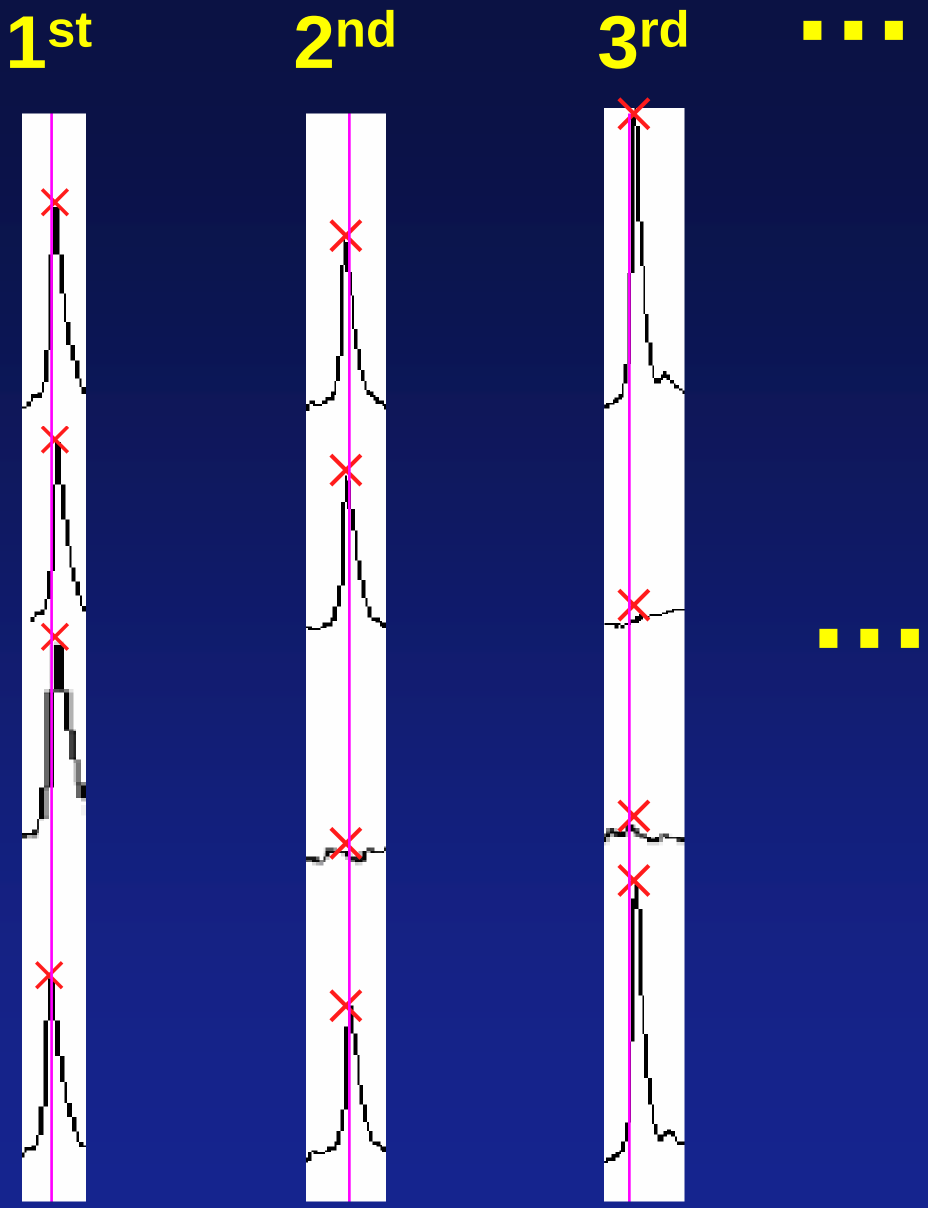
Classifier construction via. Boosting

Yutaka Yasui, Ph.D.

