Published in final edited form as:
Stat Med. 2010 June 15; 29(13): 1391–1410. doi:10.1002/sim.3876.

Assessing risk prediction models in case-control studies using semiparametric and nonparametric methods

Ying Huang[†,†] and **Margaret Sullivan Pepe**[*]

[†]Department of Biostatistics, Mailman School of Public Health, Columbia University New York, NY 10032

[*]Fred Hutchinson Cancer Research Center Public Health Sciences 1100 Fairview Avenue N., M2-B500, Seattle, WA 98109-1024

Summary

The predictiveness curve is a graphical tool that characterizes the population distribution of $Risk(Y) = P(D = 1|Y)$, where D denotes a binary outcome such as occurrence of an event within a specified time period and Y denotes predictors. A wider distribution of $Risk(Y)$ indicates better performance of a risk model in the sense that making treatment recommendations is easier for more subjects. Decisions are more straightforward when a subject's risk is deemed to be high or low. Methods have been developed to estimate predictiveness curves from cohort studies. However early phase studies to evaluate novel risk prediction markers typically employ case-control designs. Here we present semiparametric and nonparametric methods for evaluating a continuous risk prediction marker that accommodate case-control data. Small sample properties are investigated through simulation studies. The semiparametric methods are substantially more efficient than their nonparametric counterparts under a correctly specified model. We generalize them to settings where multiple prediction markers are involved. Applications to prostate cancer risk prediction markers illustrate methods for comparing the risk prediction capacities of markers and for evaluating the increment in performance gained by adding a marker to a baseline risk model. We propose a modified Hosmer-Lemeshow test for case-control study data to assess calibration of the risk model that is a natural complement to this graphical tool.

Keywords

biomarker; case-control study; classification; Hosmer-Lemeshow test; predictiveness curve; risk; ROC curve

1. Introduction

Criteria for evaluating a biomarker depend on the purpose for which it will be used. The key performance measure for a diagnostic marker is its classification accuracy, i.e. the ability to provide the correct diagnosis given a subject's true disease status. Classification accuracy of a continuous marker has been commonly assessed by the receiver operating characteristic (ROC) curve [1]. Classification, however, is not always the objective. Sometimes a marker is used mainly to predict risk of disease and to stratify the population into risk groups geared towards different treatment recommendations. Because of its popularity in the field of diagnostic testing, the ROC curve has been used frequently in this setting as well. However, as pointed out by Gail and Pfeiffer [2], Cook [3], and Pencina *and others* [4], criteria for

[†]*Corresponding author's address:* yh2441@columbia.edu.

evaluating a classification marker might be unnecessarily stringent for evaluating a risk prediction marker. In other words, the ROC curve may not be optimal when selecting a marker for risk prediction.

The predictiveness curve [5] was proposed by Pepe *and others* [6] and Huang *and others* [7] to evaluate a risk prediction marker or model. It characterizes the performance of a risk prediction model by displaying the population distribution of risk endowed by the model. Arguments for displaying the risk distribution have also appeared recently in the clinical literature [8]. A binary outcome D is considered here such as presence of disease or occurrence of an event within some specified time period. We write $D = 1$ for cases, subjects with a bad outcome and $D = 0$ for controls, subjects with a good outcome. Let Y be a vector of predictors of interest and let $Risk(Y) = P(D = 1|Y)$ be the risk calculated based on Y. The predictiveness curve displays the risk distribution through the population quantiles, $R(v)$ vs v for $v \in (0, 1)$, where $R(v)$ is the v^{th} quantile of $Risk(Y)$. Equivalently, the inverse function, $R^{-1}(p) = P\{Risk(Y) \leq p\}$, is the proportion of the population with risks less than or equal to p, the cumulative distribution function. If p_H corresponds to a high risk threshold, the capacity of the risk model to identify high risk subjects is $1 - R^{-1}(p_H)$. If p_L is a low risk threshold, $R^{-1}(p_L)$ quantifies the capacity of the model to identify low risk subjects. Better risk markers put more subjects into high and low risk categories and fewer people into the intermediate range where treatment decisions are more difficult. In other words, a risk prediction model with larger variability in population quantiles, i.e. steeper predictiveness curve, has a better capacity to stratify risk.

For cohort studies, Huang *and others* [7] developed a semiparametric estimator of the predictiveness curve. However case-control studies, being smaller and more cost efficient than cohort studies, are the design of choice in early phases of biomarker development [9,10]. Thus one objective of the current manuscript is to extend estimation to case-control designs. We describe two semiparametric methods. Large sample theory for these estimators was developed in Huang and Pepe [11] when Y is univariate. Here we consider the practical application of these methods. We examine methods for making inference in practical sample sizes and evaluate them using simulation studies. Importantly we extend the methods to accommodate multiple predictors as this often arises in real applications. In practice, robustness to modeling assumptions is always a concern. Another objective of the current paper is to develop a nonparametric estimator. We compare its performance with the semiparametric methods in simulations and in a real dataset. Moreover, we propose a measure accompanying the estimated predictiveness curve to formally test for calibration of the risk model.

We begin with models including only a single continuous marker or a pre-defined marker combination and later examine the extension to a general risk model. The problems caused by developing combinations and assessing them in the same dataset have been well recognized and the assessment of a predefined combination with independent test data is encouraged [12,13]. In these circumstances our methods apply to evaluations with the test data. For example, Buyse *and others* [14] recently reported the performance of a gene expression signature combination previously developed by van't Veer *and others* [15] and van de Vijver *and others* [16]. Other examples of well known predefined combination scores are the Framingham score for cardiovascular events [17] and the Gail score for breast cancer risk [18].

Let $\rho = P(D = 1)$ denote the prevalence of the bad outcome. We assume either that ρ is fixed at a specified value or that an estimate $\widehat{\rho}$ is available in addition to the case-control sample. For example, the prevalence is essentially known if obtained from a large population registry; alternatively, one can entertain various fixed values for ρ that might reflect

prevalences in different populations, performing a "what if" exercise that allows one to surmise in which populations the biomarker would be useful and in which populations it might not. Settings where a prevalence estimate is available includes estimates from an independent cohort study reported in the literature, or estimates calculated from a parent cohort within which the case-control study is nested [10,19]. When an estimate of ρ is obtained from an independent cohort or the parent cohort, variability in $\widehat{\rho}$ must be taken into account in computing variance of the predictiveness estimator.

We make the assumption that $P(D = 1|Y)$ is monotone increasing in Y. If the risk is decreasing in Y, the marker can be negated to satisfy this assumption. Extensions discussed in section 6 accommodate non-monotone risk functions. Under the monotone increasing risk assumption, the v^{th} quantile of the marker corresponds to the v^{th} quantile of risk which implies that $R(v) = Risk\{F^{-1}(v)\}$. For estimation purposes we therefore need to estimate the risk function, $Risk(Y) = P(D = 1|Y)$, as well as the marker distribution $F(y) = P(Y \leq y)$, and combine the two estimands to get the estimator for the risk quantile.

2. Estimation of the Risk Function

In this section, we consider estimation of the risk function as the first step in estimating the predictiveness curve. The risk can be estimated either using parametric or nonparametric methods. The former gives rise to semiparametric predictiveness curve estimates while the latter gives rise to fully nonparametric estimates.

2.1 Parametric Risk Functions: Logistic Regression

For case-control data, a logistic regression formulation of the risk model is convenient. We write it as

$$\text{logit}P(D=1|Y) = \text{logit}\{G(\theta, Y)\} = \theta_0 + \eta(\theta_1, Y) \tag{2.1}$$

where η is monotone increasing in Y. For example, $\eta(\theta_1, Y)$ can take a linear form $\theta_1 Y$ with $\theta_1 > 0$. A more general and flexible model can involve the Box-Cox type transformation [20]. That is $\eta(\theta_1, Y) = \theta_{11} Y^{(\theta_{12})}$ with $\theta_{11} > 0$, where $Y^{(\theta_{12})} = (Y^{\theta_{12}} - 1)/\theta_{12}$ when $\theta_{12} \neq 0$ and $Y^{(\theta_{12})} = \log Y$ when $\theta_{12} = 0$. In case-control studies, since the sampling rate of cases versus controls is fixed by design, the intercept term θ_0 in the risk model is not estimable. However, the odds ratio is still estimable, a fact that is routinely exploited in epidemiology [21]. The maximum likelihood estimator of the odds ratio from the retrospective likelihood can be obtained by maximizing the prospective likelihood of the case-control sample, pretending that the outcome is random and ignoring the outcome-dependent nature of the sampling [22,23].

Let n_D and $n_{\bar{D}}$ be the number of cases and controls respectively in the case-control sample. Applying the logistic regression model (2.1) to the data and then applying a shift $\log\left\{\widehat{\rho}/(1-\widehat{\rho})\, n_{\bar{D}}/n_D\right\}$ to the intercept, we obtain $\widehat{\theta} = (\widehat{\theta}_0, \widehat{\theta}_1)$, the maximum likelihood estimator of θ. This follows because the population odds is related to the sample odds as a result of the Bayes' theorem:

$$\frac{P(D=1|Y)}{P(D=0|Y)} = \frac{P(D=1|Y,S)}{P(D=0|Y,S)} \frac{P(D=0|S)}{P(D=1|S)} \frac{P(D=1)}{P(D=0)},$$

where S is the indicator of being included in the case-control sample. Therefore to calculate the population risk from the model fit to case-control data, we add the term $\log\left\{\widehat{\rho}/(1-\widehat{\rho})\, n_{\bar{D}}/n_D\right\}$ to the estimated intercept.

2.2 Nonparametric Risk Functions: Isotonic Regression

A more robust approach is to estimate the risk model nonparametrically. Again the risk is assumed to be monotone increasing in Y. We compute the nonparametric maximum likelihood estimator for the risk function subject to monotonicity using isotonic regression [24]. A heuristic explanation of the algorithm in this particular circumstance was given by Lloyd [25]. Specifically, marker data $\{y_1, \ldots, y_n\}$ are arranged in increasing order, followed by repetitive blocking and pooling of adjacent blocks until the sample proportion of cases within each block is non-decreasing. Finally, we calculate $\hat{P}(D = 1|Y = y_j, S)$, the proportion of diseased subjects within the block containing y_j. Case-control sampling again requires an adjustment to estimate the population risk function. Specifically we use the relationship

$$\frac{\widehat{P}(D=1|Y)}{\widehat{P}(D=0|Y)} = \frac{\widehat{P}(D=1|Y,S)}{\widehat{P}(D=0|Y,S)} \frac{n_{\bar{D}}}{n_D} \frac{\widehat{\rho}}{1-\widehat{\rho}}.$$

3. Estimation of the Marker Distribution and the Predictiveness Curve

In a case-control study, F cannot be estimated directly but can be estimated as a weighted average of the distributions of Y in the case and control subpopulations. Specifically, since $F = \rho F_D + (1-\rho)F_{\bar{D}}$, we estimate ρ, F_D and $F_{\bar{D}}$ and substitute the estimates to obtain the estimate of F. Two approaches to estimating F_D and $F_{\bar{D}}$ are possible under the parametric and nonparametric risk modeling assumptions.

3.1 The Semiparametric Estimators of the Predictiveness Curve

3.1.1 The Semiparametric "Empirical" Estimator—A natural strategy to estimate $F_{\bar{D}}$ and F_D is to use the corresponding empirical estimators which we denote by $\tilde{F}_{\bar{D}}$ and $\tilde{F}_D$. Estimating F with $\tilde{F} = \widehat{\rho}\tilde{F}_D + (1-\widehat{\rho})\tilde{F}_{\bar{D}}$, the resulting semiparametric "empirical" estimators of $R(v)$ and $R^{-1}(p)$ are

$$\begin{aligned} \tilde{R}(v) &= G\left\{\widehat{\theta}, \tilde{F}^{-1}(v)\right\} \quad \text{for} \quad v \in (0,1), \\ \tilde{R}^{-1}(p) &= \tilde{F}\left\{G^{-1}\left(\widehat{\theta}, p\right)\right\} \quad \text{for} \quad p \in \{R(v) : v \in (0,1)\}, \end{aligned}$$

where $G^{-1}(\theta, p) = \inf\{y : G(\theta, y) \geq p\}$.

