UW Biostatistics Working Paper Series

4-8-2005

Application of the Time-Dependent ROC Curves for Prognostic Accuracy with Multiple Biomarkers

Yingye Zheng
Fred Hutchinson Cancer Research Center, yzheng@fhcrc.org

Tianxi Cai
Harvard University, tcai@hsph.harvard.edu

Ziding Feng
University of Washington & Fred Hutchinson Cancer Research Center, zfeng@fhcrc.org

1. Introduction

Alterations in gene sequences, expression levels and protein functions can be used as molecular markers to signal the onset or progression of disease. With the rapid advances in genomic and proteomic technologies, the focus on biomarker based disease detection and risk assessment has now shifted from a single biomarker to a panel of biomarkers (Srivastava & Gopal-Srivastava, 2002). For example, a 70-gene prognosis profile has been demonstrated to be a stronger prognostic factor for breast cancer than the standard clinical and histological information (Van de Vijver, He, Van't Veer and etal. 2002). As another example, aberrant methylation pattern of 14 tumor suppressor genes has been proposed to separate different forms of hematological malignancies (Takahashi etal., 2004). Indeed, compared with a single clinical or genetic marker alone, a panel of multiple biomarkers may contain a higher level of discriminatory information, particularly across large heterogeneous patient populations and for complex multistage diseases such as cancer.

Once a molecular profile or a panel of markers for disease is identified, such information can potentially be used as diagnostic tests for diagnosing disease, monitoring the response to therapy and the early detection of disease. But prior to the use of these tests in clinical practice, the actual prognostic values of the markers need to be fully established. Statistical measures have long been available to evaluate the accuracy of a continuous marker M under the traditional diagnostic setting where M is measured concurrently with the binary disease status D. Commonly used measures include sensitivity, specificity and the receiver operating characteristic (ROC) curves (McNeil and Hanley 1984; Pepe 2003; Zhou, Obuchowski and McClish 2002) motived as follows. Suppose for any given threshold value c, a subject is classified as test positive if $M > c$ and test negative if $M \leq c$. The accuracy of this classification rule can be characterized by a pair of parameters: sensitivity or the "true positive rate" (TPR), and 1-specificity or the "false positive rate" (FPR), with

$$TPR(c) = P[M_i > c|D_i = 1], \qquad FPR(c) = P[M_i > c|D_i = 0].$$

An ROC curve displays the full spectrum of values for TPR and FPR, by considering all possible

test threshold values c. The higher the ROC curve, the better capacity of a test for distinguishing diseased from non-diseased population. The area under an ROC curve (AUC), ranges from .5 to 1, has an interpretation as the probability that the test result from a diseased subject exceeds that from a non-diseased subject, were the two subjects chosen randomly. It is recognized that ROC curves are invariant with respect to the measurement scale. This is particularly useful in a study for novel genetic markers where there is a great interest to demonstrate the incremental prognostic value of the markers over routine clinical information.

In prospective cohort studies the disease status of a subject often changes during the course of the study and there is often a time lag between when the marker is measured and the occurences of disease. To evaluate the accuracy of such marker, one needs to take the time lag into account since the accuracy may be higher when the markers are measured closer to the onset of disease. To extend the notion of diagnostic accuracy to incorporate the time domain, Heagerty et al (2000), Cai et al (2003) proposed various definitions of time dependent sensitivity, specificity and ROC curve. Among those, a more straightforward definition is the cumulative incidence based ROC curves where at any given time t, the goal is to discriminate between subjects who will by time t from those will not. Specifically, let T denote the failure time and M denote the marker value measured at baseline. Heagerty, Lumley and Pepe (2000) defines cumulative incidence based ROC curves $\mathrm{ROC}_t^{\mathbb{C}}(\cdot)$ based on the following time dependent TPR and FPR functions:

$$TPR_t^{\mathbb{C}}(c) = P[M > c|D(t) = 1], \qquad FPR_t^{\mathbb{C}}(c) = P[M > c|D(t) = 0].$$

where $D(t) = I(T \leq t)$. Using this definition an individual can be a control early in time when $t < T$ but contributes as a case when $t \geq T$. Therefore, the outcome classification is 'dynamic'. Incorporation of survival time data into ROC analysis has recently been discussed by a number of authors (Heagerty et al. 2000; Heagerty and Zheng 2005). A non-parametric method that characterizes the accuracy using disease prevalence for the case definition, $D_i(t) = 1(T_i \leq t)$, is given by Heagerty et al. (2000).

When there are multiple markers available to assist in prediction, it is of clinical interest to construct an optimal prognostic index based on available marker information. In the standard binary setting, various procedures have been proposed to construct composite scores that offer optimal ROC curves or AUCs (Baker, 2000; Pepe and Thompson, 2000; McIntosh and Pepe, 2002; Pepe, Cai and Zhang, 2005). For failure time outcomes, the task of combining a set of markers in a way that best discriminates between the two populations of interest remains to be solved. Methods developed for the binary setting are not directly applicable because the disease status at a given time t is not always ascertainable due to censoring. The most frequently used approach is to fit a Cox proportional hazards model (Cox, 1972) and use the risk score for prediction. However, it has not been shown whether the risk score obtained under the Cox model offers optimal accuracy in discriminating $D(t) = 1$ from $D(t) = 0$ over time. In addition, the proportional hazards assumption may not hold in practice. In this paper, we develop time specific composite scores under a flexible time varying logistic regression model and demonstrate that the resulting score maximizes the time specific ROC curves when the model assumption holds. We show how the existing statistical methods on ROC analysis can be adapted to provide a quantitative basis for this evaluation. In particular, we focus on statistical rules for combining markers that achieve optimality criterion, and that accommodate time-varying marker effects.

The paper is organized as follows. In section 2 we consider modified binary regression procedures to develop optimal prognostic scores. The performances of classification rules from the proposed procedures are then compared with those from several commonly used failure time regression procedures via numerical studies. Results are reported in Section 3. We use a gene-expression profile study as an example to illustrate our methods. We close in section 4 with recommendation for practice and plans for future development.

