

SISCER Module 5
Part II: Evaluating Risk Models
July 12-13, 2021
8:30-Noon PT / 11:30-3pm ET



Kathleen Kerr, PhD
Professor of Biostatistics
SISCER Director
University of Washington



Risk Model Assessment

- Risk Model Calibration
 - required for a risk model to be *valid*
 - Particularly crucial whenever a risk model will be used to convey risk to patients
- Risk Model Discrimination Performance
 - required for a risk model to be *useful*
 - performance assessment depends on what the model will be used for

CALIBRATION

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 a breast cancer diagnosis in the next 5 years is 1%”
- In order to be valid, risks must be calibrated

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

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

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

104

Mean Calibration

- Also called “calibration-in-the-large”
- Def: average predicted risk equals the prevalence
- To assess, compare event rate with average predicted risk
 - If 3% of the population are cases, then the risk model has mean calibration if the average predicted risk is 3%
- Very low bar

105

Weak Calibration

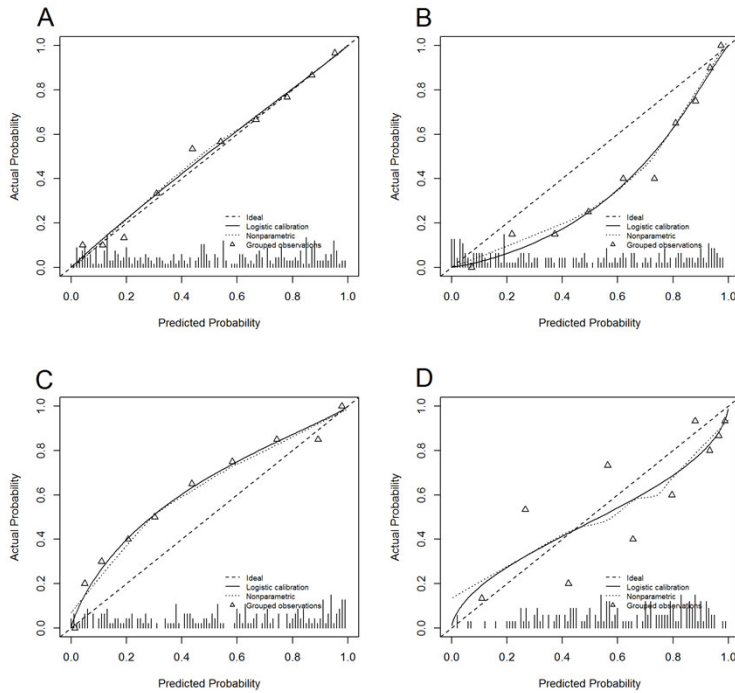
- Also called “Logistic calibration”
- Predicted risks are obtained from a previously developed model for D (e.g., based on logistic regression); the linear combination of predictors defines the “linear predictor” $L=b_0+b_1 X_1 + \dots + b_k X_k$
- Regress D on L: $\text{logit}(P(D) | L) = a + b L$
 - a is the “calibration intercept”; b is the “calibration slope”
- Def: If $a \approx 0$ and $b \approx 1$, the model is calibrated in the weak sense
- In data not used to fit the model, typically the calibration slope $b < 1$: large predicted risks are too high and low predicted risks are too low
 - (R demo)

106

Moderate Calibration

- Def: $P(D = 1 | \widehat{\text{risk}}(X_1, X_2) = r) = r$
 - here, there are two risk factors X_1 and X_2
- “collapses” data among groups of people with the same predicted risk
- Common practice to assess: 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 calibration (in the moderate sense); and 3 poorly calibrated risk models

107



Forecasts of rain: are the risks well calibrated?

From *The Signal and the Noise*, Nate Silver, The Penguin Press 2012.

FIGURE 4-7: NATIONAL WEATHER SERVICE CALIBRATION

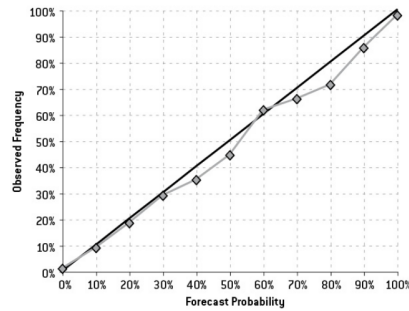


FIGURE 4-8: THE WEATHER CHANNEL CALIBRATION

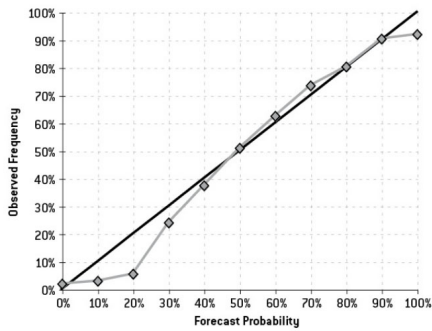
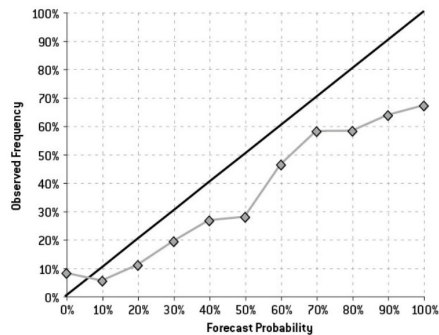


FIGURE 4-9: LOCAL TV METEOROLOGIST CALIBRATION



how NOT to assess moderate calibration

- Hosmer-Lemeshow test statistic
- p-value from Hosmer-Lemeshow test

- In small datasets, badly miscalibrated models may not give a large H-L test statistic or a significant p-value
- In large datasets, small/unimportant deviations from good calibration can produce large H-L test statistic or small p-value

110

Calibration plots

- Can be sensitive to the choice of the groups and choice of smoother and other options (beware of smooths that eliminate “outliers”)

111

Strong Calibration

- Def: $\widehat{risk}(X_1, X_2) = P(D=1 | X_1, X_2)$
- Must consider every unique combination of predictors and ask whether observed and predicted risks agree for people with that combination
- Compared to calibration in the moderate sense, does not “collapse” groups of people with the same $\widehat{risk}(X_1, X_2)$
- Typically only feasible to assess when there are a limited number of predictors and they are all categorical

112

Predicted risks for Huntington’s Disease

Level	Definition	Individuals with 1 HD Parent: risk=50%	Genotyped individuals: risk is 0% or 100%
Mean	Observed event rate equals average predicted risk		
Weak	No systematic over- or under-estimation of risks		
Moderate	Predicted risks correspond to observed event rates		
Strong	For every combination of risk factors, predicted risks correspond to observed event rates		

Calibration is not enough

- If the prevalence is ρ , a calibrated risk model assigns everyone risk ρ .
- A risk model needs to do more: stratify people into “low risk” and “high risk” groups.
 - Achieved perfectly by genotyping the *HTT* gene among those at risk for Huntington’s disease, but less perfectly for most applications.

