Risk Model Assessment

- Risk Model Calibration
  - required for a risk model to be valid
  - crucial whenever a model will be used to convey information to a patient

- Risk Model Performance
  - required for a risk model to be useful
  - performance assessment depends on what the model will be used for
Calibration

• A risk is a number of some import
  – “based on my test results, the chance (risk) I have the disease is 5%”
  – “based on my age and family history, my chance of getting breast cancer in the next 5 years is 1%”
• In order to be valid, risks must be calibrated

Assessing Mean Calibration

• Compare event rate with average predicted risk
  – If D=1 for 3% of the population, then the risk model has mean calibration if the average predicted risk is ≈3%

Assessing Weak Calibration

• Logistic calibration
• Predicted risks are obtained from a previously developed model for D (e.g., based on logistic regression); the linear combination of predictors defines the “linear predictor” \( L = b_0 + b_1 X_1 + \ldots + b_k X_k \)
• Regress D on L: \( \logit(D) = a + b L \)
• If \( a \approx 0 \) and \( b \approx 1 \), the model has weak calibration
• \( a \) is the “calibration intercept”; \( b \) is the “calibration slope”
• frequently, in data not used to fit the model the calibration slope \( b < 1 \): large predicted risks are too high and low predicted risks are too low

What does it mean for a risk model to be calibrated? A Hierarchy:

<table>
<thead>
<tr>
<th>Level</th>
<th>Definition</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Observed event rate equals average predicted risk</td>
<td>“calibration-in-the-large”</td>
</tr>
<tr>
<td>Weak</td>
<td>No systematic overestimation or underestimation of risks</td>
<td>“logistic calibration”</td>
</tr>
<tr>
<td>Moderate</td>
<td>Predicted risks correspond to observed event rates</td>
<td>Often the best we can assess with limited data</td>
</tr>
<tr>
<td>Strong</td>
<td>For every combination of risk factors, predicted risks correspond to observed event rates</td>
<td>The ideal; difficult to assess</td>
</tr>
</tbody>
</table>

Adapted from Van Calster et al, *J Clinical Epidemiology*, 2016
Assessing Moderate Calibration

• require $P(D = 1 | \text{risk}(X_1, X_2) = r) = r$
  – here, there are two predictors $X_1$ and $X_2$
• “collapses” data among groups of people with the same predicted risk
• Common practice: divide available data into deciles based on predicted risks
• Compare event rate in a decile of individuals with similar predicted risk → calibration curve
  – Next slide: 1 risk model that has good (moderate) calibration; and 3 poorly calibrated risk models

Forecasts of rain: are the risks well calibrated?


how NOT to assess moderate calibration

• Hosmer-Lemeshow test statistic
• pvalue from Hosmer-Lemeshow test

• In small datasets, badly miscalibrated models may not give a large H-L test statistic or a small pvalue
• In large datasets, small/unimportant deviations from good calibration can still lead to large H-L test statistic or small pvalue
Assessing Strong Calibration

- Must consider every unique combination of predictors and ask whether observed and predicted risks agree for people with that combination
- risk(\(X_1, X_2\)) compared to \(P(D=1|X_1, X_2)\)
- Typically only feasible to assess when there are a limited number of predictors and they are all categorical

Calibration is not enough

- The goal for a risk model requires more than good calibration: to stratify people into “low risk” and “high risk” groups.
- If the prevalence of a condition is \(\rho\), a (moderately) calibrated risk model assigns everyone risk \(\rho\). If we only cared about calibration, we would not need to identify risk factors and develop a risk prediction model.

