Harvard-MIT Division of Health Sciences and Technology
HST.951J: Medical Decision Support, Fall 2005
Instructors: Professor Lucila Ohno-Machado and Professor Staal Vinterbo

6.873/HST.951 Medical Decision Support

Spring 2004

Evaluation

Lucila Ohno-Machado

Outline

Calibration and Discrimination

- AUCs
- H-L statistic

Strategies:

- Cross-validation
- Bootstrap

- Decomposition of error
  - Bias
  - Variance

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.]

Dataset: Attributes Collected

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

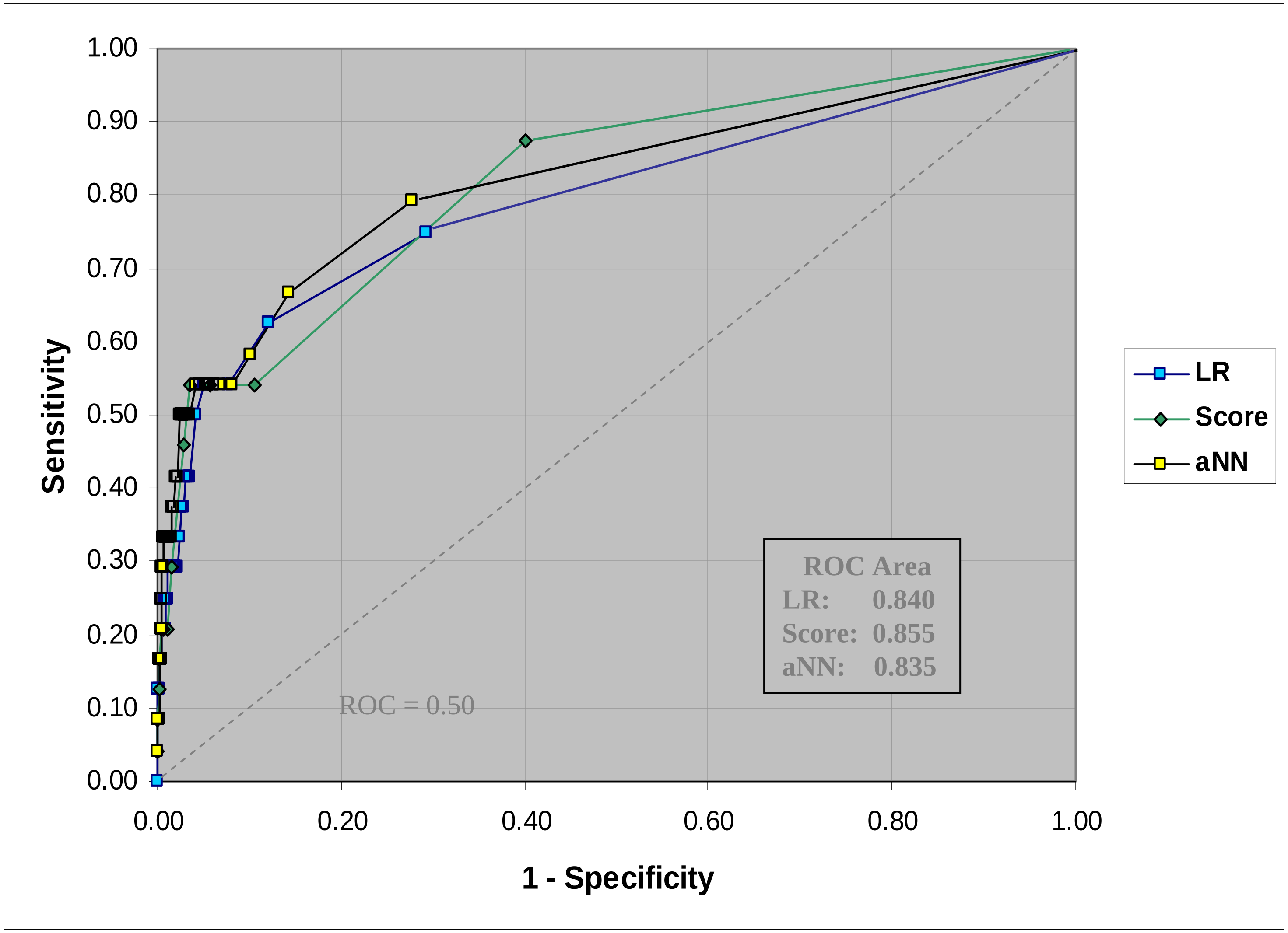
Data Source:
- Medical Record
- Clinician Derived
- Other

Study Population

	Development Set 1/97-2/99		Validation Set 3/99-12/99		
Cases	2,804		1,460		
Women	909	(32.4%)	433	(29.7%)	p=.066
Age > 74yrs	595	(21.2%)	308	(22.5%)	p=.340
Acute MI	250	(8.9%)	144	(9.9%)	p=.311
Primary	156	(5.6%)	95	(6.5%)	p=.214
Shock	62	(2.2%)	20	(1.4%)	p=.058
Class 3/4 CHF	176	(6.3%)	80	(5.5%)	p=.298
gp IIb/IIIa antagonist	1,005	(35.8%)	777	(53.2%)	p<.001
Death	67	(2.4%)	24	(1.6%)	p=.110
Death, MI, CABG (MACE)	177	(6.3%)	96	(6.6%)	p=.739

