Evaluation of Predictive Models

Assessing calibration and discrimination

Examples

Decision Systems Group,
Brigham and Women's Hospital
Harvard Medical School

Harvard-MIT Division of Health Sciences and Technology
HST.951J: Medical Decision Support

Main Concepts

- Example of a Medical Classification System
- Discrimination
  - Discrimination: sensitivity, specificity, PPV, NPV, accuracy, ROC curves, areas, related concepts
- Calibration
  - Calibration curves
  - Hosmer and Lemeshow goodness-of-fit

Example I

Modeling the Risk of Major In-Hospital Complications Following Percutaneous Coronary Interventions

Frederic S. Resnic, Lucila Ohno-Machado, Gavin J. Blake, Jimmy Pavliska, Andrew Selwyn, Jeffrey J. Popma

[Simplified risk score models accurately predict the risk of major in-hospital complications following percutaneous coronary intervention.
Am J Cardiol. 2001 Jul 1;88(1):5-9.]

Background

- Interventional Cardiology has changed substantially since estimates of the risk of in-hospital complications were developed
  - coronary stents
  - glycoprotein IIb/IIIa antagonists

- Alternative modeling techniques may offer advantages over Multiple Logistic Regression
  - prognostic risk score models: simple, applicable at bedside
  - artificial neural networks: potential superior discrimination

Objectives

- Develop a contemporary dataset for model development:
  - prospectively collected on all consecutive patients at Brigham and Women's Hospital, 1/97 through 2/99
  - complete data on 61 historical, clinical and procedural covariates
- Develop and compare models to predict outcomes
  - Outcomes: death and combined death, CABG or MI (MACE)
  - Models: multiple logistic regression, prognostic score models, artificial neural networks
  - Statistics: c-index (equivalent to area under the ROC curve)
- Validation of models on independent dataset: 3/99 - 12/99

Dataset: Attributes Collected

History	Presentation	Angiographic	Procedural	Operator/Lab
age gender diabetes iddm history CABG Baseline creatinine CRI ESRD hyperlipidemia	acute MI primary rescue CHF class angina class Cardiogenic shock failed CABG	occluded lesion type (A,B1,B2,C) graft lesion vessel treated ostial	number lesions multivessel number stents stent types (8) closure device gp 2b3a antagonists dissection post rotablator atherectomy angiojet max pre stenosis max post stenosis no reflow	annual volume device experience daily volume lab device experience unscheduled case

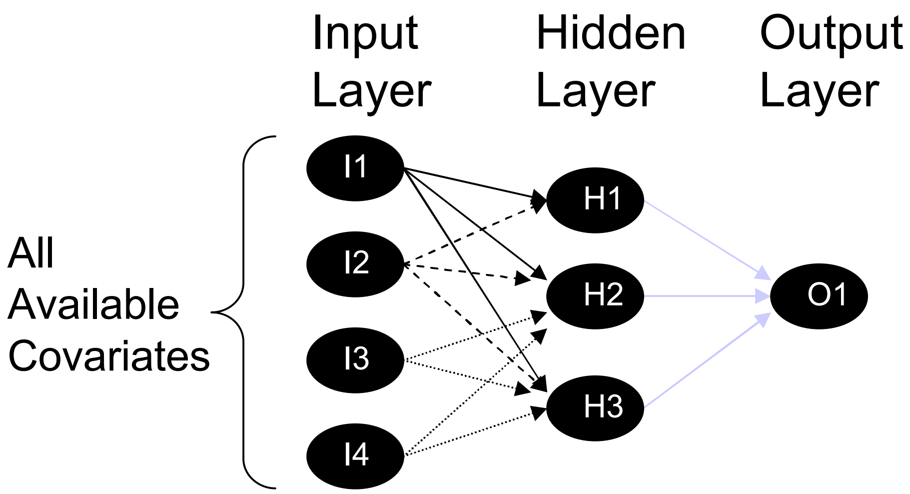
Data Source:
- Medical Record
- Clinician Derived
- Other

Logistic and Score Models for Death

	Logistic Regression Model Odds Ratio	Prognostic Risk Score Model Risk Value
Age > 74yrs	2.51	2
B2/C Lesion	2.12	1
Acute MI	2.06	1
Class 3/4 CHF	8.41	4
Left main PCI	5.93	3
IIb/IIIa Use	0.57	-1
Stent Use	0.53	-1
Cardiogenic Shock	7.53	4
Unstable Angina	1.70	1
Tachycardic	2.78	2
Chronic Renal Insuf.	2.58	2

Artificial Neural Networks

- Artificial Neural Networks are non-linear mathematical models which incorporate a layer of hidden "nodes" connected to the input layer (covariates) and the output.



Evaluation Indices

General indices

- Brier score (a.k.a. mean squared error)

$$\frac{\Sigma(e_i - o_i)^2}{n}$$

e = estimate (e.g., 0.2)

o = observation (0 or 1)

n = number of cases

Discrimination Indices

Discrimination

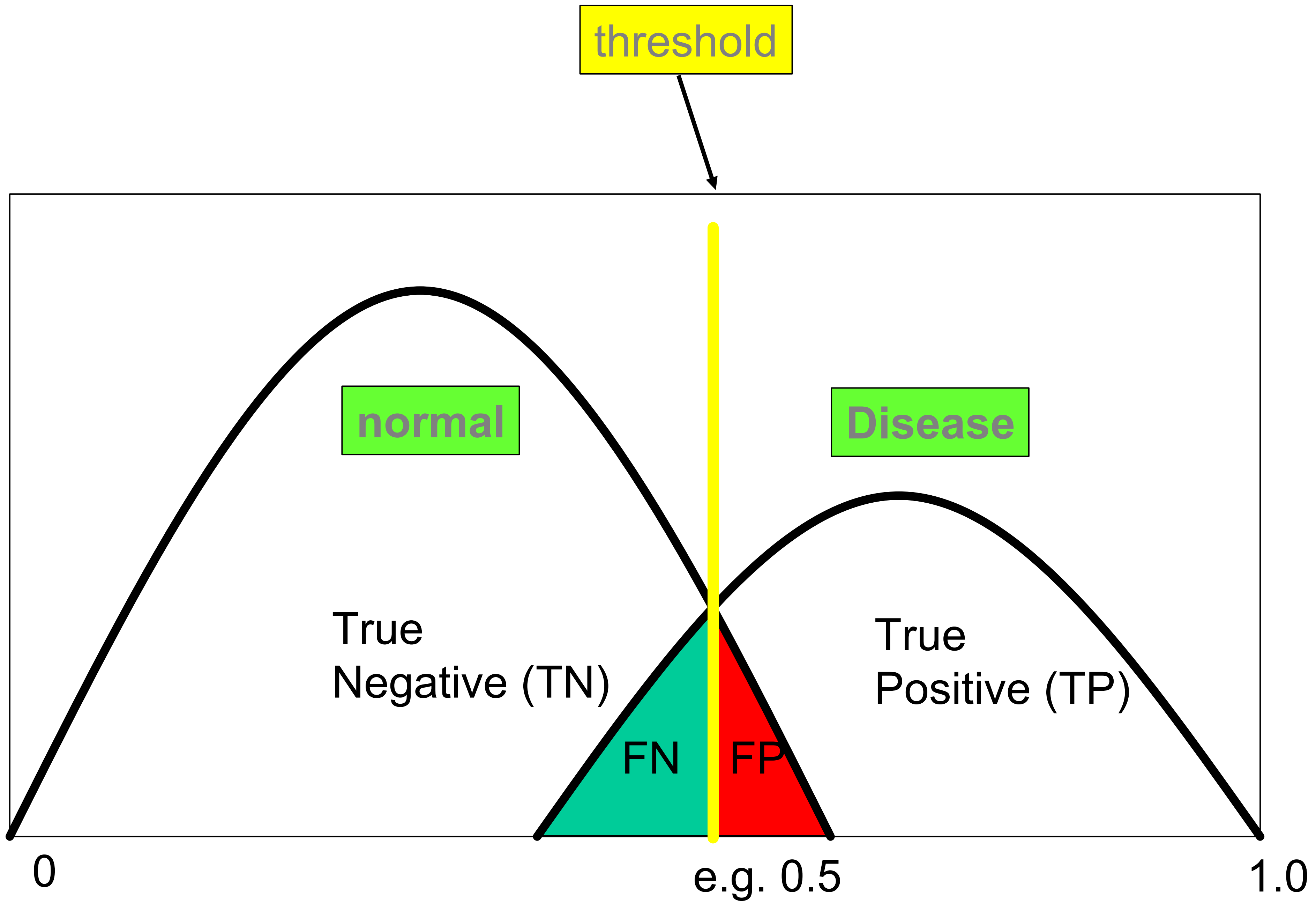
- The system can “somehow” differentiate between cases in different categories

- Binary outcome is a special case:
  - diagnosis (differentiate sick and healthy individuals)
  - prognosis (differentiate poor and good outcomes)

Discrimination of Binary Outcomes

- **Real** outcome (true outcome, also known as “gold standard”) is 0 or 1, estimated outcome is usually a number between 0 and 1 (e.g., 0.34) or a rank

- In practice, classification into category 0 or 1 is based on Thresholded Results (e.g., if output or probability > 0.5 then consider “positive”)
  - Threshold is arbitrary


threshold
normal
Disease
True
Negative (TN)
True
Positive (TP)
FN
FP
0
e.g. 0.5
1.0

