Measuring The Performance Of Medical Screening Tests: Average Precision Vs. Area Under ROC Curve

Wanhua Su†[a], Yan Yuan†[b] and Mu Zhu[c]

†These authors contributed equally to this project [a]Department of Mathematics and Statistics, Grant MacEwan University, Edmonton, Canada, [b]School of Public Health, University of Alberta, Edmonton, Canada, [c]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo , Canada

Motivation

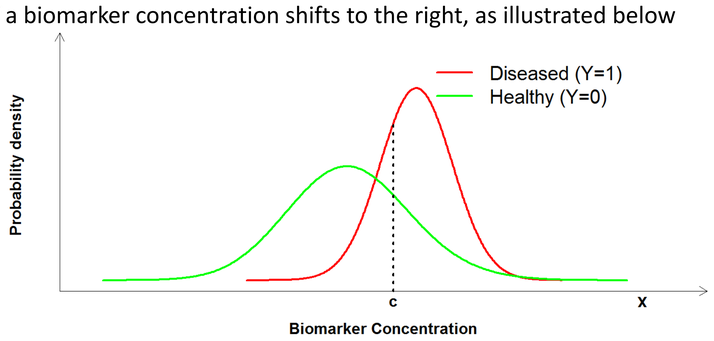
Early detection of disease in its asymptomatic stage could greatly enhance the likelihood of cure. For example, routine breast, prostate and colon cancer screening for certain age groups are recommended by medical experts in North America. The objective of medical screening tests is to detect from the underlying population the diseased asymptomatic subjects who make up a very small proportion (typically < 0.5%). Because the false positive results from screening is considered one major harm of this type of tests[1], the ability to identify diseased subjects with minimal false positive findings should be an important consideration when performance of a screening test is evaluated.

In practice, the **R**eceiver **O**perating **C**haracteristic (ROC) curve has been the primary performance measure for medical tests including the screening tests. The **A**rea **U**nder the ROC **C**urve (AUC) provides a single numerical summary of the ROC curve, thus is used as the basis of inferential statistics for comparing ROC curves. We introduce here a performance measure called **A**verage **P**recision (AP) as it emphasizes more on early detection. AP is a summary measure of the so-called hit curve, which is analogous to ROC curve.

The ROC and Hit curve

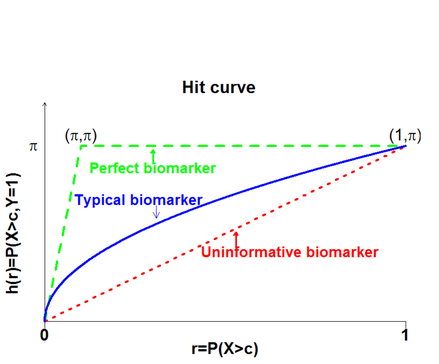
Through an example, let's look at the hit curve and the ROC curve, and the definition of AUC and AP.

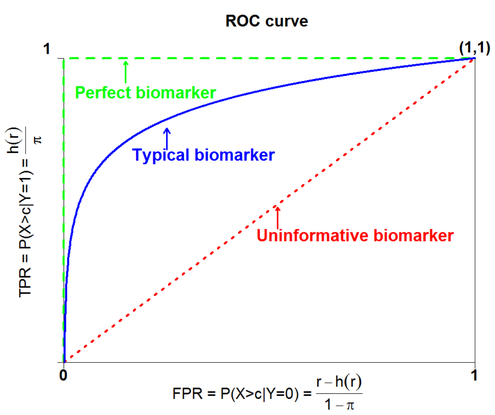
Suppose that in asymptomatic diseased (*Y*=1) subjects, the distribution of a biomarker concentration shifts to the right, as illustrated below



Using a threshold c, the hit curve plots P(*X*>c, *Y*=1) vs. P(*X*>c), and the ROC plots true positive fraction P(*X*>c|*Y=1*) *vs.* false positive fraction P(*X*>c|*Y=0*) for the entire set of thresholds.

Notations: r = P(X>c), the hit function h(r) =P(X>c, Y=1), and the proportion of diseased subjects π=P(Y=1)





$$AP \stackrel{\text{def}}{=} \int_0^1 \frac{h(r)}{r}\, d\frac{h(r)}{\pi} = \frac{1}{\pi}\int_0^1 \frac{h(r)}{r}\, dh(r)$$

$$AUC \stackrel{\text{def}}{=} \int_0^1 \frac{h(r)}{\pi}\, d\frac{r-h(r)}{1-\pi} = \frac{1}{\pi(1-\pi)}\left[\int_0^1 h(r)dr - \frac{\pi^2}{2}\right]$$

AP is the precision (i.e. P(*Y*=1 | *X*>c)) averaged over the entire hit curve, which takes value between [π, 1].
AUC is the area under the ROC curve, which takes value between [0.5, 1].

Relationship between AP and AUC

Continuous framework:

If we approximate the hit curve by a quasi-concave curve and let β be the **initial true positive rate** of the underlying test, we can show that

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

where $\widetilde{AP}(h)$ and $\widetilde{AUC}(h)$ are re-scaled to lie between 0 and 1 for any hit (or ROC) curve *h*.

This approximation implies that if two medical screening tests have the same AUC, then the AP will "reward extra points" to the one with the larger β, i.e. AP places more emphasis on the initial part of the hit (ROC) curve.

Discrete framework:

A radiology screening test typically generates ordinal test results, which use categories such as highly suspicious, possibly malignant, possibly benign, etc. We consider the setup with a multinomial distribution as described in Table 1.

Table 1: A screening/diagnostic test partitions n subjects into K groups (K distinct scores). The broken bars (¦) indicates the case where all those with scores ≥ x_k are declared to be test positive (class 1), while all those with scores < x_k are declared to be test negative (class-0).

