Biometrika (2009), **96**, 4, *pp*. 991–997
doi: 10.1093/biomet/asp040
Advance Access publication 12 October 2009

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

BY YING HUANG AND MARGARET SULLIVAN PEPE

Fred Hutchinson Cancer Research Center, Public Health Sciences, 1100 Fairview Avenue N., Seattle, Washington 98109-1024, U.S.A.

yhuang@fhcrc.org mspepe@u.washington.edu

SUMMARY

The performance of a well-calibrated risk model for a binary disease outcome can be characterized by the population distribution of risk and displayed with the predictiveness curve. Better performance is characterized by a wider distribution of risk, since this corresponds to better risk stratification in the sense that more subjects are identified at low and high risk for the disease outcome. 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 methods that accommodate case-control data. The semiparametric methods are flexible, and naturally generalize methods previously developed for cohort data. Applications to prostate cancer risk prediction markers illustrate the methods.

Some key words: Biased sampling; Biomarker; Case-control; Predictiveness curve; Risk prediction; Semiparametric method.

1. INTRODUCTION

Selecting biomarkers for medical practice 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 by helping to identify disease or risk of poor outcomes. Criteria for evaluating a biomarker depend on its purpose. In this paper, our goal is to evaluate risk prediction markers that are used to stratify the population into risk groups for whom different treatment recommendations are made.

Pepe et al. (2008a) suggested using the predictiveness curve (Bura & Gastwirth, 2001) to evaluate a risk prediction marker or model. They argued that the performance of a model to predict risk within a population relies not only on the effect of each predictor in the risk model, but also on the distribution 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 nondiseased. Let Y denote a marker of interest and let $\text{Risk}(Y) = \text{pr}(D = 1 \mid Y)$ denote the risk calculated on the basis of Y. The predictiveness curve is the curve $R(v)$ against v for $v \in (0, 1)$, where $R(v)$ is the vth quantile of $\text{Risk}(Y)$. The inverse function $R^{-1}(p) = \text{pr}\{\text{Risk}(Y) \leqslant p\}$ 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 prespecified 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. We focus on the scenario of a single continuous marker or a predefined marker combination, although our methods can be easily extended 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 test data (Ransohoff, 2007; Simon, 2005; Pepe et al., 2008b). The methods presented here accommodate such evaluations.

Let Y, Y_D and $Y_{\bar{D}}$ denote the marker measurement in the general, diseased, and nondiseased 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 $\rho = \mathrm{pr}(D = 1)$ 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 $\mathrm{pr}(D = 1 \mid Y)$ is monotone increasing in Y. Under this assumption, $R(v)$ equals $\mathrm{pr}\{D = 1 \mid Y = F^{-1}(v)\}$, the risk at the vth quantile of Y. Thus, the curve $R(v)$ against v is the same as the curve $\mathrm{pr}(D = 1 \mid Y = y)$ against $F(y)$. Therefore, estimation of the predictiveness curve can be undertaken in two steps: estimation of the risk model $\mathrm{pr}(D = 1 \mid Y = y)$ and estimation of the marker distribution $F(y)$. We develop estimators for these two entities and combine them to get a predictiveness curve estimator. We consider a case-control study with n_D cases $Y_{Di} (i = 1, \ldots, n_D)$, $n_{\bar{D}}$ controls $Y_{\bar{D}i}$ $(i = 1, \ldots, n_{\bar{D}})$, and write $\{Y_k, k = 1, \ldots, n\}$ for $\{Y_{\bar{D}1}, \ldots, Y_{\bar{D}n_{\bar{D}}}, Y_{D1}, \ldots, Y_{Dn_D}\}$ where $n = n_{\bar{D}} + n_D$. A 2008 University of Washington working paper 333 by Huang and Pepe contains proofs.