2. Linear combination of markers

We consider linearly combining a panel of markers $\mathbf{M} = [M_1, ..., M_P]^\intercal$ as

$$s_\beta(\mathbf{M}) = \sum_{p=1}^{P} \beta_p M_p,$$

Note that M_p can represent any transformation of the raw marker value, thus the composite score $s(\mathbf{M})$, although taking a linear form, can be quite flexible. Other approach for combining multiple resource of information for classification exist for binary disease outcome. For example, one may use logic rules that focus on and-or combination of markers (Etzioni, Kooperberg, Pepe and Smith, 2003). In this manuscript we restrict our attention to the class of linear predictors.

We first address the question of under what circumstances $s_\beta(\mathbf{M})$ gives rules that best discriminate between those with $D(t) = 1$ from those with $D(t) = 0$. To this end, we define the disease status at time t as $D(t) = I(T \leq t)$ and the log-odds function at t as

$$\Re^{\mathbb{C}}(t, \mathbf{M}) = \text{logit}P\{D(t) = 1 \mid \mathbf{M}\} = \text{logit}P\{D(t) = 1\} + \log \frac{P\{\mathbf{M} \mid D(t) = 1\}}{P\{\mathbf{M} \mid D(t) = 0\}}.$$

Thus $\Re^{\mathbb{C}}(t, \mathbf{M})$ is a monotone function of the likelihood ratio $P(\mathbf{M} \mid T \leq t)/P(\mathbf{M} \mid T > t)$. In the binary setting, analogous to the context of statistical hypothesis testing, It can be argued with the Neyman-Pearson lemma (Neyman and Pearson, 1933) that the likelihood ratio function of the P markers, $LR(\mathbf{M}_i) = P(\mathbf{M}_i \mid D_i = 1)/P(\mathbf{M}_i \mid D_i = 0)$ is optimal. For survival outcome, at any fixed t, using the same argument as in (McIntosh and Pepe 2002), we can show that $\Re^{\mathbb{C}}(t, \mathbf{M})$ is the optimal combination of $\mathbf{M}$ in distinguishing $D(t) = 1$ from $D(t) = 0$. That is, the $\text{ROC}_t(u)$ for $\Re^{\mathbb{C}}(t, \mathbf{M})$ is maximized at every false positive rate u. Therefore if we assume that $\Re^{\mathbb{C}}(t, \mathbf{M})$ is characterized by a linear function $\alpha + \boldsymbol{\beta}^T\mathbf{M}$, then the rules based $s_\beta(\mathbf{M}) = \boldsymbol{\beta}^T\mathbf{M} > c$, or $s_\beta(\mathbf{M}) = g(\boldsymbol{\beta}^T\mathbf{M}) > c$, with g denotes an arbitrary monotone increasing function, are optimal. We thus translate the problem of combining markers into seeking estimators that are consistent for model (2), since under that model $s_\beta(\mathbf{M}_i, t)$ gives the best predictive accuracy at time t. We give detailed estimating procedures below.

2.1 Modeling the time-varying effects of markers

Since the predictive accuracy of a marker may be higher when it is measured closer to the time of disease occurrence, we consider the following time-varying logistic regression model for the event time:

$$\text{logit}P(T \leq t \mid \mathbf{M}) = \alpha(t) + \boldsymbol{\beta}(t)^T\mathbf{M}.$$

By permitting the coefficients to be time-dependent, the model may be a more realistic representation of the relationship between biomarker process and the time to an important clinical event. If the survival status is observable at t for every subject, then we can apply a standard logistic regression to directly estimate $\boldsymbol{\beta}$ based on the disease status indicator $D(t)$. However, with censored survival data, the disease status of those subjects who are censored before t is unknown. Assuming a random censorship, we propose an estimator for $\boldsymbol{\beta}(t)$ by adapting the idea of the inverse probability weighting (IPW) approach (Horvitz and Thompson 1951).

Suppose for the ith subject, we observe $\mathbf{M}_i$, $X_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$, where C is the censoring time. To estimate $\alpha(t)$ and $\boldsymbol{\beta}(t)$ in the presence of censoring, we consider the following reweighted logistic score equation:

$$U_{IPW}\{\alpha(t), \boldsymbol{\beta}(t)\} = \sum_i^n \frac{V_i(t)}{\widehat{\pi}_i(t)} \begin{pmatrix} 1 \\ \mathbf{M}_i \end{pmatrix} [D_i(t) - \mu_i\{\alpha(t), \boldsymbol{\beta}(t)\}]$$

where $V_i(t) = 0$ if $X_i < t$ and $\delta_i = 0$, and $V_i(t) = 1$ otherwise, $\widehat{\pi}_i(t)$ is a consistent estimator of the selection probability

$$\pi_i(t) = P\{V_i(t) = 1 \mid \mathbf{M}_i\} = \begin{cases} P(C_i > X_i \mid \mathbf{M}_i, X_i) & \text{if } X_i \leq t, \delta_i = 1 \\ P(C_i > t \mid \mathbf{M}_i) & \text{if } X_i > t \end{cases},$$

$\mu_i\{\alpha(t), \boldsymbol{\beta}(t)\} = P\{D_i(t) = 1 \mid \mathbf{M}_i\} = F\{\alpha(t) + \boldsymbol{\beta}(t)^T\}$ and $F(x) = \exp(x)/\{1 + \exp(x)\}$. The proposed estimator thus uses only those cases with $V_i(t) = 1$ for the logistic regression model, but it weighs each observation by the inverse of the selection probability $\pi_i(t)$. In practice, we estimate π_i using the information from the entire sample. In the simplest case where censoring distribution

does not depend on $\mathbf{M}$, $P(C > t \mid \mathbf{M}) = P(C > t)$ can be estimated using a Kaplan-Meier estimator based on $\{X_i, 1 - \delta_i\}$. When the censoring distribution depends on $\mathbf{M}$, we can estimate covariate-specific censoring probabilities fitting a proportional hazards model to the data $\{X_i, 1 - \delta_i, \mathbf{M}_i\}$. In both cases, the estimator for $P(C > t \mid \mathbf{M})$ is consistent uniformly in $\mathbf{M}$. Under such uniform consistency assumption, one can easily establish the consistency for the estimated $\boldsymbol{\beta}(t)$ by observing that

$$E\left\{ \frac{V_i(t)}{\pi_i(t)} \middle| T_i, \mathbf{M}_i \right\} = E\left\{ \frac{I(X_i \geq t)}{P(C_i > t \mid \mathbf{M}_i)} + \frac{I(X_i \leq t)\delta_i}{P(C_i > X_i \mid X_i, \mathbf{M}_i)} \middle| T_i, \mathbf{M}_i \right\} = 1$$

The weighted logistic estimator has several advantages. The robustness of the time varying coefficient model makes it broadly applicable in practice. The estimator is easy to implement using standard software and is theoretical sound. When the model holds, the resulting composite score is the optimal score. Using the same argument as given in Eguchi and Copas (2002) for the binary setting, it is not hard to show that when the logistic model is approximately correct, the estimated composite score minimizes a weighted area between the $\mathrm{ROC}_t^{\mathbb{C}}(\cdot)$ of the linear score and of the true optimal score. Thus even if the model is mis-specified, the resulting score remains optimal in a certain sense.

