Semiparametric prognosis models in genomic studies
Shuangge Ma, Jian Huang, Mingyu Shi, Yang Li and Ben-Chang Shia
Submitted: 26th October 2009; Received (in revised form): 4th December 2009

Abstract
Development of high-throughput technologies makes it possible to survey the whole genome. Genomic studies have been extensively conducted, searching for markers with predictive power for prognosis of complex diseases such as cancer, diabetes and obesity. Most existing statistical analyses are focused on developing marker selection techniques, while little attention is paid to the underlying prognosis models. In this article, we review three commonly used prognosis models, namely the Cox, additive risk and accelerated failure time models. We conduct simulation and show that gene identification can be unsatisfactory under model misspecification. We analyze three cancer prognosis studies under the three models, and show that the gene identification results, prediction performance of all identified genes combined, and reproducibility of each identified gene are model-dependent. We suggest that in practical data analysis, more attention should be paid to the model assumption, and multiple models may need to be considered.

Keywords: genomic studies; semiparametric prognosis models; model comparison

INTRODUCTION
For complex diseases such as cancer, diabetes and obesity, extensive biomedical studies have shown that clinical and environment risk factors may not have sufficient predictive power for prognosis. In the past decade, we have witnessed unparalleled development in high-throughput technologies. Such development makes it possible to survey the whole genome and search for genomic markers that may have predictive power for disease prognosis. Gene signatures have been constructed for prognosis of breast cancer, ovarian cancer, lymphoma, obesity and many other diseases [1, 2]. To avoid ambiguity, we limit ourselves to gene expressions measured using microarrays but note that the results may also be applicable to other profiling studies.

Denote T as the survival time, which can be progression free, overall, or other type of survival. Denote C as the censoring time. Denote Z as the length d gene expression measurement. Under right censoring, one observes \( Y = \min(T, C), \Delta = I(T \leq C, Z) \). We assume that \( T \sim \phi(\beta' Z) \), where \( \beta \) is the regression coefficient. The most common choice for \( \phi \) is the Cox proportional hazards model [3, 4], while other models have also been adopted. Of note, there are a few model-free approaches. However, they have not been extensively used and will not be discussed.

Data generated in microarray studies has the ‘large d, small n’ characteristic—a typical study measures expressions of \( 10^3-10^4 \) genes on \( 10^1-10^3 \) subjects. When fitting regression models with number of genes...
larger than sample size, proper regularization is needed. In addition, among the thousands of genes profiled, only a subset are associated with disease prognosis. Identification of disease susceptibility genes can not only improve model fitting by removing noises but also lead to better understanding of the mechanisms underlying disease development. Many statistical approaches have been developed for regularized estimation and gene selection. Examples include dimension reduction approaches such as the partial least squares, singular value decomposition and principal component analysis, feature selection approaches including the filter, wrapper and embedded approaches, and hybrid approaches such as the sparse principal component analysis. Comprehensive reviews of dimension reduction and variable selection methods have been rendered by various authors [5–7].

Despite great efforts on gene selection techniques, insufficient attention has been paid to underlying prognosis models. Usually, it is assumed that a certain model (for example, the Cox model) holds, while insufficient justification is provided to support such an assumption. As shown by Fleming and Harrington [8] and Klein and Moeschberger [4], the Cox model may fail, and alternative models—such as the accelerated failure time (AFT) model and additive risk model—may fit the data better. Although various model diagnosis methods have been proposed, their validity is established under the assumption that number of covariates is much smaller than sample size. To the best of our knowledge, none of these methods has been proved valid under the ‘large d, small n’ setup.

In this article, we first review three semiparametric prognosis models. We then describe the Lasso approach, which will be used for gene identification in this study. We also describe cross validation-based approaches for evaluation of overall prediction performance (of all identified genes combined) and evaluation of reproducibility of each identified gene. We conduct simulation and show that gene identification can be unsatisfactory under model misspecification. We analyze data from three cancer prognosis studies under the three different model misspecification. We show that the gene identification results, prediction performance and reproducibility of identified genes are model-dependent. The article concludes with discussion.

It is not our intention to show the superiority of certain models or to develop a generically applicable model comparison/selection criterion. Rather, our intention is to raise the awareness of multiple available prognosis models and demonstrate that different datasets may demand different prognosis models.