Sens = TP/TP+FN

40/50 = .8

Spec = TN/TN+FP

45/50 = .9

PPV = TP/TP+FP

40/45 = .89

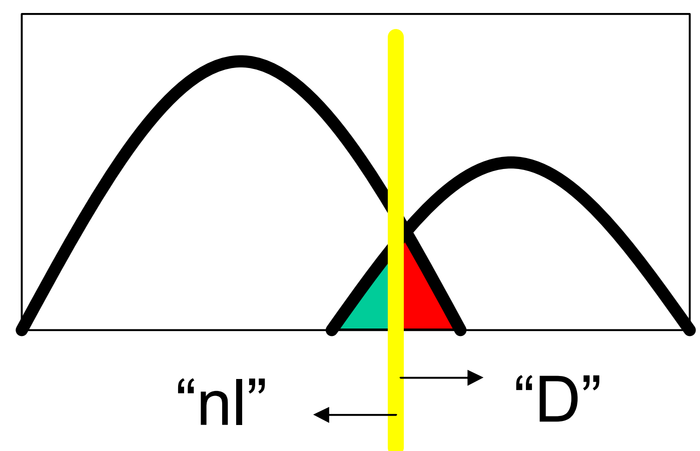
NPV = TN/TN+FN

45/55 = .81

Accuracy = TN +TP

70/100 = .85

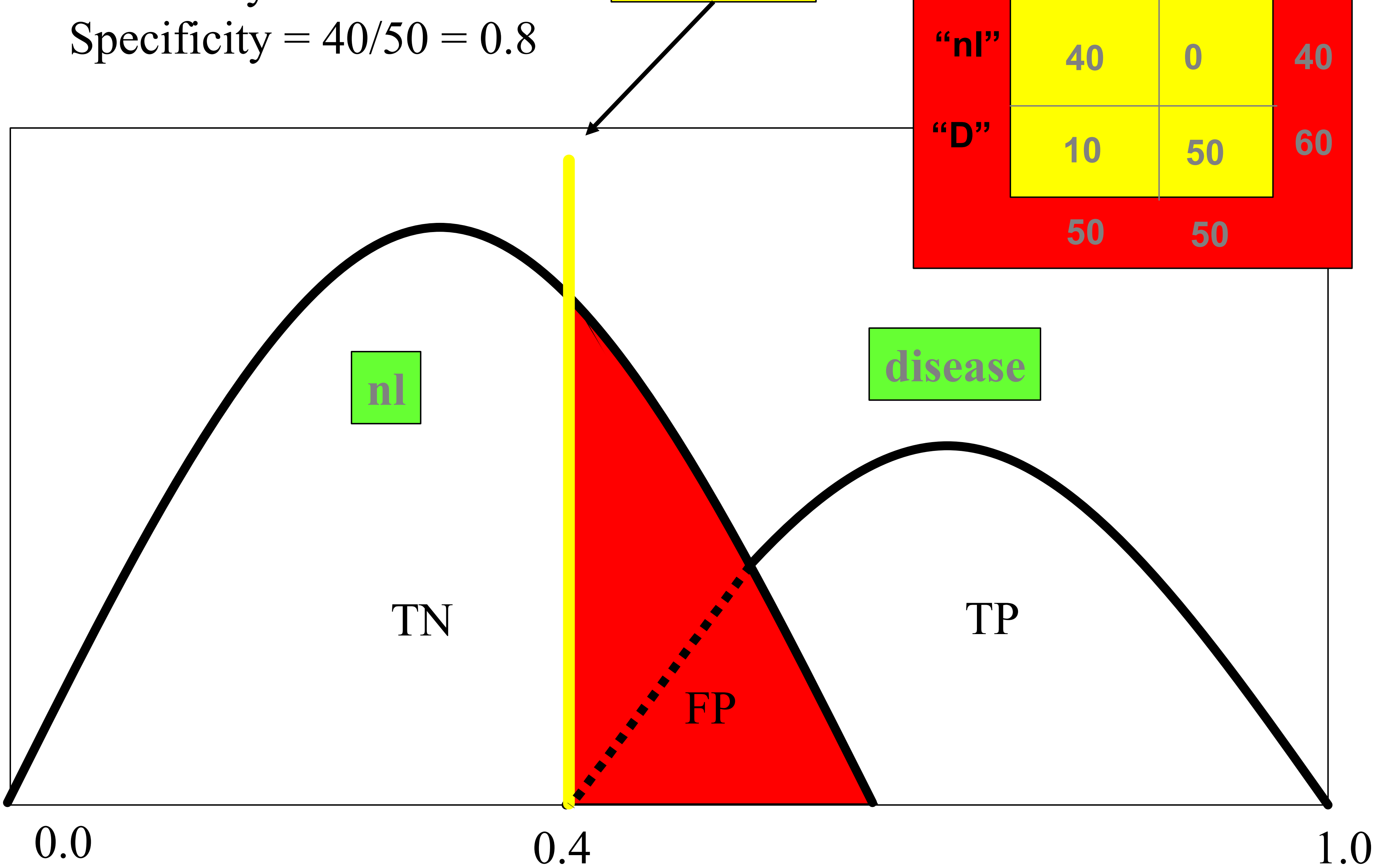
	nl	D
“nl”	45	10
“D”	5	40



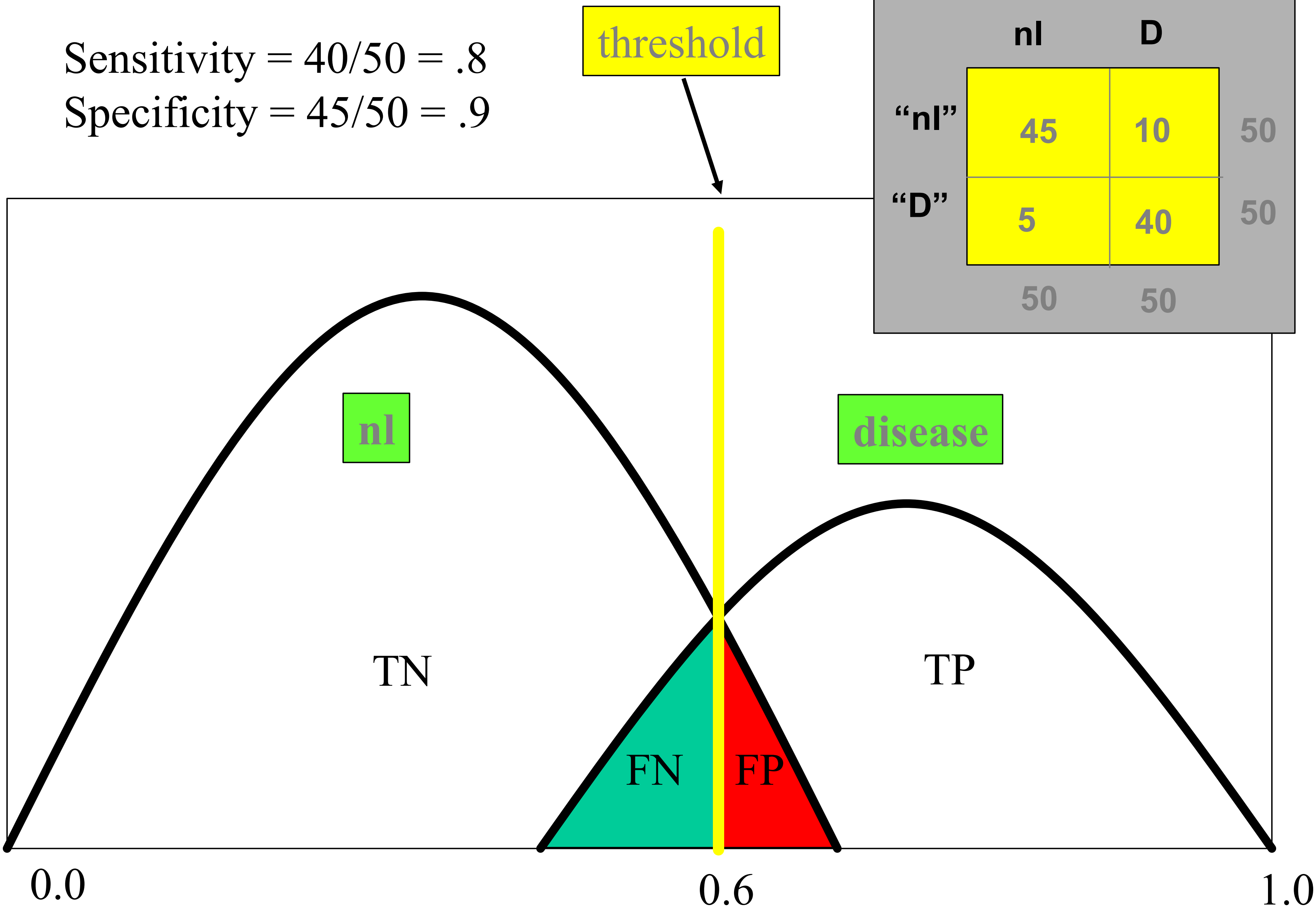
Sensitivity = 50/50 = 1

Specificity = 40/50 = 0.8

	nl	D	
"nl"	40	0	40
"D"	10	50	60
	50	50	

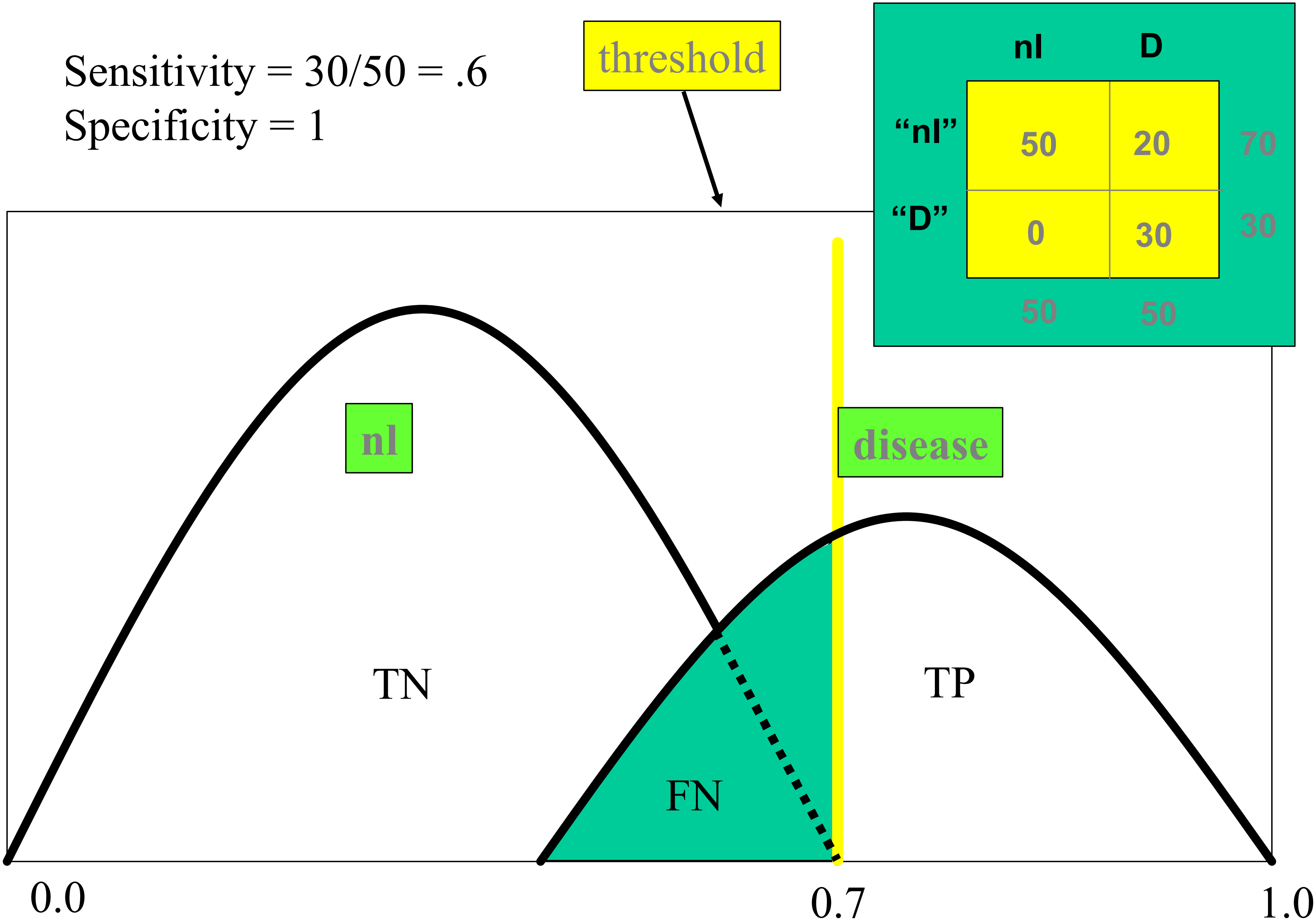


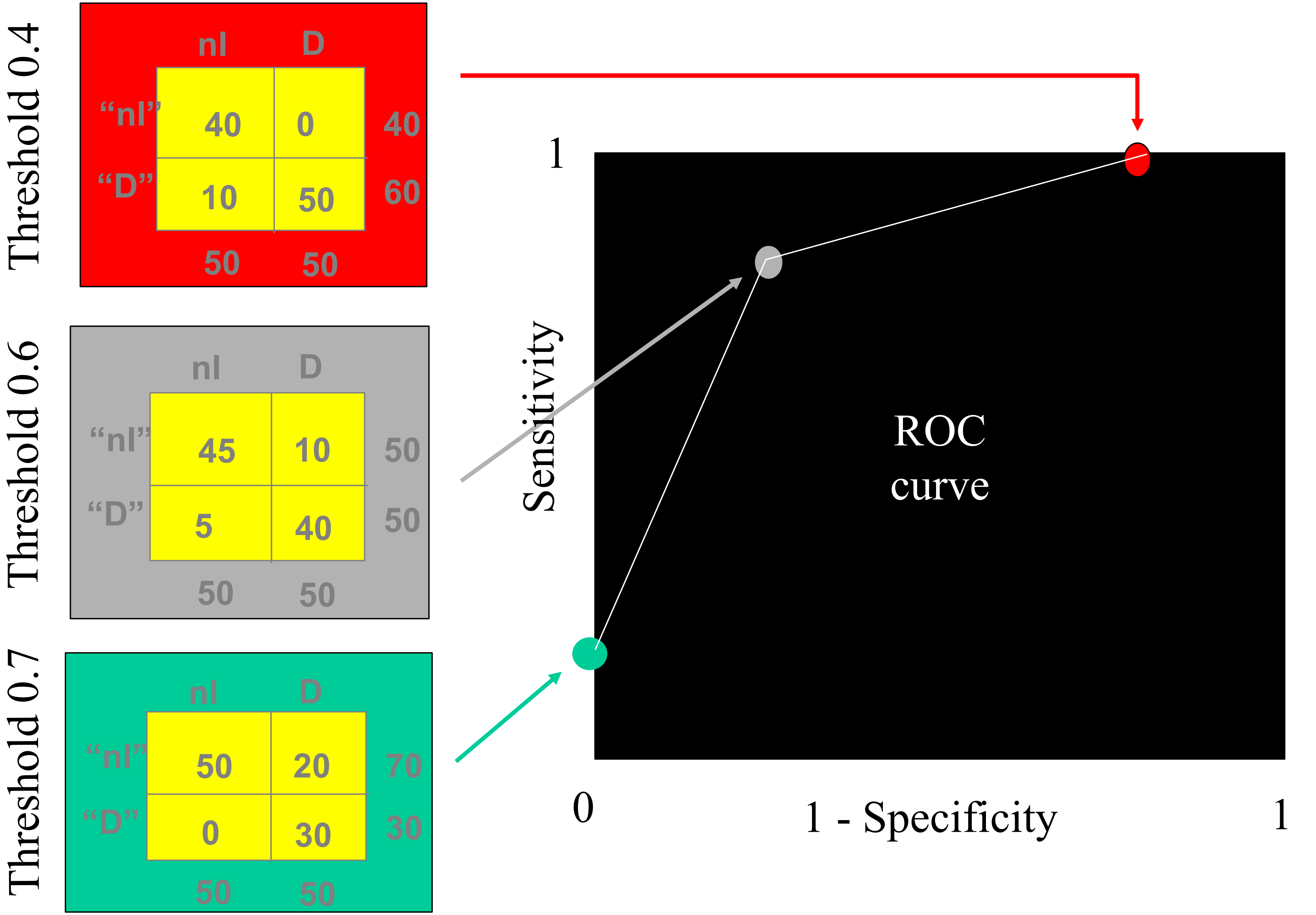
Sensitivity = 40/50 = .8
Specificity = 45/50 = .9



