Time-dependent Predictive Values of Prognostic Biomarkers with Failure Time Outcome

Yingye Zheng*, Tianxi Cai †, Margaret S. Pepe* and Wayne C. Levy‡

* Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109

† Department of biostatistics, Harvard School of Public Health, Boston, MA 02115

‡ Division of Cardiology, University of Washington, Seattle, WA 98177

ABSTRACT

In a prospective cohort study, information on clinical parameters, tests and molecular markers is often collected. Such information is useful to predict patient prognosis and to select patients for targeted therapy. We propose a new graphical approach, the positive predictive value (PPV) curve, to quantify the predictive accuracy of prognostic markers measured on a continuous scale with censored failure time outcome. The proposed method highlights the need to consider both predictive values and the marker distribution in the population when evaluating a marker, and it provides a common scale for comparing different markers. We consider both semiparametric and nonparametric based estimating procedures. In addition, we provide asymptotic distribution theory and resampling based procedures for making statistical inference. We illustrate our approach with numerical studies and datasets from the Seattle Heart Failure Study.

KEYWORDS: Prognostic accuracy, Positive predictive value, Survival analysis.

ACKNOWLEDGEMENTS: This research was supported by grant U01-CA86368 and P01-CA053996 awarded by the National Institutes of Health.

Corresponding author: Yingye Zheng e-mail: yzheng@fhcrc.org; phone: 206-667-7580;fax: 206-667-5977.

1. INTRODUCTION

A common research question in modern medicine is: can putative markers predict future progression of disease? We consider a marker to be any measurement with the potential to signal onset or progression of disease. In disease screening and prognosis, markers that predict future onset or progression of disease are sought. In epidemiology, identified risk factors for many diseases are routinely used in public health practice to classify subjects in regards to risk of future disease events. In these settings predictive markers can be used to stratify patients according to future risk of a (bad) outcome. This leads to more refined treatment or monitoring strategies. Before adopting a marker in practice, however, (i) its predictive accuracy must be quantified, and (ii) it must be compared with other potential markers, including existing prognostic systems, so that the best marker is selected for public health practice.

There are two main approaches to describing the accuracy of a dichotomous marker, Y, where the binary outcome is D (e.g., diseased $D = 1$ versus not diseased $D = 0$). The retrospective measures are the true and false positive fractions (TPF, FPF), also known as sensitivity and 1-specificity. These are often of interest in early phases of biomarker studies, since they quantify the extent to which the marker reflects the true outcome and can be calculated directly from case-control studies. However, positive and negative predictive values (PPV, NPV), the prospective measures, are of more interest to the end users of the test, the clinician and the patient, since they quantify the subject's risk of the outcome, D, given the test result Y. Calculation of the PPV and NPV is typically performed with a cohort study.

The PPV and NPV are defined for dichotomous tests. No standard definition exists when biomarker Y is continuous. We propose to follow the approach of Moskowitz and Pepe (2004b), who defined for $0 \leq v \leq 1$ $\text{PPV}(v) = P\{D = 1|F(y) \geq v\}$ and $\text{NPV}(v) = P\{D = 0|F(y) < v\}$, where F is the cumulative distribution function of

Y. They plot PPV(v) versus v, where subjects with marker values at or above the vth population percentile are considered as test positive (i.e., $F(y) \geq v$), and those below are regarded as negative. Note that NPV(v) is a function of v, PPV(v) and the prevalence ρ, i.e., $\text{NPV}(v) = 1 - \{\rho - \text{PPV}(v)(1-v)\}v^{-1}$.

The receiver operating characteristic (ROC) curve is a plot of $\text{TPF}(c) = P(Y \geq c|D = 1)$ versus $\text{FPF}(c) = P(Y \geq c|D = 0)$ for $c \in (-\infty, \infty)$, generalizing the notion of (TPF,FPF) to continuous data by thresholding the marker. The PPV curve is a natural analogue of the ROC curve for generalizing the notion of predictive value to continuous markers. Importantly, using v as the X-axis rather than the raw marker value provides a common scale for different markers that may be incomparable with respect to their raw values. Moreover, since v is the proportion of the population testing negative with the marker, it makes sense to compare the PPVs of markers when they are rescaled to have equal vs. This highlights the need to consider both the positivity probability, $1 - v$, and the associated PPV(v) when evaluating a marker.

We generalize the definition of the PPV curve to outcome variables that are event times. Specifically, for an event time T, we define for a marker Y measured at baseline

$$\text{PPV}(t, v) = P\{T < t|F(y) \geq v\}.$$

A number of approaches to summarizing the predictive accuracy of a continuous marker or covariate are available (Begg et al., 2000). Perhaps the most commonly used approach in practice is to simply report the hazard ratio estimated from a Cox regression analysis. This, however, ignores absolute risks and the distribution of subjects across risk levels, fundamental aspects of the predictive value of a marker. Other popular approaches include an R^2 summary as the proportion of variation explained by covariates (Schemper and Henderson, 2000) and the Brier score, a measure of residual variation (Graf et al., 1999). However, these measures may lack clinical relevance. The notion of explained variation or degree of separation cannot be translated into

a clinically meaningful quantity that is easily understood by clinicians and patients. Furthermore, these measures do not easily facilitate formal comparisons between two markers, and they do not distinguish between different types of errors.

We propose a new way of quantifying the predictive accuracy of prognostic markers measured on continuous scale. In contrast to other suggested measures of predictive accuracy for survival data, we seek a measure that is simple and meaningful for clinical practice, amenable to the comparison of multiple markers, and flexible in its assumptions about the underlying model and censoring mechanism.

2. ESTIMATION

We consider a prospective study where each subject denoted by the subscript i has a marker Y_i measured at the baseline. We let $F(y) = P(Y \leq y)$ denote the cumulative distribution function and $f(y)$ the corresponding density function. Also let T_i be the time to failure for subject i. We assume that T_i may be censored at time C_i, and we only observe $X_i = \min(T_i, C_i)$ and an associated censoring indicator Δ_i where $\Delta_i = 1$ if $X_i = T_i$ and 0 otherwise. Here (Y_i, X_i, Δ_i) $i = 1, \cdots, n$ are independent. In addition, we assume independent censoring such that C_i is conditionally independent of the event time T_i given marker Y_i. Although valid estimation of the PPV curve does not depend on the requirement that risk $P(T|Y = y)$ be a monotonic function of Y, the assumption is desirable in a setting where a biomarker threshold value is used for clinical decision making. For example, rising prostate specific antigen (PSA) may predict poor disease-free survival in patients with prostate cancer. By convention we assume that larger values of Y are associated with higher risks of failure.