**SEMIPARAMETRIC PROGNOSIS MODELS**

Parametric models are likely to suffer from model misspecification. Semiparametric models can be much more flexible than parametric models, while enjoying the interpretability not shared by nonparametric models.

### Cox proportional hazards model

The most widely adopted prognosis model is the Cox proportional hazards model [3, 4]. Under this model, the conditional hazard function is

\[ \lambda(t|Z) = \lambda_0(t) \exp(\beta Z) \]  

Here \( \lambda_0(t) \) is the baseline hazard, and \( \beta \) is the regression coefficient.

Assume \( n \) iid observations \((Y_i, \delta_i, Z_i), i = 1 \ldots n \). The log-partial likelihood function is

\[ R_n(\beta) = \sum_{i=1}^{n} \delta_i \{ \beta' Z_i - \log(\sum_{k \in r_i} \exp(\beta' Z_k)) \} \]  

where \( r_i = \{ k : Y_k \geq Y_i \} \) is the at-risk set at time \( Y_i \).

Examples of applying the Cox model in prognosis studies with microarrays include [9–12] and many others.

### Additive risk model

As a flexible alternative to the Cox model, the additive risk model is adopted when the covariates contribute *additively* to the conditional hazard [13, 14]. Under the additive risk model, the conditional hazard function is

\[ \lambda(t|Z) = \lambda_0(t) + \beta' Z \]  

where notations have similar meanings as under the Cox model.

Unlike for the Cox model, there is not a simple profile likelihood for the additive risk model. Straightforward application of likelihood approaches involves estimating the infinite dimensional parameter \( \lambda_0(t) \). To tackle this problem, Lin and Ying [15] proposed the following approach.

Define \( A_0(t) = \int_0^t \lambda_0(u) \, du \). For the \( ith \) subject, denote \( \{ N_i(t) = I(Y_i \leq t, \delta_i = 1); t \geq 0 \} \) and \( \{ A_i(t) = I(Y_i \geq t); t \geq 0 \} \) as the observed event
process and the at-risk process, respectively. The regression coefficient $\beta$ can be estimated by solving

\[
U(\beta) = \sum_{i=1}^{n} \int_{0}^{\infty} Z_i dN_i(t) - A_i(t) d\hat{\Lambda}(\beta,t) - A_i(t) \beta Z_i dt = 0,
\]

where $\hat{\Lambda}(\beta,t)$ is the estimate of $\Lambda_0$ satisfying

\[
\hat{\Lambda}(\hat{\beta},t) = \sum_{i=1}^{n} \int_{0}^{\infty} \frac{dN_i(t)}{\sum_{i=1}^{n} A_i(t)} - A_i(t) \hat{\beta} Z_i dt.
\]

The resulting estimator of $\beta$ satisfies the estimating equation

\[
\left[ \sum_{i=1}^{n} \int_{0}^{\infty} A_i(t) [Z_i(t) - \bar{Z}(t)]^2 dt \right] \hat{\beta} = \left[ \sum_{i=1}^{n} \int_{0}^{\infty} \{Z_i(t) - \bar{Z}(t)\} dN_i(t) \right].
\]

where $\bar{Z}(t) = \sum_{i=1}^{n} A_i(t) Z_i / \sum_{i=1}^{n} A_i(t)$. Denote

\[
L_i^j = \int_{0}^{\infty} A_i(t) (Z_i(t) - \bar{Z}(t))^2 dt,
\]

\[
R_i^j = \int_{0}^{\infty} \{Z_i(t) - \bar{Z}(t)\} dN_i(t).
\]

Denote the $(j, l)^{th}$ element of $L^j$ as $L_j^l$, and the $j^{th}$ component of $R^j$ as $R_j^l$ respectively. Equation (4) is equivalent to the following $d$ equations:

\[
\left( \sum_{i=1}^{n} L_i^j \right) \beta_1 + \ldots + \left( \sum_{i=1}^{n} L_i^d \right) \beta_d = \sum_{i=1}^{n} R_i^j = 1 \ldots d.
\]

Motivated by (5), Ma and Huang [16] proposed the following objective function

\[
R_n(\beta) = -\sum_{j=1}^{d} \left( \left( \sum_{i=1}^{n} L_i^j \right) \beta_1 + \ldots + \left( \sum_{i=1}^{n} L_i^d \right) \beta_d - \sum_{i=1}^{n} R_i^j \right)^2.
\]

Examples of using the additive risk model in prognosis studies with microarrays include [16–19] and others.