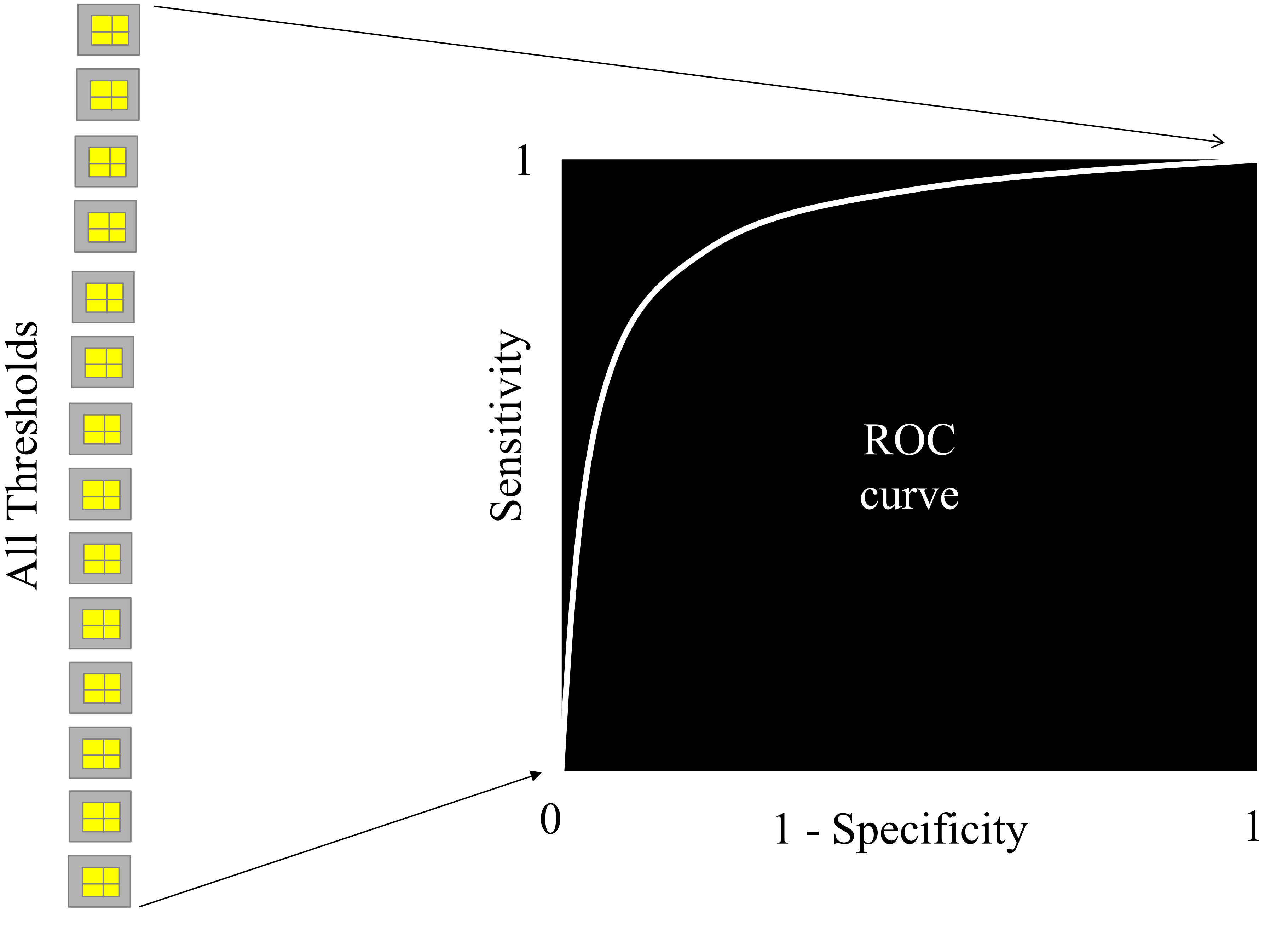
	nl	D	
"nl"	45	10	50
"D"	5	40	50
	50	50	

Sensitivity = 30/50 = .6
Specificity = 1



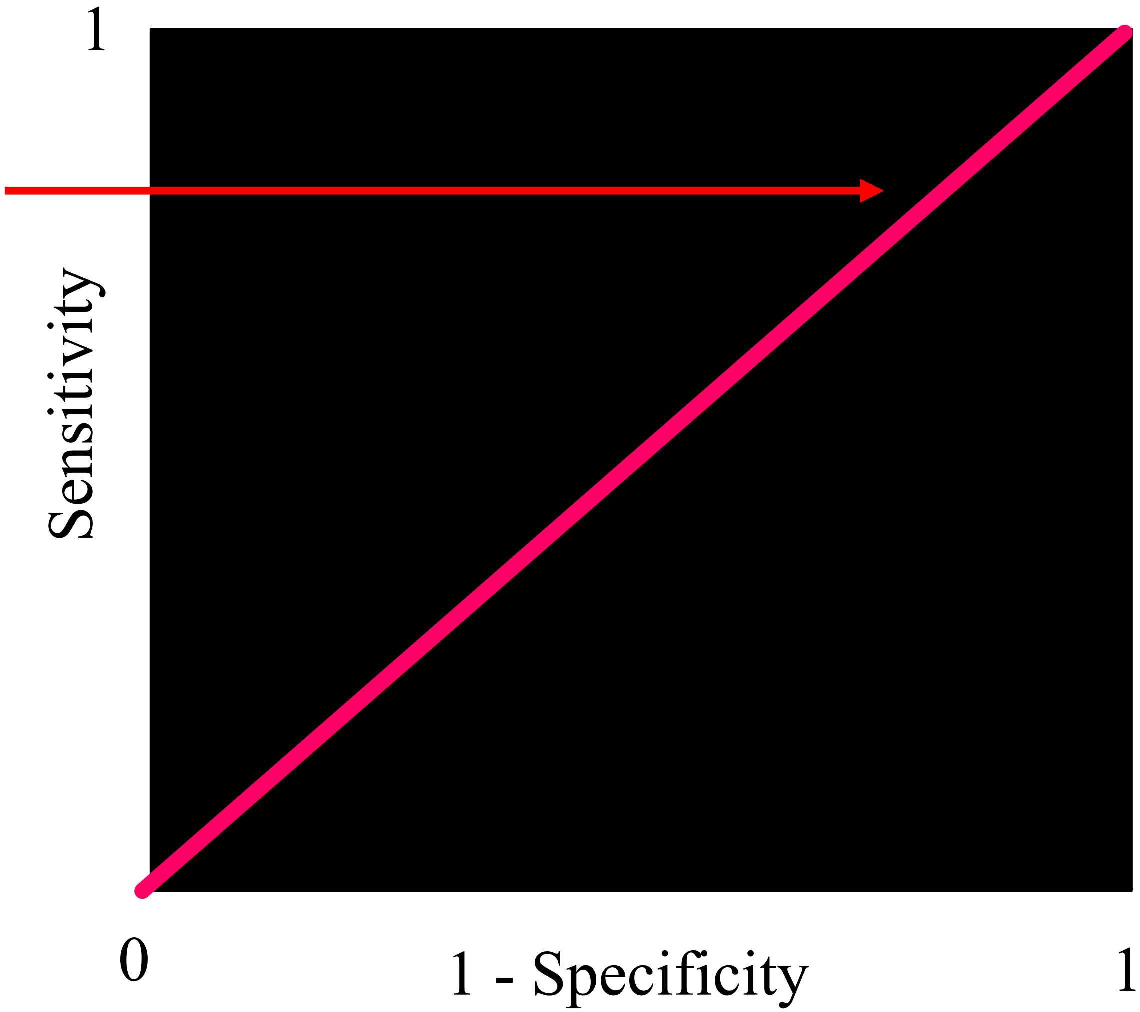


Threshold 0.4
nl
D
"nl"
40
0
40
"D"
10
50
60
50
50
Threshold 0.6
nl
D
"nl"
45
10
50
"D"
5
40
50
50
50
Threshold 0.7
nl
D
"nl"
50
20
70
"D"
0
30
30
50
50
1
Sensitivity
ROC
curve
0
1 - Specificity
1



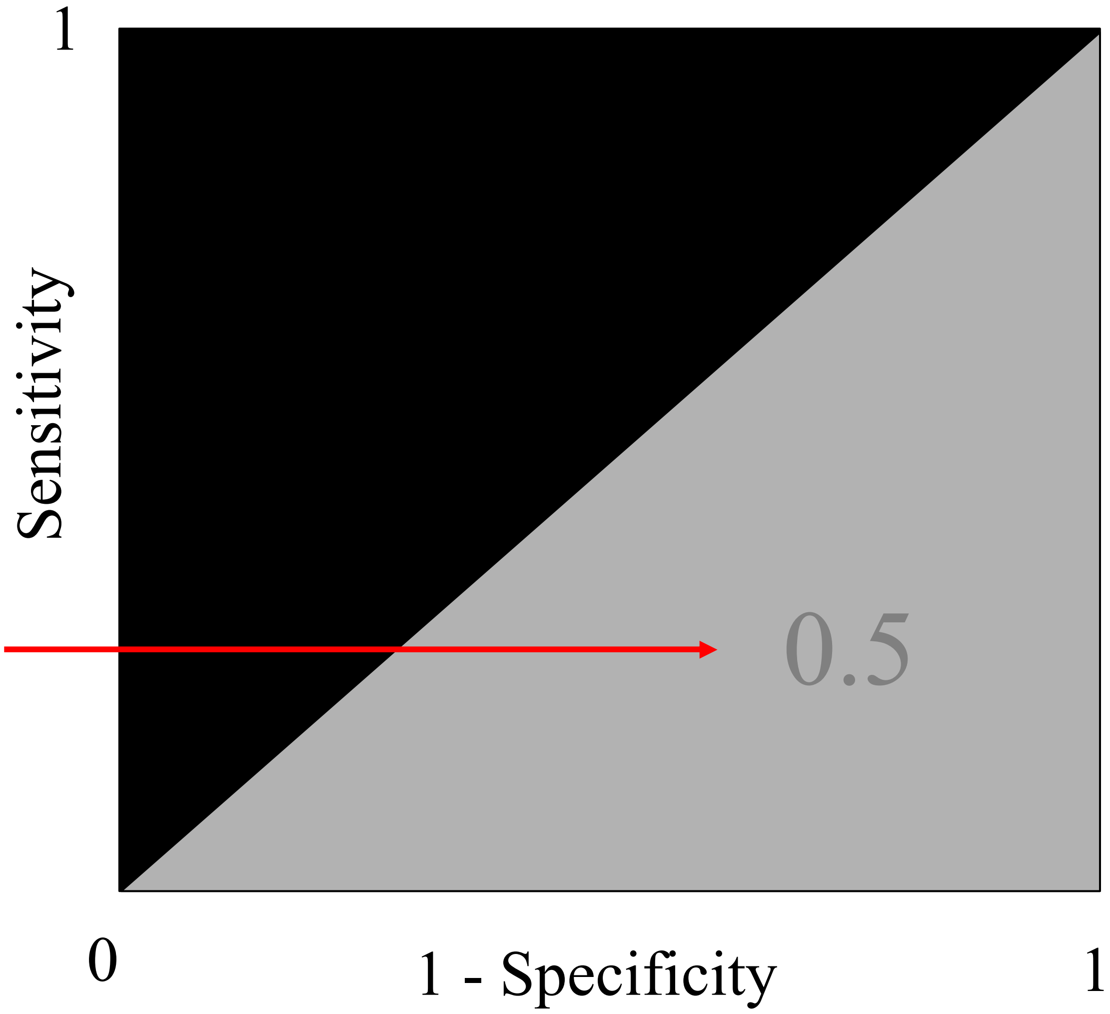
All Thresholds
1
Sensitivity
ROC
curve
0
1 - Specificity
1

45 degree line:
no discrimination

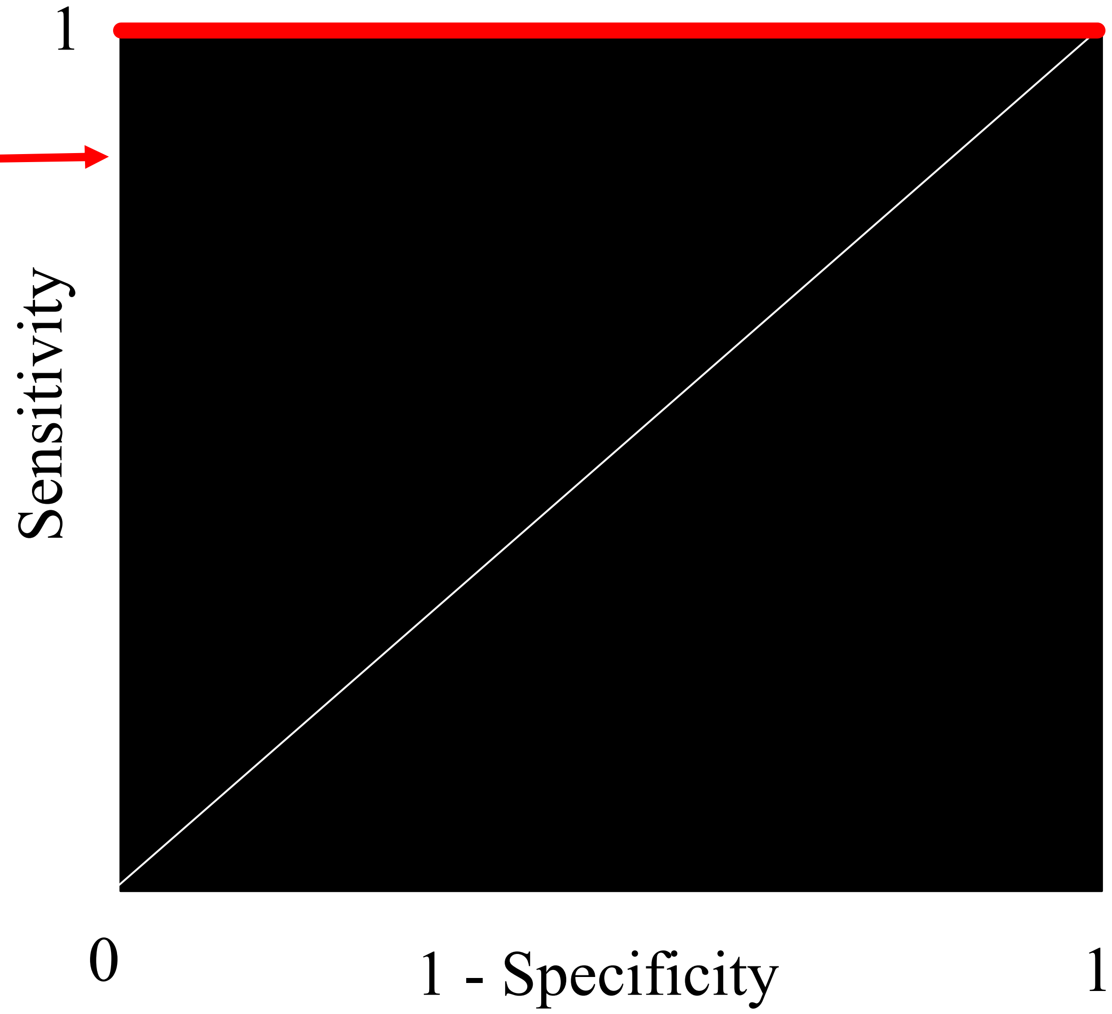


45 degree line:
no discrimination

Area under ROC:

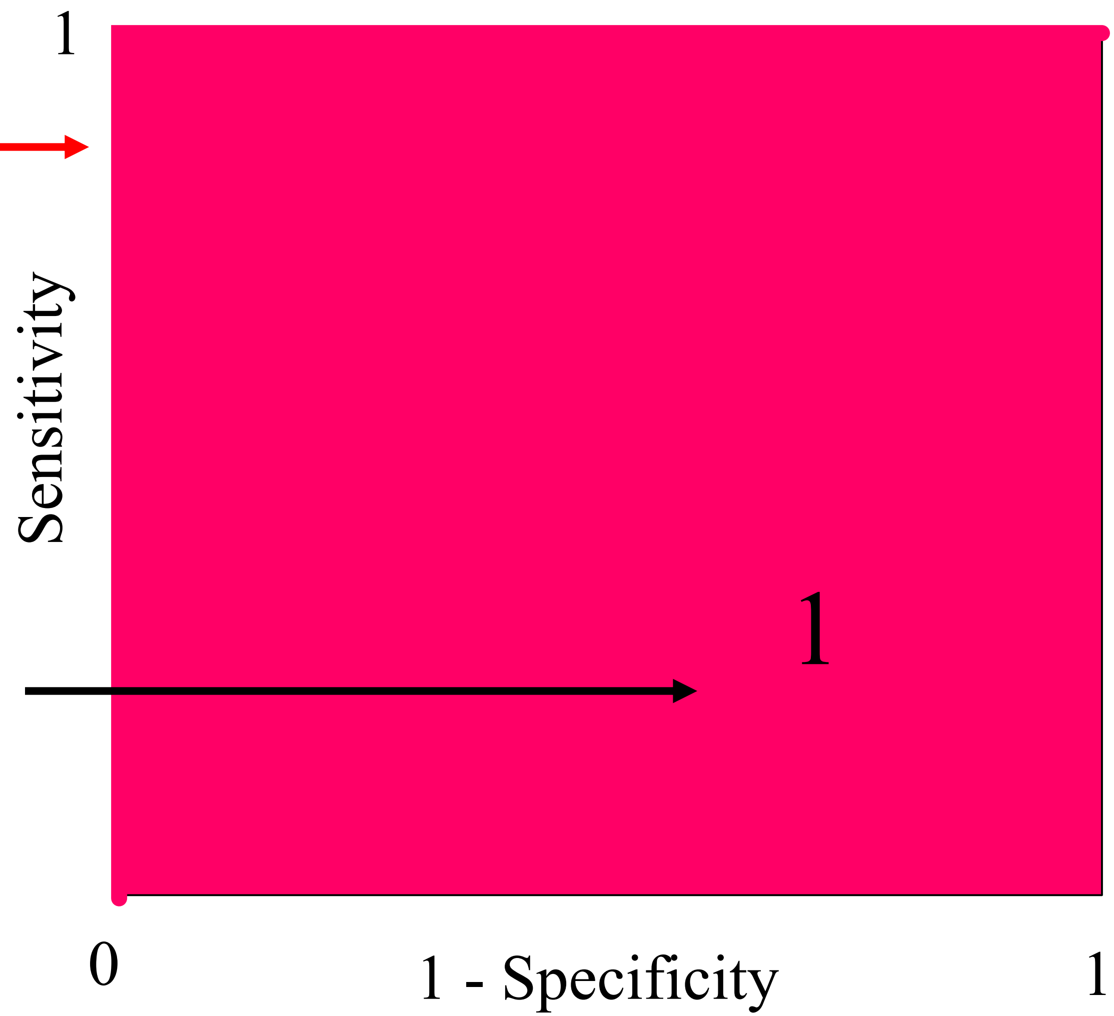


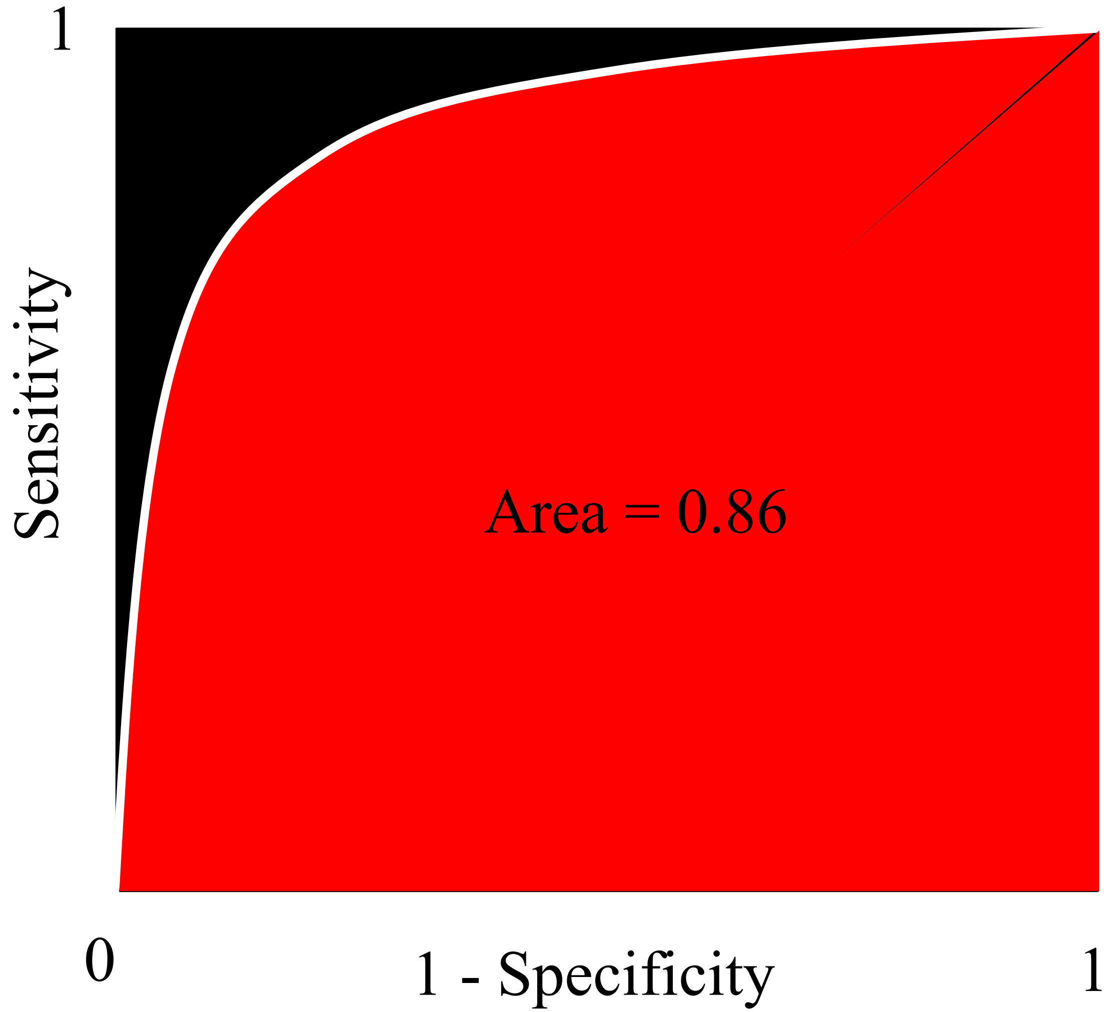
Perfect discrimination



Perfect discrimination

Area under ROC:




1
Sensitivity
Area = 0.86
0
1 - Specificity
1

What is the area under the ROC?

- An estimate of the **discriminatory performance** of the system
  - the real outcome is binary, and systems’ estimates are continuous (0 to 1)
  - all thresholds are considered
- **NOT** an estimate on how many times the system will give the “right” answer
- Usually a good way to describe the discrimination if there is no particular trade-off between false positives and false negatives (unlike in medicine…)
  - Partial areas can be compared in this case

Simplified Example

Systems' estimates for 10 patients

"Probability of being sick"

"Sickness rank"

(5 are healthy, 5 are sick):

0.3
0.2
0.5
0.1
0.7
0.8
0.2
0.5
0.7
0.9

Interpretation of the Area
divide the groups

- Healthy (real outcome is 0)

0.3

0.2

0.5

0.1

0.7

- Sick (real outcome is1)

0.8

0.2

0.5

0.7

0.9

All possible pairs 0-1

Healthy		Sick	
0.3	<	0.8	concordant
0.2		0.2	discordant
0.5		0.5	concordant
0.1		0.7	concordant
0.7		0.9	concordant

