Evaluation of Classification Models

Probability and Statistics for Data Science

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Classification

Goal: Assign one of several predefined classes based on features

Binary classification: 2 classes (*Positive* and *Negative*)

Diagnosis of Alzheimer's disease

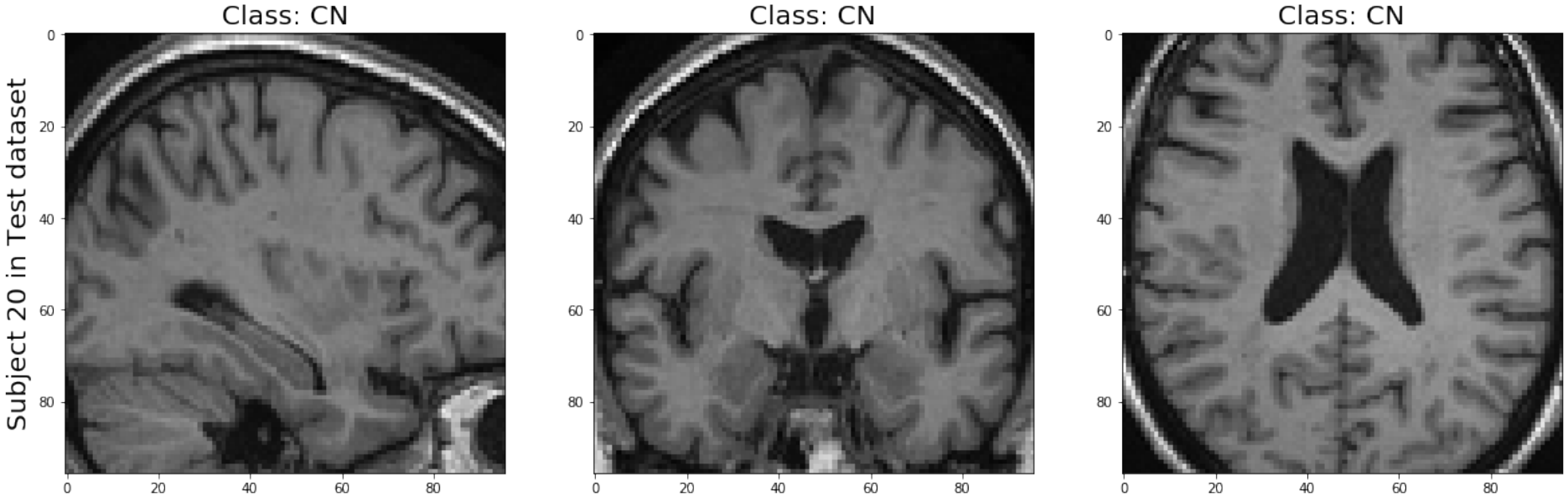
Neurodegenerative disease causing $60 - 70\%$ cases of dementia

Diagnosis via positron-emission tomography is invasive and costly

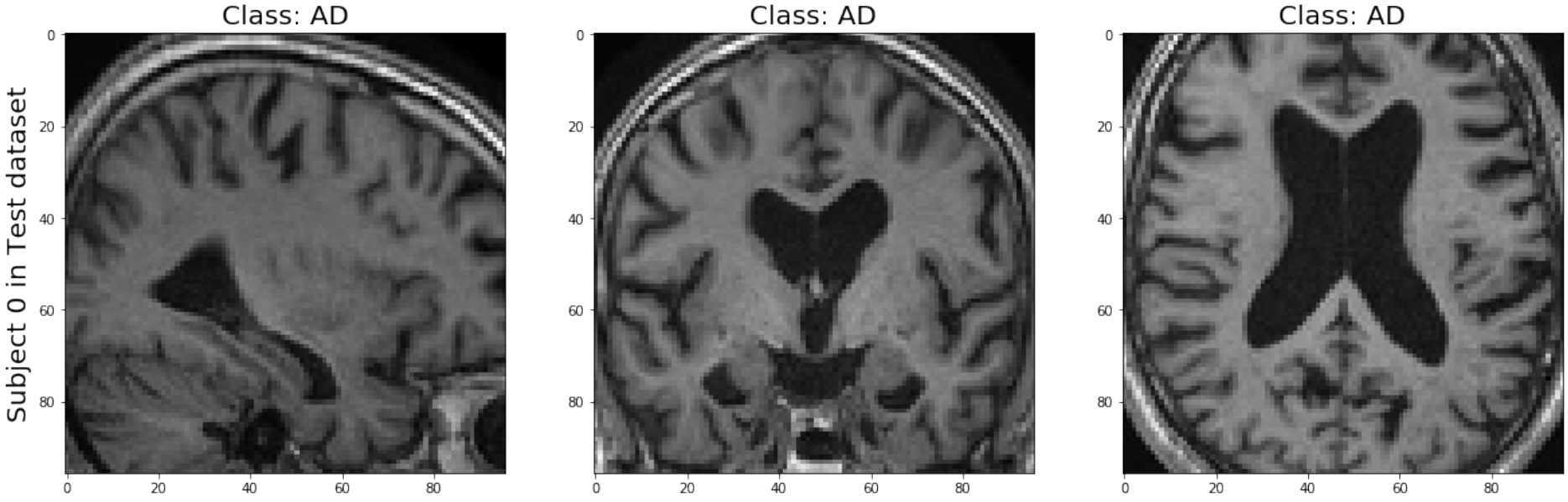
Structural MRI is non-invasive and less costly

