SISCR Module 7

Part II:
Evaluating Risk Models

Kathleen Kerr, Ph.D.
Associate Professor
Department of Biostatistics
University of Washington

Part II: Risk Model Assessment

- Risk Model Calibration
  - required for a risk model to be *valid*

- Risk Model Performance
  - required for a risk model to be *useful*

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

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

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

Level	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; difficult to assess

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

Assessing Mean Calibration

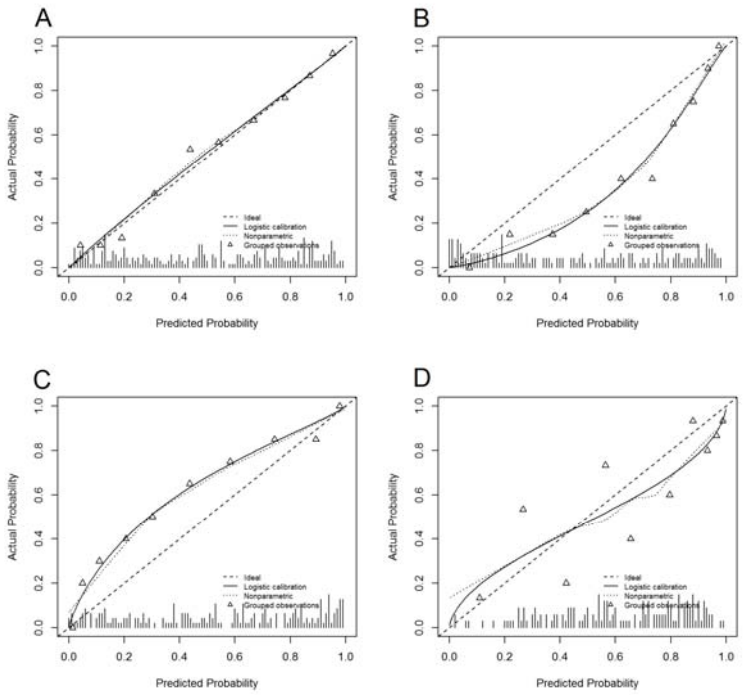
- Compare event rate with average predicted risk
  - If D=1 for 3% of the population, then the risk model has mean calibration if the average predicted risk is ≈3%
- Very low bar

Assessing Weak Calibration

- "Weak calibration" also known as "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 ≈ 0 and b ≈ 1, the model is weakly calibrated
- 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

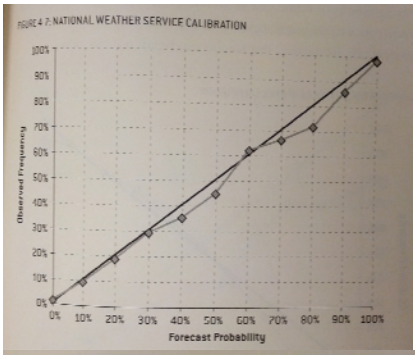
Assessing Moderate Calibration

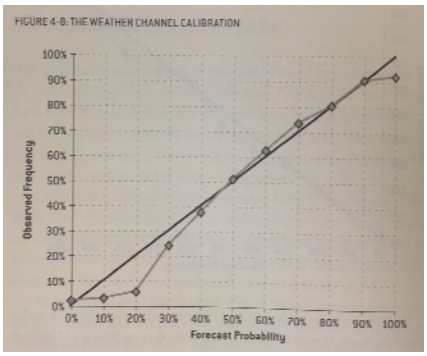
- require $P(D = 1|\widehat{risk}(X_1, X_2) = r) = r$
  - here, there are two predictors X_1 and X_2
- "collapses" data among groups of people with the same predicted risk
- Common practice: divide available data into deciles based on predicted risks
- Compare event rate in a decile of individuals with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

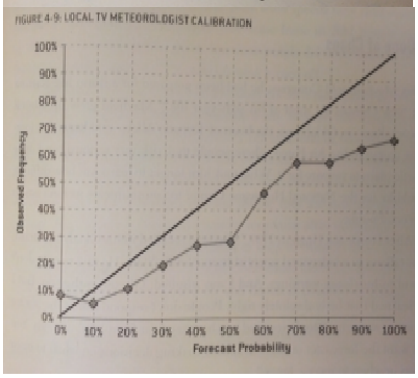


Forecasts of rain: are the risks well calibrated?

