Published in final edited form as:
Methods Inf Med. 2012 ; 51(2): 152–161. doi:10.3414/ME11-02-0019.

Identification of Breast Cancer Prognosis Markers using Integrative Sparse Boosting

Shuangge Ma[1], **Jian Huang**[2], **Yang Xie**[3], and **Nengjun Yi**[4]

[1]School of Public Health, Yale University

[2]Department of Statistics and Actuarial Science, University of Iowa

[3]Department of Clinical Sciences, UT Southwestern Medical Center

[4]Department of Biostatistics, Section on Statistical Genetics, University of Alabama at Birmingham

Summary

Objectives—In breast cancer research, it is important to identify genomic markers associated with prognosis. Multiple microarray gene expression profiling studies have been conducted, searching for prognosis markers. Genomic markers identified from the analysis of single datasets often suffer a lack of reproducibility because of small sample sizes. Integrative analysis of data from multiple independent studies has a larger sample size and may provide a cost-effective solution.

Methods—We collect four breast cancer prognosis studies with gene expression measurements. An accelerated failure time (AFT) model with an unknown error distribution is adopted to describe survival. An integrative sparse boosting approach is employed for marker selection. The proposed model and boosting approach can effectively accommodate heterogeneity across multiple studies and identify genes with consistent effects.

Results—Simulation study shows that the proposed approach outperforms alternatives including meta-analysis and intensity approaches by identifying the majority or all of the true positives, while having a low false positive rate. In the analysis of breast cancer data, 44 genes are identified as associated with prognosis. Many of the identified genes have been previously suggested as associated with tumorigenesis and cancer prognosis. The identified genes and corresponding predicted risk scores differ from those using alternative approaches. Monte Carlo-based prediction evaluation suggests that the proposed approach has the best prediction performance.

Conclusions—Integrative analysis may provide an effective way of identifying breast cancer prognosis markers. Markers identified using the integrative sparse boosting analysis have sound biological implications and satisfactory prediction performance.

Keywords

Breast cancer prognosis; Gene Expression; Integrative analysis; Sparse boosting

1. Introduction

Worldwide, breast cancer is the commonest cancer death among women. In 2008, breast cancer caused 458,503 deaths worldwide (13.7% of cancer deaths in women).

Correspondence to: Shuangge Ma 60 College ST, New Haven CT 06520, USA School of Public Health Yale University Tel: 203-785-3119 Fax: 203-785-6912 Shuangge.ma@yale.edu.

Epidemiologic studies have been extensively conducted, searching for risk factors of breast cancer. In the literature, established risk factors include age, lack of childbearing or breastfeeding, higher hormone levels, race, economic status and dietary iodine deficiency. Multiple models have been developed for the prediction of breast cancer etiology and prognosis based on those risk factors. Despite considerable effort, it has been recognized that models containing only epidemiologic risk factors and environmental exposures are not sufficient, and genomic markers can improve predictive power [1,2]. Multiple high-throughput profiling studies have been conducted, searching for genomic markers associated with breast cancer prognosis. Considerable progresses have been made. Examples include the 97-gene signature [3], which includes genes UBE2C, PKNA2, TPX2, FOXM1, STK6, CCNA2, BIRC5, MYBL2 and others, and the 70-gene signature [4], which involves the hallmarks of cancer including cell cycle, metastasis, angiogenesis, and invasion. A comprehensive review is provided in [1].

In high-throughput cancer studies with gene expression measurements, genomic markers identified from the analysis of a single dataset often suffer a lack of reproducibility. With breast cancer prognosis gene signatures, the lack of reproducibility has been noted [1]. This is reconfirmed by our data analysis (Section 5). Multiple factors may contribute to the lack of reproducibility, including the incomparability of study subjects, inherent variation in the profiling process, and insufficient accounting for the hierarchical structure. Another, perhaps more important, reason for the lack of reproducibility is the small sample sizes of individual cancer gene expression studies. Large-scale, prospective studies may provide an ideal solution. However, such studies can be extremely expensive and time-consuming. For breast cancer prognosis as well as several other cancer outcomes, there are multiple independent studies sharing comparable designs, which makes it possible to conducted pooled analysis and increase sample size [5,6,7].

Available multi-datasets analysis approaches can be classified as meta-analysis and integrative analysis approaches [7,8]. In meta-analysis, multiple datasets are analyzed separately. Then summary statistics, for example the lists of identified genes, p-values or effect sizes, are pooled across multiple datasets. In contrast, integrative analysis approaches pool and analyze raw data from multiple studies. With gene expression data, intensity approaches, a family of integrative analysis approaches, search for transformations that make gene expressions comparable across multiple studies and platforms. After transformation, multiple datasets are combined and analyzed using single-dataset methods. Intensity approaches do not demand new analysis approaches. However, they may be limited in that they need to be conducted on a case-by-case basis, and there is no guarantee that the desired transformations always exist.

The goal of this study is to identify genes associated with breast cancer relapse-free survival. For this purpose, four independent breast cancer studies are collected. We describe the relationship between survival and gene expressions using the accelerated failure time (AFT) models. Unlike the commonly adopted Cox model, the AFT model assumes a linear function for the transformed event time, and thus its regression coefficients may have a more straightforward interpretation. In addition, its estimation objective function can be simpler than Cox model's partial likelihood function and hence incurs lower computational cost. In addition, studies have shown that the AFT model can be an appropriate choice when the proportional hazard assumption has been violated. This model has been adopted in [6,10,11,12] for modeling prognosis data with gene expression measurements. We adopt a sparse boosting approach for marker selection. Boosting provides an effective way of combining multiple weak learners into a strong one. It can be appropriate for cancer gene expression data as individual genes usually have weak effects, but combined together, they may have a strong effect [13,14]. We provide a brief introduction of the most relevant boosting techniques in Section 3 and refer to [15,16] and others for comprehensive reviews.

The rest of the article is organized as follows. In Section 2, we describe the data and model setup. We describe marker selection using an integrative sparse boosting approach in Section 3. We conduct simulation study in Section 4 to examine performance of the boosting approach. In Section 5, we analyze four breast cancer studies and identify prognosis markers. The article concludes with discussion in Section 6.