2·1 *The PPV Curve*

We define the PPV curve as a plot of $\text{PPV}(t, v) = P\{T < t | F(y) \geq v\}$ versus v, for v in an open interval of $(0, 1)$. On the x-axis it shows the proportion of subjects testing positive when a positive biomarker test is defined as exceeding the threshold

corresponding to the vth percentile of Y in the population: $Y \geq F^{-1}(v)$ or equivalently $F(y) \geq v$. On the y-axis it shows the risk of an event by time t for subjects who satisfy that positivity criterion. A horizontal line corresponding to the marginal event time probability $P(T < t)$ serves as a benchmark PPV curve for completely uninformative markers. More informative markers have PPV curves that rise more steeply and reach higher levels.

Some appealing attributes of the PPV curve for practical use include its ease in interpretation and visualization of useful quantities. For example, if only subjects in the top 10^{th} percentile of risk are eligible for an intervention study, one can observe the expected proportion of such subjects with an event by time t, $\text{PPV}(t, 0.90)$. Conversely if a fraction p to have an event by time t is desired, one can observe the corresponding fraction $1 - v$ of the population that will be required to test positive with the marker, $\text{PPV}^{-1}(t; p) = v$, from a monotonic PPV curve. The PPV curve also provides a common meaningful scale for comparing multiple markers. Lastly, the PPV curve can be used to suggest thresholds that are optimal for defining biomarker positivity. Although PPV curves have been used in the applied literature (e.g. Blanks et al., 2001) they have only recently been formally considered in the statistical literature (Moskowitz and Pepe, 2004b). We extend the idea from the application to binary outcomes considered by Moskowitz and Pepe (2004b) to event time outcomes.

2·2 *Estimation: Non-parametric Approaches*

We first describe a class of nonparametric approaches. Such methods do not impose modeling assumptions on the relationship between the marker and survival and therefore will be broadly applicable to many practical settings.

Under independent censoring: We first consider the case where the censoring process C does not depend on Y. A natural estimator for $\text{PPV}(t, v)$ can be obtained by estimating the survival distribution based on the subset of subjects with

$\widehat{F}(y) \geq v$, where $\widehat{F}(y) = n^{-1}\sum_{i=1}^{n} I(Y_i < y)$ is the empirical distribution function of Y. The survival probability function can be estimated nonparametrically using either the Aalen-Nelson or Kaplan-Meier estimator. Since these two estimators are asymptotically equivalent, we only consider the Aalen-Nelson estimator. Specifically, let $\Lambda_v(t)$ be the cumulative hazard function of T among subjects with $F(Y) \geq v$, then $\text{PPV}(t, v) = 1 - \exp\{-\Lambda_v(t)\}$ can be estimated by

$$\widetilde{\text{PPV}}(t, v) = 1 - \exp\left\{-\widetilde{\Lambda}_v(t)\right\} = 1 - \exp\left\{-\int_0^t \frac{d\widetilde{N}_v(s)}{\widetilde{\pi}_v(s)}\right\}, \qquad (2{\cdot}1)$$

where $\widetilde{N}_v(s) = n^{-1}\sum_i \widehat{w}_v(Y_i)N_i(s)$, $N_i(s) = I(X_i \leq s)\Delta_i$, $\widehat{w}_v(Y_i) = I\{\widehat{F}(Y_i) \geq v\}$ and $\widetilde{\pi}_v(s) = n^{-1}\sum_i \widehat{w}_v(Y_i)I(X_i \geq s)$.

Under marker dependent censoring: Here, we allow C to depend on Y, but assume that T remains independent of C conditional on Y. In the presence of such dependence, $\widetilde{\text{PPV}}(t, v)$ is subject to bias. For example if individuals with lower marker values tend to be censored earlier then we may expect $\widetilde{\text{PPV}}(t, v)$ to be biased downward. This problem often arises in situations where a prognostic biomarker is available and the frequency of follow-up efforts is influenced by the marker value measured at baseline. For example, in many AIDS studies individual's censoring status may be related to CD4 counts, a well-accepted marker for survival. To account for marker dependent censoring, we note that

$$\text{PPV}(t, v) = 1 - \frac{P\{T \geq t, Y \geq F^{-1}(v)\}}{P\{F(Y) \geq v\}} = 1 - (1-v)^{-1}\int_{F^{-1}(v)}^{\infty} S_y(t)dF(y),$$

where $S_y(t) = P(T \geq t \mid Y = y)$ is the conditional survival function. Although T may depend on C conditional on $F(Y) \geq v$, T is independent of C given $Y = y$ and thus $S_y(t)$ can be estimated non-parametrically. In particular we consider the kernel estimator for $S_y(t)$ (Beran, 1981; Dabrowska, 1989; Akritas, 1994) :

$$\widehat{S}_y(t) = \exp\left\{-\widehat{\Lambda}_y(t)\right\} = \exp\left\{-\int_0^t \frac{d\widehat{N}_y(s)}{\widehat{\pi}_y(s)}\right\},$$

where $\widehat{N}_y(s) = n^{-1}\sum_{i=1}^n K_h(Y_i - y)N_i(s)$, $\widehat{\pi}_y(s) = n^{-1}\sum_{i=1}^n K_h(Y_i - y)I(X_i \geq s)$, and $K_h(x) = K(x/h)/h$. Here K is a given symmetric smooth kernel density function, and h is the bandwidth such that $nh^2 \to \infty$ and $nh^4 \to 0$ as $n \to \infty$. A plug-in estimator for $\mathrm{PPV}(t, v)$ based on the bivariate distribution function is

$$\widehat{\mathrm{PPV}}(t, v) = 1 - (1 - v)^{-1} \int_{\widehat{F}^{-1}(v)}^{\infty} \widehat{S}_y(t) d\widehat{F}(y). \tag{2·2}$$