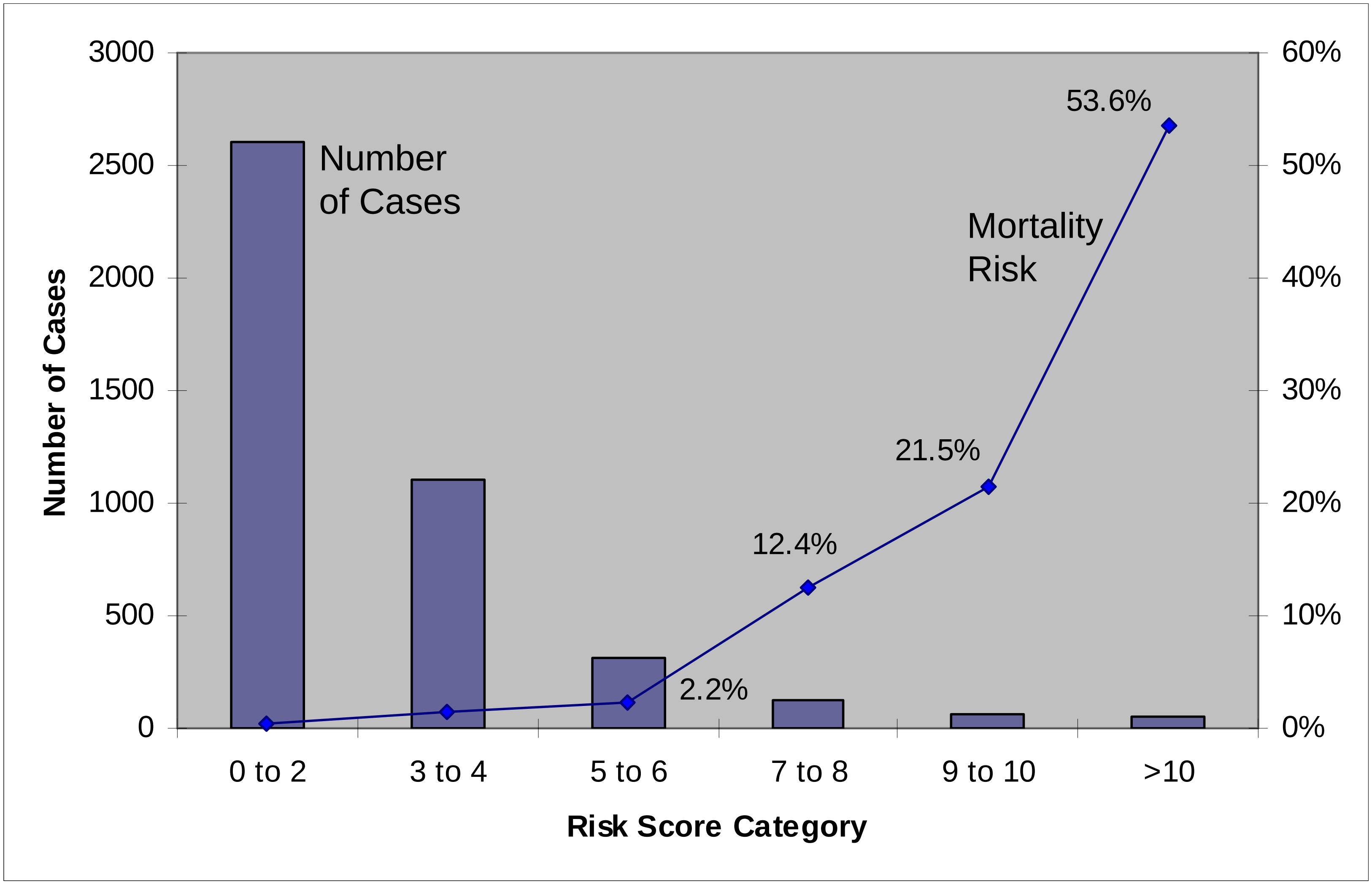
ROC Curves: Death Models

Validation Set: 1460 Cases



Risk Score of Death: BWH Experience

Unadjusted Overall Mortality Rate = 2.1%



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

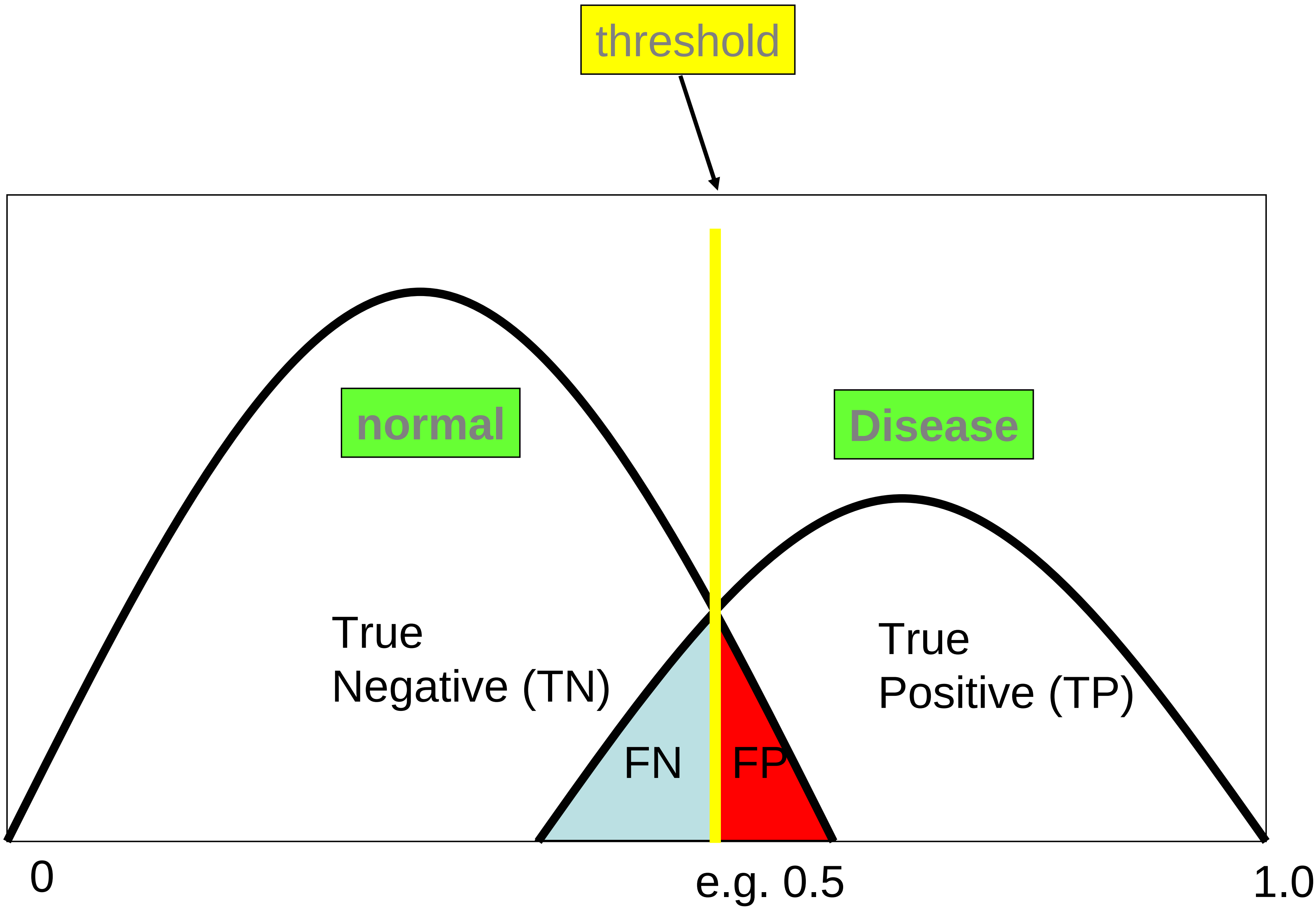
- The system can “somehow” differentiate between cases in different categories
- Binary outcome is a special case:
  - diagnosis (differentiate sick and healthy individuals)
  - prognosis (differentiate poor and good outcomes)

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)