2 Integrative Analysis of Multiple Heterogeneous Prognosis Studies

2.1 Integrative analysis

Consider the scenario where there are multiple independent studies investigating the association between the same cancer prognosis response variable and the same set of genes. With such data, our goal is to identify genes showing consistent associations across multiple studies. Such genes are more likely to represent the essential features of cancer development [17,18]. Although those studies share a certain common ground, it is inappropriate to directly combine data and analyze as if they were generated in a single study. Particularly with gene expression data, the comparability of measurements using different platforms and/or from different batches is still debatable. There is a lack of generically applicable transformation/function that links one unit increase in cDNA measurement to that in Affymetrix measurement. In addition, confounding from unmeasured clinical and environmental risk factors may lead to different strengths of association between genes and prognosis in different studies. To tackle this heterogeneity problem, recent studies suggest allowing for different statistical models in different studies [5,6,7].

2.2 Statistical modeling

Consider data from M independent prognosis studies. Assume that the same set of gene expressions (also referred to as “covariates”) are measured in all studies. Denote $T^1,...,T^M$ as the logarithm of failure times and $X^1,..., X^M$ as length-d covariates. In study $m(= 1,...,M)$, assume the AFT model

$$T^m = \alpha^m + \beta^{m'} X^m + \varepsilon^m.$$

Here α^m is the unknown intercept, β^m is the regression coefficient, β^{m}’ is the transpose of β^m, and ε^m is the random error with an unknown distribution. Denote $C^1,...,C^M$ as the logarithm of random censoring times. Under right censoring, one observation consists of (Y^m, δ^m, X^m), where $Y^m = \min(T^m, C^m)$ and $\delta^m = I(T^m \leq C^m)$.

Denote $\beta = (\beta^1,...,\beta^M$ as the $d \times M$ matrix of regression coefficient. Recent studies [5,6,7] suggest that β should have the following characteristics. First, β^m s are sparse in the sense that only a few elements are nonzero. In genome-wide studies, a large number of genes are profiled. However, only a few genes are associated with prognosis and hence have nonzero regression coefficients. Considering that so far only less than 400 genes have been identified as “cancer genes”, this sparsity property is reasonable. Second, $\beta^1,...,\beta^M$ have the same sparsity structure. That is, elements of β in the same row are either all zero or all nonzero. This characteristic reflects the fact that multiple studies share the same set of markers. Third, for a cancer marker with nonzero regression coefficients, the magnitudes of regression coefficients are allowed to differ across studies to accommodate heterogeneity.

2.3 Weighted least squares estimation

In study m (= 1,..., M), assume n^m iid observations $\left\{\left(Y_i^m, \delta_i^m, X_i^m\right), i=1\dots n^m\right\}$. Denote $Y_{(1)}^m \le \dots \le Y_{(n^m)}^m$ are the order statistics of Y_i^m s, $\delta_{(1)}^m, \dots, \delta_{(n^m)}^m$ as the associated censoring indicators and $X_{(1)}^m, \dots, X_{(n^m)}^m$ as the associated covariates. Let $\hat{F}^m$ be the Kaplan-Meier estimate of F^m, the distribution function of T^m. It can be computed as $\widehat{F}^m(y) = \sum_{i=1}^{n^m} w_i^m I\left(Y_{(i)}^m \le y\right)$. w_i^m s are the jumps in the Kaplan-Meier estimate and computed as

$$w_1^m = \frac{\delta_{(1)}^m}{n^m} \quad \text{and} \quad w_{(i)}^m = \frac{\delta_{(i)}^m}{n^m - i + 1}\prod_{j=1}^{i-1}\left(\frac{n^m - j}{n^m - j + 1}\right)^{\delta_{(j)}^m} \quad \text{for} \quad i=2\dots n^m.$$

For study m, the weighted least squares loss function is

$$R^m(\beta^m) = \frac{1}{2}\sum_{i=1}^{n^m} w_i^m \left(Y_{(i)}^m - \alpha^m - \beta^{m\prime} X_{(i)}^m\right)^2.$$

Center $X_{(i)}^m$ and $Y_{(i)}^m$ as $X_{(i)}^{m*} = \sqrt{w_i^m}\left(X_{(i)}^m - \frac{\sum w_i^m X_{(i)}^m}{\sum w_i^m}\right)$ and $Y_{(i)}^{m*} = \sqrt{w_i^m}\left(Y_{(i)}^m - \frac{\sum w_i^m Y_{(i)}^m}{\sum w_i^m}\right)$. The overall loss function is

$$R(\beta) = \sum_{m=1}^{M} R^m(\beta^m) = \sum_{m=1}^{M} \frac{1}{2}\sum_{i=1}^{n^m}\left(Y_{(i)}^{m*} - \beta^{m\prime} X_{(i)}^{m*}\right)^2.$$

In this study, it is assumed that the error distribution is unknown. Thus the likelihood based approach in [10] is not applicable. Multiple methods have been developed for estimating the AFT model with an unknown error distribution. Commonly used approaches include the Buckley-James approach, rank-based approach and others [11]. In theoretical studies with a fixed number of covariates, it is shown that none of the existing estimation methods dominates the others. We adopt the weighted least squares approach [19], which is equivalent to the inverse probability weighted approach, because of its low computational cost.

3. Marker Selection with Integrative Sparse Boosting

3.1 Boosting

Boosting approaches assemble multiple weak learners into a strong learner. Compared with alternative approaches, boosting approaches may be preferred because of their flexibility, affordable computational cost and satisfactory empirical performance [20,21]. The literature on boosting is too vast to be reviewed here. Below, we provide brief descriptions of L2 boosting [22] and sparse boosting [5,16], which have statistical framework closest to the approach we adopt. With boosting approaches, there are usually multiple choices for weak learners. In this study, we take the linear functions of gene expressions as weak learners. It is possible to adopt more complex weak learners, for example classification trees and polynomial functions. However, such weak learners can lead to high computational cost and a lack of interpretability and hence are not pursued.

L2 Boosting is designed for the analysis of a single dataset with a squared error loss function. It iteratively fits working residuals with weak leaners. It can be derived from a generic functional gradient descent algorithm using the squared error loss [22]. For the completeness of this article, we describe the L2 Boosting algorithm below and refer to published studies for more details.