2. SEMIPARAMETRIC ESTIMATORS

2·1. *Estimation of the risk model*

Suppose the risk model of interest is $\mathrm{pr}(D = 1 \mid Y) = G(\theta, Y)$, where

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

and η is some monotone increasing function of Y. Examples of $\mathrm{logit}\{G(\theta, Y)\}$ include $\theta_0 + \theta_1 Y$ with $\theta_1 > 0$, the 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 & Green, 1992). In case-control studies, the maximum likelihood estimator of the odds ratio from the retrospective likelihood can be obtained by applying the prospective logistic model to the sample (Anderson, 1972; Prentice & Pyke, 1979), and 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 $\mathrm{logit}\{\mathrm{pr}(D = 1 \mid Y, S)\} = \theta_{0S} + \eta(\theta_{1S}, Y)$ to the data and correct the intercept with disease prevalence according to Bayes' theorem: $\mathrm{logit}\{\mathrm{pr}(D = 1 \mid Y)\} = \mathrm{logit}\{\mathrm{pr}(D = 1 \mid Y, S)\} - \mathrm{logit}\{\mathrm{pr}(D = 1 \mid S)\} + \mathrm{logit}\{\mathrm{pr}(D = 1)\}$. That is, let $(\hat{\theta}_{0S}, \hat{\theta}_{1S})$ be the maximum likelihood estimators of $(\theta_{0S}, \theta_{1S})$, then the estimator of θ is $\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1)$, where $\hat{\theta}_0 = \hat{\theta}_{0S} + \log[n_{\bar{D}}\hat{\rho}/\{n_D(1 - \hat{\rho})\}]$ and $\hat{\theta}_1 = \hat{\theta}_{1S}$.

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 $\hat{\rho}F_D + (1 - \hat{\rho})F_{\bar{D}}$, substituting estimates for F_D and $F_{\bar{D}}$. Next, we examine two ways of estimating F_D and $F_{\bar{D}}$.

First, in the absence of matching, 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 $\tilde{R}(v) = G\{\hat{\theta}, \tilde{F}^{-1}(v)\}$ for $v \in (0, 1)$ and $\tilde{R}^{-1}(p) = \tilde{F}\{G^{-1}(\hat{\theta}, p)\}$ for $p \in \{R(v) : v \in (0, 1)\}$, where $G^{-1}(\theta, p) = \inf\{y : G(\theta, y) \geqslant p\}$.

However, there is a more efficient way to obtain estimators for F_D and $F_{\bar{D}}$. Observe that the risk model (1) implies the following relationship between marker densities in cases and controls:

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

where $\alpha = \theta_0 + \log\{(1 - \rho)/\rho\}$, $\beta = \theta_1$, and $LR(Y)$ is the likelihood ratio function of Y (Green & Swets, 1966). When we estimate F_D and $F_{\bar{D}}$ empirically, positive point masses are allocated only to 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. A related issue arises 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 nondiseased populations is ignored.

Incorporation of (2) can be achieved by using the semiparametric likelihood framework (Qin & Zhang, 1997, 2003). This was originally proposed by Qin & Zhang (1997) to test the logistic regression assumption under a case-control sampling plan, and used by Qin & Zhang (2003) to estimate the receiver operating characteristic curve as an alternative to the fully parametric and nonparametric approaches. Suppose $\eta(\beta, Y) = \beta^{\mathrm{T}} r(Y)$, where $r(Y)$ is a vector of functions of Y. The likelihood ratio of Y becomes $LR(Y) = \exp\{\alpha + \beta^{\mathrm{T}} r(Y)\}$. 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 $L(\alpha, \beta, F_{\bar{D}}) = \prod_{i=1}^{n_{\bar{D}}} dF_{\bar{D}}(Y_{\bar{D}i}) \prod_{j=1}^{n_D} \exp\{\alpha + \beta^{\mathrm{T}} r(Y_{Dj})\} dF_{\bar{D}}(Y_{Dj})$, subject to $\sum_{i=1}^{n} dF_{\bar{D}}(Y_i) = 1$ and $\sum_{i=1}^{n} \exp\{\alpha + \beta^{\mathrm{T}} r(Y_i)\} dF_{\bar{D}}(Y_i) = 1$.

