Memorial Sloan-Kettering Cancer Center

Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology & Biostatistics Working Paper Series

Year 2008 *Paper* 16

Optimal Cutpoint Estimation with Censored Data

Mithat Gonen[*] Camelia Sima[†]

[*] Memorial Sloan-Kettering Cancer Center, gonenm@mskcc.org

[†] Memorial Sloan-Kettering Cancer Center, simac@mskcc.org

Optimal Cutpoint Estimation with Censored Data

Mithat Gonen and Camelia Sima

Abstract

We consider the problem of selecting an optimal cutpoint for a continuous marker when the outcome of interest is subject to right censoring. Maximal chi square methods and receiver operating characteristic (ROC) curves-based methods are commonly-used when the outcome is binary. In this article we show that selecting the cutpoint that maximizes the concordance, a metric similar to the area under an ROC curve, is equivalent to maximizing the Youden index, a popular criterion when the ROC curve is used to choose a threshold. We use this as a basis for proposing maximal concordance as a metric to use with censored endpoints. Through simulations we evaluate the performance of two concordance estimates and three chi-square statistics under various assumptions. Maximizing the partial likelihood ratio test statistic has the best performance in our simulations.

Optimal Cutpoint Estimation with Censored Data

Mithat Gönen

gonenm@mskcc.org

Camelia Sima

simac@mskcc.org

Department of Epidemiology and Biostatistics

Memorial Sloan-Kettering Cancer Center

307 East 63 Street, New York, NY 10021

Summary

We consider the problem of selecting an optimal cutpoint for a continuous marker when the outcome of interest is subject to right censoring. Maximal chi square methods and receiver operating characteristic (ROC) curves-based methods are commonly-used when the outcome is binary. In this article we show that selecting the cutpoint that maximizes the concordance, a metric similar to the area under an ROC curve, is equivalent to maximizing the Youden index, a popular criterion when the ROC curve is used to choose a threshold. We use this as a basis for proposing maximal concordance as a metric to use with censored endpoints. Through simulations we evaluate the performance of two concordance estimates and three chi-square statistics under various assumptions. Maximizing the partial likelihood ratio test statistic has the best performance in our simulations.

Key words: threshold, survival, ROC curve, concordance, maximal chi-square

1 Introduction

Dichotomization of continuous markers is usually disfavored because of the inherent loss of information (Royston et al, 2006). Nevertheless it is sometimes necessary to do so. For example, when a marker is used to inform treatment decisions, such as prostate specific antigen (PSA) in recurrent prostate cancer, a threshold is necessary for the binary action treat/do not treat. Similarly, clinical trials that seek to enroll high-risk patients will need a threshold for a marker to define high-risk. In other cases dichotomization is not essential but helpful. For example, adjustment for risk status as a confounder can be performed using the marker as a continuous covariate in an analysis of covariance model, but an analysis stratified by two levels of the marker could be favored because it enables the analyst to provide concrete summary statistics or graphical summaries (such as survival curves) within each risk group. For all these reasons, choosing a threshold to dichotomize a marker has been an active area of research (see, for example, Mazumdar and Glassman (2000) and the references therein).

There are two statistical approaches to the problem of choosing an optimal threshold. One uses the receiver operating characteristic (ROC) curve and the other attempts to maximize a suitably chosen test statistic. ROC curves are commonly used to assess the ability of ordinal or continuous markers in distinguishing the two states of a binary outcome. Their use in diagnostic medicine (Begg et al 2000; Pepe, 2003) and predictive modeling (Harrell, 1996) is firmly grounded. It may at first seem odd to use the ROC curve for dichotomizing purposes since it is usually promoted as a way to assess the overall discriminatory ability of the marker while avoiding dichotomization. However, since each point on the ROC curve represents the sensitivity and (one minus) the specificity of a potential threshold, it is only natural to compare the thresholds by using a criterion that combines these two measures of predictive accuracy. When the outcome of interest is binary, there are two widely-used such criteria: distance from the ideal marker and distance from the non-informative marker, also called the *Youden index* (Youden, 1950).

Despite their intuitive appeal, the expansion of ROC-based methods to censored outcomes is not a trivial problem. When the outcome is time to event and it is subject to censoring, an ROC curve can only be defined as a function of time (in fact they are called time-dependent ROC curves) (Haegerty et al, 2000; Haegerty and Zheng, 2005; Cai et al, 2006). Since most investigators find the idea of a time-dependent threshold irksome, using a time-dependent ROC curve will require first selecting a time point, which could be an arbitrary choice.