**AFT model**

Both the Cox and additive risk models describe the survival hazard. In contrast, the AFT model describes the event time directly [13, 14]. The AFT model assumes

\[
\log(T) = \alpha + \beta' Z + \epsilon,
\]

where $\alpha$ is the intercept and $\epsilon$ is the random error with an unknown distribution. Here, the log transformation can be replaced with other monotone transformations.

Estimation of the AFT model has been extensively studied. Several available approaches, including the Buckley-James approach [20] and rank-based approaches [21], suffer from a high computational cost when there are a large number of covariates. A computationally more affordable approach is the weighted least squares approach [22].

With slight abuse of notation, define $Y = \min(\log(T), \log(1-C))$. Denote $Y_{(i)} \leq \ldots \leq Y_{(n)}$ as the ordered $Y$s. Denote $\delta_{(i)}$ and $Z_{(i)}$ as the censoring indicator and covariate associated with $Y_{(i)}$. Let $\hat{F}$ be the Kaplan-Meier estimate of the distribution function $F$ of $\log(T)$, which can be computed as $\hat{F}(y) = \sum_{i=1}^{n} w_i I(Y_{(i)} \leq y)$. Here $w_i$ are computed as

\[
w_1 = \frac{\delta(1)}{n},
\]

\[
w_i = \frac{\delta(i)}{n-i+1} \prod_{k=1}^{i-1} \left( \frac{n-k}{n-k+1} \delta(k) \right) \text{ for } i = 2 \ldots n.
\]

With the weighted least squares approach, the objective function is

\[
R_n(\beta) = -\frac{1}{2} \sum_{i=1}^{n} w_i (Y_{(i)} - \alpha - \beta' Z_{(i)})^2.
\]

Examples of using the AFT model in prognosis studies with microarrays include [9, 23–26] and others.

**Remarks**

Denote $S(t)$ as the survival function at time $t$. Denote $S^l(t)$ and $S^2(t)$ as the survival functions for covariate values $Z$ and $Z+1$. Under the Cox model, $\log(-\log(S^l(t)))-\log(-\log(S^2(t)))$ is a constant; under the additive risk model, $\log(S^l(t)))/t$ is a constant; Under the AFT model, a linear relationship between $\log(t)$ and covariates can be observed. Such observations have been used to develop graphic diagnosis tools when there are a small number of covariates [4, 8].

For all the three models, when $d<<n$, $\hat{\beta} = \arg\max R_n(\beta)$ is a consistent estimate of $\beta$. An attractive feature of $R_n(\beta)$ is that it only involves $\beta$ but not the nonparametric parameter. That is, although the three models are much more flexible than parametric models, they have similar computational cost as parametric models. We refer to [27, 28] for thorough discussions of the three models.

There are other semiparametric prognosis models, including the proportional odds model, general transformation model, accelerated hazards model.
and others. Those models are not widely used, and hence will not be considered.

GENE SELECTION AND MODEL EVALUATION

Gene selection using Lasso

Among the thousands of genes surveyed in microarray studies, only a subset are associated with disease prognosis. Hence, variable selection is needed along with model estimation.

Among the many available variable selection techniques, penalization methods have attracted special attention. The most famous penalization methods are perhaps the AIC and BIC, which penalize the number of unknown parameters in likelihood approaches and have been extensively used in biomedical studies. Examples of penalization methods in microarray prognosis studies include [9–11, 16, 23, 24, 29–31] and many others.

When there are a large number of covariates, the AIC and BIC may be unstable and suffer from high computational cost. An alternative penalization approach that has better statistical properties is the Lasso [32], where the estimate is defined as

$$\hat{\beta} = \arg \max \ R_n (\beta) \text{ subject to } \sum_{j=1}^{d} |\beta_j| \leq M \quad (9)$$

Here $R_n$ is defined in (2), (6) and (8), $\beta_j$ is the $j^{th}$ component of $\beta$, and $M$ is the data-dependent tuning parameter.

