SISCER Module 3

Part III:
Comparing Two Risk Models

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

Outline of Part III

1. How to compare two risk models
2. How to assess the incremental value of a new biomarker
3. How not to assess the incremental value of a new biomarker

1. How to compare two risk models

In a nutshell:

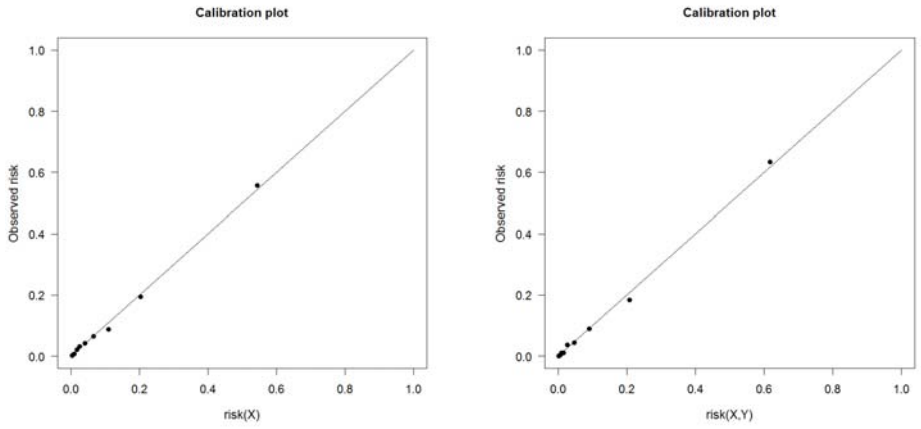
- What is your preferred measure(s) for evaluating a single risk model?
- Compare that measure(s) for two risk models.

Example

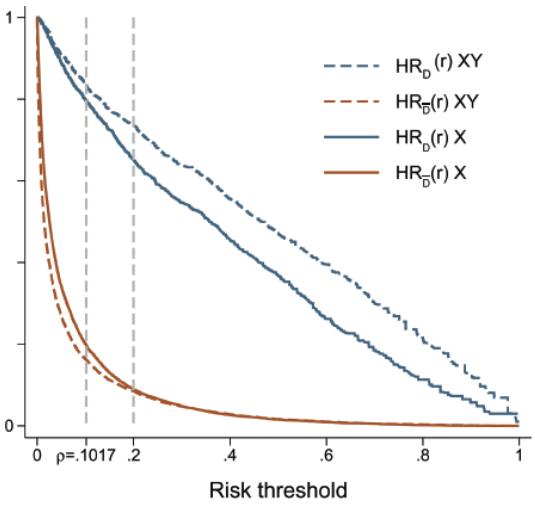
- risk(X) and risk(X,Y) for simulated data from DABS
- Both models are very well calibrated (in the moderate sense):

$$P(D=1 \mid \text{predicted risk } r) \approx r$$

(moderate calibration criterion)