While a single ROC curve with censored data is elusive, it is possible to estimate the concordance probability, a metric closely connected to the area under the ROC curve, when the outcome is subject to right censoring (Harrell et al, 1984; Gönen and Heller, 2005). In this article we will propose a criterion that uses concordance to dichotomize a continuous marker. We will show that this criterion concurs with the Youden index, thereby making available a familiar criterion to choose the optimal cutpoint with censored data.

The other approach commonly used for dichotomization is the maximization of an appropriate test statistic. This is often a two-sample test that compares the groups resulting from dichotomization. This method was first developed in the context of a binary outcome using Pearson's chi-square test (Miller and Siegmund, 1982) and is often called the maximal chi-square method. In the case of censored data one can use three test statistics routinely reported by commonly available statistical software: log-rank test, Wald test and the partial likelihood ratio test (Kalbfleisch and Prentice, 2002). Since the score test from a proportional hazards model with only one binary explanatory variable and no ties in the survival times is asymptotically the same as the log-rank test (Klein and Moeschberger, 1997), it is not given separate consideration.

The following section details on the notion of concordance probability and places it in the context of other ROC-based methods available for dichotomization. Section 3 presents and compares two estimates of concordance probability commonly used in practice, while section 4 briefly reviews the maximal chi-square methods used with censored data. We present our simulation study and its results in Section 5, and conclude with a general discussion and recommendations in Section 6.

2 Maximal concordance: Definition and Relation to Existing Criteria

Concordance is the probability that, in a pair of randomly selected patients, the ordering of markers is consistent with the order of outcomes. If X is the marker and Y is the outcome, then concordance probability can be defined as $CP = P(Y_1 > Y_2|X_1 \geq X_1)$ for two randomly selected observations. Concordance is directly linked to various rank correlation measures such as Kendall's τ, Somers' D and Kruskal's γ (Pratt and Gibbons, 1981). More importantly for our purposes, when Y is binary concordance is a monotone function of the area under the ROC curve (Begg et al, 2000). This is a key point in the subsequent development in this article and it implies that optimal cutoffs chosen by maximizing the AUC and concordance would coincide.

Although the ROC curve and the corresponding AUC are most often considered for ordinal or continuous markers, they are also defined for binary markers. Consider a marker with M distinct values $m = 1, \cdots, M$. Each value m, characterized by sensitivity Sens_m and specificity Spec_m, can be regarded as a threshold that dichotomizes marker into a binary variable Z_m. In Figure 1, let $\mathscr{C}$ be the ROC curve of the marker that we want to dichotomize. If P is the point on the ROC curve corresponding to the value m, then the pair of segments (OP, PV) can be regarded as the ROC of the binary marker Z_m, and the corresponding AUC can be computed as:

$$AUC_m = \frac{\text{Sens}_m - (1 - \text{Spec}_m) + 1}{2}$$

The threshold that provides the maximum AUC will be called the maximal concordance threshold.

We will call the thresholds obtained from the ROC-based methods *operating points*. In the remaining of this section we will describe two methods that are commonly used for choosing an operating point on the ROC curve, and explore their relationship to the proposed method of maximal concordance.

In Figure 1, the diagonal line OV represents the ROC curve of a non-informative marker, and the point B represents the ideal binary marker which is 100% sensitive and specific. The first method seeks the threshold m that maximizes the vertical distance $|PJ|$ from the curve $\mathscr{C}$ to the non-informative marker. This optimal threshold is called the Youden Index, and it is the point which, at the same level of specificity as the non-informative marker, provides the maximum excess of sensitivity (Youden, 1950). The second method chooses the threshold that minimizes the distance $|PB|$ to the ideal marker B. Previous work has shown that these two methods often disagree. In fact one can show that, if the marker distributions within negative and positive groups both belong to the same location-scale family, the two methods will agree (i.e. find the same cutpoint) if and only if the two distributions have the same scale parameter (Perkins and Schisterman, 2006).