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







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

Calibration plots to assess moderate calibration

- While I favor use of these plots, one should be aware that they can be sensitive to the choice of the groups and choice of smoother and other options (beware of smooths that eliminate "outliers")

Assessing Strong Calibration

- Must consider every unique combination of predictors and ask whether observed and predicted risks agree for people with that combination
- $\widehat{risk}(X_1, X_2)$ compared to P(D=1|X_1, X_2)
- Compared to moderate calibration, 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

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

- The goal for a risk model requires more than good calibration: to stratify people into "low risk" and "high risk" groups.
  - Achieved perfectly by genotyping the *HTT* gene, but less perfectly for most applications.
- If the prevalence of a condition is ρ, a calibrated risk model assigns everyone risk ρ. If we only cared about calibration, we would not need to identify risk factors and develop a risk prediction model.

RISK MODEL PERFORMANCE

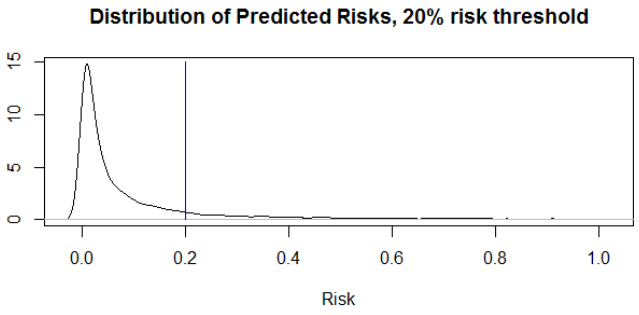
Risk Model Performance

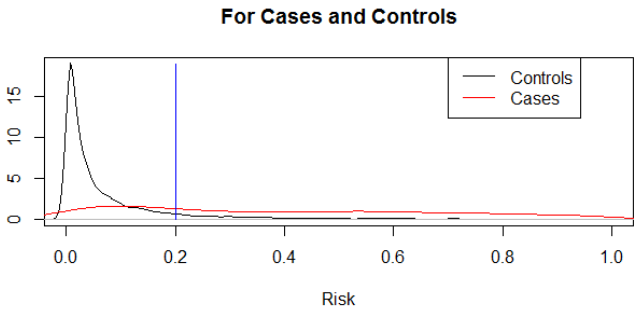
We will discuss three classes of assessment

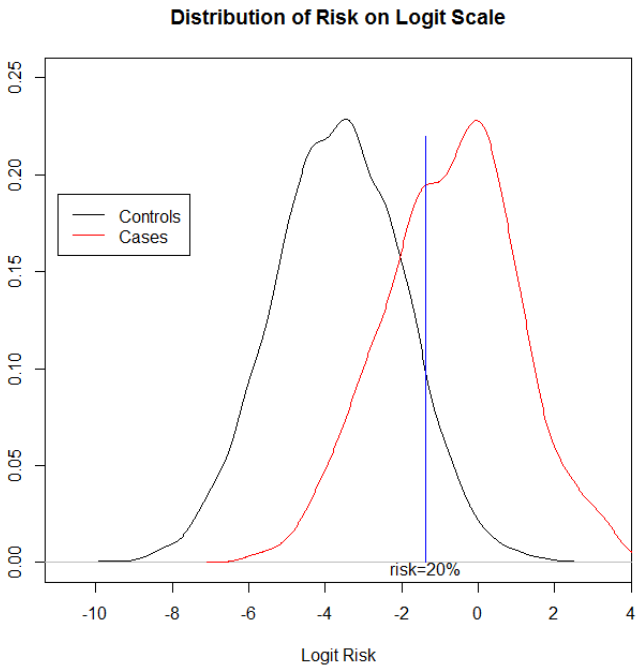
- Generic measures
  - "purely mathematical"
    - meaning: they do not directly translate to any clinical, public health, or public policy impact of using the risk model
- Assessing performance when model will be used to recommend treatment/intervention for high risk individuals
- Assessing performance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







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

MRD, AARD, AUC

- MRD = Mean Risk Difference ≡ mean(risk(X)|case)-mean(risk(X)|control)
- AARD = Above Average Risk Difference ≡ 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**. They attempt to quantify:

How well does the risk model discriminate between (separate) cases and controls?

