7-9-2008

Semiparametric and nonparametric methods for evaluating risk prediction markers in case-control studies

Ying Huang
Fred Hutchinson Cancer Research Center, ying@u.washington.edu

Margaret Pepe
University of Washington, Fred Hutch Cancer Research Center, mspepe@u.washington.edu

Semiparametric and nonparametric methods for evaluating risk prediction markers in case-control studies

Ying Huang†, Margaret Sullivan Pepe

Fred Hutchinson Cancer Research Center Public Health Sciences

1100 Fairview Avenue N., M3-A310, Seattle, WA 98109-1024

and

Fred Hutchinson Cancer Research Center Public Health Sciences

1100 Fairview Avenue N., M2-B500, Seattle, WA 98109-1024

† *Corresponding author's address:* yhuang@fhcrc.org

Summary

The performance of a well calibrated risk model, $Risk(Y) = P(D = 1|Y)$, can be characterized by the population distribution of $Risk(Y)$ and displayed with the predictiveness curve. Better performance is characterized by a wider distribution of $Risk(Y)$, since this corresponds to better risk stratification in the sense that more subjects are identified at low and high risk for the outcome $D = 1$. Although methods have been developed to estimate predictiveness curves from cohort studies, most studies to evaluate novel risk prediction markers employ case-control designs. Here we develop semiparametric and nonparametric methods that accommodate case-control data and assume apriori knowledge of $P(D = 1)$. Large and small sample properties are investigated. The semiparametric methods are flexible, substantially more efficient than the nonparametric counterparts and naturally generalize methods previously developed for cohort data. Applications to prostate cancer risk prediction markers illustrate the methods.

Key words:

Risk, classification, biomarker, predictiveness curve, case-control study, ROC curve

1. Introduction

Selecting biomarkers for medical application is an important and challenging task. Of the thousands of markers made available by modern techniques, we want to find those that can assist medical decision making and help identify disease or risk of disease early when interventions are most effective. Criteria for evaluating a biomarker should depend on its purpose. An intrinsic property of a diagnostic marker is classification accuracy, i.e. its 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 (Pepe, 2003). 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 (2005), Cook (2007), and Pencina et al. (2007), criteria for 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.

Pepe et al. (2008a) suggested using the predictiveness curve (Bura and Gastwirth, 2001) to evaluate a risk prediction marker or model. They argued that the performance of a model to predict risks within a population relies not only on the effect of each predictor in the risk model, but also on the distributions of the predictors. The predictiveness curve integrates these two factors together by displaying the population distribution of risk endowed by the risk model. Let D denote a binary outcome that we term disease here, $D = 1$ for diseased and $D = 0$ for non-diseased. Let Y denote a vector of predictors of interest and let $Risk(Y) = P(D = 1|Y)$ denote the risk calculated on the basis of Y. The predictiveness curve is the curve $R(v)$ vs v for $v \in (0, 1)$, where $R(v)$ is the v^{th} percentile of $Risk(Y)$. 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. An attractive feature of this curve is that it provides a common meaningful scale for comparing markers that may not be

comparable on their original scales. A risk prediction model with larger variability in $R(v)$ has a better capacity to stratify risk. A particularly clinically meaningful comparison can be based on $R^{-1}(p)$. Suppose there exists a pre-specified low risk threshold p_L and/or a high risk threshold p_H such that recommendation for or against treatment is clear if the estimated risk for a patient is above p_H or below p_L. A risk model which assigns more people into the low and high risk ranges (i.e. larger $R^{-1}(p_L)$ and larger $1 - R^{-1}(p_H)$) is preferred.

Pepe et al. (2001) proposed five phases for developing a biomarker. Case-control studies are conducted in phases 1, 2, and 3, since they are smaller and more cost efficient than cohort studies. Since early phase studies dominate biomarker research, it is crucial that measures of biomarker performance accommodate case-control designs. Huang et al. (2007) developed a semiparametric estimator of the predictiveness curve for cohort studies. Here we address the more common case-control design and extend estimation to include nonparametric and alternative semiparametric methods. We start with the scenario of a single continuous marker or a pre-defined marker combination and examine later the extension to a general risk model. Biomarker researchers are well aware of problems caused by developing combinations and assessing them in the same dataset and encourage the assessment of a predefined combination with independent data (Ransohoff, 2007; Simon, 2005; Pepe et al., 2008b). Examples of well known pre-defined combination scores are the Framingham score for cardiovascular events (Anderson et al., 1991) and the Gail score for breast cancer risk (Gail et al., 1989).

Let Y, Y_D, and $Y_{\bar{D}}$ denote the marker measurement in the general, diseased, and non-diseased populations respectively. Let F, F_D, and $F_{\bar{D}}$ be the corresponding distribution functions and let f, f_D, and $f_{\bar{D}}$ be the density functions. Let $\rho = P(D = 1)$ denote the disease prevalence. We assume either that ρ is known or that a prevalence estimate $\hat{\rho}$ is available in addition to the case-control sample. For example, an estimate might be obtained from a cohort study reported in the literature. Alternatively, it may be calculated from a parent cohort within which the case-control study is nested (Baker et al., 2002; Pepe et al., 2008b). In these scenarios, variability in $\hat{\rho}$ can be evaluated and taken into account in calculating the variance of the predictiveness curve estimator.

Furthermore, we assume the risk of disease $P(D=1|Y)$ is monotone increasing in Y. Under this monotone increasing risk assumption, we have $R(v) = P\{D=1|Y=F^{-1}(v)\}$, the risk at the v^{th} quantile of Y in the population. Thus the curve $R(v)$ vs v is the same as the curve $P(D=1|Y=y)$ vs $F(y)$. Therefore, estimation of the predictiveness curve can be undertaken in two steps: estimation of the risk model $P(D=1|Y=y)$, and estimation of the marker distribution $F(y)$. We develop estimators for these two entities and combine them to get a predictiveness curve estimate. We consider a case-control study with n_D cases $Y_{Di}, i=1,\ldots,n_D$, and $n_{\bar{D}}$ controls $Y_{\bar{D}i}$, $i=1,\ldots,n_{\bar{D}}$ and write $\{Y_k, k=1,\ldots,n\}$ for $\left\{Y_{\bar{D}1},\ldots,Y_{\bar{D}n_{\bar{D}}},Y_{D1},\ldots,Y_{Dn_D}\right\}$ where $n = n_{\bar{D}} + n_D$.

2. Semiparametric Estimators

2.1 *Estimation of the Risk Model*

Suppose the risk model of interest is $P(D=1|Y) = G(\theta, Y)$, where

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

and η is some monotone increasing function of Y. Examples of $\text{logit}\,\{G(\theta,Y)\}$ include $\theta_0 + \theta_1 Y$ with $\theta_1 > 0$, the ordinary linear logistic model, and $\theta_0 + \theta_1 Y^{(\theta_2)}$ with $\theta_1 > 0$, where $Y^{(\theta_2)} = (Y^{\theta_2}-1)/\theta_2$ when $\theta_2 \neq 0$ and $Y^{(\theta_2)} = \log Y$ when $\theta_2 = 0$, the logistic model with Box-Cox transformation (Cole and Green, 1992). 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 for disease is still estimable. It has been shown that the maximum likelihood estimator of the odds ratio from the retrospective likelihood can be obtained by applying the prospective logistic model to the case-control sample (Anderson, 1972; Prentice and Pyke, 1979), and that this achieves the semiparametric information bound (Bickel et al., 1993; Breslow et al., 2000; Gilbert, 2000).

Let S denote being selected into the case-control sample. We apply the standard logistic regression model $\text{logit}\{P(D=1|Y,S)\} = \theta_{0S} + \eta(\theta_{1S}, Y)$ to the data and correct the intercept with disease prevalence according to Bayes' theorem,

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

That is, let $\left(\hat{\theta}_{0S}, \hat{\theta}_{1S}\right)$ be the maximum likelihood estimators of $(\theta_{0S}, \theta_{1S})$, the estimator of θ is

$$\hat{\theta} = \begin{pmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{pmatrix} = \begin{pmatrix} \hat{\theta}_{0S} + \log\left(\frac{n_{\bar{D}}}{n_D} \frac{\hat{\rho}}{1-\hat{\rho}}\right) \\ \hat{\theta}_{1S} \end{pmatrix}.$$

2.2 *Estimation of the Marker Distribution and the Predictiveness Curve*

In a case-control study, since we do not have an independent identically distributed sample from the population, the marker distribution F cannot be estimated directly. Rather, with an estimate of disease prevalence, $\hat{\rho}$, we can estimate F according to $\hat{\rho}F_D + (1-\hat{\rho})F_{\bar{D}}$. Next we examine two ways of estimating F_D and $F_{\bar{D}}$.

2.2.1 Semiparametric "Empirical" Estimators In an unmatched case-control study, since control and case samples are representative of their corresponding distributions in the population, natural estimators for $F_{\bar{D}}$ and F_D are the empirical estimators $\tilde{F}_{\bar{D}}$ and $\tilde{F}_D$. We estimate F with $\tilde{F} = \hat{\rho}\tilde{F}_D + (1-\hat{\rho})\tilde{F}_{\bar{D}}$. The semiparametric "empirical" estimators of $R(v)$ and $R^{-1}(p)$ are

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

2.2.2 Semiparametric Maximum Likelihood Estimators Observe that the risk model (1) implies the following relationship between marker densities in cases and controls

$$f_D(Y) = \mathcal{L}\mathrm{R}(Y) f_{\bar{D}}(Y) = \exp\{\alpha + \eta(\beta, Y)\} f_{\bar{D}}(Y), \tag{2}$$

where $\alpha = \theta_0 + \log\{(1-\rho)/\rho\}$, $\beta = \theta_1$, and $\mathcal{L}\mathrm{R}(Y)$ is the likelihood ratio of Y (Green and Swets, 1966). When we estimate F_D and $F_{\bar{D}}$ empirically as in Section 2.2.1, positive point masses are allocated only to those marker values observed in the corresponding case or control sample. For a marker measured on a continuous scale, the supports for $\tilde{F}_{\bar{D}}$ and $\tilde{F}_D$ are rarely

the same. Therefore the relationship (2) is not incorporated into estimation of F_D and $F_{\bar{D}}$ in the "empirical" procedure. An issue with a similar flavor has been raised in a different problem where the task is to estimate the misclassification rates of a binary classification rule constructed from binomial regression (Lloyd, 2000). Lloyd (2000) pointed out that if the accuracy of the rule is summarized by the empirical type I and type II misclassification rates, the exponential tilt relationship (2) between densities of predictors in the diseased and non-diseased populations is ignored.