To enable discussion of the three ROC-based criteria (Youden Index, minimum distance from the ideal marker, maximal concordance) under a single analytical framework, we will let $f(x, y)$ generically represent a criterion and define a threshold optimal with respect to f if the following is satisfied:

$$m^{opt} = \text{argmax}_m \ f(\text{Sens}_m, \text{Spec}_m)$$

It can easily be shown that the three criteria discussed so far can be represented with the following choices of f:

$$\begin{aligned} f_1 &= \text{Sens}_m - (1 - \text{Spec}_m) && (1) \\ f_2 &= -\sqrt{(1 - \text{Spec}_m)^2 + (1 - \text{Sens}_m)^2} && (2) \\ f_3 &= \frac{\text{Sens}_m - (1 - \text{Spec}_m) + 1}{2} && (3) \end{aligned}$$

where the minus sign in the second definition is used to ensure compliance with the definition of m_{opt} as a maximization over f. Here f_1 is the objective function for Youden's index, f_2 for distance from ideal marker and f_3 for maximal concordance. We will use m_1^{opt}, m_2^{opt} and m_3^{opt} to refer to the optimal operating points satisfying these three criteria.

As shown by Perkins (2006), in general $m_1^{opt} \neq m_2^{opt}$. It is also evident from the above formulation that $m_1^{opt} = m_3^{opt}$. Therefore, the Youden's index and the maximal concordance criteria are equivalent.

3 Using Maximal Concordance with Censored Data

Despite the lack of a single ROC curve (and hence a single AUC), concordance probability (CP) is well-defined with censored data. In general, for two pairs of observations randomly selected from the bivariate distribution (X, T), where X is a continuous marker and T is time to event, the concordance probability is defined as

$$CP = \mathrm{pr}(T_2 > T_1 | X_2 \geq X_1).$$

While concordance probability is well-defined, its estimation is not necessarily straightforward. An estimator proposed by Harrell et. al. (1984), called the c-index, has been widely used in pratice. Letting t denote the observed survival or follow-up time, δ the censoring status and x the marker, the c-index is computed by forming all pairs $\{(t_i, x_i, \delta_i), (t_j, x_j, \delta_j)\}$ where the smaller follow-up time is a failure time:

$$CI = \frac{\sum\sum_{i<j} \{I(t_i < t_j)I(t_i > t_j)I(\delta_i = 1) + I(t_j < t_i)I(x_j > x_i)I(\delta_j = 1)\}}{\sum\sum_{i<j} \{I(t_i < t_j)I(\delta_i = 1) + I(t_j < t_i)I(\delta_j = 1)\}}$$

The relationship of CI to Kendall's τ is investigated by Pencina and D'Agostino (2004).

Harrell's c-index is simple to conceptualize and operationalize but has the deficiency of ignoring pairs where the shorter time is censored. An altenative estimate can be obtained if one is willing to assume proportional hazards (Gönen and Heller, 2005), which leads to the following definition for the CP:

$$CP = \frac{\displaystyle\iint_{\beta^{\mathrm{T}}x_1 > \beta^{\mathrm{T}}x_2} \left[1 + \exp\left\{\beta^{\mathrm{T}}(x_2 - x_1)\right\}\right]^{-1} dF(\beta^{\mathrm{T}}x_1)dF(\beta^{\mathrm{T}}x_2)}{\displaystyle\iint_{\beta^{\mathrm{T}}x_1 > \beta^{\mathrm{T}}x_2} dF(\beta^{\mathrm{T}}x_1)dF(\beta^{\mathrm{T}}x_2)},$$

where β is the vector of regression parameters in the corresponding Cox model, x is the vector of covariates and F is the distribution function of the covariate linear combination $\beta^{\mathrm{T}}X$. Concordance probability can be estimated by substituting estimators of β and F in the above expression. The partial likelihood estimator $\hat{\beta}$ presents itself naturally for β and the empirical distribution is used for F. The result is the concordance probability estimator:

$$CPE = \frac{2}{n(n-1)} \sum\sum_{i<j} \left\{ \frac{I(\hat{\beta}^{\mathrm{T}}x_{ji} < 0)}{1 + \exp(\hat{\beta}^{\mathrm{T}}x_{ji})} + \frac{I(\hat{\beta}^{\mathrm{T}}x_{ij} < 0)}{1 + \exp(\hat{\beta}^{\mathrm{T}}x_{ij})} \right\},$$

where x_{ij} represents the pairwise difference $x_i - x_j$.