Mean Risk Difference (MRD)

Also known as

- PEV = Proportion of Explained Variation = R^2 = var(E(D|X))/var(D)=var(risk(X))/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

Above Average Risk Difference (AARD)

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

= 0.797 − 0.198 = 0.599

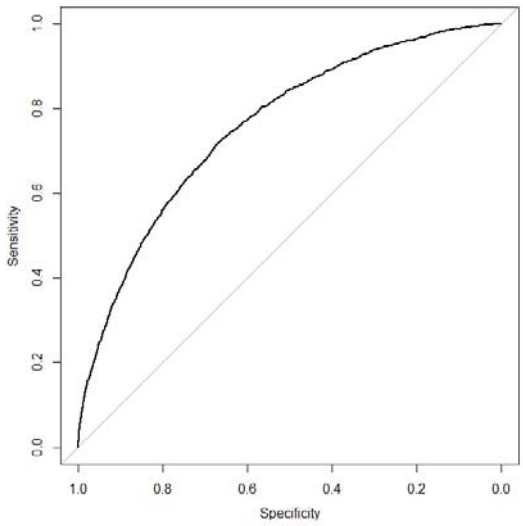
Also known as

- HR_D(ρ) − $HR_{\bar{D}}$(ρ)=TPR(ρ) − FPR (ρ)=Youden's index(ρ)
- RU(ρ)=NB(ρ)/ρ (will come to RU and NB shortly)
  - Proof: NB(r)=ρ HR_D(r)-(1- ρ)$\frac{r}{1-r}$ $HR_{\bar{D}}$(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)

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

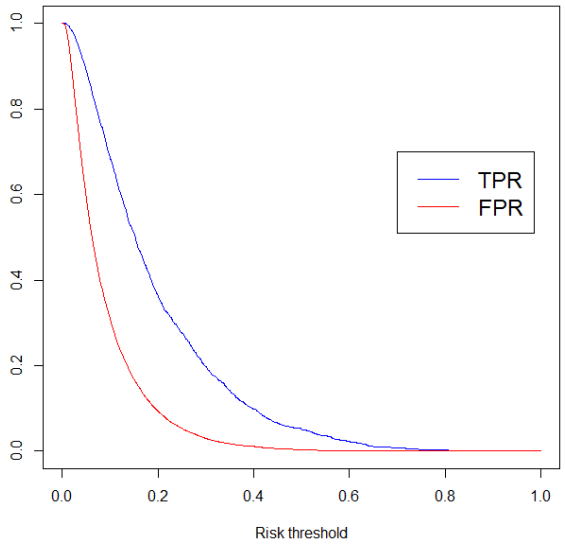
ROC Curve for a risk model



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 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
  - costs include 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

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

Vickers and Elkin, Decision Curve Analysis. *Medical Decision Making* 2006.
Pauker and Kassierer, The threshold approach to clinical decision making. *NEJM* 1980.

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·P(D=1|X)-C·P(D=0|X) > 0
- B·P(D=1|X) > C·P(D=0|X)
- $\frac{\text{P(D=1|X)}}{1-\text{P(D=1|X)}} > \frac{C}{B}$

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 1/4th the benefit of treating a case.

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 → r_H = 0.0076 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(risk(X) > r | D=1)
- Find r such that only f_0 proportion f controls are worked up or treated; f_0 = P(risk(X) > r | D=0)
- Find r such that v_0 proportion of the population is worked up or treated; v_0 = P(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(risk(X)>r_H|D=1)
=% cases in High Risk category
=TPR or sensitivity

$HR_{\bar{D}}(r_H)$ =P(risk(X)>r_H|D=0)
=% controls in High Risk category
=FPR or 1-specificity

Ideally, $HR_D(r_H)$=1 and $HR_{\bar{D}}(r_H)$ =0.

Net Benefit and the Risk Model

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

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

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

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

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

Given a risk threshold r_H that defines "high risk" for treatment recommendation:

Key summary measures:

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

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

$NB(r_H)$ = net benefit of using the model with threshold r_H

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

This expression for NB implicitly assumes r_H has been *rationally selected*, i.e. it corresponds to the benefits and costs of treatment

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%
  - $HR_{\bar{D}}(r_H)$ = 8.9%
  - $NB(r_H)$=0.046 · benefit of statins to subject who would have a CVD event without them

Standardized Net Benefit (Relative Utility)

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) = ρ

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

Standardized Net Benefit ≡ Relative Utility ≡ NB(r_H) / ρ

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

= TPR discounted by an appropriate amount of the FPR

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

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 (and no controls)
- Achieve the same net benefit to the population as 45.5% of cases and no controls called high risk.