3.1.2 The Semiparametric "Maximum Likelihood" Estimator—Let f_D and $f_{\bar{D}}$ denote density functions of the marker Y in the case and control populations respectively. Observe that the risk model (2.1) implies an exponential tilt relationship between marker densities among cases and controls

$$\mathcal{L}\mathrm{R}(Y) = f_D(Y)/f_{\bar{D}}(Y) = \exp\left\{\theta_0 + \log\left(\frac{1-\rho}{\rho}\right) + \eta(\theta_1, Y)\right\}, \tag{3.1}$$

where $\mathcal{LR}(Y)$ is called the likelihood ratio of Y. This relationship is not exploited when F_D and $F_{\bar{D}}$ are estimated empirically [26,11]. We have shown that by employing an empirical likelihood approach [27,28,29], the maximum likelihood estimators for $F_{\bar{D}}$ and F_D are

$$\begin{aligned}\widehat{F}_{\bar{D}}(y) &= \frac{1}{n_{\bar{D}}}\sum_{i=1}^{n}\frac{I(Y_i\le y)}{1+\frac{n_D}{n_{\bar{D}}}\exp\left\{\widehat{\theta}_0+\log\left(\frac{1-\widehat{\rho}}{\widehat{\rho}}\right)+\eta\left(\widehat{\theta}_1,Y_i\right)\right\}}=\frac{1}{n}\sum_{i=1}^{n}\frac{I(Y_i\le y)}{\frac{n_{\bar{D}}}{n}+\frac{n_D}{n}\widehat{\mathcal{LR}}(Y_i)},\\ \widehat{F}_D(y) &= \frac{1}{n_{\bar{D}}}\sum_{i=1}^{n}\frac{\exp\left\{\widehat{\theta}_0+\log\left(\frac{1-\widehat{\rho}}{\widehat{\rho}}\right)+\eta\left(\widehat{\theta}_1,Y_i\right)\right\}I(Y_i\le y)}{1+\frac{n_D}{n_{\bar{D}}}\exp\left\{\widehat{\theta}_0+\log\left(\frac{1-\widehat{\rho}}{\widehat{\rho}}\right)+\eta\left(\widehat{\theta}_1,Y_i\right)\right\}}=\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{\mathcal{LR}}(Y_i)I(Y_i\le y)}{\frac{n_{\bar{D}}}{n}+\frac{n_D}{n}\widehat{\mathcal{LR}}(Y_i)},\end{aligned}$$

where $\widehat{\theta}_0$ is the logistic regression intercept adjusted by $\log\left\{\widehat{\rho}/(1-\widehat{\rho})\,n_{\bar{D}}/n_D\right\}$, and $\widehat{\mathcal{LR}}$ is the maximum likelihood estimator of $\mathcal{LR}$ [11]. We use these estimators to compute $\widehat{F}=(1-\widehat{\rho})\,\widehat{F}_{\bar{D}}+\widehat{\rho}\widehat{F}_D$, and then plug $\widehat{\theta}$ and $\widehat{F}$ into G to get the semiparametric maximum likelihood estimators of $R(v)$ and $R^{-1}(p)$:

$$\begin{aligned}\widehat{R}(v) &= G\left\{\widehat{\theta},\widehat{F}^{-1}(v)\right\} \quad \text{for} \quad v\in(0,1),\\ \widehat{R}^{-1}(p) &= \widehat{F}\left\{G^{-1}\left(\widehat{\theta},p\right)\right\} \quad \text{for} \quad p\in\{R(v):v\in(0,1)\}.\end{aligned}$$

Note that the semiparametric estimators developed here for case-control studies generalizes the semiparametric estimator developed for cohort studies in Huang *and others* [7]. That is, when plugging in $\widehat{\rho}=n_D/n$ from a cohort study, both the semiparametric maximum likelihood estimator and the semiparametric "empirical" estimator of the predictiveness curve equal to the cohort version proposed earlier [11].

Asymptotic distribution theory for the two estimators can be found in Huang and Pepe [11]. As an example, consider an ordinary logistic risk model: logit $\{G(\theta,Y)\}=\theta_0+\theta_1^T r(Y)$, where $r(Y)$ is some monotone increasing function of Y. Suppose $\widehat{\rho}$ is estimated from a cohort independent of the case-control sample, or the parent cohort within which the case-control sample is nested, with the size of the cohort λ times the size of the case-control sample. Then we have

$$\begin{aligned}n\mathrm{var}\left\{\widehat{R}^{-1}(p)\right\} &\simeq \left\{F_D(p)-F_{\bar{D}}(p)\right\}^2\rho(1-\rho)/\lambda+V_{M1}\\ &+ \left\{\frac{\partial R^{-1}(p)}{\partial\theta}\right\}^T\left\{\begin{pmatrix}\frac{1}{\lambda\rho(1-\rho)} & 0\\ 0 & 0\end{pmatrix}+V_{M2}\right\}\left\{\frac{\partial R^{-1}(p)}{\partial\theta}\right\}\\ &+ 2\left\{\frac{\partial R^{-1}(p)}{\partial\theta}\right\}^T\left\{\frac{F_D(t)-F_{\bar{D}}(t)}{\lambda}+V_{M3}\right\}.\end{aligned} \tag{3.2}$$

Analytic forms for V_{M1}, V_{M2}, V_{M3} are provided in Appendix A. Similarly, we have

$$\begin{aligned}n\mathrm{var}\left\{\tilde{R}^{-1}(p)\right\} &\simeq \left\{F_D(p)-F_{\bar{D}}(p)\right\}^2\rho(1-\rho)/\lambda+V_{E1}\\ &+ \left\{\frac{\partial R^{-1}(p)}{\partial\theta}\right\}^T\left\{\begin{pmatrix}\frac{1}{\lambda\rho(1-\rho)} & 0\\ 0 & 0\end{pmatrix}+V_{E2}\right\}\left\{\frac{\partial R^{-1}(p)}{\partial\theta}\right\}\\ &+ 2\left\{\frac{\partial R^{-1}(p)}{\partial\theta}\right\}^T\left\{\frac{F_D(t)-F_{\bar{D}}(t)}{\lambda}+V_{E3}\right\},\end{aligned} \tag{3.3}$$

where analytic forms for V_{E1}, V_{E2}, V_{E3} are provided in Appendix A. Moreover, for $v = R^{-1}(p)$, var $\{\tilde{R}(v)\} \simeq \{\partial R(v)/\partial v\}^2$ var $\{\tilde{R}^{-1}(p)\}$ and var $\{\hat{R}(v)\} \simeq \{\partial R(v)/\partial v\}^2$ var $\{\hat{R}^{-1}(p)\}$. When ρ is fixed, essentially we have $\lambda \to \infty$, thus terms involving $1/\lambda$ vanish in (3.2) and (3.3).

3.2 The Nonparametric Estimator

Similar to the semiparametric approach, we can estimate F_D and $F_{\bar{D}}$ empirically with $\tilde{F}_{\bar{D}}$ and $\tilde{F}_D$ yielding $\tilde{F} = \widehat{\rho}\tilde{F}_D + (1 - \widehat{\rho})\tilde{F}_{\bar{D}}$. Substituting the non-parametric risk estimator and $\tilde{F}$ we derive the nonparametric "empirical" estimators of $R(v)$ and $R^{-1}(p)$as

$$\begin{aligned} \tilde{R}(v) &= \widehat{P}\left\{D=1|Y=\tilde{F}^{-1}(v)\right\} \qquad v \in (0,1), \\ \tilde{R}^{-1}(p) &= \tilde{F}\left[\sup\left\{y:\widehat{P}(D=1|Y=y) \le p\right\}\right] \qquad p \in \{R(v): v \in (0,1)\}. \end{aligned}$$

Alternatively, we can incorporate the estimated risk function into estimation of the marker distribution, as was done for the semiparametric procedure. Lloyd [25] showed that maximizing the joint likelihood of D and Y can be achieved by first obtaining $\hat{P}(D = 1|Y, S)$, and then estimating $f_{\bar{D}}$ and f_D based on the relationship

$$\mathcal{L}\mathrm{R}(Y) = \frac{f_D(Y)}{f_{\bar{D}}(Y)} = \frac{P(D=1|Y,S)}{P(D=0|Y,S)}\frac{n_{\bar{D}}}{n_D} \propto \frac{P(D=1|Y,S)}{P(D=0|Y,S)}.$$

In particular, let $\hat{w}(Y) = \hat{P}(D = 1|Y, S)/\hat{P}(D = 0|Y, S)$ and let κ denote $\{k : \hat{w}(Y_k) = \infty\}$. He demonstrated that by maximizing $\mathcal{L}\left(F_{\bar{D}}, F_D\right) = \prod_{i=1}^{n_D} f_{\bar{D}}\left(Y_{\bar{D}i}\right) \prod_{j=1}^{n_D} f_D\left(Y_{Dj}\right) = \prod_{i=1}^{n} f_{\bar{D}}(Y_i) \prod_{j=1}^{n_D} \widehat{w}\left(Y_{Dj}\right)/\mu$ with μ a normalizing factor, the estimators of $f_{\bar{D}}$ and f_D are

$$\widehat{f}_{\bar{D}}(Y_k) = \begin{cases} \widehat{\mu}/\left(n_D\widehat{w}(Y_k) + n_{\bar{D}}\widehat{\mu}\right) & k \notin \kappa \\ 0 & k \in \kappa \end{cases}, \quad \widehat{f}_D(Y_k) = \begin{cases} \widehat{w}(Y_k)\widehat{f}_{\bar{D}}(Y_k)/\widehat{\mu} & k \notin \kappa \\ 1/n_D & k \in \kappa \end{cases} \tag{3.4}$$

in the absence of ties. He also suggested that $\widehat{\mu}$ could be found by solving

$$\sum_{k \in \kappa} \frac{\mu}{n_D\widehat{w}(Y_k) + n_{\bar{D}}\mu} = 1, \tag{3.5}$$

which is monotone increasing in μ.

The following new result, proved in Appendix B1, shows that when $P(D = 1|Y, S)$ is estimated using isotonic regression, $\widehat{\mu}$ can be written down explicitly as a function of $n_{\bar{D}}$ and n_D.

Theorem 1 When $P(D = 1|Y, S)$ is estimated using isotonic regression, $\widehat{\mu} = n_D/n_{\bar{D}}$.

Plugging $\widehat{\mu}$ into (3.4), we have

$$\widehat{f}_{\bar{D}}(Y_k)=\begin{cases}\frac{1}{n_{\bar{D}}\{\widehat{w}(Y_k)+1\}} & k\notin\kappa\\ 0 & k\in\kappa\end{cases},\quad \widehat{f}_{D}(Y_k)=\begin{cases}\frac{\widehat{w}(Y_k)}{n_D\{\widehat{w}(Y_k)+1\}} & k\notin\kappa\\ 1/n_D & k\in\kappa\end{cases}.$$

Estimating F with $\widehat{F}=\widehat{\rho}\widehat{F}_D+(1-\widehat{\rho})\widehat{F}_{\bar{D}}$, where $\hat{F}_D$ and $\hat{F}_{\bar{D}}$ denote the corresponding cumulative distribution functions, the nonparametric maximum likelihood estimators of $R(v)$ and $R^{-1}(p)$ are

$$\begin{aligned}\widehat{R}(v) &= \widehat{P}\left\{D=1|Y=\widehat{F}^{-1}(v)\right\} \quad \text{for} \quad v\in(0,1),\\ \widehat{R}^{-1}(p) &= \widehat{F}\left[\sup\left\{y:\widehat{P}(D=1|Y=y)\le p\right\}\right] \quad \text{for} \quad p\in\{R(v):v\in(0,1)\}.\end{aligned}$$

Interestingly, we have found that even if the nonparametric "empirical" and maximum likelihood procedures described above lead to different estimators of the marker distribution F, the corresponding predictiveness curve estimators are the same (Theorem 2). This fact is not true for the semiparametric estimators. A proof can be found in Appendix B2.

Theorem 2 When the risk model is estimated nonparametrically with isotonic regression, $\hat{R}(v) = \tilde{R}(v)$ and $\hat{R}^{-1}(p) = \tilde{R}^{-1}(p)$.

3.3 Area under the Predictiveness Curve

The area under the true predictiveness curve, $\int_0^1 R(v)\,dv$, is equal to ρ [7]. This facilitates visual comparisons of predictiveness curves for two different risk models because, in a sense, it maintains them both on exactly the same scale. The steepness of curves can be compared more easily when both integrate to ρ. An analogous result holds for the nonparametric and semiparametric maximum likelihood estimators (Theorem 3) but not for the semiparametric "empirical" estimator.

Theorem 3 Let $\hat{R}(v)$ be the nonparametric or semiparametric maximum likelihood estimator of $R(v)$ for $v\in(0,1)$ using the prevalence estimator $\widehat{\rho}$. Then $\int_0^1\widehat{R}(v)\,dv$, the area under the predictiveness curve estimate equals to $\widehat{\rho}$.