Harrell's c-index CI is biased and the bias increases with censoring rate (Gönen and Heller, 2005). On the other hand, under proportional hazards, CPE is a consistent estimate of concordance since $\hat{\beta}$ itself is consistent. It is not clear, though, which concordance estimate will perform better in choosing a cutpoint with varying censoring rates and departures from proportional hazards.

4 Using Maximal Chi-Square with Censored Data

In addition to the ROC-based methods, another approach taken for dichotomizing a continuous marker is to maximize a test statistic that represents the association between the dichotomized marker and the binary outcome (Mazumdar and Glassman, 2000). It is well-known that a maximally selected test statistic has a different null distribution than one which involves no selection. For example, for a randomly selected cutpoint with a binary outcome, the null distribution of Pearson's chi-square test follows a chi-square distribution, but the maximum of test statistics over all possible thresholds follows a Brownian bridge under the null hypothesis of no association. While great effort has been devoted to deriving these distributions under a variety of conditions (Miller and Siegmund, 1982; Lausen and Schumacher, 1992; Hilsenbeck and Clark, 1996; Mazumdar and Glassman, 2000), considerably less emphasis was placed on the bias of the cutpoint estimate using this method.

In our simulation study, the maximal chi-square method will be applied with three test statistics that we will attempt to maximize: log-rank test (T_{LogR}), Wald test (T_W) and the partial likelihood ratio test (T_{PL}), where the latter two are obtained from a proportional hazards model with the repeatedly dichotomized marker as the only covariate. The estimates of the three test statistics are formally defined below (Kalbfleisch and Prentice, 2002):

$$T_{LogR} = \sum\nolimits_{i=1}^{n} \delta_i \times \left(Z_i - \frac{\sum\nolimits_{j=1}^{n} I(t_j \geq t_i) \times Z_j}{\sum\nolimits_{j=1}^{n} I(t_j \geq t_i)} \right) \tag{4}$$

$$T_W = \hat{\beta}^T \times [\hat{V}(\hat{\beta})]^{-1} \times \hat{\beta} \tag{5}$$

$$T_{PL} = 2 \times [l(\hat{\beta}) - l(0)] \tag{6}$$

where δ_i, t_i and $\hat{\beta}$ are defined as in Section 3; Z_i is the group membership indicator for subject i; $\hat{V}(\hat{\beta})$ is the estimated variance of $\hat{\beta}$; and $l(\cdot)$ is the partial likelihood.

5 Simulations

We first describe the simulation scheme used to generate data such that there is a true cutpoint in the distribution of a continuous marker that separates two groups with distinct survival outcomes. Then, we outline the approach taken to estimate the optimal cutpoint based on each of the following criteria: three variations of the maximal chi-square method (maximal log-rank test, Wald test and partial likelihood-ratio test) and two variations of the concordance-based method (maximal c-index and CPE).

5.1 Data generation

A continuous marker X is generated from a normal distribution with mean μ and variance v. We choose a cutpoint c^* to create a binary variable $Z = I(X \leq c^*)$.

Survival times T corresponding to each Z value are generated from two independent Weibull distributions, with shape and scale parameters (γ_0, λ_0), for $Z = 0$, and (γ_1, λ_1), for $Z = 1$. This way of generating data ensures that c^* is the true cutpoint that divides the sample into two groups with distinct survival outcomes. An independent censoring time U is generated from a uniform distribution ranging from 0 to τ, and the event indicator δ is defined as $I(T \leq U)$. The value of τ controls the level of censoring.

5.2 Choosing the best cutpoint

Consider a value x on the continuous distribution of X. A Cox proportional hazard model can be fit such that:

$$h(t|x) = h_0(t)e^{\beta I(X \geq x)}$$

Based on this model, the following statistics are estimated: Wald test (T_W), the partial likelihood ratio test (T_{PL}), and the concordance probability estimate (CPE).

Additionally, we estimate the log-rank statistic (T_{LogR}) for the test of survival difference between the two groups defined by $I(X \geq x)$, and the c-index (CI) as described in Harrell et. al. (1984).

The five measures listed above are estimated for each value x between the 5^{th} and the 95^{th} percentile of the X distribution. Excluding from evaluation the extreme values of the cutpoint is standard pratice to avoid singularities due to small group sizes that result from these cutpoints (Mazumdar and Glassman, 2000). The best cutpoint, according to each of the five criteria, is the one that maximizes the corresponding statistic.