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.
- Relative Utility Curves
  - Papers by Baker, e.g. "Putting Risk Prediction in Perspective" Relative Utility Curves." *JNCI*, 2009.

Net Benefit

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

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

which equals

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

- Estimate with:

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

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

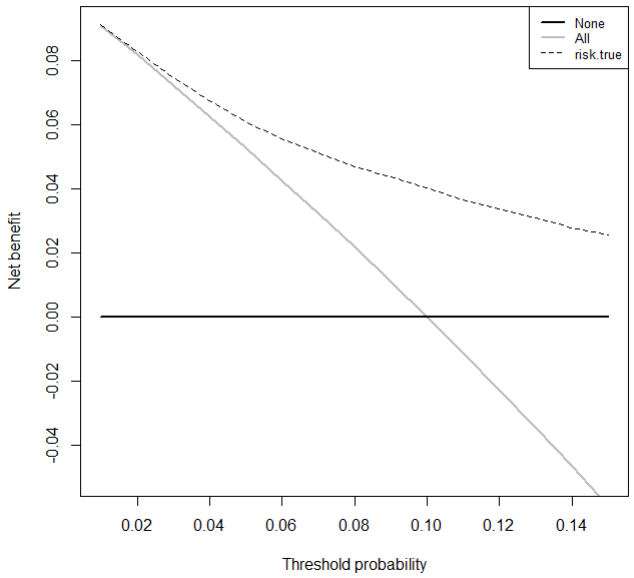
- A **Decision Curve** plots NB against the risk threshold r_H

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

$$risk(X) = P(D \mid X)$$

  - (we don't need to model risk as a function of X)



Understanding the plot

- If the policy is "treat none," then NB is:

$$\text{\# positive cases/20000 - \# positive cntls/20000} \frac{r_H}{1-r_H}$$

$$= 0 - 0 \cdot \frac{r_H}{1-r_H}$$

$$= 0$$

- Therefore the "treat none" policy has NB≡0 for any benefits and costs.

Understanding the plot

- If the policy is "treat all," then NB is:

$$\#\text{ cases}/20000 - \#\text{ cntls}/20000\,\frac{r_H}{1-r_H}$$

$$= \rho - (1-\rho)\cdot\frac{r_H}{1-r_H}$$

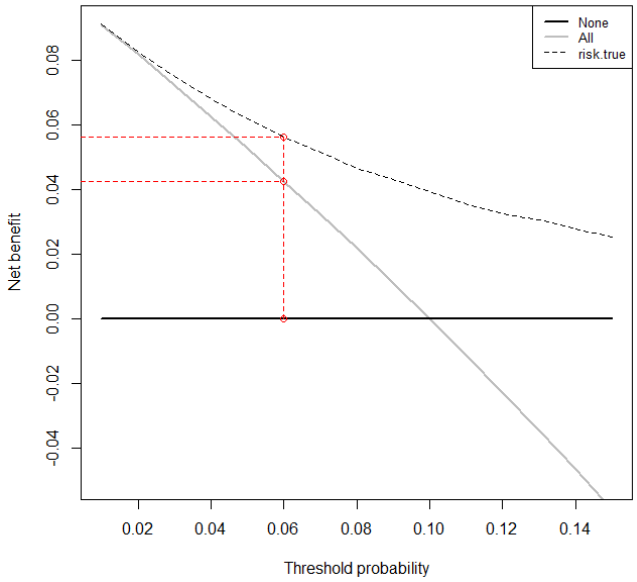
- Even though r_H is not used to decide treatment, it is still used to capture/summarize benefits and costs.
- The curve for "treat all" might look like a straight line, but it isn't.

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:

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

Interpreting the plot



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/ρ = 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/ρ = 43% of cases and no controls .

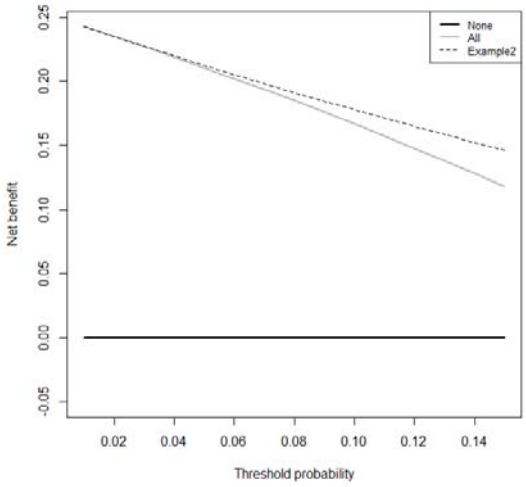
Interpreting the plot

- 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%.