The Lasso controls complexity of the model by controlling the norms of regression coefficients. It is a special case of the bridge-type penalty defined as $\sum_{j=1}^{d} |\beta_j|^\gamma$. When $\gamma = 0$, this penalty becomes the AIC/BIC-type penalty; When $\gamma = 2$, this penalty becomes the ridge penalty, which has a long history for ill-designed models but cannot conduct variable selection. We refer to Ma and Huang [6], Tibshirani [32], and Zhang and Huang [33] for more detailed discussions of the Lasso and other penalization methods. We note that, the Lasso approach is ‘independent’ of the model and objective function. It can be directly applied to the three prognosis models. We adopt the Lasso for gene selection, because it is the most extensively used approach, has affordable computational cost, and has served as benchmark in many studies. In addition, Zhang and Huang [33] and several other studies have established that, with a high probability, the Lasso is capable of selecting all genes associated with disease outcomes. This is further confirmed by our simulation study presented below. Thus, the Lasso seems to be a proper choice for gene selection. The Lasso has been used in cancer prognosis studies with microarrays under the Cox [10], additive risk [16], and AFT [23] models. The R package ‘penalized’ has been developed to compute the Lasso estimates (we refer to http://cran.r-project.org/web/packages/penalized/penalized.pdf for more information).

In this study, we use the boosting approach to compute the Lasso estimate [34]. The Lasso involves a tuning parameter $M$, which balances goodness-of-fit and sparsity and will be selected using V-fold cross validation.

Evaluation

In simulation, the true generating model is known and independent testing data can be generated. Thus, it is easy to evaluate gene identification and prediction performance.

Analysis of real data can provide more relevant information than simulation [35]. In analysis of real data, with limited understanding of genomic mechanisms underlying disease prognosis, we are unable to determine which model selects biologically the most meaningful genes. Instead, we consider the following statistical evaluation/comparison techniques.

Evaluation of prediction performance

An important goal of genomic studies is to make prediction using identified gene signatures. In addition, prediction performance can provide an indirect assessment of gene identification results—if genes identified are biologically more meaningful, prediction using those genes should be more accurate. Ideally, prediction evaluation should use independent datasets. However, for most genomic studies, it is hard to find independent studies with comparable designs.

We use the Leave-One-Out (LOO) cross validation-based approach for prediction evaluation. For $i = 1 \ldots n$, (i) remove subject $i$ from data; (ii) compute the estimate $\hat{\beta}^{(i-)}$ using the rest $n-1$ subjects. To ensure a fair evaluation, supervised gene screening (if any) and tuning parameter selection need to be conducted with the reduced data. (iii) Compute the predictive risk score $\hat{\beta}^{(i-)} Z_i$; and (iv) repeat steps (i)–(iii) over all $n$ subjects. Dichotomize the $n$ predictive risk scores at the median and create two risk groups. Compute the logrank statistic, which measures difference of survival between the two groups.
the null, the logrank statistic is $\chi^2$ distributed with degree of freedom 1. A significant logrank statistic indicates that the genes selected under a specific model have satisfactory predictive power [4, 28]. The logrank statistic and its significance level can be computed using several available software, for example Proc Lifetest in SAS or the ‘survival’ package in R.

**Evaluation of reproducibility**

The approach described above assesses performance of all identified genes combined. Reproducibility of each identified gene is also of significant interest. Ideally, reproducibility evaluation should be based on analysis of multiple independent datasets. Without such data, we again resort to cross validation.

We consider the following approach: (i) remove one subject from data; (ii) conduct regularized estimation and gene selection using the reduced data with sample size $n-1$; (iii) repeat steps (i) and (ii) over all subjects. For each gene, count $c$, the number of times it is identified in the $n$ models. The ratio $c/n$ provides a measure of reproducibility, and is referred to as the ‘occurrence index’ in the literature [16]. The reproducibility evaluation is a byproduct of the prediction evaluation, and does not incur any additional computational cost.

