Identification of Breast Cancer Prognosis Markers Using Integrative Sparse Boosting

S. Ma1; J. Huang2; Y. Xie3; N. Yi4
1School of Public Health, Yale University, New Haven, Connecticut, USA;
2Department of Statistics and Actuarial Science, University of Iowa, Iowa City, Iowa, USA;
3Department of Clinical Sciences, UT Southwestern Medical Center, Dallas, Texas, USA;
4Department of Biostatistics, Section on Statistical Genetics, University of Alabama, Birmingham, Alabama, USA

Keywords
Breast cancer prognosis, gene expression, integrative analysis, sparse boosting

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.

Correspondence to:
Shuangge Ma
School of Public Health
Yale University
60 College ST
New Haven CT 06520
USA
E-mail: Shuangge.ma@yale.edu

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). 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 conduct pooled analysis and increase sample size [5–7].
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 propor-tional hazard assumption has been violated. This model has been adopted in [6,10–12] for modeling prognosis data with gene expression measurements. We adopt a sparse boosting approach for marker selection. This approach 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 generally 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 [5–7].

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 T1, . . . , TM as the logarithm of failure times and X1, . . . , XM as length-d covariates. In study m (= 1, . . . , M) assume the AFT model

\[ T^m = \alpha^m + \beta^m X^m + \epsilon^m. \]

Here \( \alpha^m \) is the unknown intercept, \( \beta^m \) is the regression coefficient, \( \beta^m \) is the transpose of \( \beta^m \), and \( \epsilon^m \) is the random error with an unknown distribution. Denote \( C_1, \ldots, C_M \) as the logarithm of random censoring times. Under right censoring, one observation consists of \( (Y^m, \delta^m, X^m) \), where \( Y^m = \min(T^m, C^m) \) and \( \delta^m = I(T^m \leq C^m) \).

Denote \( \beta = (\beta^1, \ldots, \beta^M) \) as the \( d \times M \) matrix of regression coefficient. Recent studies [5–7] suggest that \( \beta \) should have the following characteristics. First, \( \beta \)’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, \( \beta^1, \ldots, \beta^M \) have the same sparsity structure. That is, elements of \( \beta \) 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 \( m = (1, \ldots, M) \), assume \( n^m \) iid observations \( \{(Y^m_i, \delta^m_i, X^m_i)\}, i = 1, \ldots, n^m \). Denote \( Y^m_1 \leq \ldots \leq Y^m_{n^m} \) as the order statistics of \( Y^m_1, \ldots, Y^m_{n^m} \) as the associated censoring indicators and \( X^m_1, \ldots, X^m_{n^m} \) 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

\[ \hat{F}^m(y) = \sum w^n(y_{(i)} \leq y), \]

where \( w^n \)’s are the jumps in the Kaplan-Meier estimate and computed as

\[ w^n = \int_{t_{(i)}}^{t_{(i+1)}} f(t) dt, \]

where \( t_{(i)} \) is the \( i \)th order statistic from the data. The weighted least squares estimator \( \hat{\beta} \) is then estimated by solving the following minimization problem

\[ \min_{\beta} \sum_{m=1}^M \sum_{i=1}^{n^m} w^n_i (Y^m_i \leq \hat{F}^m_{(i)}) \]
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 learners. 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

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

The L2 Boosting proceeds as follows.

**Step 1:** Initialize $k = 0$, the working residual $U^{(k)} = (u_1, \ldots, u_n) = (Y_{(1)}, \ldots, Y_{(n)})$ and $\hat{f}^{(0)}(X_i) = 0$. $\hat{f}^{(0)}(X_i)$ is an estimate of $E(Y_i | X_i = X_i)$.  

**Step 2:** Fit and update. $k = k + 1$. Compute equation (Fig. 1). Here $X_{(i)^*}$ is the $s$-th component of $X_i$.

\[ \hat{f}^{(k)}(X_i) = \arg \min_f \sum_{i=1}^{n^m} u_i \cdot v \cdot \hat{Y}_{(i)^*} X_i. \]

For personal or educational use only. No other uses without permission. All rights reserved.
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 diagnosis 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 study-specific and easily extendable than the MDL in [5, 7] and discussions in Section 2. We denote

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

Fig. 2

$$\hat{s} = \arg\min_{\lambda_1 \leq \lambda \leq \lambda_5} \sum_{m=1}^{M} \sum_{i=1}^{n_m} \left( u_i^m - y_i^m \cdot x_i^m \right)^2$$

Step 1: Initialize $k = 0$. For $m = 1 \ldots , M$ initialize the working residuals $\{\hat{r}_i^m(0) = (y_i^m, \ldots, y_i^m) = (y_i^m, \ldots, y_i^m) \}$ and $\hat{f}_m^0(0) = 0$. Here $\hat{f}_m^0(X^m)$ is an estimate of $E(Y^m|X^m)$. As only linear weak learners are used, write $\hat{f}_m^0(X^m) = \hat{\beta}_m^0 / \hat{\beta}_m^0$.