Goal: Diagnose Alzheimer's using MRI scans

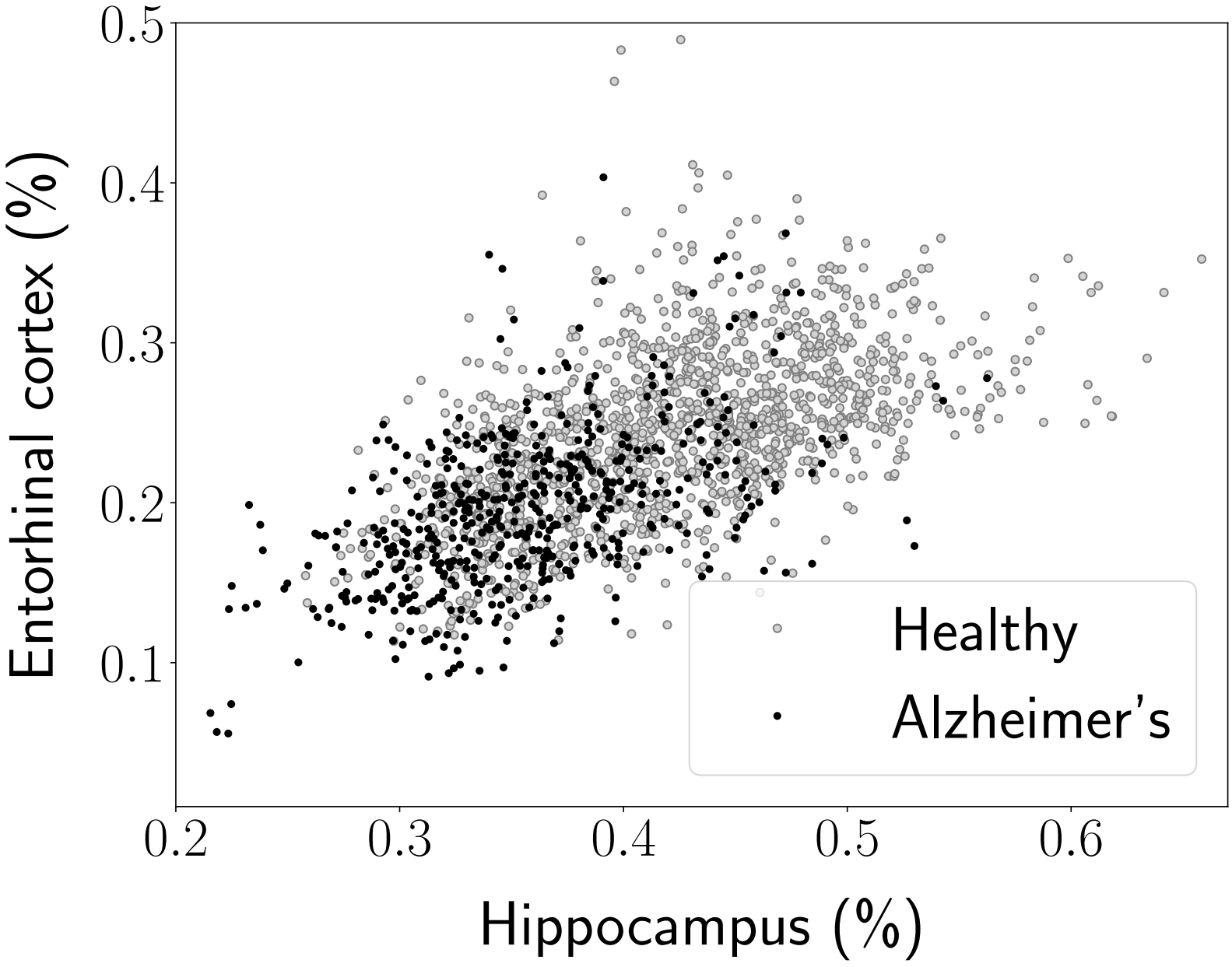
Cognitively-normal patient



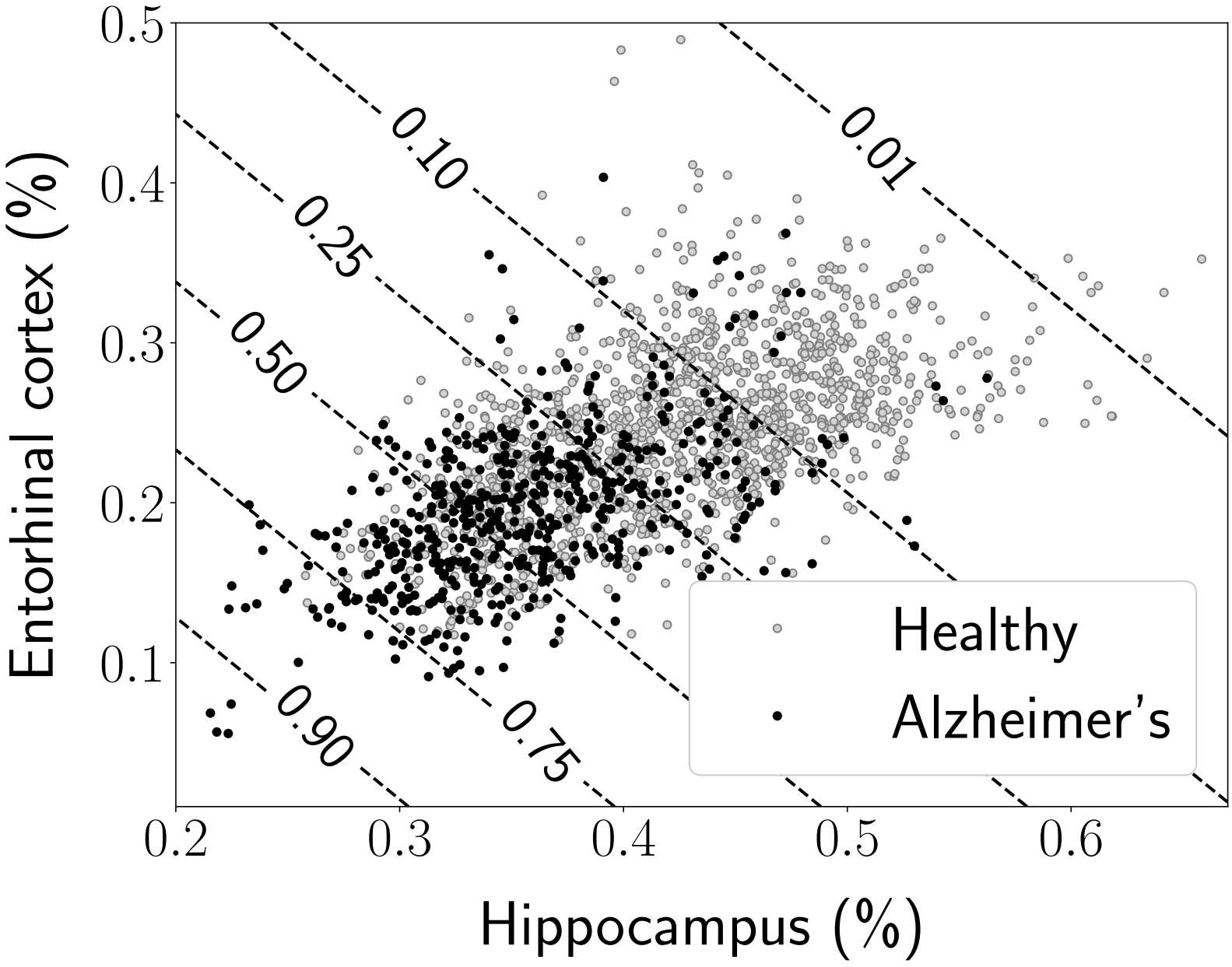
Alzheimer's patient



Alzheimer's diagnosis



Logistic regression



Plan

How should we evaluate?

- Predict Alzheimer's if estimated probability > 0.5
- Compute accuracy (fraction of correct estimates)

Accuracy is 81.0%

End of story?

Not really

Fraction of Alzheimer's patients diagnosed correctly?
We don't know!

Consider classifier that classifies *everyone as healthy*

Accuracy is 78.4%, but Alzheimer's is never detected

We need more informative metrics

Binary classification

P examples with positive labels

N examples with negative labels

True / Est.	Negative	Positive
Negative	**True Negatives** (TN)	**False Negatives** (FN)
Positive	**False Positives** (FP)	**True Positives** (TP)

$$\mathrm{TP} + \mathrm{FN} = \mathrm{P}$$
$$\mathrm{TN} + \mathrm{FP} = \mathrm{N}$$