Risk Model Performance

We will discuss three classes of assessment
- Generic measures
  - “purely mathematical”
    - meaning: they do not directly translate to any clinical, public health, or public policy impact of using the risk model
- Assessing performance when model will be used to recommend treatment/intervention for high risk individuals
- Assessing performance for prognostic enrichment of clinical trials
The Distribution of Risk

- Case (D=1) and control (D=0) risk distributions are fundamental components of all performance measures.
- When examining risk distributions, it can be useful to include any conventional thresholds for deciding who is “high risk.”
- The logit scale may be more convenient than the 0 to 1 risk scale.
- Next slide: data from DABS website.
MRD, AARD, AUC

- MRD = Mean Risk Difference $\equiv$ mean(risk(X)|case)-mean(risk(X)|control)
- AARD = Above Average Risk Difference $\equiv$ P(risk(X) > ρ | case) - P(risk(X) > ρ | control)
- AUC = Area Under the ROC Curve = P( risk_case(X) > risk_cntl(X) )

These measures are sometimes called measures of discrimination
- How well does the risk model discriminate between (separate) cases and controls?

Above Average Risk Difference (AARD)

AARD = P(risk(X) > ρ | case) - P(risk(X) > ρ | control)

= 0.797 - 0.198 = 0.599

Also known as
- $HR_D(\rho) - HR_D(\rho)=TPR(\rho) - FPR(\rho)$ = Youden's index(ρ)
- RU(ρ)=NB(ρ)/ρ (will come to RU and NB shortly)
  - Proof: NB(r)=ρ $HR_D(r)\cdot(1-\rho)$ $\frac{1}{1-\rho}$ $HR_D(r)$; set r=ρ and divide by ρ
- Half of category-free NRI comparing risk(X) with no model
- Half of 2-category NRI comparing risk(X) with no model, using risk threshold ρ (will discuss NRI statistics in section 3)

Mean Risk Difference (MRD)

Also known as
- PEV = Proportion of Explained Variation = $R^2 = \frac{\text{var}(E(D|X))}{\text{var}(D)} = \frac{\text{var}(\text{risk}(X))}{\text{var}(D)}$
- Yates’ slope
- Change in MRD for two nested models also known as IDI=Integrated Discrimination Improvement Index

For our data example, mean(risk|case)=0.391, mean(risk|cntl)=0.069; MRD=0.322

AUC for a Risk Model

- AUC not a clinically relevant measure of predictive performance
  - Arguably roughly similar to MRD in terms of clinical relevance
- Ignores the meaning of risk
- Preferable (more clinically relevant) to average TPR over a relevant range of FPR
  - pAUC
A disadvantage of ROC curves for risk models is that the curve does not show the risk threshold corresponding to each (FPR, TPR).

The next slide shows an alternative to the ROC curve that overcomes this disadvantage.

As an alternative to ROC, plot TPR and FPR versus risk threshold.

EVALUATING A RISK MODEL FOR RECOMMENDING TREATMENT
Use Risk to Decide Treatment

- Sometimes the intended use of a risk model is to determine who should be treated
  - e.g., screen high risk individuals for cancer
  - e.g., only treat cancer patients with high risk of relapse with adjuvant chemotherapy
  - e.g., only treat individuals at high risk of a heart attack with statins

- What risk threshold should define “high risk”?

Benefits and Costs of Treatment

- We will assume there is some expected benefit $B$ to treating a case
  - life extended, morbidity reduced
- We will assume there is some cost $C$ to treating a control
  - does NOT just mean monetary cost
  - side effects of treatment, stress/anxiety, toxic exposures

Choice of Risk Threshold

Classical Decision Theory Result

Suppose the default is no treatment; alternative is single treatment that offers benefit $B$ to a case and cost $C$ to a control. Then the optimal high risk threshold is

$$r_H = \frac{C}{C + B} \leftrightarrow \frac{C}{B} = \frac{r_H}{1 - r_H}$$

Choice of Risk Threshold

Classical Decision Theory Result: Outline of Proof

$$r_H = \frac{C}{C + B} \leftrightarrow \frac{C}{B} = \frac{r_H}{1 - r_H}$$

When should patients choose treatment?