2·3 *Estimation: A Semi-parametric Approach*

The proposed PPV curve can also be estimated using a regression model approach. Compared with a nonparametric estimator, parametric methods are usually more efficient when the underlying assumptions hold. In addition, marker-dependent censoring is easily accommodated. As an illustration, we assume a proportional hazards model for survival time of the form $\lambda(t|Y) = \lambda_0(t)\exp(\beta_0 Y)$. Under this model the survival function is $S(t|y) = \exp\{-\Lambda_0(t)\exp(\beta_0 y)\}$, where $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ is the cumulative baseline hazard function. A plug-in estimator for $\mathrm{PPV}(t, v)$ based on the conditional survival probability is

$$\widehat{\mathrm{PPV}}^*(t, v) = 1 - (1 - v)^{-1} \int_{\widehat{F}^{-1}(v)}^{\infty} \exp\left\{-\widehat{\Lambda}_0(t)\exp\left(\widehat{\beta} y\right)\right\} d\widehat{F}(y), \tag{2·3}$$

where $\widehat{\beta}$ is the maximum partial likelihood estimator of β_0 and $\widehat{\Lambda}_0(t)$ is the Breslow estimator of $\Lambda_0(t)$.

To construct a PPV curve, one can select a grid of points $v \in (0, 1)$, and estimate the corresponding values of $\mathrm{PPV}\{t, \widehat{F}^{-1}(v)\}$. For example, the key quantity $\int_{\widehat{F}^{-1}(v)}^{\infty} \widehat{S}_y(t) d\widehat{F}(y)$ in $\widehat{\mathrm{PPV}}(t, v)$ can be calculated by first estimating the weighted Aalen-Nelson estimator at each y, and then integrating over the range of y with $y > \widehat{F}^{-1}(v)$.

2·4 *Evaluating and Comparing Predictive Values of Markers*

A few summaries based on the PPV curve are of interest. For example, we may wish to make inference about risk of t-year mortality for these $100(1-v)\%$ individuals

testing positive, i.e., $\text{PPV}(t, v)$ at specified values for v and t; or the fraction of the population testing positive that corresponds to a PPV value of p by year t, i.e., the inverse $\text{PPV}^{-1}(t; p)$ at specified values for p and t if the curve is monotonic.

A fundamental attraction of the PPV curve is that it provides a common meaningful scale for comparing markers. We will first consider comparing the $\text{PPV}(t, v)$ of two markers, Y_1 and Y_2, at any given (t, v) or jointly over a set of points $\{(t_k, v_k),\ k = 1, ..., K\}$. Typically marker data arise from study designs where both markers are measured on each individual. Based on such paired data, one may estimate the relative predictive value, $r\text{PPV}(t, v) = \text{PPV}_{Y_1}(t, v)/\text{PPV}_{Y_2}(t, v)$, or $\Delta\text{PPV}(t, v) = \text{PPV}_{Y_1}(t, v) - \text{PPV}_{Y_2}(t, v)$ using the aforementioned PPV estimators.

3. INFERENCE IN LARGE SAMPLES

We show in Appendix A that $\widetilde{\text{PPV}}(t, v)$ is uniformly consistent for $\text{PPV}(t, v)$. Furthermore, the process $\widetilde{\mathcal{W}}_v(t) \equiv n^{\frac{1}{2}}\{\widetilde{\text{PPV}}(t, v) - \text{PPV}(t, v)\}$ is asymptotically equivalent to $n^{-\frac{1}{2}} \sum_{i=1}^{n} \eta_i(t, v)$ and converges weakly to a zero mean Gaussian process, where $\eta_i(t, v)$ is defined in (A·1) of Appendix A.

To obtain a pointwise confidence interval and a simultaneous confidence band for $\text{PPV}(t, v)$, we will use the *resampling* method (Parzen et al., 1994) which has been successfully extended to approximate the distribution of a process (see Park and Wei (2003) for details). Specifically, we first generate J independent samples of standard normal random variables, $\{\mathcal{N}_i^{(j)}, i = 1, \ldots, n\}$, for $j = 1, ..., J$. Let $\mathbb{W}_v^{(j)}(t) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \widehat{\eta}_i(t, v)\mathcal{N}_i^{(j)}$ with $\widehat{\eta}_i(t, v)$ obtained by replacing all theoretical quantities in $\eta_i(t, v)$ by their empirical counterparts. The function $\partial\Lambda(t, c)/\partial c$ in $\widehat{\eta}_i(t, v)$ can be estimated with the finite difference estimator. Conditional on the data, the process $\mathcal{W}_v(t)$ has the same limiting covariance function as that of $\mathbb{W}_v^{(j)}(t)$. Therefore, we may approximate the distribution of $\widetilde{\mathcal{W}}_v(t)$ based on the realizations of $\{\mathbb{W}_v^{(j)}(t)\}$. Now, based on a functional delta method, we construct $100(1 - \alpha)\%$

confidence intervals for $\mathrm{PPV}(t,v)$ as

$$1-\exp\left(-\widetilde{\Lambda}_v(t)\exp\left[\pm\frac{d_\alpha\widetilde{\sigma}_v(t)}{\widetilde{\Lambda}_v(t)\exp\{-\widetilde{\Lambda}_v(t)\}}\right]\right),$$

where $\widetilde{\sigma}_v(t)^2=J^{-1}\sum_{j=1}^{J}\mathbb{W}_v^{(j)}(t)^2$, d_α is the $100(1-\alpha/2)$th percentile of the standard normal for point-wise confidence intervals, and d_α is obtained as the $100(1-\alpha)$th empirical quantile of $\{\sup_{v\in\{v_a,v_b\}}|\mathbb{W}_v^{(j)}(t)/\widetilde{\sigma}_v(t)|, j=1,...,J\}$ for simultaneous confidence intervals.