Consider the data and model settings described in Section 2. Without loss of generality, consider the first study ($m = 1$). Then the least squared loss function is $R^1\left(\beta^1\right)=\frac{1}{2}\sum_{i=1}^{n^1}\left(Y_i^{1^*}-\beta^{1'}X_i^{1^*}\right)^2$. The L2 Boosting proceeds as follows.

Step 1: Initialize $k = 0$, the working residual $U^{[k]}=(u_1,\ldots,u_{n^1})=\left(Y_1^{1^*},\ldots,Y_{n^1}^{1^*}\right)$ and $\widehat{f}^{[k]}\left(X_i^1\right)=0$. $\widehat{f}^{[k]}\left(X_i^1\right)$ is an estimate of $E\left(Y_i^{1^*}\mid X^1=X_i^1\right)$;

Step 2: Fit and update. $k = k + 1$. Compute

$\widehat{s}=\arg\min_{1\le s\le d}\quad \arg\min_{\gamma}\sum_{i=1}^{n^1}\left(u_i-\gamma\times X_{i,s}^{1^*}\right)^2$. Here $X_{i,s}^{1^*}$ is the s th component of $X_i^{1^*}$.

Update $\widehat{f}^{[k]}=\widehat{f}^{[k-1]}+\nu\times\widehat{\gamma_{\widehat{s}}}X_{\widehat{s}}^1$, where $\widehat{\gamma_{\widehat{s}}}=\arg\min_{\gamma}\sum_{i=1}^{n^1}\left(u_i-\gamma\times X_{i,\widehat{s}}^{1^*}\right)^2$. ν is the step size.

Following [16] and references therein, we set $\nu=0.1$.

Update $U^{[k]}=U^{[k-1]}-\nu\times\widehat{\gamma_{\widehat{s}}}X_{\widehat{s}}^{1^*}$.

Step 3: Repeat Step 2 until a certain stopping rule is reached.

This approach iteratively selects one weak learner, which leads to the most improvement of goodness-of-fit, and updates its estimate. In Step 2, a tuning parameter is the step size. Buhlmann and Yu [16] suggests that the choice of step size is not critical, as long as it is small enough. Our limited numerical study confirms this observation. We note that it may be possible to select the step size in a data-dependent way. However, to reduce computational cost, in this study we fix the step size. In Step 3, there are multiple choices for the stopping rule, including AIC, BIC, cross validation, GCV and others. With high dimensional gene expression data, it is not clear what the optimal stopping rule is. In this study, we choose $\hat{k}$, the optimal number of iterations, to minimize a BIC criterion. In our simulation, we have experimented with a few other stopping rules but failed to identify one significantly better than BIC. As this approach adds one weaker at each iteration, when the stopping rule is appropriate, the resulted strong learner may enjoy a certain degree of sparsity.

Sparse boosting is first proposed in [16] for simple linear regression. A different sparse boosting approach is proposed in [23]. Ordinary boosting may enjoy a certain degree of sparsity, if the weak learners and stopping rule are properly chosen. However, numerical studies suggest that with high dimensional data, ordinary boosting is not “sparse enough”. That is, it may identify a considerable number of false positives. Sparse boosting may introduce further sparsity by modifying the loss function and stopping rule. Specifically, the sparse boosting loss function consists of two parts. The first part is the same as that with ordinary boosting and measures goodness-of-fit. The second part is a penalty term and measures model complexity. Choices for model complexity measure include AIC, BIC, MDL (minimum description length) and others. Weak learners are chosen in a way that balances sparsity and goodness-of-fit and differ from those chosen based only on goodness-of-fit.

3.2 Integrative sparse boosting

The boosting approaches described in Section 3.1 are designed for the analysis of a single dataset. In a recent study, we develop a sparse boosting approach for the integrative analysis of multiple heterogeneous diagnosis studies [5]. In this article, we extend the approach in [5] to prognosis studies. Two boosting algorithms, integrative L2 boosting (iBoost) and integrative sparse L2 boosting (iSBoost), will be considered. They significantly differ from existing approaches along the following aspects. First, they advance from the L2 Boosting and sparse boosting in [16,23] by analyzing multiple heterogeneous datasets. Second, in this study, we analyze censored prognosis data, whereas [5] focuses on diagnosis data with binary responses. Thus the statistical models and loss functions are significantly different. In addition, the BIC criterion is adopted for boosting and stopping. It is much more commonly adopted and easily extendable than the MDL in [5,16]. More importantly, this study is the first time the sparse boosting technique is applied to identify breast cancer prognosis markers.

3.2.1 iBoost—Consider the data and model settings described in Section 2. In study m(= 1,..., M), denote $f^m(x^{m*}) = E(Y^{m*} \mid X^{m*} = x^{m*})$. Following [5,6,7] and discussions in Section 2, we allow for study-specific f^m(m = 1...M) to accommodate heterogeneity. The iBoost algorithm proceeds as follows.

Step 1: Initialize $k = 0$. For $m = 1,...,M$, initialize the working residuals $U^{m[k]} = \left(u_1^m, \ldots, u_{n^m}^m\right) = \left(Y_1^{m^*}, \ldots, Y_{n^m}^{m^*}\right)$ and $\hat{f}^{m[k]} = 0$. Here $\hat{f}^{m[k]}(X^{m*})$ is an estimate of $E(Y^{m*} \mid X^{m*})$. As only linear weak learners are used, write $\widehat{f}^{m[k]}\left(X^{m^*}\right) = \widehat{\beta}^{m[k]\prime} X^{m^*}$;

Step 2: Fit and update. $k = k + 1$. Compute $\widehat{s} = \text{arg min}_{1 \le s \le d} \quad \text{arg min}_{\gamma_s^1,\ldots,\gamma_s^M} \sum_{m=1}^{M} \quad \sum_{i=1}^{n^m} \left(u_i^m - \gamma_s^m \times x_{i,s}^m\right)^2$. Here $x_{i,s}^m$ is the sth component of $X_i^{m^*}$.

Compute $\left(\widehat{\gamma}_{\widehat{s}}^1, \ldots, \widehat{\gamma}_{\widehat{s}}^M\right) = \text{arg min}_{\gamma_{\widehat{s}}^1,\ldots,\gamma_{\widehat{s}}^M} \sum_{m=1}^{M} \quad \sum_{i=1}^{n^m} \left(u_i^m - \gamma_{\widehat{s}}^m \times x_{i,\widehat{s}}^m\right)^2$.

