Measuring the Prediction Performance of Medical Screening Tests

Yan Yuan
School of Public Health
University of Alberta

May 8, 2015

Joint work with Dr. Wanhua Su and Dr. Mu Zhu

Brief Biography Yan Yuan

- 2011- present, Assistant Prof; School of Public Health, University of Alberta
- 2008-2011, Biostatistician, Population Health Research, Cancer Control, Alberta Health Services
- 2003 MMath, 2008 PhD in Statistics, University of Waterloo
- 1999-2001 Lab manager and research technician in Animal behavior lab, University of Guelph, Canada
- 1999, MSc in Animal behavior, Michigan State University, USA
- 1996 BSc in Biochemistry, Nanjing University, China

Outline

- Motivation
  - Detecting the Rare Events (low prevalence/incidence)

10-year cancer diagnosis per 1000 person	Colorectal cancer		Breast cancer	Prostate cancer
	Male	Female		
Age 50	6.8	5.2	23	22
Age 60	13	9	35	63

- Metrics for evaluating medical test
- The relation between two single numeric summary metrics
- Variance of AP
- Examples
- Summary and future work

Predicting the Rare Class

- Cancer screening: detect from the *asymptomatic* population the diseased subjects, who make up a very small proportion (typically < 1%).
- Risk models (for general population): CVD, diabetes, chronic pulmonary diseases
- Drug discovery: identify potential chemical compounds that are biologically active for some target (typically < 5%).
- Information retrieval

Medical Screening Tests

- Screening aims at detecting disorders at an early *asymptomatic* stage
- Its utility is determined by its ability to detect the disorder, measured by positive predictive value (PPV)
- The current evaluation metrics for medical tests
  - Sensitivity, Specificity, Diagnostic likelihood ratios, Predictive values
  - Receiver operating characteristic (ROC) curve

Motivating Data 1

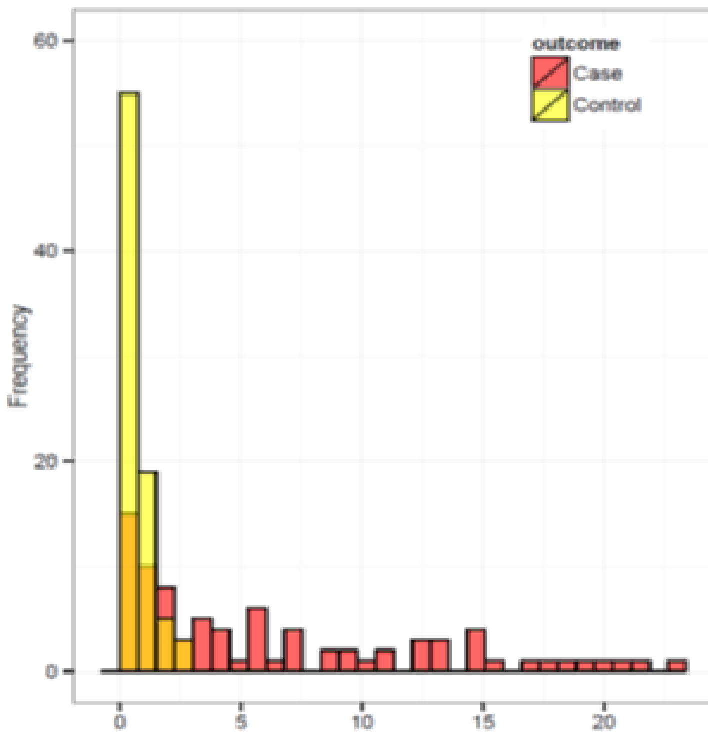
Digital Mammography Imaging Screening Trial (Pisano et al. 2005 *New England Journal of Medicine*)

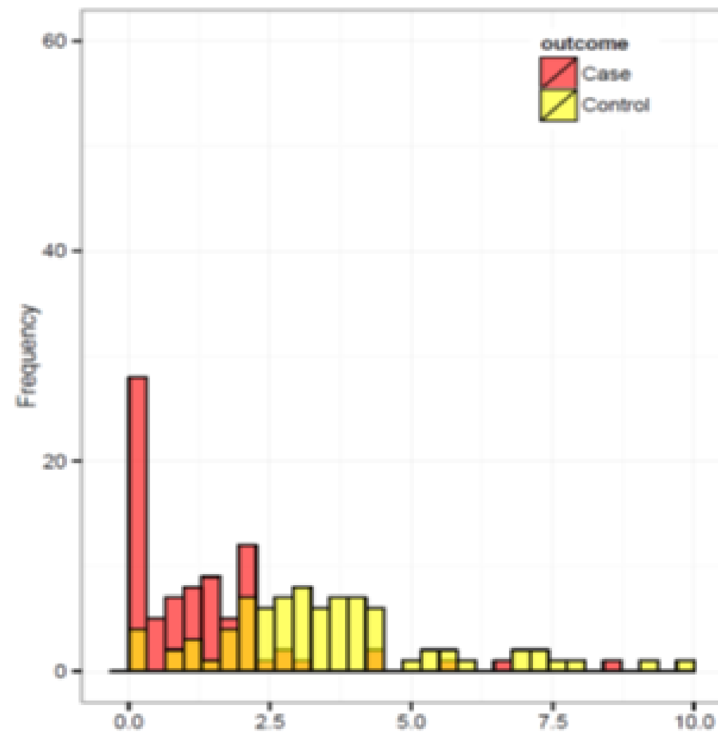
Malignancy score		7	6	5	4	3	2	1	Total
Digital	Category Total	11	29	69	1061	2224	6588	32588	42570
	Cancers	10	18	25	85	49	25	122	334
Film	Category Total	17	29	70	942	2291	6910	32486	42745
	Cancers	13	24	25	74	35	33	131	335

Motivating Data 2

Mass spectrometry data for prostate cancer (Adam *et al*. 2002 Cancer Research)

- 779 potential biomarkers were assessed in 83 late-stage prostate cancer patients and 82 normal subjects.





Performance Measures for Medical tests (classifiers)

- Threshold Dependent Measure
  - Misclassification rate
  - Sensitivity and Specificity
  - Positive and Negative Predictive Value
- Threshold Independent Measure
  - Area Under the ROC* Curve (AUC or aROC)
  - Average positive predictive value (AP)

*Receiver Operating Characteristic

AP

- Definition

$\{Y_{(1)}, Y_{(2)}, Y_{(3)}, \ldots Y_{(m)}, \ldots, Y_{(n)}\}$. where *Y* is the true binary class label.

Positive predictive value at $Y_{(m)}$:

$$PPV_m = \frac{\sum_{i=1}^{m} Y_{(i)}}{m}$$

(i.e. the proportion of class 1 subjects in the top m ranked subjects)

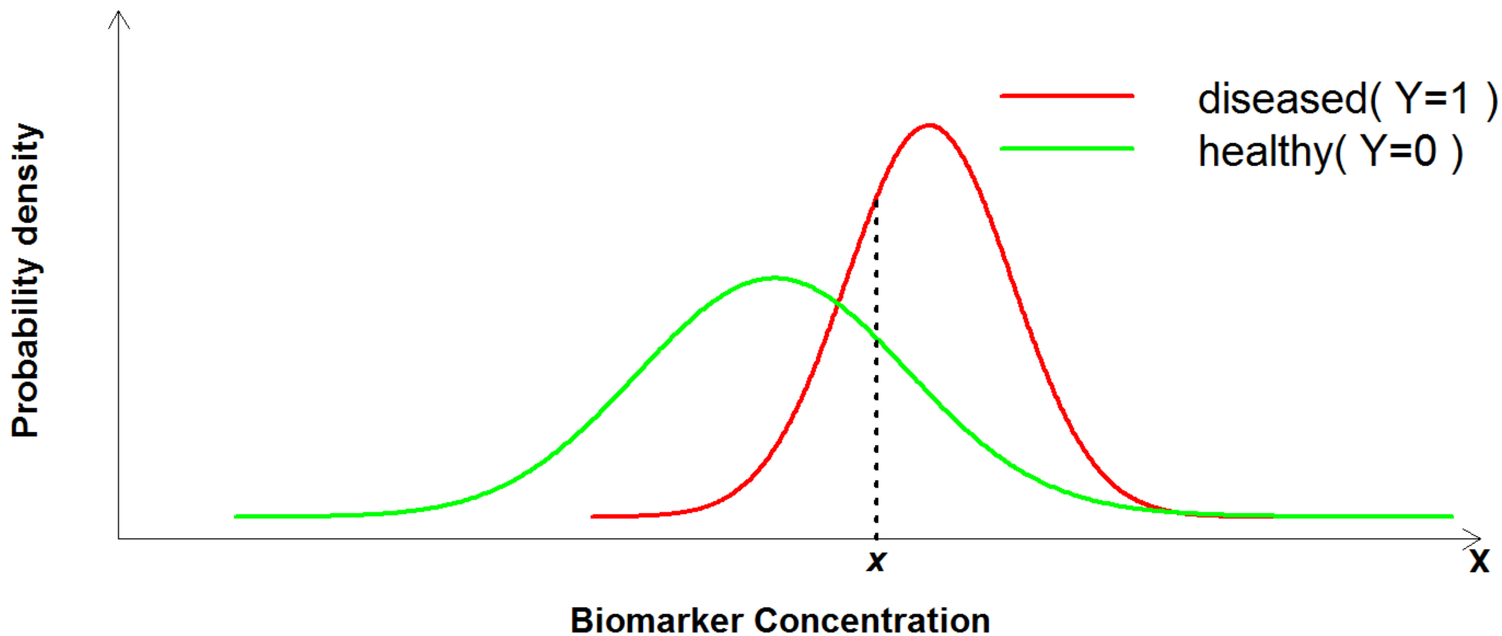
$$AP = \frac{1}{n_1} \sum_{m=1}^{n} Y_{(m)} PPV_m ,$$

where $n_1 = \sum_{m=1}^{n} Y_{(m)}$, total number of class 1 subjects

An Illustration Example

Rank	Classifier 1	Classifier 2	Classifier 3
1	1	1	0
2	1	0	0
3	1	1	1
4	0	0	1
5	0	1	1
AP	1	0.76	0.48

