Genetics: Early Online, published on November 9, 2016 as 10.1534/genetics.116.189191

Pathway-Structured Predictive Model for Cancer Survival Prediction: A Two-Stage Approach

Xinyan Zhang[1], Yan Li[1], Tomi Akinyemiju[2], Akinyemi I. Ojesina[2], Phillip Buckhaults[3], Nianjun Liu[4], Bo Xu[5], and Nengjun Yi[1,*]

[1] Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294, USA

[2] Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL 35294, USA

[3] Department of Drug Discovery and Biomedical Sciences, The South Carolina College of Pharmacy, The University of South Carolina, SC 29208, USA

[4] Department of Epidemiology and Biostatistics, School of Public Health, Indiana University Bloomington, IN 47405-4801, USA

[5] Department of Oncology, Southern Research Institute, Birmingham, AL 35205, USA

*** Corresponding author:**

Nengjun Yi

Department of Biostatistics

University of Alabama at Birmingham

Birmingham, AL 35294-0022

Phone: 205-934-4924

Fax: 205-975-2540

Email: nyi@uab.edu

Key words: cancer prognosis, hierarchical Cox model, pathway, penalized Cox regression, The Cancer Genome Atlas (TCGA)

Running title: Pathway-structured prognosis

Copyright 2016.

ABSTRACT

Heterogeneity in terms of tumor characteristics, prognosis, and survival among cancer patients has been a persistent problem for many decades. Currently, prognosis and outcome predictions are made based on clinical factors and/or by incorporating molecular profiling data. However, inaccurate prognosis and prediction may result by using only clinical or molecular information directly. One of the main shortcomings of past studies is the failure to incorporate prior biological information into the predictive model, given strong evidence of pathway-based genetic nature of cancer, i.e. the potential for oncogenes to be grouped into pathways based on biological functions such as cell survival, proliferation and metastatic dissemination.

To address this problem, we propose a two-stage approach to incorporate pathway information into the prognostic modeling using large-scale gene expression data. In the first stage, we fit all predictors within each pathway using penalized Cox model and Bayesian hierarchical Cox model. In the second stage, we combine the cross-validated prognostic scores of all pathways obtained in the first stage as new predictors to build an integrated prognostic model for prediction. We apply the proposed method to analyze two independent breast and ovarian cancer datasets from The Cancer Genome Atlas (TCGA), predicting overall survival using large-scale gene expression profiling data. The results from both datasets show that the proposed approach not only improves survival prediction compared with the alternative analyses that ignore the pathway information, but also identifies significant biological pathways.

INTRODUCTION

Over the past three decades, remarkable improvement has been achieved in cancer treatment in the United States, with the annual death rate from cancer declining 1.4% for women, 1.8% for men, and 2.3% for children ages 0-10 years from 2002 to 2011 (EDWARDS *et al.* 2014). However, the problem that has persisted in cancer treatment is the heterogeneity of prognostic prediction across patients (BARILLOT 2013). This heterogeneity is, for the most part, genetically determined and rooted in the molecular profile of patients. The Precision Medicine Initiative has been introduced by White House to expand cancer genomics research as the goal to develop better prevention and treatment methods for cancers (COLLINS and VARMUS 2015). Recent high-throughput technologies can easily and robustly generated large-scale molecular profiling data, including genomic, epi-genomic, transcriptomic, and proteomic markers offers extraordinary opportunities to integrate clinical and genomic data in prediction models, improving our understanding of inter-individual differences that may be critical for the application of precision medicine strategies (BARILLOT 2013; COLLINS and VARMUS 2015).

The recent development of molecular signatures to predict recurrence of breast, colon and prostate cancers are notable and clinically useful but may not be sufficient to achieve the goals of precision medicine (MOOK *et al.* 2007; POHL and LENZ 2008; ENG *et al.* 2013). Gene signatures across different studies with very few overlapped genes can have similar prediction results, implying that some underlying mechanisms may exist (VAN DE VIJVER *et al.* 2002; WANG *et al.* 2005; SOTIRIOU and PICCART 2007). According to the analysis of 24 pancreatic tumors by Jones *et al.* (2008), altered genes varied greatly across tumors but the pathways with the altered genes remain largely the same. It implied that one potential reason for the discrepancies may be that many genes are associated with outcomes in complex diseases, yet with a small individual contribution to marginal effect. Thus, for small sample sizes, many substantial but weak signals

may be missed (MOOTHA *et al.* 2003). It has also been revealed that the genetic nature of cancer is pathway-based, that is, oncogenes can be grouped into pathways based on biological functions such as cell survival, proliferation and metastatic dissemination (BARILLOT 2013; HUANG *et al.* 2014). Therefore, by combining weak signals from a number of genes within each pathway could increase power in prediction and prognosis methods.

Various studies have thus focused on developing pathway-based methods for cancer prognosis prediction (JONES 2008; JONES *et al.* 2008; LEE *et al.* 2008; REYAL *et al.* 2008; ABRAHAM *et al.* 2010; TESCHENDORFF *et al.* 2010; ENG *et al.* 2013; HUANG *et al.* 2014). These methods can be divided into two major categories. The first category focused on employing sophisticated statistical methods for variable selection with grouped predictors or pathways such as group lasso with an "all-in-all-out" idea, meaning that when one predictor in a group is chosen, then all variables in that group are chosen (PARK *et al.* 2007; WEI and LI 2007; JONES 2008). The second category, on the other hand, reduces the data dimension by generating pathway risk scores to be used in downstream data analysis. Abraham *et al.* (2010) adopted a gene set statistic to provide stability of prognostic signatures instead of individual genes. Huang *et al.* (2014) converted the gene matrix to a pathway matrix through "principal curve", similar to principal components analysis. Both of these two methods did not incorporate outcome when generating the pathways scores from the individual genes. Eng *et al.* (2013) proposed a method to reduce the computational complexity by incorporating a binary outcome to stand for decreased or increased risk score in each pathway which inferred potentially loss of information.

In this article, we propose a two-stage procedure to incorporate pathway information into the prognostic models using large-scale gene expression data. In the first stage, we calculate the leave-one-out cross-validated (LOOCV) prognostic score as risk score for each pathway by

fitting all predictors within each pathway using penalized Cox model and Bayesian hierarchical Cox model. In the second stage, we combine the risk scores of all pathways obtained in the first stage as new predictors to build a super prediction model. We used the proposed method in both breast cancer and ovarian cancer projects from The Cancer Genome Atlas (TCGA) to predict overall survival using gene expression profiling.

Breast cancer is the second most commonly diagnosed malignancy after skin cancer in women (HUANG *et al.* 2014). It is widely understood that breast cancer can be categorized into four clinical subtypes: Luminal A, Luminal B, Triple Negative/Basal like and Her2 and the survival/metastasis outcomes differ significantly among these four subtypes (CAREY *et al.* 2006; O'BRIEN *et al.* 2010; HAQUE *et al.* 2012). However, it is increasingly being realized that using only the clinical subtypes cannot fully discriminate breast cancers patients, and that better prediction of prognosis is needed. Ovarian cancer is the leading cause of death from gynecologic malignancies in the western world. According to statistics from American cancer society, the 5-year survival rate is 46.2% for ovarian cancer. However, the survival rates are discrepant among different cancer stages. The majority of patients are diagnosed with advanced stages, requiring a high toxicity six to eight cycles of platinum-based chemotherapy (PIGNATA *et al.* 2011). To this end, because of inaccurate prediction of response with clinical measures, approximately 30% of patients will be identified as chemo-resistance after undergoing multiple cycles of therapy with little or no benefit. Thus better prediction of response and prognosis is under great demand (BARAKAT *et al.* 2009). By applying our method to these two gene expression datasets from TCGA, we not only improve survival prediction compared with the alternative gene-based model that ignores the pathway information, but also identify significant biological relevant pathways.

MATERIALS AND METHODS

Data Collection

We downloaded two datasets with clinical information, microarray mRNA gene expression for breast cancer and ovarian cancer projects from TCGA using *R* package *TCGA-Assembler* (ZHU *et al.* 2014). Overall survival (OS) is the outcome of interest for both datasets. We downloaded and analyzed the processed level 3 (log2 lowess normalized (cy5/cy3) collapsed by gene symbol) gene expression data for both datasets. It represents up-regulation or down-regulation of a gene compared relative to the reference population (ZHAO *et al.* 2014).

Data Preprocessing and Pathway Analysis

For both datasets, samples with missing or zero overall survival time were removed. As the ratios of missing gene expression data were low, simple imputation with mean values across samples were adopted. To construct the pathways, we used genome annotation tools, KEGG (KANEHISA and GOTO 2000), to map genes to pathways. We first mapped gene symbols to Entrez ids with *R/bioconductor* package *AnnotationDbi*, and then mapped all the probes to KEGG pathways using the Bioinformatics tool *DAVID* (HUANG DA et al. 2009b; HUANG DA et al. 2009a).

Cox proportional hazards models

Cox proportional hazards model is the commonly used method for analyzing censored survival data (VAN HOUWELINGGEN and PUTTER 2012), for which the hazard function of survival time takes the form:

$$h(t \mid X) = h_0(t)\exp(X\beta) \tag{1}$$

where $h_0(t)$ is the baseline hazard function, X and β are the vectors of predictors and coefficients, respectively, and $X\beta$ is the linear predictor or called the prognostic index. The coefficients β are estimated by maximizing the partial log-likelihood:

$$pl(\beta) = \sum_{i=1}^{n} d_i \log\left(\frac{\exp(X_i\beta)}{\sum_{j \in R(t_i)} \exp(X_j\beta)} \right) \tag{2}$$

where the censoring indicator d_i takes 1 if the observed survival time t_i for individual i is uncensored and 0 if it is censored, and $R(t_i)$ is the risk set at time t_i. For molecular data, the number of coefficients is much larger than the number of individuals, and/or covariates are usually highly correlated, for which Cox regression is not directly applicable.