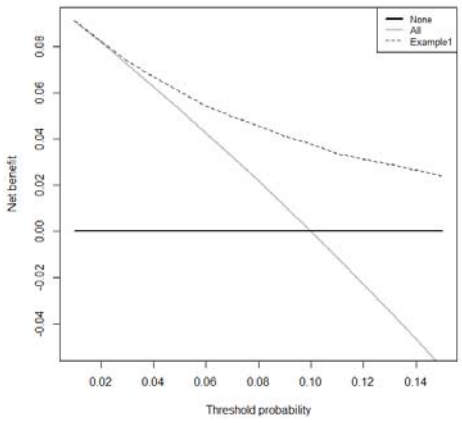
Decision Curve Example 2

- 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

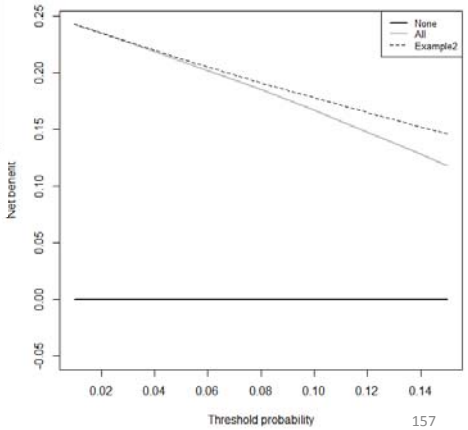
$$risk(X)=P(D|X)$$



It is more difficult for marker-based treatment to "beat" Treat-All when prevalence is high.



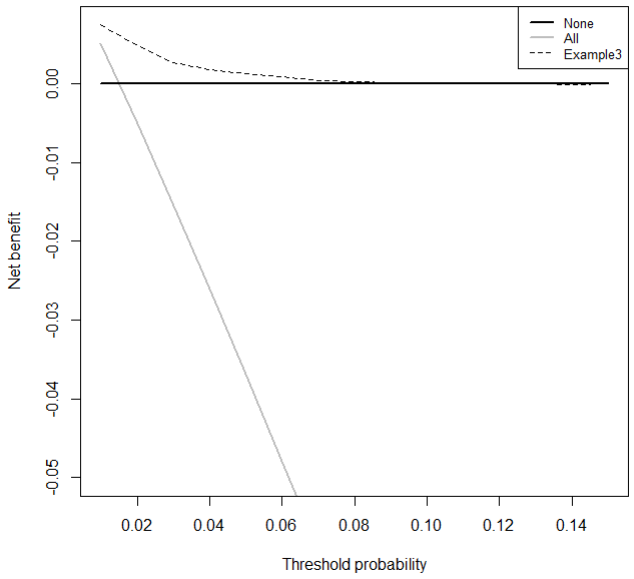
Notice the scale change between the two plots. With higher prevalence there are more Benefits and fewer Costs.



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)