114

RISK MODEL DISCRIMINATION PERFORMANCE

Risk Model Performance

We will discuss three classes of assessment

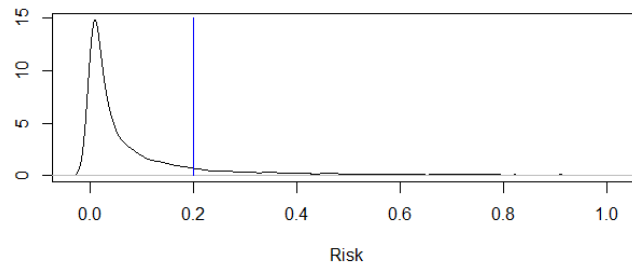
- Generic measures
 - “purely mathematical”
 - meaning: they do not directly translate to any clinical, public health, or public policy impact of using the risk model
- Assessing performance when model will be used to recommend treatment/intervention to high risk individuals
- Assessing performance for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

116

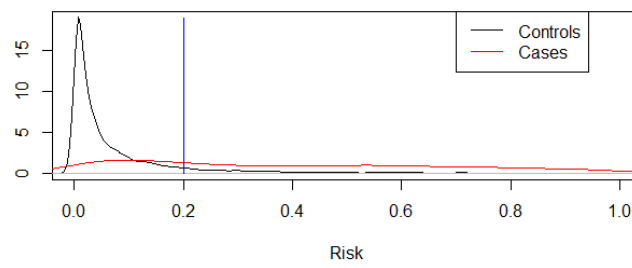
The Distribution of Risk

- Case ($D=1$) and control ($D=0$) risk distributions are fundamental components of all performance measures
- When examining risk distributions, 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: risk model for simulated data from DABS website

Distribution of Predicted Risks, 20% risk threshold

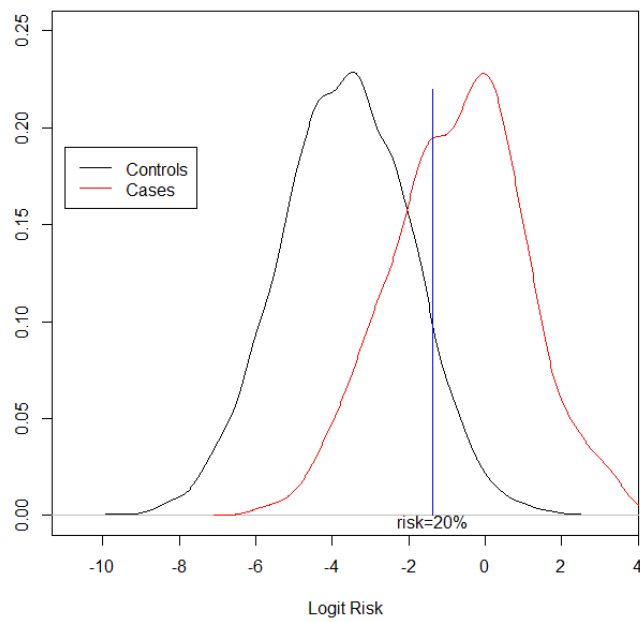


For Cases and Controls



118

Distribution of Risk on Logit Scale



119

GENERIC MEASURES OF RISK MODEL PERFORMANCE (MEASURES THAT DO NOT USE A RISK THRESHOLD)

MRD, AARD, AUC

- MRD = Mean Risk Difference
- AARD = Above Average Risk Difference
- AUC = Area Under the ROC Curve

These are measures of **discrimination**. They quantify:

How well does the risk model discriminate/separate cases and controls?

Mean Risk Difference (MRD)

$MRD \equiv \text{mean}(\text{risk}(X) | \text{case}) - \text{mean}(\text{risk}(X) | \text{control})$

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

For the DABS data example, $\text{mean}(\text{risk} | \text{case}) = 0.391$,
 $\text{mean}(\text{risk} | \text{cntl}) = 0.069$; $MRD = 0.322$

122

Above Average Risk Difference (AARD)

$AARD = P(\text{risk}(X) > \rho | \text{case}) - P(\text{risk}(X) > \rho | \text{control})$

$AARD = 0.797 - 0.198 = 0.599$ in the DABS example

- Can also write as: $HR_D(\rho) - HR_{\bar{D}}(\rho)$
 - or $TPR(\rho) - FPR(\rho)$
- Related to Net Benefit metrics (will come to these soon)
 - $NB(r) = \rho HR_D(r) - (1 - \rho) \frac{r}{1-r} HR_{\bar{D}}(r)$; set $r = \rho$ and divide by ρ

123

AUC for a Risk Model

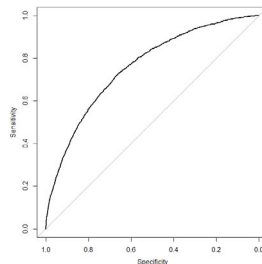
AUC = Area Under the ROC Curve = $P(\text{risk}_{\text{case}}(X) > \text{risk}_{\text{cntl}}(X))$

- Ignores the meaning of risk
- AUC not a clinically relevant measure of predictive performance
 - Arguably roughly similar to MRD in terms of clinical relevance
- Rather than AUC, it might be more clinically relevant to average TPR over a relevant range of FPR (rather than entire range)
 - pAUC

124

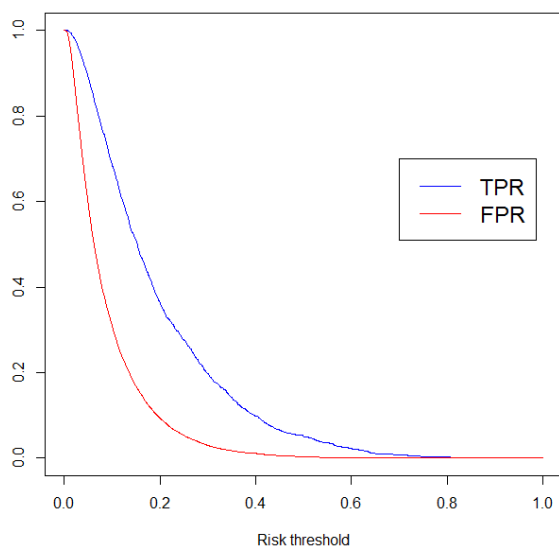
ROC Curve for a risk model

- 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 ROC alternative that addresses this disadvantage.



