Current Methods for Evaluating Prediction Performance of Biomarkers and Tests

Instructor: Margaret Pepe, University of Washington and Fred Hutchinson Cancer Research Center

Abstract. This course will describe and critique methods for evaluating the performance of markers to predict risk of a current or future clinical outcome. Examples from cancer, cardiovascular disease and kidney injury research will be presented. Course content will include:

a. Four criteria for evaluating a risk model: calibration, benefit for decision making, accurate classification,risk stratification;
b. Comparing models;
c. Comparing baseline and expanded models;
e. Hypothesis testing and confidence interval construction;
f. Relationships and differences with assessing discrimination;
g. Software.

Evaluating Prediction Performance of Biomarkers and Tests

Margaret Sullivan Pepe
Fred Hutchinson Cancer Research Center
and
University of Washington

email: mspepe@u.washington.edu
pertinent website: labs.fhcrc.org/pepe/dabs/index.html

FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE

UNIVERSITY OF WASHINGTON

Examples

1. Acute kidney injury after cardiac surgery[1]

- TRIBE-AKI consortium study
- $n = 1139$ adults, $n_C = 407$ (36%) events
- $Y =$ preoperative BNP
- $X =$ demographics, co-morbidities, surgery characteristics
- $D =$ AKI event

Patel et al. (submitted) *Circulation*

Examples

2. Critical illness assessment in out-of hospital emergency care[2]

- King County EMS 2002–2006
- $n =$145,000 without trauma or cardiac arrest
- $n_C =$ 8,000 (5%) sepsis events
- $Y =$ whole blood lactate
- $X =$ demographics, clinical measurements (blood pressure, oximetry, respiratory rate, ...), location (nursing home, ...)
- $D =$ sepsis

Seymour C. *JAMA* 2010

Examples

3. Invasive breast cancer within 5 years after treatment for DCIS[3].

- Early Detection Research Network (EDRN) ongoing multicenter study.
- One site: Western Washington population-based study.
- nested case-control design (n_C =445, n_N =890).
- Y = DCIS tissue biomarkers TBD.
- X = age, race, mammographic density, treatment
- D = invasive breast cancer within 5 years

Examples

4. 10-year risk of cardiovascular health event[4].

- Framingham study
- n =3264
- n_C =183
- Y =HDL
- X = demographics, smoking, diabetes, blood pressure, total cholesterol
- D = CVD event

Pencina *Stat in Med* 2008

Examples

5. Simulated Data available on DABS website[5].

- n =10,000
- n_C =1,017
- Y =continuous
- X = one-dimensional continuous
- For practice and illustration

Pepe *AJE* 2011

Outline

Unit 1: Evaluating a risk model — the basic concepts

Unit 2: Comparing two risk models — including the role of risk reclassification metrics

Unit 3: Estimation and inference from data

Unit 1

Evaluating a Risk Model – Basic Concepts

Notation

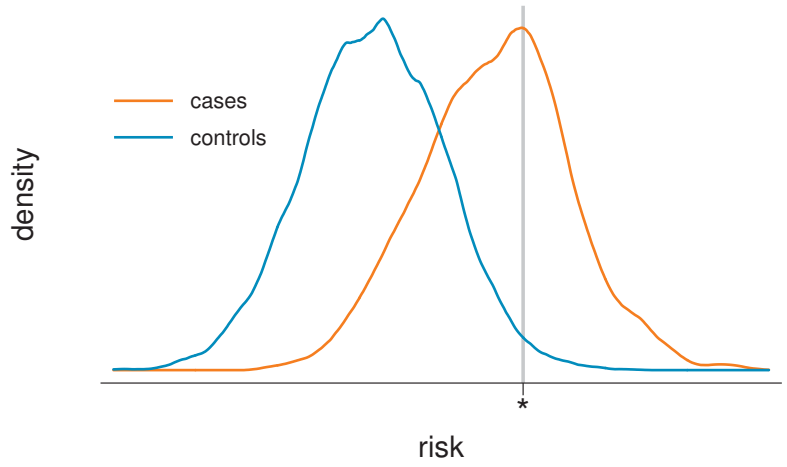
- $D = 1$ bad outcome (case, event)
  $D = 0$ good outcome (control, non-event)
- $X =$ predictors
- $r(X) = \text{Prob}(D = 1|X) = \mathit{risk}(X)$

What's the point of developing a risk model?

- To help make medical decisions.
- Offer new interventions particularly to those who might benefit (cases: bad outcome in absence of intervention).
- Offer no new intervention 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.

Fundamental Evaluations

1. Calibration — is the risk calculator correctly calculating risk?
2. Risk distributions and consequent benefit — is the risk calculator helpful to cases and controls in making appropriate medical decisions?



* threshold for intervention

(Unit 1 will expand on these basic ideas)

What is risk(X)?

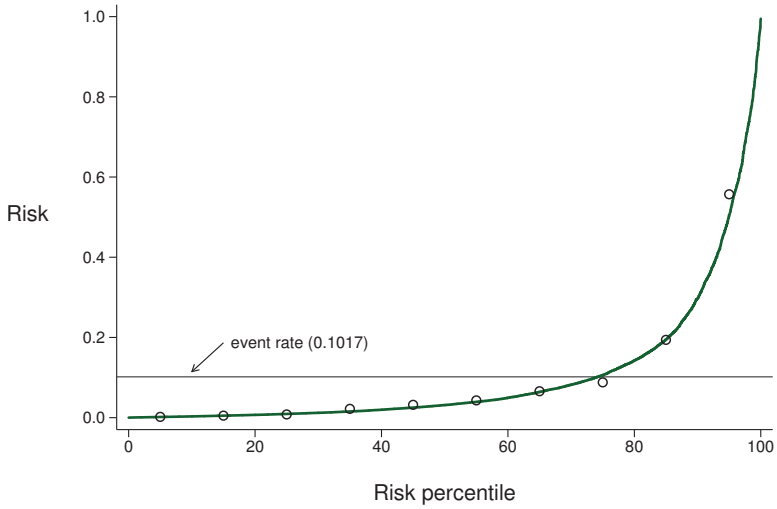
- *Not* individual level probability.
- $risk(X) = P[D = 1|X = x]$ is a *frequency* of events among the group of subjects with $X = x$.
- Individual level risks are not observable[5].
- Individual level risks are not well defined.
- (more literature on this?)

Calibration

- Is the risk calculator $r(X)$ valid?
- Among people with $r(X) = r$ is the fraction of events $\approx r$?[6]
- More lenient than asking if $r(X) = P(D = 1|X)$. We are asking if $P(D = 1|r(X) = r) = r$.
- Validity of the risk calculator is crucial.
- Otherwise, we are engaged in evaluating a score for discrimination/classification, not with the higher level task of evaluating risk prediction performance.

Calibration

- Visual assessment with the predictiveness curve[7] or calibration plot.[6]



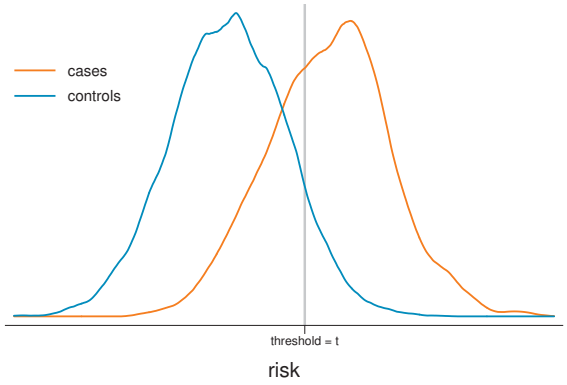
- Flexibility in choosing quantile categories on abscissa.
- Hosmer-Lemshow statistic[8] sometimes a useful adjunct, but depends on sample sizes.

Calibration

- Achievable if the number of predictors is small.
- Recalibration[9,10]:
  - (i) training set develops a score $X(W)$ where W are predictors;
  - (ii) test set for estimating $r(X) \equiv P(D = 1|X)$.
- If $P(D = 1|X) \neq P(D = 1|W)$ we lose optimality but gain 'validity.'
- Assume henceforth that risk calculators are valid, i.e.,well calibrated, in the sense that $P(D = 1|r(X) = r) \approx r$.
- some measures of performance only make sense when models are valid, others make sense regardless.

Prediction Performance

- Case (C) and Control (N) risk distributions are fundamental components of all performance measures.



- Risk categories corresponding to treatment recommendations can be overlaid to assess benefit for decision making.
- Suppose scenario of no treatment (default) versus a single treatment.
- HR_C =% cases in High Risk category: potential benefit of using $r(X)$.
- HR_N =% controls at High Risk category: potential cost of using $r(X)$.

Existing Risk Categories

- Prob[Breast Cancer] $\geq 0.8\%$ per year for tamoxifen treatment in white women 50–59 years old.[11]
- Prob[CHD event in 10 years] $\geq 20\%$ for cholesterol lowering medication (ATPIII guidelines).[12]

When Risk Categories Don't Exist

- Displaying distributions of risk allows one to overlay subjective risk categories.
- "What if?" exercises about future interventions.

Given a risk threshold t that determines recommendation for treatment:

Key summary statistics

- $HR_C(t) = P(r(X) > t|D = 1)$ ideally $HR_C(t) = 1$
- $HR_N(t) = P(r(X) > t|D = 0)$ ideally $HR_N(t) = 0$
- $NB(t) =$ net benefit of using the model and t for subjects in the population
  $= B \times P(D = 1)HR_C(t) - Cost \times P(D = 0)HR_N(t)$

where $B =$ expected benefit of treatment to a case and $Cost =$ expected cost of treatment to a control.

Notes: $t > P[D = 1]$ is only relevant when default is "no treatment."[13]

TPR$= HR_C(t)$
FPR$= HR_N(t)$

Result 1

$$Cost/B = t/(1 - t)$$

Proof

When should a patient choose treatment?

- when he/she expects to benefit
- $E(benefit|D = 1, X)P(D = 1|X) - E(cost|D = 0, X)P(D = 0|X) > 0$
- $B \times P(D = 1|X) - Cost \times P(D = 0|X) > 0$
- $P(D = 1|X)/P(D = 0|X) > Cost/B$

- a rational choice of risk threshold implies $Cost/B$
- specified $Cost/B$ implies a rational choice of threshold.

Reference: 14, 29

Example

20% risk threshold for treatment 'feels right' equivalent to

$$\text{perceived cost-benefit ratio } = \frac{t}{1-t} = \frac{.2}{.8} = .25$$

Example

Gail[11] evaluated risk models for breast cancer in terms of decisions about tamoxifien (T) prevention therapy. Values relate to Prob(bad outcome) where a bad outcome includes hip fracture, stroke, endometrial cancer, pulmonary embolism in addition to breast cancer. In 50–59 year old white women:

$$Cost/B = 0.0077 \rightarrow t = 0.0077 \text{ per year.}$$

Net Benefit (t) in the population

- $$\begin{aligned} NB(t) &= B \times P(D=1)HR_C(t) - Cost \times P(D=0)HR_N(t) \\ &= B\left\{P(D=1)HR_C(t) - \frac{t}{1-t}P(D=0)HR_N(t)\right\} \\ &= P(D=1)HR_C(t) - \frac{t}{1-t}P(D=0)HR_N(t) \end{aligned}$$

where B is the unit of measurement $\equiv$ benefit of intervention for a would be case in the absence of intervention.

- $1/NB(t) =$ number of tests needed to yield a NB of 1
  - ignoring costs of testing
  - $NB = 1$ is equivalent to 1 case and no controls classified as high risk.