5.3 Simulation results

Tables 1-3 present simulation results, under different parameter scenarios.

Tables 1 and 2 examine the situation when the survival times corresponding to the two groups meet the proportional hazards assumption ($\gamma_0 = \gamma_1$). The bias in the optimal cutpoint estimates is investigated when the true cutpoint migrates away from the center of the marker's distribution by 0.25, 0.5, 0.75 and 1 standard deviation. The proportion of censoring increases from 35% - 55% in Table 1 to 75% - 85% in Table 2.

Table 3 examines the performance of the five criteria when the two groups violate the proportional hazards assumption ($\gamma_0 \neq \gamma_1$) and have different degrees of separation (as controlled by the difference between λ_0 and λ_1). The proportion of censoring is 45% - 55%. This scenario is of particular interest because the CPE is derived assuming proportional hazards.

In all simulations, the continuous marker is generated from a normal distribution with mean $\mu = 0$ and variance $v = 4$. The sample size is $n = 100$ although, for scenarios that generate considerably biased estimates, increased sample sizes ($n = 300, 500$) are also examined.

The following messages emerge from these results:

- All five criteria perform well when the true cutpoint lies in the center of the marker's distribution, even when the proportional hazards assumption is violated.

- For all five criteria, the bias in the estimates increases as the true cutpoint migrates away from the center of the marker's distribution (Table 1). The bias decreases with increasing sample size.

- T_{PL} outperforms the other two chi-square-based methods in all scenarios considered, while the Wald test has the worst results among all methods considered.

- CI and CPE have similar performance, with CI generating slightly better estimates when the true cutpoint migrates away from the center of the marker's distribution and the censoring proportion is maintained below 60% (Table 1).

- When the percentage of censored observations exceeds 75%, the only methods that provide reasonably good estimates are T_{PL} and CPE (Table 2).

6 Lung Cancer Example

Locally advanced lung cancer is treated with surgery. Recurrence rates are high even after complete resection of the tumor hence it is commonly accepted that high-risk patients should be treated with post-surgical chemotherapy. One possible way of selecting patients for chemotherapy is the glucose uptake of the tumor in a positron emission tomography (PET) scan, frequently summarized by the standardized uptake value (SUV), a continuous variable where higher values are indicative of a more aggressive tumor.

In a study where all patients underwent a PET scan before surgery, SUV values are associated with survival (Downey et al, 2004). The median follow-up was 26 months (range: 5-81 months) and 79% of observations were censored. The interest is selecting a cutpoint for SUV above which patients will be considered high risk and candidates for post-surgical chemotherapy. In this data SUV had a median of 9 (range: 0.5-32) and a mean of 10 (standard deviation=6.8).

The CPE and T_{PL} methods selected the same cutpoint (8.93), situated in the center of the marker's distribution (50^{th} quantile). T_W statistic chose a close optimal threshold (9.8, 54^{th} quantile), while both CI and T_{LogR} selected an extreme cutpoint (2.3, 8^{th} quantile). When restricting the limits of the cutpoint search from the 5th-95th percentiles of the SUV distribution to the 10th-90th percentiles, CPE, T_{PL} and T_W produced the same result, while CI and T_{LogR} selected the most extreme possible value.

This is a situation when the selected cutpoint varies remarkably based on the method chosen. This finding is not surprising: our simulation results showed that, in the case of datasets with high censoring rates, CPE and T_{PL}, although imperfect, give the only reliable results. Based on these considerations, we recommend choosing 8.93 (which is also a clinically meaningful value) for separating low-risk from high-risk patients.

7 Discussion

Our simulations indicate that maximizing the likelihood ratio test statistic has the smallest bias under a variety of scenarios, including high censoring and violations of the proportional hazards assumption. Maximizing the log-rank and Wald statistics have considerably worse performance. At first look this may come as a surprise because applied statisticians have come to view those tests as exchangeable in larger samples. While that view applies for significance testing of the group differences (or regression coefficients), we have shown that the location of its maxima can be quite different.