Proof of Theorem 3 is presented in Appendix B3. An implication of Theorem 3 is that for the nonparametric and semiparametric maximum likelihood estimators, the two areas sandwiched between the curve and the horizontal line are equal. To see this, let $v^*=\inf\left\{v:\widehat{R}(v)\ge\widehat{\rho}\right\}$, then the area below the horizontal line at $\widehat{\rho}$ and above the estimated predictiveness curve is equal to $\int_0^{v^*}\left\{\widehat{\rho}-\widehat{R}(v)\right\}dv$, while the area above the horizontal line at $\widehat{\rho}$ and below the estimated predictiveness curve is equal to $\int_{v^*}^1\left\{\widehat{R}(v)-\widehat{\rho}\right\}dv$. According to Theorem 3,

$$\int_0^1\widehat{R}(v)\,dv=\widehat{\rho}\Rightarrow\int_0^1\left\{\widehat{R}(v)-\widehat{\rho}\right\}dv=0\Rightarrow\int_0^{v^*}\left\{\widehat{R}(v)-\widehat{\rho}\right\}dv+\int_{v^*}^1\left\{\widehat{R}(v)-\widehat{\rho}\right\}dv=0\Rightarrow\int_0^{v^*}\left\{\widehat{\rho}-\widehat{R}(v)\right\}dv=\int_{v^*}^1\left\{\widehat{R}(v)-\widehat{\rho}\right\}dv.$$

4. Simulation Studies

We conducted simulation studies in two settings to evaluate the performances of the proposed estimators. In each setting, data were generated to mimic a two-phase study. In the

first phase, a random cohort sample is obtained and the disease status of every subject is determined. In the second phase, cases and controls were selected independently from the parent cohort and biomarker data was ascertained. The size of the cohort is chosen to be five times that of the nested case-control sample.

4.1 Simulation Setting 1

In the first simulation setting, a binary outcome status was generated with $\rho = 0.2$ from a cohort of size $5n$ and marker data were generated according to $Y_{\bar{D}} \sim N(0, 1)$ and $Y_D \sim N(\mu_D, 1)$, for equal numbers of cases and controls, $n_D = n_{\bar{D}} = n/2$. The resulting risk function follows a linear logistic model. We explored sample sizes n ranging from 100 to 2,000. For each scenario, 5,000 Monte-Carlo simulations were conducted.

The semiparametric (based on the linear logistic model) and nonparametric estimators of the predictiveness curve were estimated using $\widehat{\rho}$ obtained from the cohort. Variance estimates for the semiparametric estimators were calculated using analytic formulae from the asymptotic theory which incorporates variability in $\widehat{\rho}$ (as provided in Appendix B1). Bootstrapping was also performed by separately resampling cases and controls for Y and resampling D from the parent cohort. Results pertaining to the choice $\mu_D = 1$ are presented in Tables 1 - 4 for $v = 0.1, 0.3, 0.5, 0.7$, and 0.9 and for the corresponding values of p, $p = R(v)$.

First we consider the performance of the semiparametric estimators for $R(v)$ and $R^{-1}(p)$. We see that they have minimal bias for sample sizes as small as 100 (Table 1). Variance estimators that are based on analytic formulas from asymptotic theory agree well with the empirical variance from simulations when $n_D + n_{\bar{D}} \geq 500$. This was also true for the bootstrap variance (results not shown). Coverage of the 95% Wald confidence intervals using asymptotic or bootstrap variance estimates are fairly close to the nominal level, except for a little undercoverage when $n_D + n_{\bar{D}} \leq 200$ (Table 2). The intervals shown in Table 2 assumed that the logit transform of the estimator was normally distributed and had better coverage than symmetric intervals for the untransformed estimators.

Results are also shown in Table 3 for confidence intervals employing percentiles of the bootstrap distribution. We found that these confidence intervals performed best overall. Moreover, the corresponding lower and upper confidence limits are monotone increasing in v. This is a desirable property because, by definition, the predictiveness curve itself is monotone increasing. Having lower and upper pointwise confidence limit curves that are monotone increasing is consistent with monotonicity of the predictiveness curve. To see that the pointwise confidence limits are increasing in v, let $\hat{R}_b(v)$ be the estimate of $R(v)$ based on the b^{th} bootstrap sample. We have $\hat{R}_b(v_1) \leq \hat{R}_b(v_2)$ for $v_1 \leq v_2$ according to our estimation methods. As a result, the α^{th} percentile of $\hat{R}_b(v_1)$ is always smaller than or equal to the α^{th} percentile of $\hat{R}_b(v_2)$ among the same set of bootstrap samples. In our simulations, we chose $\alpha = 0.025$ and $\alpha = 0.975$.

The nonparametric estimator of the predictiveness curve performed poorly relative to the semi-parametric estimators in this simulation setting. When sample sizes are smaller than 500, biases in estimates of $R(v)$ and $R^{-1}(p)$ are substantial, and confidence intervals suffer from undercoverage or overcoverage in many settings (Tables 2,3). Their efficiency is dramatically worse than the efficiencies of the semiparametric estimators (Table 4). This is especially true in large samples.

The two semiparametric estimators have similar performances in the simulations. Of note, the semiparametric "empirical" estimator is fairly efficient relative to the semiparametric maximum likelihood estimator (Table 4). Based on these limited simulations we recommend

use of either of the semiparametric estimators in practice with confidence intervals constructed from percentiles of the bootstrap distribution when resampling is feasible, or from the logit transform with corresponding analytic variance formulas when bootstrapping is not feasible.

4.2 Simulation Setting 2

We investigated another simulation setting where marker Y follows a standard normal distribution, and the risk quantile follows a piece-wise linear form with cutpoint at the quintiles. Specifically, $R(v)$ takes value 0, 0.1, 0.16, 0.2, 0.24, 0.3 at cutpoints $v = 0, 0.2, 0.4, 0.6, 0.8, 1$, and is linear in between. As before, we first simulate a cohort of size $5n$, and then randomly sample $n_D = n/2$ cases and $n_{\bar{D}} = n/2$ controls from the cohort. The semiparametric estimators again are obtained assuming a linear logistic model. Tables 5 and 6 present bias, efficiency (in terms of mean squared error), and coverage of the 95% percentile bootstrap confidence intervals for the semiparametric and nonparametric estimators, for case-control sample sizes varying from 500 to 2,000. The semiparametric estimators that assume an incorrect working model, have poorer performance compared to that in the first simulation setting. For estimation of both $R(v)$ and $R^{-1}(p)$, they have nonignorable biases that do not dissapate as sample size increases. As a result, coverages of their confidence intervals are often seriously below the nominal level, especially in large sample size. The performance of the nonparametric estimator, nevertheless, is fairly consistent with that in the first simulation setting. Its bias is much smaller compared to that of the semiparametric estimators and decreases as sample size increases. Consequently, for certain quantiles, the mean squared error of the nonparametric estimator is smaller than that of the semiparametric estimators. In general, for a sample size greater than 500, the confidence interval constructed from the nonparametric estimators maintain coverage close to the nominal level.

Because of its robustness, the nonparametric estimator might be preferred in large studies where bias rather than precision is the major concern. On the other hand, in practice it is important to make the semiparametric model flexible to ensure a good fit, in light of its sensitivity to the risk model assumption. Comparing the nonparametric estimator with the semiparametric estimator provides an avenue for model checking. Later in this paper we propose a goodness-of-fit test to assess calibration formally.

5. Illustration

Levels of prostate specific antigen (PSA) and recent increases in levels of PSA (PSA velocity) are markers for prostate cancer. We evaluate them as predictors of the risk that a man will be diagnosed with prostate cancer if biopsied. These markers should only be used in decisions to take a biopsy if they are sufficiently informative of this risk. Data for evaluating these markers come from the Prostate Cancer Prevention Trial, a randomized prospective study with 7 years of follow-up [30]. Subjects were at least 55 years old, had serum PSA value less than 3.0 ng/ml at baseline and were scheduled for annual blood draws to measure serum PSA. Almost all subjects had a prostate biopsy taken at the end of study. We analyze data for the 5519 men on the placebo arm of the trial that had a prostate biopsy, a PSA measure during the year prior to biopsy and at least 2 PSA values from the 3 years prior to biopsy to calculate PSA velocity. The prevalence of prostate cancer in the cohort is $\widehat{\rho}$=21.9%. We selected 250 cases and 250 controls at random from the cohort to simulate a nested case-control study. Thus the data for analysis consist of the prevalence estimate from the 5519 men in the parent cohort and PSA and PSA velocity for subjects in the case-control subset.

To implement the semiparametric methods for estimating predictiveness curves, for each marker a logistic regression risk model was employed using a Box-Cox transformation of

the marker. The two semiparametric estimators of the predictiveness curves are very similar to each other for both PSA and PSA velocity and so only the semiparametric maximum likelihood estimators are displayed in Figure 1. Also displayed in Figure 1 are the nonparametric predictiveness curve estimates. Observe that the semiparametric curves are much smoother than the nonparametric ones, but agree with them, suggesting a good-fit for the semiparametric models. Overall, PSA has a steeper predictiveness curve, indicating that it is a better marker for predicting risk of prostate cancer than PSA velocity.

For the semiparametric estimators, the asymptotic and bootstrap variance estimates for $\hat{R}(v)$ and $\hat{R^{-1}}(p)$ are similar in magnitude (results not shown). Moreover, the Wald confidence intervals for $R(v)$ and $R^{-1}(p)$ are close to those based on percentiles of the bootstrap distributions. Here we present only the latter. The pointwise 95% percentile bootstrap confidence intervals for $R(v)$ constructed from the semiparametric maximum likelihood estimators are displayed in Figure 2(a)(b). They are much narrower in comparison to those constructed from the nonparametric estimators.

We next compare the predictive capacities of the two markers in terms of the 10^{th} and 90^{th} percentiles of their risk distributions and sizes of risk strata corresponding to a low risk threshold of 10% and a high risk threshold of 30% (Table 7 (a)). Results are presented for both the semiparametric maximum likelihood estimators and for the nonparametric estimators. *P*-values employ Wald tests based on differences in $R(v)$ and $R^{-1}(p)$ with variances estimated with the bootstrap. Using the semiparametric methods, PSA appears to have a better capacity to predict high risk of prostate cancer than does PSA velocity given that it has a larger value for $R(0.9)$ as well as better capacity to predict low risk given that it has a smaller value for $R(0.1)$. In addition, PSA categorizes more people into the low and high risk ranges as can be seen from semiparametric estimates of $R^{-1}(0.1)$ and $1 - R^{-1}(0.3)$. In contrast, these comparisons are not significant based on the nonparametric methods due to their large sampling variabilities.

6. Extending Semiparametric Estimation to Multiple Markers

The semiparametric estimators can be extended naturally to accommodate multiple predictors or to settings where the monotone increasing risk assumption is not true.

6.1 Inference

We present the generalized estimators here as well as their asymptotic distribution theory. In practice, since estimation of the asymptotic variance involves both numerical differentiation and nonparametric density estimation, we rely on resampling techniques rather than on asymptotic theory for inference.

Let Y be a vector of predictors that may include different functional forms of a single predictor (e.g. a set of spline basis functions) as well as interactions among predictors. Let F_R, F_{DR}, $F_{\bar{D}R}$ indicate the cumulative distribution functions for $Risk(Y)$ in the general, case and control populations respectively. As before, we calculate $\widehat{Risk}(Y_i)$ as the predicted risk for subject i based on fitting a standard logistic regression model to case-control data with offset $\log\left\{(1-\hat{\rho})/\hat{\rho}n_D/n_{\bar{D}}\right\}$. To estimate F_R we write $F_R = \rho F_{DR} + (1-\rho)F_{\bar{D}R}$ and substitute estimates for each component. The components F_{DR} and $F_{\bar{D}R}$ can be estimated "empirically" using $\tilde{F}_{DR}(p) = \sum_{i=1}^{n_D} I\left\{\widehat{Risk}(Y_{Di}) \le p\right\}/n_D$ and $\tilde{F}_{\bar{D}R}(p) = \sum_{i=1}^{n_D} I\left\{\widehat{Risk}\left(Y_{\bar{D}i}\right) \le p\right\}/n_{\bar{D}}$.

A more efficient approach is to use the semiparametric maximum likelihood estimates that are derived using arguments similar to those provided in Huang and Pepe [11] for the single marker setting:

$$\widehat{F}_{DR}(p) = \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{\mathcal{L}\mathrm{R}}_R\left\{\widehat{Risk}(Y_i)\right\} I\left\{\widehat{Risk}(Y_i)\le p\right\}}{\frac{n_{\bar{D}}}{n}+\frac{n_D}{n}\widehat{\mathcal{L}\mathrm{R}}_R\left\{\widehat{Risk}(Y_i)\right\}},$$

$$\widehat{F}_{\bar{D}R}(p) = \frac{1}{n}\sum_{i=1}^{n}\frac{I\left\{\widehat{Risk}(Y_i)\le p\right\}}{\frac{n_{\bar{D}}}{n}+\frac{n_D}{n}\widehat{\mathcal{L}\mathrm{R}}_R\left\{\widehat{Risk}(Y_i)\right\}},$$

where $\widehat{\mathcal{L}\mathrm{R}}_R\left\{\widehat{Risk}(Y_i)\right\} = \widehat{Risk}(Y_i) / \left\{1-\widehat{Risk}(Y_i)\right\} \times (1-\widehat{\rho})/\widehat{\rho}$.