Penalized Cox Model (ridge, lasso and elastic-net Cox models). The elastic net is a widely used penalization approach to deal with high-dimensional models, which adds the elastic-net penalty to the log-likelihood function and estimates the parameters β by maximizing the penalized log-likelihood(ZOU and HASTIE 2005; HASTIE *et al.* 2009; FRIEDMAN *et al.* 2010; SIMON *et al.* 2011; HASTIE *et al.* 2015). For the Cox model described above, we estimate the parameters β by maximizing the penalized partial log-likelihood:

$$ppl_{\alpha}(\beta) = pl(\beta) - \lambda n \sum_{j=1}^{J} [\alpha \mid \beta_j \mid + (1-\alpha)\frac{1}{2}\beta_j^2] \tag{3}$$

where α ($0 \le \alpha \le 1$) is a predetermined elastic-net parameter, λ ($\lambda \ge 0$) is a penalty parameter, and $pl(\beta)$ is the partial log-likelihood of the Cox model. The penalty parameter λ controls the overall strength of penalty and the size of the coefficients; for a small λ, many coefficients can

be large, and for a large λ, many coefficients will be shrunk towards zero. The elastic net includes the lasso ($\alpha = 1$) and ridge Cox regression ($\alpha = 0$) as special cases (TIBSHIRANI 1997; GUI and LI 2005; VAN HOUWELINGEN *et al.* 2006; SIMON *et al.* 2011; VAN HOUWELINGGEN and PUTTER 2012).

The ridge, lasso and elastic net Cox models can be fitted by the cyclic coordinate descent algorithm, which successively optimizes the penalized log-likelihood over each parameter with others fixed and cycles repeatedly until convergence. The cyclic coordinate descent algorithm has been implemented in the *R* package *glmnet*. Cross-validation is the most widely used method to select an optimal value λ (e.g., an optimal Cox model) that gives minimum cross-validated error.

Bayesian hierarchical Cox model. Hierarchical modelling is another efficient approach to handling high-dimensional data, where the regression coefficients are themselves modeled (GELMAN and HILL 2007; GELMAN *et al.* 2014). Hierarchical models are more easily interpreted and handled in the Bayesian framework, where the distribution of the coefficient is the prior distribution and statistical inference is based on the posterior estimation. The commonly used prior is the double-exponential (or Laplace) prior distribution (PARK and CASELLA 2008; YI and XU 2008; YI and MA 2012):

$$\beta_j \sim DE\left(\beta_j \mid 0, s\right) = \frac{1}{2s}\exp\left(-\frac{\left|\beta_j\right|}{s}\right) \qquad (4)$$

where the inverse scale *s* is shrinkage parameter and controls the amount of shrinkage; a smaller scale *s* induces stronger shrinkage and thus forces the estimates of β_j towards the prior mean

zero. For the hierarchical Cox model with the double-exponential prior, the log posterior distribution of the parameters can be expressed as

$$\log p(\beta \mid t,d) \propto pl(\beta) - \frac{1}{s}\sum_{j=1}^{J}\left|\beta_j\right| \tag{5}$$

By maximizing the log posterior distribution, we estimate the parameters by finding the posterior modes of the parameters. We have developed an algorithm for fitting the hierarchical Cox model by incorporating an EM procedure into the usual Newton-Raphson algorithm for fitting classical Cox models. Our algorithm has been implemented in our *R* package *BhGLM* (http://www.ssg.uab.edu/bhglm/). The above hierarchical Cox model with $s = 1/(n\lambda)$ is equivalent to the lasso, if one estimates the mode of the posterior distribution.

Optimizing the penalty parameter λ and the inverse scale *s*: The penalized Cox models heavily depend on the penalty tuning parameter λ. The tuning parameter is usually chosen using *K*-fold cross-validation procedure and estimated by maximizing the cross-validated partial likelihood (CVPL) (VAN HOUWELINGEN *et al.* 2006; SIMON *et al.* 2011):

$$CV(\lambda) = \sum_{k=1}^{K}[pl\left(\hat{\beta}_{(-k)}\right) - pl_{(-k)}\left(\hat{\beta}_{(-k)}\right)] \tag{6}$$

where $\hat{\beta}_{(-k)}$ is the estimate of β from all the data except the *k*-th part of the data in *K*-fold cross-validation, $pl\left(\hat{\beta}_{(-k)}\right)$ is the partial likelihood of all the data points and $pl_{(-k)}\left(\hat{\beta}_{(-k)}\right)$ is the partial likelihood excluding the *k*-th part of the data. By subtracting the log-partial likelihood evaluated on the non-left out data from that evaluated on the full data, we can make efficient use of the death times of the left out data in relation to the death times of all the data. We choose the λ

value which maximizes $CV(\lambda)$. Based on the relationship between the lasso and the hierarchical Cox model, the inverse scale *s* in the hierarchical Cox model is chosen to be $s = 1/(n\lambda)$.

Two-Stage Approach for Pathway Integration

Intuitively we can simultaneously analyze all the genes in a single model. In some previous studies, penalized Cox models have been used to analyze molecular data with this gene-based single model approach (RAPPAPORT 2007; BOVELSTAD *et al.* 2009; JACOB 2009; ZHANG *et al.* 2013; YUAN *et al.* 2014; ZHAO *et al.* 2014). However, due to high-dimension of molecular data, fitting a model including all genes can lead to instability of predictive model and may result in decreased prediction performance with the increased model complexity.

To overcome these problems, we propose an two-stage procedure to build a prediction model, inspired by the super learner of van der Laan et al. (2007) (VAN DER LAAN *et al.* 2007; VAN HOUWELINGEN and PUTTER 2011). Suppose that genes are assigned into *G* pathways, $G_g : g = 1,\ldots,G,$ with the *g-th* pathway G_g containing J_g genes, and denote the vector of predictors in the *g-th* pathway by X^g. Overlapping is common in pathways analysis, that is, a gene could belong to multiple functional pathways. The two-stage procedure can easily deal with overlapping.

In the first stage, we separately analyze each pathway by fitting all the predictors within each pathway using the hierarchical Cox model or penalized Cox regression. For the *g-th* pathway, we fit the model: $h(t \mid X_i^g) = h_0(t)\exp(X_i^g \beta^g)$ and can obtain the estimate of prognostic score, $\eta_i^g = X_i^g \hat{\beta}^g$, for each individual. However, directly using the prognostic score in the second stage can result in over-fitting. To prevent over-fitting, we estimate cross-validated

prognostic scores using leave-one-out cross-validation (LOOCV). The LOOCV prognostic score for the *i-th* individual and the g-*th* pathway is calculated as follows; we first estimate the coefficients β^g of the predictors X^g using all other (n – 1) individuals (i.e., excluding the *i-th* individual). Denote the estimate by $\hat{\beta}^{g(-i)}$, then the LOOCV prognostic score is

$$\eta^g_{(CV,i)} = X_i^{\,g}\hat{\beta}^{g(-i)} \tag{7}$$

The purpose of using LOOCV prognostic scores instead of prognostic scores directly is to prevent overfitting in the super prediction model in the second stage. The LOOCV prognostic score for each patient is derived independently of the observed data of the patient, and hence the scores $\eta^g_{(CV,i)}$ along with the observed data of the patients can essentially be treated as a 'new dataset'. Therefore, this procedure can overcome overfitting (TIBSHIRANI and EFRON 2002; HASTIE *et al.* 2015).

In the second stage, we combine the LOOCV prognostic scores of pathways with cross-validated C-index greater than 0.5 obtained in the first stage as new predictors to build a super prediction model:

$$h(t \mid \eta_{(CV,i)}) = h_0(t)\exp\left(\sum_{g=1}^{G}\eta^g_{(CV,i)}\alpha^g\right) \tag{8}$$

To prevent the super prediction model from being non-identifiable and over-fitting, we utilize the hierarchical Cox model approach, i.e., assuming that the pathway effects, α^g, follow the double-exponential prior as described in (4). The hierarchical Cox model with the EM algorithm can produce interval estimates and p-values for the pathway effects, α^g, and thus allow us to detect significant pathways. To evaluate the prognostic performance of the super prediction model, we carry out 10-fold cross-validation over 10 repeats. The two-stage procedure to build the pathway-

structured predictive model is presented in **Figure 1**. The R source code for implementing the two-stage approach is provided in File S1.

(Insert **Figure 1** here)

Evaluating the predictive performance

To assess the prognostic utility of the fitted model, we need to evaluate the quality of the fitted model and its predictive value. There are several ways to measure the performance of a Cox model (STEYERBERG 2009; VAN HOUWELINGGEN and PUTTER 2012): 1) **Concordance Index (C-index)** (HARRELL *et al.* 1996): the standard way to measure the concordance between the observed survival times and predicted survival times. The performance is better when the C-index is greater; 2) **CVPL:** a general measurement of model quality and prediction as mentioned above as $CV(\lambda)$; 3) **Prediction Error**: the most popular measure of prediction error is the Brier score, which is defined as (VAN HOUWELINGGEN and PUTTER 2012): $Brier(y, S(t_0 \mid x)) = (y - S(t_0 \mid x))^2$, where $S(t_0 \mid x)$ is the estimated survival probability of an individual beyond t_0 given the predictor x; 4) **Pre-validated Kaplan Meier Analysis**: We compute the cross-validated prognostic score for each patient for the super prediction model in the second stage. We then divide the individuals into two groups of high and low risks based on the median of the cross-validated prognostic scores. The samples with higher scores are in high risk group, while the samples with lower scores are in low risk group. We then get the Kaplan Meier plot and log-rank test by comparing the two groups (TIBSHIRANI and EFRON 2002; HASTIE *et al.* 2015). This provides a valid assessment of the predictive performance of the model (TIBSHIRANI and EFRON 2002; HASTIE *et al.* 2015).