125

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



EVALUATING A RISK MODEL FOR RECOMMENDING TREATMENT

Risk-based decisions

- Sometimes the purpose of a risk model is to inform who should be treated
 - e.g., screen high risk individuals for cancer
 - e.g., treat individuals at high risk of a heart attack with statins
 - e.g., treat cancer patients with high risk of relapse with adjuvant chemotherapy
- What risk threshold should define “high risk”?

128

Benefits and Costs of Treatment

- Assume there is some expected benefit B to treating a case
 - life extended, morbidity reduced
- Assume there is some cost C to treating a control
 - cost includes side effects of treatment, stress/anxiety, toxic exposures, as well as monetary cost

129

Choice of Risk Threshold

Classical Decision Theory Result

Treatment offers benefit B to a case and cost C to a control. Then the optimal risk threshold r_H for selecting high-risk patients for treatment is

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

Pauker and Kassierer, The threshold approach to clinical decision making. *NEJM* 1980.

Vickers and Elkin, Decision Curve Analysis. *Medical Decision Making* 2006.

130

Choice of Risk Threshold

Outline of derivation:

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

When should patients choose treatment?

- When expected result of treatment > 0
- $E(\text{benefit} | D=1, X)P(D=1 | X) - E(\text{cost} | D=0, X)P(D=0 | X) > 0$

$$B \cdot P(D=1 | X) - C \cdot P(D=0 | X) > 0$$

$$B \cdot P(D=1 | X) > C \cdot P(D=0 | X)$$

$$\frac{P(D=1 | X)}{1 - P(D=1 | X)} > \frac{C}{B}$$

131

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.

132

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 1/4th the benefit of treating a case.

The benefit of treating a case is four times larger than the cost of treating a control.

133

Choice of Risk Threshold: Example 2

Gail (JNCI, 2009) evaluated risk models for breast cancer in terms of decisions 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

$$C/B = 0.0077 \rightarrow r_H = 0.0076 \text{ per year}$$

134

Net Benefit of a Risk Model

Let r_H be the appropriate risk threshold to define “high risk” and recommended for treatment.

Key elements:

$$HR_D(r_H) = P(r(X) > r_H \mid D=1)$$

– essentially, TPR of the risk model for the risk threshold

$$HR_{\bar{D}}(r_H) = P(r(X) > r_H \mid D=0)$$

– essentially, FPR of the risk model for the risk threshold

Next we calculate the net benefit of using the risk model and r_H to decide treatment in the population. We assume r_H has been *rationally selected*, i.e. r_H corresponds to the benefits and costs of treatment

135

Net Benefit of a Risk Model

Overall population impact of the risk model – combines $HR_D(r_H)$ and $HR_{\bar{D}}(r_H)$:

$$\begin{aligned} \text{NB}(r_H) &= B P(D=1) HR_D(r_H) - C P(D=0) HR_{\bar{D}}(r_H) \\ &= B \left\{ P(D=1) HR_D(r_H) - \frac{r_H}{1-r_H} P(D=0) HR_{\bar{D}}(r_H) \right\} \\ &= P(D=1) HR_D(r_H) - \frac{r_H}{1-r_H} P(D=0) HR_{\bar{D}}(r_H) \end{aligned}$$

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

136

Net Benefit Example: DABS simulated data

- D is CVD over 10 years
 - $P(D=1)=10.17\%$
 - Marker X
- Suppose $r_H=20\%$:
 - $HR_D(r_H) = 65.2\%$
 - $HR_{\bar{D}}(r_H) = 8.9\%$
 - $\text{NB}(r_H)=0.046 \cdot$ benefit of statins to subject who would have a CVD event without them

137

Standardized Net Benefit

$$NB(r_H) = P(D=1) HR_D(r_H) - \frac{r_H}{1-r_H} P(D=0) HR_{\bar{D}}(r_H)$$

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

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

Standardized Net Benefit $\equiv NB(r_H) / \rho$

$$= HR_D(r_H) - \frac{r_H}{1-r_H} \frac{1-\rho}{\rho} HR_{\bar{D}}(r_H)$$

= TPR discounted by an appropriate amount of FPR

Interpretation: sNB is z% \rightarrow risk model achieves z% of the benefit of a perfectly discriminating model

138

DABS Simulated Data Example, continued interpreting standardized net benefit

- $sNB=0.046/0.1017=0.455=45.5\%$
- 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.
- With this model, 65.2% of cases are above the high risk threshold; discounting for controls also classified as high risk, we achieve the equivalent of 45.5% of cases classified as high risk and 0% of controls
- Achieve the same net benefit to the population as 45.5% of cases and no controls called high risk.

139

Assessing Net Benefit Graphically

- Decision Curves
 - Proposed in: Vickers and Elkin, “Decision Curve Analysis: A Novel Method for Evaluation Prediction Models.” *Medical Decision Making*, 2006.
 - Additional ref: Kerr, Brown, Zhu, and Janes: “Assessing the Clinical Impact of Risk Prediction Models with Decision Curves: Guidance for Correct Interpretation and Appropriate Use.” *J Clinical Oncology*, 2016.
- Related to Relative Utility Curves
 - Papers by Baker, e.g. “Putting Risk Prediction in Perspective” Relative Utility Curves.” *JNCI*, 2009.