We write $\widehat{R}(v) = \widehat{F}_R^{-1}(v)$ and $\hat{R}^{-1}(p) = \hat{F}_R(p)$ for the semiparametric maximum likelihood estimators of the predictiveness curve and $\tilde{R}(v) = \tilde{F}_R(p)$ and $\tilde{R}^{-1}(p) = \tilde{F}_R(p)$ for the semiparametric "empirical" estimators. The following results are proved in Appendix B4. Here variability of $\widehat{\rho}$ is taken into account in calculating asymptotic variance of the predictiveness curve estimators, with details provided in Appendix B4.

Theorem 4 As $n \to \infty$,

i. $\sqrt{n}\left\{\widehat{R}^{-1}(p) - R^{-1}(p)\right\}$ converges to a normal random variable with mean zero and variance

$$\begin{aligned}\Sigma_{2M.R}(p) &= \mathrm{var}\left[\sqrt{n}\left\{Q_M(p) - R^{-1}(p)\right\}\right] \\ &+ \left\{\frac{\partial R^{-1}(p)}{\partial\theta}\right\}^T \mathrm{var}\left\{\sqrt{n}\left(\widehat{\theta}-\theta\right)\right\}\left\{\frac{\partial R^{-1}(p)}{\partial\theta}\right\} \\ &+ 2\left\{\frac{\partial R^{-1}(p)}{\partial\theta}\right\}^T \mathrm{cov}\left[\sqrt{n}\left(\widehat{\theta}-\theta\right), \sqrt{n}\left\{Q_M(p) - R^{-1}(p)\right\}\right],\end{aligned}$$

ii. $\sqrt{n}\left\{\widehat{R}(v) - R(v)\right\}$ converges to a normal random variable with mean zero and variance $\Sigma_{1M.R}(v) = \{\partial R(v)/\partial v\}^2 \Sigma_{2M.R}\{R(v)\}$, where

$$Q_M(p) = \widehat{\rho}\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{\mathcal{L}\mathrm{R}}_R\left\{\widehat{Risk}(Y_i)\right\} I\{Risk(Y_i)\le p\}}{\frac{n_{\bar{D}}}{n}+\frac{n_D}{n}\widehat{\mathcal{L}\mathrm{R}}_R\left\{\widehat{Risk}(Y_i)\right\}} + (1-\widehat{\rho})\frac{1}{n}\sum_{i=1}^{n}\frac{I\{Risk(Y_i)\le p\}}{\frac{n_{\bar{D}}}{n}+\frac{n_D}{n}\widehat{\mathcal{L}\mathrm{R}}_R\left\{\widehat{Risk}(Y_i)\right\}}.$$

iii. $\sqrt{n}\left\{\tilde{R}^{-1}(p) - R^{-1}(p)\right\}$ converges to a normal random variable with mean zero and variance

$$\begin{aligned}\Sigma_{2E.R}(p) &= \mathrm{var}\left[\sqrt{n}\left\{Q_E(p) - R^{-1}(p)\right\}\right] \\ &+ \left\{\frac{\partial R^{-1}(p)}{\partial\theta}\right\}^T \mathrm{var}\left\{\sqrt{n}\left(\widehat{\theta}-\theta\right)\right\}\left\{\frac{\partial R^{-1}(p)}{\partial\theta}\right\} \\ &+ 2\left\{\frac{\partial R^{-1}(p)}{\partial\theta}\right\}^T \mathrm{cov}\left[\sqrt{n}\left(\widehat{\theta}-\theta\right), \sqrt{n}\left\{Q_E(p) - R^{-1}(p)\right\}\right],\end{aligned}$$

iv. $\sqrt{n}\left\{\tilde{R}(v) - R(v)\right\}$ converges to a normal random variable with mean zero and variance $\Sigma_{1E.R}(v) = \{\partial R(v)/\partial v\}^2 \Sigma_{2E.R}\{R(v)\}$, where

$$Q_E(p) = \widehat{\rho}\frac{1}{n_D}\sum_{i=1}^{n_D} I\{Risk(Y_{Di}) \le p\} + (1-\widehat{\rho})\frac{1}{n_{\bar{D}}}\sum_{i=1}^{n_{\bar{D}}} I\left\{Risk\left(Y_{\bar{D}i}\right) \le p\right\}.$$

6.2 Illustration

We illustrate using the data described in Section 5. We compare the logistic risk model based on PSA alone with a more comprehensive risk model that combines PSA and other risk factors, namely family history of prostate cancer, digital rectal exam, and previous negative biopsies. All of these factors are highly statistically significant factors in determining risk [30].

A fundamental preliminary step in assessing the value of a risk model concerns calibration or model fit. The predictiveness curve can be helpful in this assessment [6]. We now illustrate how the assessment can be made with data from a nested case-control study by application to the model that includes other risk factors in addition to PSA. We let $\widehat{Risk}(Y)$ be risk estimates from the model employing PSA and other risk factors and partition the observations into deciles of the distribution for $\widehat{Risk}(Y)$. For $k \in \{1, \ldots, K=10\}$, we estimate $\hat{P}(D = 1|k, S)$ as the observed proportion of cases within the k^{th} group. The population risk within the k^{th} group, $P(D = 1|k)$ is then estimated according to

$$\frac{\widehat{P}(D=1|k)}{1-\widehat{P}(D=1|k)} = \frac{\widehat{P}(D=1|k,S)}{1-\widehat{P}(D=1|k,S)}\frac{n_{\bar{D}}}{n_D}\frac{1-\widehat{\rho}}{\widehat{\rho}}.$$

At the midpoints of the k^{th} group, a visual comparison can be made between these points and the predictiveness curve by superimposing the value of $\hat{P}(D = 1|k)$ on the predictiveness curve plot. This comparison of observed risk and average risk within deciles of modeled risk provides a graphical display of calibration as suggested previously [6] but now generalized to case-control data. According to Figure 3(a), there does not to appear be serious lack of fit of the risk model.

Let $\widehat{Risk}^S$ be the risk estimate from the logistic regression applied to the case-control sample without correcting for the intercept term. We propose a Hosmer-Lemeshow measure of calibration to accompany the plot:

$$T = \sum_{k=1}^{K} \frac{\left\{\widehat{P}(D=1|k) - \widehat{R}_k\right\}^2}{\widehat{R}_k^2\left(1-\widehat{R}_k\right)^2 / \left\{n_k \widehat{R}_k^S\left(1-\widehat{R}_k^S\right)\right\}}, \tag{6.1}$$

where $\widehat{R}_k^S$ is the average of $\widehat{Risk}^S$ within the k^{th} group, and $\hat{R}_k$ is a population version of it corrected for the biased sampling

$$\widehat{R}_k = \left\{\widehat{R}_k^S \frac{n_{\bar{D}}}{n_D}\frac{\widehat{\rho}}{1-\widehat{\rho}}\right\} / \left\{1 + \widehat{R}_k^S \frac{n_{\bar{D}}}{n_D}\frac{\widehat{\rho}}{1-\widehat{\rho}}\right\}.$$

Within each discretized risk group, the term in the numerator of (6.1) is the squared difference between the 'observed' and 'expected' disease proportion in the population, while the term in the denominator is the estimated variance of this difference. Comparing this measure T with a χ^2_{K-2} distribution, we obtain a p-value for the test of calibration. The calculation of $\hat{R}_k$ and justification for this test are outlined in Appendix C. Our measure is a modification of the established Hosmer-Lemeshow test for case-control study [32], which compares observed and expected disease proportion in the case-control sample. The test without modification yields a valid test with case-control data. But the modified test is more tightly linked to the predictiveness curve and the measure of calibration upon which it is based does not vary with the case-control design. In our example, the modified test yields a p-value of 0.170 for the predictiveness curve employing PSA and other risk factors, suggesting a reasonable representation of the risk distribution in the population provided by the predictiveness curve.

Figure 3(a) displays the semiparametric maximum likelihood estimators of the predictiveness curves for PSA alone and for the model that includes other factors (the semiparametric "empirical" estimators are similar). The two risk models have very similar predictiveness curves. Confidence intervals for $R(v)$ constructed with the semiparametric maximum likelihood estimators are presented in Figure 3(b). Sampling variability of estimates derived from the two risk models appears to be similar in magnitude.

Detailed results comparing the predictiveness curves of the two models are shown in Table 7(b). Briefly, the percentages of people classified into the low, high, or equivocal risk ranges are not significantly different between the two models, nor are the 10^{th} and 90^{th} percentiles of risk significantly different. Thus including other factors in the model in addition to PSA does not lead to a significant improvement in risk stratification even when these factors are all statistically significant in the multivariate logistic regression model. It reinforces our earlier argument that the risk model by itself is not enough to characterize the population performance of a risk prediction model.

An important issue pertaining to models with multiple predictors is over-fitting when the number of predictors gets large relative to the sample size. To account for potential over-fitting, we implemented 10-fold cross-validation to compare the predictiveness of the two models. Again, including other factors beside PSA in the risk model has trivial influence on risk stratification.

7. Discussion

It has been argued in both the applied [8] and biostatistical literature [2,6] that displaying the population distribution of risk is useful for evaluating the potential impact of a risk model for risk stratifying the population. The key ideas of risk stratification tables, introduced by Cook *and others* [33] and Cook [34], are closely related. In particular, the margins of the two-way risk stratification table show the population distribution of risk achieved by the two models, albeit using discrete risk categories. Janes *and others* [35] show that the key information pertinent to comparing models is contained in the margins rather than in the cells of the table. Predictiveness curves provide more complete descriptions of the marginal risk distributions since they show risk distributions over a continuum of risk thresholds that could be used to define risk categories rather than only at a few pre-specified risk thresholds.

Methods for estimating predictiveness curves from cohort studies were developed previously. However, case-control designs are often preferred in biomarker development [9] and the goal of the current paper is to develop estimation procedures for use with case-control data. Here we discussed semiparametric methods that rely on a logistic regression

form for the risk and a non-parametric method that relies on isotonic regression for estimating the risk. Another approach developed by Huang and Pepe [36] is based on the relationship between the predictiveness curve and the ROC curve. Here, we found that the nonparametric method is inefficient compared with the semiparametric methods and that valid inference requires large sample sizes. We recommend the semiparametric methods for use in practice because (i) simulations indicate that inferential procedures are adequate with realistic sample sizes, (ii) they accommodate risk models with multiple predictors, and (iii) they can be made flexible by employing flexible forms for the predictors in the logistic regression model. The last point is important to ensure good model fit by the semiparametric model. The nonparametric estimator has the advantage that it is completely robust but potentially very inefficient. Therefore it can be useful in large studies where precision is not an issue and minimum bias is desired. And it can be used for comparison with the semiparametric estimator in a single marker setting to further assess its goodness-of-fit. For a general logistic risk model allowing for multiple markers, we proposed a modified Hosmer-Lemeshow test assessing calibration of the risk model. It extends the established Hosmer-Lemeshow test for case-control data by mapping the difference between observed and established disease proportion at the case-control sample level to that at the population level. As a result, performance of the test under alternative hypothesis would potentially be less sensitive to factors such as case-control ratio. Based on limited simulation studies (results not shown), this modified test appears to have power comparable to that of the standard Hosmer-Lemeshow test and is more powerful in some settings when the proportion of cases in the case-control sample is high.

Pepe *and others* [6] proposed displaying the predictiveness curve and curves displaying true and false positive rates together for maximum information. Specifically, to evaluate a risk prediction marker, one will be interested in knowing not only $1 - R^{-1}(p)$, the proportion of the population with risk above p, but also the proportion of diseased subjects correctly classified (the true positive rate $\mathrm{TPR}(p) = P\{Risk(Y) > p|D = 1\}$) and the proportion of non-diseased subjects incorrectly classified (the false positive rate $\mathrm{FPR}(p) = P\{Risk(Y) > p|D = 0\}$), according to the classification rule '$Risk(Y) > p$'. Our semiparametric and nonparametric procedures developed in this manuscript yield estimators of F_{DR}, $F_{\bar{D}R}$ and F_R as by-products. These can be directly plugged into $\mathrm{TPR}(p) = F_{DR}\left\{F_R^{-1}(p)\right\}$ and $\mathrm{FPR}(p) = F_{\bar{D}R}\left\{F_R^{-1}(p)\right\}$ to estimate these quantities. Asymptotic theory for corresponding semiparametric estimators can be developed using techniques similar to those employed for estimators of the predictiveness curve.

