Measuring the Model Prediction Performance for Rare Events

Yan Yuan

School of Public Health

April 10, 2015

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

Outline

- Motivation
  - Predicting/Detecting the Rare Events (low prevalence/incidence)
- Metrics for evaluating model performance
  - Area under the ROC curve (AUC)
  - Average Positive Predictive Value (AP)
- Examples
- Summary and future work

Motivating Data

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 M	Category Total	11	29	69	1061	2224	6588	32588	**42570**
	Cancers	10	18	25	85	49	25	122	**334**
Film M	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.

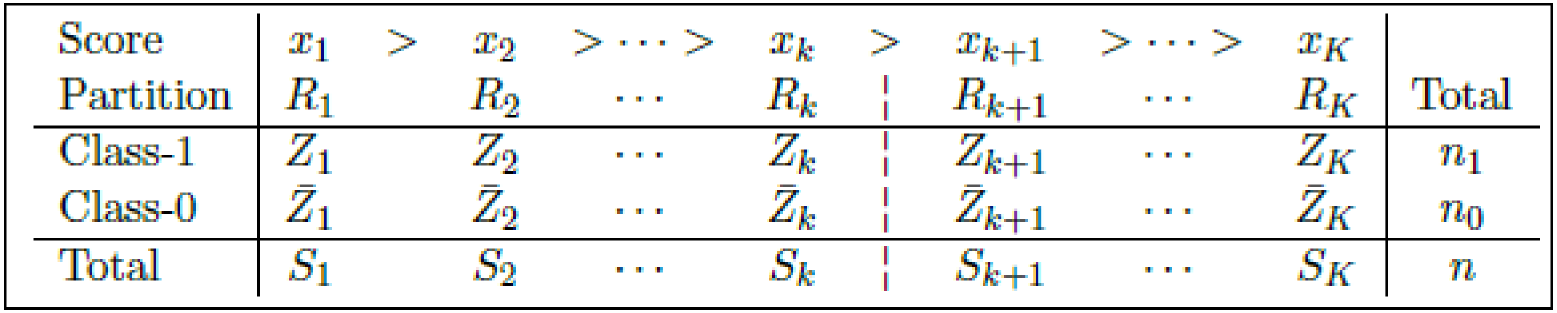
Predicting the Rare Events

- Cancer screening: detect from the asymptomatic population the diseased subjects, who make up a very small proportion (typically < 1%).
- Risk models
- Drug discovery: identify potential chemical compounds that are biologically active for some target (typically < 5%).
- Information retrieval
- Prediction of Rare events in your subject area?

Evaluating Model Performance for Predicting Rare Events

- Threshold Dependent Measure
  - Misclassification rate
  - Sensitivity and Specificity
  - Positive and Negative Predictive Value
- Threshold Independent Measure (Pre-clinical or pre-application stage)
  - Area Under the ROC* Curve (AUC)
  - Average Positive Predictive Value (AP)

*Receiver Operating Characteristic



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

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

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

$$\widehat{AUC} = \frac{n}{n_0}\left\{\underbrace{\left[\frac{S_1+S_2+\ldots+S_K}{n}\right]}_{w'_1}\left[\frac{Z_1}{n_1}\right] + \underbrace{\left[\frac{S_2+\ldots+S_K}{n}\right]}_{w'_2}\left[\frac{Z_2}{n_1}\right] + \ldots + \underbrace{\left[\frac{S_K}{n}\right]}_{w'_K}\left[\frac{Z_K}{n_1}\right] - \frac{1}{2}\left(\frac{n_1}{n_0}\right)\right\} - \frac{1}{2}\left(\frac{n_1}{n_0}\right)$$

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

Example 1: Two technology for Breast cancer screening

Malignancy score		7	6	5	4	3	2	1	Total
Digital M	Category Total	11	29	69	1061	2224	6588	32588	42570
	Cancers	10	18	25	85	49	25	122	334
Film M	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.