All possible pairs 0-1

Systems' estimates for

- Healthy

0.3

0.2

0.5

0.1

0.7

- Sick

0.8	concordant
0.2	tie
0.5	concordant
0.7	concordant
0.9	concordant

C - index

- Concordant

18

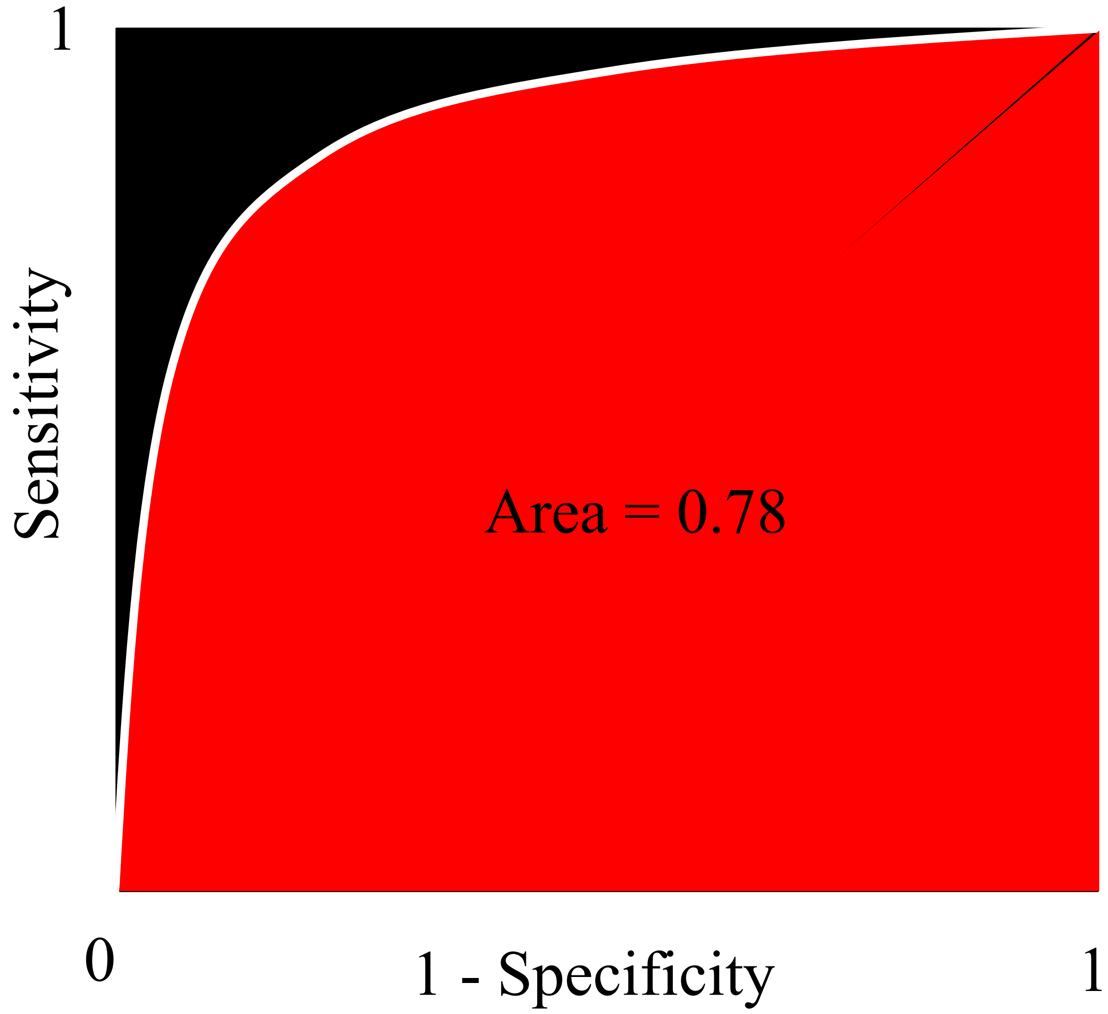
- Discordant

4

- Ties

3

$$\text{C -index} = \frac{\text{Concordant} + 1/2 \text{ Ties}}{\text{All pairs}} = \frac{18 + 1.5}{25}$$


1
Sensitivity
Area = 0.78
0
1 - Specificity
1

Calibration Indices

Discrimination and Calibration

- Discrimination measures how much the system can discriminate between cases with gold standard ‘1’ and gold standard ‘0’
- Calibration measures how close the estimates are to a **“real”** probability
- “If the system is good in discrimination, calibration can be fixed”

Calibration

- System can reliably estimate probability of
  - a diagnosis
  - a prognosis
- Probability is close to the "real" probability

What is the "real" probability?

- Binary events are YES/NO (0/1) i.e., probabilities are 0 or 1 for a given individual
- Some models produce continuous (or quasi-continuous estimates for the binary events)
- Example:
  - Database of patients with spinal cord injury, and a model that predicts whether a patient will ambulate or not at hospital discharge
  - Event is 0: doesn't walk or 1: walks
  - Models produce a probability that patient will walk: 0.05, 0.10, ...

How close are the estimates to the "true" probability for a patient?

- "True" probability can be interpreted as probability within a set of similar patients
- What are similar patients?
  - Clones
  - Patients who look the same (in terms of variables measured)
  - Patients who get similar scores from models
  - How to define boundaries for similarity?