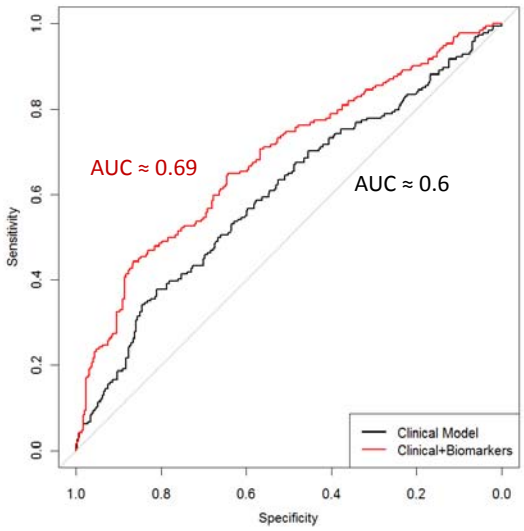
Decision Curve Example 4

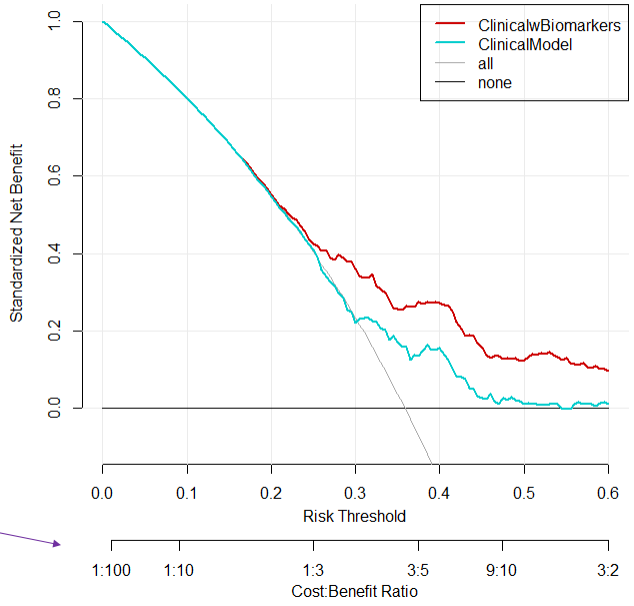
- Prospective study of 570 men scheduled for prostate biopsy.
- New marker: Urinary PCA3 (an RNA that is over-expressed in prostate cancer cells)
- Existing marker: Serum PSA
- Clinical risk factors: age, results of digital rectal exam
- n=541 men, prevalence 36%

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.

ROC Curves





Note this plot uses sNB; thus this is a "Relative Utility Plot"

2nd axis helps interpretation of "all" and "none" curves

Relative Utility Plot

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

- If all cases are high risk and no controls are high risk then NB = prevalence = ρ
  - ρ is the highest possible value for NB, so divide by ρ

    $RU(r_H) \equiv sNB(r_H) = NB(r_H)/\ \rho$

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.

- 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 smaller than 25%.

Notes on risk thresholds

- The curve for "treat all" and "treat none" always cross at the prevalence:
- $NB_{all}=1\cdot\rho - 1\cdot(1-\rho)\cdot\frac{r_H}{1-r_H}=\rho-(1-\rho)\frac{r_H}{1-r_H}=0$ if and only if $r_H=\rho$
- This also implies that in the absence of a risk model or biomarker, the rational choice is "treat all" if $r_H<\rho$ and "treat none" if $r_H>\rho$.

Notes 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.
  - $NB_{treat\text{-}all} < NB_{treat\ none}$, 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 $r_H > \rho$.
  - Some people will only draw the Decision Curve for $r_H > \rho$ when current policy is "treat none."

Notes on risk threshold

- In the absence of a risk model or biomarker, the rational choice is "treat all" if $r_H < \rho$ and "treat none" if $r_H > \rho$.
- The formulation of Decision Curves presented here implicitly uses "treat none" as the reference.
- An alternative formulation uses "treat none" as the reference for $r_H > \rho$ but uses "treat all" as the reference for $r_H < \rho$.

Alternative Formulation

- Use "treat all" as the reference for $r_H < \rho$:
- $NB(r_H) = HR_D(r_H)\,\rho - HR_{\bar{D}}(r_H)\,(1-\rho)\,\frac{r_H}{1-r_H}$

$$-\left[\rho - (1-\rho)\,\frac{r_H}{1-r_H}\right] =$$

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

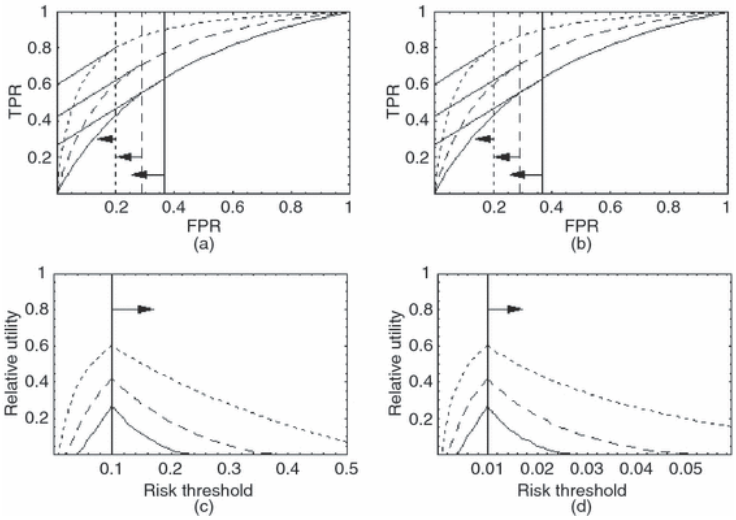
=true negative rate · $(1-\rho)\,\frac{r_H}{1-r_H}$ – false negative rate · ρ

- This formulation leads to "mountain shaped" curves with peak at ρ. Baker *et al* (2009) uses this formulation (next slide)
- Can be viewed as using the marker to "opt out" of treatment rather than to "select in" to treatment.