References: 13, 14

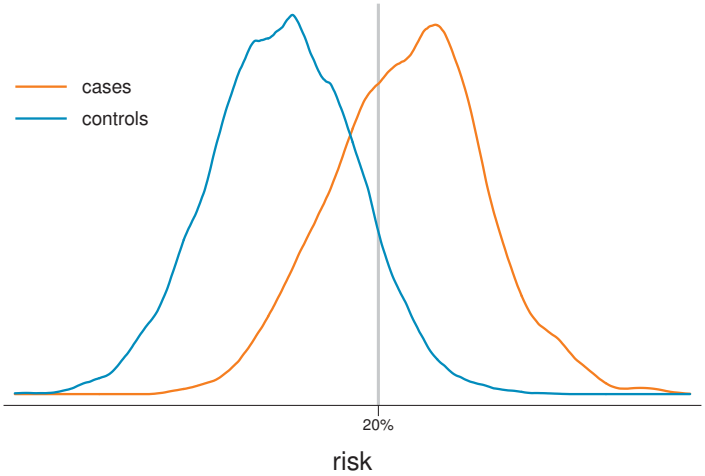
Given a risk threshold t that determines recommendation for treatment:

Key summary statistics

- $HR_C(t) = P(r(X) > t|D = 1)$ ideally 1
- $HR_N(t) = P(r(X) > t|D = 0)$ ideally 0
- $NB(t) =$ net benefit
  $= P(D = 1)HR_C(t) - \frac{t}{1-t}P(D = 0)HR_N(t)$

Example

- $t = 20\%$; predictor X; data in [5]
- $HR_C(t) = 65.2\%$
- $HR_n(t) = 8.9\%$
- $NB(t) = 0.0465\times$ benefit of statins to subjects who would get a CVD event without them.



Understanding Net Benefit

$$NB(t) = P(D=1)HR_C(t) - \frac{t}{1-t}P(D=0)HR_N(t)$$

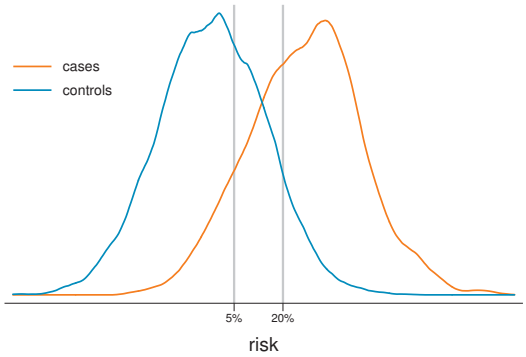
- $\rho \equiv P(D=1)$
- maximum value of $NB = P(D=1)$
  'The best we can do is treat all cases and no controls'
- relative utility[13] $= NB(t)/\rho =$ % of maximum benefit
- relative utility $= HR_C(t) - \frac{1-\rho}{\rho}\frac{t}{(1-t)}HR_N(t)$

  $=$ true positive rate discounted appropriately for the false positive rate.

In our data

- $RU = 45.6\%$
- The maximum possible benefit is to treat all 1017 cases and no controls per 10,000. We can achieve **45.6%** of this benefit using the prediction model.
- With this model 65.2% of cases are above the treatment risk threshold. Discounting for controls also classified as high risk, we achieve the equivalent of **45.6%** of cases classified for treatment.

More than Two Risk Categories



- Default = medium category intervention since $P[D = 1] = 0.102$

	low	med	high
	$< 5\%$	5-20%	$> 20\%$
Cases %	11.2	23.6	65.2
Control%	65.7	25.3	8.9

- To calculate net benefit here, we also need to specify $B_{\text{high}}/Cost_{\text{low}}$ for a case.
- e.g., $B_{\text{high}}/Cost_{\text{low}} = 10$: $NB = 0.0485 \times B_{\text{high}}$
  $NB_{\max} = 0.1017 + 0.0047 = 0.1064$
  $RU = 45.6\%$

FYRP

- Use C to denote Cost here

$$NB = \rho\{-C(\text{low})P_C(\text{low}) + B(\text{high})P_C(\text{high})\} + (1-\rho)\{B(\text{low})P_N(\text{low}) - C(\text{high})P_N(\text{high})\}$$

$$\text{Arguments in Result 1} \Rightarrow \frac{B_{\text{low}}}{C_{\text{low}}} = \frac{t_{\text{low}}}{1-t_{\text{low}}} \quad \frac{C_{\text{high}}}{B_{\text{high}}} = \frac{t_{\text{high}}}{1-t_{\text{high}}}$$

Let $\lambda = B_{\text{high}}/C_{\text{low}}$

$$NB = \rho\{-B_{\text{high}}/\lambda P_C(\text{low}) + B_{\text{high}}P_C(\text{high})\} + (1-\rho)\left\{\frac{t_{lo}}{1-t_{\text{low}}}\frac{B_{\text{high}}}{\lambda}P_N(\text{low}) - B_{\text{high}}\frac{t_{hi}}{1-t_{hi}}P_N(\text{high})\right\}$$

$$= B_{\text{high}}\left\{-\frac{\rho}{\lambda}P_C(\text{low}) + \rho P_C(\text{high}) + \frac{t_{\text{low}}}{1-t_{\text{low}}}\frac{(1-\rho)}{\lambda}P_N(\text{low}) - \frac{t_{\text{high}}}{1-t_{\text{high}}}(1-\rho)P_N(\text{high})\right\}$$

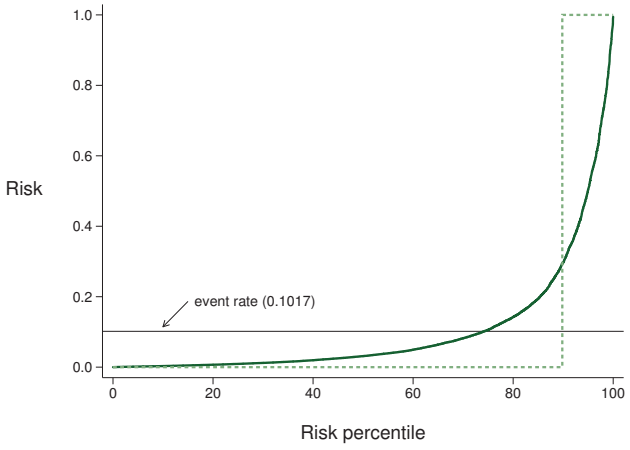
If $\lambda = 10 \Rightarrow NB = 0.0485$.

In the absence of risk categories or thresholds:

- Plots: $HR_C(t)$, $HR_N(t)$, $NB(t)$ or $RU(t)$
- Summary statistics
  - But first, about the predictiveness curve . . . .

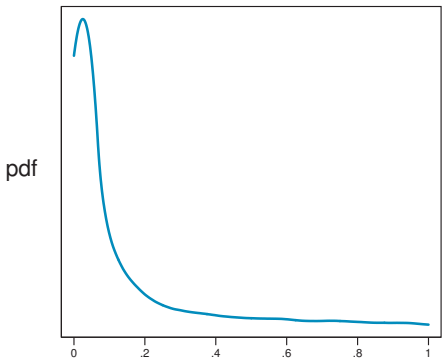
The Predictiveness Curve

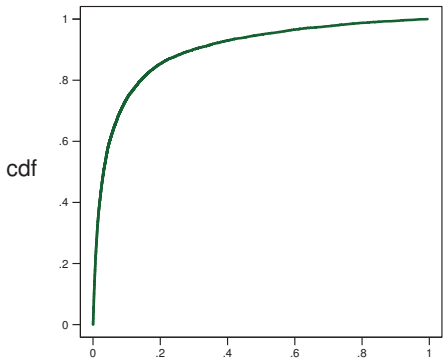
- Plots risk quantiles in the whole population.[5,15]
- Shows the population risk distribution, cases and controls together
- Shows what % of subjects in the population will be classified as high risk by the model − useful for enrollment in clinical trials for example.
- Total Gain = area between curve and horizontal (useless model) curve[16]



The Predictiveness Curve

- There are other ways of showing the population risk distribution.[17]





The Predictiveness Curve

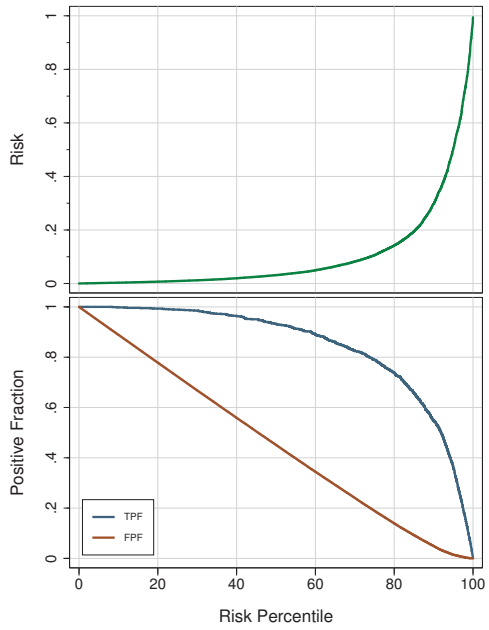
- Is equivalent to case and control risk distributions.[18]
- Is derived from case and control distributions.

$$P(risk(X) > t) = \rho P(risk(X) > t|D = 1) + (1-\rho)P(risk(X) > t|D = 0)$$

- Can be used to calculate case and control distributions.

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

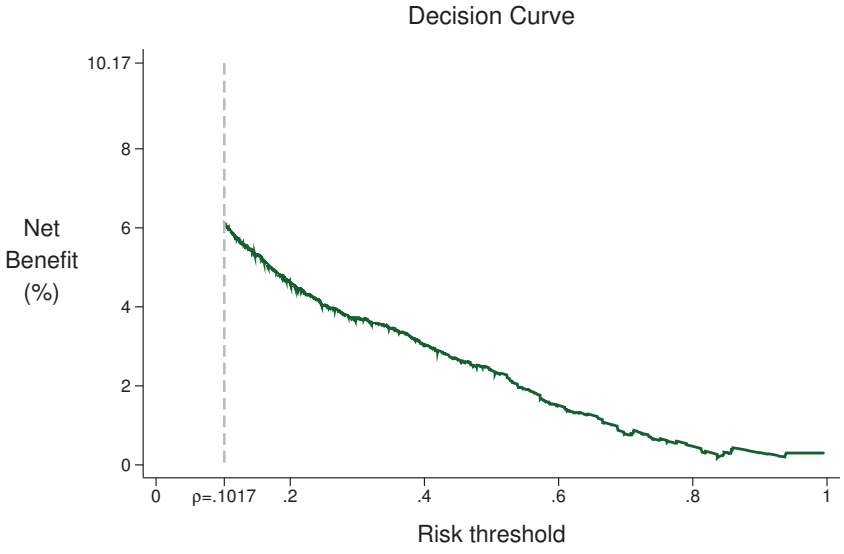
$$= rP(risk(X) = r)/P(D)$$



- The *integrated* plot[7] (2008)
- I have come to think that the lower panel of this plot is most important.
- Should be supplemented with the net benefit (decision) or relative utility curve.

Decision Curve

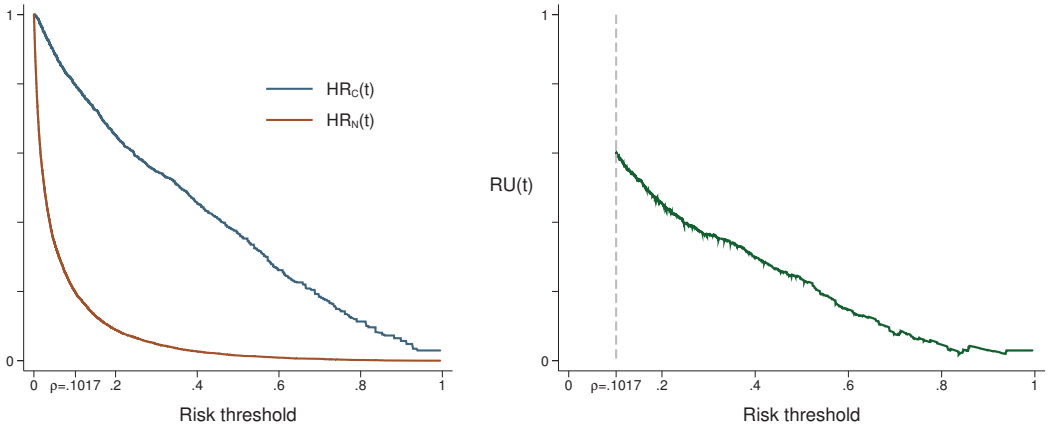
Decision Curve[14]: $NB(t) = \rho HR_C(t) - (1-\rho)\{t/(1-t)\}HR_N(t)$