Division of Public Health Sciences
Fred Hutchinson Cancer Research Center

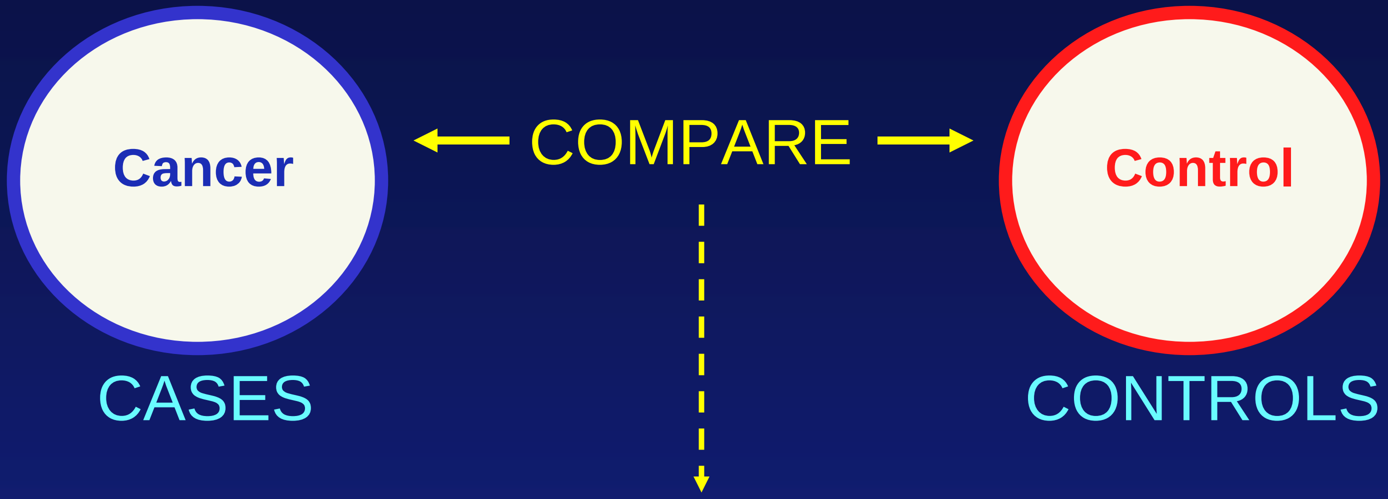


The aligned datase
for
searching
signature markers
profiles

Completion of pre-
analysis processing

Yasui et al. J. Biomed. & Biotech
(Special Issue on Proteomics) 20

Basic Study Design



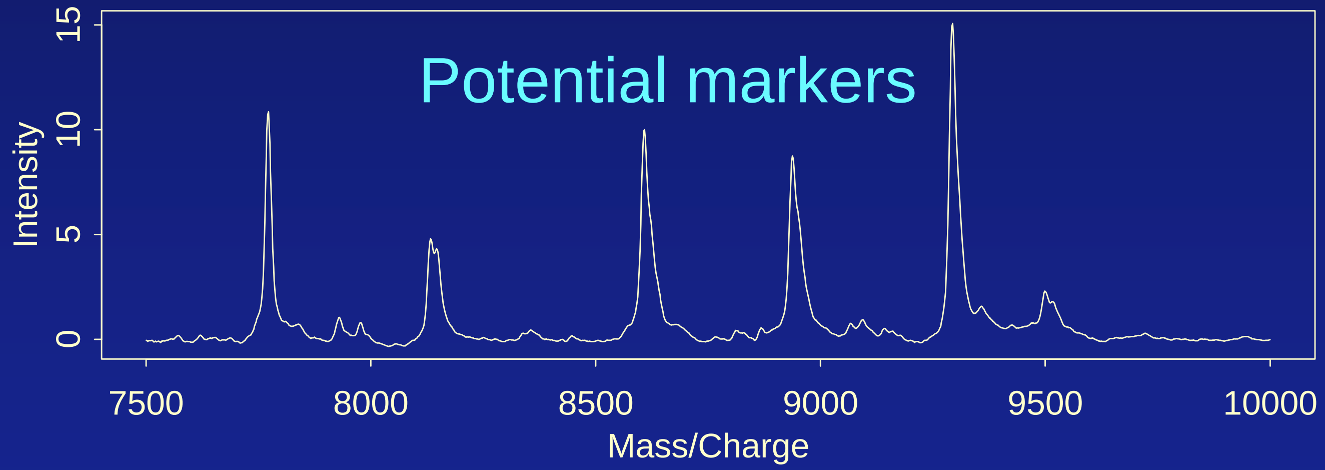
Cancer

COMPARE

Control

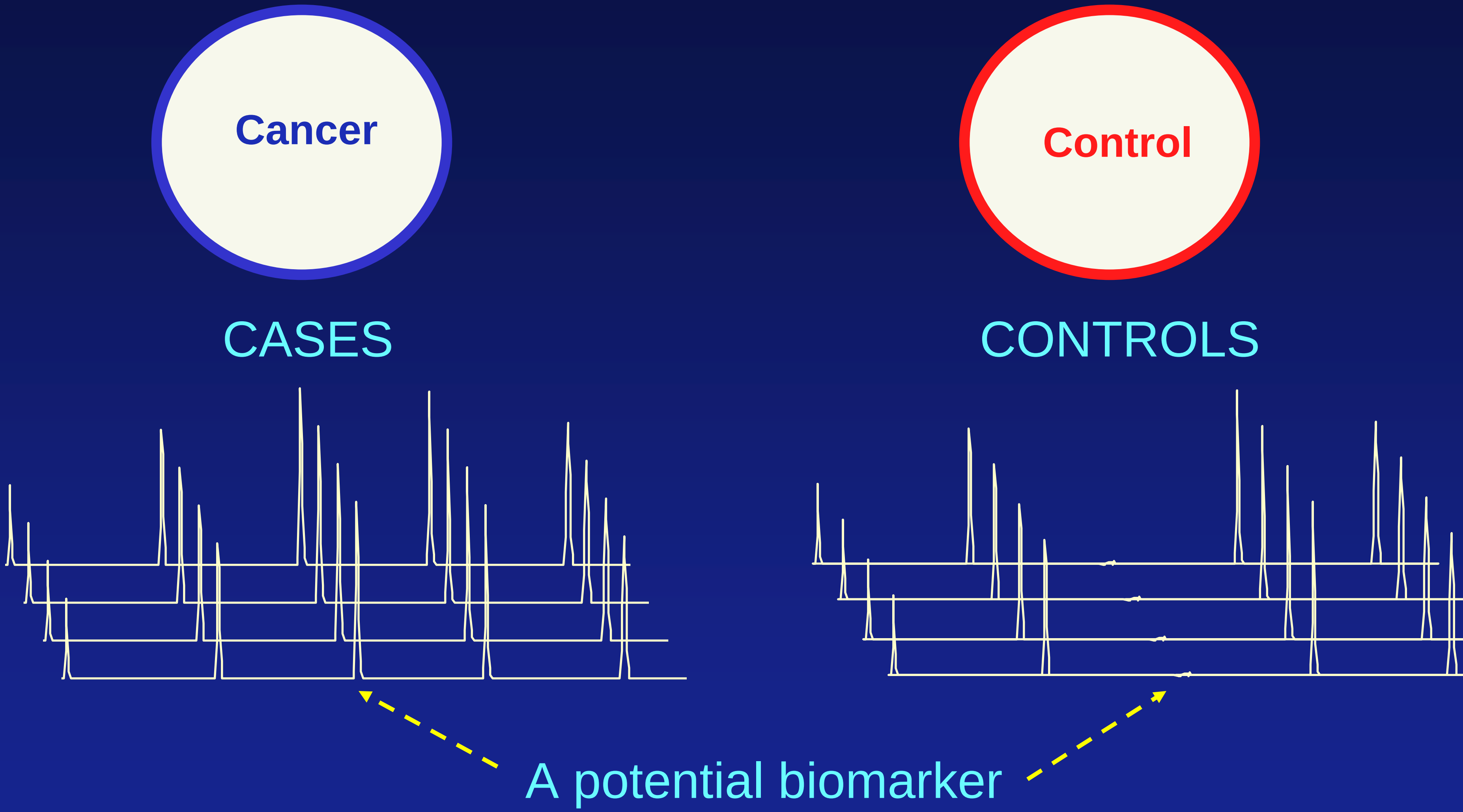
CASES

CONTROLS



Potential markers

Biomarker Discovery



Cancer

Control

CASES

CONTROLS

A potential biomarker

Phases of Biomarker Discovery & Validation

Pepe et al. JNCI 2001

→	***Pre-clinical Exploratory***	***PHASE 1***	*Promising directions identified*
- - →	***Clinical Assay Validation***	***PHASE 2***	*Clinical assay detects established disease*
	Retrospective Longitudinal	***PHASE 3***	*Biomarker detects pre-clinical disease and a “screen positive” rule defined*
	Prospective Screening	***PHASE 4***	*Extent and characteristics of disease detected by the test and the false referral rate are identified*
	Cancer Control	***PHASE 5***	*Impact of screening on reducing burden of disease on population is quantified*

100% sensitivity & specificity

in classifying cases vs. controls

$\neq$

Identification of biomarkers for cases

Three Principles of Case-Control Design

(Wacholder et al. Am J Epidemiol 1992)

1. A common study base for cases and controls

2. Controlling for confounding effects

3. Comparable accuracy and precision in exposure measurements

1. Common Study Base

○ Define a common study base (who, where, when) and sample both cases and controls from it

× Cases and controls from different institutions

× Cases from a past study, controls from an ongoing study

⇨ Disease is not the only difference between cases and controls

2. Controlling for confounding

- Balance age and race between cases and controls (or adjust for in the analysis)

- Study base = 30-75 women in Montreal in 2003
  Breast cancer cases = Tend to be older
  Controls = Younger

⇨ Markers for age, not cancer, will distinguish cases and controls

3. Comparable measurement errors

- Unify the sample collection, processing, storage, and assay methods for cases and controls.

  Balance the use of machines, technicians, chips, and wells between cases and controls.

If not,

⇨ True marker-disease relation is distorted

Use of multiple markers

in classifying disease classes

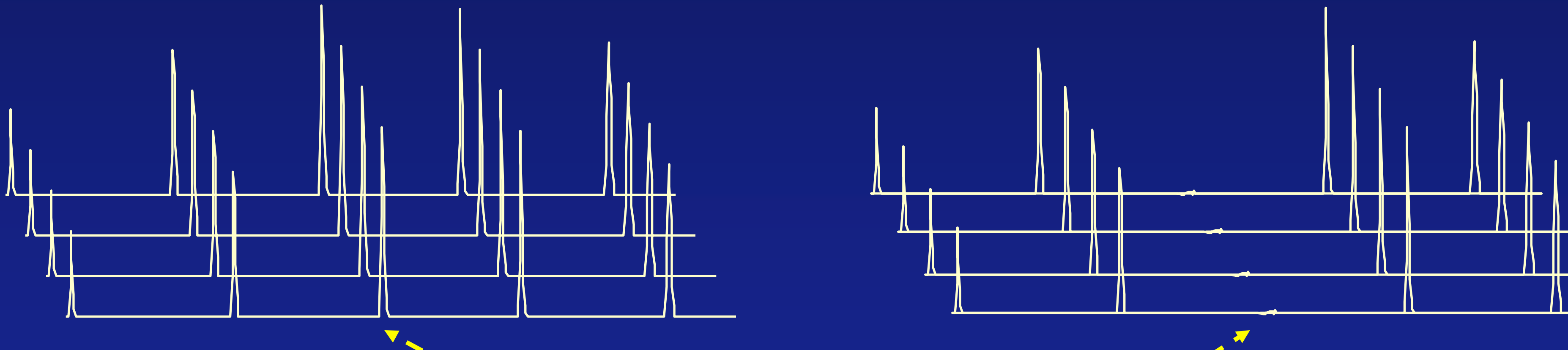
Biomarker Discovery

Cancer

Control

CASES

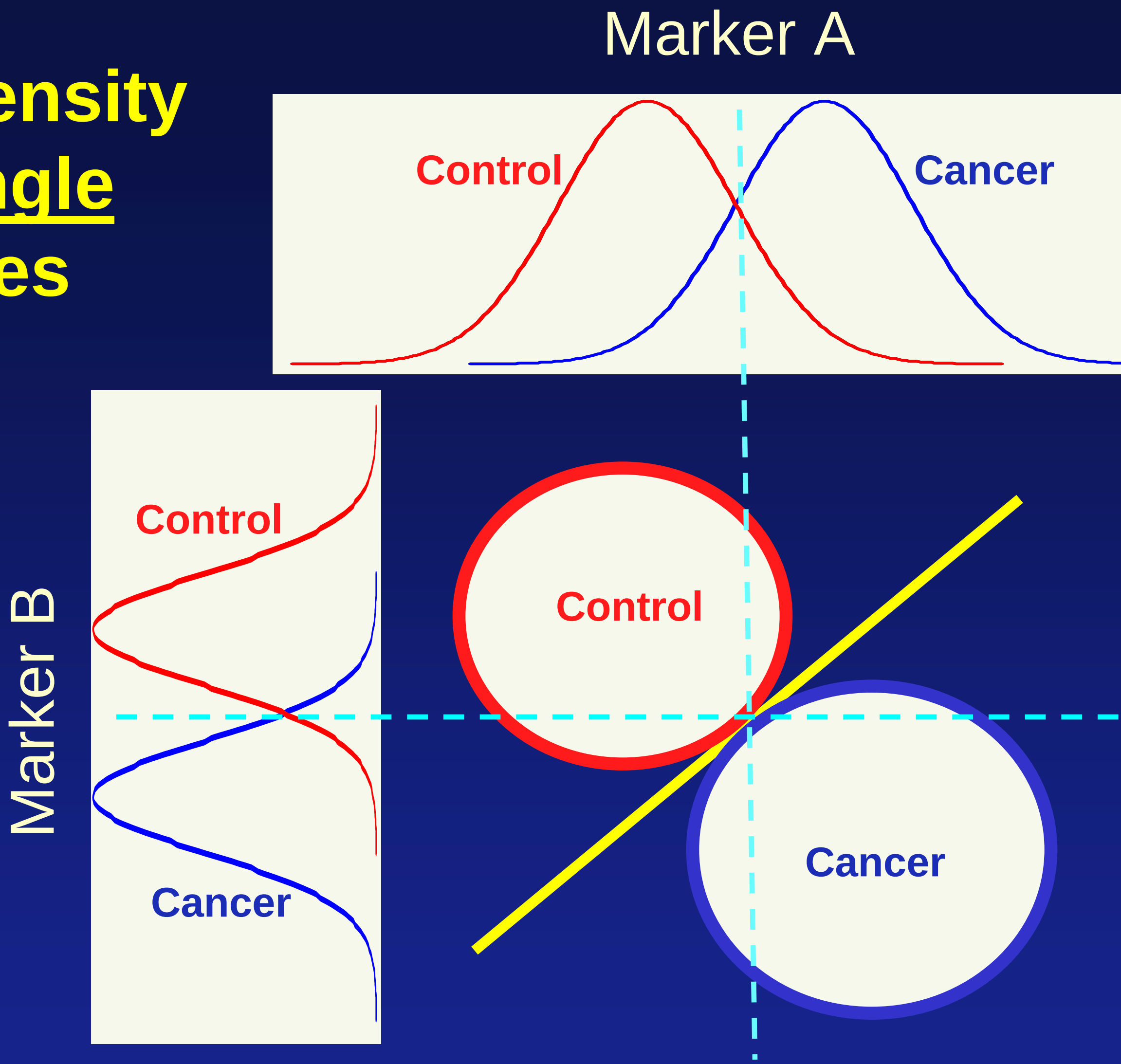
CONTROLS



A potential biomarker

Likely overlap of intensity distributions of a single marker between cases and controls

Need to combine information from multiple markers!