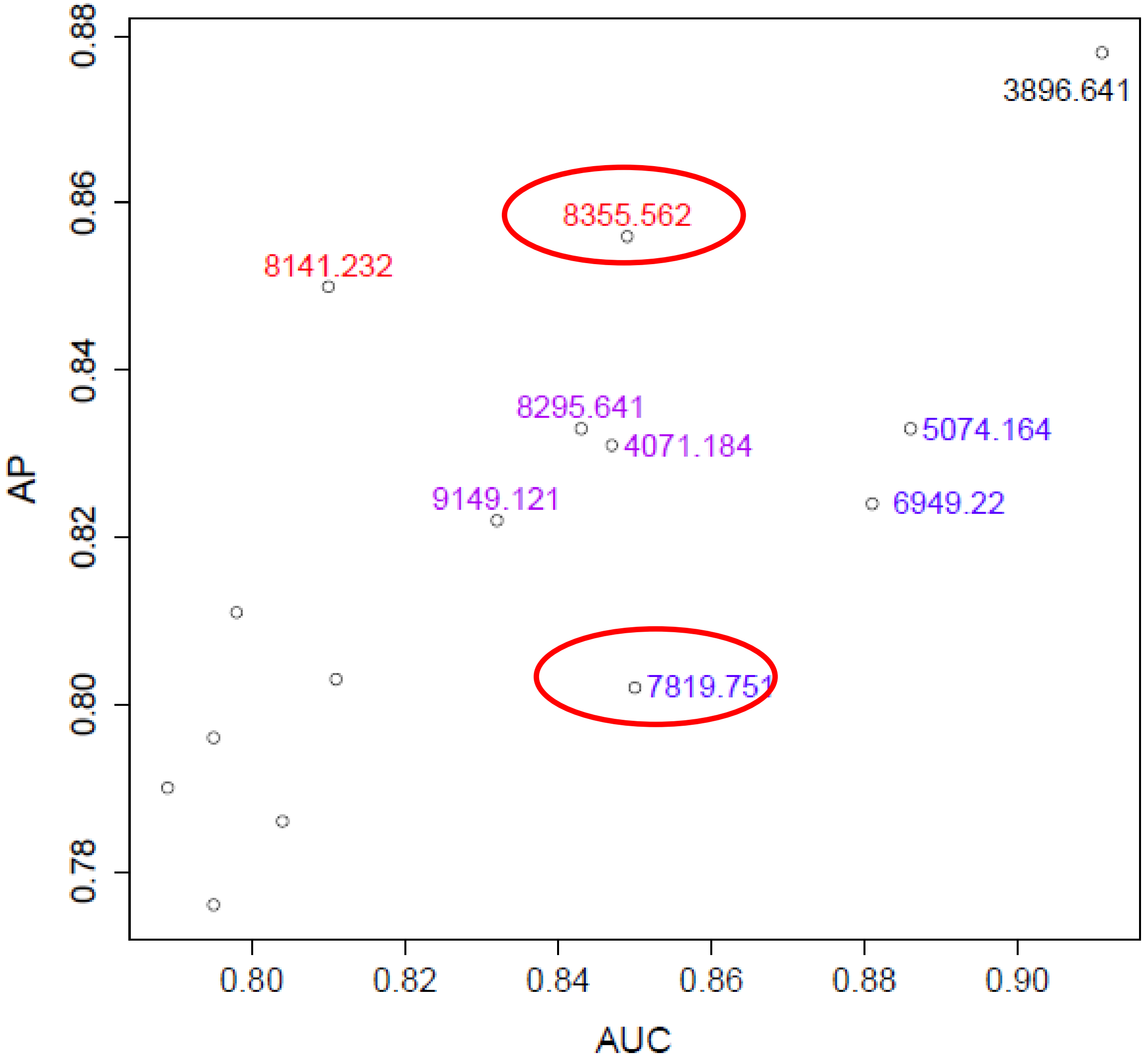
Given that 335 breast cancer diagnosed in 42,760 screening participants at 15 months follow-up, the prevalence π is 0.783%.