Accuracy

Fraction of examples that are correctly classified

$$\text{Accuracy} := \frac{\text{TN} + \text{TP}}{\text{N} + \text{P}}$$

AD diagnostics

Est. \ True	Healthy	AD
Healthy	1579	356
AD	24	86

$$\text{Accuracy} := \frac{\text{TN} + \text{TP}}{\text{N} + \text{P}} = 0.81$$

True positive rate (TPR), a.k.a. recall, sensitivity, hit rate

Fraction of positive examples that are correctly classified

$$\text{TPR} := \frac{\text{TP}}{\text{P}}$$

AD diagnostics

Est. \ True	Healthy	AD
Healthy	1579	356
AD	24	86

$$\text{TPR} := \frac{\text{TP}}{\text{P}} = 0.19$$

False positive rate (FPR)

Fraction of negative examples that are incorrectly classified

$$\text{FPR} := \frac{\text{FP}}{\text{N}}$$

$1 - \text{FPR}$ is known as specificity, a.k.a. true negative rate, selectivity

$$\text{Specificity} := \frac{\text{TN}}{\text{N}}$$

AD diagnostics

Est. \ True	Healthy	AD
Healthy	1579	356
AD	24	86

$$\text{FPR} := \frac{\text{FP}}{\text{N}} = 0.01$$

Precision, a.k.a. positive predictive value

Fraction of examples predicted as positive that are correctly classified

$$\text{Precision} := \frac{\text{TP}}{\text{TP} + \text{FP}}$$

AD diagnostics

Est. \ True	Healthy	AD
Healthy	1579	356
AD	24	86

$$\text{Precision} := \frac{\text{TP}}{\text{TP} + \text{FP}} = 0.78$$

F1 score

TPR and precision are fractions with the same numerator (TP)

We can combine them with a harmonic mean

$$\text{F1 score} := \frac{2 \cdot \text{TPR} \cdot \text{Precision}}{\text{TPR} + \text{Precision}}$$