Incorporating (2) can be achieved by using the semiparametric likelihood framework (Qin and Zhang, 1997, 2003). This was originally proposed by Qin and Zhang (1997) to test the logistic regression assumption under a case-control sampling plan, and used by Qin and Zhang (2003) to estimate the ROC curve as an alternative to the fully parametric and nonparametric approaches.

Suppose $\eta(\beta, Y) = \beta^T r(Y)$,where $r(Y)$ is a vector of functions of Y. The likelihood ratio of Y becomes $\mathcal{L}\mathrm{R}(Y) = \exp\left\{\alpha + \beta^T r(Y)\right\}$. Here we focus on Y being a single marker, but this method applies also when Y is a vector of markers. The semiparametric likelihood for observing the case-control data is

$$\begin{aligned} \mathcal{L}(\alpha, \beta, F_{\bar{D}}) &= \prod_{i=1}^{n_{\bar{D}}} dF_{\bar{D}}(Y_{\bar{D}i}) \prod_{j=1}^{n_D} \exp\left\{\alpha + \beta^T r(Y_{Dj})\right\} dF_{\bar{D}}(Y_{Dj}) \qquad (3) \\ &= \left\{\prod_{i=1}^{n} dF_{\bar{D}}(Y_i)\right\} \left[\prod_{j=1}^{n_D} \exp\left\{\alpha + \beta^T r(Y_{Dj})\right\}\right], \end{aligned}$$

subject to $\sum_{i=1}^{n} dF_{\bar{D}}(Y_i) = 1$ and $\sum_{i=1}^{n} \exp\left\{\alpha + \beta^T r(Y_i)\right\} dF_{\bar{D}}(Y_i) = 1$. Refer to Qin and Zhang (1997, 2003) for details about solving this restricted maximum likelihood using the Lagrange Multiplier method. As a result, the maximum likelihood estimators for $F_{\bar{D}}$ and F_D are

$$\begin{aligned} \hat{F}_{\bar{D}}(y) &= \frac{1}{n_{\bar{D}}} \sum_{i=1}^{n} \frac{I(Y_i \le y)}{1 + \frac{n_D}{n_{\bar{D}}} \exp\left\{\hat{\alpha} + \hat{\beta}^T r(Y_i)\right\}} = \frac{1}{n} \sum_{i=1}^{n} \frac{I(Y_i \le y)}{\frac{n_{\bar{D}}}{n} + \frac{n_D}{n} \widehat{\mathcal{L}\mathrm{R}}(Y_i)}, \\ \hat{F}_D(y) &= \frac{1}{n_{\bar{D}}} \sum_{i=1}^{n} \frac{\exp\left\{\hat{\alpha} + \hat{\beta}^T r(Y_i)\right\} I(Y_i \le y)}{1 + \frac{n_D}{n_{\bar{D}}} \exp\left\{\hat{\alpha} + \hat{\beta}^T r(Y_i)\right\}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\widehat{\mathcal{L}\mathrm{R}}(Y_i) I(Y_i \le y)}{\frac{n_{\bar{D}}}{n} + \frac{n_D}{n} \widehat{\mathcal{L}\mathrm{R}}(Y_i)}, \end{aligned}$$

where $\hat{\alpha} = \hat{\theta} - \log\{\hat{\rho}/(1-\hat{\rho})\}$, $\hat{\beta} = \hat{\theta}$, and $\widehat{\mathcal{L}\mathrm{R}}$ is the maximum likelihood estimator of $\mathcal{L}\mathrm{R}$.

We use these estimators to compute $\hat{F} = (1-\hat{\rho})\hat{F}_{\bar{D}} + \hat{\rho}\hat{F}_D$. Then we plug $\hat{\theta}$ and $\hat{F}$ into G to get the semiparametric model-based estimators of $R(v)$ and $R^{-1}(p)$:

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

Note that an intrinsic property of the predictiveness curve is that the area under the curve is equal to ρ since $\int_0^1 R(v)dv = P(D=1) = \rho$. This is not necessarily true, though, for an estimated predictiveness curve due to estimation error. However, it can be shown that the area under the semiparametric maximum likelihood estimator $\hat{R}(v)$ is always equal to $\hat{\rho}$ (see Huang (2007) for a proof). This property facilitates visual comparison between two estimated curves. This result does not hold for $\tilde{R}(v)$. An intuitive explanation is that the "empirically" estimated marker distribution does not take advantage of the structure imposed by the risk model.

2.3 *Estimation in a Cohort Design*

The semiparametric methods were developed for case-control designs but can nevertheless be applied to a cohort study as well by plugging in the sample prevalence $\hat{\rho} = n_D/n$. Let $\hat{\alpha}$, $\hat{\beta}$ be the MLE of α, β by applying the logistic regression model $\text{logit}\{P(D=1|Y)\} = \alpha + \beta^T r(Y) + \log\left(\frac{n_{\bar{D}}}{n_D}\right)$ to the cohort sample. The last term is included here in order to make notation, and definition of α in particular, consistent with the previous subsection. For $y \in \mathcal{R}$, the semiparametric "empirical" estimator of F becomes

$$\tilde{F}(y) = \frac{n_D}{n}\frac{1}{n_D}\sum_{i=1}^{n_D} I(Y_{Di} \le y) + \frac{n_{\bar{D}}}{n}\frac{1}{n_{\bar{D}}}\sum_{i=1}^{n_{\bar{D}}} I(Y_{\bar{D}i} \le y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \le y),$$

while the semiparametric maximum likelihood estimator is

$$\begin{aligned} \hat{F}(y) &= \frac{n_D}{n}\frac{1}{n_{\bar{D}}}\sum_{i=1}^{n}\frac{\exp\left\{\hat{\alpha} + \hat{\beta}^T r(Y_i)\right\} I(Y_i \le y)}{1 + \frac{n_D}{n_{\bar{D}}}\exp\left\{\hat{\alpha} + \hat{\beta}^T r(Y_i)\right\}} + \frac{n_{\bar{D}}}{n}\frac{1}{n_{\bar{D}}}\sum_{i=1}^{n}\frac{I(Y_i \le y)}{1 + \frac{n_D}{n_{\bar{D}}}\exp\left\{\hat{\alpha} + \hat{\beta}^T r(Y_i)\right\}} \\ &= \frac{1}{n}\sum_{i=1}^{n} I(Y_i \le y). \end{aligned}$$

That is, $\hat{F}$ and $\tilde{F}$ calculated from a cohort sample are the same as the empirical distribution function. This is also true for a case-control sample where the proportion of cases is equal to $\hat{\rho}$. Consequently the two semiparametric estimators of the predictiveness curve developed in Section 2 when applied to a cohort study is the same as the semiparametric estimator developed in Huang et al. (2007). That is our methods generalize those methods to case-control designs. Of course, the asymptotic theory presented next for the case-control design does not apply to a cohort design since n_D and $n_{\bar{D}}$ are fixed in the former but random in the latter.

3. Asymptotic Theory for Semiparametric Estimators

We present asymptotic theory for the semiparametric estimators defined in Section 2 as well as some consequent attractive properties. We assume the following conditions hold:

(i) $G(s, Y)$ is differentiable with respect to s and Y at $s = \theta$, $Y = F^{-1}(v)$;

(ii) $\partial G^{-1}(s, p)/\partial s$ exists at $s = \theta$;

(iii) for $0 < a < b < 1$, F has continuous positive density f on $[F^{-1}(a) - \epsilon, F^{-1}(b) + \epsilon]$ for some $\epsilon > 0$;

(iv) $\hat{\rho}$ is either estimated from a cohort or is equal to the true ρ.

Asymptotic theory for the semiparametric maximum likelihood predictiveness curve estimator is presented in Theorems 1 and 2 (proof given in Huang and Pepe (2008)). Note that variability in $\hat{\rho}$ is incorporated into asymptotic variances of the predictiveness curve estimators.

Theorem 1 As $n \to \infty$, $\sqrt{n}\left\{\hat{R}(v) - R(v)\right\}$ converges to a normal random variable with mean zero and variance

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

■

Theorem 2 As $n \to \infty$, $\sqrt{n}\left\{\hat{R}^{-1}(p) - R^{-1}(p)\right\}$ converges to a normal random variable with

mean zero and variance

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

■

Observe that when $v = R^{-1}(p)$, $\left\{\partial R^{-1}(p)/\partial\theta\right\}\left\{\partial R(v)/\partial v\right\} = \left\{\partial v/\partial\theta\right\}\left\{\partial R(v)/\partial v\right\} = \partial R(v)/\partial\theta$, thus $\Sigma_{1M}(v) = \left\{\partial R(v)/\partial v\right\}^2 \Sigma_{2M}(p)$. That is the variance of $\hat{R}(v)$ and its inverse are related by a factor equal to the derivative of $R(v)$. Intuitively a perturbation in $R(v)$ can be approximated by $R'(v)$ times a perturbation in $R^{-1}(p)$. Using an analogous approach, we can prove asymptotic theory for the semiparametric "empirical" estimators.

Theorem 3 As $n \to \infty$, $\sqrt{n}\left\{\tilde{R}(v) - R(v)\right\}$ converges to a normal random variable with mean zero and variance

$$\begin{aligned}\Sigma_{1E}(v) &= \left\{\frac{\partial R(v)}{\partial v}\right\}^2 \text{var}\left(\sqrt{n}\left[\tilde{F}\left\{F^{-1}(v)\right\} - v\right]\right) + \left(\frac{\partial R(v)}{\partial\theta}\right)^T \text{var}\left\{\sqrt{n}(\hat{\theta}-\theta)\right\}\left(\frac{\partial R(v)}{\partial\theta}\right) \\ &+ 2\left(\frac{\partial R(v)}{\partial\theta}\right)^T \text{cov}\left(\sqrt{n}(\hat{\theta}-\theta), \sqrt{n}\left[\tilde{F}\left\{F^{-1}(v)\right\} - v\right]\right)\left\{\frac{\partial R(v)}{\partial v}\right\}.\end{aligned}$$

■

Theorem 4 As $n \to \infty$, $\sqrt{n}\left\{\tilde{R}^{-1}(p) - R^{-1}(p)\right\}$ converges to a normal random variable with mean zero and variance

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

■

Again, when $v = R^{-1}(p)$, we have $\Sigma_{1E}(v) = \left\{\partial R(v)/\partial v\right\}^2 \Sigma_{2E}(p)$. Analytical forms for components for these variances can be found in Huang and Pepe (2008).