2.2 Modeling the constant effects of markers

When the effects of markers on T are constant over time, or when there is an interest to obtain a single classification rule to predict the overall risk, one may wish to instead consider a proportional odds model:

$$\mathrm{logit} P(T \leq t \mid \mathbf{M}) = \alpha(t) + \boldsymbol{\beta}^T(t)\mathbf{M} = \alpha(t) + \boldsymbol{\beta}^T\mathbf{M}. \tag{1}$$

Under this model, the estimator from the time-varying logistic model at any given time t is a consistent estimator of $\boldsymbol{\beta}$. However, these estimators do not account for the fact that $\boldsymbol{\beta}(t) = \boldsymbol{\beta}$ and thus may not be efficient. To estimate $\boldsymbol{\beta}$ under this model, non-parametric maximum likelihood methods and generalized estimating equation based methods has been proposed (Cheng, Wei and

Ying, 1995, 1997; Murphy, Rossini and Van der Vaart, 1997). However these methods are either computationally intensive or not efficient. We propose an alternative approach that is relatively simple yet efficient. Our proposal is to extend the weighted logistic score estimator proposed by Newey (2004) for non-censored data to the setting where T is subject to censoring.

Specifically, we partition the time axis into J nonoverlapping intervals with cut-off points $t_1, ..., t_j, ..., t_{J-1}$ and let $D_{ij} = I(t_{j-1} \leq T_i < t_j)$, $\boldsymbol{\theta} = (\alpha_1 = \alpha(t_1), ..., \alpha_{J-1} = \alpha(t_{J-1}), \boldsymbol{\beta}^T)^T$, $\alpha_0 = -\infty$ and $\alpha_J = \infty$. Newey and Smith (2004) showed that the optimal estimating equation for $\boldsymbol{\theta}$ is

$$\sum_{j=1}^{J-1} \sum_{i=1}^{n} \frac{\partial \log \frac{\nabla \mu_{ij}(\boldsymbol{\theta})}{\nabla \mu_{i,j+1}(\boldsymbol{\theta})}}{\partial \boldsymbol{\theta}} \{D_i(t_j) - \mu_i(\alpha_j, \boldsymbol{\theta})\} = 0. \tag{2}$$

where $\mu_i(\alpha_j, \boldsymbol{\beta}) = F\{\alpha_j + \boldsymbol{\beta}^T \mathbf{M}_i\}$ and $\nabla \mu_{ij}(\boldsymbol{\theta}) = E\{D_{ij} \mid \mathbf{M}_i, \boldsymbol{\theta}\} = F\{\alpha_j + \boldsymbol{\beta}^T \mathbf{M}_i\} - F\{\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{M}_i\}$. Here the estimator is optimal in the sense that as $J \to \infty$, its asymptotic variance approaches the semiparametric bound. To account for censoring, we modify 2 and consider the following inverse probability reweighted estimating equation:

$$\sum_{j=1}^{J-1} \sum_{i=1}^{n} \frac{V_i(t_j)}{\widehat{\pi}_i(t_j)} \frac{\partial \log \frac{\nabla \mu_{ij}(\boldsymbol{\theta})}{\nabla \mu_{i,j+1}(\boldsymbol{\theta})}}{\partial \boldsymbol{\theta}} \{D_i(t_j) - \mu_i(\alpha_j, \boldsymbol{\beta})\} = 0. \tag{3}$$

To obtain estimates for $\boldsymbol{\theta}$ in practice, one may solve a set of $J - 1 + P$ estimating equations simultaneously. For easy implementation, we adapt a two-step estimating procedure to obtain efficient estimates for $\boldsymbol{\theta}$. We first fit separate logistic models at each cut-off points and construct an initial estimator for $\boldsymbol{\theta}$, $\widetilde{\boldsymbol{\theta}}$. Then we solve (3) with weights $\partial \log \frac{\nabla \mu_{ij}(\boldsymbol{\theta})}{\nabla \mu_{i,j+1}(\boldsymbol{\theta})} / \partial \boldsymbol{\theta}$ evaluated at $\widetilde{\boldsymbol{\theta}}$. The resulting two-step estimators are consistent and asymptotically normal, and can be shown to have equivalent efficiency as the one associated with equation (3). We give the detailed estimating procedure in Appendix.

In practice, a linear composite score can be generated by any regression model for censored event time. For example Cox proportional hazards model could be a more common choice (Fan, Au, Heagerty etal., 2002). We focus on proportional odds model here because of its theoretical

merits. Similar to its counterpart in the binary setting, it yields optimal linear function when the log-odds function is indeed linear. When the true log-odds function is unknown, it may still give good discriminant function in most of the situation as it approximately minimizes the weighted area between the ROC curve for linear function and for linear score. In contrast, the optimality properties for other regression models are less clear theoretically, and their performance in practice should be compared to the models that is known to achieve optimality criterion. We carry out a number of numerical studies to evaluate the performance of the proposed approaches for optimal marker combination.

3. Simulation Study

In this section, we evaluate the performance of the composite scores obtained from the proposed procedures under the time varying logistic model and the proportional odds model. We also compare these scores to the scores derived from fitting two commonly used survival models: 1) the Cox proportional hazards model with a maximum partial likelihood estimator; and 2) the Accelerated Failure Time (AFT) model (Tsiatis, 1990) with a Gehan type estimator.