Under a specific model, the Lasso can identify a set of genes using all observations. We are interested in evaluating whether those genes have reasonably high occurrence indexes, and whether their occurrence indexes are higher than those not identified. This evaluation can reveal whether a specific prognosis model (with the Lasso) can identify the most reproducible genes.

**NUMERICAL STUDIES**

**Simulation**

We conduct simulation and investigate performance of different models under different data generating models. We generate 500 genes, whose marginal distributions are normal. Expressions of genes $i$ and $j$ have correlation coefficient $\rho^{|i-j|}$, with $\rho = 0.1$ (weak correlation) or 0.6 (strong correlation). We generate a training set and a testing set, each with sample size 100. The event times are generated from the Cox, additive risk and AFT models. The censoring times are generated from exponential distributions. We adjust the censoring distribution so that the censoring rate is about 50%. Among the 500 genes, the first 15 have nonzero regression coefficients, i.e. are associated with disease prognosis. For each training set, we apply the three models and use Lasso for gene selection. We then make prediction for subjects in the testing set. Specifically, for each subject in the testing set, we compute the predictive risk score. We then create two risk groups by dichotomizing the risk scores, and compute the logrank statistic.

We compute summary statistics based on 500 replicates and present the results in Table 1. We can see that when the assumed model matches the true underlying model, a majority or all of the true positives can be identified. For example, when genes have weakly correlated expressions and both the assumed and true models are Cox models (row 1 of Table 1), the Lasso identifies all the 15 true positives. The logrank statistic is highly significant, indicating satisfactory prediction performance. In contrast, under model misspecification, the Lasso fails to identify the true positives. For example, when the true model is the Cox model but the additive risk model is assumed (row 2), only 4 true positives are identified.

<table>
<thead>
<tr>
<th>True model</th>
<th>Correlation</th>
<th>Assumed model</th>
<th>Positive</th>
<th>True positive</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cox</td>
<td>0.1</td>
<td>Cox</td>
<td>27</td>
<td>15</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Additive</td>
<td>36</td>
<td>4</td>
<td>0.153</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AFT</td>
<td>50</td>
<td>5</td>
<td>0.306</td>
</tr>
<tr>
<td>0.6</td>
<td></td>
<td>Cox</td>
<td>18</td>
<td>15</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Additive</td>
<td>38</td>
<td>8</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AFT</td>
<td>40</td>
<td>4</td>
<td>0.020</td>
</tr>
<tr>
<td>Additive</td>
<td>0.1</td>
<td>Cox</td>
<td>40</td>
<td>5</td>
<td>0.418</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Additive</td>
<td>25</td>
<td>13</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AFT</td>
<td>10</td>
<td>0</td>
<td>0.731</td>
</tr>
<tr>
<td>0.6</td>
<td></td>
<td>Cox</td>
<td>40</td>
<td>5</td>
<td>0.231</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Additive</td>
<td>22</td>
<td>13</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AFT</td>
<td>10</td>
<td>0</td>
<td>0.402</td>
</tr>
<tr>
<td>AFT</td>
<td>0.1</td>
<td>Cox</td>
<td>27</td>
<td>10</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Additive</td>
<td>25</td>
<td>8</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AFT</td>
<td>26</td>
<td>14</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>0.6</td>
<td></td>
<td>Cox</td>
<td>17</td>
<td>9</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Additive</td>
<td>13</td>
<td>5</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AFT</td>
<td>22</td>
<td>15</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Positive: number of genes identified; true positive: number of identified genes truly associated with disease; P-value: P-value of the logrank statistic.
Cancer microarray studies

**DLBCL study**
Rosenwald *et al.* [36] reported a diffuse large B-cell lymphoma (DLBCL) prognosis study. This study retrospectively collected tumor biopsy specimens and clinical data for 240 patients with untreated DLBCL. The median follow up is 2.8 years, with 138 observed deaths. Lymphochip cDNA microarray is used to measure the expressions of 7399 genes. Raw data and detailed experiment protocol are available at http://llmpp.nih.gov/DLBCL/.