The uniform consistency of $\widehat{\mathrm{PPV}}(t,v)$ follows directly from the uniform consistency of $\widehat{\Lambda}_y(t)$ and $\widehat{F}(y)$. To obtain interval estimates of $\mathrm{PPV}(t,v)$, we show in Appendix B that $\widehat{\mathcal{W}}_v(t)\equiv n^{\frac{1}{2}}\left\{\widehat{\mathrm{PPV}}(t,v)-\mathrm{PPV}(t,v)\right\}$ is asymptotically equivalent to $n^{-\frac{1}{2}}\sum_{i=1}^{n}\xi_i(t,v)$, and converges weakly to a zero-mean Gaussian process, where $\xi_i(t,v)$ is defined in (B·1) of Appendix B. The distribution of $\widehat{\mathcal{W}}_v(t)$ can be approximated via the resampling methods by estimating all the unknown quantities in $\xi_i(t,v)$ empirically. Note that the density function $f(\cdot)$ in $\xi_i(t,v)$ can be estimated using a kernel estimator. Confidence intervals for $\mathrm{PPV}(t,v)$ can be constructed accordingly.

In Appendix C, we show that $\widehat{\mathcal{W}}_v^*(t)\equiv n^{\frac{1}{2}}\left\{\widehat{\mathrm{PPV}}^*(t,v)-\mathrm{PPV}(t,v)\right\}$ is asymptotically equivalent to $n^{-\frac{1}{2}}\sum_{i=1}^{n}\zeta_i(t,v)$, and converges weakly to a zero-mean Gaussian process, where $\zeta_i(t,v)$ is defined in (C·1). Subsequent inference procedures follow that of $\widehat{\mathrm{PPV}}(t,v)$. To make inference about $\mathrm{PPV}^{-1}(t;p)$, we note that by the stochastic equicontinuity of $\widehat{\mathcal{W}}_v(t)$, $n^{\frac{1}{2}}\{\widehat{\mathrm{PPV}}^{-1}(t;p)-\mathrm{PPV}^{-1}(t;p)\}$ is asymptotically equivalent to $\partial\mathrm{PPV}(t,v)/\partial v\widehat{\mathcal{W}}_v(t)$. Thus the distribution of $\widehat{\mathrm{PPV}}^{-1}(t;p)$ can also be derived similarly based on that of $\widehat{\mathcal{W}}_v(t)$.

To test whether two markers measured simultaneously on the same subject have significantly different predictive values, we test the hypothesis

$$H_0: r\mathrm{PPV}(t,v)\equiv\frac{\mathrm{PPV}_{Y_1}(t,v)}{\mathrm{PPV}_{Y_2}(t,v)}=1.$$

To obtain a confidence interval for $r\text{PPV}(t,v)$, we consider its log-transformation and note that by a continuous mapping theorem, $n^{\frac{1}{2}}\{\log\{\widehat{r\text{PPV}}(t,v)\}-\log\{r\text{PPV}(t,v)\}$ is asymptotically equivalent to $n^{-\frac{1}{2}}\sum_{i=1}^{n}\{\xi_i^{(Y_1)}(t,v)/\text{PPV}_{Y_1}(t,v)-\xi_i^{(Y_2)}(t,v)/\text{PPV}_{Y_2}(t,v)\}$ whose distribution can be approximated using the resampling method as well.

Simulation studies were performed to examine the finite sample properties of the proposed procedures and to investigate the impact of model assumptions on the two classes of estimators. The results suggest that our methods provide reasonably unbiased estimates and our nonparametric estimators are quite robust. See JASA supplemental web site for details on simulation results.

4. EXAMPLE: THE SEATTLE HEART FAILURE MODEL FOR PREDICTION OF SURVIVAL IN HEART FAILURE

We illustrate our methods with an example in the context of predicting survival among patients with heart failure. Heart failure is a serious condition with highly variable outcome. Often clinicians need to counsel patients about prognosis and to make decisions about medications, transplantation and end of life care. The Seattle Heart Failure Model (SHFM), a multivariate Cox model, was derived in a cohort of heart failure patients and prospectively validated in 5 additional cohorts with nearly 10,000 heart failure patients. The model incorporates 13 variables relating to clinical status and laboratory parameters with higher values of the SHFM score being more indicative of worse prognosis. Levy et al. (2006) have provided a complete description.

First we wish to quantify the accuracy of the SHFM score for predicting t-year survival. We consider data from the **Val-HeFT** study, a cohort independent of the original derivation trial. **Val-HeFT** is a randomized trial in 5,010 patients in 16 countries. The median follow-up was 2 years, with 976 death observed over the course of the study. Since there was no prespecified cutoff value for defining a positive result, time-dependent PPV curves provide graphical displays that characterize the risk of death

by t year among the $(1-v)$*100% of the population with a positive test, across a full spectum of $v \in (0,1)$. We considered PPV curves based on the three proposed estimators. A proportional hazards model of the form $\lambda(t|Y) = \lambda_0(t)exp(\beta$ SHFM score) was used for $\widehat{\text{PPV}}^*(t,v)$. For illustration we randomly selected 1000 patients from this study. For $\widehat{\text{PPV}}(t,v)$, let c denote the standard deviation of SHFM scores, the bandwidth h was chosen to be $c/n^{1/3} \approx 0.07$. The estimates are presented in Table 1. Figure 1 constructs PPV curves (left panel) and NPV curves (right panel) at $t = 1$ for v ranging from 0.05 to 0.95. Starting from the point $P(T < 1) = 0.08$, the PPV curve increases steeply, an indication that the SHFM score is informative for identifying patients at greater risk of death by the first year. For example, at $v = 0.5$, the corresponding $\widetilde{\text{PPV}}(t = 1, v = 0.5) = 0.14$ (95%CI: $(0.11, 0.17)$) whereas $\widetilde{\text{PPV}}(t = 1, v = 0.95) = 0.29$ (95%CI: $(0.19, 0.40)$) at $v = 0.95$. In other words, if the score is used to refer patients for a novel therapy, among patients whose scores are in the top 5% of the population, on average 29% would have failed by year 1; however among patients whose scores are in the top 50% of the population, on average only 14% would have failed by year 1. Such information may be helpful for clinicians to determine how many patients with heart failure are eligible for more aggressive therapy such as cardiac resynchronization. As shown in Figure 1, there existed a substantial discrepancy between the PPV curve calculated based on a Cox regression model and those based on nonparametric procedures, suggesting that the Cox model may not fit the data well. In this example, consideration of a smoothed estimator may be advantageous as it may be more robust.