For all the simulation studies presented below, the censoring time was generated from a normal with mean μ_c and unit variance with where μ_c was chosen to induce about about 30% of censoring. The censoring distribution was estimated using the Kaplan-Meier estimator. To estimate $\boldsymbol{\beta}$ in the proportional odds model, the time axis was divided into 10 non-overlapping intervals. For each simulated dataset, we estimate the regression coefficient $\boldsymbol{\beta}$ under all four models and evaluate the accuracy of the composite score, $\widehat{s}(\mathbf{M}) = \widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{M}$, in distinguishing $D(t) = 1$ from $D(t) = 0$ for three time points, $t = t_1, t_2$ and t_3, chosen as the 25%, 50% and 75% percentile of X. The accuracy measure we use is the time specific AUC, $\mathrm{AUC}_t = \int_0^1 \mathrm{ROC}_t^{\mathbb{C}}(u)du$.

Time Varying Logistic Model We first investigated the performance of various methods under the time varying logistic regression model. We generate a panel of five marker $\mathbf{M} = (M_1, \ldots, M_5)^T$

from the exponentially transformed multivariate normal with zero mean, unit variance and correlation 0.3. The failure time T is generated from

$$T = 10 \exp\left\{ \frac{\epsilon}{c_0 + \sum_p^5 c_p M_p} \right\}$$

with $P(\epsilon \leq x) = \frac{e^x}{1+e^x}$. Under this configuration, T follows a time-varying logistic model:

$$\text{logit}P\{D(t) = 1 \mid \mathbf{M}\} = c_0 \log(t/10) + \sum_{p=1}^{5} c_p \log(t/10) M_p.$$

In Table 1, we present the average AUC_t of the scores derived from all four models for $t = t_1$, t_2 and t_3 at sample size of $n = 200$ and $n = 1000$. For comparison, shown also are the AUC_t's for the optimal score using the likelihood ratio function of $\mathbf{M}$. Results suggest that across all the time points, the combined scores based on our weighted logistic regression yield highest AUC values, on average the same as these from the true model. In contrast, the scores obtained from other models have lower accuracy. For example, when $n = 200$ at t_1, the average AUCs are .82 for the logistic estimator, .73 under the Cox model, 0.65 under the AFT model and 0.58 under the proportional odds model. This is consistent with our theoretical speculation that the time varying logistic estimator, by modeling directly the likelihood ratio function, is advantageous over the other regression models.

Proportional Odds Model We now assess the robustness and the relative efficiency of the four procedures under the proportional odds model. We generate the first two markers from generated from a $(50\%, 50\%)$ mixture of $M_1 \sim \text{Uniform}(0,\, 2), M_2 \sim \text{Uniform}(0,\, 2) + 5M_1^2$, and $M_1 \sim \text{Weibull}(1,\, 0.5), M_2 \sim \text{Weibull}(1,\, 0.5) + 5M_1^2$; and the remaining three markers are generated independently from standard normal. The survival time was generated from a proportional odds model

$$h(T) = -0.8M_1 - 0.2M_2 + \epsilon,$$

with $P(\epsilon \leq x) = \frac{e^x}{1+e^x}$ and $h^{-1}(x) = 10\left\{2\Phi\left(x/5\right) - 1\right\}^3 + 10$. Thus, M_3, M_4 and M_5 have no effect on the survival time. Since the marker effects are constant over time and the risk ratio function is linear, we expect the proportional odds model to perform well as it is the right model under such an assumption. The weighted logistic regression still offers consistent estimates, However it can be less efficient. Our results shown in Table 2 are consistent with what we expect. Although both consistent, the estimator of $\boldsymbol{\beta}$ from the proportional odds model is indeed more efficient than the estimator from the time-specific logistic regression model. For example, at sample size of $n = 1000$, the efficiency of the time-specific logistic estimator for β_1 ranges from 44% to 62% relative to the estimator from the proportional odds model.

We then assessed the performance of the four methods in terms of the area under the ROC curves. Table 3 shows the estimated AUCs (top rows) and the corresponding standard errors (bottom rows). In contrast to the time-specific logistic model and the proportional odds model, we found that the Cox model and the additive failure time model were biased when estimating the effects of markers (results not shown). Surprisingly, it appears that the bias in these models does not translate into meaningful reductions in the performance of the corresponding linear predictors, as the estimated AUC are essentially the same across almost all situations for the logistic, the proportional odds and the Cox model. The performance of AFT model again is slightly worse. In this setting we find the proportional odds model is still advantageous compared to the other models, as the resulting estimates are considerably less variable, as is evident from the smallest standard errors.

Mixture Model when the Time Varying Logistic Model Fails We now investigate the performance of these procedures when the data are generated from a mixture model and thus none of these four models hold. Specifically, the failure times were derived from a mixture model: $s(\boldsymbol{\beta}, \mathbf{M}_i) = \sum_{p=1}^{3} \beta_p M_{ip}$ if $\mathcal{I} = 0$ and $s(\boldsymbol{\beta}, \mathbf{M}_i) = \sum_{p=4}^{5} \beta_p M_{ip}$ if $\mathcal{I} = 1$ with $\mathcal{I} \sim$ Bernoulli(0.9).

Under this configuration, the likelihood ratio function can not be characterized by a simple linear function $\boldsymbol{\beta}^T\mathbf{M}_i$, and clearly neither of the four models holds in this scenario. Therefore this setting allows us to evaluate the classification performance under the misspecified models. The results, shown in Table 1, suggest that although not optimal, the score derived from the time varying logistic procedure has higher accuracy than the other scores. For example, at $t = t_2$ with a sample size of $n = 1000$, the average AUC_t is 0.87 from the time varying logistic method, which is higher than 0.76 from the Cox model, 0.72 from the AFT model and 0.75 from the proportional odds model. We do not expect the proportional odds model to do well in this setting as it was not designed to estimate any time-vary effect.