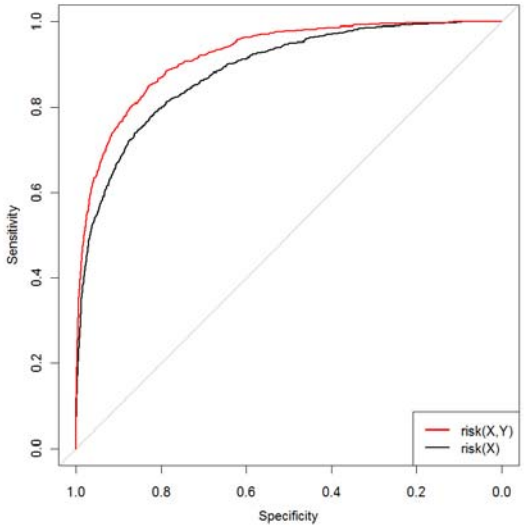


High risk classification for cases and controls

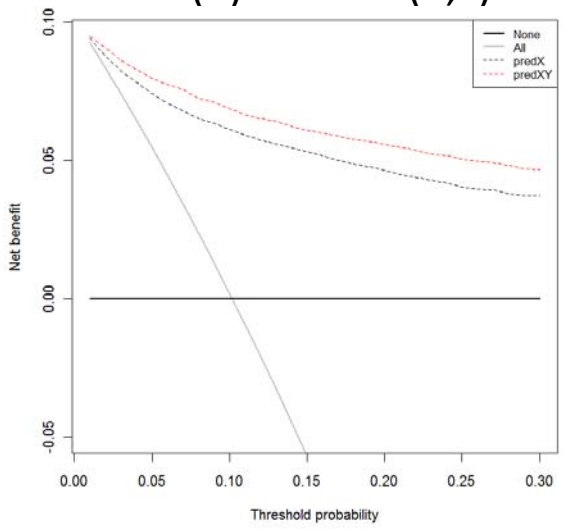


$\rho = 0.1017$

Compare ROC Curves



Decision Curves – compare the NB of risk(X) and risk(X,Y)



(Also Recall: Prostate Cancer Example in Part 2b)

Most appealing summary measures

assuming r_H=20% is an appropriate high risk threshold; and ρ=0.1017

		risk(X)	risk(X,Y)	Δ
Proportion of Cases High Risk	$HR_D(r_H)$	65.2%	73.5%	8.4%
Proportion of Controls High Risk	$HR_{\bar{D}}(r_H)$	8.9%	8.4%	-0.5%
% of maximum possible benefit	sNB or RU	45.5%	55.0%	9.5%

Less appealing summary measures

	risk(X)	risk(X,Y)	Δ	comments
AUC	0.884	0.920	0.036	ΔAUC is most popular metric
MRD	0.322	0.416	0.094*	ΔMRD is also known as IDI
AARD	0.599	0.673	0.074	
ROC(0.20)	0.672	0.758	0.087	Sensitivity at fixed specificity

2. Incremental Value of New Biomarkers

- *Incremental Value* or *Prediction Increment*: the improvement in prediction from using a new marker in addition to existing markers.
- Kattan (2003): "Markers should be judged on their ability to improve an already optimized prediction model."

A common approach:
2-stage approach for evaluating incremental value

- Use a regression model to estimate $P(D|\ X, Y\ )$ where X is the established predictor(s) and Y is the new marker

  e.g., $\text{logit}\ P(D=1|X, Y)=\beta_0+\beta_X X+\beta_Y Y$

  Test H_0: $\beta_Y=0$
- If the null hypothesis is rejected, then examine $AUC_{X,Y}$ and test

$$H_0: AUC_{X,Y} = AUC_X$$

Empirical argument against the two-stage approach:

Vickers et al. BMC Medical Research Methodology 2011, 11:13
http://www.biomedcentral.com/1471-2288/11/13

BMC Medical Research Methodology

DEBATE Open Access

One statistical test is sufficient for assessing new predictive markers

Andrew J Vickers[1*], Angel M Cronin[2], Colin B Begg[1]

Theoretical argument:

Research Article

Statistics in Medicine

Received 19 December 2011, Accepted 11 December 2012 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5727

Testing for improvement in prediction model performance

Margaret Sullivan Pepe,[a*†] Kathleen F. Kerr,[b] Gary Longton[a] and Zheyu Wang[b]

Equivalent Null Hypotheses

- Pepe *et al* (2013) prove the following null hypotheses are equivalent:
  - risk(X,Y)=risk(X)
  - $AUC_{X,Y}=AUC_X$
  - $ROC_{X,Y}(\cdot)=ROC_X(\cdot)$
  - $ROC_{Y|X}$ is the 45° line
  - IDI = 0
  - $NRI^{>0}=0$
  - (and a few others)

This is the null hypothesis when testing $\beta_Y=0$

In the two-stage approach, this test is done after the first test

- To say that these null hypotheses are the same is NOT to say that the associated statistical tests are the same.
- However, it doesn't make sense to test the same null hypothesis twice.
  - first, with a well-developed, powerful test
  - second, with an under-developed test with poor power (p-value from software should not be trusted, may be excessively conservative)
  - Illogical scientific approach

More details about why the AUC-based test is wrong:

Research Article

Statistics in Medicine

Received 22 December 2010, Accepted 6 January 2012 Published online 13 March 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5328

Misuse of DeLong test to compare AUCs for nested models

Olga V. Demler,[a*†] Michael J. Pencina[a] and Ralph B. D'Agostino, Sr.[b]