Definition of AUC and AP



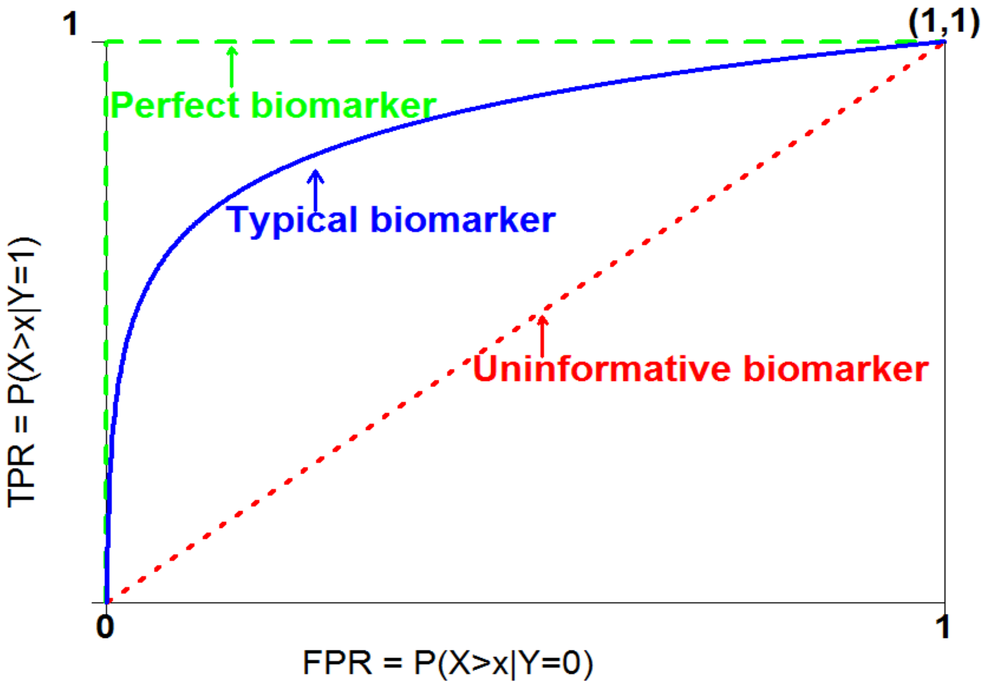
Notations

$\pi \quad = P(Y=1)$

$s \quad = P(X > x) = G_X(x)$

$h(s) = P(X > x, Y=1) = \pi F_1(x)$

ROC curve



$$\pi \quad = P(Y=1)$$

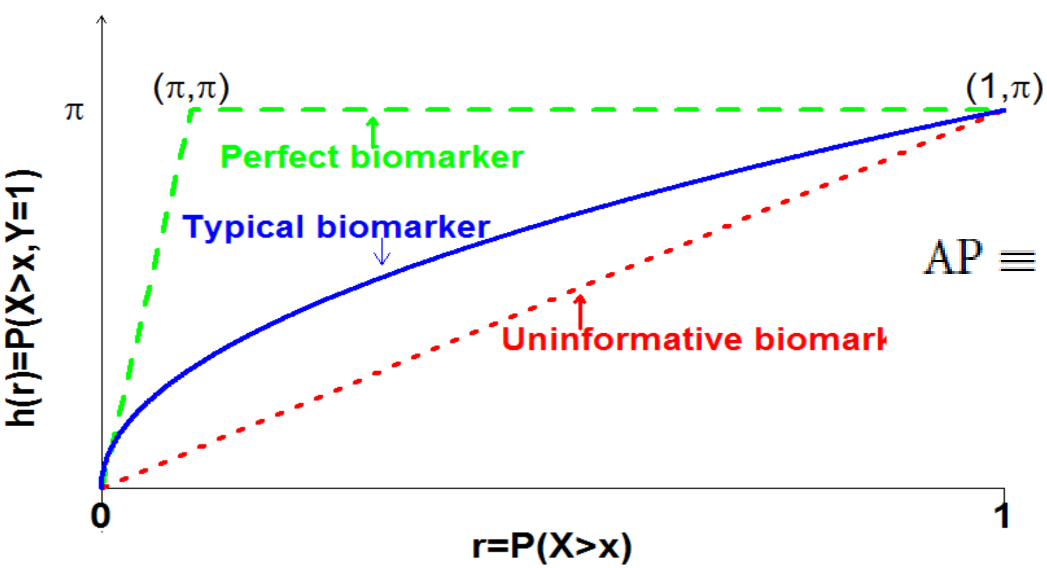
$$s \quad = P(X > x)$$

$$h(s) = P(X > x, Y=1) = \pi\, S_1(x)$$

$$\mathrm{AUC} \equiv \int_0^1 \mathrm{TPF}(s)\,\mathrm{dFPF}(s)$$

$$= \frac{1}{\pi(1-\pi)}\left[\int_0^1 h(s)ds - \frac{\pi^2}{2}\right]$$

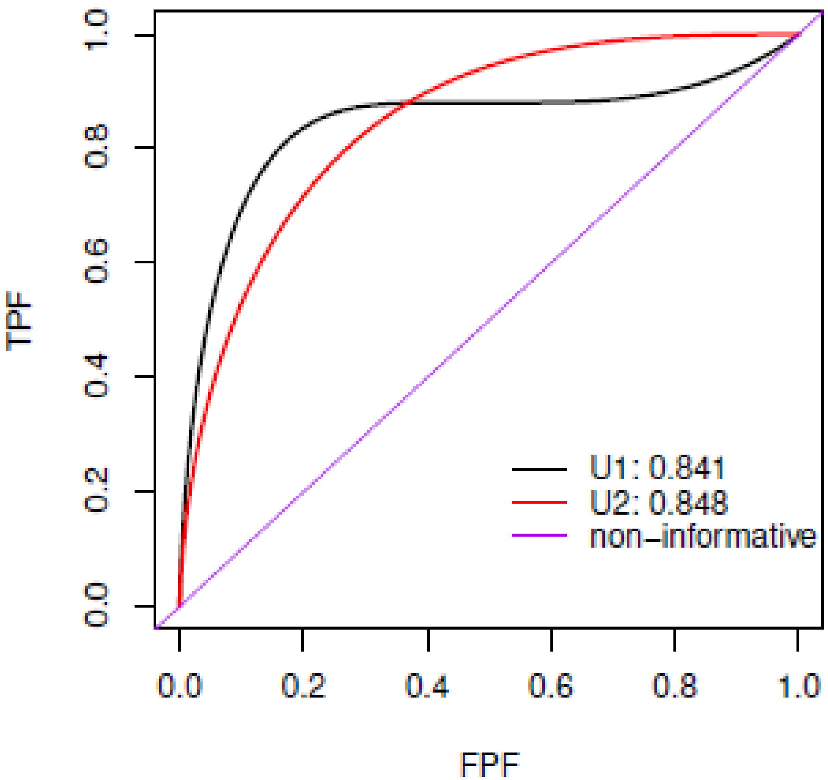
Hit curve

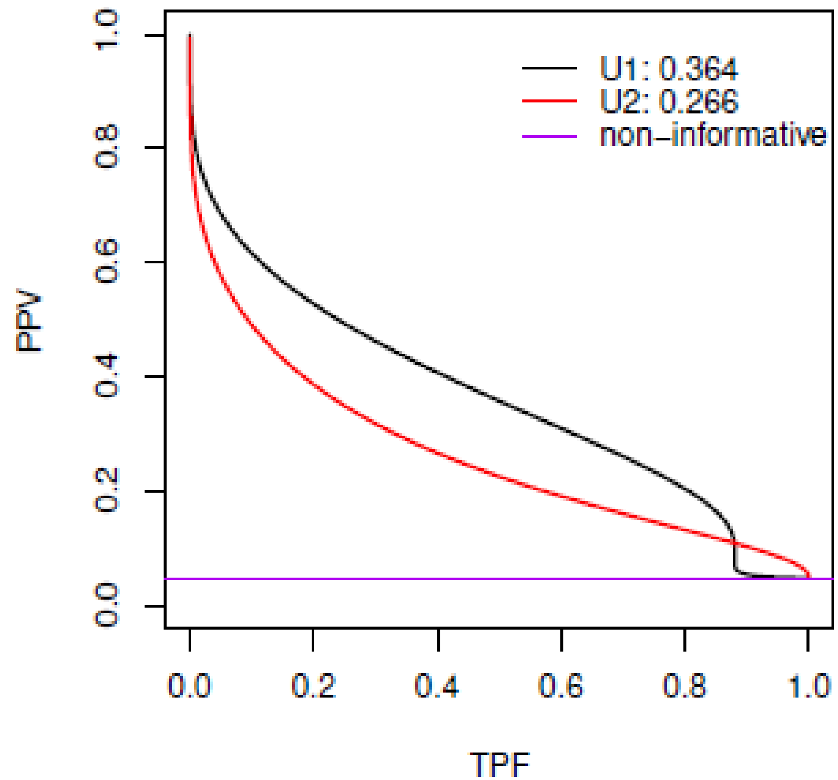


$$\mathrm{AP} \equiv \int_0^1 \mathrm{PPV}(s)\,\mathrm{dTPF}(s) = \frac{1}{\pi}\int_0^1 \frac{h(s)}{s}\,\mathrm{dh}(s).$$

aROC vs aPR

$$\log(T_i) = 7.2 - 1.1U_{i1} - 2.5U_{i2} - 1.5log(U_{i1}^2) + \epsilon_T,$$





Example 1

Digital Mammography Imaging Screening Trial (Pisano et al. 2005 *New England Journal of Medicine*)

Malignancy score		7	6	5	4	3	2	1	Total
Digital	Category Total	11	29	69	1061	2224	6588	32588	**42570**
	Cancers	10	18	25	85	49	25	122	**334**
Film	Category Total	17	29	70	942	2291	6910	32486	**42745**
	Cancers	13	24	25	74	35	33	131	**335**

Ordinal Data

- Radiologist reading of an image
- Clinical symptom
- Psychology questionnaire

Score	x_1	$>$	x_2	$> \cdots >$	x_k	$>$	x_{k+1}	$> \cdots >$	x_K	
Partition	R_1		R_2	$\cdots$	R_k		R_{k+1}	$\cdots$	R_K	Total