Note that estimating F using the maximum likelihood method in a case-control design is a special case of the biased sampling problem. Vardi (1985) developed a nonparametric maximum likelihood estimator of the distribution function in a biased sampling model with known selection weights, for which the large sample theory was provided in Gill et al. (1988). Gilbert et al. (1999) extended this method to allow the weight functions to depend on an unknown finite dimensional parameter θ. Gilbert (2000) demonstrated that the maximum likelihood estimators for θ and $F_{\bar{D}}$ are semiparametric efficient, which implies that our semiparametric maximum likelihood estimators are efficient.

3.1 *Comparison of Efficiency between the Two Semiparametric Estimators*

It can be shown that the asymptotic covariance between $\hat{\theta}$ and the estimator of F is the same for the two semiparametric procedures (Huang and Pepe, 2008). This is expected according to the convolution theorem (van der Vaart, 1998, theorem 25.20) given the fact that $\hat{\theta}$ is the semiparametric efficient estimator. Thus the difference in asymptotic variance between $\hat{R}(v)$ and $\tilde{R}(v)$ or between $\hat{R}^{-1}(p)$ and $\hat{R}^{-1}(p)$ is completely attributed to the difference in asymptotic variance between $\hat{F}$ and $\tilde{F}$. The latter can be shown to be positively proportional to the asymptotic variance of $\sqrt{n}\left\{\hat{F}_{\bar{D}} - \tilde{F}_{\bar{D}}\right\}$ (Huang and Pepe, 2008). Thus, as expected, $\hat{R}(v)$ and $\hat{R}^{-1}(p)$ are asymptotically more efficient than $\tilde{R}(v)$ and $\tilde{R}^{-1}(p)$.

4. Nonparametric Estimator

To this point we have estimated the predictiveness curve semiparametrically by assuming a parametric risk model but leaving the control distribution unspecified. A more robust approach is to estimate the risk model nonparametrically under the monotone increasing risk assumption.

4.1 *Estimation of the Risk Model Using Isotonic Regression*

We compute the nonparametric maximum likelihood estimator for the risk function subject to monotonicity using isotonic regression (Barlow et al., 1972). A heuristic explanation of the algorithm in this particular circumstance was given by Lloyd (2002). 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. We then estimate the risk function in the population according to Bayes' theorem

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

4.2 *Estimation of the Marker Distribution and the Predictiveness Curve*

We can estimate F_D and $F_{\bar{D}}$ empirically with $\tilde{F}_{\bar{D}}$ and $\tilde{F}_D$ and calculate $\tilde{F} = \hat{\rho}\tilde{F}_D + (1-\hat{\rho})\tilde{F}_{\bar{D}}$. The nonparametric "empirical" estimators of $R(v)$ and $R^{-1}(p)$ are

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

Alternatively, we can incorporate the estimated risk function into estimation of the marker distribution, as was done for the semiparametric procedure. Lloyd (2002) 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 $\hat{w}(Y) = \hat{P}(D=1|Y,S)/\hat{P}(D=0|Y,S)$. Let κ denote $\{k:\hat{w}(Y_k)=\infty\}$, Lloyd (2002) showed that by maximizing, $\mathcal{L}(F_{\bar{D}}) = \prod_{i=1}^{n_{\bar{D}}} f_{\bar{D}}(Y_{\bar{D}i})\prod_{j=1}^{n_D} f_D(Y_{Dj}) = \prod_{i=1}^{n} f_{\bar{D}}(Y_i)\prod_{j=1}^{n_D}\frac{\hat{w}(Y_{Dj})}{\mu}$ with μ a normalizing factor, the estimators of $f_{\bar{D}}$ and f_D are

$$\hat{f}_{\bar{D}}(Y_k) = \begin{cases} \hat{\mu}/(n_D\hat{w}(Y_k)+n_{\bar{D}}\hat{\mu}) & k\notin\kappa \\ 0 & k\in\kappa \end{cases}, \qquad \hat{f}_D(Y_k) = \begin{cases} \hat{w}(Y_k)\hat{f}_{\bar{D}}(Y_k)/\hat{\mu} & k\notin\kappa \\ 1/n_D & k\in\kappa \end{cases},$$

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

$$\sum_{k\notin\kappa}\mu/\left\{n_D w(Y_k)+n_{\bar{D}}\mu\right\} = 1, \tag{4}$$

which is monotone increasing in μ. We found that when $P(D = 1|Y, S)$ is estimated using isotonic regression, $\hat{\mu}$ can be written down explicitly as a function of $n_{\bar{D}}$ and n_D, as presented in the following new result, proved in Appendix.

Theorem 5

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

Plugging $\hat{\mu}$ into (4), we have

$$\hat{f}_{\bar{D}}(Y_k) = \begin{cases} \frac{1}{n_{\bar{D}}\{\hat{w}(Y_k)+1\}} & k \notin \kappa \\ 0 & k \in \kappa \end{cases}$$

and

$$\hat{f}_D(Y_k) = \begin{cases} \frac{\hat{w}(Y_k)}{n_D\{\hat{w}(Y_k)+1\}} & k \notin \kappa \\ 1/n_D & k \in \kappa \end{cases}.$$

Calculating F with $\hat{F} = \rho\hat{F}_D + (1-\rho)\hat{F}_{\bar{D}}$, the nonparametric model-based estimators of $R(v)$ and $R^{-1}(p)$ are

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

Interestingly, even if the nonparametric "empirical" and model-based procedures described above lead to different estimators of the marker distribution F, the corresponding predictiveness curve estimators are the same (Theorem 6), a fact that is not true for the semiparametric estimators. A proof can also be found in Appendix.

Theorem 6

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

Another appealing property of the nonparametric predictiveness curve is that the area under the curve is always equal to ρ, as shown in Huang (2007). Finally, note that the nonparametric

method can be applied to a cohort design. In that case, the step adjusting for biased sampling in risk estimation is no longer needed.

5. Simulation Studies

We simulate a case-control study under a linear logistic risk model with equal number of $Y_{\bar{D}} \sim N(0,1)$ and $Y_D \sim N(\mu_D, 1)$. Assume $\hat{\rho}$ can be obtained from a phase-one cohort, in which the case-control sample is nested and which is five times the size of the case-control sample. For each simulated sample, variance estimates of the predictiveness estimators are calculated based on analytic formulae, with variability in $\hat{\rho}$ incorporated. Bootstrapping is also performed. Separate resampling of cases and controls is employed, together with resampling of D from the parent cohort. Results for $\rho = 0.2$ and $\mu_D = 1$ are presented in Tables 1 - 3 for $v = 0.1$, 0.3, 0.5, 0.7, and 0.9 and the corresponding $p = R(v)$. We explore varying sample sizes from 100 to 2,000. For each scenario, 5,000 Monte-Carlo simulations are conducted.

The semiparametric estimators for $R(v)$ have minimal bias for sample sizes as small as 100, while the nonparametric estimator has considerable bias at $v = 0.1$ even when $n = 2000$ (Table 1). Asymptotic variances of the semiparametric estimators agree well with the empirical variance from simulations. For this particular simulation setting, the semiparametric "empirical" estimator is fairly efficient relative to the semiparametric maximum likelihood estimator. Both are much more efficient than the nonparametric one, especially when n is large (Table 2). Coverage of the semiparametric 95% Wald confidence intervals using asymptotic or bootstrap variance estimates are fairly close to the nominal level, except for a little undercoverage when $n \leq 200$. Coverage of the nonparametric 95% confidence intervals for $R(v)$ is not as good as its semiparametric counterpart; undercoverage exists even when n is as large as 1000 for small v. A logit transformation lessens the problem of undercoverage in all cases, but creates overcoverage in the nonparametric setting when n is small and v is close to the boundary (results not shown). Confidence intervals using percentiles of the bootstrap distribution seem to have reasonable coverage in all settings, except when $v = 0.1$ and $n \leq 100$ for the nonparametric estimator (Table 3). Results for $R^{-1}(p)$ follow a similar pattern.

Table 1

Bias of the semiparametric and nonparametric estimators and their asymptotic variances in case-control studies for the linear logistic model. Here $n_D = n_{\bar{D}}, n = n_D + n_{\bar{D}}$. Size of the phase-one cohort for estimating $\hat{\rho}$ is $5n$.

		$v = 0.1$	$v = 0.3$	$v = 0.5$	$v = 0.7$	$v = 0.9$
$R(v)$		0.045	0.094	0.15	0.24	0.43
% bias in $\hat{R}(v)$						
$n = 100$	SPMLE[a]	4.47	−0.50	−1.39	−0.82	0.82
	SPE[b]	5.13	−0.40	−1.18	−0.53	0.94
	NPMLE[c]	−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 asymptotic variance of $\hat{R}(v)$						
$n = 100$	SPMLE	0.14	7.51	3.03	−3.79	−4.82
	SPE	3.54	6.50	4.51	−4.30	−4.15
$n = 500$	SPMLE	−0.49	0.25	0.13	−0.85	−0.47
	SPE	−0.91	0.75	0.13	−1.35	−2.06
$n = 2000$	SPMLE	−0.36	0.38	−0.33	−0.35	−0.57
	SPE	−0.04	0.62	−0.06	−0.11	−0.77
		$p = 0.045$	$p = 0.094$	$p = 0.15$	$p = 0.24$	$p = 0.43$
$R^{-1}(p)$		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 asymptotic variance of $\hat{R}^{-1}(p)$						
$n = 100$	SPMLE	11.89	6.92	−3.35	−5.99	9.07
	SPE	11.74	8.07	−3.14	−4.93	9.16
$n = 500$	SPMLE	3.54	0.09	0.002	−2.47	1.62
	SPE	3.50	−0.002	−0.40	−2.71	0.34
$n = 2000$	SPMLE	0.35	0.67	0.33	0.47	0.62
	SPE	0.33	0.71	0.96	1.56	1.24

[a]: semiparametric maximum likelihood estimator
[b]: semiparametric "empirical" estimator
[c]: nonparametric maximum likelihood estimator

Table 2

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 case-control studies for the linear logistic model.