Estimate	"True"
0.3	0
0.2	0
0.5	1
0.1	0

- 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

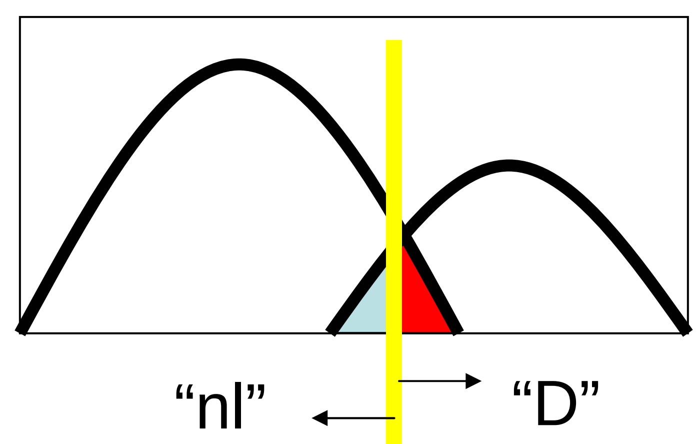
Spec = TN/TN+FP

PPV = TP/TP+FP

NPV = TN/TN+FN

Accuracy = TN +TP

	nl	D
“nl”	TN	FN
“D”	FP	TP
	-	+



Sens = TP/TP+FN

40/50 = .8

Spec = TN/TN+FP

45/50 = .9

PPV = TP/TP+FP

40/45 = .89

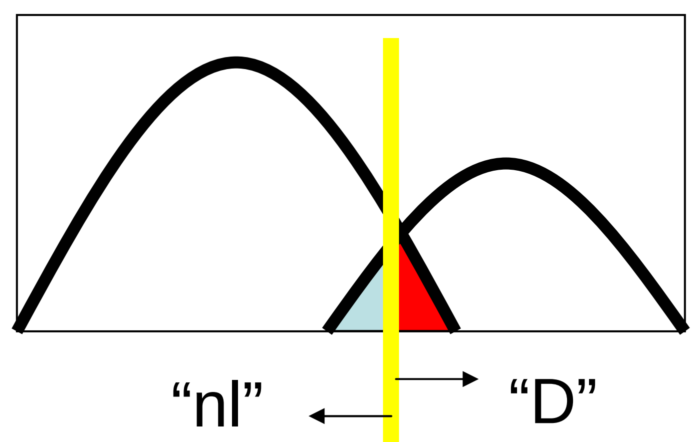
NPV = TN/TN+FN

45/55 = .81

Accuracy = TN +TP

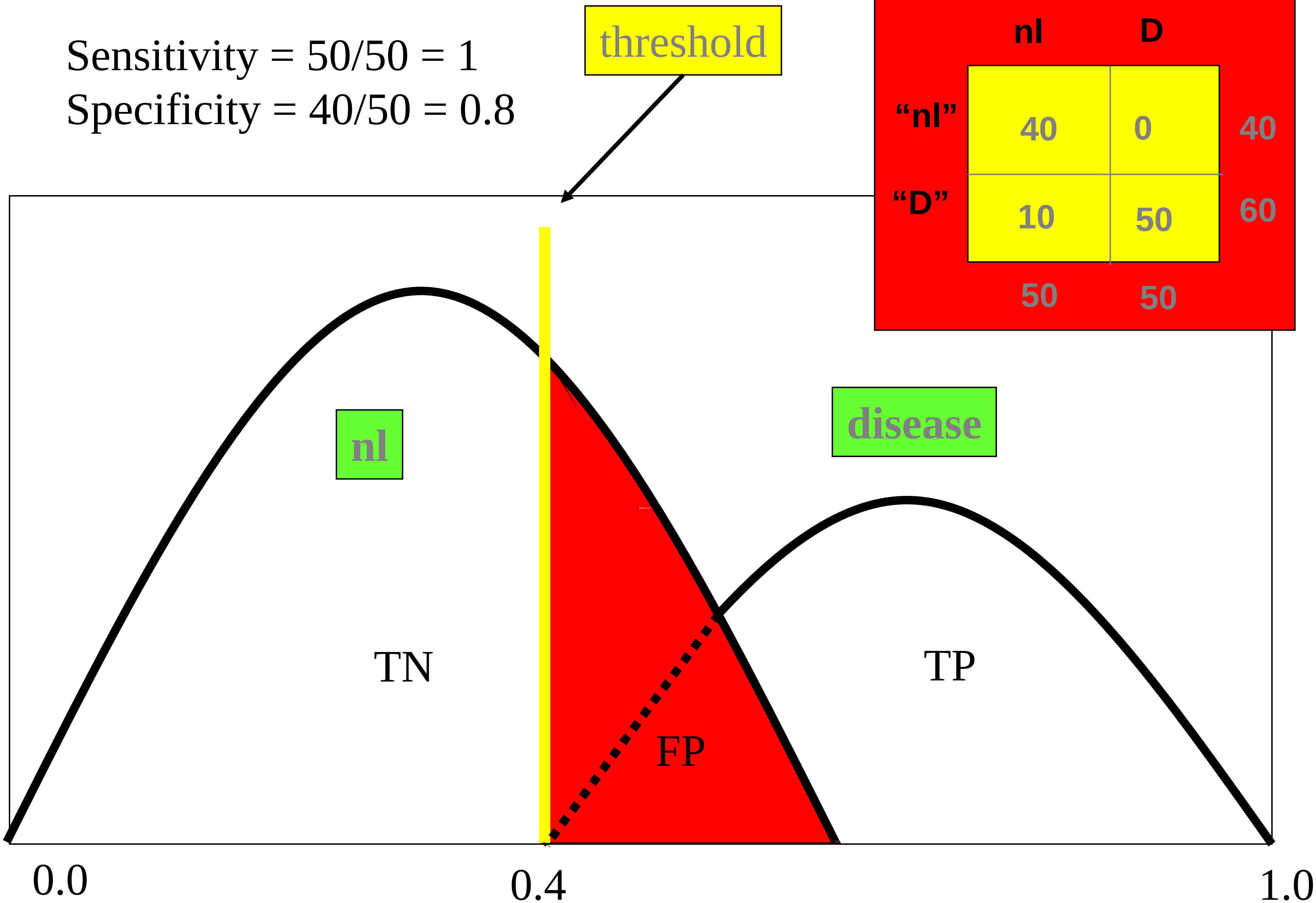
85/100 = .85

	nl	D
"nl"	**45**	**10**
"D"	**5**	**40**



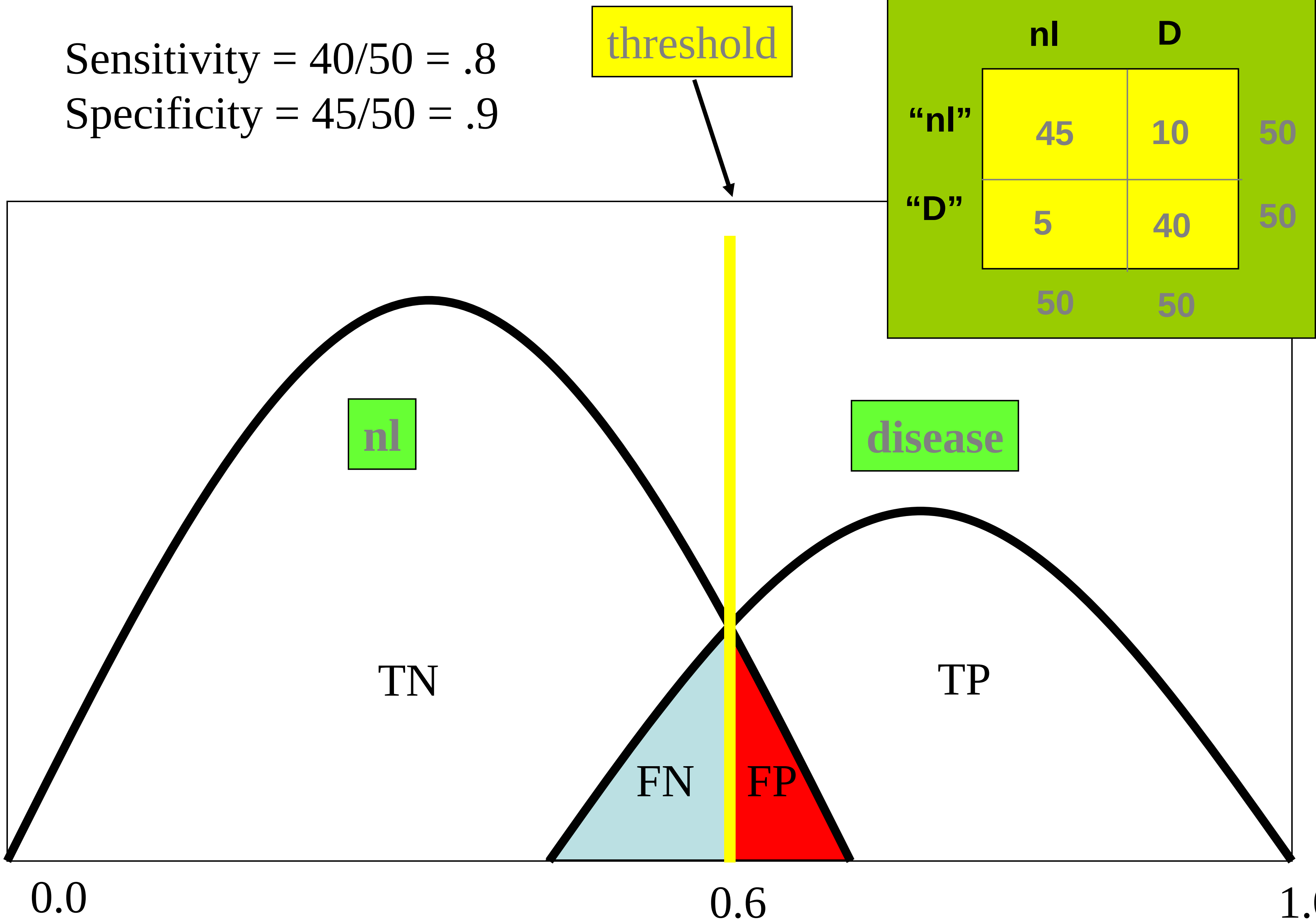
Sensitivity = 50/50 = 1

Specificity = 40/50 = 0.8



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

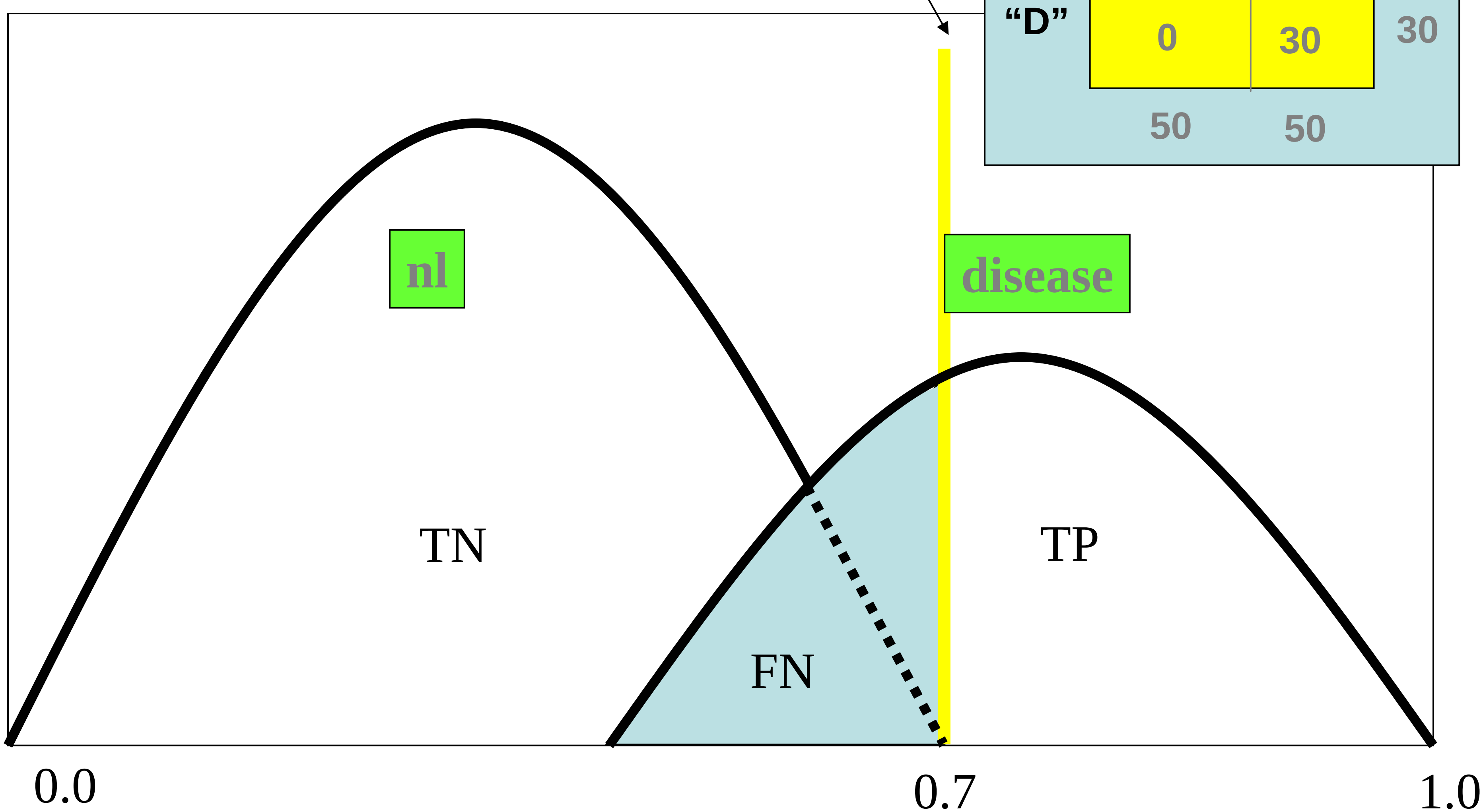
Sensitivity = 40/50 = .8
Specificity = 45/50 = .9