Solving this restricted maximum likelihood using the Lagrange multiplier method, the resulting maximum likelihood estimators for $F_{\bar{D}}$ and F_D are

$$\hat{F}_{\bar{D}}(y) = \frac{1}{n_{\bar{D}}} \sum_{i=1}^{n} \frac{I(Y_i \leqslant y)}{1 + \frac{n_D}{n_{\bar{D}}} \exp\{\hat{\alpha} + \hat{\beta}^{\mathrm{T}} r(Y_i)\}} = \frac{1}{n} \sum_{i=1}^{n} \frac{I(Y_i \leqslant y)}{\frac{n_{\bar{D}}}{n} + \frac{n_D}{n} \hat{LR}(Y_i)},$$

$$\hat{F}_D(y) = \frac{1}{n_{\bar{D}}} \sum_{i=1}^{n} \frac{\exp\{\hat{\alpha} + \hat{\beta}^{\mathrm{T}} r(Y_i)\} I(Y_i \leqslant y)}{1 + \frac{n_D}{n_{\bar{D}}} \exp\{\hat{\alpha} + \hat{\beta}^{\mathrm{T}} r(Y_i)\}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{LR}(Y_i) I(Y_i \leqslant y)}{\frac{n_{\bar{D}}}{n} + \frac{n_D}{n} \hat{LR}(Y_i)},$$

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

We use these estimators to compute $\hat{F} = (1 - \hat{\rho})\hat{F}_{\bar{D}} + \hat{\rho}\hat{F}_D$. Then we insert $\hat{\theta}$ and $\hat{F}$ into G to get the semiparametric maximum likelihood estimators of $R(v)$ and $R^{-1}(p)$: $\hat{R}(v) = G\{\hat{\theta}, \hat{F}^{-1}(v)\}$ for $v \in (0, 1)$, $\hat{R}^{-1}(p) = \hat{F}\{G^{-1}(\hat{\theta}, p)\}$ for $p \in \{R(v) : v \in (0, 1)\}$.

An intrinsic property of the predictiveness curve is that the area under the curve is equal to ρ since $\int_0^1 R(v) dv = \mathrm{pr}(D = 1) = \rho$. The analogue is not necessarily true though for an estimated predictiveness curve. However, it can be shown that the area under the semiparametric maximum likelihood estimator $\hat{R}(v)$ is always equal to $\hat{\rho}$. See an unpublished 2007 University of Washington dissertation by Huang for a proof. This property facilitates visual comparison between two estimated curves, for example predictiveness curves for different markers. 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*

Our 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

maximum likelihood estimators of α, β by applying the logistic regression model $\text{logit}\{\text{pr}(D = 1 \mid Y)\} = \alpha + \beta^{\mathrm{T}} r(Y) + \log(n_D/n_{\bar{D}})$ to the cohort sample. The last term is included here in order to make the notation, and the definition of α in particular, consistent with the previous subsection. For $y \in \mathcal{R}$, the semiparametric empirical estimator of F becomes $\tilde{F}(y) = n_D(n_D n)^{-1} \sum_{i=1}^{n_D} I(Y_{Di} \leqslant y) + n_{\bar{D}}(n_{\bar{D}} n)^{-1} \sum_{i=1}^{n_{\bar{D}}} I(Y_{\bar{D}i} \leqslant y) = n^{-1} \sum_{i=1}^{n} I(Y_i \leqslant y)$, while the semiparametric maximum likelihood estimator $\hat{F}(y)$ can be easily shown to also equal $n^{-1} \sum_{i=1}^{n} I(Y_i \leqslant y)$. That is, $\hat{F}$ and $\tilde{F}$ calculated from a cohort sample are both equal to 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 § 2·2 when applied to a cohort study are the same as the semiparametric estimator developed by Huang et al. (2007). That is, our methods generalize earlier methods to case-control designs.

3. ASYMPTOTIC THEORY FOR THE SEMIPARAMETRIC ESTIMATORS

We present asymptotic theory for the semiparametric estimators defined in § 2·2 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 the 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 set equal to a known constant ρ.

Asymptotic theory for the semiparametric maximum likelihood predictiveness curve estimator is presented in Theorems 1 and 2.

THEOREM 1. *As $n \to \infty$, $n^{1/2}\{\hat{R}(v) - R(v)\}$ converges to a normal random variable with mean zero and variance*

$$\Sigma_{1M}(v) = \left\{\frac{\partial R(v)}{\partial v}\right\}^2 \text{var}\big(n^{1/2}[\hat{F}\{F^{-1}(v)\} - v]\big) + \left\{\frac{\partial R(v)}{\partial \theta}\right\}^{\mathrm{T}} \text{var}\{n^{1/2}(\hat{\theta} - \theta)\} \left\{\frac{\partial R(v)}{\partial \theta}\right\}$$
$$+ 2\left\{\frac{\partial R(v)}{\partial \theta}\right\}^{\mathrm{T}} \text{cov}\big(n^{1/2}(\hat{\theta} - \theta), n^{1/2}[\hat{F}\{F^{-1}(v)\} - v]\big) \left\{\frac{\partial R(v)}{\partial v}\right\},$$

and $n^{1/2}\{\hat{R}^{-1}(p) - R^{-1}(p)\}$ converges to a normal random variable with mean zero and variance $\Sigma_{2M}(p) = \Sigma_{1M}(v)/\{\partial R(v)/\partial v\}^2$ for $v = R^{-1}(p)$.