In summary, the simulation studies demonstrate that the operating characteristics of the four procedures differ in regards to their attained predictive accuracies, and the differences depend on whether the effects of biomarkers vary with time. When the underlying marker effects vary with time, and the likelihood ratio function is of a linear function of the markers at each time point, we found that the time varying logistic regression indeed offers the optimal prognostic score. When the optimal likelihood ratio score is not a monotone function of $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{M}$, the time varying logistic regression still produces a better linear score compared to other procedures. This is again consistent with , as it seeks to somehow minimize the difference between the ROC curve from the true likelihood ratio function and that from the logistic model all the way across the curve. Under the setting where the marker effects are constant over time, we found that the classification performances are somewhat comparable for the four methods. Nevertheless the proportional odds model is attractive because it is the most efficient procedure in both estimating the marker effect and predictive accuracy, hence it is the most powerful especially when sample size is small. Finally, we observed that in general the Cox proportional hazards model performed better when compared to the AFT model, and in many cases, particularly under the time-invariant marker model, it even yields linear scores that classify as accurately as the those from a proportional odds model. This is however not so surprising as

the two models are not fundamentally different in their operating characteristics within the class of transformational model, when the effect of markers are indeed proportional in nature (Dabrowska & Doksum, 1988, Box & Cox, 1982). These results are encouraging.

4. Example

We analyze a publicly available cDNA microarray dataset from a study of breast cancer reported by Van de Vijver et al. (2002). The gene expression measurement is the logarithm of the intensity ratios between the red and the green fluorescent dyes, where green dye is used for the reference pool and red is used for the experimental tissue. The data consists of a previously established 70-gene prognosis profile from 295 breast cancer patients who were diagnosed with breast cancer between 1984 and 1995. The median survival time is 3.8 years for these patients. By separating the patients into groups with good and poor-prognosis signature, the authors showed that the prognosis profile was a strong predictor for disease outcome, providing survival information beyond that of the standard and histologic criteria. Here we use the data to evaluate the prognostic values of the gene prognosis profile by applying the time-dependent ROC curve methods. Particularly, we aim to assess how well the genes selected can distinguish between subjects who die and subjects who do not in a follow-up interval $[0, t]$, with t chosen to be 2, 5 and 8 years after diagnosis for illustration.

First, are all 70 genes necessary for prognosis prediction? It would be advantageous to decrease the number of predictor genes for clinical applications. The issues of model selection, particularly from high-dimensional data, although relevant here, are beyond the scope of the current paper. We used the data to simply illustrate that different choices of regression models can lead to linear composite score with different classification performance. We therefore consider a small set of 6 genes that were selected using a forward stepwise procedure. We suppose the expression levels of the six genes are given, and the goal is to estimate the linear score based on these six genes.

We obtain optimal linear composite scores based on these 6 genes using the aforementioned four procedures and then assess the capacity these scores have in discriminating between patients would

or would not survive t years after breast cancer diagnosis, for $t = 2$, 5 and 8. The estimated AUCs and their 95% confidence intervals for the composite scores obtained through these procedures are reported in Table 4. To illustrate the benefit from combining the marker panel, we also show in Table 4 the accuracies of a single gene, which univariately has the highest AUC among all 70 genes. As expected, great improvement in the accuracy was achieved when consider all six genes together. For example, at $t = 5$ years, the estimated AUC is 0.75 (95% CI = $(0.66, 0.83)$) if a single gene is used. With a panel of six genes, the corresponding accuracy increases to 0.84 (95% CI = $(0.80, 0.89)$) based on the linear scores of a logistic regression model. Furthermore, if some clinical information is available, by comparing ROC curve based on the clinical information only with the ROC curve using both the clinical and gene profile information, we can decide if the gene profile provides improved discrimination for cumulative mortality at different follow-up times.

We found that for this study, the linear scores from the four methods yield comparable area under the curves for distinguishing long term survivors when $t = 5$ and $t = 8$. The weighted logistic regression, however, appeared to be more accurate and more efficient classification rules for distinguishing short term survivors, in this case, at $t = 2$ years after diagnosis. For example, at the second year, when 80% of the surviving women are screened negative, 74% of the women who died are screen positive based a threshold derived from the logistic regression, whereas with the same specificity (80%), the associated sensitivity is 68% from that of the Cox model.

Figure 1 compares the ROC curves at different failure times $t = 2$, $t = 5$ and $t = 8$ years for the logistic regression model (panel (a)) and the Cox model (panel (b)) respectively. Based on the ROC curves for logistic model, it seems that the operating characteristics of the set of genes selected may vary over time: they are more sensitive at identifying women that die within the first few years but their discriminatory power for cumulative mortality decreases when the goal is to identify long terms survivors. In the contrary, for the Cox model, it is puzzling at a first glance that the ROC curve at $t = 5$ dominates the ROC curve at $t = 2$. This is not too surprising, as

the estimated AUC from a Cox model at $t = 2$ was not as efficient as that of a logistic regression, a finding that is consistent with our simulation results under the time-varying marker effects. We conclude that the logistic model, adopting different linear functions at different times, appears to provide better separation for ROC curves at different cut-off times, compared with the Cox model that assumes a time invariant linear function for marker combination.

5. Discussion

In this paper, we develop new procedures for combining markers to optimally predict future disease status which may vary over time. These procedures may be used to construct prognostic indices that can potentially be used for early detection of disease and risk assessment. It seems a natural step to extend the time-dependent ROC methodology to studies based on multiple biomarkers such as a gene expression profile for survival prediction (Li and Gui, 2004). Regression models for failure time data are routinely used in practice as a basis for constructing optimal prognostic scores. In this manuscript we provide justification for their use by demonstrating that the derived scores achieve optimal accuracy in the view of time specific ROC curves. We showed that in the limit, the linear score $\widehat{\boldsymbol{\beta}}(t)^{\top}\mathbf{M}$ is the achieves optimal accuracy if the logistic model holds and is an optimal linear combination of $\mathbf{M}$ if the model only holds approximately. The robustness property may not hold if the model is severely mis-specified. In the binary setting, robust optimization procedures through maximizing AUC has been recently proposed (Pepe, Cai and Zhang, 2005). Further development on more robust methods for time-dependent ROC curve methodology is warranted. However, we note that the robustness of our proposed procedures can alternatively be improved by including quadratic or higher order polynomial functions of the predictors.