Comparison to gene-based model

To demonstrate the advantages of our approach, we compared our two-stage pathway-structured predictive model to the previously used gene-based model. For the gene-based model, we used the penalized Cox models and the hierarchical Cox model to simultaneously fit all the genes that are mapped into pathways as one single model. 10-fold cross-validation over 10 repeats was used to evaluate the predictive performance of the gene-based model. Then we compared the prediction performances of the pathway-structured predictive model and the gene-based model.

SIMULATION STUDY

We carried out a simulation study to evaluate the false positive rate of the proposed two-stage approach. We simulated 10 groups (or called pathways) of 500 genes with 50 genes in each group. The sample size was set to be 500, and the correlation between genes within each group was 0.6. To assess the false positive rate under the null, the coefficient β_j for each gene was set to be 0. Following Simon et al. (2011), we generated "true" survival time T_i for each individual from the exponential distribution: $T_i \sim \exp\left(\sum_{j=1}^{m} x_{ij}\beta_j\right)$, and then generated censoring time C_i for each individual from the exponential distribution: $C_i \sim \exp(r_i)$, where r_i were randomly sampled from a standard normal distribution. The observed censored survival time t_i was set to be the minimum of the "true" survival and censoring times, $t_i = \min(T_i, C_i)$, and the censoring indicator d_i was set to be 1 if $C_i > T_i$ and 0 otherwise.

The simulation was repeated 1000 times to evaluate the performance of two-stage approach in controlling false positive rate. Our simulation results showed that the false positive rate of our two-stage approach was ~0.01 at the significance level of 0.05. This indicates that our two-stage approach well controls the false positive rate.

REAL DATA APPLICATION RESULTS

Data Summary

For breast cancer data, 54 samples were removed for missing or zero overall survival time. 17815 features across 533 samples were profiled for gene expression, which included a total of 1571 missing observations. For our analysis, only 505 samples were kept for whom survival time and gene expression were both available. Among these 505 patients, only 65 were dead and thus the event rate was 12.9%. For ovarian cancer data, the final sample size was 538 with 17814 features of gene expression profiled, which included a total of 746 missing observations. Among these 538 patients, 278 were dead and thus the event rate was 51.7%.

Simple imputation with mean values across samples was adopted to fill the missing gene expression values for both cancer datasets. In pathway analysis, we mapped gene symbols to Entrez ids with *R/Bioconductor* package *AnnotationDbi*.

For breast cancer, 3181 probes were mapped to 109 pathways. For ovarian cancer, 4887 probes were mapped to 193 pathways with at least five genes in each pathway. Therefore, we used the expression data of 3181 and 4887 genes to predict overall survival for breast and ovarian cancers respectively.

Building Pathway-structured Predictive Model with Two-Stage Approach

In the first stage, we calculated the LOOCV prognostic score for each pathway and each patient by fitting all the genes in that pathway. The procedure was then repeated for all the pathways. For breast cancer data, the predictive model was built with ridge, lasso and elastic-net Cox models and hierarchical Cox model. For ovarian cancer data, the predictive model was built with lasso Cox model and hierarchical Cox model. For each pathway, the tuning parameters λ in the penalized Cox models were estimated by 10-fold cross-validation over 10 repeats. For the inverse scale s in the hierarchical Cox model, we set $s = 1/(n\lambda)$ based on the relationship between lasso and hierarchical Cox model. For breast cancer data, we tested three inverse scales: $s = 1/(n\lambda)$, $1/(n\lambda) + 0.03$, 0.08. For ovarian cancer data, we only used $s = 1/(n\lambda)$ for the hierarchical Cox model.

In the second stage, we used pathways with cross-validated C-index greater than 0.5 obtained in the first stage to build a super prognostic model. A super prognostic model was built with the LOOCV prognostic scores obtained in the first stage as new predictors with the hierarchical Cox model approach. 10-fold cross-validation over 10 repeats was carried out to validate the super prognostic models. To select the prior scale for hierarchical Cox model, we calculated CVPL from 10-fold cross-validation for different prior scales, s = 0.08, 0.10, 0.12, 0.14, 0.16, 0.18, 0.20, and 0.22, and then the scale with highest CVPL was chosen.

For breast cancer, **Table 1** shows the numbers of pathways with C-index greater than 0.5, CVPL, and C-index for all the two-stage models. It can be seen that the hierarchical Cox models with $s = 1/(n\lambda)$ or $1/(n\lambda) + 0.03$ generated larger CVPL and C-index, and thus had better prediction performance than the other approaches. For ovarian cancer data, CVPL were -1652.928 (3.753) and -1650.489 (2.599), and C-index were 0.718 (0.003) and 0.722 (0.004), for

two-stage hierarchical-hierarchical Cox model with $s = 1/(n\lambda)$ and two-stage lasso-hierarchical Cox model, respectively.

(Insert **Table 1** here)

Pathway-structured Predictive Model Superior to Gene-based Model in Prediction Performance

To compare the two-stage approach with the single model analysis, we simultaneously fit all the genes that were mapped into pathways using the lasso and the hierarchical Cox model. CVPL and C-index for the joint analyses are presented in **Table 2**. For breast cancer data, CVPL were -364.845 (0.949) and -363.554 (0.626), and C-index were 0.507 (0.023) and 0.572 (0.023), for the joint lasso and joint hierarchical Cox models, respectively. For ovarian cancer data, CVPL were -1716.385 (1.565) and -1714.236 (0.631), and C-index were 0.541 (0.011) and 0.573 (0.006), for the joint lasso and joint hierarchical Cox models, respectively. Our results of the joint analyses were similar to those of Zhao et al. (2014) and Yuan et al. (2014). Therefore, the two-stage approach provided much larger CVPL and lower C-index than the gene-based models, and thus significantly outperformed the gene-based models for both datasets.

(Insert **Table 2** here)

The predictive performances of all the models were also assessed by the prediction error. **Figure 2 and 3** present the Brier prediction error curves for breast and ovarian cancer data, respectively. For breast cancer data, the pathway-structured predictive models consistently performed better than the joint lasso. Among all the pathway-structured predictive models, the two-stage ridge-hierarchical Cox model had the highest prediction error. The two-stage hierarchical-hierarchical Cox models with $s = 1/(n\lambda)$ or $1/(n\lambda) + 0.03$ and the two-stage lasso-

hierarchical Cox model had lower prediction error than the best fitted pathway. For ovarian cancer data, all the pathway-structured predictive models consistently performed better than the joint lasso and the best fitted pathway.

(Insert **Figure 2** and **Figure 3** here)

Risk Group Stratification

We also demonstrated that our pathway-structured predictive models were superior to the gene-based models by evaluating their ability in stratifying samples. For both breast cancer and ovarian cancer datasets, the Kaplan-Meier curves were drawn for the low risk and high risk groups to compare among the joint lasso, joint hierarchical Cox model, two-stage lasso-hierarchical (L-H) model and two-stage hierarchical-hierarchical (H-H) model. Log-rank tests were carried out for each model. The Kaplan-Meier curves and p-values of log-rank tests are shown in **Figure 4 and 5**, for breast and ovarian cancer datasets, respectively. For both breast and ovarian cancer datasets, our pathway-structured predictive models generated significantly larger differences in Kaplan Meier curves between low-risk group and high-risk group (p-value = 4.69e-11, p-value = 5.81e-11, p-value<0.001, p-value<0.001, respectively). Both gene-based models did not show significant differences of Kaplan-Meier curves between two groups (p-value = 0.242, p-value = 0.231, p-value = 0.0504, p-value = 0.0539, respectively).

(Insert **Figure 4** and **Figure 5** here)

Biological Relevance of Associated Pathways

We evaluated the ability of the pathway-structured predictive models to identify prognostic cancer relevant pathways by comparing with previously published results. **Figure 6**

and 7 show the estimated effects and p-values of the identified significant pathways for the two-stage lasso-hierarchical Cox model and the two-stage hierarchical-hierarchical Cox model with $s = 1/(n\lambda)$ for breast and ovarian cancer datasets, respectively. For breast cancer, we compared all the detected significant pathways with fifteen core cancer pathways discussed in Eng *et al.* (2013) and found that five out of fifteen were consistent for the two-stage hierarchical-hierarchical Cox model. Besides that, metabolic pathways such as *Pyrimidine metabolism*, *Glycerolipid metabolism* and *Metabolism of xenobiotics by cytochrome P450* have been identified in both two-stage Lasso-hierarchical Cox model and two-stage hierarchical-hierarchical Cox model. They have been reported as important pathways in regulating breast cancer in previous literature (MURRAY *et al.* 1993; SCHRAMM *et al.* 2010; MERDAD *et al.* 2015). Moreover, *Ribosome* and *Proteasome* are respective functional as protein builders and degraders which also play an essential role in breast cancer. For ovarian cancer, several significant pathways are as well consistent with those fifteen core cancer pathways discussed in Eng *et al.* (2013). Besides that, metabolic pathways such as *Nicotinate and nicotinamide metabolism* and *fatty acid metabolism* have been identified in the two-stage hierarchical-hierarchical Cox model. They have been discussed as important pathways in regulating ovarian cancer in previous literature (TANIA *et al.* 2010; VERMEERSCH *et al.* 2014).