THEOREM 2. *As $n \to \infty$, $n^{1/2}\{\tilde{R}(v) - R(v)\}$ converges to a normal random variable with mean zero and variance*

$$\Sigma_{1E}(v) = \left\{\frac{\partial R(v)}{\partial v}\right\}^2 \text{var}\big(n^{1/2}[\tilde{F}\{F^{-1}(v)\} - v]\big) + \left\{\frac{\partial R(v)}{\partial \theta}\right\}^{\mathrm{T}} \text{var}\{n^{1/2}(\hat{\theta} - \theta)\} \left\{\frac{\partial R(v)}{\partial \theta}\right\}$$
$$+ 2\left\{\frac{\partial R(v)}{\partial \theta}\right\}^{\mathrm{T}} \text{cov}\big(n^{1/2}(\hat{\theta} - \theta), n^{1/2}[\tilde{F}\{F^{-1}(v)\} - v]\big) \left\{\frac{\partial R(v)}{\partial v}\right\},$$

and $n^{1/2}\{\tilde{R}^{-1}(p) - R^{-1}(p)\}$ converges to a normal random variable with mean zero and variance $\Sigma_{2E}(p) = \Sigma_{1E}(v)/\{\partial R(v)/\partial v\}^2$ for $v = R^{-1}(p)$.

Theorems 1 and 2 state that variance of $\hat{R}(v)$ and $\tilde{R}(v)$ and their 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 $R^{-1}(p)$.

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 for F in a biased sampling model with known selection weights, for which the large sample theory was provided by Gill et al. (1988). Gilbert et al. (1999) extended this method to allow the weight functions to depend

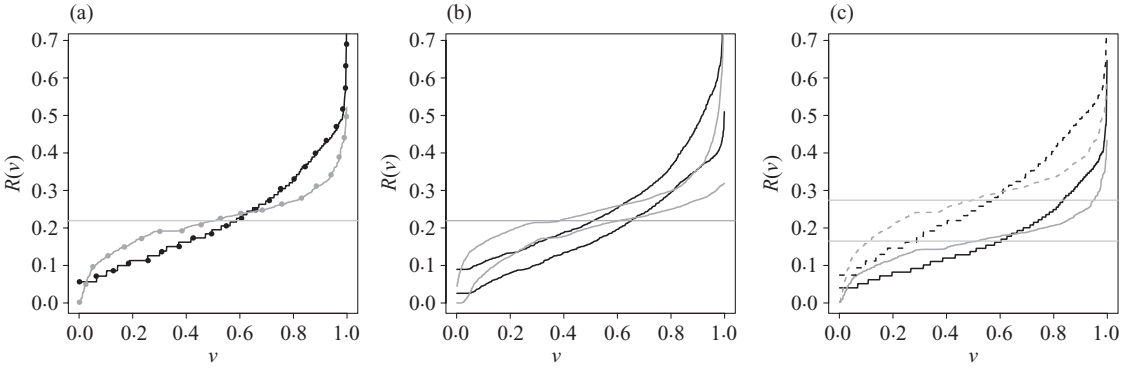

Fig. 1. (a) The semiparametric maximum likelihood estimators (solid) and semiparametric empirical estimators (dotted) of predictiveness curves for PSA (black) and PSA velocity (grey) for predicting prostate cancer; (b) the 95% pointwise confidence intervals constructed from percentiles of the bootstrap distribution based on the semiparametric maximum likelihood estimators of predictiveness curves for PSA (black) and PSA velocity (grey); and (c) the semiparametric maximum likelihood estimators of predictiveness curves for PSA (black) and PSA velocity (grey) when $\rho = 0{\cdot}165$ (solid) and $\rho = 0{\cdot}274$ (dashed). The horizontal lines indicate disease prevalences plugged in.

on an unknown finite-dimensional parameter θ. Gilbert (2000) demonstrated that the maximum likelihood estimators for θ and $F_{\bar{D}}$ are semiparametric efficient. The efficiency of our semiparametric maximum likelihood estimators follows.