		$v = 0.1$	$v = 0.3$	$v = 0.5$	$v = 0.7$	$v = 0.9$
$R(v)$		0.045	0.094	0.15	0.24	0.43
Asymptotic	SPE[a]	0.99	0.97	0.97	0.90	0.91
$n = 100$	SPE	1.02	0.97	0.97	0.90	0.91
	NPMLE[b]	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
		$p = 0.045$	$p = 0.094$	$p = 0.15$	$p = 0.24$	$p = 0.43$
$R^{-1}(p)$		0.1	0.3	0.5	0.7	0.9
Asymptotic	SPE	0.99	0.97	0.97	0.90	0.91
$n = 100$	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

[a]: semiparametric "empirical" estimator
[b]: nonparametric maximum likelihood estimator

Based on these limited simulations, we recommend use of the percentile bootstrap confidence intervals for the predictiveness curve because they have good coverage and because the corresponding lower and upper confidence bands are also monotone increasing in v.

6. Illustration

We illustrate our methods using a simulated case-control dataset from the Prostate Cancer Prevention Trial, a randomized prospective study of men with PSA < 3.0 ng/mL and 55 years and older who were followed up for 7 years with annual PSA measurements. Thompson et al. (2006) identified 5519 men on the placebo arm of the trial who had undergone prostate biopsy and had a PSA and digital rectal exam (DRE) during the year prior to biopsy and at least 2 PSA values from the 3 years prior to biopsy, and evaluated prostate cancer risk as a function of PSA, PSA velocity and several other variables including age, family history, DRE and prior prostate biopsy. We randomly sampled 250 cases and 250 controls from this study cohort to form the case-control sample. Sample disease prevalence from the study cohort is $\hat{\rho} = 21.9\%$ and is used for estimation

Table 3

Coverage of 95% *percentile bootstrap confidence intervals based on the semiparametric and nonparametric estimators in case-control studies for the linear logistic model (%).*

		$v = 0.1$	$v = 0.3$	$v = 0.5$	$v = 0.7$	$v = 0.9$
$R(v)$		0.045	0.094	0.15	0.24	0.43
$n = 100$	SPMLE[a]	94.24	94.50	95.08	95.86	94.80
	SPE[b]	94.58	94.64	95.28	96.36	94.98
	NPMLE[c]	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

[a]: semiparametric maximum likelihood estimator
[b]: semiparametric "empirical" estimator
[c]: nonparametric maximum likelihood estimator

of the predictiveness curve.

We compare PSA and PSA velocity as risk prediction markers for prostate cancer utilizing the predictiveness curve technique. A logistic regression risk model with a Box-Cox transformation of the marker is employed. The two semiparametric estimators of the predictiveness curves are fairly similar to each other for both PSA and PSA velocity and are much smoother than the nonparametric ones. The semiparametric maximum likelihood and nonparametric estimators are displayed in Figure 1(a). PSA has more steep predictiveness curves, suggesting that it is a better marker for predicting risk of prostate cancer.

The pointwise 95% percentile bootstrap confidence intervals for $R(v)$ constructed from semiparametric maximum likelihood estimators are displayed in Figure 1(b)(c), with variability in $\hat{\rho}$ incorporated. They are much narrower compared to those constructed from nonparametric estimators.

Table 4(a) presents results comparing PSA and PSA velocity with respect to risk percentiles, $R(v)$, for $v = 10\%$ and 90%, and risk stratum sizes, $R^{-1}(p)$, for a low risk threshold 10% and a high risk threshold 30%. P-values for comparing markers are based on bootstrap variance estimates. Using the semiparametric methods we conclude that PSA is a significantly better risk prediction marker than PSA velocity. Specifically, it is better for predicting high risk (larger $R(0.9)$), better for predicting low risk (smaller $R(0.1)$), and it classifies more people into the low and high risk ranges. In contrast, these conclusions cannot be drawn with the nonparametric methods due to their large sampling variability.

In practice, there may not always be a cohort for estimating prevalence. Oftentimes an investigator plugs in a specific prevalence value and treats it as known. We illustrate application of a sensitivity analysis using our example. We study $\rho = 0.165$ and $\rho = 0.274$ which correspond to a 25% change in $\hat{\rho} = 0.219$. The corresponding predictiveness curves are displayed in Figure 1(d). Note that the comparison of predictiveness curves with respect to steepness is not sensitive to perturbation in prevalence. PSA appears overall to be a better risk prediction marker than PSA velocity in the sense that the risk percentiles vary more. Comparison at particular risk thresholds, on the other hand, are affected by prevalence. For example, when $\rho = 0.165$, based on

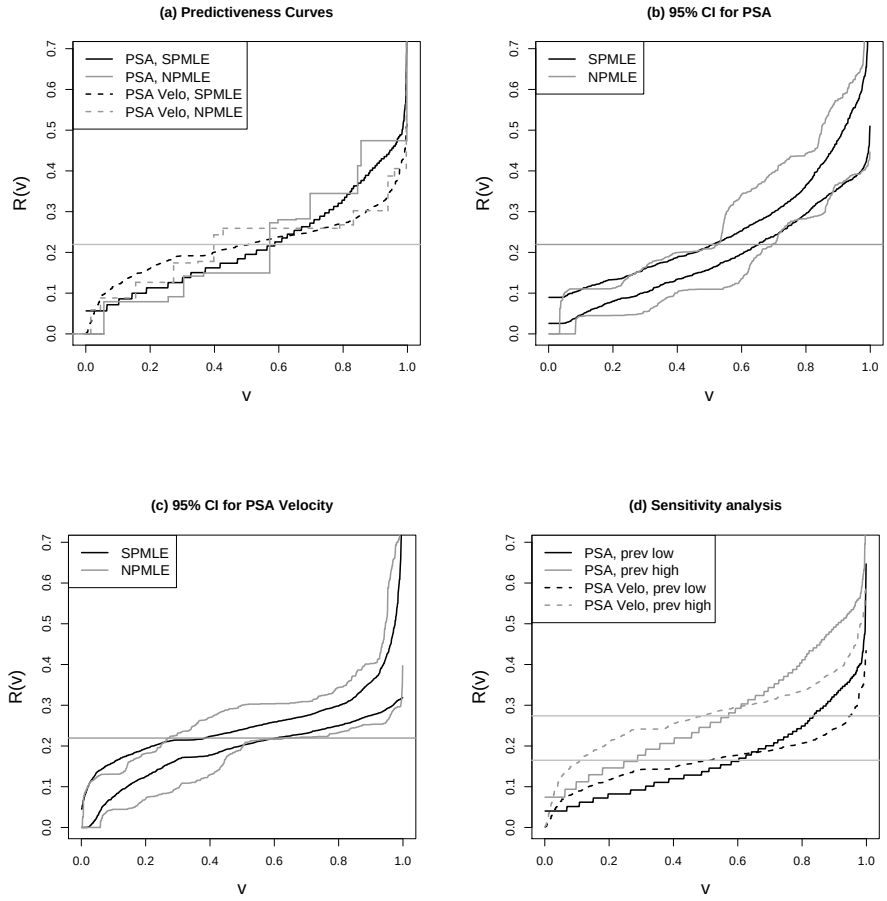

Figure 1. (a) The predictiveness curves for PSA and PSA velocity for predicting prostate cancer;(b)(c) their 95% pointwise confidence intervals constructed from percentiles of the bootstrap distribution; and (d) sensitivity analysis. The horizontal lines indicate disease prevalences plugged in. SPMLE: semiparametric maximum likelihood estimator; NPMLE: nonparametric maximum likelihood estimator.

Table 4

Comparisons between (a) PSA and PSA velocity and (b) between PSA and PSA plus other risk factors for predicting risk of prostate cancer.

(a)

Measure	Method	PSA		PSA Velocity		pvalue
		Est	95% CI	Est	95% CI	
$R(0.1)$	NPMLE[a]	0.079	(0.043,0.110)	0.088	(0.044, 0.131)	0.730
	SPMLE[b]	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

(b)

Measure	Method	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

[a]: nonparametric maximum likelihood estimator
[b]: semiparametric maximum likelihood estimator

the semipmarametric maximum likelihood procedure, PSA assigns significantly more people into low risk range than PSA velocity, with estimates of $R^{-1}(0.1)$ being 31.3% and 13.5% respectively (p-value< 0.001). PSA is also a significantly better marker for predicting high risk than PSA velocity, with estimates of $1 - R^{-1}(0.3)$ being 12.5% and 2.9% respectively (p-value < 0.001). In contrast, when $\rho = 0.263$, estimates of $R^{-1}(0.1)$ become 9.7% and 3.8% for PSA and PSA velocity, and estimates of $1 - R^{-1}(0.3)$ are 38.9% and 36.9% respectively. None of the comparisons are significant (p-value=0.192 and 0.736). The comparison with respect to percentage classified into the equivocal risk range is significant when $\rho = 0.165$ (p-value < 0.001) but not when $\rho = 0.274$ (p-value=0.375).