- When expected result of treatment > 0
- $E(benefit|D=1,X)P(D=1|X) - E(cost|D=0,X)P(D=0|X) > 0$
- $B \cdot P(D=1|X) - C \cdot P(D=0|X) > 0$
- $B \cdot P(D=1|X) > C \cdot P(D=0|X)$
- $\frac{P(D=1|X)}{1-P(D=1|X)} > \frac{C}{B}$
Choice of Risk Threshold

Specifying a Cost-Benefit ratio C/B implies a rational choice of risk threshold.
Equivalently, a risk threshold is rational when it corresponds to the Cost/Benefit ratio.

Choice of Risk Threshold: Example 1

20% risk threshold for treatment is equivalent to
\[
\frac{C}{C + B} = 0.2
\]
\[
\frac{C}{B} = \frac{0.2}{1 - 0.2} = \frac{0.2}{0.8} = 0.25
\]
The cost of treating a control equals 1/4\textsuperscript{th} the benefit of treating a case.

Choice of Risk Threshold: Example 2

Gail (JNCI, 2009) evaluated risk models for breast cancer in terms of decision about prophylactic tamoxifen use in 50-59 year old white women. Tamoxifen can reduce the risk of breast cancer but increases the risk of other serious diseases. Under some strong assumptions, he estimated
\[C/B = 0.0077 \rightarrow r_H = 0.0076 \text{ per year}\]

Choice of Risk Threshold: Other Methods

Choose threshold \( r \) satisfying some performance criterion
- Find \( r \) such that \( t_0 \) proportion of cases are detected and treated; \( t_0 = P(\text{risk}(X) > r | D=1) \)
- Find \( r \) such that only \( f_0 \) proportion of controls are worked up or treated; \( f_0 = P(\text{risk}(X) > r | D=0) \)
- Find \( r \) such that \( v_0 \) proportion of the population is worked up or treated; \( v_0 = P(\text{risk}(X) > r) \)
These approaches might be used when budget or resource constraints drive the choice of risk threshold.
Proportion of Cases and Controls High Risk

“High Risk” designation is based on risk, not based on marker.

\[
HR_D(r_H) = P(\text{risk}(X) > r_H | D=1)
\]

= % cases in High Risk category

= TPR or sensitivity

\[
HR_D(r_H) = P(\text{risk}(X) > r_H | D=0)
\]

= % controls in High Risk category

= FPR or 1-specificity

Ideally, \( HR_D(r_H) = 1 \) and \( HR_D(r_H) = 0 \).

Net Benefit and the Risk Model

Overall population impact of the risk model – combines \( HR_D(r_H) \) and \( HR_D(r_H) \):

\[
NB(r_H) = B \left( P(D=1) \; HR_D(r_H) - C \; P(D=0) \; HR_D(r_H) \right)
\]

\[
= B \left( P(D=1) \; HR_D(r_H) - \frac{r_H}{1-r_H} \; P(D=0) \; HR_D(r_H) \right)
\]

In the last expression, Net Benefit is interpreted “in units of B”

B = expected benefit of treatment for a case

C = expected cost of treatment for a control

Example (DABS data)

- D is CVD over 10 years
  - \( P(D=1) = 10.17\% \)
  - Marker X

- Suppose \( r_H = 20\% \):
  - \( HR_D(r_H) = 65.2\% \)
  - \( HR_D(r_H) = 8.9\% \)
  - \( NB(r_H) = 0.046 \cdot \text{benefit of statins to subject who would have a CVD event without them} \)
Standardized Net Benefit (Relative Utility)

\[ NB(r_h) = P(D=1) \cdot HR_D(r_h) - \frac{r_h}{1-r_h} \cdot P(D=0) \cdot HR_D(r_h) \]

Maximum value of \( NB \) is \( P(D=1) = \rho \)

- The best we can do is treat all cases and no controls

\[ NB(r_h) = HR_D(r_h) - \frac{r_h}{1-r_h} \cdot HR_D(r_h) \]