314

- Hypothesis testing has limited value
  - much more important to quantify the improvement offered by the new predictor
  - the strength of evidence to establish whether a new predictor is *useful* far exceeds what is needed to establish statistical significance

315

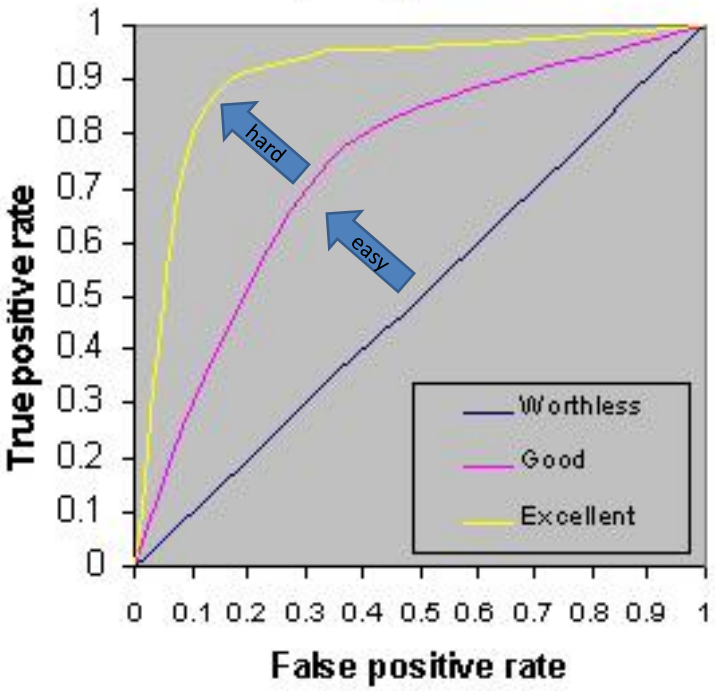
Testing Vs. Estimation

- A statistical test examines the evidence that a marker has *any* incremental value.
- However, the real challenge is finding markers that offer clinically important improvements in prediction.
- Quantifying incremental value is much more important (and more challenging) than hypothesis testing.
  - This comes down to deciding how we value a risk model

3. How not to assess incremental value

- Most common approach is to examine increase in AUC
- Since AUC is not a clinically meaningful measure, how do we know whether the increase in AUC is “enough”?

- ΔAUC ($AUC_{X,Y}$ compared to AUC_X). Some investigators consider this metric to be "insensitive" (Cook, 2007)
  - This might mean that a favorite biomarker produced a disappointing ΔAUC.
  - "Sensitivity" of ΔAUC is probably not the problem. The real problems are
    - The scale of AUC is such that an increase of 0.02 is "large"
    - p-values computed for ΔAUC are wrong; incorrect methodology tends to produce too-large p-values
    - It's fundamentally hard to improve upon a risk model that has moderately good predictive ability



A new approach: Reclassification
(Cook, *Circulation* 2007)

- Proposed that a new marker is useful if it re-classifies lots of people
  - reclassification table, next slide

TABLE 3. Comparison of Observed and Predicted Risks Among Women in the Women's Health Study*

	Model With HDL 10-Year Risk (%)				
Model Without HDL 10-Year Risk (%)	0 to <5%	5 to <10%	10 to <20%	20%+	% Reclassified
0% to <5%					
Total, n	22655	696	6	0	...
%†	97.0	3.0	0.0	0.0	3.0
Observed 10-year risk (%)‡	1.5	5.9	0.0	...	...
5% to <10%					
Total, n	593	1712	291	0	...
%	22.8	66.0	11.2	0.0	34.0
Observed 10-year risk (%)	3.7	7.6	14.7	...	...
10% to <20%					
Total, n	3	214	512	76	...
%	0.4	26.6	63.6	9.4	36.4
Observed 10-year risk (%)	0.0	7.5	10.7	23.3	...
20%+					
Total, n	0	0	41	102	
%	0.0	0.0	28.7	71.3	28.7
Observed 10-year risk (%)	...	...	15.8	32.5	...

*This comparison uses models that include Framingham risk factors with and without HDL. All estimated and observed risks represent 10-year risk of cardiovascular disease.
†Percent classified in each risk stratum by the model with HDL.
‡Observed proportion of participants developing cardiovascular disease in each category.