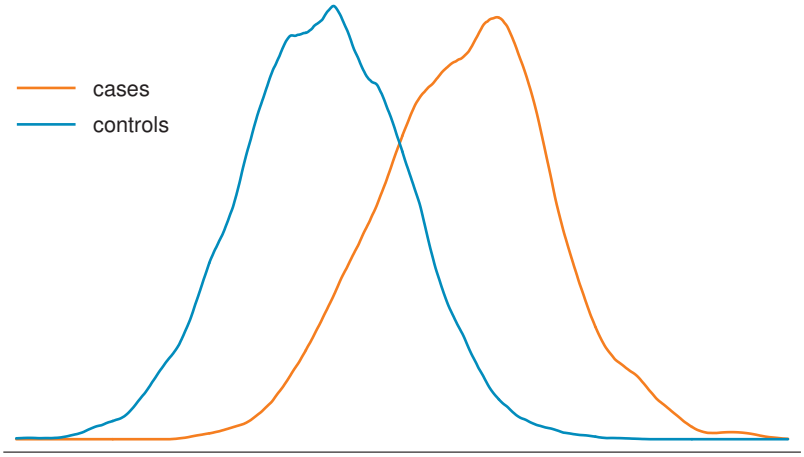
- Considering B fixed, larger risk threshold (t) corresponds to larger *Cost*.
- Larger *Cost* implies smaller *NB* (decreasing $NB(t)$).
- Relative utility curve $= NB(t)$/maximum possible $NB(t)$
  $= NB(t)/\rho$ on (0,1) scale
- Recall default $=$ no treatment here.
- Relative utility defined more generally.[13]

$\rho = 0.1017$

Data Displays without Risk Categories



Summary Measures without Risk Categories



- MRD = Mean Risk Difference $\equiv$ mean($risk(X)$|case)-mean($risk(X)$|control)
- $AARD$ = Above Average Risk Difference $\equiv P(risk(X) > \rho|\text{case}) - P(risk(X) > \rho|\text{control})$

Mean Risk Difference (MRD)

Also known as

- IDI = Integrated Discrimination Improvement relative to no model[4,19]
  $= \int \{P(risk(X) > r|D = 1) - P(risk(X) > r|D = 0)\}\, dr$
- PEV = Proportion of Explained Variation
  $= \text{var}(E(D|X))/\text{var}(D)$
  $= \text{var}(risk(X))/\text{var}(D)$
- R^2 = PEV (There are other more complex R^2measures[18])
- Yates' Slope

For our data

- mean($risk$|case) = 0.391
- mean($risk$|control) = 0.069
- MRD= .322

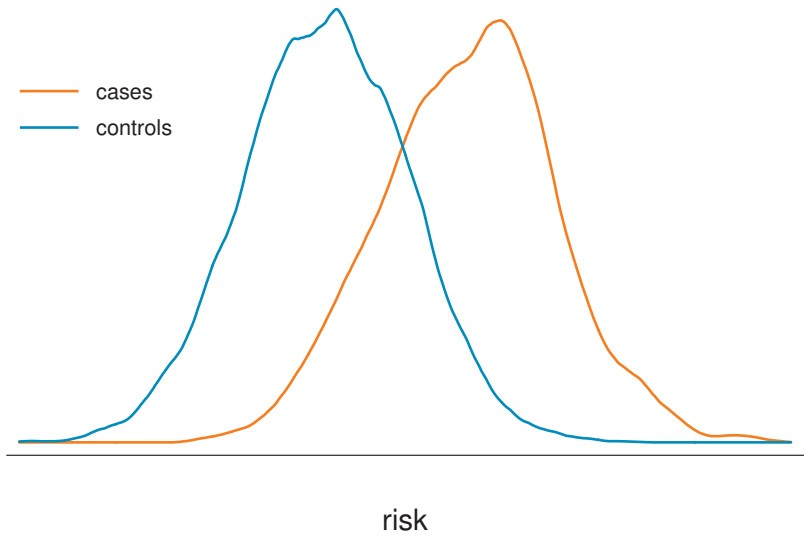
Above Average Risk Difference (AARD)

$AARD = P(risk(X) > \rho | D = 1) - P(risk(X) > \rho | D = 0) = 0.797 - 0.198 = .599$

Also known as

- $HR_C(\rho) - HR_N(\rho) = TPR(\rho) - FPR(\rho) =$ Youden's index (ρ)
- $RU(\rho) = NB(\rho)/\rho$
  Proof: $NB(t) = \rho HR_C(t) - (1-\rho)\frac{t}{1-t} HR_N(t)$
- Standardized Total Gain $\equiv TG/2\rho(1-\rho)$. 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 $< \rho$.

Summary Measures without Risk Categories

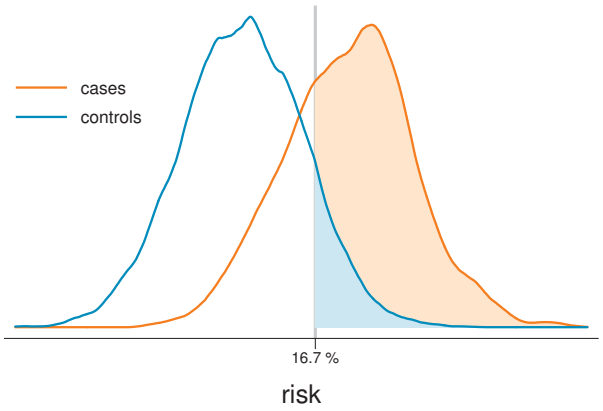


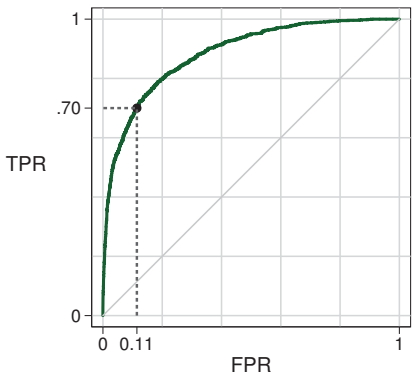
MRD = 0.322

$AARD$ = 0.599

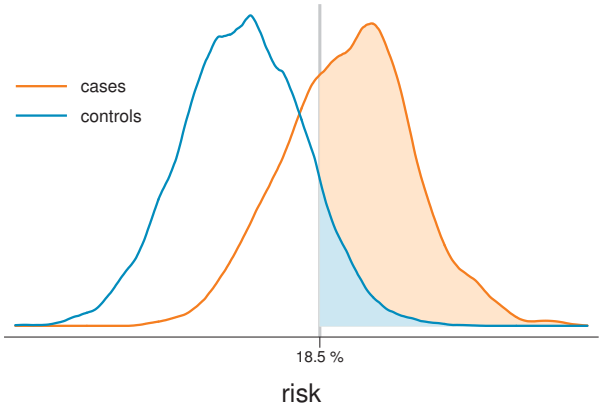
ROC Type Statistics as Summary Measures

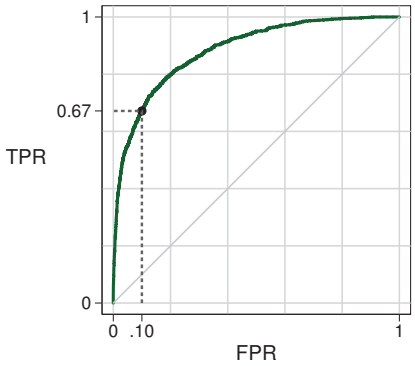
- May be useful when no clinically relevant risk thresholds exist.
- $\text{ROC}(f_0) = P(risk(X) > t|D = 1)$ where t: $f_0 = P(risk(X) > t|D = 0)$
- $\text{ROC}^{-1}(t_0) = P(risk(X) > t|D = 0)$ where t: $t_0 = P(risk(X) > t|D = 1)$
- $PCF = \mathcal{L}(v_0) = P(risk(X) > t|D = 1)$ where t: $v_0 = P(risk(X) > t)$
- $PNF = \mathcal{L}^{-1}(w_0) = P(risk(X) > t)$ where t: $w_0 = P(risk(X) > t|D = 1)$
  $= \rho w_0 + (1 - \rho) ROC^{-1}(w_0)$
- $\mathcal{L}$ is related to the Lorenz curve.[22]
- Report the risk threshold corresponding to the criterion as well.





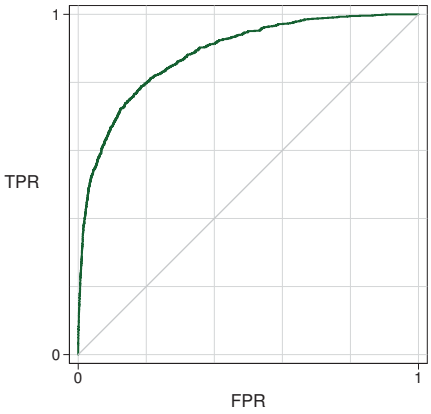
$\text{ROC}^{-1}(0.7) = 0.115$
$\mathcal{L}(0.7) = 0.175$





$\text{ROC}(0.1) = 0.672$
$\mathcal{L}^{-1}(0.15) = 0.656$

ROC Plot is Inadequate



- Because it does not display absolute risk scale.[7]
- Preferable to display two risk distributions rather than just the nonparametric distance between them.

Area Under the ROC Curve = 0.884

- Mann-Whitney U-statistic (Wilcoxon) for comparing risk distributions for cases versus controls.
- Average TPR over all possible FPR
- $P(risk(X_C) > risk(X_N))$
- Not a clinically relevant measure of prediction performance
- Not worse than MRD in terms of clinical relevance (is it?).
- ROC(f_0) or other point measures may be more clinically relevant than AUC.
- AUC gives too much weight to low risk ranges and ignores the meaning of risk.[23]
- Average TPR over a relevant range of FPR preferable to total AUC=partial AUC(f)/$f = 0.510$ at $f = 0.10$.

Summary of Unit #1

- The statistical meaning of $risk(X)$: an observable frequency of events.
- Calibration≡ "is the risk calculator valid?" is crucial.
- Medical decisions based on risk:
  - correspondence between risk categories and $cost(FP)/benefit(TP)$
  - case and control frequencies in risk categories are the essential components of net benefit gained by using risk model (excluding cost of testing here)
- Without risk categories
  - display case and control risk distributions, and relative utility (or decision curves) that show benefit of using the model under various decision rules
  - statistics: ROC or Lorenz points
  - more statistics: MRD, $AARD$, AUC; are they useful for comparing models?

Unit 2

Comparing Risk Models

In a Nutshell

- Choose your favorite measure(s) of prediction performance.
- Compare that measure(s) for the two risk models.

Example

- $r(X, Y)$ versus $r(X)$ for data in Pepe *AJE* 2011.
- Both models are very well calibrated in the weak sense[18] that $P(D = 1|r) \approx r$.

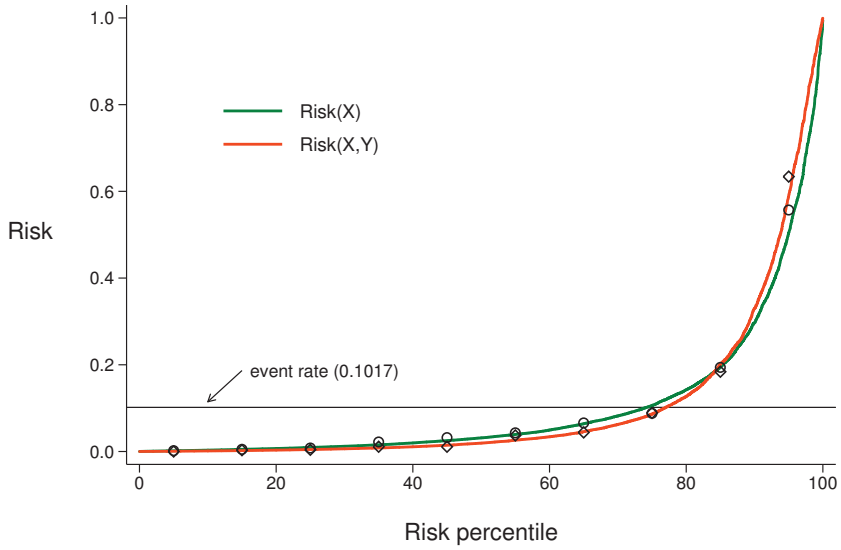
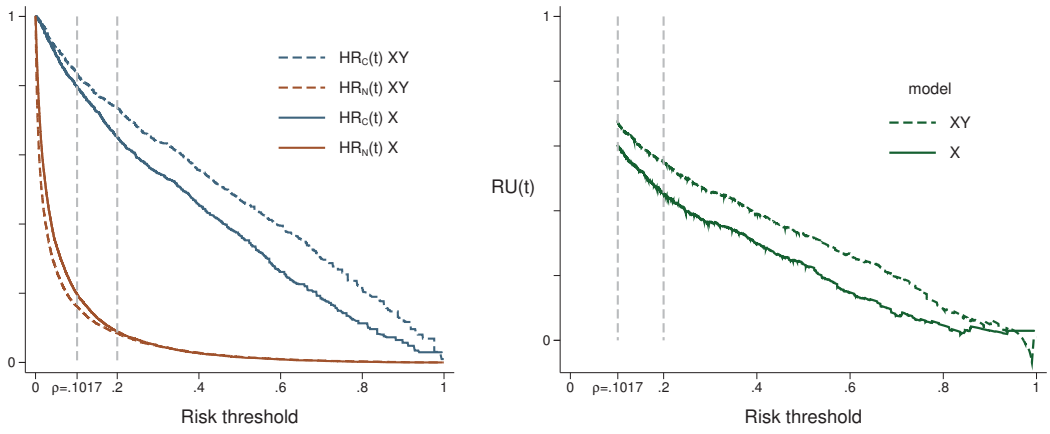


Figure 2. Plots showing high risk classification for cases (C) and controls (N) under the baseline model (X) and the expanded model (X, Y). A comparison of net benefit is also shown through the relative utility plot.