(Insert **Figure 6** and **Figure 7** here)

Incorporating Clinical Factors with Pathway Information to Predict Cancer Survival

The most widely used method for integrating clinical factors and gene expression data is to include all the predictors in a single model. However, previous studies rarely found meaningful improvement in terms of prediction performance compared to models with only traditional clinical factors or a single type of molecular data. There are many drawbacks with

these previous methods: 1) they create a huge model with too many predictors which cannot be easily handled and thus can reduce the prediction accuracy, 2) models including different types of predictors can result in different coefficient estimates compared with the models with only clinical or molecular predictors, and 3) the predictive values of clinical and molecular factors obtained from such models cannot be easily interpreted, and cannot easily be applied to other data sets.

To overcome the limitations stated above, we incoporated clinical factors with pathway predictors in the second stage of the two-stage approach. Comparing to the methods incoporating clinical factors and molecular data directly, this approach reduces the unstablility of the model drastically. Besides, as we could apply different prior scales for hierarchical cox regression to clinical factors and pathways, the contribution of clincial factors in prediction will be preserved in our predictive model. We applied this approach in both two TCGA breast anc ovarian cancer datasets to incorporate age of patients at diagnosis and tumor stages with pathway matrix information to predict cancer survival. Our results showed that the prediction performance can be improved after combined clincial factors with pathway information. The results are presented in details in File S2.

DISCUSSION

The heterogeneity of prognostic prediction in cancers has been a persisted problem for decades (BARILLOT 2013). It is now realized that cancer is a disease of genome and can be understood by identifying the abnormal genes and proteins that are associated with the risk of developing cancer. Some statistical methods have been used to analyze genomic data with gene-

based approach to search for gene signature and to predict prognosis (RAPPAPORT 2007; BOVELSTAD *et al.* 2009; JACOB 2009; ZHANG *et al.* 2013; YUAN *et al.* 2014; ZHAO *et al.* 2014). Due to the complicated genetic nature of cancer and potentially underpowered statistical analysis of gene-based approach, it has been suggested that the complexity of cancer should be handled based on pathway-centric instead of gene-centric perspectives (JONES 2008). Oncogenes and tumor suppressor genes have been well studied and can be arranged into signaling pathways according to their biological functions such as cell survival, proliferation and metastatic dissemination. Other studies have investigated methods to analyze high-throughput cancer genomics data based on functional units, i.e. pathways (GOEMAN and BUHLMANN 2007; LEE *et al.* 2008; REYAL *et al.* 2008; ABRAHAM *et al.* 2010; TESCHENDORFF *et al.* 2010).

Our two-stage approach is developed to incorporate the functional structure of pathways to predict survival for cancer patients. The proposed method has two outstanding features in reducing the large-scale molecular matrix to a predictable pathway-based matrix: (i) it incorporates the correlation with survival information in calculating risk scores for each pathway; (ii) we use LOOCV prognostic score as risk score, which not only prevents overfitting and can be easily carried out, but also gives an unbiased view on the contribution of the different information from pathways to the super prediction model. Meanwhile, pathway-structured predictive models perform consistently better than the gene-based models using lasso or hierarchical Cox model in terms of C-index, CVPL and reduced prediction error in two cancer datasets. Our pathway-structured predictive models also show remarkable performance in discriminating the prognostic effects between different patients compared with gene-based methods based on the Kaplan-Meier analysis results. It is noteworthy that different penalized Cox models and hierarchical Cox models with double exponential prior have been compared in

generating pathway-based matrix, which suggests that lasso and hierarchical Cox model are more stable in preserving the pathway information for super prediction model.

Our approach is meanwhile capable to identify important pathways. Among those pathways identified for both breast and ovarian cancers, we have identified in total seven out of fifteen core cancer pathways. *Mitogen-activated protein kinase (MAPK)* pathways are important in controlling fundamental cellular processes, i.e. growth, proliferation, differentiation, migration and apoptosis (DHILLON *et al.* 2007). When abnormally activated, *MAPK* pathways can lead to the progression of cancer (USSAR and VOSS 2004; MCCUBREY *et al.* 2007). Another pathway, *the mammalian target of rapamycin (mTOR),* also plays an essential role in the regulation of cell proliferation, growth, differentiation, migration and survival. Similarly to *MAPK* pathways, the dysregulation of *mTOR signaling* happens in various human tumors, resulting in higher susceptibility to inhibitors of *mTOR* (HUANG and HOUGHTON 2003). The *Hedgehog* pathway regulates many fundamental processes including stem cell maintenance, cell differentiation, tissue polarity and cell proliferation. It has been demonstrated that inappropriate activation of *Hedgehog* pathway occurs in various cancers such as brain, gastrointestinal, lung, breast and prostate cancer (GUPTA *et al.* 2010). The *JAK-STAT* pathway is also identified by our approach. This pathway regulates in various cellular processes such as stem cell maintenance, apoptosis and the inflammatory response and was found frequently dysregulated in diverse types of cancer (THOMAS *et al.* 2015). Furthermore, *cell cycle* pathway and *Cell adhesion molecules* pathway have also been reported to play an essential role in cancer progression and the potential to find cancer therapy (OKEGAWA *et al.* 2004).

However, there are some issues that need to be addressed in the future based on our results. We implemented our approach in two independent microarray gene expression data from

TCGA breast and ovarian cancer projects. There are data from other platforms, such as RNA-Seq data. Our model can be directly adopted in RNA-Seq data by applying additional constraints in normalization of the RNA-Seq data. The second issue is the potential loss of gene information, as only 18-30% genes are mapped into pathways in both two datasets. One plausible solution is to calculate a risk score for the unmapped genes as one additional group, yet requiring pre-filtering with univariate analysis approach or variance-filter to reduce the number of unmapped genes to a certain feasible level. Meanwhile, our method can easily incorporate clinical factors with the pathway information, yielding great potential for application in clinical research. In the future, we will also apply our approach in other levels of genomic data, e.g. DNA methylation, miRNA and copy number alterations, for more than 30 types of cancer, with a combination with clinical biomarkers into our pathway-structured predictive model to better predict cancer survival in clinical research.

Key Points

- Current development of molecular signatures to predict for breast, colon and prostate cancers are notable but may not be sufficient to achieve the goal of precision medicine. It has been revealed that the genetic nature of cancer is pathway-based, that is, oncogenes can be grouped into pathways based on biological functions.
- However, current methods have shortcomings in incorporating pathway information into predictive modeling. We proposed a two-stage procedure to incorporate pathway information into the predictive modeling using large-scale gene expression data and applied the proposed

method to analyze two independent breast and ovarian cancer datasets from The Cancer Genome Atlas (TCGA) project for predicting overall survival.

- The results show that the proposed approach not only improves survival prediction compared with the alternative gene-based methods that ignore the pathway information, but also identifies significant biological pathways.
- The approach can be extended to data from other platforms such as RNA-Seq data or other molecular level of data including DNA methylation, miRNA and copy number alterations, for various types of cancer.

Acknowledgements

We thank two reviewers and the associate editor for their constructive suggestions and comments that have improved the manuscript. This work was supported in part by the research grants: NIH R01GM069430, NIH U01CA158428, NIH R03DE024198, NIH R03DE025646, R01CA133093, R01ES016354, the Alabama Innovation Fund, Avon Foundation grant 02-2014-030, and V Foundation Scholar Award V2015-009.

REFERENCES

Abraham, G., A. Kowalczyk, S. Loi, I. Haviv and J. Zobel, 2010 Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context. BMC Bioinformatics 11**:** 277.

Barakat, R. R., M. Markman and M. Randall, 2009 *Principles and practice of gynecologic oncology*. Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia.

Barillot, E., 2013 *Computational systems biology of cancer*. CRC Press, Boca Raton, FL.

Bovelstad, H. M., S. Nygard and O. Borgan, 2009 Survival prediction from clinico-genomic models--a comparative study. BMC Bioinformatics 10**:** 413.

Carey, L. A., C. M. Perou, C. A. Livasy, L. G. Dressler, D. Cowan *et al.*, 2006 Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. JAMA 295**:** 2492-2502.

Collins, F. S., and H. Varmus, 2015 A new initiative on precision medicine. N Engl J Med 372**:** 793-795.

Dhillon, A. S., S. Hagan, O. Rath and W. Kolch, 2007 MAP kinase signalling pathways in cancer. Oncogene 26**:** 3279-3290.

Edwards, B. K., A. M. Noone, A. B. Mariotto, E. P. Simard, F. P. Boscoe *et al.*, 2014 Annual Report to the Nation on the status of cancer, 1975-2010, featuring prevalence of comorbidity and impact on survival among persons with lung, colorectal, breast, or prostate cancer. Cancer 120**:** 1290-1314.

Eng, K. H., S. Wang, W. H. Bradley, J. S. Rader and C. Kendziorski, 2013 Pathway index models for construction of patient-specific risk profiles. Stat Med 32**:** 1524-1535.

Friedman, J., T. Hastie and R. Tibshirani, 2010 Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33**:** 1-22.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin, 2014 *Bayesian data analysis*. Taylor & Francis.

Gelman, A., and J. Hill, 2007 Data analysis using regression and hierarchical/multilevel models, pp. Cambridge University Press: Cambridge, UK.

Goeman, J. J., and P. Buhlmann, 2007 Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 23**:** 980-987.

Gui, J., and H. Li, 2005 Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics 21**:** 3001-3008.

Gupta, S., N. Takebe and P. Lorusso, 2010 Targeting the Hedgehog pathway in cancer. Ther Adv Med Oncol 2**:** 237-250.

Haque, R., S. A. Ahmed, G. Inzhakova, J. Shi, C. Avila *et al.*, 2012 Impact of breast cancer subtypes and treatment on survival: an analysis spanning two decades. Cancer Epidemiol Biomarkers Prev 21**:** 1848-1855.