If a PPV value p at t is considered for clinical decision making, what percentage of the population will be selected based on the SHFM score? We address this question by studying $\text{PPV}^{-1}(t = 1; p)$. For this study, if the goal is to achieve a PPV value of 0.25 by year one, then it requires that approximately 15% (95%CI: $(0.06, 0.24)$)(i.e.,

1-PPV$^{-1}(t = 1; p = 0.25)$) of the population test positive, i.e., we choose the $85th$ percentile of the marker (score) as the threshold for defining positivity.

One imminent question here is whether the SHFM provides improved prognostic potential over existing heart failure models. The Toronto heart failure model (THFM) was derived in hospitalized patients using information identified shortly after hospital presentation (Lee et al., 2003). It is of interest to compare the capacities of the two models for predicting 1-, 2- and 3-year mortality risk in populations that reflect broad range of systolic heart failure. Both prognostic scores appear to be significant predictors of survival from Cox models: hazards ratio (HR) for SHFM is 2.15 (95%CI: (2.01,2.31)), and 1.04 (95%CI: (1.03,1.04)) for THFM. The R^2s calculated from the Cox models are 0.075 and 0.042 for SHFM and THFM respectively.

We compare the predictive accuracies of the SHFM score and THFM score using data from the entire cohort of 5010 patients. In Table 2, we list for selected v and for t = 1, 2, and 3 year the estimated rPPV and their 95% pointwise confidence intervals and simultaneous confidence band calculated over the region $v = [0.05, 0.95]$. All $r\text{PPV}(t, v)$s with $v \geq 0.90$ are significantly higher than 1, however only those for $t = 2$ remain significant if the 95% confidence bands are considered. We conclude here that the SHFM is more predictive of 1-, 2-, and 3-year mortality risks than THFM when a small fraction of the population is selected for further treatment.

5. DISCUSSION

In this paper we have introduced a graphical approach for quantifying and comparing the prognostic accuracies of continuous markers with censored failure time outcome. Observe that a marker may be useful for prediction but perform poorly for classification. Gail and Pfeiffer (2005) noted that performance criteria of markers for selecting patients for cancer prevention interventions are not the same as those required of markers for cancer screening. Our motivating applications are concerned with predic-

tion and risk stratification. Therefore it is appropriate to evaluate them in terms of their prospective accuracy parameters, PPV and NPV. Much work in the literature has focused on evaluating the performance of a marker as a classifier, i.e., with respect to its retrospective accuracy; however clinically meaningful methods for quantifying the prospective prognostic accuracy have not been well developed (Moskowitz and Pepe, 2004a). The work presented here offers such a method.

Several approaches for estimating the PPV and NPV curves are studied. The semiparametric approach is more efficient than the nonparametric procedures, but can be sensitive to modeling assumptions about how the marker is related to survival. The two nonparametric approaches are more flexible, and the kernel smoothing estimator we considered also takes into account marker dependent censoring. These methods will be broadly applicable to many practical settings.

There are two considerations one must take into account when adopting the PPV curve in practice. First, PPVs and NPVs depend on prevalence of the outcome; consequently they reflect characteristics of the cohort that gave rise to the curves. It is therefore important to assure that the research cohort indeed constitutes a random sample of the general population of interest where the clinical decision rules will be applied (see Pepe et al. (2007) for a discussion of this issue). The Val-HeFT study consists of participants from 16 countries. It is conceivable that prospective accuracy might be different when applied to an individual country and it should be further evaluated in subcohorts with different heart failure rates. Second, in this paper we considered only the predictive performance of a baseline marker. Frequently in practice repeated measurements are collected for monitoring disease progression, and the 'updated' prediction of risk as a function of current and past marker information is of interest. Estimating such a quantity requires more deliberation. Further investigation on adopting the notion of PPV curve for longitudinal markers is warranted.

APPENDIX

Throughout we assume that the joint density of T, C and Y is continuously differentiable and the marker Y is bounded. We consider $v \in [p_l, p_u] \subset (0,1)$ and $t \in [\tau_1, \tau_2]$, where τ_1 and τ_2 are given constants such that $P(X < \tau_1) > 0$ and $P(X > \tau_2) > 0$. In addition, we assume that the first and second order derivatives of $F(y)$ are bounded away from 0 for $y \in (-\infty, \infty)$. $\Lambda(t,c)$ is continuously differentiable with $\sup_{s,c}\{\Lambda(t,c) + \dot{\Lambda}(s,c)\} < \infty$, where $\dot{\Lambda}(s,c) = \partial\Lambda(s,c)/\partial c$.

A. Asymptotic Properties of $\widetilde{\mathbf{PPV}}(t,v)$

Since $\widetilde{\mathrm{PPV}}(t,v) = 1 - \exp\{\widetilde{\Lambda}_v(t)\}$ is a smooth monotonic transformation of $\widetilde{\Lambda}_v(t)$, we first derive the asymptotic properties for $\widetilde{\Lambda}_v(t)$. To this end, we define

$$\bar{N}(s,c) = n^{-1}\sum_{Y_i \geq c} I(X_i \leq s)\Delta_i \quad \bar{\pi}(s,c) = n^{-1}\sum_{Y_i \geq c} I(X_i \geq s), \quad \bar{\Lambda}(t,c) = \int_0^t \frac{d\bar{N}(s,c)}{\bar{\pi}(s,c)}$$

$A(s,c) = E\{\bar{N}(s,c)\}$, $\pi(s,c) = E\{\bar{\pi}(s,c)\}$, and $\Lambda(t,c) = \int_0^t dA(s,c)/\pi(s,c)$. It follows from a uniform law of large numbers (Pollard, 1990) that $\sup_{t,c}|\bar{\Lambda}(t,c) - \Lambda(t,c)| \to 0$, almost surely. This, together with the uniform consistency of $\widehat{c}_v = \widehat{F}^{-1}(v)$ for $c_v = F^{-1}(v)$ and a continuous mapping theorem, implies that $\sup_{t,v}|\widetilde{\Lambda}_v(t) - \Lambda_v(t)| \to 0$ almost surely. To derive the large sample distribution for $\widetilde{\Lambda}_v(t)$, we write

$$n^{\frac{1}{2}}\{\widetilde{\Lambda}_v(t) - \Lambda_v(t)\} = n^{\frac{1}{2}}\left\{\bar{\Lambda}(t,c_v) - \Lambda(t,c_v)\right\} + n^{\frac{1}{2}}\left\{\bar{\Lambda}(t,\widehat{c}_v) - \bar{\Lambda}(t,c_v)\right\}.$$