$\rho = 0.1017$



My Favorite Summary Measures

		$t = 20\%, \rho = .1017$		
		$risk(X)$	$risk(X, Y)$	Δ
Cases $> t$	$HR_C(t)$	65.2%	73.5%	8.4%
Controls $> t$	$HR_N(t)$	8.9%	8.4%	-0.5%
% of max benefit	$RU(t)$	45.5%	55.0%	9.5%

- Hypothesis testing and confidence intervals in Unit 3.

Not My Favorite Summary Measures

	$risk(X)$	$risk(X, Y)$	Difference
AUC	0.884	0.920	0.036
MRD/IDI	0.322	0.416	0.094
$AARD$	0.599	0.673	0.074
$ROC(0.20)$	0.672	0.758	0.087
$\mathcal{L}^{-1}(0.70)$	0.174	0.134	-0.040

- Hypothesis testing and confidence intervals in Unit 3.

Risk Reclassification Tables

- Showing numbers of subjects and event rates (%)

	$risk(X, Y)$			
$r(X)$	$\leq 5\%$	$5 - 20\%$	$> 20\%$	Total
$\leq 5\%$	5558	437	25	6020
	1.30	8.71	16.00	1.89
$5 - 20\%$	1036	1095	386	2517
	2.03	9.59	29.53	9.54
$> 20\%$	40	329	1094	1463
	0.00	10.03	57.59	45.32
Total	6634	1861	1505	10,000
	1.40	9.46	49.70	10.17

- Cook and Ridker[23,24]
- Motivated by clinically relevant risk categories in cardiovascular health research.

Event and non-Event Risk Reclassification Tables[4]

Events

	$risk(X, Y)$			
$r(X)$	$< 5\%$	$5 - 20\%$	$\geq 20\%$	Total
$\leq 5\%$	72	38	4	114 (11.2%)
$5 - 20\%$	21	105	114	240 (23.6%)
$\geq 20\%$	0	33	630	663 (65.2%)
Total	93	176	748	1017
	(9.1%)	(17.3%)	(73.5%)	

Event and non-Event Risk Reclassification Tables

Non-Events

	$risk(X, Y)$			
$r(X)$	$< 5\%$	$5 - 20\%$	$\geq 20\%$	Total
$\leq 5\%$	5486	399	21	5906 (65.8%)
$5 - 20\%$	1015	990	272	2277 (25.3%)
$\geq 20\%$	40	296	464	800 (8.9%)
Total	6541	1685	757	8983
	(72.8%)	(18.8%)	(8.4%)	

- **Useful to see how *r(X,Y)* performs within strata defined by *r(X)***

	$\rho(X)$	cases	controls	% of max benefit
Population	Event Rate	$HR_C(0.20)$	$HR_N(0.20)$	$RU(0.20)$
low risk $r(X)$	1.89%	0.035	0.004	-1.7%
med risk $r(X)$	9.54%	0.475	0.119	19.3%
high risk $r(X)$	45.32%	0.950	0.580	25.4%

$*$ Assume default is 'no treatment' for low and medium risk strata.

$$RU(t) = HR_C(t) - \frac{(1-\rho(X))}{\rho(X)} \frac{t}{1-t} HR_N(t)$$

$*$ Assume default is 'treatment' for high risk stratum.

$$RU(t) = (1 - HR_N(t)) - \frac{\rho(X)}{1-\rho(X))} \frac{1-t}{t} (1 - HR_C(t))$$

Population	$\rho(X)$ Event Rate	cases $HR_C(0.20)$	controls $HR_N(0.20)$	% of max benefit $RU^*(0.20)$
low risk $r(X)$				
med risk $r(X)$	9.54%	0.475	0.119	19.3%
high risk $r(X)$ %				

- Calculations based on middle rows of event and non-event reclassification tables.

Population	$\rho(X)$ Event Rate	cases $HR_C(0.20)$	controls $HR_N(0.20)$	% of max benefit $RU^*(0.20)$
low risk $r(X)$				
med risk $r(X)$				
high risk $r(X)$	45.32%	0.950	0.580	25.4%

$*$ Assume default is 'treatment' for high risk stratum.

$$RU(t) = (1 - HR_N(t)) - \frac{\rho(X)}{1 - \rho(X))} \frac{1 - t}{t} (1 - HR_C(t))$$

Population	$\rho(X)$ Event Rate	cases $HR_C(0.20)$	controls $HR_N(0.20)$	% of max benefit $RU = \frac{NB}{\rho}(0.20)$
low risk $r(X)$	1.89%	0.035	0.004	-1.7%
med risk $r(X)$				
high risk $r(X)$				

Risk Reclassification Tables

- Valuable for evaluating $r(X, Y)$ within subpopulations defined by $r(X)$
- These are separate single model evaluations in each subpopulation.

Risk Reclassification Tables

- Not helpful for overall comparison of use of $r(X)$ versus use of $r(X, Y)$ in the entire population.
- Margins show risk distributions for each model. Performance for each model determined by the corresponding marginal distributions of risk.
- Comparison of performance determined by comparison of marginal distributions.
- Reclassification statistics do not calculate performance for each model and compare the two values.

Event and non-Event Risk Reclassification Tables

Events

	$risk(X, Y)$			
$r(X)$	$< 5\%$	$5 - 20\%$	$\geq 20\%$	Total
$\leq 5\%$	72	38	4	114 (11.2%)
$5 - 20\%$	21	105	114	240 (23.6%)
$\geq 20\%$	0	33	630	663 (65.2%)
Total	93	176	748	1017
	(9.1%)	(17.3%)	(73.5%)	

Event and non-Event Risk Reclassification Tables

Non-Events

	$risk(X, Y)$			
$r(X)$	$< 5\%$	$5 - 20\%$	$\geq 20\%$	Total
$\leq 5\%$	5486	399	21	5906 (65.8%)
$5 - 20\%$	1015	990	272	2277 (25.3%)
$\geq 20\%$	40	296	464	800 (8.9%)
Total	6541	1685	757	8983
	(72.8%)	(18.8%)	(8.4%)	

Some Problems with Interior Cells of Risk Reclassification Tables

(1) Much reclassification does not imply improved performance.

Example of RR tables with the same margins but different % reclassification. $r(X)$

	$risk(X, Y)$			
	< 5	$5 - 20$	$> 20\%$	Total
< 5	10	10	0	20
$5 - 20$	5	20	10	35
> 20	5	5	35	45
Total	20	35	45	100

% reclassification= 35%

	$risk(X, Y)$			
	< 5	$5 - 20$	$> 20\%$	Total
< 5	20	0	0	20
$5 - 20$	0	35	0	35
> 20	0	0	45	45
Total	20	35	45	100

% reclassification= 0%

- The comparison of performances of models $r(X)$ versus $r(X, Y)$ is found by comparing vertical versus horizontal margins.

Some Problems with Interior Cells of Risk Reclassification Tables

(2) The NRI statistics[4,21] do not capture 'improved risk reclassification' when # categories > 2.

Definitions

$$\begin{aligned} NRI \text{ (event)} &= P[r_{cat}(X, Y) > r_{cat}(X) | D = 1] \\ &\quad - P[r_{cat}(X, Y) < r_{cat}(X) | D = 1] \end{aligned}$$

$$\begin{aligned} NRI \text{ (non-event)} &= P[r_{cat}(X, Y) < r_{cat}(X) | D = 0] \\ &\quad - P[r_{cat}(X, Y) > r_{cat}(X) | D = 0] \end{aligned}$$

$$NRI = NRI \text{ (event)} + NRI \text{ (non-event)}$$

Example where $NRI > 0$ but no performance improvement

Events

$r(X)$ \ $r(X, Y)$	low	med	high	
low	10	10	0	20
med	5	20	10	35
high	5	5	35	45
	20	35	45	100

Non-Events

$r(X)$ \ $r(X, Y)$	low	med	high	
low	500	100	0	600
med	100	200	0	300
high	0	0	100	100
	600	300	100	900

NRI (event) $+ NRI$ (non-event) $= (20\% - 15\%) + 0\% = 5\%$

but

% cases and % controls in each risk category are the same for $r(X, Y)$ and $r(X)$ models.

Example where $NRI = 0$ but there is performance change (improvement)

Events		$r(X, Y)$ low	$r(X, Y)$ med	$r(X, Y)$ high	
$r(X)$	low	10	10	0	20
	med	15	20	10	45
	high	0	5	30	35
		25	35	40	100

Non-Events		$r(X, Y)$ low	$r(X, Y)$ med	$r(X, Y)$ high	
	low	500	100	0	600
	med	100	200	0	300
	high	0	0	100	100
		600	300	100	900

- NRI(event)$= 10\% + 10\% + 0\% - 5\% - 0\% - 15\% = 0$
- % cases at high risk increased from 35 to 40%

NRI with 2 Risk Categories

Events		$r(X, Y) < 20$	$r(X, Y) \geq 20$	
$r(X)$	< 20	236 (23.2%)	118 (11.6%)	354
	≥ 20	33 (3.2%)	630 (61.9%)	663
		269	748	1017

NRI(event)$= (118 - 33)/1017 = 8.4\%$
$\Delta HR_C = 748/1017 - 663/1017 = 73.5\% - 65.2\% = 8.4\%$

- NRI(event)$= \Delta HR_C$
  NRI(non-event)$= -\Delta HR_N$
  $NRI = \Delta HR_C - \Delta HR_N$
- uses the table margins

Continuous NRI

Definitions

- Cont-NRI(event)$\equiv P[risk(X, Y) > risk(X)|D = 1] - P[risk(X, Y) < risk(X)|D = 1]$
  $= 2P[risk(X, Y) > risk(X)|D = 1] - 1$
- Cont-NRI(non-event) $\equiv 2P[risk(X, Y) < risk(X)|D = 0] - 1$
- Cont-NRI=NRI(event) $+NRI$(non-event)

Concerns

- Possible for Cont-NRI(event)> 0 even when distribution of $risk(X)$ and $risk(X, Y)$ are the same for cases and vice versa. Similar concerns for Cont-NRI(non-event).
- Other continuous summary indices such as ΔAUC, ΔMRD, IDI, and $AARD$ do not have this undesirable property because they use the table margins.

Cook-Ridker Analysis Strategy[24]

- Is problematic.
  (i) % reclassification
    * Not a key summary in general (unless there is very little reclassification).
  (ii) % 'correct' reclassification
    * Defined as an off diagonal cell is correctly reclassified if % events in the cell is closer to $r(X, Y)$ category label than to $r(X)$ 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 $X = \phi$.