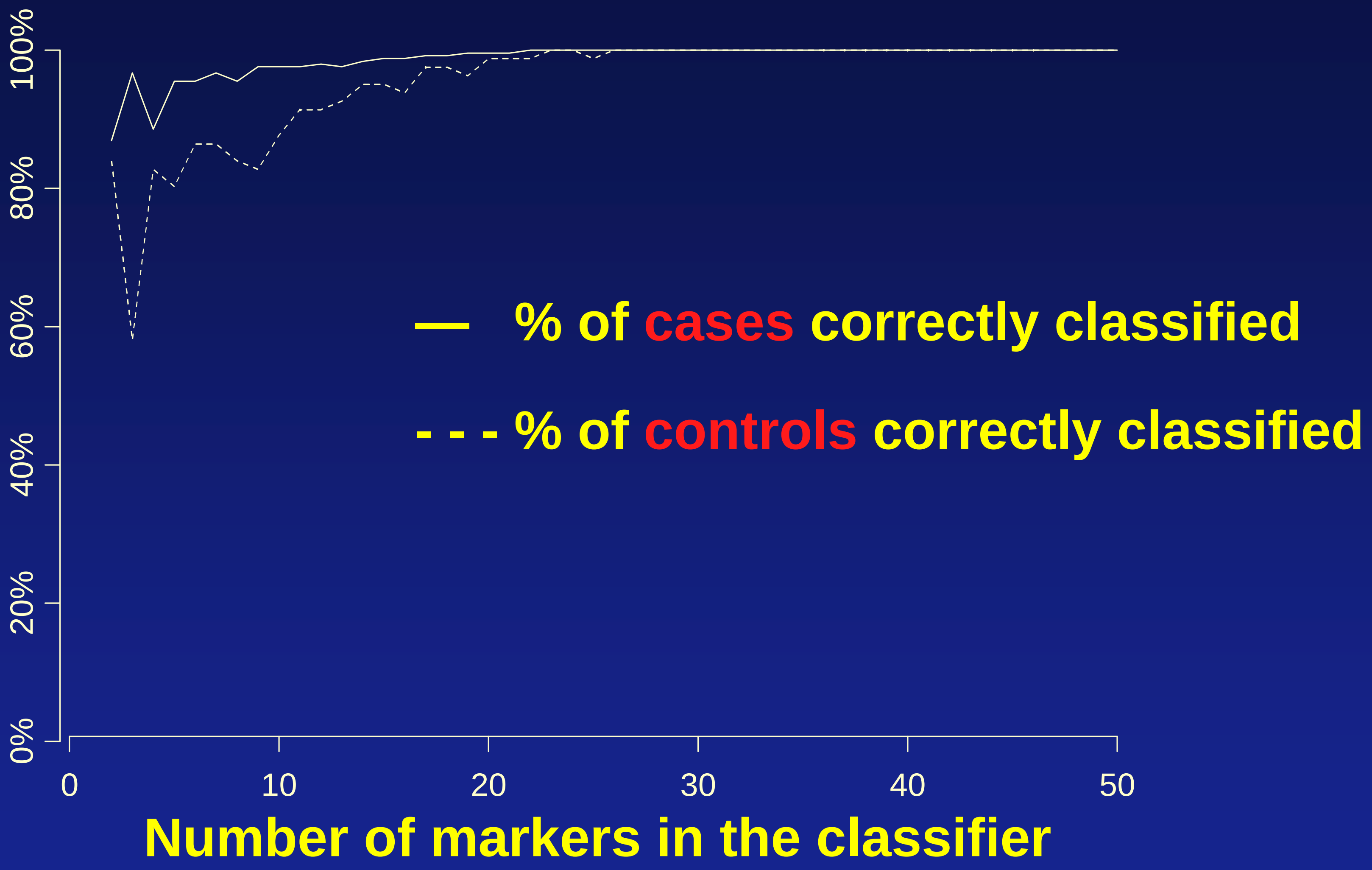


Building Classifiers

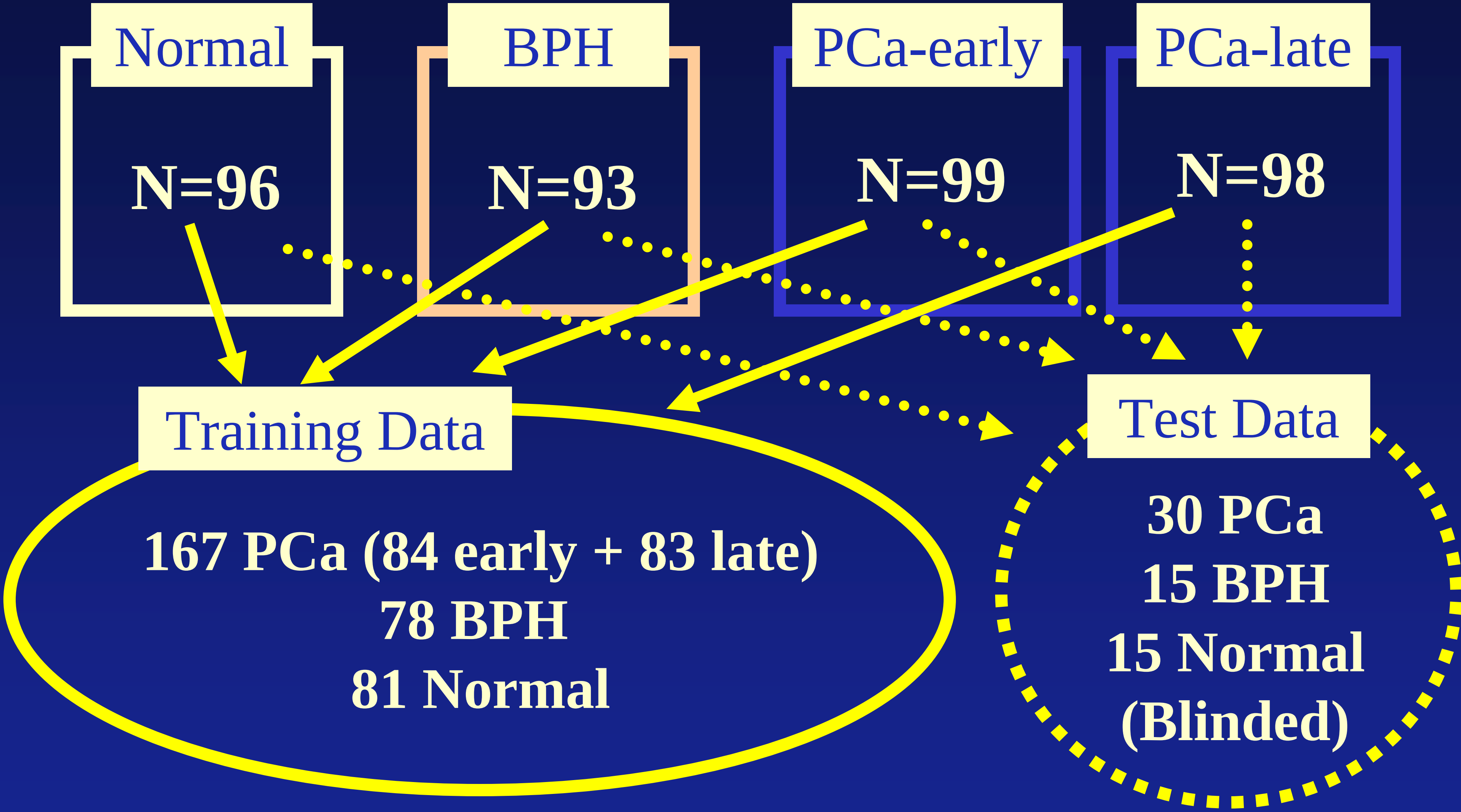
- Classical Discriminant Analysis
- Logistic Regression
- CART
- Neural Network
- Support Vector Machine
- Boosting

…

ancer vs. control classification in a given datase



The design of the EVMS biomarker analysis



Normal

N=96

BPH

N=93

PCa-early

N=99

PCa-late

N=98

Training Data

167 PCa (84 early + 83 late)
78 BPH
81 Normal

Test Data

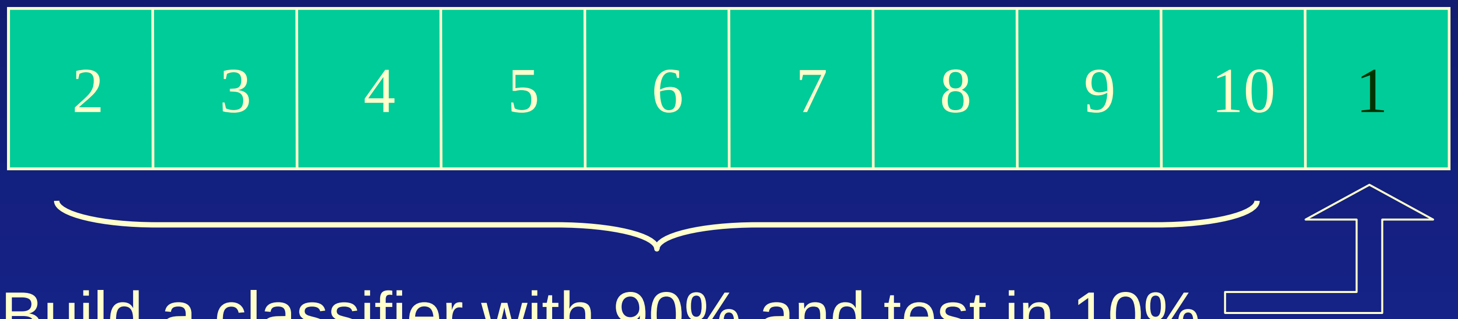
30 PCa
15 BPH
15 Normal
(Blinded)

How to assess over-fitting in the training set ?

- Cross-validation of the training data

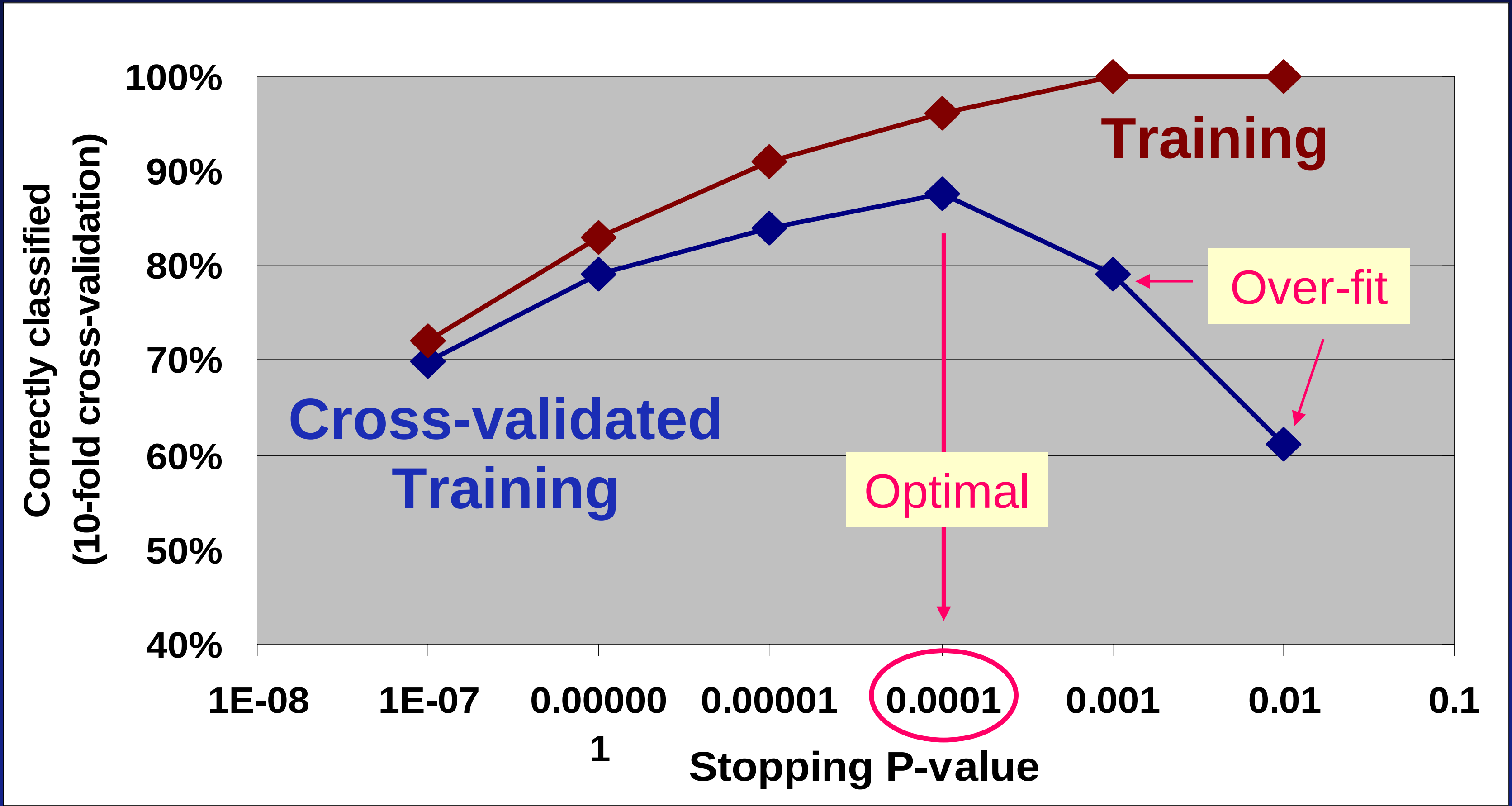
Use 90% to form the marker set & 10% to test