Update $\widehat{f}^{m[k]} = \widehat{f}^{m[k-1]} + \nu \times \widehat{\gamma}_{\widehat{s}}^m X_{\widehat{s}}^m, m = 1, \ldots, M$, where ν is the step size as in L2 boosting;

Update $U^{m[k]} = U^{m[k-1]} - \nu \times \widehat{\gamma}_{\widehat{s}}^m X_{\widehat{s}}^m$.

Step 3: Repeat Step 2 for K iterations.

Step 4: At iteration k, compute $S^{[k]} = \sum_{m=1}^{M} \sum_{i=1}^{n^m} \left(u_i^m\right)^2 + \log\left(\sum_{m=1}^{M} n^m\right) \sum_{j=1}^{d} I\left(\|\widehat{\beta}_j^{m[k]}\| \neq 0\right)$. The optimal number of iterations is computed as $\hat{k} = \text{arg min}_{1 \le k \le K} S^{[k]}$.

Rationale With multiple independent datasets and without making assumptions on gene effects across studies, the overall loss function is taken as the sum of individual loss functions. iBoost is a boosting approach in that at each iteration, it searches for one weak learner and then updates its estimates. It differs from existing boosting approaches in that when choosing the weak learners, iBoost evaluates the overall effects of genes across M studies. The selected gene has the strongest overall effect, however, not necessarily the strongest effects in individual datasets. When a gene is selected, it is selected in all M studies, leading to the same set of genes identified as prognosis markers and the same sparsity structure for all M models. The regression coefficients $\widehat{\gamma}_{\widehat{s}}^m$ s are allowed to differ for different m, which is in line with discussions in Section 2. In Step 4, we adopt a BIC criterion for stopping.

3.2.2 iSBoost—

iSBoost takes a different approach for selecting weak learners. Particularly, it modifies Step 2 of iBoost as

$$\widehat{s}=\arg\min_{1\le s\le d}\quad \arg\min_{\gamma_s^1,\dots,\gamma_s^M}\sum_{m=1}^{M}\sum_{i=1}^{n^m}\left(u_i^m-\gamma_s^m\times x_{i,s}^m\right)^2+\log\left(\sum_{m=1}^{M}n^m\right)\sum_{j=1}^{d}I\left(\|\widehat{\beta}_j^{m[k]}\|\neq 0\right).$$

Other steps remain the same.

This modification has been motivated by the following considerations. iBoost takes the goodness-of-fit as the weak learner selection criterion. Thus there is a risk that it may introduce genes that can only improve goodness-of-fit by a small amount but add model complexity by selecting more genes. In classic statistical analysis as well as [16], it has been suggested that model complexity should be considered along with goodness-of-fit in model/variable selection. With iSBoost, we use a BIC-type objective function as the criterion for weak learner selection to encourage sparsity. As a different weak learner selection rule is adopted, the sequence of weak learners selected by iSBoost may differ from that selected by iBoost. Specifically, with penalty on the number of selected genes, iSBoost tends to select fewer genes.

4. Simulation Study

For simplicity of notation, we have assumed that multiple studies have matched gene sets. When different gene sets are measured in different studies, we can adopt a simple rescaling approach. Assume that gene s is only measured in the first M_s studies. We modify Step 2 of iBoost as

$$\sum_{m=1}^{M}\quad\sum_{i=1}^{n^m}\left(u_i^m-\gamma_s^m\times x_{i,s}^m\right)^2\to\sum_{m=1}^{M_s}\quad\sum_{i=1}^{n^m}\left(u_i^m-\gamma_s^m\times x_{i,s}^m\right)^2\times\frac{\sum_{m=1}^{M}n^m}{\sum_{m=1}^{M_s}n^m}.$$

Other quantities can be rescaled in a similar manner.

We simulate four independent datasets, each with 100 subjects. We simulate 50 and 100 gene clusters, with 20 genes per cluster. Thus the total numbers of gene expressions simulated are 1000 and 2000. Gene expressions have marginally normal distributions. Genes in different clusters have independent expressions. For genes within the same clusters, their expressions have the following correlation structures: (i) auto-regressive correlation, where expressions of genes j and k have correlation coefficient $\rho^{|j-k|}$; (ii) banded correlation, where expressions of genes j and k have correlation coefficient max(0,1–|j–k|ρ); and (iii) compound symmetry, where expressions of genes j and k have correlation coefficient ρ when $j\neq k$. Under each correlation scenario, we consider two different values of ρ. Within each of the first 4 clusters, there are 5 genes associated with prognosis. Thus there are a total of 20 prognosis-associated genes, and the rest are noises. For prognosis-associated genes, we generate their nonzero regression coefficient from $Unif\,[-1,-0.5]\cup Unif\,[0.5,1]$. 20% prognosis-associated and 10% noisy genes are measured only in the first two studies. We generate the logarithm of event times from the AFT models with standard normal random errors. The censoring times are generated independently from exponential distributions. We adjust the censoring parameters so that the overall censoring rate is close to 50%. The simulated data closely mimics observed data. Genes have the pathway structure, where genes within the same pathways tend to have correlated expressions, and genes within different pathways tend to have weakly correlated or independent expressions. Only some pathways are associated with prognosis, and within those pathways, only a few genes are associated with prognosis.

To better quantify performance of the proposed approach, we also consider the following alternative approaches. (a) Meta-analysis. We analyze each dataset separately. The lists of identified genes are combined using a vote counting approach. When analyzing each dataset, we consider the following three approaches. The first is the elastic net (Enet), which is a penalization approach. It contains a Lasso component for penalized marker selection and a ridge component to accommodate correlations among genes. It has two tuning parameters, which are selected using 4-fold cross validation. Computation with Enet is realized using the R package *glmnet*. The second is the L2 boosting approach described in Section 3.1 (denoted as "Boost" in Table I). With this approach, the number of iterations is the only tuning parameter and selected using a BIC criterion. The third is the sparse L2 boosting approach (denoted as "SBoost" in Table I). This approach is the single-dataset counterpart of the proposed iSBoost approach, with the weak learner selection and stopping rule determined by a BIC criterion; (b) An intensity approach. In simulation, the similarity among the four datasets is much higher than that encountered in practice. We adopt an intensity approach, transform gene expressions to achieve comparability, combine the four datasets, and analyze as if they were from a single study. For the combined dataset, we analyze using the Enet, Boost and SBoost approaches; and (c) Integrative analysis. We extend the approach in [7] and use a group Enet approach for penalized selection.