Estimates and Outcomes

- Consider pairs of
  - estimate and true outcome

  0.6 and 1

  0.2 and 0

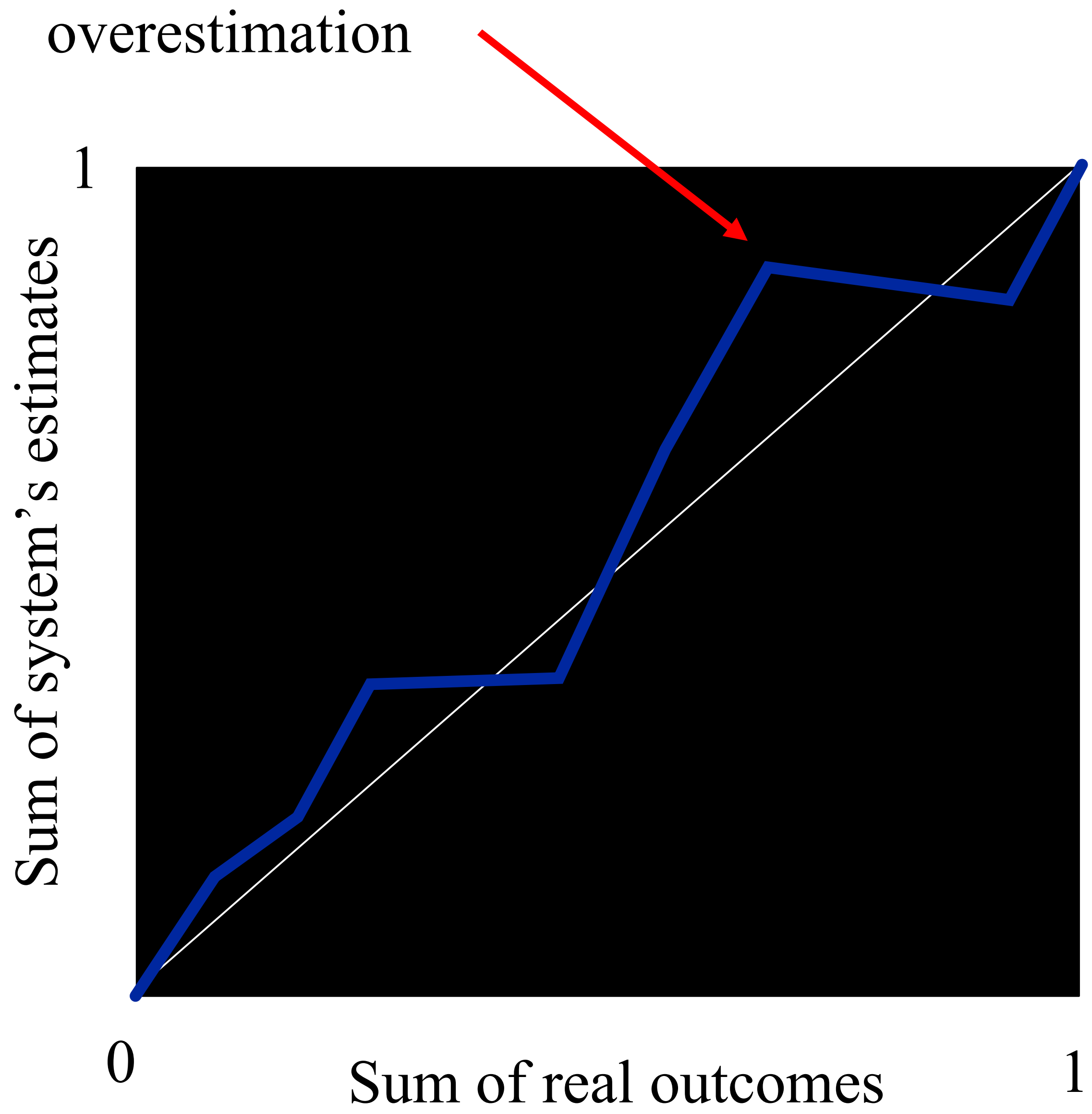
  0.9 and 0

  - And so on…

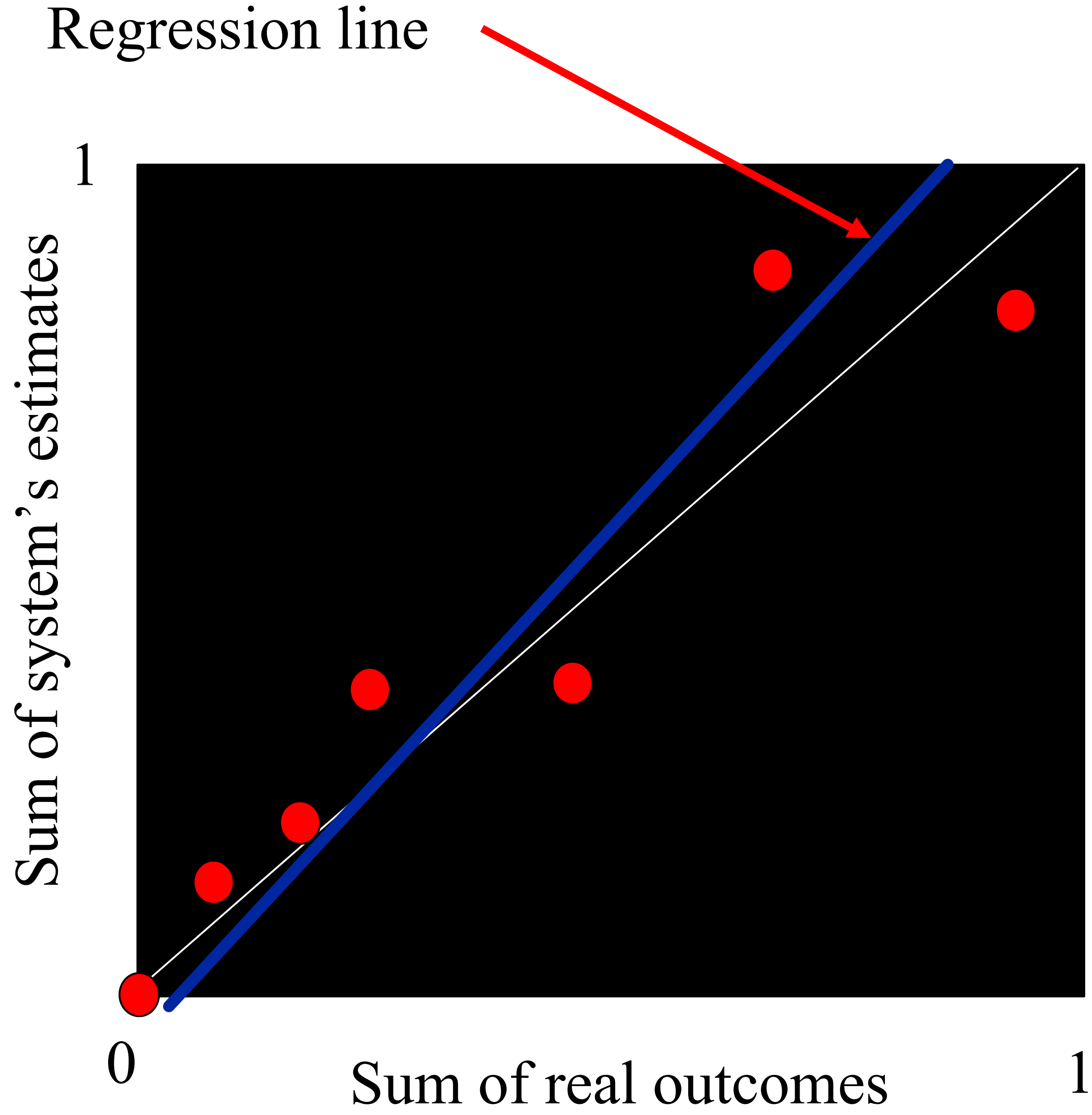
Calibration

Sorted pairs by systems' estimates		Real outcomes	
0.1		**0**	
0.2		**0**	
0.2	**sum of group = 0.5**	**1**	**sum = 1**
0.3		**0**	
0.5		**0**	
0.5	**sum of group = 1.3**	**1**	**sum = 1**
0.7		**0**	
0.7		**1**	
0.8		**1**	
0.9	**sum of group = 3.1**	**1**	**sum = 3**

Calibration Curves



Linear Regression and 45^0 line



Goodness-of-fit

Sort systems' estimates, group, sum, **chi-square**

Estimated		Observed	
0.1		0	
0.2		0	
0.2	sum of group = 0.5	1	sum = 1
0.3		0	
0.5		0	
0.5	sum of group = 1.3	1	sum = 1
0.7		0	
0.7		1	
0.8		1	
0.9	sum of group = 3.1	1	sum = 3

$\chi 2 = \Sigma$ **[(observed - estimated)²/estimated]**

Hosmer-Lemeshow C-hat

Groups based on *n*-iles (e.g., terciles), *n*-2 d.f. training, *n* d.f. test

Measured Groups Estimated	Observed	**"Mirror groups"** Estimated	Observed
0.1	**0**	**0.9**	**1**
0.2	**0**	**0.8**	**1**
0.2 sum = 0.5	**1 sum = 1**	**0.8 sum = 2.5**	**0 sum = 2**
0.3	**0**	**0.7**	**1**
0.5	**0**	**0.5**	**1**
0.5 sum = 1.3	**1 sum = 1**	**0.5 sum = 1.7**	**0 sum = 2**
0.7	**0**	**0.3**	**1**
0.7	**1**	**0.3**	**0**
0.8	**1**	**0.2**	**0**
0.9 sum = 3.1	**1 sum = 3**	**0.1 sum=0.9**	**0 sum = 1**

Hosmer-Lemeshow H-hat

Groups based on *n* fixed thresholds (e.g., 0.3, 0.6, 0.9), *n*-2 d.f.

Measured Groups Estimated	Observed	**"Mirror groups"** Estimated	Observed
0.1	**0**	**0.9**	**1**
0.2	**0**	**0.8**	**1**
0.2	**1**	**0.8**	**0**
0.3 sum = 0.8	**0 sum = 1**	**0.7 sum = 3.2**	**1 sum = 2**
0.5	**0**	**0.5**	**1**
0.5 sum = 1.0	**1 sum = 1**	**0.5 sum = 1.0**	**0 sum = 1**
0.7	**0**	**0.3**	**1**
0.7	**1**	**0.3**	**0**
0.8	**1**	**0.2**	**0**
0.9 sum = 3.1	**1 sum = 3**	**0.1 sum=0.9**	**0 sum = 1**

Covariance decomposition

- Arkes et al, 1995

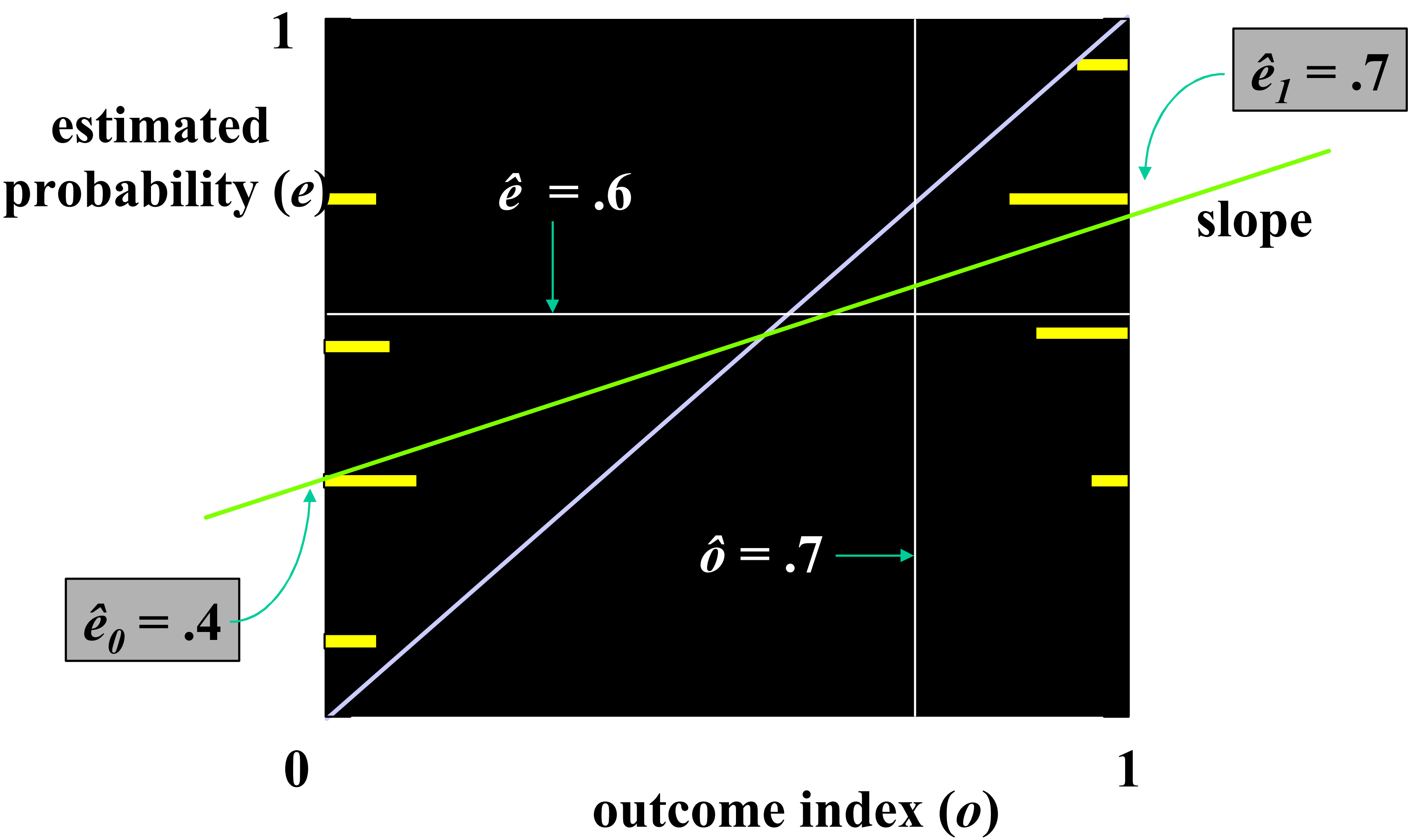
Brier =

$d(1-d) + bias^2 + d(1-d)slope(slope-2) + scatter$

- where d = prior
- bias is a calibration index
- slope is a discrimination index