Drew in the recent literature for seeking optimal classification rules for binary outcome, we have showed that a good linear discriminant function can be estimated directly from a time varying logistic regression procedure by appealing to the Neyman-Pearson lemma. Other failure time regression models, for example Cox regression models, may provide good approximation of the true

underlying object function under certain condition, however their theoretical merits in regard to optimality are difficult to evaluate. Therefore the time varying logistic regression can be used as a benchmark for assessing the performance of these other models in practice. Another advantage of the approach is that it is more likely to capture the time-varying nature of the markers, since it allows different linear functions of the markers for different target times. When the effect of markers are approximately constant over time, such a model can be used to build more efficient estimating procedures. The estimating procedures for the proportional odds model we suggested is easy to implement, and yield predictor and accuracy summaries that are efficient. In our numerical study we also found that Cox regression model can be quite robust. Thus under the time-vary effect, one may consider Cox regression model when there is an interest to report just a single combination rule to achieve good predictive accuracy over time. The implication of these results are very important in practice.

There are other issues need to be addressed when the source of markers is from recently developed high-throughput technologies. One issue arises is how to select a small set of markers from a pool of hundreds or thousands candidate genes. The other issue relates to the fact that the crude retrospective error rate for classification can be optimistic for assessing the performance of a classifier in a small sample. the area under the time-dependent ROC curves offers a measure of the goodness of fit for multiple regression model, and it has been shown that under complex data structure of nonlinear regression it could be more sensible to the traditional measures of R^2 (Heagerty and Zheng, 2005). When the objective of the analysis is to search for markers that maximize the prognostic accuracy at time t, $AUC(t)$, after being properly adjusted for overfitting, can be considered as an objective function for both model fitting and selection.

Appendix A. Two-step estimator for proportional odds model

We note that model (2.1), althrough inefficient, is however consistent under model (1). We therefore first fit $J-1$ weighted logistic regression models at $t_1, ..., t_{J-1}$ respectively and obtain corresponding $\tilde{\alpha_j}, \tilde{\boldsymbol{\beta}}_j$. Let $\tilde{\boldsymbol{\beta}} = \sum_{j=1}^{J-1} \tilde{\boldsymbol{\beta}}_j$, then $\tilde{\boldsymbol{\theta}} = (\tilde{\alpha}_1, ..., \tilde{\alpha}_{J-1}, \tilde{\boldsymbol{\beta}}^T)$ can be employed as initial values for solving equation (3).

As an initial step, let

$$\begin{aligned}
\omega_{1j}(\mathbf{M}_i, \boldsymbol{\theta}) &= \frac{\partial log \frac{\nabla \mu_{j-1}(\mathbf{M}_i, \boldsymbol{\theta})}{\nabla \mu_j(\mathbf{M}_i, \boldsymbol{\theta})}}{\partial \alpha_j} = -\omega_{0i}(t_j) \frac{f(\alpha_j + \boldsymbol{\beta}^T \mathbf{M}_i)}{F(\alpha_j + \boldsymbol{\beta}^T \mathbf{M}_i) - F(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{M}_i)} \\
\omega_{2j}(\mathbf{M}_i, \boldsymbol{\theta}) &= \frac{\partial log \frac{\nabla \mu_j(\mathbf{M}_i, \boldsymbol{\theta})}{\nabla \mu_{j+1}(\mathbf{M}_i, \boldsymbol{\theta})}}{\partial \alpha_j} = \frac{f(\alpha_j + \boldsymbol{\beta}^T \mathbf{M}_i)\{F(\alpha_{j+1} + \boldsymbol{\beta}^T \mathbf{M}_i) - F(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{M}_i)\}}{\{F(\alpha_j + \boldsymbol{\beta}^T \mathbf{M}_i) - F(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{M}_i)\}\{F(\alpha_{j+1} + \boldsymbol{\beta}^T \mathbf{M}_i) - F(\alpha_j + \boldsymbol{\beta}^T \mathbf{M}_i)\}} \\
\omega_{3j}(\mathbf{M}_i, \boldsymbol{\theta}) &= \frac{\partial log \frac{\nabla \mu_{j+1}(\mathbf{M}_i, \boldsymbol{\theta})}{\nabla \mu_{j+2}(\mathbf{M}_i, \boldsymbol{\theta})}}{\partial \alpha_j} = -\frac{f(\alpha_j + \boldsymbol{\beta}^T \mathbf{M}_i)}{F(\alpha_{j+1} + \boldsymbol{\beta}^T \mathbf{M}_i) - F(\alpha_j + \boldsymbol{\beta}^T \mathbf{M}_i)},
\end{aligned}$$

where $f(x)$ is the derivative of $F(x)$. We have from model (3), for any given $\boldsymbol{\beta}$, $\hat{\boldsymbol{\theta}}(\boldsymbol{\beta}) = (\hat{\alpha}(\boldsymbol{\beta})^T, \boldsymbol{\beta}^T)^T$ is the solution to

$$\begin{aligned}
\sum_i^n \omega_{0i}(t_j) \Big[\omega_{1j}(\mathbf{M}_i, \tilde{\boldsymbol{\theta}})\{D_i(t_{j-1}) - \mu_i(t_{j-1}, \boldsymbol{\theta})\} &+ \omega_{2j}(\mathbf{M}_i, \tilde{\boldsymbol{\theta}})\{D_i(t_j) - \mu_i(t_j, \boldsymbol{\theta})\} \\
&+ \omega_{3j}(\mathbf{M}_i, \tilde{\boldsymbol{\theta}})\{D_i(t_{j+1}) - \mu_i(t_{j+1}, \boldsymbol{\theta})\} \Big] = 0,
\end{aligned}$$

for $j = 1, ..., J-1$. At the second step, we can obtain an efficient $\hat{\boldsymbol{\beta}}$ as the solution to

$$\sum_{j=1}^{J-1} \sum_{i=1}^{n} W(t_j, M_i, \tilde{\boldsymbol{\theta}}) \mathbf{M}_i \Big\{ D_i(t_j) - \mu_i(t_j, \hat{\boldsymbol{\theta}}(\boldsymbol{\beta})) \Big\} = 0,$$

with

$$W(t_j, M_i, \tilde{\boldsymbol{\theta}}) = \omega_{0i}(t_j) \left\{ \frac{f(\hat{\alpha}_j + \boldsymbol{\beta}^T \mathbf{M}_i) - f(\hat{\alpha}_{j-1} + \boldsymbol{\beta}^T \mathbf{M}_i)}{F(\hat{\alpha}_j + \boldsymbol{\beta}^T \mathbf{M}_i) - F(\hat{\alpha}_{j-1} + \boldsymbol{\beta}^T \mathbf{M}_i)} - \frac{f(\hat{\alpha}_{j+1} + \boldsymbol{\beta}^T \mathbf{M}_i) - f(\hat{\alpha}_j + \boldsymbol{\beta}^T \mathbf{M}_i)}{F(\hat{\alpha}_{j+1} + \boldsymbol{\beta}^T \mathbf{M}_i) - F(\hat{\alpha}_j + \boldsymbol{\beta}^T \mathbf{M}_i)} \right\}$$

REFERENCES

Baker, S. G. (2000), "Identifying combinations of cancer markers for further study as triggers of early intervention," *Biometrics*, 56(4), 1082–1087.