7. Extension of Semiparametric Estimation

The semiparametric estimators can be extended naturally to accommodate multiple predictors or to settings where the monotone increasing risk assumption is not true. Let $F_R, F_{DR}, F_{\bar{D}R}$ indicate the cumulative distribution functions of $Risk(Y)$ in the general, diseased, and nondiseased

populations respectively. We first calculate $Risk(Y_i)$ as the predicted risk for subject i based on a a standard logistic regression model with offset $\log\{(1-\hat{\rho})/\hat{\rho}\}$. Then estimate F_R according to $\hat{\rho}F_{DR} + (1-\hat{\rho})F_{\bar{D}R}$. Note that F_{DR} and $F_{\bar{D}R}$ can be estimated "empirically" using

$$\begin{aligned}
\tilde{F}_{DR}(p) &= \frac{1}{n_D}\sum_{i=1}^{n_D} I\left\{\widehat{Risk}(Y_{Di}) \le p\right\}, \\
\tilde{F}_{\bar{D}R}(p) &= \frac{1}{n_{\bar{D}}}\sum_{i=1}^{n_{\bar{D}}} I\left\{\widehat{Risk}(Y_{\bar{D}i}) \le p\right\},
\end{aligned}$$

or based on the semiparametric maximum likelihood method

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

where $\widehat{\mathcal{L}\mathrm{R}}_R\left\{\widehat{Risk}(Y_i)\right\} = \widehat{Risk}(Y_i)/\left\{1-\widehat{Risk}(Y_i)\right\} \times (1-\hat{\rho})/\hat{\rho}$. Compute semiparametric maximum likelihood estimators $\hat{R}(v) = \hat{F}_R^{-1}(v)$ and $\hat{R}^{-1}(p) = \hat{F}_R(p)$ and semiparametric "empirical" estimators $\tilde{R}(v) = \tilde{F}_R(p)$ and $\tilde{R}^{-1}(p) = \tilde{F}_R(p)$. Following similar arguments as in the single marker setting, asymptotic distribution theory presented in Theorems 7 and 8 can be derived. In practice, since estimation of asymptotic variance involves both numerical differentiation and nonparametric density estimation, we rely on resampling techniques for inference.

Theorem 7

As $n \to \infty$, $\sqrt{n}\left\{\hat{R}^{-1}(p) - R^{-1}(p)\right\}$ converges to a normal random variable with mean zero and variance

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

and $\sqrt{n}\left\{\hat{R}(v)-R(v)\right\}$ converges to a normal random variable with mean zero and variance

$$\Sigma_{1M.R}(v) \;=\; \left\{\frac{\partial R(v)}{\partial v}\right\}^2 \Sigma_{2M.R}\{R(v)\},$$

where

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

■

Theorem 8

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

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

and $\sqrt{n}\left\{\hat{R}(v)-R(v)\right\}$ converges to a normal random variable with mean zero and variance

$$\Sigma_{1E.R}(v) \;=\; \left\{\frac{\partial R(v)}{\partial v}\right\}^2 \Sigma_{2E.R}\{R(v)\},$$

where

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

■

Components of these asymptotic variance terms can be calculated similar to those in the single marker setting by replacing the marker distribution with the risk distribution. Note that when Y has only one component and under the monotone increasing risk assumption, the variance

expressions in Theorems 7 and 8 reduce to those in Theorems 1-4, i.e. $Q_M(p) = \hat{F}\{G^{-1}(p)\}$ and $Q_E(p) = \tilde{F}\{G^{-1}(p)\}$.

7.1 *Illustration*

We illustrate using the simulated case-control sample described in Section 6. We compare the logistic risk model based on PSA alone with a more comprehensive risk model that combines Box-Cox transformation of PSA and other risk factors, including family history of prostate cancer, DRE, and previous negative biopsies. All of these factors are highly statistically significant (Thompson et al., 2006).

Figure 2(a) displays the semiparametric maximum likelihood estimators of the predictiveness curves for PSA alone and for PSA plus other factors (the semiparametric "empirical" estimators are similar), for $\hat{\rho} = 21.9\%$. The two risk models have very similar predictiveness curves. To assess model fit, 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 $k \in \{1, \ldots, 10\}$, 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 estimated according to

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

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 plot. This comparison of observed risk and average risk within deciles of modeled risk is the basis for the Hosmer-Lemeshow test (Hosmer and Lemeshow, 1980). Our approach provides a graphical display that generalizes to case-control data. In Figure 2(a), there seems to be a slight lack of fit in the middle of the curve but good fit at both ends which are the regions of primary interest. Confidence intervals for $R(v)$ constructed with the semiparametric maximum likelihood estimators are presented in Figure 2(b). Variabilities from the two risk models appear to be similar in magnitude.

Detailed results comparing the two models for predictiveness are shown in Table 4(b). Briefly, the percentages of people classified into the low, high, or equivocal risk ranges are not significantly

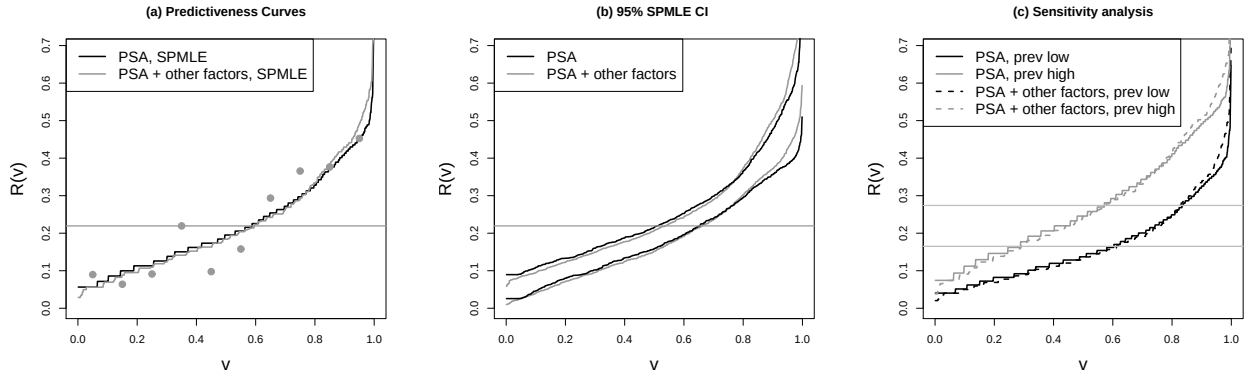

Figure 2. (a) The semiparametric predictiveness curves for PSA and PSA plus other factors for predicting risk prostate cancer,the dots are average risk within deciles of modeled risk based on the latter model; (b) their 95% pointwise confidence intervals using percentiles of the bootstrap distribution; and (c) sensitivity analysis. The horizontal lines indicate disease prevalences. SPMLE: semiparametric maximum likelihood estimator; NPMLE: nonparametric maximum likelihood estimator.

different between the two models, nor are the 10^{th} and 90^{th} percentiles of risk. 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.

Again, we conducted sensitivity analysis with fixed $\rho = 0.165$ and $\rho = 0.274$, the corresponding predictiveness curves are displayed in Figure 2(c). Different choices of ρ do not change our conclusion about the comparison between the two risk models.

8. Concluding Remarks

In this article we have developed flexible semiparametric and nonparametric estimators of the predictiveness curve for case-control studies. This is particularly valuable for evaluating a risk prediction marker or model early in its development when case-control designs are most common. Both estimators are easy to compute: risk models can be estimated utilizing standard statistical procedures, and risk distributions can be calculated easily based on analytic formulae. The nonparametric estimator is highly robust but demands very large sample sizes for reasonable precision. This begs for introduction of smoothing techniques in future research. The semipara-

metric estimators, in contrast, have satisfactory finite-sample performance. Another approach to estimate the predictiveness curve is based on its relationship with the ROC curve. This is currently under investigation (Huang and Pepe, 2007).

In terms of comparison between the two semiparametric estimators, their validity both rely upon assumptions about the risk model. If the risk model is misspecified, bias can be introduced into both estimators. This, however, may not be a big concern since the risk model can be made highly flexible using techniques such as regression splines. Given a well specified model, the semiparametric maximum likelihood estimator is more efficient than the its "empirical" counterpart. Asymptotic relative efficiency of the former versus the latter is a complicated function of the disease prevalence, separation between cases and controls, case-control sampling ratio, and the quantile of interest. In the examples, we showed that when the disease prevalence is medium, the two estimators have similar efficiency. We note that for rare diseases, using the model-based approach may achieve considerable efficiency gains compared to the "empirical" approach for certain quantiles (Huang, 2007).

An important use of the asymptotic theory is to guide study design. To design an efficient case-control study for evaluating a risk model, the optimal case-control sampling ratio is dictated by the disease prevalence, separation between cases and controls, and the measure that are of primary interest. A detailed study can be found in Huang (2007).

Comparing markers or models for their risk stratification capacity is of great significance in medical practice. Researchers are often interested in whether additional risk factors which may be hard to measure can lead to a significant improvement in utility compared to an existing model. More research on methods to evaluate incremental value is warranted. Methods described here might be adapted for such purposes. An important issue pertaining to the general risk model methods is overfitting when the number of predictors gets large relative to the sample size. The effectiveness of cross-validation techniques for addressing biases needs to be investigated.

Pepe et al. (2008a) 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 $\text{TPR}(p) = P(Risk(Y) > p|D = 1)$) and the proportion of non-diseased subjects incorrectly classified (the false positive rate $\text{FPR}(p) = P(Risk(Y) > p|D = 0)$), according to the classification rule '$Risk(Y) > p$'. Our semiparametric and nonparametric procedures developed in this manuscript yield estimators of $F_D, F_{\bar{D}}$ and F as by-products, which can be directly plugged into $\text{TPR}(p) = F_D\{F^{-1}(p)\}$ and $\text{FPR}(p) = F_{\bar{D}}\{F^{-1}(p)\}$ for estimation of these quantities. Asymptotic theory for corresponding semiparametric estimators can be developed in a similar fashion.

Acknowledgments

The authors are grateful for support provided by NIH grants GM-54438 and NCI grants CA86368.

References

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59(1),** 19-35.

Anderson, K.M. and Odell, P.M. and Wilson, P.W. and Kannel, W.B. (1991). Cardiovascular disease risk profiles. *American Heart Journal* **121,** 293-298.

Baker, S. G., Kramer, B. S., and Srivastava, S. (2002). Markers for early detection of cancer: Statistical guidelines for nested case-control studies. *BMI Medical Research Methodology*, **2:4**.

Barlow, R. E. and Bartholomew, D. J. and Bremner, J. M. and Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression.* Wiley, London.

Bickel, P. J. and Klaassen, C. A. J. and Ritov, Y. and Wellner, J. A. (2002). *Efficient and Adaptive Estimation for Semiparametric Models.* Springer.

Billingsley, P. (1999). *Convergence of Probability Measures.* Wiley, New York.

Breslow, N. E. and Robins, J. M. and Wellner, J. A. (2000). On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli* **6(3),**447-455.

Bura, E. and Gastwirth, J. L. (2001). The binary regression quantile plot: assessing the importance of predictors in binary regression visually. *Biometrical Journal* **43 (1),** 5-21.

Cole, T. J. and Green, P. J. (1992). Smoothing reference centile curves: The LMS method and penalized likelihood. *Statistics in Medicine* **11(165),** 1305-1319.

Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* **115,** 928-935.

Gail, M.H. and Brinton, L.A. and Byar, D.P. and Corle, D.K. and Green, S.B. and Schairer, C. and Mulvihill, J.J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually . *JNCI* **81(24),** 1879-1886.

Gail, M. H. and Pfeiffer, R. M. (2005). On criteria for evaluating models of absolute risk. *Biostatistics* **6(2),** 227-239.

Gilbert, P. B. (2000). Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. *Annals of Statistics* **28(1),** 151-194.

Gilbert, P. B. and Lele, S. and Vardi, Y. (1999). Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika* **86,** 27-43.

Gill, R. D. and Vardi, Y. and Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Annals of Statistics* **16,** 1069-1112.

Green, D. M. and Swets, J. A. (1966). *Signal detection theory and psychophysics.* Wiley, New York.

Hosmer, D.W. and Lemesbow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics-Theory and Methods* **9(10),** 1043-1069.

Huang, Y. (2007). Evaluating the predictiveness of continuous biomarkers. *UW thesis.*

Huang, Y. and Pepe, M. S. and Feng, Z. (2007). Evaluating the predictiveness of a continuous marker. *Biometrics* **63(4)**, 1181-1188.

Huang, Y. and Pepe, M. S. (2007). A parametric ROC model based approach for evaluating the predictiveness of continuous markers in case-control studies. *UW Biostatistics Working Paper Series* **318**.

Huang, Y. and Pepe, M. S. (2008). Semiparametric and nonparametric methods for evaluating risk prediction markers in case-control studies. *UW Biostatistics Working Paper Series* **318**.

Lloyd, C. J. (2000). Maximum likelihood estimation of misclassification rates of a binomial regression. *Biometrika* **87(3),** 700-705.

Lloyd, C. J. (2002). Estimation of a Convex ROC Curve. *Statistics & Probability Letters* **59,** 99-111.

Pencina, M.J. and D'Agostino Sr., R.B. and D'Agostino Jr., R.B. and Vasan, R.S. (2007). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine* (In Press).

Pepe, M. S. (2000). An interpretation for the ROC curve and inference using GLM procedures. *Biometrics* **56,** 352-359.

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

Pepe, M. S. and Etzioni, R. and Feng, Z. and Potter, J. D. and Thompson, M. L. and Thornquist, M. and Winget, M. and Yasui, Y. (2001) Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute* **93(14),** 1054-1061.

Pepe, M. S. and Feng, Z. and Huang, Y. and Longton, G. M. and Prentice, R. and Thompson, I. M. and Zheng, Y. (2008a). Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology* **167(3),** 362-368.

Pepe, M. S. and Feng, Z. and Janes, H. and Bossuyt, P. M. and Potter, J. D. (2008b). Pivotal evaluation of the accuracy of a classification biomarker: the PRoBE study design. *Journal of the National Cancer Institute*, In Press.

Prentice, R. L. and Pyke, R. (1979). Logistic Disease Incidence Models and Case-Control Studies. *Biometrika* **66(3),** 403-411.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* **22(1),** 300-325.

Qin, J. and Zhang, J. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika* **84(3),** 609-618.

Qin, J. and Zhang, J. (2003). Using logistic regression procedures for estimating receiver operating characteristic curves. *Biometrika* **93(3),** 585-596.

Ransohoff, D. F. (2007). How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. *Journal of Clinical Epidemiology* **60,** 1205-1219.

Simon, R. (2005). Roadmap for developing and validating therapeutically relevant genomic classifiers. *Journal of Clinical Oncology* **23(29),** 7332-7341.

Thompson, I. M. and Pauler Ankerst, D. and Chi, C. (2006). Assessing prostate cancer risk: results from the prostate cancer prevention trial. *Journal of the National Cancer Institute* **98,** 529-534.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Vardi, Y. (1985). Empirical distributions in selection bias models. *Annals of Statistics* **13,** 178-203.

Zhang, B. (2000). Quantile estimation under a two-sample semi-parametric model. *Bernoulli* **6(3),** 491-511.

9. Appendix

9.1 *Components for Asymptotic Variances of Semiparametric Estimators in Theorems 1-4*

Consider $\hat{\rho}$ 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. Denote

$$\begin{aligned}
\hat{F}^{\star}(t) &= \rho\hat{F}_D(t) + (1-\rho)\hat{F}_{\bar{D}}(t) \\
\tilde{F}^{\star}(t) &= \rho\tilde{F}_D(t) + (1-\rho)\tilde{F}_{\bar{D}}(t), \\
\hat{\theta}^{\star} &= \begin{pmatrix} \hat{\theta}_{0S} + \log\left(\frac{n_{\bar{D}}}{n_D}\frac{\rho}{1-\rho}\right) \\ \hat{\theta}_{1S} \end{pmatrix},
\end{aligned}$$

then the following Lemma holds.

Lemma 0

$$\begin{aligned}
&\mathrm{var}\left[\sqrt{n}\left\{\hat{F}(t) - F(t)\right\}\right] = \left\{F_D(t) - F_{\bar{D}}(t)\right\}^2 \rho(1-\rho)/\lambda + \mathrm{var}\left[\sqrt{n}\left\{\hat{F}^{\star}(t) - F(t)\right\}\right], \\
&\mathrm{var}\left[\sqrt{n}\left\{\tilde{F}(t) - F(t)\right\}\right] = \left\{F_D(t) - F_{\bar{D}}(t)\right\}^2 \rho(1-\rho)/\lambda + \mathrm{var}\left[\sqrt{n}\left\{\tilde{F}^{\star}(t) - F(t)\right\}\right], \\
&\mathrm{var}\left\{\sqrt{n}\left(\hat{\theta} - \theta\right)\right\} = \begin{pmatrix} \frac{1}{\lambda\rho(1-\rho)} & 0 \\ 0 & 0 \end{pmatrix} + \mathrm{var}\left\{\sqrt{n}\left(\hat{\theta}^{\star} - \theta\right)\right\} \\
&\mathrm{cov}\left[\sqrt{n}\left(\hat{\theta} - \theta\right), \sqrt{n}\left\{\hat{F}(t) - F(t)\right\}\right] = \frac{F_D(t) - F_{\bar{D}}(t)}{\lambda} + \mathrm{cov}\left[\sqrt{n}\left(\hat{\theta}^{\star} - \theta\right), \sqrt{n}\left\{\hat{F}^{\star}(t) - F(t)\right\}\right].
\end{aligned}$$

■

Suppose

$$\mathrm{logit}\left\{G(\theta, Y)\right\} = \theta_0 + \theta_1^T r(Y) = \alpha + \log\left(\frac{\rho}{1-\rho}\right) + \beta^T r(Y).$$

Let $\hat{\alpha}$, $\hat{\beta}$ be the MLE of α, β based on the semiparametric likelihood (3). Let $\eta = n_D/n_{\bar{D}}$. The following Lemmas are needed to calculate the asymptotic variances of the semiparametric predictiveness curve estimators. Lemma 1 characterizes the distribution of $\hat{\theta}^{\star}$, a result proved in Prentice and Pyke (1979), Qin and Zhang (1997) and Zhang (2000). Lemmas 2 to 7 present

asymptotic theory for $\hat{F}^\star$, $\tilde{F}^\star$ and the correlation between them and $\hat{\theta}^\star$. Proofs of Lemmas 2-7 are given in Appendix B.

Lemma 1 Let

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

$$A(t) = \begin{pmatrix} A_0(t) & A_1(t)^T \\ A_1(t) & A_2 \end{pmatrix},$$

and $A = A(\infty)$. If A^{-1} exists,

$$\sqrt{n}\begin{pmatrix} \hat{\alpha}-\alpha \\ \hat{\beta}-\beta \end{pmatrix} \xrightarrow{d} N(0,\Sigma),$$

where

$$\Sigma = \frac{1+\eta}{\eta}\left\{A^{-1} - \begin{pmatrix} 1+\eta & 0 \\ 0 & 0 \end{pmatrix}\right\}.$$

■

Lemma 2 As $n\to\infty$, $\sqrt{n}\left\{\hat{F}^\star(t)-F(t)\right\}$ converges weakly in $D[-\infty,\infty]$ to $W(t)$, a Gaussian process with mean 0 and covariance function specified by

$$
\begin{aligned}
& E\{W_M(s)W_M(t)\} \\
= \; & (1-\rho)^2(1+\eta)\{F_{\bar{D}}(s\wedge t) - F_{\bar{D}}(s)F_{\bar{D}}(t)\} + \rho^2\frac{1+\eta}{\eta}\{F_D(s\wedge t) - F_D(s)F_D(t)\} \\
& - \left(\frac{1+\eta}{\eta}\right)\{\rho-(1-\rho)\eta\}^2\left\{A_0(s\wedge t) - \left(A_0(s), A_1(s)^T\right)A^{-1}\begin{pmatrix} A_0(t) \\ A_1(t)\end{pmatrix}\right\}.
\end{aligned}
$$

■

Lemma 3

$$
\sqrt{n}\left(\hat{F}^{\star -1}(v) - F^{-1}(v)\right) \xrightarrow{d} W\{F^{-1}(v)\}/f\{F^{-1}(v)\}
$$

on $D[a,b]$, where W is the Gaussian process with continuous sample path as specified in Lemma 2. ■

Lemma 4 As $n\to\infty$, $\sqrt{n}\left\{\tilde{F}^{\star}(t) - F(t)\right\}$ converges weakly in $D[-\infty,\infty]$ to $W_E(t)$, a Gaussian process with mean 0 and covariance function specified by

$$
\begin{aligned}
& E\{W_E(s)W_E(t)\} \\
= \; & (1-\rho)^2(1+\eta)\{F_{\bar{D}}(s\wedge t) - F_{\bar{D}}(s)F_{\bar{D}}(t)\} + \rho^2\frac{1+\eta}{\eta}\{F_D(s\wedge t) - F_D(s)F_D(t)\}.
\end{aligned}
$$

■

Lemma 5

$$
\sqrt{n}\left(\tilde{F}^{\star -1}(v) - F^{-1}(v)\right) \xrightarrow{d} W_E\{F^{-1}(v)\}/f\{F^{-1}(v)\}
$$

on $D[a,b]$, where W_E is the Gaussian process with continuous sample path as specified in Lemma 4. ■