FYRP

Calibration Reminders

- $r(X)$ and $r(X, Y)$ are well calibrated is a basic assumption

$$P(D = 1|r(X)) = r(X) \text{ and } P(D = 1|r(X, Y)) \approx r(X, Y)$$

- Additional requirement for risk reclassification tables is that within categories of $r(X)$, $r(X, Y)$ is well calibrated*

$$P[D = 1|r(X, Y), r_{cat}(X)) \approx P(D = 1|r(X, Y)) \approx r(X, Y)$$

* This follows directly if $r(X, Y) = P(D = 1|X, Y)$.

FYRP

Implications for Cook and Ridker Analysis Strategy

- $P(D = 1|r(X, Y) \in A, r(X) \in B) =$
  $= E\{P(D = 1|r(X, Y))|r(X, Y) \in A, r(X) \in B\}$
  $= E\{r(X, Y)|r(X, Y) \in A, r(X) \in B)$
- Event rate in off diagonal cell = average $(r(X, Y)|cell)$
  - in interval A, not in interval B
  - H_0 regarding $r(X, Y)$ model holds
  - H_0 regarding $r(X)$ model does not hold

Summary of Unit 2

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.

Unit 3

Inference

- Not just for Pepe's preferred statistics.
- Reviewers may demand certain analyses such as ΔAUC, ΔMRD, NRI, etc.

Analysis Strategy for a Single Model

(i) Calibration assessment via predictiveness curve or calibration curve.

(ii) Plots of distributions of risk in cases and in controls and relative utility curves (or decision curves).

(iii) Single threshold statistics: $HR_C(t)$, $HR_N(t)$, $RU(t)$, $HR_C(t) - HR_N(t)$.

- analogues for several thresholds or risk categories.

(iv) Continuous summary statistics: AUC, $pAUC$, MRD, $AARD$

(v) Given specified criterion for % cases classified as high risk (t_c) or % controls classified as high risk (t_n), or % population classified as high risk (t_p): $ROC(t_n)$, $ROC^{-1}(t_c)$, $L(t_p)$, $L^{-1}(t_c)$.

Analysis Strategy for Comparison of Two Models

(i) Plots of case and control risk distributions and of relative utility aligned to enable comparisons between models.

(ii) Single threshold comparison statistics along with event and non-event reclassification tables: $\Delta HR_c(t)$, $\Delta HR_N(t)$, $\Delta RU(t)$, $\Delta HR_C(t) - \Delta HR_N(t) = NRI(t)$, % reclassification, risk recalibration statistics.

- Analogues with multiple risk categories.

(iii) Continuous comparison statistics: ΔAUC, $\Delta pAUC$, ΔMRD (a.k.a. IDI), $\Delta AARD$(a.k.a. ΔTG), *NRI* (event, non-event, overall).

(iv) $\Delta ROC(t_n)$, $\Delta ROC^{-1}(t_c)$, $\Delta \mathcal{L}(t_p)$, $\Delta \mathcal{L}^{-1}(t_c)$

Table 1. Estimates of performance with 95% confidence intervals. (Stata programs `incrisk, incroc`)

	Risk Model		
	$r(X)$	$r(X,Y)$	Comparison
$HR_C(0.20)$	65.2%(62.0,68.1)	73.5%(70.9,76.0)	8.4%(6.1,10.8)
$HR_N(0.20)$	8.9%(8.2,9.7)	8.4%(7.7,9.1)	−0.5%(1.0,0.0)
$RU(0.20)$	45.5%(41.9,48.8)	54.9%(51.9,58.0)	9.4%(7.0,12.2)
$HR_C(.2) - HR_N(.2)$	56.3%(53.2,59.1)	65.1%(62.5,67.6)	8.8%(6.5,11.5)
Cat-NRI (0.2)	—	—	0.088(0.65,0.115)
% reclassification	—	—	7.8%(6.8,8.7)
AUC	0.884(0.873,0.895)	0.920(0.91,0.929)	0.036(0.030,0.043)
MRD/*IDI*	0.322(0.298,0.350)	0.416(0.391,0.441)	0.094(0.081,0.109)
AARD	0.599(0.574,0.628)	0.671(0.650,0.696)	0.073(0.050,0.098)
Cont-NRI (event)	—	—	0.388(0.339,0.430)
Cont-NRI (non-event)	—	—	0.411(0.369,0.447)
Cont-NRI	—	—	0.799(0.724,0.865)
$ROC(0.10)$	0.672(0.643,0.703)	0.758(0.732,0.786)	0.087(0.059,0.111)
$ROC^{-1}(0.70)$	0.115(0.098,0.133)	0.070(0.057,0.081)	−0.044(0.058,0.034)
$\mathcal{L}(0.15)$	0.658(0.634,0.687)	0.735(0.711,0.761)	0.077(0.055,0.097)
$\mathcal{L}^{-1}(0.70)$	0.174(0.158,0.191)	0.135(0.121,0.145)	−0.040(0.053,0.030)

Estimation

Simple and Most Common Approach

- Cohort data = $\{D_i, X_i, Y_i; i = 1, \ldots n\}$
- Fit models (e.g., with logistic regression); $\widehat{r}(X)$ and $\widehat{r}(X, Y)$
- Empirical estimates of performance measures substituting $\widehat{r}(X_i)$ and $\widehat{r}(X_i, Y_i)$ for $r(X_i)$ and $r(X_i, Y_i)$.

Issues

- Optimistic bias due to developing models and assessing performance with the same data.
- Variability in $\widehat{r}$ must be acknowledged in assessing variability of estimated performance measures.

Variability in $\widehat{r}$

- Is a substantial component of var($\widehat{M}(\widehat{r})$) where $\widehat{M}$ denotes estimate of performance measure.
- *Often* ignored in variance expressions.[4,21,22,24]
- But *not always* ignored in variance expressions.[20]
- Addressed in bootstrap resampling for variance and confidence interval calculations by refitting the risk models in each resampled dataset.

Table 2. Confidence interval widths with and without refitting the model in each bootstrap sample. Confidence limits are 2.5^{th} and 97.5^{th} percentiles of the bootstrap distribution. Threshold $t = 20\%$.

	Estimate	Without Refitting	With Refitting	% Increase
ΔAUC	0.036	0.012	0.013	8.3%
Cont-NRI	0.799	0.151	0.138	20.0%
Δ*MRD/IDI*	0.094	0.020	0.028	40.0%
$\Delta AARD$	0.073	0.045	0.048	6.6%
Cat-NRI (0.20)	0.088	0.049	0.050	2.0%
Reclassification %	7.800	1.060	1.910	80.2%
$\Delta NB(0.20)\%$	0.960	0.550	0.520	-5.5%

Optimistic Bias (Overfitting Bias)

The Problem

- The fitting procedures find $\widehat{r}(X)$ to maximize an estimated measure of prediction performance (or one that is related to measures of prediction performance).
- Well documented that performance estimates are better than true performance.
- A major problem when there are many candidate predictors.

Optimistic Bias (Overfitting Bias)

Partial Solutions

- Cross validation: include $\widehat{r}_{-i}(X_i)$ in the estimate of performance, where $-i$ denotes all but i^{th} observation used to fit the model.
  - But the fitted model $\widehat{r}(X)$ may still be problematic, poorly calibrated.
- Training-test dataset split: fit $\widehat{r}(X)$ on training set and evaluate on the test set.
  - But the fitted model $\widehat{r}(X)$ may still be problematic.
  - Probably not well calibrated on the test set if $dim(X) large$.

Optimistic Bias (Overfitting Bias)

Solutions

- Use shrinkage procedures in fitting $\widehat{r}(X)$ to the training set and evaluate on the test set.
  - Requires artistry/expertise.
- Derive preliminary $\widehat{r}^*(X)$ from the training set, then fit a model to $\widehat{r}(X) \equiv P[D = 1|S(X)]$ where $S(X) = \widehat{r}^*(X)$. Evaluate performance with the test set. This is recalibration.[9,10]
  - Since S is one-dimensional, overfitting is not a big problem (see below).
  - If $P[D = 1|S(X)] \neq P[D = 1|X]$, prediction performance is not optimal.

- For low dimensional predictors overfitting is not a big problem.
- Evidenced by comparing cross-validated performance with performance not cross-validated for our data.

	$r(X)$		$r(X, Y)$		Δ	
	CV	not CV	CV	not CV	CV	not CV
$HR_C(0.20)$	0.650	0.652	0.735	0.735	0.086	0.084
$HR_N(0.20)$	0.089	0.089	0.084	0.084	-0.005	-0.005
$RU(0.20)\%$	45.3	45.5	54.9	54.9	9.69	9.41
$Cat - NRI(0.2)$	—	—	—	—	0.091	0.088
AUC	0.883	0.884	0.920	0.920	0.036	0.036
MRD	0.322	0.322	0.416	0.416	0.094	0.094
$AARD$	0.599	0.599	0.673	0.671	0.074	0.073

Hypothesis Testing

- Testing the null hypothesis of no performance improvement is redundant if Y has already been shown to be a risk factor.[25]

$$
\begin{aligned}
& H_0^1 \ : \ r(X, Y) = r(X) \text{ with probability } 1 \\
\Leftrightarrow\ & H_0^2 \ : \ ROC_{X,Y}(f) = ROC_X(f) \quad \forall \quad f \\
\Leftrightarrow\ & H_0^3 \ : \ AUC_{X,Y} = AUC_X \\
\Leftrightarrow\ & H_0^4 \ : \ MRD_{X,Y} = MRD_X \\
\Leftrightarrow\ & H_0^5 \ : \ HR^{case}_{X,Y}(t) = HR^{case}_X(t) \text{ and } HR^{control}_{X,Y}(t) = HR^{control}_X(t) \quad \forall \quad t
\end{aligned}
$$

  etc.

- Presented yesterday.

Hypothesis Testing

- Properties of tests of H_0 using ΔAUC, ΔMRD, *NRI* are very poor.[25,26,27]
  - Do not have 5% rejection rate under H_0.

Table 3: Data simulated under the null hypothesis of H_0 : no improvement. Shows are rejection rates (%) compared with nominal level of 5%. Sample size: 1000 subjects, 10% prevalence

	without adjustment for var($\widehat{\beta}$)	bootstrap including refitting $\widehat{\beta}$
ΔAUC	0.1%	2.0%
ΔMRD	3.3%	4.1%
Cat-NRI (ρ)	11.9%	0.0%
Cont-NRI	6.3%	2.5%

Hypothesis Testing

Redundancy and poor statistical properties of tests based on performance measures.
$\Rightarrow$ use standard tests of $H_0^1 : r(X, Y) = r(X)$ to address the null hypothesis of no performance improvement.

- If H_0^1 is rejected, estimate performance measures with confidence intervals.

Example

- Test of $\beta_Y = 0$ in logit$P(D = 1|X, Y) = \beta_0 + \beta_X X + \beta_Y Y$ rejected with $p =< 0.001$ using Wald test.

	estimate	confidence interval
ΔHR_C	8.4%	(6.1,10.5)
$\Delta HR_N(0.2)$	-0.5%	$-$(1.0,0.0)
$\Delta RU(0.2)$	9.4%	(7.0,12.2)
ΔAUC	0.036	(0.030,0.043)
ΔMRD	0.94	(0.081,0.109)
$\Delta AARD$	0.073	(0.050,0.098)
Cont-NRI	0.799	(0.724,0.865)

Hypothesis Testing

- Consider H_0: performance improvement $\geq$ minimal level. e.g., $\Delta NB \geq .01$ means that we require that by measuring Y on 100 subjects we gain the equivalent of correctly identifying 1 case without identifying any controls (i.e., 1 unit of benefit).
- May be useful in sample size calculations

  P[lower confidence limit $>$ minimal|anticipated performance] $\geq 80\%$

  $\Rightarrow$ sample size calculated with simulations.

Nested Case-Control Designs

- EDRN DCIS study
- (X_i, D_i) $i = 1, \ldots n$ for a cohort
  (Y_i) for n_C cases and n_N controls chosen randomly from the cohort.
- When ascertaining Y is costly, this can be cost efficient.