It is important to remember that our simulations were based on a true-cutpoint model and the results should be interpreted in this context. If one believes that the underlying model follows this assumption, then the question of whether a test statistic or concordance should be used is somewhat moot because the overarching goal would be to estimate the true cutpoint. In many instances, however, one might be interested in estimating a cutpoint even if the data generating process does not have a true cutpoint. For example the relationship between most biomarkers and disease progression is thought to be smooth. Nevertheless, a dichotomization of the biomarker may still be necessary if one will use the biomarker for identifying high-risk patients for a clinical trial. In this case the metric used to select the cutpoint (concordance or test statistic) is relevant directly to how the dichotomized version of the marker will be used. In this example of selecting patients for a clinical trial, concordance may be more appropriate because the dichotomized marker is essentially being used to predict who is at high risk. In contrast, consider the situation when a marker is used for stratifying patients during randomization. This is essentially a problem of association since the data will be analyzed by a stratified model and maximum power is achieved when stratifying factors are highly associated with the outcome. In this case maximizing a test statistic may be preferred.

Our results can provide guidance even when the true-cutpoint model is not thought to hold. Based on the foregoing discussion, investigators would identify their optimality criterion in conjunction with the purpose of the use of the cutpoint. If maximal concordance is desired then an argument for the use of c-index can be made because it exhibited smaller bias than CPE in most of the situations covered in our simulation study, except for the case when the censoring level is high.

All estimators had a certain amount of underestimation the degree of which is a function of the distance of the true cutpoint from the center of the marker distribution. Similar observations were made by Lausen and Schumacher (1992), although in a narrower context. An intuitive explanation for this underestimation is that, as the true cutpoint moves away from the center of the distribution, one of the dichotomized groups gets smaller, resulting in poorer estimation of its outcome profile.

In general our results point to the difficulty of finding the true cutpoint. Most statistical tools are devised with smooth relationships in mind: a regression model, for example, can be considered as a way of "smoothing" noisy data. On the other hand a true-cutpoint model is non-differentiable at the cutpoint and hence non-smooth. Using methods that are originally devised for smoothing the data to detect singular points is fundamentally difficult and partly explains the persistent bias we observed in our simulations. As long as one is aware of this inherent bias, using the partial likelihood ratio statistic or concordance probability estimate are the best available strategies.

References

Begg C. B., Cramer L. D., Venkatraman E. S., Rosai J. (2000). Comparing tumour staging and grading systems : a case study and a review of the issues, using thymoma as a model. *Statistics in Medicine* 19: 1997–2014.

Cai T., Pepe M.S., Zheng Y., Lumley T., Jenny N.S. (2006) The sensitivity and specificity of markers for event times. Biostatistics 7:182–97

Cox, D.R. (1972). Regression models and life tables (with Discussion). *J. R. Statist. Soc.* B **34**, 187–220.

Downey, R.J., Akhurst, T., Gönen, M., Vincent, A., Bains, M.S., Larson, S.M. & Rusch, V. (2004). Preoperative F-18 fluorodeoxyglucose-positron emission tomography maximal standardized uptake value predicts survival after lung cancer resection. *J. Clin. Oncol.* **22**, 3255-60.

Gönen M., Heller G. (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* 92:965–970.

Haegerty P.J., Lumley T., Pepe M.S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 56:337–344.

Haegerty P.J., Zheng Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* 61:92–105.

Harrell F.E. Jr, Lee K.L., Califf R.M., Pryor D.B., Rosati R.A. (1984) Regression modelling strategies for improved prognostic prediction. *Stat Med.* 3:143–52.

Harrell FE Jr, Lee KL, Mark DB. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med.* 15:361–87.

Hilsenbeck S.G., Clark G.M. (1996) Practical p-value adjustment for optimally selected cutpoints. *Stat Med.* 15:103–12.

Kalbfleisch J. D., Prentice, R. L. (2002) *The Statistical Analysis of Failure Time Data.* Wiley, New York.

Klein J.P., Moeschberger M.L. (2005). *Survival Analysis: Techniques for Censored and Truncated Data.* Springer, New York.

Lausen, B., Schumacher, M. (1992). Maximally Selected Rank Statistics. *Biometrics* 48:73-85

Mazumdar M., Glassman J.R. (2000). Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Stat Med* 19: 113–132.

Miller and Siegmund (1982). Maximally selected chi square statistics. *Biometrics* 38:1011–1016.

Pencina, M.J. & D'Agostino, R.B. (2004). Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statist. Med.* **23**, 2109-23.

Pepe M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press, New York.

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

Pratt, J.W. & Gibbons, J.D. (1981). *Concepts of Nonparametric Theory.* New York: Springer-Verlag.