In our simulation, data is generated under the AFT model. We first examine performance of the Cox model, which is the most commonly adopted model. Under the simulation scenarios described in rows 1 and 2 of Table I, with the iBoost and iSBoost algorithms for marker selection, the Cox model identifies 8 and 7 true positives, respectively. The unsatisfactory performance of the Cox model under model mis-specification is not surprising and has been previously noted [24]. We do not further pursue the Cox model in simulation.

Simulation suggests that the proposed approach is computationally affordable. Analysis of one replicate takes less than five minutes on a regular desktop PC. Summary statistics based on 100 replicates are shown in Table I. We can see that with the meta-analysis approaches, the majority or all of the true positives can be identified. However, a large number of false positives are also identified. With the simulated data, the degree of heterogeneity is significantly less than that with real data, which favors intensity approaches. Intensity approaches significantly outperform meta-analysis approaches, with considerably fewer false positives and sometimes more true positives. Performance can further improve with integrative analysis. Among the three integrative analysis approaches, loosely speaking, iSBoost imposes the strongest control on model complexity. Thus it is sensible it identifies the fewest genes. Under the auto-regressive and banded correlation structures, it identifies all of the true positives with almost no false positives. Under the compound symmetry structure, it still can identify the majority of true positives, with a reasonable number of false positives. Such a result is intuitively reasonable, as under the compound symmetry structure, one gene is correlated with many genes, which makes it difficult to identify truly important genes.

5. Identification of breast cancer prognosis markers

With gene expression data, preprocessing and normalization are needed prior to analysis. With Affymetrix data, a floor and a ceiling may be added, and then measurements are log2 transformed. With both Affymetrix and cDNA data, we fill in missing expressions with means across samples. We then standardize each gene expression to have zero mean and unit variance. A significant advantage of the proposed integrative analysis is that it does not require the full comparability of measurements from different studies. Thus, cross-study/platform transformation or normalization is not needed.

We analyze four breast cancer prognosis studies with gene expression measurements. We provide brief descriptions of the four studies in Table II and refer to the original publications [4,25,26,27] for more detailed information. Among the four datasets, two used cDNA, one used oligonucleotide arrays, and one used Affymetrix chips for profiling. We match genes in the four studies using their Unigene Cluster IDs. Although the proposed analysis can accommodate partially matched gene sets, we focus on the 2,555 genes that are measured in all four studies to increase reliability.

With the iSBoost approach, 44 genes are identified as associated with breast cancer prognosis (Table III). Searching published literature suggests that quite a few of the identified genes have sound biological implications, which may partly support the effectiveness of the proposed approach. Particularly, gene IGFBP1 is a member of the insulin-like growth factor binding protein (IGFBP) family and encodes a protein with an IGFBP domain and a thyroglobulin type-I domain. It has been identified as a breast cancer marker in [28,29]. Gene DCAF8 encodes a WD repeat-containing protein that interacts with the Cul4-Ddb1 E3 ligase macromolecular complex. Its involvement in breast cancer pathogenesis is proposed in [30]. The amino acid sequence of the protein encoded by gene CIB2 is similar to that of KIP/CIB, calcineurin B and calmodulin. This suggests that the encoded protein may be a Ca2+-binding regulatory protein that interacts with DNA-dependent protein kinase catalytic subunit (DNAPKcs). It is one the of breast cancer prognosis markers identified in [4]. Cyclic AMP-dependent protein kinase A (PKA; encoded by gene PRKAR1B) is an essential enzyme in the signaling pathway of the second messenger cAMP. Through phosphorylation of target proteins, PKA controls many biochemical events in the cell including regulation of metabolism, ion transport and gene transcription [31]. Gene LFNG is a member of the fringe gene family which also includes radical and manic fringe genes. They all encode evolutionarily conserved glycosyltransferases that act in the Notch signaling pathway to define boundaries during embryonic development. Its involvement in breast cancer prognosis has been established in [32,33]. Gene NUCB1, also known as CALNUC, encodes a member of a small calcium-binding EF-hand protein family. Calnuc protein may be a tumor-associated antigen (TAA) that induces autoantibody response in human cancers [34]. Gene IRAK1 encodes the interleukin-1 receptor-associated kinase 1, one of two putative serine/threonine kinases that become associated with the interleukin-1 receptor (IL1R) upon stimulation. This gene is partially responsible for IL1-induced upregulation of the transcription factor NF-kappa B. It is identified as a breast cancer prognosis marker in [4,27]. The protein encoded by gene SMARCC2 is a member of the SWI/SNF family of proteins, whose members display helicase and ATPase activities and are thought to regulate transcription of certain genes by altering the chromatin structure around those genes. The encoded protein is part of the large ATP-dependent chromatin remodeling complex SNF/SWI and contains a predicted leucine zipper motif typical of many transcription factors. Its implication in breast cancer prognosis is suggested by [34]. The proteins encoded by gene FGF2 and FGF7 are members of the fibroblast growth factor (FGF) family. FGF family members possess broad mitogenic and cell survival activities and are involved in a variety of biological processes, including embryonic development, cell growth, morphogenesis, tissue repair, tumor growth and invasion. These proteins have been implicated in diverse biological processes, such as limb and nervous system development, wound healing and tumor growth [35,36]. The protein encoded by gene GSN binds to the “plus” ends of actin monomers and filaments to prevent monomer exchange. The encoded calcium-regulated protein functions in both assembly and disassembly of actin filaments. Defects in this gene are a cause of familial amyloidosis Finnish type (FAF). It is one of the identified breast cancer markers in [35]. The protein encoded by gene MAP2K1 is a member of the dual specificity protein kinase family, which acts as a mitogen-activated protein (MAP) kinase. MAP kinases, also known as extracellular signal-regulated kinases (ERKs), act as an integration point for multiple biochemical signals. This protein kinase lies upstream of MAP kinases and stimulates the enzymatic activity of MAP kinases upon a wide variety of extra- and intracellular signals. As an essential component of MAP kinase signal transduction pathway, this kinase is involved in many cellular processes such as proliferation, differentiation, transcription regulation and development. This gene has been implicated in multiple types of cancers, including lung cancer, bladder cancer, colorectal cancer, endometrial cancer and pancreatic cancer. Gene MST1 is an identified breast cancer marker in [25]. Gene MGP is identified as a breast cancer marker in [36]. Gene SF3B4 encodes one of four subunits of the splicing factor 3B. The protein encoded by this gene cross-links to a region in the pre-mRNA immediately upstream of the branchpoint sequence in pre-mRNA in the prespliceosomal complex A. It also may be involved in the assembly of the B, C and E spliceosomal complexes. In addition to RNA-binding activity, this protein interacts directly and highly specifically with subunit 2 of the splicing factor 3B. Involvement of this gene in breast cancer prognosis is suggested in [34]. The protein encoded by gene ARHGEF9 is a Rho-like GTPase that switches between the active (GTP-bound) state and inactive (GDP-bound) state to regulate CDC42 and other genes. Defects in this gene are a cause of startle disease with epilepsy (STHEE), also known as hyperekplexia with epilepsy. Its implication in breast cancer progression has been inferred in [35]. Gene IMPA2 has also been identified in [4]. The protein encoded by gene WASF1, a member of the Wiskott-Aldrich syndrome protein (WASP)-family, plays a critical role downstream of Rac, a Rho-family small GTPase, in regulating the actin cytoskeleton required for membrane ruffling [32]. Members of the Rab protein family, encoded by gene RAB2A, are nontransforming monomeric GTP-binding proteins of the Ras superfamily that contain four highly conserved regions involved in GTP binding and hydrolysis. Rabs are prenylated, membrane-bound proteins involved in vesicular fusion and trafficking [35]. Vasodilator-stimulated phosphoprotein (VASP) is a member of the Ena-VASP protein family. Ena-VASP family members contain an EHV1 N-terminal domain that binds proteins containing E/DFPPPPXD/E motifs and targets Ena-VASP proteins to focal adhesions. VASP is associated with filamentous actin formation and likely plays a widespread role in cell adhesion and motility. VASP may also be involved in the intracellular signaling pathways that regulate integrinextracellular matrix interactions [34]. We note that although some of the identified genes have been previously identified as breast cancer prognosis markers, this study might be the first time they are identified in an integrative analysis context. In addition, there are new findings that need to be further studied.