140

Net Benefit

- If there is agreement on a rational risk threshold r_H for recommending treatment, we have seen that Net Benefit is:

$$HR_D(r_H) \rho - HR_{\bar{D}}(r_H) (1-\rho) \frac{r_H}{1-r_H}$$

which equals

$$P(\text{case \& high risk}) - P(\text{cntl \& high risk}) \frac{r_H}{1-r_H}$$

- Estimate with:

$$\widehat{NB} = \frac{\# \text{ positive cases}}{n} - \frac{\# \text{ positive cntls}}{n} \frac{r_H}{1-r_H}$$

141

Net Benefit → Decision Curves

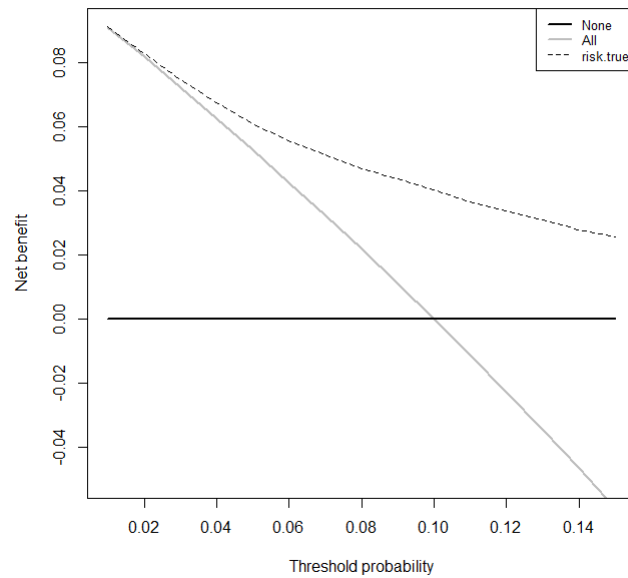
- A (rationally-chosen) risk threshold r_H encapsulates the benefits (B) of treating a case compared to the harm/cost (C) of treating a control
- A **Decision Curve** plots NB against the risk threshold r_H

142

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
$$\text{risk}(X) = P(D | X)$$
 - (we don't need to model risk as a function of X)

143



144

Understanding the plot

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

$$\begin{aligned}
 & \# \text{ positive cases}/20000 - \# \text{ positive cntls}/20000 \frac{r_H}{1-r_H} \\
 &= 0 - 0 \cdot \frac{r_H}{1-r_H} \\
 &= 0
 \end{aligned}$$

- Therefore the “treat none” policy has $NB \equiv 0$ for any benefits and costs.

145

Understanding the plot

- If the policy is “treat all,” then NB is:

$$\begin{aligned} & \# \text{ cases}/20000 - \# \text{ cntls}/20000 \frac{r_H}{1-r_H} \\ & = \rho - (1-\rho) \cdot \frac{r_H}{1-r_H} \end{aligned}$$

- Even though r_H is not used to decide treatment under the “treat all” policy, it is used to capture/summarize benefits and costs.
- The curve for “treat all” might look like a straight line, but it isn’t.

146

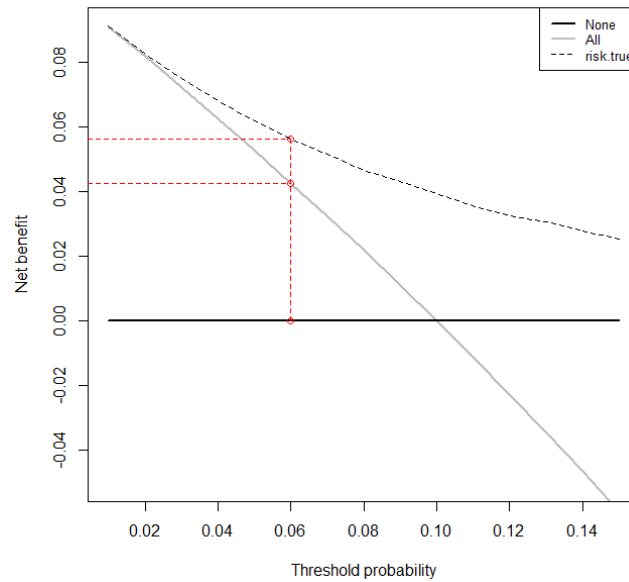
Understanding the plot

- If the policy is to use the risk model to recommend treatment, then NB is estimated by considering each risk threshold and the number of cases and controls that exceed the threshold:

$$\frac{\# \text{ positive cases}}{n} - \frac{\# \text{ positive cntls}}{n} \frac{r_H}{1-r_H}$$

147

Interpreting the plot



148

Interpreting the plot

- Suppose the 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 .

149

Interpreting the plot

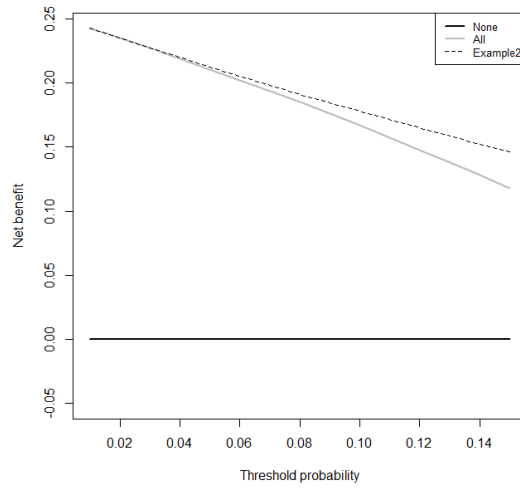
- It is challenging to interpret Net Benefit. The main use of these plots may be to examine whether a risk model has the potential to add value – examine whether NB is higher than “treat all”/“treat none” – for a range of plausible risk thresholds
- If there is consensus on the risk threshold, the plot is unnecessary (potentially distracting)
 - E.g., if clinicians agree that patients should be treated with statins if 10-year risk of CVD is at least 20%.