Repeat 10 times and summarize



Build a classifier with 90% and test in 10%

Logistic regression with forward variable selection with various stopping p-values



Use of the test set

Enable unbiased assessment of classification erro

if no modification/selection of the classifier-construction method is made with the test set

e.g., Construct 2 classifiers with the training set and report the one with the better test-set performance

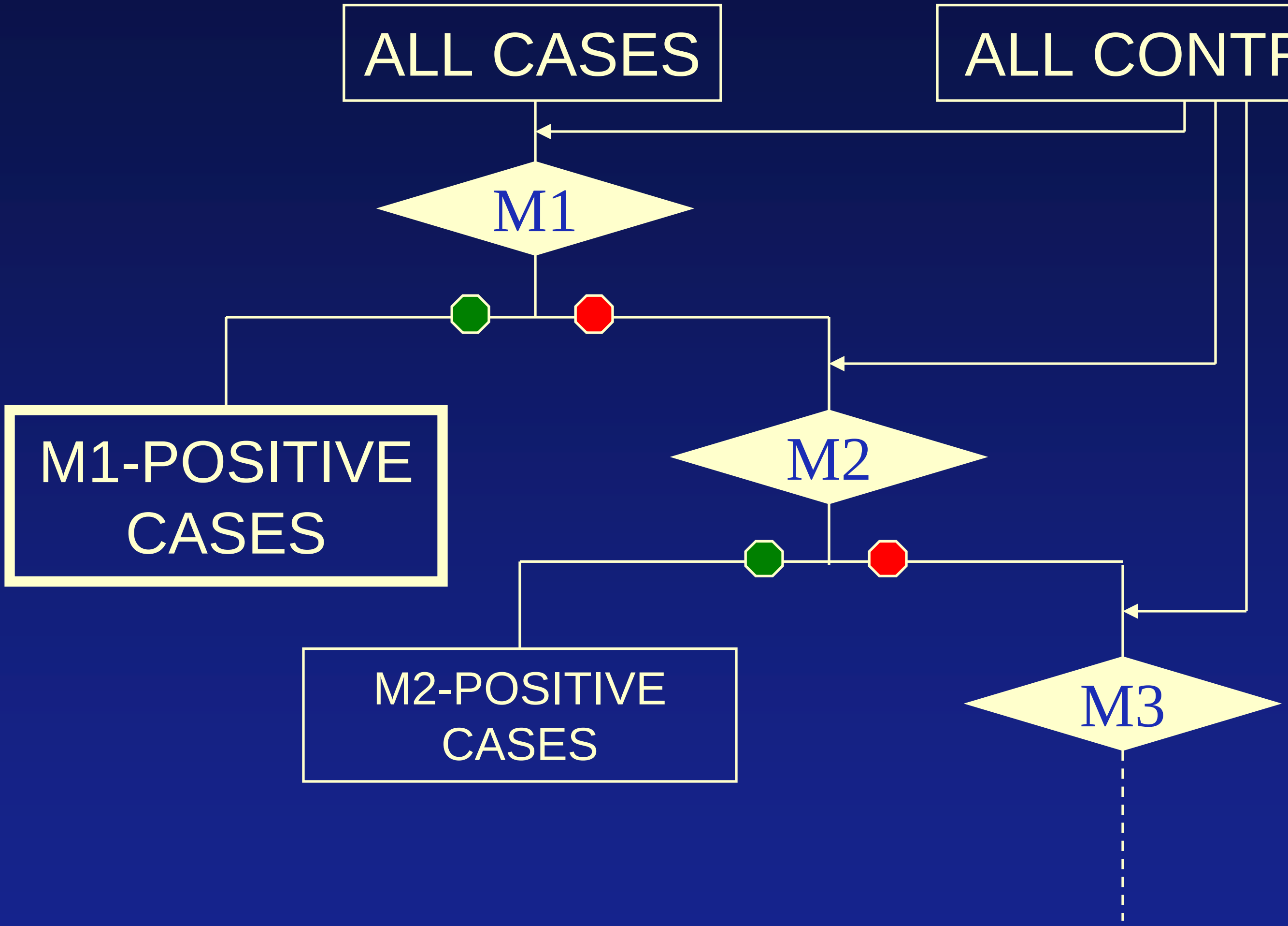
(2 feature selection methods, stepwise stopping, etc

Boosting for supervised and partially supervised learning

Method for classifier building and

its modification for partially-incorrect class labels

Heterogeneity / subtypes within cancer



Real AdaBoost Algorithm ($y^* = 1$ vs. $y^* = -1$)

Friedman, Hastie, Tibshirani (Annals of Statistics, 2000)

1. Let $w_i \equiv 1/N$ for $i = 1, 2, \ldots, N$
2. Repeat for $m = 1, 2, \ldots, M$
   - Fit a classifier with weights $\{w_i\}$ to get
     $p_m(x) = \Pr(Y^*=1|x, \{w_i\})$
   - $\boxed{\text{Set } w_i = w_i \times \exp\{-\,0.5\, y_i^* \times \text{logit } p_m(x_i)\}}$
   - Renormalize $\{w_i\}$ such that $\sum_i w_i = 1$
3. The final classifier:
   $\eta_M(x) = \text{logit } p_1(x) + \text{logit } p_2(x) + \ldots + \text{logit } p_M(x) > c$

Real AdaBoost with Logistic Regression

$$(\alpha_m, \beta_m, X^{(m)}) = \arg\min_{(\alpha,\beta,X)} \sum_i \underline{e^{-\frac{y_i^*}{2}\sum_{j=1}^{m-1}(\hat{\alpha}_j+\hat{\beta}_j X_i^{(j)})}} \; \underline{\ln\{1+e^{-y_i^*(\alpha+\beta X_i)}\}}$$

$$e^{-\frac{y_i^*}{2}\sum_{j=1}^{m-1}(\hat{\alpha}_j+\hat{\beta}_j X_i^{(j)})}$$

$$\ln\{1+e^{-y_i^*(\alpha+\beta X_i)}\}$$

Weights

Negative log-likelihoo

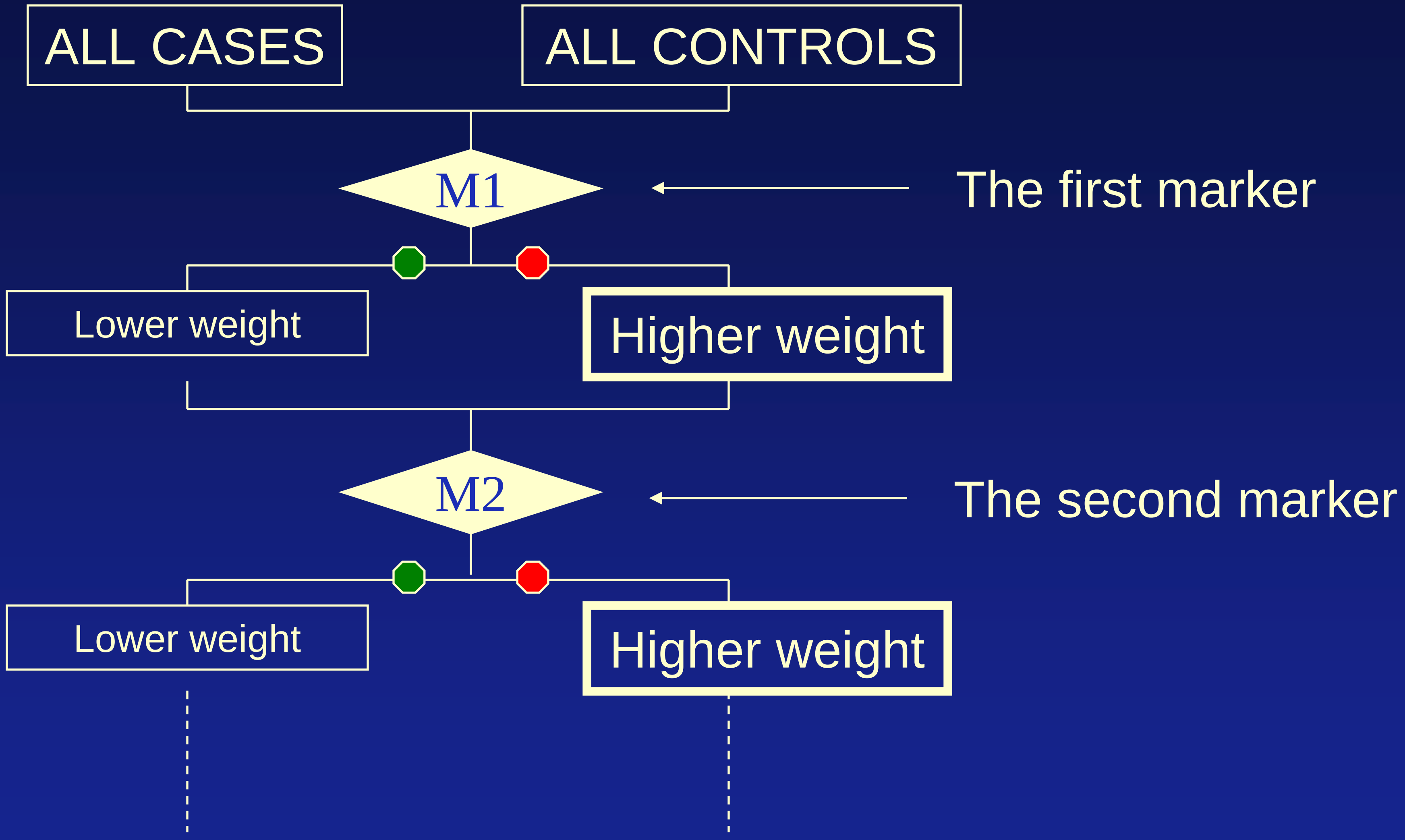
${}_i \;= w_i \times \exp\{-\,0.5\ y_i \times \text{logit}\ p_{m-1}(x_i)\}$