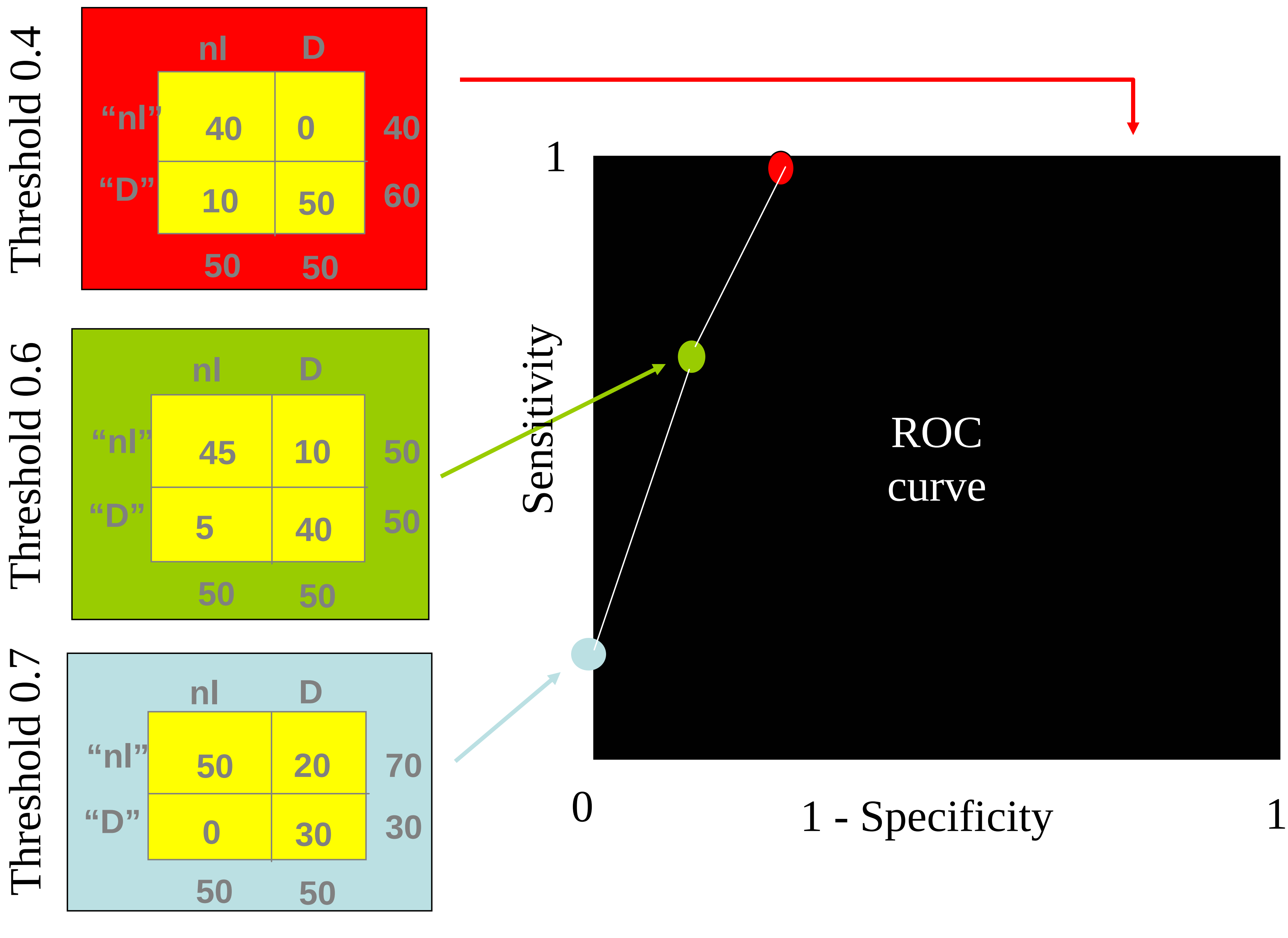
	nl	D	
“nl”	45	10	50
“D”	5	40	50
	50	50	

Sensitivity = 30/50 = .6

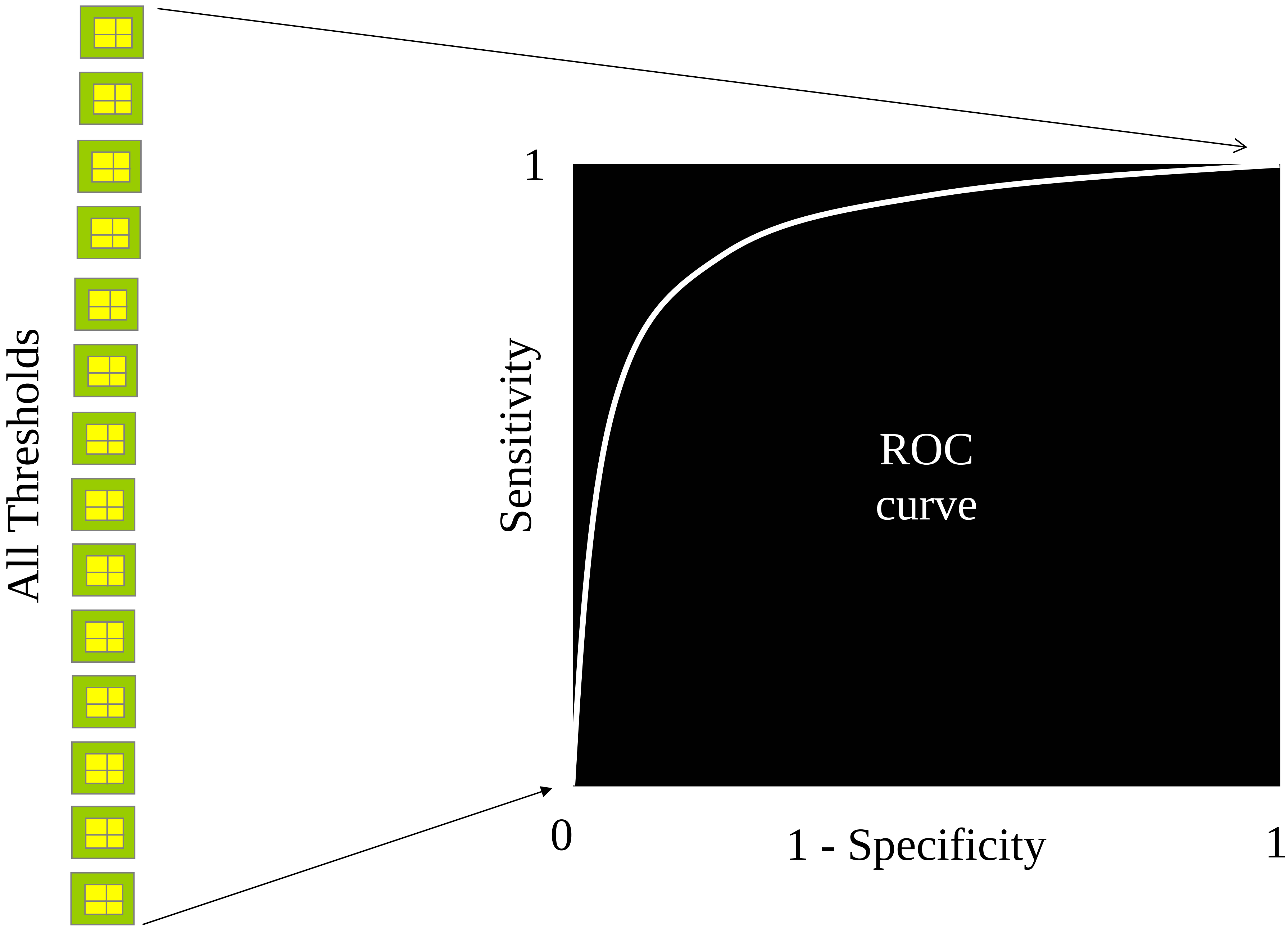
Specificity = 1



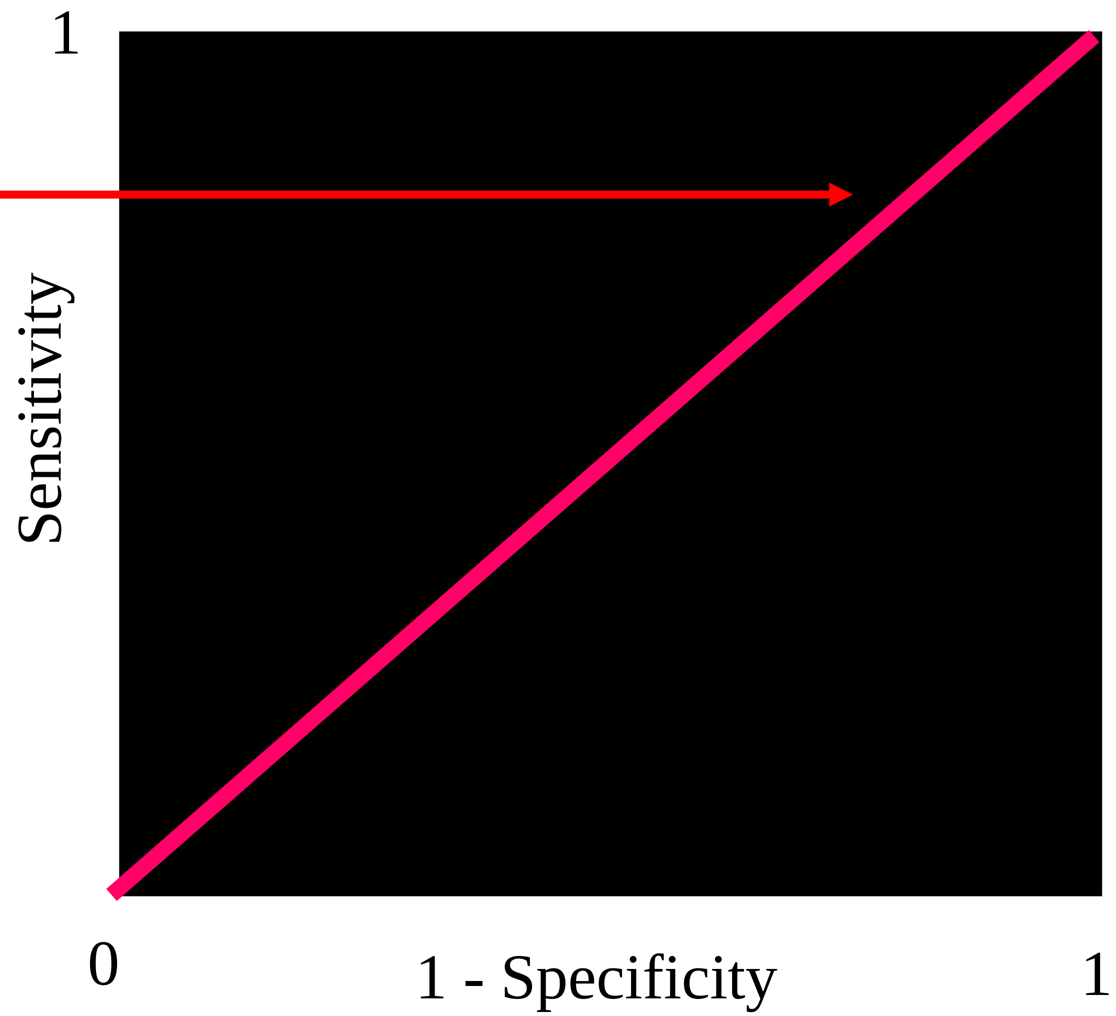
	nl	D	
"nl"	50	20	70
"D"	0	30	30
	50	50	