Alzheimer's diagnostics

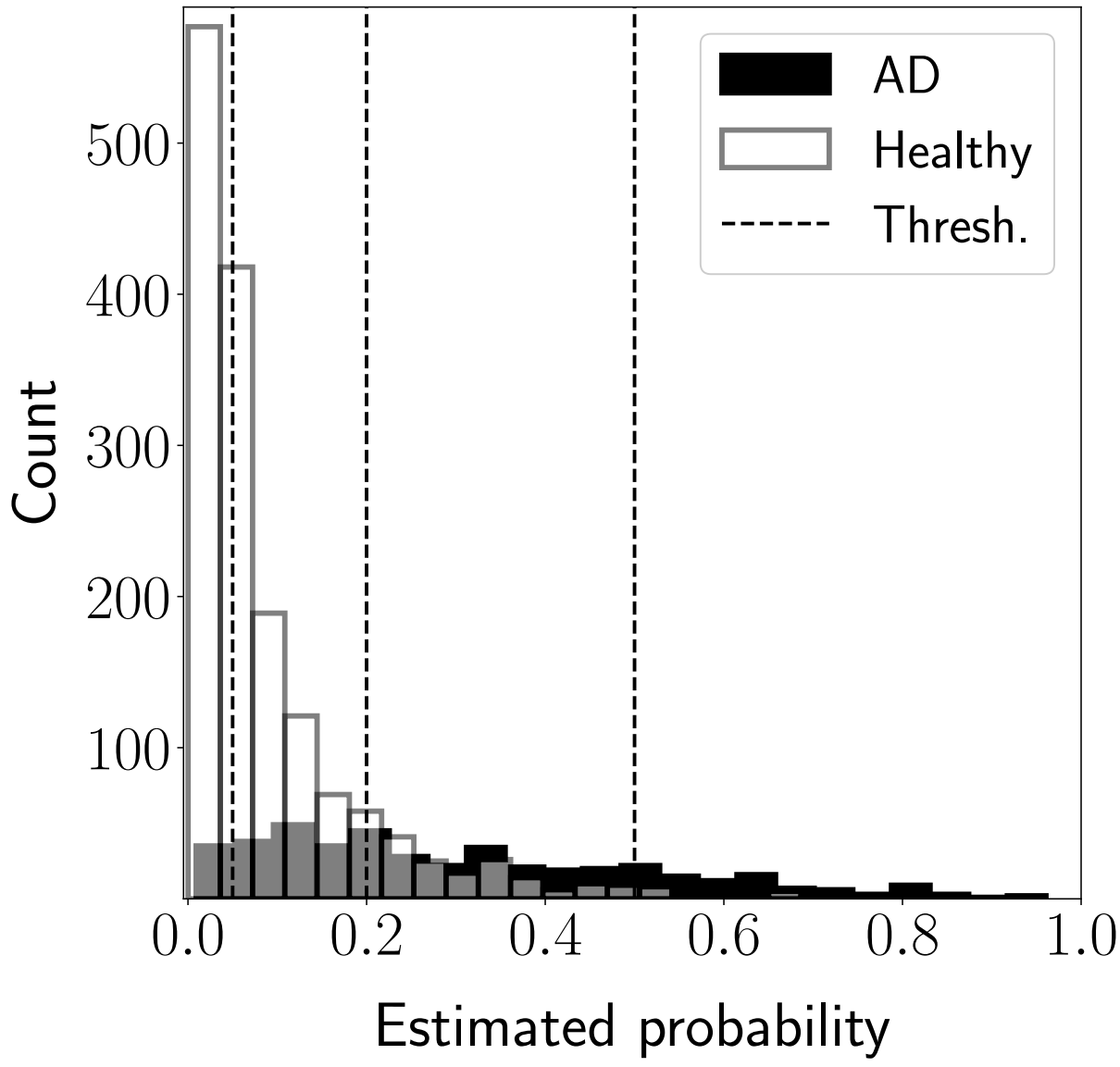
$$\text{F1 score} := 0.31$$

Classifiers output estimated probabilities

- Naive Bayes
- Gaussian discriminant analysis
- Logistic regression
- Classification trees
- Neural networks