Harrell, F. E., Jr., K. L. Lee and D. B. Mark, 1996 Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15**:** 361-387.

Hastie, T., R. Tibshirani and J. Friedman, 2009 *The Elements of Statistical Learning*. Springer-Verlag, New York, NY, USA.

Hastie, T., R. Tibshirani and M. Wainwright, 2015 *Statistical Learning with Sparsity - The Lasso and Generalization*. CRC Press, New York.

Huang da, W., B. T. Sherman and R. A. Lempicki, 2009a Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37**:** 1-13.

Huang da, W., B. T. Sherman and R. A. Lempicki, 2009b Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4**:** 44-57.

Huang, S., and P. J. Houghton, 2003 Targeting mTOR signaling for cancer therapy. Curr Opin Pharmacol 3**:** 371-377.

Huang, S., C. Yee, T. Ching, H. Yu and L. X. Garmire, 2014 A novel model to combine clinical and pathway-based transcriptomic information for the prognosis prediction of breast cancer. PLoS Comput Biol 10**:** e1003851.

Jacob, L. e. a., 2009 Group lasso with overlap and graph lasso.

Jones, D., 2008 Pathways to cancer therapy. Nat Rev Drug Discov 7**:** 875-876.

Jones, S., X. Zhang, D. W. Parsons, J. C. Lin, R. J. Leary *et al.*, 2008 Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 321**:** 1801-1806.

Kanehisa, M., and S. Goto, 2000 KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28**:** 27-30.

Lee, E., H. Y. Chuang, J. W. Kim, T. Ideker and D. Lee, 2008 Inferring pathway activity toward precise disease classification. PLoS Comput Biol 4**:** e1000217.

McCubrey, J. A., L. S. Steelman, W. H. Chappell, S. L. Abrams, E. W. Wong *et al.*, 2007 Roles of the Raf/MEK/ERK pathway in cell growth, malignant transformation and drug resistance. Biochim Biophys Acta 1773**:** 1263-1284.

Merdad, A., S. Karim, H. J. Schulten, M. Jayapal, A. Dallol *et al.*, 2015 Transcriptomics profiling study of breast cancer from Kingdom of Saudi Arabia revealed altered expression of Adiponectin and Fatty Acid Binding Protein4: Is lipid metabolism associated with breast cancer? BMC Genomics 16 Suppl 1**:** S11.

Mook, S., L. J. Van't Veer, E. J. Rutgers, M. J. Piccart-Gebhart and F. Cardoso, 2007 Individualization of therapy using Mammaprint: from development to the MINDACT Trial. Cancer Genomics Proteomics 4**:** 147-155.

Mootha, V. K., C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag *et al.*, 2003 PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34**:** 267-273.

Murray, G. I., R. J. Weaver, P. J. Paterson, S. W. Ewen, W. T. Melvin *et al.*, 1993 Expression of xenobiotic metabolizing enzymes in breast cancer. J Pathol 169**:** 347-353.

O'Brien, K. M., S. R. Cole, C. K. Tse, C. M. Perou, L. A. Carey *et al.*, 2010 Intrinsic breast tumor subtypes, race, and long-term survival in the Carolina Breast Cancer Study. Clin Cancer Res 16**:** 6100-6110.

Okegawa, T., R. C. Pong, Y. Li and J. T. Hsieh, 2004 The role of cell adhesion molecule in cancer progression and its application in cancer therapy. Acta Biochim Pol 51**:** 445-457.

Park, M. Y., T. Hastie and R. Tibshirani, 2007 Averaged gene expressions for regression. Biostatistics 8**:** 212-227.

Park, T., and G. Casella, 2008 The Bayesian Lasso. Journal of the American Statistical Association 103**:** 681-686.

Pignata, S., L. Cannella, D. Leopardo, C. Pisano, G. S. Bruni *et al.*, 2011 Chemotherapy in epithelial ovarian cancer. Cancer Lett 303**:** 73-83.

Pohl, A., and H. J. Lenz, 2008 Individualization of therapy for colorectal cancer based on clinical and molecular parameters. Gastrointest Cancer Res 2**:** S38-41.

Rappaport, F. e. a., 2007 Classification of microarray data using gene networks. BMC Bioinformatics 8.

Reyal, F., M. H. van Vliet, N. J. Armstrong, H. M. Horlings, K. E. de Visser *et al.*, 2008 A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the proliferation, immune response and RNA splicing modules in breast cancer. Breast Cancer Res 10**:** R93.

Schramm, G., E. M. Surmann, S. Wiesberg, M. Oswald, G. Reinelt *et al.*, 2010 Analyzing the regulation of metabolic pathways in human breast cancer. BMC Med Genomics 3**:** 39.

Simon, N., J. Friedman, T. Hastie and R. Tibshirani, 2011 Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. Journal of Statistical Software 39**:** 1-13.

Sotiriou, C., and M. J. Piccart, 2007 Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? Nat Rev Cancer 7**:** 545-553.

Steyerberg, E. W., 2009 *Clinical Prediction Models: A Practical Approch to Development, Validation, and Updates.* Springer, New York.

Tania, M., M. A. Khan and Y. Song, 2010 Association of lipid metabolism with ovarian cancer. Curr Oncol 17**:** 6-11.

Teschendorff, A. E., S. Gomez, A. Arenas, D. El-Ashry, M. Schmidt *et al.*, 2010 Improved prognostic classification of breast cancer defined by antagonistic activation patterns of immune response pathway modules. BMC Cancer 10**:** 604.

Thomas, S. J., J. A. Snowden, M. P. Zeidler and S. J. Danson, 2015 The role of JAK/STAT signalling in the pathogenesis, prognosis and treatment of solid tumours. Br J Cancer 113**:** 365-371.

Tibshirani, R., 1997 The lasso method for variable selection in the Cox model. Stat Med 16**:** 385-395.

Tibshirani, R. J., and B. Efron, 2002 Pre-validation and inference in microarrays. Stat Appl Genet Mol Biol 1**:** Article1.

Ussar, S., and T. Voss, 2004 MEK1 and MEK2, different regulators of the G1/S transition. J Biol Chem 279**:** 43861-43869.

van de Vijver, M. J., Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart *et al.*, 2002 A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347**:** 1999-2009.

van der Laan, M. J., E. C. Polley and A. E. Hubbard, 2007 Super learner. Statistical applications in genetics and molecular biology 6.

van Houwelingen, H., and H. Putter, 2011 *Dynamic prediction in clinical survival analysis*. CRC Press.

van Houwelingen, H. C., T. Bruinsma, A. A. Hart, L. J. Van't Veer and L. F. Wessels, 2006 Cross-validated Cox regression on microarray gene expression data. Stat Med 25**:** 3201-3216.

van Houwelinggen, H. G., and H. Putter, 2012 *Dynamic Prediction in Clinical Survival Analysis*. CRC Press.

Vermeersch, K. A., L. Wang, J. F. McDonald and M. P. Styczynski, 2014 Distinct metabolic responses of an ovarian cancer stem cell line. BMC Syst Biol 8**:** 134.

Wang, Y., J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look *et al.*, 2005 Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365**:** 671-679.

Wei, Z., and H. Li, 2007 Nonparametric pathway-based regression models for analysis of genomic data. Biostatistics 8**:** 265-284.

Yi, N., and S. Ma, 2012 Hierarchical Shrinkage Priors and Model Fitting for High-dimensional Generalized Linear Models. Stat Appl Genet Mol Biol.

Yi, N., and S. Xu, 2008 Bayesian LASSO for quantitative trait loci mapping. Genetics 179**:** 1045-1055.

Yuan, Y., E. M. Van Allen, L. Omberg, N. Wagle, A. Amin-Mansour *et al.*, 2014 Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nat Biotechnol 32**:** 644-652.

Zhang, W., T. Ota, V. Shridhar, J. Chien, B. Wu *et al.*, 2013 Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. PLoS Comput Biol 9**:** e1002975.

Zhao, Q., X. Shi, Y. Xie, J. Huang, B. Shia *et al.*, 2014 Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. Brief Bioinform 16**:** 291-303.

Zhu, Y., P. Qiu and Y. Ji, 2014 TCGA-assembler: open-source software for retrieving and processing TCGA data. Nat Methods 11**:** 599-600.

Zou, H., and T. Hastie, 2005 Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 67**:** 301–320.

Table 1. Prediction Performance Comparison between Different Two-stage Models for Breast Cancer Data

Model	Number of Pathways	CVPL	C-index
Lasso-Hierarchical Cox Model	75	-340.058 (4.215)	0.725 (0.014)
Ridge-Hierarchical Cox Model	88	-353.444 (2.373)	0.640 (0.021)
Elastic Net-Hierarchical Cox Model (α=0.5)	74	-340.893 (2.435)	0.711 (0.012)
Hierarchical-Hierarchical Cox Model $s = 1/(n\lambda)$	77	-337.170 (7.187)	0.760 (0.015)
Hierarchical-Hierarchical Cox Model $s = 1/(n\lambda) + 0.03$	77	-333.358 (1.971)	0.748 (0.013)
Hierarchical-Hierarchical Cox Model $s = 0.08$	70	-347.459 (2.796)	0.692 (0.014)

Table 2. Prediction performance comparison between two gene-based models

Model	Joint Lasso	Joint Hierarchical Cox Model
Breast Cancer Data		
Number of Genes	3181	3181
CVPL	-364.845 (0.949)	-363.554 (0.626)
C-index	0.507 (0.023)	0.572 (0.023)
Ovarian Cancer Data		
Number of Genes	4887	4887
CVPL	-364.845 (0.949)	-363.554 (0.626)
C-index	0.507 (0.023)	0.572 (0.023)

Figures Legend

Figure 1. Flowchart of the two-stage prognostic model.