Step 2: Fit and update. $k = k + 1$. Compute as demonstrated in Figure 2. Here $\hat{x}_j^m$ is the $s$-th component of $X_j^m$.

Compute

$$\hat{y}_1^1, \ldots, \hat{y}_s^M = \arg\min_{\lambda_1 \leq \lambda \leq \lambda_5} \sum_{m=1}^{M} \sum_{i=1}^{n_m} \left( u_i^m - \hat{y}_i^m \cdot x_i^m \right)^2.$$  

Update $\hat{y}_m^k(1) = \hat{y}_m^k(k-1) + \nu \cdot \hat{y}_m^k \cdot x_i^m$, where $\nu$ is the step size as in L2 boosting;

Update $U^k_\nu = U^k_{\nu - 1} - \nu \cdot \hat{y}_m^k \cdot X_i^m$.

Step 3: Repeat Step 2 for $K$ iterations.

Step 4: At iteration $k$, compute as demonstrated in Figure 3. The optimal number of iterations is computed as $k = \arg\min_{\lambda_1 \leq \lambda \leq \lambda_5} S^k$.

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 $\gamma_m$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 (Fig. 4). 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 iS-Boost, 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.

Fig. 3

$$S^k = \sum_{m=1}^{M} \sum_{i=1}^{n_m} \left( \log(\nu) \sum_{j=1}^{d} I(\| \beta_j^k \| \neq 0) \right)$$

Fig. 4

$$\hat{s} = \arg\min_{\lambda_1 \leq \lambda \leq \lambda_5} \sum_{m=1}^{M} \sum_{i=1}^{n_m} \left( u_i^m - \hat{y}_i^m \cdot x_i^m \right)^2 + \log(\nu) \sum_{j=1}^{d} I(\| \beta_j^k \| \neq 0)$$
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 $M_1$ studies. We modify Step 2 of iBoost as can be seen in Figure 5. 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 $\max(0, 1 - |j - k|/p)$; and (iii) compound symmetry, where expressions of genes $j$ and $k$ have correlation coefficient $\rho$ when $j \neq k$. Under each correlation scenario, we consider two different values of $\rho$.

Within each of the first four clusters, there are five 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 $\text{Unif}[-1, -0.5] \cup \text{Unif}[0.5, 1]$. 20% prognosis-associated genes 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 1). 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 1). 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 1, 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 misspecification 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 1. 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 2 and refer to the original publications [4, 25–27] 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 3). 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 (DNA-PKcs). It is one of the 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 LFGN 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 GSF2 and GSF7 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