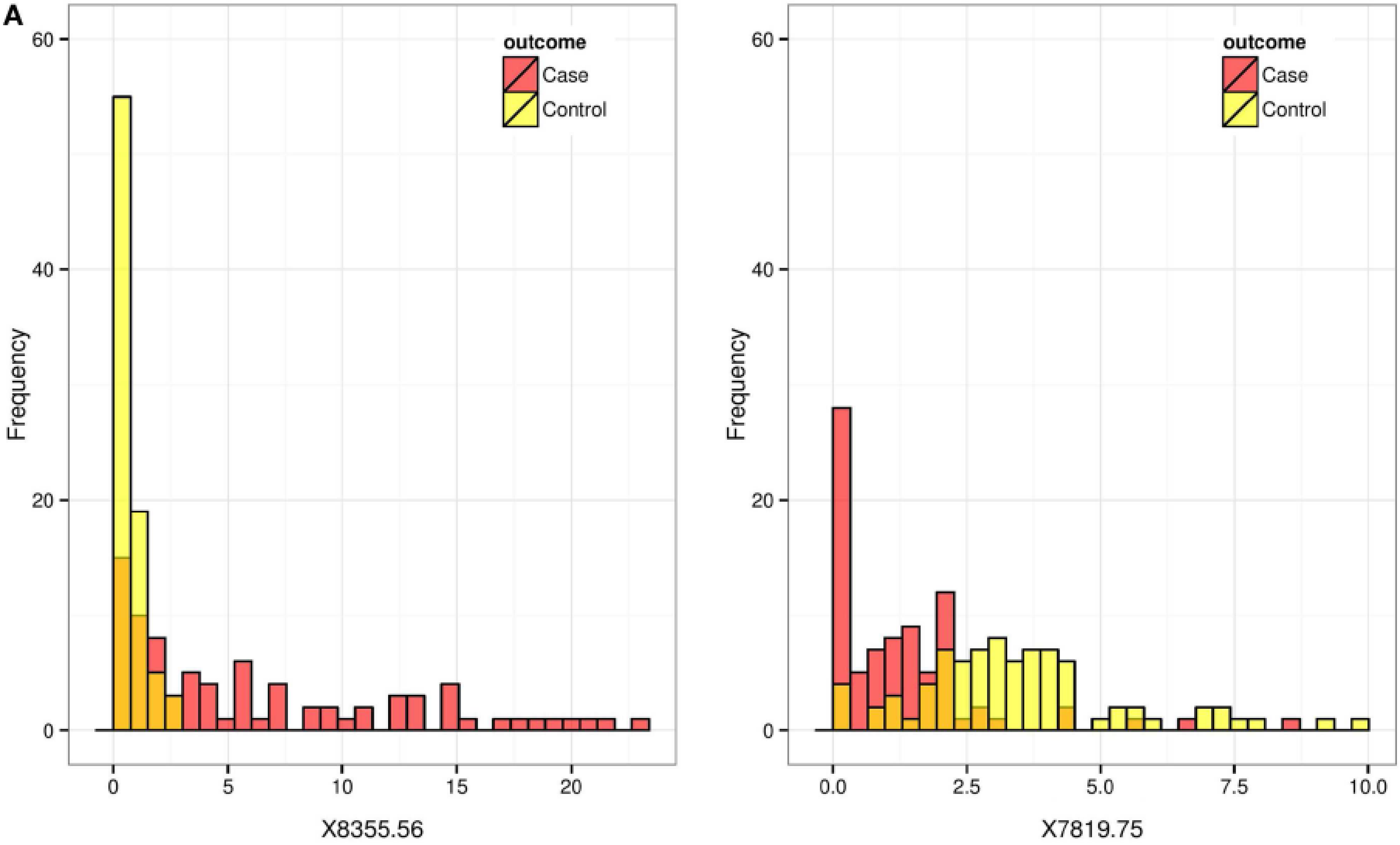
	Seven-point Malignancy Scale	
	$\widehat{AUC}$ (s.e.)	$\widehat{AP}$ (s.e.)
Film mammography	0.735 (0.012)	0.166 (0.022)
Digital mammography	0.753 (0.012)	0.144 (0.021)

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

Example 2: Biomarkers for prostate cancer

779 potential biomarkers were assessed in 83 late-stage prostate cancer patients and 82 normal subjects. (Adam *et al*. 2002 Cancer Research)




A
outcome
Case
Control
60
40
20
0
Frequency
0
5
10
15
20
X8355.56
outcome
Case
Control
60
40
20
0
Frequency
0.0
2.5
5.0
7.5
10.0
X7819.75

A Thought Experiment

- The biomarker study is based on a case-control study (# disease ≈ # non-disease); its goal is to identify potential screening markers.
- How 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 $\pi = 0.5$	n_0 X 1 $\pi = 0.5$	n_0 X 10 $\pi = 0.1$	n_0 X 100 $\pi = 0.01$
8355.562	**0.849**	0.856	0.606	0.571
7819.751	**0.850**	0.802	0.370	0.062

Summary and future work

- AP is a single numerical measure, similar to AUC
  - Connection between AP and AUC
  - Empirical estimation of AP and its asymptotic variance
- Assessing risk prediction and survival models