Statistical Considerations for Evaluating Prognostic Biomarkers: Choosing Optimal Threshold

Zheng Zhang

Abstract The use of biomarker is increasingly popular in cancer research and various imaging biomarkers have been developed recently as prognostic markers. In practice, a threshold or cutpoint is required for dichotomizing continuous markers to distinguish patients with certain conditions or responses from those who are without. Two popular ROC based methods to establish "optimal" threshold are based on Youdan index J and closest top-left criterion. We have shown in this paper the importance to acknowledge the inherent variance of such estimates. In addition, a purely data-driven approach to search for optimal threshold can produce estimates that are not necessarily meaningful due to the large variance in such estimates. Instead, we propose to estimate the threshold through pre-specified criterion, such as a fixed level of specificity. The confidence intervals of the threshold and sensitivity at the pre-specified specificity are much narrower compared to the quantities measured through either Youdan index J or closest top left criterion. We suggest to estimate the threshold at a pre-specified level of specificity, and the sensitivity at that threshold, all the estimates should be accompanied by appropriate 95 % confidence intervals.

Keywords Biomarker • ROC • Threshold • Optimal • Youdan index

1 Introduction

From various clinical studies conducted during the past decade, a large collection of biomarkers have been studied on their abilities to predict important clinical outcomes such as treatment response, progression-free survival and overall survival in patients who were diagnosed with cancer and under treatment. One group of such markers have been derived from advanced imaging procedures, such as rCBV from dynamic susceptibility contrast-enhanced (DSC) MR perfusion (Paulson and Schmainda 2008), K^{trans} from dynamic contrast-enhanced (DCE) MR perfusion (Sourbron and Buckley 2013) and ADC values from diffusion-weighted

Z. Zhang (✉)
Department of Biostatistics, Center for Statistical Sciences, Brown University School of Public Health, Providence, RI 02912, USA

J. Lin et al. (eds.), *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, ICSA Book Series in Statistics,
DOI 10.1007/978-3-319-42568-9_2

imaging (DWI) (Bihan et al. 2006). Those markers are usually measured at several time-points throughout the study, such as pre-treatment, mid-treatment and post-treatment. The most frequently used marker values are those either measured at pre-treatment, or changes in marker values from pre-treatment measurements to various after treatment measurements. Due to the continuous nature of those values, the clinical usefulness of such markers often depends on whether a threshold can be determined to classify the marker. For example, a marker value above that threshold would predict a favorable outcome (better response to treatment, longer survival, etc.) and a marker value below that threshold would predict an unfavorable outcome. For all the possible thresholds that can be found, we would want to determine whether there is an optimal threshold that offers the best predictive performance.

2 A Brief Review of the ROC Curve

The receiver operating characteristic (ROC) curve (Swets and Pickett 1982; Pepe 2003) is a popular statistical tool to define predictive accuracy, hence it provides a pathway to determine the optimal threshold. The ROC curve is a collection of pairs of sensitivities and specificities, each pair is determined by a unique threshold. Assuming a test is done to diagnose a disease, the ROC curve is a plot of sensitivity versus 1-specificity, where sensitivity is the probability of the test value to correctly identify disease and specificity is the probability of it to correctly identify non-disease cases. The ROC curve can be written as a function of $t \in (0, 1)$, by letting $\bar{D}$ and D denote non-diseased and diseased populations and $S_{\bar{D}}$ and S_D be the survivor functions for test result Y from $\bar{D}$ and D, respectively, such as $S_D(c) = P[Y \geq c|D]$, $S_{\bar{D}}(c) = P[Y \geq c|\bar{D}]$, then the ROC curve is defined as $ROC(t) = S_D(S_{\bar{D}}^{-1}(t)), \quad t \in (0, 1)$.

To estimate the ROC curve empirically from test results $Y = \{Y_{D,i}, Y_{\bar{D},j}\}$, $i = 1, \ldots, n_D, j = 1, \ldots, n_{\bar{D}}$, $N = n_D + n_{\bar{D}}$, define $\widehat{sen}(c) = \sum_{i=1}^{n_D} I[Y_{D_i} \geq c]/n_D$ and $\widehat{1 - spec}(c) = \sum_{j=1}^{n_{\bar{D}}} I[Y_{\bar{D}_j} \geq c]/n_{\bar{D}}$, then the empirical ROC curve is a plot of $\widehat{sen}(c)$ versus $\widehat{1 - spec}(c)$ for all possible cut points c on the real line.