Nested Case-Control Designs

Estimation of performance improvement measures requires estimating

1. $\rho = P(D = 1)$
2. $r(X)$
3. case and control distributions of $r(X)$:
   $P(r(X) > t|D = 1)$ and $P(r(X) > t|D = 0)$
4. r(X,Y)
5. the case and control distributions of $r(X, Y)$:
   $P(r(X, Y) > t|D = 1)$ and $P(r(X, Y) > t|D = 0)$

1., 2., and 3. are estimable from cohort data. 5. is estimable from case-control data when $\widehat{r}(X, Y)$ is available.

Nested Case-Control Designs

Estimation of $r(X, Y)$

- requires cohort and case-control data

$$\text{logit}P[D = 1|X, Y] = \text{logit}\widehat{\rho}_{cohort} - \text{logit}\widehat{\rho}_{case-control} + \text{logit}P[D = 1|X, Y, \text{ case-control}]$$

Nested *Matched* Case-Control Designs

- Select controls to match cases on $g(X)$
- Efficient for testing $H_0 : r(X, Y) = r(X)$
- Estimating $r(X, Y)$ requires adjustment for matching

$$\text{logit}r(X, Y) = \{\text{logit}r(X) - \text{logit}r_{cc}(X)\} + \text{logit}r_{cc}(X, Y)$$

  where cc denotes case-control sample.

- Estimating $P(r(X, Y) > t|D)$ requires re-weighting the distributions in case-control matching strata.

$$P(r(X, Y) > t|D) = \sum_{g(X)} \widehat{P}(r(X, Y) > t|D, g(X))P(g(X)|D)$$

- Work in progress. . . .

Summary of Unit 3

(0) Fit complex models on a training set. Recalibrate on test set.

(1) Acknowledge variability in $\widehat{r}$ when calculating confidence intervals or p-values regarding performance measures.

(2) Testing H_0: no improvement, best achieved with standard methods for testing if Y is a risk factor.

(3) Calculate CI for Δ performance measures and test hypotheses concerning meaningful gain in performance.

(4) Methods can accommodate nested matched case-control designs.

(5) Stata software on DABS website (`incrisk, incroc` and `predcurve` programs).

Important Issues not Addressed in this Course

1. Event time outcomes with censoring

- Define $D \equiv$ [event within T years].
- Concepts should remain the same.
- Estimation is complicated by censoring.
- See work by Y. Zheng, T. Cai and others.

Important Issues not Addressed in this Course

2. What if benefit and/or cost of treatment depend on (X, Y)?

- We assumed $B(X, Y) = B$ and $C(X, Y) = C$ in calculating net benefit.
- Otherwise risk thresholds depend on (X, Y).
- Analyses are more complicated but concepts of net benefit and % cases and % controls classified as high risk are the same.
- A RCT will generally be needed to determine (X, Y) specific decision rules and overall benefit of using them.
- Methods for evaluating treatment selection markers is an active field.[28]

Important Issues not Addressed in this Course

3. Decision rules incorporating sampling variability in $\widehat{r}$

- e.g., treat if lower confidence limit of $\widehat{r} > t$
- All methods apply to $\mathit{ll}(\widehat{r})$, e.g., AUC, MRD, $HR_C(t)$, $HR_N(t)$, $RU(t)$, etc.

Software

DABS website

- Diagnostics And Biomarkers Statistical Center
- Google or Bing “DABS center” to locate the website
- datasets and software
- contributions are welcome

Stata programs

- `incrisk`
- `incroc`
- `predcurve`

Stata Code

Preliminary Help files for three commands: `predcurve, incrisk, incroc`.

To obtain code and current help files within Stata, type:

```
net from http://labs.fhcrc.org/pepe/stata
```

and install the `risk_prediction` and `pcvsuite` packages.

Title

predcurve — Predictiveness curve

Syntax

predcurve disease_var test_var1 [test_var2] [*if*] [*in*] [, *options*]

options	Description
Risk estimation	
link(*function***)**	**logit** (default) or **probit** link function.
riskl(*p***)**	lower risk threshold.
riskh(*p***)**	upper risk threshold.
covar(*varlist***)**	covariables to include in the risk model.
ci	include bootstrap confidence intervals for risk, TPR, and FPR estimates.
Case-control adjustment	
rho(#)	the population disease prevalence should be specified when case-control data are used. A cohort design is assumed by default.
Graph	
ylim(#)	upper limit for the y-axis scale.
class	plot TPR & FPR curves in a second panel.
offset(#)	specify CI offset from risk thresholds. default is .004
class_offset(#)	specify x axis offset for FPR and TPR CI's. default is .25.
Bootstrap options	
nsamp(#)	number of bootstrap samples; default is 1000.
cluster(*varlist***)**	variables identifying bootstrap resampling clusters
level(#)	set confidence level; The default is **level(95)** or as set by set level
New variable options (pending)	
genprvvars	generate new variables to hold model p_hat.

Description

predcurve plots predictiveness curves, i.e. plots of estimated disease risk vs. the risk distribution (empirical cdf). Risk estimates are based on a generalized linear binary model of disease risk as a function of the specified continuous marker variable(s), *test var1* [and *test var2*]. *disease var* is the disease indicator used for the model dependent variable. Additional covariables specified with **covar(***varlist***)** can be included with the marker variable in the risk model.

Risk percentiles and reference lines for specified high and/or low risk thresholds, **riskh(***p***)** and **riskl(***p***)**, are optionally included on the plot.

An additional plot of the true and false positive fractions, TPR and FPR, as functions of the risk distribution is optionally included as are estimates and reference lines corresponding to specified risk thresholds.

Risk calculations assume a cohort sampling design by default. A correction for case-control data will be employed if the population prevalence of disease is specified with **rho(#)**, and bootstrap samples for optional CI calculation will be drawn separately from cases and controls.

Options

Risk estimation

link(*function***)** Specifies the binary GLM link function. *function* options include **logit** (the default) and **probit**.

riskl(*p***)** lower risk threshold. Calculate the percentage of subjects with risk < p; also calculate the percentages of cases (TPR) and controls (FPR) with risk ≥ p if **class** is specified.

riskh(*p***)** upper risk threshold. Calculate the percentage of subjects with risk > p; also calculate the percentages of cases (TPR) and controls (FPR) with risk ≥ p if **class** is specified.

covar(*varlist***)** covariables to include in the risk model.

Case-control adjustment

rho(#) The population disease prevalence should be specified when case-control data are used. rho will be used for risk calculation adjustment, and bootstrap sampling for optional CI's will done separately from case and control samples. Only the **logit** GLM link may be used with case-control data; including the **link(probit)** with the **rho(#)** option will return an error. Risk calculations assume a cohort sampling design by default.

Graph options

class specifies that TPR and FPR curves should be plotted.

ylim(#) set the upper limit for y-axis scale. Must be between 0 and 1. If not specified, this is 10% larger than the largest observed risk estimate.

ci indicates that bootstrap percentile-based confidence intervals for risk percentiles at specified thresholds **riskh(***p***)** or **riskl(***p***)** be calculated. CI's for TPR and FPR will be calculated for the specified thresholds if the **class** option is included.

offset(#) specifies the offset from specified risk thresholds for placement of 2nd CI if 2 markers are specified, for avoidance of superimposed interval bars. The argument must be between 0 and .02; default is **offset(.004)**. **#** is a proportion of the yaxis range if **ylim(#)** is included.

class_offset(#) specifies the x-axis (risk percentile) offset for placement of TPR and FPR CI's in order to avoid superimposed interval bars. The argument must be between 0 and 2; default is **offset(.25)**. relevant only if both the **class** and **ci** options are specified.

Bootstrap options

These options are relevant only if the **ci** option is specified.

nsamp(#) specifies the number of bootstrap samples for risk model estimation be performed to obtain confidence intervals. The default is 1000 replications.

cluster(*varlist***)** specifies variables identifying bootstrap resampling clusters. See the cluster option of the bootstrap command.

level(#) specifies the confidence level, as a percentage, for confidence intervals. The default is **level(95)** or as set by set level.

New variable options

genprvvars (*pending*) generate new variables, ph#, for each marker in the *test varlist* to hold predicted probabilities for each subject based on the binary risk model fit. New variable names are numbered (#) according to marker variable order in the *test_varlist*

replace requests that existing variables *ph#* be overwritten by **genpcv** or **genrocvars**.

Saved results

If risk threshold options, **riskl(***p***)** or **riskh(***p***)** are specified, **predcurve** saves the following in **r()**:

Scalars

r(*rpc_l_#*)	lower threshold risk percentile estimate, marker number **#**.
r(*rpc_u_#*)	upper threshold risk percentile estimate, marker number **#**.

If the class option, **class**, is additionally specifed, TPR and FPR estimates for the specified thresholds are returned in:

r(*tpr_l_#*)	lower threshold TPR estimate, marker number **#**.
r(*tpr_u_#*)	upper threshold TPR estimate, marker number **#**.
r(*fpr_l_#*)	lower threshold FPR estimate, marker number **#**.
r(*fpr_u_#*)	upper threshold FPR estimate, marker number **#**.

If the **ci** option is included, bootstrap postestimation results left behind by bstat are available.

Returned matrices include:

e(ci_percentile)	2 x k matrix of bootstrap percentile CI's, where k = (# markers) * (# threholds) * (# estimators), and rows correspond to upper and lower bounds. Matrix column names indicate estimator, marker #, and threhold.

Examples

```
. use http://labs.fhcrc.org/pepe/data/janssens_c, clear

. predcurve d logscr

. predcurve d logscr bmi

. predcurve d logscr bmi, riskh(.40)

. predcurve d logscr bmi, riskh(.40) riskl(.10) class

. predcurve d logscr, cov(age hyper bmi bruit vas gender) riskh(.40) riskl(.10) class
    ci

. logistic d age hyper bmi bruit vasc gender, coef
. predict mod1, xb
. la var mod1 "risk(X)"

. logistic d age hyper bmi bruit vasc gender logscr, coef
. predict mod2, xb
. la var mod2 "risk(X,Y)"

. predcurve d mod1 mod2, riskh(.40) riskl(.10) class
```

```
. predcurve d mod1 mod2, riskh(.40) riskl(.10) class ylim(.5)
```

References

Authors

Gary Longton, Fred Hutchinson Cancer Research Center, Seattle, WA.
glongton@fhcrc.org

Margaret Pepe, Fred Hutchinson Cancer Research Center and University of Washington, Seattle, WA.
mspepe@u.washington.edu

Statistics/Data Analysis

help incrisk (version 1.0.5)

Title

incrisk — Incremental value of one or more markers or predictors relative to a list of existing predictors. Evaluation is with respect to various risk improvement measures.

Syntax

incrisk *disease_var* [*if*] [*in*], **x(***varlist***) y(***varlist***)** [*options*]

options	Description
Models	
* **x(***varlist***)**	base model predictor variables
* **y(***varlist***)**	additional marker or predictor variables
Optional Statistics	
cook	Cook Reclassification calibration statistic.
GLM alternatives	
glm(*type***)**	specify GLM model type; **logit** (default), **poisson**, or **logbin**
Risk classification	
rcat(*numlist***)**	specify up to 4 thresholds for risk category classification.
categorical	include categorical risk classification tables and statistics even if **rcat()** is not specified. Define two categories with single threshold *rho* = disease prevalence.
rho(*r***)**	population disease prevalence; sample prevalence is used by default. Required if case-control sampling is specified.
Cross-validation	
nsplit(*k***)**	specify number of random partitions for k-fold cross-validation; default is **nsplit(10)**
nocv	omit cross-validation
Summary statistics	
noauc	omit calculation of the area under ROC curve (AUC)
lz(*fp***)**	Lorenz measure at specified population proportion *fp*; default is **lz(.15)**
lzinv(*fc***)**	inverse Lorenz measure at specified case proportion *fc*; default is **lzinv(.70)**
Sampling variability	
nsamp(*#***)**	specify number of bootstrap samples; default is **nsamp(1000)**
nobstrap	omit bootstrap sampling
ccsamp	specify case-control rather than cohort bootstrap sampling
cluster(*varlist***)**	specify variables identifying bootstrap resampling clusters
resfile(*filename***)**	save bootstrap results in *filename*
replace	overwrite specified bootstrap results file if it already exists
level(*#***)**	specify confidence level; default is **level(95)**
New variable	
genxb	generate new variable, **xb***#*, to hold predicted values for model *#*
replace_xb	overwrite existing **xb***#* variables

* **x(***varlist***)** and **y(***varlist***)** are required.