It follows from standard empirical processes theory (Pollard, 1990) that $n^{\frac{1}{2}}\{\bar{\Lambda}(t,c) - \Lambda(t,c)\}$ is asymptotically equivalent to $n^{-\frac{1}{2}}\sum_i \eta_{i1}(t,c)$ and converges weakly to a zero mean Gaussian process in (t,c), where

$$\eta_{i1}(t,c) = \int_0^t I(Y_i \geq c)\pi(s,c)^{-1}\{dN_i(s) - \pi(s,c)^{-1}I(X_i \geq s)dA(s,c)\}.$$

It follows that $n^{\frac{1}{2}}\{\bar{\Lambda}(t,\widehat{c}_v) - \bar{\Lambda}(t,c_v)\} = \dot{\Lambda}(t,c_v)n^{\frac{1}{2}}(\widehat{c}_v - c_v) + o_p(1)$, where $\dot{\Lambda}(s,c) = \partial\Lambda(s,c)/\partial c$. Here and throughout, the $o_p(1)$ is uniform in t and v. This, together

with the weak convergence of the quantile process $n^{\frac{1}{2}}(\widehat{c}_v - c_v)$ in v, implies that $\widetilde{\mathcal{W}}_v(t)$ is asymptotically equivalent to $n^{-\frac{1}{2}} \sum_i \eta_i(t, v)$, where

$$\eta_i(t,v) = \left\{1 - \widetilde{\mathrm{PPV}}(t,v)\right\} \left[\eta_{1i}(t, c_v) - \frac{\dot{\Lambda}(t, c_v)}{f(c_v)} \{I(Y_i \le c_v) - v\}\right], \qquad (\mathrm{A}{\cdot}1)$$

and $f(y) = dF(y)/dy$. It then follows from a functional central limit theorem (Pollard, 1990) that $\widetilde{\mathcal{W}}_v(t)$ converges weakly to a zero-mean Gaussian process.

B. Asymptotic Properties of $\widehat{\mathrm{PPV}}(t, v)$

We require the same conditions as specified in Du and Akritas (2002). Briefly, $K(\cdot)$ is a twice continuously differentiable symmetric probability density function with bounded second derivative. To derive the large sample distribution for $\widehat{\mathrm{PPV}}(t, v)$, we write

$$\widehat{\mathcal{W}}_v(t) = n^{\frac{1}{2}}\{\widehat{\mathrm{PPV}}(t,v) - \mathrm{PPV}(t,v)\} = \{\widehat{\mathcal{W}}_{1v}(t) + \widehat{\mathcal{W}}_{2v}(t)\}/(1-v),$$

where $\widehat{\mathcal{W}}_{1v}(t) = n^{\frac{1}{2}} \int_{\widehat{c}_v}^{\infty} \{e^{-\widehat{\Lambda}_y(t)} - e^{-\Lambda_y(t)}\} d\hat{F}(y)$ and $\widehat{\mathcal{W}}_{2v}(t) = n^{\frac{1}{2}}\{\int_{\widehat{c}_v}^{\infty} S_y(t) d\hat{F}(y) - \int_{c_v}^{\infty} S_y(t) dF(y)\}$. To approximate the distribution of $\widehat{\mathcal{W}}_{1v}(t)$, we note that

$$\sup_y \left|\widehat{F}(y) - F(y)\right| + \sup_{t,y} \left|\widehat{\Lambda}_y(t) - \Lambda_y(t)\right| + \sup_v |\widehat{c}_v - c_v| = o_p(n^{-\frac{1}{4}}).$$

This, together with a Taylor series expansion and Lemma A.3 of Bilias et al. (1997), implies that $\widehat{\mathcal{W}}_{1v}(t) = -n^{\frac{1}{2}} \int_{c_v}^{\infty} S_y(t)\{\widehat{\Lambda}_y(t) - \Lambda_y(t)\} dF(y) + o_p(1)$. Furthermore, from the asymptotic expansions for $\widehat{\Lambda}_y(t)$ in Du and Ariktas (2002), we have

$$\widehat{\mathcal{W}}_{1v}(t) = -n^{\frac{1}{2}} \int_{c_v}^{\infty} S_y(t) \left\{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - Y_i}{h}\right) M_y(t; X_i, \Delta_i)\right\} dy$$

where $M_y(t, X_i, \Delta_i) = \int_0^t \left\{\frac{dN_i(s)}{\pi_y(s)} - \frac{dA_y(s)}{\pi_y(s)^2} I(X_i \ge s)\right\}$, $\pi_y(s) = P(X \ge s \mid Y = y)$ and $A_y(s) = E\{N(s) \mid Y = y\}$. Now, by a change variable $\psi = \frac{y - Y_i}{h}$ and $nh^4 = o_p(1)$,

$$\begin{aligned}
\widehat{\mathcal{W}}_{1v}(t) &= -n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_{-\infty}^{\infty} I\left(Y_i \ge c_v\right) K(\psi) S_{Y_i}(t) M_{Y_i}(t; X_i, \Delta_i) d\psi + O(n^{\frac{1}{2}} h^2) + o_p(1) \\
&= -n^{-\frac{1}{2}} \sum_{i=1}^{n} \xi_{i1}(t, v) + o_p(1)
\end{aligned}$$

where $\xi_{i1}(t,v) = I(Y_i \geq c_v) S_{Y_i}(t) M_{Y_i}(t; X_i, \Delta_i)$. For $\widehat{\mathcal{W}}_{2v}(t)$, we note that

$$\widehat{\mathcal{W}}_{2v}(t) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_{c_v}^{\infty} S_y(t) d\{I(Y_i \leq y) - F(y)\} - n^{\frac{1}{2}} \{\widehat{c}_v - c_v\} S_{c_v}(t) f(c_v) + o_p(1)$$

$$= n^{-\frac{1}{2}} \sum_{i=1}^{n} \left[S_{Y_i}(t) I(Y_i > c_v) - \int_{c_v}^{\infty} S_y(t) dF(y) + S_{c_v}(t)\{1(Y_i \leq c_v) - v\} \right] + o_p(1).$$

It follows that $\widehat{\mathcal{W}}_v(t) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \xi_i(t,v) + o_p(1)$,

$$\xi_i(t,v) = \frac{\xi_{i1}(t,v) + \{S_{Y_i}(t) - S_{c_v}(t)\} I(Y_i > c_v)}{1-v} + S_{c_v}(t) - \mathrm{PPV}(t,v). \qquad \text{(B·1)}$$

This, together with a functional central limit theorem, implies that $\widehat{\mathcal{W}}_v(t)$ converges weakly to a zero-mean Gaussian process.