**MCL study**
Rosenwald *et al.* [37] reported a study using microarray gene expression analysis in mantle cell lymphoma (MCL). Among 101 untreated patients with no history of previous lymphoma, 92 were classified as having MCL based on established morphologic and immunophenotypic criteria. Survival times of 64 patients were available and 28 patients were censored. The median survival time was 2.8 years (range 0.02–14.05 years). Lymphochip DNA microarrays were used to quantify mRNA expressions in the lymphoma samples from the 92 patients. Gene expression data that contains expression values of 8810 cDNA elements is available at http://llmpp.nih.gov/MCL.

**Breast cancer study**
Breast cancer is the second leading cause of death from cancer among women in the United States. Despite major progress in breast cancer treatment, the ability to predict metastasis of the tumor remains limited. van’t Veer *et al.* [38] reported a breast cancer prognosis study investigating the time to distant metastasis. 97 lymph node-negative breast cancer patients, 55 years old or younger, participated in this study. Among them, 46 developed distant metastases within 5 years. Expression levels for 24 481 gene probes were collected. Data is available at http://www.rri.com/publications/2002/vantveer.html.

**Data processing**
For all the three datasets, we fill in missing expressions using medians across all samples. We normalize gene expressions to have median zero and variance one. In addition, to reduce computational cost, we conduct the following supervised screening. Using uncensored observations, we compute the correlation coefficient between event time and each gene expression. We select 500 genes with the largest absolute values of correlation coefficients. Since the number of genes associated with prognosis is expected to be much smaller than 500, the supervised screening should lead to very few false negatives. Of note, to ensure a fair evaluation, the supervised screening needs to be conducted for each reduced dataset with one observation removed.

**Data analysis results**
We analyze the three datasets under the three prognosis models. We use the Lasso for regularized estimation and gene selection. The tuning parameter $M$ is selected using 5-fold cross validation.

**Gene identification**
We show the summary of gene identification results under different prognosis models in Table 2. We can see that, under each model, a small number of genes are identified. The most interesting observation is that gene sets identified under different models have little or even no overlap. For example, for the DLBCL data, under the Cox and additive risk models, 39 and 25 genes are identified, respectively. However, there are only 5 genes in common. We conclude from Table 2 that ‘gene identification results are model-dependent’.

It has been hypothesized that although genes identified using different approaches (or under different models) are different, they may belong to the same pathways and/or represent the same biological processes. Our knowledge of genes’ biological functions is still very limited. For example, when searching the KEGG, we are only able to find pathway information for less than half of the genes for the three datasets. In addition, genes, as opposed to pathways or biological processes, are still the most commonly used functional units in genomic studies.

<table>
<thead>
<tr>
<th></th>
<th>Cox</th>
<th>Additive</th>
<th>AFT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DLBCL</strong></td>
<td>39</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Additive</td>
<td></td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td>AFT</td>
<td></td>
<td>49</td>
<td></td>
</tr>
<tr>
<td><strong>MCL</strong></td>
<td>37</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>Additive</td>
<td>20</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>AFT</td>
<td></td>
<td>55</td>
<td></td>
</tr>
<tr>
<td><strong>Breast</strong></td>
<td>21</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Additive</td>
<td>13</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>AFT</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Thus, we focus on comparing identification results at the gene level.

**Evaluation of prediction**

We use the LOO approach to evaluate prediction performance of genes identified under different models. The logrank statistics are shown in Table 3. We can see that for the DLBCL and MCL data, prediction under all the three models is satisfactory ($P$-value < 0.05). For the Breast cancer data, prediction under the Cox and additive models is satisfactory ($P$-value < 0.05). The most interesting observation is that, if we use the logrank statistic—which measures prediction performance—as the criterion for model selection, ‘the three datasets have three different optimal models’.

**Evaluation of reproducibility**

We use the LOO approach to evaluate reproducibility, and show the summary statistics of occurrence indexes in Table 4. We can see that the genes identified using all observations have reasonably high occurrence indexes. In contrast, genes not identified have much lower occurrence indexes. We also see that all models miss a few genes with high occurrence indexes. In addition, ‘if occurrence index, which is a measure of reproducibility, is chosen as the criterion for model selection, the Cox model outperforms the other two’.

