveness curve
- Integrated plot (Pepe 2011)
- Decision curve
- Relative utility curve
Summary Measures (when have no risk categories or thresholds)

- MRD: Mean risk difference; \( \text{mean}(\text{risk}(X) | \text{case}) - \text{mean}(\text{risk}(X) | \text{control}) \)
- AARD: Above average risk difference; 
  \( P(\text{risk}(X) > \rho | \text{case}) - P(\text{risk}(X) > \rho | \text{control}) \)
- ROC type statistics
MRD

Also known as:
- IDI: integrated discrimination improvement relative to no model (Pencina 2007)
- PEV: Proportion of explained variation
- $R^2 =$ PEV, there are other more complex $R^2$ measures (Gail 2005)
- Yate’ Slope
Above Average Risk Difference (AARD)

\[
AARD = P(risk(X) > \rho | D = 1) - P(risk(X) > \rho | D = 0) = 0.797 - 0.198 = 0.599
\]

Also known as

- \[ HR_C(\rho) - HR_N(\rho) = TPR(\rho) - FPR(\rho) = \text{Youden's index} (\rho) \]

- \[ RU(\rho) = NB(\rho) / \rho \]

Proof: \[ NB(t) = \rho HR_C(t) - (1 - \rho) \frac{t}{1-t} HR_N(t) \]

- Standardized Total Gain \( TG/2\rho(1-\rho)\). Not intuitive result!\(^{20}\)

- \(0.5\times\) continuous NRI\(^{21}\) comparing \(risk(X)\) with no model

- \(0.5\times\) categorized NRI comparing \(risk(X)\) with no model using risk categories \(> \rho\) and \(< \rho\).
ROC Type Statistics as Summary Measures

- May be useful when no clinically relevant risk thresholds exist.
- $\text{ROC}(f_0) = P(\text{risk}(X) > t | D = 1)$ where $t$: $f_0 = P(\text{risk}(X) > t | D = 0)$
- $\text{ROC}^{-1}(t_0) = P(\text{risk}(X) > t | D = 0)$ where $t$: $t_0 = P(\text{risk}(X) > t | D = 1)$
- $\mathcal{L}(v_0) = P(\text{risk}(X) > t | D = 1)$ where $t$: $v_0 = P(\text{risk}(X) > t)$
- $\mathcal{L}^{-1}(w_0) = P(\text{risk}(X) > t)$ where $t$: $w_0 = P(\text{risk}(X) > t | D = 1)$
  $\quad = \rho w_0 + (1 - \rho) \text{ROC}^{-1}(w_0)$
- $\mathcal{L}$ is the Lorenz curve.\textsuperscript{22}
- Report the risk threshold corresponding to the criterion as well.
Some performance measures (Steyerberg et. al, 2010)

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Measure</th>
<th>Visualization</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall performance</td>
<td>$R^2$, Brier</td>
<td>Validation graph</td>
<td>Better with lower distance between $Y$ and $\hat{Y}$. Captures calibration and discrimination aspects</td>
</tr>
<tr>
<td>Discrimination</td>
<td>$c$ statistic</td>
<td>ROC curve</td>
<td>Rank order statistic; interpretation for a pair of subjects with and without the outcome</td>
</tr>
<tr>
<td>Discrimination slope</td>
<td></td>
<td>Box plot</td>
<td>Difference in mean of predictions between outcomes; easy visualization</td>
</tr>
<tr>
<td>Calibration</td>
<td>Calibration-in-the-large</td>
<td>Calibration or validation graph</td>
<td>Compare mean ($y$) versus mean ($\hat{y}$); essential aspect for external validation</td>
</tr>
<tr>
<td>Calibration slope</td>
<td></td>
<td></td>
<td>Regression slope of linear predictor; essential aspect for internal and external validation; related to “shrinkage” of regression coefficients</td>
</tr>
<tr>
<td>Hosmer-Lemeshow test</td>
<td></td>
<td></td>
<td>Compares observed to predicted by decile of predicted probability</td>
</tr>
<tr>
<td>Reclassification</td>
<td>Reclassification table</td>
<td>Cross-table or scatter plot</td>
<td>Compare classifications from 2 models (one with, one without a marker) for changes</td>
</tr>
<tr>
<td>Reclassification statistic</td>
<td></td>
<td></td>
<td>Compare observed outcomes to predicted risks within cross-classified categories</td>
</tr>
<tr>
<td>Net reclassification index (NRI)</td>
<td></td>
<td></td>
<td>Compare classifications from 2 models for changes by outcome for a net calculation of changes in the right direction</td>
</tr>
<tr>
<td>Integrated discrimination index (IDI)</td>
<td>Box plots for 2 models (one with, one without a marker)</td>
<td></td>
<td>Integrates the NRI over all possible cut-offs; equivalent to difference in discrimination slopes</td>
</tr>
<tr>
<td>Clinical usefulness</td>
<td>Net benefit (NB)</td>
<td>Cross-table</td>
<td>Net number of true positives gained by using a model compared to no model at a single threshold (NB) or over a range of thresholds (DCA)</td>
</tr>
<tr>
<td>Decision curve analysis (DCA)</td>
<td>Decision curve</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Favorite summary measures**