Metrics depend on threshold used to determine positive / negative estimates

Alzheimer's diagnostics



Alzheimer's diagnostics

Threshold = 0.05

Est. \ True	H	AD
H	774	31
AD	829	411

Accuracy = 0.58
TPR = 0.93
FPR = 0.52
Precision = 0.33
F1 score = 0.49

Threshold = 0.2

Est. \ True	H	AD
H	1399	173
AD	204	269

Accuracy = 0.82
TPR = 0.61
FPR = 0.13
Precision = 0.57
F1 score = 0.59

Threshold = 0.5

Est. \ True	H	AD
H	1579	356
AD	24	86

Accuracy = 0.81
TPR = 0.19
FPR = 0.01
Precision = 0.78
F1 score = 0.31

Goal

Quantify discriminative ability of probability estimates

TPR - FPR tradeoff

Threshold = 0.05

Est. \ True	H	AD
H	774	31
AD	829	411

Accuracy = 0.58

TPR= 0.93

FPR= 0.52

Precision = 0.33

F1 score = 0.49

Threshold = 0.2

Est. \ True	H	AD
H	1399	173
AD	204	269

Accuracy = 0.82

TPR= 0.61

FPR= 0.13

Precision = 0.57

F1 score = 0.59

Threshold = 0.5

Est. \ True	H	AD
H	1579	356
AD	24	86

Accuracy = 0.81

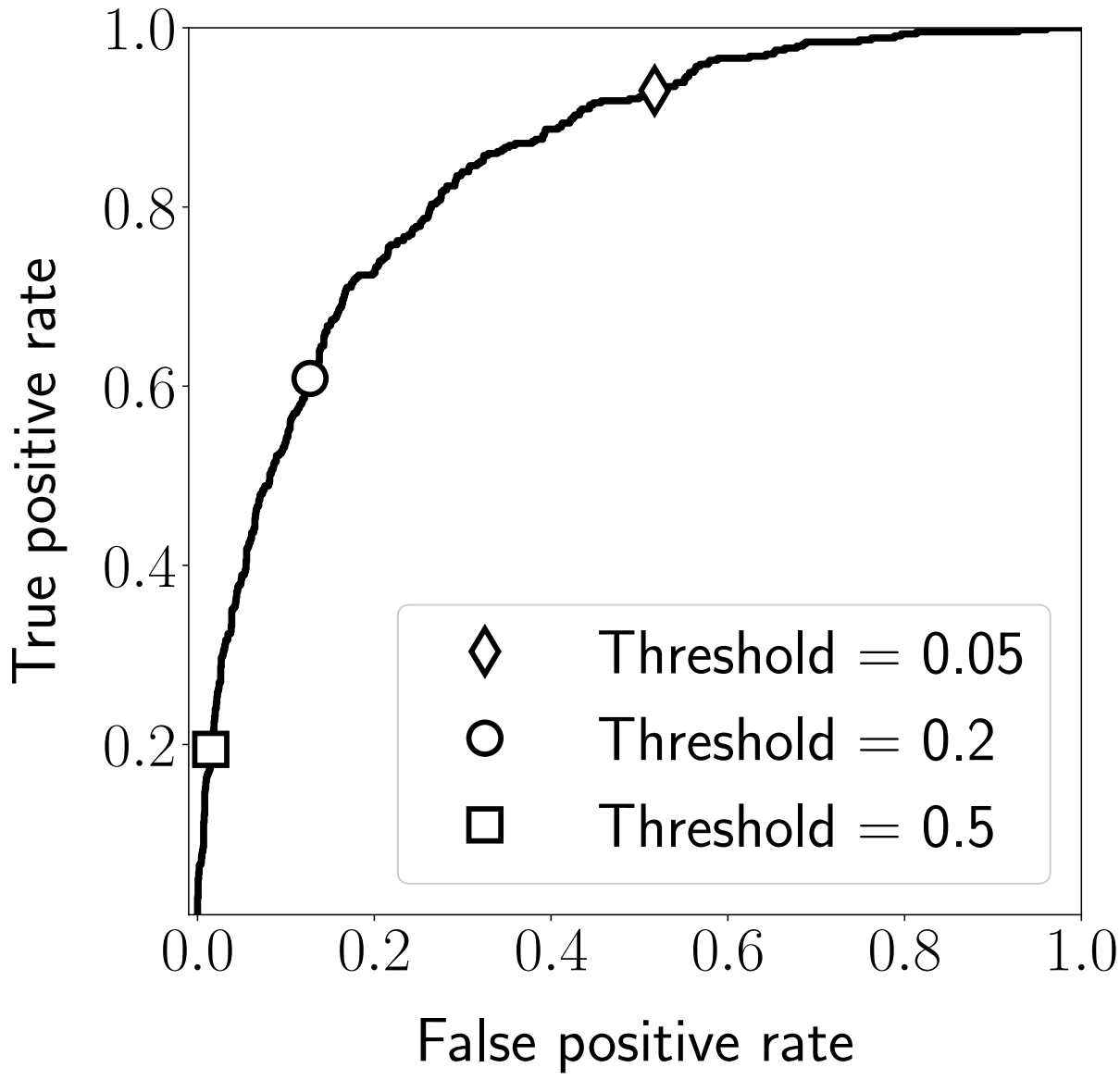
TPR= 0.19

FPR= 0.01

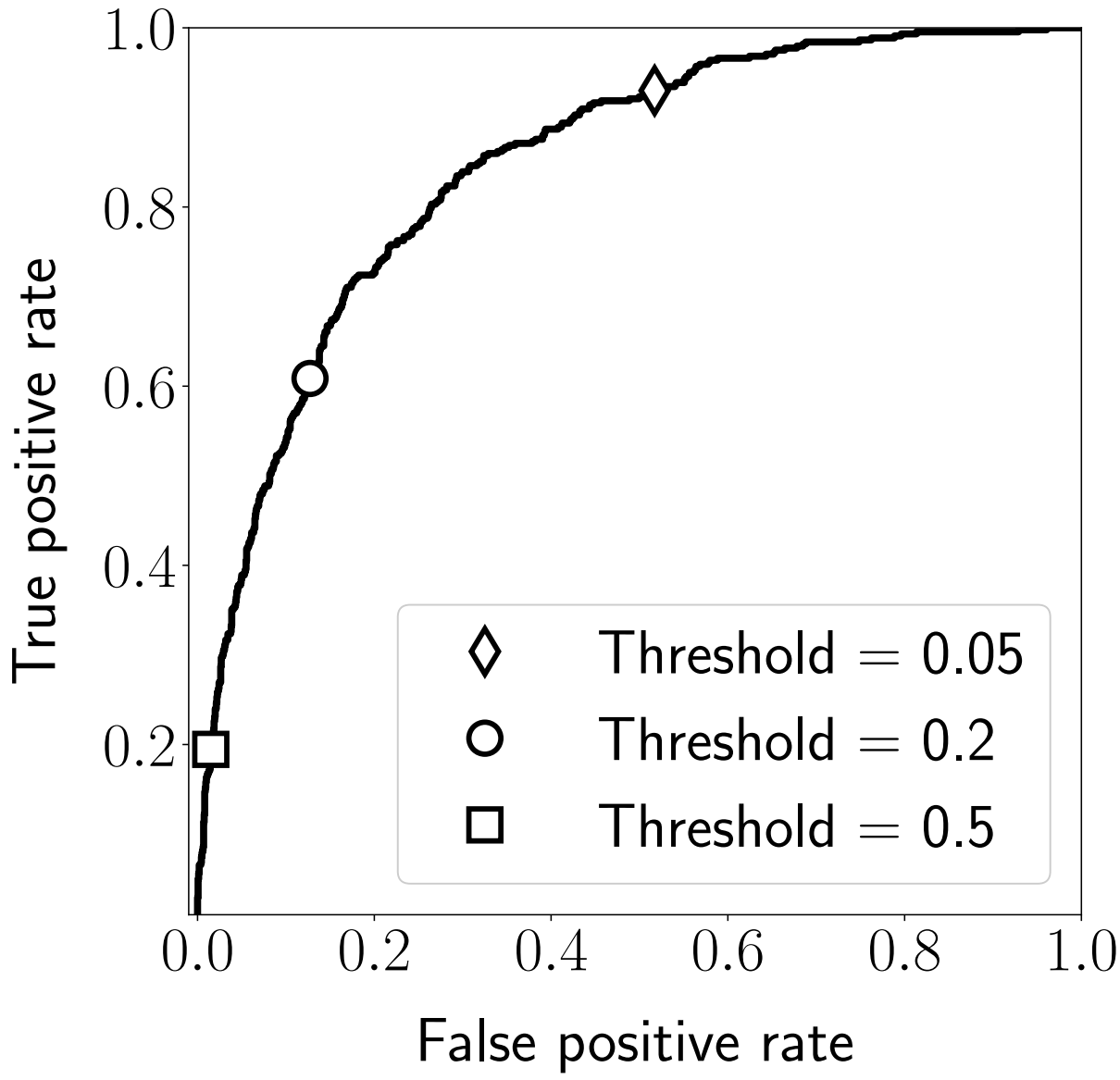
Precision = 0.78

F1 score = 0.31

Receiver operating characteristic (ROC) curve



Area under ROC curve (AUROC or AUC) = 0.847



Concordance or c-index

Threshold-free measure of discrimination performance

Fraction of negative - positive examples such that estimated probability is higher for positive example

Equal to AUC

Is discrimination all we care about?