Reclassification Tables: Considerations

- Original proposal did not account for whether reclassification was in the "correct" direction
- Does not teach us about the performance of either risk(X) or risk(X, Y)
- If presented separately for cases and controls, a reclassification table can be interesting
  - Still, table doesn't directly help us assess whether a new biomarker offers clinically meaningful improvements in risk prediction

Reclassification Tables: Considerations

- Lots of reclassification does not imply improved performance.

		$r(X, Y)$ Low	Med	High	Total
$r(X)$	Low	10	10	0	20
	Med	5	20	10	35
	High	5	5	35	45
	Total	20	35	45	100

% reclassification= 35%

Net Reclassification Index (NRI)

- Proposed in 2008
  - Pencina, D'Agostino, D'Agostino, Vasan, *Statistics in Medicine*, 2008
- Followed on the heels of Cook's paper
- NRI is really a family of statistics

NRI terminology

event	person with the condition or destined to have the condition ("case")
nonevent	not an event ("control")
old	risk model with established predictors ("baseline")
new	risk model with established predictors and new predictor ("expanded")

Net Reclassification Improvement (NRI)

NRI = P(up | event) - P(down | event) + P(down | nonevent) - P(up | nonevent)

"up" means an individual moves to a higher risk category using new model compared to old

"down" means an individual moves to a lower risk category in new model compared to old

Original NRI (categorical NRI): fixed risk categories

- 2 categories (low risk, high risk)
- 3 categories (low risk, medium risk, high risk)
- 4 categories (e.g., Cook's paper)
- Etc.

Net Reclassification Improvement (NRI)

NRI = P(up | event) - P(down | event) + P(down | nonevent) - P(up | nonevent)

NRI_e NRI_{ne}

The *NRI* is the sum of the "event *NRI*" and the "nonevent *NRI*":

Net Reclassification Improvement (NRI)

$$NRI = \underbrace{P(\text{up} \mid \text{event}) - P(\text{down} \mid \text{event})}_{NRI_e^{>0}} + \underbrace{P(\text{down} \mid \text{nonevent}) - P(\text{up} \mid \text{nonevent})}_{NRI_{ne}^{>0}}$$

The ``category-free NRI'' interprets this formula for any upward or downward movement in predicted risk. Denote $NRI^{>0}$

Net Reclassification Indices for Evaluating Risk Prediction Instruments

A Critical Review

Kathleen F. Kerr,[a] Zheyu Wang,[a] Holly Janes,[b] Robyn L. McClelland,[a] Bruce M. Psaty,[c] and Margaret S. Pepe[b]

Epidemiology • Volume 25, Number 1, January 2014

Numerous problems with NRI statistics

- Difficult to Interpret (often mis-interpreted)
  - Not a proportion but often interpreted as such
- Why a simple sum of a summary of non-event and events?
  - In most applications, most of the population are non-events
- Does not contrast model performance measures
- 3+ categorical NRI and category-free NRI weights reclassifications indiscriminately
- Not a "proper scoring rule" – can make overfit or poorly calibrated models look good

330

Research Article

Statistics in Medicine

Received 4 July 2012, Accepted 26 February 2013 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5804

A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index

Jørgen Hilden and Thomas A. Gerds*†

JNCI JOURNAL OF THE NATIONAL CANCER INSTITUTE

Net Risk Reclassification *P* Values: Valid or Misleading?

Margaret S. Pepe, Holly Janes and Christopher I. Li

+ Author Affiliations

Correspondence to: Margaret S. Pepe, PhD, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, Seattle, WA 98109 (mspepe@u.washington.edu).

Received September 26, 2013.
Revision received December 24, 2013.
Accepted January 23, 2014.