Lemma 6

$$\begin{aligned}
&\operatorname{cov}\left[\sqrt{n}\begin{pmatrix}\hat{\alpha}-\alpha\\ \hat{\beta}-\beta\end{pmatrix}, \sqrt{n}\left\{\hat{F}^{\star}(t)-F(t)\right\}\right] = \operatorname{cov}\left[\sqrt{n}\begin{pmatrix}\hat{\alpha}-\alpha\\ \hat{\beta}-\beta\end{pmatrix}, \sqrt{n}\left\{\tilde{F}^{\star}(t)-F(t)\right\}\right] \\
= &\frac{1+\eta}{\eta}\left[\{\rho-\eta(1-\rho)\}A^{-1}\begin{pmatrix}A_0(t)\\ A_1(t)\end{pmatrix} - \begin{pmatrix}\rho F_D(t)-\eta(1-\rho)F_{\bar{D}}(t)\\ 0\end{pmatrix}\right] + o_p(1)
\end{aligned}$$

■

Lemma 7

$$\begin{aligned}
&\operatorname{cov}\left[\sqrt{n}\begin{pmatrix}\hat{\alpha}-\alpha\\ \hat{\beta}-\beta\end{pmatrix}, \sqrt{n}\left\{\hat{F}^{\star-1}(v)-F^{-1}(v)\right\}\right] = \operatorname{cov}\left[\sqrt{n}\begin{pmatrix}\hat{\alpha}-\alpha\\ \hat{\beta}-\beta\end{pmatrix}, \sqrt{n}\left\{\tilde{F}^{\star-1}(v)-F^{-1}(v)\right\}\right] \\
= &\frac{1}{f\left\{F^{-1}(v)\right\}}\left\{\frac{1+\eta}{\eta}\{\rho-\eta(1-\rho)\}A^{-1}\begin{pmatrix}A_0\left\{F^{-1}(v)\right\}\\ A_1\left\{F^{-1}(v)\right\}\end{pmatrix}\right. \\
- &\left.\frac{1+\eta}{\eta}\begin{pmatrix}\rho F_D\{F^{-1}(v)\}-\eta(1-\rho)F_{\bar{D}}\{F^{-1}(v)\}\\ 0\end{pmatrix}\right\} + o_p(1)
\end{aligned}$$

■

9.2 *Proofs of Lemmas and Theorems for Semiparametric Estimators*

9.2.1 Proof of Lemma 0

$$\begin{aligned}
&\sqrt{n}\left\{\hat{F}(t)-F(t)\right\} \\
= &\sqrt{n}\left\{\hat{F}(t)-F^{\star}(t)\right\} + \sqrt{n}\left\{\hat{F}^{\star}(t)-F(t)\right\} \\
\simeq &\sqrt{n}\{\hat{\rho}F_D(t)+(1-\hat{\rho})F_{\bar{D}}(t)-F(t)\} + \sqrt{n}\left\{\hat{F}^{\star}(t)-F(t)\right\}. \qquad (5)
\end{aligned}$$

Since the two terms in (5) are asymptotically uncorrelated, we have

$$\begin{aligned} & \operatorname{var}\left[\sqrt{n}\left\{\hat{F}(t)-F(t)\right\}\right] \\ \simeq\ & \operatorname{var}\left[\sqrt{n}\left\{F_D(t)-F_{\bar{D}}(t)\right\}^2\operatorname{var}(\hat{\rho})\right]+\operatorname{var}\left[\sqrt{n}\left\{\hat{F}^{\star}(t)-F(t)\right\}\right] \\ =\ & \left\{F_D(t)-F_{\bar{D}}(t)\right\}^2\rho(1-\rho)/\lambda+\operatorname{var}\left[\sqrt{n}\left\{\hat{F}^{\star}(t)-F(t)\right\}\right]. \end{aligned}$$

The proofs of other results in this Lemma follow similar arguments.

9.2.2 Proof of Lemma 2 Let

$$\begin{aligned} H_1(t) &= \frac{1}{n_{\bar{D}}}\sum_{i=1}^{n}\frac{I(Y_i\le t)}{1+\eta\exp\left\{\alpha+\beta^T r(Y_i)\right\}}, \\ H_2(t) &= \frac{1}{n_{\bar{D}}}\sum_{i=1}^{n}\frac{\exp\left\{\alpha+\beta^T r(Y_i)\right\}I(Y_i\le t)}{1+\eta\exp\left\{\alpha+\beta^T r(Y_i)\right\}}, \\ H_3(t) &= \frac{1}{n}\left(A_0(t),A_1(t)^T\right)S^{-1}\begin{pmatrix}\frac{\partial l(\alpha,\beta)}{\partial\alpha}\\ \frac{\partial l(\alpha,\beta)}{\partial\beta}\end{pmatrix}. \end{aligned}$$

By first-order Taylor expansion,

$$\begin{aligned} & \hat{F}^{\star}(t)=(1-\rho)\hat{F}_{\bar{D}}(t)+\rho\hat{F}_D(t) \\ =\ & (1-\rho)\frac{1}{n_{\bar{D}}}\sum_{i=1}^{n}\frac{I(Y_i\le t)}{1+\eta\exp\left\{\hat{\alpha}+\hat{\beta}^T r(Y_i)\right\}}+\rho\frac{1}{n_{\bar{D}}}\sum_{i=1}^{n}\frac{\exp\left\{\hat{\alpha}+\hat{\beta}^T r(Y_i)\right\}I(Y_i\le t)}{1+\eta\exp\left\{\hat{\alpha}+\hat{\beta}^T r(Y_i)\right\}} \\ =\ & (1-\rho)\frac{1}{n_{\bar{D}}}\sum_{i=1}^{n}\frac{I(Y_i\le t)}{1+\eta\exp\left\{\alpha+\beta^T r(Y_i)\right\}}+\rho\frac{1}{n_{\bar{D}}}\sum_{i=1}^{n}\frac{\exp\left\{\alpha+\beta^T r(Y_i)\right\}I(Y_i\le t)}{1+\eta\exp\left\{\alpha+\beta^T r(Y_i)\right\}} \\ +\ & \left\{\rho-(1-\rho)\eta\right\}\left(\hat{\alpha}-\alpha,\hat{\beta}^T-\beta^T\right)\begin{pmatrix}\frac{1}{n_{\bar{D}}}\sum_{i=1}^{n}\frac{\exp\left\{\alpha+\beta^T r(Y_i)\right\}I(Y_i\le t)}{\left[1+\eta\exp\left\{\alpha+\beta^T r(Y_i)\right\}\right]^2}\\ \frac{1}{n_{\bar{D}}}\sum_{i=1}^{n}\frac{r(Y_i)\exp\left\{\alpha+\beta^T r(Y_i)\right\}I(Y_i\le t)}{\left[1+\eta\exp\left\{\alpha+\beta^T r(Y_i)\right\}\right]^2}\end{pmatrix}+o_p(n^{-1/2}) \\ =\ & (1-\rho)H_1(t)+\rho H_2(t)+\frac{\left\{\rho-(1-\rho)\eta\right\}}{n}\left(A_0(t),A_1(t)^T\right)S^{-1}\begin{pmatrix}\frac{\partial l(\alpha,\beta)}{\partial\alpha}\\ \frac{\partial l(\alpha,\beta)}{\partial\beta}\end{pmatrix}+o_p\left(n^{-1/2}\right). \end{aligned}$$

We have

$$\begin{aligned}
\mathrm{cov}\{H_1(s), H_1(t)\} &= \frac{1}{n_{\bar{D}}}\left\{F_{\bar{D}}(s\wedge t) - F_{\bar{D}}(s)F_{\bar{D}}(t)\right\} \\
&+ \frac{\eta}{n_{\bar{D}}}\left\{F_{\bar{D}}(s)A_0(t) + F_{\bar{D}}(t)A_0(s)\right\} - \frac{\eta}{n_{\bar{D}}}A_0(s\wedge t) - \frac{\eta(1+\eta)}{n_{\bar{D}}}A_0(s)A_0(t), \\
\mathrm{cov}\{H_1(s), H_2(t)\} &= \frac{1}{n_{\bar{D}}}A_0(s\wedge t) - \frac{1}{n_{\bar{D}}}F_{\bar{D}}(s)A_0(t) - \frac{1}{n_{\bar{D}}}F_D(t)A_0(s) + \frac{1+\eta}{n_{\bar{D}}}A_0(s)A_0(t), \\
\mathrm{cov}\{H_1(s), H_3(t)\} &= \frac{1}{n_{\bar{D}}}A_0(t)\left\{F_{\bar{D}}(s) - (1+\eta)A_0(s)\right\},
\end{aligned}$$

$$\begin{aligned}
\mathrm{cov}\{H_2(s), H_2(t)\} &= \frac{1}{n_D}\left\{F_D(s\wedge t) - F_D(s)F_D(t)\right\} + \frac{1}{n_D}\left\{F_D(s)A_0(t) + F_D(t)A_0(s)\right\} \\
&- \frac{1}{n_D}A_0(s\wedge t) - \frac{1+\eta}{n_D}A_0(s)A_0(t), \\
\mathrm{cov}\{H_2(s), H_3(t)\} &= \frac{1}{n_D}A_0(t)\left\{(1+\eta)A_0(s) - F_D(s)\right\}, \\
\mathrm{cov}\{H_3(s), H_3(t)\} &= \frac{1}{n_D}\left(A_0(s), A_1(s)^T\right)A^{-1}\begin{pmatrix} A_0(t) \\ A_1(t)\end{pmatrix} - \frac{1+\eta}{n_D}A_0(s)A_0(t).
\end{aligned}$$

It can be shown that

$$E\left[(1-\rho)H_1(t) + \rho H_2(t) + \{\rho - (1-\rho)\eta\}H_3(t) - F(t)\right] = (1-\rho)F_{\bar{D}}(t) + \rho F_D(t) - F(t) = 0$$

and that

$$\begin{aligned}
&\mathrm{cov}\left(\sqrt{n}\left[(1-\rho)H_1(s) + \rho H_2(s) + \{\rho - (1-\rho)\eta\}H_3(s)\right],\right. \\
&\left.\sqrt{n}\left[(1-\rho)H_1(t) + \rho H_2(t) + \{\rho - (1-\rho)\eta\}H_3(t)\right]\right) \\
= \;& (1-\rho)^2(1+\eta)\left\{F_{\bar{D}}(s\wedge t) - F_{\bar{D}}(s)F_{\bar{D}}(t)\right\} + \rho^2\frac{1+\eta}{\eta}\left\{F_D(s\wedge t) - F_D(s)F_D(t)\right\} \\
&- \left(\frac{1+\eta}{\eta}\right)\{\rho - (1-\rho)\eta\}^2\left\{A_0(s\wedge t) - \left(A_0(s), A_1(s)^T\right)A^{-1}\begin{pmatrix} A_0(t) \\ A_1(t)\end{pmatrix}\right\}.
\end{aligned}$$

Thus according to standard central limit theorem, the finite-dimensional distribution of $\sqrt{n}\left\{\hat{F}^{\star}(t)-F(t)\right\}$ converges weakly to that of $W(t)$. The asymptotic tightness of the process can be proved by employing the tightness criteria in Billingsley (1999), and has been shown in Gilbert (2000).