150

Decision Curve Example 2

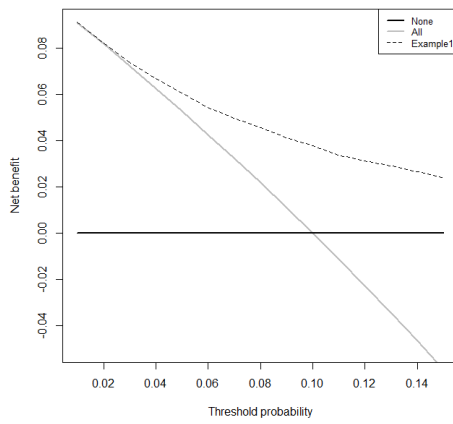
- Simulated data on 20,000 patients and a single marker X
- Marker is Normal(0,1) in controls
- Marker is Normal(1,1) in cases
- 25% of population are cases
- Use Bayes rule to calculate
$$\text{risk}(X) = P(D|X)$$

151

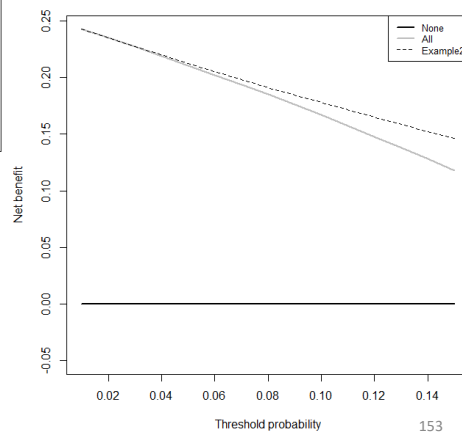


It is more difficult for risk-based treatment to “beat” Treat-All when prevalence is high.

152



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

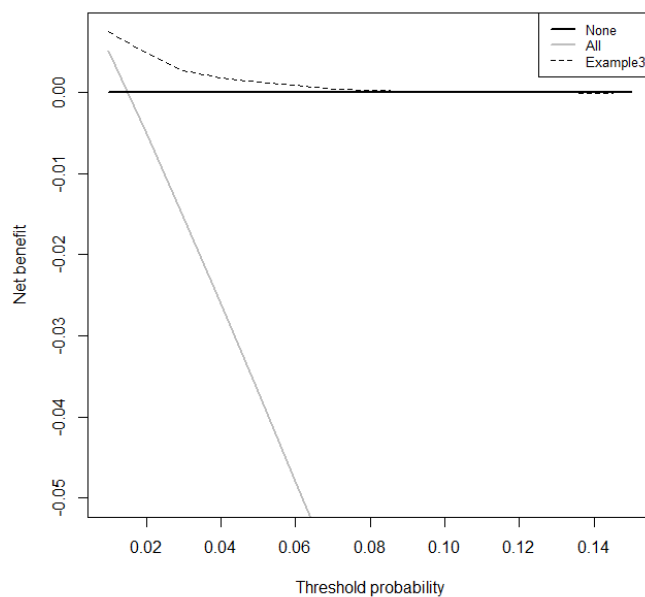


153

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
 $\text{risk}(X) = P(D|X)$

154



155

Decision Curve Example 4

- Prospective study of 570 men scheduled 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%

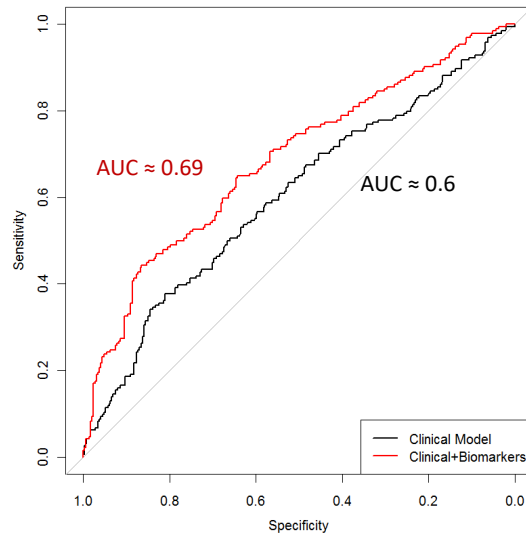
156

Decision Curve Example 4

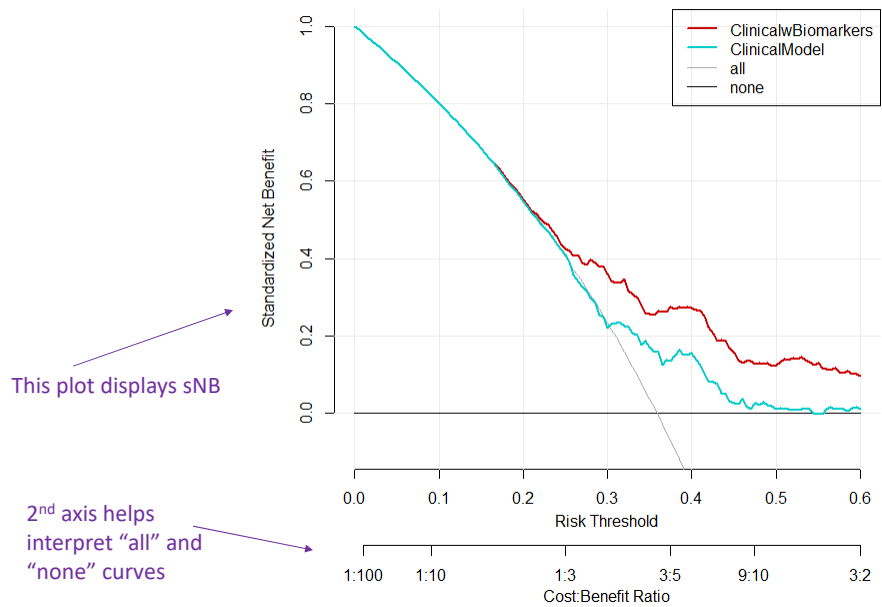
- Here, we compare
 - clinical model (using age and DRE results)
 - biomarker-aided prediction: (additionally use Serum PSA and PCA3 to predict risk of disease)
- I used logistic regression to estimate risk for each set of predictors.

157

ROC Curves



158



159

- It is tempting to use Decision Curves to choose r_H to maximize 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.

160

- Decision curves are potentially useful when there is no consensus on an appropriate treatment threshold, to compare the performance of different risk models across a range of plausible thresholds.
- In the prostate cancer example, the risk model that used biomarkers only offers higher Net Benefit than the clinical model if r_H exceeds about 25%.
 - It is likely that patients and clinicians would say r_H is much smaller than 25%.

161

Notes on Decision Curves and Net Benefit

- The curve for “treat all” and “treat none” always cross at the prevalence:
- $NB_{\text{treat-all}} = 1 \cdot \rho - 1 \cdot (1-\rho) \cdot \frac{r_H}{1-r_H} = \rho - (1-\rho) \frac{r_H}{1-r_H} = 0$ if and only if $r_H = \rho$
- Moreover:
 - “treat all” beats “treat none” if $r_H < \rho$
 - “treat none” beats “treat all” if $r_H > \rho$.