### Table 3  Analysis of breast cancer prognosis studies. Gene identified using iSBoost and their estimates.

<table>
<thead>
<tr>
<th>Gene</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
<th>Gene</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGFBP1</td>
<td>−0.001</td>
<td>−0.001</td>
<td>−0.004</td>
<td>0.000</td>
<td>FGF7</td>
<td>−0.019</td>
<td>0.005</td>
<td>−0.003</td>
<td>0.001</td>
</tr>
<tr>
<td>DCAF8</td>
<td>0.014</td>
<td>0.010</td>
<td>0.006</td>
<td>0.001</td>
<td>SF3B4</td>
<td>0.012</td>
<td>−0.003</td>
<td>0.008</td>
<td>0.001</td>
</tr>
<tr>
<td>CIB2</td>
<td>−0.011</td>
<td>−0.001</td>
<td>−0.001</td>
<td>0.002</td>
<td>RAD50</td>
<td>0.010</td>
<td>0.007</td>
<td>0.010</td>
<td>0.007</td>
</tr>
<tr>
<td>UQCRB</td>
<td>0.000</td>
<td>0.005</td>
<td>0.004</td>
<td>0.006</td>
<td>LIN37</td>
<td>−0.009</td>
<td>−0.002</td>
<td>−0.006</td>
<td>0.000</td>
</tr>
<tr>
<td>PRKAR1B</td>
<td>−0.005</td>
<td>−0.003</td>
<td>−0.001</td>
<td>0.000</td>
<td>RBM4</td>
<td>0.000</td>
<td>0.012</td>
<td>−0.006</td>
<td>−0.008</td>
</tr>
<tr>
<td>EDEM1</td>
<td>0.006</td>
<td>0.007</td>
<td>0.007</td>
<td>0.005</td>
<td>SLCO1A2</td>
<td>0.013</td>
<td>−0.004</td>
<td>−0.007</td>
<td>0.001</td>
</tr>
<tr>
<td>LFNG</td>
<td>−0.002</td>
<td>0.019</td>
<td>0.017</td>
<td>−0.010</td>
<td>KCNQ2</td>
<td>−0.003</td>
<td>0.000</td>
<td>0.002</td>
<td>−0.009</td>
</tr>
<tr>
<td>GSTA4</td>
<td>0.000</td>
<td>0.001</td>
<td>−0.002</td>
<td>−0.006</td>
<td>ARHGEF9</td>
<td>0.021</td>
<td>0.006</td>
<td>0.014</td>
<td>0.001</td>
</tr>
<tr>
<td>NUCB1</td>
<td>0.004</td>
<td>0.008</td>
<td>0.007</td>
<td>0.001</td>
<td>IMPA2</td>
<td>−0.018</td>
<td>−0.019</td>
<td>−0.006</td>
<td>−0.005</td>
</tr>
<tr>
<td>IRAK1</td>
<td>−0.004</td>
<td>−0.006</td>
<td>−0.013</td>
<td>−0.003</td>
<td>EPB42</td>
<td>0.007</td>
<td>0.003</td>
<td>0.004</td>
<td>0.001</td>
</tr>
<tr>
<td>SMARCC2</td>
<td>0.006</td>
<td>0.001</td>
<td>−0.007</td>
<td>−0.001</td>
<td>EXOSC10</td>
<td>−0.026</td>
<td>−0.004</td>
<td>−0.003</td>
<td>0.009</td>
</tr>
<tr>
<td>AGPAT1</td>
<td>0.003</td>
<td>0.022</td>
<td>0.011</td>
<td>0.013</td>
<td>WASF1</td>
<td>−0.015</td>
<td>−0.011</td>
<td>−0.001</td>
<td>0.006</td>
</tr>
<tr>
<td>Transcribed locus</td>
<td>−0.004</td>
<td>−0.006</td>
<td>0.002</td>
<td>0.010</td>
<td>BRP44</td>
<td>−0.033</td>
<td>−0.007</td>
<td>0.000</td>
<td>−0.005</td>
</tr>
<tr>
<td>PAR5</td>
<td>0.006</td>
<td>0.000</td>
<td>−0.004</td>
<td>−0.007</td>
<td>RAB2A</td>
<td>−0.008</td>
<td>−0.009</td>
<td>−0.010</td>
<td>−0.005</td>
</tr>
<tr>
<td>TYW1</td>
<td>0.002</td>
<td>0.003</td>
<td>0.016</td>
<td>0.003</td>
<td>GSK3B</td>
<td>−0.018</td>
<td>−0.005</td>
<td>−0.002</td>
<td>−0.003</td>
</tr>
<tr>
<td>Transcribed locus</td>
<td>−0.011</td>
<td>−0.036</td>
<td>−0.002</td>
<td>−0.014</td>
<td>GLG1</td>
<td>−0.004</td>
<td>−0.005</td>
<td>−0.019</td>
<td>−0.005</td>
</tr>
<tr>
<td>FGF2</td>
<td>0.016</td>
<td>0.001</td>
<td>0.007</td>
<td>0.003</td>
<td>CLCN7</td>
<td>0.011</td>
<td>0.002</td>
<td>0.019</td>
<td>0.004</td>
</tr>
<tr>
<td>GSN</td>
<td>−0.028</td>
<td>−0.011</td>
<td>−0.003</td>
<td>−0.002</td>
<td>IL16</td>
<td>−0.002</td>
<td>0.014</td>
<td>−0.008</td>
<td>0.002</td>
</tr>
<tr>
<td>KIAA1024</td>
<td>0.008</td>
<td>0.008</td>
<td>0.009</td>
<td>0.005</td>
<td>CSNK2A2</td>
<td>−0.005</td>
<td>−0.008</td>
<td>−0.011</td>
<td>−0.008</td>
</tr>
<tr>
<td>MAP2K1</td>
<td>−0.002</td>
<td>−0.003</td>
<td>0.006</td>
<td>0.004</td>
<td>CHIT1</td>
<td>0.003</td>
<td>−0.007</td>
<td>0.022</td>
<td>−0.004</td>
</tr>
<tr>
<td>MST1</td>
<td>0.001</td>
<td>0.002</td>
<td>0.008</td>
<td>0.018</td>
<td>VASP</td>
<td>−0.030</td>
<td>−0.006</td>
<td>−0.011</td>
<td>−0.003</td>
</tr>
<tr>
<td>MGP</td>
<td>−0.008</td>
<td>−0.005</td>
<td>−0.004</td>
<td>−0.003</td>
<td>ABCD4</td>
<td>−0.007</td>
<td>−0.001</td>
<td>−0.001</td>
<td>−0.002</td>
</tr>
</tbody>
</table>
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 spliceosomal 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 ARH-GEF9 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 non-transforming monomeric GTP-binding proteins of the Ras superfamily that contain four highly conserved regions involved in GTP binding and hydrolysis. Rab proteins 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 integrin-extracellular 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 4 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 $\hat{\beta}'X$, 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 4. 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. Inte-
grative 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 estimator 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 of studies.

Acknowledgements

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