- scatter is a variance index

Covariance Graph

PS= .2 bias= -0.1 slope= .3 scatter= .1



Logistic and Score Models for MACE

	Logistic Regression Model: Odds Ratio	Risk Score Model: Risk Value
Age > 74yrs	1.42	0
B2/C Lesion	2.44	2
Acute MI	2.94	2
Class 3/4 CHF	3.56	3
Left main PCI	2.34	2
IIb/IIIa Use	1.43	0
Stent Use	0.56	-1
Cardiogenic Shock	3.68	3
USA	2.60	2
Tachycardic	1.34	0
No Reflow	2.73	2
Unscheduled	1.48	0
Chronic Renal Insuff.	1.64	1

Model Performance

Development Set (2804 consecutive cases) 1/97-2/99
Validation Set (1460 consecutive cases) 3/99-12/99

	Death	MACE
Multiple Logistic Regression		
c-Index Training Set	0.880	0.806
c-Index Test Set	0.898	0.851
c-Index Validation Set	0.840	0.787
Prognostic Score Model		
c-Index Training Set	0.882	0.798
c-Index Test Set	0.910	0.846
c-Index Validation Set	0.855	0.780
Artificial Neural Network		
c-Index Training Set	0.950	0.849
c-Index Test Set	0.930	0.870
c-Index Validation Set	0.835	0.811

Model Performance

Validation Set: 1460 consecutive cases 3/1/99-12/31/99

	Death	MACE
Multiple Logistic Regression		
c-Index Validation Set	**0.840**	**0.787**
Hosmer-Lemeshow	16.07*	24.40*
c-Index Test Set	0.898	0.851
Prognostic Score Models		
c-Index Validation Set	**0.855**	**0.780**
Hosmer-Lemeshow	11.14*	10.66*
c-Index Test Set	0.910	0.846
Artificial Neural Networks		
c-Index Validation Set	**0.835**	**0.811**
Hosmer-Lemeshow	7.17*	20.40*
c-Index Test Set	0.930	0.870

* indicates adequate goodness of fit (prob >0.5)

Conclusions

- **In this data set, the use of stents and gp IIb/IIIa antagonists are associated with a decreased risk of in-hospital death.**
- **Prognostic risk score models offer advantages over complex modeling systems.**
  - **Simple to comprehend and implement**
  - **Discriminatory power approaching full LR and aNN models**
- **Limitations of this investigation include:**
  - **the restricted scope of covariates available**
  - **single high volume center’s experience limiting generalizability**

Example

Comparison of Practical Prediction Models for Ambulation Following Spinal Cord Injury

Todd Rowland, M.D.