Figure 2. Breast cancer data: Brier prediction error curves for two-stage hierarchical-hierarchical Cox model ($s = 1/n\lambda$ and $s_2 = 1/n\lambda + 0.03$), two-stage lasso-hierarchical Cox model, two-stage ridge-hierarchical Cox model, best fitted pathway, and joint lasso.

Figure 3. Ovarian cancer data: Brier prediction error curves for two-stage hierarchical-hierarchical Cox model ($s = 1/n\lambda$), two-stage lasso-hierarchical Cox model, best fitted pathway, and joint lasso.

Figure 4. Kaplan-Meier curves of breast cancer data for low-risk and high-risk groups from joint lasso, joint hierarchical Cox model, two-stage lasso-hierarchical (L-H) model and two-stage hierarchical-hierarchical (H-H) model ($s = 1/n\lambda$). P-values are calculated using log-rank test. Red dash line is for all tumors, green solid line is for low-risk group, and blue solid line is for high-risk group.

Figure 5. Kaplan-Meier curves of ovarian cancer data for low-risk and high-risk groups from joint lasso, joint hierarchical Cox model, two-stage lasso-hierarchical (L-H) model and two-stage hierarchical-hierarchical (H-H) model ($s = 1/n\lambda$). P-values are calculated using log-rank test. Red dash line is for all tumors, green solid line is for low-risk group, and blue solid line is for high-risk group.

Figure 6. Significant pathways for breast cancer data: estimated coefficients (points and short lines) and p-values (right side) of two-stage lasso-hierarchical Cox model and two-stage hierarchical-hierarchical Cox model ($s = 1/n\lambda$), and summaries of the identified pathways. Starred pathways are consistent with core cancer pathways.

Figure 7. Significant pathways for ovarian cancer data: estimated coefficients (points and short lines) and p-values (right side) of two-stage lasso-hierarchical Cox model and two-stage hierarchical-hierarchical Cox model ($s = 1/n\lambda$), and summaries of the identified pathways. Starred pathways are consistent with core cancer pathways.

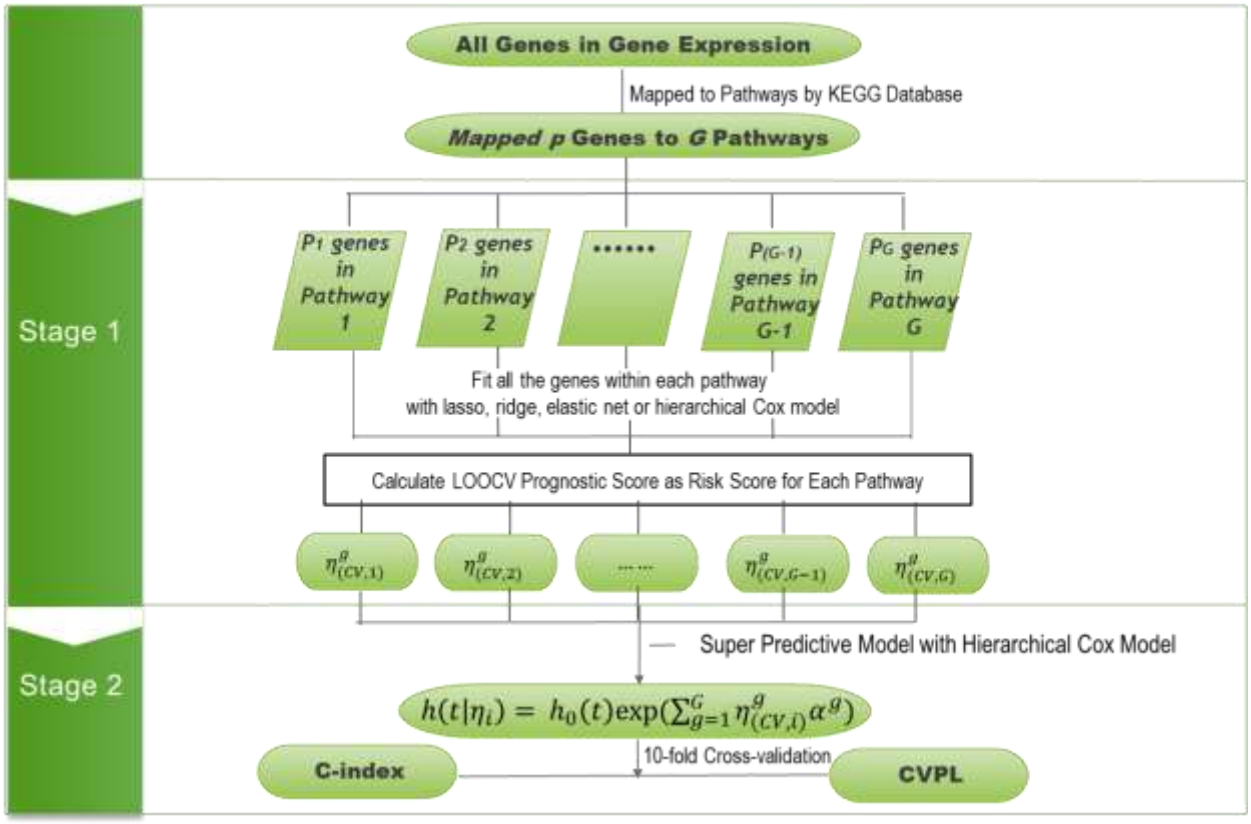
All Genes in Gene Expression
Mapped to Pathways by KEGG Database
Mapped p Genes to G Pathways
Stage 1
P1 genes in Pathway 1
P2 genes in Pathway 2
.....
P(G-1) genes in Pathway G-1
PG genes in Pathway G
Fit all the genes within each pathway
with lasso, ridge, elastic net or hierarchical Cox model
Calculate LOOCV Prognostic Score as Risk Score for Each Pathway
Stage 2
Super Predictive Model with Hierarchical Cox Model
$h(t|\eta_i) = h_0(t)\exp(\sum_{g=1}^{G}\eta_{(CV,i)}^{g}\alpha^{g})$
10-fold Cross-validation
C-index
CVPL

Figure 1

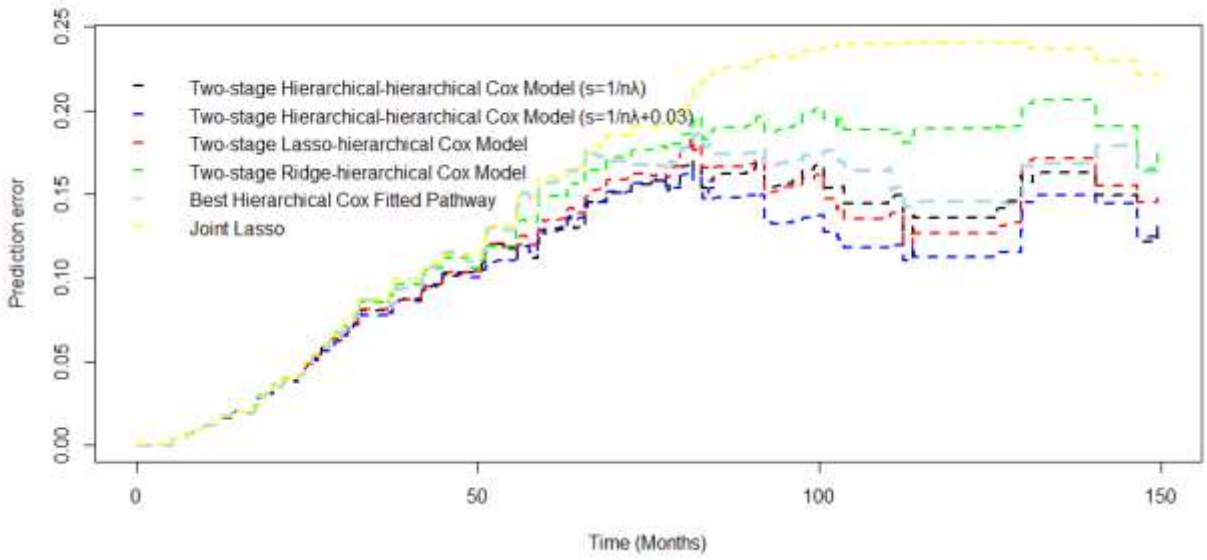
Two-stage Hierarchical-hierarchical Cox Model (s=1/nλ)
Two-stage Hierarchical-hierarchical Cox Model (s=1/nλ+0.03)
Two-stage Lasso-hierarchical Cox Model
Two-stage Ridge-hierarchical Cox Model
Best Hierarchical Cox Fitted Pathway
Joint Lasso
Prediction error
0.00
0.05
0.10
0.15
0.20
0.25
0
50
100
150
Time (Months)