162

Alternative Formulation

- Most published Decision Curves use treat-none as the reference
- These are “opt-in Decision Curves”
 - implicitly assume the default is treat-none
 - high risk patient *opt into* treatment
- But what if standard is treat-all and we envision opting *low-risk* patients *out* of treatment?
 - benefits accrue from controls who avoid cost C at the expense of cases who miss out on B

163

Alternative Formulation

Opt-out Decision Curves use treat-all as the reference and display the “opt-out Net Benefit”

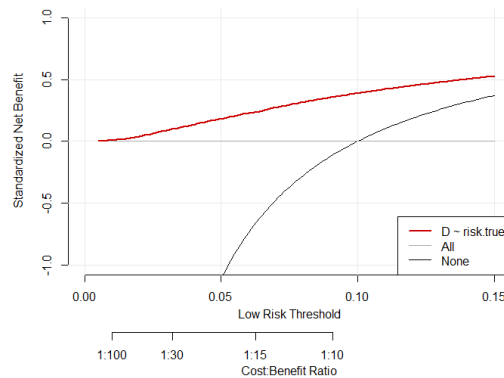
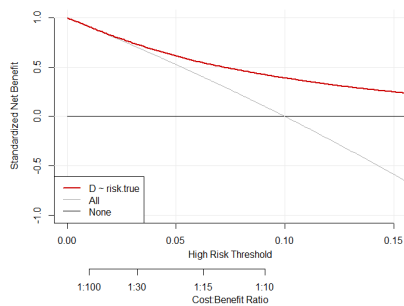
$$NB^{opt-out} = (1 - \rho)LR_{\bar{D}}(r_H) - \rho \frac{1 - r_H}{r_H} LR_D(r_H)$$

$$sNB^{opt-out} = LR_{\bar{D}}(r_H) - \frac{\rho}{1-\rho} \frac{1-r_H}{r_H} LR_D(r_H)$$

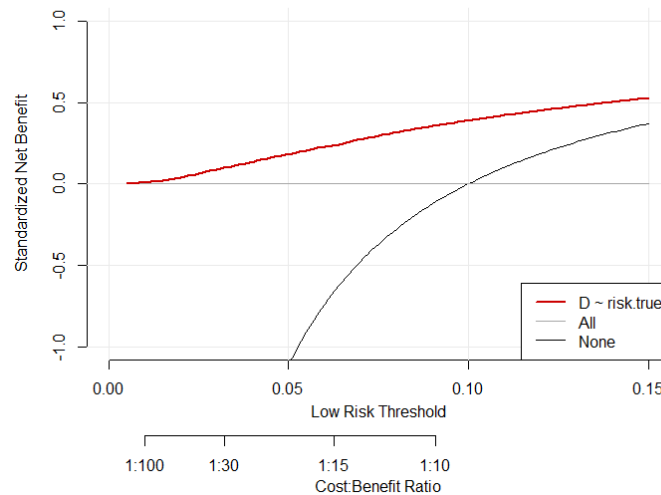
true negative rate

false negative rate

Example 1 revisited: Opt-in and Opt-out Decision Curves



Example 1 revisited: Opt-out Decision Curve




166

Opt-out Decision Curves

- More useful than opt-in decision curves when current default is treat-all.
- Better-suited to evaluate the evidence for switching from treat-all to risk-based treatment

Brief Report

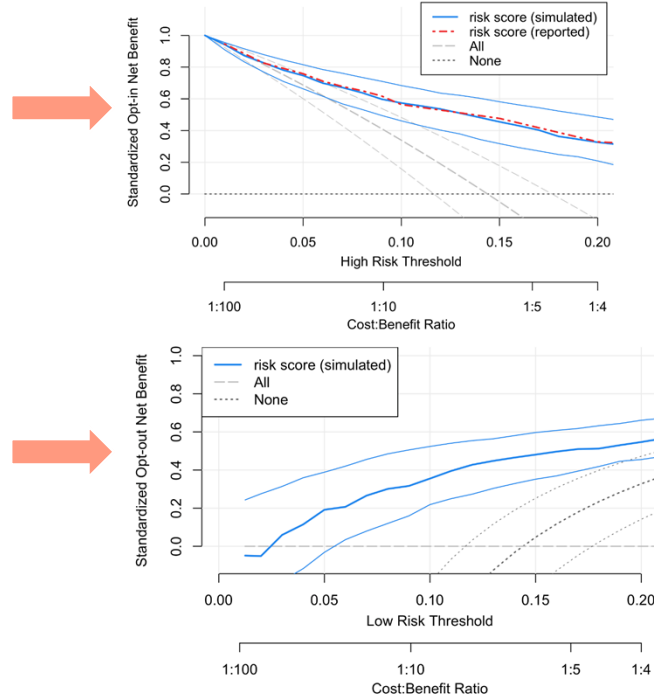
Assessing the Clinical Impact of Risk Models for Opting Out of Treatment

Kathleen F. Kerr , Marshall D. Brown, Tracey L. Marsh, and Holly Janes

MDM
Medical Decision Making

Medical Decision Making
2019, Vol. 39(2) 86–90
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0272989X18819479
journals.sagepub.com/home/mdm





168

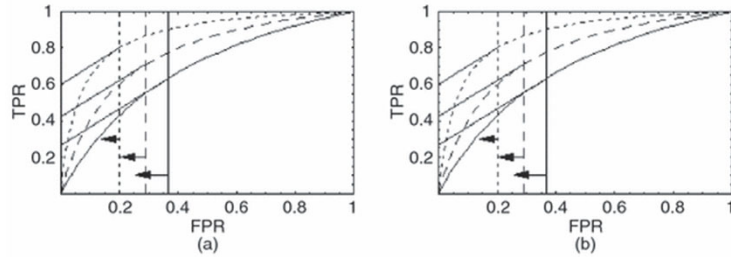
Relative Utility Curves

- Baker's Relative Utility Curves use treat-none as the reference above ρ and treat-all as the reference below ρ .
 - Thus the reference policy changes at ρ
 - Creates “hill shaped” curves that crest at ρ
 - Relative Utility is related to standardized Net Benefit