The area under the ROC curve (AUC) is commonly used to determine the discrimination power of the test. It is defined as

$$AUC = P(Y_D > Y_{\bar{D}}) \tag{1}$$

The empirical AUC is estimated as a Mann-Whitney U-Statistics

$$\widehat{AUC} = \sum_{i=1}^{N} \sum_{j=1}^{N} \{I[Y_{D,i} > Y_{\bar{D},j}] + \frac{1}{2} I[Y_{D,i} = Y_{\bar{D},j}]\}/N^2 \tag{2}$$

3 Criteria Based on the ROC Curve

The criteria based on the ROC curve seek to maximize sensitivity and specificity simultaneously. Two such criteria are frequently used: The first one is called Youdan index J (Youdan 1950), which is the threshold corresponding to the point on the ROC curve that has the longest distance to the identity (diagonal) line. Hence this threshold is chosen to maximize the sum of sensitivity and specificity. Intuitively, this point is the point on the ROC curve that is the furthest away from the curve corresponds to a "useless" test. First define the distance from a point on the ROC curve to the diagonal line as D and c is the threshold corresponding to that point, then $D = \sqrt{(sen(c) + spec(c) - 1)^2/2}$ and Youdan index J is $J = sen(c) + spec(c) - 1$.

The second criterion, "closest top left" criterion (Perkins and Schisterman 2006) identifies the point on the ROC curve that had the shortest distance to the top-left corner (a point that confers the perfect test). This criterion seeks to minimize the sum of squares of false positive rate and false negative rate. Intuitively, this point is the point on the ROC curve that is closest to point with perfect sensitivity and perfect specificity. Here the distance $D = \sqrt{(1 - sen(c))^2 + (1 - spec(c))^2}$.

4 Issues When Reporting the Optimal Threshold

The optimal thresholds determined through either Youdan index J or "closest top left" criteria that were reported in the medical or statistical literature have seldom been accompanied by any measures of uncertainty. We should be aware that since either threshold is estimated from the ROC curve, there are inherent variances associated with the threshold estimates. This motivated our simulation studies to assess the variability in threshold estimation.

5 Simulation Study

We had simulated data from normal distribution with 100 or 200 subjects, evenly distributed between diseased and non-diseased subjects. The parameters of the normal distribution are chosen with AUC of 0.760 or 0.814. The ROC curve and its AUC are estimated empirically and the variabilities of the estimations are evaluated through 1000 bootstrap samples. We report empirical AUC, optimal thresholds and their associated sensitivities and specificities. For each quantity, we will calculate the exact 95 % bootstrap confidence intervals (CI).

We first generated the data as $Y_{\bar{D}} \sim N(0, 1)$ and $Y_D \sim N(1, 1)$ so that the true AUC is 0.760.

Table 1 shows the simulation results. For $N = 200$, we found empirical AUC to be 0.761(95 % CI: 0.693 to 0.827). The optimal threshold is 0.448(95 % CI:

Table 1 Thresholds and the associated accuracy measures

	Youdan	Top-left	Spec=0.70	Spec=0.90
N=(50,50), AUC=0.760				
Threshold	0.428(−0.232,1.104)	0.496(0.108,0.897)	0.512(0.176,0.869)	1.244(0.805,1.712)
Sensitivity	0.75(0.52,0.94)	0.72(0.58,0.86)	0.69(0.50,0.86)	0.41(0.20,0.64)
Specificity	0.70(0.44,0.92)	0.72(0.58,0.86)	–	–
N=(100,100), AUC=0.760				
Threshold	0.448(−0.112,1.005)	0.491(0.176,0.794)	0.517(0.263,0.786)	1.265(0.969,1.592)
Sensitivity	0.73(0.54,0.90)	0.71(0.60,0.82)	0.69(0.55,0.81)	0.40(0.24,0.56)
Specificity	0.70(0.50,0.87)	0.71(0.60,0.81)	–	–
N=(50,50), AUC=0.814				
Threshold	0.568(0.206,0.951)	0.445(0.187,0.714)	0.256(0.077,0.438)	0.627(0.404,0.877)
Sensitivity	0.70(0.52,0.86)	0.74(0.62,0.86)	0.77(0.64,0.88)	0.65(0.48,0.80)
Specificity	0.89(0.72,1.00)	0.83(0.70,0.94)	–	–
N=(100,100), AUC=0.814				
Threshold	0.593(0.300,0.893)	0.445(0.251,0.634)	0.259(0.123,0.392)	0.633(0.475,0.805)
Sensitivity	0.68(0.54,0.80)	0.73(0.63,0.81)	0.77(0.67,0.86)	0.64(0.53,0.73)
Specificity	0.89(0.77,0.98)	0.83(0.73,0.91)	–	–