= TPR discounted by an appropriate amount of the FPR

Interpretation: Relative utility is \( z\% \rightarrow \) risk model achieves the same standardized net benefit that we would achieve by detecting \( z\% \) of cases and no controls

Assessing Net Benefit Graphically

- Decision Curves
- Relative Utility Curves

Example, continued

- Relative Utility = 0.046/0.1017 = 0.455 = 45.5%
- The maximum possible benefit is to detect and treat all 1017 cases and no controls per 10,000. We can achieve 45.5% of this benefit using the risk model based on the marker X.
- With this model, 65.2% of cases are above the high risk threshold but discounting for controls also classified as high risk, we achieve the equivalent of 45.5% of cases classified as high risk.
- Achieve the same net benefit to the population as 45.5% of cases and no controls called high risk.

Net Benefit

- If there is agreement on a rational risk threshold \( r_h \) for recommending treatment, we have seen that Net Benefit is:
  \[ HR_D(r_h) \cdot \rho - HR_D(r_h) \cdot (1-\rho) \cdot \frac{r_h}{1-r_h} \]
  which equals
  \[ P(\text{case & high risk}) - P(\text{cntl & high risk}) \cdot \frac{r_h}{1-r_h} \]
- Estimate with:
  \[ \hat{NB} = \frac{\# \text{positive cases}}{n} - \frac{\# \text{positive cntls}}{n} \cdot \frac{r_h}{1-r_h} \]
**Net Benefit → Decision Curves**

- A (rationally-chosen) risk threshold \( r_H \) encapsulates the benefits (B) of treating a case compared to the harm/cost (C) of treating a control

- A **Decision Curve** plots NB against the risk threshold \( r_H \)

**Decision Curve Example 1**

- Simulated data on 20,000 patients and a single marker \( X \)
- Marker is Normal(0,1) in controls
- Marker is Normal(1,1) in cases
- 10% of population are cases
- Using Bayes rule calculate

  \[
  \text{risk}(X) = P(D \mid X)
  \]

  – (we don’t need to model risk as a function of \( X \))

**Understanding the plot**

- If the policy is “treat none,” then NB is:

  \[
  \frac{\# \text{ positive cases} / 20000 - \# \text{ positive cntls} / 20000}{1 - r_H} \cdot \frac{r_H}{1 - r_H}
  \]

  \[
  = 0 - 0 \cdot \frac{r_H}{1 - r_H}
  \]

  \[
  = 0
  \]

- Therefore the “treat none” policy has NB\(\equiv0\) for any benefits and costs.
Understanding the plot

- If the policy is "treat all," then NB is:
  \[
  \frac{\text{# cases}/20000 - \text{# cntls}/20000}{r_H} \cdot \frac{r_H}{1-r_H} 
  \]
  \[
  = \rho - (1-\rho) \cdot \frac{r_H}{1-r_H}
  \]
- Even though \( r_H \) is not used to decide treatment, it is still used to encapsulate benefits and costs.
- The curve for "treat all" might look like a straight line, but it isn’t.

Interpreting the plot

Suppose our risk threshold is 6%
- The NB for using the risk model is 0.055
  - The same sNB as a rule that treated \( 0.055/\rho = 55\% \) of cases and no controls.
- The NB for the "treat all" strategy is 0.043
  - The same sNB as a rule that treated \( 0.043/\rho = 43\% \) of cases and no controls.
Interpreting the plot

• It is challenging to interpret Net Benefit. The main use of these plots may be to examine whether a risk model has the potential to add value -- examine whether NB is higher than “treat all”/“treat none” -- for a range of plausible risk thresholds
• If there is consensus on the risk threshold, the plot is unnecessary (potentially distracting)
  – E.g., if clinicians agree that patients should be treated with statins if 5-year risk of CVD is at least 20%.