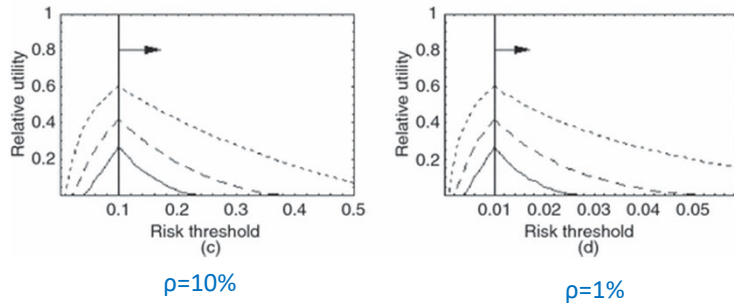
169

Baker, Cook, Vickers, & Kramer (2009)

ROC Curves:



Relative Utility Curves:



Journal of the Royal Statistical Society: Series A (Statistics in Society)
 Volume 172, Issue 4, pages 729-748, 2009
<http://onlinelibrary.wiley.com/doi/10.1111/j.1467-985X.2009.00592.x/full#f1>

Key Assumptions of all Decision Curves/Relative Utility Curves

- Expected benefit of treatment B is the same for all cases; expected cost of treatment C is the same for all controls
 - Biomarker does not predict treatment effect
- Risk threshold r_H is rational – reflects the Cost:Benefit Ratio
- Reminder: under these assumptions, curves show the Net Benefit of the risk model *to the population*

Don't forget uncertainty

- To focus on interpretation, I showed Decision Curves without confidence intervals
- Unfortunately, Decision Curves often appear in the literature without any acknowledgement of uncertainty
- As in any other inference from biomedical data, we should acknowledge the uncertainty in our inferences
 - Confidence intervals in plots and/or tables of Net Benefit (*R* demo)

172

**EVALUATING A RISK MODEL FOR
PROGNOSTIC ENRICHMENT OF
CLINICAL TRIALS**

Prognostic Enrichment

- Sometimes the intended use of a risk model is to identify patients at high risk for inclusion in a clinical trial
 - Temple (2010) called this “Prognostic Enrichment”

Temple, Enrichment of Clinical Study Populations, *Clinical Pharmacology and Therapeutics*, 2010

174

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

175

Prognostic Enrichment Biomarker

- Suppose a biomarker has some ability to identify ADPKD 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

176

Prognostic Enrichment Biomarker

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

177

Prognostic Enrichment Biomarker

Trial sample size: calculated based on statistical testing and clinical parameters

- Based on the desired power $0 < 1 - \beta < 1$, Type I error rate $0 < \alpha < 1$, event rate without intervention $0 < \pi < 1$, and event rate with intervention $0 < \tau < 1$, the sample size SS across the two arms of the trial for a two-sided test is $SS = 2 \times \frac{\left(\phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{2\left(\frac{\pi + \tau}{2}\right)\left(1 - \frac{\pi + \tau}{2}\right)} + \phi^{-1}(1 - \beta) \sqrt{\pi(1 - \pi) + \tau(1 - \tau)} \right)^2}{(\pi - \tau)^2}$,

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

178

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)

179

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

$$TC = C1 \times SS + C2 \times \frac{SS}{1-t} = SS(C1 + \frac{C2}{1-t})$$

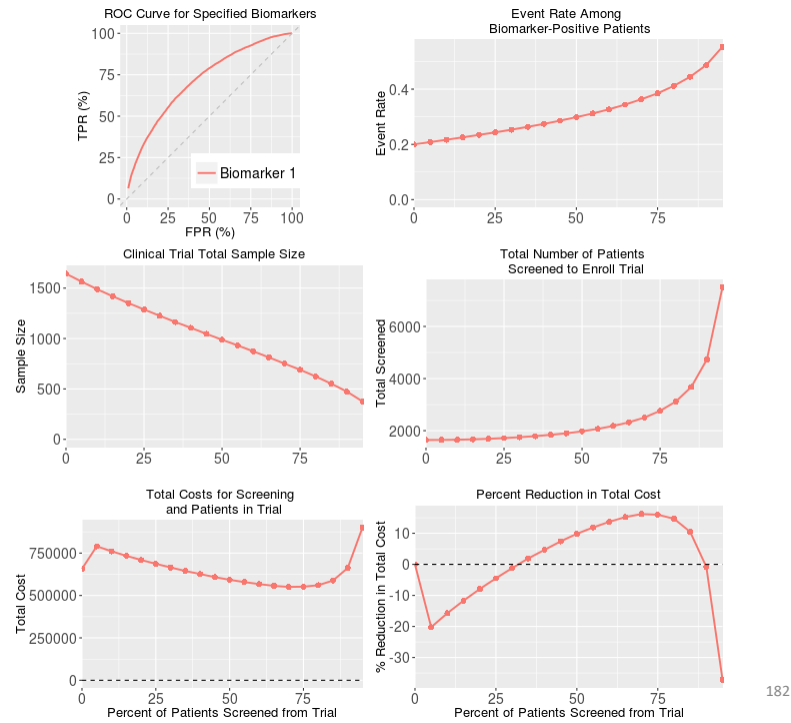
- However, when $t=0$ no screening is needed so in this special case $TC = C1 \times SS$