$= w_i \times \exp\{-\,0.5\ y_i \times (\alpha_{m-1}+\beta_{m-1}\ x_i)\}$

$= \exp\{-\,0.5\ y_i \sum_{j=1,\ldots,(m-1)} (\alpha_j+\beta_j\ x_i)\}$

Yasui et al. (Biostatistics, 200

Boosting algorithm



ALL CASES

ALL CONTROLS

M1

The first marker

Lower weight

Higher weight

M2

The second marker

Lower weight

Higher weight

Performance of the boosting classifier

(1st stage: Abnormal vs. Normal)

	Correct classification	
	Training dataset	Test dataset
Cancer/BPH	245/245 (100%)	44/45 (97.8%)
Normal	81/ 81 (100%)	15/15 (100%)

Why does this work?

AdaBoost = “Best off-the-shelf classifier”

(Brieman)

$$\alpha_m, \beta_m, X^{(m)}) = \arg\min \sum \mathrm{e}^{-\frac{y_i^*}{2}\sum_{j=1}^{m-1}(\hat{\alpha}_j+\hat{\beta}_j X_i^{(j)})} \ln\{1+\mathrm{e}^{-y_i^*(\alpha+\beta X$$

osting = Stage-wise minimization of a loss function

$$= \underset{(\alpha,\beta,X)}{\arg\min} \sum_i L_i^*(y_i^*, \eta_{\underset{\%}{\phi_m}}(\underset{\%}{X}_i^{(m)}))$$

$$(\alpha_m, \beta_m, X^{(m)}) = \arg\min_{(\alpha,\beta,X)} \sum_i L_i^*(y_i^*, \eta_{\underset{\%}{\phi}_m}(\underset{\%}{X}_i^{(m)}))$$

$$= \arg\min_{(\underset{\%}{\theta}=(\alpha,\beta),X)} \sum_i L_i^*(y_i^*, \eta_{(\underset{\%}{\theta},\underline{\underset{\%}{\phi}_{m-1}})}(X, \underline{\underset{\%}{X}_i^{(m-1)}}))$$

$$\underset{\%}{\phi}_{(m-1)} = (\underset{\%}{\theta}_1, \ldots, \underset{\%}{\theta}_{(m-1)})$$

$$\underset{\%}{X}^{(m-1)} = (X^1, \ldots, X^{(m-1)}$$

Previous stages' parameters

Previous stages' biomarkers

FIXED

$$(\alpha_m,\beta_m,X^{(m)}) \;=\; \underset{(\underset{\%}{\theta}=(\alpha,\beta),X)}{\arg\min} \sum_i L_i^*\big(y_i^*,\eta_{(\underset{\%}{\theta},\underbrace{\underset{\%}{\phi}_{m-1}}_{\text{fixed}})}(X,\underbrace{\underset{\%}{X}_i^{(m-1)}}_{\text{fixed}})\big)$$

Boosting = Stage-wise minimization of a loss function L^* given previously selected biomarkers $X^{(m-1)}$ and their parameters $\phi_{(m-1)}$

Classifier changes slightly at each stage = Slow learning

$$(\alpha_m, \beta_m, X^{(m)}) = \underset{(\alpha,\beta,X)}{\arg\min} \sum_i e^{-\frac{y_i^*}{2}\left[\left\{\sum_{j=1}^{m-1}(\hat{\alpha}_j+\hat{\beta}_j X_i^{(j)})\right\} + (\alpha+\beta X_i)\right]}$$

$$\boxed{= \underset{(\alpha,\beta,X)}{\arg\min} \sum_i e^{-\frac{y_i^*\ \eta_{\underset{\%}{\phi_m}}(\underset{\%}{X_i^{(m)}})}{2}}}$$

$$= \underset{(\alpha,\beta,X)}{\arg\min} \sum_i L_i^*(y_i^*, \eta_{\underset{\%}{\phi_m}}(\underset{\%}{X_i^{(m)}}))$$

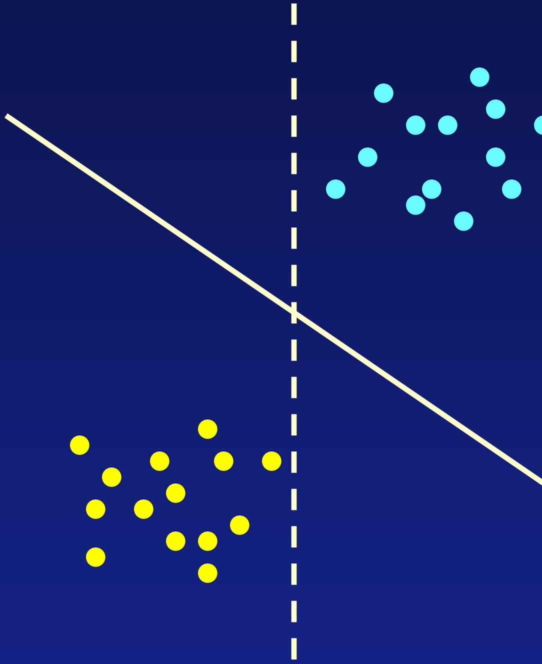
Does this form of the loss function make sense?

Large margin classifiers

$\text{Margin}_i \equiv y_i\ \eta(x_i)$

> 0 if $\eta(x_i)$ is correct

< 0 if $\eta(x_i)$ is wrong



- Higher confidence in classification
- Increased generalizability

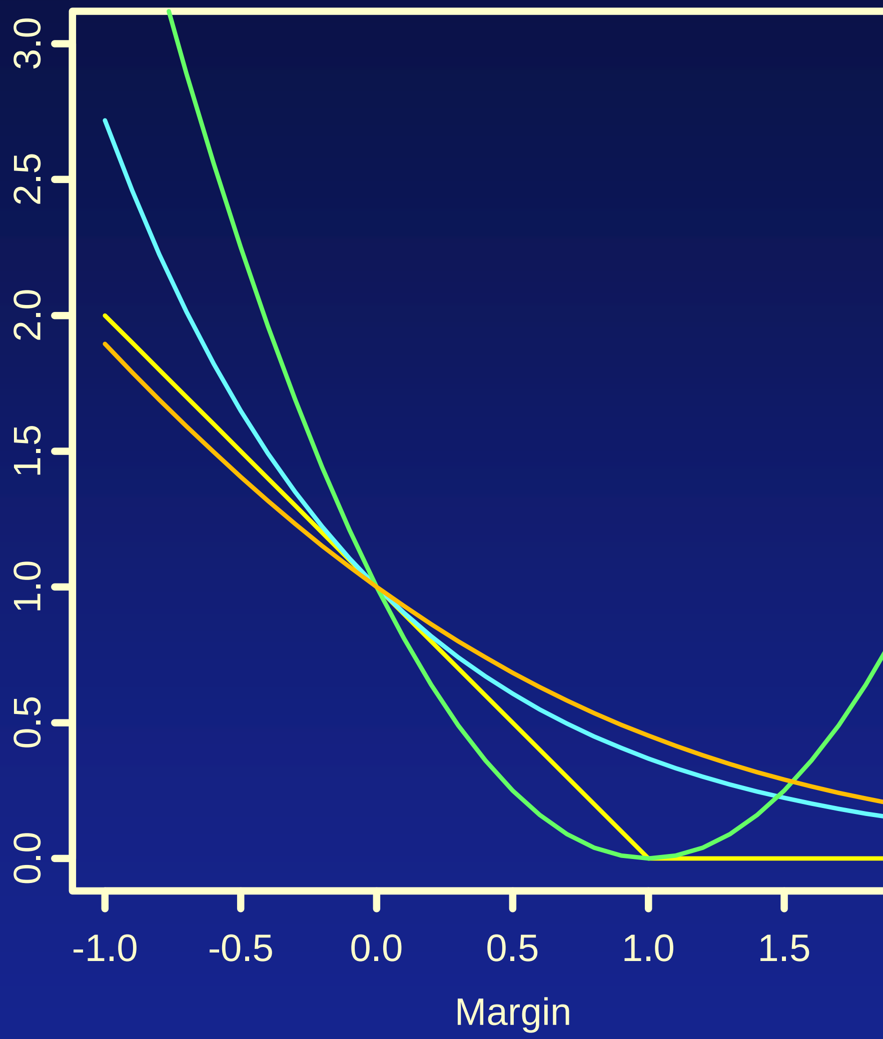
Large margin classifiers

SVM = min $\sum_i$ max(0,1-margin$_i$)

AdaBoost = min $\sum_i$ $e^{-margini}$

Logistic = min $\sum_i$ log(1+$e^{-margini}$)

NN = min $\sum_i$ (1-margin$_i$)^2



Discrete AdaBoost Algorithm ($y^* = 1$ vs. $y^* = -1$)

1. Let $w_i \equiv 1/N$ for $i = 1, 2, \ldots, N$
2. Repeat for $m = 1, 2, \ldots, M$
   - Fit a base classifier $f_m(x_i) \in \underline{\{-1,1\}}$ (e.g., a decision tree) with weights $\{w_i\}$
   - $ERR_m = \sum w_i 1\{y_i \neq f_m (x_i)\}$
   - $C_m=\log\{(1-ERR_m)/ERR_m\}$
   - Set $w_i = w_i \times \exp\{- 0.5\ \underline{C_m\, y_i^* \times f_m(x_i)}\}$
   - Renormalize $\{w_i\}$ such that $\sum_i w_i = 1$
3. The final classifier: $C_1f_1(x) + C_2f_2(x) + \ldots + C_Mf_M(x) > c$

It worked well for Cancer/BPH vs. Normal

But …

Performance of the boosting classifier

(2nd stage: Cancer vs. BPH)

Correct classification

	Training dataset	Test dataset
Cancer	160/167 (95.8%)	28/30 (93.3%)
BPH	70/ 78 (89.7%)	7/15 (46.7%)

European Prostate Cancer Detection Study

Protocol: Biopsy 1,051 men with PSA 4-10 ng/ml

If negative, take another biopsy 6 weeks later

If negative again, take another 8 weeks later

Cancer detection: 231 were detected by Biopsy 1

83 were detected by Biopsy 2

36 were detected by Biopsy 3

119 cance
missed b
Biopsy 1

∴ A single biopsy can miss > 1/3 of cancers in PSA 4-10 patients

Cancer label = 100% correct

Non-cancer label < 100% correct

≡ Partially Supervised Learning

How can we "learn" from potentially partially mislabeled data?