Decision Curve Example 2

• Simulated data on 20,000 patients and a single marker X
• Marker is Normal(0,1) in controls
• Marker is Normal(1,1) in cases
• 25% of population are cases
• Using Bayes rule calculate
  \[ \text{risk}(X) = P(D | X) \]

Notice the scale change between the two plots. With higher prevalence there are more Benefits and fewer Costs.
Decision Curve Example 3

- Simulated data on 20,000 patients and a single marker X
- Marker is Normal(0,1) in controls
- Marker is Normal(1,1) in cases
- 1.5% of population are cases
- Using Bayes rule calculate risk(X)=P(D | X)

Decision Curve Example 4

- Prospective study of 570 men schedule for prostate biopsy.
- New marker: Urinary PCA3 (an RNA that is over-expressed in prostate cancer cells)
- Existing marker: Serum PSA
- Clinical risk factors: age, results of digital rectal exam
- n=541 men, prevalence 36%

Decision Curve Example 4

- Here, we compare
  - clinical model (using age and DRE results)
  - biomarker-aided prediction: (additionally use Serum PSA and PCA3 to predict risk of disease

- I used logistic regression to estimate risk for each set of predictors.
Relative Utility Plot

\[ NB = HR_D(r_H) \rho - HR_B(r_H) (1-\rho) \cdot \frac{r_H}{1-r_H} \]

- If all cases are high risk and no controls are high risk then \( NB = \text{prevalence} = \rho \)
  - \( \rho \) is the highest possible value for \( NB \), so divide by \( \rho \)
  - \( RU(r_H) \equiv sNB(r_H) = NB(r_H)/\rho \)

Relative Utility plot uses \( sNB \) rather than \( NB \) on the vertical axis (\( r_H \) remains on the horizontal axis).

- It is tempting to try to use Decision Curves (or relative utility plot) to choose \( r_H \) to maximizes Net Benefit. This is **wrong**.
  - Net Benefit depends on benefits and harms, captured by \( r_H \).
  - The data used to make the plot contain no information of the benefit of treatment to cases or the harms of treatment to controls.
  - \( r_H \) must be selected from other considerations (data?), then used to evaluate the relative merits of policies.
One last note on risk thresholds...

• If current policy in the absence of a risk model is “treat none”, that should mean that the benefits of treating all cases do not surpass the costs of treating all controls.
  \[ NB_{treat\text{-}all} < NB_{treat\text{ none}}, \text{ i.e., } \rho - (1 - \rho) \cdot \frac{r_H}{1-r_H} < 0, \]
  which implies that \( \rho < r_H \).

• Therefore, if current policy is “treat none” then it is only rationally consistent to consider \( r_H > \rho \).

Prognostic Enrichment

• Sometimes the intended use of a risk model is to identify patients at high risk for inclusion in a clinical trial
  – I am calling this “Prognostic Enrichment” following Temple, 2010 (although this term is not widely used)
Prognostic Enrichment: Example

- ADPKD patients: 20% will experience substantial decline in renal function in one year (D)
- new therapy believed to reduce the risk of D
- Designing a trial to have 90% power to detect a 30% reduction in the risk of D would require 1643 patients
  - possibly prohibitively expensive

Prognostic Enrichment Biomarker

- Suppose a biomarker has some ability to identify patients at higher risk of D
- For example, suppose that 40% of biomarker-positive patients will experience D (compared to 20% of all ADPKD patients)
- Conducting the trial in biomarker-positive patients requires 651 patients to have 90% power to detect a 30% reduction in the risk of D
  - may be much more practical

Prognostic Enrichment Biomarker

Let’s examine the impact of using the biomarker on:

- trial sample size
- total number of patients to screen to enroll trial
  - proxy for calendar time to enroll trial
- total cost of patient screening & patients in trial

Prognostic Enrichment Biomarker

Trial sample size: key point is that sample size is calculated based on statistical testing and clinical parameters