Class-1	Z_1		Z_2	$\cdots$	Z_k		Z_{k+1}	$\cdots$	Z_K	n_1
Class-0	$\bar{Z}_1$		$\bar{Z}_2$	$\cdots$	$\bar{Z}_k$		$\bar{Z}_{k+1}$	$\cdots$	$\bar{Z}_K$	n_0
Total	S_1		S_2	$\cdots$	S_k		S_{k+1}	$\cdots$	S_K	n

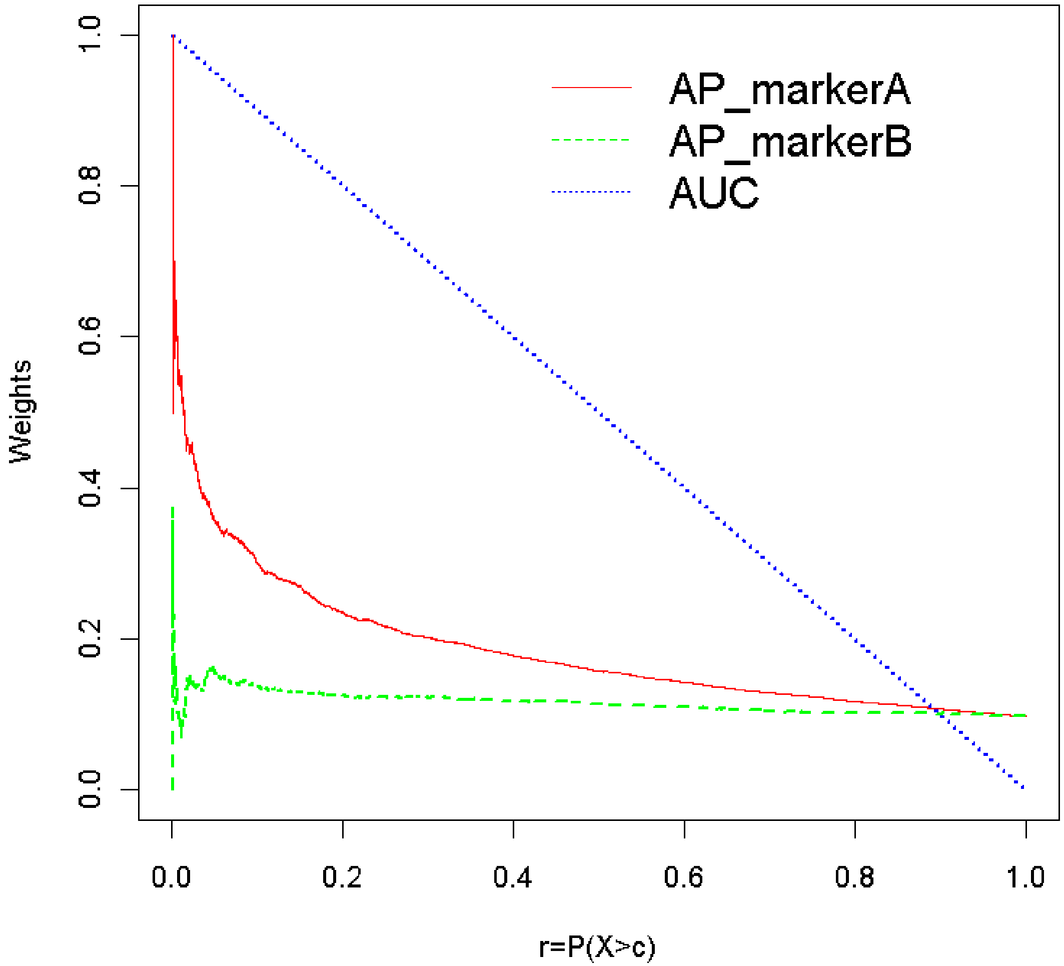
$$\widehat{AP} = \underset{w_1}{\left[\frac{Z_1}{S_1}\right]}\left[\frac{Z_1}{n_1}\right] + \underbrace{\left[\frac{Z_1+Z_2}{S_1+S_2}\right]}_{w_2}\left[\frac{Z_2}{n_1}\right] + \cdots + \underbrace{\left[\frac{Z_1+Z_2+\cdots+Z_K}{S_1+S_2+\cdots+S_K}\right]}_{w_K}\left[\frac{Z_K}{n_1}\right]$$

$$= \sum_{k=1} w_k \left[\frac{Z_k}{n_1}\right].$$

$$\widehat{AUC} = \frac{n}{n_0}\left\{\underbrace{\left[\frac{S_1+S_2+\ldots+S_K}{n}\right]}_{w_1'}\left[\frac{Z_1}{n_1}\right] + \underbrace{\left[\frac{S_2+\ldots+S_K}{n}\right]}_{w_2'}\left[\frac{Z_2}{n_1}\right] + \ldots + \underbrace{\left[\frac{S_K}{n}\right]}_{w_K'}\left[\frac{Z_K}{n_1}\right] - \frac{1}{2}\left(\frac{n_1}{n_0}\right)\right\} - \frac{1}{2}\left(\frac{n_1}{n_0}\right)$$

$$= \frac{n}{n_0}\sum_{k=1} w_k'\left[\frac{Z_k}{n_1}\right] - \frac{1}{2}\left(\frac{n_1}{n_0}\right)$$

A Simulated Example



Weights, w_k for AP and w'_k for AUC, in a simulated example.

$f_0(x)$ ~ $N(0, 1)$ and $f_1(x)$ ~ $N(\Delta, 1)$ where $\Delta_A = 1$ and $\Delta_B = 0.25$; $\pi = 0.1$.