It can be shown that the asymptotic covariance between $\hat{\theta}$ and the estimator of F is the same for the two semiparametric procedures. 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 the asymptotic variances of $\hat{F}$ and $\tilde{F}$. The latter can be shown to be positively proportional to the asymptotic variance of $n^{1/2}(\hat{F}_{\bar{D}} - \tilde{F}_{\bar{D}})$. 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. 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, the prostate specific antigen, less than 3·0 ng/mL, and 55 years of age and older who were followed up for seven years with annual PSA measurements. Thompson et al. (2006) identified 5519 men on the placebo arm who had undergone prostate biopsy, had a PSA and digital rectal exam during the year prior to biopsy and at least two PSA values from the three years prior to biopsy. Sample disease prevalence from the study cohort is $\hat{\rho} = 21{\cdot}9\%$. We randomly sampled 250 cases and 250 controls from this cohort to form a case-control sample for illustration.

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 displayed in Fig. 1(a) are fairly similar to each other for both markers. Their variance estimates are also similar. The pointwise 95% bootstrap percentile confidence intervals for $R(v)$ constructed from the semiparametric maximum likelihood estimators are displayed in Fig. 1(b), with variability in $\hat{\rho}$ incorporated.

PSA has steeper predictiveness curves, suggesting that it is a better marker for predicting risk of prostate cancer. Table 1 presents values for the risk percentiles of PSA and PSA velocity in the population, $R(v)$, for $v = 10\%$ and 90%. In addition, risk stratum sizes, $R^{-1}(p)$, for a low risk threshold of 10% and a high risk threshold of 30% are presented. P-values for comparing markers are based on the 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 as quantified by larger $R(0{\cdot}9)$,

Table 1. *Comparisons between PSA and PSA velocity for the predicting risk of prostate cancer using the semiparametric maximum likelihood method*

	PSA		PSA velocity		
Measure	Estimate	95% Confidence interval	Estimate	95% Confidence interval	*P*-value
$R(0{\cdot}1)$	0·072	(0·046, 0·109)	0·122	(0·075, 0·159)	0·027
$R(0{\cdot}9)$	0·413	(0·356, 0·476)	0·313	(0·275, 0·356)	<0·001
$R^{-1}(0{\cdot}1)$	0·188	(0·073, 0·291)	0·060	(0·020, 0·142)	0·021
$1 - R^{-1}(0{\cdot}3)$	0·244	(0·191, 0·296)	0·129	(0·030, 0·197)	0·009
$R^{-1}(0{\cdot}3) - R^{-1}(0{\cdot}1)$	0·568	(0·443, 0·724)	0·811	(0·668, 0·935)	0·004

better for predicting low risk, i.e. smaller $R(0{\cdot}1)$, and it classifies more people into the low and high risk ranges.

In practice, there may not always be a cohort for estimating prevalence. Often an investigator plugs in a specific prevalence value and treats it as known. We illustrate application of a sensitivity analysis using our example. Consider two values $\rho = 0{\cdot}165$ and $\rho = 0{\cdot}274$, which correspond to a 25% change from $\hat{\rho} = 0{\cdot}219$. The corresponding predictiveness curves are displayed in Fig. 1(c). 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. Comparisons at particular risk thresholds, on the other hand, are affected by prevalence. For example, when $\rho = 0{\cdot}165$, based on the semiparametric maximum likelihood procedure, PSA assigns significantly more people into the low risk range than PSA velocity, with estimates of $R^{-1}(0{\cdot}1)$ being 31·3% and 13·5% respectively, p-value $< 0{\cdot}001$. PSA is also a significantly better marker for predicting high risk than PSA velocity, with estimates of $1 - R^{-1}(0{\cdot}3)$ being 12·5% and 2·9% respectively, p-value $< 0{\cdot}001$. In contrast, when $\rho = 0{\cdot}263$, estimates of $R^{-1}(0{\cdot}1)$ become 9·7% and 3·8% for PSA and PSA velocity, and estimates of $1 - R^{-1}(0{\cdot}3)$ are 38·9% and 36·9% respectively. Neither of the comparisons is significant with p-values being 0·192 and 0·736, respectively. The comparison with respect to the percentage classified into the equivocal risk range is significant when $\rho = 0{\cdot}165$, p-value $< 0{\cdot}001$, but not when $\rho = 0{\cdot}274$, p-value $= 0{\cdot}375$.