−0.112 to 1.005) using Youdan's index and 0.491(95 % CI: 0.176 to 0.794) using the closest top left criterion. The estimated sensitivity is 0.73(95 % CI: 0.54 to 0.90) or 0.71(95 % CI: 0.60 to 0.82) and the estimated specificity is 0.70(95 % CI: 0.50 to 0.87) or 0.71(95 % CI: 0.60 to 0.81), respectively.

We next simulated data as $Y_{\bar{D}} \sim N(0, 0.5)$ and $Y_D \sim N(1, 1)$ so that the true AUC is 0.814. For N=200, the empirical AUC was estimated to be 0.813(95 % CI: 0.743 to 0.872). The optimal threshold is 0.593(95 % CI: 0.300 to 0.893) using Youdan's index and 0.445(95 % CI: 0.251 to 0.634) using the closest top left criterion. The estimated sensitivity is 0.68(95 % CI: 0.54 to 0.80) or 0.73(95 % CI: 0.63 to 0.81) and the estimated specificity is 0.89(95 % CI: 0.77 to 0.98) or 0.83(95 % CI: 0.73 to 0.91), respectively.

Optimal threshold based on the Youdan index tends to have wider confidence intervals than the threshold estimated through the top-left corner criterion. Compared to the same quantities estimated from the top left corner criterion, the associated sensitivity at the Youdan's threshold is lower, but the associated specificity is higher, and both have wider confidence intervals.

However, the utility of "optimal threshold" is debatable. As shown above, the optimal thresholds and their associated sensitivities and specificities all have large variance and are hard to interpret. We instead propose to estimate the threshold corresponding to a pre-specified criterion, such as a fixed specificity. As shown in Table 1, we had estimated he threshold values corresponding to the fixed specificity level of 70 % or 90 %, and the associated sensitivities at those thresholds. For N=200 and AUC=0.814, the threshold is 0.259(95 % CI 0.123 to 0.392) at 70 %

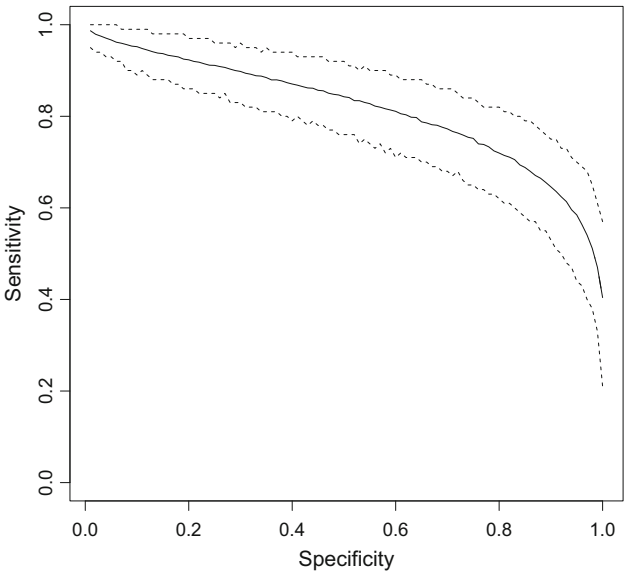

Fig. 1 Sensitivity-Specificity plot. The estimated sensitivities at the fixed specificity levels with the 95 % confidence intervals are shown

specificity and the associated sensitivity is 0.77(95 % CI 0.67 to 0.86). For 90 % specificity, the threshold is 0.633(95 % CI 0.475 to 0.805), the associated sensitivity is 0.64(95 % CI 0.53 to 0.73).