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

R_k= the k^{th} region in the space of test scores; S_k = total number of subjects in R_k; Z_k = total number of class-1 subjects in R_k; $\bar{Z}_k$ = total number of class-0 subjects in R_k.

We can show that

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

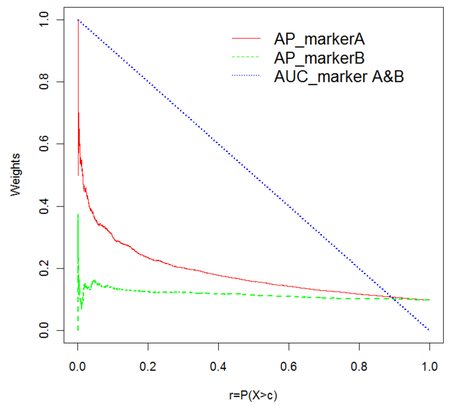
$$= \sum_{k=1}^{K} w_k \left[\frac{Z_k}{n_1}\right]$$

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

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

These weights, w_k and w'_k, again show that AP places an emphasis on the ability of a test to give the diseased subjects a high score.

Figure 1: Weights for AP and AUC in a simulated example. The concentration of markers A and B in healthy subjects ~ N(0,1), the concentration of markers A and B in diseased subjects ~ N(1, 1) and N(0.25, 1), respectively. The proportion of diseased (π) subjects is 0.1.



By the simulation design, marker A is better than marker B for detecting the diseased subjects. Figure 1 shows that unlike AUC, the weights of AP favors marker A. Thus, AP favors marker A more than AUC does.

Asymptotic Variance of AP

In order to use AP as an evaluation metric, we need to know its variance. The data in each row of Table 1 follows a multinomial distribution, and the two multinomial distributions are independent given π, the proportion of diseased subject. Therefore, the log-likelihood function is

$$\ell(p_k, q_k, \pi) = \sum_{k=1}^{K} z_k \log p_k + \sum_{k=1}^{K} \bar{z}_k \log q_k + [n_1 \log \pi + (n-n_1)\log(1-\pi)],$$

subject to the constraints

$$\sum_{k=1}^{K} p_k = 1 \text{ and } \sum_{k=1}^{K} q_k = 1,$$

AP is a function of these parameters, i.e. AP = f(p_k, q_k, π). We obtain the asymptotic variance of AP by using the delta method,

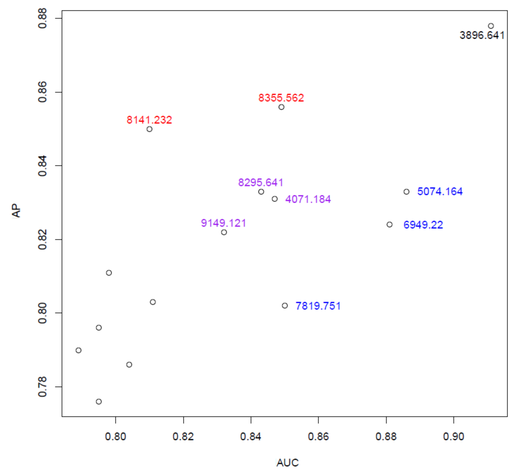
$$\text{var(AP)} = (\nabla f)^T \mathrm{J}^{-1} \nabla f,$$

where J is the expected information matrix.

Examples

Continuous score for potential biomarkers: a case-control study (π=0.5)

Figure 2: AP vs. AUC of the top 15 potential biomarkers for prostate cancer.[2]



We see that AP and AUC rank the biomarkers differently.

Ordinal score for a screening test (π=0.00784)

Table 2: The film vs. digital mammography for breast cancer screening[3]

Breast cancer diagnosis (455 day follow-up)	**Malignancy scores**	
	$\widehat{AUC}$ (s.e.)	$\widehat{AP}$ (s.e.)
Film mammography	0.735 (0.012)	0.166 (0.017)
Digital mammography	0.753 (0.012)	0.144 (0.018)

Discussion:

- Among tests that have similar AUCs, the test that finds the largest proportion of diseased subjects with minimal false positives in the early part of the hit (ROC) curve will have the highest AP. It is consistent with the goal of screening. Therefore, AP is a more relevant performance measure for a screening test than the AUC.
- AUC is $Pr(X_D > X_{\bar{D}} \mid$ a randomly selected pair of diseased and healthy subjects), where X_D and $X_{\bar{D}}$ are the test scores for the diseased and healthy subjects in the randomly selected pair, respectively. It is a conditional probability, and ignores an important parameter π, the proportion of diseased subjects in the asymptomatic population.
- The relative small numerical value of AP for a screening test (Example 2) is advantageous. Large values of AUC give clinicians and patients a false sense of accuracy of the test results, which aggravates the harm of a screening test.[4] Thus, the relative small valued single summary measure AP offers a useful alternative to summarize the test performance.

References:

1. Fletcher SW (2011) Breast Cancer Screening: A 35-year perspective. *Epidemiologic Reviews*. 33: 165-175.
2. Wang Z. and Change Y. (2011) Marker selection via maximizing the partial area under the ROC curve of linear risk scores. *Biostatistics*. 12(2): 369-385.
3. Pisano et al. (2005) Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening. *The New England Journal of Medicine*. 353(17):773-1783.
4. Gigerenzer, et al. (2008) Helping Doctors and Patients Make Sense of Health Statistics. *Psychological Science in the Public Interest.* 8(2):53-90.

Acknowledgement: The authors would like to thank Maoji Li for her technical assistance in making the graphs, checking the R code for the results presented.