Finally, for interested readers, fitting of the logistic regression and isotonic regression models can be performed using standard statistical software such as R. R programs for estimating the corresponding predictiveness curves and their asymptotic variances are available from the authors upon request.

Acknowledgments

The authors are grateful for support provided by NIGMS grant GM-54438 and NCI grant CA86368.

Appendix A: Analytic Forms of the Asymptotic Variances for the Semiparametric Estimators of the Predictiveness Curve (for the Example in Section 3.1)

Let $\alpha = \theta_0 - \log\{\rho/(1-\rho)\}$, $\beta = \theta_1$, and let $\eta = n_D/n_{\bar{D}}$, we have

$$
\begin{aligned}
V_{M1} &= (1-\rho)^2(1+\eta)\left[F_{\bar{D}}\left\{G^{-1}(\theta,p)\right\} - F_{\bar{D}}\left\{G^{-1}(\theta,p)\right\}^2\right] + \rho^2\tfrac{1+\eta}{\eta}\left[F_D\left\{G^{-1}(\theta,p)\right\} - F_D\left\{G^{-1}(\theta,p)\right\}^2\right] - \left(\tfrac{1+\eta}{\eta}\right)\{\rho-(1-\rho)\eta\}^2\left\{A_0\left\{G^{-1}(\theta\right.\right.\\
V_{E1} &= (1-\rho)^2(1+\eta)\left[F_{\bar{D}}\left\{G^{-1}(\theta,p)\right\} - F_{\bar{D}}\left\{G^{-1}(\theta,p)\right\}^2\right] + \rho^2\tfrac{1+\eta}{\eta}\left[F_D\left\{G^{-1}(\theta,p)\right\} - F_D\left\{G^{-1}(\theta,p)\right\}^2\right],\\
V_{M2} &= V_{E2} = \tfrac{1+\eta}{\eta}\left\{A^{-1} - \begin{pmatrix} 1+\eta & 0 \\ 0 & 0 \end{pmatrix}\right\},\\
V_{M3} &= V_{E3} = \tfrac{1+\eta}{\eta}\left(\{\rho-\eta(1-\rho)\}A^{-1}\begin{bmatrix} A_0\left\{G^{-1}(\theta,p)\right\} \\ A_1\left\{G^{-1}(\theta,p)\right\}\end{bmatrix} - \begin{bmatrix} \rho F_D\left\{G^{-1}(\theta,p)\right\} - \eta(1-\rho)F_{\bar{D}}\left\{G^{-1}(\theta,p)\right\} \\ 0\end{bmatrix}\right),
\end{aligned}
$$

where

$$
\begin{aligned}
A_0(t) &= \int_{-\infty}^{t}\frac{\exp\{\alpha+\beta^T r(y)\}}{1+\eta\exp\{\alpha+\beta^T r(y)\}}dF_{\bar{D}}(y),\\
A_1(t) &= \int_{-\infty}^{t}\frac{r(y)\exp\{\alpha+\beta^T r(y)\}}{1+\eta\exp\{\alpha+\beta^T r(y)\}}dF_{\bar{D}}(y),\\
A_2(t) &= \int_{-\infty}^{t}\frac{r(y)r(y)^T\exp\{\alpha+\beta^T r(y)\}}{1+\eta\exp\{\alpha+\beta^T r(y)\}}dF_{\bar{D}}(y),\\
A &= \begin{pmatrix} A_0 & A_1^T \\ A_1 & A_2\end{pmatrix},
\end{aligned}
$$

with $A_0 = A_0(\infty)$, $A_1 = A_1(\infty)$, $A_2 = A_2(\infty)$.

Appendix B: Proof of Theorems

B1: Proof of Theorem 1

Suppose there are m pooled groups after isotonic regression with $\hat{w}(Y) < \infty$. In the i^{th} group, there are m_i observations, among which m_{Di} are cases. Then for subject k ($k \notin \kappa$) belonging to the i^{th} group, $\hat{w}(Y_k) = m_{Di}/(m_i - m_{Di})$.

Plugging $\widehat{\mu} = n_D/n_{\bar{D}}$ into (3.5) results in

$$
\sum_{k\notin\kappa}\frac{\widehat{\mu}}{n_D\widehat{w}(Y_k)+n_{\bar{D}}\widehat{\mu}} = \sum_{k\notin\kappa}\frac{\frac{n_D}{n_{\bar{D}}}}{n_D\widehat{w}(Y_k)+n_{\bar{D}}\frac{n_D}{n_{\bar{D}}}} = \sum_{i=1}^{m}\frac{\frac{n_D}{n_{\bar{D}}}m_i}{n_D\frac{m_{Di}}{m_i-m_{Di}}+n_{\bar{D}}\frac{n_D}{n_{\bar{D}}}} = \sum_{i=1}^{m}\frac{\frac{n_D}{n_{\bar{D}}}m_i(m_i-m_{Di})}{n_D m_{Di}+n_D(m_i-m_{Di})} = \sum_{i=1}^{m}\frac{m_i-m_{Di}}{n_{\bar{D}}} = \frac{n_{\bar{D}}}{n_{\bar{D}}} = 1.
$$

Since the term on the left-hand side of (3.5) is monotone increasing in μ, $\widehat{\mu} = n_D/n_{\bar{D}}$ is the unique solution.

B2: Proof of Theorem 2

At the end of the isotonic regression, the estimated risks are constant within each block of marker values. Suppose there are m blocks with m_i subjects and m_{Di} cases in the i^{th} block. Let $y_{(1)}, ..., y_{(n)}$ be the marker values in the case-control sample ordered increasingly, with $y_{(i1)}, ..., y_{(im_i)}$ belonging to the i^{th} block, then $\hat{P}(D = 1|Y)$ is constant for $Y \in \{y_{(i1)}, ..., y_{(im_i)}\}$. Because the quantile function F^{-1} is defined to be left continuous by convention, the nonparametric estimator $\hat{R}(v)$ or $\tilde{R}(v)$ vs v is a step function where a jump is ready to be made (but not yet) at every v corresponding to the largest element in a block, i.e. $v = \hat{F}\{y_{(im_i)}\}$ or $v = \tilde{F}\{Y_{(im_i)}\}$ for $i = 1, ..., m$.

Therefore, to show the equivalence between the two predictiveness curve estimators, all we need to show is that the sets of *v*′*s* where jumps are about to happen is the same between the two curves. In other words, we want to show that $\tilde{F}\{y_{(im_i)}\} = \hat{F}\{y_{(im_i)}\}$ for $i = 1, ..., m$.

Notice that

$$\begin{aligned}\tilde{F}\{y_{(im_i)}\} &= \widehat{\rho}\frac{1}{n_D}\sum_{j=1}^{n_D} I\left\{Y_{Dj} \le y_{(im_i)}\right\} + (1-\widehat{\rho})\frac{1}{n_{\bar{D}}}\sum_{j=1}^{n_{\bar{D}}} I\left\{Y_{\bar{D}j} \le y_{(im_i)}\right\} \\ &= \widehat{\rho}\frac{1}{n_D}\sum_{l=1}^{i} m_{Dl} + (1-\widehat{\rho})\frac{1}{n_{\bar{D}}}\sum_{l=1}^{i}(m_l - m_{Dl})\end{aligned}$$

and that

$$\begin{aligned}\widehat{F}\{y_{(im_i)}\} &= \widehat{\rho}\frac{1}{n}\sum_{j\notin\kappa, j=1}^{n}\frac{\frac{m_{Dj}}{m_j - m_{Dj}}\frac{n_{\bar{D}}}{n_D} I\left\{Y_j \le y_{(im_i)}\right\}}{\frac{n_{\bar{D}}}{n} + \frac{n_D}{n}\frac{m_{Dj}}{m_j - m_{Dj}}\frac{n_{\bar{D}}}{n_D}} + \widehat{\rho}\sum_{j\in\kappa, j=1}^{n}\frac{I\left\{Y_j \le y_{(im_i)}\right\}}{n_D} + (1-\widehat{\rho})\frac{1}{n}\sum_{j=1}^{n}\frac{I\left\{Y_j \le y_{(im_i)}\right\}}{\frac{n_{\bar{D}}}{n} + \frac{n_D}{n}\frac{m_{Dj}}{m_j - m_{Dj}}\frac{n_{\bar{D}}}{n_D}} \\ &= \widehat{\rho}\frac{1}{n_D}\sum_{l\notin\kappa, l=1}^{i} m_{Dl} + \widehat{\rho}\frac{1}{n_D}\sum_{l\subset\kappa, l=1}^{i} m_l + (1-\widehat{\rho})\frac{1}{n_{\bar{D}}}\sum_{l=1}^{i}(m_l - m_{Dl}) \\ &= \widehat{\rho}\frac{1}{n_D}\sum_{j\notin\kappa, l=1}^{i} m_{Dl} + \widehat{\rho}\frac{1}{n_D}\sum_{l\subset\kappa, l=1}^{i} m_{Dl} + (1-\widehat{\rho})\frac{1}{n_{\bar{D}}}\sum_{l=1}^{i}(m_l - m_{Dl}) \\ &= \widehat{\rho}\frac{1}{n_D}\sum_{l=1}^{i} m_{Dl} + (1-\widehat{\rho})\frac{1}{n_{\bar{D}}}\sum_{l=1}^{i}(m_l - m_{Dl}).\end{aligned}$$

Consequently under the monotone increasing risk model assumption, the nonparametric "empirical" and model-based approaches lead to the same estimator of the predictiveness curve.

B3: Proof of Theorem 3

For a continuous marker *Y*, it has been shown that there is a one-to-one relationship between the predictiveness curve and the ROC curve [36]. That is, suppose $P(D = 1|Y)$ is monotone increasing in *Y*, $R(v)$ vs *v* can be represented as

$$\frac{\rho \mathrm{ROC}'(t)}{\rho \mathrm{ROC}'(t) + (1-\rho)} \quad \text{vs} \quad 1 - (1-\rho)t - \rho \mathrm{ROC}(t), \qquad t \in (0,1).$$

A similar result can be proved for the semiparametric maximum likelihood estimator. Suppose the unique marker value within the case-control sample is $\{y_1, ..., y_n\}$ in increasing order. For a marker value y_i,

$$\begin{aligned} v_i &= \widehat{F}(y_i) = \widehat{\rho}\widehat{F}_D(y_i) + (1-\widehat{\rho})\widehat{F}_{\bar{D}}(y_i), \\ \widehat{R}(v_i) &= \widehat{P}(D=1|Y=y_i) = \frac{\widehat{P}(D=1, Y=y_i)}{\widehat{P}(D=1)} \\ &= \frac{\widehat{\rho}\widehat{P}(Y=y_i|D=1)}{\widehat{\rho}\widehat{P}(Y=y_i|D=1) + (1-\widehat{\rho})\widehat{P}(Y=y_i|D=0)} = \frac{\widehat{\rho}\widehat{\mathcal{L}R}(y_i)}{\widehat{\rho}\widehat{\mathcal{L}R}(y_i) + (1-\widehat{\rho})}. \end{aligned}$$

Next we generate the ROC curve, $\widehat{\mathrm{ROC}}(t)$, corresponding to the semiparametric maximum likelihood estimator of the predictiveness curve: (1) we order the support of the marker

decreasingly; (2) we estimate the pair of TPF(c), FPF(c) where $c = \{y_n, ..., y_1, -\infty\}$ (here we define positive as $Y > c$ instead of $Y \geq c$ to accommodate the convention that $\hat{F}$ is right continuous); (3) we connect neighboring points by a straight line and define $\widehat{\mathrm{ROC}}'(t)$ to be the right-hand derivative of $\widehat{\mathrm{ROC}}(t)$. Suppose $\hat{F}_{\bar{D}}(y_i) = 1 - t_i$, since $\widehat{\mathcal{L}\mathrm{R}}\left\{\hat{F}^{-1}(v_i)\right\} = \widehat{\mathcal{L}\mathrm{R}}\left\{\hat{F}_{\bar{D}}^{-1}(1-t_i)\right\} = \widehat{\mathrm{ROC}}'(t_i)$, we have that $\hat{R}(v)$ vs v can be represented as

$$\frac{\widehat{\rho\mathrm{ROC}}'(t)}{\widehat{\rho\mathrm{ROC}}'(t) + (1-\widehat{\rho})} \quad \text{vs} \quad 1 - (1-\widehat{\rho})\,t - \widehat{\rho\mathrm{ROC}}(t), \qquad t \in (0,1).$$