We also analyze the same data using multiple alternative approaches. Specifically, we consider both Cox and AFT models. As simulation has shown the satisfactory performance of the sparse boosting, under the Cox model, we only analyze using the sparse boosting approach. Under the AFT model, both Enet and two boosting approaches are considered. Table IV shows that different approaches may identify significantly different sets of genes.

With practical data, it is difficult to objectively evaluate marker identification accuracy. As an alternative, we evaluate prediction performance, which may provide an indirect evaluation of gene identification accuracy. Particularly, it is expected that if the identified genes are more meaningful, prediction using those genes is more accurate. We conduct evaluation using a cross-validation based approach. We first split each dataset randomly into a training set and a testing set with sizes 3:1. We then analyze the training set (including tuning parameter selection, marker identification and estimation) and use the training set models to make prediction for the testing set subjects. Using the predicted risk score $\beta^{m\prime} X^m$, we generate two risk groups with equal sizes. The logrank statistics, which measure the survival difference between the two groups, are computed. For each random split, we compute the mean logrank statistic over four datasets. To avoid biased evaluation caused by an extreme split, we repeat the whole process 100 times, compute the mean logrank statistics and present the results in Table IV. iSBoost under the AFT model has the best prediction performance with the logrank statistic equal to 5.611 (chi-squared distributed with degree of freedom one, p-value 0.018). We also compute the correlations between predicted risk scores generated using iSBoost and those using alternative approaches and then calculate the averages among splits. The mean correlations are shown in Table IV. We can see that the predicted risk scores are moderately correlated, with correlation coefficients ranging from 0.209 to 0.679, suggesting that the iSBoost models are significantly different from those under alternative approaches.

6. Discussion

The identification of breast cancer prognosis markers is of great importance. Integrative analysis provides an effective way of analyzing multiple heterogeneous datasets and identifying reproducible markers. In this study, we analyze four breast cancer prognosis studies with gene expression measurements. We adopt the AFT model to describe survival and the integrative sparse boosting approach for marker selection. Simulation shows satisfactory performance of iSBoost. In data analysis, we show that iSBoost may identify genes significantly different from those using alternative approaches. The identified genes have important biological implications and satisfactory prediction performance.

Although not as popular as the Cox model, the AFT model provides a flexible alternative and has been adopted in multiple cancer prognosis studies. To the best of our knowledge, with extremely high-dimensional data, there is still a lack of model diagnostics tool. With the four breast cancer datasets, the AFT and Cox models identify different sets of genes and different predicted risk scores. The AFT model has slightly better prediction performance. As its regression coefficients have simple interpretations, the AFT model can be preferred. However, we cautiously note that the model will have to be further validated using prospective, independent data. The adopted AFT model assumes an unknown error distribution and can be more flexible than the parametric models in [10]. The weighted least squares estimation may be computationally more affordable than the approach in [11].

In this study, the loss function has a least squares form. The integrative sparse boosting algorithms can also accommodate other types of loss functions. We propose using BIC as the model complexity measure. It is more commonly adopted than MDL [16]. Gene expressions can be highly correlated. It has been pointed out that de-correlation may be needed in marker selection. Under the penalization framework, the Enet approach uses a ridge penalty for de-correlation. We note that the proposed iSBoost does not have an explicit de-correlation component. In our simulation and data analysis, there are correlated but very few extremely highly correlated genes. iSBoost is shown to outperform Enet under such scenarios. It is possible to extend iSBoost, for example adding a ridge penalty to the loss function. However, such an extension will incur higher computational cost and mix the boosting and penalization framework and is not pursued.

In data analysis, we show that iSBoost may identify genes different from alternatives. We are unable to objectively determine which approaches identify gene sets that are "more meaningful". A cross validation based approach is used for prediction evaluation. Although it does not use completely independent data, it compares all approaches on the same basis and is expected to be reasonably fair. In the literature, quite a few different evaluation approaches have been suggested. However, most of them are ad hoc, and there is a lack of consensus. We note that ultimately the identified markers need to be validated using mechanisms studies.