Royston, P., Altman, D.G., Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine* 25: 127–141.

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

Table 1: Data generated under proportional hazards assumption. Censoring Rate: 35-55%

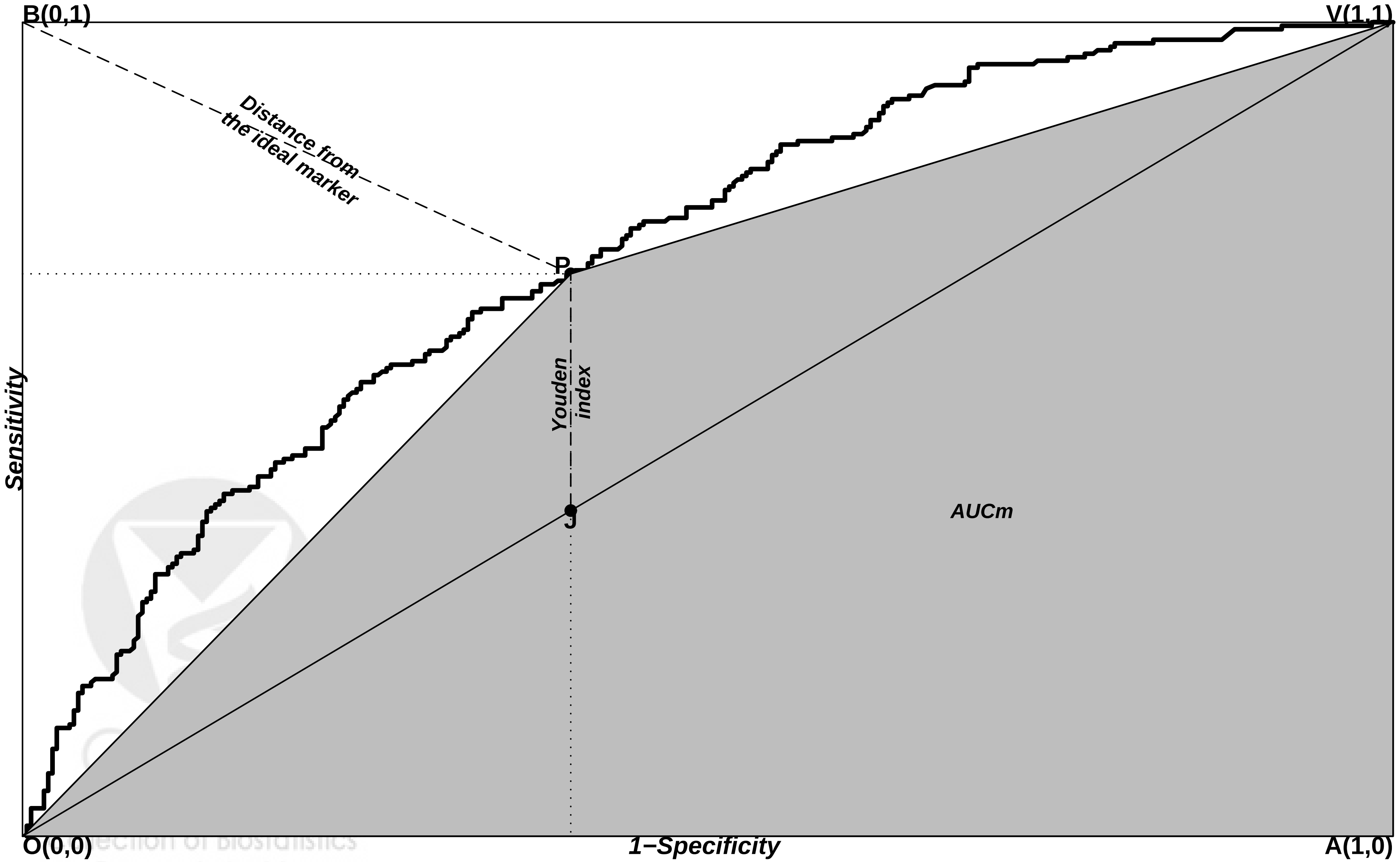
$(\gamma_1, \lambda_1), (\gamma_0, \lambda_0)$	True cutpoint	Sample Size	Selected Optimal Cutpoint (Bias)				
			T_{LogR}	T_W	T_{PL}	CI	CPE
(1,3), (1,1)	0	100	0.08 (0.08)	-0.06 (-0.06)	-0.02 (-0.02)	-0.07 (-0.07)	-0.01 (-0.01)
	0.5	100	0.42 (-0.08)	0.41 (-0.09)	0.47 (-0.03)	0.42 (-0.08)	0.46 (-0.04)
	1	100	0.9 (-0.1)	0.81 (-0.19)	0.97 (-0.03)	0.88 (-0.12)	0.81 (-0.19)
	1.5	100	1.36 (-0.14)	1.08 (-0.42)	1.46 (-0.04)	1.3 (-0.2)	1.28 (-0.22)
	1.5	300	1.47 (-0.03)	1.38 (-0.12)	1.5 (0)	1.47 (-0.03)	1.42 (-0.08)
	2	100	1.83 (-0.17)	1.15 (-0.85)	1.89 (-0.11)	1.55 (-0.45)	1.35 (-0.65)
	2	300	1.97 (-0.03)	1.72 (-0.28)	1.99 (-0.01)	1.94 (-0.06)	1.75 (-0.25)
	2	500	1.98 (-0.02)	1.77 (-0.23)	1.99 (-0.01)	1.97 (-0.03)	1.85 (-0.15)