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

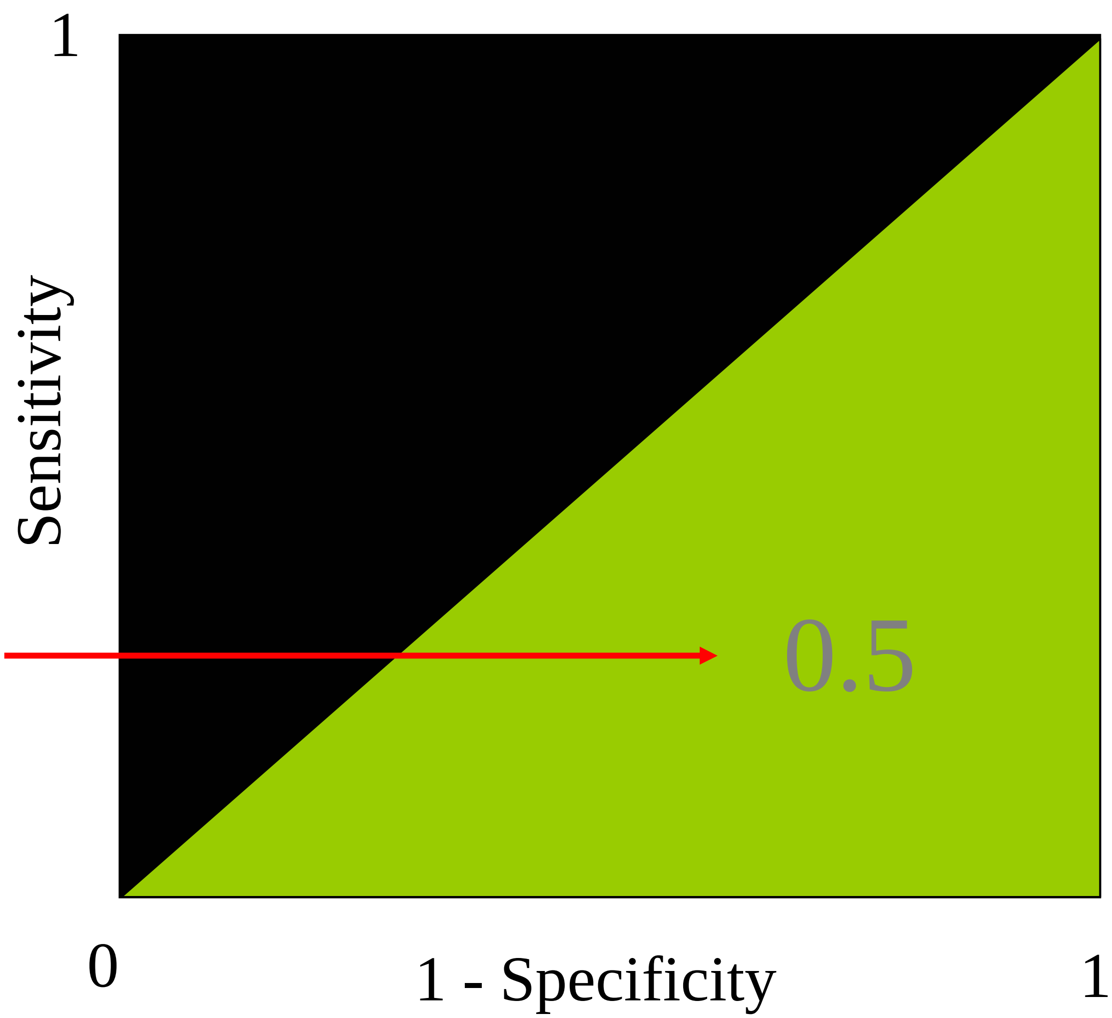

All Thresholds
1
Sensitivity
ROC
curve
0
1 - Specificity
1

45 degree line:
no discrimination

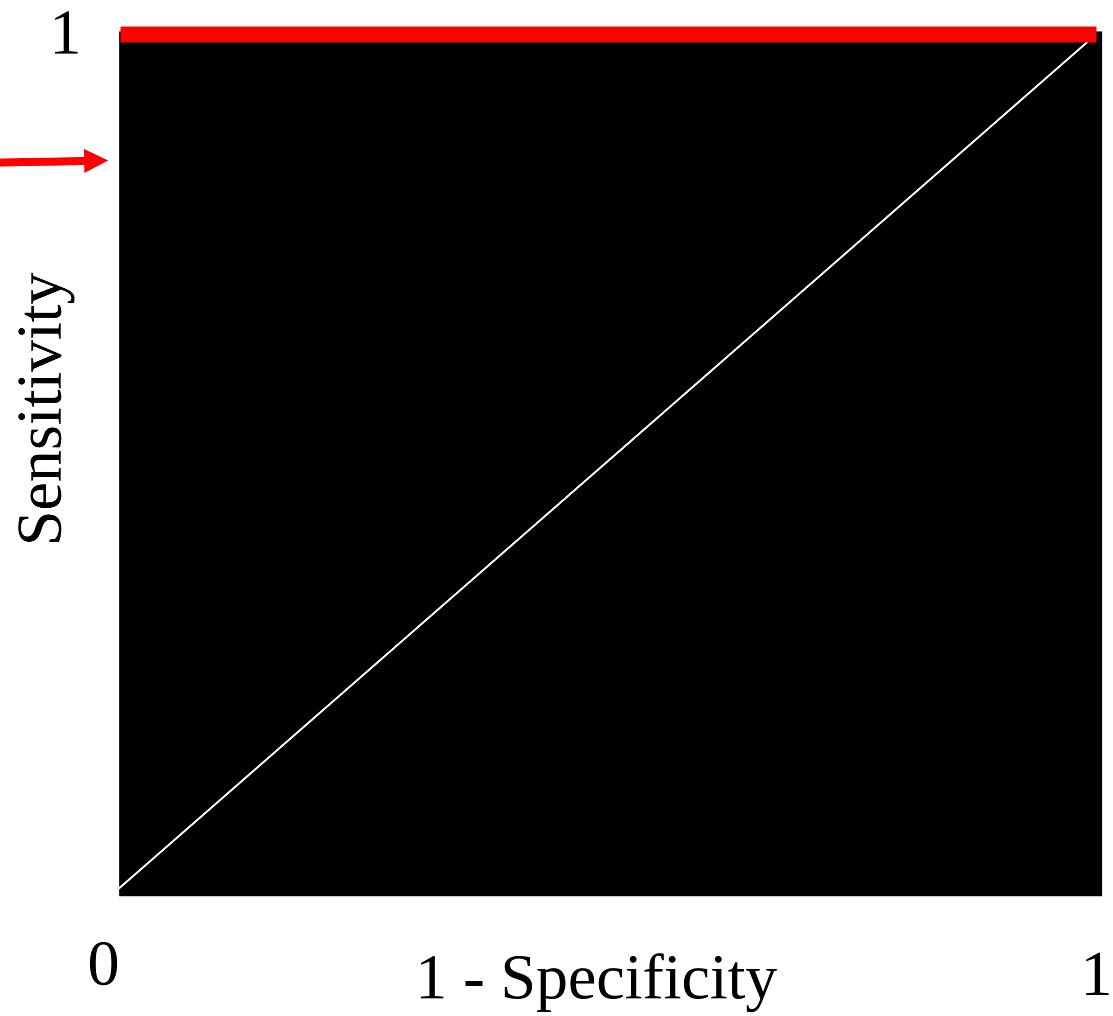


45 degree line:
no discrimination

Area under ROC:

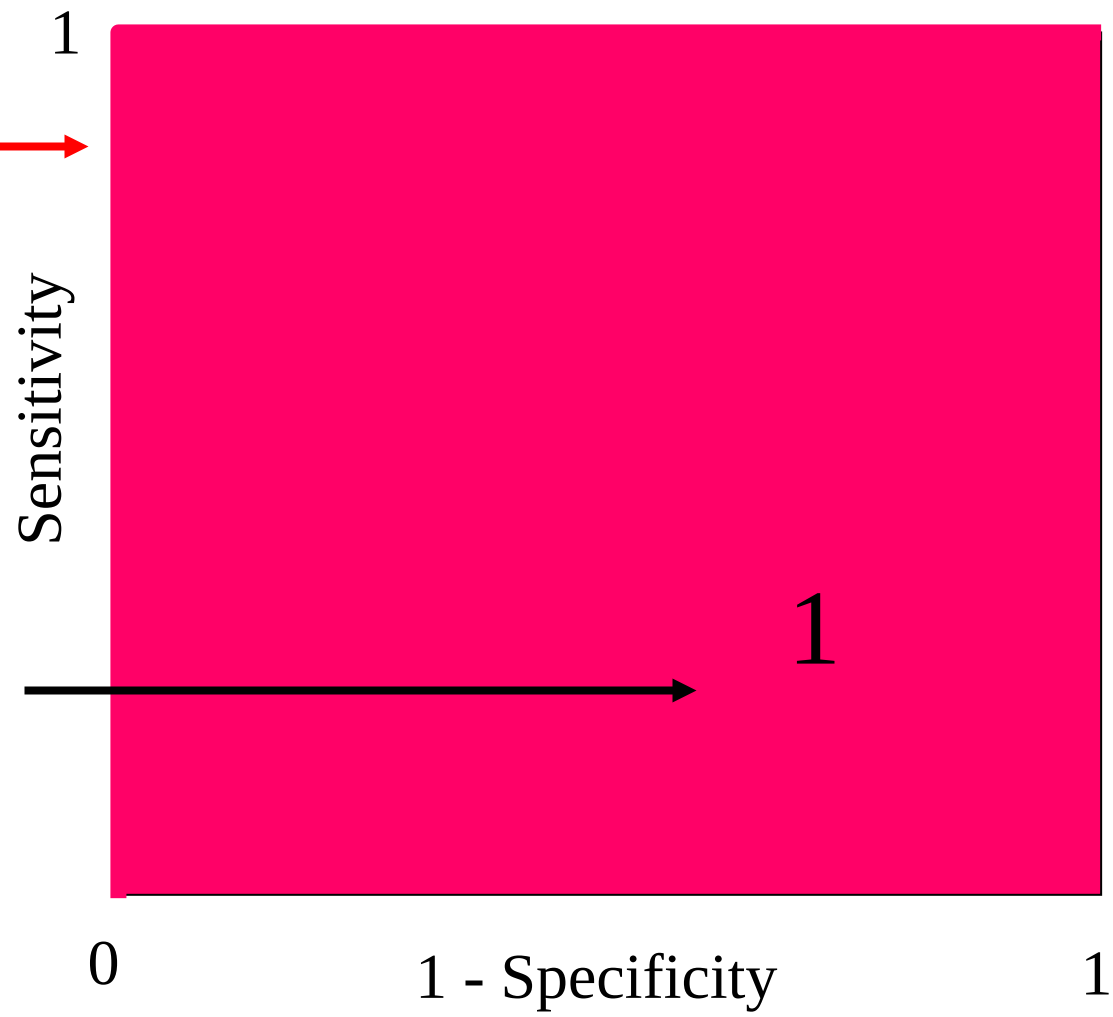


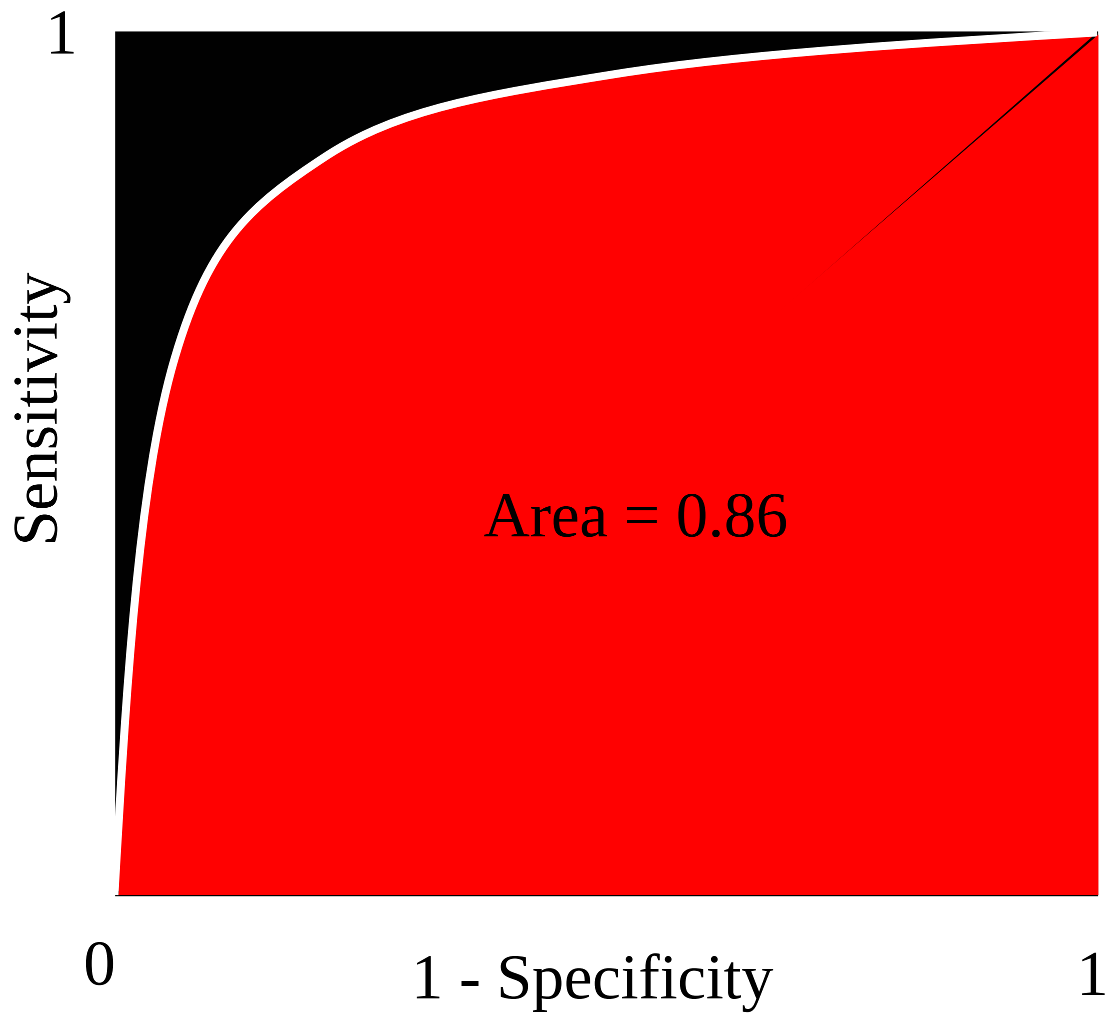
Perfect discrimination



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
- 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