- Based on the desired power 0<1-β<1, Type I error rate 0<α<1, event rate without intervention 0<τ<1, and event rate with intervention 0<ϕ<1, the sample size N across the two arms of the trial for a two-sided test is
  \[ N = \left( \frac{\phi^{-1}(1-\beta)}{\sqrt{2\left(\pi(1-\pi)^2 + (\phi^{-1}(1-\beta) - \phi^{-1}(\pi))^2\right)}} \right)^2 \]

where \( \pi \neq \tau \) and \( \phi^{-1}(x) \) is the quantile function of the standard Normal distribution such that \( \phi^{-1}(x) = z \) where \( P[Z\leq z]=x \). For a one-sided test the formula is the same except replacing \( \phi^{-1}(1 - \frac{\pi}{2}) \) with \( \phi^{-1}(1 - \alpha) \).
Prognostic Enrichment Biomarker

Total number of patients to screen to enroll trial
• Suppose we use threshold $t$ to decide eligibility for the trial. That is, the fraction $t$ of patients at lowest risk for D are screened from the trial.
• That implies that $1/(1-t)$ patients must be screened to identify one patient eligible for the trial.
• Therefore total patients screened = (Trial Sample Size) / (1-$t$)

Prognostic Enrichment Biomarker

total cost of patient screening & patients in trial
• Let $C1$ be the cost of running a patient through the trial and let $C2$ be the cost of screening a patient for the trial using the biomarker
• Total Cost with screening threshold $t$ is
  \[ TC = C1 \times SS + C2 \times \frac{SS}{1-t} = SS(C1 + \frac{C2}{1-t}) \]
• However, when $t=0$ no screening is needed so in this special case $TC = C1 \times SS$

Prognostic Biomarker 1

• Event rate without prognostic enrichment: 20%
• AUC of biomarker: 0.72
• Cost to measure biomarker: $100
• Cost to run one patient through trial: $400
• Specifying trial design to have 90% power to detect a 30% reduction in event rate using $\alpha=0.025$ with one-sided testing
Prognostic Biomarker 2

- Event rate without prognostic enrichment: 20%
- AUC of biomarker: 0.92
- Cost to measure biomarker: $100
- Cost to run one patient through trial: $400
- Specifying trial design to have 90% power to detect a 30% reduction in event rate using $\alpha = 0.025$ with one-sided testing

Prognostic Enrichment – Other Important Considerations

- Generalizability
  - by definition, the intervention will not be tested on patients screened out of the trial
  - this may lead to investigators to err on the side of less stringent screening
- Ethics
  - In oncology, the primary motivation for prognostic enrichment is traditionally not cost. Rather, therapies are often toxic and only ethical to test on patients with poor prognosis
  - The “event-rate in biomarker positive patients” becomes a quantity of primary interest
  - Such ethical considerations may lead investigators to err on the side of more stringent screening.
Insight into the utility of markers for prognostic enrichment

- Sometimes unimpressive markers look like they could be helpful for prognostic enrichment
  - e.g. prognostic biomarker 1 had modest AUC, 0.72
- This is because the biggest "gains" in reduced sample size are at the low end of the event rate (next slide)
  - Detecting a 30% reduction in the event rate requires much larger sample sizes if the event rate is 10% (vs 7%) compared to 20% (14%)
    - "a little bit of enrichment can go a long way"

Summary of Section 2

- In order for a risk model to be valid it must be well-calibrated
  - Otherwise cannot interpret predicted risks as risks
  - Recommend graphical assessment (moderate calibration)
- Risk model discrimination
  - Can use ROC curve but more informative to use an alternative that shows the risk threshold
  - Presented AUC and other numeric measures
Summary of Section 2

• Evaluating a risk model (or biomarker) for prognostic enrichment of a clinical trial. Key considerations:
  – trial sample size
  – total patients screened to enroll trial/calendar time to enroll
  – cost savings of smaller trial vs. cost of screening
  – generalizability
  – ethics of eligibility criteria