Figure 2

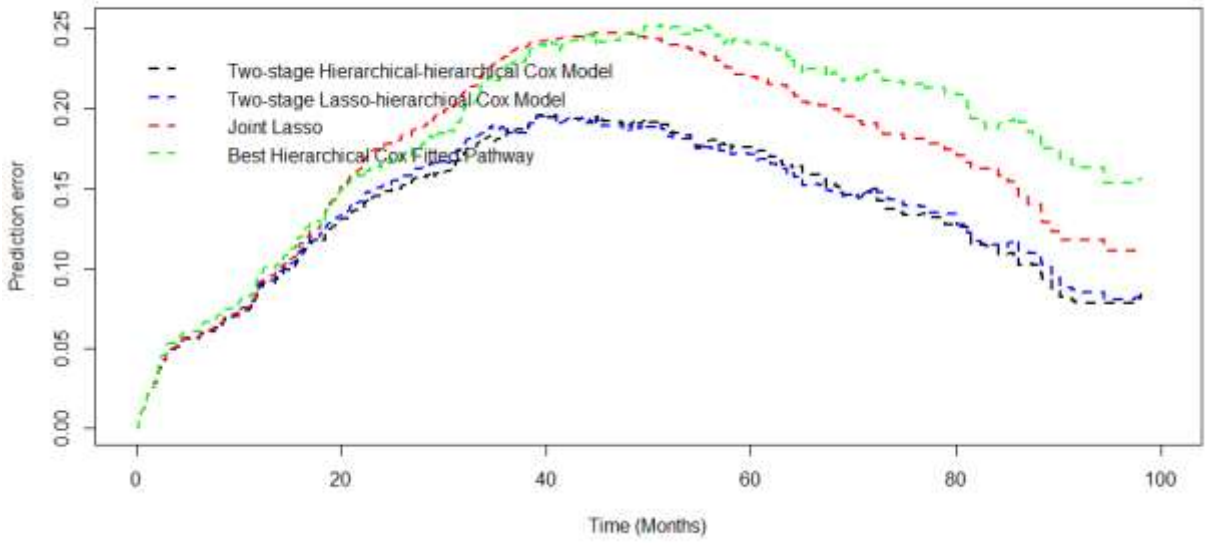
Two-stage Hierarchical-hierarchical Cox Model
Two-stage Lasso-hierarchical Cox Model
Joint Lasso
Best Hierarchical Cox Fitted Pathway
Prediction error
0.00
0.05
0.10
0.15
0.20
0.25
0
20
40
60
80
100
Time (Months)

Figure 3

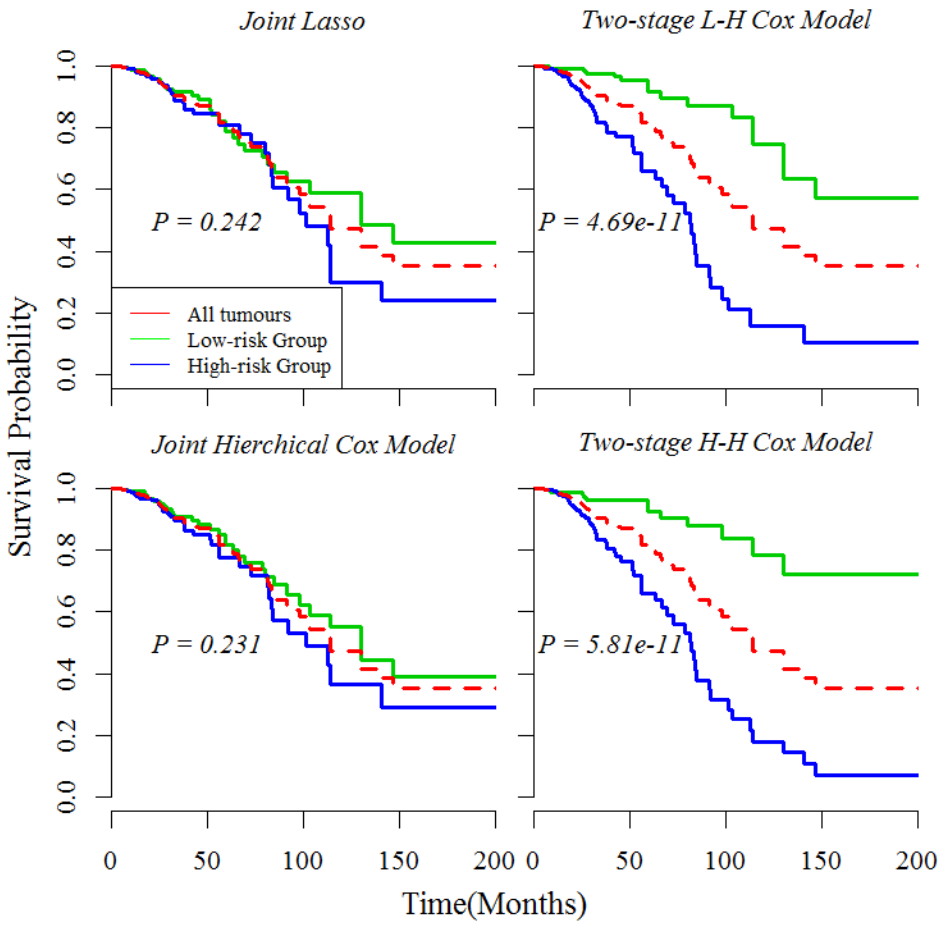
Joint Lasso
Two-stage L-H Cox Model
Joint Hierchical Cox Model
Two-stage H-H Cox Model
P = 0.242
P = 4.69e-11
P = 0.231
P = 5.81e-11
All tumours
Low-risk Group
High-risk Group
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
Time(Months)

Figure 4

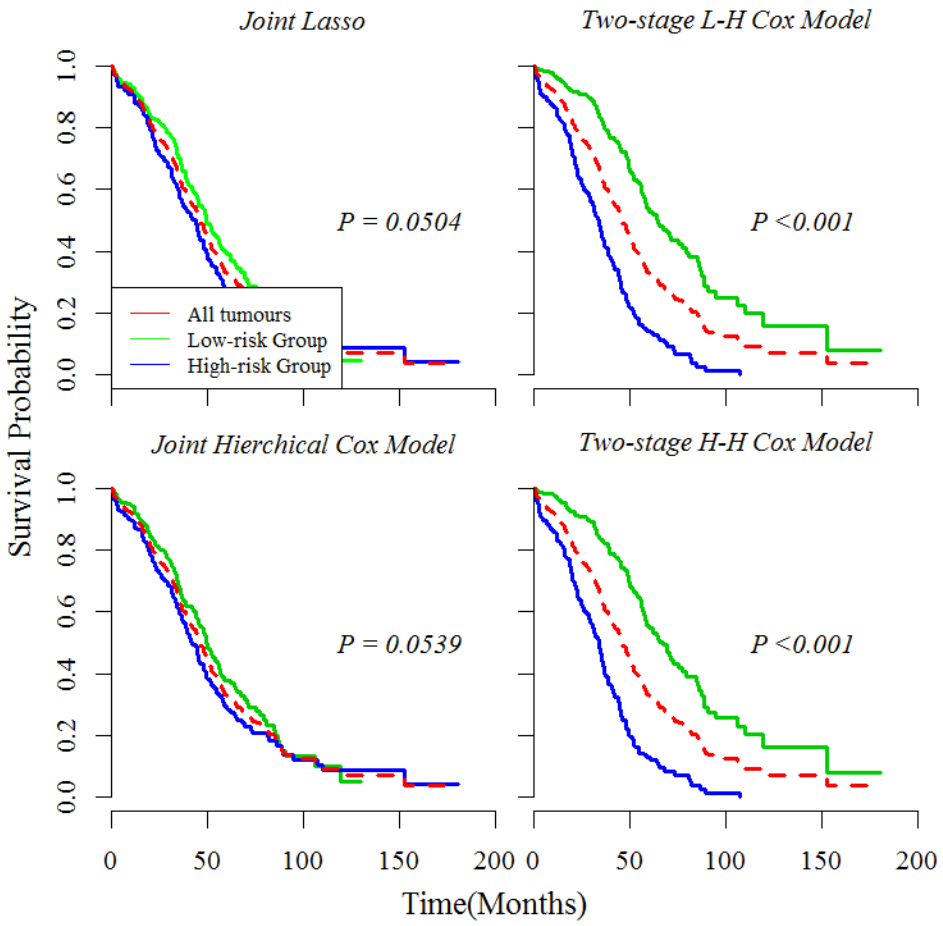
Joint Lasso
Two-stage L-H Cox Model
Joint Hierchical Cox Model
Two-stage H-H Cox Model
P = 0.0504
P <0.001
P = 0.0539
P <0.001
All tumours
Low-risk Group
High-risk Group
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
Time(Months)