Description

incrisk compares predicted risks from base (X) and enhanced (XY) models of disease outcome with respect to various performance improvement measures including Net Reclassification Improvement (NRI), Integrated Discrimination Improvement (IDI), and changes in each of the Mean Risk Difference (MRD), Lorenz measure (L and L^(-1)), Above Average Risk Difference (AARD), standardized Total Gain (TG), and Area under the ROC curve (AUC).

If categorical risk classification is specified, additional measures include the categorical version of the Net Reclassification Improvement (NRI), Reclassification percent (RC), changes in percent of cases and controls classified as high risk (HRc and HRn), Net Benefit (NB), and optionally, the Cook Reclassification Calibration statistic (RCC).

Base model variables are specified with **x(***varlist***)**. The enhanced model additionally includes variables specified with **y(***varlist***)**. *disease_var* is the 0/1 event or disease indicator variable.

All estimators are calculated empirically from model based predicted values. The model based predicted values used for classification and for calculation of the improvement mesasures are obtained by k-fold cross-validation unless specified otherwise with the **nocv** option.

Bootstrap standard errors and confidence intervals (CIs) for the requested statistics are calculated. Percentile CI's are displayed. Normal-based, and bias-corrected CIs are available (see estat bootstrap and **[R] bootstrap**).

Many of the standard e-class bootstrap results left behind by **bstat** are available after running **incrisk**, in addition to the r-class results listed below.

Options

Models

x(*varlist***)** specifies the variables to be included in the base model. **x()** is required.

y(*varlist***)** specifies additional variables to be included in the enhanced model. **y()** is required.

Optional Statistics

cook includes the Cook Reclassification calibration statistic in the output. Entries for the mean model-based risk from base and enhanced models are additionally included in the Reclassification table. Valid only when either of the **rcat()** or **categorical** options are also specified.

GLM alternatives

glm(*type***)** specifies alternate GLM model types for estimation of disease risk.

logit, the default, uses logistic regression models (GLM binomial family and logit link) and resulting linear predictor for calculation of improvement measures and risk classification.

poisson uses poisson regression models (GLM poisson family and log link) and the resulting linear predictor.

logbin uses a GLM model with binomial family and log link instead of a logit link. Convergence problems are often encountered with this option.

Risk Classification

rcat(*numlist***)** specifies between 1 and 4 thresholds for risk category classification into 2-5 categories. Values must be between 0 and 1. If **categorical** is specified and **rcat()** is not, two categories with single threshold at either the sample disease prevalence, *rho_hat*, or if specified, the population prevalence, **rho()**, are used.

categorical specifies that categorical risk classification tables and statistics are to be included when **rcat()** is not specified. Two categories are defined by threshold *rho* = disease prevalence.

rho(*r***)** specifies the population disease prevalence to be used for default categorical risk classfication, Total Gain statistic calculation, and model-based risk estimation. Value must be between 0 and 1. Required when case-control bootstrap sampling is specified with **ccsamp**. The sample disease prevalence is used by default.

Cross-validation

nsplit(*k***)** specifies the number of random partitions of the data that are to be used to obtain predicted values via k-fold cross-validation. The default is **nsplit(10)**. Specifying **nsplit(***1***)** is equivalent to specifying **nocv**.

nocv specifies that cross-validation is not used. Instead, within sample predicted values are obtained from a single model fit to the full sample. **nocv** overides any specification for **nsplit(***k***)**.

Summary statistics

noauc omits calculation and report of the ROC area under the curve (AUC).

lz(*fp***)** specifies *fp* for calculation of the Lorenz measure, the proportion of cases with risk above the threshold exceeded by the proportion *fp* of subjects in the population. The argument must be between 0 and 1. The default is **lz(.15)**.

lzinv(*fc***)** specifies *fc* for calculation of the inverse Lorenz measure, the proportion of the population with risk above the threshold exceeded by the proportion *fc* of cases. The argument must be between 0 and 1. The default is **lzinv(.70)**.

Sampling variability

nsamp(#) specifies the number of bootstrap samples to be drawn for estimation of standard errors and CIs. The default is **nsamp(1000)**.

nobstrap omits bootstrap sampling and estimation of standard errors and CIs. If **nsamp()** is specified, **nobstrap** will override it.

ccsamp specifies that bootstrap samples be drawn separately from cases and controls rather than sampling from the combined sample; cohort sampling is the default.

cluster(*varlist***)** specifies variables identifying bootstrap resampling clusters. See the **cluster()** option in **[R] bootstrap**.

resfile(*filename***)** creates a Stata file (a **.dta** file) with the bootstrap results for the included statistics. **bstat** can be run on this file to view the bootstrap results again.

replace specifies that if the specified file already exists, then the existing file should be overwritten.

level(#) specifies the confidence level for CIs as a percentage. The default is **level(95)** or as set by **set level**.

New variable

genxb generates new variables **xb1** [and **xb2**], to hold the x-validation generated model predicted values from models specified in **mod1(***varlist***)** [and **mod2(***varlist***)**].

replace_xb requests that existing variables **xb1** [and **xb2**] be overwritten by **genxb**.

Saved results

Examples

load simulated data

```
. use http://labs.fhcrc.org/pepe/dabs/risk_reclass_b
```

incrisk with continuous measures only, and with default settings otherwise.

```
. incrisk d, x(x) y(y)
```

include categorical measures with two categories defined by the sample disease prevalence

```
. incrisk d, x(x) y(y) categorical
```

include categorical measures with 4 categories defined by specified thresholds

```
. incrisk d, x(x) y(y) rcat(.15 .25 .5) cook nsamp(100)
```

list returned results

```
. return list
```

save bootstrap results to a file and display the saved results

```
. incrisk d, x(x) y(y) rcat(.15 .25.5) cook nocv resfile(result1) replace

. bstat using result1
```

References

Pepe MS. slides for a shortcourse, *Current Methods for Evaluating Prediction Performance of Biomarkers and Tests* , can be obtained at the *FHCRC DABS site*

Authors

Gary Longton
Fred Hutchinson Cancer Research Center
Seattle, WA
glongton@fhcrc.org

Margaret Pepe
Fred Hutchinson Cancer Research Center and University of Washington
Seattle, WA
mspepe@u.washington.edu

Also See

Online: **predcurve**, **incroc** (if installed)

Statistics/Data Analysis

help incroc (version 1.0.2)

Title

incroc — Incremental value of a marker relative to a list of existing predictors. Evaluation is with respect to receiver operating characteristic (ROC) summary statistics. ROC statistics are based on percentile value (PV) calculations

Syntax

incroc *disease var* [*if*] [*in*], **mod1(***varlist***)** [**mod2(***varlist***)** *options*]

options	Description
Models	
* **mod1(***varlist***)**	model 1 predictor variables
mod2(*varlist***)**	model 2 predictor variables
Cross-validation	
nsplit(*k***)**	specify number of random partitions for k-fold cross-validation; default is **nsplit(10)**
nocv	omit cross-validation
Summary statistics	
auc	area under ROC curve (AUC)
pauc(*f***)**	partial AUC (pAUC) for false positive rate range FPR < *f*
roc(*f***)**	ROC at specified FPR = *f*
rocinv(*t***)**	ROC^(-1)(t) at specified true positive rate TPR = *t*
Standardization method	
pvcmeth(*method***)**	specify PV calculation method; **empirical** (default) or **normal**
tiecorr	correction for ties
Covariate adjustment	
adjcov(*varlist***)**	specify covariates to adjust for
adjmodel(*model***)**	specify model adjustment; **stratified** (default) or **linear**
Sampling variability	
nsamp(*#***)**	specify number of bootstrap samples; default is **nsamp(1000)**
nobstrap	omit bootstrap sampling
noccsamp	specify cohort rather than case-control bootstrap sampling
nostsamp	draw samples without respect to covariate strata
cluster(*varlist***)**	specify variables identifying bootstrap resampling clusters
resfile(*filename***)**	save bootstrap results in *filename*
replace	overwrite specified bootstrap results file if it already exists
level(*#***)**	specify confidence level; default is **level(95)**
New variable	
genxb	generate new variable, **xb***#*, to hold predicted values for model *#*
replace	overwrite existing **xb***#* variables

* **mod1(***varlist***)** is required.

Description

incroc compares linear predictors from two logit models of disease outcome with respect to one or more ROC statistics: the AUC, the pAUC for FPR < *f*, the ROC at FPR = *f*, and the inverse ROC at TPR = *t* (Pepe Section 4.3, Dodd and Pepe). Model variables are specified with **mod1(***varlist***)** and **mod2(***varlist***)**. Though any two lists of model variables can be specified, **incroc** was motivated by interest in the incremental value of a marker relative to other predictors, implying nested variable lists differing by the inclusion of the marker variable. *disease_var* is the 0/1 disease indicator variable.

Alternatively, a single model can be specified with **mod1(***varlist***)**, in which case the requested ROC statistics are returned without comparison statistics.

The model predicted values used for ROC estimation are obtained by k-fold cross-validation.

All ROC statistics are calculated by using PVs of the disease case measures relative to the corresponding linear predictor distribution among controls (Pepe and Longton, Huang and Pepe).

Optional covariate adjustment can be achieved either by stratification or with a linear regression approach (Janes and Pepe, 2008; Janes and Pepe, in press).

Bootstrap standard errors and confidence intervals (CIs) for the requested statistics and linear predictor differences are calculated. Percentile, normal-based, and bias-corrected CIs are displayed (see estat bootstrap and **[R] bootstrap**).

Wald test results for predictor comparisons are based on the bootstrap standard errors for the difference between predictors.

Many of the standard e-class bootstrap results left behind by **bstat** are available after running **incroc**, in addition to the r-class results listed below.

Options

Models

mod1(*varlist***)** specifies the variables to be included in the linear predictor of the the first logit model. **mod1()** is required.

mod2(*varlist***)** specifies the variables to be included in the linear predictor of the second logit model.

Cross-validation

nsplit(*k***)** specifies the number of random partitions of the data that are to be used to obtain predicted values via k-fold cross-validation. The default is **nsplit(10)**. Specifying **nsplit(***1***)** is equivalent to specifying **nocv**.

nocv specifies that cross-validation is not used. Instead, within sample predicted values are obtained from a single model fit to the full sample. **nocv** overides any specification for **nsplit(***k***)**.

Summary statistics

auc compares predictors with respect to the AUC. This is the default if no summary statistics are specified. +

pauc(*f***)** includes a comparison with respect to the pAUC for FPR < specified *f*. The argument must be between 0 and 1. A tie correction is included in the PV calculation if this option is included among the specified summary statistics options and the empirical PV calculation method is used.

roc(*f***)** compares predictors with respect to the ROC at the specified FPR = *f*. The argument must be between 0 and 1.

rocinv(*t***)** compares predictors with respect to the inverse ROC, ROC^(-1)(t), at the specified TPR = *t*. The argument must be between 0 and 1.

Standardization method

pvcmeth(*method***)** specifies how the PVs are to be calculated. *method* can be one of the following:

empirical, the default, uses the empirical distribution of the test measure among controls (D=0) as the reference distribution for the calculation of case PVs. The PV for the case measure y_i is the proportion of control measures Y_Db < y_i.

normal models the test measure among controls with a normal distribution. The PV for the case measure y_i is the standard normal cumulative distribution function of (y_i - mean)/sd, where the mean and the standard deviation are calculated by using the control sample.

tiecorr indicates that a correction for ties between case and control values is included in the empirical PV calculation. The correction is important only in calculating summary indices such as the AUC. The tie-corrected PV for a case with the model predicted value y_i is the proportion of control values Y_Db < y_i plus one half the proportion of control values Y_Db = y_i, where Y_Db denotes controls. By default, the PV calculation includes only the first term, i.e., the proportion of control values Y_Db < y_i. This option applies only to the empirical PV calculation method.

Covariate adjustment

adjcov(*varlist***)** specifies the variables to be included in the adjustment.

adjmodel(*model***)** specifies how the covariate adjustment is to be done. *model* can be one of the following:

stratified PVs are calculated separately for each stratum defined by *varlist* in **adjcov()**. This is the default if **adjmodel()** is not specified and **adjcov()** is. Each case-containing stratum must include at least two controls. Strata that do not include cases are excluded from calculations.

linear fits a linear regression of the predictor distribution on the adjustment covariates among controls. Standardized residuals based on this fitted linear model are used in place of the predicted values for cases and controls.