For the semiparametric maximum likelihood estimator, $\hat{P}(Y = y_i|D = 0) > 0$, thus we have $\widehat{\mathcal{L}\mathrm{R}}(y_i) = \widehat{P}(Y{=}y_i|D{=}1)\,/\,\widehat{P}(Y{=}y_i|D{=}0) < \infty$. That is, the derivative of the corresponding curve is always finite. Therefore, the semiparametric maximum likelihood estimator of the predictiveness curve corresponds to an ROC curve which is continuous and piecewise differentiable everywhere. Moreover the ROC curve is concave since $P(D|Y)$ is monotone increasing in Y. We have

$$\begin{aligned}\int_0^1 \widehat{R}(v)\,dv = \int_{t=1}^{t=0} \frac{\widehat{\rho\mathrm{ROC}}'(t)}{\widehat{\rho\mathrm{ROC}}'(t) + (1-\widehat{\rho})}\, d\left\{1 - (1-\widehat{\rho})\,t - \widehat{\rho\mathrm{ROC}}(t)\right\} &= -\int_{t=1}^{t=0} \frac{\widehat{\rho\mathrm{ROC}}'(t)}{\widehat{\rho\mathrm{ROC}}'(t) + (1-\widehat{\rho})}\left\{(1-\widehat{\rho}) + \widehat{\rho\mathrm{ROC}}'(t)\right\} dt \\ &= \widehat{\rho}\int_{t=0}^{t=1} \widehat{\mathrm{ROC}}'(t)\,dt = \widehat{\rho}\left\{\widehat{\mathrm{ROC}}(1) - \widehat{\mathrm{ROC}}(0)\right\} = \widehat{\rho}\end{aligned}$$

For the nonparametric maximum likelihood estimator of the predictiveness curve, we can obtain the corresponding ROC curve similarly. This ROC curve is piecewise differentiable with finite derivative everywhere if $\hat{P}(D = 1|Y = y) < 1$ for every y in the support of the marker. However, when we use isotonic regression to estimate the risk model, the estimated risk could be 1 if the largest marker measure comes from the case sample. This would lead to a vertical line from (0, 0) to $(0, n_\kappa/n_D)$ in the corresponding ROC curve, where n_κ is the number of observations in κ. Nevertheless, the area under the estimated predictiveness curve is still equal to $\widehat{\rho}$ in this scenario because

$$\begin{aligned}\int_0^1 \widehat{R}(v)\,dv = 1 \times \sum_{k\in\kappa} \widehat{f}(Y_k) + \int_{t=1}^{t=0^+} \frac{\widehat{\rho\mathrm{ROC}}'(t)}{\widehat{\rho\mathrm{ROC}}'(t) + (1-\widehat{\rho})}\, d\left\{1 - (1-\widehat{\rho})\,t - \widehat{\rho\mathrm{ROC}}(t)\right\} &= \widehat{\rho} n_\kappa \frac{1}{n_D} + (1-\widehat{\rho}) \times 0 - \int_{t=1}^{t=0^+} \frac{\widehat{\rho\mathrm{ROC}}'(t)}{\widehat{\rho\mathrm{ROC}}'(t) + (1-\widehat{\rho})}\left\{(1-\widehat{\rho}) + \widehat{\rho\mathrm{R}}\right. \\ &= \widehat{\rho} n_\kappa \frac{1}{n_D} + \widehat{\rho}\int_{t=0+}^{t=1} \widehat{\mathrm{ROC}}'(t)\,dt = \rho n_\kappa \frac{1}{n_D} + \widehat{\rho}\left\{\widehat{\mathrm{ROC}}(1) - \widehat{\mathrm{ROC}}\right. \\ &= \widehat{\rho} n_\kappa \frac{1}{n_D} + \widehat{\rho}\left(1 - \frac{n_\kappa}{n_D}\right) = \widehat{\rho}.\end{aligned}$$

B4: Proof of Theorem 4

For the semiparametric maximum likelihood estimator,

$$\sqrt{n}\left\{\widehat{R}^{-1}(p) - R^{-1}(p)\right\} = \sqrt{n}\left\{\widehat{F}_R(p) - F_R(p)\right\} = \sqrt{n}\left\{(1-\widehat{\rho})\,\widehat{F}_{\bar{D}R}(t) + \widehat{\rho}\widehat{F}_{DR}(t) - F_R(t)\right\} = A + B,$$

where

$$
\begin{aligned}
A &= \sqrt{n}\left[\widehat{\rho}\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{\mathcal{LR}}_R\left\{\widehat{Risk}(Y_i)\right\}I\{Risk(Y_i)\le p\}}{\frac{n_-}{n}+\frac{n_D}{n}\widehat{\mathcal{LR}}_R\left\{\widehat{Risk}(Y_i)\right\}}+(1-\widehat{\rho})\frac{1}{n}\sum_{i=1}^{n}\frac{I\{Risk(Y_i)\le p\}}{\frac{n_-}{n}+\frac{n_D}{n}\widehat{\mathcal{LR}}_R\left\{\widehat{Risk}(Y_i)\right\}}-R^{-1}(p)\right]\\
&= \sqrt{n}\left\{Q_M(p)-R^{-1}(p)\right\},\\
B &= \sqrt{n}\left[\widehat{\rho}\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{\mathcal{LR}}_R\left\{\widehat{Risk}(Y_i)\right\}I\left\{\widehat{Risk}(Y_i)\le p\right\}}{\frac{n_-}{n}+\frac{n_D}{n}\widehat{\mathcal{LR}}_R\left\{\widehat{Risk}(Y_i)\right\}}+(1-\widehat{\rho})\frac{1}{n}\sum_{i=1}^{n}\frac{I\left\{\widehat{Risk}(Y_i)\le p\right\}}{\frac{n_-}{n}+\frac{n_D}{n}\widehat{\mathcal{LR}}_R\left\{\widehat{Risk}(Y_i)\right\}}\right]\\
&\quad -\sqrt{n}\left[\widehat{\rho}\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{\mathcal{LR}}_R\left\{\widehat{Risk}(Y_i)\right\}I\{Risk(Y_i)\le p\}}{\frac{n_-}{n}+\frac{n_D}{n}\widehat{\mathcal{LR}}_R\left\{\widehat{Risk}(Y_i)\right\}}+(1-\widehat{\rho})\frac{1}{n}\sum_{i=1}^{n}\frac{I\{Risk(Y_i)\le p\}}{\frac{n_-}{n}+\frac{n_D}{n}\widehat{\mathcal{LR}}_R\left\{\widehat{Risk}(Y_i)\right\}}\right]\\
&=^{*} \sqrt{n}\left[\rho\frac{1}{n}\sum_{i=1}^{n}\frac{\mathcal{LR}_R\{Risk(Y_i)\}I\left\{\widehat{Risk}(Y_i)\le p\right\}}{\frac{n_-}{n}+\frac{n_D}{n}\mathcal{LR}_R\{Risk(Y_i)\}}+(1-\rho)\frac{1}{n}\sum_{i=1}^{n}\frac{I\left\{\widehat{Risk}(Y_i)\le p\right\}}{\frac{n_-}{n}+\frac{n_D}{n}\mathcal{LR}_R\{Risk(Y_i)\}}\right]\\
&\quad -\sqrt{n}\left[\rho\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{\mathcal{LR}}_R\left\{\widehat{Risk}(Y_i)\right\}I\{Risk(Y_i)\le p\}}{\frac{n_-}{n}+\frac{n_D}{n}\widehat{\mathcal{LR}}_R\left\{\widehat{Risk}(Y_i)\right\}}+(1-\rho)\frac{1}{n}\sum_{i=1}^{n}\frac{I\{Risk(Y_i)\le p\}}{\frac{n_-}{n}+\frac{n_D}{n}\widehat{\mathcal{LR}}_R\left\{\widehat{Risk}(Y_i)\right\}}\right]\\
&= \sqrt{n}\left\{F_{\widehat{R}}(t)-F_R(t)\right\}+o_p(1)=\sqrt{n}\frac{\partial F_R(t)}{\partial\theta}\left(\widehat{\theta}-\theta\right)+o_p(1),
\end{aligned}
$$

where * holds under appropriate equicontinuity conditions [37].

Suppose $\widehat{\rho}$ is estimated from a cohort independent of the case-control sample, or the parent cohort where the case-control sampled is nested within. Assume the size of the cohort is λ times the size of the case-control sample. Let $Q_M^{\star}, \widehat{\theta}^{\star}$ be the counterparts of Q_M, $\widehat{\theta}$ if we plug in true ρ where $\widehat{\rho}$ is originally used. Note that $\widehat{\mathcal{LR}}_R\left\{\widehat{Risk}\right\}$ is independent of $\widehat{\rho}$ for a logistic regression risk model. Then, we have

$$
\begin{aligned}
&\mathrm{var}\left[\sqrt{n}\left\{Q_M(p)-R^{-1}(p)\right\}\right]=\left\{F_{DR}(p)-F_{\bar{D}R}(p)\right\}^2\rho(1-\rho)/\lambda+\mathrm{var}\left[\sqrt{n}\left\{Q_M^{\star}(p)-R^{-1}(p)\right\}\right],\\
&\mathrm{var}\left\{\sqrt{n}\left(\widehat{\theta}-\theta\right)\right\}=\begin{pmatrix}\frac{1}{\lambda\rho(1-\rho)} & 0\\ 0 & 0\end{pmatrix}+\mathrm{var}\left\{\sqrt{n}\left(\widehat{\theta}^{\star}-\theta\right)\right\},\\
&\mathrm{cov}\left[\sqrt{n}\left(\widehat{\theta}-\theta\right),\sqrt{n}\left\{Q_M(p)-R^{-1}(p)\right\}\right]=\frac{F_{DR}(p)-F_{\bar{D}R}(p)}{\lambda}+\mathrm{cov}\left[\sqrt{n}\left(\widehat{\theta}^{\star}-\theta\right),\sqrt{n}\left\{Q_M^{\star}(p)-R^{-1}(p)\right\}\right].
\end{aligned}
$$

To show this, use the first result as an example, note that

$$
\begin{aligned}
&\mathrm{var}\left[\sqrt{n}\left\{Q_M(p)-R^{-1}(p)\right\}\right]\\
&\quad=\mathrm{var}\left[\sqrt{n}\left\{Q_M(p)-Q_M^{\star}(p)\right\}\right]\\
&\quad+\mathrm{var}\left[\sqrt{n}\left\{Q_M^{\star}(p)-R^{-1}(p)\right\}\right]\simeq \sqrt{n}\left\{\widehat{\rho}F_{DR}(p)+(1-\widehat{\rho})F_{\bar{D}R}(p)-F_R(p)\right\}\\
&\quad+\mathrm{var}\left[\sqrt{n}\left\{Q_M^{\star}(p)-R^{-1}(p)\right\}\right] \qquad (*).
\end{aligned}
$$

Since the two terms in (*) are asymptotically uncorrelated, we have

$$\begin{aligned}\mathrm{var}\left[\sqrt{n}\left\{Q_M(p)-R^{-1}(p)\right\}\right] &\simeq \mathrm{var}\left[\sqrt{n}\left\{F_{DR}(p)-F_{\bar{D}R}(p)\right\}^2\mathrm{var}(\widehat{\rho})\right]\\ &+\mathrm{var}\left[\sqrt{n}\left\{Q_M^{\star}(p)-R^{-1}(p)\right\}\right]\\ &= \left\{F_D(t)-F_{\bar{D}}(t)\right\}^2\rho(1-\rho)\\ &/\lambda+\mathrm{var}\left[\sqrt{n}\left\{Q_M^{\star}(p)-R^{-1}(p)\right\}\right].\end{aligned}$$

Finally, results for semiparametric "empirical" estimators can be derived following similar arguments.