Baker, Cook, Vickers, & Kramer (2009)

ROC Curves:

Relative Utility Curves:



ρ=10%

ρ=1%

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

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
  - I am calling this "Prognostic Enrichment" following Temple, 2010 (although this term is not widely used)

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

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<π<1, and event rate with intervention 0<τ<1, the sample size SS across the two arms of the trial for a two-sided test is $SS = 2 \times \frac{\left(\phi^{-1}(1-\frac{\alpha}{2})\sqrt{2\left(\frac{\pi+\tau}{2}\right)\left(1-\frac{\pi+\tau}{2}\right)}+\phi^{-1}(1-\beta)\sqrt{\pi(1-\pi)+\tau(1-\tau)}\right)^2}{(\pi-\tau)^2}$,

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 $\phi^{-1}(1-\frac{\alpha}{2})$ 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

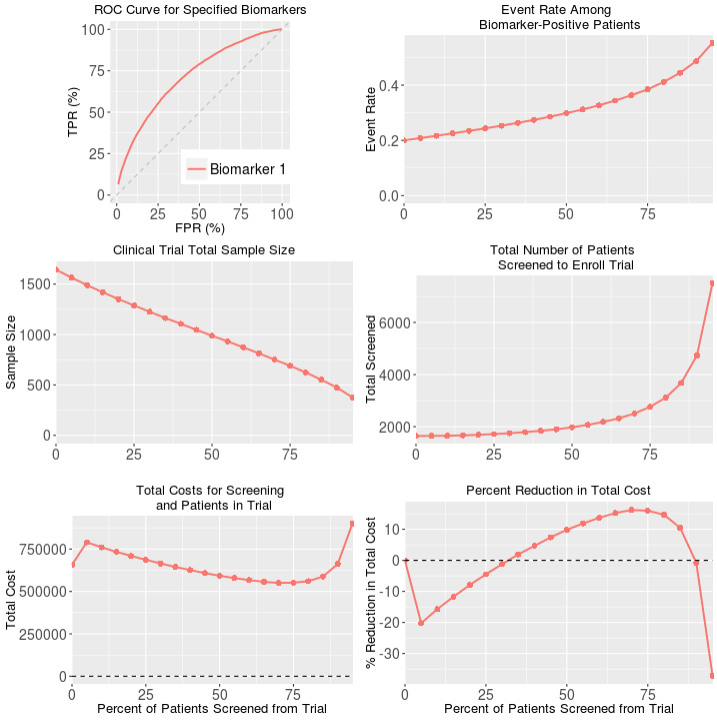
- Let *C1* be the cost of running a patient through the trial and let *C2* be the cost of screening a patient for the trial using the biomarker
- Total Cost with screening threshold *t* is