MLE of AP

Score	x_1	$>$	x_2	$> \cdots >$	x_k	$>$	x_{k+1}	$> \cdots >$	x_K	
Partition	R_1		R_2	$\cdots$	R_k		R_{k+1}	$\cdots$	R_K	Total
Class-1	Z_1		Z_2	$\cdots$	Z_k		Z_{k+1}	$\cdots$	Z_K	n_1
Class-0	$\bar{Z}_1$		$\bar{Z}_2$	$\cdots$	$\bar{Z}_k$		$\bar{Z}_{k+1}$	$\cdots$	$\bar{Z}_K$	n_0
Total	S_1		S_2	$\cdots$	S_k		S_{k+1}	$\cdots$	S_K	n

Data in the 2 X K table follow

$$
\begin{aligned}
(Z_1, Z_2, ..., Z_K)|n_1 &\sim \text{multinomial}(n_1; p_1, p_2, ..., p_K), \\
(\bar{Z}_1, \bar{Z}_2, ..., \bar{Z}_K)|n_1 &\sim \text{multinomial}(n - n_1; q_1, q_2, ..., q_K), \\
n_1 &\sim \text{binomial}(n, \pi),
\end{aligned}
$$

where

$$
p_k = \int_{R_k} f_1(x)dx, \quad q_k = \int_{R_k} f_0(x)dx,
$$

Asymptotic Variance of AP

$$\widehat{\mathrm{AP}} = g(\widehat{p}_k, \widehat{q}_k, \widehat{\pi}) = \sum_{k=1}^{K} \left[\widehat{p}_k \left(\frac{\widehat{\pi} \sum_{k' \leq k} \widehat{p}_{k'}}{\widehat{\pi} \sum_{k' \leq k} \widehat{p}_{k'} + (1 - \widehat{\pi}) \sum_{k' \leq k} \widehat{q}_{k'}} \right) \right]$$

Apply the Delta method, we get

$$\widehat{var}\left(\widehat{AP}\right) \approx (\nabla g)^T \hat{J}^{-1} (\nabla g)$$

Example 1

Digital Mammography Imaging Screening Trial (Pisano et al. 2005 *New England Journal of Medicine*)

Malignancy score		7	6	5	4	3	2	1	Total
Digital	Category Total	11	29	69	1061	2224	6588	32588	42570
	Cancers	10	18	25	85	49	25	122	334
Film	Category Total	17	29	70	942	2291	6910	32486	42745
	Cancers	13	24	25	74	35	33	131	335

42,760 screening participants underwent two screening technology, 335 were diagnosed with breast cancer at 15 months follow-up.

Given that 335 breast cancer diagnosed in 42,760 screening participants at 15 months follow-up, the prevalence π is 0.00783.

	Seven-point Malignancy Scale	
	$\widehat{AUC}$ (s.e.)	$\widehat{AP}$ (s.e.)
Film mammography	0.735 (0.012)	0.166 (0.022)
Digital mammography	0.753 (0.012)	0.144 (0.021)

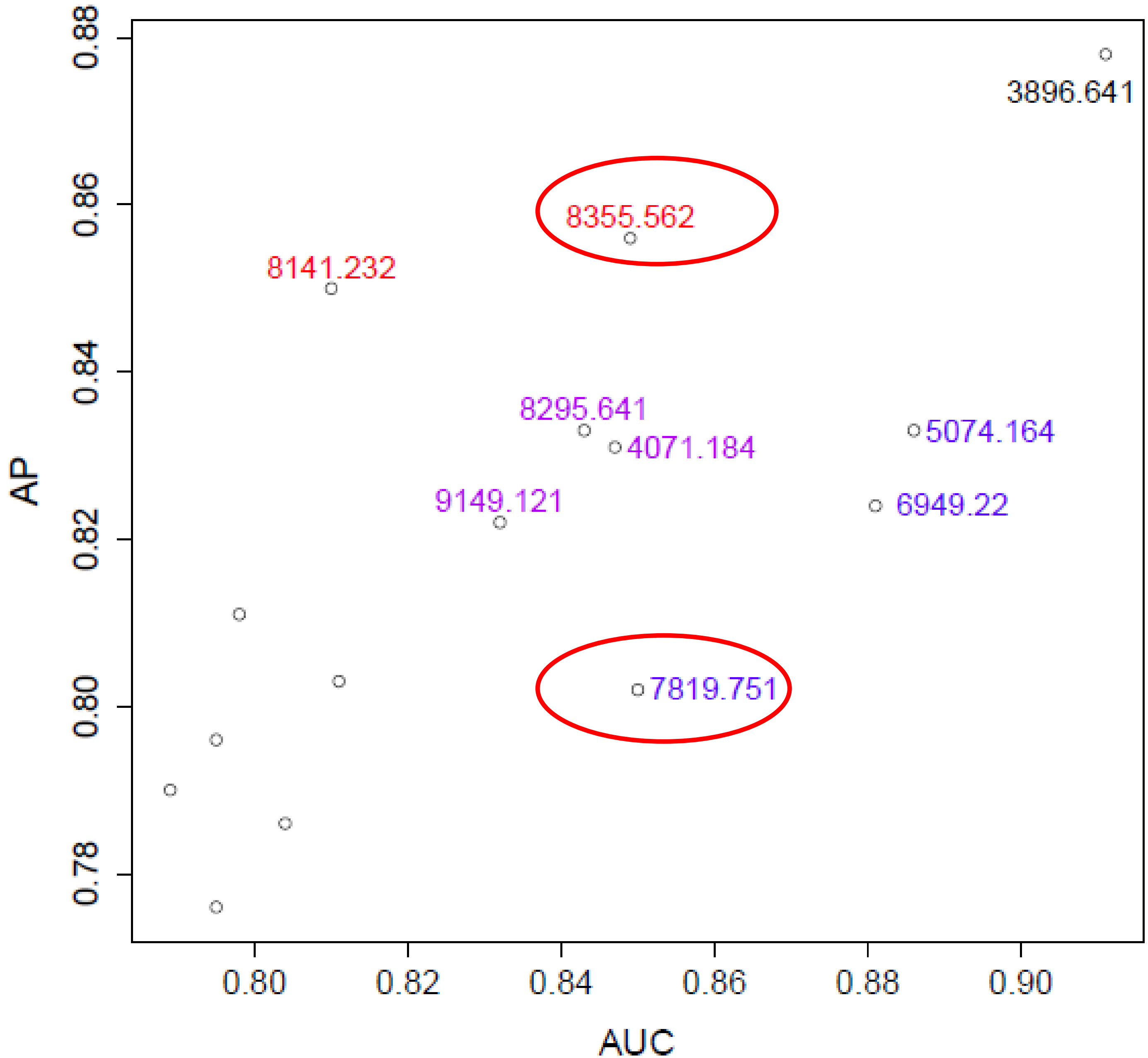
Remark: Resampling method can be used for the inference of the difference in AP when we have paired data.

Example 2

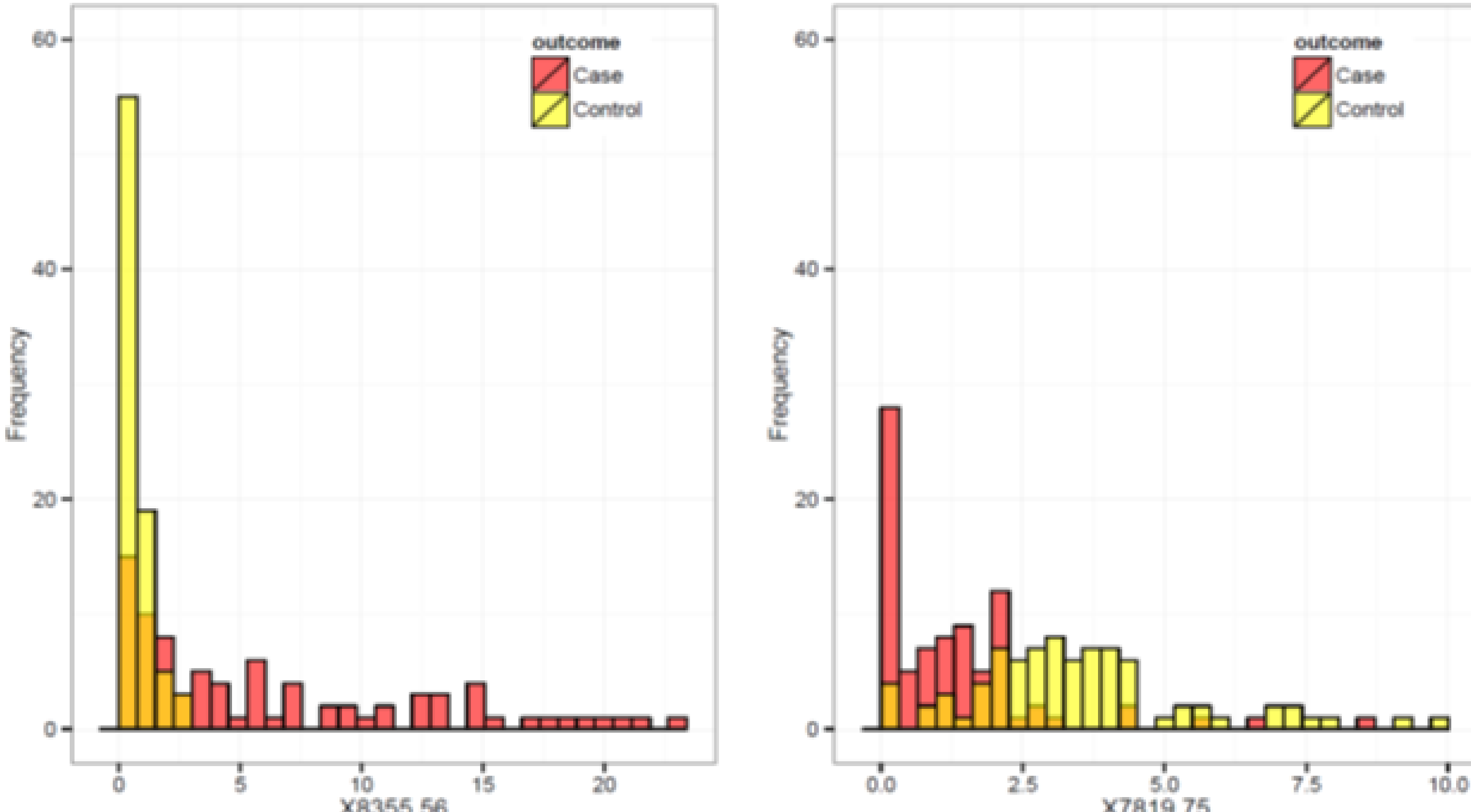
Mass spectrometry data for prostate cancer (Adam *et al*. 2002 Cancer Research)

- 779 potential biomarkers were assessed in 83 late-stage prostate cancer patients and 82 normal subjects.