### My Favorite Summary Measures

$t = 20\%, \rho = .1017$

<table>
<thead>
<tr>
<th></th>
<th>risk($X$)</th>
<th>risk($X, Y$)</th>
<th>$\Delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cases $&gt; t$</strong></td>
<td>$HR_C(t)$</td>
<td>65.2%</td>
<td>73.5%</td>
</tr>
<tr>
<td><strong>Controls $&gt; t$</strong></td>
<td>$HR_N(t)$</td>
<td>8.9%</td>
<td>8.4%</td>
</tr>
<tr>
<td><strong>% of max benefit</strong></td>
<td>$RU(t)$</td>
<td>45.5%</td>
<td>55.0%</td>
</tr>
</tbody>
</table>
Cook-Ridker analysis strategy (Cook & Ridker, 2009)

Cook-Ridker Analysis Strategy\textsuperscript{24}

- Is problematic.
  (i) \% reclassification
    * Not a key summary in general (unless there is very little reclassification).
  (ii) \% ‘correct’ reclassification
    * Defined as \% events in a cell closer to $r(X, Y)$ category label than $r(X)$ label.
    * This will be 100\% in large samples if $Y$ is a risk factor.
  (iii) Reclassification calibration statistical tests
    * Not traditional use of the term calibration
    * Compare event rate in off diagonal cell with average ($r(X)$) and average ($r(X, Y)$).
    * $H_0$ regarding $r(X)$ model always rejected in large samples if $Y$ is a risk factor. RCC statistic is equal to Pearson chi-squared test when $X = \phi$.
    * $H_0$ regarding $r(X, Y)$ model rejected at nominal level in large samples. RCC statistic is equal to Hosmer-Lemeshow test when $X = \phi$. 
Comments on comparing risk models

To compare two risk models

- Choose your favorite measure(s) of prediction performance, report for each model and compare

- risk reclassification analyses that do not focus on marginal performance of each model (NRI, reclassification calibration, % reclassification) can be misleading.

Stratifying on risk categories from a baseline model

- Evaluating $r(X, Y)$ within baseline risk strata can be helpful to identify subpopulations where information on $Y$ may be most useful.

- Risk reclassification tables have a role for this purpose.
Null hypothesis about improvement in prediction performance

To evaluate the incremental value of $Y$ for prediction over use of $X$ alone.

$$H_0^1 : \text{risk}(X, Y) = \text{risk}(X)$$ (1)
Null hypothesis about improvement in prediction performance

To test if discrimination provided by $\text{risk}(X,Y)$ is better than provided by $\text{risk}(X)$.

$$H_0^2 : \text{ROC}(X,Y)(\cdot) = \text{ROC}_X(\cdot)$$
Null hypothesis about improvement in prediction performance

In ROC analysis, the AUC is typically used as the basis of a test statistic.

\[ H_0^3 : AUC_{(X,Y)} = AUC_X \]  (3)
Null hypothesis about improvement in prediction performance

In the ROC framework, another approach is to assess if, condition on $X$, the ROC curve for $Y$ is equal to the null ROC curve (Janes & Pepe, 2009). This is particularly relevant when controls are matched by design to cases on $X$ (Janes & Pepe, 2008)

$$H_0^4 : \text{ROC}(Y|X)(f) = f, f \in (0, 1) \forall X.$$  \hspace{1cm} (4)
Null hypothesis about improvement in prediction performance

The predictiveness curve, $R(\cdot)$, displays the quantiles of the risk distribution (Hunag et al., 2007). Risk stratification tables (Cook, 2007) are essentially discretized versions of the risk distribution. The NRI is a summary of the classified risks for subjects with and without the outcome event.

$$H_0^5 : R_{(X,Y)}(\cdot) = R_X(\cdot)$$  \hspace{1cm} (5)
Null hypothesis about improvement in prediction performance

The IDI (Pencina et al. 2008) can be interpreted as the difference in risk variances (Pepe et al. 2008).

\[ H^6_0 : IDI = 0 \]
Null hypothesis about improvement in prediction performance

All 6 null hypotheses are equivalent.

\[ H_0^1 \iff H_0^2 \iff H_0^3 \iff H_0^4 \iff H_0^5 \iff H_0^6 \]  \hspace{1cm} (7)
A single test on the regression coefficient for Y in the risk model, given the model has approximately correct forms.
MORE ... MUCH more
Thanks!