C. Asymptotic Properties of $\widehat{\mathbf{PPV}}^*(t,v)$

We assume the same regularity conditions as in Andersen and Gill (1982), who showed that $n^{\frac{1}{2}}(\widehat{\beta} - \beta_0)$ is asymptotically normal and $n^{\frac{1}{2}}(\widehat{\Lambda}_0(t) - \Lambda_0(t))$ converges weakly to a Gaussian process. Similar to the derivation for $\widehat{\mathrm{PPV}}(t,v)$, following standard empirical processes theory (Pollard, 1990), we can show that the process $\widehat{W}_v^*(t) = n^{1/2}\{\widehat{\mathrm{PPV}}^*(v,t) - \mathrm{PPV}(v,t)\} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \zeta_i(t,v) + o_p(1)$, with

$$\zeta_i(t,v) = \frac{I(Y_j > c_v)\{\zeta_{i1}(\beta_0, t, Y_j) + S_{Y_i}(t) - S_{c_v}(t)\}}{1-v} + S_{c_v}(t) - \mathrm{PPV}(t,v), \text{(C·1)}$$

where

$$\zeta_{i1}(\beta_0, t, y, \Delta_i, X_i) = S_y(t) e^{\beta_0 y} \left[\int_0^t \frac{dM_i(u)}{s_0(u,\beta_0)} + \{\Lambda_0(t) y + \mathcal{H}(t,\beta_0)\} \mathcal{I}^{-1}(\beta_0) \int_0^{\infty} \left\{ Y_i - \frac{r_1(u,\beta_0)}{r_0(u,\beta_0)} \right\} dM_i(u) \right],$$

$M_i(t) = N_i(t) - \int_0^t I(X_i \geq u) e^{\beta Y_i} d\Lambda_0(u)$, $r_b(t,\beta) = E\{I(X_i \geq t) Y_i^b e^{\beta Y_i}\}$, $\mathcal{H}(\beta,t)$ is the limit of $\partial \widehat{\Lambda}_0(t,\beta)/\partial\beta$, and $\mathcal{I}(\beta) = \int_0^{\infty} \{r_2(\beta,u)/r_0(\beta,u) - r_1^2(\beta,u)/r_0^2(\beta,u)\} dE\{N(u)\}$.

Table 1

Estimates (95% Confidence Intervals) of $NPV(t,v)$, $PPV(t,v)$ and $PPV^{-1}(p)$ with Various Sample Percentiles (v) and Average Risk Probabilities (p) Evaluated at $t = 1$ Year after Enrollment.

			NPV		
v	.1	.3	.5	.7	.9
$\widetilde{\text{NPV}}(t,v)$	.97(.94, .99)	.96(.94, .98)	.96(.94, .98)	.96(.94,.97)	.93(.92, .95)
$\widehat{\text{NPV}}(t,v)$	.98(.94, 1.00)	.96(.94, .99)	.96(.94, .98)	.96(.94, .97)	.94(.92, .95)
$\widehat{\text{NPV}}^{*}(t,v)$	.97(.96, .98)	.96(.95, .97)	.95(.94, .96)	.94(.93, .96)	.93(.91, .94)
			PPV		
v	.1	.3	.5	.7	.9
$\widetilde{\text{PPV}}(t,v)$	.09(.08, .11)	.11(.09, .13)	.14(.11, .17)	.19(.15, .24)	.29(.19, .40)
$\widehat{\text{PPV}}(t,v)$	.09(.07, .11)	.11(.09, .13)	.13(.10, .16)	.19(.14, .23)	.29(.20, .38)
$\widehat{\text{PPV}}^{*}(t,v)$	.09(.08, .11)	.11(.09, .13)	.13(.10, .15)	.16(.13, .19)	.23(.18, .27)
			$\text{PPV}^{-1}(t,p)$		
p	.10	.15	.20	.25	.30
$\widehat{\text{PPV}}^{-1}(t,p)$	.23 (.01, .45)	.59 (.46, .73)	.75(.64, .86)	.85(.76, .94)	.90(.83, .96)

Table 2

Estimates (95% Confidence Intervals), [95% Confidence bands] of $rPPV(t,v)$ at t=1, 2, 3 Years with Various Sample Percentiles (v).

	v= .80	v=.85	v=.90	v=.95
t = 1	1.11 (.98, 1.26)	1.14 (.98, 1.32)	1.26 (1.05, 1.51)	1.40 (1.11, 1.76)
	[.91, 1.36]	[.89, 1.44]	[.93, 1.70]	[.96, 2.05]
t = 2	1.08 (.99, 1.19)	1.14 (1.02, 1.27)	1.25 (1.09, 1.43)	1.31 (1.09, 1.58)
	[.95, 1.23]	[.98, 1.32]	[1.03, 1.51]	[1.00, 1.72]
t = 3	1.05 (.95, 1.16)	1.06 (.95, 1.19)	1.15 (.99, 1.34)	1.26 (1.06,1.49)
	[.90, 1.21]	[.90, 1.26]	[.92, 1.43]	[.98,1.61]