Box, G. E. P., and Cox, D. R. (1982), "An analysis of transformations revisited, rebutted," *Journal of the American Statistical Association*, 77, 209–10.

Cheng, S. C., Wei, L. J., and Ying, Z. (1995), "Analysis of transformation models with censored data," *Biometrika*, 82, 835–845.

Cox, D. R. (1972), "Regression models and life-tables (with discussion)," *Journal of the Royal Statistical Society, Series B, Methodological*, 34, 187–220.

Dabrowska, D., and Doksum, K. (1988), "Estimation and Testing in a Two-Sample Generalized Odds-Rate Model," *Journal of the American Statistical Association*, 83, 744–9.

Eguchi, S., and Copas, J. (2002), "A class of logistic-type discriminant functions," *Biometrika*, 89(1), 1–22.

Etzioni, R., Kooperberg, C., Pepe, M., Smith, R., and Gann, P. (2003), "Combining biomarkers to detect disease with application to prostate cancer," *Biostatistics*, 4, 523–538.

Fan, V., Au, D., Heagerty, P., Deyo, R., McDonell, M., and Fihn, S. (2002), "Validation of case-mix measures derived from self-reports of diagnoses and health," *Journal of Clinical Epidemiology*, 55, 371–380.

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000), "Time-dependent ROC curves for censored survival data and a diagnostic marker," *Biometrics*, 56(2), 337–344.

Heagerty, P., and Zheng, Y. (2005), "Survival Model Accuracy and ROC Curves," *Biometrics*, .

Horvitz, D., and Thompson, D. J. (1951), "A generalization of sampling withut replacement from a finie universe," *Journal of the American Statistical Association*, 47, 663–685.

Li, H., and Gui, J. (2004), "Partial Cox Regression Analysis for High-Dimensional Microarray Gene Expression Data," *Bioinformatics*, 20, i208–i215.

Lin, D. Y., and Ying, Z. (1994), "Semiparametric analysis of the additive risk model," *Biometrika*, 81, 61–71.

McIntosh, M. W., and Pepe, M. S. (2002), "Combining several screening tests: Optimality of the risk score," *Biometrics*, 58(3), 657–664.

McNeil, B., and Hanley, J. (1984), "Statistical approaches to the analysis of receiver operating characteristic curves," *Medical Decision Making*, 4(1), 137–150.

Murphy, S. A., Rossini, A. J., and van der Vaart, A. W. (1997), "Maximum likelihood estimation in the proportional odds model," *Journal of the American Statistical Association*, 92, 968–976.

Newey, W. K. (2004), "Efficient semiparametric estimation via moment restrictions," *Econometrika*, 72, 1877–1897.

Neyman, J., and Pearson, E. S. (1933), "On the problem of the most efficient tests of statistical hypothesis," *Philosophical Transactions of the Royal society of London, Series A*, 231, 289–337.

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

Pepe, M. S., Cai, T., and Zhang, Z. (2003), "Robust binary regression for optimally combining predictors," *University of Washington Biostatistics Working Paper Series*, http://www.bepress.com/uwbiostat/paper198, 1–29.

Srivastava, S., and Gopal-Srivastava, R. (2002), "Biomarkers in Cancer Screening: a Public Health Perspective," *Journal of Nutrition*, 132, 2471s–2475s.

Takahashi, T., Shivapurkar, N., reddy, J., Shigematsu, H., Miyajima, K., and etal. (2004), "DNA methylation profiles of lymphoid and hematopoietic malignancies," *Clinical Cancer Research*, 10, 2928–2935.

Tsiatis, A. A. (1990), "Estimating regression parameters using linear rank tests for censored data," *The Annals of Statistics*, 18, 354–72.

Van de Vijver, M. J., He, Y. D., Van't Veer, L. J., and etal. (2002), "A gene-expression signatiure as a predictor of survival in breast cancer," *The New England Journal of Medicine*, 347, 1999–2009.

Zhou, X.-H., Obuchowski, N., and McClish, D. (2002), *Statistical Methods in Diagnostic Medicine*, New York: Wiley.

Table 1: Estimated AUC (standard errors) under time-varying marker effects from weighted logistic regression, Cox proportional hazards model (Cox), additive failure time model (AFT) and the proportional odds model. Results are based on 200 simulated datasets for each sample size.

	$n = 200$			$n = 1000$		
	t_1	t_2	t_3	t_1	t_2	t_3
			Time Varying Logistic Model			
True	.89	.85	.72	.89	.85	.72
Logistic	.91(.03)	.89(.04)	.75(.05)	.88(.03)	.85(.03)	.71(.02)
Cox	.77(.11)	.74(.09)	.64(.07)	.83(.05)	.80(.04)	.68(.02)
AFT	.76(.11)	.73(.10)	.66(.06)	.75(.12)	.73(.10)	.65 (.06)
Proportional Odds	.78 (.12)	.74(.07)	.73(.11)	.75(.10)	.72(.06)	.65(.06)
			Mixture Model			
True	.93	.92	.85	.93	.92	.85
Logistic	.93(.01)	.92(.03)	.82(.06)	.92(.03)	.87(.04)	.76(.03)
Cox	.78 (.13)	.76(.12)	.70 (.09)	.82(.11)	.76(.09)	.73(.04)
AFT	.80 (.11)	.76(.11)	.68 (.08)	.76(.12)	.72(.09)	.67 (.06)
Proportional Odds	.83(.11)	.80(.11)	.71(.08)	.80(.12)	.75(.09)	.68(.07)

Note: True AUCs are the AUCs estimated using the true parameters for simuation.