Simulations

- X is predictive (to varying degrees)
- new marker Y is noise

Bivariate Normal Simulation Model

$$\text{Among controls: } \begin{pmatrix} X \\ Y \end{pmatrix} \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix})$$

$$\text{Among cases: } \begin{pmatrix} X \\ Y \end{pmatrix} \sim N(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix})$$

$$\text{logit}P(D=1|X=x) = \text{logit}(\rho) - \frac{1}{2}\mu_X^2 + \mu_X x$$

$$\text{logit}P(D=1|X=x, Y=y) = \text{logit}(\rho) - \frac{\mu_X^2 + \mu_Y^2 - 2r\mu_X\mu_Y}{2(1-r^2)} + \frac{\mu_X - r\mu_Y}{1-r^2}x + \frac{\mu_Y - r\mu_X}{1-r^2}y$$

In our simulations, *Y* is useless, so $\mu_Y = 0$ and $r = 0$

- Performance of model with useless marker added: ΔAUC is negative, on average

prev	AUC_X	N-train	N-test	Δ AUC	NRI
0.1	0.6	250	25,000	-1.23 (2.6)	
0.1	0.7	250	25,000	-0.88 (1.29)	
0.1	0.8	250	25,000	-0.46 (0.64)	
0.1	0.9	250	25,000	-0.23 (0.33)	
0.5	0.6	50	5,000	-1.36 (3.45)	
0.5	0.7	50	5,000	-1.65 (2.49)	
0.5	0.8	50	5,000	-1.01 (1.61)	
0.5	0.9	50	5,000	-0.62 (0.93)	

- Performance of model with useless marker added: $NRI^{>0}$ is positive, on average

prev	AUC_X	N-train	N-test	Δ AUC	NRI
0.1	0.6	250	25,000	-1.23 (2.6)	0.15 (2.83)
0.1	0.7	250	25,000	-0.88 (1.29)	0.93 (5.21)
0.1	0.8	250	25,000	-0.46 (0.64)	3.13 (9.36)
0.1	0.9	250	25,000	-0.23 (0.33)	7.56 (16.08)
0.5	0.6	50	5,000	-1.36 (3.45)	0.59 (5.11)
0.5	0.7	50	5,000	-1.65 (2.49)	2.5 (9)
0.5	0.8	50	5,000	-1.01 (1.61)	7.24 (14.77)
0.5	0.9	50	5,000	-0.62 (0.93)	17.6 (28.28)

MESA example: Polonsky et al, JAMA 2010
Adding CACS to Framingham risk factors to predict CHD events

- Risk categories 0-3%, 3-10%, >10%
- model with CACS reclassifies 26% of the sample
- estimated 3-category NRI_{event} = 0.23
- estimated 3-category $NRI_{nonevent}$ = 0.02

These are summaries of the reclassification tables (next slide)

How do we interpret these NRI statistics? Do they help us understand the clinical or public health benefit of incorporating CACS into the model?

Nonevents

Old Model	Model with CACS 0-3%	3-10%	>10%	Total
0–3%	58%	7%	1%	
	3276	408	5	65%
3–10%	12%	14%	4%	
	697	791	244	31%
>10%	1%	1%	3%	
	30	63	155	4%
Total	71%	22%	7%	5669

Events

Old Model	Model with CACS 0-3%	3-10%	>10%	Total
0–3%	16%	11%	0%	
	34	22	1	27%
3–10%	7%	25%	23%	
	15	52	48	55%
>10%	1%	3%	13%	
	2	7	28	18%
Total	24%	39%	37%	209

Risk	Old risk model		New risk model (model with CACS)	
Category	nonevent	event	nonevent	event
0-3%	67.1%	27.3%	70.6%	24.4%
3–10%	30.6%	55.0%	22.3%	38.8%
>10%	4.4%	17.7%	7.1%	36.8%
Total	5669	209	5669	209
	100%	100%	100%	100%

Summary

- The best way to compare two risk models is to compare them on a measure of performance you care about
  - e.g., Net Benefit of using the risk model to recommend treatment
- The same principle applies to assessing the incremental contribution of a new marker Y to risk prediction: is the performance of risk(X,Y) better than the performance of risk(X)?
- We don't need special metrics for comparing two risk models

Summary

- Often $AUC_{X,Y}$ will not be much larger than AUC_X. This, in itself, is not a reason to discard AUC.
  - Better reason to seek alternatives: AUC is not a clinically meaningful measure of risk model performance

Summary

- NRI statistics do not help us assess the incremental value of new markers
  - despite ~3500 citations of original 2008 paper
- NRI statistics have many of the same problems as ΔAUC, and some new problems
  - Not interpretable
  - potential to make useless new markers look promising