### Table 3: Evaluation of prediction: Logrank statistics under different models

<table>
<thead>
<tr>
<th></th>
<th>Cox</th>
<th>Additive</th>
<th>AFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLBCL</td>
<td>10.0</td>
<td>6.03</td>
<td>12.50</td>
</tr>
<tr>
<td>MCL</td>
<td>18.10</td>
<td>13.10</td>
<td>4.80</td>
</tr>
<tr>
<td>Breast</td>
<td>4.23</td>
<td>10.00</td>
<td>3.78</td>
</tr>
</tbody>
</table>

A logrank statistic greater than 3.84 is significant at the 0.05 level.

**DISCUSSION**

A major goal of genomic studies is to identify disease markers. Although great effort has been devoted to development of gene selection techniques, insufficient attention has been paid to the assumed prognosis models. In this article, we review the three most extensively used semiparametric prognosis models. We conduct simulation and find that model misspecification can lead to unsatisfactory gene identification results. We analyze three datasets under the three models, and find that gene identification results, prediction performance and reproducibility are model-dependent. Thus, we suggest that in practical data analysis, researchers should examine multiple prognosis models.

To the best of our knowledge, there is still no consensus on how to select the optimal model under the ‘large d, small n’ setup. The biological implications of identified genes may be the ultimate selection criterion. However, genomic research is still at the discovery stage. Our limited knowledge has forced us to consider statistical, as opposed to

### Table 4: Occurrence Index under different models: median [range]

<table>
<thead>
<tr>
<th></th>
<th>Cox</th>
<th>Additive</th>
<th>AFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLBCL</td>
<td>Selected</td>
<td>0.996 [0.213, 1]</td>
<td>0.458 [0.008, 0.979]</td>
</tr>
<tr>
<td>Not selected</td>
<td>0 [0, 0.696]</td>
<td>0 [0, 0.554]</td>
<td>0 [0, 0.592]</td>
</tr>
<tr>
<td>MCL</td>
<td>Selected</td>
<td>0.978 [0.152, 1]</td>
<td>0.902 [0.239, 0.978]</td>
</tr>
<tr>
<td>Not selected</td>
<td>0 [0, 0.609]</td>
<td>0 [0, 0.467]</td>
<td>0 [0, 0.652]</td>
</tr>
<tr>
<td>Breast</td>
<td>Selected</td>
<td>0.990 [0.835, 1]</td>
<td>0.484 [0.021, 0.928]</td>
</tr>
<tr>
<td>Not selected</td>
<td>0 [0, 0.814]</td>
<td>0 [0, 0.608]</td>
<td>0 [0, 0.474]</td>
</tr>
</tbody>
</table>

Selected: genes selected using all observations.
biological, comparison and evaluation. In this article, we use the LOO approaches for evaluation of prediction and reproducibility. The logrank statistic and the occurrence index are proposed as the evaluation/comparison criteria.

Among the three prognosis models, the Cox and AFT models belong to the family of transformation models. Theoretically speaking, it is possible to fit a general transformation model and then determine whether the Cox or the AFT model is more proper. In an unpublished study, we experimented with this approach and found that the estimation results were unsatisfactory even with a moderate number of covariates. In addition, such an approach cannot determine whether the additive risk model holds.

Our presentation is in the context of microarray studies and we focus on analysis of cancer data. We expect similar conclusions to hold for other profiling platforms and other diseases. We adopt the Lasso for gene selection. We expect that adopting other penalties or other selection methods may change some minor details but not the model-dependent nature of gene selection and prediction performance.

Our literature review suggests that prognosis models other than the Cox model and penalized gene selection have been extensively used in statistical and bioinformatics studies. However, we are unable to find such applications in medical studies. The unfortunate lag between the development of new statistical technologies and their medical applications has been well noted [39]. Given their satisfactory performance demonstrated in statistical studies, we expect the additive risk model, AFT model, and Lasso approach to be used in medical studies in the near future.

**Key Points**

- Multiple prognosis models have been used in genomic studies.
- Our numerical studies show that the gene identification results, predictive performance and reproducibility of identified genes are model-dependent.
- Researchers should explore multiple prognosis models in practical data analysis.
- There is a critical need for model diagnosis and selection methods for high-dimensional genomic data.

**Acknowledgements**
The authors thank the editor and three referees for their insightful comments, which have led to significant improvement of this article.

**FUNDING**
National Institutes of Health [LM009754 and CA120988 to S.M. and J.H.].

**References**