Acknowledgments

We would like to thank the editors and three referees for very careful review and insightful comments, which have led to a significant improvement of the paper. This study has been supported by CA120988, CA152301 and CA142774 from NIH, USA and Research of Longitudinal Data Analysis Methodology and Its Application (2009JJD910002) from Key Research Institute of Humanities and Social Sciences Program, China.

References

1. Cheang M, van de Rijin M, Nielson TO. Gene expression profiling of breast cancer. Annual Review of Pathology: Mechanisms of Disease. 2008; 3:67–97.
2. Knudsen, S. Cancer Diagnostics with DNA Microarrays. Wiley; 2006.
3. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. JNCI. 2006; 98:262–72. [PubMed: 16478745]
4. van't Veer LJ, Dai H, Vijver MJ van de, He YD, Hart AA, Mao M, Peterse HL, Kooy K van der, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; 415:530–36. [PubMed: 11823860]
5. Huang Y, Huang J, Shia BC, Ma S. Identification of cancer genomic markers via integrative sparse boosting. Biostatistics. 2011 In press.
6. Ma S, Huang J, Wei F, Xie Y, Fang K. Integrative analysis of multiple cancer prognosis studies with gene expression measurements. Statistics in Medicine. 2011 In press.
7. Ma S, Huang J, Song X. Integrative analysis and variable selection with multiple high-dimensional datasets. Biostatistics. 2011 In press.
8. Guerra, R.; Goldstein, DR. Meta-Analysis and Combining Information in Genetics and Genomics. Chapman and Hall/CRC; 2009.
9. Wei LJ. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Statistics in Medicine. 1992; 11:1871–79. [PubMed: 1480879]
10. Schmid M, Hothorn T. Flexible boosting of accelerated failure time models. BMC Bioinformatics. 2008; 9:269. [PubMed: 18538026]
11. Wang Z, Wang CY. Buckley-James boosting for survival analysis with high dimensional biomarker data. Statistical Applications in Genetics and Molecular Biology. 2010; 9(1):24.
12. Datta S, Le-Rademacher J, Datta S. Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. Biometrics. 2007; 63:259–71. [PubMed: 17447952]
13. Dettling M. BagBoosting for tumor classification with gene expression data. Bioinformatics. 2004; 20:3583–93. [PubMed: 15466910]
14. Dettling M, Buhlmann P. Boosting for tumor classification with gene expression data. Bioinformatics. 2003; 19:1061–69. [PubMed: 12801866]
15. Buhlmann P, Hothorn T. Boosting algorithms: regularization, prediction and model fitting (with discussion). Stat Sci. 2007; 22:477–505.
16. Buhlmann P, Yu B. Sparse boosting. Journal of Machine Learning Research. 2006; 7:1001–24.

17. Rhodes D, Chinnaiyan AM. Bioinformatics strategies for translating genome-wide expression analyses into clinically useful cancer markers. Annals of the New York Academy of Sciences. 2004; 1020:32–40. [PubMed: 15208181]

18. Rhodes D, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM. Large-scale meta-analysis of cancer microarray data identified common transcriptional profiles of neoplastic transformation and progression. PNAS. 2004; 101:9309–14. [PubMed: 15184677]

19. Stute W. Consistent estimation under random censorship when covariables are available. Journal of Multivariate Analysis. 1993; 45:89–103.

20. Berk, RA. Statistical Learning from a Regression Perspective. Springer; 2008.

21. Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning. Springer; 2009.

22. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of Statistics. 2001; 29:1189–232.

23. Zhang, J.; Ramadge, PJ. Sparse boosting.. 2009 International Conference on Acoustics, Speech and Signal Processing.; 2009;

24. Ma S, Huang J, Shi M, Li Y, Shia BC. Semiparametric prognosis models in genomic studies. Briefings in Bioinformatics. 2010; 11:385–393. [PubMed: 20123942]

25. Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, Bild A, Iversen ES, Liao M, Chen CM, West M, Nevins JR, Huang AT. Gene expression predictors of breast cancer outcomes. Lancet. 2003; 361:1590–6. [PubMed: 12747878]

26. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, Rijn M van de, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. PNAS. 2001; 98:10869–74. [PubMed: 11553815]

27. Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL, Liu ET. Breast cancer classification and prognosis based on gene expression profiles from a population based study. PNAS. 2003; 100:10393–8. [PubMed: 12917485]

28. Gu F, Schumacher FR, Canzian F, Allen NE, et al. Eighteen insulin-like growth factor pathway genes, circulating levels of IGF-I and its binding protein, and risk of prostate and breast cancer. Cancer Epidemiology, Biomarkers, Prevention. 2010; 19:2877–87.

29. He C, Kraft P, Chasman DI, Buring JE, et al. A large-scale candidate gene association study of age at menarche and age at natural menopause. Human Genetics. 2010; 128:515–27. [PubMed: 20734064]

30. Turashvili G, Bouchal J, Baumforth K, Wei W, et al. Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. BMC Cancer. 2007; 7:55. [PubMed: 17389037]

31. Hennessy BT, Gonzalez-Angulo AM, Stemke-Hale K, Gilcrease MZ, et al. Characterization of a naturally occurring breast cancer subset enriched in epithelial-to-mesenchymal transition and stem cell characteristics. Cancer Research. 2009; 69:4116–24. [PubMed: 19435916]

32. Creighton CJ, Massarweh S, Huang S, Tsimelzon A, et al. Development of resistance to targeted therapies transforms the clinically associated molecular profile subtype of breast tumor xenografts. Cancer Research. 2008; 68:7493–501. [PubMed: 18794137]

33. Kreike B, van Kouwenhove M, Horlings H, Weigelt B, et al. Gene expression profiling and histopathological characterization of triple-negative/basal-like breast carcinomas. Breast Cancer Research. 2007; 9:R65. [PubMed: 17910759]

34. Kulasingam V, Diamandis EP. Proteomics analysis of conditioned media from three breast cancer cell lines: a mine for biomarkers and therapeutic targets. Mol Cell Proteomics. 2007; 6:1997–2011. [PubMed: 17656355]

35. Emery LA, Tripathi A, King C, Kavanah M, et al. Early dysregulation of cell adhesion and extracellular matrix pathways in breast cancer progression. Am J Pathol. 2009; 175:1292–302. [PubMed: 19700746]

36. Poola I, DeWitty RL, Marshalleck JJ, Bhatnagar R, et al. Identification of MMP-1 as a putative breast cancer predictive marker by global gene expression analysis. Nat Med. 2005; 11:481–3. [PubMed: 15864312]

Table I

Simulation study based on 100 replicates. Correlation structures include auto (auto-regressive), band (banded) and comp (compound symmetry). P: number of covariates identified; TP: number of true positives.