Appendix C: The modified Hosmer-Lemeshow test for case-control data

Let observations in a case-control sample by divided into K groups according to distribution of $\widehat{Risk}^S$, the unmodified Hosmer-Lemeshow test for case-control study is defined as

$$HL=\sum_{k=1}^{K}\frac{\left\{\widehat{P}(D=1|k,S)-\widehat{R}_k^S\right\}^2}{\widehat{R}_k^S\left(1-\widehat{R}_k^S\right)/n_k}.$$

Based on Bayes' theorem, we have $\widehat{Risk}=g\left(\widehat{Risk}^S\right)$ and $\widehat{R}_k=g\left(\widehat{R}_k^S\right)$ for

$$g(x)=\frac{x}{1-x}\frac{n_D}{n_{\bar{D}}}\frac{\widehat{\rho}}{1-\widehat{\rho}}/\left(1+\frac{x}{1-x}\frac{n_D}{n_{\bar{D}}}\frac{\widehat{\rho}}{1-\widehat{\rho}}\right).$$

Then for $k = 1, \ldots, K$, we have $\widehat{P}(D=1|k)-\widehat{R}_k=g\left(\widehat{P}(D=1|k,S)-g\left(\widehat{R}_k^S\right)\right.$. Its variance under H_0 can be shown to be approximately equal to $\widehat{R}_k^2\left(1-\widehat{R}_k\right)^2/\left\{n_kR_k^S\left(1-R_k^S\right)\right\}$ through Delta method. Therefore under H_0, $\left\{\widehat{P}(D=1|k,S)-\widehat{R}_k^S\right\}^2/\left\{\widehat{R}_k^S\left(1-\widehat{R}_k^S\right)/n_k\right\}$ and $\left\{\widehat{P}(D=1|k)-\widehat{R}_k\right\}^2/\left[\widehat{R}_k^2\left(1-\widehat{R}_k\right)^2/\left\{n_k\widehat{R}_k^S\left(1-\widehat{R}_k^S\right)\right\}\right]$ are asymptotically equivalent, which proves the asymptotic equivalence between T and HL.

References

1. Pepe, MS. Oxford University Press; Oxford, United Kingdom: 2003. The Statistical Evaluation of Medical Tests for Classification and Prediction..
2. Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. Biostatistics. 2005; 6(2): 227–239. [PubMed: 15772102]
3. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007; 115:928–935. [PubMed: 17309939]
4. Pencina MJ, D'Agostino Sr RB, D'Agostino Jr RB, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Statistics in Medicine. 2008; 27(2):157–172. [PubMed: 17569110]
5. Bura E, Gastwirth JL. The binary regression quantile plot: assessing the importance of predictors in binary regression visually. Biometrical Journal. 2001; 43(1):5–21.

6. Pepe MS, Feng Z, Huang Y, Longton GM, Prentice R, Thompson IM, Zheng Y. Integrating the predictiveness of a marker with its performance as a classifier. American Journal of Epidemiology. 2008; 167(3):362–368. [PubMed: 17982157]
7. Huang Y, Pepe MS, Feng Z. Evaluating the predictiveness of a continuous marker. Bio-metrics. 2007; 63(4):1181–1188.
8. Stern RH. Evaluating new cardiovascular risk factors for risk stratification. Journal of Clinical Hypertension. 2008; 10:485–488. [PubMed: 18550940]
9. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M, Yasui Y. Phases of biomarker development for early detection of cancer. Journal of the National Cancer Institute. 2001; 93(14):1054–1061. [PubMed: 11459866]
10. Baker SG, Kramer BS, Srivastava S. Markers for early detection of cancer: Statistical guidelines for nested case-control studies. BMI Medical Research Methodology. 2002; 2:4.
11. Huang Y, Pepe MS. Semiparametric methods for evaluating risk prediction markers in case-control studies. Biometrika. 2009 In Press.
12. Ransohoff DF. How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. Journal of Clinical Epidemiology. 2007; 60:1205–1219. [PubMed: 17998073]
13. Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. Journal of Clinical Oncology. 2005; 23(29):7332–7341. [PubMed: 16145063]
14. Buyse M, Loi S, van't Veer L, Viale G, Delorenzi M, Glas AM, Saghatchian d'Assignies M, Bergh J, Lidereau R, Ellis P, Harris A, Bogaerts J, Therasse P, Floore A, Amakrane M, Piette F, Rutgers E, Sotiriou C, Cardoso F, Piccart MJ. Validation and Clinical Utility of a 70-Gene Prognostic Signature for Women With Node-Negative Breast Cancer. Journal of the National Cancer Institute. 2006; 98(17):1183–1192. [PubMed: 16954471]
15. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; 415(6871):530–536. [PubMed: 11823860]
16. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AAM, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R. A gene-expression signature as a predictor of survival in breast cancer. New England Journal of Medicine. 2002; 247(25):1999–2009. [PubMed: 12490681]
17. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. American Heart Journal. 1991; 121:293–298. [PubMed: 1985385]
18. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. Journal of the National Cancer Institute. 1989; 81(24):1879–1886. [PubMed: 2593165]
19. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal evaluation of the accuracy of a classification biomarker: standards for study design. Journal of the National Cancer Institute. 2008; 100(20):1432–1438. [PubMed: 18840817]
20. Cole TJ, Green PJ. Smoothing reference centile curves: The LMS method and penalized likelihood. Statistics in Medicine. 1992; 11(165):1305–1319. [PubMed: 1518992]
21. Breslow NE. Statistics in epidemiology: the case-control study. JASA. 1996; 91:14–28. [PubMed: 12155399]
22. Anderson JA. Separate sample logistic discrimination. Biometrika. 1972; 59(1):19–35.
23. Prentice RL, Pyke R. Logistic Disease Incidence Models and Case-Control Studies. Biometrika. 1979; 66(3):403–411.
24. Barlow, RE.; Bartholomew, DJ.; Bremner, JM.; Brunk, HD. Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression. Wiley; London: 1972.
25. Lloyd CJ. Estimation of a Convex ROC Curve. Statistics & Probability Letters. 2002; 59:99–111.
26. Lloyd CJ. Maximum likelihood estimation of misclassification rates of a binomial regression. Biometrika. 2000; 87(3):700–705.

27. Owen AB. Empirical likelihood ratio confidence intervals for a single functional. Biometrika. 1988; 75(2):237–249.
28. Qin J, Zhang J. A goodness-of-fit test for logistic regression models based on case-control data. Biometrika. 1997; 84(3):609–618.
29. Qin J, Zhang J. Using logistic regression procedures for estimating receiver operating characteristic curves. Biometrika. 2003; 93(3):585–596.
30. Thompson IM, Pauler Ankerst D, Chi C. Assessing prostate cancer risk: results from the prostate cancer prevention trial. Journal of the National Cancer Institute. 2006; 98:529–534. [PubMed: 16622122]
31. Hosmer DW, Lemesbow S. Goodness of fit tests for the multiple logistic regression model. Communications in Statistics-Theory and Methods. 1980; 9(10):1043–1069.
32. Hosmer, DW.; Lemeshow, S. 2nd Edition. John Willey & Sons; 2000. Applied logistic regression.
33. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. Ann Intern Med. 2006; 145:21–9. [PubMed: 16818925]
34. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clin Chem. 2008; 54:17–23. [PubMed: 18024533]
35. Janes H, Pepe MS, Gu W. Assessing the value of risk predictions by using risk stratification tables. Annals of Internal Medicine. 2008; 149:751–760. [PubMed: 19017593]
36. Huang Y, Pepe MS. A parametric ROC model based approach for evaluating the predictiveness of continuous markers in case-control studies. Biometrics. 2009 In Press.
37. van der Vaart, AW.; Wellner, JA. Weak Convergence and Empirical Process. Spring-Verlag; New York: 1996.

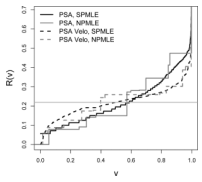

Figure 1.
The predictiveness curves for PSA and PSA velocity for predicting prostate cancer.

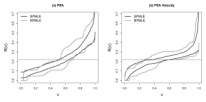

Figure 2.
The 95% pointwise confidence intervals constructed from percentiles of the bootstrap distribution for the predictiveness curves of PSA and PSA velocity. SPMLE: semiparametric maximum likelihood estimator; NPMLE: nonparametric maximum likelihood estimator.

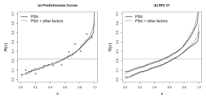

Figure 3.
(a) The semiparametric maximum likelihood estimates of predictiveness curves for PSA and PSA plus other factors for predicting risk prostate cancer,the dots are average risk within deciles of modeled risk based on the latter model; (b) their 95% pointwise confidence intervals using percentiles of the bootstrap distribution. The horizontal lines indicate disease prevalences.

Table 1

Bias of the semiparametric and nonparametric estimators for the linear logistic model. Also shown are biases of variance estimators based on asymptotic theory. Here $n_D = n_{\bar{D}} = n/2$. *Size of the phase-one cohort for estimating* $\widehat{\rho}$ *is* 5n. SPMLE denotes the semiparametric maximum likelihood estimator, SPE denotes the semiparametric "empirical" estimator, NPMLE denotes the nonparametric maximum likelihood estimator.

		$v = 0.1$	$v = 0.3$	$v = 0.5$	$v = 0.7$	$v = 0.9$
$R(v)$		0.045	0.094	0.15	0.24	0.43
% bias in $\hat{R}(v)$						
$n = 100$	SPMLE	4.47	–0.50	–1.39	–0.82	0.82
	SPE	5.13	–0.40	–1.18	–0.53	0.94
	NPMLE	–35.35	–9.42	–5.44	–3.27	2.86
$n = 500$	SPMLE	1.14	–0.06	–0.34	–0.13	0.16
	SPE	1.12	–0.06	–0.30	–0.07	0.15
	NPMLE	–13.15	–3.21	–1.86	–1.38	0.58
$n = 2000$	SPMLE	0.16	–0.13	–0.13	–0.07	0.11
	SPE	0.15	–0.13	–0.10	–0.06	0.10
	NPMLE	–4.72	–1.59	–0.83	–0.47	0.35
% bias in variance estimate of $\hat{R}(v)$						
$n = 100$	SPMLE	-4.44	7.91	7.28	1.54	-4.23
	SPE	-1.27	7.03	9.02	1.22	-4.35
$n = 500$	SPMLE	-1.99	1.47	3.63	2.11	-0.35
	SPE	–2.50	2.07	3.88	-1.92	2.48
$n = 2000$	SPMLE	–1.39	1.33	1.91	1.20	-0.49
	SPE	–1.13	1.65	2.31	1.65	-1.06
		$p = 0.045$	$p = 0.094$	$p = 0.15$	$p = 0.24$	$p = 0.43$
$R^{-1}(p)$		0.1	0.3	0.5	0.7	0.9
% bias in $\hat{R}^{-1}(p)$						
$n = 100$	SPMLE	12.68	0.39	–0.16	0.55	0.11
	SPE	12.90	0.38	–0.34	0.50	0.12
	NPMLE	80.50	15.00	5.86	2.06	–0.79
$n = 500$	SPMLE	2.43	0.02	0.002	0.11	0.03
	SPE	2.53	0.02	–0.04	0.06	0.04
	NPMLE	29.78	5.84	2.11	0.86	–0.23

$n = 2000$	SPMLE	0.94	0.15	0.05	0.04	–0.01	
	SPE	0.94	0.14	0.03	0.04	–0.01	
	NPMLE	12.84	2.47	0.98	0.34	–0.16	
% bias in variance estimate of $R^{-1}(p)$							
$n = 100$	SPMLE	15.14	-6.80	-14.03	-11.01	9.52	
	SPE	14.88	-5.62	-13.62	-9.67	8.81	
$n = 500$	SPMLE	7.40	-6.14	-5.92	-5.86	5.54	
	SPE	7.34	–6.12	-6.05	-5.87	3.58	
$n = 2000$	SPMLE	2.75	-2.87	-3.08	-1.84	3.66	
	SPE	2.70	-2.75	-2.36	-0.56	4.00	

Table 2

Coverage of 95% Wald confidence intervals based on the semiparametric and nonparametric estimators in nested case-control studies for the linear logistic model, assuming the logit transform of the estimator is normally distributed. SPMLE denotes the semiparametric maximum likelihood estimator, SPE denotes the semiparametric "empirical" estimator, and NPMLE denotes the nonparametric maximum likelihood estimator.