- If correct labels y_i^*s are available:

$$(\alpha_m, \beta_m, X^{(m)}) = \underset{(\alpha,\beta,X)}{\arg\min} \sum_i \underbrace{\mathrm{e}^{-\frac{y_i^*}{2}\sum_{j=1}^{m-1}(\hat{\alpha}_j + \hat{\beta}_j X_i^{(j)})}}_{\text{weights}} \; \underbrace{\ln\{1 + \mathrm{e}^{-y_i^*(\alpha + \beta X_i)}\}}_{\text{- log-likelihood}}$$

High (low) weights for incorrectly (correctly) classified observation

Results of (m-1)th classification ⇒ Who should “speak louder” at mth stage

- If correct labels y_i^*s are NOT available:

⇒ We cannot determine whether the (m-1)th classification was correct or not

⇒ Unclear who should speak louder at the mth stage

PROPOSAL

Let the observations that are likely to be misclassified at (m-1)th stage speak louder at mth stage

$$\boxed{\Pr[y_i^* = \underline{-1} \mid \underset{\%}{\phi}_{(m-1)}, \underset{\%}{X}^{(m-1)}, \underset{\%}{y}] \quad \times} \quad e^{-\frac{\overset{\overset{-1}{\downarrow}}{y_i^*}}{2}\sum_{j=1}^{m-1}(\hat{\alpha}_j+\hat{\beta}_j X_i^{(j)})} \qquad \ln\{1+e^{-\overset{\overset{-1}{\downarrow}}{y_i^*}(\alpha+\beta X_i}$$

Loss if $y_i^* = \underline{-1}$

+

$$\boxed{\Pr[y_i^* = \underline{1} \mid \underset{\%}{\phi}_{(m-1)}, \underset{\%}{X}^{(m-1)}, \underset{\%}{y}] \quad \times} \quad e^{-\frac{\overset{\overset{1}{\downarrow}}{y_i^*}}{2}\sum_{j=1}^{m-1}(\hat{\alpha}_j+\hat{\beta}_j X_i^{(j)})} \qquad \ln\{1+e^{-\overset{\overset{1}{\downarrow}}{y_i^*}(\alpha+\beta X_i}$$

Loss if $y_i^* = \underline{1}$

- If correct labels y_i^*s are available:

$$\underset{(\underset{\%}{\theta}=(\alpha,\beta),X)}{\arg\min} \sum_i L_i(\underset{\%}{\theta}, X; \underset{\%}{\phi}_{(m-1)}, \underset{\%}{X}^{(m-1)}, y_i^*)$$

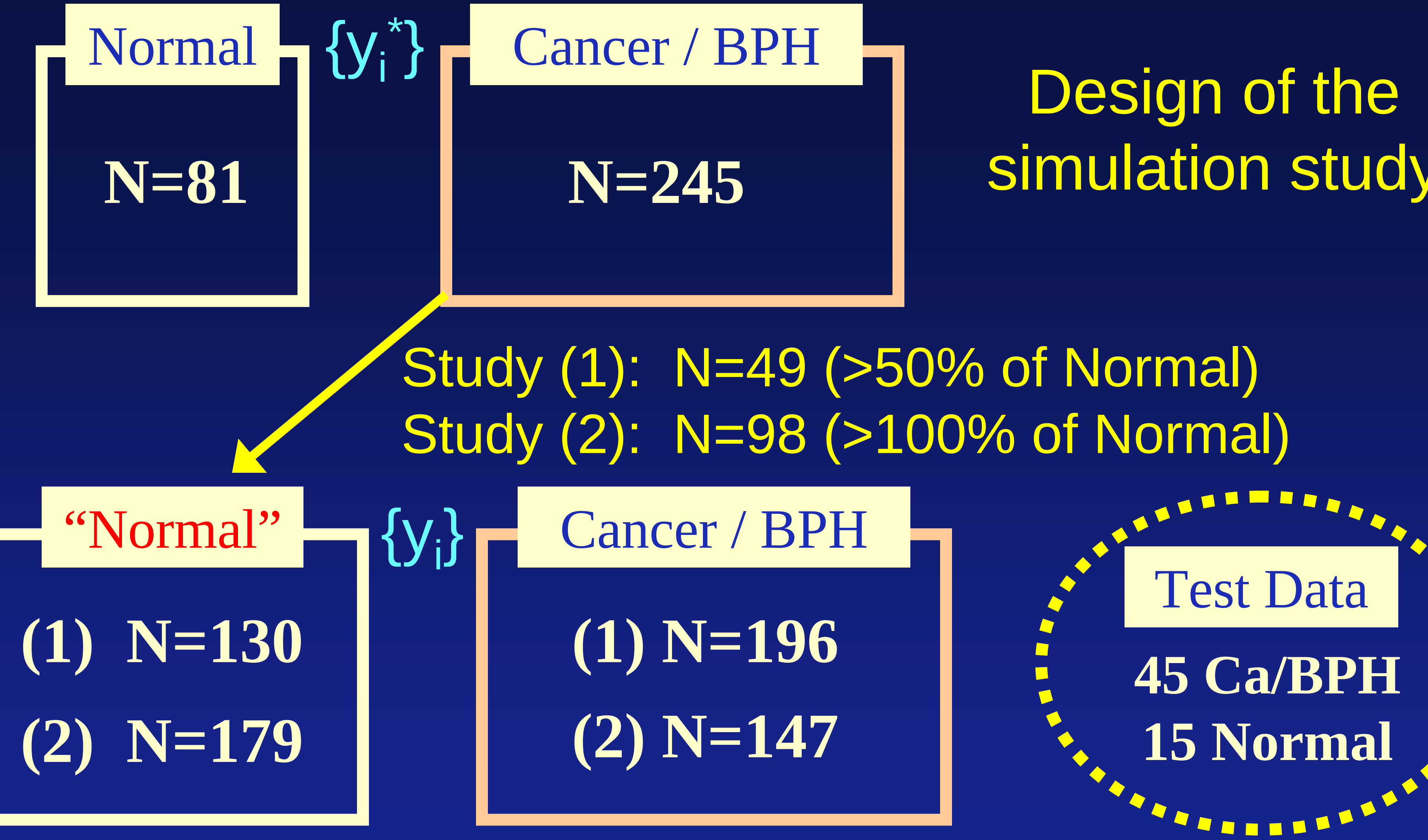
- If correct labels y_i^*s are NOT available:

$$\underset{(\underset{\%}{\theta}=(\alpha,\beta),X)}{\arg\min} \sum_i \sum_{y_i^*=-1}^{y_i^*=1} L_i(\underset{\%}{\theta}, X; \underset{\%}{\phi}_{(m-1)}, \underset{\%}{X}^{(m-1)}, y_i^*)\,\Pr[y_i^* \mid \underset{\%}{\phi}_{(m-1)}, \underset{\%}{X}^{(m-1)},$$

$$\underset{(\underset{\%}{\theta}=(\alpha,\beta),X)}{\arg\min} \sum_i E[L_i(\underset{\%}{\theta}, X; \underset{\%}{\phi}_{(m-1)}, \underset{\%}{X}^{(m-1)}, y_i^*) \mid \underset{\%}{\phi}_{(m-1)}, \underset{\%}{X}^{(m-1)}, \underset{\%}{y}] \Rightarrow \mathrm{E}$$

Yasui et al. (Biometrics, 2004

Design of the simulation study



Normal

N=81

$\{y_i^*\}$

Cancer / BPH

N=245

Study (1): N=49 (>50% of Normal)

Study (2): N=98 (>100% of Normal)

“Normal”

(1) N=130

(2) N=179

$\{y_i\}$

Cancer / BPH

(1) N=196

(2) N=147

Test Data

45 Ca/BPH

15 Normal

Questions in the simulation study

Q1: Can we recover the cancer/BPH samples that were incorrectly labeled as “normal”?

Q2: How do the classifiers constructed from the incorrectly labeled training dataset perform when tested against the test dataset?