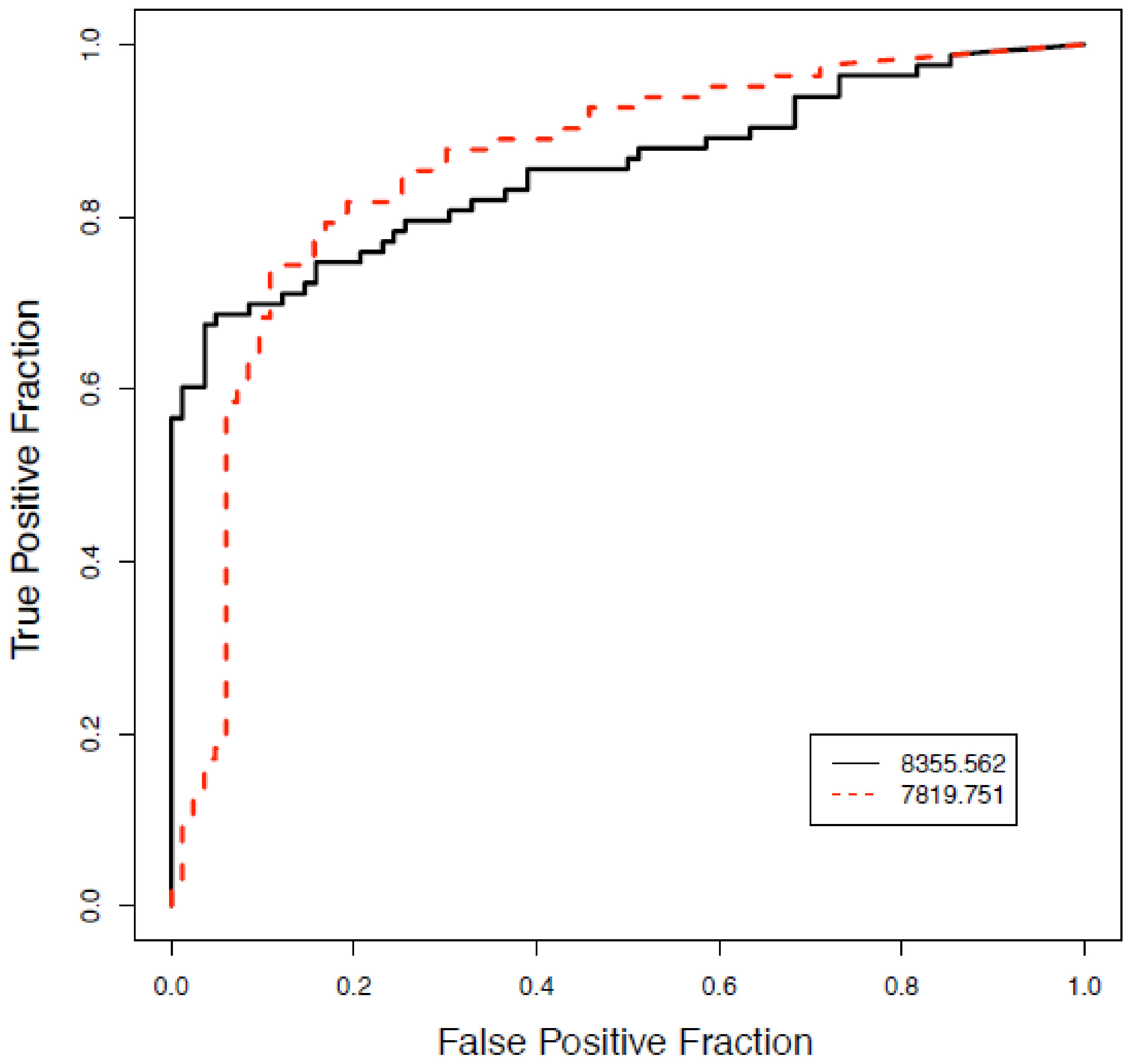
		Standard Error of AP		
Biomarker	AP	Asymptotic	P-Bootstrap	NP-Bootstrap
3896.641	0.878	0.0345	0.0344	0.0344
8355.562	0.856	0.0336	0.0339	0.0340
8141.232	0.850	0.0319	0.0324	0.0321
8295.641	0.833	0.0328	0.0327	0.0327
5074.164	0.833	0.0403	0.0405	0.0403
4071.184	0.831	0.0368	0.0364	0.0366
6949.220	0.824	0.0414	0.0415	0.0413
9149.121	0.822	0.0378	0.0380	0.0378

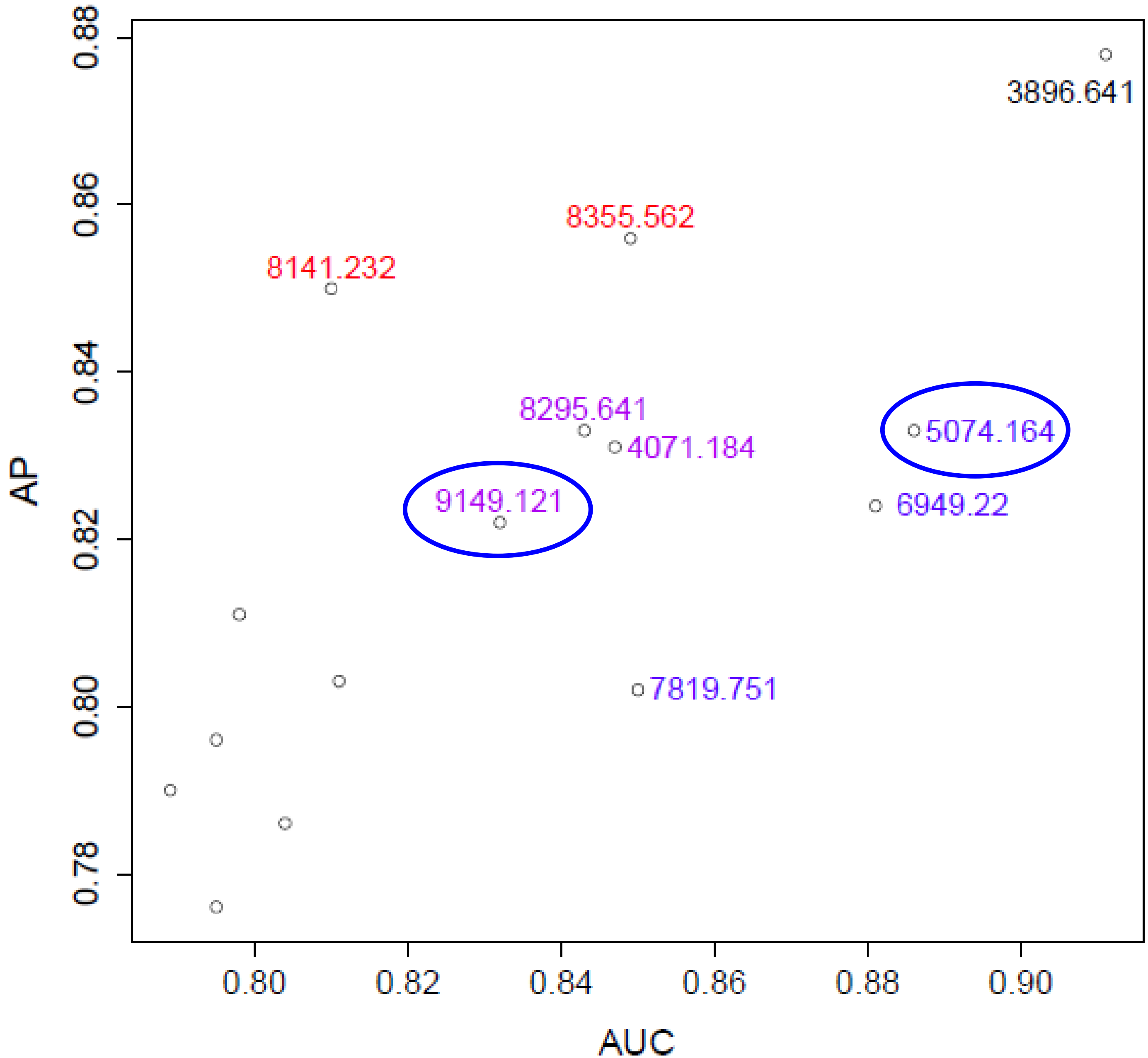


0.88
0.86
0.84
0.82
0.80
0.78
AP
0.80
0.82
0.84
0.86
0.88
0.90
AUC
3896.641
8355.562
8141.232
8295.641
4071.184
5074.164
9149.121
6949.22
7819.751