			Meta-analysis						Intensity approach						Integrative analysis					
			Enet		Boost		SBoost		Enet		Boost		SBoost		Enet		iBoost		iSBoost	
#cov	corr	ρ	P	TP	P	TP	P	TP	P	TP	P	TP	P	TP	P	TP	P	TP	P	TP
1000	auto	0.3	152	20	151	20	118	18	93	20	95	20	30	20	85	20	138	20	20	20
		0.7	127	20	210	20	83	19	77	20	80	20	28	20	37	20	137	20	21	20
	band	0.2	103	19	209	20	72	18	80	20	76	20	26	20	33	20	138	20	22	20
		0.33	136	20	191	20	90	18	72	19	91	20	29	20	56	20	142	20	22	20
	comp	0.3	123	18	193	20	90	16	79	20	122	20	36	20	102	20	159	20	40	19
		0.7	98	20	212	20	48	12	82	20	106	19	30	13	48	20	133	20	37	17
2000	auto	0.3	205	20	143	18	173	16	91	20	115	20	43	20	117	20	176	20	21	20
		0.7	224	19	196	20	144	18	78	20	96	20	36	20	104	20	168	20	21	20
	band	0.2	208	17	231	20	129	17	67	20	89	20	32	20	118	20	166	20	22	20
		0.33	199	20	178	20	150	16	84	18	94	20	37	20	120	20	173	20	22	20
	comp	0.3	190	19	185	17	150	14	68	20	149	20	43	18	112	20	199	20	41	18
		0.7	211	20	258	19	79	11	101	20	134	18	36	12	94	20	172	19	42	16

Table II

Breast cancer prognosis studies.

Reference	Platform	Gene	Sample
Huang et al. (2003)	Affymetrix	12625	71
Sorlie et al. (2001)	cDNA	8102	58
Sotiriou et al. (2003)	cDNA	7650	98
Van't Veer et al. (2002)	Oligonucleotide	24481	78

Table III

Analysis of breast cancer prognosis studies. Gene identified using iSBoost and their estimates.

Gene	D1	D2	D3	D4	Gene	D1	D2	D3	D4
IGFBP1	-0.001	-0.001	-0.004	0.000	FGF7	-0.019	0.005	-0.003	0.001
DCAF8	0.014	0.010	0.006	0.001	SF3B4	0.012	-0.003	0.008	0.001
CIB2	-0.011	-0.001	-0.001	0.002	RAD50	0.010	0.007	0.010	0.007
UQCRB	0.000	0.005	0.004	0.006	LIN37	-0.009	-0.002	-0.006	0.000
PRKAR1B	-0.005	-0.003	-0.001	0.000	RBM4	0.000	0.012	-0.006	-0.008
EDEM1	0.006	0.007	0.007	0.005	SLCO1A2	0.013	-0.004	-0.007	0.001
LFNG	-0.002	0.019	0.017	-0.010	KCNQ2	-0.003	0.000	0.002	-0.009
GSTA4	0.000	0.001	-0.002	-0.006	ARHGEF9	0.021	0.006	0.014	0.001
NUCB1	0.004	0.008	0.007	0.001	IMPA2	-0.018	-0.019	-0.006	-0.005
IRAK1	-0.004	-0.006	-0.013	-0.003	EPB42	0.007	0.003	0.004	0.001
SMARCC2	0.006	0.001	-0.007	-0.001	EXOSC10	-0.026	-0.004	-0.003	0.009
AGPAT1	0.003	0.022	0.011	0.013	WASF1	-0.015	-0.011	-0.001	0.006
Transcribed locus	-0.004	-0.006	0.002	0.010	BRP44	-0.003	-0.007	0.000	-0.005
PAR5	0.006	0.000	-0.004	-0.007	RAB2A	-0.008	-0.009	-0.010	-0.005
TYW1	0.002	0.003	0.016	0.003	GSK3B	-0.018	-0.005	-0.002	-0.003
Transcribed locus	-0.011	-0.036	-0.002	-0.014	GLG1	-0.004	-0.005	-0.019	-0.005
FGF2	0.016	0.001	0.007	0.003	CLCN7	0.011	0.002	0.019	0.004
GSN	-0.028	-0.011	-0.003	-0.002	IL16	-0.002	0.014	-0.008	0.002
KIAA1024	0.008	0.008	0.009	0.005	CSNK2A2	-0.005	-0.008	-0.011	-0.008
MAP2K1	-0.002	-0.003	0.006	0.004	CHIT1	0.003	-0.007	0.022	-0.004
MST1	0.001	0.002	0.008	0.018	VASP	-0.030	-0.006	-0.011	-0.003
MGP	-0.008	-0.005	-0.004	-0.003	ABCD4	-0.007	-0.001	-0.001	-0.002

Table IV

Analysis of breast cancer prognosis studies. Gene: number of genes identified using different approaches; Overlap: number of overlapped genes with those identified with iSBoost; Corr: correlation coefficient between predictive risk scores and those with iSBoost.

		Gene	Overlap	Corr	Logrank
Meta-analysis	Cox-SBoost	53	7	0.209	3.443
	AFT-Enet	70	12	0.354	3.012
	AFT-Boost	182	35	0.409	1.240
	AFT-SBoost	69	17	0.387	4.578
Intensity approach	Cox-SBoost	22	3	0.274	2.887
	AFT-Enet	28	9	0.412	2.773
	AFT-Boost	127	16	0.366	1.717
	AFT-SBoost	20	4	0.447	2.570
Integrative analysis	Cox-iSBoost	51	13	0.404	5.125
	AFT-Enet	38	11	0.523	3.306
	AFT-iBoost	117	31	0.679	2.290
	AFT-iSBoost	44	--	--	5.611