We had further estimated the thresholds and the associated sensitivities within the range of specificities of 0–99 % for N=200 and AUC=0.814, and plotted the estimated sensitivities versus the specificities in Fig. 1, which we named as a sensitivity-specificity plot, and included the point-wise 95 % confidence intervals for the estimated sensitivities. Similarly, Fig. 2 is a threshold-specificity plot, which presented both the estimated thresholds and the corresponding point-wise 95 % confidence intervals. Using both figures, we wanted to demonstrate the approach of finding the thresholds and the associated performance matrix for a prognostic biomarker.

6 Discussion

In estimating the optimal threshold, it is important to acknowledge the inherent variance of such estimates. In addition, a purely data-driven approach to search for optimal threshold can produce estimates that are not necessarily meaningful due to the large variance in such estimates. Instead, we propose to estimate the threshold through pre-specified criterion, such as a fixed level of specificity. The confidence intervals of the threshold and sensitivity at the pre-specified specificity are much

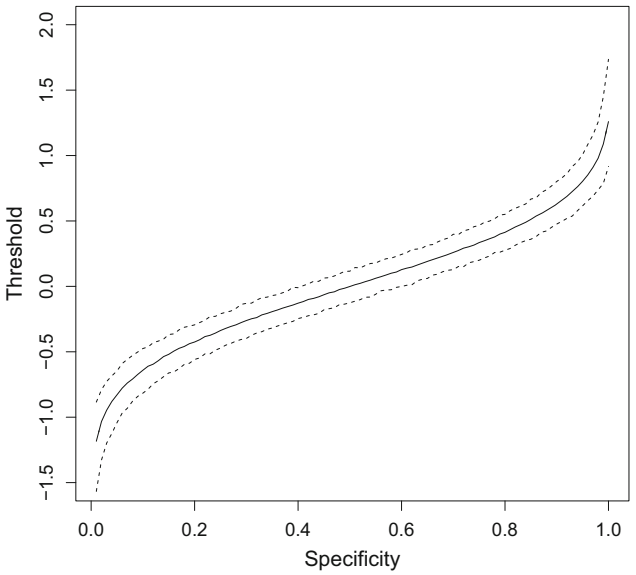

Fig. 2 Threshold-Specificity plot. The estimated thresholds at the fixed specificity levels with the 95 % confidence intervals are shown

narrower compared to the quantities measured through either Youdan index or closest top left criterion. We hereby suggest to estimate the threshold at a pre-specified level of specificity, and then estimate the sensitivity at that threshold, and all the estimates should be accompanied by appropriate 95 % confidence intervals.

References

Bihan, D. L., Urayama, S., Aso, T., Hanakawa, T. and Fukuyama, H. (2006). Direct and fast detection of neuronal activation in the human brain with diffusion MRI. *PNAS*, 103(21): 8263–8268.

Paulson, E. S. and Schmainda, K. M. (2008). Comparison of Dynamic Susceptibility-weighted Contrast-enhanced MR method: recommendations for measuring relative cerebral blood volume in brain tumors. *Radiology*, 249(2): 601–613.

Pepe, M. S. (2003). The statistical evaluation of medical tests for classification and prediction. Oxford University Press, United Kingdom.

Perkins, N. J. and Schisterman, E. F. (2006). The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am. J. Epidemiol.*, 163(7): 670–675.

Sourbron, S. P. and Buckley, D. L. (2013). Classic models for dynamic contrast-enhanced MRI. *NMR Biomed.*, 26: 1004–1027.

Swets, J. A. and Pickett, R. M. (1982). Evaluation of diagnostic systems: method from signal detection theory. Academic Press.

Youdan, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3: 32–35.

http://www.springer.com/978-3-319-42567-2

Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics

Selected Papers from the 2015 ICSA/Graybill Applied Statistics Symposium, Colorado State University, Fort Collins

Lin, J.; Wang, B.; Hu, X.; Chen, K.; Liu, R. (Eds.)
2016, XV, 359 p. 68 illus., 44 illus. in color., Hardcover
ISBN: 978-3-319-42567-2