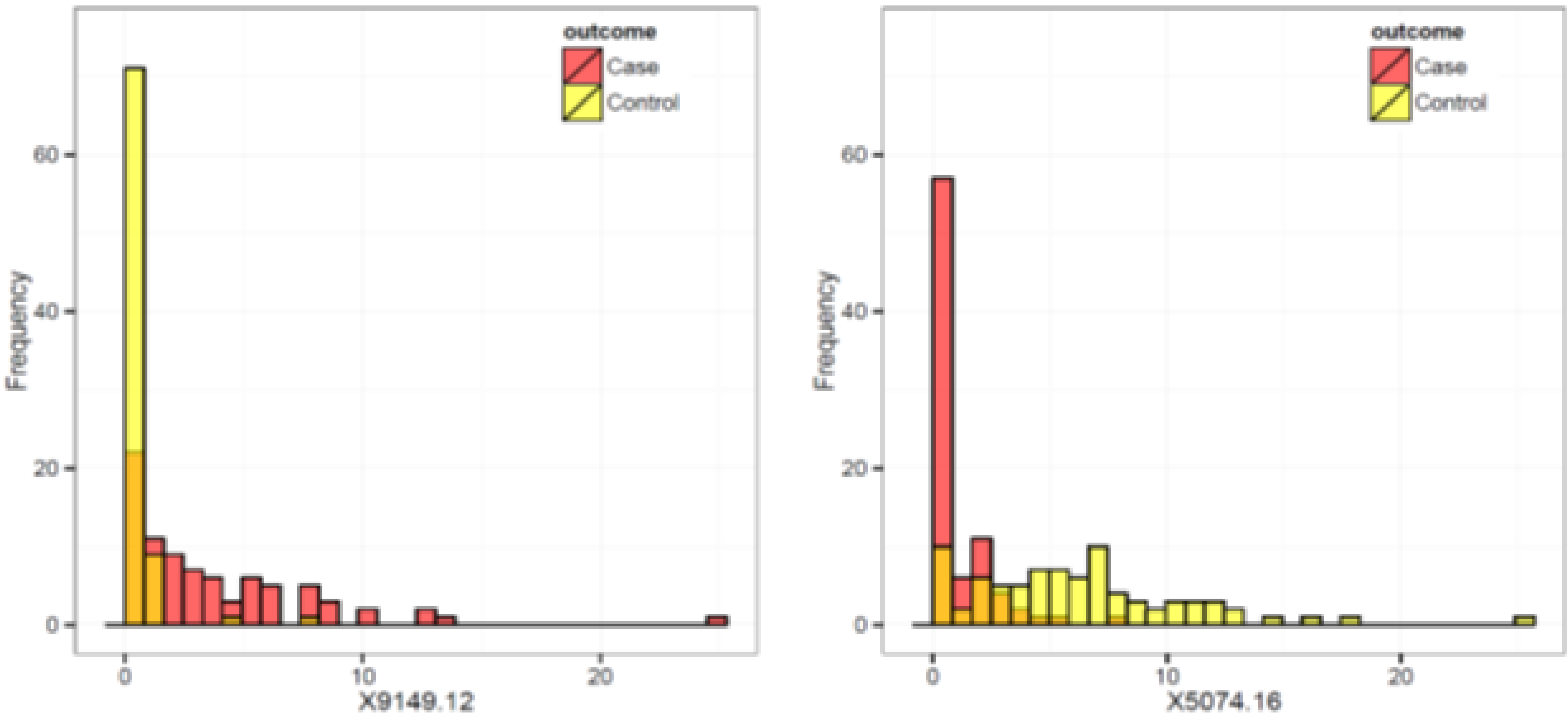


Pair A: similar AUC scores but different AP scores

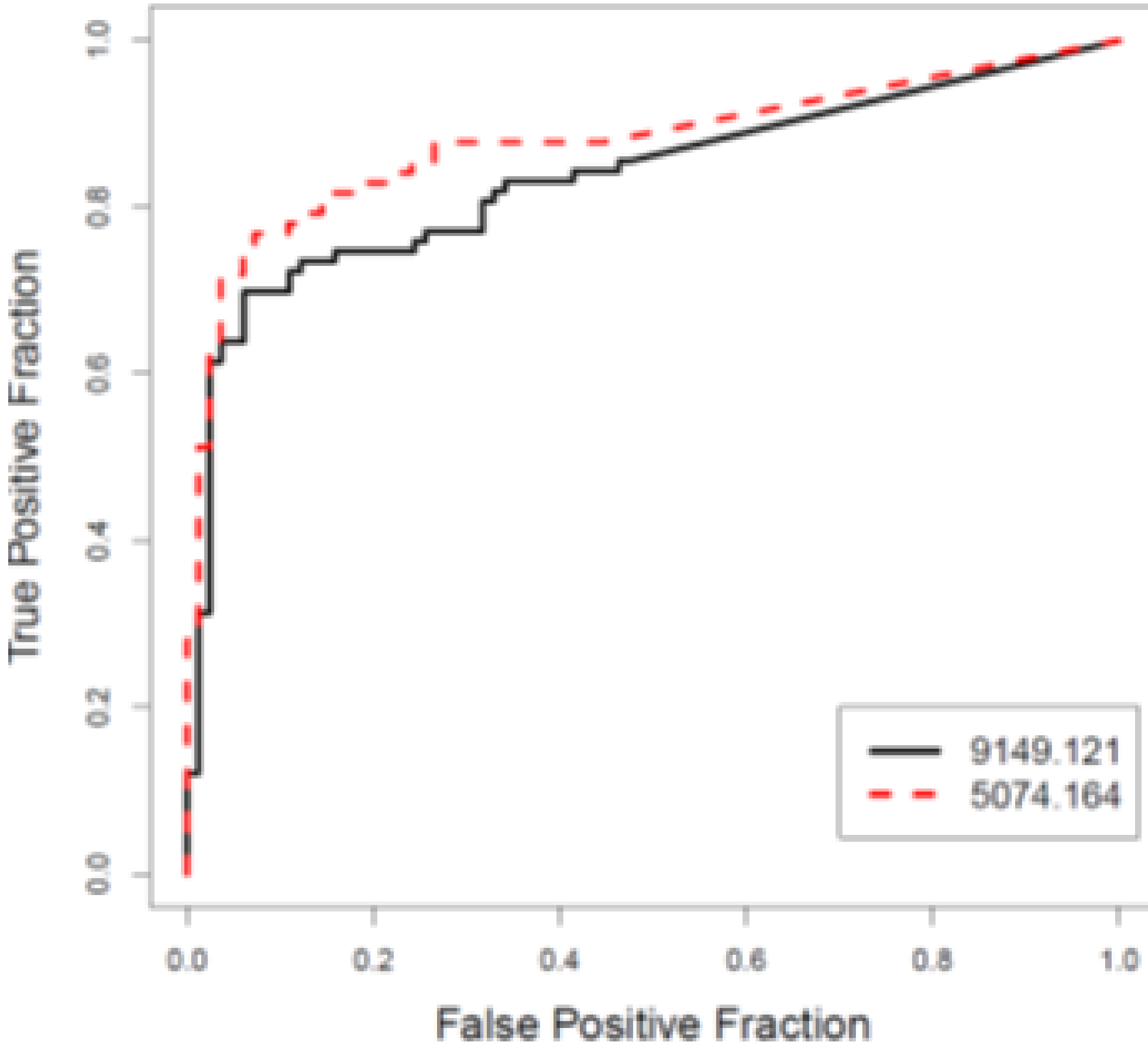

True Positive Fraction
False Positive Fraction
0.0
0.2
0.4
0.6
0.8
1.0
8355.562
7819.751



3896.641
8355.562
8141.232
8295.641
4071.184
5074.164
9149.121
6949.22
7819.751
AP
AUC
0.78
0.80
0.82
0.84
0.86
0.88
0.80
0.82
0.84
0.86
0.88
0.90



Pair B: similar AP scores but different AUC scores


True Positive Fraction
False Positive Fraction
0.0
0.2
0.4
0.6
0.8
1.0
9149.121
5074.164

A Thought Experiment

- The biomarker study is based on a case-control study with the goal to identify potential **screening** markers.
- How AUC, AP and the ranking of biomarkers is affected when the prevalence is much lower as in a screening setting?