		$v = 0.1$	$v = 0.3$	$v = 0.5$	$v = 0.7$	$v = 0.9$
$R(v)$		0.045	0.094	0.15	0.24	0.43
		Based on asymptotic variance estimate				
$n = 100$	SPMLE	95.82	95.64	95.94	96.22	95.50
	SPE	95.78	95.58	96.38	96.04	95.44
$n = 500$	SPMLE	94.62	95.14	95.88	95.40	94.80
	SPE	94.64	95.30	95.78	95.42	94.52
$n = 2000$	SPMLE	95.00	95.38	95.32	95.06	94.96
	SPE	94.92	95.22	95.44	95.62	95.50
		Based on bootstrap variance estimate				
$n = 100$	SPMLE	96.02	95.14	95.68	96.68	96.26
	SPE	96.04	95.02	96.02	96.86	96.04
	NPMLE	96.49	97.56	96.64	96.36	98.10
$n = 500$	SPMLE	95.06	94.94	95.30	95.28	95.18
	SPE	94.96	95.02	95.34	95.60	94.86
	NPMLE	96.57	94.76	95.16	95.30	96.04
$n = 2000$	SPMLE	95.20	95.18	94.78	95.22	94.94
	SPE	95.02	95.22	95.04	95.36	95.58
	NPMLE	94.58	94.08	94.84	94.62	94.90
		$p = 0.045$	$p = 0.094$	$p = 0.15$	$p = 0.24$	$p = 0.43$
$R^{-1}(p)$		0.1	0.3	0.5	0.7	0.9
		Based on asymptotic variance estimate				
$n = 100$	SPMLE	91.39	93.55	95.30	95.94	93.13
	SPE	91.22	93.97	95.50	96.26	93.63
$n = 500$	SPMLE	95.40	95.10	94.98	94.86	95.76
	SPE	95.32	94.86	94.94	94.82	95.76
$n = 2000$	95.40	94.82	94.80	94.94	95.28	
	SPE	95.28	95.10	94.80	95.30	95.56

		Based on bootstrap variance estimate				
$n = 100$	SPMLE	91.73	94.75	96.16	97.10	93.85
	SPE	91.62	94.99	96.28	96.98	94.06
	NPMLE	72.82	90.81	96.40	97.52	90.68
$n = 500$	SPMLE	95.16	95.32	95.48	95.32	94.78
	SPE	95.10	95.36	95.60	95.24	95.34
	NPMLE	83.21	93.08	96.26	96.66	93.20
$n = 2000$	SPMLE	94.86	94.92	95.04	95.10	94.84
	SPE	94.80	95.12	95.36	95.58	94.90
	NPMLE	89.08	94.14	95.22	95.68	94.58

Table 3

Coverage of 95% percentile bootstrap confidence intervals based on the semiparametric and nonparametric estimators in nested case-control studies for the linear logistic model. SPMLE denotes the semiparametric maximum likelihood estimator, SPE denotes the semiparametric "empirical" estimator, and NPMLE denotes the nonparametric maximum likelihood estimator.

		$v = 0.1$	$v = 0.3$	$v = 0.5$	$v = 0.7$	$v = 0.9$
$R(v)$		0.045	0.094	0.15	0.24	0.43
$n = 100$	SPMLE	94.24	94.50	95.08	95.86	94.80
	SPE	94.58	94.64	95.28	96.36	94.98
	NPMLE	79.74	94.30	96.58	97.54	98.02
$n = 500$	SPMLE	94.18	94.18	94.42	95.88	94.22
	SPE	94.66	94.16	94.60	96.02	94.54
	NPMLE	93.04	95.92	97.18	97.40	98.20
$n = 2000$	SPMLE	94.38	94.52	94.66	94.72	94.22
	SPE	94.58	94.38	95.08	95.08	94.76
	NPMLE	94.80	96.66	97.44	97.80	97.96
		$p = 0.045$	$p = 0.094$	$p = 0.15$	$p = 0.24$	$p = 0.43$
$R^{-1}(p)$		0.1	0.3	0.5	0.7	0.9
$n = 100$	SPMLE	94.24	94.54	95.04	95.90	94.80
	SPE	94.56	94.72	95.28	96.38	95.08
	NPMLE	79.74	94.36	96.62	97.60	98.06
$n = 500$	SPMLE	94.12	94.06	95.02	94.90	94.22
	SPE	94.22	94.46	95.40	95.18	94.62
	NPMLE	91.44	96.02	97.20	97.92	98.20
$n = 2000$	SPMLE	94.38	94.48	94.70	94.74	94.12
	SPE	94.58	94.42	95.12	95.10	94.74
	NPMLE	94.80	96.62	97.46	97.80	97.92

Stat Med. Author manuscript; available in PMC 2011 June 15.

Table 4

Efficiency (ratio of observed variances in simulation studies) of the semiparametric "empirical" estimator and nonparametric estimator relative to the semiparametric maximum likelihood estimator of the predictiveness curve in nested case-control studies for the linear logistic model. SPE denotes the semiparametric "empirical" estimator, NPMLE denotes the nonparametric maximum likelihood estimator. Here n = ∞ *denotes the asymptotic variance.*

		$v = 0.1$	$v = 0.3$	$v = 0.5$	$v = 0.7$	$v = 0.9$
$R(v)$		0.045	0.094	0.15	0.24	0.43
$n = 100$	SPE	1.02	0.97	0.97	0.90	0.91
	NPMLE	0.45	0.41	0.27	0.18	0.29
$n = 500$	SPE	0.98	0.98	0.96	0.90	0.89
	NPMLE	0.25	0.25	0.16	0.10	0.18
$n = 2000$	SPE	0.99	0.98	0.96	0.90	0.91
	NPMLE	0.17	0.16	0.10	0.06	0.11
$n = \infty$	SPE	0.99	0.97	0.97	0.90	0.91
		$p = 0.045$	$p = 0.094$	$p = 0.15$	$p = 0.24$	$p = 0.43$
$R^{-1}(p)$		0.1	0.3	0.5	0.7	0.9
$n = 100$	SPE	0.99	0.99	0.96	0.91	0.91
	NPMLE	0.47	0.47	0.32	0.21	0.38
$n = 500$	SPE	0.99	0.98	0.95	0.90	0.90
	NPMLE	0.28	0.27	0.17	0.10	0.20
$n = 2000$	SPE	0.99	0.98	0.96	0.91	0.91
	NPMLE	0.18	0.16	0.10	0.06	0.12
$n = \infty$	SPE	0.99	0.97	0.97	0.90	0.91

Table 5

Performances of the semiparametric and nonparametric estimators for R(v) when the predictiveness curve is piecewise linear. Here $n_D = n_{\bar{D}} = n/2$. *Size of the phase-one cohort for estimating* $\widehat{\rho}$ *is* 5n. SPMLE denotes the semiparametric maximum likelihood estimator, SPE denotes the semiparametric "empirical" estimator, NPMLE denotes the nonparametric maximum likelihood estimator.

		$v = 0.1$	$v = 0.3$	$v = 0.5$	$v = 0.7$	$v = 0.9$
$R(v)$		0.05	0.13	0.18	0.22	0.27
% bias in $\hat{R}(v)$						
$n = 500$	SPMLE	57.46	-9.06	-14.25	-9.24	4.72
	SPE	51.92	-10.74	-14.35	-8.02	7.70
	NPMLE	-13.25	-4.58	-2.55	-0.80	1.92
$n = 1000$	SPMLE	57.48	-8.87	-14.07	-9.13	4.57
	SPE	51.88	-10.53	-14.16	-7.94	7.54
	NPMLE	-8.42	-2.62	-1.38	-0.66	0.78
$n = 2000$	SPMLE	57.63	-8.74	-13.97	-9.11	4.46
	SPE	51.97	-10.38	-14.05	-7.90	7.36
	NPMLE	-4.54	-1.14	-0.71	-0.50	0.15
Efficiecy[a] related to MLE						
$n = 500$	SPE	1.19	0.81	0.98	1.22	0.66
	NPMLE	1.68	0.35	1.14	0.70	0.44
$n = 1000$	SPE	1.21	0.77	0.98	1.26	0.56
	NPMLE	2.57	0.43	1.65	0.97	0.48
$n = 2000$	SPE	1.22	0.75	0.99	1.29	0.50
	NPMLE	4.10	0.56	2.55	1.46	0.52
Coverage of 95% percentile boostrap CI						
$n = 500$	SPMLE	30.16	79.22	16.70	46.18	90.06
	SPE	37.60	74.26	17.82	58.78	85.18
	NPMLE	92.44	96.18	96.64	96.78	97.56
$n = 1000$	SPMLE	5.92	65.32	1.64	18.60	87.36
	SPE	10.60	56.44	2.08	33.80	75.56
	NPMLE	94.52	96.54	96.84	96.94	97.62
$n = 2000$	SPMLE	0.06	43.04	0.00	2.16	79.06
	SPE	0.40	29.86	0.00	8.54	58.44

NPMLE	95.60	96.90	97.20	97.32	97.46

[a] efficiency in terms of mean squared error

Table 6

Performances of the semiparametric and nonparametric estimators for $R^{-1}(p)$ when the predictiveness curve is piecewise linear. Here $n_D = n_{\bar{D}} = n/2$. *Size of the phase-one cohort for estimating* $\widehat{\rho}$ *is* 5n. SPMLE denotes the semiparametric maximum likelihood estimator, SPE denotes the semiparametric "empirical" estimator, NPMLE denotes the nonparametric maximum likelihood estimator.

		$p = 0.05$	$p = 0.13$	$p = 0.18$	$p = 0.22$	$p = 0.27$
$R^{-1}(p)$		0.1	0.3	0.5	0.7	0.9
% bias in $\hat{R}^{-1}(p)$						
$n = 500$	SPMLE	-74.16	21.00	24.42	9.86	-2.01
	SPE	-71.15	23.93	23.28	8.27	-3.28
	NPMLE	19.68	7.85	5.52	1.53	-1.70
$n = 1000$	SPMLE	-76.78	21.13	24.07	9.65	-2.03
	SPE	-73.94	24.04	22.94	8.07	-3.32
	NPMLE	13.36	4.86	3.24	1.19	-0.99
$n = 2000$	SPMLE	-78.39	21.13	23.99	9.57	-2.04
	SPE	-75.62	24.02	22.83	7.97	-3.33
	NPMLE	8.28	2.99	1.59	1.05	-0.24
Efficiecy[a] related to SPMLE						
$n = 500$	SPE	1.05	1.20	1.17	1.21	1.17
	NPMLE	4.09	0.41	0.29	0.82	0.74
$n = 1000$	SPE	1.05	1.25	1.21	1.22	1.18
	NPMLE	6.68	0.32	0.29	1.05	0.73
$n = 2000$	SPE	1.06	1.29	1.23	1.23	1.19
	NPMLE	9.92	0.28	0.31	1.38	0.70
Coverage of 95% percentile boostrap CI						
$n = 500$	SPMLE	30.28	79.30	16.74	46.14	90.04
	SPE	37.64	74.22	17.78	58.78	85.20
	NPMLE	92.40	96.14	96.64	96.78	97.58
$n = 1000$	SPMLE	5.94	65.32	1.64	18.56	87.40
	SPE	10.64	56.48	2.06	33.96	75.60
	NPMLE	94.44	96.54	96.78	96.96	97.64
$n = 2000$	SPMLE	0.06	42.94	0.00	2.16	79.06

Stat Med. Author manuscript; available in PMC 2011 June 15.

SPE	0.42	29.88	0.00	8.56	58.44	
NPMLE	95.62	96.96	97.24	97.30	97.46	

[a] efficiency in terms of mean squared error

Table 7

Comparisons between (a) PSA and PSA velocity and (b) between PSA and PSA plus other risk factors for predicting risk of prostate cancer. SPMLE denotes the semiparametric maximum likelihood estimator, NPMLE denotes the nonparametric maximum likelihood estimator.

Measure	Method	(a) PSA		PSA Velocity		pvalue
		Est	95% CI	Est	95% CI	
$R(0.1)$	NPMLE	0.079	(0.043,0.110)	0.088	(0.044, 0.131)	0.730
	SPMLE	0.072	(0.046, 0.109)	0.122	(0.075, 0.159)	0.027
$R(0.9)$	NPMLE	0.474	(0.369, 0.577)	0.302	(0.253, 0.402)	0.005
	SPMLE	0.413	(0.356, 0.476)	0.313	(0.275, 0.356)	< 0.001
$R^{-1}(0.1)$	NPMLE	0.304	(0.050, 0.39)	0.155	(0.019, 0.305)	0.178
	SPMLE	0.188	(0.073, 0.291)	0.06	(0.020, 0.142)	0.021
$1 - R^{-1}(0.3)$	NPMLE	0.302	(0.140, 0.447)	0.168	(0.007, 0.501)	0.274
	SPMLE	0.244	(0.191, 0.296)	0.129	(0.030, 0.197)	0.009
$R^{-1}(0.3) - R^{-1}(0.1)$	NPMLE	0.393	(0.210, 0.664)	0.677	(0.301, 0.933)	0.100
	SPMLE	0.568	(0.443, 0.724)	0.811	(0.668, 0.935)	0.004

Measure	Method	(b) PSA		PSA + other factors		pvalue
		Est	95% CI	Est	95% CI	
$R(0.1)$	SPMLE	0.072	(0.045, 0.109)	0.070	(0.039, 0.094)	0.798
$R(0.9)$	SPMLE	0.413	(0.356, 0.476)	0.429	(0.372, 0.502)	0.223
$R^{-1}(0.1)$	SPMLE	0.188	(0.073, 0.291)	0.204	(0.109, 0.310)	0.595
$1 - R^{-1}(0.3)$	SPMLE	0.244	(0.191, 0.296)	0.243	(0.203, 0.281)	0.952
$R^{-1}(0.3) - R^{-1}(0.1)$	SPMLE	0.568	(0.443, 0.724)	0.554	(0.436, 0.662)	0.667