Sampling variability

nsamp(#) specifies the number of bootstrap samples to be drawn for estimation of standard errors and CIs. The default is **nsamp(1000)**.

nobstrap omits bootstrap sampling and estimation of standard errors and CIs. If **nsamp()** is specified, **nobstrap** will override it.

noccsamp specifies that bootstrap samples be drawn from the combined sample rather than sampling separately from cases and controls; case-control sampling is the default.

nostsamp draws bootstrap samples without respect to covariate strata. By default, samples are drawn from within covariate strata when stratified covariate adjustment is requested via the **adjcov()** and **adjmodel()** options.

cluster(*varlist***)** specifies variables identifying bootstrap resampling clusters. See the **cluster()** option in **[R] bootstrap**.

resfile(*filename***)** creates a Stata file (a **.dta** file) with the bootstrap results for the included statistics. **bstat** can be run on this file to view the bootstrap results again.

replace specifies that if the specified file already exists, then the existing file should be overwritten.

level(#) specifies the confidence level for CIs as a percentage. The default is **level(95)** or as set by **set level**.

New variable

genxb generates new variables **xb1** [and **xb2**], to hold the x-validation generated model predicted values from models specified in **mod1(***varlist***)** [and **mod2(***varlist***)**].

replace requests that existing variables **xb1** [and **xb2**] be overwritten by **genxb**.

Saved results

incroc saves the following in **r()**, where *stat* is one or more of **auc**, **pauc**, **roc**, or **rocinv**, corresponding to the requested summary statistics :

Scalars

r(*stat***1)**	statistic estimate for first model predictor
r(*stat***2)**	statistic estimate for second model predictor
r(*stat***delta)**	estimate difference, *stat***2** - *stat***1**
r(se_*stat***1)**	bootstrap standard-error estimate for first predictor statistic
r(se_*stat***2)**	bootstrap standard-error estimate for second predictor statistic
r(se_*stat***delta)**	bootstrap standard-error estimate for the difference, *stat***2** - *stat***1**

Examples

Use Norton neonatal audiology dataset

```
. use http://labs.fhcrc.org/pepe/book/data/nnhs2, clear
```

Single model

```
. incroc_v102 d if ear == 1, mod1(currage gender)
```

Clustered bootstrap sampling; observations for both ears

```
. incroc_v102 d, mod1(currage gender) cluster(id) auc pauc(.20) roc(.20)
```

Nested models

```
. incroc_v102 d if ear == 1 & site < 4, mod1(currage gender) mod2(currage gender
      y1) adjcov(site)adjmodel(strat) auc pauc(.20) roc(.20)
```

Predictors obtained without cross-validation; save and plot predicted values

```
. incroc_v102 d if ear == 1 & site < 4, mod1(currage gender) mod2(currage gender
      y1) nocv adjcov(site) adjmodel(strat) genxb
. roccurve d xb1 xb2
```

References

Dodd L and Pepe MS. 2003. Partial AUC estimation and regression. *Biometrics* 59:614-623.

Huang Y, Pepe MS. Biomarker evaluation using the controls as a reference population. *Biostatistics* (in press)

Janes H, Pepe MS. 2008. Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting. *American Journal of Epidemiology* 168: 89-97.

Janes H, Pepe MS. Adjusting for covariate effects on classification accuracy usin covariate-adjusted roc curve. *Biometrika* (in press)

Pepe MS, Longton G. 2005. Standardizing markers to evaluate and compare their performances. *Epidemiology* . 16(5):598-603.

Pepe MS. 2003. *The Statistical Evaluation of Medical Tests for Classification an Prediction.* Oxford University Press.

Authors

Gary Longton
Fred Hutchinson Cancer Research Center
Seattle, WA
glongton@fhcrc.org

Margaret Pepe
Fred Hutchinson Cancer Research Center and University of Washington
Seattle, WA
mspepe@u.washington.edu

Holly Janes
Fred Hutchinson Cancer Research Center
Seattle, WA
hjanes@fhcrc.org

Also See

Online: **roccurve**, **comproc**, **rocreg** (if installed)

References

Current Methods for Evaluating Prediction Performance of Biomarkers and Tests

1. Patel UD, Garg AX, Krumholz HM, Shilpak MG, Coca SG, Sint K, Theissen-Philibrook H, Koyner JL, Swaminathan M, Passik CS, Parikh CR for the TRIBE-AKI Consortium. Pre-operative serum brain narituretic peptide and risk of acute kidney injury after cardiac surgery. *Circulation* (submitted).

2. Seymour CW, Kahn JM, Cooke CR, Watkins TR, Heckbert SR, Rea TD. Prediction of critical illness during out-of-hospital emergency care. *Journal of the American Medical Association* 2010 **304**:747–54.

3. Kerlikowske K, Molinaro AM, Gauthier ML, Berman HK, Waldman F, Bennington J, Sanchez H, Jimenez C, Stewart K, Chew K, Ljung BM, Tlsty TD. Biomarker expression and risk of subsequent tumors after initial ductal carcinoma in situ diagnosis. *Journal of the National Cancer Institute* 2010 **102**:627–637.

4. Pencina MJ, D'Agostino RB, D'Agostino Jr, RB, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine* 2008 **27**:157–172.

5. Pepe MS. Problems with risk reclassification methods for evaluating prediction models. *American Journal of Epidemiology* 2011 **173**:1327–1335.

6. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010 **21**:128–138.

7. Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, et al. Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology* 2008 **167**:362-368.

8. Hosmer DW, Lemeshow S. Goodness of Fit Tests for the Multiple Logistic Regression Model. *Communications in Statistics, Theory and Methods*. 1980 **A9**:1043–1069.

9. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in Medicine* 2004 **23**:2567–2586.

10. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer 2009.

11. Gail MH. Value of adding single-nucleotide polymorphism genotypes to a breast cancer risk model. *Journal of the National Cancer Institute* 2009 **101**:959–963.

12. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III). *Journal of the American Medical Association* 2001 **285**:4286–2497.

13. Baker SG. Putting risk prediction in perspective: Relative utility curves. *Journal of the National Cancer Institute* 2009 **101**:1538–1542.

14. Vickers AJ, Elkin, EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making* 2006 **26**:565–574.

15. Huang Y, Pepe MS, Feng Z. Evaluating the predictiveness of a continuous marker. *Biometrics* 2007 **63**:1181–1188.

16. Bura E, Gastwirth JL. The Binary Regression Quantile Plot: Assessing the Importance of Predictors in Binary Regression Visually. *Biometrical Journal* 2001 **43**:5–21.

17. Stern RH. Nocebo Responses to Antihypertensive Medications. *Journal of Clinical Hypertension* 2008 **10**:723-725.

18. Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. *Biostatistics* 2005 **6**:227–239.

19. Pepe MS, Feng Z, Gu JW. Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina et al., *Statistics in Medicine* 2008 **27**:173–181.

20. Gu W, Pepe M. Measures to summarize and compare the predictive capacity of markers. *International Journal of Biostatistics* 2009 **5**:Article 27.

21. Pencina MJ, D'Agostino RB Sr, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in Medicine* 2011 15;30(1):11-21.

22. Pfeiffer RM, Gail MH. Two criteria for evaluating risk prediction models. *Biometrics* 2011 **67**:1057–1065.

23. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007 **115**:928–935.

24. Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. *Annals of Internal Medicine* 2009 **150**:795–802.

25. Pepe MS, Kerr KF, Longton GM, Wang Z. Testing for improvement in prediction model performance. UW Biostatistics Working Paper Series. 2011 Working Paper 379.
http://www.bepress.com/uwbiostat/paper379

26. Kerr KF, McClelland RL, Brown ER, Lumley T. Evaluating the incremental value of new biomarkers with integrated discrimination improvement. *American Journal of Epidemiology* 2011 **174**:364–374.

27. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. *BMC Medical Research Methodology*. 2011 **11**:13.

28. Janes H, Pepe MS, Bossuyt PM, Barlow WE. Measuring the performance of markers for guiding treatment decisions. *Annals of Internal Medicine* 2011 **154**:253–259.

29. Pauker SG, Kassierer JP. The threshold approach to clinical decision making. *New England Journal of Medicine* **302**:1109–1117.

Glossary

$AARD$ =above average risk difference=% cases above ρ $-$ % controls above ρ =standardized total gain statistic where total gain is the area between the predictiveness curve and the horizontal useles test curve=Youden's index at $\rho = .5\times$continuous NRI comparing $r(X)$ to null model without predictors= $.5 \times cat - NRI$ comparing to null model and using two risk categories defined by threshold ρ.

$AUC =$ area under the ROC curve

$B =$ expected benefit of treatment to a subject who would be a case in the absence of treatment

$(C, N) =$ subscripts for cases=C and controls $= N$

$Cat - NRI$= $Cat - NRI$(event)+$cat - NRI$(non-event), which is a number in $(0, 2)$

$Cat - NRI$(event)=proportion of cases above diagonal of a risk reclassification table $-$ proportion of cases below diagonal of the table.

$Cat - NRI$(non-event)=proportion of controls below versus above the diagonal of the non-event risk reclassification table

$Cont - NRI = Cont - NRI(event) + cont - NRI(non - event)$

$Cont-NRI$(event)= Prob($risk(X, Y) > risk(X)|case$)$-$Prob($risk(X, Y) < risk(X)|case$) $=$ net proportion of cases whose risks are increased by including Y.

$Cont - NRI$(non-event)= Prob($risk(x, Y) < risk(X)|control$) $-$ Prob($risk(X, Y) > risk(X)|control$) =net proportion of controls whose risks are decreased by including Y

Cost $=$ expected cost of treatment to a control (including toxicities, expense, etc.)

HR_C =% cases whose risks are in the high risk category

$HR_C(t) = P(r(X) > t|D = 1)$ where t is a risk threshold

HR_N =% controls whose risks are in the high risk category

$HR_N(t) = P(r(X) > t|D = 0)$

IDI =integrated discrimination improvement= MRD for the larger model$-MRD$ for the smaller model

$\mathcal{L}^{-1}(f_c)$ =proportion of the population with risk above the threshold exceeded by f_c of cases $= (1-\rho)^* ROC^{-1}(f_c) + \rho^* f_c$

$\mathcal{L}(f_p)$ =proportion of cases with risk above the threshold exceeded by f_p subjects in the population $= ROC(f_n)$ where f_n is defined by $f_p = p^* ROC(f_n) + (1-p)^* f_n$

MRD =mean risk $(cases)-$ mean risk $(controls) = IDI$ relative to a baseline model without any predictors=Yates slope= proportion of variation explained= R^2

$n =$ total number of subjects in the cohort

$n_C =$ number of cases (events)

$n_N =$ number of controls (non-events)

NB=net benefit$= B \times P(D=1) HR_C(t) - CostP(D=0) HR_N(t) = \rho^* HR_C - (1-\rho)^* (t/(1-t))^* HR_N$ where t is the risk threshold of treating a case has benefit $B \equiv 1$

partial-$AUC(f_0) f_0 = \{$area under the ROC over the false positive range $(0, f_0)/f_0$ =average TPR over FPR$\in (0, f_0)$

Predictiveness curve: a plot of the quantiles of $risk(X)$ in the population

$\rho = P(D=1)$

$RU(t)$ =relative utility $= NB(t)/\rho =$ proportion of maximum possible utility (ρ is max possible NB when no treatment is the default)

reclassification %=total % of subjects for whom risk category based on $r(x)$ is not equal to the risk category based on $r(X, Y)$

$risk(X) = P(D=1|X) =$ frequency of events among subjects with covariate value $X = r(X)$

$risk(X, Y) = P(D=1|X, Y) = r(X, Y)$

$ROC^{-1}(f_c)$ =proportion of controls with risks above the threshold exceeded by a proportion f_c of controls

$ROC(f_n) =$proportion of cases with risks above the threshold exceeded by a proportion f_n of controls

Total Gain $=$ area between the predictiveness curve and the horizontal line at ρ. It can be standardized by dividing by $2\rho(1-\rho)$. See *AARD*.