Inflate the controls by replicating them

Biomarker	AUC	AP		
		n_0 X 1	n_0 X 10	n_0 X 100
8355.562	**0.849**	0.856	0.606	0.571
7819.751	**0.850**	0.802	0.370	0.062
5074.164	0.886	**0.833**	0.306	0.043
9149.121	0.832	**0.822**	0.512	0.225

Example 3

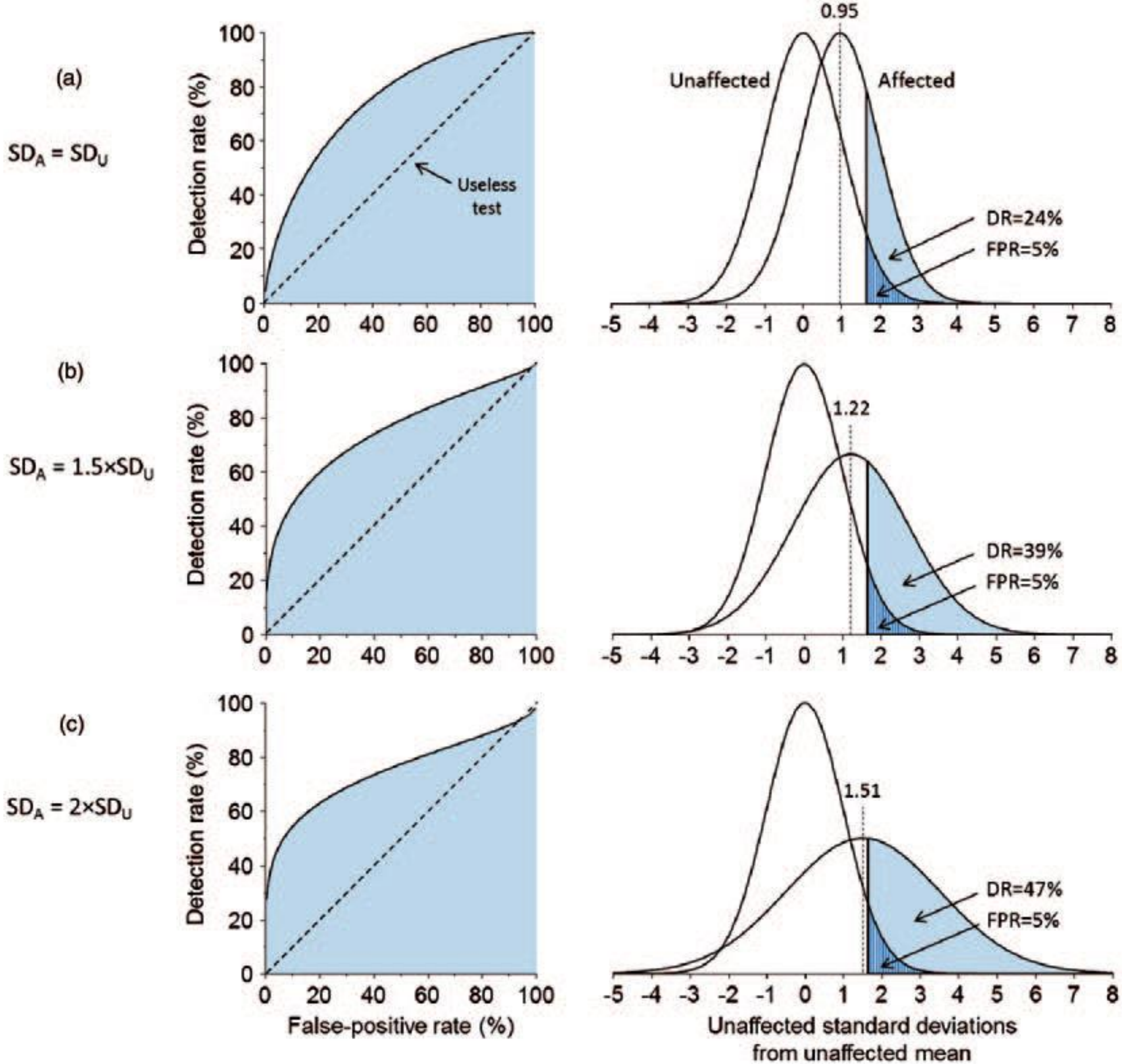

(a)
SD_A = SD_U
(b)
SD_A = 1.5×SD_U
(c)
SD_A = 2×SD_U
Detection rate (%)
False-positive rate (%)
Useless test
0.95
1.22
1.51
Unaffected
Affected
DR=24%
FPR=5%
DR=39%
FPR=5%
DR=47%
FPR=5%
Unaffected standard deviations from unaffected mean

Table 5 | AUC, AP, DR, and FPF for three tests from Wald and Bestwick [(10), Figure 2].

	AUC[a]	AP			DR at FPF 0.05[a]	FPF at DR 50%[a]
		$\pi = 0.5$	$\pi \approx 0.09$	$\pi \approx 0.01$		
$SD_A = SD_U$	0.75	0.74	0.26	0.04	0.24	0.17
$SD_A = 1.5SD_U$	0.75	0.79	0.42	0.16	0.39	0.11
$SD_A = 2SD_U$	0.75	0.81	0.51	0.29	0.47	0.07