Classifier that assigns:

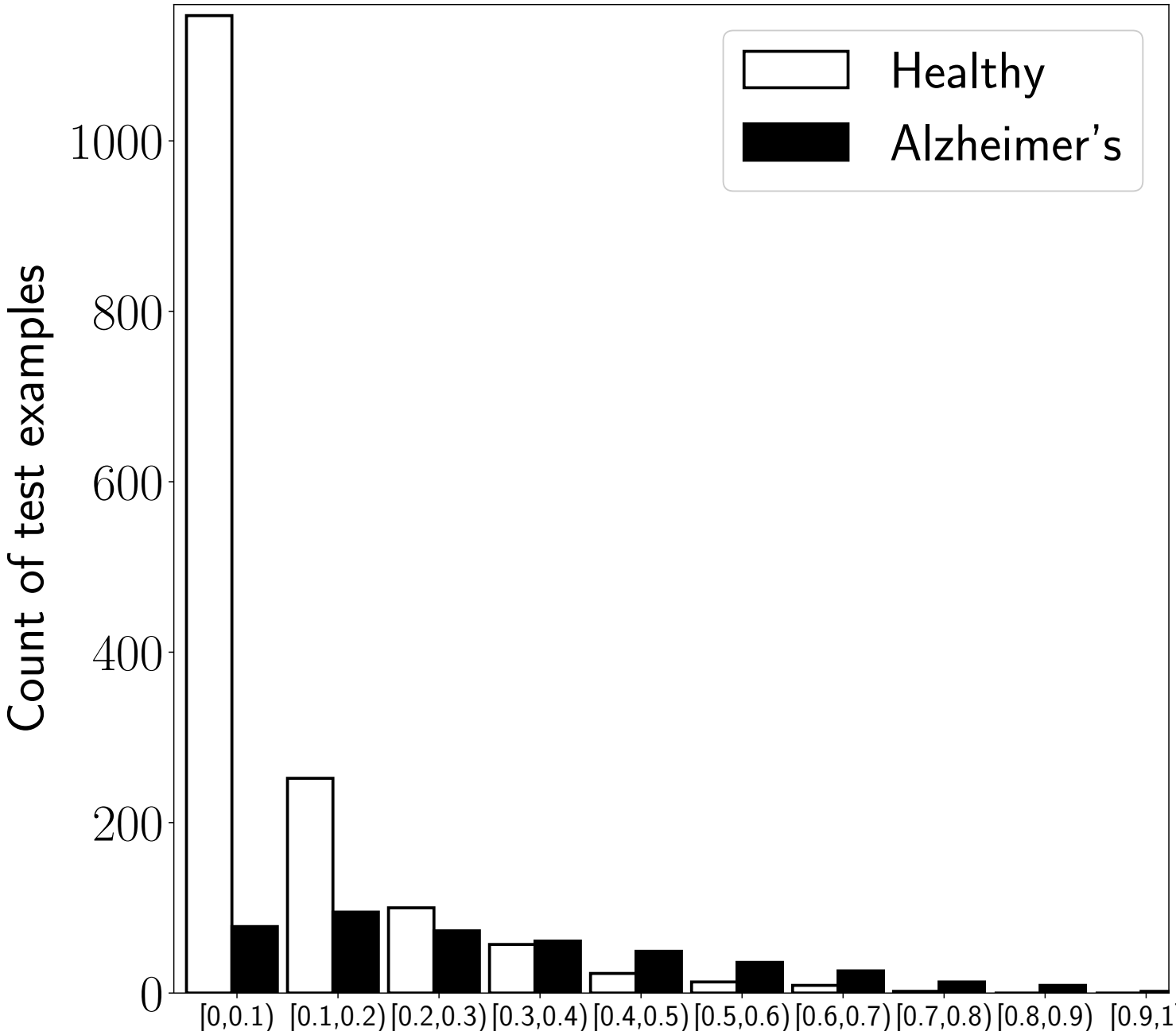
- 0.8 $\rightarrow$ healthy subjects
- 0.9 $\rightarrow$ AD patients