180

Prognostic Biomarker 1

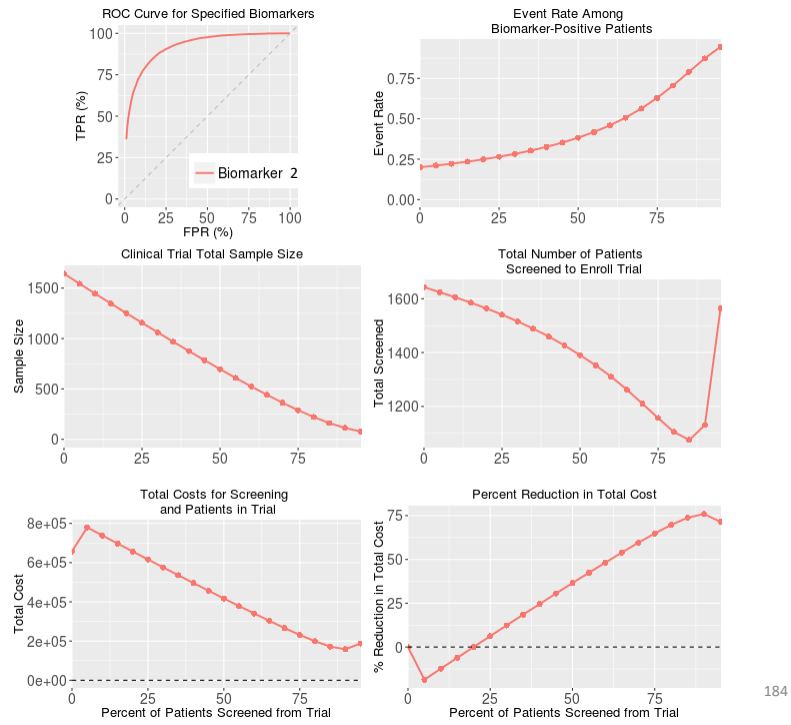
- Event rate without prognostic enrichment: 20%
- AUC of biomarker: **0.72**
- Cost to measure biomarker: \$100
- Cost to run one patient through trial: \$400
- Specifying trial design to have 90% power to detect a 30% reduction in event rate using $\alpha=0.025$ with one-sided testing

181

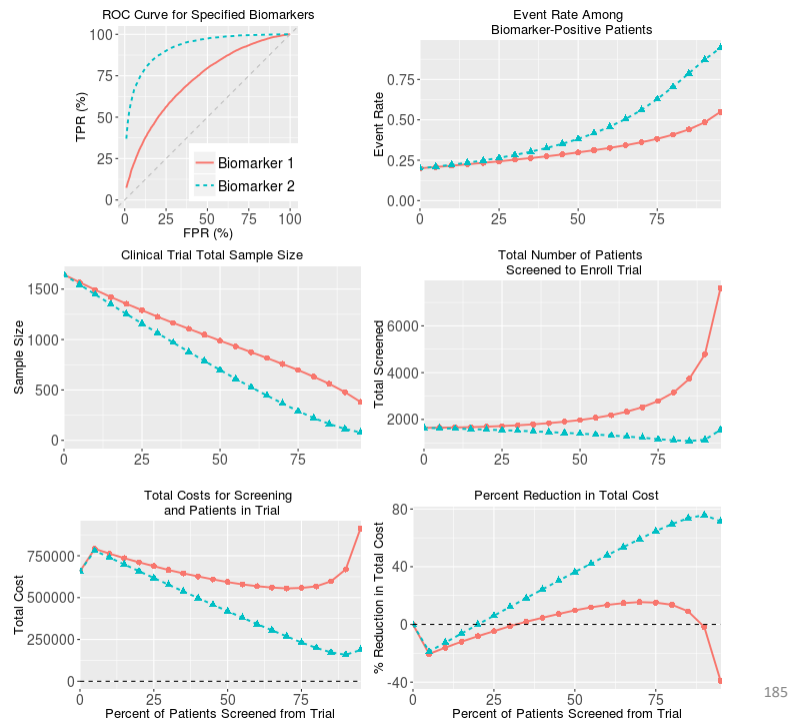


Prognostic Biomarker 2

- 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



184



185

Prognostic Enrichment – Other Important Considerations

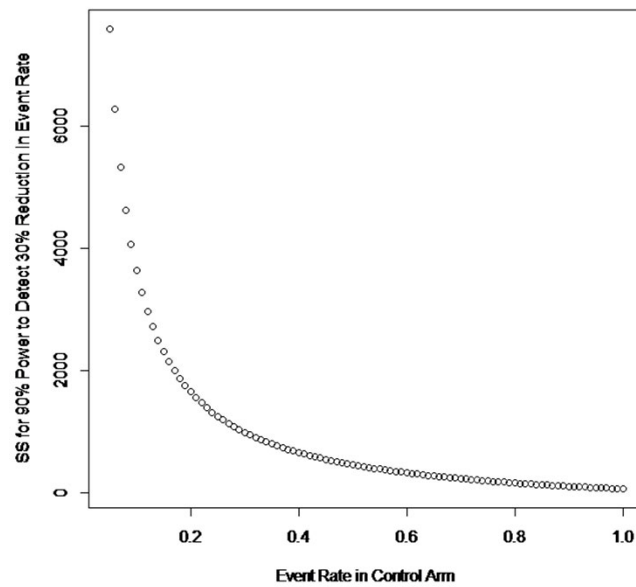
- Generalizability
 - by definition, the intervention will not be tested on patients screened out of the trial
 - this may lead to investigators to err on the side of less stringent screening
- Ethics
 - In oncology, the primary motivation for prognostic enrichment is traditionally not cost. Rather, therapies are often toxic and only ethical to test on patients with poor prognosis (very likely to have the bad clinical outcome)
 - 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.

186

Insight into the utility of markers for prognostic enrichment

- Sometimes (nominally) unimpressive markers are helpful for prognostic enrichment
 - e.g. prognostic biomarker 1 had modest AUC, 0.72
- This is because the biggest “gains” in reducing 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% (vs 14%)
 - “a little bit of enrichment can go a long way”

187



188

Summary of Part II

- 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)
 - Recommend assessing strong calibration when possible
- Risk model discrimination
 - Can use ROC curve; more informative to use an alternative that shows the risk threshold
 - Presented AUC and other numeric measures

189

Summary of Part II

- Decision Curves
 - Potentially useful to evaluate risk-based treatment policies over a range of plausible risk thresholds
 - Challenging to interpret values of NB
 - Aids the assessment of the population impact of treatment policies
 - If standard is treat-all and motivation is to opt low-risk patients out of treatment, use opt-out Decision Curves
- I prefer standardized Net Benefit over Net Benefit
 - Maximum sNB always 1.0 (or 100%)
 - Slightly easier to interpret

190

Summary of Part II

- 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

191



Misconceptions about Biomarkers and Risk Models



- A large odds ratio implies that a biomarker is useful for prediction. ❌
- A data analyst can identify the optimal threshold from an ROC curve. ❌
- A data analyst can identify the optimal risk threshold from a Decision Curve. ❌
- The best biomarker to improve a risk model is the one with strongest association with the outcome.
- To improve prediction, a new biomarker should be independent of existing predictors
- To assess whether to add new biomarker to a risk model, multiple stages of hypothesis testing are needed.
- We can often use biomarkers to identify which patients will benefit from treatment.