Estimates per class

- 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

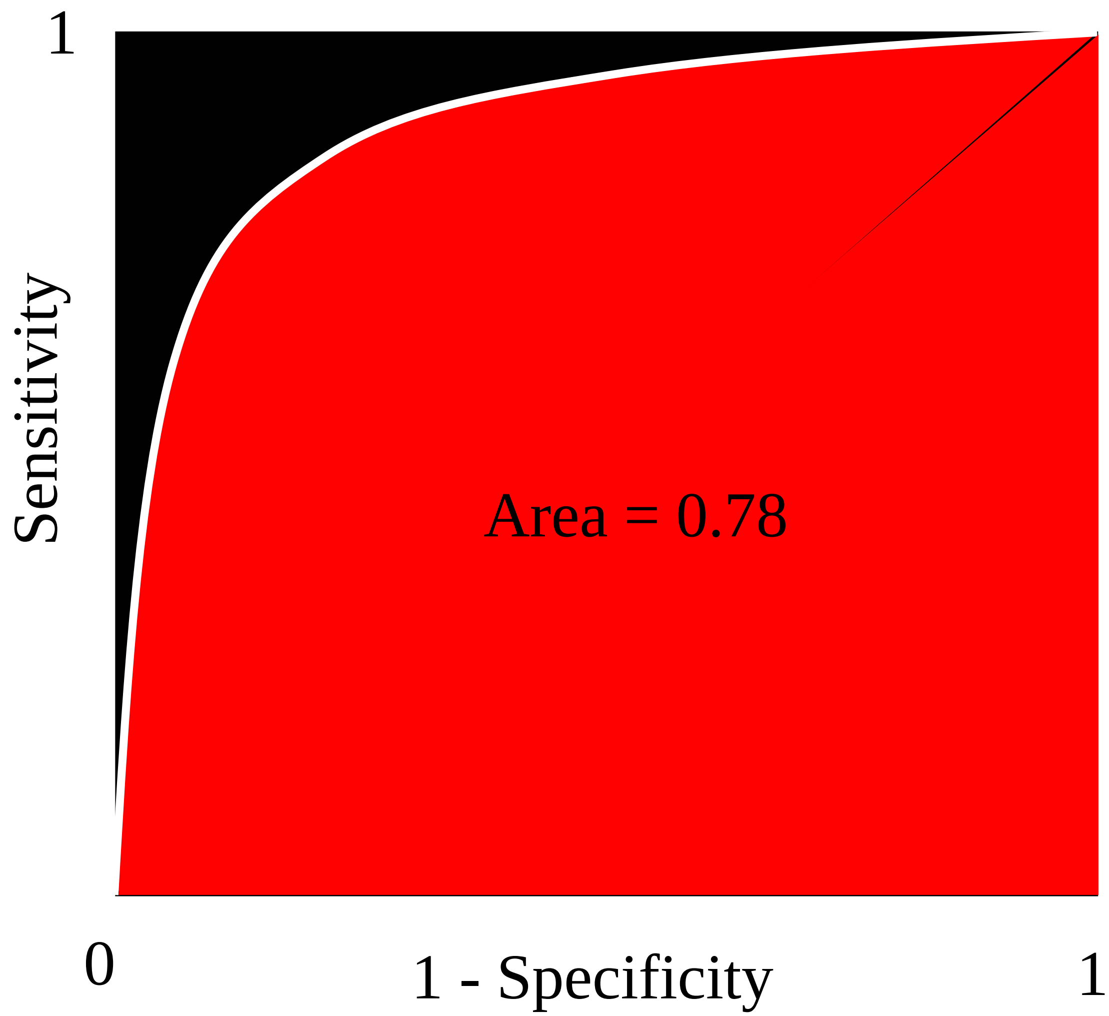
C - index

- Concordant
18

- Discordant
4

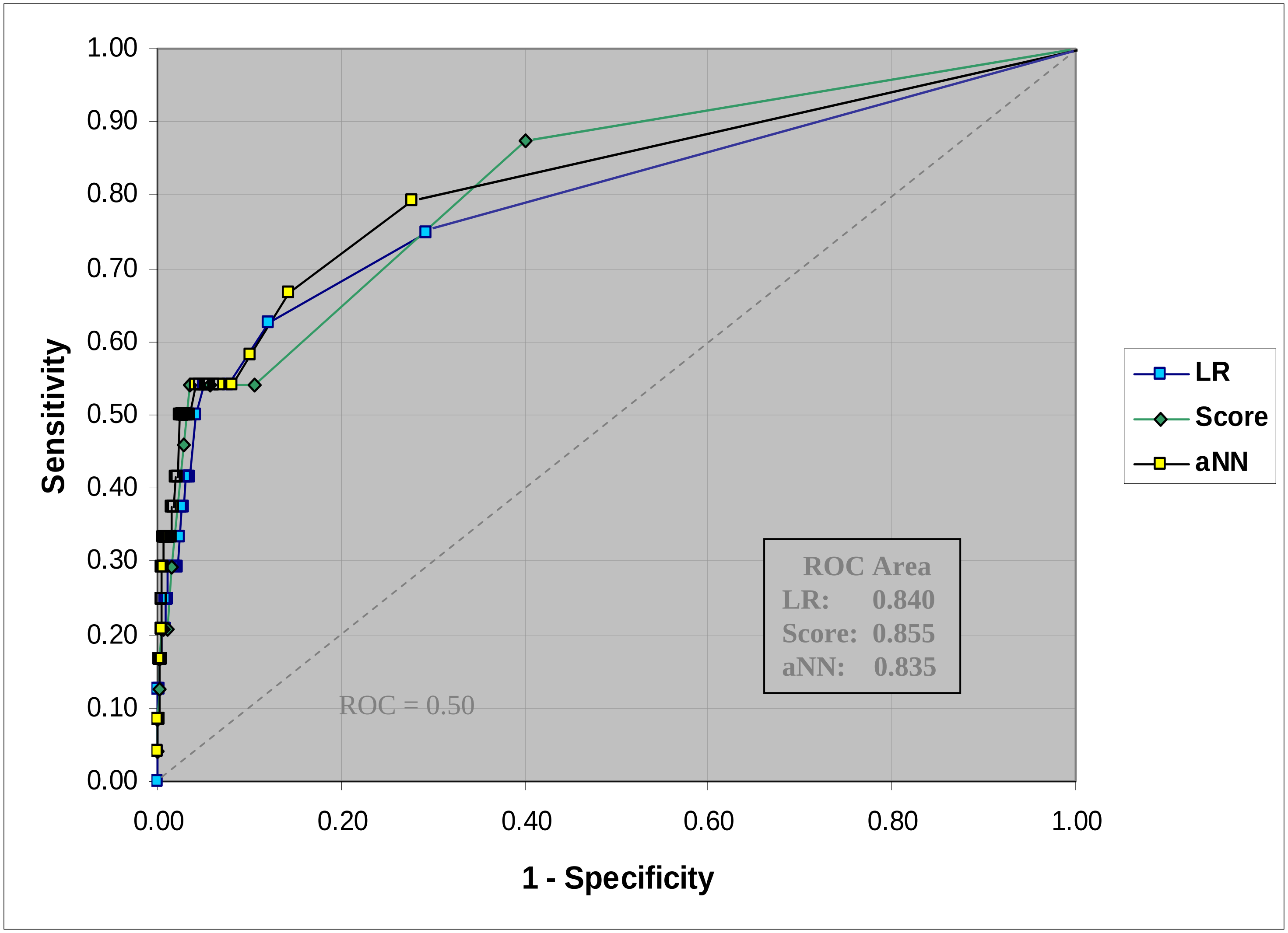
- Ties
3

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


1
Sensitivity
Area = 0.78
0
1 - Specificity
1

ROC Curves: Death Models

Validation Set: 1460 Cases



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

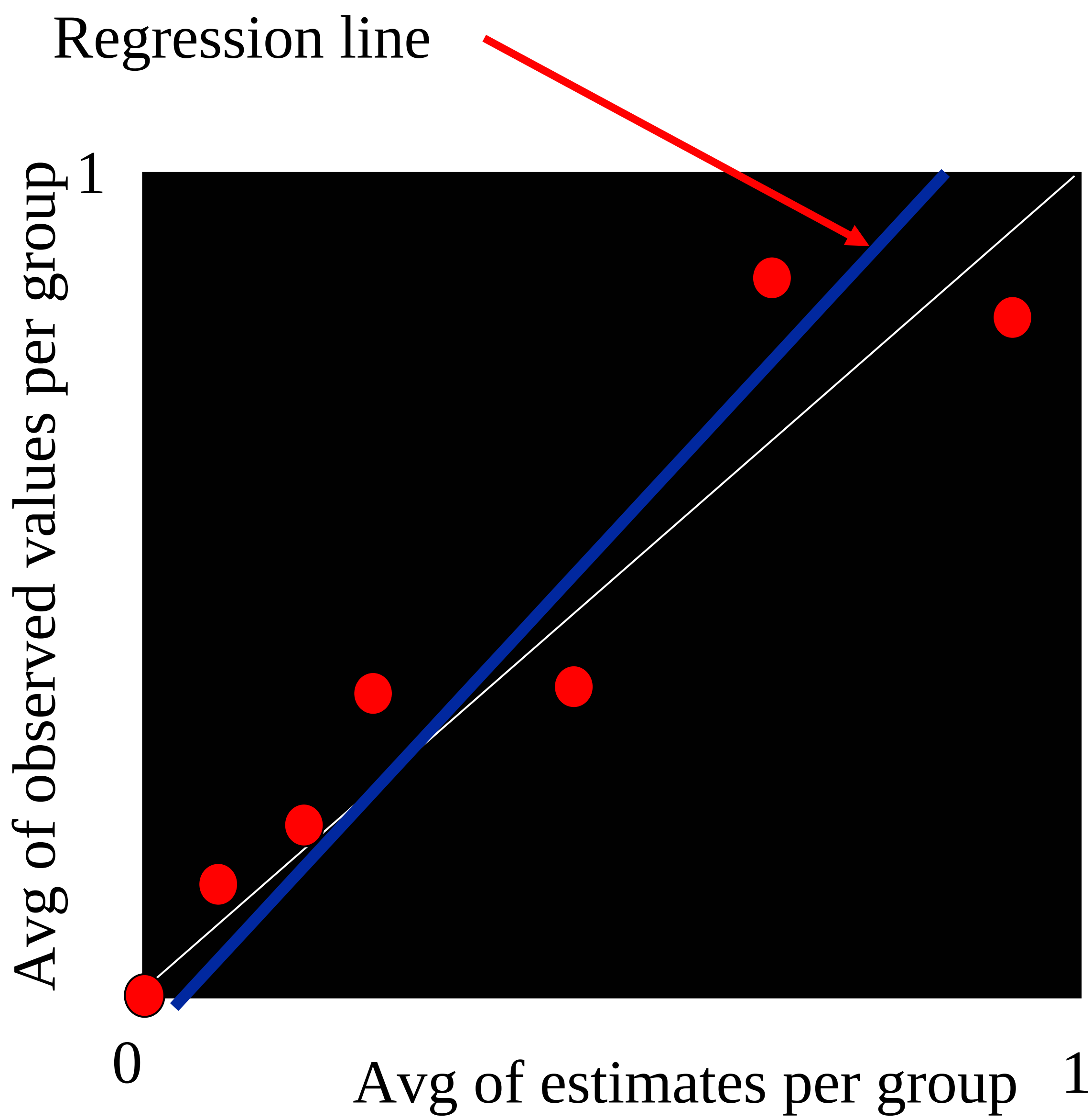
  0.2 and 0

  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**

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	
0.1		**0**	
0.2		**0**	
0.2	**sum = 0.5**	**1**	**sum = 1**
0.3		**0**	
0.5		**0**	
0.5	**sum = 1.3**	**1**	**sum = 1**
0.7		**0**	
0.7		**1**	
0.8		**1**	
0.9	**sum = 3.1**	**1**	**sum = 3**

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	
0.1		**0**	
0.2		**0**	
0.2		**1**	
0.3	**sum = 0.8**	**0**	**sum = 1**
0.5		**0**	
0.5	**sum = 1.0**	**1**	**sum = 1**
0.7		**0**	
0.7		**1**	
0.8		**1**	
0.9	**sum = 3.1**	**1**	**sum = 3**

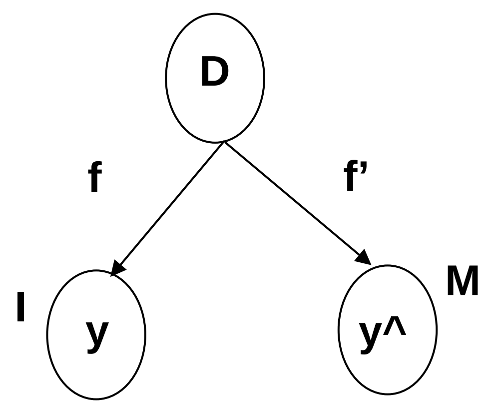
Decomposition of Error

The "ideal" model generates data *D*.

A "learned" model is learned from *D*.

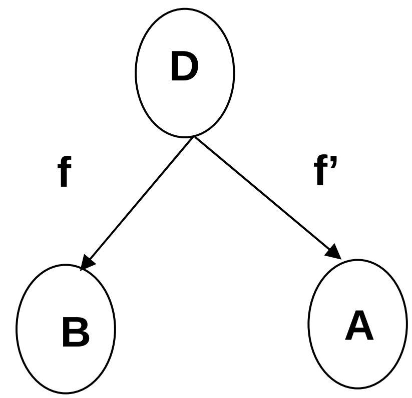
Once learned, model *M* is fixed.

After learning, *I* and *M* are conditionally independent given *D*.