Figure 5

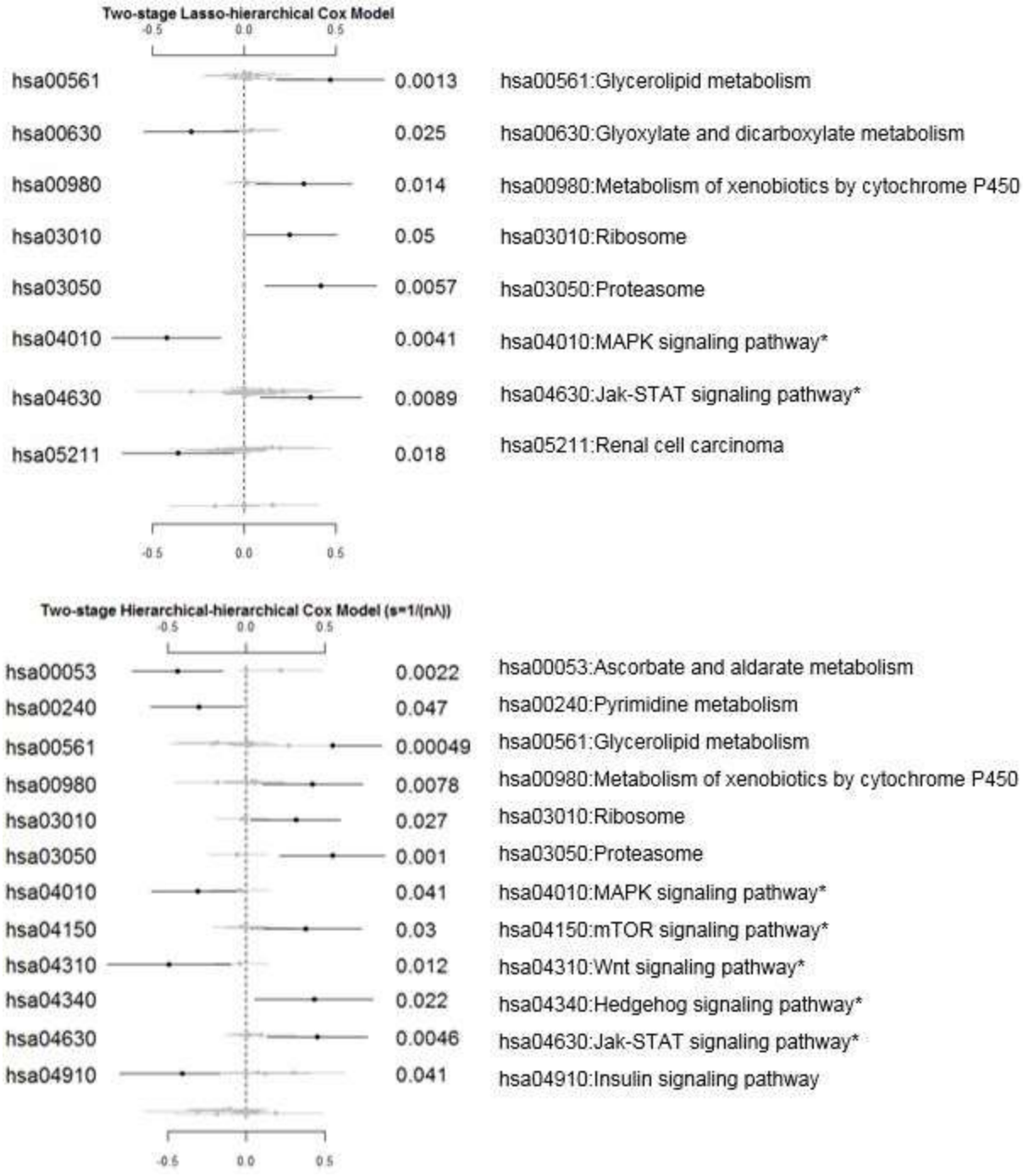
Two-stage Lasso-hierarchical Cox Model
-0.5
0.0
0.5
hsa00561
0.0013
hsa00561:Glycerolipid metabolism
hsa00630
0.025
hsa00630:Glyoxylate and dicarboxylate metabolism
hsa00980
0.014
hsa00980:Metabolism of xenobiotics by cytochrome P450
hsa03010
0.05
hsa03010:Ribosome
hsa03050
0.0057
hsa03050:Proteasome
hsa04010
0.0041
hsa04010:MAPK signaling pathway*
hsa04630
0.0089
hsa04630:Jak-STAT signaling pathway*
hsa05211
0.018
hsa05211:Renal cell carcinoma
-0.5
0.0
0.5
Two-stage Hierarchical-hierarchical Cox Model (s=1/(nλ))
-0.5
0.0
0.5
hsa00053
0.0022
hsa00053:Ascorbate and aldarate metabolism
hsa00240
0.047
hsa00240:Pyrimidine metabolism
hsa00561
0.00049
hsa00561:Glycerolipid metabolism
hsa00980
0.0078
hsa00980:Metabolism of xenobiotics by cytochrome P450
hsa03010
0.027
hsa03010:Ribosome
hsa03050
0.001
hsa03050:Proteasome
hsa04010
0.041
hsa04010:MAPK signaling pathway*
hsa04150
0.03
hsa04150:mTOR signaling pathway*
hsa04310
0.012
hsa04310:Wnt signaling pathway*
hsa04340
0.022
hsa04340:Hedgehog signaling pathway*
hsa04630
0.0046
hsa04630:Jak-STAT signaling pathway*
hsa04910
0.041
hsa04910:Insulin signaling pathway
-0.5
0.0
0.5

Figure 6

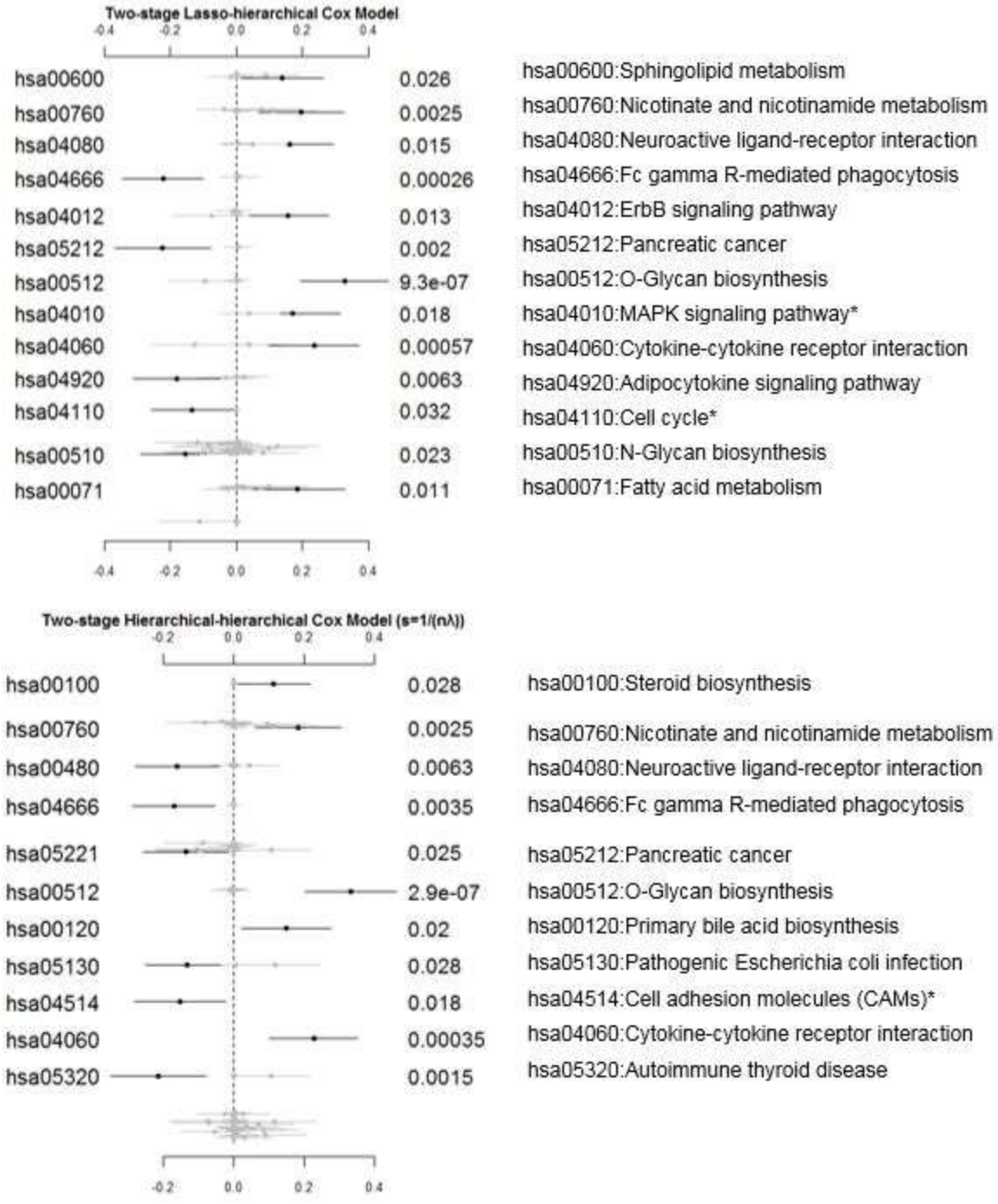
Two-stage Lasso-hierarchical Cox Model
-0.4
-0.2
0.0
0.2
0.4
hsa00600
hsa00760
hsa04080
hsa04666
hsa04012
hsa05212
hsa00512
hsa04010
hsa04060
hsa04920
hsa04110
hsa00510
hsa00071
0.026
0.0025
0.015
0.00026
0.013
0.002
9.3e-07
0.018
0.00057
0.0063
0.032
0.023
0.011
hsa00600:Sphingolipid metabolism
hsa00760:Nicotinate and nicotinamide metabolism
hsa04080:Neuroactive ligand-receptor interaction
hsa04666:Fc gamma R-mediated phagocytosis
hsa04012:ErbB signaling pathway
hsa05212:Pancreatic cancer
hsa00512:O-Glycan biosynthesis
hsa04010:MAPK signaling pathway*
hsa04060:Cytokine-cytokine receptor interaction
hsa04920:Adipocytokine signaling pathway
hsa04110:Cell cycle*
hsa00510:N-Glycan biosynthesis
hsa00071:Fatty acid metabolism
Two-stage Hierarchical-hierarchical Cox Model (s=1/(nλ))
-0.2
0.0
0.2
0.4
hsa00100
hsa00760
hsa00480
hsa04666
hsa05221
hsa00512
hsa00120
hsa05130
hsa04514
hsa04060
hsa05320
0.028
0.0025
0.0063
0.0035
0.025
2.9e-07
0.02
0.028
0.018
0.00035
0.0015
hsa00100:Steroid biosynthesis
hsa00760:Nicotinate and nicotinamide metabolism
hsa04080:Neuroactive ligand-receptor interaction
hsa04666:Fc gamma R-mediated phagocytosis
hsa05212:Pancreatic cancer
hsa00512:O-Glycan biosynthesis
hsa00120:Primary bile acid biosynthesis
hsa05130:Pathogenic Escherichia coli infection
hsa04514:Cell adhesion molecules (CAMs)*
hsa04060:Cytokine-cytokine receptor interaction
hsa05320:Autoimmune thyroid disease

Figure 7

Legend of Supplements

File S1. R code for the two-stage pathway approach.

File S2. Results for combining clinical factors with pathway matrix.