AUC? 1! Perfectly discriminative

Among examples assigned a probability estimate of 0.8, how many have AD? 0

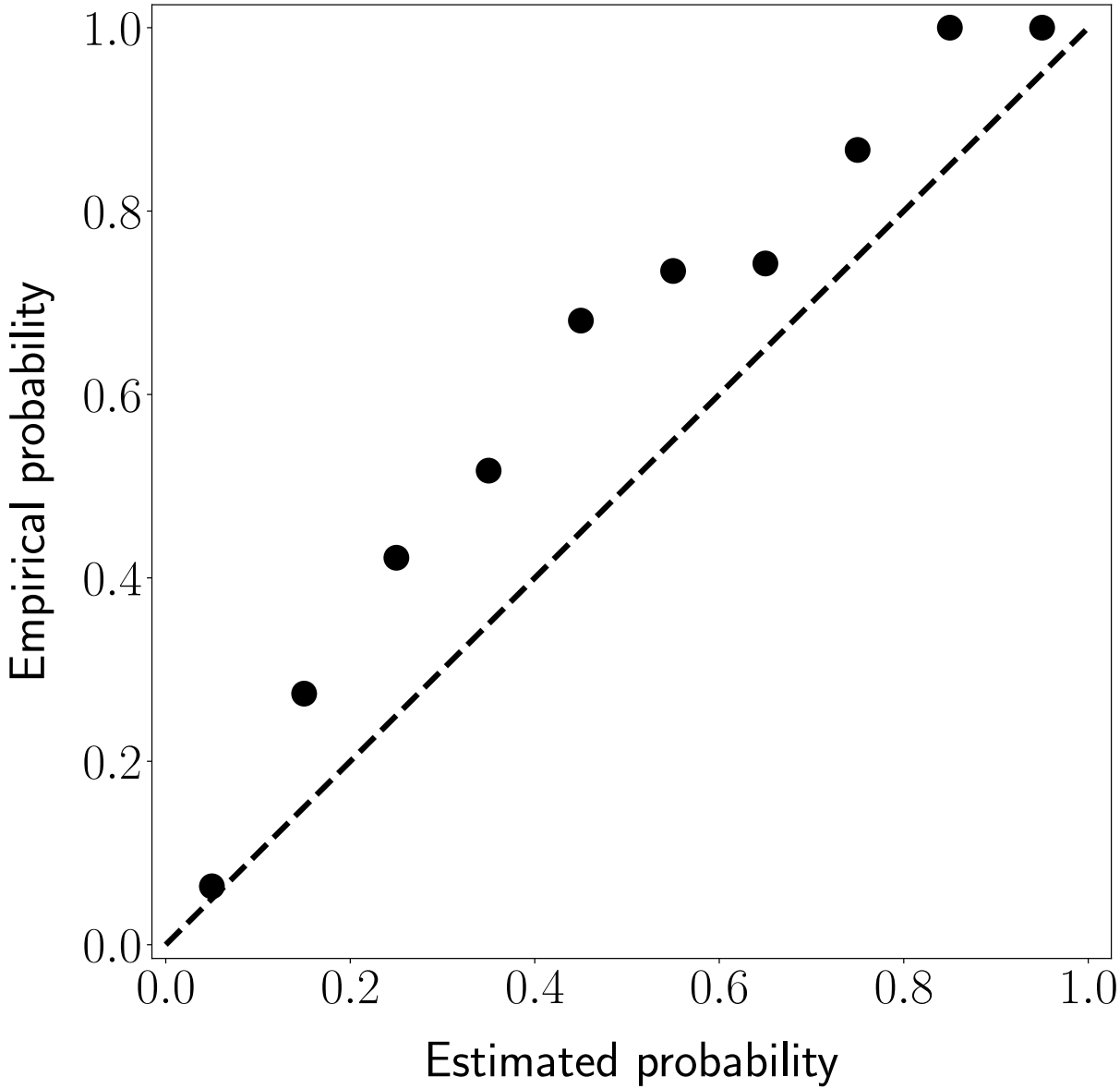
Probability estimates are totally wrong, the model is uncalibrated

Estimated probabilities vs empirical probabilities



Reliability diagram

Evaluates model calibration



Is calibration enough?

For dataset with 21.6% of AD patients

Calibration of model that assigns 0.216 to every data point?

Only one bin, empirical probability = 0.216

Perfectly calibrated, is this a good classifier?

It's terrible! Completely non-discriminative (ignores the features)

Brier score

Given dataset (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)

$$\text{Brier score} := \frac{1}{n} \sum_{i=1}^{n} (y_i - p_{\text{est}}(x_i))^2$$

Evaluates both discrimination and calibration

Perfectly discriminative, but uncalibrated model: 0.504

Perfectly calibrated, but undiscriminative model: 0.169

Our model: 0.131

What have we learned?

How to evaluate binary classification models:

- Accuracy, TPR, FPR, Precision, F1 score
- AUC / concordance
- Calibration
- Brier score