9.2.3 Proof of Lemma 3 Let

$$
\begin{aligned}
r_{n1}(v) &= -\left\{F^{-1}\hat{F}^{\star}F^{-1}(v)-F^{-1}FF^{-1}(v)+V_n(v)\right\},\\
r_{n2}(v) &= F^{-1}F\hat{F}^{\star-1}(v)-F^{-1}\hat{F}^{\star}\hat{F}^{\star-1}(v)+F^{-1}\hat{F}^{\star}F^{-1}(v)-F^{-1}FF^{-1}(v),\\
r_{n3}(v) &= F^{-1}\hat{F}^{\star}\hat{F}^{\star-1}(v)-F^{-1}(v),
\end{aligned}
$$

then $\hat{F}^{\star-1}(v)-F^{-1}(v)=V_n(v)+r_{n1}(v)+r_{n2}(v)+r_{n3}(v)$. Following arguments similar to that in Zhang (2000), it can be shown that

$$
\begin{aligned}
sup_{a\le t\le b}|r_{n1}(v)| &= o_p(n^{-1/2}),\\
sup_{a\le t\le b}|r_{n2}(v)| &= o_p(n^{-1/2}),\\
sup_{a\le t\le b}|r_{n3}(v)| &= O_p(n^{-1}),
\end{aligned}
$$

and the convergence of $\hat{F}^{\star-1}(v)-F^{-1}(v)$ follows.

9.2.4 Proof of Lemma 4

$$
E\left\{\tilde{F}^{\star}(t)-F(t)\right\}=(1-\rho)E\left\{\tilde{F}_{\bar{D}}(t)\right\}+\rho E\left\{\tilde{F}_{D}(t)\right\}-F(t)=(1-\rho)F_{\bar{D}}(t)+\rho F_{D}(t)-F(t)=0,
$$

$$
\begin{aligned}
& \operatorname{cov}\left[\sqrt{n}\left\{\tilde{F}^{\star}(s)-F(s)\right\},\sqrt{n}\left\{\tilde{F}^{\star}(t)-F(t)\right\}\right]\\
= \;& n(1-\rho)^2\operatorname{cov}\left\{\tilde{F}_{\bar{D}}(s),\tilde{F}_{\bar{D}}(t)\right\}+n\rho^2\operatorname{cov}\left\{\tilde{F}_{D}(s),\tilde{F}_{D}(t)\right\}\\
= \;& (1-\rho)^2(1+\eta)\left\{F_{\bar{D}}(s\wedge t)-F_{\bar{D}}(s)F_{\bar{D}}(t)\right\}+\rho^2\frac{1+\eta}{\eta}\left\{F_{D}(s\wedge t)-F_{D}(s)F_{D}(t)\right\}
\end{aligned}
$$

9.2.5 Proof of Lemma 5 Proof of Lemma 5 follows arguments similar to those in the proof of Lemma 3 (omitted).

9.2.6 Proof of Lemma 6

$$\begin{aligned} & \operatorname{cov}\left[\sqrt{n}\begin{pmatrix}\hat{\alpha}-\alpha\\ \hat{\beta}-\beta\end{pmatrix}, \sqrt{n}\left\{\hat{F}^{\star}(t)-F(t)\right\}\right] \\ = \; & \operatorname{cov}\left(\sqrt{n}H_4, \sqrt{n}\left[(1-\rho)H_1(t)+\rho H_2(t)+\{\rho-(1-\rho)\eta\}H_3(t)\right]\right)+o_p(1), \end{aligned}$$

where

$$H_4 \;=\; \frac{1}{n}S^{-1}\begin{pmatrix}\frac{\partial l(\alpha,\beta)}{\partial \alpha}\\ \frac{\partial l(\alpha,\beta)}{\partial \beta}\end{pmatrix}.$$

We have

$$\begin{aligned} \operatorname{cov}\{H_1(t), H_4\} &= \frac{1}{n_{\bar{D}}}\begin{pmatrix}F_{\bar{D}}(t)-(1+\eta)A_0(t)\\ 0\end{pmatrix}. \\ \operatorname{cov}\{H_1(t), H_4\} &= \frac{1}{n_{\bar{D}}}\begin{pmatrix}F_{\bar{D}}(t)-(1+\eta)A_0(t)\\ 0\end{pmatrix}, \\ \operatorname{cov}\{H_2(t), H_4\} &= \frac{1}{n_D}\begin{pmatrix}(1+\eta)A_0(t)-F_D(t)\\ 0\end{pmatrix}, \\ \operatorname{cov}\{H_3(t), H_4\} &= \frac{1}{n_D}\left(A_0(t), A_1(t)^T\right)A^{-1}-\left(\frac{1+\eta}{n_D}A_0(t), 0\right). \end{aligned}$$

The result follows by plugging in covariances between individual terms.

Similarly,

$$\begin{aligned} & \operatorname{cov}\left[\sqrt{n}\begin{pmatrix}\hat{\alpha}-\alpha\\ \hat{\beta}-\beta\end{pmatrix}, \sqrt{n}\left\{\tilde{F}^{\star}(t)-F(t)\right\}\right] \\ = \;& n(1-\rho)\operatorname{cov}\{H_4, \tilde{F}_{\bar{D}}(t)\} + n\rho \times \operatorname{cov}\{H_4, \tilde{F}_D(t)\} + o_p(1). \end{aligned}$$

We have

$$\begin{aligned} \operatorname{cov}\left\{H_4, \tilde{F}_{\bar{D}}(t)\right\} &= \frac{1+\eta}{n}\left\{-A^{-1}\begin{pmatrix}A_0(t)\\ A_1(t)\end{pmatrix} + \begin{pmatrix}F_{\bar{D}}(t)\\ 0\end{pmatrix}\right\}, \\ \operatorname{cov}\left\{H_4, \tilde{F}_D(t)\right\} &= \frac{1}{n}\frac{1+\eta}{\eta}\left\{A^{-1}\begin{pmatrix}A_0(t)\\ A_1(t)\end{pmatrix} - \begin{pmatrix}F_D(t)\\ 0\end{pmatrix}\right\}. \end{aligned}$$

9.2.7 Proof of Lemma 7 Lemma 7 follows immediately from Lemmas 3, 5, and 6.

9.2.8 Proof of Theorem 1 By Taylor's expansion,

$$\begin{aligned} & \sqrt{n}\left\{\hat{R}(v)-R(v)\right\} = \sqrt{n}\left[G\left\{\hat{\theta}, \hat{F}^{-1}(v)\right\} - G\left\{\theta, F^{-1}(v)\right\}\right] \\ = \;& \left\{\frac{\partial G(s,y)}{\partial y}|_{s=\theta, y=F^{-1}(v)}\right\}^T \sqrt{n}\left\{\hat{F}^{-1}(v)-F^{-1}(v)\right\} + \left(\frac{\partial G(s,y)}{\partial s}|_{s=\theta, y=F^{-1}(v)}\right)^T \sqrt{n}(\hat{\theta}-\theta) + o_p(1). \end{aligned}$$

The result follows according to the delta method.

9.2.9 Proof of Theorem 2

$$\begin{aligned} & \sqrt{n}\left\{\hat{R}^{-1}(p)-R^{-1}(p)\right\} = \sqrt{n}\left[\hat{F}\left\{G^{-1}\left(\hat{\theta}, p\right)\right\} - F\{G^{-1}(\theta, p)\}\right] \\ = \;& \sqrt{n}\left[\hat{F}\left\{G^{-1}(\theta, p)\right\} - F\{G^{-1}(\theta, p)\}\right] + \sqrt{n}\left[F\left\{G^{-1}\left(\hat{\theta}, p\right)\right\} - F\{G^{-1}(\theta, p)\}\right] + R_n, \end{aligned}$$

where

$$R_n = \sqrt{n}\left[\hat{F}\left\{G^{-1}\left(\hat{\theta},p\right)\right\} - \hat{F}\left\{G^{-1}(\theta,p)\right\}\right] - \left(\sqrt{n}\left[F\left\{G^{-1}\left(\hat{\theta},p\right)\right\} - F\{G^{-1}(\theta,p)\}\right]\right) = o_p(1)$$

by equicontinuity of the process $\sqrt{n}\left(\hat{F}-F\right)$.

The result follows according to the delta method.

Proof of Theorems3-4 follow similar arguments.

9.3 *Results about Nonparametric Predictiveness Curve Estimator*

9.3.1 Proof of Theorem 5 Suppose there are m pooled groups after isotonic regression with $w(Y) < \infty$. 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 $\hat{\mu} = n_D/n_{\bar{D}}$ into (4) results in

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

Since the term on the left-hand side of (4) is monotone increasing in μ, $\hat{\mu} = n_D/n_{\bar{D}}$ is the only solution.

9.3.2 Proof of Theorem 6 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 $Y \in \left\{y_{(i1)},...,y_{(im_i)}\right\}$. 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.

$v = \hat{F}\left\{y_{(im_i)}\right\}$ or $v = \tilde{F}\left\{y_{(im_i)}\right\}$ 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

$$\tilde{F}\left\{y_{(im_i)}\right\} = \hat{F}\left\{y_{(im_i)}\right\} \qquad \text{for } i = 1, ..., m.$$

Notice that

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

and

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

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