Table 2: Sample average (empirical standard error) of the estimated β_p under the proportional odds model when β is estimated from the time specific logistic regression at given $t = t_1$, t_2 and t_3 and when β is estimated from the proportional odds model. True $\boldsymbol{\beta}$ are 0.8,0.2,0,0 and 0 respectively.

	$n = 200$				$n = 1000$			
	$\widehat{\beta}(t_1)$	$\widehat{\beta}(t_2)$	$\widehat{\beta}(t_3)$	$\widehat{\beta}$	$\widehat{\beta}(t_1)$	$\widehat{\beta}(t_1)$	$\widehat{\beta}(t_1)$	$\widehat{\beta}$
M_1	1.00(1.43)	.87(1.31)	.80(1.20)	.95(1.05)	.79(.60)	.80(.55)	.76(.51)	.79(.40)
M_2	.19(.14)	.21(.15)	.21(.16)	.20(.10)	.20(.07)	.20(.07)	.20(.06)	.20(.04)
M_3	.00(.23)	-.01(.18)	.00(.18)	.00(.14)	.00(.10)	.00(.08)	.00(.07)	.00(.06)
M_4	.01(.21)	.00(.19)	-.01(.17)	.00(.14)	.00(.09)	.00(.07)	.00(.07)	.00(.06)
M_5	.00(.23)	.00(.18)	.00(.18)	.00(.15)	.00(.10)	.00(.08)	-.01(.07)	.00(.06)

Table 3: Estimated AUC and their standard errors under time-independent marker effects from the proportional odds model, weighted logistic regression, Cox proportional hazards model(Cox) and additive failure time model(AFT).

		$n = 200$			$n = 1000$		
		t_1	t_2	t_3	t_1	t_2	t_3
Estimates	Proportional Odds	.86	.80	.76	.86	.80	.77
	Logistic	.84	.79	.76	.85	.80	.77
	Cox	.85	.80	.76	.86	.80	.77
	AFT	.85	.79	.75	.85	.80	.76
Std. Err. $\times 10^2$	Proportional Odds	.39	.66	.83	.06	.12	.16
	Logistic	.66	.67	.91	.14	.15	.21
	Cox	.42	.70	.89	.09	.16	.22
	AFT	.43	.83	1.09	.18	.36	.50

Table 4: Estimated AUC(95% CI) at 2, 5 and 8 years after diagnosis using a 6-gene classifier with linear composite scores derived from different regression models. For comparison, shown also are the estimates of AUC from a single gene with the best predictive accuracy.

	$t = 2$ years	$t = 5$ years	$t = 8$ years
a single gene	.75(.56, .87)	.75(.66, .83)	.68(.59, .76)
Logistic	.85(.80, .91)	.84(.80, .89)	.77(.71, .84)
Cox	.78(.62, .87)	.84(.78, .88)	.77(.71, .84)
AFT	.81(.70, .88)	.84(.81, .89)	.78(.72, .84)
Proportional Odds	.78(.59, .87)	.83(.68, .88)	.77(.65, .84)

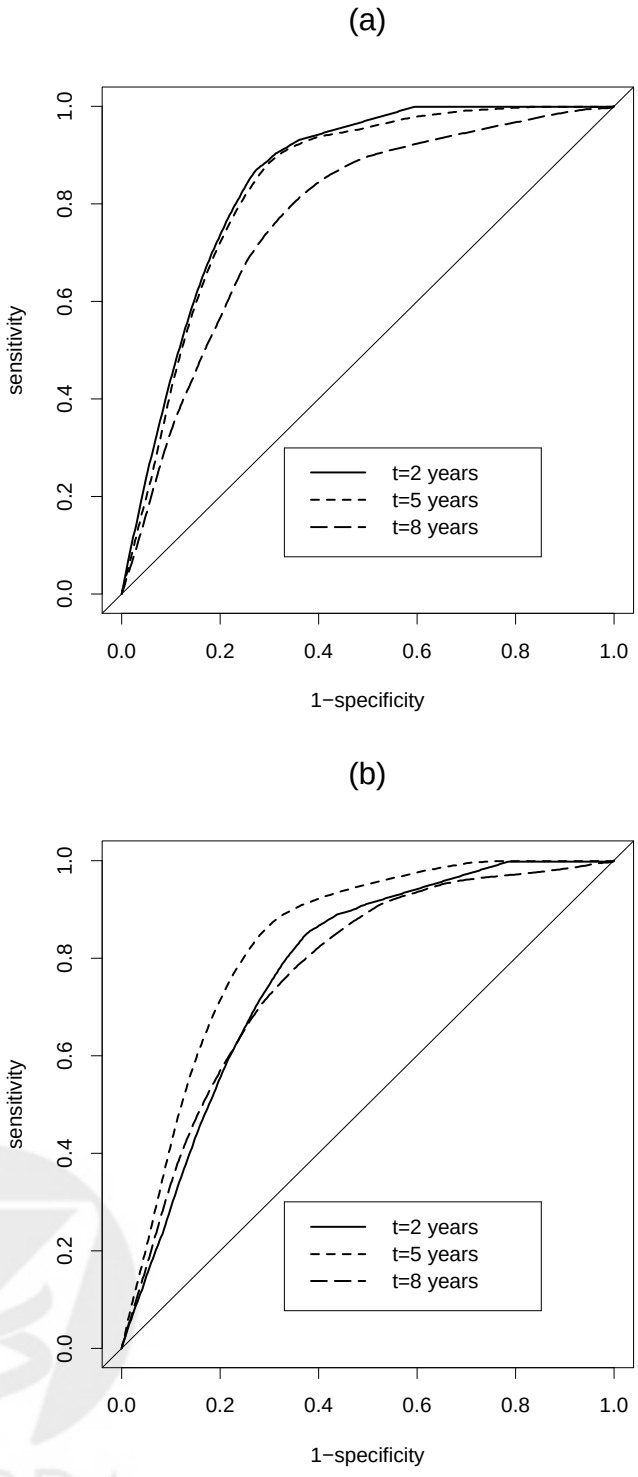

Figure 1: (a) ROC curves at $t = 2$, $t = 5$, and $t = 8$ based on the linear function of 6 genes selected using the logistic regression model. (b) ROC curves at $t = 2$, $t = 5$ and $t = 8$ based on the linear function of 6 genes selected using the Cox regression model.