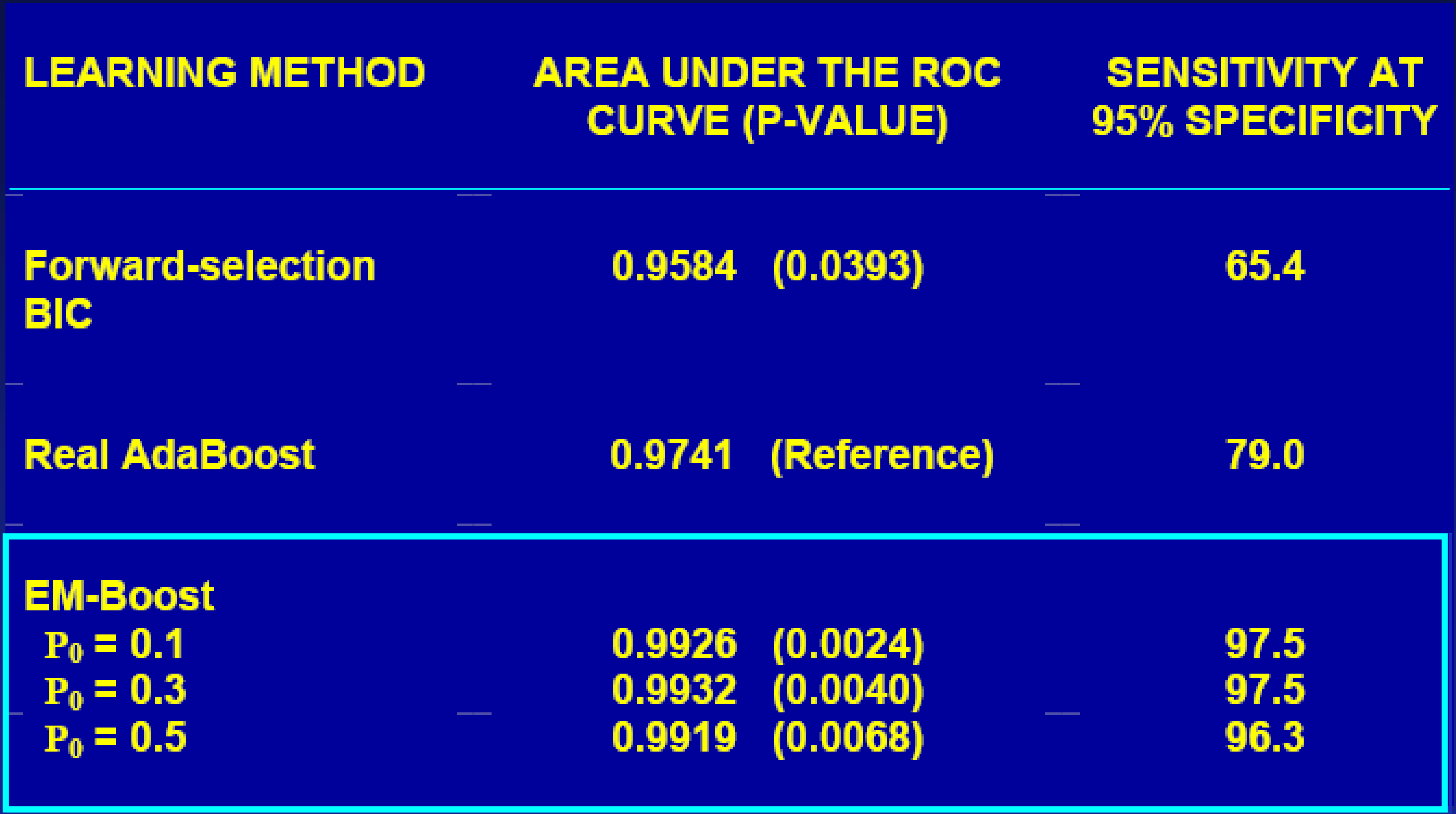
Learning methods compared

(1) Forward-selection logistic regression with BIC as the model-selection criteria

(2) Real AdaBoost with logistic regression (stopped at m=100th iterations)

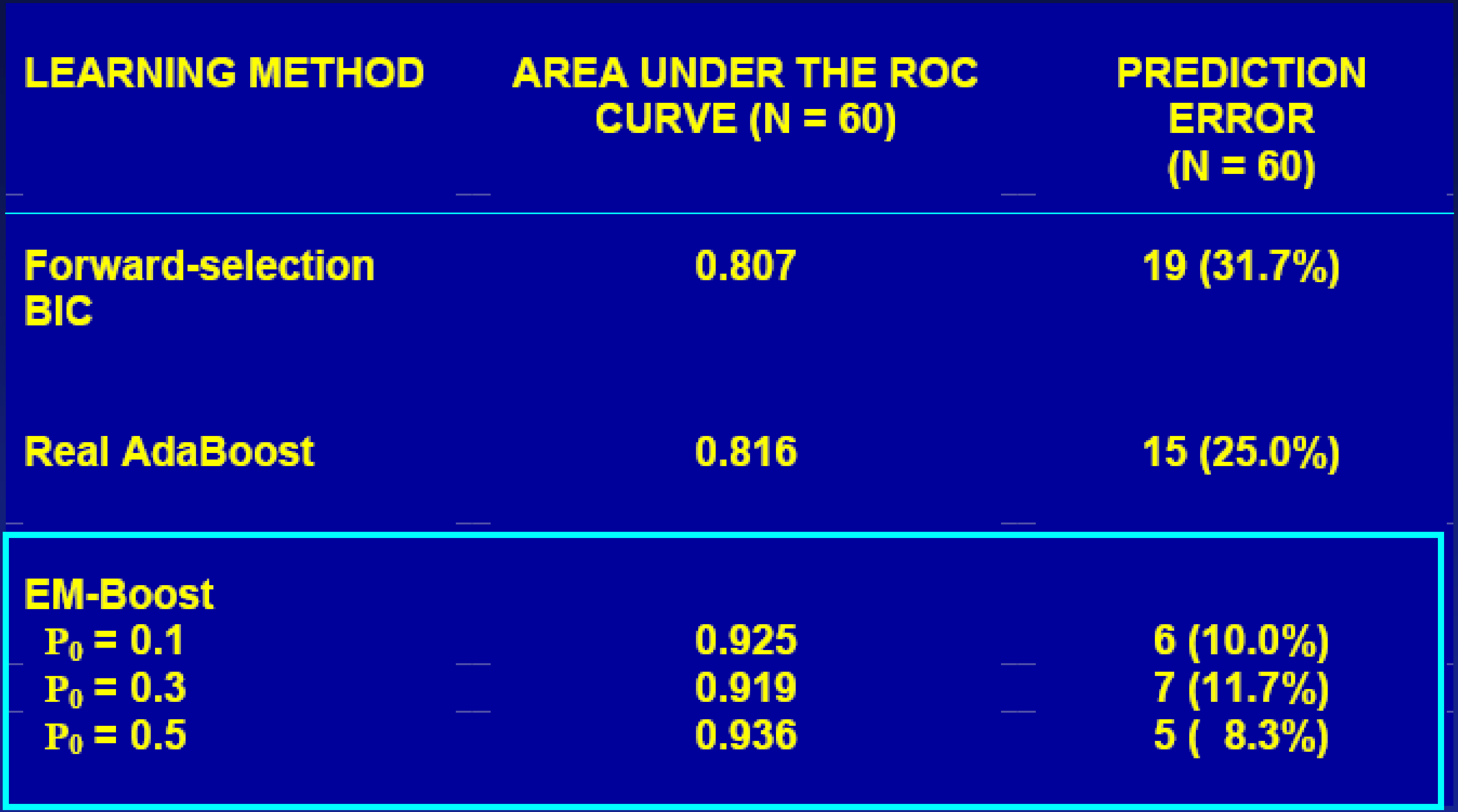
(3) EM-Boost with $P_0 = 0.1, 0.3, 0.5$ (stopped at m=100th iterations)

Study (1): Training Dataset Results



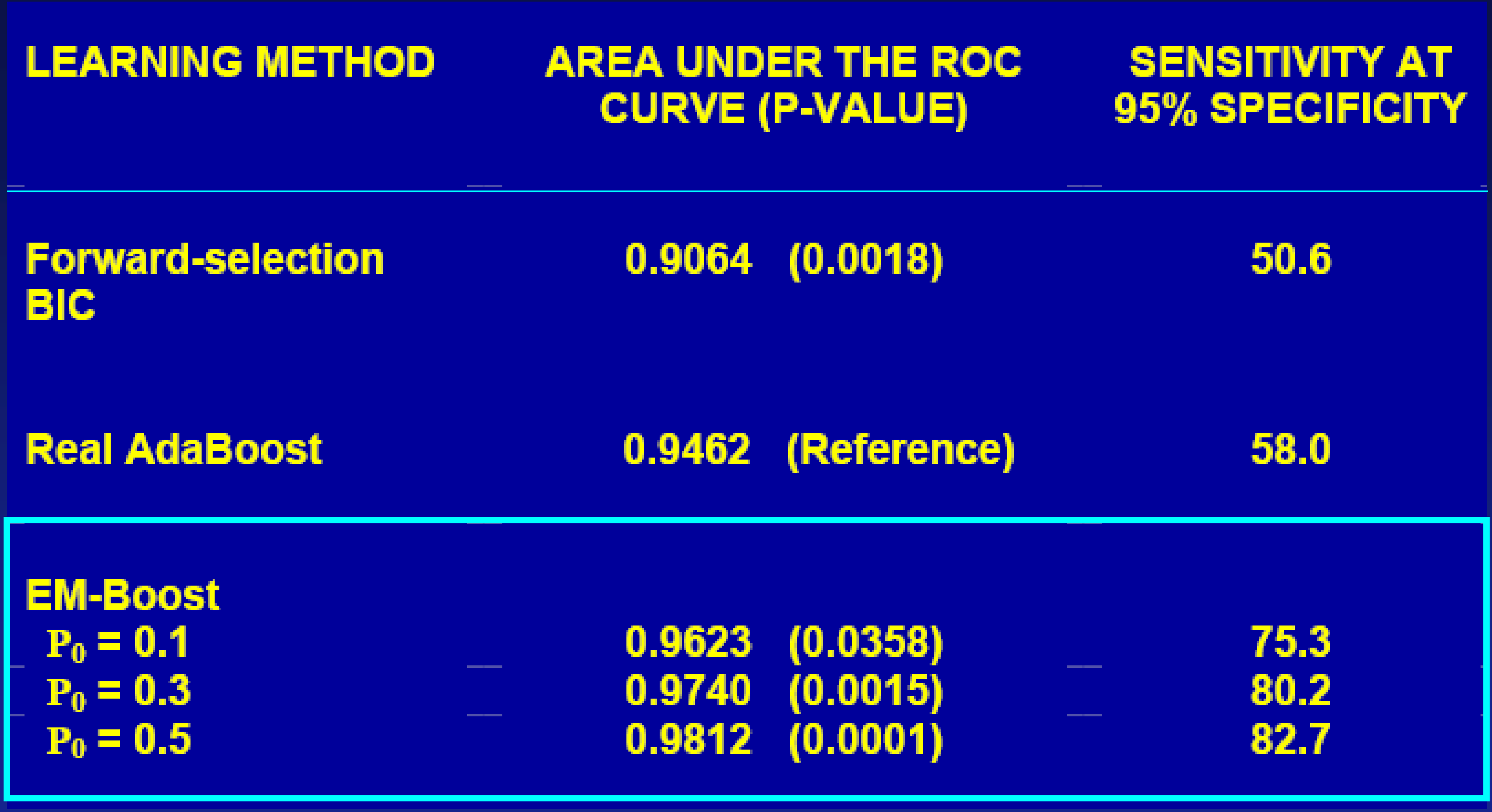
LEARNING METHOD	AREA UNDER THE ROC CURVE (P-VALUE)	SENSITIVITY AT 95% SPECIFICITY
Forward-selection BIC	0.9584 (0.0393)	65.4
Real AdaBoost	0.9741 (Reference)	79.0
EM-Boost		
P_0 = 0.1	0.9926 (0.0024)	97.5
P_0 = 0.3	0.9932 (0.0040)	97.5
P_0 = 0.5	0.9919 (0.0068)	96.3