5. CONCLUDING REMARKS

In this paper, we have developed flexible semiparametric 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 semiparametric 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. There are other approaches under development for estimating the predictiveness curve, including a nonparametric approach and an approach based on its relationship with the receiver operating characteristic curve (Huang & Pepe, 2009).

The validity of both semiparametric estimators relies 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 its empirical counterpart. Asymptotic relative efficiency of the former versus the latter is a complicated function of the disease prevalence, the separation between cases and controls, the case-control sampling ratio and the quantile of interest. In our example, the two estimators have similar variance when the disease prevalence is medium. It is shown in the 2007 dissertation by Huang that for rare diseases, using the model-based approach may achieve considerable efficiency gains for certain quantiles.

An important use of 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, the separation between cases and controls and the performance measure that is of primary interest. A detailed study can be found in the 2007 dissertation by Huang.

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 with an existing model. More research on methods to evaluate incremental value is warranted. Methods described here can be easily extended and adapted for such purposes.

Acknowledgement

The authors are grateful for support provided by grants from the U.S. National Institutes of Health and the National Cancer Institute.

References

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

Baker, S. G., Kramer, B. S. & Srivastava, S. (2002). Markers for early detection of cancer: statistical guidelines for nested case-control studies. *BMI Med. Res. Methodol.* **2**, 4–11.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. & Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore, MD: Johns Hopkins University Press.

Breslow, N. E., Robins, J. M. & Wellner, J. A. (2000). On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli* **6**, 447–55.

Bura, E. & Gastwirth, J. L. (2001). The binary regression quantile plot: assessing the importance of predictors in binary regression visually. *Biomet. J.* **43**, 5–21.

Cole, T. J. & Green, P. J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood. *Statist. Med.* **11**, 1305–19.

Gilbert, P. B. (2000). Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. *Ann. Statist.* **28**, 151–194.

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

Gill, R. D., Vardi, Y. & Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.* **16**, 1069–1112.

Green, D. M. & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Huang, Y. & Pepe, M. S. (2009). A parametric ROC model based approach for evaluating the predictiveness of continuous markers in case-control studies. *Biometrics*, doi: 10.1111/j.1541-0420.2009.01201.x

Huang, Y., Pepe, M. S. & Feng, Z. (2007). Evaluating the predictiveness of a continuous marker. *Biometrics* **63**, 1181–88.

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

Pepe, M. S., Etzioni, R., Feng, Z., Potter, J. D., Thompson, M. L., Thornquist, M., Winget, M. & Yasui, Y. (2001) Phases of biomarker development for early detection of cancer. *J. Nat. Cancer Inst.* **93**, 1054–61.

Pepe, M. S., Feng, Z., Huang, Y., Longton, G. M., Prentice, R., Thompson, I. M. & Zheng, Y. (2008a). Integrating the predictiveness of a marker with its performance as a classifier. *Am. J. Epidemiol.* **167**, 362–68.

Pepe, M. S., Feng, Z., Janes, H., Bossuyt, P. M. & Potter, J. D. (2008b). Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. *J. Nat. Cancer Inst*. **100**, 1432–38.

Prentice, R. L. & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–11.

Qin, J. & Zhang, J. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika* **84**, 609–18.

Qin, J. & Zhang, J. (2003). Using logistic regression procedures for estimating receiver operating characteristic curves. *Biometrika* **93**, 585–96.

Ransohoff, D. F. (2007). How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. *J. Clin. Epidemiol.* **60**, 1205–19.

Simon, R. (2005). Roadmap for developing and validating therapeutically relevant genomic classifiers. *J. Clin. Oncol.* **23**, 7332–41.

Thompson, I. M., Pauler Ankerst, D. & Chi, C. (2006). Assessing prostate cancer risk: results from the prostate cancer prevention trial. *J. Nat. Cancer Inst.* **98**, 529–34.

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

Vardi, Y. (1985). Empirical distributions in selection bias models. *Ann. Statist.* **13**, 178–203.

[*Received March* 2008. *Revised January* 2009]