Decision Systems Group

Brigham and Womens Hospital

Study Rationale

- Patient's most common question: "Will I walk again"
- Study was conducted to compare logistic regression , neural network, and rough sets models which predict ambulation at discharge based upon information available at admission for individuals with acute spinal cord injury.
- Create simple models with good performance

- 762 cases training set
- 376 cases test set
  - univariate statistics compared to make sure sets were similar (e.g., means)

SCI Ambulation Classification System

Admission Info (9 items)

system days
injury days
age
gender
racial/ethnic group
level of neurologic fxn
ASIA impairment index
UEMS
LEMS

Ambulation (1 item)

Yes - 1
No - 0

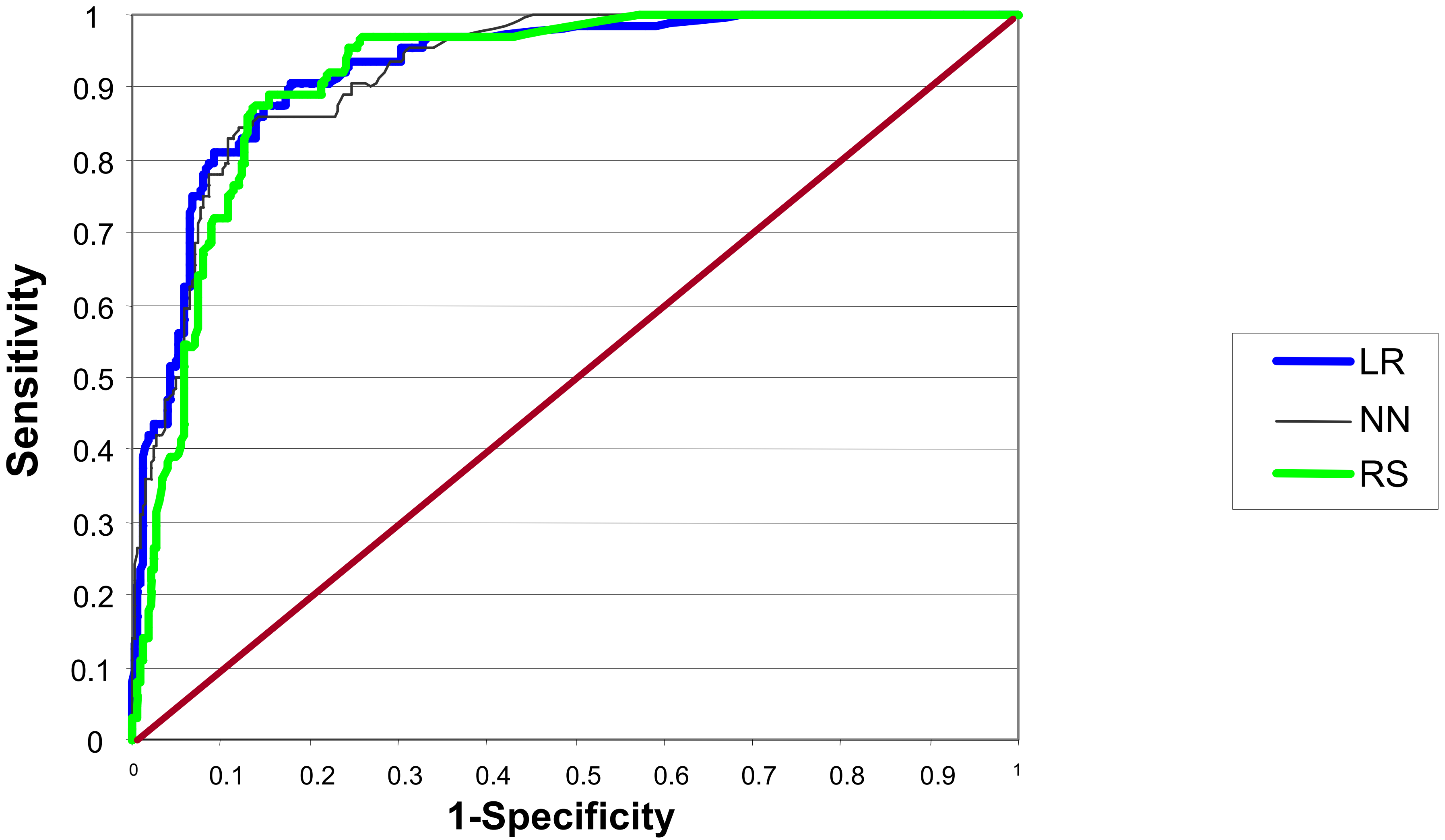
Thresholded Results

	Sens	Spec	NPV	PPV	Accuracy
• LR	0.875	0.853	0.971	0.549	0.856
• NN	0.844	0.878	0.965	0.587	0.872
• RS	0.875	0.862	0.971	0.566	0.864

Brier Scores

	Brier
• LR	0.0804
• NN	0.0811
• RS	0.0883

ROC Curves

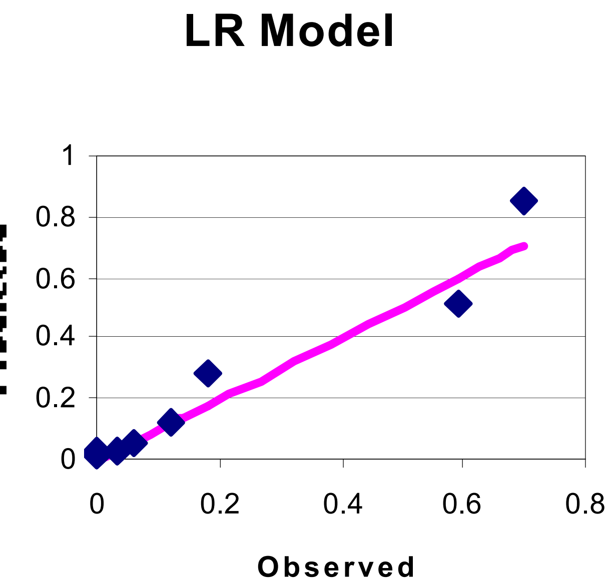


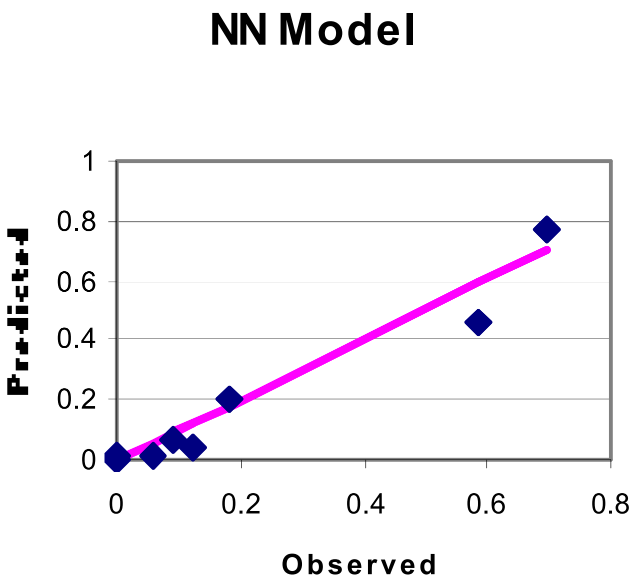
Sensitivity
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1-Specificity
LR
NN
RS

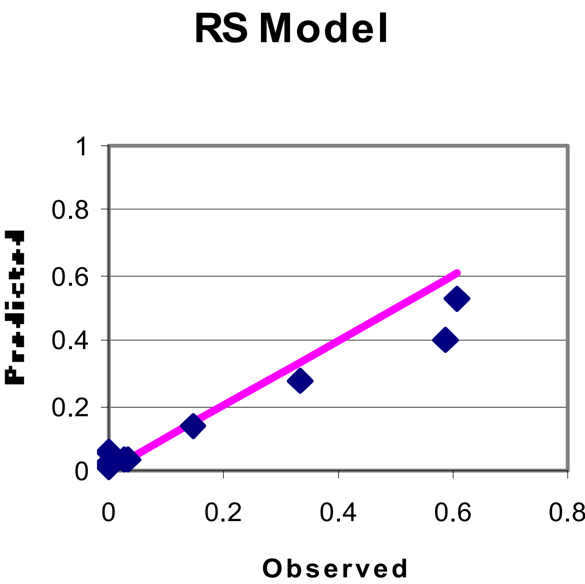
Areas under ROC Curves

Model	ROC Curve Area	Standard Error
Logistic Regression	0.925	0.016
Neural Network	0.923	0.015
Rough Set	0.914	0.016

Calibration curves







Results: Goodness-of-fit

- Logistic Regression: H-L $p = 0.50$
- Neural Network: H-L $p = 0.21$
- Rough Sets: H-L $p < .01$
- $p > 0.05$ indicates reasonable fit

Conclusion

- For the example, logistic regression seemed to be the best approach, given its simplicity and good performance
- Is it enough to assess discrimination and calibration in one data set?