Study (1): Test Dataset Results



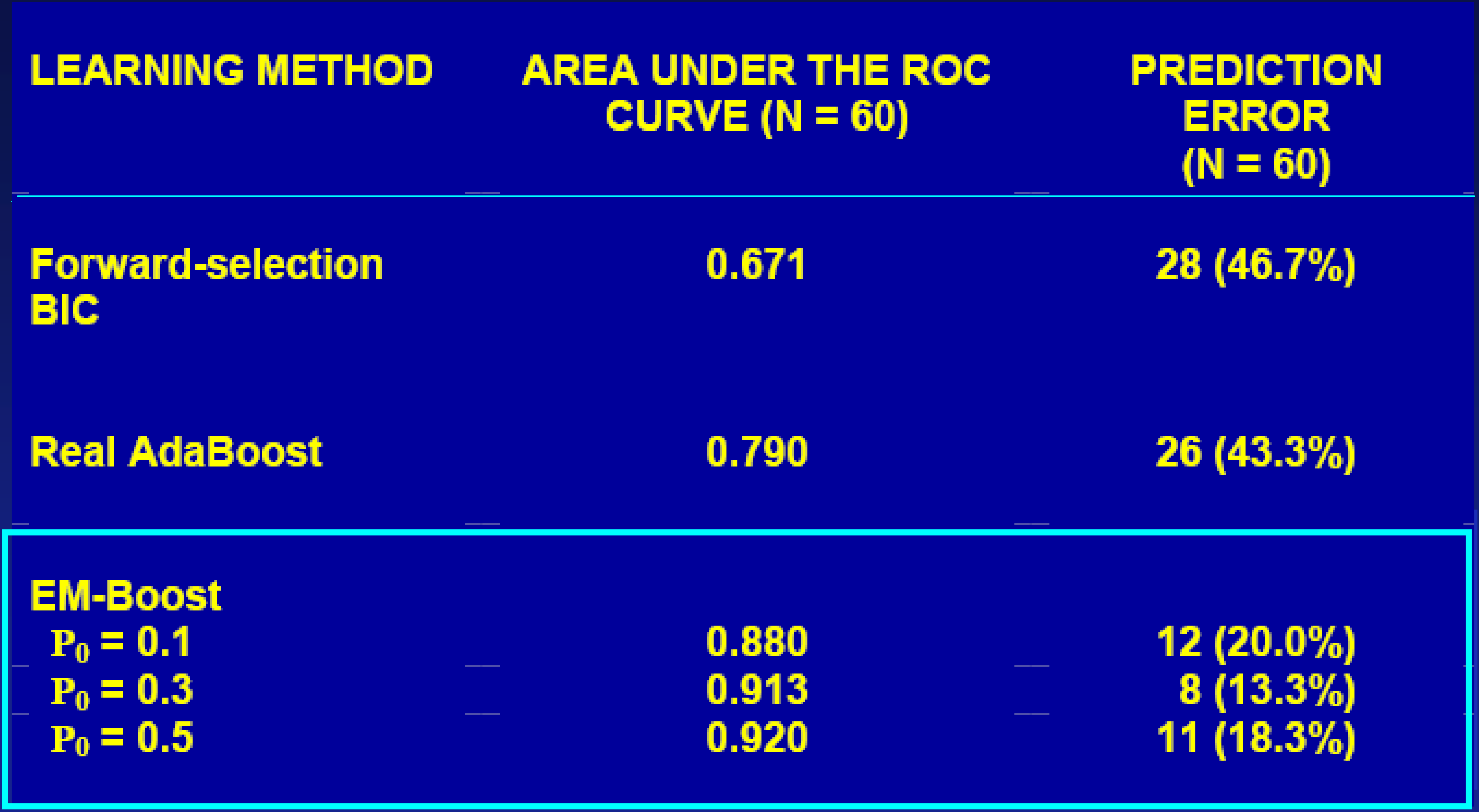
LEARNING METHOD	AREA UNDER THE ROC CURVE (N = 60)	PREDICTION ERROR (N = 60)
Forward-selection BIC	0.807	19 (31.7%)
Real AdaBoost	0.816	15 (25.0%)
EM-Boost		
$P_0 = 0.1$	0.925	6 (10.0%)
$P_0 = 0.3$	0.919	7 (11.7%)
$P_0 = 0.5$	0.936	5 (8.3%)

Study (2): Training Dataset Results



LEARNING METHOD	AREA UNDER THE ROC CURVE (P-VALUE)	SENSITIVITY AT 95% SPECIFICITY
Forward-selection BIC	0.9064 (0.0018)	50.6
Real AdaBoost	0.9462 (Reference)	58.0
EM-Boost		
P_0 = 0.1	0.9623 (0.0358)	75.3
P_0 = 0.3	0.9740 (0.0015)	80.2
P_0 = 0.5	0.9812 (0.0001)	82.7

Study (2): Test Dataset Results



LEARNING METHOD	AREA UNDER THE ROC CURVE (N = 60)	PREDICTION ERROR (N = 60)
Forward-selection BIC	0.671	28 (46.7%)
Real AdaBoost	0.790	26 (43.3%)
EM-Boost		
$P_0 = 0.1$	0.880	12 (20.0%)
$P_0 = 0.3$	0.913	8 (13.3%)
$P_0 = 0.5$	0.920	11 (18.3%)

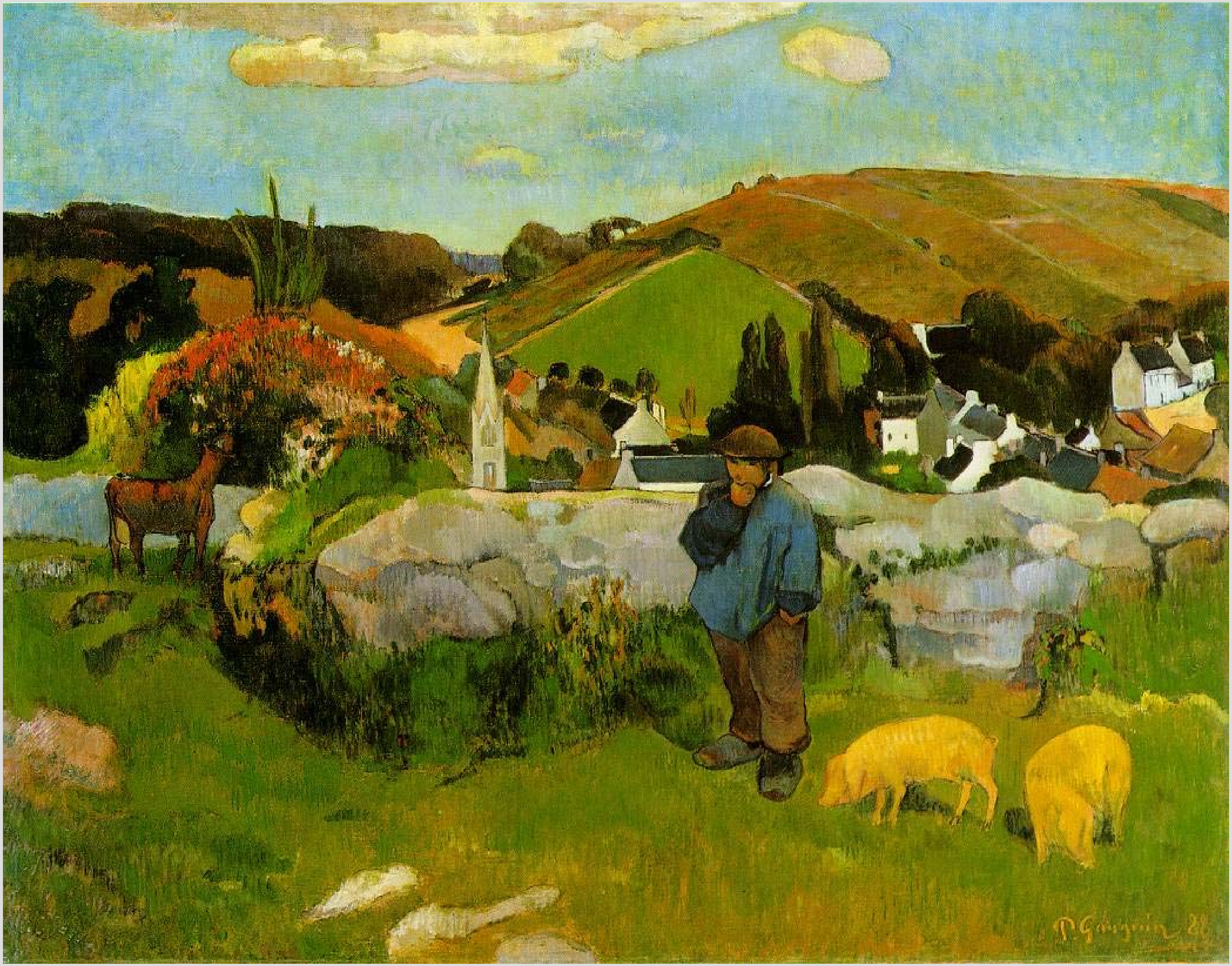
Summary

- Pre-analysis processing is crucial for a proper analysis

- Avoiding overfitting is the key in classifier building with multiple biomarkers

- In biomedical applications, imperfect class labels are common

- EM-Boost modifies the boosting algorithm to accommodate potential mislabeling: allows “learning” in partially supervised settings


P Gauguin 88

$$\underline{\Pr[y_i^* \mid \underset{\%}{\phi}_{(m-1)}, \underset{\%}{X}^{(m-1)}, \underset{\%}{y}]}$$

$$\begin{cases} \Pr[y_i^* = 1 \mid \underset{\%}{\phi}_{(m-1)}, \underset{\%}{X}^{(m-1)}, \underline{y_i = 1}] = 1 \\ \Pr[y_i^* = -1 \mid \underset{\%}{\phi}_{(m-1)}, \underset{\%}{X}^{(m-1)}, \underline{y_i = 1}] = 0 \end{cases}$$

$$\pi_i^{(m)} = \Pr[y_i^* = 1 \mid \underset{\%}{\phi}_{(m-1)}, \underset{\%}{X}^{(m-1)}, \underline{y_i = -1}]$$

$$\ln \frac{\pi_i^{(m)}}{1-\pi_i^{(m)}} = \ln \frac{\pi_i^{(m-1)}}{1-\pi_i^{(m-1)}} + \beta_{m-1}(X_i^{(m-1)} - \overline{X^{(m-1)}})$$

$$= \ln \frac{\boxed{\pi_i^{(0)}}}{1-\pi_i^{(0)}} + \sum_{j=1}^{m-1} \beta_j (X_i^{(j)} - \overline{X^{(j)}})$$

Initial value: P_0