Table 2: Data generated under proportional hazards assumption. Censoring Rate: 75-85%

$(\gamma_1, \lambda_1), (\gamma_0, \lambda_0)$	True cutpoint	Sample Size	Selected Optimal Cutpoint (Bias)				
			T_{LogR}	T_W	T_{PL}	CI	CPE
(1,3), (1,1)	0	100	-0.3 (-0.3)	-0.43 (-0.43)	-0.09 (-0.09)	0.18 (0.18)	0.02 (0.02)
	0.5	100	0.06 (-0.44)	-0.08 (-0.58)	0.37 (-0.13)	0.21 (-0.29)	0.42 (-0.08)
	1	100	0.51 (-0.49)	0.23 (-0.77)	0.84 (-0.16)	0.46 (-0.54)	0.83 (-0.17)
	1.5	100	0.58 (-0.92)	0.54 (-0.96)	1.16 (-0.34)	0.7 (-0.8)	1.06 (-0.44)
	2	100	0.98 (-1.02)	0.24 (-1.76)	1.6 (-0.4)	0.83 (-1.17)	1.34 (-0.66)

Table 3: Data violate proportional hazards assumption. Censoring Rate: 45-55%

$(\gamma_1, \lambda_1), (\gamma_0, \lambda_0)$	True cutpoint	Sample Size	Selected Optimal Cutpoint (Bias)				
			T_{LogR}	T_W	T_{PL}	CI	CPE
(1, 3), (1.5, 1)	0	100	-0.05 (-0.05)	-0.04 (-0.04)	-0.02 (-0.02)	-0.05 (-0.05)	-0.002 (-0.002)
(1, 3), (2, 1)	0	100	-0.04 (-0.04)	-0.01 (-0.01)	-0.004 (-0.004)	-0.03 (-0.03)	0.02 (0.02)
(1, 2), (1.5, 1)	0	100	-0.06 (-0.06)	-0.005 (-0.005)	-0.05 (-0.05)	-0.03 (-0.03)	0.08 (0.08)
(1, 2), (2, 1)	0	100	-0.005 (-0.005)	0.13 (0.13)	0.09 (0.09)	0.01 (0.01)	0.08 (0.08)
(1, 3), (1.5, 1)	1	100	0.93 (-0.07)	0.81 (-0.19)	0.98 (-0.02)	0.89 (-0.11)	0.91 (-0.09)
(1, 3), (2, 1)	1	100	0.95 (-0.05)	0.75 (-0.25)	0.99 (-0.01)	0.91 (-0.09)	0.93 (-0.07)
(1, 2), (1.5, 1)	1	100	0.91 (-0.09)	0.61 (-0.39)	0.99 (-0.01)	0.82 (-0.18)	0.84 (-0.16)
(1, 2), (2, 1)	1	100	1 (0)	0.56 (-0.44)	1.02 (0.02)	0.91 (-0.09)	0.9 (-0.1)


B(0,1)
V(1,1)
Distance from
the ideal marker
P
Sensitivity
Youden
index
J
AUCm
O(0,0)
1−Specificity
A(1,0)