Decomposition of Error

A and B binary (y-hat and y-ideal)



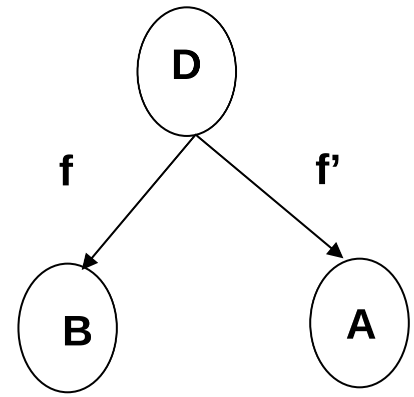
$$=1-\sum_{A=B} P(AB \mid D) =$$

$$=1-\sum_{A=B} P(AB \mid D) = 1-\sum_{A=B} P(A \mid D) P(B \mid D) =$$

Decomposition of Error

A represents classification from learned model

B represents classification from "ideal"



$$= 1 - \sum_{A=B} P(A \mid D) P(B \mid D) =$$

$$= 1 - \sum P(A) P(B) =$$

$$= \left[\frac{1}{2} + \frac{1}{2} \right] - \sum P(A) P(B) + 0 + 0 + 0$$

$$= \frac{1}{2} + \frac{1}{2} - \sum P(A) P(B) + \left[\sum P(AB) - \sum P(AB) \right] + \left[\frac{1}{2} \sum P(A)^2 - \frac{1}{2} \sum P(A)^2 \right] + \left[\frac{1}{2} \sum P(B)^2 - \frac{1}{2} \sum P(B)^2 \right] =$$

Decomposition of Error

$$=\frac{1}{2}+\frac{1}{2}-\sum P(A)P(B)+\sum P(AB)-\sum P(AB)+\frac{1}{2}\sum P(A)^2-\frac{1}{2}\sum P(A)^2+\frac{1}{2}\sum P(B)^2-\frac{1}{2}\sum P(B)^2=$$

$$=-1\underbrace{[\sum P(A)P(B)-\sum P(AB)]}_{\mathbf{=0}}+\frac{1}{2}[1-\sum P(A)^2]+\frac{1}{2}[1-\sum P(B)^2]+\frac{1}{2}\left[\sum P(A)^2-\sum P(AB)+\sum P(B)^2\right]=$$

$$=\frac{1}{2}\left\{\left[\underbrace{1-\sum P(A)}_{\textbf{variance}}{}^2\right]+\left[\underbrace{1-\sum P(B)}_{\textbf{error}}{}^2\right]+\underbrace{\sum\left[P(A)-P(B)\right]^2}_{\textbf{bias}}\right\}$$