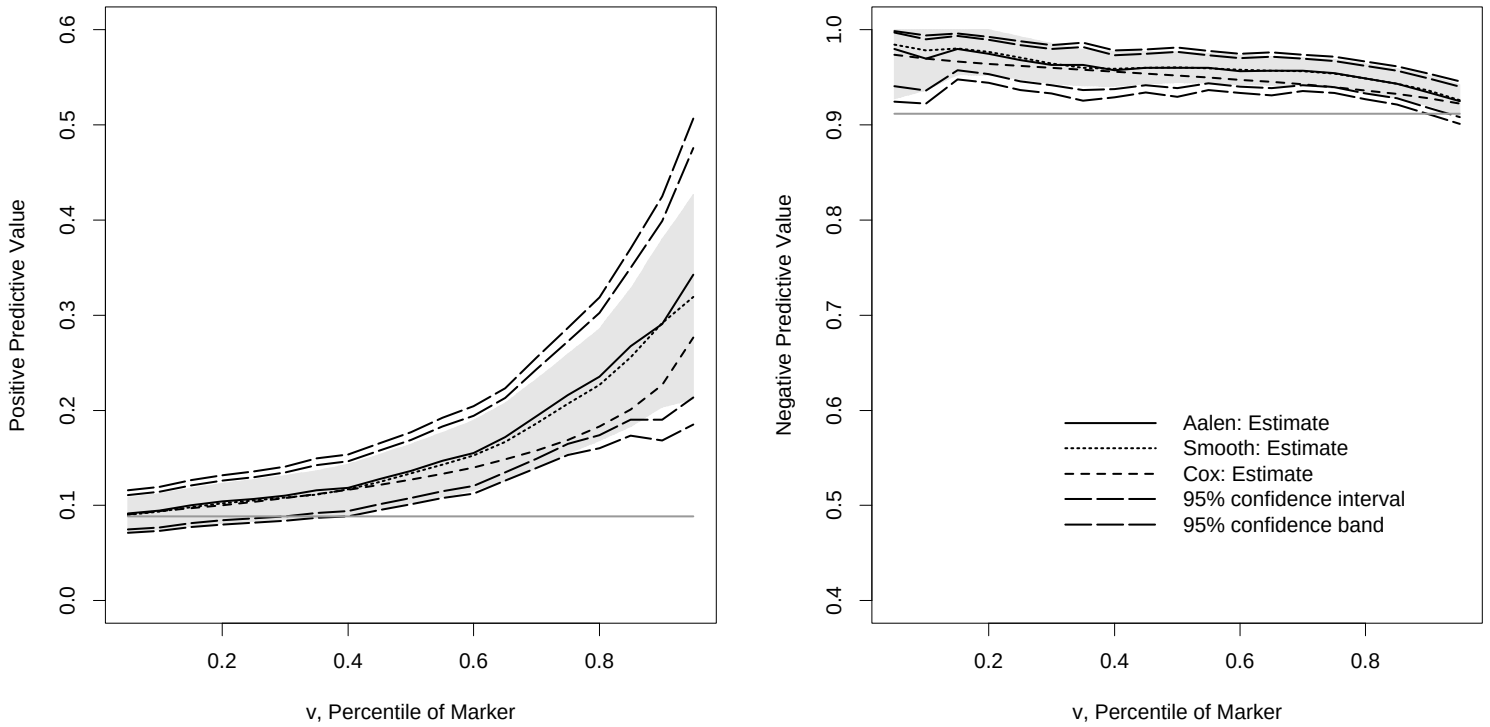

Figure 1. PPV (left panel) and NPV (right panel) curves and 95% confidence intervals and bands for $v \in (0.05, 0.95)$ and at $t = 1$ year after enrollment in Seattle Heart Failure Study. Solid lines: estimates from Aalen estimator; dotted lines: smooth estimator; short dashed lines: the Cox estimator. long dashed lines, 95% confidence intervals and confidence bands (outer curves) for Aalen estimator; Shaded areas are confidence interval for the smooth estimator. Horizontal lines are for $P(T < 1\ year)$ in the PPV plot and $P(T \geq 1\ year)$ in the NPV plot.

References

Akritas, M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *The Annals of Statistics* **22**, 1299–1327.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study (Com: p1121-1124). *The Annals of Statistics* **10**, 1100–1120.

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

Beran, R. (1981). Nonparametric regression with randomly censored survival data. *Technical report[Univ. California, Berkeley]* .

Bilias, Y., Gu, M. and Ying, Z. (1997). Towards a general asymptotic theory for Cox model with staggered entry. *The Annals of Statistics* **25**, 662–682.

Blanks, R., Moss, S. and Wallis, M. (2001). Monitoring and evaluating the UK national health service breast screening programme: evaluating the variation in radiological performance between individual programmes using ppv-referral diagrams. *Journal of medicical Screen* **8**, 24–28.

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate. *The Annals of Statistics* **17**, 1157–1167.

Du, Y. and Akritas, M. G. (2002). Uniform strong representation of the conditional kaplan-meier process. *Mathematical Methods of Statistics* **11**, 152–182.

Gail, M. and Pfeiffer, R. (2005). On criteria for evaluating models of absolute risk. *Biostatistics* **6**, 227–39.

Graf, E., Schmoor, C., Sauerbrei, W. and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* **18**, 2529–2545.

Lee, D., Austin, P., Rouleau, J., Liu, P., Naimark, D. and Tu, J. (2003). Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model. *JAMA* **290**, 2581–7.

Levy, W., Mozaffarian, D., Linker, D. et al. (2006). The Seattle heart failure model: prediction of survival in heart failure. *Circulation* **113**, 1424–33.

Moskowitz, C. and Pepe, M. (2004a). Quantifying and comparing the accuracy of binary biomarkers when predicting a failure time outcome. *Statistics in Medicine* **23**, 1555–1570.

Moskowitz, C. and Pepe, M. (2004b). Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. *Biostatistics* **5**, 113–127.

Park, Y. and Wei, L. J. (2003). Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika* **90**, 717–23.

Parzen, M. I., Wei, L. J. and Ying, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika* **81**, 341–350.

Pepe, M., Feng, Z., Huang, Y. et al. (2007). Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology.* **In press**.

Pollard, D. (1990). *Empirical processes: theory and applications.* Institute of Mathematical Statistics.

Schemper, M. and Henderson, R. (2000). Predictive accuracy and explained variation in Cox regression. *Biometrics* **56**, 249–255.