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

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

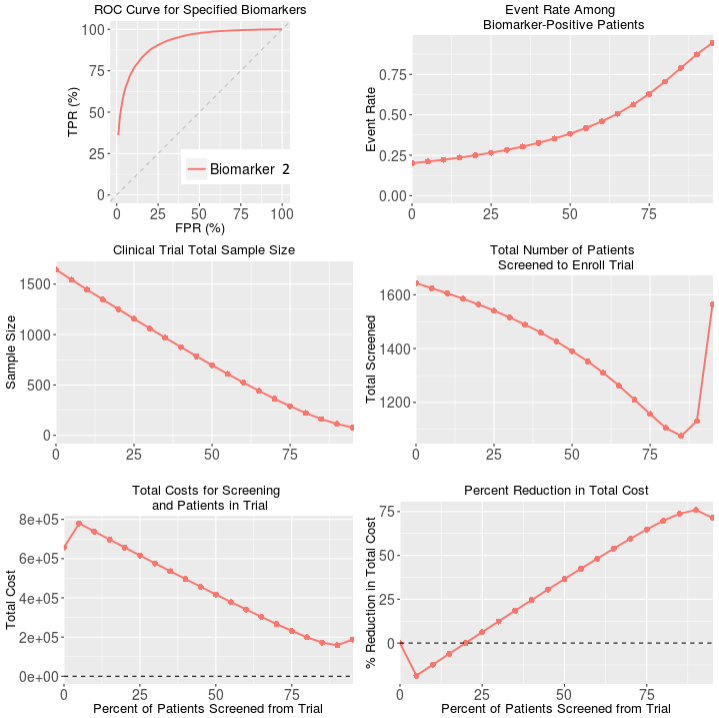
Prognostic Biomarker 1

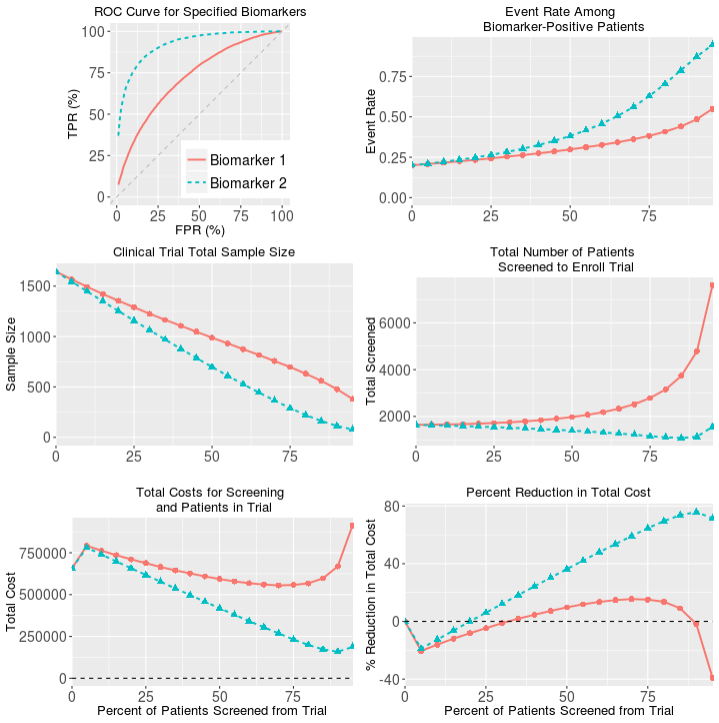
- Event rate without prognostic enrichment: 20%
- AUC of biomarker: 0.72
- Cost to measure biomarker: $100
- Cost to run one patient through trial: $400
- Specifying trial design to have 90% power to detect a 30% reduction in event rate using α= 0.025 with one-sided testing



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 α= 0.025 with one-sided testing





184

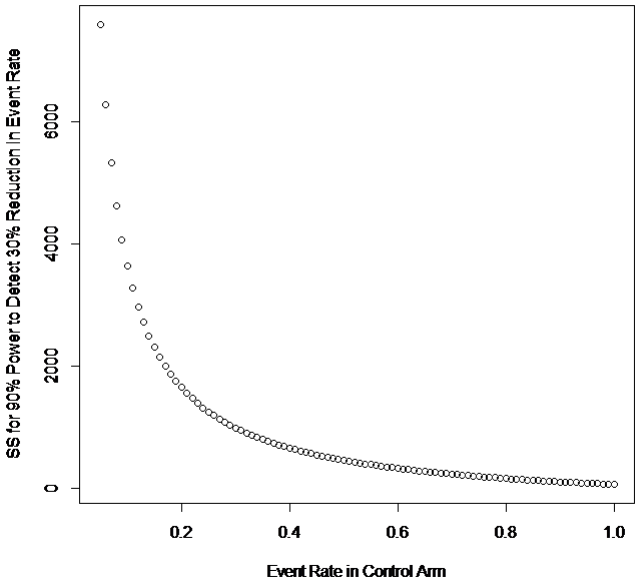
Prognostic Enrichment – Other Important Considerations

- 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.

185

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 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 (but usually not)
- 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 Part II

- Decision Curves
  - Potentially useful to evaluate a risk model over a range of plausible risk thresholds
    - Challenging to interpret values of NB
  - Aids the assessment of the population impact of treatment policies
- Relative Utility Plots
  - Similar to Decision Curves, but RU perhaps easier to interpret
  - Maximum RU always 1.0 (or 100%)

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