Continuous Version

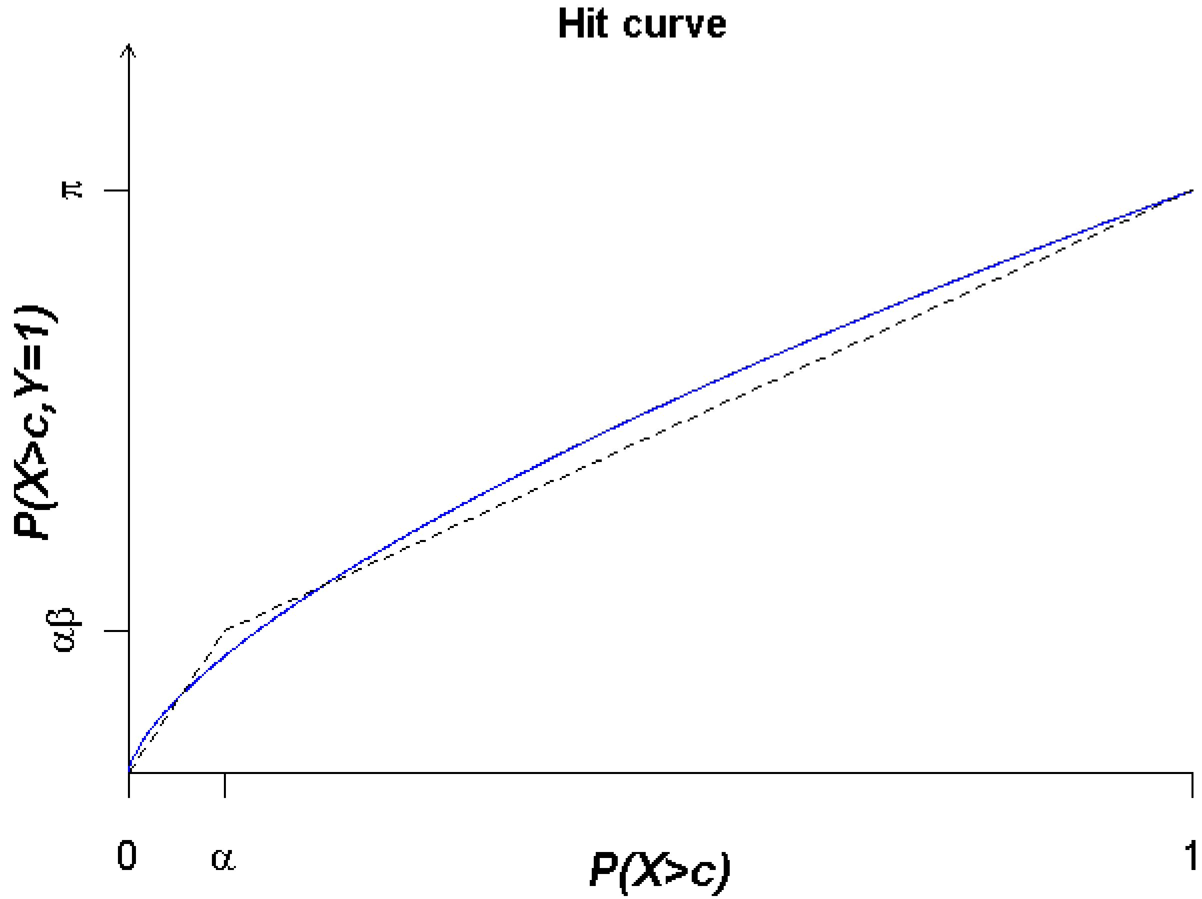
$$s = P(X > x)$$

$$h(s) = P(X > x, Y = 1)$$

$$\mathrm{AUC} \equiv \int_0^1 \mathrm{TPF}(s)\,\mathrm{dFPF}(s)$$

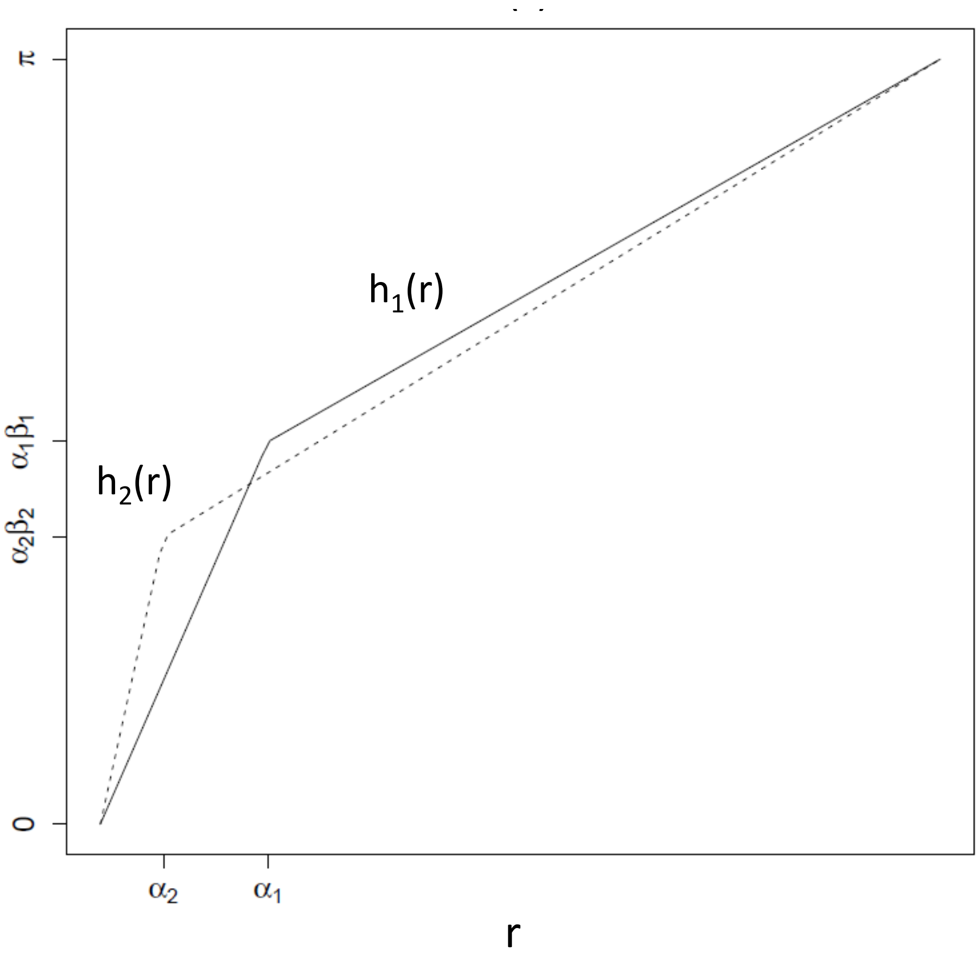
$$= \frac{1}{\pi(1-\pi)}\left[\int_0^1 h(s)ds - \frac{\pi^2}{2}\right]$$

$$\mathrm{AP} \equiv \int_0^1 \mathrm{PPV}(s)\,\mathrm{dTPF}(s) = \frac{1}{\pi}\int_0^1 \frac{h(s)}{s}\mathrm{dh}(s).$$



Approximate the hit curve by a piecewise linear curve, let β be the initial true positive rate of the underlying test

$$h(r) = \begin{cases} \beta r, & r \in [0, \alpha] \\ \frac{\pi-\alpha\beta}{1-\alpha}(r-\alpha) + \alpha\beta, & r \in (\alpha, 1] \end{cases}$$



Theorem 1: If two hit curves, $h_1(r)$ and $h_2(r)$, both belong to the piecewise linear family, and are parameterized respectively by (α_1, β_1) and (α_2, β_2), then AUC(h_1) = AUC(h_2) if and only if

$$(\beta_1 - \pi)\,\alpha_1 = (\beta_2 - \pi)\,\alpha_2$$

Theorem 2: If a hit curve, h(r), belongs to the piecewise linear family, then

$$\widetilde{AP}(h) \approx \beta \times \widetilde{AUC}(h)$$

where AP and AUC are re-scaled to lie between 0 and 1 for any hit curve *h*

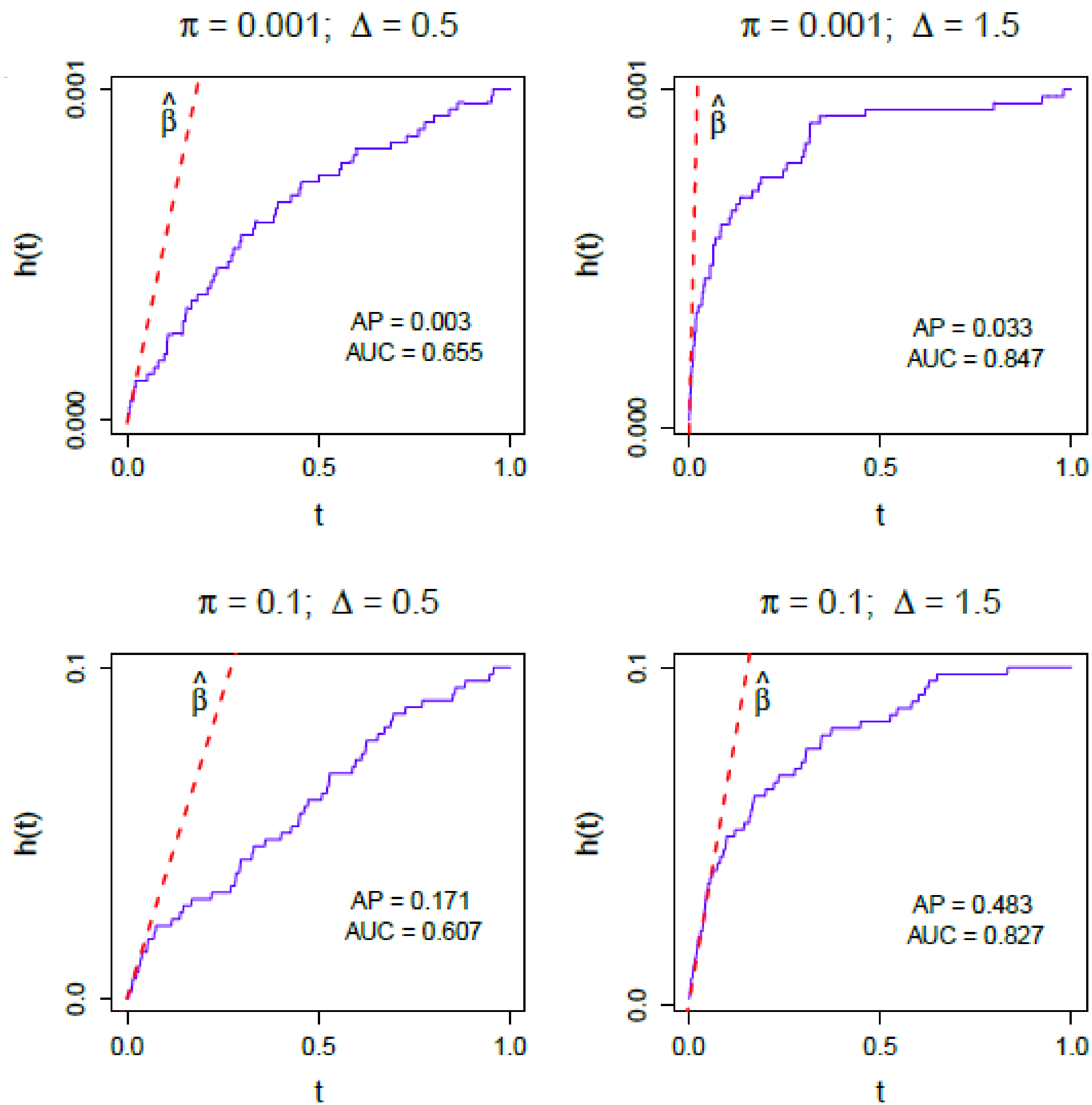
$$\widetilde{\mathrm{AP}} \equiv \frac{\mathrm{AP} - \pi}{1 - \pi}$$

$$\widetilde{\mathrm{AUC}} \equiv \frac{\mathrm{AUC} - 1/2}{1 - 1/2} = 2\mathrm{AUC} - 1.$$

Simulation Study

- Non-diseased subjects (Y=0), $f_0(x) \sim N(0, 1)$
- Diseased subjects (Y=1), $f_1(x) \sim N(\Delta, 1)$

- Simulation settings:
  - $\Delta = 0.5$ or 1.5
  - $\pi = 0.001$ and $n = 50{,}000$ or $\pi = 0.1$ and $n = 500$



$$\widehat{\beta} = \frac{\widetilde{\mathrm{AP}}(h)}{\widetilde{\mathrm{AUC}}(h)} = \frac{(\mathrm{AP}(h) - \pi)/(1-\pi)}{2\mathrm{AUC}(h) - 1}$$

Summary

- AP is a single numerical measure measuring prediction performance
- Connection between AP and AUC
- Estimation of AP and its asymptotic variance
- Practical relevance

Future work

- Assessing survival/risk prediction models with AP(t)